remote sensing

MDPI

Article

Severe Precipitation Recognition Using Attention-UNet of Multichannel Doppler Radar

Weishu Chen [1,2,3], Wenjun Hua [1,2,3], Mengshu Ge [1,2,3], Fei Su [1,2,3,*], Na Liu [4], Yujia Liu [4] and Anyuan Xiong [4]

[1] Key Laboratory of Interactive Technology and Experience System, Ministry of Culture and Tourism, Beijing University of Posts and Telecommunications, Beijing 100876, China
[2] Beijing Key Laboratory of Network System and Network Culture, Beijing University of Posts and Telecommunications, Beijing 100876, China
[3] School of Artificial Intelligence, Beijing University of Posts and Telecommunications, Beijing 100876, China
[4] National Meteorological Information Center, Beijing 100081, China
* Correspondence: sufei@bupt.edu.cn

Abstract: Quantitative precipitation estimation (QPE) plays an important role in meteorology and hydrology. Currently, multichannel Doppler radar image is used for QPE based on some traditional methods like the $Z - R$ relationship, which struggles to capture the complicated non-linear spatial relationship. Encouraged by the great success of using Deep Learning (DL) segmentation networks in medical science and remoting sensing, a UNet-based network named Reweighted Regression Encoder–Decoder Net (RRED-Net) is proposed for QPE in this paper, which can learn more complex non-linear information from the training data. Firstly, wavelet transform (WT) is introduced to alleviate the noise in radar images. Secondly, a wider receptive field is obtained by taking advantage of attention mechanisms. Moreover, a new Regression Focal Loss is proposed to handle the imbalance problem caused by the extreme long-tailed distribution in precipitation. Finally, an efficient feature selection strategy is designed to avoid exhaustion experiments. Extensive experiments on 465 real processes data demonstrate that the superiority of our proposed RRED-Net not only in the threat score (TS) in the severe precipitation (from 17.6% to 39.6%, $\geq$20 mm/h) but also the root mean square error (RMSE) comparing to the traditional Z-R relationship-based method (from 2.93 mm/h to 2.58 mm/h, $\geq$20 mm/h), baseline models and other DL segmentation models.

Keywords: QPE; deep learning; precipitation; Doppler radar; wavelet; long-tailed distribution; attention; feature selection

check for updates

Citation: Chen, W.; Hua, W.; Ge, M.; Su F.; Liu, N.; Liu, Y.; Xiong, A. Severe Precipitation Recognition Using Attention-UNet of Multichannel Doppler Radar. *Remote Sens.* **2023**, *15*, 1111. https://doi.org/10.3390/rs15041111

Academic Editor: Kenji Nakamura

Received: 15 January 2023
Revised: 13 February 2023
Accepted: 15 February 2023
Published: 17 February 2023

Copyright: © 2023 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

Quantitative precipitation estimation (QPE) [1] is crucial for weather-dependent decision-making, including retail, agriculture, and the aviation industry. Particularly, QPE for severe precipitation events gains more attention because they rarely happen but affect human life critically. In the real application, Doppler weather radar is utilized for QPE [1]. Traditional QPE methods are based on the relationship between the radar reflectivity (Z) and precipitation rate (R) [2,3]. The radar reflectivity is strongly related to the precipitation rate in general. The exponential $Z - R$ equation $Z = aR^b$ is used for rainfall estimation, where a and b are hyper-parameters that are dependent on precipitation types, seasons, and districts. This $Z - R$ relationship has been used to estimate surface precipitation in meteorological operations. The National Mosaic and Multi-Sensor Quantitative Precipitation Estimation (NMQ) [3], for example, is broadly used for meteorological practices and research. However, the $Z - R$ relationship ignores the spatial correlation of the precipitation field, which will affect precipitation accuracy.

With the development of automatic rain gauges positioned around the world, it is possible to use data-driven-based machine learning methods to solve the QPE task by

using the rain gauge data as precipitation label [4]. A lot of traditional machine learning algorithms have explored the relationship between radar reflectivity and surface rain gauge measurement, such as support vector machine [1,5,6] and random forest [7–10]. Compared to the $Z - R$ relationship, these models can gather complex non-linear relationships, and preliminarily represent the spatial correlation. However, these extracted statistical features cannot capture the different spatial scale correlations in the precipitation field. Moreover, errors may exist during reflectance measurement, introducing noise to the Doppler radar data. Therefore, to model the noisy weather system, a more appropriate model rather than a point-wise linear or exponential model is needed.

Deep learning (DL) exhibits high flexibility in learning complex non-linear information from massive data. In recent years, we have witnessed the great success of DL in the fields of computer vision, and natural language processing [11,12]. To exploit DL on more specific tasks, the structure and parameters of the network need to be carefully designed. In the meteorological field, convolutional long short-term memory network [13], which captures the spatiotemporal relationship of the time series, achieves good performance in precipitation forecasting [14,15] and prediction of radar echoes [16,17]. Another DL framework, Generative adversarial network [18], is utilized to improve the spatial resolution of low-resolution atmospheric images [19]. Motivated by these applications on meteorological tasks, we find QPE and the image segmentation tasks in computer vision share similarities in some places.

Image segmentation tasks, such as cell segmentation and satellite image analysis, have benefited from the development of DL [20–23]. The capability of these DL networks in deriving different receptive field features of images inspires us that these networks possess the potential for QPE. Among them, UNet [21] adds the skip connections to the fully convolutional network (FCN) [20]. It combines the high-resolution features in the downsampling path with the upsampled output, thus, enlarging the receptive field and preserving the details in high-resolution feature maps simultaneously. Other networks, like the series of Deeplab [22,24,25], have more parameters that obtain better performance than UNet in the natural image segmentation tasks. QPE can be viewed as a pixel-to-pixel prediction task from the view of computer vision, which is similar to the image segmentation task. The similarities of these two tasks enlighten us about the capabilities of the image segmentation networks in digesting the hidden precipitation information in Doppler radar images. However, applying these networks to QPE by rote will not obtain satisfactory results, because there are differences that can not be ignored between QPE and image segmentation tasks.

First, Doppler radar signals contain various kinds of noise. The Doppler shift in weather radar return introduces the colored noise [26]. Clutter, weather, and water vapor emit radiation at all wavelengths, leading to thermal noise [27]. Thus, the receiver noise can produce a spurious linear relation between the spectral width and radar range [28]. Secondly, traditional segmentation networks like UNet rely on piles of convolution layers to gather information from the neighbor pixels, which restrains the ability to obtain the spatial relationship from further pixels. Thirdly, the precipitation follows so-called "long-tailed" distribution in the most region, where there are massive pixels representing low rainfall intensity while few pixels stand for severe precipitation in a precipitation image. The long-tailed problem in QPE caused by the small probability of the extreme precipitation occurrence is intractable because of the severe imbalance between "head" (low precipitation) and "tail" (high precipitation). Furthermore, it is important to handle this long-tailed problem due to the huge influence on human daily life brought by the severe precipitation.

In this study, a UNet-based network named Reweighted Regression Encoder-Decoder Net (RRED-Net) is proposed for QPE. Firstly, we introduce wavelet transform (WT) to address the first problem, as mentioned before. As for the long-tailed distribution problem in precipitation, the proposed new RF (Rregression Focal) loss can dynamically scale the MSE (mean square Error) loss in high precipitation. Moreover, there are 8 different Doppler radar features that contribute more or less to QPE. Distinct feature combinations have different

effects. Therefore, we design a feature selection method combining information gain [29] and Chi square [30] to avoid the brutal exhaustive selection. Experiments show that the proposed model performs better in cases using features more relevant to precipitation as inputs.

The rest of this paper is organized as follows. Section 2 describes our network architecture in detail, including wavelet transform, attention block, and the solution for long-tailed distribution. Section 3 introduces the dataset and the feature selection process. Section 4 expounds on the metrics we use in experiments. Experimental results are described in Section 5, and are discussed broadly in the following Section 6. Section 7 concludes this work and presents the goals for future work.

2. Method

As we know, the goal of multichannel Doppler radar QPE is to use multichannel Doppler radar images to generate precipitation images in a local region for a short time. This problem can be regarded as a cross-modal recognition problem from the perspective of machine learning, inspired by the spatiotemporal prediction problem definition in [13].

Suppose there's an observable system over a local region represented by an $M \times N$ grid, including M rows and N columns. For each cell in the grid, there are C measurements that are not independent. Thus, this observation can be represented by a tensor $X \in \mathbb{R}^{C \times M \times N}$. In the same region, another system $Y \in \mathbb{R}^{M \times N}$ is related to X. The problem is to generate the most likely Y given X:

$$\hat{Y} = \arg\max p(Y|X) \tag{1}$$

In our work, X stands for the 2D multichannel Doppler radar images after necessary feature selection and data preprocessing. Y denote the 2D precipitation images. $M \times N$ represents the size of Doppler radar images and labels. C stands for the number of selected Doppler radar features.

In this section, we first give the details of the 2D WT denoising method and the theoretical rationale for its efficiency in the QPE task. Then, the criss-cross attention module [31] is adopted in our RRED-Net to capture dense contextual information for feature conversion quickly. Finally, the RF loss is introduced for alleviating the long-tailed distribution in QPE task.

2.1. Denoise

Wavelet transform is a mathematical method that changes signals into a new domain for processing or further analysis. Harr 2D wavelet is employed to separate the high-frequency noise, which is illustrated in Figure 1. 2D WT is the extension of 1D WT, which iteratively utilizes wavelet decompsition filters to compute the wavelet coefficients. An original image is transformed into four images, including **A**pproximation (low frequency), **H**orizontal-high frequency, **V**ertical-high frequency and **D**iagonal-high frequency. The low-frequency image is retained, while other high-frequency images are discarded during processing.

