Supplementary Materials: Solar Multimodal Transformer: Intraday Solar Irradiance Predictor using Public Cameras and Time Series

Yanan Niu[1], Roy Sarkis[1], Demetri Psaltis[1], Mario Paolone[1], Christophe Moser[1], Luisa Lambertini[1, 2]
[1]EPFL, Lausanne, 1015, Vaud, Switzerland
[2]Universita' della Svizzera Italiana (USI), Lugano, 6900, Ticino, Switzerland
yanan.niu@epfl.ch

1. Additional implementation details

1.1. Dataset details

The public camera images are downloaded in real-time. Due to webcam settings, a panorama is created every ten minutes as the camera completes a 360° rotation, which normally takes around 1 or 2 minutes. The camera operates during daylight hours; however, frames are sometimes missing or unusable due to misoperation or optical distortion, resulting in gaps in our dataset. For the five datasets used in this study, besides the urban open space dataset which consists of high-resolution RGB frames ($3 \times 2048 \times 10752$), the rest are of lower resolution ($3 \times 171 \times 900$). Examples are shown in Fig. 1 and Fig. 2. All of the panoramas are then resized to $3 \times 224 \times 224$ for the forecasting task. For privacy purposes, all individuals appearing in the images are blurred using OpenCV's YOLO v3. The blurring of individuals does not affect the results, as the proportion of blurred sections is trivial in the overall image, according to our experiments.

The corresponding GHI measurements for datasets "urban streetscape," "valley," "lake," and "mountain" are collected from nearby weather stations operated by MeteoSwiss. For the "urban open space" dataset, the measurements are collected using an SP-2306 all-seasons pyranometer from Apogee[1], located 125 meters away from the camera; these measurements are assumed to be unbiased and noiseless. The device continuously collects data, and measurements are averaged into ten-minute intervals to align with the granularity of the camera frames. According to the clear sky model, the GHI values during the night should theoretically be zero. However, measurements can be slightly above zero due to environmental influences. We correct the nighttime values to zero to mitigate irrelevant bias. Additionally, the highest daily GHI measurement, $y^{\text{max}}_{\text{day}(t)}$, sometimes exceeds the daily maximum clear sky GHI value, $C^{\text{max clear sky}}_{\text{day}(t)}$. This indicates that the scaled GHI according to our GHI normalization step is not strictly between 0 and 1 and can occasionally be higher than 1.

1.2. Pretrained models

The pre-train-and-fine-tune approach has proven effective in fields such as CV, NLP, and multimodal tasks, providing a robust initialization or even enabling good zero-shot performance in downstream applications [1]. However, we opt to train our models from scratch due to the unique nature of our task. Unlike downstream tasks such as visual question answering or image-text retrieval where desired features are determined by the objects in the images, the relevant features for solar forecasting remain undefined. Furthermore, the static nature of public camera images, characterized by subtle changes in optical flow such as clouds, the Sun, and shadow movements, further limits the added value of pretrained models for our task, see the attention analysis in the paper.

1.3. Implementation details of models

LSTNet. Code is adopted from the GitHub repository of Lai *et al.*[2]. The default settings of parameters are applied, using the past 24 hours of historical GHI to predict the value 2 hours into the future. The highway lookback window is set to 2 hours.

Hybrid SMT. For both CNN + SMT and U-net + SMT, the latent spaces of CNN or U-net are flattened and projected into a series of vectors before being fed into the transformer module of SMT. Specifically, for both models, the latent spaces are dimensioned at $C_{latent} \times H_{latent} \times W_{latent}$. To transition from feature maps to vectors, the latent space of each image is flattened to $C_{latent} \times (H_{latent} \times W_{latent})$, organized into column patches of $1 \times C_{latent} \times (H_{latent} \times W_{latent})$, and finally projected to $C_{latent} \times D$,

[1] https://www.apogeeinstruments.com/

[2] https://github.com/laiguokun/LSTNet

Figure 1. An example of the original panorama, dataset "urban open space".

Figure 2. Examples from the datasets "urban streetscape", "valley", "lake", and "mountain", from top to bottom.

