




## Article

# Spatiotemporal Feature Fusion Transformer for Precipitation Nowcasting via Feature Crossing

Taisong Xiong<sup>1,2</sup>, Weiping Wang<sup>3,\*</sup>, Jianxin He<sup>1,2</sup>, Rui Su<sup>3</sup>, Hao Wang<sup>1,4,5</sup>  and Jinrong Hu<sup>6</sup>

<sup>1</sup> College of Meteorological Observation, Chengdu University of Information Technology, Chengdu 610225, China; xts@cuit.edu.cn (T.X.)

<sup>2</sup> The Key Laboratory of Atmospheric Sounding, China Meteorological Administration, Chengdu 610225, China

<sup>3</sup> Jiangxi Atmospheric Observation Technology Center, Nanchang 330000, China; suruihao@163.com

<sup>4</sup> China Meteorological Administration Radar Meteorology Key Laboratory, Nanjing 210000, China

<sup>5</sup> Wenjiang National Climatology Observatory, Sichuan Provincial Meteorological Service, Chengdu 611130, China

<sup>6</sup> School of Computer Science, Chengdu University of Information Technology, Chengdu 610225, China

\* Correspondence: ww13707090967@163.com

**Abstract:** Precipitation nowcasting plays an important role in mitigating the damage caused by severe weather. The objective of precipitation nowcasting is to forecast the weather conditions 0–2 h ahead. Traditional models based on numerical weather prediction and radar echo extrapolation obtain relatively better results. In recent years, models based on deep learning have also been applied to precipitation nowcasting and have shown improvement. However, the forecast accuracy is decreased with longer forecast times and higher intensities. To mitigate the shortcomings of existing models for precipitation nowcasting, we propose a novel model that fuses spatiotemporal features for precipitation nowcasting. The proposed model uses an encoder–forecaster framework that is similar to U-Net. First, in the encoder, we propose a spatial and temporal multi-head squared attention module based on MaxPool and AveragePool to capture every independent sequence feature, as well as a global spatial and temporal feedforward network, to learn the global and long-distance relationships between whole spatiotemporal sequences. Second, we propose a cross-feature fusion strategy to enhance the interactions between features. This strategy is applied to the components of the forecaster. Based on the cross-feature fusion strategy, we constructed a novel multi-head squared cross-feature fusion attention module and cross-feature fusion feedforward network in the forecaster. Comprehensive experimental results demonstrated that the proposed model more effectively forecasted high-intensity levels than other models. These results prove the effectiveness of the proposed model in terms of predicting convective weather. This indicates that our proposed model provides a feasible solution for precipitation nowcasting. Extensive experiments also proved the effectiveness of the components of the proposed model.

**Keywords:** precipitation nowcasting; spatiotemporal feature fusion; multi-head squared attention; cross-feature fusion strategy



**Citation:** Xiong, T.; Wang, W.; He, J.; Su, R.; Wang, H.; Hu, J. Spatiotemporal Feature Fusion Transformer for Precipitation Nowcasting via Feature Crossing. *Remote Sens.* **2024**, *16*, 2685. <https://doi.org/10.3390/rs16142685>

Academic Editors: Mohamed Lamine Mekhalfi, Yakoub Bazi, Edoardo Pasolli and Mawloud Guermoui

Received: 5 June 2024

Revised: 14 July 2024

Accepted: 19 July 2024

Published: 22 July 2024



**Copyright:** © 2024 by the authors. Licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license (<https://creativecommons.org/licenses/by/4.0/>).

## 1. Introduction

Severe weather has caused many serious losses, including losses of human lives and property. It results in extreme thunderstorms, tornadoes, or blizzards. Severe storms, which are packed with strong winds, hail, heavy rain, and lightning often wreak havoc in a short time. To effectively reduce the damage caused by severe weather, it is important to forecast its occurrence in advance. Precipitation nowcasting involves forecasting weather conditions 0–2 h ahead and is a very important approach to forecasting severe weather [1,2]. Hence, precipitation nowcasting has received much attention since the 1970s.

The methods for precipitation nowcasting are divided into two categories. One is numerical weather prediction (NWP) [3] systems, and the other is radar echo extrapolation [4]. NWP cannot meet the needs of precipitation nowcasting because it requires many more computational resources to simulate the physical rules of atmospheric circumstances. At the same time, NWP hardly simulates small-scale processes, which are turbulence and convection. Therefore, methods based on radar echo extrapolation have become mainstream. Classical methods, such as tracking radar echoes using correlation [5,6], have been proposed and have obtained better results. However, these methods cannot comprehensively learn the inherent movement of radio echo maps under the conditions of few echo maps. Much more work is needed to improve these precipitation nowcasting methods.

In recent years, deep learning [7] has been successful in many fields, such as image recognition [8,9] and semantic segmentation [10]. Some models based on deep learning have also been proposed for application to precipitation nowcasting. In [11], the authors regarded precipitation nowcasting as a spatiotemporal sequence forecasting problem. Inspired by the temporal features of the recurrent neural network (RNN) and the spatial features of the convolutional neural network (CNN), they proposed convolutional long short-term memory (ConvLSTM) [11] and applied it to precipitation nowcasting. To the best of our knowledge, the model in [11] was regarded as a pioneering work in deep learning in the field of meteorology. ConvLSTM captures spatiotemporal features in sequences of radar echo maps by adding convolution operations into long short-term memory (LSTM). ConvLSTM has demonstrated its superiority over traditional optical flow methods. Compared with conversional physics-based methods, deep-learning methods based on data-driven approaches have demonstrated their potential for meteorological prediction. ConvLSTM only processes locally invariant scenarios; however, the movement of radar echoes varies [12]. To learn the temporal dynamics and spatial correlations of the sequences of radar echo maps, a recurrent architecture model called PredRNN [13] was proposed. Its memory states can interact across various layers. However, these models lack the ability to capture global dependencies because of the limitations of convolution operations. To learn the long dependencies of spatial and temporal domains, self-attention memory [14] was proposed. Similar to the work in [14], a contextual self-attention convolutional LSTM [15] was proposed and applied to precipitation nowcasting. These deep-learning models [16] are all RNN-based approaches. Their network architecture is shown in Figure 1. The accuracies of high intensities of nowcasting results obtained with these models based on RNN were relatively lower. The main cause was the average operations of CNN. Furthermore, the accumulated errors induced lower prediction accuracies with longer forecasting times.

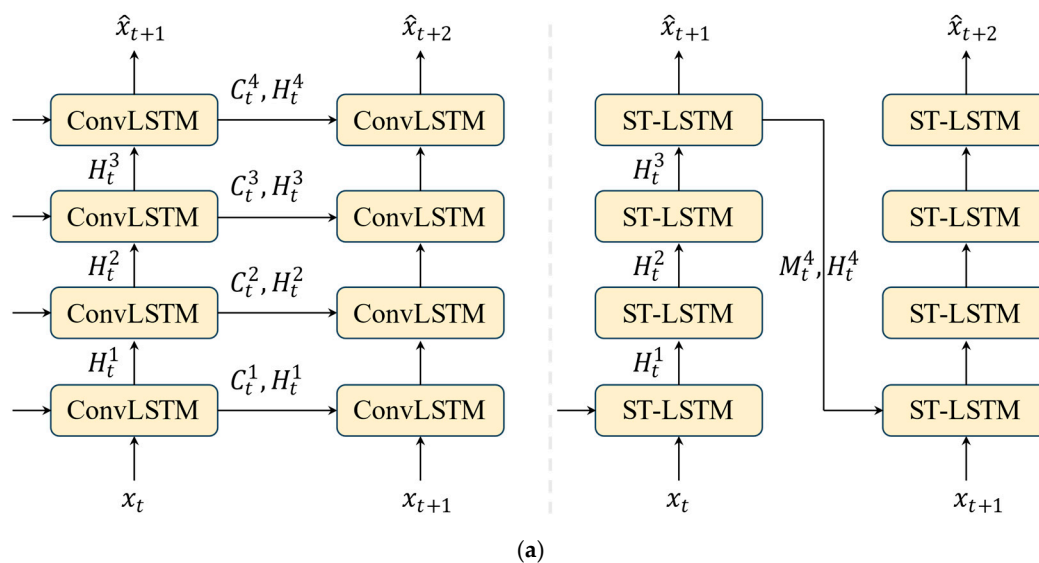
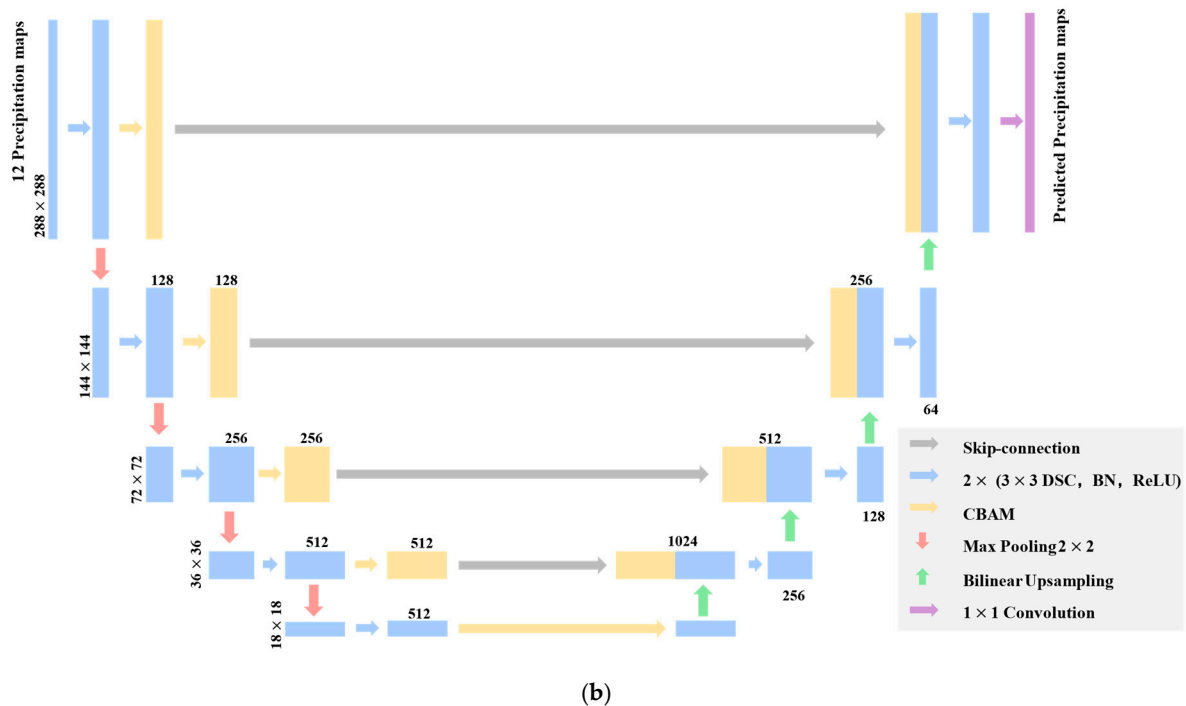


Figure 1. Cont.