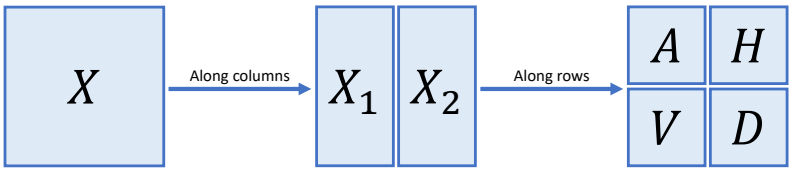

Figure 1. Illustration of 2D wavetlet transform (WT).

Because of the *strong region-related* and *weak non-linear* characteristics in Doppler radar images, the Harr wavelet transform is effective. Strong region-related is that, except for the undetectable area, the intensity of the Doppler radar image pixel is always continuous in a local region due to the physical limitation. Weak non-linear is inferred from our observation. As shown in Figure 2, the distributions of Doppler radar image patches are unimodal, while

the patches of natural images are multimodal or unimodal. Therefore, we can model the intensity of the Doppler radar image pixel as a normal distribution.

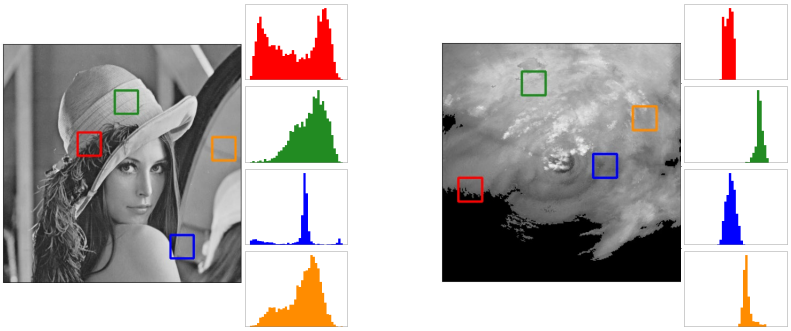

Figure 2. Distribution of random patches in a natural photo and a Doppler radar image.

Suppose p and q represent the intensities of two close pixels individually. The noisy observation of p' is

$$p' = p + n \tag{2}$$

q' is denoted as the noisy observation of q. n is the noise. On account of the weak non-linear characteristic, their relationship can be modeled by:

$$q' \sim \mathcal{N}(p, \sigma) \tag{3}$$

where σ is positively correlated to the p and the distance between these two pixels. The low frequency q_l is extracted by Harr wavelet transform when p' and q are adjacent points:

$$p_l = \frac{p' + q'}{2} \tag{4}$$

Equations (3) and (4) are combined as:

$$p_l = p + \frac{n + \sigma\mathcal{N}}{2} \tag{5}$$

where $\mathcal{N}$ is the standard normal distribution. Owing to the strong region-related, the $|\sigma|$ is poor and likely to overpass the $|n|$. Therefore, comparing p' in the Equation (2) and p_l in the Equation (5), the noise in the p' is relieved by Harr wavelet transform.

2.2. Network Architecture

QPE can be viewed as a pixel-to-pixel segmentation task in computer vision. UNet, one of the widely used networks in image segmentation tasks, exploits downsampling and upsampling to extract features in different receptive fields. At the same time, it utilizes a skip connection to rebuild the detailed information which is lost in downsampling. Inspired by the successful application of UNet and its variants in medical image segmentation [23,32,33], we build our RRED-Net for QPE based on the UNet. Furthermore, our model can capture the global information during the transformation between radar features and precipitation features by utilizing attention block, which avoids the defects of tiny receptive fields in the traditional methods.

Non-local [34] is an attention block in the field of computer vision. Compared to the convolution network, the non-local block enables each pixel to connect with any other pixels. The non-local block can be formulated as:

$$Q, K, V = w_q * X, w_k * X, w_v * X \tag{6}$$

$$Z = w_0 * \text{softmax}(Q \otimes K^T) \otimes V + X \tag{7}$$

where Q, K, V are attention vectors **Q**uery, **K**ey, **V**alue respectively. w_q, w_k, w_v are the corresponding transformation matrix for Q, K, V. $\otimes$ is the matrix multiplication after channel transformation. The structure of non-local is represented in Figure 3.

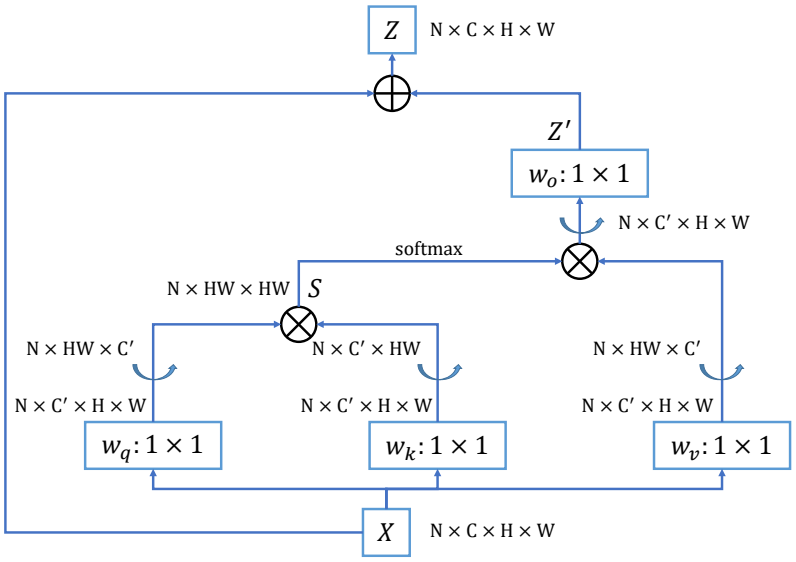

Figure 3. The structure of non-local [34] attention. N, C, H, and W stand for the batch size, channels, height, and weight of the input image X, respectively. w_q, w_k, w_v, and w_o stand for the different convolution kernels. S and Z' are the middle layers. Z is the output of the non-local attention.

However, as the caculation of non-local takes every pair of pixels into consideration, its computational cost is intolerable. For $X \in \mathbb{R}^{H \times W}$, both the space and time complexity are $O(HW \times HW)$. Many methods have tried to reduce the the computation complexity of attention blocks, including axial attention [35] and criss-cross attention, whose computaional complexity are $O((H+W) \times HW)$ and $O(max(H,W) \times HW)$, respectively. In our work, crisscross attention is applied to aggregate information in both horizontal and vertical directions rather than one direction in the axial attention module.

The difference between non-local, axial attention, and criss-cross attention is the way they extract attention vectors, as shown in Figure 4. Non-local exploits all pixels in the image to calculate Q, K, and V, while the other two employ a so-called **Affinity** operation, which focuses on the pixel of interest.

For each position u, we can get a vector $Q_u \in \mathbb{R}^{C'}$ from Q. Meanwhile, the set $\Omega_u \in \mathbb{R}^{(H+W-1) \times C'}$ is extractedfrom K, which has the same column or the same row as u. The ith element of Ω_u is defined as the $\Omega_{i,u} \in C'$. The **Affinity** operation in criss-cross attention is defined as follows:

$$S_{i,u} = \text{softmax}(Q_u \Omega_{i,u}^{\intercal}) \tag{8}$$

where $S_{i,u} \in S$ is the degree of correlation between Q_u and $\Omega_{i,u}$, $i = [1, ..., H+W-1]$. Then, a set $\Phi_u \in \mathbb{R}^{(H+W-1) \times C}$ is obtained from V_u just like how the Ω_u is extracted from K. Thus, giving the input X, the output Z of criss cross attention is formulated as follows:

$$Z_u = \sum_{i=0}^{H+W-1} S_{i,u} \Phi_{i,u} + X_u \tag{9}$$

where X_u and Z_u are vectors obtaining from the same position u in the spatial dimension of X and Z, respectively. Two consecutive criss-cross attention modules are used to harvest the contextual information of all the pixels in the input image.

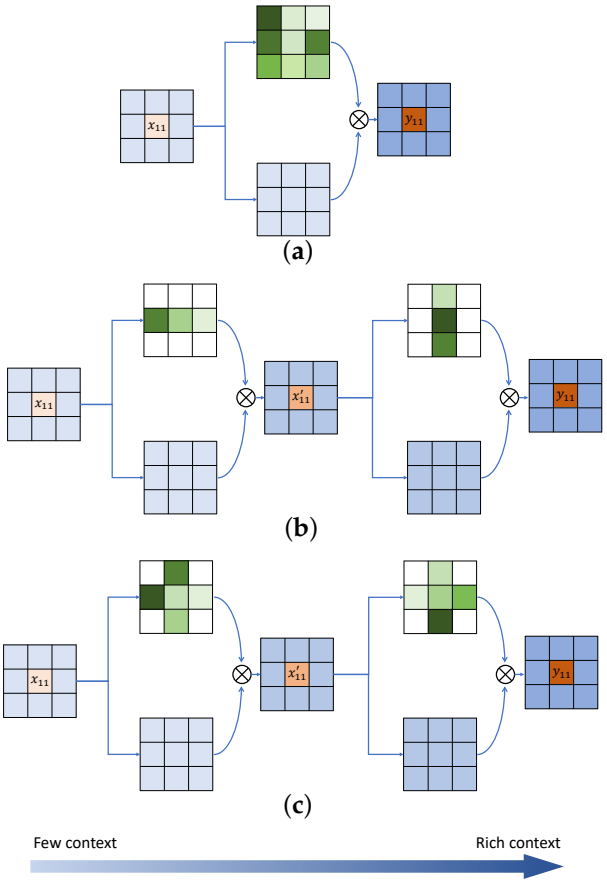

Figure 4. Implementation of different attention mechanisms. (**a**) is non-local attention. (**b**) is axial attention. (**c**) is criss-cross attention.

As shown in Figure 5, the attention modules are injected in the feature converter where the high-level representation extracted from Doppler radar images is translated to the representation of precipitation images. In Section 5.5, performances of axial attention and criss-cross attention will be evaluated and compared. Both downsampling and upsampling blocks contain three convolution blocks in RRED-Net. Each convolution block contains two 3 × 3 convolution layers with stride 1, which is followed by a batch normalization layer and a ReLU activation layer in sequence (3 × 3 1sConv, BN, ReLU). After each convolution block, there's a downsampling block (3 × 3 1sConv, BN, ReLU) or an upsampling block (4 × 4 2sDeconv, BN, ReLU). Skip connections are utilized between the corresponding convolution blocks in the downsampling module and upsampling module. In the feature converter module, there are three convolution blocks and two criss-cross modules.

