Improved RepVGG Ground-Based Cloud Image Classification with Attention Convolution

Chaojun Shi[1, 2], Leile Han[1], Ke Zhang[1, 2], Hongyin Xiang[1, 2], Xingkuan Li[1], Zibo Su[1], and Xian Zheng[1]

[1]Department of Electronic and Communication Engineering, North China Electric Power University, Baoding 071003, China
[2]Hebei Key Laboratory of Power Internet of Things Technology, North China Electric Power University, Baoding Hebei 071003, China

Correspondence to: Chaojun Shi (scj@ncepu.edu.cn) and Hongyin Xiang (66283880@qq.com)

Abstract. Atmospheric clouds greatly impact the Earth's radiation, hydrological cycle, and climate change. Accurate automatic recognition of cloud shape based on ground-based cloud image is helpful to analyze solar irradiance, water vapor content, and atmospheric motion, and then predict photovoltaic power, weather trends, and severe weather changes. However, the appearance of clouds is changeable and diverse, and its classification is still challenging. In recent years, convolution neural network (CNN) has made great achievements in ground-based cloud image classification. However, traditional CNNs poorly associate long-distance clouds, making the extraction of global features of cloud images quite problematic. This study attempts to mitigate this problem by elaborating a ground-based cloud image classification method based on the improved RepVGG convolution neural network and attention mechanism. Firstly, the proposed method increases the RepVGG residual branch and obtains more local detail features of cloud images through small convolution kernels. Secondly, an improved channel attention module is embedded after the residual branch fusion, effectively extracting the global features of cloud images. Finally, the linear classifier is used to classify the ground cloud images. Finally, the warm-up method is applied to optimize the learning rate in the training stage of the proposed method, making it lightweight in the inference stage and thus avoiding overfitting and accelerating the model's convergence. The proposed method is validated on MGCD and GRSCD ground-based cloud image datasets containing 7 cloud categories, with the respective classification accuracy rate values of 98.15% and 98.07%, outperforming those of ten most advanced methods used as the reference. The results obtained are considered instrumental in ground-based cloud image classification.

1. Introduction

In meteorology, cloud is an aerosol consisting of a visible mass of water droplets, ice crystals, their aggregates or other particles suspended in the atmosphere. Clouds of different types cover over 70% of the Earth surface (Qu et al., 2021; Gyasi and Swarnalatha, 2023; Fabel et al., 2022). Cloud analysis plays a crucial role in meteorological observation because clouds can affect the Earth's water cycle, climate change, and solar irradiance (Gorodetskaya et al., 2015; Goren et al., 2018; Zheng et al., 2019). Cloud observation methods mainly include satellite observation (Norris et al., 2016; Zhong et al., 2017; Li et al., 2023) and ground observation (Calbó and Sabburg, 2008; Nouri et al., 2019; Lin et al., 2023). Satellite observation refers to the distribution, movement, and change of clouds observed by high-resolution remote sensing satellites from above. When observing local sky regions, satellite observations have low performance and are unable to obtain sufficient resolution to describe the characteristics of different cloud layers in detail (Long et al., 2023; Sarukkai et al., 2020). Compared with satellite observation, ground-based observation opens up a new way to monitor and understand regional sky conditions. Typical ground-based cloud observation instruments include All-Sky Imager (ASI) (Shi et al., 2019; Cazorla et al., 2008), Total Sky Imager (TSI) (Long et al., 2006; Tang et al., 2021), etc. The relevant equipment and ground-based cloud images are shown in Figure 1.

(a) (b)

Figure 1: Two kinds of ground-based cloud images and their observation equipment: (a) ASI ground-based cloud image and its observation equipment (Cazorla et al., 2008; Shi et al., 2019); (b) TSI ground-based cloud image and its observation equipment (Long et al., 2006).

Ground-based cloud observation can obtain more obvious cloud characteristics by observing the information at the bottom of the cloud, which is conducive to assisting the prediction of local photovoltaic power generation. Clouds play an important role in maintaining the atmospheric radiation balance by absrbing short-wave and the ground not to solar radiation (Taravat et al., 2015). Pv power prediction is affected by multiple factors such as cloud genus, cloud cover change, solar irradiance, and solar cell performance in local areas, among which cloud genus is an important factor affecting PV power prediction (Zhu et al., 2022). Therefore, it is of great significance to accurately obtain sky cloud information through cloud observation and then accurately classify clouds for accurate prediction of photovoltaic power generation (Alonso-Montesinos et al., 2016). The traditional ground-based cloud observation method is mainly visual observation, which relies heavily on the experience of observers, cannot achieve standardization. Therefore, ground-based cloud automatic observation has been widely concerned by scholars. In recent years, with the development of digital image acquisition devices, many ground-based whole-sky cloud image acquisition devices have emerged the world, providing massive data support for automatic ground-based cloud observation (Pfister et al., 2003).

Ground-based cloud image classification is an important part of the foundation of automatic cloud observation and is the key to climate change and photovoltaic power prediction. The classification of ground-based cloud images mainly classifies each cloud image taken from the ground into the corresponding cloud genus by extracting cloud image features, such as cirrus, cumulus, stratus, nimbostratus, etc. According to different cloud image feature extraction methods, the ground-based cloud image classification method is divided into based on traditional machine learning method and based on deep learning method (Simonyan and Zisserman, 2015; Krizhevsky et al., 2017; Hu et al., 2018). Most of the ground-based cloud image classification methods based on traditional machine learning classify cloud images by artificially designing cloud image features, while the ground-based cloud image classification methods based on deep learning mainly classify cloud images through self-learning cloud image features of deep neural network (DNN) (Wu et al., 2019).