**Figure 1.** The network architectures of the model for precipitation nowcasting. (a) Models based on RNN, (b) models based on FCN.

To reduce the accumulated errors and obtain higher forecasting accuracies, another precipitation nowcasting category was proposed based on a full convolutional network (FCN). A representational model was proposed in [17], and this was regarded as an image-to-image translation problem. Unlike models based on LSTM, which explicitly model temporal features, the model in [17] uses a U-Net CNN [18] to capture the changes in radar echo maps. The network framework of the model is shown in Figure 1. Similarly, RainNet [19] combines U-Net and SegNet, and obtains better results than those of a model based on optical flow. To reduce the model's parameter size and obtain comparable performance, a small attention-U-Net model (SmaAt-U-Net) [20], which integrated depth-wise-separable convolutions [21] and an attention mechanism, was proposed. The models based on FCNs are completely data-driven. They do not consider the inherent physical rules and temporal features of precipitation nowcasting.

To learn the physical rules of the movement of radar echo maps and improve precipitation nowcasting, two methods were proposed in [22,23]. PhyDNet [22] uses a semantic latent space to disentangle prior physical knowledge of radar echo sequences. In [24], an L-CNN model was presented based on a CNN and input data were transformed into Lagrangian coordinates. NowcastNet [25] produced precipitation nowcasting 3 h ahead, and combined condition-learning and physical-evolution schemes.

To mitigate the blur when forecasting radar echo maps, generative adversarial networks (GANs) [26] have been adopted. A deep generative model of rainfall [27] was proposed for precipitation nowcasting. It integrated optical flow, and qualitative and quantitative experimental results were obtained. To focus on the local spatial variability in the representation of radar echo maps and alleviate the effect of spatial blurry, an attentional GAN (AGAN) [28] was proposed.

The aforementioned models obtained results superior to those of traditional methods. However, these models paid little attention to high intensity, which is the main cause of severe weather. Simultaneously, the phenomenon of gradient explosion or gradient disappearance [29] may occur in these models based on RNNs during their training processes. Furthermore, the error increases as the forecasting time increases.

Recently, applications of the transformer [30,31] have extended from NLP to computer vision. The strength of the transformer is that it can effectively model global features and long-range dependencies. The vision transformer has gradually become the backbone network for vision applications. Variants [32,33] of the vision transformer have been proposed to improve computational efficiency. Some models based on the transformer have been proposed and applied to spatiotemporal prediction learning. They have obtained better results than those of state-of-the-art models. A Swin spatiotemporal fusion model (SwinSTFM) [34] was proposed for forecasting remote-sensing images. Hu et al. proposed a Swin transformer-based variational RNN (SwinVRNN) [35] to deterministically forecast future states. Furthermore, they integrated a perturbation module to generate inference in the stochastic latent variable of meteorological data. SwinRDM [36], which improved SwinVRNN using a diffusion model, was proposed to forecast the temperature at a geopotential of 500 hPa and resolution of 2 m 5 days ahead. A three-dimensional (3D) Earth-specific transformer [37] was proposed for medium-range global weather forecasting. FuXi [38] proposed a model based on the Swin transformer V2, which offers 15-day global forecasts at a temporal resolution of 6 h. Rainformer [29] mainly consists of two units that capture the local and global features. Simultaneously, a gate fusion unit in Rainformer is used to ingeniously fuse the local and global features. Another space–time transformer called Earthformer [39] was proposed for Earth system forecasting. Earthformer includes a cuboid attention mechanism that decomposes the input tensor into non-overlapping cuboids and performs a self-attention operation in these cuboids in parallel. However, the forecasting time of the two models is very short: one is only 9 frames (45 min) and the other is 12 frames (60 min). A temporal–spatial parallel transformer, which uses the past 20 frames to predict the future 20 frames was proposed in [40]. To reduce the computational complexity of the attention mechanism, lightweight attention [41] was proposed and integrated into a hierarchical transformer for precipitation nowcasting. Using multiple meteorological elements, the preformer [42] captures global spatiotemporal dependencies to forecast future precipitation.

Although progress has been made to mitigate the error of the higher intensity of echo maps and decrease the error rate with a longer forecasting time, improvement is needed for precipitation nowcasting. To overcome the two shortcomings of precipitation nowcasting, we propose a spatiotemporal feature fusion transformer (STFFT) for precipitation nowcasting via feature crossing. STFFT uses an encoder–forecaster framework that is similar to the U-Net architecture. The encoder and forecaster of the proposed model both consist of several transformers. The transformer in the encoder network is constructed by a unique independent temporal attention module and a global spatial and temporal feedforward network (GSTFFN) to explicitly capture the spatiotemporal features of radar echo sequences. In the decoder architecture, we propose a new attention module and feedforward network (FFN), both of which we construct by fusing crossing features. The mechanism that fuses cross-channel attention can effectively capture the correlations between these radar echo sequences. The proposed encoder–forecaster architecture is a new network architecture that consists of both an RNN module and an FCN module. The encoder architecture can effectively learn local and global temporal features, as well as their interaction information in various timespans. Furthermore, two novel attention modules and two FFNs that are integrated into the encoder–forecaster architecture extract the global and long-distance features of the radar echo sequences.

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

- (1) We propose an encoder–forecaster framework for precipitation nowcasting; the encoder explicitly processes the temporal sequence data and the forecaster processes the sequence data in total. The framework efficiently integrates the merits of the models based RNN and FCN.
- (2) A component with MaxPool and AvgPool operations [43] is integrated with the attention model which can effectively capture the features of high intensities. At the same time, the GSTFFN strengthens the spatial–temporal features. These operations

will effectively mitigate the error rates of forecasting higher intensities and longer forecasting times for the proposed model.

- (3) Based on the strategy of feature crossing, a cross-channel attention is proposed in the forecaster to effectively simulate the movements of these radar echo sequences.
- (4) The forecaster, which is similar to that of the models based on FCN, effectively reduces accumulated errors and improves the forecasting accuracies of longer nowcasting times.

The rest of this paper is organized as follows: The data used in this study are introduced in Section 2. The proposed model is described in detail in Section 3. The comprehensive experimental results and analyses are given in Section 4. The summary and conclusions are given in Section 5.

## 2. Dataset

We trained and tested the proposed model and state-of-the-art models on the public precipitation nowcasting dataset proposed in [44], which is called SEVIR, and consists of more than 10,000 weather events whose time span and time steps are 4 h and 5 min, respectively. SEVIR contains a series of spatially and temporally aligned image sequences. These sequences belong to five categories which are NEXRAD vertically integrated liquid (VIL) mosaics, three channels (C02, C09, and C13) from the GOES-16 advanced baseline imager, and GOES-16 Geostationary Lightning Mapper (GLM) flashes. The SEVIR dataset can be applied for front detection, synthetic radar generation, precipitation nowcasting, etc. For precipitation nowcasting, we chose the weather events captured over a WSR-88D (NEXRAD) radar mosaic of VIL and its event count was 20,393. The storage format of SEVIR is HDF5 file. The VIL images in SEVIR are grayscale and their intensity values are in the range 0–255. The conversion of the pixels values to the true units of vertically integrated liquid ( $\text{kg}/\text{m}^2$ ) is given in Equation (1). The non-linear scaling rule is more convenient for storing VIL images. Examples of the SEVIR VIL images are shown in Figure 2. This example shows the information for 5, 20, 30, 40, 50, and 60 min. The original sequence length was 49 frames, and the total number of frames in the input and output was 24. We sampled each frame in sequence, and the stride was 12. The configuration of the training, validation, and test datasets is provided in Table 1.

$$f(x) = \begin{cases} 0 & \text{if } X \leq 5 \\ \frac{(X-2)}{90.66} & \text{if } 5 < X \leq 18 \\ \frac{\exp(X-83.9)}{38.9} & \text{if } x > 18 \end{cases} \quad (1)$$

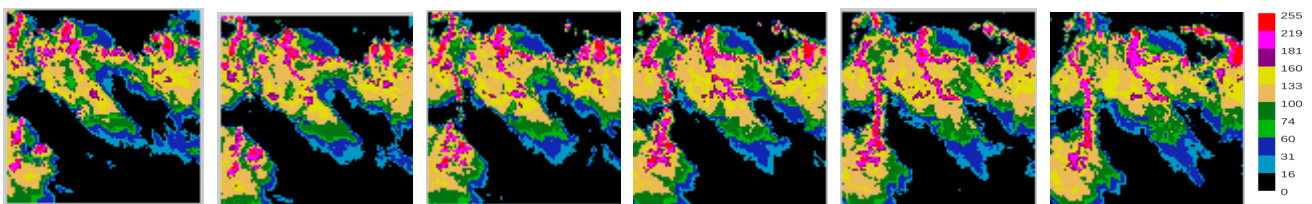


Figure 2. Sample VIL images in SEVIR.

Table 1. Configuration of the datasets.

