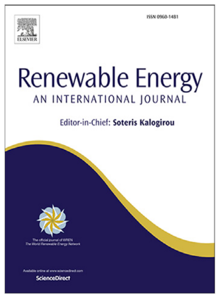
Journal Pre-proof



Advanced multimodal fusion method for very short-term solar irradiance forecasting using sky images and meteorological data: A gate and transformer mechanism approach

Liwenbo Zhang, Robin Wilson, Mark Sumner, Yupeng Wu

PII: S0960-1481(23)00858-3
DOI: https://doi.org/10.1016/j.renene.2023.118952
Reference: RENE 118952

To appear in: *Renewable Energy*

Received date : 22 February 2023
Revised date : 9 June 2023
Accepted date : 22 June 2023

Please cite this article as: L. Zhang, R. Wilson, M. Sumner et al., Advanced multimodal fusion method for very short-term solar irradiance forecasting using sky images and meteorological data: A gate and transformer mechanism approach, *Renewable Energy* (2023), doi: https://doi.org/10.1016/j.renene.2023.118952.

Advanced Multimodal Fusion Method for Very Short-Term Solar Irradiance Forecasting using Sky Images and Meteorological Data: A Gate and Transformer Mechanism Approach

Liwenbo Zhang*, Robin Wilson, Mark Sumner, Yupeng Wu*

Faculty of Engineering, University of Nottingham, Nottingham, NG7 2RD, the United Kingdom

Abstract

Cloud dynamics are the main factor influencing the intermittent variability of short-term solar irradiance, therefore affect the solar farm output. Sky images have been widely used for short-term solar irradiance prediction with encouraging results due to the spatial information they contain. At present, there is little discussion on the most promising deep learning methods to integrate images with quantitative measures of solar irradiation. To address this gap, we optimise the current mainstream framework using gate architecture and propose a new transformer-based framework in an attempt to achieve better prediction results. It was found that compared to the classical CNN model based on late feature-level fusion, the transformer framework model based on early feature-level prediction improves the balanced accuracy of Ramp Event by 9.43% and 3.91% on the 2-minute and 6-minute scales, respectively. However, based on the results, it can be concluded that for the single picture-digital bimodal model, the spatial information validity of a single picture is difficult to achieve beyond 10 minutes. This work has the potential to contribute to the interpretability and iterability of deep learning models based on sky images.

Keywords: Solar energy, Forecasting, Computer vision, Deep learning, Vision Transformer, Sky images

1. Introduction

As solar power generation grows, its inherent variability presents the grid with issues related to reserve costs, dispatchability and ancillary generation, and grid reliability in general [1]. Accurate forecasting of solar irradiance at different time scales is a prerequisite for effective utilisation of solar energy and a critical step in the grid integration and management of solar farms [2, 3]. Reliable solar forecasting tools improve the economics of PV power generation and reduce the negative impact of PV uncertainty on grid stability [4].

Changes in cloud cover are the leading cause of rapid changes in solar irradiance. Since the prediction models based on statistical numerical regression used in very short-term forecast models does not include information on fast moving clouds, alternative or additional data inputs that account for these rapidly changing meteorological phenomena are required if accuracy at this time scale is to be improved.

Ground-based sky imagery represents one such exogenous data source and plays a crucial role in solar energy forecasting due to its ability to provide information on cloud distribution and motion.

*Corresponding author

Email addresses: `liwenbo.zhang@nottingham.ac.uk` (Liwenbo Zhang), `yupeng.wu@nottingham.ac.uk` (Yupeng Wu)

Solar irradiation models informed by cloud motion data offer the potential to deliver accurate forecasts of very short-term solar irradiation, and thus provide valuable supporting information for grid management and informing the market around power supply and demand [5].

Currently, sky images taken by fish-eye cameras contain rich spatio-temporal features and thus are widely accepted by the academic community as exogenous data for intra-hourly level sky modelling [6, 7, 8]. The main methods for predicting solar irradiance based on sky images can be divided into two categories. The first is a sky modelling approach based on classical image analysis. To determine spatial features, methods such as red-blue ratio (RBR) or red-blue difference (RBD) [9, 10, 11], 3D cross correlation [12], or image feature correlation [13] are used to identify cloud pixels in the sky image. To determine temporal features, the most common approach is to use the cross correlation method [10], which calculates the cloud motion vector by comparing two consecutive cloud maps. In addition to cross correlation, other methods include optical flow [6, 14] and ray tracing [15]. The optical flow method determines the velocity of feature pixels based on the intensity of two consecutive images and uses this to calculate the position of the cloud in relation to the ground projection of the cloud at the approaching time point. The ray-tracing approach uses multiple images of the sky taken simultaneously from different positions, combined with ground shadow maps to model clouds in 3D. The advantage of this approach is that the 3D model solves the problem of individual site images not being able to determine the height of the cloud base [12], while both the cross correlation and optical flow methods require additional instrumentation to measure the height of the cloud base to determine the correct ground projection of the cloud [16]. Image-based forecasts determine the impact on solar irradiance estimates by combining the estimates of cloud position with estimates of cloud transmittance, and general methods used to determine the latter include fixed transmittance [6, 10], cloud density-based transmittance [17, 7] and cloud height-based transmittance approaches [18]. However, these modelling approaches to image analysis are still limited by the complex physical properties of clouds. For example, cloud motion is assumed to involve shifting only and does not account for cloud generation and dissipation. Additionally, cloud transmittance depends on the transparency of the cloud, but it is not currently feasible to measure the transmittance of all cloud types directly. Therefore, this approach remains of limited use in improving the accuracy of future irradiance forecasts [19]. At present this approach is based on decision-level fusion, i.e. solar irradiation forecasts and RAMP forecasts are made independently of each other and only influence each other when combined in the final stage as shown in Figure 1 (a).

The second approach uses deep learning methods [20, 21, 22, 23, 24, 25, 26, 27, 28, 29]. This usually employs a combination of convolutional neuron networks (CNN) [30] and recurrent neural networks [31] (RNN) based methods to predict solar irradiance information for future time periods. The widely used CNN-based computer vision models, such as ResNet [32] and VGGNet [33], can extract feature information from a dataset containing many sky images using deep convolutional neuron networks to obtain spatial dimensional perception capability. After extracting the spatial information of the images, various methods can be used to obtain time-series based information. These include, pre-processing by stacking a time series of images [21], convolution processes using 3D-CNN with an extra temporal dimension [23], convolution-based long and short-term memory (LSTM) network [20], convolution followed by feature-based LSTM networks [22, 28], directly using regression algorithms for continuous results [21, 23], or combine feature engineering techniques with LSTM techniques [26]. By combining the architecture of two networks and fitting them using a large amount of data, a network model with both spatial and temporal feature perception can be obtained. This stitching model can be used to map the relationship between specific features in continuous input image data and forecast targets. This type of model has been applied to short-term forecast intervals for different forecast resolutions. In contrast to models based on image analysis, current deep learning models can be mainly categorised as late feature fusion models, where the image and numerical values respectively abstract features as a high-dimensional vector in their respective models and concatenate the two vectors at the end of their respective operations, as shown in Figure 1(b). The tandem high-dimensional vector can be thought of as a joint feature extract based on the two modalities, and the final prediction is based on the extraction of available information from that vector.

While deep learning networks have been shown to deliver predictions with greater accuracy than those based on feature engineering in the field of ground-based sky picture solar prediction, due to its black box nature, researchers cannot assess the relationships between variables that affect performance. For example, using sky images as exogenous data to aid solar prediction has been shown to improve model performance at time scales ranging from 2 minutes ahead [34] to 1 hour ahead [35]. It is obvious that the images play a different role at these two different time scales but the features it identifies are not understood.

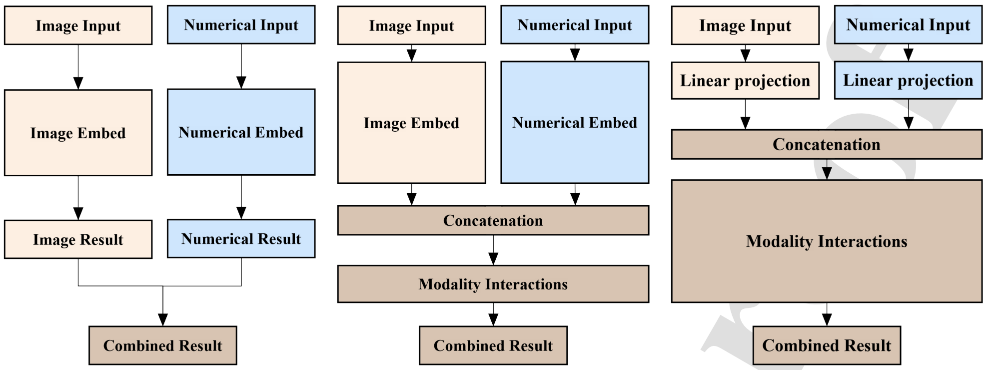
The research carried out by Paletta et al [20]. highlighted that prevailing image- and numerical-based forecasting models show a propensity towards reactive, rather than anticipatory, predictions. This predilection represents a significant challenge in current prediction models. More specifically, these models did not anticipate the timing of imminent solar ramp events from sky images as anticipated by the researchers.

We argue in this paper that solar irradiance forecasting using ground-based images from which numerical features are extracted that describe the solar field can be categorised as a general multi-modal learning domain, rather than a purely computer vision domain. That is, the model is forecasting through use of a deep learning network based on two or more heterogeneous data sources with complementary information.

As shown in Figure 1, for the broad field of image-informed multi-modal learning, besides the two aforementioned architectures, i.e. decision-level and late feature-level fusion of image information, the fusion methods also include: data-level fusion (not shown in Figure) and early feature-level fusion. Of these, early feature-level fusion and late feature-level fusion both extract feature fusion within the model, with early fusion focusing on modal interactions and late fusion focusing on feature extraction [36]. In deep learning models used for solar forecasting, two architectures are currently applied, namely late feature-level fusion [20, 37, 22, 38] and decision-level fusion [39, 21]. In the work of Paletta et al. [20], the use of numerical data as additional inputs fused with a computer vision model improved the 2-minute forecast skill (FS), which rose from -3.4% to 12.9% and the 10-minute FS, which rose from 18.8% to 23.9%.

However, the literature suggests that the interest of researchers is currently focused on the image feature side to improve overall forecasting power through a more robust image network. This approach neglects both the role that the numerical component plays in the model and whether it interacts effectively with the image component. For example, the numerical regression-based fully connected Multi-Layer neural network module (MLP) has been added to forecasting models by default due to the use of PV logarithms as an additional numerical input in the work of Sun et. al. [37] and significantly improved the performance of the model.

Another potential area of research responds to the fact that the image-numerical bimodal model currently in use is not modal interaction friendly. The prevailing image feature framework is the convolutional neuron network (CNN), where specific features of an image are extracted by sliding convolutional modules through the image and gradually constructing a high-dimensional vector



(a) Decision-level fusion (b) Late feature-level fusion (c) Early feature-level fusion

Figure 1: Schematic diagram of the model architecture for the different stages of fusion.

representation of the image by multi-layer superposition. This architecture means that it is not possible to extract features present in the 3D image and use these directly with complementary data held in a 1D array. Therefore, if data features of different dimensions are extracted simultaneously by convolutional computation, i.e. early feature-level fusion, this must be done by projecting the 1D data to a higher dimension and concatenating it with another, a process that may lead to distortion of the low-dimensional data. Venugopal et al. [39] compared CNN networks against PV output-based regression predictions with different fusion methods. Their results showed that late feature-level fusion and decision-level fusion achieved better prediction performance, but data-level fusion and early feature-level fusion failed to effectively interact information across modalities to achieve results beyond the baseline.

Multimodal learning, adopts a unique feature extraction approach, where its transformer architecture enables data from different modalities to be fed into the encoder in parallel to achieve early feature-level fusion, as shown in Figure 1(c). It can effectively address the challenges of inherent data misalignment arising from the variable sampling rate and establishing cross-modal element correlations of each modality's sequence [36]. Thus, the transformer-based model is widely used in the multimodal learning fields of image-language interpretation [40], image-sentiment recognition [41], the joint expression of video-audio-text [42, 43], etc. These applications share commonality with the mixed-mode data feeds available for irradiation forecasting. The original contributions of this study are:

1. To present two new approaches for picture-numerical bimodal model interaction. Namely, an improvement of the later feature-level fusion method by means of a gate architecture and a new early feature-level fusion method based on the Transformer architecture.
2. To assess the performance of the model 2, 6, and 10 minute forecasting horizons by scoring its quantitative statistical performance using the Smart Persistence Model (SPM)-based FS metric and the qualitative performance of the model using the Ramp Events (RE)-based Balanced Precision (BP) metric.
3. To show contradictions in the quantitative and qualitative performance of late feature-level fusion models in terms of single image and numerical fusion. In particular, the widely used CNN model based on late feature-level fusion obtained higher FS while resulting in lower BP.

