Contextual Sa-Attention Convolutional LSTM for Precipitation Nowcasting: A Spatiotemporal Sequence Forecasting View

Taisong Xiong, Jianxing He, Hao Wang, Xiaowen Tang, Zhao Shi, and Qiangyu Zeng

***Abstract*—Precipitation nowcasting is an important tool for nowcasting weather. In recent years, progress has been achieved in some models based on deep learning for precipitation nowcasting. However, these models do not consider the contextual relationships between the input data and the output of a network and their deficiency in capturing the information of prediction objects. To overcome these shortcomings, in this study, we propose a model that performs convolution operation on input data and the output of a Long short-term memory (LSTM) networks. Second, a self-attention operation is added to capture the local and global dependencies of the hidden state of LSTM. The proposed network structure is inserted in an encoding–forecasting network framework and applied to spatiotemporal sequence forecasting. Third, the outputs of the precede sequence are also regarded as the inputs of according LSTM layer and this operation effectively captures temporal feature of sequence data. Comprehensive experiments are conducted on the KTH action dataset and Hong Kong observation 07 radar echo maps dataset. The visual and quantitative prediction results demonstrate the accuracy and efficacy of the proposed model.**

***Index Terms*—Deep learning, long short-term memory (LSTM), precipitation nowcasting, radar echo maps, spatiotemporal sequence forecasting.**

I. Introduction

NOWCASTING convective precipitation has been a focus of weather forecasting for many years. Its goal is to forecast the local weather conditions over a comparatively short period from the present to a few hours ahead. The convective weather can cause severe damage to properties and can pose a threat to human beings. The meteorological disasters can be effectively avoided when the convective weather is accurately forecasted several hours ahead. However, to precisely forecast the convective weather remains a challenge because of its rapid evolution and complex interactions with its surroundings [1]. Numerical weather prediction (NWP) models have been applied to nowcasting convective precipitation for many years and achieved improvements. However, to meet the needs of precipitation nowcasting, NWP models must overcome many challenges [2].

The conventional and representative extrapolation techniques based on radar echo maps are thunderstorm identification, tracking, analysis, and nowcasting [3] and tracking radar echoes by correlation [4]. They achieve success and are widely applied to precipitation nowcasting. However, these techniques have some inherent defects and need to be improved [2]. Simultaneously, optical flow methods have been applied to extrapolate radar echoes and obtain accurate prediction results [5]–[7]. The trajectory and location of a radar echo cannot be precisely predicted because these models utilize several radar echo maps and cannot sufficiently learn the inherent features of each map. Furthermore, the accuracy decreases rapidly when the prediction timespan increases. How to obtain more precise and longer radar echo extrapolation results is a key research in the meteorology community.

In recent years, deep learning [8] has been applied to many fields and attracted increasingly more researchers. These fields include image recognition [9]–[11], sematic segmentation [12], [13] object detection [14], [15], etc. One feature of deep learning is that it needs a lot of data to sufficiently train a network.

At the same time, deep learning has also been applied to natural language processing (NLP) [16], [17] and sequence learning [18], [19]. Long short-term memory (LSTM) [20], [21] can effectively learn the inherent relationships of long distance objects. The LSTM encoder–decoder framework [18] effectively solves the sequence-to-sequence learning problems using temporally concatenated LSTM. An LSTM encoder–predictor model is proposed in [19] to predict the future video sequences. However, only the temporal relationships, and not the spatial coherence of video sequences, are considered for the model in [19] when it predicts the future video frames. Video

Manuscript received February 2, 2021; revised May 24, 2021 and October 9, 2021; accepted November 5, 2021. Date of publication November 17, 2021; date of current version December 15, 2021. This work was supported in part by the National Key Research and Development Plan under Grant 2018YFC1506100, Grant 2018YFC1506102, and Grant 2018YFC1506104, in part by the Scientific Research Foundation of the Education Department of Sichuan Province under Grant 17ZB0096, in part by the Application Basic Research of Sichuan Department of Science and Technology under Grant 2019YJ0316, in part by the Sichuan Science and Technology program under Grant 2019JDJQ0002 and Grant 2022ZDYF1935, in part by the Special Funds for the Central Government to Guide Local Technological Development under Grant 2020ZYD051, in part by the Open Project for the Research Center for meteorological disaster prediction, early warning, and emergency management under Grant ZHYJ17-YB05, and in part by the 2018 China Meteorological Administration Soft Science Independent Project (09). *(Corresponding author: Hao Wang.)*

The authors are with the Chengdu University of Information Technology, Chengdu 610225, China, and also with the Key Laboratory of Atmospheric Sounding, CMA, Chengdu 610225, China (e-mail: xts@cuit.edu.cn; hjx@cuit.edu.cn; wh@cuit.edu.cn; xtang@cuit.edu.cn; sxz@cuit.edu.cn; zqy@cuit.edu.cn).

Digital Object Identifier 10.1109/JSTARS.2021.3128522

sequence owns both temporal and spatial features. Spatiotemporal sequences widely exist in the real world. For example, the traffic flow, video prediction, etc. The mathematical form of spatiotemporal sequence forecasting [22] is given in the following equation:

$$\hat{X}_{t+1:t+T} = \arg\max_{X_{t+1:t+T}} p(X_{t+1:t+L}|X_{1:t}) \tag{1}$$

where $X_{1:t}$ is a spatiotemporal sequence with length of t as a matrix $X_{1:T} = [X_1, X_2, \ldots, X_T]$. $X_t \in R^{K\times M\times N}$, where K denotes the length and $M \times N$ denotes the two dimensions. Spatiotemporal sequence forecast the future L sequence given the current t sequence. Similarly, the radar echo extrapolation is to forecast the future $K(K > 1)$ radar echo maps given the current L echo maps. Therefore, the radar echo extrapolation can also be regarded as spatiotemporal sequence forecasting.

Inspired by the successful application in predicting video sequences [19] for LSTM and the spatial features representation of the convolutional neural network (CNN), a novel convolutional LSTM (ConvLSTM) model that utilizes an encoder–forecaster framework to apply to precipitation nowcasting is proposed in [23]. More improved forecast results, as shown in [23], are obtained by ConvLSTM compared to those of LSTM and traditional algorithm, real-time optical flow by variational methods for echoes of radar [24]. Following ConvLSTM, a deep predictive coding network [25] is proposed for video prediction based on ground truth video frames, where each ConvLSTM layer produces an error term at each time step. The error term is propagated in the entire network. A predictive recurrent neural network [26] is proposed using a unified memory pool to learn spatial appearances and temporal coherence of video sequences. A memory in memory recurrent neural network (RNN) block [27] is proposed that can effectively leveraging differential information between neighboring hidden states to predict the video frames. To effectively represent the location-variant relationship of video sequence, a trajectory GRU (TrajGRU) model is presented in [28] and applied to precipitation nowcasting to obtain better prediction results. However, a TrajGRU model is very complicated and its implementation is not easy. A model combining convolutional and LSTM [29] was proposed to improve the accuracy of wind speed predictions based on WRF. A self-attention (SA) ConvLSTM model was presented in [30] and obtains impressive results to predict the remaining useful life of rolling bearings. A temporal recurrent U-Net model with attention mechanism [31] was given to predict high-resolution rainfall. To predict and compress the video frames, a model integrated ConvLSTM with Gan was presented in [32]. A spatial-temporal gating network [33] with multiple physical properties was proposed for video prediction, including precipitation radar images. Based on ConvLSTM, a spatiotemporal memory unit [34] was proposed to forecasting temperature.

A multiscale LSTM with an attention mechanism model [35] was proposed for monthly precipitation prediction. A ConvLSTM with star-shaped bridge model [36], combing ConvLSTM and dual attention model [37] have demonstrated their effectiveness for precipitation nowcasting. But the extrapolation time of the model in [36] is only 60 min away.

The aforementioned models independently perform operations on the input data, hidden states, and model outputs. However, the input data and the model outputs should have inherent interactions. To effectively represent their relationships and capture the contextualized representation of the input data and model outputs, inspired by the model in [38], we propose a model in this study, which performs convolution operation between the input data of current network and the outputs of its preceding network before they are fed into the LSTM gates. This step can effectively learn the interactive representations of the input data and model outputs. To effectively model the spatiotemporal contextualized relationships of LSTM and improve the effects of models, we first extend the ConvLSTM integrating convolution operation between the input data and model outputs. The convolution operation in the proposed model is known as the contextual convolution operation, the proposed model is referred to as the contextual ConvLSTM (CConvLSTM). Second, to effectively represent the local and global features of hidden state of LSTM, an SA mechanism is introduced in the LSTM and the model is referred to as contextual SA ConvLSTM (CSAConvLSTM). Then, we adopt an encoding-forecasting framework with six layers to apply to spatiotemporal sequence forecasting. Third, to utilize the temporal relationships of sequences, we adopt the outputs of precede sequence's down-sample and up-sample layer to concatenate the outputs of peers of current sequence to regard as the inputs of LSTM. Furthermore, a $1 * 1$ convolutional operation is added to reduce the dimensions of inputs of LSTM and capture the temporal features.

The main contributions of the proposed model are summarized as follows.