Items	Training	Validation	Test
Sequences	35,718	9060	12,159

## 3. Methods

### 3.1. Problem Statement

As described in [11], precipitation nowcasting can be regarded as a spatiotemporal sequence forecasting problem. We suppose that we have known the radar echo maps for the past half hour and can predict information approximately 90 min ahead; that is, the

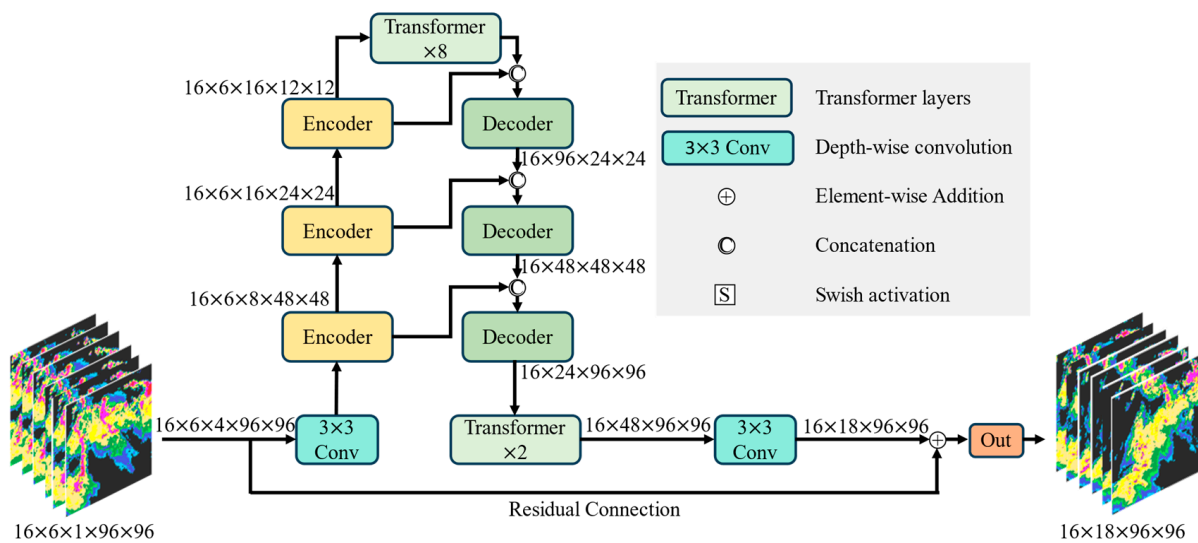
objective of our model is to predict the next 18 frames (90 min) using our knowledge of the past 6 frames (30 min). A radar echo sequence that consists of  $N$  frames is denoted by  $Y_1, Y_2, \dots, Y_N$ . The precipitation nowcasting process of the proposed model can be defined as in Equation (2):

$$\hat{Y}_{t:t+17} = \underset{Y_{t:t+17}}{\operatorname{argmax}} F\left(|Y_t, Y_{t+1}, \dots, Y_{t+17} | Y_{-6}, Y_{-5}, \dots, Y_{-1}\right) \quad (2)$$

where the subscript of  $Y$  denotes the timespan which is 5 min. The objective of our model is to simulate the function  $F$  to obtain  $\hat{Y}_{t:t+17}$ , which closely approximates the ground truth  $Y_{t:t+17}$ . For precipitation nowcasting, the objective of  $Y_n$  is two-dimensional radar echo maps. We provide detailed information about the proposed model in the following section.

### 3.2. Network Architecture

To effectively simulate the function  $F$  in Equation (2), we propose a new encoder–forecaster model based on a transformer. The network architecture of the proposed model, which is similar to that of the network in [23], is shown in Figure 3. Our network consists of three encoders and forecasters, which are the same as those in [23]. However, our processes are different from those in [23]. The encoder consists of  $N$ -stacked transformer layers and a downsampling layer follows every layer except the last one. The forecaster consists of  $N$ -stacked transformer layers and an upsampling layer is inserted in front of transformer. We provide detailed descriptions of the model in the following section.



**Figure 3.** Network architecture of the proposed model.

### 3.3. Spatiotemporal Encoder

For the encoder, the transformer consists of an attention module and a feedforward network (FFN). The framework of the encoder is shown in Figure 4. The attention module is a spatiotemporal multi-head squared attention (STMSA) based on MaxPool and Average-Pool. The FFN is a GSTFFN. Our proposed STMSA was inspired by [40,41]. The STMSA processes independent temporal feature maps. There are six independent temporal units in the proposed model, which capture the spatiotemporal features of an individual radar echo. The processed feature maps of the six frames are then concatenated and fed into the GSTFFN. The function of the GSTFFN is to learn the global spatial and temporal features using the Conv3D operation.

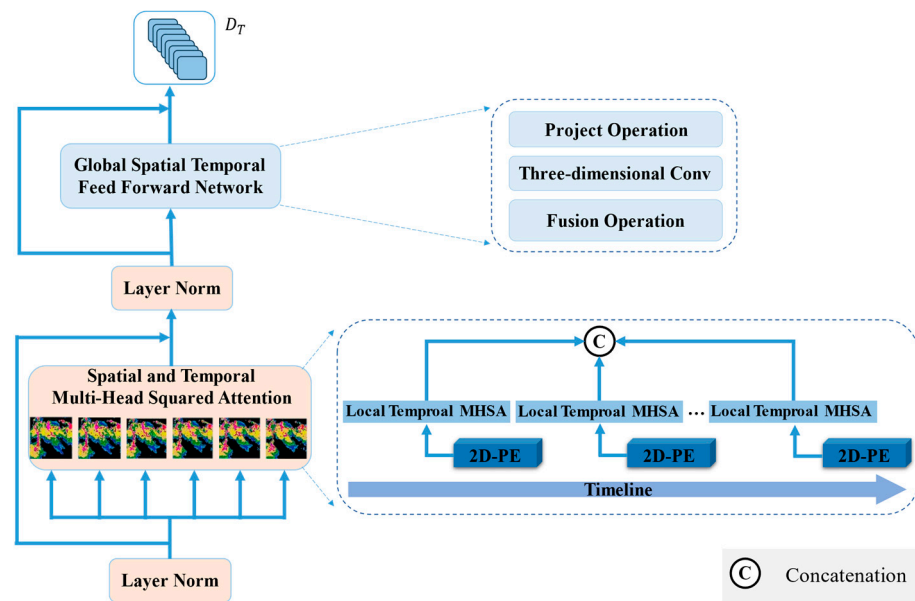


Figure 4. Network architecture of the encoder.

In the STMSA, the input data are first projected into three subplaces using Conv2D, whose kernel size is 3, and the dimension of its output is the same as that of the input. To effectively capture the high intensity of the meteorological radar echoes, we give a unit of high-intensity feature representations (UHIFR). In UHIFR, the operations of MaxPool and AveragePool are conducted and applied to the query and key variables. The obtained values are multiplied according to their feature maps. The results are then concatenated and projected into one subspace whose dimension is equal to that of the original subspace. The STMSA mainly captures the spatiotemporal features of the individual timespan and cannot learn the global features of the whole sequence. The network architecture of the three components in the encoder are shown in Figure 5. All operations of the STMSA are defined in Equations (3)–(9):

$$\hat{Y} = W_1 \text{reshape}(\text{Attention}(\hat{Q}, \hat{K}, \hat{V})) + Y \tag{3}$$

$$\text{Attention}(\hat{Q}, \hat{K}, \hat{V}) = \text{Sigmoid}\left(\hat{Q} \cdot \frac{\hat{K}^T}{\beta}\right) \cdot \hat{V} \tag{4}$$

$$\ddot{Q} = W_3^Q Y_{LN} \tag{5}$$

$$\hat{Q} = \text{reshape}(W_1^Q(\text{concat}(\text{MaxPool}(\ddot{Q}) \cdot \ddot{Q}, \text{AvgPool}(\ddot{Q}) \cdot \ddot{Q}))) \tag{6}$$

$$\ddot{K} = W_3^K Y_{LN} \tag{7}$$

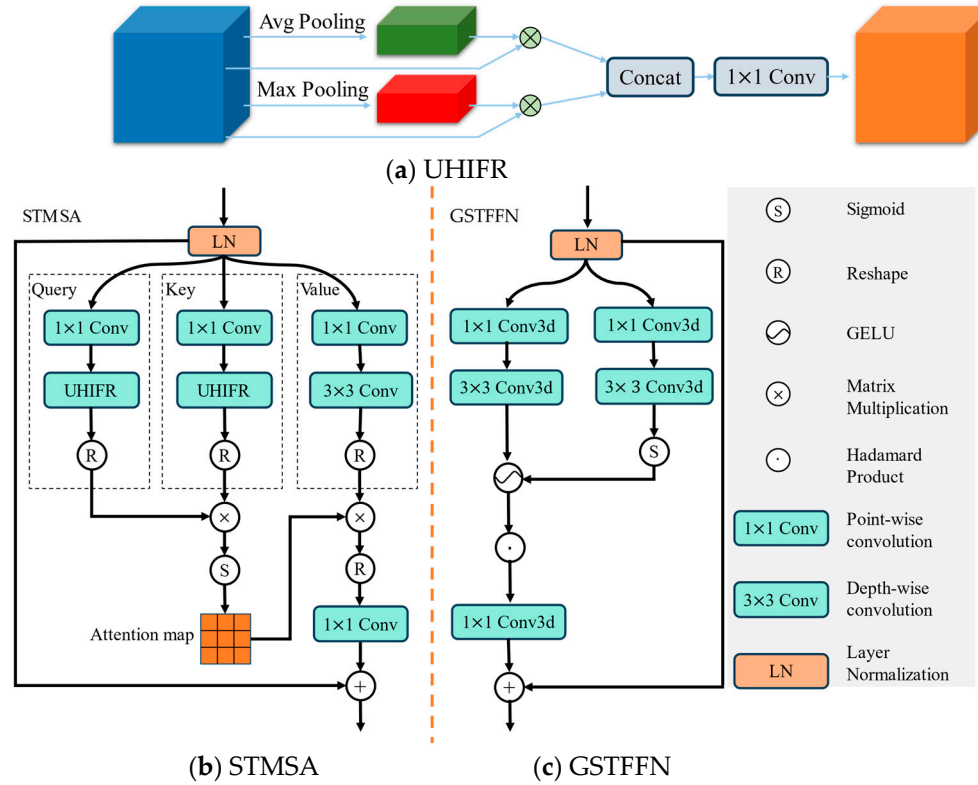
$$\hat{K} = \text{reshape}(W_1^K(\text{concat}(\text{MaxPool}(\ddot{K}) \cdot \ddot{K}, \text{AvgPool}(\ddot{K}) \cdot \ddot{K}))) \tag{8}$$

$$\hat{V} = \text{reshape}(W_3^V Y_{LN}) \tag{9}$$

where  $\hat{Q}, \hat{K}$ , and  $\hat{V} \in R^{C \times H \times W}$ ;  $Y \in R^{C \times H \times W}$ ; and C, H, and W denote the channel, height, and width of the feature maps, respectively. The subscripts of the weight matrix W, 3 and 1, represent  $3 \times 3$  and  $1 \times 1$  convolutional operations, respectively. The input of STMSA is first subjected to layer normalization (LN). LN is a primary component of the transformer. Its definition is given in Equation (10):

$$Y = \frac{X - \mu(X)}{\sqrt{\text{conv}(X) + \epsilon}} * \beta \tag{10}$$

where  $X$  and  $Y$  represent the input and output of the LN layer, respectively.  $\mu$  and  $conv$  denote the mean and variance of  $X$ , respectively. The constant  $\epsilon$  is a very small value and its function is to prevent the denominator from zero.  $\beta$  is used to improve the expression of the models.