Early ground-based cloud image classification studies relied on manual classification methods, which focused on features such as texture, structure, and color, combined with traditional machine learning methods to classify ground-based cloud images. These methods include a decision tree, K-nearest neighbor (KNN) classifier, support vector machine (SVM), etc. (Singh and Glennen, 2005) proposed a method for automatically training the texture function of a cloud classifier. In this method, five feature extraction methods including autocorrelation, co-occurrence matrix, edge frequency, Laws texture analysis, and original length are used respectively. Compared with other cloud classification methods, this method has the advantages of high accuracy and fast classification speed, but its classification ability for mixed clouds is insufficient. (Heinle et al., 2010) described cloud images by using spectral features (mean value, standard deviation, skewness, and difference) and texture features (energy, entropy, contrast, homogeneity, and cloud cover), and combined with a KNN classifier, divided ground cloud images into seven categories. In addition, (Zhuo et al., 2014) reported that the spatial distribution of contour lines could represent the structural information of cloud shapes, used the central description pyramid to simultaneously extract the texture and structural features of ground-based cloud images, and used SVM and KNN to classify cloud images. It can be seen that the traditional classification method of ground-based cloud images based on machine learning mainly uses hand-designed texture, structure, color, shape, and other features to extract, and obtains high-dimensional feature expression of ground-based cloud images through single feature or fusion feature. Traditional machine learning methods mostly describe the features from the perspective of digital signal analysis and mathematical statistics, but ignore the representation and interpretation of the visual features of the cloud image itself.

In recent years, under the background of cross-integration of different disciplines and artificial intelligence, the ground-based cloud image classification method based on deep learning has become a research hotspot with its superior classification performance. Aiming at the unique characteristics of ground-based cloud images, (Shi et al., 2017) proposed Deep Convolutional Activations-Based Features (DCAFs) to classify ground-based cloud images, and the results are better than the artificially designed cloud image features. Alternatively, (Ye et al., 2017) used CNN to extract cloud image features and proposed a local pattern mining method based on ground-based cloud images to optimize the local features of cloud images and improve the classification accuracy of cloud images. (Zhang et al., 2018a) put the wake cloud as a new genus of cloud into the ground-based cloud image database for the first time, proposed a simple convolutional neural network model called CloudNet, and applied it to the ground-based cloud image classification task, effectively improving the accuracy of ground-based cloud image classification. More recently, (Wang et al., 2020) proposed the CloudA network, an optimized iteration of the AlexNet convolutional neural network, which reduces the number of parameters through a simplified network architecture. The classification accuracy on the Singapore Whole-Sky Imaging Categories (SWIMCAT) ground-based cloud image dataset exceeded the traditional ground-based cloud image classification methods. (Liu et al., 2020b) proposed Multi-Evidence and Multi-Modal Fusion Networks (MMFN) by fusing heterogeneous features, local visual features, and multi-mode information, which significantly improves the classification accuracy of cloud images. Aiming at the problem that the traditional neural network has insufficient ability to classify the ground-based cloud images within and between genera, (Zhu et al., 2022) proposed to use of an improved combined convolutional neural network to classify the cloud images, and the classification accuracy is greatly improved compared with the traditional neural network. Alternatively, (Yu et al., 2021) used two sub-convolutional neural networks to extract features of ground-based cloud images and used weighted sparse representation coding to classify them, which solved the problem of occlusion in multi-mode ground-based cloud image data and greatly improved the robustness of cloud images classification. (Liu et al., 2020a) introduced a ground-based cloud image classification method based on a graph convolution network (GCN). However, the weight assigned by GCN failed to accurately reflect the importance of connection nodes, thus reducing the discrimination of aggregated cloud image features. To make up for this deficiency,(Liu et al., 2022) proposed a context attention network for ground-based cloud classification and publicly released a new cloud classification dataset. In addition, (Liu et al., 2020c) further combined CNN and GCN to propose a multimodal ground-based cloud image classification method based on heterogeneous deep feature learning. Alternatively, (Wang et al., 2021) elaborated a ground-based cloud image classification method based on Transfer Convolutional Neural Network (TCNN) by combining deep learning and transfer learning. (Li et al., 2022) further enhanced the classification performance of ground-based cloud images based on the improved Vision Transformer combined with the EfficientNet-CNN. The performance of the above-mentioned ground-based cloud image classification methods based on deep learning has significantly improved compared to traditional machine learning methods.

CNN plays an important role in the field of target detection, image classification, and image segmentation, especially in the tasks of power line fault detection (Zhao et al., 2016), face recognition (Meng et al., 2021), and medical image segmentation (Zhang et al., 2021), and has been widely used and achieved great achievements. Ground-based cloud image classification is an emerging task in the field of image classification and has achieved rapid and considerable development based on the CNN method. However, it still has some shortcomings such as shallow network level of ground-based cloud image classification method, limited ground-based cloud image classification performance, and small ground-based cloud image classification dataset, which cannot verify the generalization ability of large-scale ground-based cloud image classification dataset.