From which we speculate on, and attempt to demonstrate, a link between this and the poor sensitivity of its architecture to images.

4. To demonstrate that for the end-to-end single picture-numerical bimodal model, the main variability of the model, both architecturally and algorithmically, was most pronounced for the 2 minutes ahead forecast. This variability fades with longer forecasting horizons. At 10-minutes ahead forecast, the validity of the image information is extremely low and all models have degenerated into a mean reversion model that relies primarily on irradiance and clear sky irradiance.

The remainder of the paper is structured as follows: Section 2 presents the overall experimental approach, including Data pre-processing, model architecture, and evaluation methods; Section 3 presents results that show quantitative and qualitative evaluation results for all models and discusses the results; and Section 4 presents our conclusions and recommendations for future work.

2. Methodology

Figure 2 illustrates the methodology adopted in this study. The approach to building a deep learning solar forecasting model based on image-numerical fusion comprised three stages. The first was a data pre-processing stage, which aligned, filtered, sampled, and grouped the raw data into a format suitable for training a deep learning model. The second was a training stage, where the training dataset was fed into the model and the weights within the model were fixed by back propagation. Following this, the model was evaluated on a validation set to assess the performance trained in training dataset. Through continuous iteration, the model that achieves the optimal result on the validation set, i.e. the model with the least loss, is saved to end the training process. The final stage involved use of a test dataset to obtain a forecast for comparison with ground truth data, in order to quantify the final performance of the different models studied in this paper.

Clear sky index (CSI), i.e. the solar irradiance as a percentage of the clear sky irradiance, was chosen as the target for forecasts rather than the GHI, reflecting consensus within the solar forecasting community around its ability to improve the accuracy of solar irradiance forecasts made using numerical regression algorithms [44], including those that involve image-numerical multi-modality approaches. Additionally, use of CSI as a forecast target has a beneficial inductive bias compared to the direct forecast of irradiance, i.e., the model assumes a priori knowledge of the clear sky background. Forecasts generate an atmospheric transmission rate (or attenuation rate) based on the clear sky background, which is also consistent with traditional image analysis methods when harnessed for use in irradiance forecasting.

The reach of the forecast target was informed by the approach of Kong et al. [45]. A forecast resolution of 4 minutes and forecast span of 10 minutes were selected, and the input data set was used in three different models to generate independent solar irradiance forecasts, each over 2-, 6-, and 10-minute time horizons. Results were compared to quantify the relative forecasting performance of the models under 3 different forecast horizons.

As shown in Figure 2, Section 2.1 the data pre-processing explains the process of going from raw data to trainable data. Section 2.2 describes the process of the five main supervised image-numerical multimodality models in this paper along with other standard model architectures. Section 2.3 evaluation matrix introduces the two main criteria for model prediction performance evaluation.

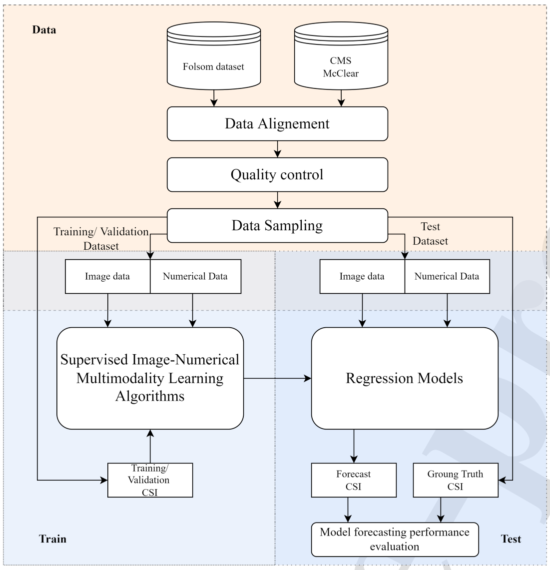


Figure 2: Overview of the solar forecasting framework.

2.1. Data pre-processing

Data for the experiments were obtained from the Folsom, California [46] public database, supplemented by clear sky irradiance values from the McClear [47] clear sky irradiance model. Output from the latter was generated using the timestamps of corresponding Folsom data points.

Inputs to each of the models comprised a set of time synchronised data that included clear sky irradiance (GHI, DNI, and DHI), measured irradiance (GHI, DNI, DHI), weather data (dry bulb air temperature, humidity, relative air pressure, wind speed, and wind direction) measured at ground base stations, and solar geometry (solar zenith and solar azimuth angles).

Data alignment and Quality control The initial stage of data pre-processing involved image compression, alignment of images to numerical data, quality control, and data normalisation. The Folsom dataset provides raw image data (1536 pixels $\times$ 1536 pixels), solar irradiance data, and weather data. These data first went through a process of temporal alignment using timestamps and the corresponding clear sky irradiance was then sourced from the McClear clear sky model Following this, quality control filters were applied to screen each piece of data.

For numerical data, a quality control strategy following Yang‘s [48] work was used to reject data outliers, with decisions being made on the basis of identifying extremely-rare limits [49], a diffuse ratio test [49], and other filters [5].

Images were down-sampled to 128 pixels $\times$ 128 pixels, a resolution considered to be the smallest resolution that can be maintained for sky information, using the bilinear method to match the input format of the ANN. In addition, the image dataset showed occasional time shifts possibility due to cumulative errors resulting from continuous shooting. Data points that showed significant offsets (more than 15 seconds from the timestamps) were removed. Finally, to balance the weights of all inputs, all RGB channels and numerical data of the images were normalized to the interval

$[0, 1]$, except for the solar altitude angle which was normalized to $[-1, 1]$ after a trigonometric transformation.

Segmentation and resampling of dataset The Folsom dataset provides numerical and image data for three years from 2014-2016. In this study, the 2014 data was used as the training set, the 2015 data as the validation set, and the 2016 data as the test set. Following the data alignment and quality control stage these contained 195k, 233k, 228k data points respectively. Within these datasets, the sample size for sunny periods was much larger than that for non-sunny days, the former accounting for approximately 60% of the entire dataset. As may be inferred from the cumulative distribution of CSI on left side of Figure 3, the dataset is unbalanced, with a clustering of CSI values between approximately 0.9 and 1.05. Recent research [50] suggests that unbalanced datasets can generate models biased towards non-critical conditions – in the case of the Folsom dataset, the sunny periods. To guard against potential bias, simple algorithm was used to filter out consecutive data points within sunny period. Specifically, a data point was excluded if the preceding 5 and following 10 points where 'sunny' as defined by the limits of the data clustering, i.e., a CSI greater than 0.9 and less than 1.05. The right side of Figure 3 shows the data distribution after resampling, suggesting it is better balanced. The remaining datasets contains 86K, 100K and 94K data points respectively.

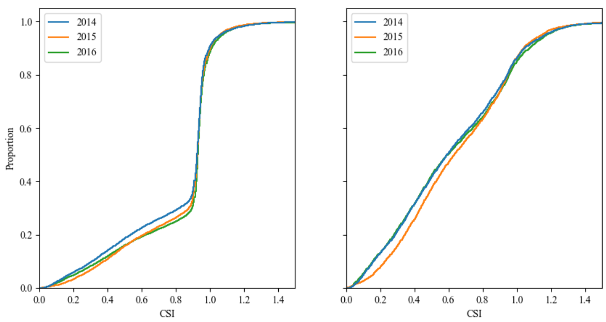


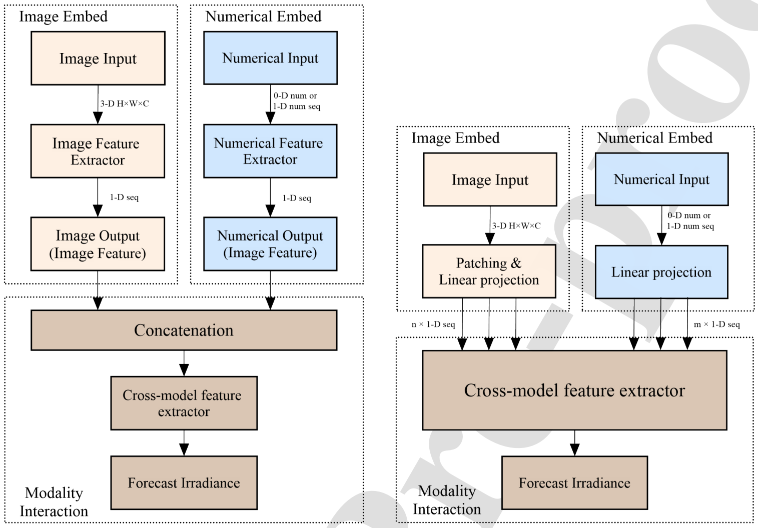
Figure 3: Data before (left) and after (right) resampling CSI distribution

Due to computer memory and training time constraints, it was verified that a quarter of the data was randomly sampled(in Appendix A, Figure A.14). The final training, validation and test datasets used for analysis contained approximately 21k, 25k and 23k data points respectively. The detailed monthly distribution of the final data is shown in Appendix A, Figure A.15

Due to the computer memory and training time constraints, only a quarter of the training data were used, these being randomly sampled from the training dataset. The final training, validation, and test datasets used in the analysis contained approximately 21k, 25k, and 23k data points respectively. A specific data sampling test and validation of sample rate are presented in Figure B.6

2.2. Development of deep-learning based irradiance forecast model

We propose or utilise models and architectural methods aimed at enhancing or optimising the interaction or fusion between patterns, balancing the predictive role of image patterns in multimodal models. In this section, we introduce the mainstream architecture of the current image-to-text multimodal prediction model, namely the late fusion architecture at the feature level, and propose a balancing mechanism using gate mechanisms to dynamically balance the outputs between modalities. Next, we present our novel model, which is based on an attention-based Transformer architecture, enabling early fusion at the feature level.



(a) (b)

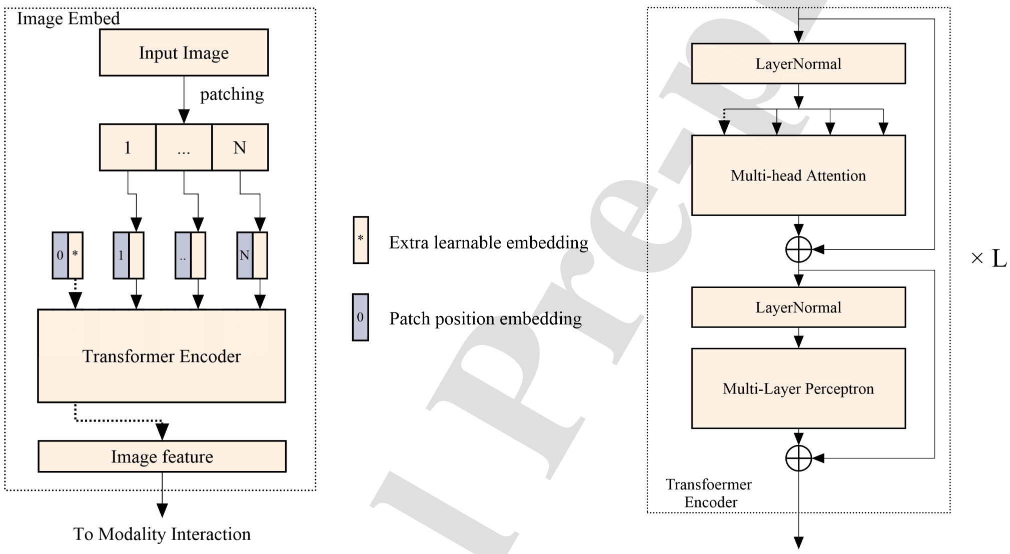
Figure 4: Schematic diagram of the numerical-image bimodality model. (a) Late Feature-level fusion [37]. (b) Early Feature-level fusion.

2.2.1. Bimodal model based on late feature-level fusion

Currently, mainstream deep learning-based image-numerical bimodal models are based on late-stage feature-level fusion architectures [23, 22, 20, 37, 45], as illustrated in Figure 4(a). The architecture consists of three main components: an image embedding process that extracts the input image features as high-dimensional vectors; a numerical embedding process that extracts the input numerical features as high-dimensional vectors; and a modal interaction module that extracts the joint features from the two vectors after a process of concatenation, which ultimately derives the forecasting results.

CNN - Current Image embedding Among the sky image-based PV forecast models, CNN and other variants based on convolutional computation, are currently the dominant image feature extractors due to their excellent image resolution performance [45, 23, 20]. These extract features from images in a continuous convolutional scan, building a weighting system from detailed to macroscopic images by sequentially expanding the receptive field size of the model through a multi-layer repetitive architecture. In this study, the most widely accepted ResNet-18 model [32] was used as a baseline model for CNN image extractors.