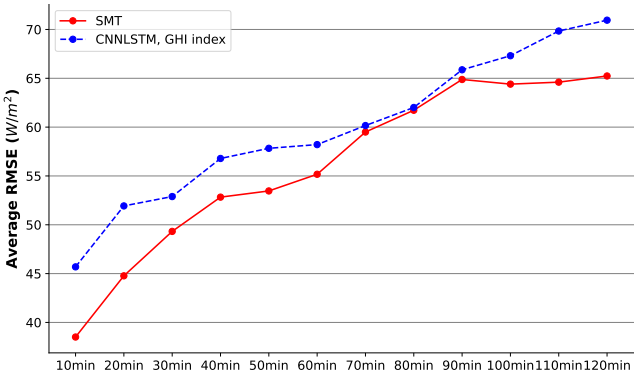

Figure 3. Comparison of different forecasting horizons: SMT *vs.* CNNLSTM

where D represents the consistent embedding size across all transformer encoder layers (192 in this instance).

2. Additional results

2.1. Different forecasting horizons

In this paper, we report a forecasting horizon of 2 hours; however, other horizons are expected to yield similar results in model comparisons. Shorter horizons may achieve higher prediction accuracy due to reduced uncertainty. Conversely, as the forecasting horizon increases, the corresponding RMSE also increases, as expected. This trend underscores the challenges associated with longer-term forecasts. A comparison between the SMT, which uses historical GHI data from the past 24 hours, and the CNNLSTM model, which utilizes a single frame and scaled GHI, is presented in Fig. 3.

2.2. Extra examples of attention analysis

More visualizations of the attention analysis on dataset "urban open space" are presented in Fig. 4. The first row of examples demonstrates a scenario where the Sun is clearly visible. Both the direction of the Sun and its $180°$ opposite direction can always be accurately tracked, and the last attention panel shows a specific focus returning to the image itself, as explained in the paper. This indicates that data fusion predominantly occurs in the shallow blocks of the transformer. Even in cases where the Sun remains distinguishable despite the presence of heavy clouds, as exemplified in the second row, it can still be accurately tracked.

Regarding the experiments using row patches, as explained in the paper, the pixels from the ground remain important, as seen in all the cases from Fig. 5. The GHI time series data consistently plays a crucial role in prediction outcomes. The row of pixels where the Sun is located can be precisely captured as long as it is identifiable in the image, even in the presence of clouds. Furthermore, the upper and lower rows tend to be more relevant across examples, as those pixels are closer to the camera's location compared to the middle part of the panorama and better represent localized weather conditions.

References

[1] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. Learning transferable visual models from natural language supervision, 2 2021. 1

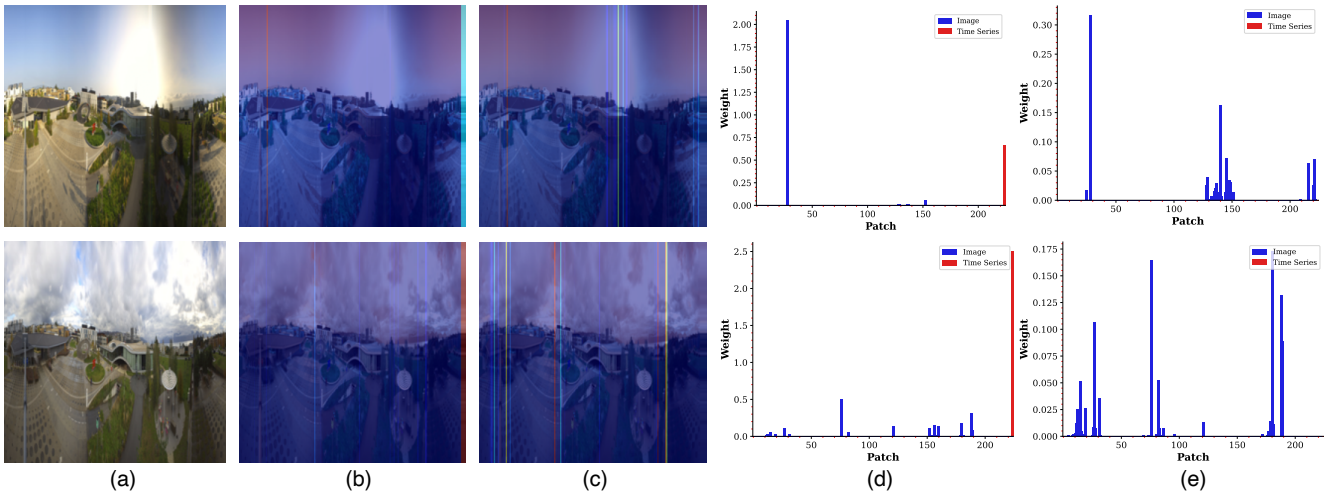

Figure 4. Attention analysis using column patches for SMT (cont.)