To solve the above problems, the current study improved the RepVGG (Ding et al., 2021) and used it as a basis for elaborating a new classification method for ground-based cloud images called CloudRVE (Cloud Representative Volume Element Network). In this method, the ground-based cloud image was incorporated into the CNN model, and its image features were extracted. Multi-branch convolution layer and channel attention module were used to capture local and global features of the cloud image simultaneously time to enhance the classification performance of ground-based cloud images. The method's application to the multi-modal ground-based cloud dataset named MGCD (Liu et al., 2020a) and ground-based remote sensing cloud database (GRSCD) (Liu et al., 2020b) . The main contributions of this paper are as follows:

(1) This study elaborated the Improved RepVGG ground-based cloud image classification method with attention convolution called CloudRVE. It broadened the residual structure and comprehensively combined the attention mechanism's abilities to extract the cloud image's global features and describe in detail its local features in the classification process.

(2) In particular, the Efficient Channel Attention network (ECA) was improved and incorporated into the feature extraction process of ground-based cloud images, which optimization occurred through local cross-channel interaction without dimensionality reduction. Besides, the structural re-parameterization in the inference stage was performed, reducing the model complexity, improving the feature extraction performance, and enhancing the network's learning ability of ground-based cloud image features.

(3) The comparative analysis of experimental results on the ground-based cloud image classification dataset MGCD proved that the proposed method outperformed ten other state-of-the-art methods in classification accuracy. Its application to GRSCD dataset further verified its generalization ability. Finally, the proposed method's training process optimization and dynamical adjustment of its learning rate were provided by the warm-up method, and the respective recommendations were drawn.

The rest of this paper is organized as follows. Section 2 elaborates on the structure and composition of the proposed CloudRVE method for classifying ground cloud images. Section 3 briefly introduces the ground cloud image classification datasets used in this paper and the model evaluation indices. Section 4 provides the experimental results and discusses the feasibility and effectiveness of the proposed method. Finally, Section 5 concludes the study and outlines future research directions and practical application of the research results.

2. Methods

2.1 Overview of Method

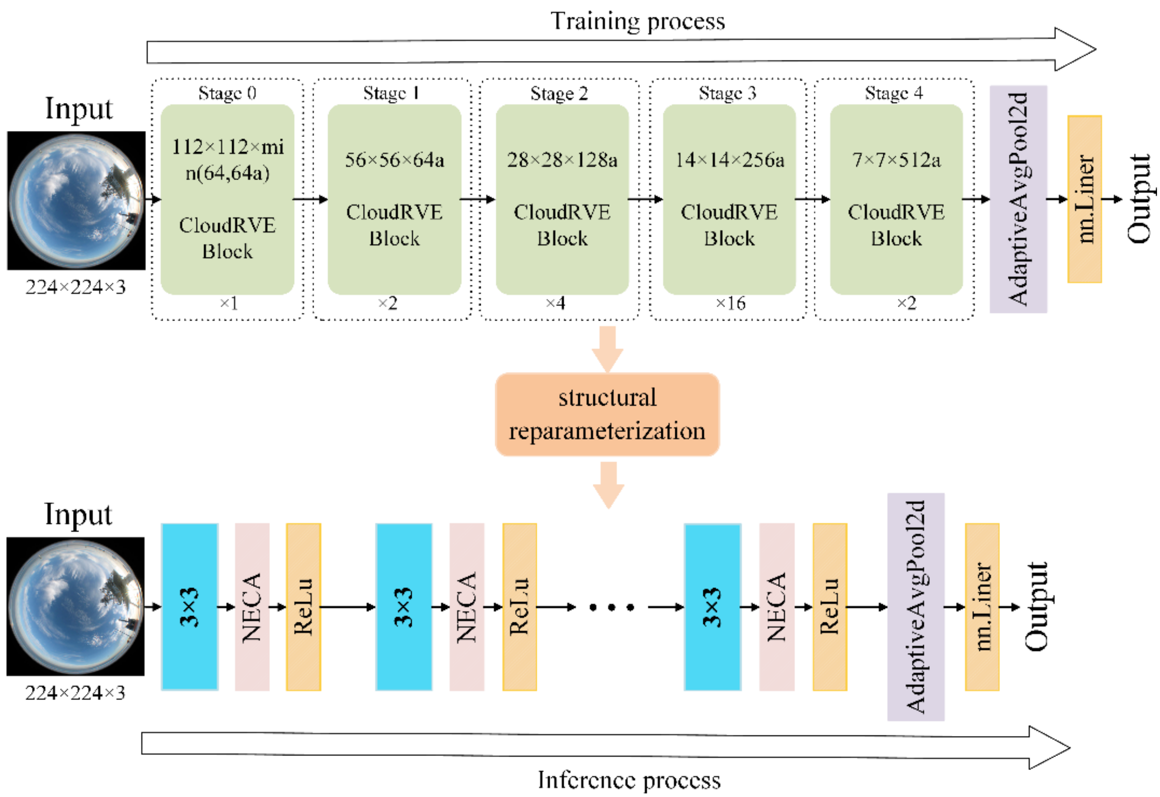


Figure 2: CloudRVE network framework. Ground-based cloud images come from Kiel-F datasets (Kalisch and Macke, 2008).

This section shows the overall architecture of the proposed RepVGG-based improved classification method, as shown in Figure 2. In the CloudRVE training process, CloudRVE Block with a multi-branch topology structure is used to extract features of ground-based cloud images. The multi-branch topology structure has rich gradient information and a complex network structure, which can effectively improve the characterization ability of local feature information of ground-based cloud images. Feature maps extracted by CloudRVE Block enter the New Efficient Channel Attention (NECA) network and learn the feature relationships between sequences to obtain the global feature representation of an image. In addition, the warm-up method is incorporated into the CloudRVE training process to dynamically optimize the learning rate and accelerate the model parameter convergence to enhance the model training effect. CloudRVE inference process uses the single branch topology structure of VGG-style (Simonyan and Zisserman, 2015), and through structural re-parameterization, the multi-branch convolutional layer and batch normalization (BN) (Ioffe and Szegedy, 2015) are converted into a 3×3 convolutional layer, increasing its inference speed. The CloudRVE training process and inference process use the linear classifier to classify the ground-based cloud images to get the final result. The specific framework parameter information of the model is shown in Table 1, where *a* and *b* are magnification factors used to control the network width. The specific contents of each part are as follows.