ViT - Proposed Image embedding As mentioned above, methods based on Transformer architecture are emerging as a widely used backbone network for a variety of tasks, and amongst these, the Vision Transformer (ViT) has been developed to undertake image feature extraction [51]. Unlike the convolution-based scanning adopted by CNN models, ViT-based vision models build a weighted system by extracting interconnections between patches within images. As a result, such models can establish relationships between pixels at different areas within the image. This paper postulates that since the main feature of the sky image in short-term solar forecasts is primarily the relative relationship between regions occupied by cloud, clear sky and the sun, the relative importance of fine-grain texture and detail in the image is lower and ViT models, based on multiple self-attention, are able to extract the more important larger-scale features in sky images more efficiently.



(a) ViT architecture (b) Transformer Encoder

Figure 5: Schematic diagram of Vision Transformer (ViT) image embedding.

For a module that acts only as an image feature extractor, based on the work [32], the computational process can be expressed as

$$\mathbf{z}_{i0} = \left[\mathbf{x}_{\text{class}}\,; \mathbf{x}_p^1\mathbf{E}; \cdots ; \mathbf{x}_p^N\mathbf{E}\right] + \mathbf{E}_{\text{pos}} \qquad \mathbf{E} \in \mathbb{R}^{(P^2\cdot C)\times D}, \mathbf{E}_{\text{pos}} \in \mathbb{R}^{(N+1)\times D} \tag{1}$$

$$\mathbf{z}_{il}' = \text{MSA}\left(\text{LN}\left(\mathbf{z}_{il-1}\right)\right) + \mathbf{z}_{il-1}, \qquad l = 1 \ldots L \tag{2}$$

$$\mathbf{z}_{il} = \text{MLP}\left(\text{LN}\left(\mathbf{z}_{il}'\right)\right) + \mathbf{z}_{il}', \qquad l = 1 \ldots L \tag{3}$$

$$\hat{\mathbf{z}}_i = \text{LN}\left(\mathbf{z}_{i\,L}^0\right) \tag{4}$$

As shown in Figure 5(a), the image input $\mathbf{x} \in \mathbb{R}^{H\times W\times C}$ is divided into N patches of side length P and stitched into a 2D sequence $\mathbf{x}_p \in \mathbb{R}^{N\times(P^2\cdot C)}$. Following this, the pixels of each patch are projected linearly onto D dimensions via transfor embedding, a learnable latent vector $\mathbf{E} \in \mathbb{R}^{(P^2\cdot C)\times D}$. Following the process described by Devlin et al. [52], the input after reshaping is stitched with an additional learnable class token, $\mathbf{x}_{\text{class}}$, and embedded with a learnable position component $\mathbf{E}_{\text{pos}} \in \mathbb{R}^{(N+1)\times D}$, which describes the spatial relationships between patches. Eventually, the image part of the input is represented as $\mathbf{z}_{i0} \in \mathbb{R}^{(N+1)\times D}$. This input is added to a standard Transformer module, shown in Figure 5(b), i.e., a module based on a Multiheaded Self-Attentive (MSA) process [53] and a Multi-Layer Perceptron (MLP) process, iterated L times. Ultimately, the learnable class token, xclass, is extracted, and after Layer Normalisation (LN), is output as a high-dimensional vector $\hat{\mathbf{z}}_i$, representing the image feature.

ANN - Current Modality interaction embedding Currently, multilayer feedforward Artificial Neural Networks (ANN), also known as MLP, are widely used as one-dimensional vector feature extractors in models with numerical inputs [54]. ANNs are also used widely in the modal fusion phase of image-numerical bi-modal solar forecasting models [37, 23, 22, 20]. As mentioned above, when ANNs are used as a cross-modal feature extractor, as shown in Figure 6(a), the direct concatenation that takes place before feature extraction fails to make effective connections between the input parameters, and the interaction of the inter-model outputs is completely dependent on the subsequent adaptive of the network architecture to such outputs. Also, due to the heterogeneity of the different data, models based on ANNs face multiple challenges when performing mapping (converting image information into irradiance data) and fusion forecasting (combining information from two modalities to predict ramp events). These challenges include instances where information from different modalities have different predictive power and noise topology, or instances where models are unable to capture features from one of the modalities.

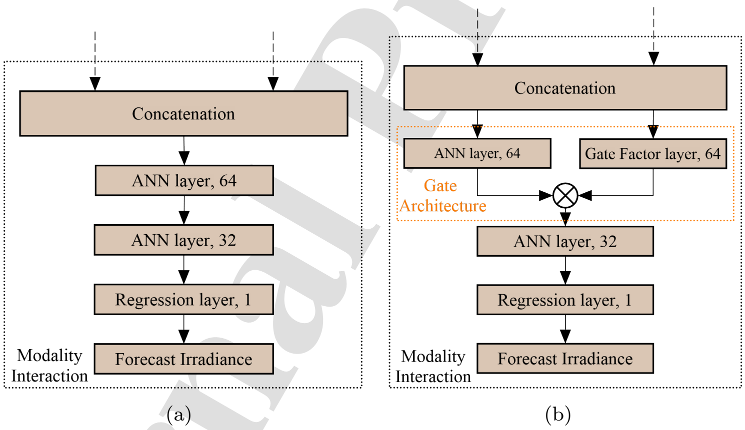


Figure 6: Schematic diagram of modality interaction in late feature-level fusion models. (a) ANN feature extractor (b) Gated-ANN feature extractor.

ANN with gate architecture - Proposed Modality interaction embedding In order to improve the attention given to target features in the both modality processed by the MLP and to suppress feature activation in irrelevant regions, this paper proposes that addition of a layer based on attention gate architecture, as shown in Figure 6(b). It is implemented by a mechanism similar to the gated recurrent unit in the LSTM [31], by controlling the weighting of the parameters through the layers. The gate architecture generates a gating coefficient for each node in ANN with the same dimensionality as the input feature and then converts this into an attention weight map multiplied by the original feature. The attention gate performs the task of focussing the model's attention on essential regions of the input data and neglecting irrelevant regions. The simplicity of this approach makes it possible to improve feature extraction without significant an increase in computing cost.

2.2.2. Transformer-based early feature-level fusion

As mentioned above, the MSA-based ViT model finds application beyond image processing. Because the MSA module inputs are a series of 1D multidimensional vectors or tensors, it is possible to input image and numerical data in parallel. As an alternative to CNNs, such backbone networks have been shown to offer outstanding capabilities in several fields dealing with multi-modality tasks, such as image and text [55], video and text [56], etc. However, there is, as yet, no such work applied to the field of solar energy forecasting. Therefore, inspired by Kim et al. [57], this paper speculates that multi-modality input short-term irradiance forecast models that combine sky images and measurement logs can also be constructed using the Transformer module as the backbone network to replace both the CNN visual layer and the MLP numerical regression computational layer to construct input data with early feature-level fusion.

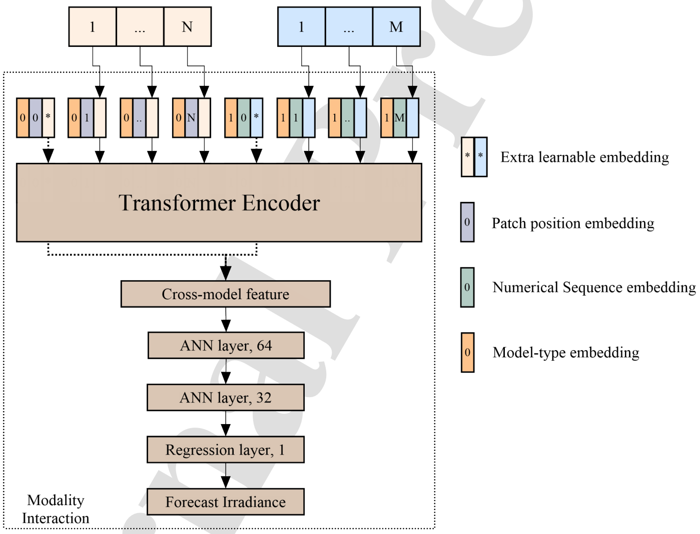


Figure 7: Schematic diagram of image/text bimodal transformer architecture.

The proposed early feature and fusion model is based on the Transformer architecture shown in Figure 7. The main inputs to the model comprise image data and numerical data. For the image data, input follows the patching process illustrated in Fig 5(a). For the numerical data, a standard unbiased MLP for numeric features is used to up dimension the numeric information to D, $\mathrm{MLP}(\mathbf{y}) \in \mathbb{R}^{1\times D}$, and provide a learnable class token. The numerical data are divided into five groups based on type: solar irradiance, clear sky solar irradiance, sun angle, ground wind conditions, and weather parameters (dry bulb air temperature, humidity and relative air pressure).

As with image processing similar to the ViT process, the image part of the inpus is represented as $\mathbf{z}_{i0}$. Meanwhile, the learnable class token for numerical data, $\mathbf{y}_{\text{class}}$, combined with learnable position embedding $\mathbf{E}_{\text{seq}} \in \mathbb{R}^{(M+1)\times D}$ is used to describe the position relationships within the data sequence. The numerical part of the input is represented as $\mathbf{z}_{n0} \in \mathbb{R}^{(M+1)\times D}$. Finally, $\mathbf{z}_{i0}$ and $\mathbf{z}_{n0}$ are embedded separately in the model type embedding process as $\mathbf{z}_i^{\text{type}}$ and $\mathbf{z}_n^{\text{type}}$, before the process of concatenation to generate $\mathbf{z}_0 \in \mathbb{R}^{(M+N+2)\times D}$. The vector $\mathbf{z}_0$ is iteratively updated through L-depth transformer layers up until the final sequence $\mathbf{z}_l$. The final $\hat{\mathbf{z}}$ representing the forecast vector is generated by a linear projection of the two learnable vectors $\mathbf{z}_{i\,L}^0$ and $\mathbf{z}_{n\,L}^0$ in series with hyperbolic tangent activation.

The overall data processing can be described as

$$\mathbf{z}_{i0} = \left[\mathbf{x}_{\text{class}}; \mathbf{x}_p^1\mathbf{E}; \cdots; \mathbf{x}_p^N\mathbf{E}\right] + \mathbf{E}_{\text{pos}} \qquad \mathbf{E} \in \mathbb{R}^{(P^2\cdot C)\times D}, \mathbf{E}_{pos} \in \mathbb{R}^{(N+1)\times D} \tag{5}$$

$$\mathbf{z}_{n0} = \left[\mathbf{y}_{\text{class}}; \text{MLP}(\mathbf{y}^1); \cdots; \text{MLP}(\mathbf{y}^M)\right] + \mathbf{E}_{\text{seq}} \qquad \mathbf{E}_{\text{seq}} \in \mathbb{R}^{(M+1)\times D} \tag{6}$$

$$\mathbf{z}_0 = \left[\mathbf{z}_{i0} + \mathbf{z}_i^{\text{type}}; \mathbf{z}_{n0} + \mathbf{z}_n^{\text{type}}\right] \tag{7}$$

$$\mathbf{z}'_l = \text{MSA}\left(\text{LN}\left(\mathbf{z}_{l-1}\right)\right) + \mathbf{z}_{l-1}, \qquad l = 1 \ldots L \tag{8}$$

$$\mathbf{z}_l = \text{MLP}\left(\text{LN}\left(\mathbf{z}'_l\right)\right) + \mathbf{z}'_l, \qquad l = 1 \ldots L \tag{9}$$

$$\hat{\mathbf{z}} = \text{LN}\left(\left[\mathbf{z}_{i\,L}^0; \mathbf{z}_{n\,L}^0\right]\right) \tag{10}$$

For all experiments presented in this paper, hidden size D of 192, later depth L of 12, patch size P of 8, MLP size of 192, and number of attention heads of 12 are used.

2.2.3. Smart Persistent Model

This paper uses the Smart Persistent Model (SPM) as the benchmark for evaluating the performance of alternative modelling approaches. In contrast to the Persistent Model (PM), which assumes that solar irradiance remains constant throughout the forecast interval, the SPM assumes instead that the clear sky index remains constant. This offers the advantage that potential seasonal and temporal factors are added to the model as default preconditions and can be expressed as follows:

$$\hat{\mathbf{z}}_{\text{SPM}}(T + \Delta T) = \frac{\mathbf{z}(T)}{\mathbf{z}_{\text{clear}}(T)} \cdot \mathbf{z}_{\text{clear}}(T + \Delta T)$$

Implicit in the use of a SPM is the requirement for a clear sky model as a reference for clear sky irradiance. In this paper, the McClear model [47] is used for clear sky irradiance generation.

