Efficient Forecasting of Geostationary Infrared Brightness Temperature Sequences: A Benchmark and a Lightweight Model

Kuai Dai[1], **Hui Su**[1*], **Xutao Li**[2], **Chengxing Zhai**[1]

[1]The Hong Kong University of Science and Technology
[2]Harbin Institute of Technology, Shenzhen
daikuai_hit@163.com, cehsu@ust.hk, lixutao@hit.edu.cn, cxzhai@ust.hk

Abstract

Forecasting geostationary infrared brightness temperature sequences from historical observations is a significant and challenging task. By analyzing these predictions, cloud evolution, convective activity, and atmospheric radiative states can be revealed in advance, offering high potential value in domains such as weather nowcasting, energy management, and disaster monitoring. Recently, artificial intelligence techniques have provided valuable insights into this task. However, as a nascent research area, the lack of a standardized, high-quality benchmark has significantly impeded progress. Moreover, training existing deep learning models for this task remains computationally expensive due to the complexity of their network architectures and modeling mechanisms. To address these challenges, we introduce a new benchmark, FY4ABT, and propose a lightweight prediction model, WavePredNet. Specifically, FY4ABT comprises three sub-datasets designed to respectively evaluate prediction performance under short-term, medium-term, and long-term scenarios. Meanwhile, WavePredNet effectively captures multi-scale dynamics, including both low- and high-frequency components with low computational costs while delivering exceptional performance.

Code — https://github.com/Applied-IAS/WavePredNet
Datasets — https://zenodo.org/records/17577328

Introduction

Geostationary satellite infrared brightness temperature values reflect the radiative temperature of cloud tops and Earth's surface. Compared with traditional radar and automatic station observations (Turner, Zawadzki, and Germann 2004; Ravuri et al. 2021; Zhang et al. 2023), the satellite brightness temperature data offer broad and continuous coverage, making them vital for monitoring atmospheric convection, cloud evolution, and surface thermal conditions in data-sparse regions such as developing countries. Accurate forecasting of brightness temperature sequence provides valuable information for weather (Lebedev et al. 2019; Leinonen et al. 2023), energy (Hatanaka et al. 2023; Xia et al. 2024), and agriculture (Tarasiou, Chavez, and Zafeiriou 2023; Benson et al. 2024; Garnot and Landrieu 2021) applications. As a result, this task has attracted increasing attention in recent years (Shukla, Kishtawal, and Pal 2013; Lee et al. 2019; Xu et al. 2019).

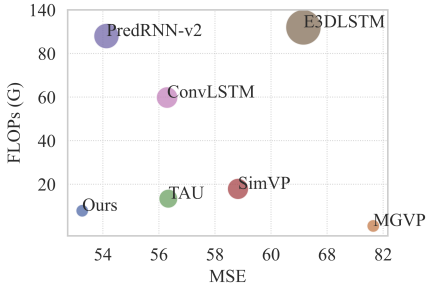

Figure 1: The FLOPs and MSE scores of our WavePredNet and state-of-the-art AI prediction models for 4-hour brightness temperature sequence forecasting results. Notably, the point size denotes the model's parameter scale.

Existing brightness temperature sequence forecasting approaches can be broadly classified into two main taxonomies, namely, advection-based methods (Bowler, Pierce, and Seed 2006; Zach, Pock, and Bischof 2007; Pulkkinen et al. 2019) and deep generative methods (Shi et al. 2015; Wang et al. 2017; Guen and Thome 2020). Advection-based methods such as STEPS (Bowler, Pierce, and Seed 2006) and pySTEPs (Pulkkinen et al. 2019) employ physical advection equations to model the motion field of two consecutive historical frames, and then warp the estimated motion field with the latest input frame to produce the next predicted frame. These approaches are relatively simple to implement and require minimal computational resources, making them widely deployed in practical scenarios. However, the methods tend to produce inaccurate prediction results due to idealized model assumptions that are easily disrupted under complex atmospheric conditions. In contrast to advection-based methods, deep generative methods can effectively leverage large amounts of training data with advanced neural networks (Shi et al. 2015; Wang et al. 2019; Gao et al. 2022; Tan et al. 2023a) guided by deterministic loss functions such as mean squared error (MSE) and mean absolute error (MAE). This data-driven learning manner enables statistically optimal results, and better models

*The corresponding author.

spatial features and temporal variations of brightness temperature sequences, resulting in more accurate prediction results, especially in long-term conditions. Hence, deep generative methods have become the mainstream approach for brightness temperature sequence forecasting.

Though previous works have made notable progress in brightness temperature sequence forecasting, challenges persist in this task, which can be summarized into two main aspects. On the one hand, a high-quality and standardized dataset is urgently needed to comprehensively evaluate existing prediction models. The lack of such a benchmark significantly hampers progress in the field. On the other hand, training existing deep generative prediction models for brightness temperature sequence forecasting is computationally expensive, due to the complicated network architecture and the high-dimensional nature of sequence data. The high computational costs are not economic and severely limit the feasibility of lightweight deployments, such as space-borne operations. Furthermore, brightness temperature sequences contain rich appearance features and complex motion patterns, placing high demands on the spatiotemporal modeling capabilities of deep learning prediction models. As illustrated in Figure 1, our lightweight model, WavePredNet, achieves the best performance with significantly lower FLOPs and fewer parameters. While one exhibits even lower computational cost, its forecasting accuracy is significantly inferior to our method.

In this work, we aim to address the two aforementioned problems by presenting a standard dataset and an efficient prediction model. As for the new benchmark, we first collect and process the data of 2019 & 2020 from the Advanced Geostationary Radiation Imager (AGRI) in the Chinese FengYun-4A geostationary satellite, and then extract the long-wave infrared channel (10.7 μm waveband) to construct the FY4ABT series datasets. In particular, it contains three sub-datasets for 1-hour (short-term), 2-hour (medium-term), and 4-hour (long-term) brightness temperature prediction. Notably, each frame of the sequence is 128×128-size, each pixel denotes an area of 16×16 square kilometers, and the interval between two consecutive frames is 1 hour. With the evaluation benchmark, we can comprehensively validate the prediction performance of models under short-term, medium-term, and long-term prediction requests. In terms of the methodology, we carefully design and present a lightweight predictive framework WavePredNet, which can efficiently model the complex temporal variations of brightness temperature sequences by innovatively introducing Wavelet Transform techniques. Specifically, WavePredNet consists of a spatial encoder, a Temporally-Cascaded Wavelet Transform Block (TCWTB), and a spatial decoder. The spatial encoder and decoder account for modeling the appearance features of each frame, while the TCWTB aims to accurately capture the temporal patterns of sequences. With the specially designed temporally-cascaded architecture, the TCWTB can finely capture temporal evolution patterns in a progressive manner. In addition, as a key designed structure of the TCWTB, the MultiScale Dual-Branch WaveConv (MSDBWC) structure can model low- and high-frequency elements at various scale levels with low computation cost, which efficiently preserves multi-grained dynamics. Comprehensive experiments conducted on the FY4ABT benchmark show that, compared with state-of-the-art models, our model WavePredNet consistently delivers the best performance across the short-term, medium-term, and long-term prediction, with low computation costs. In addition, corresponding ablation studies show the effectiveness of key components and structures in WavePredNet.

Overall, the contributions of this paper can be summarized as follows:

- To the best of our knowledge, we present the first standardized evaluation benchmark, FY4ABT, for brightness temperature sequence forecasting.
- The proposed temporally-cascaded wavelet transform block and multiscale dual-branch waveconv structure effectively model complicated temporal patterns of brightness temperature sequences, with low computation costs.
- Extensive experiments across short-term, medium-term, and long-term conditions demonstrate the superiority of WavePredNet over state-of-the-art prediction models in both performance and efficiency.

Related Work

Brightness Temperature Sequence Forecasting

Dataset Challenge. Recent studies (Xu et al. 2019; Dai et al. 2022, 2023) have explored short-term forecasting using satellite cloud imagery, particularly for convective weather scenarios. The approaches deliver impressive prediction results in cloud appearance and motion, yet they focus on perceptual quality rather than overall physical accuracy. In contrast, forecasting brightness temperature sequences, especially in the long-wave infrared band, offers broader applications such as radiative transfer modeling, surface temperature analysis, and disaster detection. Unlike cloud prediction, brightness temperature forecasting emphasizes the accurate prediction of continuous and physically meaningful spatial fields. However, the existing related works (Hartman et al. 2021; Jiang et al. 2022) rely on proprietary or closed datasets, which impede progress in the field. To address this gap, we present FY4ABT, a benchmark dataset constructed from the FengYun-4A geostationary satellite, which can comprehensively validate and analyze the forecasting models.

SpatioTemporal Forecasting Models. In the past few years, AI-based spatiotemporal techniques have provided valuable insights for brightness temperature sequence forecasting. According to the forecasting architecture, the models are categorized into two groups, namely, recurrent-based models (Shi et al. 2015; Wang et al. 2017, 2022) and recurrent-free models (Gao et al. 2022; Tan et al. 2023a). First, among the recurrent-based models, PredRNN (Wang et al. 2017) designs a unified memory pool to simultaneously memorize the spatial features and temporal variations. Based on PredRNN, its enhanced version PreRNN++ (Wang et al. 2018a) proposes a causal LSTM module and gradient highway unit to strengthen the short-term dynamics modeling and alleviate the vanishing gradient issue, respectively. SA-ConvLSTM (Lin et al. 2020) introduces the self-attention

mechanism to learn global spatiotemporal dynamics. PhyDNet (Guen and Thome 2020) proposes to disentangle spatiotemporal dynamics into known physical dynamics and unknown factors, and designs a special recurrent cell to learn physical dynamics. Based on PredRNN and PredRNN++, PredRNN-V2 (Wang et al. 2022) further presents a memory decoupling scheme and a reverse scheduled sampling strategy to effectively model spatiotemporal patterns. Second, for the recurrent-free methods, SimVP (Gao et al. 2022) designs an inception-unet module to model multi-scale temporal patterns. Inspired by this, TAU (Tan et al. 2023a) presents a temporal attention module, decomposed into intra-frame statical attention and the inter-frame dynamical attention, to effectively model long-term temporal evolution.

Overall, recurrent-based models can accurately capture temporal patterns but come with high computational costs, while recurrent-free models are more efficient but provide relatively coarse temporal modeling. Therefore, balancing computational cost and predictive accuracy for brightness temperature sequence forecasting remains a significant challenge. In this work, we aim to develop an efficient and powerful prediction model to address this issue.

Wavelet Transform

Wavelet Transform (WT) (Daubechies 1992) has wide applications in the computer vision domain, such as image understanding, segmentation, super-resolution, and compression (Zhao et al. 2024). Compared with CNN, WT can effectively deal with time-frequency analysis and has advantages in capturing non-stationary signals (Wang et al. 2021). As a common paradigm, WT operation is usually incorporated with CNN to enlarge the receptive field and preserve appearance details (Finder et al. 2024). Specifically, (Liu et al. 2018) presents a multi-level wavelet-CNN (MWCNN) architecture for image restoration, which can effectively enlarge the receptive field with low computation cost. (Alaba and Ball 2022) designs a wavelet-based 3D object detection mode (WCNN3D), which shows the advantages of WT in lightweight and small object detection tasks. (Yao et al. 2022) integrates WT and self-attention within a ViT architecture, effectively mitigating information loss caused by downsampling operations prior to the self-attention layers. Similarly, (Xu et al. 2023) presents a Harr wavelet-based downsampling (HWD) module to replace the conventional downsampling operations, which significantly promotes the prediction accuracy in image segmentation.

Though the previous works demonstrate the effectiveness of wavelet transform in modeling images, the potential of WT in modeling spatiotemporal sequences, especially in brightness temperature sequence forecasting, has yet to be explored. In this work, we introduce the wavelet transform to model the complex temporal patterns of brightness temperature sequences, which can gain performance improvement with fewer computational costs.

Problem Definition

Formally, brightness temperature sequence forecasting can be regarded as a special case of spatiotemporal sequence predictive learning. Given the past t-length brightness temperature sequence $X^{1,t} = \{x_1, x_2, \cdots, x_t\} \in \mathbb{R}^{t\times c\times m\times n}$, we aim to accurately predict the next k-frame sequence $\hat{Y}^{t+1,k} = \{\hat{x}_{t+1}, \hat{x}_{t+2}, \cdots, \hat{x}_{t+k}\} \in \mathbb{R}^{k\times m\times n}$, which should be as similar to the ground-truth sequence $Y^{t+1,k} = \{x_{t+1}, x_{t+2}, \cdots, x_{t+k}\} \in \mathbb{R}^{k\times m\times n}$ as possible. c, m, and n denote the channel, width, and length of the brightness temperature frame at each time step, respectively.

The aforementioned mapping procedure $X^{1,t} \rightarrow Y^{t+1,k}$ can be learned with a deep learning prediction model P with parameters θ by modeling spatial features and temporal variations of brightness temperature sequences. The optimal parameters θ^* is obtained through the mini-batch gradient descent algorithm, which is defined as follows:

$$\theta^* = \arg\min_{\theta} \mathcal{L}(P_\theta(X^{1,t}), Y^{t+1,k}), \tag{1}$$

$\mathcal{L}(\cdot)$ is the loss function that guides the training process of the prediction model P.

Methodology

Overall Layout

Figure 2 (a) illustrates the overall framework of WavePredNet, which comprises a spatial encoder, a Temporally-Cascaded Wavelet Transform Block (TCWTB), and a spatial decoder. First, the spatial encoder extracts spatial features from each frame in the input sequence. Next, the encoded features are concatenated along the channel dimension before being passed into the temporally-cascaded wavelet transform block, which can effectively model the complex temporal variations of brightness temperature sequences, with its temporally-cascaded architecture and a specially designed module, MultiScale Dual-Branch WaveConv (MSDBWC). Finally, the hidden states generated by the TCWTB are reshaped along the batch dimension, and the spatial decoder is responsible for decoding these features into the predicted frames.

The spatial encoder contains four convolution layers to model spatial features. Notably, the second and fourth convolution layers with strides of 2 are employed to aggregate the spatial features and simultaneously reduce the computation cost. Correspondingly, the spatial decoder comprises four deconvolution layers to decode the spatial features into predicted frames, with the second and fourth layers configured with a stride of 2. In addition, spatial features produced by the first convolution layer in the spatial encoder are passed to the final convolution layer in the spatial decoder via a skip connection, to better preserve the appearance details. Next, we introduce the temporally-cascaded wavelet transform block in detail.

Temporally-Cascaded Wavelet Transform Block Given the two consecutive input frames $\{x_{t-1}, x_t\}$, the spatial encoder transforms the frames into feature maps $\{F_{t-1} \in \mathbb{R}^{C\times H\times W}, F_t \in \mathbb{R}^{C\times H\times W}\}$. Concatenate the features along the channel dimension, and we can get the input elements $F^{t-1,t} \in \mathbb{R}^{(2\times C)\times H\times W}$ for the temporally-cascaded wavelet transform block (TCWTB).

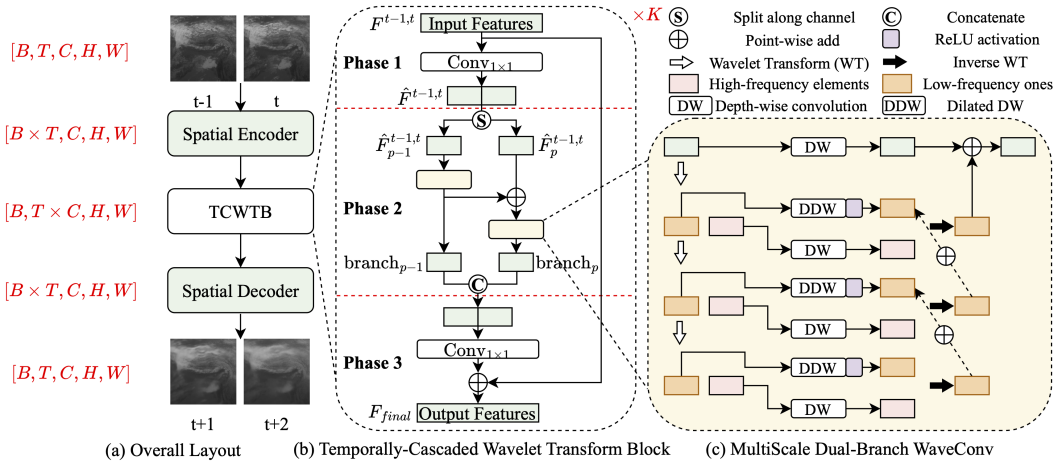

Figure 2: (a) The overall layout of WavePredNet. The red font highlights the tensor dimension transformation process, where B, T, C, H, and W denote the batch, temporal, channel, height, and width dimensions, respectively. (b) The structure of the temporally-cascaded wavelet transform block. For simplicity, only the two consecutive branches are presented here. (c) The architecture of the multiscale dual-branch waveconv.

As shown in Figure 2 (b), the TCWTB contains three phases in total. First, the concatenated feature $F^{t-1,t}$ is fed into a combination of point-wise convolution layer and activation layers to reduce channel size and obtain enhanced features $\hat{F}^{t-1,t} = \mathrm{ReLU}(\mathrm{BN}(\mathrm{Conv}_{1\times1}(\hat{F}))) \in \mathbb{R}^{\hat{C}\times H\times W}$. The $\mathrm{ReLU}(\cdot)$ and $\mathrm{BN}(\cdot)$ denote the activation function and batch normalization layer, respectively. The $\hat{C}$ represents the reduced channel size.

Second, $\hat{F}^{t-1,t}$ are divided into p equal parts $\{\hat{F}_1^{t-1,t} \in \mathbb{R}^{\frac{\hat{C}}{p}\times H\times W}, \hat{F}_2^{t-1,t} \in \mathbb{R}^{\frac{\hat{C}}{p}\times H\times W}, \cdots, \hat{F}_p^{t-1,t} \in \mathbb{R}^{\frac{\hat{C}}{p}\times H\times W}\}$, along the channel dimension. The TCWTB then adopts a temporally-cascaded architecture to capture the evolution patterns of these split features. It consists of p branches, each dedicated to modeling its respective split part, with the output of one branch passed to the next through a point-wise addition operation. This design enables the block to efficiently model temporal patterns in a progressive manner. Compared with directly modeling temporal variations across the entire feature maps through a convolution layer, the new structure can more finely model the temporal patterns. Moreover, each branch incorporates a specially designed convolutional structure, the multiscale dual-branch waveconv, which leverages the wavelet transform to effectively capture low- and high-frequency temporal patterns at different levels. Similarly, this new structure further promotes the performance and efficiency in modeling temporal patterns. In particular, the modeling procedure of the $(p-1)$-th and p-th branches is formally defined as follows:

$$\mathrm{branch}_{p-1} = \mathrm{MSDBWC}(\hat{F}_{p-1}^{t-1,t}), \quad (2)$$

$$\mathrm{branch}_{p} = \mathrm{MSDBWC}(\mathrm{branch}_{p-1} \oplus \hat{F}_{p}^{t-1,t}). \quad (3)$$

The $\mathrm{MSDBWC}(\cdot)$ denotes the designed multiscale dual-branch waveconv operation, where the $\oplus$ denotes the point-wise add operation, and branch_{p-1} represents the output of $(p-1)$-th branch.

Third, the output results of previous branches are concatenated and fed into a point-wise convolution layer, to obtain entire temporal variations $F_{entire} = \mathrm{BN}(\mathrm{ReLU}(\mathrm{Conv}_{1\times1}(\mathrm{branch}_{p-1} \text{ concat } \mathrm{branch}_p))) \in \mathbb{R}^{2\times C\times H\times W}$. To further enhance the modeling of temporal dynamics, a skip connection is employed, enabling the construction of a deeper temporal modeling network. Hence, the final output of one TCWTB is computed as $F_{final} = \mathrm{ReLU}(F_{entire} \oplus F^{t-1,t})$. In addition, by stacking the K TCWTB modules, we can further enhance the temporal modeling ability. Next, we elaborate on the computation procedure of multiscale dual-branch waveconv.

MultiScale Dual-Branch WaveConv The architecture of the multiscale dual-branch waveconv is shown in Figure 2 (c). Overall, its basic structure is similar to a standard UNet (Ronneberger, Fischer, and Brox 2015) model and can extract and aggregate dense temporal features at different levels. First, the wavelet transform is utilized to recursively decompose the low-frequency components, extracting both low- and high-frequency elements across multiple levels. Second, at each level, two dedicated branches are specifically designed to model low- and high-frequency temporal dynamics, respectively. Finally, the inverse wavelet transform is employed to integrate the temporal dynamics across all levels. Due to the efficiency of wavelet transform, our carefully designed MSDBWC module can effectively model complex temporal variations with low computation costs.

Specifically, each MSDBWC module consists of two parts: one, referred to as the base branch, is designed to model temporal patterns from the perspective of spatial features, while the other focuses on modeling multiscale temporal patterns from the frequency angle. Assume $\hat{F}_{p-1}^{t-1,t}$ represents the input of the MSDBWC module in the $(p-1)$-th

branch, its computation procedure is formally defined as follows:

$$\mathrm{MSDBWC}(\hat{F}_{p-1}^{t-1,t}) = \mathrm{P}_b(\hat{F}_{p-1}^{t-1,t}) \oplus \mathrm{P}_w(\hat{F}_{p-1}^{t-1,t}). \quad (4)$$

The $\mathrm{P}_b(\cdot)$ is a depth-wise convolution layer. The $\mathrm{P}_w(\cdot)$ represents multiscale wavelet transform structure. Its recurrent computation process is as follows:

$$\mathrm{lowfreq}_1, \mathrm{highfreq}_1 = \mathrm{WT}(\hat{F}_{p-1}^{t-1,t}), \quad (5)$$

$$\mathrm{lowfreq}_i, \mathrm{highfreq}_i = \mathrm{WT}(\mathrm{lowfreq}_{i-1}), (i \geq 2), \quad (6)$$

$$\mathrm{lowtemporal}_i = \mathrm{ReLU}(\mathrm{DDWConv}(\mathrm{lowfreq}_i)), \quad (7)$$

$$\mathrm{hightemporal}_i = \mathrm{DWConv}(\mathrm{highfreq}_i), \quad (8)$$

$$\mathrm{temporl}_i = \mathrm{IWT}(\mathrm{lowtemporal}_i \oplus \mathrm{temporal}_{i+1} \text{ Concat } \mathrm{hightemporal}_i). \quad (9)$$

The $\mathrm{WT}(\cdot)$ denotes the wavelet transform operation, and $\mathrm{lowfreq}_1$ and $\mathrm{highfreq}_1$ represent the initial low- and high-frequency elements, respectively. The $\mathrm{DDWConv}(\cdot)$ denotes the dilated depth-wise convolution layer (Yu and Koltun 2016), and the combination of $\mathrm{ReLU}(\cdot)$ and $\mathrm{DDWConv}(\cdot)$ can better capture temporal patterns of low-frequency elements. We employ the standard depth-wise convolution layer to model temporal patterns of high-frequency ones. This design difference can better model the temporal dynamics of low- and high-frequency elements, respectively. The $\mathrm{temporl}_i$ is the output of the i-th layer in $\mathrm{P}_w(\cdot)$. The $\mathrm{IWT}(\cdot)$ is the inverse wavelet transform to fuse the learned low- and high-frequency temporal patterns. By literally conducting the formulae (6) - (9), the designed multiscale wavelet transform structure can effectively model temporal dynamics at various scale levels.

Training

Due to its architecture, the predicted sequence length of WavePredNet is equal to the input sequence length at a single prediction step. However, WavePredNet is capable of generating a sequence of arbitrary length in an autoregressive manner. For instance, to produce k predicted frames, WavePredNet can generate the k frames in $\lceil \frac{k}{t} \rceil$ recurrent prediction steps, using the given t frames. Specifically, we utilize the combination of MSE and MAE loss functions to optimize the proposed WavePredNet.

Experiments

Experimental Setup

FY4ABT Series Datasets We present FY4ABT series datasets by collecting and processing AGRI-L1-4KM data from the Chinese FengYun-4A (FY-4A) geostationary satellite. It scans the China region and full disk area in 4 minutes and 15 minutes, respectively, to achieve continuous observation of the China region. With the advanced geostationary radiation imager (AGRI), the FY-4A satellite can produce 14-channel data in different wavebands. Specifically, channels 1-3, 4-6, 7-8, 9-10, and 11-14 denote visible & near-infrared type, short-wave infrared type, mid-wave infrared type, water vapor type, and long-wave infrared type, respectively. Here, we utilize the 12th channel (10.7 μm waveband) to construct brightness temperature sequences, as this band provides stable observations. Moreover, it is the most consistent and comparable channel across other geostationary satellites such as GOES and Himawari series.

Dataset	(C, H, W)	N_{train}	N_{val}	N_{test}	T	L
FY4ABT-S		443100	23180	48440		1
FY4ABT-M	$[1, 128, 128]$	332325	17385	36330	2	2
FY4ABT-L		110775	5795	12110		4

Table 1: (C, H, W) denotes the shape of each processed brightness temperature frame. N_{train}, N_{val}, and N_{test} denote the sample size of training, validation, and test datasets, respectively. T and L are the input and output lengths.

In particular, each pixel in original satellite data denotes the $4km \times 4km$-size region, and its value indicates the temperature brightness. First, the brightness temperature images are made by mapping the temperature brightness into standard gray values, ranging in [0, 255], and the higher gray value denotes lower temperature brightness. Then, we extract a series of 24-frame brightness temperature sequences $\{x_{t-105}, x_{t-90}, ..., x_{t-15}, x_t, x_{t+15}, x_{t+30}, ..., x_{t+240}\}$ with an interval of 15 minutes. Additionally, we use 95% of the samples extracted from the 2019 data as the training set, with the remaining 5% serving as the validation set, and 10% of samples from the 2020 data are used as the test set. To save the computational cost, we first crop 512×512-size parts of the whole brightness temperature data and then resize the 512×512-size sequences into 128×128 ones, and finally resample the sequences with an interval of 1 hour. In this manner, we obtain the resampled brightness temperature sequences $\{x_{t-105}, x_{t-45}, x_{t+15}, x_{t+75}, x_{t+135}, x_{t+195}\}$ with a period of 6 hours, covering local areas of approximately 512×512 square kilometers. In this work, we aim to predict the next brightness temperature frames with the historical 2-hour sequences. Based on previous studies (Mecikalski and Bedka 2006; Han et al. 2019) on atmospheric evolution, especially for convective cloud initiation and development, we create three sub-datasets, FY4ABT-S (predicting the next 1 hour), FY4ABT-M (2 hours), and FY4ABT-L (4 hours), to validate the forecasting ability of our method under short-, medium-, and long-term conditions, respectively. The basic statistics of the datasets are shown in Table 1.

Baselines To comprehensively validate the performance of the proposed WavePredNet, we compare it against diverse and strong baselines, including classic Non-DL prediction methods and state-of-the-art DL spatiotemporal predictive models. First, we compare WavePredNet with two typical non-deep learning baselines, Persistence (Trebing, Stańczyk, and Mehrkanoon 2021) and pySTEPs (Pulkkinen et al. 2019), both of which are widely used in practical weather nowcasting scenarios. Persistence (Trebing, Stańczyk, and Mehrkanoon 2021) adopts the last frame of the input sequence as the prediction results at the next time step, which is established on the assumption that weather conditions won't change abruptly during a short

Prediction Method	Performance short-term (1h)				Performance medium-term (2h)				Performance long-term (4h)						
	MSE ↓	PSNR ↑	SSIM ↑	LPIPS ↓	MSE	PSNR	SSIM	LPIPS	MSE	PSNR	SSIM	LPIPS	FLOPs	#param	Memory
Persistence	62.459	25.702	0.802	**0.0565**	88.080	24.350	0.761	**0.0754**	132.330	22.670	0.710	**0.1056**	-	-	-
pySTEPs	46.681	26.854	0.836	0.0663	69.369	25.350	0.790	0.0838	110.221	23.504	0.732	0.1110	-	-	-
ConvLSTM	27.680	29.182	0.877	0.2279	39.601	27.710	0.842	0.2936	56.291	26.278	0.814	0.2909	59.779G	15.08M	0.638GB
E3DLSTM	33.815	28.415	0.854	0.2635	44.880	27.184	0.821	0.3170	66.656	25.573	0.788	0.3643	0.131T	52.92M	1.600GB
PredRNN-V2	29.098	29.420	0.883	0.1879	38.397	27.949	0.847	0.2612	54.128	26.485	0.816	0.2984	0.123T	23.59M	1.032GB
SimVP	26.213	29.500	0.888	0.1634	40.378	27.640	0.835	0.3431	58.821	26.216	0.809	0.3057	17.862G	14.22M	1.032GB
TAU	24.097	29.742	0.893	0.1835	34.838	28.379	0.860	0.2234	56.335	26.379	0.812	0.3193	13.371G	9.97M	1.320GB
MGVP	37.707	27.813	0.856	0.1443	54.834	26.307	0.813	0.2143	81.648	24.578	0.763	0.3213	0.864G	0.60M	0.542GB
WavePredNet (ours)	**24.035**	**29.849**	**0.894**	0.1484	**34.342**	**28.478**	**0.862**	0.2089	**53.263**	**26.631**	**0.818**	0.3013	7.832G	0.53M	0.580GB

Table 2: Quantitative prediction results of baseline methods and WavePredNet. The MSE, SSIM, and PSNR are low-level metrics to evaluate prediction accuracy, while the LPIPS score is a high-level metric to judge the visual quality of prediction results. Notably, the MSE score measures the image-level error between the prediction (normalized to 0-1) and the ground truth. The FLOPs represents the computational cost, and #params indicates the number of trainable parameters. Memory is the consumed training GPU memory with batch size of 1. Bold and underline denote the best and the second-best results.

period. pySTEPs (Pulkkinen et al. 2019) is an advection-based prediction approach that utilizes physical equations to model complicated motion fields of two consecutive frames and then extrapolates the states into the future. Second, we evaluate WavePredNet against state-of-the-art deep learning spatiotemporal predictive models, including competitive recurrent-based architectures: ConvLSTM (Shi et al. 2015), E3DLSTM (Wang et al. 2018b), PredRNN-V2 (Wang et al. 2022), recurrent-free architectures: SimVP (Gao et al. 2022), TAU (Tan et al. 2023a), and a recent baseline MGVP (Zhong et al. 2024).

Implementation Details We utilize the Adam optimizer with a learning rate of 1e-3 to train the proposed WavePredNet. Notably, we employ the Haar wavelet to implement the wavelet transform operations. In particular, the channel of spatial features for each frame is 64, and the reduced channel size $\hat{C}$ is 64. The input sequence's split number, the TCWTB's stacking depth, and the MSDBWC module's wavelet-decomposing level are set to 2, 4, and 2, respectively. The kernel size of each point-wise convolution layer is 3×3, and the dilated size of the dilated point-wise convolution layer is set to 3. In addition, we employ a standard framework OpenSTL (Tan et al. 2023b) to train and evaluate baselines. All the experiments are conducted on a GPU server of CPU i9-10920X @ 3.50GHz with 64G memory and two RTX-3090 cards.

Comparison Results

As shown in Table 2, we have the following findings here. First, WavePredNet consistently delivers the best performance across all MSE, PSNR, and SSIM metrics under short-, medium-, and long-term prediction scenarios. This demonstrates its superior ability to accurately model and predict the temporal patterns of sequences under different forecast horizons. Second, while the traditional methods (Persistence and pySTEPs) achieve the best and second-best LPIPS scores, respectively, their performance on the other three metrics is significantly inferior to the deep learning models. This suggests that while visually appealing (low LPIPS), their predictions are inaccurate. Notably, WavePredNet also achieves near-best performance on the LPIPS score among the deep learning approaches, indicating better preservation of visual details. Third, although some deep learning baselines show competitive performance, the computational cost remains a significant limiting factor. Specifically, while the recurrent-based baseline PredRNN-V2 achieves the second-best performance for long-term prediction, and the recurrent-free baseline TAU for short- and medium-term predictions, WavePredNet consistently outperforms both with substantially lower computational requirements. For the 4-hour prediction, WavePredNet achieves a 1.60% improvement in MSE score over PredRNN-V2, using 6.37% FLOPs and 2.25% parameters. Similarly, it achieves a 5.45% MSE improvement over TAU, using 58.56% FLOPs and 5.32% parameters. At last, though the recent baseline MGVP takes a minor computational cost, it delivers quite inaccurate prediction results. All the aforementioned observations consistently highlight the superiority of our model in both performance and efficiency.

Figure 3 presents the representative 4-hour prediction samples. First, as shown in the error map, the baseline TAU delivers the most inaccurate results among the three deep learning models. Second, compared to the baseline PredRNN-V2, our proposed WavePredNet better preserves appearance details. Additionally, by comparing the error maps, we can see that WavePredNet more accurately predicts the locations and intensity of clouds (low brightness temperature). Figure 4 shows metric curves w.r.t lead time. First, E3DLSTM and MGVP are significantly inferior to other models during the 4-hour lead time, even at the 1-hour time step. This suggests the limitation of E3DLSTM and MGVP in modeling brightness temperature sequences. Second, our proposed WavePredNet outperforms the other deep

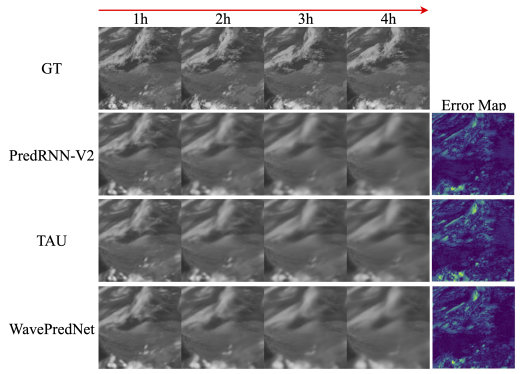

Figure 3: Visualization of representative 4-hour prediction results. The error map measures the difference between the last predicted frame and the ground truth.

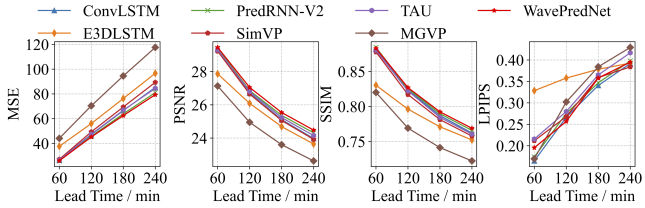

Figure 4: The forecasting curves w.r.t 4-hour lead time.

learning models in MSE, PSNR, and SSIM scores, and the advantages become more obvious as lead time increases. For the LPIPS score, WavePredNet delivers competitive results that are close to the best. The previous observations suggest that WavePredNet can better preserve appearance details and more accurately predict long-term temporal patterns.

Ablation Study

We conduct ablation experiments to verify the effectiveness of the two key structures, temporal-cascaded wavelet transform block (TCWTB) and multiscale dual-branch waveconv (MSDBW), by answering the following four questions.

(1) How does the number of temporal cascades in TCWTB affect the performance? This element determines the granularity for extracting temporal patterns of brightness temperature sequences. The different numbers denote different-grained temporal modeling. By setting various split numbers (1, 2, 4, 8), we obtain corresponding variants of WavePredNet. Figure 5 (a) shows the 4-hour prediction results of these variants. We can see that a too-small split number, such as 2, or a too-large value, such as 8, degrades prediction accuracy, even underperforming non-split (number 1). Notably, a split number of 4 yields the best results. The 2-branch setting lacks sufficient temporal decomposition to improve over the baseline, while the 8-branch setting suffers from over-fragmentation and results in suboptimal performance. This observation highlights the significance of a suitable branch setting for the cascaded setting.

(2) How does the decomposing level of MSDBW affect the performance? The large decomposing level represents finer-grained spatial modeling. Overall, increasing the split

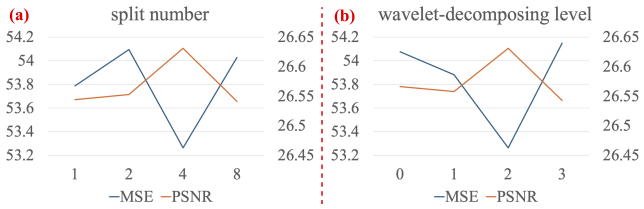

Figure 5: The ablation experiment results in terms of the split number and wavelet-decomposing level.

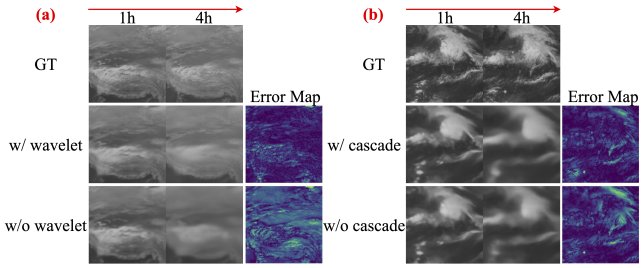

Figure 6: Visualization comparison in ablating the wavelet-decomposing and temporally-cascaded structures.

number leads to better results. The 4-hour prediction results for variants employing different wavelet decomposition levels (0, 1, 2, 3) are illustrated in Figure 5 (b). While increasing the decomposition level generally improves predictive accuracy, excessively high levels, such as 3, lead to performance degradation. Notably, a decomposition level of 2 achieves optimal results. This observation demonstrates the effectiveness of the multiscale dual-branch waveconv structure.

(3) Is the temporally cascaded structure necessary? To explain the function of this architecture, we remove the temporal structure in a pre-trained WavePredNet model, whose corresponding prediction samples are shown in Figure 6 (a). Intuitively, without the cascaded structure, the accuracy for predicting pixel locations and intensity significantly degrades. This is reasonable, as without the structure, our model cannot progressively increase the temporal receptive field, leading to inaccurate predictions. This observation demonstrates the significance of the temporal cascaded structure in capturing temporal patterns.

(4) Is the wavelet-decomposition useful? Similarly, we remove the wavelet-decomposition architecture, and Figure 6 (b) shows comparison samples. Without this structure, predictions become much blurrier, and the motion patterns are also much more inaccurate. This highlights the importance of the designed architecture in capturing multi-scale dynamics across low- and high-frequency components.

Conclusion

In this work, we present an evaluation benchmark, FY4ABT, and a novel, efficient model, WavePredNet, for brightness temperature sequence forecasting. Comprehensive experiments validate the superiority of WavePredNet and the effectiveness of its key structures. We hope our work will inspire further research and development in this task.

Acknowledgments

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

References

Alaba, S. Y.; and Ball, J. E. 2022. Wcnn3d: Wavelet convolutional neural network-based 3d object detection for autonomous driving. *Sensors*, 22(18): 7010.

Benson, V.; Robin, C.; Requena-Mesa, C.; Alonso, L.; Carvalhais, N.; Cortés, J.; Gao, Z.; Linscheid, N.; Weynants, M.; and Reichstein, M. 2024. Multi-modal Learning for Geospatial Vegetation Forecasting. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 27788–27799.

Bowler, N. E.; Pierce, C. E.; and Seed, A. W. 2006. STEPS: A probabilistic precipitation forecasting scheme which merges an extrapolation nowcast with downscaled NWP. *Quarterly Journal of the Royal Meteorological Society: A Journal of the Atmospheric Sciences, Applied Meteorology and Physical Oceanography*, 132(620): 2127–2155.

Dai, K.; Li, X.; Ma, C.; Lu, S.; Ye, Y.; Xian, D.; Tian, L.; and Qin, D. 2023. Learning Spatial–Temporal Consistency for Satellite Image Sequence Prediction. *IEEE Transactions on Geoscience and Remote Sensing*, 61: 1–17.

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

Daubechies, I. 1992. Ten lectures on wavelets. *Society for Industrial and Applied Mathematics*.

Finder, S. E.; Amoyal, R.; Treister, E.; and Freifeld, O. 2024. Wavelet Convolutions for Large Receptive Fields. In *Proceedings of the European Conference on Computer Vision*.

Gao, Z.; Tan, C.; Wu, L.; and Li, S. Z. 2022. Simvp: Simpler yet better video prediction. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 3170–3180.

Garnot, V. S. F.; and Landrieu, L. 2021. Panoptic Segmentation of Satellite Image Time Series with Convolutional Temporal Attention Networks. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 4872–4881.

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

Han, D.; Lee, J.; Im, J.; Sim, S.; Lee, S.; and Han, H. 2019. A novel framework of detecting convective initiation combining automated sampling, machine learning, and repeated model tuning from geostationary satellite data. *Remote Sensing*, 11(12): 1454.

Hartman, C. M.; Chen, X.; Clothiaux, E. E.; and Chan, M.-Y. 2021. Improving the analysis and forecast of Hurricane Dorian (2019) with simultaneous assimilation of GOES-16 all-sky infrared brightness temperatures and tail Doppler radar radial velocities. *Monthly Weather Review*, 149(7): 2193–2212.

Hatanaka, Y.; Glaser, Y.; Galgon, G.; Torri, G.; and Sadowski, P. 2023. Diffusion models for high-resolution solar forecasts. *arXiv preprint arXiv:2302.00170*.

Jiang, Y.; Cheng, W.; Gao, F.; Zhang, S.; Liu, C.; and Sun, J. 2022. CSIP-Net: Convolutional Satellite Image Prediction Network for Meteorological Satellite Infrared Observation Imaging. *Atmosphere*, 14(1): 25.

Lebedev, V.; Ivashkin, V.; Rudenko, I.; Ganshin, A.; Molchanov, A.; Ovcharenko, S.; Grokhovetskiy, R.; Bushmarinov, I.; and Solomentsev, D. 2019. Precipitation Nowcasting with Satellite Imagery. In *Proceedings of the ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, 2680–2688.

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

Leinonen, J.; Hamann, U.; Sideris, I. V.; and Germann, U. 2023. Thunderstorm Nowcasting With Deep Learning: A Multi-Hazard Data Fusion Model. *Geophysical Research Letters*, 50(8): e2022GL101626.

Lin, Z.; Li, M.; Zheng, Z.; Cheng, Y.; and Yuan, C. 2020. Self-Attention ConvLSTM for Spatiotemporal Prediction. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 34, 11531–11538. ISBN 2374-3468.

Liu, P.; Zhang, H.; Zhang, K.; Lin, L.; and Zuo, W. 2018. Multi-level wavelet-CNN for image restoration. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops*, 773–782.

Mecikalski, J. R.; and Bedka, K. M. 2006. Forecasting convective initiation by monitoring the evolution of moving cumulus in daytime GOES imagery. *Monthly Weather Review*, 134(1): 49–78.

Pulkkinen, S.; Nerini, D.; Pérez Hortal, A. A.; Velasco-Forero, C.; Seed, A.; Germann, U.; and Foresti, L. 2019. Pysteps: An open-source Python library for probabilistic precipitation nowcasting (v1. 0). *Geoscientific Model Development*, 12(10): 4185–4219.

Ravuri, S.; Lenc, K.; Willson, M.; Kangin, D.; Lam, R.; Mirowski, P.; Fitzsimons, M.; Athanassiadou, M.; Kashem, S.; Madge, S.; et al. 2021. Skilful precipitation nowcasting using deep generative models of radar. *Nature*, 597(7878): 672–677.

Ronneberger, O.; Fischer, P.; and Brox, T. 2015. U-net: Convolutional networks for biomedical image segmentation. In *Proceedings of the Medical Image Computing and Computer-Assisted Intervention*, 234–241.

Shi, X. J.; Chen, Z. R.; Wang, H.; Yeung, D. Y.; Wong, W. K.; and Woo, W. C. 2015. Convolutional LSTM Network: A Machine Learning Approach for Precipitation

Nowcasting. In *Proceedings of the Advances in Neural Information Processing Systems*, volume 28, 802–810.

Shukla, B. P.; Kishtawal, C. M.; and Pal, P. K. 2013. Prediction of satellite image sequence for weather nowcasting using cluster-based spatiotemporal regression. *IEEE Transactions on Geoscience and Remote Sensing*, 52(7): 4155–4160.

Tan, C.; Gao, Z.; Wu, L.; Xu, Y.; Xia, J.; Li, S.; and Li, S. Z. 2023a. Temporal attention unit: Towards efficient spatiotemporal predictive learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 18770–18782.

Tan, C.; Li, S.; Gao, Z.; Guan, W.; Wang, Z.; Liu, Z.; Wu, L.; and Li, S. Z. 2023b. Openstl: A comprehensive benchmark of spatio-temporal predictive learning. *Proceedings of the Advances in Neural Information Processing Systems*, 36: 69819–69831.

Tarasiou, M.; Chavez, E.; and Zafeiriou, S. 2023. Vits for sits: Vision transformers for satellite image time series. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 10418–10428.

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

Turner, B. J.; Zawadzki, I.; and Germann, U. 2004. Predictability of precipitation from continental radar images. Part III: Operational nowcasting implementation (MAPLE). *Journal of Applied Meteorology and Climatology*, 43(2): 231–248.

Wang, T.; Lu, C.; Sun, Y.; Yang, M.; Liu, C.; and Ou, C. 2021. Automatic ECG classification using continuous wavelet transform and convolutional neural network. *Entropy*, 23(1): 119.

Wang, Y.; Gao, Z.; Long, M.; Wang, J.; and Philip, S. Y. 2018a. Predrnn++: Towards a resolution of the deep-in-time dilemma in spatiotemporal predictive learning. In *Proceedings of the International Conference on Machine Learning*, 5123–5132.

Wang, Y.; Jiang, L.; Yang, M.-H.; Li, L.-J.; Long, M.; and Fei-Fei, L. 2018b. Eidetic 3d lstm: A model for video prediction and beyond. In *Proceedings of the International Conference on Learning Representations*.

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

Wang, Y.; Wu, H.; Zhang, J.; Gao, Z.; Wang, J.; Philip, S. Y.; and Long, M. 2022. Predrnn: A recurrent neural network for spatiotemporal predictive learning. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 45(2): 2208–2225.

Wang, Y.; Zhang, J.; Zhu, H.; Long, M.; Wang, J.; and Yu, P. S. 2019. Memory in memory: A predictive neural network for learning higher-order non-stationarity from spatiotemporal dynamics. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 9154–9162.

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

Xu, G.; Liao, W.; Zhang, X.; Li, C.; He, X.; and Wu, X. 2023. Haar wavelet downsampling: A simple but effective downsampling module for semantic segmentation. *Pattern Recognition*, 143: 109819.

Xu, Z.; Du, J.; Wang, J.; Jiang, C.; and Ren, Y. 2019. Satellite Image Prediction Relying on GAN and LSTM Neural Networks. In *Proceedings of the IEEE International Conference on Communications*, 1–6.

Yao, T.; Pan, Y.; Li, Y.; Ngo, C.-W.; and Mei, T. 2022. Wavevit: Unifying wavelet and transformers for visual representation learning. In *Proceedings of the European Conference on Computer Vision*, 328–345.

Yu, F.; and Koltun, V. 2016. Multi-Scale Context Aggregation by Dilated Convolutions. In *Proceedings of the International Conference on Learning Representations*.

Zach, C.; Pock, T.; and Bischof, H. 2007. A duality based approach for realtime tv-l 1 optical flow. In *Pattern Recognition: 29th DAGM Symposium*, 214–223.

Zhang, Y.; Long, M.; Chen, K.; Xing, L.; Jin, R.; Jordan, M. I.; and Wang, J. 2023. Skilful nowcasting of extreme precipitation with NowcastNet. *Nature*, 619(7970): 526–532.

Zhao, C.; Cai, W.; Dong, C.; and Hu, C. 2024. Wavelet-based fourier information interaction with frequency diffusion adjustment for underwater image restoration. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 8281–8291.

Zhong, Y.; Liang, L.; Tang, B.; Zharkov, I.; and Neumann, U. 2024. Motion graph unleashed: A novel approach to video prediction. *Proceedings of the Advances in Neural Information Processing Systems*, 37: 111022–111046.