1) A convolutional operation is added between the input and the hidden state of LSTM to capture the contextual feature.
2) An SA mechanism is introduced in the LSTM and performed on the hidden state to represent the local and global features of hidden state.
3) The outputs of downsample and upsample layer of precede sequence are concatenated with the outputs of current peers and feeded as the inputs for LSTM layer. To reduce the filter number of inputs of LSTM, a $1 * 1$ convolutional operation is inserted.

The KTH dataset, with six action classes, and Hong Kong observation 07 (HKO-07), which is a radar echo maps dataset, are applied to verify the accuracy and efficiency of the proposed model. The proposed model obtains better prediction and extrapolation results on the KTH and HKO-07 datasets, respectively, compared to some state-of-the-art models based on deep learning.

The rest of this article is organized as follows. The relative work on video prediction is introduced in Section II. The description of the proposed model is given in detail in Section III. Comprehensive experiments which verify the effectiveness of the proposed model are given in Section IV. The conclusion and the future work are given in Section V.

II. Related Work

An end-to-end multilayer LSTM encoder–decoder framework [18] is applied to machine translation whose input data is one dimension. An encoder–decoder–forecaster framework [19] is proposed to decoder the input video frames and predict the future video frames. However, the model in [19] can only learn the temporal coherence of video sequence. To capture the spatial features of video frames, an extension of the model in [19], ConvLSTM [23] is proposed to apply to precipitation nowcasting and obtains better experimental results. Following this work, some models based on CNN and LSTM are proposed successively to predict the video frames. Because of the location-invariant of the convolution operation, TrajGRU with location-variant filter is proposed [28] to effectively capture the spatial features of video frames. The models in [23] and [28] both selected the encoding–forecasting framework to predict the video frames. Upsample and downsample layers are concurrently inserted into the encoding and forecasting networks, respectively. These layers effectively capture the spatial coherence of video sequence. Furthermore, the encoding and forecasting networks share the same weights, which productively decreases the parameter number of the model. To more effectively represent the temporal features, updating the gate of an LSTM and taking the output of the preceding last layer as the input of the first layer, enhances the modeling ability of the models [26], [39].

Some impressive research works were obtained, which were in parallel with our model. A unit model [40] applied to convective precipitation nowcasting. However, the period of its prediction is only 30 min. Integrating attention modules and depthwise separable convolutions with unit, a SmaAt-UNet [41] also applied to precipitation nowcasting on precipitation map and cloud cover datasets. Using the encoder–forecaster framework, an axial attention memory is aggregated and embedded into ConvGRU [42] and applied to weather forecasting. However, the period of radar echo extrapolation is only 60 min. To improve the performance on medium-to-heavy rain prediction, a deep generative model was presented in [43]. The model in [43] obtains impressive results on usefulness and accuracy. However, it is still challenging to get high accuracy on heavy rain prediction at long lead times.

The aforementioned models for video prediction are deterministic. Another type of model for video prediction based on variational autoencoder [44], is stochastic video prediction [45], [46]. The stochastic video prediction models generate the next frame by feeding the preceding video frame back to these models. These models can capture the diverse features of the video at the cost of high computation [47]. However, it is very difficult to obtain a satisfactory result for these models [27]. In our work, we focus on capturing the deterministic features of spatiotemporal sequence using the convolution operation on the input data and model output.

III. Proposed Model

In this section, we propose a model for spatiotemporal sequence forecasting based on ConvLSTM. One benefit of LSTM is that it can effectively avoid gradient vanishing and exploding, which often happen in an RNN model. However, the spatial features are ignored by the traditional LSTM when it is applied to image or video data, while ConvLSTM effectively captures the spatial features of images. In LSTM, we use x, c, and h, which denotes the input, cell state, and model output, respectively. Subsequently, the results of c and h are obtained by the following formula:

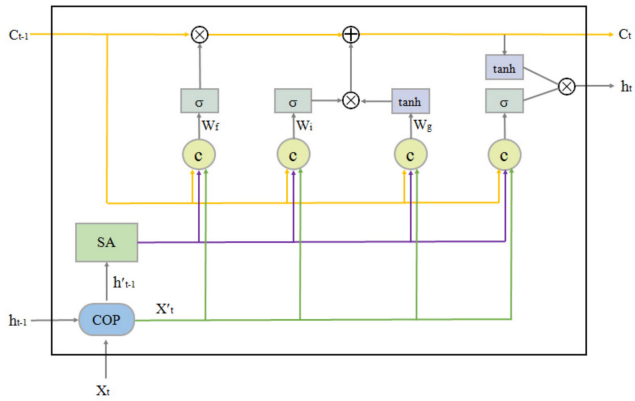


Fig. 1. Architecture of the proposed CSAConvLSTM.

$$\text{LSTM}(x, c_{\text{prev}}, h_{\text{prev}}) = (c, h). \tag{2}$$

Inspired by ConvLSTM and mogrifier LSTM, we propose a model which effortlessly integrates the convolution operation on the input data and output of the LSTM with the ConvLSTM. The added operation can effectively enhance the contextual representation abilities of the proposed model. We named the proposed model contextual ConvLSTM (CConvLSTM) because the contextual features are captured by convolution operation in LSTM. The convolution operation flow between the input data and the output of network is described in Fig. 2.

The ConvLSTM processes two-dimensional data, which are spatial dimensions (row and column). To facilitate the subsequent extension, we provide the standard ConvLSTM formulae [23]

$$i_t = \sigma\left(W^{ix} * X_t + W^{hx} * H_{t-1} + W^{ic} \circ C_{t-1} + b_i\right) \tag{3}$$

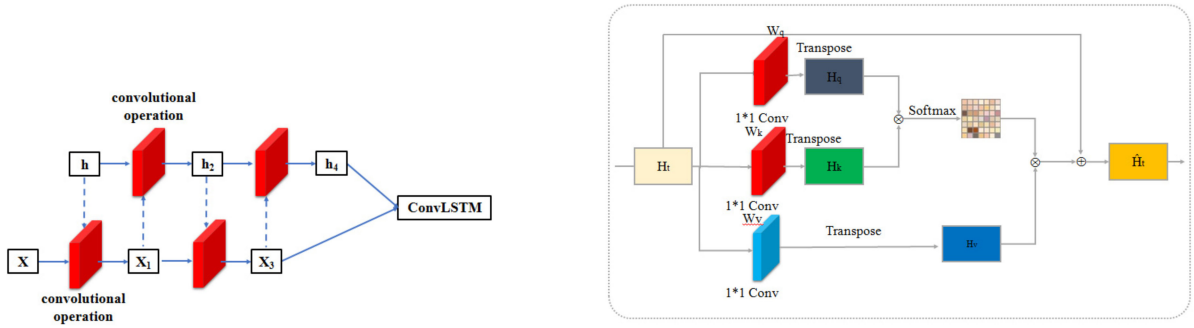
$$f_t = \sigma\left(W^{fx} * X_t + W^{fh} * H_{t-1} + W^{fc} \circ C_{t-1} + b_f\right) \tag{4}$$

$$C_t = f_t \circ C_{t-1} + i_t \circ \tanh\left(W^{cx} * X_t + W^{ch} * H_{t-1} + b_c\right) \tag{5}$$

$$o_t = \sigma\left(W^{ox} * X_t + W^{oh} * H_{t-1} + W^{oc} \circ C_t + b_o\right) \tag{6}$$

$$H_t = o_t \circ \tanh\left(C_t\right) \tag{7}$$

where $*$ denotes the convolution operator, $\circ$ represents the Hadamard product, σ represents the sigmoid function, and W^{**} and b_* are the weight matrices and biases, respectively. The CConvLSTM is an extension of ConvLSTM. As seen from Fig. 1, the input, x and output, h, of the proposed model are first converted into convolutional operations in one alternating fashion and then fed into the ConvLSTM. This operation can learn the contextual features of the video sequence and effectively



(a) (b)

Fig. 2. Architecture of two operation. (a) Contextual operation. (b) SA operation.

capture the inherent features of the spatiotemporal sequence. The later experimental results demonstrate the function. The mathematic formula of the contextual convolution operation is given by

$$\begin{aligned}&\text{ContextualConvLSTM}\left(x, c_{\text{prev}}, h_{\text{prev}}\right)\\&\quad=\text{ConvLSTM}\left(x^{\uparrow}, c_{\text{prev}}, h_{\text{prev}}^{\uparrow}\right)\end{aligned}\tag{8}$$

where $\uparrow$ is the highest index of x^i and h^i_{prev}. We extend the model to exploit convolutions in the transition of inputs and hidden states. The sequence of x^i and h^i_{prev} can be obtained by the following formulae:

$$x^i = 2 \times \sigma\left(W^{ixh} h^{i-1}_{\text{prev}}\right) \times x^{i-2} \quad \text{for} \quad i \in [1 \ldots r] \quad \text{and} \quad i \quad \text{is} \quad \text{odd} \tag{9}$$

$$h^i_{\text{prev}} = 2 \times \sigma\left(W^{ihx} x^{i-1}\right) \times h^{i-2}_{\text{prev}} \quad \text{for} \quad i \in [1 \ldots r] \quad \text{and} \quad i \quad \text{is} \quad \text{even} \tag{10}$$

where the hyper parameter r is the number of "round," $r \in N$, and its value is set at 4 in this study. The advantage of the CConvLSTM is that it can capture the contextual information by adding operation on the input data and the hidden state. Simultaneously, CConvLSTM learns the connections between the input data and the output of the network, and effectively captures the inherent features of video sequence.