2.2.4. AutoML - Additional Machine Learning Benchmarks

As part of the process of evaluating the performance of image-numerical multi-modal learning an additional predictive regression model based on only the numerical input data was created to serve as an additional benchmark. This made use of the AutoGluon [58] tool, which was used to train a forecast model and is based on the idea of automated machine learning (AutoML). AutoGluon can automate model selection, hyper-parameter tuning and model integration. The final model was generated by integrating one or more of neural networks: LightGBM boosting trees [59], CatBoost boosting trees [60], random forests, extreme randomization trees, and kNearest Neighbours, and based on multilayer stack resembling and repeated k-fold bagging strategy to increase the final accuracy [58]. In the presentation and discussion of the results, this model is referred to using the abbreviation NUM.

2.2.5. Summary of models and criteria for evaluating performance

A summary of the models used in this paper is provided in Table 1. The SPM, NUM, and CNN-L models represent benchmarks for persistence, numerical-based machine learning, and combined image-numerical based deep approaches, respectively. ViT represents the image backbone network based on Transformer architecture proposed here as the alternative to the use of a CNN. The terms appended to CNN and ViT define the approach taken to fusion where -L represents late feature-level fusion architecture, -LG represents extra gate architecture, and -E represents feature-level fusion architecture. More detailed models architecture is presented in Appendix B.

Table 1: Irradiance Forecasting models explored through this paper

Models	Inputs		Encoder architecture		Fusion	Reference
	Numerical	Images	Numerical	Images		
SPM	✓		Persistence	/	/	
NUM	✓		AutoGluon	/	/	[58]
CNN-L	✓	✓	ANN	Res-18	Late	[37, 20, 45]
CNN-LG	✓	✓	ANN	Res-18	Late, Gated	[31]
ViT-L	✓	✓	ANN	ViT-Base-patch8-128	Late	[51]
ViT-LG	✓	✓	ANN	ViT-Base-patch8-128	Late, Gated	[51, 31]
ViT-E	✓	✓	Transformer	ViT-Base-patch8-128	Early	

2.3. Evaluation Matrix

Two evaluation criteria were used to evaluate the performance these models. The first involved quantifying the error between the predicted irradiance $\hat{\mathbf{z}}$ and the ground truth data $\mathbf{z}^*$. Standard metrics widely used by the solar forecasting community, and adopted in this paper, include FS based on metrics such as RMSE, MAE or MSE to measure the running accuracy of the forecast. The second criterion was based on BP, which quantifies forecasting ability in the presence of a Ramp Event, i.e., a sudden rise or fall in irradiance due to sudden changes in cloud cover.

Forecast Skill As statistical indicators such as RMSE, MAE or MSE tend to behave in a homo-trending manner in solar forecasting. The Forecast Skill (FS), adopted in this paper used the Smart Persistent Model (SPM) clear-sky model to represent the baseline performance and only RMSE to quantify error, as follows:

$$\text{Forecast Skill} = (1 - \frac{RMSE_{\text{Model}}}{RMSE_{\text{Baseline}}}) \times 100\%$$

Balanced precision Although FS can quantify the general error between model forecasts and ground truth, it does not demonstrate the ability of models to forecast ramp events. These qualitative behaviours are of particular importance in PV generation as the rapid power fluctuations that result, increase the system frequency stabilisation cost. Balanced precision (BP) is a metric developed for ramp events [61], which defines a ramp as a rapid solar irradiance event with a rate of change exceeding 10% of the maximum installed capacity. This paper uses a modified version of the metric where periods exhibiting a rate of change in GHI exceeding 100 $W/m^2/min$ are defined as ramp events – this is to reflect the fact that for the database used, there is not a grid to as a reference., Following the suggestions of Kong et al. [45], this paper also defines the ramp direction.

For each forecast, data can be classified into three categories based on the magnitude and direction of change in solar irradiance, i.e., positive ramp events where cloud cover diminishes, negative ramp events where cloud cover grows, and periods of relatively consistent irradiation, implying an absence of ramp events. After categorising the forecast data to identify ramp events, BP may be defined as:

$$\text{Balanced Precision} = \frac{1}{2}\sum_{c\in C}\frac{\mathscr{T}_c}{\mathscr{N}_c}$$

Where $\mathscr{T}_c$ represents successfully forecast events in the positive or negative ramp category and $\mathscr{N}_c$ represents the total sample in the positive or negative ramp category.

3. Results and Discussion

Modelling was undertaken using a PC with a 3.8 GHz AMD Ryzen 9 3900X CPU and a GeForce RTX 2080 SUPER GPU on the Tensorflow 2.5 [62] platform with Keras [63] built in. To reduce errors introduced by random nature in modelling, including the randomness in observation order and the randomness in random number generator in training, five replicate trials were carried out for each image model.

3.1. Results

3.1.1. Quantitative solar irradiance forecasting

Results for the criteria used to evaluate the quantitative capabilities of the five image-numerical models (CNN-L, CNN-LG, ViT-L, ViT-LG, ViT-E) and two numerical models (SPM and NUM) are summarised in Table 2.

Table 2: GHI forecast results. The errors are expressed as mean ± standard deviation. Forecast skill was calculated relative to the SPM model.

Models	2 min		6 min		10 min	
	RMSE (W/m^2) ↓	FS (%) ↑	RMSE (W/m^2) ↓	FS (%) ↑	RMSE (W/m^2) ↓	FS (%) ↑
SPM	85.62	N/A	117.57	N/A	129.67	N/A
NUM	77.31	**9.70**	98.69	16.06	113.14	12.75
CNN-L	79.37±0.55	7.29±0.64	98.68±0.45	16.07±0.38	105.15±0.49	18.9±0.37
CNN-LG	79.89±0.66	6.68±0.76	98.54±0.64	**16.18±0.54**	104.15±0.37	**19.68±0.29**
ViT-L	82.77±0.82	3.32±0.96	99.97±0.65	14.97±0.55	105.28±1.27	18.81±0.98
ViT-LG	85.16±1.34	0.53±1.56	101.29±0.8	13.84±0.67	105.26±0.45	18.82±0.34
ViT-E	81.45±0.68	4.87±0.79	98.68±0.72	16.06±0.61	104.91±0.7	19.09±0.53

It may be seen that all models outperformed the SPM model which was used as the FS baseline predictive power. The AutoML-based NUM model achieved the best forecast results at the 2-minute horizon; the CNN model with a gate architecture achieved the best results for the 6-minute and 10-minute forecasts. Overall, there was a large difference in model FS levels at the 2-minute horizon, and this difference diminished as the forecast horizon was extended. In particular, the models based on ViT as the graphical feature extractor were all inferior to the CNN-based models in FS.

It is worth noting that for the late feature level fusion models, the effect of gate architecture is not significant, with the difference in FS being less than 1% across all models, with the exception of the ViT-LG model, which delivers significantly lower FS at the 2-minute time horizon. The ViT-E model, where the numerical and image inputs share a single encoder, outperforms both the ViT-L and ViT-LG models, where features are extracted separately and then fused, at all forecast time horizons. As shown by the linear regression curves in Figure 8, the errors in all models manifest as an overestimation of irradiance at lower irradiance and an underestimation at higher irradiance.

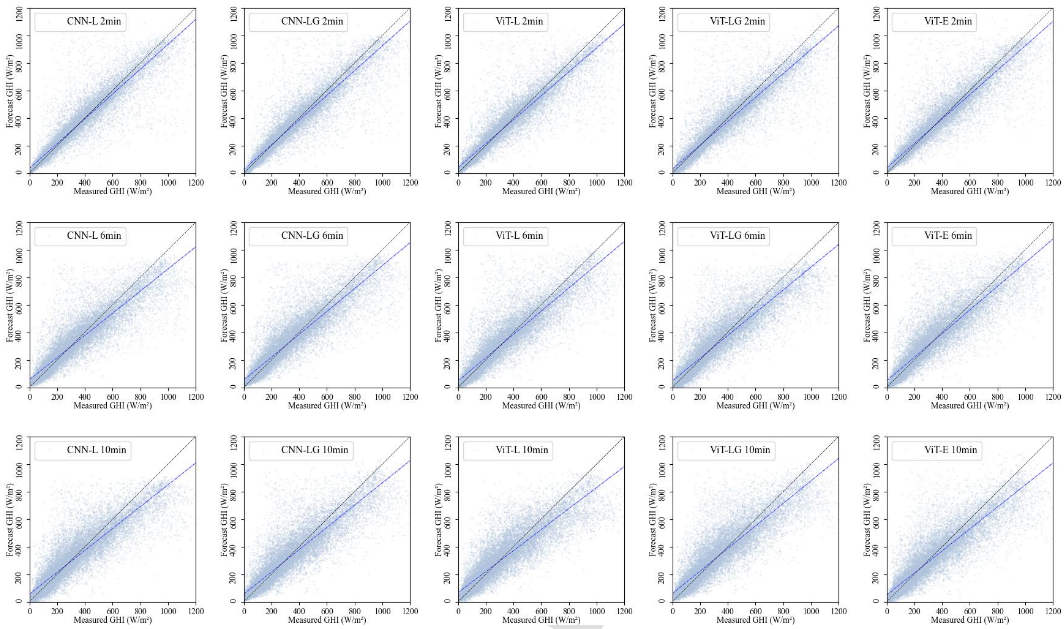


Figure 8: Forecasts using the image-numerical bimodal models over three time horizons. The blue dashed line is the predicted linear regression and the black dashed line is the expected regression (predicted value = actual value)

3.1.2. Qualitative solar irradiation (Ramp Event) forecasting

Table 3 presents the qualitative results for all models in terms of how often Ramp Events were accurately predicted, and Figure 9 illustrates performance as a confusion matrix. It may be seen that models based on the ViT framework achieve the best performance across all time horizons. It may also be seen that the qualitative results exhibit a similar trend to the quantitative results, i.e., the variability between models decreases as the forecast time horizon increases. In the case of qualitative results, however, the variability is more pronounced. At all horizons, the BP of the ViT-based models was greater than or equal to that of the CNN-based models. Additionally, the performance of the models with gate architectures exceeded or equalled that of the non-gated models. Interestingly, the BP of the widely used CNN-L fusion framework was even lower than that of the purely numerical forecast-based model NUM for 2-minute forecast. Even after the addition of the gate architecture enhanced the model's BP ability, its performance was still lower than that of NUM. Finally, it may be seen that models successfully captured falling RE more frequently than rising RE, the exception being the ViT frame model over the 2-minute horizon.

Table 3: Ramp Event forecasting results. For image-numerical models, results are expressed as the mean ± standard deviation of the results of five replicate trials.

Horizon	Models	Increase RE ↑	Decrease RE ↑	BP (%) ↑
2 min	SPM	0/1131	4/1071	0.19
	NUM	135/1131	214/1071	15.96
	CNN-L	62.6±62/1131	171.8±34.9/1071	10.78±3.41
	CNN-LG	96.2±58.2/1131	188.6±29.7/1071	13.05±1.94
	ViT-L	226.8±52.5/1131	180.8±55/1071	18.46±1.02
	ViT-LG	241±29.6/1131	185.4±34.9/1071	19.31±1.1
	ViT-E	239.4±18.8/1131	206.2±28.6/1071	**20.21±2.01**
6 min	SPM	0/1979	23/2028	0.57
	NUM	421/1979	697/2028	27.82
	CNN-L	518±84.7/1979	659.8±95.3/2028	29.35±2.26
	CNN-LG	537.4±91.5/1979	759.4±59.7/2028	32.3±1.03
	ViT-L	548.8±63.3/1979	752.6±33.2/2028	32.42±1.35
	ViT-LG	609.2±25.8/1979	752.2±55.6/2028	**33.93±1.78**
	ViT-E	671.8±28.7/1979	660.6±27.8/2028	33.26±0.9
10 min	SPM	0/2483	42/2603	0.81
	NUM	212/2483	426/2603	12.45
	CNN-L	808±61.7/2483	1101±74.9/2603	37.42±1.52
	CNN-LG	819.8±33.5/2483	1072.8±85.6/2603	37.11±1.52
	ViT-L	788±76.4/2483	1133.8±123.1/2603	**37.64±1.58**
	ViT-LG	852.4±93.5/2483	1050±93.2/2603	37.33±2.55
	ViT-E	819.6±140.4/2483	1060.6±148.6/2603	36.87±2.55

3.1.3. Comparison of model variability

Figure 10 shows the combined FS and BP performance for all models. As the SPM model has little RE predictive power, it can be approximated as being at the origin of the coordinate system and is not plotted in the figure. As observed in the work of Paletta et al., [34], the effect of architecture used in different models fed by the same inputs gradually decreases as the size of