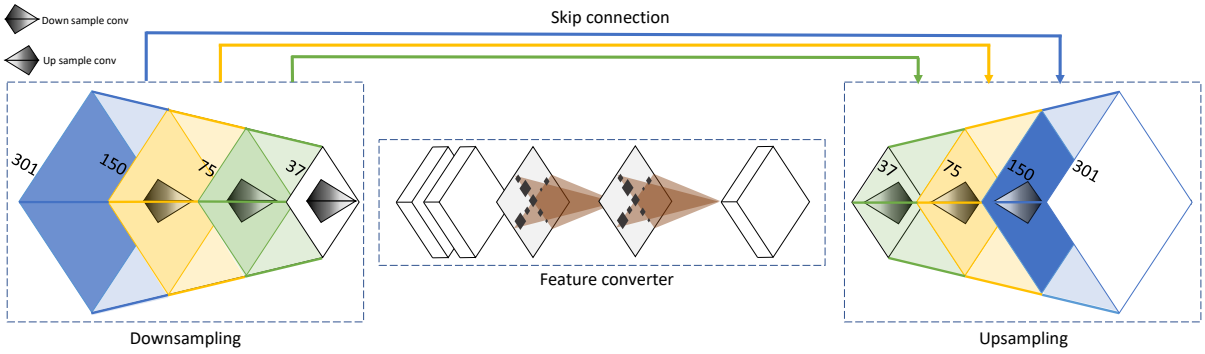

Figure 5. The network structure of the RRED-Net.

2.3. Long-Tailed Learning

In real life, precipitation always suffers from serious long-tailed effects. Figure 6 demonstrates the extreme imbalance between low precipitation and severe precipitation in our dataset for Anhui province in China from April to October, from 2016 to 2018. The precipitation higher than 3 mm/6 min only accounts for 0.9% of all the samples. In fact, long-tailed distribution also affects other meteorology tasks like nowcasting [14,15]. Zero precipitation and small precipitation occupy the vast majority of our dataset, leaving a barrier to accomplishing the accurate heavy precipitation recognition, which is rarer, but essentially affects human activities. In this work, a new loss function named regression focal loss (RF loss) is proposed to solve this kind of data imbalance problem. It belongs to re-weighting related methods.

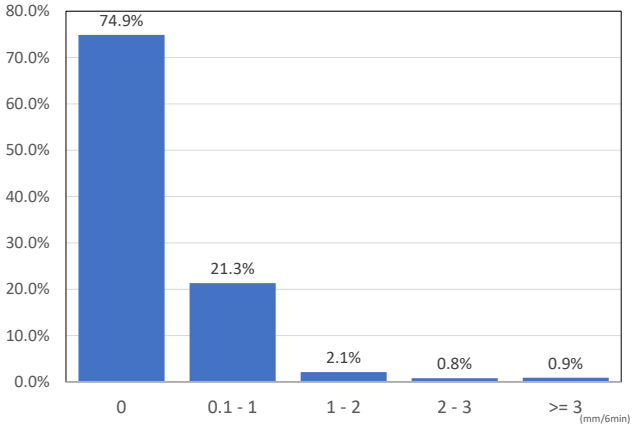

Figure 6. The occurrence frequencies of different rainfall intensity in our dataset, in Anhui province of China, from April to October from 2016 to 2018 (except October in 2018).

Based on the precipitation intensity, the precipitation is divided into 5 levels: 0 mm, 0.1–1 mm, 1–2 mm, 2–3 mm, and others. Each pixel is reweighted according to the probability p_i ($i \in \{1, 2, 3, 4, 5\}$) of its level:

$$w = \frac{1}{\sqrt[3]{p_i}} \tag{10}$$

where w is the weight of each pixel. This method is proved to be effective in Section 5. Furthermore, inspired by the focal loss [36], we also propose a new loss function, RF loss, to enhance the hard examples learning:

$$\mathrm{RF} = \sum (1 - r)^{\beta} \times w \tag{11}$$

$$r = \frac{\sum(|x_h - y| \leq d)}{\sum(x_h)} \tag{12}$$

where x_h is the high precipitation pixel in label images, y is the corresponding prediction, d is the error threshold , and β is a hyper-parameter.

The RF loss divides the high precipitation samples into hard samples and easy samples for modeling. If the recognition error for a single pixel exceeds the threshold d, it is considered as a hard sample. All the low precipitation pixels and well-recognized high precipitation pixels are regarded as easy samples. If there is no hard pixel, r is set as p. Intuitively, factor r reduces the learning contribution from easy examples and strengthens the influence in which an example has a low loss. So it owns the capability to dig into information in the hard samples. Therefore, this new loss helps us improve the performance of the high precipitation recognition. In our work, $d = 2$, $p = 0.9$, and $\beta = 0.25$.

It is also noted that there are side effects of changing the natural frequency of precipitation occurrence. The performance of zero precipitation and low precipitation slightly reduces after utilizing the long-tailed method. However, as we mentioned before, the model becomes more practical because the accuracy of high precipitation recognition is more significant. In fact, by combining other methods mentioned before, the zero and low precipitation performances are still better than the traditional method and baseline model.

3. Dataset and Features Selection

The radar data as the inputs of the model, and surface rainfall as the labels are taken from the dataset OpefBD [37]. QpefBD consists of 231,978 samples of 3185 heavy precipitation events that occurred from April to October 2016 to 2018 (except October 2018) in six provinces in central and eastern China. Our derived dataset includes 465 precipitation events in Anhui province from 2016 to 2018. An event includes a series of samples, each of which is at 6 minutes intervals. There are eight gridded products of weather radars and labels of precipitation intensity in a sample, as shown in Table 1 and Figure 7. The Doppler radar products are collected by 43 S-band Doppler weather radars with a wavelength of 10 cm and single polarization, and cover a rectangular are of $3° \times 3°$ centered on the radar station with 301×301 grids. The resolution is $0.01° \times 0.01°$. The label of precipitation intensity is a 2D precipitation image that shares the same size and resolution as the Doppler radar images. The precipitation image is interpolated by the precipitation observations collected at 15,652 weather stations. We use the Doppler radar images in a sample to predict the corresponding precipitation image. The testing set includes 95 events that happened in May, July, and September 2018, and other 370 events are used as the training set. The size of the radar image and precipitation image are both 301×301. The unit of precipitation images is millimeters per 6 min (mm/6min). All radar data are normalized, and the outliers are removed. Moreover, all the precipitation labels are normalized before rooted three to increase the influence of high precipitation in samples.

It should be noted that we use "label", which is commonly used in the field of computer vision, to refer to the the precipitation image throughout the paper. Similarly, we use "multichannel Doppler radar" to refer to the Doppler radar products, because the products occupy multiple image channels as input. We also call each Doppler radar product as Doppler radar "features" with the perspective of computer vison.

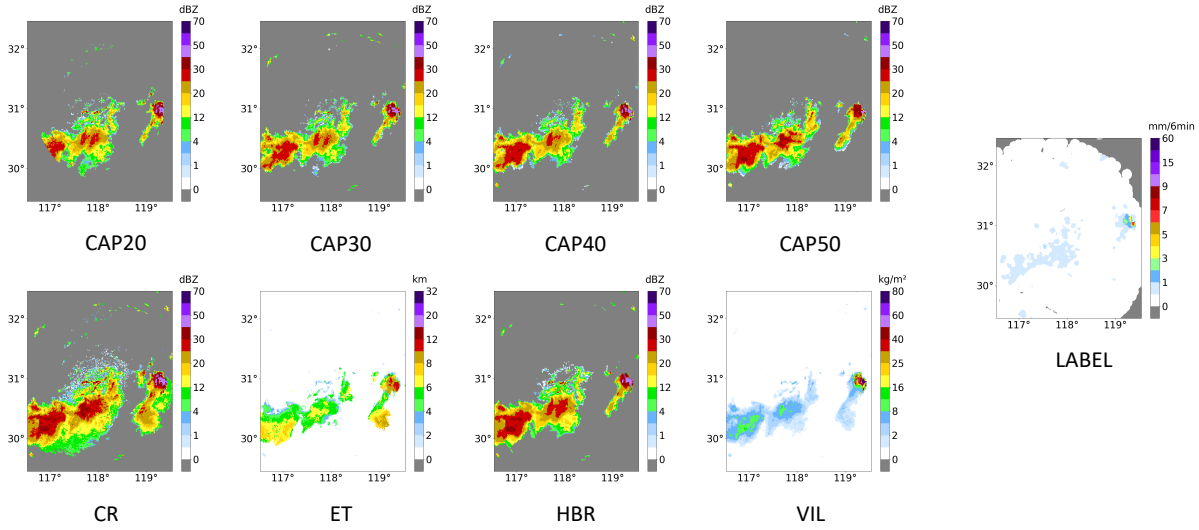

Figure 7. A visualization example of Doppler radar features and correspoding precipitation label, which comes from 4:01 Beijing Time, 14 July 2016. The left side of the figure is different channels (products) of Doppler radar, while the right side of the figure is the rainfall label.

Table 1. The full names and units of the input radar features.