An SA operation similar to in(111111111111) followed the contextual convolutional operation is added on the hidden state of LSTM. The operation is represented by the given formula

$$h'_{t-1} = \text{SA}(h_{t-1}) \tag{11}$$

where the process of SA operation is represented as the following formulae and illustrated in Fig. 2

$$h_q = W_q h_{t-1} \tag{12}$$

$$h_k = W_k h_{t-1} \tag{13}$$

$$h_v = W_v h_{t-1}. \tag{14}$$

In the abovementioned three formulae. The $W_q \in R^{C \times \overline{C}}$, $W_k \in R^{C \times \overline{C}}$, and $W_v \in R^{C \times C}$. Their values are learned by using 1×1 convolutions. The value of $\overline{C}$ is set to $C/8$ in our experiments. The size of H_q, H_k, H_v is all $(1, H * W)$, where H and W are the height and width of images.

The model which integrates contextual convolutionary operation and SA operation is referred to CSAConvLSTM. The architecture of the CSAConvLSTM is described in Fig. 1. In Fig. 1, the COP represents the contextual convolution operation and SA represents SA operation. The process of COP and SA is illustrated in Fig. 2. "C" represents the concatenation of feature maps. σ and tanh represent the sigmoid function and tanh function, respectively. The $+$ in Fig. 1 represents the sum of two matrix. And $\times$ represents the element-wise matrix production.

To effectively predict the future frames of radar echo maps, an encoding–forecasting network similar to [28] is improved to apply to spatiotemporal sequence forecasting. The encoding–forecasting network framework is illustrated in Fig. 3. The encoding network is composed of three downsample layers and three RNN layers. The three downsample layers reduce the size of feature maps and capture the spatial features of feature maps, while the three RNN layers learn the temporal features of radar echo sequences. The forecasting network is composed of three upsample layers and three RNN layers. The upsample layers enlarge the size of feature maps and capture the spatial features of feature maps, and the three RNN layers in the forecasting network have the same function as in the encoding network. Consider the temporal relationships of sequence data. We also take the outputs of downsample layers or upsample layers as the inputs of following sequence of peer RNN layers. The inputs of RNN layers are the concatenation of outputs of preceding and current downsample layers or upsample layers. To reduce the dimensions of inputs for RNN layers and capture the temporal features of sequence data, a $1 * 1$ convolutional operation is performed on the inputs to reduce half of feature maps. Because of the limitation of GPU, the SA only is add in the second and third layer of RNN.

IV. Experiments

To verify the effectiveness of the proposed model, we compared the proposed model with three state-of-the-art models, which is based on deep learning. The experiments were conducted on two public natural dataset: the KTH action [48] and HKO-7 [28] datasets. The three models are ConvLSTM,

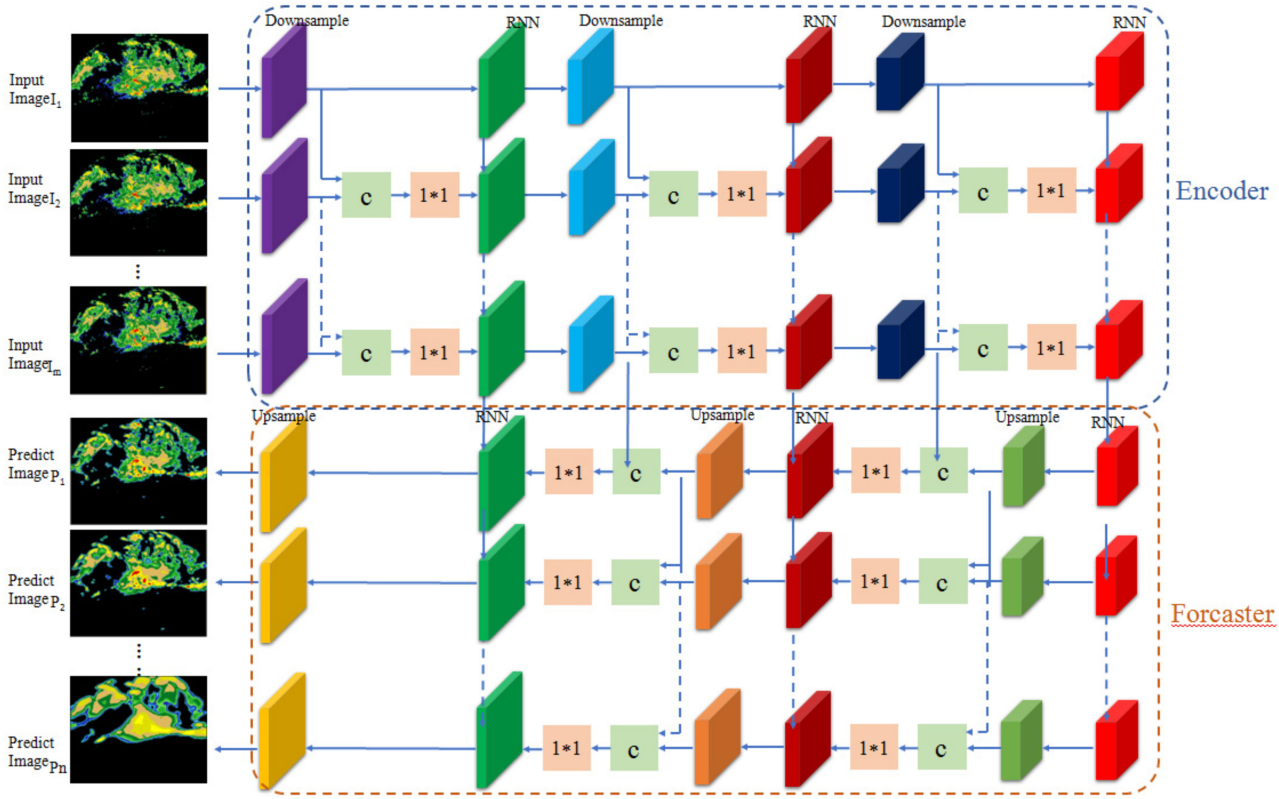


Fig. 3. Architecture of the proposed encoding–forecasting network for spatiotemporal sequence forecasting.

ConvGRU, and trajGRU. The experimental results on KTH action dataset were measured by the following two criteria: the structural similarity index measure (SSIM) and the peak signal to noise ratio (PSNR) [49]. The larger values of SSIM and PSNR indicated better results. Three quantifiable criteria were chosen for the HKO-7 dataset: the probability of detection (POD), false alarm rate (FAR), and critical success index (CSI). The prediction results (we used prediction to represent it) were obtained by all the models. The real information was represented by the ground truth. For the kth class ($k = 1, 2, \ldots, 5$), we calculated the hits (prediction=k, ground truth=k), falsealarms (prediction = k, ground truth != k), and misses (prediction=j, ground truth=k, j!=k). The formulae of the three criteria are defined in (11)–(13). Larger CSI and POD values indicated better results. On the contrary, smaller FAR values indicated better prediction results

$$\text{CSI} = \frac{\text{hits}}{\text{hits} + \text{falsealarms} + \text{misses}} \tag{15}$$

$$\text{POD} = \frac{\text{hits}}{\text{hits} + \text{misses}} \tag{16}$$

$$\text{FAR} = \frac{\text{falsealarms}}{\text{hits} + \text{falsealarms}}. \tag{17}$$

All the models were implemented using Pytorch Framework whose version is 1.1.0 and the experiments were conducted on NVIDIA Tesla P40. The Adam is chosen as the optimizer, and its learning rate is set to 0.0001, and betas are set (0.5, 0.999), and the learning rate is updated every 20 000 or 40 000 in HKO-7 and KTH, respectively.

A. KTH Action Dataset

First, we adopted the KTH action dataset [48] to verify the effectiveness of the proposed model. There were six types of human actions (hand clapping, boxing, jogging, hand waving, running, and waling) in this dataset. These actions were performed by 25 people in four different scenarios. In this experimental process, the batch of all the models was set to 8 and the training process loops to 200 000. To effectively train all the models, a random selection policy was adopted to choose training samples. For the training sets, 1 600 000 subset video sequences with a length of 30 were randomly chosen from the videos of the first 20 persons. The test set was composed of 160 000 subset video sequences, which were randomly chosen from the videos of the last 5 persons. The resolution of each frame was resized into 128×128 pixels. The key to correctly predict long-term frames is to effectively capture the spatiotemporal features of video sequences. The network framework that was adopted is illustrated in Fig. 2. The configurations of network parameters for all the models were same. The kernel size of the three downsample layers in the encoding network was $3 \times 3 \times 8$, $3 \times 3 \times 192$, and $3 \times 3 \times 192$ and had a stride of 2 pixels. The kernel size of the three upsample layers in forecasting network is $3 \times 3 \times 192$, $3 \times 3 \times 192$, and $3 \times 3 \times 8$ and had a stride of 2 pixels. An 1×1 convolution network layer