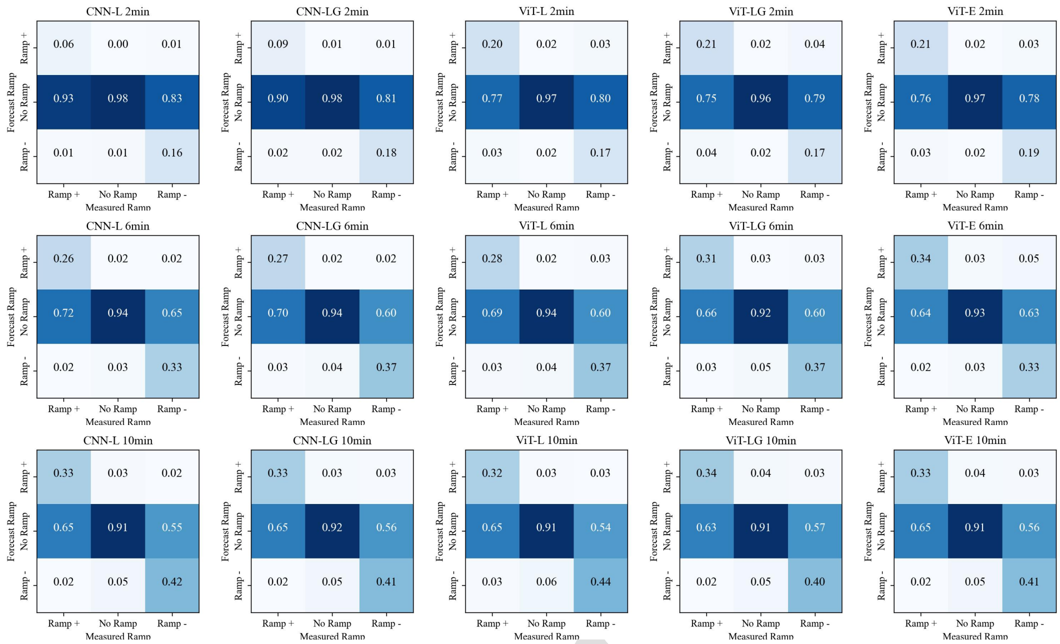


Figure 9: Confusion matrix of Ramp predictive power for 5 different image-numerical models on 3 time horizon

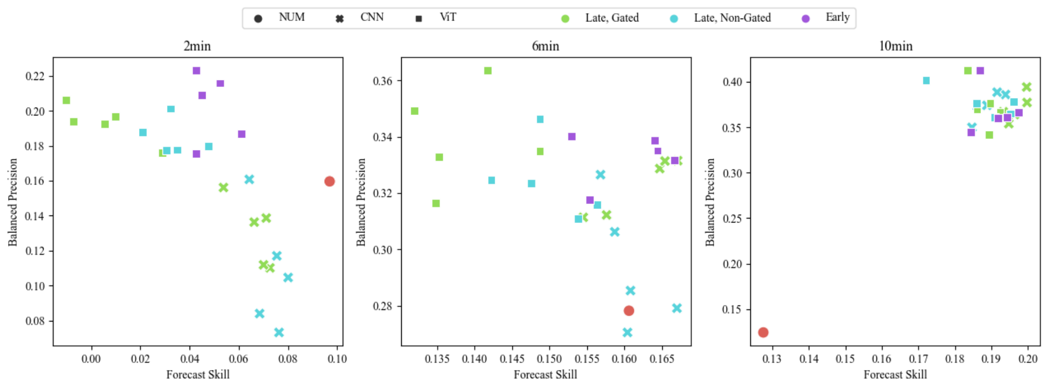


Figure 10: FS and BP results for all models over different time horizons.

the forecast horizon grows. For the bimodal frameworks studied here, it is difficult to identify any significant variability in the models at the 10 minute time horizon.

In reflecting upon performance, it is worth distinguishing between the relative importance of quantitative verses qualitative measures. In the field of solar forecasting, the merit of a model is usually determined using quantitative error, i.e., FS. The optimal strategy for such models fitted by statistical errors for rapidly changing cloudy weather is often based on mean reversion. However, for very short-term solar forecasting (10 minutes or less), the ability to capture Ramp events is more important as the information may be used to inform grid operability.

Such ramp forecasts require the model to predict the occurrence of sudden and large changes in irradiance, as opposed to consistent predictions of absolute irradiance, and metrics that quantify performance in terms of statistical error, e.g., RMSE, tend to penalise the former qualities. The 2- and 6-minute results from Figure 10 show that the models with high BP performance, i.e., ViT-L and ViT-LG, perform poorly when performance is expressed as FS, while the opposite is true for CNN models. The early feature-level fusion model, ViT-E, maintained relatively strong BP performance in the 2- and 6-minute predictions compared to the late model, and both delivered the best FS. It is posited here that there are two main reasons for this, namely the ability of the model to abstract image features, and the dual-modality strategy the model adopts to accommodate the visual and numerical inputs.

3.1.4. Impact of images in bimodal models

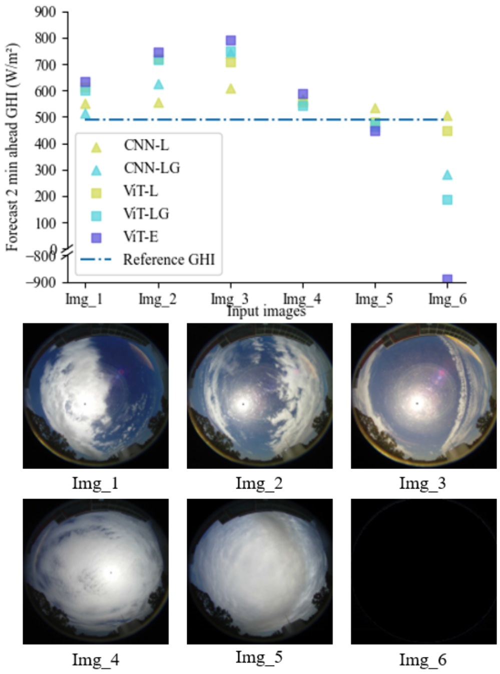


Figure 11: Image sensitivity testing for a 2-minute time horizon. Image 1 is the original image input and Image 2 to Image 6 are replacement inputs. The upper panel shows the 2-minute ahead prediction from the 5 image-numerical bimodal models. The blue dashed line represents the output from the SPM model.

To explore the sensitivity of different models to the image input, randomly selected images were used as inputs to the models on 17 June at 18:35, while keeping the numerical input unchanged. The condition of the sky at this time is shown in Image 1 of Figure 11, as are the replacement images used in the analysis - Images 2 to 5, are taken from the same day but with different sky conditions and Image 6, which is fabricated and comprises only black pixels. The output from this analysis is plotted in Figure 11 and shows that model based on ViT as an image feature extractor are more significantly affected by the image input than those based on CNN under complex sky conditions. In addition, most of the models with gate architecture (light blue in the figure) are more sensitive to images than those based on late fusion (light brown in the figure). Furthermore, the ViT-E model is always the most sensitive to images. Interestingly, when fed a picture without any information, the output of CNN-L is almost unaffected, while ViT-E deviates significantly from the refence GHI value. These results suggest that the widely used CNN-L architecture is relatively insensitive to image inputs. In particular, the model is extremely insensitive to the incorrect input. This may be explained by the findings of Paletta et al., [20] who suggest, after evaluating multiple graphical models, that fusion models always behave like a smarter SPM. i.e., the model lacks interaction between image and numerical inputs, including alignment, translation, and co-representation. This makes the model dependent on the numerical inputs and relatively insensitive to the image-based output. To address this shortcoming, methods that use an image feature extractor that is more effective at of parsing images, such as ViT, or enhancing the interaction between image and numerical data, such as a gate architecture, can be considered as more effective approaches.

3.1.5. Interaction of image and numerical data in ViT-E

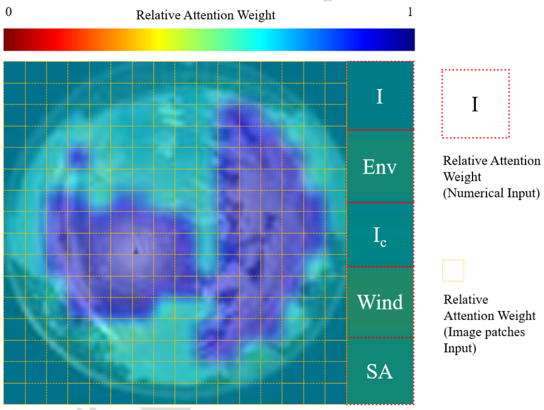


Figure 12: ViT-E model visualisation indicating relative attention weights. The colour of the heat map within each patch reveals its relative value in terms of average attention across all heads.

To understand how the Self-Attention mechanism processes image-numerical information across modalities, the attention layer of the ViT-E model was abstracted and overlaid with the input for visualisation, as shown in Figure 12. The visualised heat map consists of two main parts: on the left side are the relative attention weights corresponding to the 256 patches in the image input, and on the right side are the relative attention weights corresponding to five sets of numerical inputs, in order from top to bottom: irradiance, ambient environment, clear sky irradiance, wind condition, and solar angle. Figure 13(a) shows the GHI prediction from the ViT-E model for three different forecast horizons for the 17 June. A sample of five images, including that used in Figure 12,

479 representing a range of sky conditions were extracted and processed to visualise the model attention
480 weights as described above, and are shown in Figure 13(b).

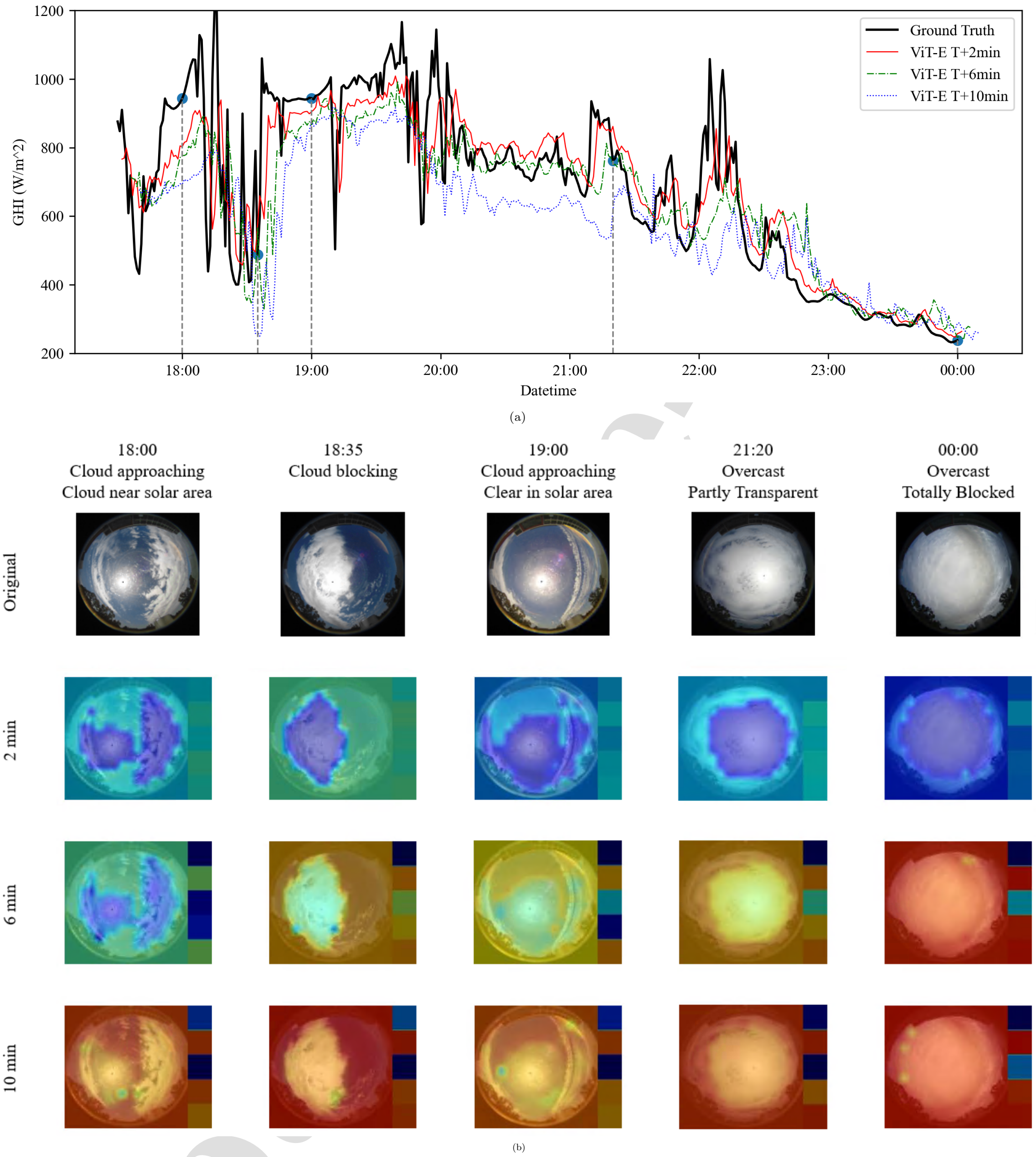


Figure 13: (a) GHI predictions from 17 June, based on ViT-E 2-, 6-, and 10-minute forecasts. (b) Attention map of the ViT-E model based on five representative GHI conditions from Figure 13 (a).