Feature	Full Name	Unit
VIL	Vertically integrated liquid	kg/m²
HBR	Hybrid scan reflectivity	dBZ
CR	Composite reflectivity	dBZ
ET	Echo tops (18 dBZ)	km
CAPPIα	Constant altitude plan position indicator at the height with α hundred meters	dBZ

As there are 8 Doppler radar features in the dataset, as shown in Table 1 (CAPPIα includes CAPPI20, CAPPI30, CAPPI40, and CAPPI50), the selection of features will affect the model performances as shown in Section 5.2. Exhaustion for the best solution will cost $2^8 - 1 = 255$ times for every experiment. The information gain [29] and Chi-square gain [30] are applied for fast and skillful feature combination selection. The information gains can be fomulated as:

$$\mathrm{H}(X) = -\sum_{i=1}^{n} p(x_i) \log p(x_i) \tag{13}$$

$$\mathrm{H}(Y|X) = -\sum_{x \in X} p(x) \sum_{y \in Y} \log p(y|x) \tag{14}$$

$$\mathrm{Gain}(X) = \mathrm{H}(X) - \mathrm{H}(Y|X) \tag{15}$$

where $\mathrm{H}(X)$ is information entropy, $\mathrm{H}(Y|X)$ is conditional entropy, and $\mathrm{Gain}(X)$ is information gain. The larger information gain means this feature contributes more to recognition. In our work, X stands for the 2D multichannel Doppler radar images. Y denotes the corresponding 2D precipitation images. It is noted that information gain is often used in classification systems [38,39]. To utilize the information gain in our regression task, the precipitation is divided into 30 classes, i.e., 0 mm, 0.1 mm, 0.2 mm ... 2.8 mm, and $\geq$2.9 mm. Based on these 30 precipitation classes, the conditional entropy and information gain are calculated for each Doppler radar feature, as shown in the Table 2.

For the Chi-square test, precipitation and each radar are divided into 30 intervals equally, as shown in Table 3. Therefore, all the radar pixels in a sample can be classified into 900 bins according to their radar reflectivity and corresponding rain intensity. $n_{i,j}$ is defined as the number of pixels whose radar reflectivity is in the ith interval and rain intensity is in the jth interval. R_i, P_j and N are calculated by $\sum_j n_{i,j}$, $\sum_i n_{i,j}$ and $\sum_i \sum_j n_{i,j}$. The Chi-square gain χ^2 of each radar feature can be formulated as

$$\chi^2 = \sum_{i,j} \frac{\left(n_{i,j} - \frac{R_i P_j}{N}\right)^2}{\frac{R_i P_j}{N}} \tag{16}$$

If the radar feature is independent of the precipitation, we can get:

$$n_{i,j} - \frac{N_i N_j}{N} \approx 0. \tag{17}$$

Therefore, the larger $n_{i,j}$ means a stronger relationship between the radar feature and precipitation.

Table 2. Conditional entropy, information gain, and Chi-square of each radar features.

Feature	Conditional Entropy	Information Gain	Chi Square (10^7)
VIL	1.3521	0.3467	1.62
HBR	1.3496	0.3467	1.85
CR	1.3347	0.3641	1.77
ET	1.4145	0.2843	1.16
CAPPI20	1.5047	0.1941	1.12
CAPPI30	1.4095	0.2893	1.55
CAPPI40	1.3693	0.3295	1.56
CAPPI50	1.3929	0.3059	1.39

Table 3. Bins for calculating the Chi square.

	j = 1	**j = 2**	**...**	**j = 30**	**sum$_i$**
i = 1	$n_{1,1}$	$n_{1,2}$	...	$n_{1,30}$	R_1
i = 2	$n_{2,1}$	$n_{2,2}$	...	$n_{2,30}$	R_2
...	...	...	...	...	
i = 30	$n_{30,1}$	$n_{30,2}$	...	$n_{30,30}$	R_{30}
sum$_j$	P_1	P_2	...	P_{30}	N

Table 2 displays the conditional entropy, information gain, and Chi-square of all the features. Sorted by information gain, the top 5 features are as follows: CR > HBR > VIL > CAPPI40 > CAPPI50. Sorted by Chi square, the top 5 are as follows: HBR > CR > VIL > CAPPI40 > CAPPI30. Both evaluation criteria for the feature selection manifest that CR, HBR, VIL, and CAPPI40 may significantly contribute to the recognition accuracy. Experiments in Section 5.2 prove the effectiveness of our feature selection method. Based on this observation, these four features in the following experiments are selected.

4. Evaluation Metrics

The following metrics are applied to evaluate the performance of our model.

(1) Root-mean-square error (RMSE) and Bias: RMSE and bias are used to measure the difference between the prediction and the real precipitation label, which are the most used evaluation metrics in regression tasks.

$$\text{RMSE} = \frac{1}{N}\sum_{n=1}^{N}\sqrt{(\hat{y}_n - y_n)} \tag{18}$$

$$\text{Bias} = \frac{1}{N}\sum_{n=1}^{N}(\hat{y}_n - y_n) \tag{19}$$

where y_n and $\hat{y}_n$ are real precipitation label and forcasted precipitation, respectively. However, due to the long-tailed distribution in the precipitation images, it is noted that these two metrics are easily affected by the large-scale low-intensity precipitation and can hardly reflect the quality of high precipitation recognition.

(2) Probability of detection (POD), false alarm ratio (FAR), and threat score (TS): Besides the regression errors illustrated above, it is also significant to describe the recognition performance at certain precipitation levels. This can be evaluated by POD, FAR, and TS. All these three metrics are based on binary metrics, which are true positive (TP), false negative (FN), and false positive (FP). TP is the number of correct recognition samples, which means both observation and estimation belong to the level of interest. FN is the number of misreporting samples where the model incorrectly indicates the absence of this precipitation level condition when it is actually present. FP is the number of vacancy forecasts where the model incorrectly indicates the precipitation level.

Therefore, POD is defined as follows:

$$\mathrm{POD} = \mathrm{TP}(\mathrm{TP} + \mathrm{FN})^{-1} \tag{20}$$

POD represents the ratio of correct recognition of a certain precipitation level. A higher POD usually means better QPE performance. However, POD is generally not used alone as an evaluation metric since it does not work in some cases. For example, if model only predicts high precipitation, its heavy rainfall POD is near 1. The deficiencies of POD are often made up by FAR, which is defined as follows:

$$\mathrm{FAR} = \mathrm{FP}(\mathrm{TP} + \mathrm{FP})^{-1} \tag{21}$$

According to the definition, the FAR represents the ratio of misreporting of a certain rain level. Generally, a model with high POD and low FAR shows good performance.

TS is another metric apart from POD and FAR:

$$\mathrm{TS} = \mathrm{CSI} = \mathrm{FP}(\mathrm{TP} + \mathrm{FN} + \mathrm{FP})^{-1} \tag{22}$$

TS is the combination of POD and FAR, which considers both FN and FP's negative impact on the model. A higher TS means a better prediction model. It is noted that TS is also called the critical success index (CSI).

5. Experiments

5.1. Training Setting

Our proposed method is implemented in PyTorch [40]. All the models are initialized from scratch with random weights at batch size 64. During training, we use the Adam optimizer [41] with a learning rate of 0.001 at the beginning. The learning rate is divided by 10 every 5 epochs from the 10th epcoh to the 25th epoch, with 40 training epochs in total. All the experiments are conducted on a machine with Tesla V100 GPU (32GB memory), 20-core Intel Xeon E5-2698 V4 CPU (2.20 GHz).

5.2. Validation of the Features Selection

UNet [21] is utilized as our baseline model and select one of CR, CAPPI20, and ET as a feature to train the baseline model to evaluate the effectiveness of the feature selection method described in Section 3. As shown in Table 4, the influence of CR feature on rainfall is greater than ET and CAPPI20 based on the information gain and Chi square.

Table 4. Ranking of influence of each radar product on precipitation.

	Influence Degree of Each Radar Feature on Precipitation
Information gain	CR > HBR > VIL > CAPPI40 > CAPPI50 > CAPPI30 > ET > CAPPI20
Chi-square	HBR > CR > VIL > CAPPI40 > CAPPI30 > CAPPI50 > CAPPI20 > ET

From Table 5, the CR feature achieves the best performance in all indicators, which suggests that CR possesses more 'understandable' information for QPE. This is consistent with our conclusion in Section 3, that the higher the information gain and Chi-square value, the stronger the relationship between the radar feature and precipitation.

Table 5. Comparison of using single radar feature as input.

Feature	**RMSE** (mm/6 min)	**BIAS** (mm/6 min)	**0 mm** POD/FAR/TS(%)	**≥0.1 mm** POD/FAR/TS(%)	**≥1 mm** POD/FAR/TS(%)	**≥2 mm** POD/FAR/TS(%)	**≥3 mm** POD/FAR/TS(%)
CR	0.4508	0.0340	94.0/7.2/87.6	71.7/24.4/58.2	72.4/53.5/39.5	58.0/55.2/33.8	40.1/54.3/27.2
CAPPI20	0.4614	0.0115	92.6/8.3/85.5	70.0/27.6/55.2	70.0/27.6/55.2	65.6/50.9/39.0	33.0/50.9/24.6
ET	0.4782	0.0163	93.4/7.8/86.6	71.0/25.4/57.2	58.8/58.5/32.1	39.2/59.7/24.8	22.6/58.8/17.1

Table 6 illustrates the results of various combinations of feature selections. From the results, 'CR+HBR+VIL+CAPPI40' has the best performance on TS in all the precipitation levels. Other Doppler radar features (ET, CAPPI20, CAPPI30, CAPPI50) adding slightly reduce the TS, which claims that these features hinder the information representation ability of the model of their weaker relationship with the precipitation. Meanwhile, deleting any feature of 'CR+HBR+VIL+CAPPI40' results in decreasing of TS.

Table 6. Comparison of different radar feature combinations as input. Bold indicates that the best indicator in the current list.

