Supplementary File for High-Fidelity Full-Sky Video Prediction for Photovoltaic Ramp Event Forecasting

Siyuan Wang, *Member, IEEE*, Fengqi You, *Senior Member, IEEE*

TABLE OF CONTENTS

I. DETAILED HYPERPARAMETERS FOR MODEL TRAINING

A. Hyperparameters settings of PhyDNet module

TABLE I
HYPERPARAMETERS OF PHYDNET

Hyper-parameter	Value / Setting
Batch size	32
Learning rate	1e-4
Optimizer	Adam
Number of epochs	300
LR scheduler	ReduceLROnPlateau
Scheduler patience	3 epochs
Scheduler factor	0.3
Nonlinearity	LeakyReLU (0.2)
Normalization	GroupNorm

B. *Hyperparameters settings of video conditional diffusion module*

TABLE II
HYPERPARAMETERS OF VIDEO CONDITIONAL DIFFUSION

Hyper-parameter	Value / Setting
Training steps	100000
Batch size	4
Trainer	AdamW
Learning rate	2e-4
Diffusion timesteps	1000
Beta schedule	Cosine
Base channel dim	32
Input convolution kernel	7×7×7
Sinusoidal embedding dim	32
Number of resolutions	4
Mid-block attention heads	4
Cross-attention heads	4
ResNetBlock3D norm	RMSNorm
Activation	SiLU

C. *Hyperparameters settings of RaPVFormer module*

TABLE III
HYPERPARAMETERS OF RAPVFORMER

Hyper-parameter	Value / Setting
Image encoder output dim	128
Transformer model dim	256
Transformer heads	4
PV embedding dim	16
Activation functions	SiLU
Transformer feedforward dim	512
Transformer dropout	0.1
Optimizer	Adam
Learning rate	1e-4
ω_P in loss function	1.0
ω_S in loss function	0.2
ω_R in loss function	0.5
LR Scheduler	ReduceLROnPlateau
Scheduler patience	3 epochs
Scheduler factor	0.3

II. METRICS FOR PREDICTED FULL-SKY VIDEO FRAMES EVALUATION

A. Peak Signal-to-Noise Ratio (PSNR)

PSNR [1] is a traditional full-reference image quality metric that measures the pixel-wise similarity between the predicted frame $\hat{I}_t$ and the ground truth frame I_t. It is derived from the Mean Squared Error (MSE) between two images.

$$\mathrm{PSNR}\left(I_t,\hat{I}_t\right) := 10\cdot\log_{10}\left(\frac{L^2}{\mathrm{MSE}\left(I_t,\hat{I}_t\right)}\right) \tag{1}$$

where L is the maximum pixel value and equals 255 for 8-bit images, and the MSE is defined as:

$$\mathrm{MSE}\left(I_t,\hat{I}_t\right) := \frac{1}{H\times W\times C}\sum_{i=1}^{H\times W\times C}\left(\hat{I}_t^{(i)}-I_t^{(i)}\right)^2 \tag{2}$$

where H, W, C represent the height, width and number of channels of the images, respectively. $I_t^{(i)}$ and $\hat{I}_t^{(i)}$ are the pixel values at position i in the ground-truth and predicted images, respectively.

A higher PSNR value indicates better image reconstruction quality.

B. Structural Similarity Index (SSIM)

SSIM [2] measures the similarity by comparing their luminance, contrast, and structural information in local image patches. Unlike PSNR, which relies purely on pixel-wise error, SSIM is more consistent with human visual perception. The SSIM is computed as:

$$\mathrm{SSIM}\left(I_t,\hat{I}_t\right) := \frac{(2\mu_{I_t}\mu_{\hat{I}_t}+C_1)(2\sigma_{I_t\hat{I}_t}+C_2)}{(\mu_{I_t}^2+\mu_{\hat{I}_t}^2+C_1)(\sigma_{I_t}^2+\sigma_{\hat{I}_t}^2+C_2)} \tag{3}$$

where μ_{I_t}, $\mu_{\hat{I}_t}$ are the mean pixel values; $\sigma_{I_t}^2$, $\sigma_{\hat{I}_t}^2$ are the variances; $C_1=(K_1L)^2$ and $C_2=(K_2L)^2$ are constants for stability; L is the maximum pixel value and equals 255 for 8-bit frames; $K_1=0.01$ and $K_2=0.03$ are default constants.

Higher SSIM values (closer to 1) indicate better structural similarity between the predicted and the ground-truth frames.

C. Learned Perceptual Image Patch Similarity (LPIPS)

LPIPS [3] measures the perceptual similarity by comparing deep feature representations extracted from a pretrained convolutional neural network such as VGG. Different from PSNR and SSIM, which operate in the pixel or structural domain, LPIPS reflects human perceptual judgments more accurately. LPIPS is defined as:

$$\mathrm{LPIPS}\left(I_t,\hat{I}_t\right) := \sum_l \frac{1}{H_lW_l}\sum_{h=1}^{H_l}\sum_{w=1}^{W_l}\left\| w_l\odot\left(\boldsymbol{f}_l\left(\hat{I}_t\right)_{hw}-\boldsymbol{f}_l\left(I_t\right)_{hw}\right)\right\|_2^2 \tag{4}$$

where $\boldsymbol{f}_l(\bullet)$ denotes the feature map at the l-th layer of the network; H_l and W_l are the height and width of the feature map at layer l; w_l is a learned weight tensor used to adjust the channel importance; $\odot$ denotes element-wise multiplication.