It may be seen from Figure 13(b), that the longer the forecasting horizon, the lower the attention weight of the model to the image-side input and the higher the attention weight to the numerical input. In the 2-minute ahead prediction, different levels of cloud cover and sun position significantly affect the attention of the model. For scenarios with low cloud-sun correlation, such as those with significant areas of clear sky in region around the sun, or those where the sun is totally obscured by cloud, the model assigns weights to both numerical and image models in a balanced manner. For scenarios with high cloud-sun correlation, such as cloud approaching or cloud blocking part of the sun, the model assigns more attention to the images. In the 6-minute ahead model, although the distribution of attention weights for the images reflects that of the 2-minute ahead model, the weighting of the numerical data is the most important part of the model. This trend of assigning a gradually decreasing weighting to images continues in the 10-minute ahead model, where the model becomes primarily dependent on irradiance and clear sky irradiance numerical inputs rather than the images.

This pattern of behaviour offers an explanation for the variability in model performance observed in Figure 10 where accuracy of the forecast declines as the prediction window is lengthened. That is, the impact of the details in the pictures on the prediction decreases as the prediction scale is lengthened. Although other potentially valuable information visible in the images (e.g., air mass) might still benefit the predictive capabilities of the model and thus outperform models without an image input, enhancing the feature extraction capability for the images for these longer time horizon forecasts is unlikely to deliver better model performance. This observation is matches that made in relation to models based on the classical image analysis method for forecasting GHI [64], i.e., the gain offered by including image data in predictions is more pronounced for time horizons below five minutes, and gradually decreases for those beyond five minutes.

We believe that the trend is a good explanation for the reason for model performance variability in Fig. 10 declines as the prediction window is lengthened. That is, the impact of the details in the pictures on the prediction is gradually decreasing as the prediction scale is lengthened. Although other potentially visible information in the images (e.g., air mass) can still enable the model to benefit in prediction and thus outperform the model without image input, enhancing the model's feature extraction capability for the images at this point no longer leads to better model performance. This is similar to the model based on the classical image analysis method for forecasting GHI [64], i.e., the gain of image data on prediction is more pronounced within five minutes, while it starts to gradually decrease after five minutes.

The results from this study suggest that there are advantages to using the transformer framework for combined image-numerical ultra-short-term solar forecasting. Specifically, the model extracts features based on the association between each of each input elements, i.e., image patches and numerical features, and dynamically assigns the impact of each element on the final prediction based on these features. This functional advantage is not conferred by ANN-based architectures as model fusion feature extractors.

In addition, as shown in Figure 13 (b), the 10-minute forecast irradiance has a similar weighting to the clear irradiance. In other words, clear sky irradiance is of equal importance to prevailing irradiance for solar irradiance prediction. The advantages of using CSI, i.e. the ratio of GHI to clear GHI, rather than using GHI directly as a prediction target [44], are intuitively demonstrated.

4. Discussion

Despite deep learning methods having demonstrated superior effectiveness over other approaches in terms of results, this study illustrates that the currently implemented intra-hour solar power forecasting deep model architectures can still yield diametrically opposing performances. It has been evidenced that different architectures and modal fusion methods can significantly influence the predictive capability of the model. As seen in Figure 10, the quantitative and qualitative performance of different models are not uniform. Models leveraging Convolutional Neural Networks (CNNs) as the image feature extraction algorithm show insensitivity to changes in the image modal input, whereas architectures based on attention mechanisms lack precision in quantitative results.

On the one hand, algorithmically, as we proposed in Section 2.2, we speculate that this disparity might be determined by the underlying algorithms of the network backbone architectures. Attention mechanisms excel in inferring through relative relationships among image pixels, thus they are more sensitive than convolutional computations that extract image details in sky image analysis. On the other hand, from the evaluation perspective, we believe that the intrinsic contradiction between qualitative and quantitative analyses results in models exhibiting markedly different patterns.

In quantitative analyses, models are expected to achieve larger FS, in other words, smaller RMSE. This constraint makes the model more sensitive to numerical data, showing a trend for mean prediction [20]. Under such circumstances, the model tends to be conservative when dealing with rapid extreme changes, like ramp events, as observed in Figure 11. In qualitative analyses, models are expected to capture more REs and further predict their trends. In this process, mean prediction sensitive to numerical values causes the model to miss most REs. However, the ViT-L series architecture, which is more sensitive to image analysis, tends to over-predict REs and loses quantitative performance. In addition, the attention model ViT-E, which is based on early fusion and accepts inputs from different modals, can achieve a more balanced quantitative and qualitative result.

Furthermore, in Section 3.1.5, the manifestation of weights within the model indicates that the importance of ground-based sky image information for solar power deep networks gradually decreases with the extension of the forecast horizon. Particularly for ramp event prediction, which is of great interest for intra-hour forecasting, a longer forecast horizon tends to homogenise different models, eventually displaying similar performances. We speculate that this phenomenon may be associated with the limited presence time of low-level rapid clouds in sky images, which are responsible for rapid RE changes. This conclusion aligns with cloud observation findings based on image analysis methods [65].

5. Conclusions

Accurate short-term forecasting is essential for predicting solar power output, and thus for effective grid management. This study found that the modal interaction component has been under-appreciated in previous studies of deep learning models for solar forecasting that combine images with numerical inputs. Also, there is ambivalence between the quantitative and qualitative performance of late feature-level fusion models for single image and numerical fusion in such models. Therefore, this project proposed the ViT-E model as being complementarity in quantitative and qualitative forecast performance by varying the modal interactions to achieve relatively superior performance. In addition, the study explored the weighting of image inputs in this class of model. The results show that the longer the forecast duration in a single image forecast, the less importance the image accounts for, and at forecasts of up to 10-minute horizons, the features that can be extracted from the image input by current vision models are minimal. As mentioned in [66], the accuracy of the model is as important as its interpretability in advancing its understanding and development. This study reveals a potential shortcoming in current multimodal solar prediction: model validation relies only on performance improvements for the results, and there is a lack of interaction studies between the actual performance of the different modes of the model, such as ablation experiments. Transformer-like models have full potential in hybrid modelling for solar energy prediction due to the intuitive interpretability of their framework. Furthermore, in future work, we propose to use the RNN framework in combination with the Transformer framework for Seq2sqe models with dynamic picture data streams as a framework to drive the current prediction framework.

Acknowledgement

This work was supported by the Engineering and Physical Sciences Research Council, UK [grant number EP/W028581/1].

Reference

References

[1] Rich H Inman, Hugo TC Pedro, and Carlos FM Coimbra. Solar forecasting methods for renewable energy integration. *Progress in energy and combustion science*, 39(6):535–576, 2013.

[2] Anna-Lena Klingler and Lukas Teichtmann. Impacts of a forecast-based operation strategy for grid-connected pv storage systems on profitability and the energy system. *Solar Energy*, 158:861–868, 2017.

[3] Dazhi Yang, Stefano Alessandrini, Javier Antonanzas, Fernando Antonanzas-Torres, Viorel Badescu, Hans Georg Beyer, Robert Blaga, John Boland, Jamie M Bright, and Carlos FM Coimbra. Verification of deterministic solar forecasts. *Solar Energy*, 210:20–37, 2020.

[4] Utpal Kumar Das, Kok Soon Tey, Mehdi Seyedmahmoudian, Saad Mekhilef, Moh Yamani Idna Idris, Willem Van Deventer, Bend Horan, and Alex Stojcevski. Forecasting of photovoltaic power generation and model optimization: A review. *Renewable and Sustainable Energy Reviews*, 81:912–928, 2018.

[5] Christian A Gueymard and Jose A Ruiz-Arias. Extensive worldwide validation and climate sensitivity analysis of direct irradiance predictions from 1-min global irradiance. *Solar Energy*, 128:1–30, 2016.

[6] Chi Wai Chow, Bryan Urquhart, Matthew Lave, Anthony Dominguez, Jan Kleissl, Janet Shields, and Byron Washom. Intra-hour forecasting with a total sky imager at the uc san diego solar energy testbed. *Solar Energy*, 85(11):2881–2893, 2011.

[7] Ricardo Marquez and Carlos FM Coimbra. Intra-hour dni forecasting based on cloud tracking image analysis. *Solar Energy*, 91:327–336, 2013.

[8] Dazhi Yang, Weixing Li, Gokhan Mert Yagli, and Dipti Srinivasan. Operational solar forecasting for grid integration: Standards, challenges, and outlook. *Solar Energy*, 224:930–937, 2021.

[9] Qingyong Li, Weitao Lu, Jun Yang, and James Z Wang. Thin cloud detection of all-sky images using markov random fields. *IEEE Geoscience and remote sensing letters*, 9(3):417–421, 2011.

[10] K Stefferud, J Kleissl, and J Schoene. Solar forecasting and variability analyses using sky camera cloud detection & motion vectors. In *2012 IEEE Power and Energy Society General Meeting*, pages 1–6. IEEE, 2012.

[11] Handa Yang, Ben Kurtz, Dung Nguyen, Bryan Urquhart, Chi Wai Chow, Mohamed Ghonima, and Jan Kleissl. Solar irradiance forecasting using a ground-based sky imager developed at uc san diego. *Solar Energy*, 103:502–524, 2014.

[12] Zhenzhou Peng, Dantong Yu, Dong Huang, John Heiser, Shinjae Yoo, and Paul Kalb. 3d cloud detection and tracking system for solar forecast using multiple sky imagers. *Solar Energy*, 118:496–519, 2015.

[13] Yinghao Chu, Mengying Li, Hugo TC Pedro, and Carlos FM Coimbra. A network of sky imagers for spatial solar irradiance assessment. *Renewable Energy*, 187:1009–1019, 2022.

[14] Samuel R West, Daniel Rowe, Saad Sayeef, and Adam Berry. Short-term irradiance forecasting using skycams: Motivation and development. *Solar Energy*, 110:188–207, 2014.

[15] Bijan Nouri, Pascal Kuhn, Stefan Wilbert, Christoph Prahl, Robert Pitz-Paal, Philippe Blanc, Thomas Schmidt, Zeyad Yasser, Lourdes Ramirez Santigosa, and Detlev Heineman. Nowcasting of dni maps for the solar field based on voxel carving and individual 3d cloud objects from all sky images. In *AIP Conference Proceedings*, volume 2033, page 190011. AIP Publishing LLC, 2018.

[16] Guang Wang, Ben Kurtz, and Jan Kleissl. Cloud base height from sky imager and cloud speed sensor. *Solar Energy*, 131:208–221, 2016.

[17] Lydie Magnone, Fabrizio Sossan, Enrica Scolari, and Mario Paolone. Cloud motion identification algorithms based on all-sky images to support solar irradiance forecast. In *2017 IEEE 44th Photovoltaic Specialist Conference (PVSC)*, pages 1415–1420. IEEE, 2017.

[18] Bijan Nouri, Stefan Wilbert, Luis Segura, P Kuhn, Natalie Hanrieder, A Kazantzidis, Thomas Schmidt, L Zarzalejo, Philipp Blanc, and Robert Pitz-Paal. Determination of cloud transmittance for all sky imager based solar nowcasting. *Solar Energy*, 181:251–263, 2019.

[19] Julien Nou, Rémi Chauvin, Julien Eynard, Stéphane Thil, and Stéphane Grieu. Towards the intrahour forecasting of direct normal irradiance using sky-imaging data. *Heliyon*, 4(4):e00598, 2018.

[20] Quentin Paletta, Guillaume Arbod, and Joan Lasenby. Benchmarking of deep learning irradiance forecasting models from sky images–an in-depth analysis. *Solar Energy*, 224:855–867, 2021.

[21] Haoran Wen, Yang Du, Xiaoyang Chen, Enggee Lim, Huiqing Wen, Lin Jiang, and Wei Xiang. Deep learning based multistep solar forecasting for pv ramp-rate control using sky images. *IEEE Transactions on Industrial Informatics*, 17(2):1397–1406, 2020.

[22] Jinsong Zhang, Rodrigo Verschae, Shohei Nobuhara, and Jean-François Lalonde. Deep photovoltaic nowcasting. *Solar Energy*, 176:267–276, 2018.

[23] Xin Zhao, Haikun Wei, Hai Wang, Tingting Zhu, and Kanjian Zhang. 3d-cnn-based feature extraction of ground-based cloud images for direct normal irradiance prediction. *Solar Energy*, 181:510–518, 2019.

[24] Jane Oktavia Kamadinata, Tan Lit Ken, and Tohru Suwa. Sky image-based solar irradiance prediction methodologies using artificial neural networks. *Renewable Energy*, 134:837–845, 2019.