Feature [1]	RMSE (mm/6 min)	BIAS (mm/6 min)	0 mm POD/FAR/TS(%)	≥0.1 mm POD/FAR/TS(%)	≥1 mm POD/FAR/TS(%)	≥2 mm POD/FAR/TS(%)	≥3 mm POD/FAR/TS(%)
CR	**0.4508**	0.0340	94.0/7.2/**87.6**	71.7/24.4/58.2	72.4/53.5/39.5	58.0/55.2/33.8	40.1/54.3/27.2
+HBR	0.5554	0.0372	89.8/11.5/80.4	75.8/21.9/62.5	73.6/49.9/42.5	58.4/53.0/35.2	39.9/52.4/27.7
+VIL	0.5502	0.0282	90.4/11.9/80.6	74.7/21.1/62.2	72.5/49.1/42.7	56.5/51.6/35.2	38.5/51.0/27.5
+CAPPI40	0.5453	0.0291	90.4/11.8/80.7	74.7/21.2/**62.2**	72.8/48.6/**43.3**	57.0/51.3/**35.6**	40.4/51.1/**28.4**
+ET	0.5445	0.0217	91.1/12.4/80.7	73.1/20.3/61.6	71.4/48.3/43.0	56.1/50.9/35.4	40.5/51.4/28.3
+CAPPI50	0.5340	**0.0195**	91.3/11.8/81.4	73.4/20.6/61.7	71.2/47.5/**43.3**	55.9/50.6/35.5	39.0/49.8/28.1

[1] '+' means we add the new feature compared to the previous line. For example, '+HBR' represents features 'CR+HBR' and '+VIL' represents features 'CR+HBR+VIL'.

5.3. Comparison with Other DL Networks

Table 7 displays the performance of four models, including the proposed RRED-Net and other three different popular segmentation networks. The **#params** is the number of parameters of each model. RMSE and bias roughly judge the overall consistency between predictions and labels. The 6 min precipitation is divided into 5 levels, and the POD, FAR, and TS (namely CSI) are calculated to underline each model's performance at high rainfall. Smaller RMSE, bias closing to zero, larger POD, smaller FAR, and larger TS indicate a better model. UNet* stands for using MSE as a loss function without reweighting, while other models reweight the label as Equation (10) before sending them to the loss function.

Table 7. Comparison of RRED-Net and other segmentation networks. Bold indicates that the best indicator in the current list.

Feature	#params	RMSE (mm/6 min)	BIAS (mm/6 min)	0 mm POD/FAR/TS(%)	≥0.1 mm POD/FAR/TS(%)	≥1 mm POD/FAR/TS(%)	≥2 mm POD/FAR/TS(%)	≥3 mm POD/FAR/TS(%)
UNet*	∼ 9.4 M	0.5472	−0.0770	94.8/15.1/**81.1**	64.5/14.4/58.2	45.7/31.7/37.7	30.9/35.7/26.3	20.4/38.6/18.1
UNet	∼ 9.4 M	**0.5428**	0.0252	90.5/11.7/80.8	74.7/21.1/62.3	72.4/48.4/**43.1**	55.9/50.5/35.6	37.9/49.6/27.6
CED	∼ 11.7 M	0.5465	0.0342	89.7/11.5/80.4	75.6/22.2/62.6	73.5/50.0/42.3	57.3/51.9/35.4	38.0/49.9/27.5
DeepLabV3	∼ 80 M	0.5603	−**0.0083**	91.6/13.0/80.6	71.3/19.8/60.6	61.9/44.9/41.2	45.6/49.6/31.5	31.0/52.4/23.1
RRED-Net	∼ 10.6 M	0.5524	0.0354	90.7/11.8/80.8	75.8/21.9/**62.7**	72.9/48.8/**43.1**	60.8/54.1/**35.7**	45.7/54.5/**29.6**

Surprisingly, the largest network, DeepLabV3, performs the worst, which suggests that QPE is easy to overfit because of the small dataset. Notice that the baseline network, UNet, shows improvements in RMSE and bias and little performance difference compared to the rough network CED (Convolutional Encoder and Decoder) while slightly reducing the parameters. Therefore, the proposed RRED-Net is designed based on UNet architecture, consisting of skip connections. RRED-Net achieves nice improvements, especially in the high precipitation recognition with tolerable degradation in RMSE and bias than UNet.

We also give the visualization of 6-min rainfalls of prediction and truth as examples, as shown in Figure 8. Correspoding composite reflectivities (CR) are also shown. The precipitation belt, especially for severe precipitation, matches well between the ground truth and the model's output. Notably, the output is evidently smoother than the label. It is because our proposed RF loss is an MSE-based loss, which tends to avoid abrupt prediction changes in the neighborhood.

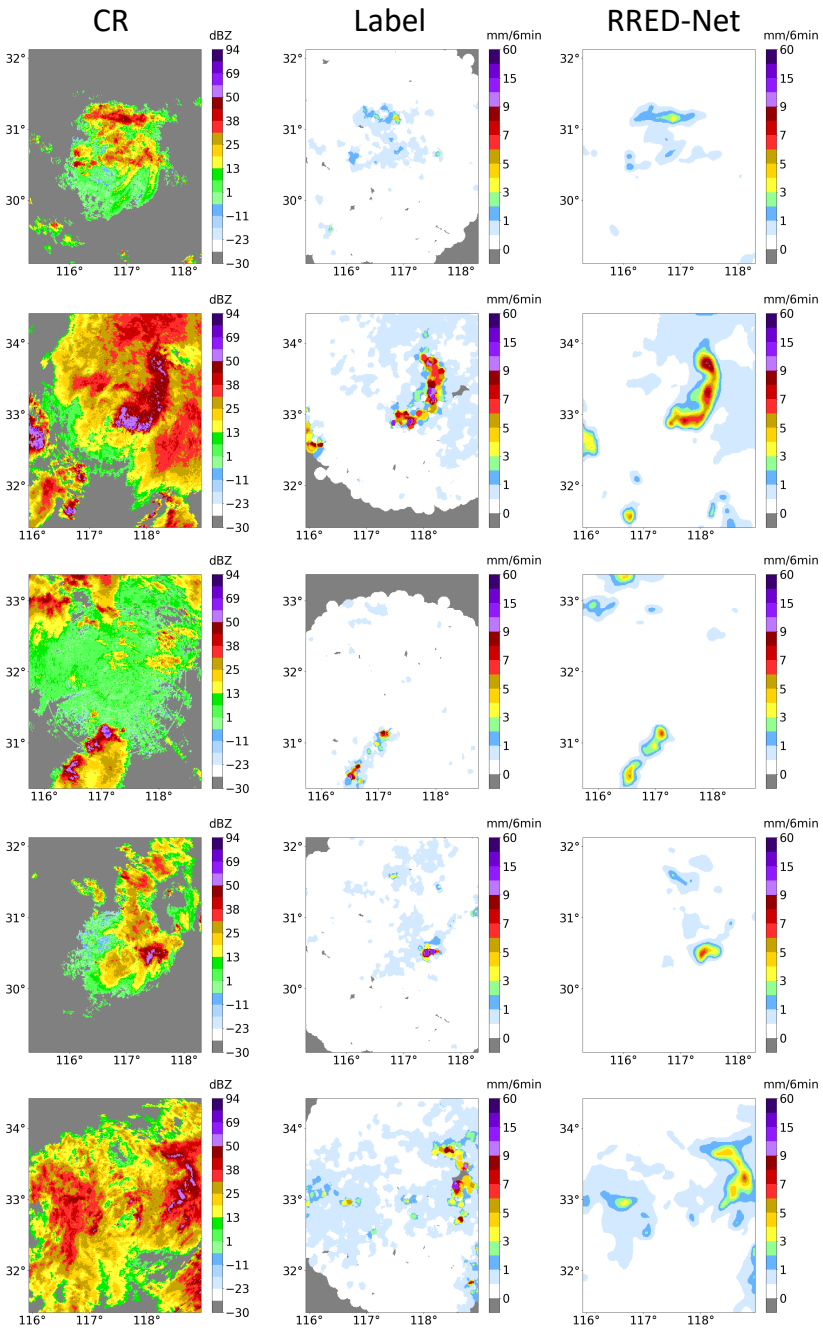

Figure 8. Visualization examples. The precipitation images come from 18:20–18:26 BT (Beijing Time), 2 July 2018; 20:38–20:42 BT, 26 July 2018; 22:43–20:48 BT, 4 July 2018; 09:30–09:36 BT, 26 July 2018; 10:40–10:46 BT, 5 July 2018 from up to down, respectively. CR Doppler radar, predictions of Label, and RRED-Net from left to right, respectively.

5.4. Comparison with Z−R Based Method

Some visualization examples of RRED-Net and $Z - R$ relationship-based methods are given in Figure 9. The red box of the first and the second examples demonstrates that the RRED-Net performs better in predicting high precipitation compared to the $Z - R$ based method. The third and fourth examples show some misreports of the precipitation given by the $Z - R$ based method, while the RRED-Net gives better results. It is obvious that our proposed method could generate the prediction by extracting useful information from other pixels in which the Doppler radar data is absent.