Table 1. The details of CloudRVE training architecture.

Stage	Blocks of each stage	Output size	Output channels
0	1	224×224	Min (64, 64a)
1	2	112×112	64a
2	4	56×56	128a
3	14	28×28	256a
4	1	14×14	512b

2.2 Broadening the CloudRVE Block of Residual Structure

CNN is a deep learning model including convolution calculation, including feedforward neural network, which has representation learning ability, similar to artificial neural network multilayer perceptron (Shi et al., 2017). In 2014, the most representative convolution neural network VGG came out, which adopted a single-branch topology structure, greatly improved the image processing effect and model inference speed, and became a new direction for scholars to learn and develop. With the in-depth study of the VGG, its potential in image processing is close to saturation. Scholars realize that the VGG has some shortcomings such as simple network structure, few network layers, and large parameters, which makes it difficult to extract high-order features of images and has limited image-processing performance. Therefore, improving network complexity and increasing the number of network layers has become a new research direction. The ResNet developed by (He et al., 2016) differed from the traditional neural network stacked by convolution layer and pooling layer. The network was stacked by residual modules, which not only increased the complexity of the network structure and reduced the number of network parameters, but also perfectly solved the problem of gradient disappearance or gradient explosion caused by increasing the number of network layers, which could extract abstract image features with semantic information and effectively improve image-processing performance. By improving the complexity and depth of the network, the ResNet could train the CNN model with higher accuracy, but there were numerous redundancies in its residual network, impeding the network inference speed and reducing the accuracy of image processing results (Szegedy et al., 2015). Therefore, increasing the complexity and depth of the network, weakening its influence on inference speed, and improving the classification effect of ground-based cloud images become the key goals of this study.

To improve the classification effect of the ground-based cloud images, the CloudRVE training process is composed of CloudRVE blocks that adopt the multi-branch topology. The CloudRVE Block contains four branches and the improved channel attention module NECA. Its main branch contains a convolutional layer with a convolution kernel size of 3×3, which can inspect the input images with a larger neighborhood scope and extract global features easily.

Ground-based cloud images contain abundant cloud shape and cloud amount information, while a large convolution kernel tends to ignore cloud boundary features, resulting in inadequate feature extraction from ground-based cloud images. Therefore, the two bypass branches of CloudRVE Block adopt the convolution layer with the convolution kernel size of 1×1, which can not only extract fine cloud boundary features and abstract cloud cover features but also keep the output dimension consistent with the input dimension, facilitating the multi-branch ground-based cloud image feature fusion. The third bypass branch of CloudRVE Block adopts the Identity branch, whose purpose is to take the input as the output and change the learning objective to the residual result approaching 0 so that the accuracy does not decline with the deepening of the network. In addition, each branch is connected to the BN layer, not only to avoid overfitting but also to prevent gradient disappearance or explosion. The specific structure of CloudRVE Block is shown in Figure 3. The input feature maps pass through three branches with a convolutional layer and BN layer at the same time. The output obtained by the input feature maps is summed with the Identity branch and input into the NECA module to obtain the final output feature.

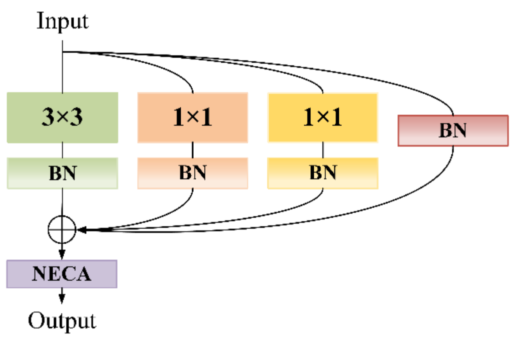


Figure 3: CloudRVE Block structure.

2.3 NECA Module Focusing on Full Image Features

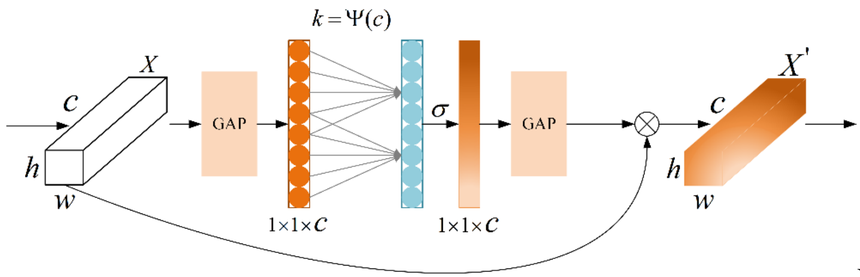
The attention mechanism is to let the neural network have the information processing way to distinguish the key points and to capture the connection between global information and local information flexibly. Its purpose is to enable the model to obtain the target region that needs to be focused on, put more weight on this part, highlight significant useful features, and suppress and ignore irrelevant features. The NECA (New Efficient Channel Attention) is an implementation form of channel attention mechanism, which can strengthen channel features without changing the size of the input feature maps. It adopts a local cross-channel interaction strategy without dimensionality reduction so that the 1×1 convolution layer can replace the full connection layer to learn channel attention information, which can effectively avoid the negative impact of dimensionality reduction on channel attention learning. The network performance is guaranteed and the complexity of the model is significantly reduced.