**Figure 5.** Component network architecture of the encoder. (a) UHIFR, (b) STMSA, (c) GSTFFN.

Following the STMSA, the GSTFFN is a convolutional network architecture that is also a component of the conventional transformer. The GSTFFN is designed to capture the global and long-range spatiotemporal features of radar echo sequences. The GSTFFN unit has five operational steps. Their definitions are given in Equations (11)–(14):

$$\hat{Y} = \text{TFFN}(Y_1, Y_2) + Y \quad (11)$$

$$Y_1 = W_1^{3D1} W_3^{3D1} (Y_{LN}) \quad (12)$$

$$Y_2 = W_1^{3D2} W_3^{3D2} (Y_{LN}) \quad (13)$$

$$\text{GSTFFN}(Y_1, Y_2) = W_1^{3D} (\text{Gelu}(Y_1) \cdot \text{Sigmoid}(Y_2)) \quad (14)$$

where  $W_1^{3D1}$ ,  $W_3^{3D1}$ ,  $W_1^{3D2}$ ,  $W_3^{3D2}$ , and  $W_1^{3D}$  are 3D convolutions, the subscript represents the convolutional kernel, and the number following 3D in the superscript represents the number of objects. The 3D convolution can effectively incorporate the global spatiotemporal features of the input vertically integrated liquid (VIL) sequences. The input of the GSTFFN is projected into two subspaces after the LN operation. The first subspace is activated by the GELU function, which exhibits strong nonlinear responses. The other subspace is activated by the sigmoid function. The two subspaces are combined into one subspace using matrix multiplication and then conducted using 3D convolution with a kernel size of 1. The 3D convolution can effectively capture the local and integrated spatiotemporal features.

### 3.4. Forecaster

The forecaster consists of three decoder layers. The decoder consists of N-stacked transformer layers and one upsampling layer. In the transformer, we propose a cross-feature



fusion strategy to construct a novel multi-head squared cross-feature fusion attention (MHSFFA) module and cross-feature fusion FFN (CAFFFN) to apply for precipitation nowcasting. The network architectures of the MHSFFA and the CAFFFN are shown in Figure 6. The operations of MHSFFA are given in Equations (15)–(19):

$$\hat{Y} = W_1 \text{reshape}(\text{Attention}(\hat{Q}, \hat{K}, \hat{V})) + Y \tag{15}$$

$$\text{Attention}(\hat{Q}, \hat{K}, \hat{V}) = \text{Sigmoid}\left(\hat{Q} \cdot \frac{\hat{K}^T}{\beta}\right) \cdot \hat{V} \tag{16}$$

$$\bar{Q} = W_3^Q W_1^Q(Y_{LN}) \quad \bar{K} = W_3^K W_1^K(Y_{LN}) \tag{17}$$

$$\bar{\bar{Q}} = \bar{Q} \text{ogelu}(\bar{K}) \quad \bar{\bar{K}} = \bar{K} \text{osigmoid}(\bar{\bar{Q}}) \tag{18}$$

$$\text{Attention}(\hat{Q}, \hat{K}, \hat{V}) = \text{sigmoid}\left(\hat{Q} \cdot \frac{\hat{K}^T}{\beta}\right) \cdot \hat{V} \tag{19}$$

where o stands for the Hadamard product. The MHSFFA is applied to whole datasets, whereas the STMSA of the encoder is applied to an individual radar echo feature and the results are then concatenated for the whole dataset. Therefore, MHSFFA can effectively capture the global and long-distance spatial-temporal features. Simultaneously, the fusion of the query and key in the MHSFFA can help to model the interactions between spatiotemporal sequences.

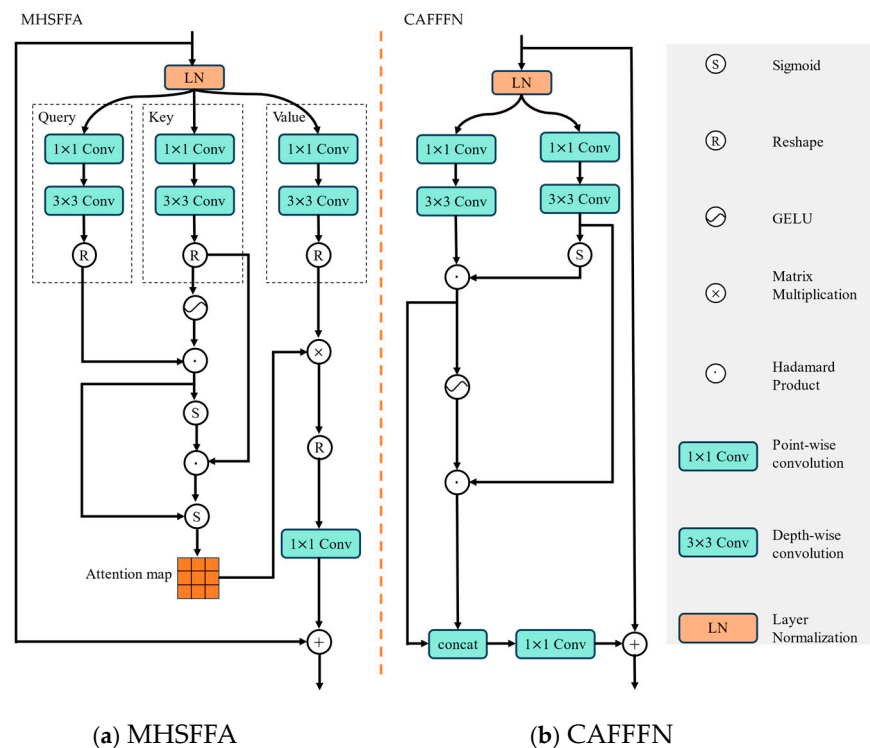


Figure 6. Network architecture of MHSFFA and CAFFFN in decoder.

To effectively capture the spatiotemporal features of VIL images, the input of the CAFFFN is projected into two subspaces. The weights of the two subspaces are obtained using the GELU and sigmoid activation functions. The two subspaces interact as a result of being multiplied by the other’s weight. Then, the two subspaces are concatenated so that

they can be projected into one subspace whose dimension is the same as the input's. All operations in the CAFFFN are defined in Equations (20)–(23):

$$\hat{Y} = \text{CAFFFN}(Y_1, Y_2) + Y \quad (20)$$

$$\bar{Y}_1 = W_3^1 W_1^1(Y_{LN}) \quad \bar{Y}_2 = W_3^2 W_1^2(Y_{LN}) \quad (21)$$

$$Y_1 = \bar{Y}_1 * \text{sigmoid}(\bar{Y}_2) \quad Y_2 = \bar{Y}_2 * \text{gelu}(Y_1) \quad (22)$$

$$\text{CAFFFN}(Y_1, Y_2) = W_1(\text{concat}(Y_1, Y_2)) \quad (23)$$

where  $Y$  and  $\hat{Y}$  denote the input and output of the CAFFFN module, respectively.  $Y_{LN}$  is the result of the LN of  $Y$ . Their shapes are  $R^{C \times H \times W}$ , where the superscripts denote the channel, height, and width of the feature maps. The subscript of  $W$  represents the size of the convolutional kernel. We compute the attention maps from both the spatial and temporal perspectives. The convolution operations first project the input into two latent spaces. Then, the interaction relationships for these two subspaces are created through cross-feature fusion, which is defined in Equations (20)–(23). Finally, the two subspaces are concatenated to be projected into one subspace.

## 4. Experiment

### 4.1. Implementation Details

We implemented all models using PyTorch and conducted the experiments on a machine with an NVIDIA A40 graphics processing unit (GPU). We chose the AdamW [45] optimizer to train all models. We set the learning rate to 0.001 and chose cosine annealing as the learning rate scheduler. Simultaneously, the max training epoch was equal to 100. We set the early stopping rule with patience to 20 and monitored the critical success index (CSI) value of the validation set in the training process to prevent the overfitting of the model. For a fair comparison, we applied the mean square error (MSE) loss to all models.

The time span of the inputs was 30 min, which is six images, and the time span of the outputs was an hour and a half, which indicated that the output was 18 frames. Some configurations of the compared models were different from those in the original papers because of the different experimental contexts. The configuration details of all models are provided in Table 2. We chose the encoder–forecaster framework of ConvLSTM in [12], whose downsampling sizes were 5, 3, and 2, and the configuration in our experiments was 2, 2, and 2. Following the training of the models, we chose each model that obtained the highest CSI value from the validation set. Then, we applied the model with the best CSI to the test dataset. To comprehensively verify the correctness and effectiveness of the proposed model, we compared it with the five models: RainNet [19], ConvLSTM [12], SmaAt-UNet [20], SimVP [46], and LPT-QPN [23].