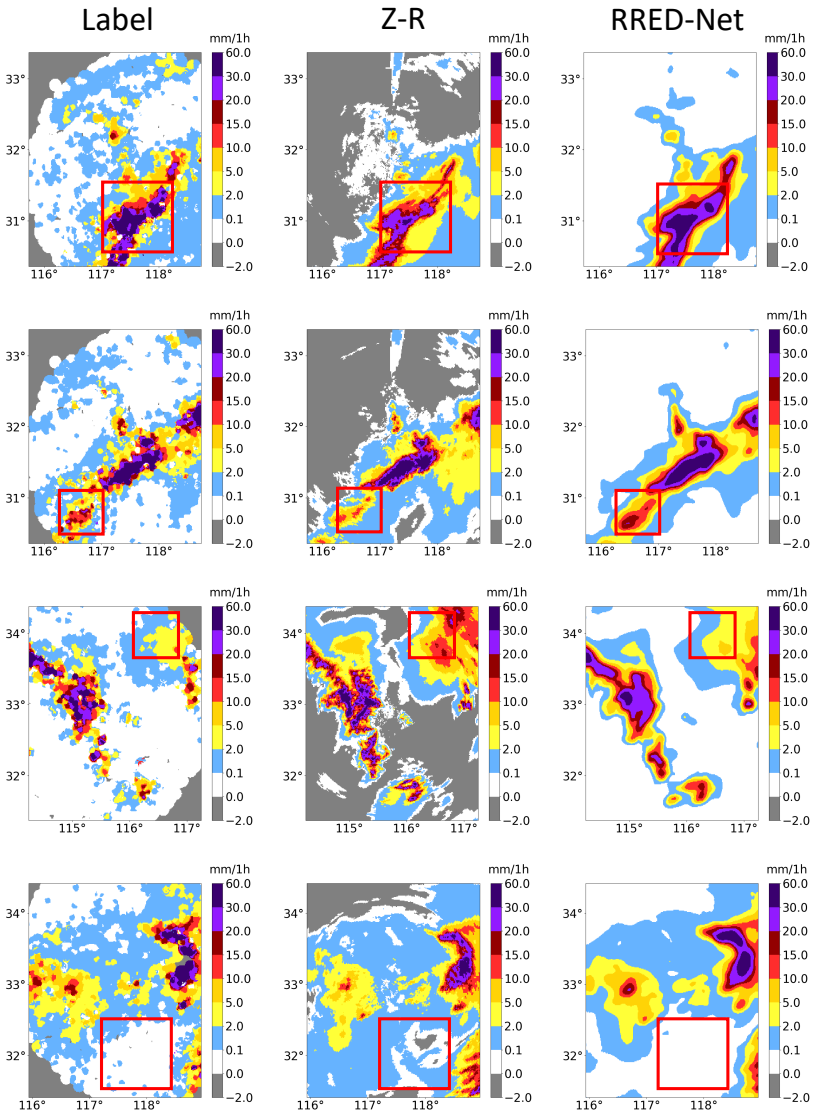

Figure 9. Visualization examples. An hour precipitation images measured and predicted during 21:41–22:40 BT (Beijing Time), 5 July 2018; 19:47–20:46 BT, 5 July 2018; 19:43–20:42 BT, 26 July 2018; 11:08–12:07 BT, 5 July 2018, from up to down, respectively. Label, the $Z - R$ relationship from left to right and the predictions of RRED-Net, respectively. The red box highlights parts of the images.

Moreover, compared with the traditional $Z - R$ relationship-based method, RRED-Net performs better (see Table 8) in nearly all the experimental indicators. Especially in the severe precipitation ($\geq$20 mm/h), RRED-Net doubles the TS from 17.6% to 39.6%. The parameter for $Z - R$ relationship $Z = aR^b$ we used are $a = 300$ and $b = 1.4$ [42].

Table 8. Comparison of $Z - R$ relationship based traditional method and our RRED-Net.

	RMSE (mm/h)	BIAS (mm/h)	0 mm TS(%)	$\geq$0.1 mm TS(%)	$\geq$5 mm TS(%)	$\geq$10 mm TS(%)	$\geq$20 mm TS(%)
$Z - R$	2.9354	-0.5355	52.6	67.2	37.7	27.9	17.6
RRED-Net	2.5775	0.2436	81.2	67.0	57.1	51.0	39.6

5.5. Ablation Studies

Ablation experiments are carried out to shed more light on the performance gains incurred by the different modules in RRED-Net. Wavelet transform (WT), criss-cross attention (CC), and regression focal loss (RF) are investigated separately. Experiment results show that all these methods can improve the performance of TS compared to the baseline.

Table 9 presents the ablation results on the RMSE, bias, and TS of five different levels. As can be observed, all three modules provide TS accuracy enhancement at the highest precipitation level. The complete model, RRED-Net, achieves the best performance at this level. These results prove that these modules contribute to better high rainfall recognition. Any of the proposed components can contribute to the high rainfall recognition. Specifically, the model with criss-cross attention gains increases at low precipitation as well. The regression focal loss has negative impacts on zero precipitation. This is reasonable as the regression focal loss sacrifices the low rainfall recognition accuracy by design. Moreover, the TS of precipitation greater than 1 mm also declined due to more misidentification of low precipitation as high precipitation. After using RF loss, FAR increased from 48.4% to 51.9%, while POD increased from 72.4% to 74.5%. By combining all these three modules, RRED-Net achieves the best TS of severe precipitation while retaining the TS of the light precipitation.

Table 9. Ablation studies on different components. UNet baseline, baseline with WT, baseline with CC, baseline with RF and RRED-Net from up to down, respectively.

WT	CC	RF	RMSE (mm/6 min)	BIAS (mm/6 min)	0 mm TS(%)	≥0.1 mm TS(%)	≥1 mm TS(%)	≥2 mm TS(%)	≥3 mm TS(%)
			0.5428	0.0252	80.8	62.3	43.1	35.6	27.6
✓			0.5428	0.0276	80.7	62.1	43.3	35.7	28.1
	✓		0.5538	0.0364	81.0	62.5	43.1	35.7	29.4
		✓	0.5667	0.0603	80.3	63.5	41.8	35.1	29.0
✓	✓	✓	0.5524	0.0354	80.8	62.4	43.1	35.7	29.6

Axial attention and criss-cross attention are also evaluated when designing the model, as shown in Table 10. Experimental results indicate that criss-cross attention is slightly better in RMSE, BIAS, and TS of all the levels. Therefore, criss-cross attention is chosen as our attention block.

Table 10. Comparison of axial attention and criss-cross attention (CC).

	RMSE (mm/6 min)	BIAS (mm/6 min)	0 mm TS(%)	≥0.1 mm TS(%)	≥1 mm TS(%)	≥2 mm TS(%)	≥3 mm TS(%)
Axial	0.5551	0.0458	80.3	62.2	42.7	35.6	29.1
CC	0.5538	0.0364	81.0	62.5	43.1	35.7	29.4

Furthermore, our proposed RF loss and an elaborately designed re-sampling strategy are compared, as shown in Table 11. Re-sampling is one of the methods to deal with class imbalance [43]. From Table 11, it is found that simply oversampling or downsampling the complete image will result in bad performance. Thus, all the precipitation images are split into smaller patches (96 × 96, overlapping rate 0.7) in the training and testing dataset. In the training stage, low precipitation patches are dropped at a fixed probability (0.2). During testing, the small patches are spliced into a complete image. As for the overlapped region, the indicators are calculated respectively in different patches and then averaged. Our proposed RF loss function gets worse performance in RMSE and BIAS. However, RF loss shows a 1.0% improvement in the high rainfall recognition compared to the resampling strategy. That is because the resampling strategy is easily affected by the recognition error in the edge area of the patch.

Table 11. Comparison of RF loss and re-sampling method.

	RMSE (mm/6 min)	BIAS (mm/6 min)	0 mm TS(%)	≥0.1 mm TS(%)	≥1 mm TS(%)	≥2 mm TS(%)	≥3 mm TS(%)
Resample	0.5518	0.0421	79.9	62.1	42.4	35.0	28.0
RF	0.5667	0.0603	80.3	63.5	41.8	35.1	29.0

How the factors in RF loss affect the performance of dealing with the long-tailed distribution is also explored. As for the d, which represents the allowable error range, we observe that bigger d is beneficial to the low precipitation recognition but harmful

to the high rainfall recognition. Figure 10a–c show that the recognition performance of low precipitation becomes better as d gets bigger, The best high precipitation performance happens when d is 2, and falls into a decline as d gets bigger, as shown in Figure 10f,g. p controls the simple sample's weight in the RF loss. It is noted that setting p as 1 is a better solution for severe rainfall recognition without too much loss of recognition accuracy of low rainfall. As for the β, along with β, the POD of high precipitation also manifests an upward trend. However, because the false alarm rate also increased, its TS score did not increase significantly. On the contrary, both the bias and the RMSE increase gradually. Therefore, supported by the experiments, a smaller β to balance the POD and the FAR of the model is recommended.

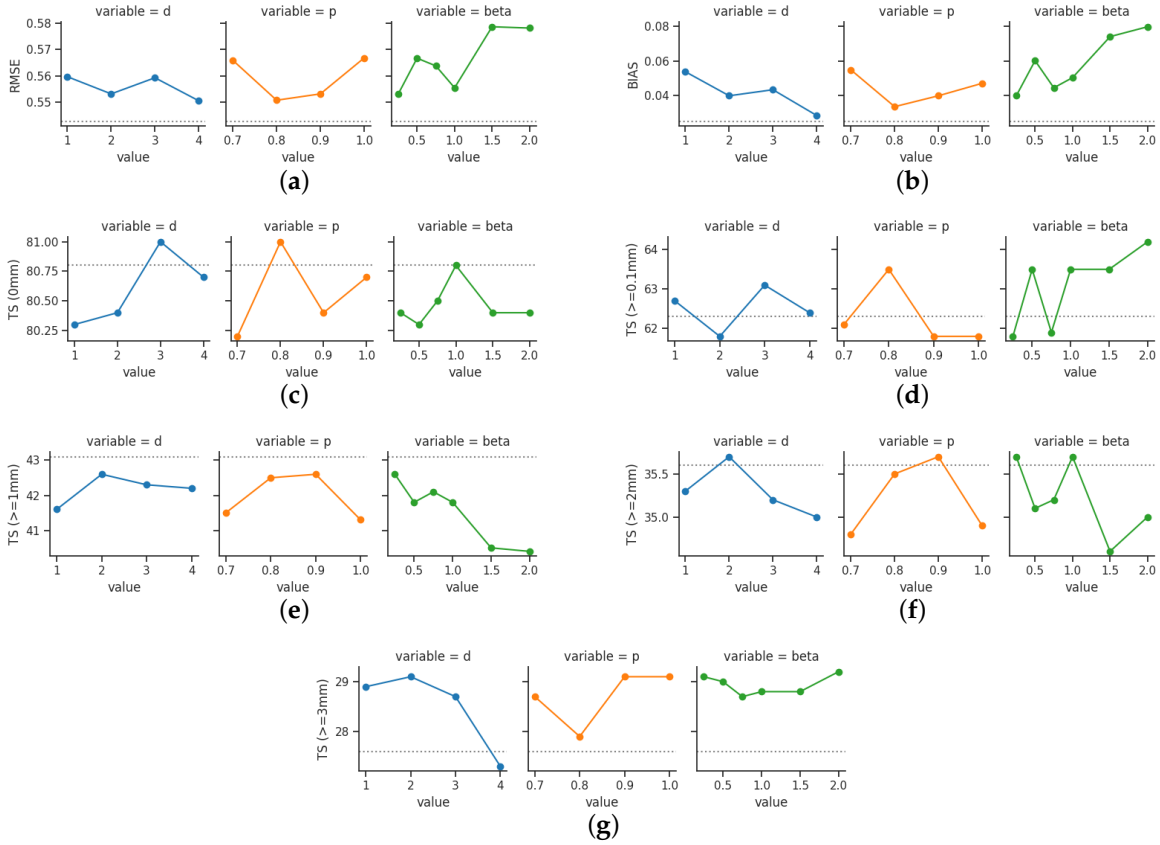

Figure 10. The effect of factors (d, p, and β (beta) in Equation (12)) in regression focal loss on the model performance. (**a**–**g**) represents RMSE, bias, TS (0 mm), TS (≥0.1 mm/ 6min), TS (≥1 mm/ 6min), TS (≥2 mm/6 min), TS (≥3 mm/6 min), respectively. Gray dotted horizontal lines indicate the performance of the weighted model without using RF loss. The default of d, p, and β are 2, 0.9, and 0.25, respectively.