The ground-based cloud image samples in Figure 2 were taken by the all-sky imager and could cover the sky in this area. However, the ground-based cloud images contain not only the valid area of the whole sky but also the black invalid area. Therefore, the NECA module abandons the traditional global maximum pooling and adopts double global average pooling. The global average pooling formulas are as follows:

$$\gamma_{gap} = \frac{1}{wh}\sum_{i=1,j=1}^{w,h} X_{ij\cdot} \ , X \in R^{w\times h\times c} \ , \tag{1}$$

$$\eta_{gap} = \sigma(V_k^{gap}\gamma_{gap}) \ , V_k^{gap} \in R^{c\times c} \ , \tag{2}$$

where X and X' represent the input and output feature maps, respectively, whereas w, h, and c are the width, height, and number of channels of the input feature map. The NECA module adopts a double global average pool, which can effectively improve its noise suppression ability and enhance its channel feature extraction ability, which can avoid the black invalid part of the feature calculation. The NECA module structure is shown in Figure 4.



N

Figure 4: NECA model structure.

Here b and r are fixed values, and their values are set to 1 and 2, respectively, while k represents the convolution kernel size and has a corresponding relationship with c. As the network deepens, the number of channels c increases by the power of 2. Therefore, k should not be a fixed value, but a dynamic change and its relationship are as follows:

$$C = \phi(k) = 2^{(\gamma * k - b)} \tag{3}$$

$$K = \psi(C) = \left|\frac{log_2(c)}{r} - \frac{b}{r}\right|_{odd} \tag{4}$$

2.4 Inference Process from Multi-Branch to Single-Branch

The residual module is crucial to the CloudRVE training process. Its multi-branch topology can improve CloudRVE Block’s ability to extract ground cloud image features and solve optimization problems such as gradient disappearance and gradient explosion caused by increasing network depth. However, the multi-branch topology will occupy more memory for the CloudRVE reasoning process, resulting in insufficient utilization of hardware computing power and slower reasoning speed. If the single-branch topology is adopted, the computing load is reduced and the inference time is saved, thus reducing memory consumption. Therefore, the single-branch topology structure is adopted in the CloudRVE inference stage, and the trained CloudRVE Block needs to be transformed into a single-branch topology model through structural re-parameterization. The conversion process mainly includes the fusion of the convolutional layer and BN layer, the conversion of the BN layer into a convolutional layer, and the fusion of the multi-branch convolutional layer. We use $W_{(3)} \in R^{C_1 \times C_2 \times 3 \times 3}$ as 3×3 convolution layers, and use C_1, C_2 as input channels and output channels respectively, and use $W_{(1)} \in R^{C_1 \times C_2 \times 1 \times 1}$ as 1×1 convolution layers. In addition, we use $\mu_{(3)}$, $\sigma_{(3)}$, $\gamma_{(3)}$, $\beta_{(3)}$ to represent the mean value, standard deviation, learning scaling factor, and deviation of the BN layer of the main branch, and use $\mu_{(1)}$, $\sigma_{(1)}$, $\gamma_{(1)}$, $\beta_{(1)}$ to represent the parameters of the BN layer of the by-pass branch containing 1×1 convolution layer, and use $\mu_{(0)}$, $\sigma_{(0)}$, $\gamma_{(0)}$, $\beta_{(0)}$ to represent the parameters of the BN layer of the identity branch, and use $M_{(1)} \in R^{N \times C_1 \times H_1 \times W_1}$, $M_{(2)} \in R^{N \times C_2 \times H_2 \times W_2}$ to represent the input and output. The CloudRVE Block structure reparameterization calculation process is as follows:

$$\begin{aligned} M_{(2)} = {} & BN(M_{(1)} * W_{(3)}, \mu_{(3)}, \sigma_{(3)}, \gamma_{(3)}, \beta_{(3)}) + BN(M_{(1)} * W_{(1)}, \mu_{(1)}, \sigma_{(1)}, \gamma_{(1)}, \beta_{(1)}) \\ & + BN(M_{(1)} * W_{(1)}, \mu_{(1)}, \sigma_{(1)}, \gamma_{(1)}, \beta_{(1)}) + BN(M_{(1)}, \mu_{(0)}, \sigma_{(0)}, \gamma_{(0)}, \beta_{(0)}) \end{aligned} \tag{5}$$

The input feature map is inputted into the NECA module through the 3×3 convolution layer completed by fusion. The process is shown in Figure 5.

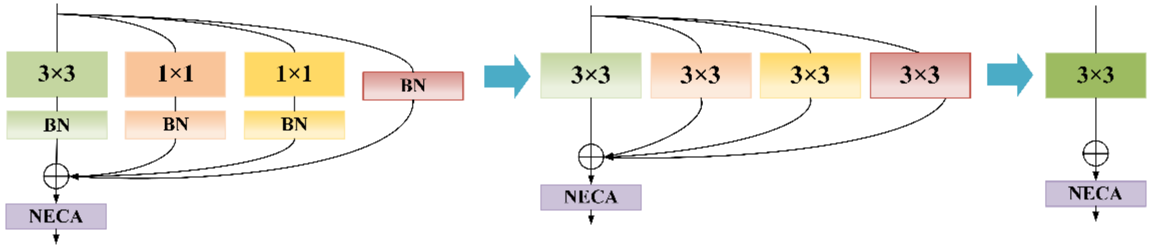


Figure 5: Re-parameterization process of CloudRVE Block structure.