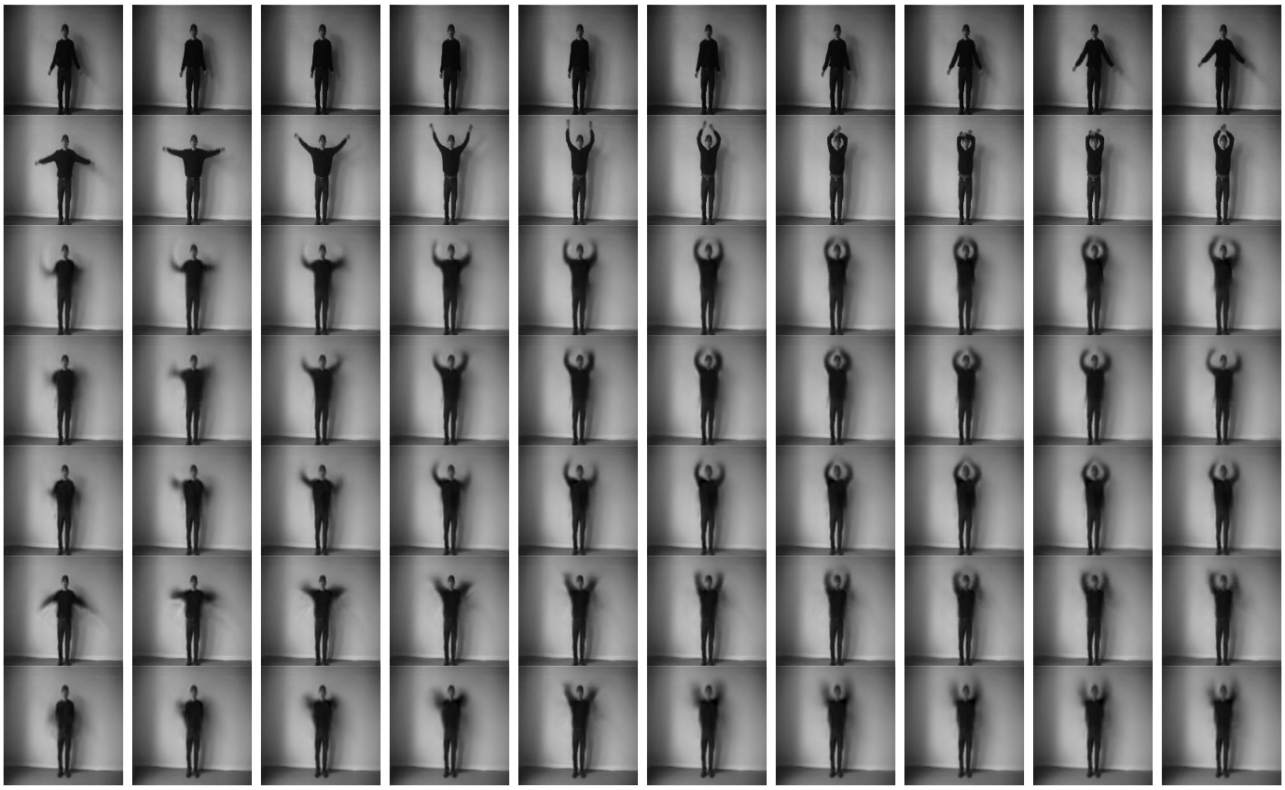


Fig. 4. Third row: ConvLSTM (SSIM=0.903, PSNR=24.91); the forth row: ConvGRU (SSIM=0.913, PSNR=26.39); the fifth row: TrajGRU (SSIM=0.897, PSNR=24.97); the sixth row: CConvLSTM (SSIM=0.924, PSNR=27.63); and the seventh row: CSAConvLSTM (SSIM=0.909, PSNR=25.39).

was added to capture the spatial feature and translates the number of feature maps. The parameters of the corresponding RNN layers in the encoding network were the same to those of the forecasting network. The kernel size of the three RNN layers was $3 \times 3 \times 192$, $3 \times 3 \times 192$, and $3 \times 3 \times 64$ and had a stride of 1. The input sizes of the three RNN layers were identical to those of their outputs.

A sample video sequence prediction result from the KTH is given in Fig. 4. The first row shows the first 10 frames of the video. The second row shows the following ground truth 20 frames, which is the 12th to the 30th frame with an interval of one frame. The last four rows show the prediction results obtained by four models, which correspond to the ground truth shown in the second row. From the prediction results, we can see that the hands of human beings obtained by ConvLSTM, ConvGRU, and TrajGRU are loose and the background is blurred. The proposed model obtains better results than those of the other models. The values of SSIM and PSNR obtained by CConvLSTM are both larger than those of the other models. The results of CSAConvLSTM are second to ConvGRU, but better than ConvLSTM and TrajGRU. It shows that the proposed model effectively capture the spatiotemporal features of video sequences and can more accurately predict the video results than other models.

Another video sequence of a man beginning to walk is shown in Fig. 5. The input 10 frames and the ground truth 10 frames are given in the first and second rows, respectively. The prediction results obtained by all the models are shown in the last four rows. The frame span is similar to the first example of KTH dataset in Fig. 3. The man is blurred and his movement trajectory is not correctly predicted by ConvLSTM, ConvGRU, and TrajGRU. The handkerchief on the man's shoulder is also loos for the three models. However, the proposed model effectively retained the detail of the handkerchief and accurately predicted the trajectory direction of the man. Furthermore, the structures of human joints were also predicted by the proposed model, which indicates that the proposed model can effectively capture the spatial feature of videos. This may be the function of convolution operation on the input and the hidden state in the proposed model. Simultaneously, the larger SSIM and PSNR values obtained by CConvLSTM and CSAConvLSTM showed better prediction results compared to those obtained by the other three models. The visual prediction results showed that the proposed model can effectively learn the spatial and temporal features of the video sequences and predict better results. This demonstrates that the proposed model can effectively learn the details of objects and correctly predict these objects.

To comprehensively verify the effectiveness and robustness of the proposed model. We randomly selected 160 000 video sequences from the test dataset. The quantifiable averages, training time, and memory usage of all the models are given in Table I. The training time is the number of 100 loops with a batch size of 8. The two quantifiable criteria obtained by the proposed model were both larger than those of the other models. This shows that the prediction results obtained by the proposed model are more similar to the true video than those of the other

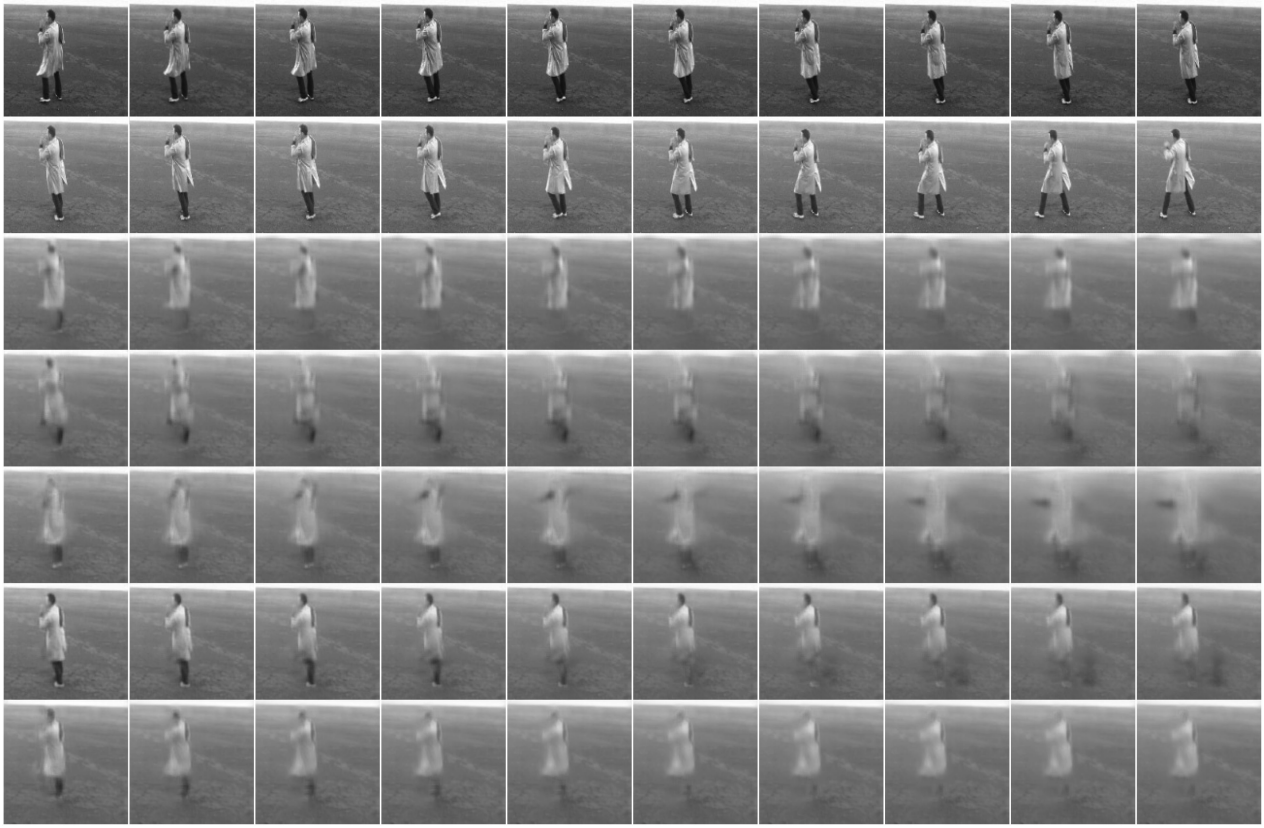


Fig. 5. Third row: ConvLSTM (SSIM=0.800, PSNR=25.01); the forth row: ConvGRU (SSIM=0.777, PSNR=23.78); the fifth row: TrajGRU (SSIM=0.765, PSNR=24.01); the sixth row: CConvLSTM (SSIM=0.827, PSNR=26.36); and the seventh row: CSAConvLSTM (SSIM=0.812, PSNR=25.59).