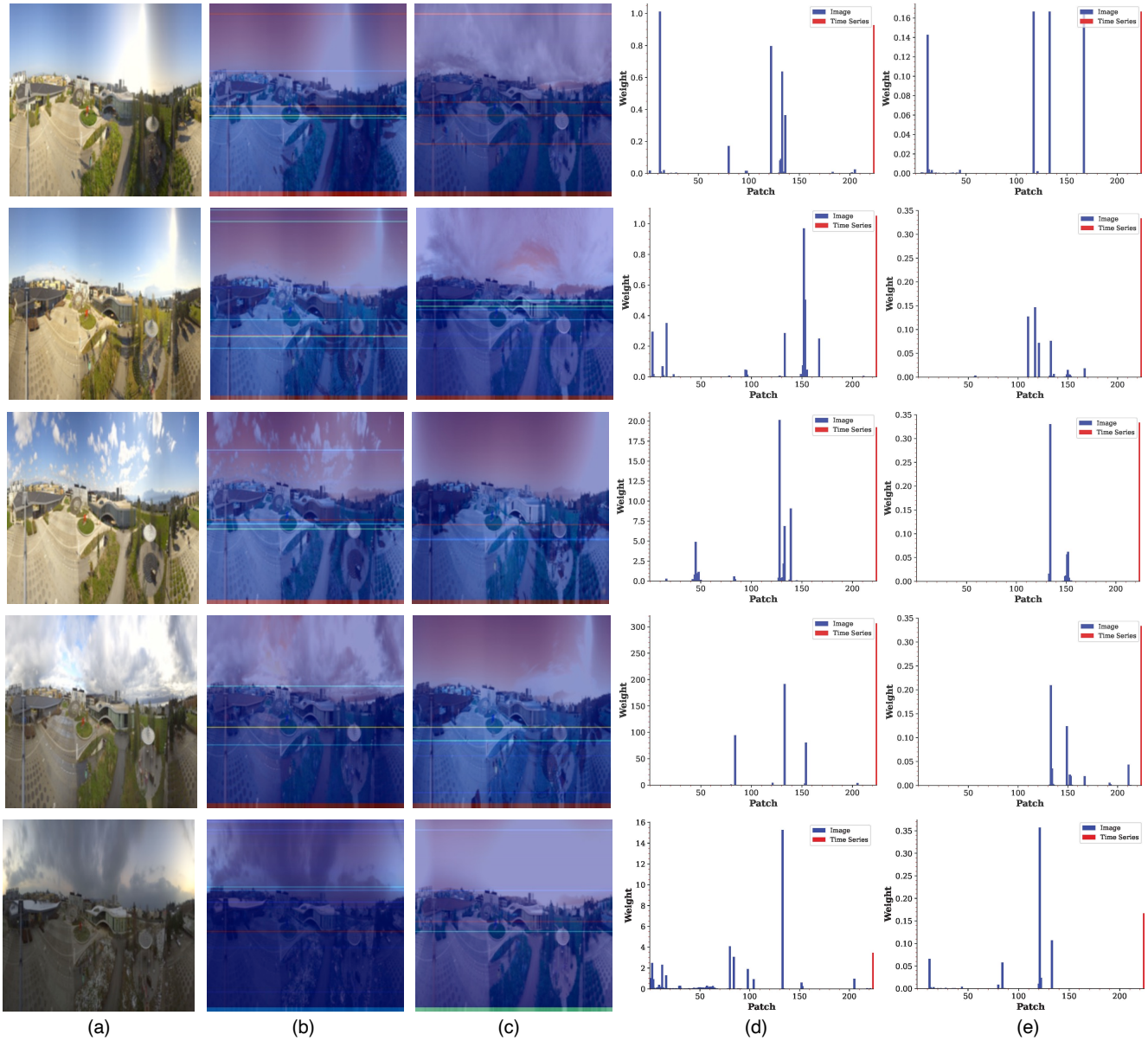

Figure 5. Attention analysis using row patches for SMT: Similarly, the last row of pixels in panel (b,c) indicating the relative importance of time series, same as the last vector in panel (d,e).