A lower LPIPS score indicates higher perceptual similarity. A perfect LPIPS score of 0 means the predicted and ground-truth frames are perceptually identical in the chosen feature space.

D. VGG-based Cosine Similarity (VGGCS)

VGGCS [4] is a semantic similarity metric that evaluates how close the predicted frame is to the ground-truth frame in the deep feature space of a pretrained VGG network. It captures high-level semantic content such as object shapes and scene layouts.

The cosine similarity is computed as:

$$\mathrm{VGGCS}\left(I_t, \hat{I}_t\right) := \frac{\boldsymbol{f}\left(I_t\right) \cdot \boldsymbol{f}\left(\hat{I}_t\right)}{\left\|\boldsymbol{f}\left(I_t\right)\right\|_2 \cdot \left\|\boldsymbol{f}\left(\hat{I}_t\right)\right\|_2} \tag{5}$$

where $\boldsymbol{f}(\bullet)$ is flattened feature vector extracted from a specific VGG layer.

A higher VGGCS value (closer to 1) indicates greater semantic similarity between the predicted and ground-truth frames.

E. Temporal Optical Flow Consistency (TOF)

TOF [5] evaluates the temporal motion consistency of predicted video frames by comparing optical flows between adjacent frames in the predicted sequence and the corresponding flows in the ground-truth sequence. It is defined as:

$$\mathrm{TOF} := \frac{1}{T-1} \sum_{t=1}^{T-1} \left\| \phi\left(\hat{I}_{t+1}, \hat{I}_t\right) - \phi\left(I_{t+1}, I_t\right) \right\|_2 \tag{6}$$

where $\phi(A, B)$ denotes the optical flow from frame B to A, estimated using a pretrained flow network, such as Recurrent All-Pairs Field Transforms (RAFT) [5].

A lower TOF value indicates more accurate and temporally consistent motion patterns.

F. Temporal Feature Change Distance (TFCD)

TFCD measures how similarly the visual features evolve over time in both sequences. It leverages high-level features extracted from a pretrained network, such as VGG, to assess how well the semantic content is preserved over time. It is calculated as:

$$\mathrm{TFCD} := \frac{1}{T-1} \sum_{t=1}^{T-1} \left\| \left(\boldsymbol{f}(\hat{I}_{t+1}) - \boldsymbol{f}(\hat{I}_t) \right) - \left(\boldsymbol{f}(I_{t+1}) - \boldsymbol{f}(I_t) \right) \right\|_2 \tag{7}$$

where $\boldsymbol{f}(\bullet)$ is flattened feature vector extracted from a specific VGG layer.

A lower TFCD value indicates that the generated video exhibits more temporally consistent and realistic motion relative to the real reference.

III. Dataset Availability

The sky images and photovoltaic power generation dataset (SKIPP'D) [6] across 2017 to 2019 for short-term solar forecasting are available at TABLE I.

TABLE IV
DATA SOURCE OF DIFFERENT YEARS

Year	Data source
2017	https://purl.stanford.edu/sm043zf7254
2018	https://purl.stanford.edu/fb002mq9407
2019	https://purl.stanford.edu/jj716hx9049

REFERENCES

[1] R. C. Gonzalez and R. E. Woods, *Digital image processing*, 2nd ed. Prentice Hall, 2002.

[2] Z. Wang, A. C. Bovik, H. R. Sheikh, and E. P. Simoncelli, "Image quality assessment: From error visibility to structural similarity," *IEEE Trans. Image Process.*, vol. 13, no. 4, pp. 600–612, 2004, doi: 10.1109/TIP.2003.819861.

[3] R. Zhang, P. Isola, A. A. Efros, E. Shechtman, and O. Wang, "The unreasonable effectiveness of deep features as a perceptual metric," in *Proceedings of the IEEE conference on computer vision and pattern recognition (CVPR)*, 2018, pp. 586–595.

[4] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," in *Proceedings of the international conference on learning representations (ICLR)*, 2015.

[5] Z. Teed and J. Deng, "RAFT: Recurrent all-pairs field transforms for optical flow," in *European conference on computer vision (ECCV)*, Springer, 2020, pp. 402–419.

[6] Y. Nie, X. Li, A. Scott, Y. Sun, V. Venugopal, and A. Brandt, "SKIPP'D: A SKy Images and Photovoltaic Power Generation Dataset for short-term solar forecasting," *Sol. Energy*, vol. 255, pp. 171–179, May 2023, doi: 10.1016/j.solener.2023.03.043.