TABLE I
COMPARISONS OF DIFFERENT CRITERIA FOR KTH DATASET

Algorithm	SSIM	PSNR	Training Time	Memory usage
ConvLSTM	0.839	27.95	126S	8121MB
ConvGRU	0.833	27.60	114S	7569MB
TrajGRU	0.833	27.75	389S	17291MB
CConvLstm	0.861	29.51	205S	16023MB
CSAConvLSTM	0.840	27.91	238S	15643MB

models. Additionally, the proposed model effectively learned the inherent features of videos. The prediction results obtained by ConvGRU were relatively inferior to those of ConvLSTM because ConvGRU is the simplified version of ConvLSTM. Therefore, the training time and memory usage of ConvGRU were both smaller than those of ConvLSTM. Because of the convolution operation of input data and hidden state, the training time and memory usage of the proposed model were larger than those of ConvLSTM and ConvGRU, however, these values were both smaller than those of TrajGRU. This demonstrates that the efficiency of the proposed model is higher than TrajGRU. These experimental results effectively prove the correctness and efficiency of the proposed model.

To advance quantitative comparisons of generating prediction frames, the quantitative criteria of frame-by-frame on test set are shown in Fig. 6. We can see that the prediction results obtained by all the models gradually became substandard with the increment of prediction steps. However, the effects of the proposed model were always superior compared to any of the other models. The effects of the prediction results obtained by TrajGRU and ConvGRU were similar, while ConvLSTM outperformed ConvGRU and TrajGRU. However, compared with ConvLSTM, CConvLSTM obtained better prediction results for every frame. Another model CSAConvLSTM are near to the ConvLSTM. This experiment demonstrated that the proposed model obtained better prediction results than any of the other models. The proposed model effectively captured the spatial and temporal features of video sequence and predicted better results. This may show that the convolution operation of the input data and hidden state for the proposed model plays an important role in learning the spatial and temporal features of the video sequence. Additionally, the proposed model obtained state-of-the-art prediction results on the KTH action dataset compared to other models.

B. HKO-7 Dataset

It is more challenging to predict the trajectory and intensities of radar echo maps because many complicated factors affect their trajectory and intensities. To verify the effect on radar echo extrapolation of the proposed model, we selected a public radar echo dataset, HKO-7, which is collected by HKO. HKO-7 contains radar CAPPI reflectivity images from 2009 to 2015.

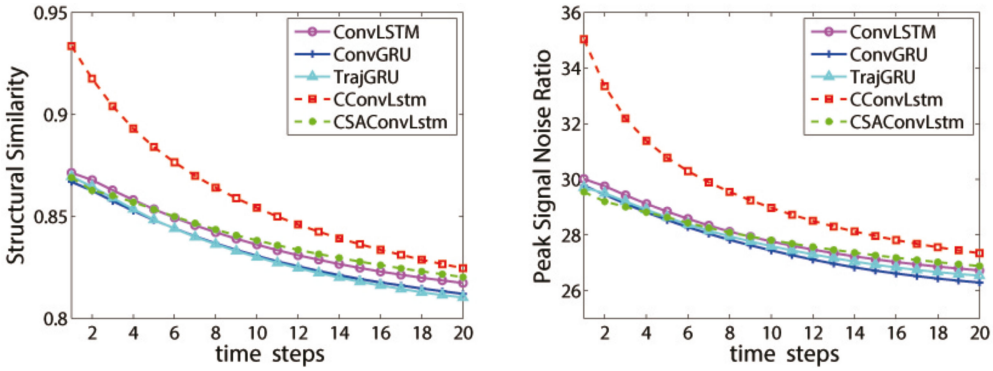


Fig. 6. SSIM and PSNR values of KTH at each period for all the models.

TABLE II
COMPARISONS OF DIFFERENT CRITERIA FOR HKO-07 DATASET

Algorithm	MSE	MAE	B-MSE	B-MAE	Training Time	Memory usage
ConvLSTM	2821	7307	5816	14921	261S	8855MB
ConvGRU	2815	7277	5966	15135	242S	8229MB
TrajGRU	2586	6799	5784	14736	455S	16131MB
CConvLstm	2667	6827	5855	14682	296S	14103MB
CSAConvLSTM	2642	6837	5727	14589	308S	14589MB

The training set is composed of rain sequences from 2009 to 2014. The kernel size of the three successively downsample layers in the encoding network was $7 \times 7 \times 8$ with a stride of 5 pixels, $5 \times 5 \times 64$ with a stride of 3 pixels, and $3 \times 3 \times 192$ with a stride of 2 pixels. The kernel size of the three upsample layers in the forecasting network was $4 \times 4 \times 192$ with a stride of 2 pixels, $5 \times 5 \times 192$ with a stride of 5 pixels, and $7 \times 7 \times 8$ with a stride of 5 pixels. In the training process, the batch size was set to 4 and all the models terminate the training after 100 000 loops. The test set is composed of radar echo maps from the year 2015. To verify the effectiveness of the proposed model, 6042 radar echo sequences were selected. The batch was set to 4 in the radar echo extrapolation experiment. Seven quantitative criteria were adopted for the prediction results. For a comprehensive comparison with the prediction results, we not only selected traditional mean absolute error (MAE) and mean squared error (MSE), but also the balanced MAE (B-MAE) and balanced MSE (B-MSE), which are defined in [28]. The B-MAE and B-MSE provide larger weights for strong radar echo intensity values. The more accurate extrapolation on strong echo intensity values indicates a more accurate prediction of strong convective weather. The accurate forecast of a strong convective weather can effectively decrease the intensity of a meteorological disaster. Therefore, it is a natural law to provide larger weights for strong echo intensity values.

The MSE, MAE, B-MSE, and B-MAE obtained by all models, training time, memory usage of all the models are given in Table II. The training time is the number of times the training model ran 100 iterations. From Table II, we can see that the training time of the proposed models are longer than that of ConvLSTM and ConvGRU because of its convolution operation of the input data and model output. The proposed model obtained a lower MSE, MAE, B-MAE, and B-MSE compared to ConvLSTM and ConvGRU, which indicates that it obtained better prediction results than ConvLSTM and ConvGRU. The prediction results obtained by TrajGRU were slightly better than those of the proposed models for MAE and MSE, but it is second to CSAConvLSTM for B-MAE and B-MSE. These indicate that CSAConvLSTM obtained better results than TrajGRU on hight rainfall intensity. At the same time, the efficiency of TrajGRU was inferior to that of the proposed models.

Visual comparison of the prediction effect of the HKO-07 dataset, a sequence of prediction radar echo frames is visualized in Fig. 7. The first row shows the five input successive echo frames. The second row shows the ground truth frames, which is generated at 6 min, 30 min, 60 min, 90 min, and 120 min. The prediction echo frames corresponding to the ground truth are shown in rows three to six, which were obtained by ConvLSTM, ConvGRU, TrajGRU, and the proposed model. We can see that the three large radar echoes always exist during the evolution process. But the details of the largest echo obtained by the proposed model were well retained compared to other models. Therefore, the proposed model can more accurately predict the shape and intensities of the radar echo than other models. The prediction results show the effectiveness of the proposed model. A longer prediction time, resulted in a gradual decrease of the proportion of the strong echo intensities of the prediction echo frames. Therefore, the proposed model can effectively process long-term variations of echo sequence. This phenomenon demonstrates that the correctness of the predicted strong echo intensities decreases rapidly.

To quantitatively compare the prediction results shown in Fig. 7, we provided the CSI values of the predicted frames obtained by all the models at 6 min, 60 min, and 120 min in Tables III–V , respectively. From the data given in the three tables, we can see that the CSI values obtained by the proposed