6. Discussion

The objective of better high precipitation recognition in QPE is reached through our proposed RRED-Net combining DL network and meteorological images, i.e., Doppler radar features and ground precipitation intensity images in this work. Two interconnected elements are vital to implementing this method successfully. One is the preprocessing method for the Doppler radar features and the labels, and the other is the solution for the long-tailed distribution in the ground precipitation intensity images.

6.1. Preprocess Method

In Section 3, the details of the feature selection method are elaborated. The experiments shown in Tables 5 and 6 prove that the feature selection scheme not only avoids exhaustive experiments but also improves the QPE performance by measuring the statistical information of radar and precipitation. Moreover, when training the DL model, the training

images are often preprocessed to expand the training set (such as random rotation and random resize) and highlight or suppress certain attributes (such as a filter or histogram equalization). Some of them apply to the QPE task, while others are not. For example, random rotation is not suitable for QPE, because the position relationship of pixels in the precipitation label strictly corresponds to the geographical position relationship in reality. Random rotation will destroy the model's learning of spatial position information, such as the top pixel of a label image being north of the bottom one. This phenomenon is also called inductive bias [44] in DL. Unlike the position, the intensity of rainfall needs to be transformed to highlight the severe precipitation. Before normalization, the label images are rooted three. The labels are in $[0, 60]$ mm. About 4% of samples are below 1 mm, which is in $[0, \frac{1}{60}]$ after normalization. After rooted three, the labels below 1 mm will be transformed to $[0, \frac{1}{\sqrt[3]{60}}]$, which is much smoother for the model to learn the distribution of the precipitation label. Doppler radars are also denoised as described in Section 2.1. The Haar wavelet is appropriate for noise removal as no observable improvements are found when using other wavelets like Daubechies wavelet. Strong region-related and weak non-linear characteristics are also given in Section 2.1 to explain the effectiveness of the Harr wavelet in the QPE task. On the contrary, injecting noise in the input, activation function, or gradient of a DL network can also be seen as a form of data augmentation [45–48]. We have also tried to add Gaussian noise to the input, i.e., Doppler radar features. It turns out that the TS of high precipitation slightly improves (from 27.6% to 28.0%, $\geq$3 mm) while TS of low precipitation degrades a bit (from 62.3% to 62.1%, $\geq$0.1 mm; from 43.1% to 41.9, $\geq$1 mm). In summary, using the common preprocessing method in DL for the QPE task needs to depend on the knowledge of meteorology, attributes of Doppler radar features, and precipitation labels.

6.2. Long-Tailed Distribution in QPE

In real applications, training samples usually exhibit a long-tailed distribution, where a small portion of classes have substantial sample points, but the others are associated with only a few samples [43]. This class imbalance in the number of training samples makes it very challenging to train deep network-based recognition models. Due to the sparsity of samples with high precipitation rates, QPE also suffers from the long-tailed distribution problem. What is more, high precipitation is typically paramount in human lives [15], which enables us to sacrifice low precipitation performance to improve high precipitation performance. Figure 10 shows how the loss function affects the performance of the model on different precipitations. Regardless of the choice of hyperparameters, the RMSE and bias decrease after using RF loss, as shown in Figure 10a,b. It is reasonable for RF loss to focus more on the rarer samples, which are difficult to be recognized, instead of the large amount of 'easy' samples. If relaxing the definition restrictions on hard samples, for example, increasing the factor d, the TS of greater than 3 mm (see Figure 10g) will decrease while RMSE and bias will become better. It is because the real hard samples will be 'diluted' by the not-so-hard samples. Dropping β can get the same result because it means turning down the weight of the hard samples in RF loss. It is intractable for obtaining high performance on the tail without sacrificing the performance on the head. Using pretrained models [49] might be a good choice without introducing more data to the training set, which does not apply to the QPE task since mainstream pretrained models are based on natural images rather than meteorological images. Another solution is to improve the overall performance first and then balance the performance of the head and tail. For example, in our work, attention block and denoising are applied to achieve this goal. As shown in Figure 8, compared with the traditional $Z - R$ relationship-based method, the improvement of our model on high precipitation is particularly obvious (from 17.6% to 39.6%, TS of $\geq$20 mm, an hour), while ensures the overall performance (from 2.94 mm/h to 2.58 mm/h, RMSE).

7. Conclusions

In this paper, RRED-Net, including denoising, DL, and long-tailed learning, addresses the problems in QPE is proposed. Considering the noise in Doppler radar, we utilize wavelet transform to denoise the signal. Compared to the traditional method, our proposed RRED-Net is more capable of processing complex non-linear information in Doppler radar. Feature converter module enables our RRED-Net to capture broader spatial context than other DL methods. Moreover, the proposed method promotes severe precipitation recognition, which is harder and more critical, without sacrificing the low precipitation recognition performance, which makes our work more practical. In the ablation study, the advantages and disadvantages of different components of our method are also discussed, which may provide a good insight into how to exploit DL methods in meteorology tasks. For future work, we plan to extend the current research in the following two directions: (1) design a post-process module to generate more vivid results, (2) introduce auxiliary information like positional information and time information for better precipitation recognition.

Author Contributions: Conceptualization, W.C., W.H. and M.G.; data curation, N.L. and Y.L.; formal analysis, W.C., W.H. and M.G.; investigation, N.L. and Y.L.; methodology, W.C. and W.H.; project administration, A.X. and F.S.; resources, A.X. and F.S.; software, N.L. and Y.L.; supervision, A.X.; validation, W.C. and W.H.; visualization, W.C. and W.H.; writing—original draft preparation, W.C.; writing—review and editing, M.G., A.X. and F.S.; All authors have read and agreed to the published version of the manuscript.

Funding: This research was funded by the Chinese National Natural Science Foundation under grant number 62076033, U1931202.

Data Availability Statement: Not applicable.

Acknowledgments: The authors thank all colleagues who participated in the deriving data.

Conflicts of Interest: The authors declare no conflict of interest.

References

1. Zhang, C.; Wang, H.; Zeng, J.; Ma, L.; Guan, L. Short-term dynamic radar quantitative precipitation estimation based on wavelet transform and support vector machine. *J. Meteorol. Res.* **2020**, *34*, 413–426. [CrossRef]
2. Crosson, W.L.; Duchon, C.E.; Raghavan, R.; Goodman, S.J. Assessment of rainfall estimates using a standard ZR relationship and the probability matching method applied to composite radar data in central Florida. *J. Appl. Meteorol. Climatol.* **1996**, *35*, 1203–1219. [CrossRef]
3. Zhang, J.; Howard, K.; Langston, C.; Vasiloff, S.; Kaney, B.; Arthur, A.; Van Cooten, S.; Kelleher, K.; Kitzmiller, D.; Ding, F.; et al. National Mosaic and Multi-Sensor QPE (NMQ) system: Description, results, and future plans. *Bull. Am. Meteorol. Soc.* **2011**, *92*, 1321–1338. [CrossRef]
4. Peng, X.; Li, Q.; Jing, J. CNGAT: A Graph Neural Network Model for Radar Quantitative Precipitation Estimation. *IEEE Trans. Geosci. Remote Sens.* **2021**, *60*, 1–14. [CrossRef]
5. Sehad, M.; Lazri, M.; Ameur, S. Novel SVM-based technique to improve rainfall estimation over the Mediterranean region (north of Algeria) using the multispectral MSG SEVIRI imagery. *Adv. Space Res.* **2017**, *59*, 1381–1394. . [CrossRef]
6. Wang, Y.; Tang, L.; Chang, P.L.; Tang, Y.S. Separation of convective and stratiform precipitation using polarimetric radar data with a support vector machine method. *Atmos. Meas. Tech.* **2021**, *14*, 185–197. [CrossRef]
7. Kuang, Q.; Yang, X.; Zhang, W.; Zhang, G. Spatiotemporal modeling and implementation for radar-based rainfall estimation. *IEEE Geosci. Remote Sens. Lett.* **2016**, *13*, 1601–1605. [CrossRef]
8. Li, X.; Yang, Y.; Mi, J.; Bi, X.; Zhao, Y.; Huang, Z.; Liu, C.; Zong, L.; Li, W. Leveraging machine learning for quantitative precipitation estimation from Fengyun-4 geostationary observations and ground meteorological measurements. *Atmos. Meas. Tech.* **2021**, *14*, 7007–7023. [CrossRef]
9. Kühnlein, M.; Appelhans, T.; Thies, B.; Nauß, T. Precipitation estimates from MSG SEVIRI daytime, nighttime, and twilight data with random forests. *J. Appl. Meteorol. Climatol.* **2014**, *53*, 2457–2480. [CrossRef]
10. Min, M.; Bai, C.; Guo, J.; Sun, F.; Liu, C.; Wang, F.; Xu, H.; Tang, S.; Li, B.; Di, D.; et al. Estimating summertime precipitation from Himawari-8 and global forecast system based on machine learning. *IEEE Trans. Geosci. Remote Sens.* **2018**, *57*, 2557–2570. [CrossRef]
11. LeCun, Y.; Bengio, Y.; Hinton, G. Deep learning. *Nature* **2015**, *521*, 436–444. [CrossRef] [PubMed]