[25] Hugo TC Pedro, Carlos FM Coimbra, Mathieu David, and Philippe Lauret. Assessment of machine learning techniques for deterministic and probabilistic intra-hour solar forecasts. *Renewable Energy*, 123:191–203, 2018.

[26] Ardan Hüseyin Eşlik, Emre Akarslan, and Fatih Onur Hocaoğlu. Short-term solar radiation forecasting with a novel image processing-based deep learning approach. *Renewable Energy*, 200:1490–1505, 2022.

[27] M Caldas and R Alonso-Suárez. Very short-term solar irradiance forecast using all-sky imaging and real-time irradiance measurements. *Renewable Energy*, 143:1643–1658, 2019.

[28] Stavros-Andreas Logothetis, Vasileios Salamalikis, Stefan Wilbert, Jan Remund, Luis F Zarzalejo, Yu Xie, Bijan Nouri, Evangelos Ntavelis, Julien Nou, Niels Hendrikx, et al. Benchmarking of solar irradiance nowcast performance derived from all-sky imagers. *Renewable Energy*, 199:246–261, 2022.

[29] D Anagnostos, T Schmidt, S Cavadias, D Soudris, J Poortmans, and F Catthoor. A method for detailed, short-term energy yield forecasting of photovoltaic installations. *Renewable Energy*, 130:122–129, 2019.

[30] Yann LeCun, Léon Bottou, Yoshua Bengio, and Patrick Haffner. Gradient-based learning applied to document recognition. *Proceedings of the IEEE*, 86(11):2278–2324, 1998.

[31] Sepp Hochreiter and Jürgen Schmidhuber. Long short-term memory. *Neural computation*, 9(8):1735–1780, 1997.

[32] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778, 2014.

[33] Karen Simonyan and Andrew Zisserman. Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556*, 2014.

[34] Quentin Paletta, Anthony Hu, Guillaume Arbod, and Joan Lasenby. Eclipse: Envisioning cloud induced perturbations in solar energy. *Applied Energy*, 326:119924, 2022.

[35] Cong Feng and Jie Zhang. Solarnet: A sky image-based deep convolutional neural network for intra-hour solar forecasting. *Solar Energy*, 2020.

[36] Tadas Baltrušaitis, Chaitanya Ahuja, and Louis-Philippe Morency. Multimodal machine learning: A survey and taxonomy. *IEEE transactions on pattern analysis and machine intelligence*, 41(2):423–443, 2018.

[37] Yuchi Sun, Vignesh Venugopal, and Adam R Brandt. Short-term solar power forecast with deep learning: Exploring optimal input and output configuration. *Solar Energy*, 188:730–741, 2019.

[38] Zhao Zhen, Jiaming Liu, Zhanyao Zhang, Fei Wang, Hua Chai, Yili Yu, Xiaoxing Lu, Tieqiang Wang, and Yuzhang Lin. Deep learning based surface irradiance mapping model for solar pv power forecasting using sky image. *IEEE Transactions on Industry Applications*, 56(4):3385–3396, 2020.

[39] Vignesh Venugopal, Yuchi Sun, and Adam R Brandt. Short-term solar pv forecasting using computer vision: The search for optimal cnn architectures for incorporating sky images and pv generation history. *Journal of Renewable and Sustainable Energy*, 11(6):066102, 2019.

[40] Jun Yu, Jing Li, Zhou Yu, and Qingming Huang. Multimodal transformer with multi-view visual representation for image captioning. *IEEE transactions on circuits and systems for video technology*, 30(12):4467–4480, 2019.

[41] Jian Huang, Jianhua Tao, Bin Liu, Zheng Lian, and Mingyue Niu. Multimodal transformer fusion for continuous emotion recognition. In *ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 3507–3511. IEEE, 2020.

[42] Hassan Akbari, Liangzhe Yuan, Rui Qian, Wei-Hong Chuang, Shih-Fu Chang, Yin Cui, and Boqing Gong. Vatt: Transformers for multimodal self-supervised learning from raw video, audio and text. *Advances in neural information processing systems*, 34:24206–24221, 2021.

[43] Yao-Hung Hubert Tsai, Shaojie Bai, Paul Pu Liang, J Zico Kolter, Louis-Philippe Morency, and Ruslan Salakhutdinov. Multimodal transformer for unaligned multimodal language sequences. In *Proceedings of the conference. Association for Computational Linguistics. Meeting*, volume 2019, page 6558. NIH Public Access, 2019.

[44] Dazhi Yang. Choice of clear-sky model in solar forecasting. *Journal of Renewable and Sustainable Energy*, 12(2):026101, 2020.

[45] Weicong Kong, Youwei Jia, Zhao Yang Dong, Ke Meng, and Songjian Chai. Hybrid approaches based on deep whole-sky-image leaing to photovoltaic generation forecasting. *Applied Energy*, 280:115875, 2020.

[46] Hugo TC Pedro, David P Larson, and Carlos FM Coimbra. A comprehensive dataset for the accelerated development and benchmarking of solar forecasting methods. *Journal of Renewable and Sustainable Energy*, 11(3):036102, 2019.

[47] Mireille Lefevre, Armel Oumbe, Philippe Blanc, Bella Espinar, Benoît Gschwind, Zhipeng Qu, Lucien Wald, Marion Schroedter-Homscheidt, Carsten Hoyer-Klick, and Antti Arola. Mcclear: a new model estimating downwelling solar radiation at ground level in clear-sky conditions. *Atmospheric Measurement Techniques*, 6(9):2403–2418, 2013.

[48] Dazhi Yang and Christian A Gueymard. Ensemble model output statistics for the separation of direct and diffuse components from 1-min global irradiance. *Solar Energy*, 208:591–603, 2020.

[49] Chuck N Long and Yan Shi. An automated quality assessment and control algorithm for surface radiation measurements. *The Open Atmospheric Science Journal*, 2(1), 2008.

[50] Yuhao Nie, Ahmed S Zamzam, and Adam Brandt. Resampling and data augmentation for short-term pv output prediction based on an imbalanced sky images dataset using convolutional neural networks. *Solar Energy*, 224:341–354, 2021.

[51] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, and Sylvain Gelly. An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv preprint arXiv:2010.11929*, 2020.

[52] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*, 2018.

[53] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Lukasz Kaiser, and Illia Polosukhin. Attention is all you need. *Advances in neural information processing systems*, 30, 2017.

[54] Abdul Rahim Pazikadin, Damhuji Rifai, Kharudin Ali, Muhammad Zeesan Malik, Ahmed N Abdalla, and Moneer A Faraj. Solar irradiance measurement instrumentation and power solar generation forecasting based on artificial neural networks (ann): A review of five years research trend. *Science of The Total Environment*, 715:136848, 2020.

[55] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, and Jack Clark. Learning transferable visual models from natural language supervision. In *International Conference on Machine Learning*, pages 8748–8763. PMLR, 2021.

[56] Valentin Gabeur, Chen Sun, Karteek Alahari, and Cordelia Schmid. Multi-modal transformer for video retrieval. In *European Conference on Computer Vision*, pages 214–229. Springer, 2020.

[57] Wonjae Kim, Bokyung Son, and Ildoo Kim. Vilt: Vision-and-language transformer without convolution or region supervision. In *International Conference on Machine Learning*, pages 5583–5594. PMLR, 2021.

[58] Nick Erickson, Jonas Mueller, Alexander Shirkov, Hang Zhang, Pedro Larroy, Mu Li, and Alexander Smola. Autogluon-tabular: Robust and accurate automl for structured data. *arXiv preprint arXiv:2003.06505*, 2020.

[59] Guolin Ke, Qi Meng, Thomas Finley, Taifeng Wang, Wei Chen, Weidong Ma, Qiwei Ye, and Tie-Yan Liu. Lightgbm: A highly efficient gradient boosting decision tree. *Advances in neural information processing systems*, 30, 2017.

[60] Anna Veronika Dorogush, Vasily Ershov, and Andrey Gulin. Catboost: gradient boosting with categorical features support. *arXiv preprint arXiv:1810.11363*, 2018.

[61] Mohamed Abuella and Badrul Chowdhury. Forecasting of solar power ramp events: A post-processing approach. *Renewable Energy*, 133:1380–1392, 2019.

[62] Martín Abadi, Paul Barham, Jianmin Chen, Zhifeng Chen, Andy Davis, Jeffrey Dean, Matthieu Devin, Sanjay Ghemawat, Geoffrey Irving, Michael Isard, et al. {TensorFlow}: a system for {Large-Scale} machine learning. In *12th USENIX symposium on operating systems design and implementation (OSDI 16)*, pages 265–283, 2016.

[63] Francois Chollet et al. Keras, 2015.

[64] Ricardo Marquez and Carlos FM Coimbra. Intra-hour dni forecasting based on cloud tracking image analysis. *Solar Energy*, 91:327–336, 2013.

[65] Zhenzhou Peng, Dantong Yu, Dong Huang, John Heiser, Shinjae Yoo, and Paul Kalb. 3d cloud detection and tracking system for solar forecast using multiple sky imagers. *Solar Energy*, 118:496–519, 2015.

[66] Markus Reichstein, Gustau Camps-Valls, Bjorn Stevens, Martin Jung, Joachim Denzler, Nuno Carvalhais, et al. Deep learning and process understanding for data-driven earth system science. *Nature*, 566(7743):195–204, 2019.

771 # Appendix

772 ## Appendix A. Random sampling

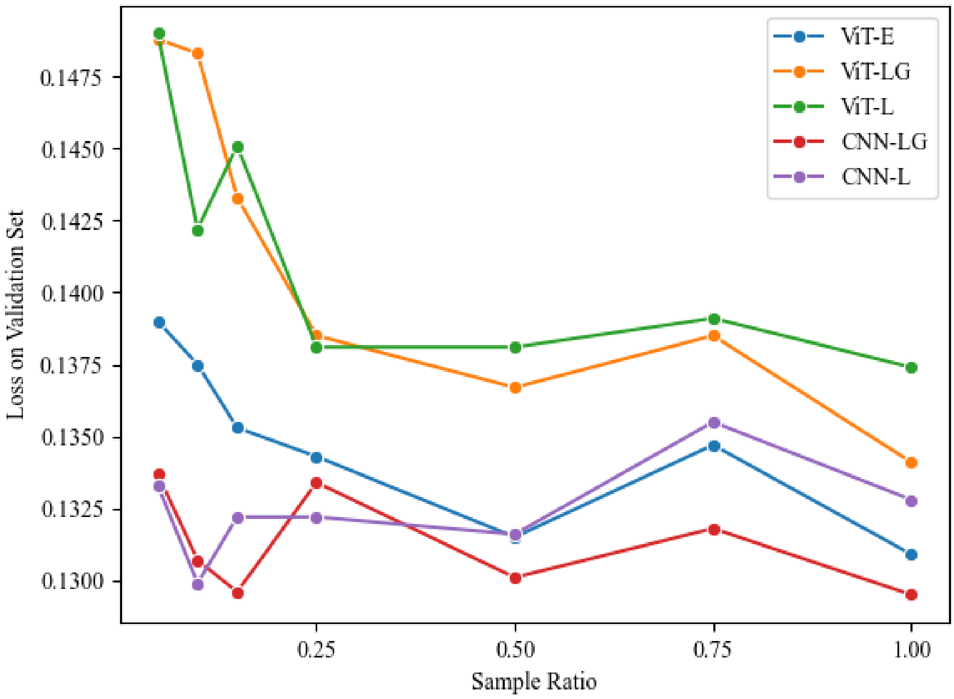


Figure A.14: Sampling rate validation experiments. The training set was used to train five different models with sampling rates of 0.05, 0.1, 0.15, 0.25, 0.5, 0.75 and 1.0. The models were then validated under the same validation set. The model loss tends to flatten out above 0.25 sample ratio.