2.4.1 Fusion of Convolutional Layer and BN Layer

This section first describes the fusion of the main branch 3×3 convolution layer with the BN layer and then describes the transformation of the bypass branch 1×1 convolution layer into the 3×3 convolution layer and fusion with the BN layer. In the inference stage, the number of convolutional kernel channels in the convolution layer is the same as the number of channels in the input feature map, and the number of convolutional kernel channels in the output feature map is the same. The main parameters of the BN layer include mean μ, variance σ^2, learning ratio factor γ, and deviation β. Of these, μ and σ^2 are obtained statistically in the training stage, while γ and β are obtained by learning in the training stage. The calculation of the i channel of the input BN layer is performed as follows:

$$y_i = \frac{x_i - u_i}{\sqrt{{\sigma_i}^2 + \varepsilon}} \times \gamma_i + \beta_i \, , \tag{6}$$

where x is the input and ε is the constant approaching 0. The calculation process of the i channel input BN in the feature map can be expressed as follows:

$$bn(M,\mu,\sigma,\gamma,\beta)_{:,i,:,:} = \left(M_{:,i,:,:} - \mu_i\right)\frac{\gamma_i}{\sigma_i} + \beta_i = \frac{\gamma_i}{\sigma_i}M_{:,i,:,:} + \beta_i - \frac{\gamma_i}{\sigma_i}\mu_i \,, \tag{7}$$

where M is the output feature map obtained by weighted summation of the convolution layer, input to BN layer and ignore x. Therefore, we can multiply γ_i/σ_i to the i convolution kernel of the 3×3 convolution layer:

$$W'_{i,:,:,:} = \frac{\gamma_i}{\sigma_i}W_{i,:,:,:} \tag{8}$$

$$\boldsymbol{b}'_i = \boldsymbol{\beta}_i - \frac{\boldsymbol{\mu}_i\boldsymbol{\gamma}_i}{\boldsymbol{\sigma}_i} \tag{9}$$

The i convolution kernel weight of the fusion of the 3 × 3 convolution layer and BN layer is obtained, and the specific fusion process is shown in Figures 6 and 7. The input channel C_1 and output channel C_2 make two, and the stride is one. In the convolution layer, the input feature map is calculated by convolution to obtain the output feature map with the number of channels 2. Figure 8 shows that the number of channels in the BN layer is 2, and the output feature map of the convolution layer is used as the input feature map of the BN layer. The output feature map with the number of channels being 2 is obtained via equation (2).

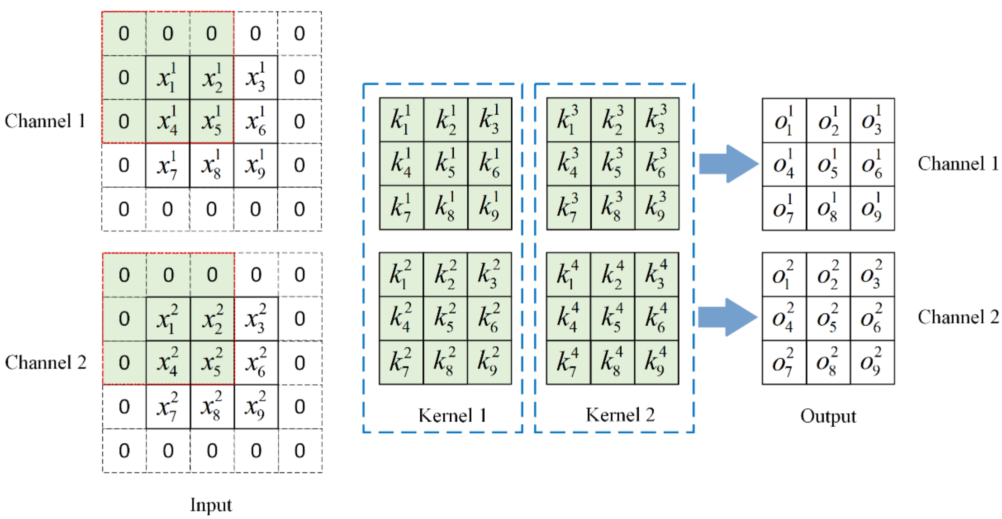


Figure 6: Input feature map through convolution layer process. For visualization, we assume that C_1=C_2=2.

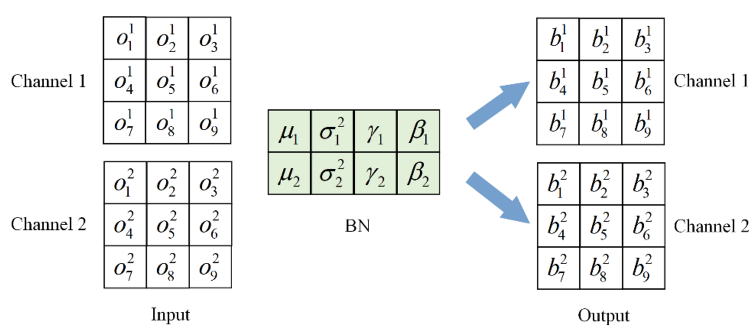


Figure 7: Convolutional layer output feature map through the BN layer process.

In addition, to ensure that the size of the output feature map is consistent with that of the input feature map, the input feature map should be converted into 5 × 5 size by padding operation. The concrete convolution is as follows:

$$o_1^1 = x_1^1 \cdot k_5^1 + x_2^1 \cdot k_6^1 + x_4^1 \cdot k_8^1 + x_5^1 \cdot k_9^1 + x_1^2 \cdot k_5^2 + x_2^2 \cdot k_6^2 + x_4^2 \cdot k_8^2 + x_5^2 \cdot k_9^2 \tag{10}$$

The specific calculation process of the input feature map through the BN layer is

$$b_1 = \frac{(x_1^1 \cdot k_5^1 + x_2^1 \cdot k_6^1 + x_4^1 \cdot k_8^1 + x_5^1 \cdot k_9^1 + x_1^2 \cdot k_5^2 + x_2^2 \cdot k_6^2 + x_4^2 \cdot k_8^2 + x_5^2 \cdot k_9^2) - \mu_1}{\sqrt{\sigma^2 + \varepsilon}} \cdot \gamma_1 + \beta_1 \tag{11}$$

Re-arranging equation (7) yields

$$b_1 = (x_1^1 \cdot k_5^1 + x_2^1 \cdot k_6^1 + x_4^1 \cdot k_8^1 + x_5^1 \cdot k_9^1 + x_1^2 \cdot k_5^2 + x_2^2 \cdot k_6^2 + x_4^2 \cdot k_8^2 + x_5^2 \cdot k_9^2) \cdot \frac{\gamma_1}{\sqrt{\sigma^2 + \varepsilon}} + (\beta_1 - \frac{\mu_1}{\sqrt{\sigma^2 + \varepsilon}}) \tag{12}$$

$$c = \frac{\gamma_1}{\sqrt{\sigma^2 + \varepsilon}} \ ; \qquad d = \beta_1 - \frac{\gamma_1 \cdot \mu_1}{\sqrt{\sigma^2 + \varepsilon}} \tag{13}$$

In equation (8), c and d are constants and are multiplied to the first convolution kernel of the convolution layer to obtain the parameters of the first convolution kernel after the convolution layer and BN layer are fused. Other fused convolution kernel parameters are calculated similarly. The convolution layer and BN layer are fused by the bypass branch containing a 1 × 1 convolution layer. The convolution layer is first converted to 3 × 3 size by padding operation and then fused with the BN layer by repeating the above steps. The convolution layer padding process is shown in Figure 8.

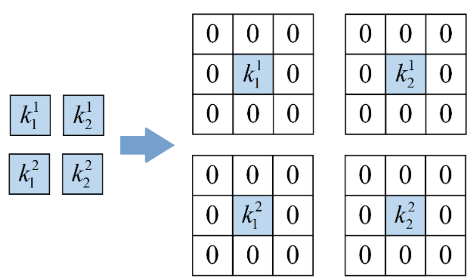


Figure 8: 1 × 1 convolution layer transformed into 3 × 3 convolution layer.

2.4.2 Converting the BN Layer to the Convolution Layer

The identity bypass branch has only a BN layer, its function is to ensure the identity mapping of the input feature map and output feature map. To realize the identical mapping between the input feature map and the output feature map in the fusion process, a 3 × 3 convolution layer with 2 convolution kernels and 2 convolution kernel channels needs to be designed. Secondly, the input feature map needs to be converted into a 5 × 5 feature map by padding operation. The specific process is shown in Figure 9. The output feature map is obtained by convolution calculation of the input feature map, and its parameters and sizes are consistent with those of the input feature map. Finally, the fusion process of the 3 × 3 convolution layer and BN layer is repeated to obtain a new 3 × 3 convolution layer.

2.4.3 Multi-Branch Convolution Layer Fusion

The structure re-parameterization transforms each branch into a 3 × 3 convolution layer by construction and fusion, which facilitates the fusion of multi-branch convolution layers into a single-branch 3 × 3 convolution. We use I and O to represent the input and output, respectively, while K_i and B_i are the convolution kernel weight and bias of the i branch. The multi-branch fusion calculation process is as follows:

$$O = (I \otimes K_1 + B_1) + (I \otimes K_2 + B_2) + (I \otimes K_3 + B_3) + I \otimes (K_1 + K_2 + K_3) + (B_1 + B_2 + B_3) \quad (14)$$

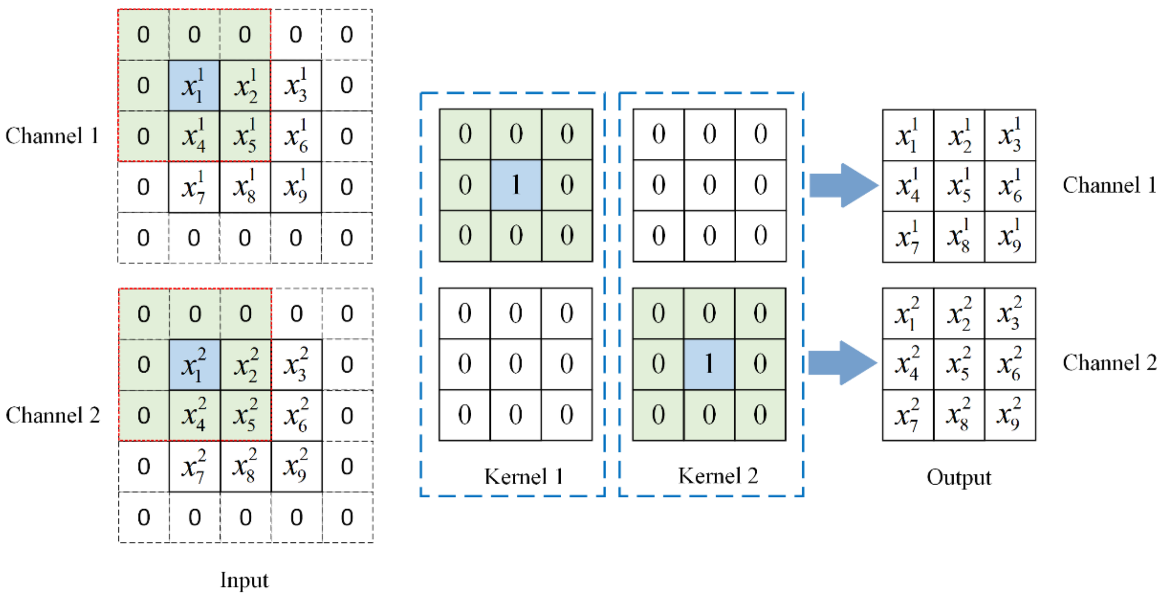


Figure 9: Identity branch Identity mapping process.