12. Bengio, Y.; Courville, A.; Vincent, P. Representation Learning: A Review and New Perspectives. *IEEE Trans. Pattern Anal. Mach. Intell.* **2013**, *35*, 1798–1828. [CrossRef] [PubMed]
13. 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*, 521.
14. Sønderby, C.K.; Espeholt, L.; Heek, J.; Dehghani, M.; Oliver, A.; Salimans, T.; Agrawal, S.; Hickey, J.; Kalchbrenner, N. Metnet: A neural weather model for precipitation forecasting. *arXiv* **2020**, arXiv:2003.12140. [CrossRef]
15. Ravuri, S.; Lenc, K.; Willson, M.; Kangin, D.; Lam, R.; Mirowski, P.; Fitzsimons, M.; Athanassiadou, M.; Kashem, S.; Madge, S.; et al. Skilful precipitation nowcasting using deep generative models of radar. *Nature* **2021**, *597*, 672–677. [CrossRef]
16. Wang, Y.; Zhang, J.; Zhu, H.; Long, M.; Wang, J.; Yu, P.S. Memory in memory: A predictive neural network for learning higher-order non-stationarity from spatiotemporal dynamics. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Long Beach, CA, USA, 15–20 June 2019; pp. 9154–9162.
17. 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*, 573.
18. Goodfellow, I.J.; Pouget-Abadie, J.; Mirza, M.; Xu, B.; Warde-Farley, D.; Ozair, S.; Courville, A.; Bengio, Y. Generative Adversarial Networks. *arXiv* **2014**, arXiv:1406.2661. [CrossRef]
19. Leinonen, J.; Nerini, D.; Berne, A. Stochastic Super-Resolution for Downscaling Time-Evolving Atmospheric Fields With a Generative Adversarial Network. *IEEE Trans. Geosci. Remote Sens.* **2021**, *59*, 7211–7223. [CrossRef]
20. Long, J.; Shelhamer, E.; Darrell, T. Fully convolutional networks for semantic segmentation. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Boston, MA, USA, 7–12 June 2015; pp. 3431–3440.
21. 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, Munich, Germany, 5–9 October 2015; Springer: Berlin/Heidelberg, Germany, 2015; pp. 234–241. [CrossRef]
22. Chen, L.C.; Papandreou, G.; Kokkinos, I.; Murphy, K.; Yuille, A.L. Deeplab: Semantic image segmentation with deep convolutional nets, atrous convolution, and fully connected crfs. *IEEE Trans. Pattern Anal. Mach. Intell.* **2017**, *40*, 834–848. [CrossRef]
23. Zhou, Z.; Rahman Siddiquee, M.M.; Tajbakhsh, N.; Liang, J. Unet++: A nested u-net architecture for medical image segmentation. In *Deep Learning in Medical Image Analysis and Multimodal Learning for Clinical Decision Support*; Springer: Berlin/Heidelberg, Germany, 2018; pp. 3–11. [CrossRef]
24. Chen, L.C.; Papandreou, G.; Kokkinos, I.; Murphy, K.; Yuille, A.L. Semantic image segmentation with deep convolutional nets and fully connected crfs. *arXiv* **2014**, arXiv:1412.7062. [CrossRef]
25. Chen, L.C.; Papandreou, G.; Schroff, F.; Adam, H. Rethinking atrous convolution for semantic image segmentation. *arXiv* **2017**, arXiv:1706.05587. [CrossRef]
26. Chen, W.; Zhou, G.; Giannakis, G.B. Velocity and acceleration estimation of Doppler weather radar/lidar signals in colored noise. In Proceedings of the 1995 International Conference on Acoustics, Speech, and Signal Processing, Detroit, MI, USA, 9–12 May 1995; IEEE: Piscataway, NJ, USA, 1995; Volume 3, pp. 2052–2055. [CrossRef]
27. Dixon, M.; Hubbert, J. The separation of noise and signal components in Doppler radar returns. In Proceedings of the Proc. Seventh European Conf. on Radar in Meteorology and Hydrology, Toulouse, France, 24–29 June 2012.
28. Gordon, W.B. An effect of receiver noise on the measurement of Doppler spectral parameters. *Radio Sci.* **1997**, *32*, 1409–1423. [CrossRef]
29. Kent, J.T. Information Gain and a General Measure of Correlation. *Biometrika* **1983**, *70*, 163–173. [CrossRef]
30. McHugh, M. The Chi-square test of independence. *Biochem. Medica* **2013**, *23*, 143–149. [CrossRef] [PubMed]
31. Huang, Z.; Wang, X.; Huang, L.; Huang, C.; Wei, Y.; Liu, W. Ccnet: Criss-cross attention for semantic segmentation. In Proceedings of the IEEE/CVF International Conference on Computer Vision, Seoul, Republic of Korea, 27–28 October 2019; pp. 603–612.
32. Isensee, F.; Jaeger, P.F.; Kohl, S.A.; Petersen, J.; Maier-Hein, K.H. nnU-Net: A self-configuring method for deep learning-based biomedical image segmentation. *Nat. Methods* **2021**, *18*, 203–211. [CrossRef]
33. Gu, Z.; Cheng, J.; Fu, H.; Zhou, K.; Hao, H.; Zhao, Y.; Zhang, T.; Gao, S.; Liu, J. Ce-net: Context encoder network for 2d medical image segmentation. *IEEE Trans. Med. Imaging* **2019**, *38*, 2281–2292. [CrossRef]
34. Wang, X.; Girshick, R.; Gupta, A.; He, K. Non-local neural networks. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Salt Lake City, UT, USA, 17–23 June 2018; pp. 7794–7803.
35. Ho, J.; Kalchbrenner, N.; Weissenborn, D.; Salimans, T. Axial attention in multidimensional transformers. *arXiv* **2019**, arXiv:1912.12180. [CrossRef]
36. Lin, T.Y.; Goyal, P.; Girshick, R.; He, K.; Dollár, P. Focal loss for dense object detection. In Proceedings of the IEEE International Conference on Computer Vision, Venice, Italy, 22–29 October 2017; pp. 2980–2988.
37. Xiong, A.; Liu, N.; Liu, Y.; Zhi, S.; Wu, L.; Xin, Y.; Shi, Y.; Zhan, Y. QpefBD: A Benchmark Dataset Applied to Machine Learning for Minute-Scale Quantitative Precipitation Estimation and Forecasting. *J. Meteorol. Res.* **2022**, *36*, 93–106. [CrossRef]
38. Omuya, E.O.; Okeyo, G.O.; Kimwele, M.W. Feature selection for classification using principal component analysis and information gain. *Expert Syst. Appl.* **2021**, *174*, 114765. [CrossRef]
39. Shang, C.; Li, M.; Feng, S.; Jiang, Q.; Fan, J. Feature selection via maximizing global information gain for text classification. *Knowl.-Based Syst.* **2013**, *54*, 298–309. [CrossRef]
40. Paszke, A.; Gross, S.; Chintala, S.; Chanan, G.; Yang, E.; DeVito, Z.; Lin, Z.; Desmaison, A.; Antiga, L.; Lerer, A. Automatic Differentiation in Pytorch. 2017. Available online: https://openreview.net/forum?id=BJJsrmfCZ (accessed on 14 January 2023).

41. Kingma, D.P.; Ba, J. Adam: A method for stochastic optimization. *arXiv* **2014**, arXiv:1412.6980 [CrossRef] .
42. Radar observation of the atmosphere. L. J. Battan (The University of Chicago Press) 1973. PP X, 324; 125 figures, 21 tables. £7·15. *Q. J. R. Meteorol. Soc.* **1973**, *99*, 793–793. [CrossRef]
43. Zhang, Y.; Kang, B.; Hooi, B.; Yan, S.; Feng, J. Deep long-tailed learning: A survey. *arXiv* **2021**, arXiv:2110.04596. [CrossRef]
44. Battaglia, P.W.; Hamrick, J.B.; Bapst, V.; Sanchez-Gonzalez, A.; Zambaldi, V.; Malinowski, M.; Tacchetti, A.; Raposo, D.; Santoro, A.; Faulkner, R.; et al. Relational inductive biases, deep learning, and graph networks. *arXiv* **2018**, arXiv:1806.01261. [CrossRef]
45. An, G. The Effects of Adding Noise During Backpropagation Training on a Generalization Performance. *Neural Comput.* **1996**, *8*, 643–674. [CrossRef]
46. Vincent, P.; Larochelle, H.; Lajoie, I.; Bengio, Y.; Manzagol, P.A.; Bottou, L. Stacked denoising autoencoders: Learning useful representations in a deep network with a local denoising criterion. *J. Mach. Learn. Res.* **2010**, *11*, 3371–3408.
47. Gulcehre, C.; Moczulski, M.; Denil, M.; Bengio, Y. Noisy activation functions. In Proceedings of the International Conference on Machine Learning, PMLR, York City, NY, USA, 19–24 June 2016; pp. 3059–3068.
48. Neelakantan, A.; Vilnis, L.; Le, Q.V.; Sutskever, I.; Kaiser, L.; Kurach, K.; Martens, J. Adding gradient noise improves learning for very deep networks. *arXiv* **2015**, arXiv:1511.06807. [CrossRef]
49. Tian, C.; Wang, W.; Zhu, X.; Wang, X.; Dai, J.; Qiao, Y. VL-LTR: Learning Class-wise Visual-Linguistic Representation for Long-Tailed Visual Recognition. *arXiv* **2021**, arXiv:2111.13579.

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.