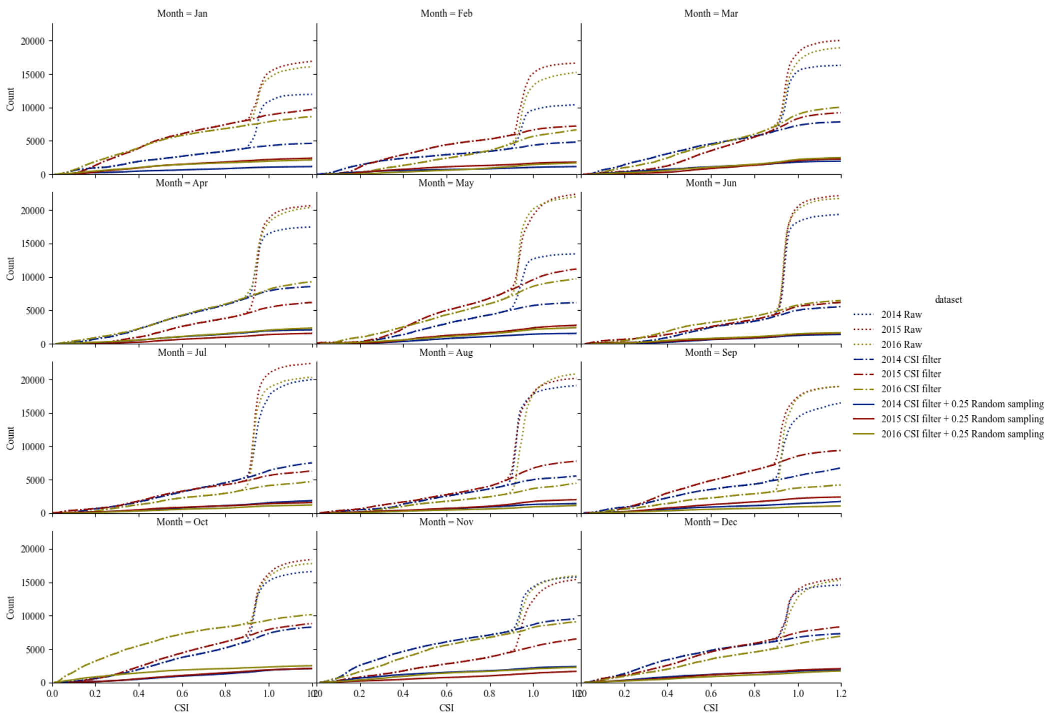


Figure A.15: Monthly CSI distribution of raw data, compared to Clear sky filtered data and 25% randomly sampled filtered data.

773 ## Appendix B. Model details

Table B.4: Hyperparameters of the Adam optimizer for training models

Hyperparameters	**CNN-L**	**CNN-LG**	**ViT-L**	**ViT-LG**	**ViT-E**
Learning rate	0.01	0.1	0.0008	0.0008	0.0008
Optimizer	SGD	SGD	SGD	SGD	SGD
Optimizer momentum	0.9	0.9	0.9	0.9	0.9
Loss	MSE	MSE	MSE	MSE	MSE
Weight decay	0.0001	0.0001	0.0001	0.0001	0.0001
Batch size	64	64	8	8	8
Training epochs	80	80	80	80	80
Warm up percentage	25%	25%	0	0	0
Learning rate decay	Cosine	Cosine	Cosine	Cosine	Cosine
Early stop	True	True	True	True	True
Early stop tolerance	20	20	20	20	20

Table B.5: The details of ViT-E model

Block	**Layer**	**Resolution**	**Channels**
Image Inputs	-	$128 \times 128 \times 3$	1
Image Patch Embedding	`Conv 8` $\times$ `8`	$128 \times 128 \times 3 \rightarrow 8 \times 8 \times 3$	$1 \rightarrow 256$
Image Class Token	`Transfer Embedding Projection` `Class Token Concat`	$8 \times 8 \times 3 \rightarrow 192$ 192	$256 \rightarrow 256$ $256 \rightarrow 257$
Position Embedding	`Position Embedding`	192	257
Numerical Imputs	-	14 (3 + 3 + 3 + 2 + 3)	1
Numerical Class Token	`Numerical Projection (MLP)` `Class Token Concat`	$14 \rightarrow 192$ 192	5 $5 \rightarrow 6$
Sequence Embedding	`Sequence Embedding`	192	6
Concatenation	`Concat`	192	263 (257 + 6)
Attention Block $\times$ 12	`LayerNorm` `Multi-Head Attention` $\times$ 12 `Add` (residual connection) `LayerNorm` `Multi-Head Attention` $\times$ 12 `Add` (residual connection)	192 192 192 192 192 192	263 263 263 263 263 263
Layer Normalization	`LayerNorm`	192	263

Regression Head	`Extract Class Token`
	`MLP`
	`MLP`
	`MLP`
	`MLP`
		384
		768
		512
		64
		1
			1
			1
			1
			1
			1

Table B.6: The details of ViT-LG model.

Block	Layer	Resolution	Channels
Image Inputs	-	$128 \times 128 \times 3$	1
Image Patch Embedding	`Conv8` $\times$ `8`	$128 \times 128 \times 3 \rightarrow 8 \times 8 \times 3$	$1 \rightarrow 256$
Image Class Token	`Transfoer Embedding Projection` `Class Token Concat`	$8 \times 8 \times 3$ $8 \times 8 \times 3$	$256 \rightarrow 256$ $256 \rightarrow 257$
Position Embedding	`Position Embedding`	$8 \times 8 \times 3$	257

Image Attention Block $\times$ 12	`LayerNorm`
	`Multi-Head Attention` $\times$ 12
	`Add (residual connection)`
	`LayerNorm`
	`Multi-Head Attention` $\times$ `12`
	`Add (residual connection)`
		192
		192
		192
		192
		192
		192
			257
			257
			257
			257
			257
			257

Image Feature Vectorization	`Extract Class Token`
	`MLP`
	`MLP`
		192
		768
		64
			1
			1
			1

Numerical Inputs	-	14 (3 + 3 + 3 + 2 + 3)	1
Numerical Feature Vectorization	`MLP` `MLP`	$14 \rightarrow 16$ 16	1 1
Concatenation	`Concat`	80 (64 + 16)	1

Regression Head	`MLP`
	`Gate MLP`
	`Gate Multiply`
	`MLP`
	`MLP`
	`MLP`
		80
		80
		80
		64
		16
		1
			1
			1
			1
			1
			1
			1

Table B.7: The details of ViT-L model.

Block	Layer	Resolution	Channels
Image Inputs	-	$128 \times 128 \times 3$	1
Image Patch Embedding	`Conv 8` $\times$ `8`	$128 \times 128 \times 3 \rightarrow 8 \times 8 \times 3$	$1 \rightarrow 256$
Image Class Token	`Transfer Embedding Projection` `Class Token Concat`	$8 \times 8 \times 3$ $8 \times 8 \times 3$	$256 \rightarrow 256$ $256 \rightarrow 257$
Position Embedding	`Position Embedding`	$8 \times 8 \times 3$	257

Image Attention Block $\times$ 12	`LayerNorm`
	`Multi-Head Attention` $\times$ `12`
	`Add` (residual connection)
	`LayerNorm`
	`Multi-Head Attention` $\times$ `12`
	`Add`(residual connection)
		192
		192
		192
		192
		192
		192
			257
			257
			257
			257
			257
			257

Image Feature Vectorization	`Extract Class Token`
	`MLP`
	`MLP`
		192
		768
		64
			1
			1
			1

Numerical Inputs	-	14 (3 + 3 + 3 + 2 + 3)	1
Numerical Feature Vectorization	`MLP` `MLP`	$14 \rightarrow 16$ 16	1 1
Concatenation	`Concat`	80 (64 + 16)	1
Regression Head	`MLP` `MLP` `MLP` `MLP`	80 64 16 1	1 1 1 1

Table B.8: The details of CNN-LG model.

Block	**Layer**	**Resolution**	**Channels**
Image Inputs	-	$128 \times 128 \times 3$	1
ResNet Block Conv 1	`Conv 7 × 7`	$128 \times 128 \times 3 \to 64 \times 64 \times 3$	$1 \to 64$
	`Max Pooling 3 × 3`	$64 \times 64 \times 3 \to 32 \times 32 \times 3$	64
ResNet Block Conv 2 × 2	`Conv 3 × 3`	$32 \times 32 \times 3$	64
	`BatchNormal`	$32 \times 32 \times 3$	64
	`Conv 3 × 3`	$32 \times 32 \times 3$	64
	`BatchNormal`	$32 \times 32 \times 3$	64
	`Add` (residual connection)	$32 \times 32 \times 3$	64
ResNet Block Conv 3 × 2	`Conv 3 × 3`	$32 \times 32 \times 3 \to 16 \times 16 \times 3$	$64 \to 128$
	`BatchNormal`	$16 \times 16 \times 3$	128
	`Conv 3 × 3`	$16 \times 16 \times 3$	128
	`BatchNormal`	$16 \times 16 \times 3$	128
	`Add`(residual connection)	$16 \times 16 \times 3$	128
ResNet Block Conv 4 × 2	`Conv 3 × 3`	$16 \times 16 \times 3 \to 8 \times 8 \times 3$	$128 \to 256$
	`BatchNormal`	$8 \times 8 \times 3$	256
	`Conv 3 × 3`	$8 \times 8 \times 3$	256
	`BatchNormal`	$8 \times 8 \times 3$	256
	`Add` (residual connection)	$8 \times 8 \times 3$	256
ResNet Block Conv 5 × 2	`Conv 3 × 3`	$8 \times 8 \times 3 \to 4 \times 4 \times 3$	$256 \to 512$
	`BatchNormal`	$4 \times 4 \times 3$	512
	`Conv 3 × 3`	$4 \times 4 \times 3$	512
	`BatchNormal`	$4 \times 4 \times 3$	512
	`Add`(residual connection)	$4 \times 4 \times 3$	512
Image Feature Transformation	`Global Average Pooling`	512	1
	`MLP`	64	1
Numerical Inputs	-	14 (3 + 3 + 3 + 2 + 3)	1
Numerical Feature Transformation	`MLP`	$14 \to 16$	1
	`MLP`	16	1
Concatenation	`Concat`	80 (64 + 16)	1
Regression Head	`MLP`	80	1
	`Gate MLP`	80	1
	`Gate Multiply`	80	1
	`MLP`	64	1
	`MLP`	16	1
	`MLP`	1	1

Table B.9: The details of CNN-L model.

Block	**Layer**	**Resolution**	**Channels**
Image Inputs	-	$128 \times 128 \times 3$	1
ResNet Block Conv 1	`Conv 7 × 7`	$128 \times 128 \times 3 \rightarrow 64 \times 64 \times 3$	$1 \rightarrow 64$
	`Max Pooling 3 × 3`	$64 \times 64 \times 3 \rightarrow 32 \times 32 \times 3$	64
ResNet Block Conv 2×2	`Conv 3 × 3`	$32 \times 32 \times 3$	64
	`BatchNormal`	$32 \times 32 \times 3$	64
	`Conv 3 × 3`	$32 \times 32 \times 3$	64
	`BatchNormal`	$32 \times 32 \times 3$	64
	`Add` (residual connection)	$32 \times 32 \times 3$	64
ResNet Block Conv 3×2	`Conv 3 × 3`	$32 \times 32 \times 3 \rightarrow 16 \times 16 \times 3$	$64 \rightarrow 128$
	`BatchNormal`	$16 \times 16 \times 3$	128
	`Conv 3 × 3`	$16 \times 16 \times 3$	128
	`BatchNormal`	$16 \times 16 \times 3$	128
	`Add` (residual connection)	$16 \times 16 \times 3$	128
ResNet Block Conv 4×2	`Conv 3 × 3`	$16 \times 16 \times 3 \rightarrow 8 \times 8 \times 3$	$128 \rightarrow 256$
	`BatchNormal`	$8 \times 8 \times 3$	256
	`Conv 3 × 3`	$8 \times 8 \times 3$	256
	`BatchNormal`	$8 \times 8 \times 3$	256
	`Add` (residual connection)	$8 \times 8 \times 3$	256
ResNet Block Conv 5×2	`Conv 3 × 3`	$8 \times 8 \times 3 \rightarrow 4 \times 4 \times 3$	$256 \rightarrow 512$
	`BatchNormal`	$4 \times 4 \times 3$	512
	`Conv 3 × 3`	$4 \times 4 \times 3$	512
	`BatchNormal`	$4 \times 4 \times 3$	512
	`Add` (residual connection)	$4 \times 4 \times 3$	512
Image Feature Transformation	`Global Average Pooling`	512	1
	`MLP`	64	1
Numerical Inputs	-	$14\ (3 + 3 + 3 + 2 + 3)$	1
Numerical Feature Transformation	`MLP`	$14 \rightarrow 16$	1
	`MLP`	16	1
Concatenation	`Concat`	$80\ (64 + 16)$	1
Regression Head	`MLP`	80	1
	`MLP`	64	1
	`MLP`	16	1
	`MLP`	1	1

Highlights

Highlights

- A novel deep learning framework for solar irradiance forecasting.
- The model is better able to forecast upcoming critical events.
- Model creates cross-modal ties between images and meteorological attributes.
- Visualizing the model inference process through intra-model weights

Liwenbo Zhang: Conceptualization, Methodology, Software, Formal analysis, Investigation, Writing – original draft, Writing — review & editing, Visualization.

Robin Wilson: Conceptualization, Resources, Writing – review & editing, Supervision, Project administration

Mark Sumner: Conceptualization, Resources, Writing – review & editing, Supervision, Project administration, Funding acquisition.

Yupeng Wu: Conceptualization, Resources, Writing – review & editing, Supervision, Project administration, Funding acquisition.

Declaration of interests

☒ The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

☐The authors declare the following financial interests/personal relationships which may be considered as potential competing interests: