Focal Frame Loss: A Simple but Effective Loss for Precipitation Nowcasting

Zhifeng Ma, Hao Zhang, and Jie Liu, *Fellow, IEEE*

***Abstract*—Precipitation nowcasting is an important but hard problem. Currently, with the landing of deep learning, it has been treated as an image prediction problem based on radar echo maps. However, deep learning models suffer from poor performance and blurred prediction results. Lots of improvement works enhance the model by adding complex modules, which increases insufferable training memory and time overhead. Others tempt to add more limitations or guidances on loss, but they usually have little effect in such an extremely complex and difficult task. In this article, we propose a simple but effective loss named focal frame loss (FFL), which assigns different weights to the images in the prediction sequence to focus on the images that are relatively difficult to predict. Experiments on two large-scale radar datasets show that FFL can greatly improve the performance of multiple popular models without introducing additional training costs.**

***Index Terms*—Deep learning, low overhead, precipitation nowcasting, sequence prediction.**

I. Introduction

NOWADAYS, precipitation nowcasting, usually up to 2 h [1], plays an important role in many fields such as agriculture [2], travel [3], transport [4], fieldwork [5], etc. It provides critical guidance for planning and scheduling in production and daily life. Accurate and high-resolution precipitation nowcasting has become a hot research topic in meteorology and hydrology communities [6]. However, predicting the short-term rainfall in a region is challenging as it relies on a mass of meteorological factors such as temperature, humidity, wind, pressure in the region, and complex atmospheric physical mechanisms.

Traditional methods for precipitation nowcasting can be roughly categorized into two classes, which are numerical weather prediction (NWP [7], [8], [9]) based methods and radar echo extrapolation based methods. Numerical models usually require an integration period to spin up the deduction processes and are greatly affected by the initial condition field, which leads to poor prediction performance in the first few hours [10]. Furthermore, solving the complex atmosphere equations consumes huge amounts of computing resources, which prevents real-time and high-resolution predictions [11]. The echoes detected by weather radar reveal the intensity of precipitation, and extrapolation (or prediction) of radar echo maps has become one of the important means for precipitation nowcasting [12]. Another prominent strategy is using optical flow methods to extrapolate the radar echo maps [13], [14]. Precipitation prediction is made along the optical flow field, which is calculated based on the last few radar maps. These methods usually cannot give accurate prediction results in practice since they can only capture the simple linear change, while plenty of complex nonlinear changes exist in nowcasting [15].

In recent years, with the development of deep learning, many studies [16], [17], [18], [19], [20], [21] introduce deep learning into precipitation nowcasting by formulating it as a radar echo spatiotemporal sequence prediction problem [22], [23], [24]. Nevertheless, spatio-temporal prediction is challenging attributing to the high nonlinearity in temporal dynamics as well as complex location-characterized patterns in spatial domains, especially in fields like precipitation nowcasting [25]. Some works [26], [27], [28] use the convolutional model UNet [29] with a simple structure for precipitation nowcasting. However, convolutions have natural shortcomings in capturing temporal trends due to their inability to cope with complex temporal nonlinear changes. ConvLSTM [16] is the pioneer recurrent model to solve this problem, which utilizes convolution and LSTM [30] to model the spatial variation and temporal dynamics, respectively. However, the performance of ConvLSTM is not satisfactory, and the predicted frames usually suffer from the blur problem [31].

Previous works can be divided into two ways to enhance the capabilities of the basic deep learning model. One is using more powerful structures. For instance, PredRNN [18], MIM [19], and MotionRNN [20] enhance ConvLSTM's ability in capturing complex meteorological changes by adding more complex modules. However, they bring in a lot of model parameters and intermediate variables, which greatly increase the time and memory cost of training. Others [32], [33] attempt to add more limitations or guidances based on the basic mean absolute error (MAE) or mean square error (MSE) loss. These methods introduce less overhead, but they usually have limited improvements in such a complex scenario of precipitation nowcasting as they do not take into account the nature of the problem.

Manuscript received 24 March 2022; revised 3 July 2022 and 19 July 2022; accepted 25 July 2022. Date of publication 28 July 2022; date of current version 26 August 2022. This work was supported in part by the National Key R&D Program of China under Grant 2021ZD0110905, in part by the Programs for Science and Technology Development of Heilongjiang Province under Grant 2021ZXJ05A03, in part by the National Natural Science Foundation of China under Grant 62106061 and Grant 61972114, in part by the Fundamental Research Funds for the Central Universities under Grant AUGA5710010521, and in part by the National Natural Science Foundation of Heilongjiang Province under Grant YQ2019F007. *(Corresponding author: Hao Zhang.)*

Zhifeng Ma and Hao Zhang are with the Faculty of Computing, Harbin Institute of Technology, Harbin 150001, China (e-mail: 20b903027@stu.hit.edu.cn; zhh1000@hit.edu.cn).

Jie Liu is with the International Institute for Artificial Intelligence, Harbin Institute of Technology Shenzhen, Shenzhen 518055, China (e-mail: jieliu@hit.edu.cn).

Digital Object Identifier 10.1109/JSTARS.2022.3194522

In this article, we find that the prediction difficulty of each frame in the sequence is different in the radar map prediction task. Lots of relatively easy predict images in the generated sequence contribute no useful learning signal and dominate the gradient. It prevents the model from learning potential dynamics from difficult samples. We should pay more attention to the frames that are hard to predict. Based on this simple idea, we propose the focal frame loss (FFL) based on the common MSE and MAE loss for precipitation nowcasting, which spontaneously assigns larger weights to those frames that are relatively difficult to predict. It significantly improves the predictive performance for multiple popular models without introducing additional overhead. The main contributions of this article are as follows:

1) We find the phenomenon that the prediction difficulty of each image varies in the radar sequence prediction task, which is ignored by commonly used loss functions.
2) We design a simple but effective loss named FFL for precipitation nowcasting.
3) We verify FFL on two large-scale radar datasets. The experiments show that FFL can greatly improve the performances of current popular models without introducing additional training cost.

II. Related Work

Model Classification: Most works treat precipitation nowcasting as a radar sequence prediction (or video prediction) task. The mainstream sequence prediction models can be divided into two categories: the convolutional neural network (CNN) and the recurrent neural network (RNN). The CNN models are dominated by UNet [34] and its variants [26], [35], [36]. However, CNN implicitly assumes complex changes in spatial appearance and may therefore fall short in learning long-term dependencies [21]. The RNN models are dominated by ConvLSTM [16] and its variants, such as TrajGRU [17], PredRNN [18], PredRNN++ [22], E3D-LSTM [23], MIM [19], CubicLSTM [37], SA-ConvLSTM [24], MotionRNN [20], etc. These RNN models are getting wider and deeper [38]. Although they alleviate the prediction ambiguity problem to some extent, it also brings a significant increase in computational cost.

Loss Function: Using an L_2 loss, to a lesser extent L_1, produces blurry predictions, increasingly worse when predicting further in the future [31]. Many previous works try to deal with the inherently blurry predictions obtained from the standard MSE or MAE loss function in the precipitation nowcasting task. Tran et al. [32] found that using a combination of structural similarity (SSIM) [39] with MSE and MAE yields better prediction quality. Song et al. [36] designed a loss function combining root mean squared error (RMSE) and intersection over union (IOU) [40] to better capture significant raining dynamics. [33] adds gradient difference loss (GDL) [31] to the basis of MSE and MAE, which is expected to guide the model to match the gradients of pixel values and to alleviate the image blurring tendency of predicted frames. However, these improvements are just designed for the overall clarity of the image, which just plays an auxiliary role. MSE and MAE are still playing a leading role.

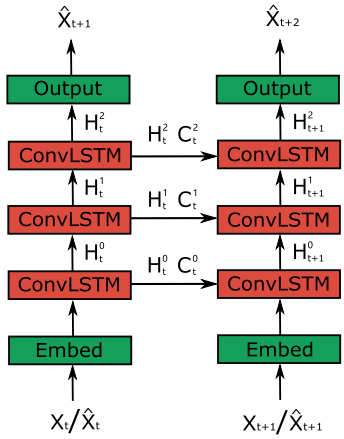

Fig. 1. Architecture of ConvLSTM. For timestamp $\{0, \ldots, m-1\}$, The input to the model is the ground truth frame. For timestamp $\{m, \ldots, m+n-2\}$, The input to the model is the model's predicted frame at the previous moment.

III. Preliminary

A. Formulation of Precipitation Nowcasting Problem

In this article, precipitation nowcasting in a local region (e.g., Los Angeles) is translated to predicting the future radar echo map (e.g., 0–2 h) based on the observed radar echo map sequence. For a radar dynamical system, from the spatial view, if we need to record C measurements of a certain local area ($H \times W$ grid points) at any time, we can express it in the form of a tensor $X_t \in \mathbb{R}^{C \times H \times W}$. From the temporal view, we can express the radar echo map sequence as a sequence of tensors $\{X_0, \ldots, X_{m-1}, X_m, \ldots, X_{m+n-1}\}$. Let $X = \{X_0, \ldots, X_{m-1}\}$ be the observation frames and $Y = \{X_m, \ldots, X_{m+n-1}\}$ be the predicted frames. The precipitation nowcasting problem is to predict the most probable length-n sequence $\hat{Y}$ in the future given the length-m observation sequence X. In this article, we train a neural network parameterized by θ to solve such a task. Specifically, we use stochastic gradient descent to find a set of parameters θ^* that maximizes the likelihood of producing the true target sequence Y given the input data X

$$\theta^* = \arg\max_{\theta} \; P(Y|X;\theta). \tag{1}$$

B. Convlstm

ConvLSTM [16] (see Fig. 1) has convolutional structures in both the input-to-state and the state-to-state transitions. which can model the spatial and temporal variation of radar sequence simultaneously. ConvLSTM is formulated as

$$\begin{aligned} i_t &= \sigma(W_{ix} * X_t + W_{ih} * H_{t-1}^l) \\ f_t &= \sigma(W_{fx} * X_t + W_{fh} * H_{t-1}^l) \\ g_t &= \tanh(W_{gx} * X_t + W_{gh} * H_{t-1}^l) \\ C_t^l &= f_t \circ C_{t-1}^l + i_t \circ g_t \\ o_t &= \sigma(W_{ox} * X_t + W_{oh} * H_{t-1}^l) \\ H_t^l &= o_t \circ \tanh(C_t^l) \end{aligned} \tag{2}$$

where W_{**} are the parameters that the model needs to learn; X_t represents input; H_t^l and C_t^l represent hidden state and cell state of layer l at time t, respectively; i, f, g, and o stand for input gate, forget gate, input modulation gate, and output gate, respectively; "$*$" is the convolution operation; "$\circ$" is the Hadamard product; and σ denotes the sigmoid activation function.

C. Focal Loss

In the object detection task, the extreme foreground-background class imbalance and the different contributions of easy and hard examples to the loss are the main reasons for the accuracy of the one-stage object detector is not as good as that of the two-stage object detector. Based on this, Lin et al. [41] proposed the famous focal loss (FL), which alleviates the above problems by modifying the cross-entropy function.

Here, we take the binary classification task as an example, and extending the FL to the multiclass case is straightforward. If we have

$$p_t = \begin{cases} p, & y = 1 \\ 1 - p, & \text{otherwise} \end{cases} \tag{3}$$

and

$$\alpha_t = \begin{cases} \alpha, & y = 1 \\ 1 - \alpha, & \text{otherwise} \end{cases} \tag{4}$$

then

$$\text{FL}(p_t) = -\alpha_t (1 - p_t)^{\gamma} \log(p_t) \tag{5}$$

where $y \in \{\pm 1\}$ specifies the ground-truth class, $p \in [0, 1]$ is the model's estimated probability for the class with label $y = 1$, and weighting factor $\alpha \in [0, 1]$ is introduced for class with label $y = 1$. For notational convenience, we define p_t and α_t. In FL, weighting factor α_t is introduced to address the class imbalance problem between positive and negative examples, and modulating factor $(1 - p_t)^{\gamma}$ is proposed to down-weight easy examples and thus focus training on hard examples.

IV. Focal Frame Loss

A. Problem of MAE or MSE Loss

For a consecutive radar sequence, m frames for the observations and n frames for the predictions. X_t and $\hat{X}_t$ represent the true target frame and its predicted frame, respectively. $\Delta X_t = |\, X_t - \hat{X}_t \,|$ represents the sum of the absolute difference of all pixels at frame t. Then, the MAE and MSE loss can be formulated as

$$\begin{aligned} \text{MAE} &= \frac{\sum_{t=1}^{m+n-1} \Delta X_t}{m + n - 1} \\ \text{MSE} &= \frac{\sum_{t=1}^{m+n-1} \Delta {X_t}^2}{m + n - 1} \end{aligned} \tag{6}$$

where $m + n$ represents the total sequence length. Here, we follow the frame reconstruction loss setting in PredRNN [18], MIM [19], and MotionRNN [20], which computes frame loss at each timestamp [21] and is better than the regular Seq2Seq structure [42] that only penalizes predicted frames. Specifically, they have the same structure as ConvLSTM (shown in Fig. 1), which input the real frames or predicted frames $\{X_0, \ldots, X_{m-1}, \hat{X}_m, \ldots, \hat{X}_{m+n-2}\}$, then output predicted frames $\{\hat{X}_1, \ldots, \hat{X}_m, \hat{X}_{m+1}, \ldots, \hat{X}_{m+n-1}\}$, and finally calculate the average loss for all output frames.

From (6), we can see that the loss weight of each frame in the sequence is the same, which is 1. However, intuitively, the prediction difficulty between each frame should be different. In the beginning, the inputs of the model are real frames $\{X_0, \ldots, X_{m-1}\}$, and the model can gradually learn sequence trends. After that, the inputs of the model are the predicted frames $\{\hat{X}_m, \ldots, \hat{X}_{m+n-2}\}$, which are inaccurate, and errors will accumulate over time [43], resulting in increasingly inaccurate predictions. That is to say, frames $\{\hat{X}_1, \ldots, \hat{X}_m\}$ with correct prior knowledge $\{X_0, \ldots, X_{m-1}\}$ are progressively easier to be predicted and frames $\{\hat{X}_{m+1}, \ldots, \hat{X}_{m+n-1}\}$ with inaccurate prior knowledge $\{\hat{X}_m, \ldots, \hat{X}_{m+n-2}\}$ are progressively harder to be predicted. These difficult frames are exactly what we want. It is unreasonable to assign the same penalty weight to each frame with different prediction difficulties. Therefore, the loss functions used by these advanced models in recent years have a problem.

To verify our opinion, we use ConvLSTM and $L_1 + L_2$ loss to experiment on the HKO-7 dataset [17], using 10 frames for the observations and 10 frames for the predictions. As shown in Fig. 2, we sample frames at intervals. We can see that Δ at time $\{t = 1, t = 3, t = 5, t = 7, t = 9\}$ is getting smaller and smaller, while Δ at time $\{t = 11, t = 13, t = 15, t = 17, t = 19\}$ is getting bigger and bigger. This is consistent with our analysis: The prediction difficulty varies for each frame.

B. Focal Frame Loss

In general, the prediction difficulty of each frame is different in radar spatiotemporal sequence prediction tasks, frames without the correct prior knowledge are difficult to predict and become more difficult over time, and simply averaging the losses for these frames results in the model penalizing the easy-to-predict and hard-to-predict frames the same. Lots of relatively easy predict frames in the generated sequence contribute no useful learning signal and dominate the gradient. In this article, inspired by [41] (see Section III-C), we introduce FFL to solve this problem, which redistributes the loss weight of each frame. In detail, let $\Delta X_t = |\, X_t - \hat{X}_t \,|$ be the sum value of absolute frame difference and $\Delta X = \{\Delta X_1, \Delta X_2, \ldots, \Delta X_{m+n-1}\}$ be the difference set of the predicted sequence, then ΔX_t can be regarded as the degree of difficulty of predicting the frame at time t. If ΔX_t is relatively large, it proves that the current frame is more difficult to be predicted, and we should focus on it, vice versa. Therefore, we can assign a weight

$$W_t = \left(\frac{\Delta X_t - \min(\Delta X)}{\max(\Delta X) - \min(\Delta X)} + \varepsilon \right)^k \tag{7}$$

to the tth frame according to the frame difference, where $\min(\Delta X)$ and $\max(\Delta X)$ represent the minimum and maximum difference value in a prediction sequence, respectively,

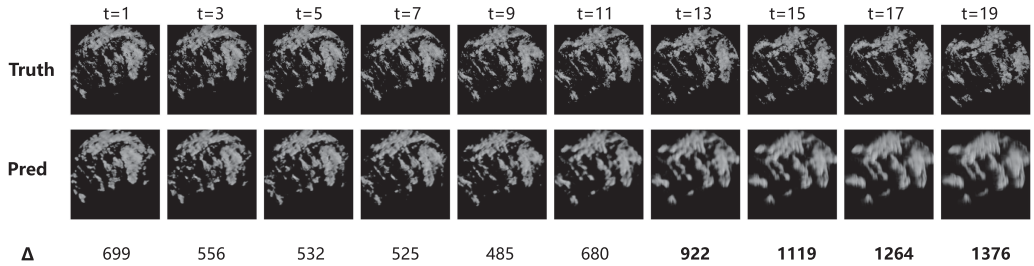

Fig. 2. Framewise difference of the ConvLSTM model on the HKO-7 dataset [17]. Δ denotes the difference value between the true image and the predicted image. The larger the Δ, the worse the prediction result and the harder it is to predict.

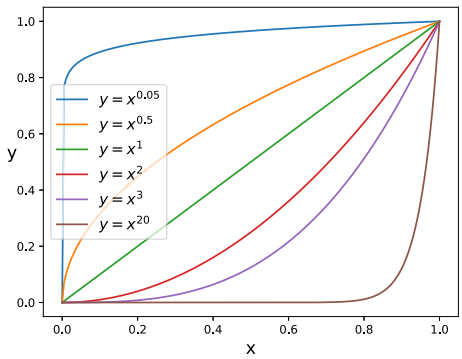

Fig. 3. Function curves of the function $y = x^k$ as k change.

k is a non-negative number ($k \geq 0$), and ε is the bias ($\varepsilon > 0$). That is, we normalize the difference of the predicted sequence and then map it to the weight W_t using the function $y = x^k$ (see Fig. 3), where the larger k is, the less weight is given to the frame that is easier to predict (smaller ΔX_t). Additionally, ε is introduced to prevent the most predictable frame ($\min(\Delta X)$) from having a weight value of 0. Combined with MAE or MSE, we propose focal frame mean absolute error (FF-MAE) and focal frame mean square error (FF-MSE), which are defined as

$$\text{FF-MAE} = \frac{\sum_{t=1}^{m+n-1} W_t \Delta X_t}{m+n-1}$$
$$\text{FF-MSE} = \frac{\sum_{t=1}^{m+n-1} W_t {\Delta X_t}^2}{m+n-1}. \tag{8}$$

In (8), we can see that the loss weight of each frame in the sequence is different, which is W_t. Specially, when $k = 0$, $W_t = 1$, FF-MAE, and FF-MSE degenerate into MAE and MSE, respectively. FF-MAE and FF-MSE loss care about frames that are difficult to predict. When ΔX_t is relatively small, it means that the frame is easier to predict than other frames in the sequence, and we should pay less attention to it. Therefore, W_t is relatively small, resulting in a smaller loss contribution of this frame, and vice versa.

In this article, we set the hyperparameter $k = 2$ and $\varepsilon = 0.01$. We experiment and analyze the setting of the value of k in Section V-D. It is worth mentioning that our proposed loss is different from FL [41] (see Section III-C): FL extends cross-entropy loss for classification tasks, while ours is combined with the L_1 or L_2 loss for regression tasks.

TABLE I
OVERVIEW OF THE HKO-7 AND DWD-12 DATASETS

Datasets	Radars	Years	Train Samples	Test Samples	Interval
HKO-7	1	7	192,000	31,365	6 min
DWD-12	17	12	605,880	123,930	5 min

V. EXPERIMENTS

A. Datasets

In this article, we use the HKO-7 [17] and DWD-12 [26] datasets (see Table I) to verify the performance of FFL, where HKO-7 is produced by the Hong Kong Observatory (HKO) while DWD-12 is produced by the German Weather Service (DWD).

For the HKO-7 dataset [17], only one Doppler radar is used to collect the data, and the radar echo map is collected from a height of 2 km every 6 min. We only use rainy days data, with 812 days for training, 50 days for validation, and 131 days for testing. The raw radar images have a resolution of 480 × 480 pixels covering a 512 × 512 km area centered in Hong Kong. The conversion relationships between radar echo reflectivity intensity (dBZ) and pixel value (P: 0-255) and rainfall intensity value (R : mm/h) are

$$P = \lfloor 255 \times \frac{\text{dBZ} + 10}{70} + 0.5 \rfloor$$
$$\text{dBZ} = 10 \times \lg(58.53 \times R^{1.056}). \tag{9}$$

For the DWD-12 dataset [26], 17 Doppler radars are used for collecting the data, and the spatial and temporal resolution of the product is 1 × 1 km and 5 min, respectively. It has a spatial extent of 900 × 900 km, covering the whole area of Germany. The raw radar images have a resolution of 900 × 900 pixels. We use the data from 2006 to 2014 for training, the data in 2015 for verification, and the data from 2016 to 2017 for testing. The conversion relationships between radar echo reflectivity intensity (dBZ) and pixel value (P: 0-255) and rainfall intensity value (R : mm/h) are

$$P = \lfloor 255 \times \frac{\text{dBZ} + 10}{70} + 0.5 \rfloor$$
$$\text{dBZ} = 10 \times \lg(256 \times R^{1.42}). \tag{10}$$

TABLE II
QUANTITATIVE STUDY ABOUT FFL ON THE HKO-7 DATASET

Models	#Params	Memory	Time	CSI ↑			HSS ↑			SSIM ↑	MSE ↓	B-MSE ↓
				$R \geq 0.5$	$R \geq 5$	$R \geq 30$	$R \geq 0.5$	$R \geq 5$	$R \geq 30$			
ConvLSTM [16]	0.89M	7.77GB	1.94H	0.666	0.531	0.317	0.779	0.677	0.465	0.785	75.144	158.658
ConvLSTM+FFL	0.89M	7.77GB	1.97H	**0.696**	**0.560**	**0.364**	**0.803**	**0.702**	**0.518**	**0.803**	**67.133**	**132.148**
TrajGRU [17]	1.06M	22.55GB	12.43H	0.665	0.531	0.310	0.779	0.676	0.457	0.782	77.244	161.084
TrajGRU+FFL	1.06M	22.55GB	12.62H	**0.703**	**0.569**	**0.369**	**0.809**	**0.710**	**0.523**	**0.805**	**65.668**	**133.039**
PredRNN [18]	1.80M	16.72GB	4.75H	0.680	0.552	0.342	0.791	0.695	0.494	0.792	69.312	144.775
PredRNN+FFL	1.80M	16.72GB	4.79H	**0.703**	**0.571**	**0.373**	**0.809**	**0.711**	**0.526**	**0.805**	**64.833**	**135.582**
MIM [19]	3.68M	31.77GB	9.11H	0.680	0.551	0.344	0.791	0.694	0.496	0.792	70.516	144.404
MIM+FFL	3.68M	31.77GB	9.12H	**0.707**	**0.571**	**0.373**	**0.812**	**0.711**	**0.526**	**0.806**	**63.999**	**137.116**
MotionRNN [20]	3.78M	33.84GB	15.64H	0.682	0.552	0.341	0.793	0.696	0.492	0.793	69.233	145.180
MotionRNN+FFL	3.78M	33.84GB	15.77H	**0.706**	**0.572**	**0.375**	**0.811**	**0.712**	**0.528**	**0.807**	**65.159**	**126.347**

TABLE III
QUANTITATIVE STUDY ABOUT FFL ON THE DWD-12 DATASET

Models	#Params	Memory	Time	POD ↑					GDL ↓	PSNR ↑	MAE ↓	B-MAE ↓
				$R \geq 0.5$	$R \geq 2$	$R \geq 5$	$R \geq 10$	$R \geq 30$				
ConvLSTM [16]	0.89M	7.77GB	6.24H	0.717	0.657	0.542	0.271	0.058	323.552	29.978	173.131	144.857
ConvLSTM+FFL	0.89M	7.77GB	6.36H	**0.771**	**0.741**	**0.676**	**0.398**	**0.200**	**292.655**	**30.907**	**142.735**	**122.269**
TrajGRU [17]	1.06M	22.55GB	39.09H	0.639	0.534	0.384	0.184	0.036	352.996	29.364	236.271	189.545
TrajGRU+FFL	1.06M	22.55GB	40.02H	**0.772**	**0.732**	**0.672**	**0.399**	**0.200**	**291.667**	**30.946**	**142.694**	**121.581**
PredRNN [18]	1.80M	16.72GB	15.14H	0.738	0.683	0.581	0.324	0.117	310.140	30.355	157.520	133.157
PredRNN+FFL	1.80M	16.72GB	15.31H	**0.774**	**0.748**	**0.680**	**0.398**	**0.202**	**289.753**	**30.977**	**142.402**	**121.895**
MIM [19]	3.68M	31.77GB	29.08H	0.749	0.700	0.604	0.350	0.135	306.174	30.487	153.564	130.670
MIM+FFL	3.68M	31.77GB	29.18H	**0.764**	**0.729**	**0.658**	**0.381**	**0.239**	**289.752**	**30.966**	**142.982**	**121.931**
MotionRNN [20]	3.78M	33.84GB	50.53H	0.740	0.686	0.583	0.326	0.129	311.583	30.359	159.424	134.794
MotionRNN+FFL	3.78M	33.84GB	50.59H	**0.775**	**0.741**	**0.690**	**0.419**	**0.196**	**289.615**	**30.991**	**142.075**	**121.096**

B. Metrics

We use the rainfall intensity thresholds 0.5, 2, 5, 10, and 30 mm/h to calculate the critical success index (CSI) [44], the heidke skill score (HSS) [17], and the probability of detection (POD) [16]. We first convert the pixel values in prediction and ground-truth radar images to 0 or 1 by threshold τ mm/h. In detail, we use (9) or (10) to convert the pixel values to rainfall R. If $R \geq \tau$, the pixel value will be 1. In other cases, the pixel value will be 0. Then, we can calculate TP (prediction $=1$, truth $=1$), FN (prediction $=0$, truth $=1$), FP (prediction $=1$, truth $=0$), and TN (prediction $=0$, truth $=0$) separately. In the end, the HSS, CSI, and POD scores can be calculated as

$$\begin{aligned} \text{HSS} &= \frac{\text{TP} \times \text{TN} - \text{FN} \times \text{FP}}{(\text{TP}+\text{FN})(\text{FN}+\text{TN}) + (\text{TP}+\text{FP})(\text{FP}+\text{TN})} \\ \text{CSI} &= \frac{\text{TP}}{\text{TP}+\text{FN}+\text{FP}} \\ \text{POD} &= \frac{\text{TP}}{\text{TP}+\text{FN}} \end{aligned} \tag{11}$$

respectively.

The peak signal to noise ratio (PSNR), SSIM [39], the GDL [31], MSE, MAE, the Balanced Mean Squared Error (B-MSE) [17], and the Balanced Mean Absolute Error (B-MAE) [17] are also adopted in our experiments. Among them, PSNR, SSIM, and GDL measure the quality of forecast image, CSI, HSS, and POD measure the accuracy of rainfall forecast, MAE, MSE, B-MAE, and B-MSE not only can measure the quality of forecast image, but also can measure the accuracy of rainfall forecast.

TABLE IV
SENSITIVITY ANALYSIS OF THE HYPERPARAMETER (k) USING PREDRNN ON THE HKO-7 DATASET

Model	k	SSIM ↑	MSE ↓
PredRNN	0	0.792	69.312
PredRNN	0.5	0.794	68.225
PredRNN	2	**0.805**	**64.833**
PredRNN	4	0.794	67.735
PredRNN	20	0.784	72.748

C. Implementation Details

To make fair comparisons, we apply the same experimental settings for all models. All models use a similar structure like ConvLSTM (see Fig. 1), which is stacked with three RNN layers. We use B-MSE + B-MAE [17] to solve the rainfall imbalance problem of the datasets. On this basis, we introduce FFL for precipitation nowcasting. We multiply the loss by a constant 0.001 to make the model converge. We use AdamHD [45] as the optimizer and set the initial learning rate to 0.0005. In the training phase, the mini-batch is set to 4, the training process is stopped after 30 epochs for both HKO-7 and DWD-12 datasets.

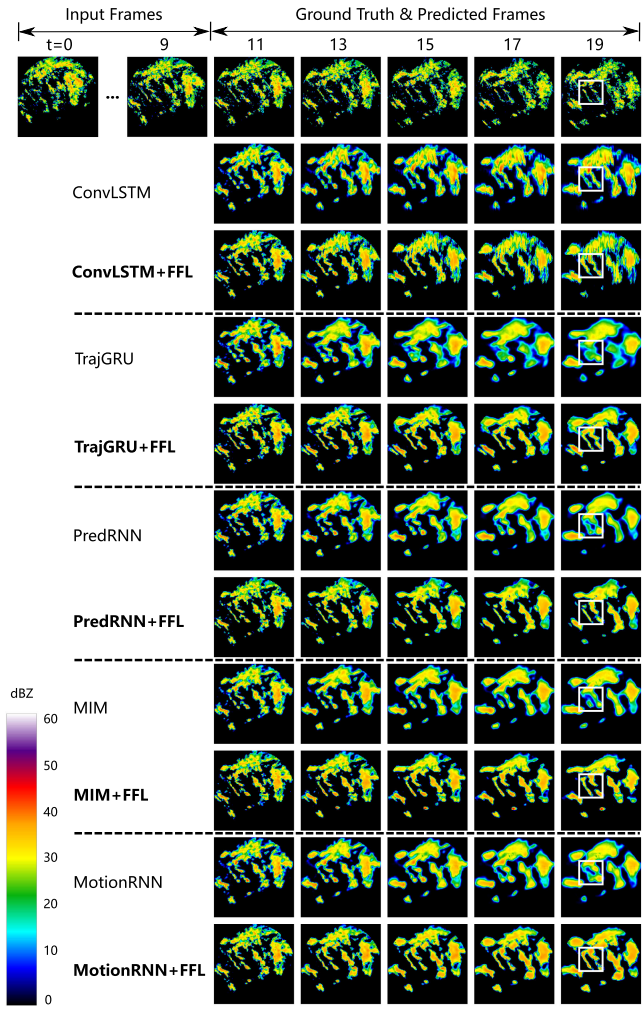

Fig. 4. Qualitative comparison about FFL on the HKO-7 dataset.

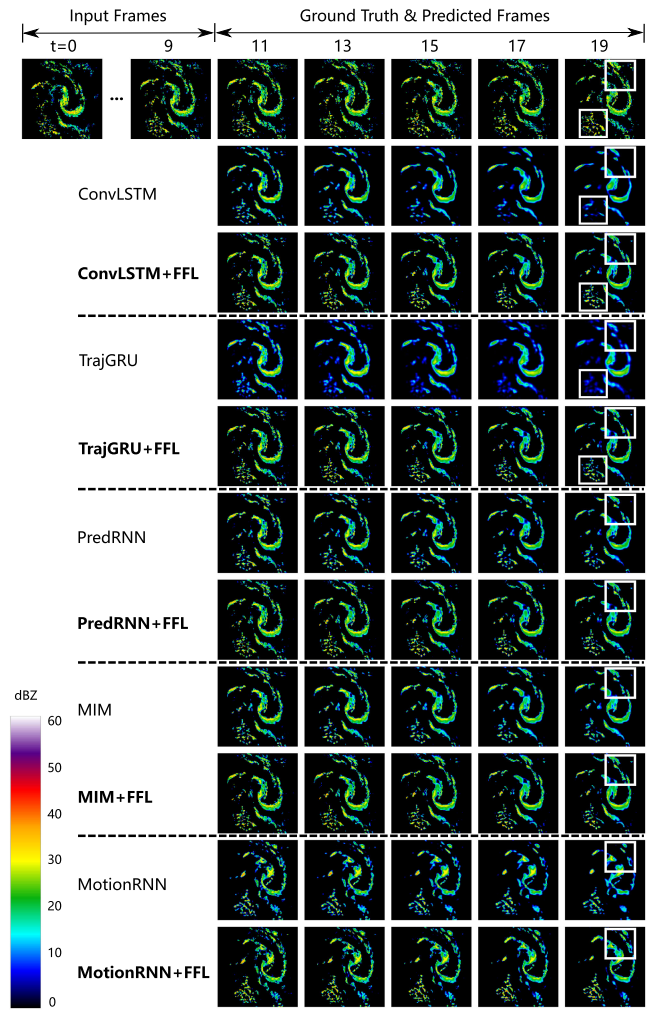

Fig. 5. Qualitative comparison about FFL on the DWD-12 dataset.

We take 20 consecutive frames as a batch, the first 10 frames are used as observations, and the last 10 frames are used to validate model predictions. The channel of the hidden state is 64. To quickly verify the proposed loss function, we resize the radar images of the two datasets to 120×120. Both the convolutional layer and the RNN layer have a kernel size of 3. All experiments are tested on NVIDIA Tesla A100 GPU.

D. Quantitative Analysis

Hyperparameter: We show the change of model performance when FFL takes different values of k in Table IV. We use PredRNN as the base model, training on the HKO-7 dataset. The experimental results demonstrate that with the increase of k, the model performance gradually increases but then decreases, where $k = 2$ is the best. Although the introduction of the FFL loss makes the training process focus on those hard-to-predict frames, giving excessive weight to difficult frames will cause the model to fail to learn the motion trend of the entire sequence, which is the reason why the model performance degrades when k is large.

We conduct experiments on five advanced rainfall prediction models from 2015 to 2021, namely ConvLSTM [16], TrajGRU [17], PredRNN [18], MIM [19], and MotionRNN [20]. We conduct comparative experiments on many metrics for these models with and without FFL. As shown in Tables II and III, the number of parameters of these models has become larger and larger, occupying more and more memory resources and training time, while FFL hardly increases the memory and time consumption of training models but has a huge improvement for each model. Specifically, with the inclusion of FFL, the MAE of MotionRNN decreases by 10.9% from 159.424 to 142.075, and the POD ($R \geqslant 30$) of MotionRNN increases by 51.9% from 0.129 to 0.196 in the DWD-12 dataset. What's more shocking is that FFL improves the performance of models to such an extent that using it on the basic model will result in the basic model (ConvLSTM) with a stronger predictive ability than the advanced model (MotionRNN).

E. Qualitative Analysis

Overall, from Figs. 4 and 5, we see that FFL mitigates the blurriness of predicted images, and even the last frame still retains a high degree of clarity. This is consistent with our loss design philosophy, focusing on those frames that are not easy to

predict. In addition, the models with FFL can learn more details in the prediction. For example, as indicated by the white box in Fig. 4, the models without FFL except ConvLSTM predict a deformed shape of the cloud, but the models using FFL all can accurately predict. As indicated by the white box in Fig. 5, with the help of FFL, the most basic ConvLSTM and TrajGRU models can also predict clear and more detailed images like PredRNN and MIM. Furthermore, with the help of FFL, it can make PredRNN, MIM, and MotionRNN learn more details and have a stronger predictive ability even when the prediction results of PredRNN, MIM, and MotionRNN are already good enough.

VI. Conclusion

In this article, given that current mainstream RNN models for short-term precipitation prediction produce blurry images when making long-term predictions, we propose FFL to focus on those frames that are not easy to be predicted. The design inspiration of FFL comes from FL, which is designed for the problem of different sample classification difficulties in the object detection task. We extend FL to radar sequence prediction tasks and propose FF-MSE and FF-MAE by combining MSE and MAE. We have done exhaustive experiments for FFL with five popular models on two large-scale radar echo datasets. It exhibits good characteristics that FFL greatly improves the performance of common RNN models without introducing additional training overhead.

Our proposed loss function alleviates the prediction blur problem to some certain extent, but the last predicted frame still has a huge gap with the real frame. Recent works are using generative adversarial networks (GANs) and variational autoencoder networks (VAEs) for video prediction, which model future uncertainty by introducing random factors into the training process. Using these models as base models rather than RNN models may further improve prediction performance. In addition to focusing on the improvement of the model structure level, we should also take into account the influencing factors at the data level. Actually, precipitation is a complex microphysical process that is affected by many factors such as humidity, temperature, and topography in the environment. It is unreasonable to consider only a single radar mode without considering changes of other elements in the atmospheric system. Therefore, the precipitation prediction model based on multimodal data fusion deserves further research in the future.

References

[1] S. Ravuri et al., "Skilful precipitation nowcasting using deep generative models of radar," *Nature*, vol. 597, no. 7878, pp. 672–677, 2021.

[2] C. K. Sønderby et al., "MetNet: A neural weather model for precipitation forecasting," 2020, *arXiv:2003.12140*.

[3] M. Montopoli, F. S. Marzano, E. Picciotti, and G. Vulpiani, "Spatially-adaptive advection radar technique for precipitation mosaic nowcasting," *IEEE J. Sel. Topics Appl. Earth Observ. Remote Sens.*, vol. 5, no. 3, pp. 874–884, Jun. 2012.

[4] J. Zhang, Y. Zheng, and D. Qi, "Deep spatio-temporal residual networks for citywide crowd flows prediction," in *Proc. AAAI Conf. Artif. Intell.*, 2017, pp. 1655–1661.

[5] F. Chirigati, "Accurate short-term precipitation prediction," *Nature Comput. Sci.*, vol. 1, no. 11, pp. 709–709, 2021.

[6] M. R. Ehsani, A. Zarei, H. V. Gupta, K. Barnard, E. Lyons, and A. Behrangi, "NowCasting-Nets: Representation learning to mitigate latency gap of satellite precipitation products using convolutional and recurrent neural networks," *IEEE Trans. Geosci. Remote Sens.*, vol. 60, pp. 1–21, 2022, Art. no. 4706021.

[7] P. Bauer, A. Thorpe, and G. Brunet, "The quiet revolution of numerical weather prediction," *Nature*, vol. 525, no. 7567, pp. 47–55, 2015.

[8] A. C. Lorenc, "Analysis methods for numerical weather prediction," *Quart. J. Roy. Meteorological Soc.*, vol. 112, no. 474, pp. 1177–1194, 1986.

[9] R. Kimura, "Numerical weather prediction," *J. Wind Eng. Ind. Aerodynamics*, vol. 90, no. 12–15, pp. 1403–1414, 2002.

[10] R. Reinoso-Rondinel, M. Rempel, M. Schultze, and S. Trömel, "Nationwide radar-based precipitation nowcasting–A localization filtering approach and its application for Germany," *IEEE J. Sel. Topics Appl. Earth Observ. Remote Sens.*, vol. 15, pp. 1670–1691, 2022.

[11] C. Luo, X. Li, and Y. Ye, "PFST-LSTM: A spatiotemporal LSTM model with pseudoflow prediction for precipitation nowcasting," *IEEE J. Sel. Topics Appl. Earth Observ. Remote Sens.*, vol. 14, pp. 843–857, 2021.

[12] J. Jing, Q. Li, X. Peng, Q. Ma, and S. Tang, "HPRNN: A hierarchical sequence prediction model for long-term weather radar echo extrapolation," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process.*, 2020, pp. 4142–4146.

[13] P. Cheung and H. Yeung, "Application of optical-flow technique to significant convection NowCast for terminal areas in Hong Kong," in *Proc. 3rd WMO Int. Symp. Nowcasting Very Short-Range Forecasting*, 2012, pp. 6–10.

[14] W.-C. Woo and W.-K. Wong, "Operational application of optical flow techniques to radar-based rainfall nowcasting," *Atmosphere*, vol. 8, 2017, Art. no. 48.

[15] L. Tian, X. Li, Y. Ye, P. Xie, and Y. Li, "A generative adversarial gated recurrent unit model for precipitation nowcasting," *IEEE Geosci. Remote Sens. Lett.*, vol. 17, no. 4, pp. 601–605, Apr. 2020.

[16] X. Shi, Z. Chen, H. Wang, D. Y. Yeung, W. K. Wong, and W. C. Woo, "Convolutional LSTM network: A machine learning approach for precipitation nowcasting," in *Proc. Adv. Neural Inf. Process. Syst.*, 2015, vol. 28, pp. 802–810.

[17] X. Shi et al., "Deep learning for precipitation nowcasting: A benchmark and a new model," in *Proc. Adv. Neural Inf. Process. Syst.*, 2017, vol. 30, Art. no. 5618.

[18] Y. Wang et al., "PredRNN: Recurrent neural networks for predictive learning using spatiotemporal LSTMs," in *Proc. Adv. Neural Inf. Process. Syst.*, 2017, vol. 30, pp. 879–888.

[19] Y. Wang, J. Zhang, H. Zhu, M. Long, J. Wang, and P. S. Yu, "Memory in memory: A predictive neural network for learning higher-order non-stationarity from spatiotemporal dynamics," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2019, pp. 9154–9162.

[20] H. Wu, Z. Yao, J. Wang, and M. Long, "MotionRNN: A flexible model for video prediction with spacetime-varying motions," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2021, pp. 15435–15444.

[21] Y. Wang et al., "PredRNN: A recurrent neural network for spatiotemporal predictive learning," *IEEE Trans. Pattern Anal. Mach. Intell.*, to be published, doi: 10.1109/TPAMI.2022.3165153.

[22] Y. Wang, Z. Gao, M. Long, J. Wang, and P. S. Yu, "PredRNN++: Towards a resolution of the deep-in-time dilemma in spatiotemporal predictive learning," in *Proc. 35th Int. Conf. Mach. Learn.*, 2018, pp. 5123–5132.

[23] Y. Wang, L. Jiang, M. H. Yang, L. J. Li, M. Long, and L. Fei-Fei, "Eidetic 3D LSTM: A model for video prediction and beyond," in *Proc. Int. Conf. Learn. Representations*, 2018.

[24] Z. Lin, M. Li, Z. Zheng, Y. Cheng, and C. Yuan, "Self-attention ConvLSTM for spatiotemporal prediction," in *Proc. AAAI Conf. Artif. Intell.*, 2020, pp. 11531–11538.

[25] H. Lin, Z. Gao, Y. Xu, L. Wu, L. Li, and S. Z. Li, "Conditional local convolution for spatio-temporal meteorological forecasting," in *Proc. AAAI Conf. Artif. Intell.*, 2021, pp. 7470–7478.

[26] G. Ayzel, T. Scheffer, and M. Heistermann, "RainNet v1.0: A convolutional neural network for radar-based precipitation nowcasting," *Geoscientific Model Develop.*, vol. 13, no. 6, pp. 2631–2644, 2020.

[27] K. Trebing, T. Stanczyk, and S. Mehrkanoon, "SmaAt-UNet: Precipitation nowcasting using a small attention-Unet architecture," *Pattern Recognit. Lett.*, vol. 145, pp. 178–186, 2021.

[28] X. Pan et al., "Improving nowcasting of convective development by incorporating polarimetric radar variables into a deep-learning model," *Geophysical Res. Lett.*, vol. 48, no. 21, 2021, Art. no. e2021GL095302.

[29] L. Han, H. Liang, H. Chen, W. Zhang, and Y. Ge, "Convective precipitation nowcasting using U-Net model," *IEEE Trans. Geosci. Remote Sens.*, vol. 60, pp. 1–8, 2022, Art. no. 4103508.

[30] S. Hochreiter and J. Schmidhuber, "Long short-term memory," *Neural Computation*, vol. 9, no. 8, pp. 1735–1780, 1997.
[31] M. Mathieu, C. Couprie, and Y. LeCun, "Deep multi-scale video prediction beyond mean square error," in *Proc. Int. Conf. Learn. Representations*, 2016.
[32] Q.-K. Tran and S.-k. Song, "Computer vision in precipitation nowcasting: Applying image quality assessment metrics for training deep neural networks," *Atmosphere*, vol. 10, no. 5, 2019, Art. no. 244.
[33] B.-Y. Yan et al., "FDNet: A deep learning approach with two parallel cross encoding pathways for precipitation nowcasting," 2021, *arXiv:2105.02585*.
[34] O. Ronneberger, P. Fischer, and T. Brox, "U-Net: Convolutional networks for biomedical image segmentation," in *Proc. Int. Conf. Med. Image Comput. Comput.-Assist. Interv.*, Springer, 2015, pp. 234–241.
[35] S. Agrawal et al., "Machine learning for precipitation nowcasting from radar images," 2019, *arXiv:1912.12132*.
[36] K. Song et al., "Deep learning prediction of incoming rainfalls: An operational service for the city of Beijing China," in *Proc. IEEE Int. Conf. Data Mining Workshops*, 2019, pp. 180–185.
[37] H. Fan, L. Zhu, and Y. Yang, "Cubic LSTMs for video prediction," in *Proc. AAAI Conf. Artif. Intell.*, 2019, pp. 8263–8270.
[38] Z. Ma, H. Zhang, and J. Liu, "MS-RNN: A flexible multi-scale framework for spatiotemporal predictive learning," 2022, *arXiv:2206.03010*.
[39] 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, Apr. 2004.
[40] H. Rezatofighi, N. Tsoi, J. Gwak, A. Sadeghian, I. Reid, and S. Savarese, "Generalized intersection over union: A metric and a loss for bounding box regression," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2019, pp. 658–666.
[41] T.-Y. Lin, P. Goyal, R. Girshick, K. He, and P. Dollár, "Focal loss for dense object detection," in *Proc. Int. Conf. Comput. Vis.*, 2017, pp. 2980–2988.
[42] I. Sutskever, O. Vinyals, and Q. V. Le, "Sequence to sequence learning with neural networks," in *Proc. Adv. Neural Inf. Process. Syst.*, 2014, vol. 27, pp. 3104–3112.
[43] S. Oprea et al., "A review on deep learning techniques for video prediction," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 44, no. 6, pp. 2806–2826, Jun. 2022.
[44] T. Xiong, J. He, H. Wang, X. Tang, Z. Shi, and Q. Zeng, "Contextual Sa-attention convolutional LSTM for precipitation nowcasting: A spatiotemporal sequence forecasting view," *IEEE J. Sel. Topics Appl. Earth Observ. Remote Sens.*, vol. 14, pp. 12479–12491, 2021.
[45] A. G. Baydin et al., "Online learning rate adaptation with hypergradient descent," 2017, *arXiv:1703.04782*.

Zhifeng Ma received the master's degree in applied statistics from Lanzhou University, Lanzhou, China, in 2020. He is currently working toward the Ph.D. degree in computer science and technology with the Harbin Institute of Technology, Harbin, China.

His research interests include deep learning, precipitation nowcasting, and spatiotemporal predictive learning.

Hao Zhang received the Ph.D. degree in information security from the University of Science and Technology of China, Hefei, China, in 2014.

He is currently an Associate Researcher with the Harbin Institute of Technology, Harbin, China. His research interests include deep learning application, federated learning, and pervasive computing.

Jie Liu (Fellow, IEEE) received the Ph.D. degree in electrical engineering and computer science from the University of California, Berkeley, CA, USA, in 2001.

He is currently a Chair Professor with the Harbin Institute of Technology, Shenzhen, China. His research interests include artificial intelligence, control engineering, Internet of Things, and computer system.