**Table 2.** Implementation details for the baseline models.

Model	Modified Details	Official Configuration	Our Adaptations
SmaAt-UNet	Input length	12	6
	Output length	1	18
ConvLSTM	Loss function	Balanced MSE	MSE
	Input length	5	6
	Output length	20	18
SimVP	Input length	10	6
	Output length	10	18
LPT-QPN	Input length	5	6
	Output length	20	18

#### 4.2. Evaluation Metrics

To comprehensively verify the superiority and effectiveness of the proposed model, we adopted two types of criteria. The first type was based on the image quantity: the MSE and mean absolute error (MAE). The equations for the MSE and MAE are defined in Equations (24) and (25):

$$MSE = \frac{1}{H * W} \sum_{h=1}^H \sum_{w=1}^W (\hat{Y}_{h,w} - y_{h,w})^2 \quad (24)$$

$$MAE = \frac{1}{H * W} \sum_{h=1}^H \sum_{w=1}^W |\hat{Y}_{h,w} - y_{h,w}| \quad (25)$$

where  $H$  and  $W$  represent the height and width of the image, respectively.  $y$  and  $\hat{y}$  represent the ground truth and predicted results, respectively. The MAE stands for the average differences between the prediction results and the ground truth. The MSE represents the mean squared error between the prediction results and the ground truth. Lower values of MSE and MAE represent better results.

The other type of criteria consisted of meteorological evaluation metrics that we used to evaluate the performance of precipitation nowcasting: the false alarm ratio (FAR) [47], probability of detection (POD), bias score (BIAS) [48], and CSI [49]. True and pred represent the ground truth and predicted results, respectively. HITS = (true = 1, pred = 1), FALSES = (true = 0, pred = 1), and MISSES = (truth = 1, pred = 0). The four criteria are defined in Equations (26)–(29):

$$FAR = \frac{FALSE}{HITS + FALSES} \quad (26)$$

$$POD = \frac{HITS}{HITS + FALSES} \quad (27)$$

$$CSI = \frac{TP}{TP + FALSES + MISSES} \quad (28)$$

$$BIAS = \frac{HITS + FALSE}{HITS + MISSES} \quad (29)$$

where the FAR represents the proportion of predicted pixels that did not occur. As the opposite of FAR, the POD computes the fraction of observed pixels that were correctly predicted. The CSI measures the consistency between the predicted and real observed outcomes, considering both the accuracy of the predictions and the rate of occurrence rate of events. The values of the FAR, POD, and CSI are between 0 and 1. The higher the values of the POD and CSI, the better the performance, and the lower the value of the FAR, the worse the performance. BIAS measures the deviation of predictions. A value of BIAS that is larger than 1 indicates that the forecast result is stronger than the observation. The prediction result is the best if the value of BIAS is equal to 1. The prediction result is weaker when the values of BIAS are lower than 1.

#### 4.3. Quantitative Performance

The quantitative results for the POD, CSI, BIAS, FAR, MSE, and MAE of all models are provided in Table 3. The best and the suboptimal performances for every criterion is represented with bold and underlined text, respectively. To verify the proposed model, we only adopted the MaxPool and AveragePool operation in the encoder, which are denoted by WM (with MaxPool) and WA (with AveragePool). The quantitative results are provided in Table 3, which shows that our proposed model obtained the best results for most criteria compared to the other models. Furthermore, the proposed model obtained relatively stable results compared with the other models for these criteria. The network architecture of the proposed model obtained the best results compared with the other two models for all

criteria except the MSE, which was larger for the MaxPool model by only 0.004. It shows that the concatenation of the MaxPool and AveragePool effectively improved the superiority of STFFT. The experimental results demonstrate the effectiveness of the proposed model. They also demonstrate that the correctness of the concatenation of MaxPool and AveragePool.

**Table 3.** Quantitative mean values for all criteria.

Model	POD	CSI	BIAS	FAR	MSE ( $10^{-3}$ )	MAE ( $10^{-3}$ )
RainNet	0.3768	0.3111	0.9050	0.6232	5.183	34.06
SmaAt-UNet	<u>0.4258</u>	0.3432	<b>1.1437</b>	0.5742	5.288	34.977
ConvLSTM	0.3970	0.3332	0.9357	0.6030	4.943	<u>31.525</u>
LPT-QPN	0.4257	0.3444	<u>1.0853</u>	0.5743	5.111	33.56
SimVP	0.4082	0.3452	0.9481	0.5918	<u>4.648</u>	<b>30.319</b>
STFFT-WA	0.4233	<u>0.3464</u>	1.0378	0.5767	4.908	33.796
STFFT-WM	0.4181	0.3417	1.0366	0.5819	<b>4.889</b>	33.796
STFFT	<b>0.4269</b>	<b>0.3522</b>	1.0162	<b>0.5731</b>	4.893	32.162

A quantitative comparison of the experimental results of various precipitation thresholds defined in [39] for the POD and CSI is provided in Tables 4 and 5, respectively. The best value is represented in bold for every level. From the two tables, we can see that the values of the POD and CSI obtained with all models decreased with an increase in the intensity levels. This shows that it is a difficult task to correctly forecast high-intensity echoes. Our models obtained the best results in most levels. There was a small difference between the proposed model and the other models in worse situations. In particular, the proposed model obtained better results for higher-level intensities. For the highest level, the difference between our models and the suboptimal model was near one percent. This shows that our proposed model can more effectively forecast the strong convective weather than the other models. These tables show that our proposed model obtained better experimental results when compared to the other models. These results demonstrate that the proposed model effectively forecasts convective weather.

**Table 4.** Quantitative comparison of the POD for various levels.

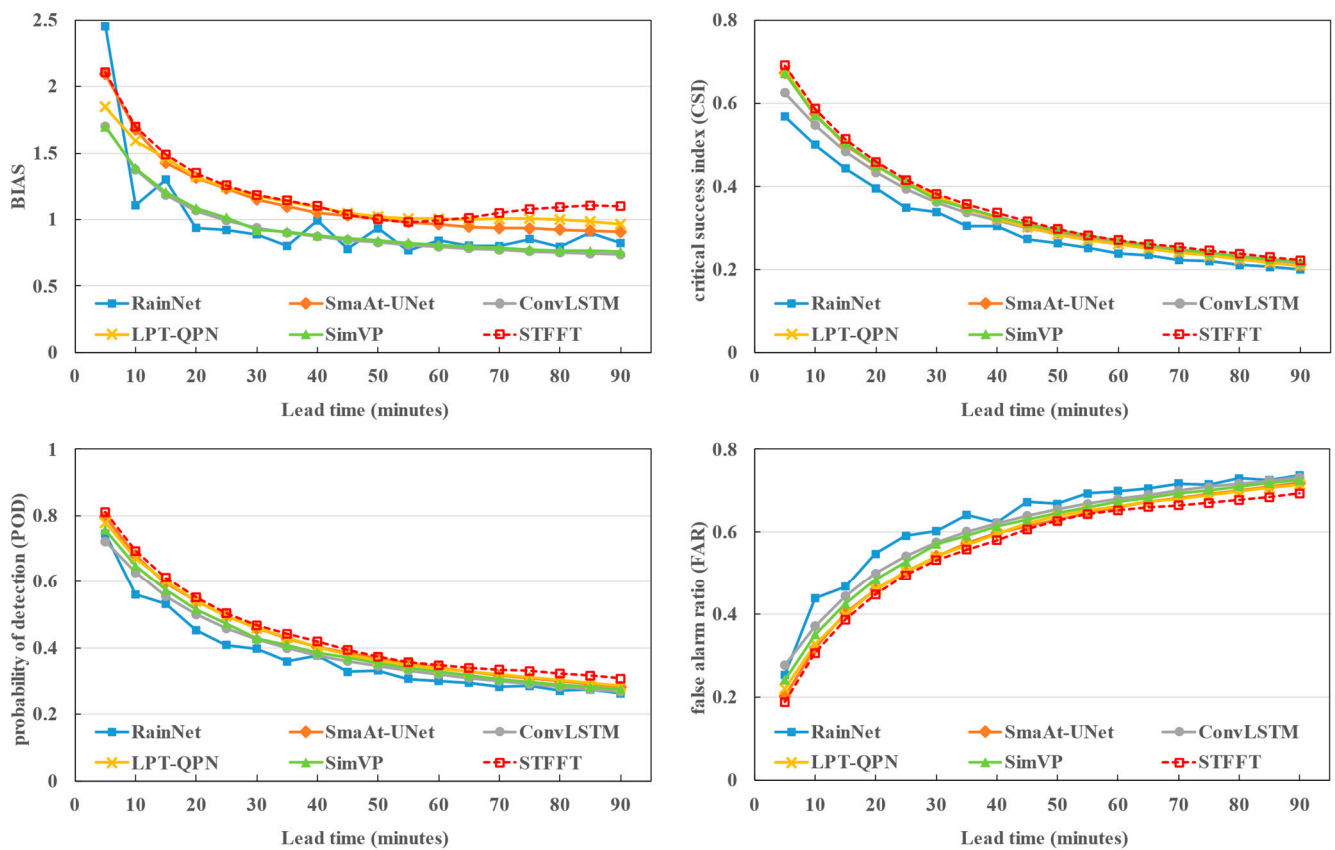
Model	POD-M	POD-16	POD-74	POD-133	POD-160	POD-181	POD-219
RainNet	0.3768	0.8953	0.7044	0.3150	0.1731	0.1212	0.0515
SmaAt-UNet	0.4258	0.8999	0.7351	0.3881	<b>0.2438</b>	<b>0.1917</b>	0.0964
ConvLSTM	0.3970	0.8800	0.7108	0.3597	0.2128	0.1560	0.0629
LPT-QPN	0.4257	0.9049	<b>0.7489</b>	<b>0.4123</b>	0.2290	0.1684	0.0909
SimVP	0.4082	0.8891	0.7247	0.3705	0.2209	0.1670	0.0771
STFFT-WA	0.4233	0.9087	0.7311	0.3856	0.2257	0.1766	<b>0.1120</b>
STFFT-WM	0.4181	<b>0.9129</b>	0.7404	0.3803	0.2129	0.1638	0.0986
STFFT	<b>0.4269</b>	0.9037	0.7408	0.3817	0.2376	0.1891	0.1085

To further verify the effectiveness of the proposed model, we provide the values of the BIAS, CSI, POD, and FAR obtained with RainNet, SmaAt-UNet, ConvLSTM, LPT-QPN, SimVP, and STFFT over an hour and a half, with a time span of 5 min. The values of the four criteria are shown in Figure 7. We can see that the performance of all models decreased with longer forecast times. This indicates that it is very difficult to forecast longer times information. However, our proposed model, STFFT, obtained the best results at each stage for the four criteria. This shows that our results were relatively stable and that our model

provided a better forecast than those of the other models. The results shown in Figure 7 further prove the effectiveness and stability of STFFT.

**Table 5.** Quantitative comparison of the CSI for various levels.

Model	CSI-M	CSI-16	CSI-74	CSI-133	CSI-160	CSI-181	CSI-219
RainNet	0.3111	0.6777	0.5984	0.2673	0.1595	0.1145	0.0492
SmaAt-UNet	0.3432	0.6891	0.6114	0.3106	0.2068	0.1632	0.0779
ConvLSTM	0.3332	<b>0.7104</b>	0.6131	0.2994	0.1867	0.1371	0.0526
LPT-QPN	0.3444	0.6848	0.6157	<b>0.3227</b>	0.2035	0.1555	0.0844
SimVP	0.3478	0.7109	<b>0.6236</b>	0.3142	0.2038	0.1583	0.0758
STFFT-WA	0.3464	0.6865	0.6173	0.3129	0.2009	0.1611	<b>0.0995</b>
STFFT-WM	0.3417	0.6840	0.6208	0.3108	0.1928	0.1517	0.0904
STFFT	<b>0.3522</b>	0.6957	0.6196	0.3147	<b>0.2121</b>	<b>0.1720</b>	0.0990

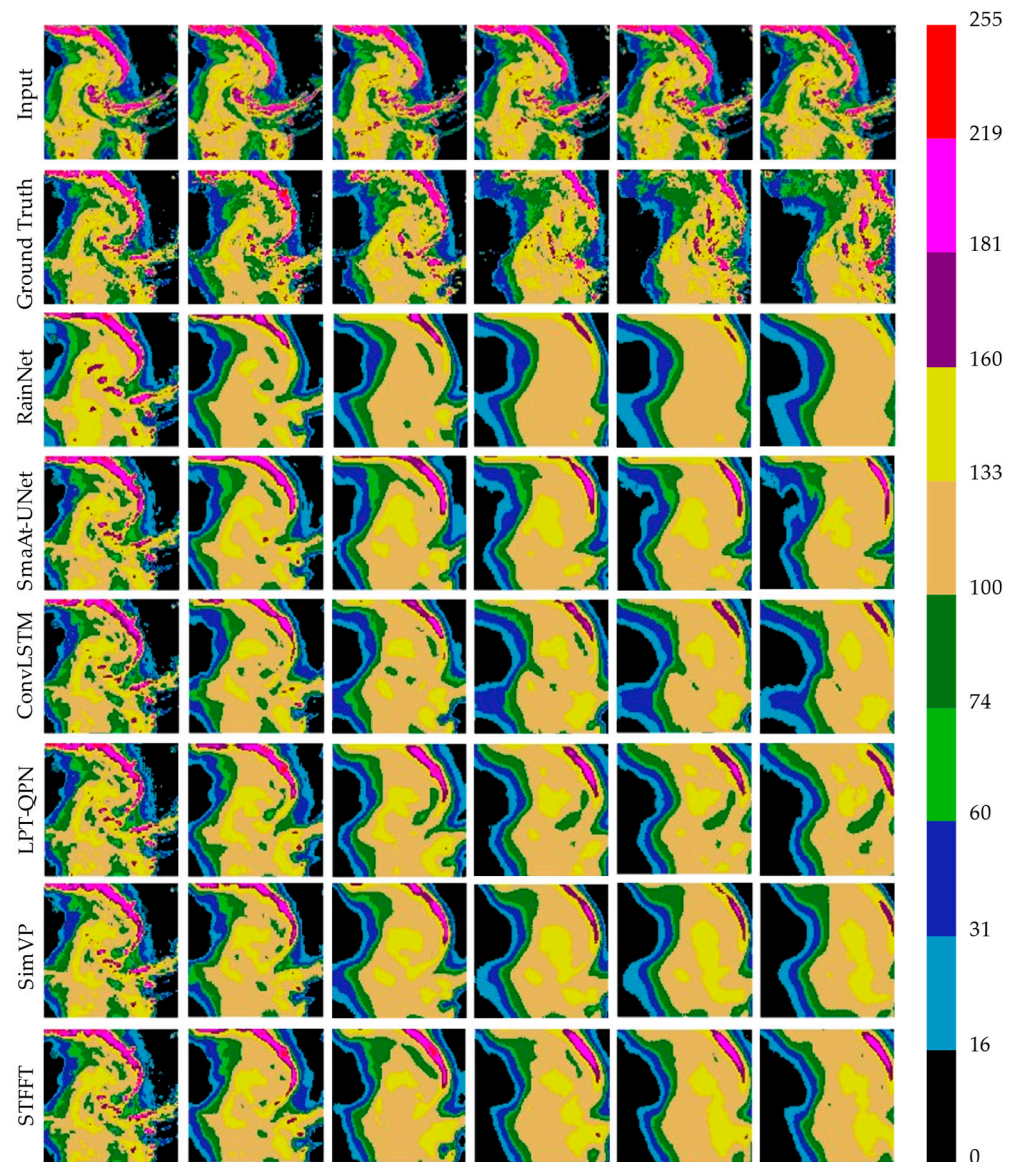


**Figure 7.** The values of the BIAS, CSI, POD, and FAR at each period for all models.

#### 4.4. Visual Performance

To verify the visual effectiveness of the nowcasting results from all models, we selected two representative cases from the test set to provide the visual and quantitative results. The first nowcasting results are shown in Figure 8, and their corresponding CSI values are given in Table 6. The first row shows the six successive input VIL frames. The second row shows the ground-truth VIL frames, which were generated at 5, 25, 45, 65, 75, and 90 min. The nowcasting results from RainNet, SmaAt-UNet, ConvLSTM, LPT-QPN, SimVP, and the proposed model are shown from the third line to the eighth line. In Figure 8, we can see that the nowcasting results were closer to the ground truths with a shorter forecasting time. With a longer forecasting time, the shape and contour of the nowcasting

results become blurred. More high-intensity pixels were obtained with the proposed model than with the other models. At the same time, the details of forecasting with our model were maintained better than those of other models. This shows that our proposed model obtained better nowcasting results than other models. The quantitative results given in Table 6 also demonstrate the effectiveness of the proposed model. The proposed model obtained the best results at all time levels except the first forecasting time. The difference between the proposed model and the best model in the first 5 min was only 0.0012%. The experimental results show that the proposed model obtained better nowcasting results with the longer forecasting times and high intensities when compared with the other models.



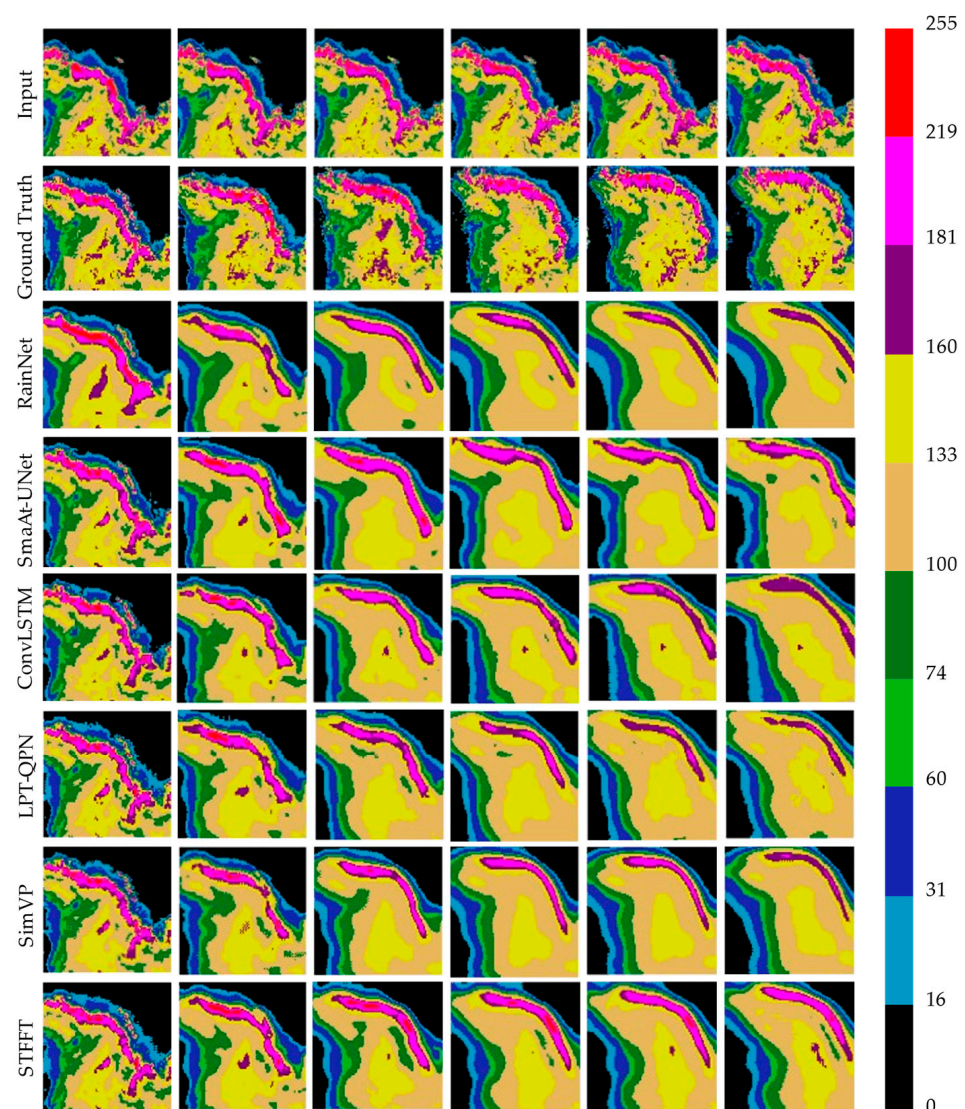
**Figure 8.** A visual example of the nowcasting results from all of the models.

A second prediction example is shown in Figure 9. The organization of Figure 9 is the same as that in Figure 8. The proportion of high intensities is very large in the sequence shown in Figure 9. The evolution process shows that the high-intensity echoes gradually decreased. In Figure 9, we can see that the nowcasting results from SmaAt-UNet, ConvLSTM, and our proposed model are closest to the ground truth. However, the shape, contour, and intensities of the prediction results obtained with our model are closest to the ground truth. This shows that the proposed model obtains the best nowcasting results when compared to the other models. The corresponding quantitative values given in Table 7

also prove this. The nowcasting results from all models demonstrate that our proposed model can capture not only the global spatial features, but also the longer temporal features of echo sequences.

**Table 6.** Quantitative comparison of the CSI values of the visual example.

Model	Mean	5 min	25 min	45 min	65 min	75 min	90 min
RainNet	0.4217	0.5832	0.4621	0.4094	0.3632	0.3311	0.3161
SmaAt-UNet	0.4773	0.6321	0.5190	0.4640	0.4145	0.3976	0.3794
ConvLSTM	0.4589	0.6302	0.5105	0.4311	0.4036	0.3911	0.3618
LPT-QPN	0.4867	<b>0.6445</b>	0.5384	0.4774	0.4306	0.3980	0.3659
SimVP	0.4699	0.6370	0.5278	0.4620	0.4010	0.3813	0.3541
STFFT	<b>0.5236</b>	0.6433	<b>0.5563</b>	<b>0.5240</b>	<b>0.4753</b>	<b>0.4571</b>	<b>0.4238</b>



**Figure 9.** A second visual example of the nowcasting results from all of the models.

**Table 7.** Quantitative comparison of the CSI values of the second visual example.

Model	Mean	5 min	25 min	45 min	65 min	75 min	90 min
RainNet	0.5108	0.6937	0.5337	0.4905	0.4710	0.4189	0.4040
SmaAt-UNet	0.5620	0.7706	0.5841	0.5666	0.5157	0.4840	0.4710
ConvLSTM	0.5378	0.7246	0.5789	0.5248	0.5028	0.4764	0.4272
LPT-QPN	0.5322	0.7529	0.5632	0.5315	0.4830	0.4381	0.4128
SimVP	0.5548	<b>0.7713</b>	0.5659	0.5531	0.5284	0.4995	0.4402
STFFT	<b>0.5783</b>	0.7603	<b>0.5952</b>	<b>0.5630</b>	<b>0.5596</b>	<b>0.5130</b>	<b>0.4918</b>

## 5. Summary and Conclusions

In this study, we propose a model called STFFT for precipitation nowcasting. In the proposed model, we use an encoder–forecaster framework which explicitly processes the temporal sequences in the encoder. The framework efficiently captures the global and long-distance information of spatiotemporal features for VIL sequences. In the encoder, the proposed STMSA learns local spatiotemporal features. The STMSA, which integrates the MaxPool and AveragePool operations, learns the higher intensities of the sequences and more efficiently forecasts severe weather. Then, the GSTFFN captures the global and long-distance spatial–temporal features and the model architecture effectively improves the results of longer forecasting times. In the decoder, the proposed MHSFFA and CAFFFN units based on a cross-feature fusion strategy capture the interactions between each two features. This design improves the accuracies of nowcasting longer times and at high intensities.

Based on the above analysis, we can obtain the following conclusions:

1. The components of an encoder consisting of STMSA and GSTFFN can effectively capture the global and long-distance spatial–temporal features; furthermore, the UHIFR integrated with STMSA strengthens the ability of model to learn the features of high-intensity pixels.
2. Based on the cross-feature fusion strategy, the MHSFFA and CAFFFN units in the decoder not only effectively simulate the movements of radar echoes by capturing the interactions of the echo sequences, but also more precisely nowcast the longer time information.
3. The quantitative and qualitative experiments demonstrate the effectiveness of the proposed model. In particular, the proposed model obtains better results for higher intensities and longer nowcasting times, which demonstrates that it pays more attention to such intensities and can capture the longer-distance features. The experimental results also demonstrate the superiority of our proposed model in forecasting severe weather and longer times information.

However, improvements are required for the proposed model. First, the input data are a simple source, which is only an image. In future research, we will consider multi-modal data [50], which include not only radar observations and satellite images, but also ground station observations and terrain data. At the same time, NWP [51] may provide a benefit supplement for the precipitation nowcasting. Furthermore, we will design a simpler network architecture to reduce the use of memory and GPU. Finally, the loss functions [52] play a very important role in training a deep-learning network. We will consider various loss functions that capture not only spatial features [53] but also the physical rule.

**Author Contributions:** Conceptualization, T.X., W.W. and J.H. (Jianxin He), methodology, R.S. and H.W., software, T.X. and H.W., validation, R.S. and J.H. (Jinrong Hu), data curation, R.S. and H.W., writing—original draft preparation, T.X., writing—review and editing, W.W., J.H. (Jianxin He) and J.H. (Jinrong Hu), supervision J.H. (Jianxin He), funding acquisition, T.X., J.H. (Jianxin He) and H.W. All authors have read and agreed to the published version of the manuscript.



**Funding:** This work was sponsored by the National Natural Science Foundation of China (U2342216), the Sichuan Provincial Central Leading Local Science and Technology Development Special Project (2023ZYD0147), the Project of the Sichuan Department of Science and Technology (2023NSFSC0244, 2023NSFSC0245), and the Open Grants of China Meteorological Administration Radar Meteorology Key Laboratory (2023LRM-A01), and was jointly funded by the Sichuan Science and Technology Program (No. 2023YFQ0072).

**Data Availability Statement:** The data presented in this study are publicly available on <https://registry.opendata.aws/sevir/>, (accessed on 1 June 2024).

**Conflicts of Interest:** The authors declare no conflicts of interest.

## References

1. Camporeale, J.E. The challenge of machine learning in space weather: Nowcasting and forecasting. *Space Weather* **2019**, *17*, 1166–1207. [CrossRef]
2. Fang, W.; Shen, L.; Sheng, V.S. VRNet: A Vivid Radar Network for Precipitation Nowcasting. *IEEE Trans. Geosci. Remote Sens.* **2024**, *62*, 1–11. [CrossRef]
3. Sun, J.; Xue, M.; Wilson, J.W.; Zawadzki, I.; Ballard, S.P.; Onvlee-Hooimeyer, J.; Joe, P.; Barker, D.M.; Li, P.W.; Golding, B.; et al. Use of NWP for nowcasting convective precipitation: Recent progress and challenges. *Bull. Am. Meteorol.* **2014**, *95*, 409–426. [CrossRef]
4. Reyniers, M. *Quantitative Precipitation Forecasts Based on Radar Observations: Principles, Algorithms and Operational Systems*; Institut Royal Météorologique de Belgique: Bruxelles, Belgium, 2008.
5. Rinehart, R.E.; Garvey, E.T. Three-dimensional storm motion detection by conventional weather radar. *Nature* **1987**, *273*, 287–289. [CrossRef]
6. Rinehart, R.E. A pattern recognition technique for use with conventional weather radar to determine internal storm motions. *Atmos. Technol.* **1981**, *13*, 119–134.
7. LeCun, Y.; Bengio, Y.; Hinton, G. Deep learning. *Nature* **2015**, *521*, 436–444. [CrossRef] [PubMed]
8. Krizhevsky, A.; Sutskever, I.; Hinton, G.E. Imagenet classification with deep convolutional neural networks. *Adv. Neural Inf. Process. Syst.* **2012**, *25*, 1097–1105. [CrossRef]
9. He, K.; Zhang, X.; Ren, S.; Sun, J. Deep residual learning for image recognition. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Las Vegas, NV, USA, 27–30 June 2016; Volume 39, pp. 770–778.
10. Long, J.; Shelhamer, E.; Darrell, T. Fully convolutional networks for semantic segmentation. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Boston, MA, USA, 7–12 June 2015; pp. 3431–3440.
11. Shi, X.; Chen, Z.; Wang, H.; Yeung, D.Y.; Wong, W.K.; Woo, W.C. Convolutional LSTM network: A machine learning approach for precipitation nowcasting. *Adv. Neural Inf. Process. Syst.* **2015**, *28*, 802–810.
12. Shi, X.; Gao, Z.; Lausen, L.; Wang, H.; Yeung, D.Y.; Wong, W.K.; Woo, W.C. Deep learning for precipitation nowcasting: A benchmark and a new model. *Adv. Neural Inf. Process. Syst.* **2017**, *30*, 5617–5627.
13. Wang, Y.; Long, M.; Wang, J.; Gao, Z.; Yu, P.S. PredRNN: Recurrent neural networks for predictive learning using spatiotemporal LSTMs. *Adv. Neural Inf. Process. Syst.* **2017**, *30*, 879–888.
14. Lin, Z.; Li, M.; Zheng, Z.; Cheng, Y.; Yuan, C. Self-attention ConvLSTM for spatiotemporal prediction. In Proceedings of the AAAI Conference on Artificial Intelligence, New York, NY, USA, 7–12 February 2020; Volume 34, pp. 11531–11538.
15. Xiong, T.; He, J.; Wang, H.; Tang, X.; Shi, Z.; Zeng, Q. Contextual Sa-Attention Convolutional LSTM for Precipitation Nowcasting: A Spatiotemporal Sequence Forecasting View. *IEEE J. Sel. Top. Appl. Earth Obs. Remote Sens.* **2021**, *14*, 12479–12491. [CrossRef]
16. Tan, C.; Li, S.; Gao, Z.; Guan, W.; Wang, Z.; Liu, Z.; Li, S.Z. OpenSTL: A Comprehensive Benchmark of Spatio-Temporal Predictive Learning. *Adv. Neural Inf. Process. Syst.* **2023**, *36*, 69819–69831.
17. Agrawal, S.; Barrington, L.; Bromberg, C.; Burge, J.; Gazen, C.; Hickey, J. Machine learning for precipitation nowcasting from radar images. *arXiv* **2019**, arXiv:1912.12132.
18. Ronneberger, O.; Fischer, P.; Brox, T. U-net: Convolutional networks for biomedical image segmentation. In Proceedings of the International Conference on Medical Image Computing and Computer-Assisted Intervention, Toronto, ON, Canada, 5–9 October 2015; pp. 234–241.
19. Ayzel, G.; Scheffer, T.; Heistermann, M. RainNet v1.0: A convolutional neural network for radar-based precipitation nowcasting. *Geosci. Model Dev.* **2020**, *13*, 2631–2644. [CrossRef]
20. Trebing, K.; Stanczyk, T.; Mehrkanoon, S. SmaAt-UNet: Precipitation nowcasting using a small attention-UNet architecture. *Pattern Recognit. Lett.* **2021**, *145*, 178–186. [CrossRef]
21. Chollet, F. Xception: Deep learning with depthwise separable convolutions. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Honolulu, HI, USA, 21–26 July 2017; pp. 1251–1258.
22. Guen, V.L.; Thome, N. Disentangling physical dynamics from unknown factors for unsupervised video prediction. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Seattle, WA, USA, 13–19 June 2020; pp. 11474–11484.

23. Li, D.; Deng, K.; Zhang, D.; Liu, Y.; Leng, H.; Yin, F.; Song, J. LPT-QPN: A Lightweight Physics-informed Transformer for Quantitative Precipitation Nowcasting. *IEEE Trans. Geosci. Remote Sens.* **2023**, *61*. [[CrossRef](#)]
24. Ritvanen, J.; Harnist, B.; Aldana, M.; Mäkinen, T.; Pulkkinen, S. Advection-free convolutional neural network for convective rainfall nowcasting. *IEEE J. Sel. Top. Appl. Earth Obs. Remote Sens.* **2023**, *16*, 1654–1667. [[CrossRef](#)]
25. Zhang, Y.; Long, M.; Chen, K.; Xing, L.; Jin, R.; Jordan, M.I.; Wang, J. Skilful nowcasting of extreme precipitation with NowcastNet. *Nature* **2023**, *619*, 526–532. [[CrossRef](#)]
26. Tian, L.; Li, X.; Ye, Y.; Xie, P.; Li, Y. A generative adversarial gated recurrent unit model for precipitation nowcasting. *IEEE Geosci. Remote Sens. Lett.* **2019**, *17*, 601–605. [[CrossRef](#)]
27. Ravuri, S.; Lenc, K.; Willson, M.; Kangin, D.; Lam, R.; Mirowski, P.; Mohamed, S. Skilful precipitation nowcasting using deep generative models of radar. *Nature* **2021**, *597*, 672–677. [[CrossRef](#)]
28. Gong, A.; Li, R.; Pan, B.; Chen, H.; Ni, G.; Chen, M. Enhancing spatial variability representation of radar nowcasting with generative adversarial networks. *Remote Sens.* **2023**, *15*, 3306. [[CrossRef](#)]
29. Bai, C.; Sun, F.; Zhang, J.; Song, Y.; Chen, S. Rainformer: Features extraction balanced network for radar-based precipitation nowcasting. *IEEE Geosci. Remote Sens. Lett.* **2022**, *19*, 1–5.
30. Vaswani, A.; Shazeer, N.; Parmar, N.; Uszkoreit, J.; Jones, L.; Gomez, A.N.; Polosukhin, I. Attention is all you need. *Adv. Neural Inf. Process. Syst.* **2017**, *30*, 6000–6010.
31. Dosovitskiy, A.; Beyer, L.; Kolesnikov, A.; Weissenborn, D.; Zhai, X.; Unterthiner, T.; Dehghani, M.; Minderer, M.; Heigold, G.; Gelly, S.; et al. An image is worth 16x16 words: Transformers for image recognition at scale. In Proceedings of the International Conference on Learning Representations, Vienna, Austria, 4 May 2021; p. 11929.
32. Liu, Z.; Lin, Y.; Cao, Y.; Hu, H.; Wei, Y.; Zhang, Z.; Lin, S.; Guo, B. Swin transformer: Hierarchical vision transformer using shifted windows. In Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), Montreal, BC, Canada, 11–17 October 2021; pp. 10012–10022.
33. Dong, X.; Bao, J.; Chen, D.; Zhang, W.; Yu, N.; Yuan, L.; Chen, D.; Guo, B. CSWin Transformer: A General Vision Transformer Backbone with Cross-Shaped Windows. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), New Orleans, LA, USA, 18–24 June 2022; pp. 12124–12134.
34. Chen, G.; Jiao, P.; Hu, Q.; Xiao, L.; Ye, Z. SwinSTFM: Remote Sensing Spatiotemporal Fusion Using Swin Transformer. *IEEE Trans. Geosci. Remote Sens.* **2022**, *60*, 1–18. [[CrossRef](#)]
35. Hu, Y.; Chen, L.; Wang, Z.; Li, H. SwinVRNN: A Data-Driven Ensemble Forecasting Model via Learned Distribution Perturbation. *J. Adv. Model. Earth Syst.* **2023**, *15*, e2022MS003211. [[CrossRef](#)]
36. Chen, L.; Du, F.; Hu, Y.; Wang, Z.; Wang, F. SwinRDM: Integrate SwinRNN with Diffusion Model towards High-Resolution and High-Quality Weather Forecasting. In Proceedings of the AAAI Conference on Artificial Intelligence, Washington, DC, USA, 7–14 February 2023; Volume 37, pp. 322–330.
37. Bi, K.; Xie, L.; Zhang, H.; Chen, X.; Gu, X.; Tian, Q. Accurate medium-range global weather forecasting with 3D neural networks. *Nature* **2023**, *619*, 533–538. [[CrossRef](#)] [[PubMed](#)]
38. Chen, L.; Zhong, X.; Zhang, F.; Cheng, Y.; Xu, Y.; Qi, Y.; Li, H. FuXi: A cascade machine learning forecasting system for 15-day global weather forecast. *NPJ Clim. Atmos. Sci.* **2023**, *6*, 190. [[CrossRef](#)]
39. Gao, Z.; Shi, X.; Wang, H.; Zhu, Y.; Wang, Y.B.; Li, M.; Yeung, D.Y. Earthformer: Exploring space-time transformers for earth system forecasting. In Proceedings of the 36th International Conference on Neural Information Processing Systems, New Orleans, LA, USA, 28 November–9 December 2022; Volume 35, pp. 25390–25403.
40. Chen, S.; Shu, T.; Zhao, H.; Zhong, G.; Chen, X. TempEE: Temporal–Spatial Parallel Transformer for Radar Echo Extrapolation Beyond Autoregression. *IEEE Trans. Geosci. Remote Sens.* **2023**, *61*. [[CrossRef](#)]
41. Li, W.; Zhou, Y.; Li, Y.; Song, D.; Wei, Z.; Liu, A.A. Hierarchical Transformer with Lightweight Attention for Radar-based Precipitation Nowcasting. *IEEE Geosci. Remote Sens. Lett.* **2024**, *21*, 1–5. [[CrossRef](#)]
42. Jin, Q.; Zhang, X.; Xiao, X.; Wang, Y.; Xiang, S.; Pan, C. Preformer: Simple and Efficient Design for Precipitation Nowcasting with Transformers. *IEEE Geosci. Remote Sens. Lett.* **2023**, *21*, 1–5.
43. Woo, S.; Park, J.; Lee, J.Y.; Kweon, I.S. CBAM: Convolutional Block Attention Module. In Proceedings of the European Conference on Computer Vision (ECCV), Munich, Germany, 8–14 September 2018; pp. 3–19.
44. Veillette, M.S.; Samsi, S.; Mattioli, C.J. SEVIR: AStormEvent Imagery Dataset for Deep Learning Applications in Radar and Satellite Meteorology. *Adv. Neural Inf. Process. Syst.* **2020**, *33*, 22009–22019.
45. Loshchilov, I.; Hutter, F. Decoupled weight decay regularization. In Proceedings of the 7th International Conference on Learning Representations (ICLR), Toulon, France, 24–26 April 2017; pp. 1–8.
46. Gao, Z.; Tan, C.; Wu, L.; Li, S.Z. SimVP: Simpler yet Better Video Prediction. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), New Orleans, LA, USA, 18–24 June 2022; pp. 3160–3170.
47. Barnes, L.R.; Schultz, D.M.; Gruntfest, E.C.; Hayden, M.H.; Benight, C.C. Corrigendum: False Alarm Rate or False Alarm Ratio? *Weather Forecast.* **2009**, *24*, 1452–1454. [[CrossRef](#)]
48. Guo, S.; Sun, N.; Pei, Y.; Li, Q. 3D-UNet-LSTM: A Deep Learning-Based Radar Echo Extrapolation Model for Convective Nowcasting. *Remote Sens.* **2023**, *15*, 1529. [[CrossRef](#)]
49. Schaefer, J.T. The Critical Success Index as an Indicator of Warning Skill. *Weather Forecast.* **1990**, *5*, 570–575. [[CrossRef](#)]

50. Ma, Z.; Zhang, H.; Liu, J. MM-RNN: MM-RNN: A Multimodal RNN for Precipitation Nowcasting. *IEEE Trans. Geosci. Remote Sens.* **2023**, *61*. [[CrossRef](#)]
51. Gilewski, P. Application of Global Environmental Multiscale (GEM) Numerical Weather Prediction (NWP) Model for Hydrological Modeling in Mountainous Environment. *Atmosphere* **2022**, *13*, 1348. [[CrossRef](#)]
52. Yang, S.; Yuan, H. A Customized Multi-Scale Deep Learning Framework for Storm Nowcasting. *Geophys. Res. Lett.* **2023**, *50*, e2023GL103979. [[CrossRef](#)]
53. Hu, J.; Yin, B.; Guo, C. METEO-DLNet: Quantitative Precipitation Nowcasting Net Based on Meteorological Features and Deep Learning. *Remote Sens.* **2024**, *16*, 1063. [[CrossRef](#)]

**Disclaimer/Publisher's Note:** The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.