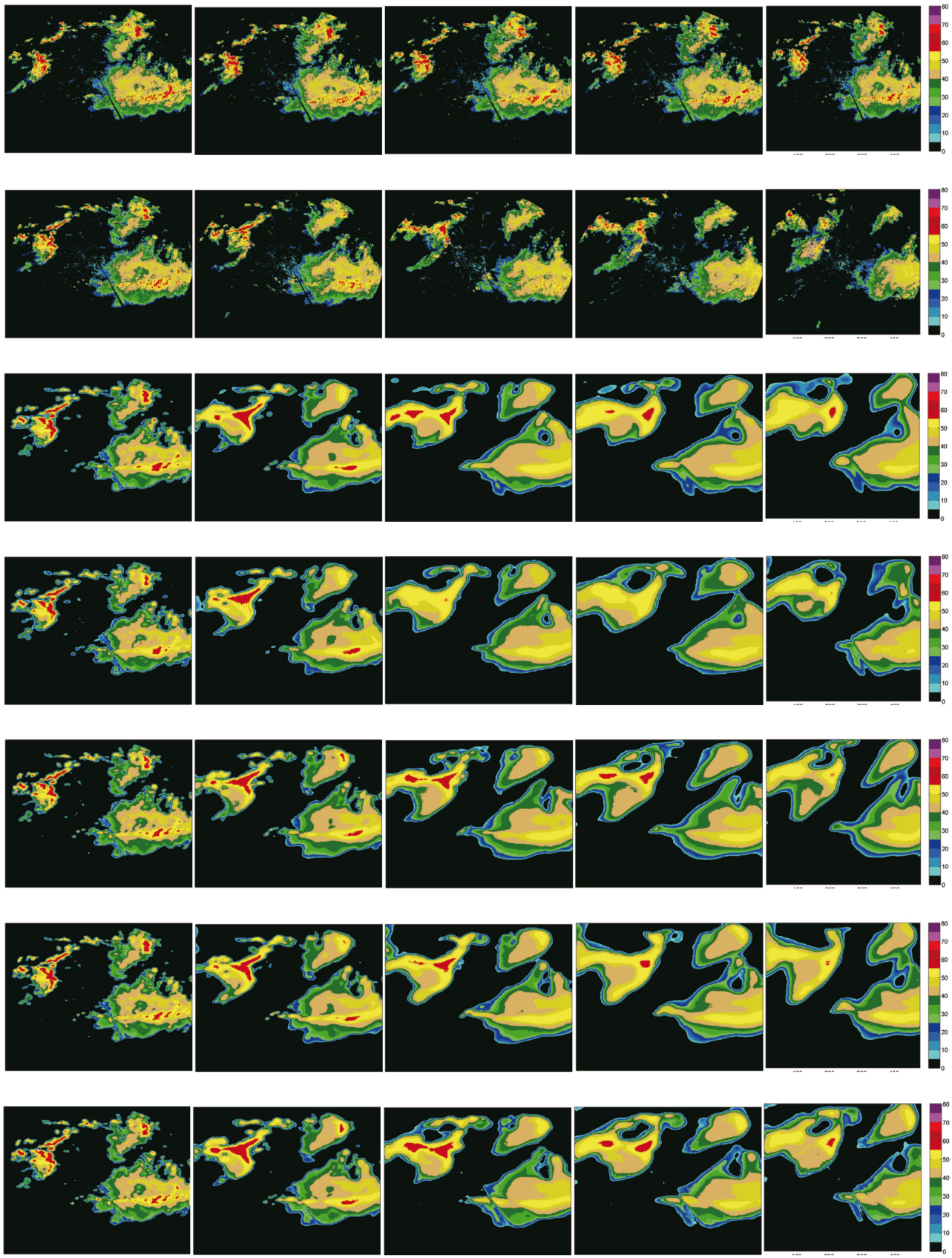


Fig. 7. Visualization of prediction example 1 for HKO-07 dataset.

TABLE III
6 MIN PREDICTION CSI VALUES FOR HKO-07 OF EXAMPLE 1

Algorithm	0.5≤ r<2	2≤ r<5	5≤ r<10	10≤ r<30	r≥30
ConvLSTM	0.8839	0.8428	0.7640	0.6750	0.5641
ConvGRU	0.8788	0.8402	0.7616	0.6769	0.5664
TrajGRU	0.8939	0.8479	0.7846	0.7068	0.5903
CConvLstm	0.8940	0.8505	0.7853	0.7056	0.5884
CSAConvLSTM	0.8896	0.8535	0.7879	0.7009	0.5877

TABLE IV
60 MIN PREDICTION CSI VALUES FOR HKO-07 OF EXAMPLE 1

Algorithm	0.5≤ r<2	2≤ r<5	5≤ r<10	10≤ r<30	r≥30
ConvLSTM	0.6341	0.5805	0.4998	0.4079	0.1997
ConvGRU	0.6498	0.5911	0.5125	0.3885	0.1794
TrajGRU	0.6503	0.6065	0.5441	0.4594	0.2433
CConvLstm	0.6543	0.6083	0.5227	0.4029	0.2030
CSAConvLSTM	0.6483	0.6037	0.5298	0.3942	0.1985

TABLE V
120 MIN PREDICTION CSI VALUES FOR HKO-07 OF EXAMPLE 1

Algorithm	0.5≤ r<2	2≤ r<5	5≤ r<10	10≤ r<30	r≥30
ConvLSTM	0.5414	0.4713	0.3518	0.2598	0.0312
ConvGRU	0.5442	0.4910	0.4063	0.2736	0.0631
TrajGRU	0.5795	0.5442	0.4563	0.3137	0.0712
CConvLstm	0.6155	0.5417	0.4340	0.3044	0.0613
CSAConvLSTM	0.5642	0.5216	0.4190	0.2803	0.0568

model is higher than those of ConvLSTM and ConvGRU. Additionally, the CSI values obtained by the proposed model was only slightly lower than those of TrajGRU for the two highest intensity levels. However, the training efficiency and memory usage of the proposed model is better than those of TrajGRU.

The second prediction example is given in Fig. 8. There are two main radar echoes in the echo maps. The evolution process shows that two echoes are gradually merged into one. The layouts of Fig. 8 are similar to that of the Fig. 6. From the prediction results shown in Fig. 8, we can see that all the models accurately predict the unification of the radar echoes. For the visual effects, we can see from Fig. 8 that prediction results obtained by the proposed model are more similar to the ground truth echo maps. Furthermore, the distribution of the strong echo intensities and the echo details for the prediction results, obtained by the proposed model, are well retained. The prediction results demonstrate that the proposed model captures the spatial and temporal inherence and predicts better results than other models. To further verify the correctness and effectiveness of the proposed model, we provided the quantitative criterion CSI values for the prediction echo maps shown in Fig. 8 in Tables VI–VIII. The effectiveness of the proposed model outperforms ConvLSTM and ConvGRU for all the rain levels. Additionally, the CSI values obtained by the proposed model was only slightly lower than TrajGRU on one or two rain levels at some time-span. These prediction results prove that the proposed model effectively captures the temporal dynamic features and the spatial contextual relationships of the radar echo sequence.

For a comprehensive comparison of the experimental results, the rainfall can be divided into several levels. We selected the same rule as in [28] and divided the rainfall intensity into five levels, between 0.5 and 2, 2 and 5, 5 and 10, 10 and 30, and above 30, which are classified as the first, second, third, fourth, and, and fifth classes, respectively. For the test dataset,

TABLE VI
6 MIN PREDICTION CSI VALUES FOR HKO-07 OF EXAMPLE 2

Algorithm	0.5≤ r<2	2≤ r<5	5≤ r<10	10≤ r<30	r≥30
ConvLSTM	0.8930	0.8332	0.6688	0.6020	0.4986
ConvGRU	0.8800	0.8283	0.6578	0.5850	0.4823
TrajGRU	0.8981	0.8362	0.6766	0.6304	0.5486
CConvLstm	0.8992	0.8385	0.6821	0.6370	0.5366
CSAConvLSTM	0.8999	0.8427	0.6912	0.6307	0.5392

TABLE VII
60 MIN PREDICTION CSI VALUES FOR HKO-07 OF EXAMPLE 2

Algorithm	0.5≤ r<2	2≤ r<5	5≤ r<10	10≤ r<30	r≥30
ConvLSTM	0.7630	0.6676	0.4608	0.4268	0.2787
ConvGRU	0.7673	0.6634	0.4614	0.3649	0.2720
TrajGRU	0.7718	0.6611	0.4737	0.4491	0.3102
CConvLstm	0.7910	0.6872	0.4799	0.4282	0.3057
CSAConvLSTM	0.7573	0.6429	0.4496	0.3827	0.2925

TABLE VIII
120 MIN PREDICTION CSI VALUES FOR HKO-07 OF EXAMPLE 2

Algorithm	0.5≤ r<2	2≤ r<5	5≤ r<10	10≤ r<30	r≥30
ConvLSTM	0.6783	0.5962	0.3937	0.3165	0.2989
ConvGRU	0.6895	0.5939	0.3967	0.2675	0.2481
TrajGRU	0.7021	0.6369	0.4326	0.3445	0.3021
CConvLstm	0.6937	0.6464	0.4543	0.3201	0.3107
CSAConvLSTM	0.6530	0.6412	0.4388	0.3112	0.2245

TABLE IX
COMPARISON CSI FOR HKO

Algorithm	0.5≤ r<2	2≤ r<5	5≤ r<10	10≤ r<30	r≥30
ConvLSTM	0.5490	0.4692	0.3650	0.2780	0.1820
ConvGRU	0.5468	0.4686	0.3623	0.2748	0.1775
TrajGRU	0.5553	0.4816	0.3831	0.2937	0.1926
CConvLSTM	0.5548	0.4774	0.3761	0.2864	0.1908
CSAConvLSTM	0.5575	0.4815	0.3785	0.2909	0.1938

we selected 6042 echo sequences that were generated in the year 2015. Three common precipitation nowcasting metrics, CSI, POD, and FAR, were adopted to quantify the experimental results. The values of the quantitative criteria obtained by all the models are given in Tables IX–XI. The CSI values obtained by CConvLSTM were higher than that of ConvLSTM and ConvGRU, and slightly lower than that of TrajGRU. The CSI values obtained by CSAConvLSTM are all higher than those of CConvLSTM. Furthermore, The POD values obtained by CConvLSTM on the largest raining level were only inferior to that of ConvLSTM. The POD values obtained by the CSAConvLSTM are all highter than those of CConvLSTM. The FAR values obtained by CConvLSTM were lower than that of ConvLSTM and ConvGRU, and slightly higher than that of TrajGRU. The FAR values obtained by CSAConvLSTM are litter than those of CConvLSTM. The three criteria indicate that the results obtained by CSAConvLSTM are better than CConvLSTM. From the data given in the three tables, we can see that there is not only one model that can obtained the best results for the three criteria. Comprehensive comparisons of the effectiveness, memory usage, and correctness show that the proposed model is relatively superior to the other three models when they are applied to precipitation nowcasting. The experimental results show that the accuracy, generality, and efficiency of the proposed model were better than the other models.

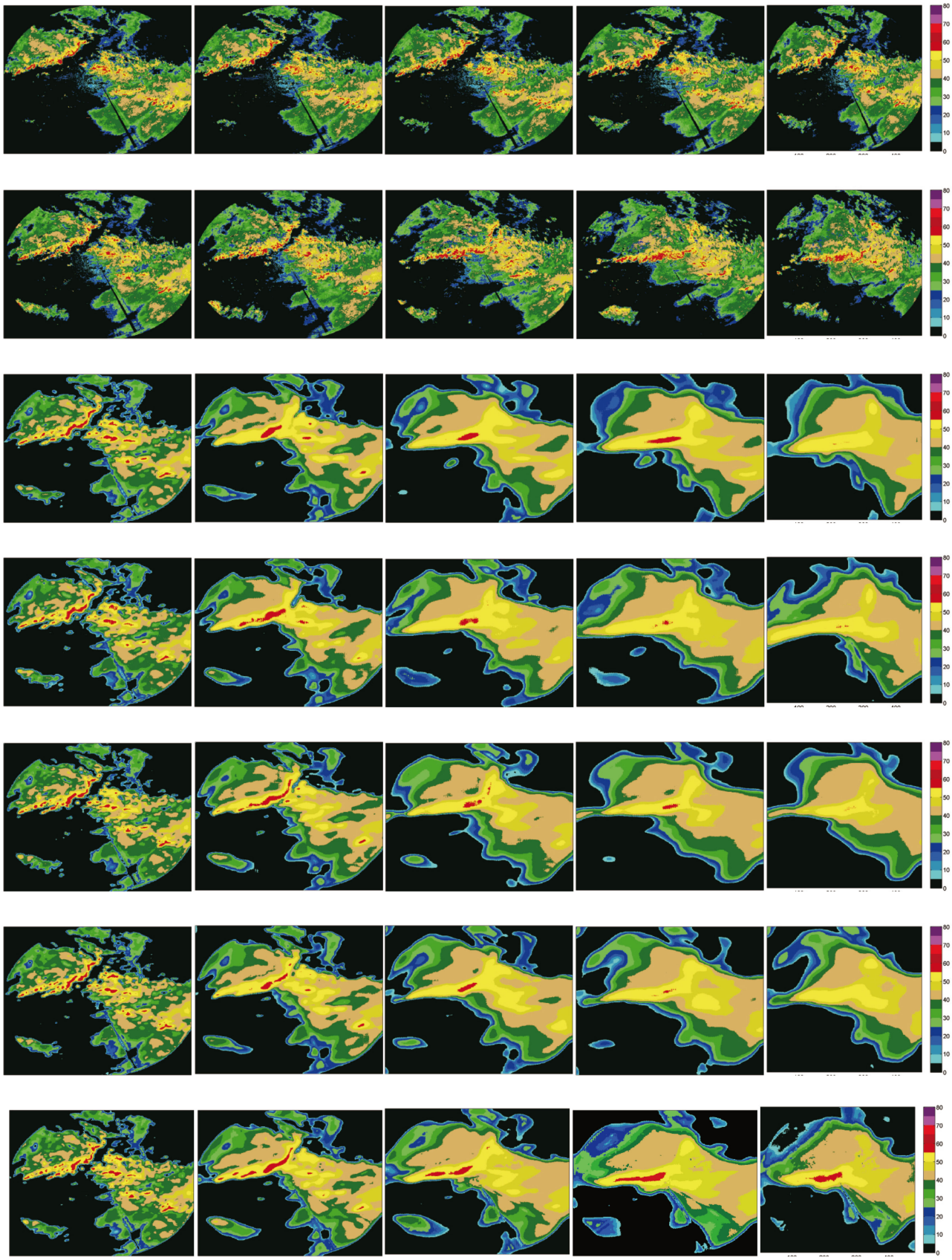


Fig. 8. Visualization of prediction example 2 for HKO-07 dataset.

TABLE X
COMPARISON POD FOR HKO

Algorithm	0.5≤ r<2	2≤ r<5	5≤ r<10	10≤ r<30	r≥30
ConvLSTM	0.7623	0.7893	0.7732	0.6638	0.4167
ConvGRU	0.7525	0.7927	0.7764	0.6675	0.3737
TrajGRU	0.7483	0.7724	0.7473	0.6272	0.3615
CConvLstm	0.7501	0.7739	0.7554	0.6466	0.4004
CSAConvLSTM	0.7521	0.7818	0.7684	0.6649	0.4155

TABLE XI
COMPARISON FAR FOR HKO

Algorithm	0.5≤ r<2	2≤ r<5	5≤ r<10	10≤ r<30	r≥30
ConvLSTM	0.3430	0.4686	0.5954	0.6812	0.7635
ConvGRU	0.3372	0.4700	0.5987	0.6855	0.7560
TrajGRU	0.3240	0.4458	0.5661	0.6521	0.7231
CConvLstm	0.3266	0.4519	0.5773	0.6667	0.7448
CSAConvLSTM	0.3239	0.4504	0.5783	0.6655	0.7452

V. CONCLUSION

In this study, we proposed a model to apply to spatiotemporal sequence forecasting that performs a convolution operation on the previous output of the network and the current input to capture the contextual relationships of the data. This operation provides a mechanism for the input data to directly interact with the previous context. At the same time, an SA operation is performed on the hidden state of LSTM. Furthermore, the outputs of preceding downsample layers and upsample layers are also feeded as the inputs of the current RNN layers. The proposed model is applied to video prediction and can automatically learn the inherent spatiotemporal features of video sequences. Two public datasets were used to verify the effectiveness and correctness of the proposed model. One is a natural video sequence prediction and the other is the extrapolation of the radar echo. Nine criteria were chosen to quantify the experimental results of video prediction. Comprehensive visual and quantitative experimental results illustrated that the proposed model is superior to some state-of-the-art models based on deep learning.

Although some improvement were achieved for the proposed model, however, further improvements are still required in future studies. For the natural video prediction sequences, some details of the objects were lost and resulted in a deterioration in the prediction results with the increment in the video frames. The radar echo extrapolation requires a lot of improvement for the prediction results. First, the quantitative criteria were relatively low for effective weather forecasting applications, especially for strong echo intensities. Second, the accuracy of the prediction results decreased with longer prediction time spans.

To effectively solve the aforementioned problems, some scenarios need be considered in our future studies. The reason for a lower accuracy during the prediction of strong echo intensities may be the average properties of convolution operations. Generative adversarial networks (GANs) [50]–[52] can preserve the sharpness of the prediction video sequences. Therefore, GANs may retain the strong echo intensities when they are applied to the radar echo extrapolation. To improve the accuracy of the precipitation nowcasting, the model should effectively capture the temporal coherence in long-term echo sequences. The transformer [53], [54] can effectively capture the long distance semantic relationships between worlds when it is applied to NLP. Transformers [55], [56] have successfully applied to computer vision since 2020. At present, transformers [57] are also used for video understanding. However, there are no reports about transformers applied to video prediction. Therefore, video prediction and weather forecasting based on transformers are one of our future research fields. The transformer may be applied to the radar echo extrapolation to effectively capture the long term temporal relationships among the echo sequences. Therefore, the transformer may obtain better long time prediction results.

ACKNOWLEDGMENT

The authors would like to the anonymous reviewers for their valuable suggestions and comments, which greatly helped us to improve both the presentation quality and the technical content of this article. The authors would also like to thank Editage (www.editage.cn) for English language editing which have helped to improve this article.

REFERENCES

[1] K. Zhou, Y. Zheng, B. Li, W. Dong, and X. Zhang, "Forecasting different types of convective weather: A deep learning approach," *J. Meteorological Res.*, vol. 33, no. 5, pp. 797–809, Oct. 2019.
[2] J. Sun *et al.*, "Use of NWP for nowcasting convective precipitation: Recent progress and challenges," *Bull. Amer. Meteorological Soc.*, vol. 95, no. 95, pp. 409–426, 2014.
[3] M. Dixon and G. Wiener, "TITAN: Thunderstorm identification, tracking, analysis, and nowcasting a radar-based methodology," *J. Atmos. Ocean. Technol.*, vol. 10, pp. 785–797, 1993.
[4] R. E. Rinehart and E. T. Garvey, "Three-dimensional storm motion detection by conventional weather radar," *Nature*, vol. 273, no. 5660, pp. 287–289, 1987.
[5] P. Cheung and H. Y. 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.
[6] U. Germann and I. Zawadzki, "Scale-dependence of the predictability of precipitation from continental radar images," *Part I: Description. Methodol., Monthly Weather Rev.*, vol. 130, no. 12, pp. 2859–2873, 2002.
[7] H. Sakaino, "Spatio-temporal image pattern prediction method based on a physical model with timevarying optical flow," *IEEE Trans. Geosci. Remote Sens.*, vol. 51, no. 5, pp. 3023–3036, May 2013.
[8] Y. LeCun, Y. Bengio, and G. Hinton, "Deep learning," *Nature*, vol. 512, no. 7553, pp. 436–444, 2015.
[9] A. Krizhevsky, I. Sutskever, and G. Hinton, "Imagenet classification with deep convolutional neural networks," in *Proc. Adv. Neural Inf. Process. Syst.*, Dec. 2012, pp. 1097–1105.
[10] L. Dong *et al.*, "Very high resolution remote sensing imagery classification using a fusion of random forest and deep learning technique subtropical area for example," *IEEE J. Sel. Topics Appl. Earth Observ. Remote Sens.*, vol. 13, no. 2, pp. 113–127, Feb. 2020.
[11] N. Kothari, S. Meher, and G. Panda, "Improved spatial information based semisupervised classification of remote sensing images," *IEEE J. Sel. Topics Appl. Earth Observ. Remote Sens.*, vol. 13, no. 1, pp. 329–340, Jan. 2020.
[12] J. Long, E. Shelhamer, and T. Darrell, "Fully convolutional networks for semantic segmentation," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 39, no. 4, pp. 640–651, Apr. 2017.
[13] S. Zhang, J. Liew, Y. Wei, S. Wei, and Y. Zhao, "Interactive object segmentation with inside-outside guidance," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2020, pp. 12234–12244.
[14] R. Girshick, J. Donahue, T. Darrell, and J. Mali, "Rich feature hierarchies for accurate object detection and semantic segmentation," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2014, pp. 1–8.
[15] Y. Yuan *et al.*, "Using an attention-based LSTM encoder CDecoder network for near real-time disturbance detection," *IEEE J. Sel. Topics Appl. Earth Observ. Remote Sens.*, vol. 13, no. 4, pp. 1819–1832, Apr. 2020.
[16] D. Bahdanau, "Neural machine translation by jointly learning to align and translate," 2014, *arXiv:1409.0473*.

[17] T. Young, D. Hazarika, S. Poria, and E. Cambria, "Recent trends in deep learning based natural language processing," *IEEE Comput. Intell. Mag.*, vol. 13, no. 3, pp. 55–75, Aug. 2018.

[18] I. Sutskever, O. Vinyals, and Q. V. Le, "Sequence to sequence learning with neural networks," in *Proc. Adv. Neural Inf. Process. Syst.*, 2014, pp. 3104–3112.

[19] N. Srivastava, E. Mansimov, and R. Salakhutdinov, "Unsupervised learning of video representations using LSTMs," in *Proc. Int. Conf. Mach. Learn.*, 2015, pp. 843–852.

[20] S. Hochreiter and J. Schmidhuber, "Long short-term memory," *Neural Comput.*, vol. 9, no. 8, pp. 1735–1780, 1997.

[21] Z. Ling *et al.*, "Self-attention ConvLSTM for spatiotemporal prediction," in *Proc. AAAI Conf. Artif. Intell.*, 2020, pp. 11531–11538.

[22] X. Shi and D. Yeung, "Machine learning for spatiotemporal sequence forecasting: A survey," 2018, *arXiv:1808.06865*.

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

[24] W. C. Woo and W. K. Wong, "Application of optical flow techniques to rainfall nowcasting," in *Proc. 27th Conf. Severe Local Storms*, Nov. 2014, pp. 156–161.

[25] W. Lotter, G. Kreiman, and D. Cox, "Deep predictive coding networks for video prediction and unsupervised learning," in *Proc. Int. Conf. Learn. Representations*, Apr. 2017, pp. 1–18

[26] Y. Wang, M. Long, J. Wang, Z. Gao, and S. Y. Philip, "PredRNN: Recurrent neural networks for predictive learning using spatiotemporal LSTMs," in *Proc. Adv. Neural Inf. Process. Syst.*, 2017, pp. 879–888.

[27] Y. Wang *et al.*, "2019: Memory in memory: A predictive neural network for learning higher-order non-stationarity from spatiotemporal dynamics," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2019, pp. 9154–9162.

[28] X. Shi, Z. Gao, L. Lausen, H. Wang, and D. Yeung, "Deep learning for precipitation nowcasting: A benchmark and a. new model," in *Proc. Adv. Neural Inf. Process. Syst.*, 2017, pp. 5617–5627.

[29] E. Christoforou *et al.*, "Spatio-temporal deep learning for day-ahead wind speed forecasting relying on WRF predictions," *Energy Syst.*, to be published, doi: 10.1007/s12667-021-00480-6.

[30] B. Li, B. Tang, L. Deng, and M. Zhao, "Self-attention ConvLSTM and its application in RUL prediction of rolling bearings," *IEEE Trans. Instrum. Meas.*, vol. 70, no. 6, pp. 351–361, Jun. 2021.

[31] R. Adewoyin *et al.*, "TRU-NET: A deep learning approach to the high-resolution prediction of rainfall," *Mach. Learn.*, vol. 110, pp. 2035–2062, 2021.

[32] B. Liu, Y. Chen, S. Liu, and H. Kim, "Deep learning in latent space for video prediction and compression," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2021, pp. 701–710.

[33] Y. Aoyagi, N. Murata, and H. Sakino, "Spatio-temporal predictive network for videos with physical properties," 2021, pp. 9154–9162.

[34] L. Shi *et al.*, "SA-JSTN: Self-attention joint spatiotemporal network for temperature forecasting," *IEEE J. Sel. Topics Appl. Earth Observ. Remote Sens.*, vol. 14, no. 9, pp. 9475–9485, Sep. 2021.

[35] L. Tao, X. He, J. Li, and D. Yang, "A multiscale extended short-term memory model with an attention mechanism for improving monthly precipitation prediction," *J. Hydrol.*, vol. 602, pp. 126815–126827, 2021.

[36] L. Chen, Y. Cao, L. Ma, and J. Zhang, "A deep learning-based methodology for precipitation nowcasting with radar," *Earth Space Sci.*, vol. 7, Jan. 26, 2020, Art. no. e2019EA000812.

[37] C. Luo *et al.*, "A novel LSTM model with interaction dual attention for radar echo extrapolation," *Remote Sens.*, vol. 13, pp. 164–181, 2021.

[38] G. Melis, T. Kocisky, and P. Blunsom, "Mogrifier LSTM," in *Proc. Int. Conf. Learn. Representations*, 2020, pp. 1–13.

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

[40] L. Han *et al.*, "Convective precipitation nowcasting using U-Net model," *IEEE Trans. Geosci. Remote Sens.*, to be published, doi: 10.1109/TGRS.2021.3100847.

[41] 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.

[42] T. Yu, Q. Kuang, and R. Yang, "ATMConvGRU for weather forecasting," *IEEE Geosci. Remote Sens. Lett.*, to be published, doi: 10.1109/LGRS.2021.3109259.

[43] S. Ravuri *et al.*, "Skillful precipitation nowcasting using deep generative models of radar," *Nature*, vol. 597, pp. 672–677, 2021.

[44] D. Kingma and M. Welling, "Auto-encoding variational Bayes," 2013, *arXiv:1312.6114*.

[45] E. Denon and R. Fergus, "Stochastic video generation with a learned prior," in *Proc. Int. Conf. Mach. Learn.*, 2018, pp. 1174–1183.

[46] K. Gregor, G. Papamakarios, F. Buesing, and T. Weber, "Temporal difference variational auto-encoder," in *Proc. Int. Conf. Learn. Representations*, 2019, pp. 1–16.

[47] J. Franceschi, E. Delasalles, M. Chen, S. Lamprier, and P. Gallinari, "Stochastic latent residual video prediction," in *Proc. Int. Conf. Mach. Learn.*, 2020, pp. 3233–3246.

[48] C. Schuldt, I. Laptev, and B. Caputo, "Recognizing human actions: A local SVM approach," in *Proc. IEEE Conf. Pattern Recognit.*, vol. 3, 2004, pp. 32–36.

[49] 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.

[50] J. Goodfellow *et al.*, "Generative adversarial nets," in *Proc. Adv. Neural Inf. Process. Syst.*, 2014, pp. 2672–2680.

[51] E. Denton *et al.* "Deep generative image models using a Laplacian pyramid of adversarial networks," in *Proc. Adv. Neural Inf. Process. Syst.*, 2015, pp. 1486–1494.

[52] H. Zhang, L. Goodfellow, D. Metaxas, and A. Odena, "Self-attention generative adversarial networks," in *Proc. Int. Conf. Mach. Learn.*, 2019, pp. 7354–7363.

[53] A. Vaswani *et al.*, "Attention is all you need," in *Proc Adv. Neural Inf. Process. Syst.*, 2017, pp. 6000–6010.

[54] Y. Tay, D. Dehghani, D. Bahri, and D. Metzler, "Efficient transformers: A survey," Sep. 2020, *arXiv:2009.06732v2*.

[55] A. Dosovitskiy *et al.*, "An image is worth 16×16 words: Transformers for image recognition at scale," in *Proc. Int. Conf. Learn. Representations*, May. 2021, pp. 1–21.

[56] R. Strudel, R. Garcia, I. Laptev, and C. Schmid, "Segmenter: Transformer for semantic segmentation," in *Proc IEEE Conf. Comput. Vis.*, 2021, pp. 7262–7272.

[57] G. Bertasius, H. Wang, and L. Torresani, "Is space-time attention all you need for video understanding?," 2021, *arXiv: 2102.0509v4*.