2.5 Warm-Up Method

In this paper, the warm-up method (He et al., 2019) is used to optimize the learning rate in the model training process, so that the learning rate varies in different training stages. In the initial stage of model training, a small learning rate is selected, which is due to the random initialization of model weights and no prior knowledge of ground-based cloud image data, and the model will quickly adjust parameters according to the input. If a large learning rate is adopted at this time, the model will be overfitted and the prediction accuracy of the model will be affected. After training the model for some time, the learning rate linearly increases to a preset large value and the model has some prior knowledge, which can avoid overfitting and accelerate the convergence speed of the model. Finally, the model distribution is relatively stable, so it is difficult to learn new features from ground-based cloud image data, and the learning rate linearly approaches to zero, keeping the model stable and easily obtaining local optima.

3. Dataset and Experimental Settings

This section introduces two kinds of ground-based cloud image classification datasets, MGCD and GRSCD, and describes the relevant experimental Settings. Subsection 3.1 describes MGCD and GRSCD in detail, and Subsection 3.2 details experimental setting parameters and model evaluation indices.

3.1 Ground-Based Cloud Image Dataset

3.1.1 Introduction to MGCD Dataset

Multi-modal Ground-based Cloud image Dataset (MGCD) is the first ground-based cloud image classification dataset composed of ground-based cloud images and multi-modal information, which was collected by the School of Electronics and Communication Engineering of Tianjin Normal University and the Meteorological Observation Center of Beijing Meteorological Bureau of China from 2017 to 2018. There are 8000 ground-based cloud images in MGCD, and 4000 ground-based cloud images in the training set and testing set, including altocumulus (Ac), cirrus (Ci), clear sky (Cl), cumulonimbus (Cb), cumulus (Cu), stratocumulus (Sc), and mix (Mx). In addition, cloud images with a cloud cover of less than 10% are classified as clear sky, and each sample contains a captured ground cloud iamge and a set of multimodal cloud information. Among them, the ground-based cloud images are collected by an all-sky camera with a fisheye lens, and its data storage format is JPEG with a resolution of 1024 × 1024 pixels; Multimodal information is collected by weather stations, including temperature, humidity, pressure, and wind speed, and these four elements are stored in the same vector. Figure 10 is a partial sample of the MGCD dataset, and the specific information is shown in Table 2.

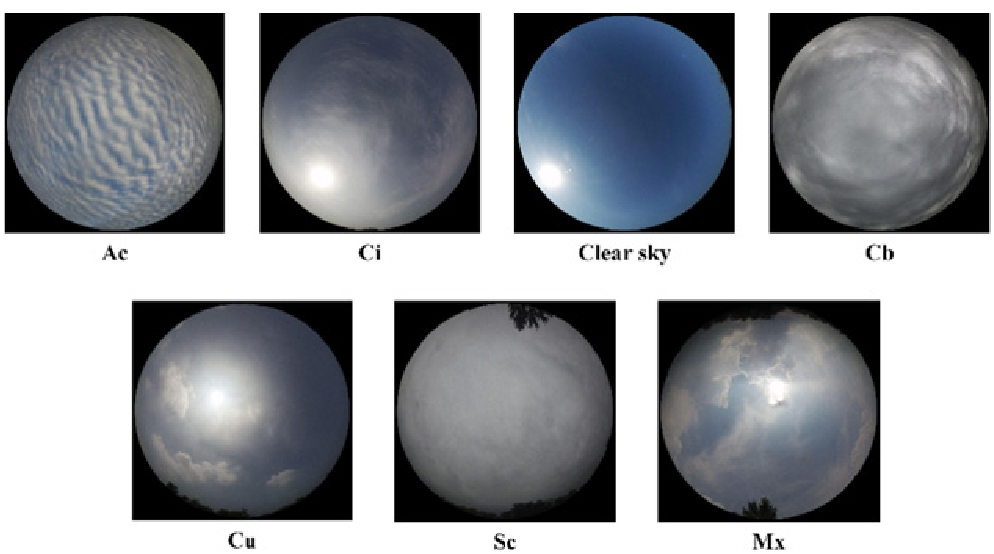


Figure 10: Sample legend of MGCD dataset (Liu et al., 2020a).

Table 2. MGCD dataset-specific information.

No	Class	Training	Testing	Total
1	Ac	365	366	731
2	Ci	662	661	1323
3	Cl	669	669	1338
4	Cb	593	594	1187
5	Cu	719	719	1438
6	Sc	482	481	963
7	Mx	510	510	1020
	Total	4000	4000	8000

3.1.2 Introduction to GRSCD Dataset

Ground remote sensing cloud dataset (GRSCD) is a ground-based cloud image classification dataset composed of ground-based cloud images and multimodal information. It was collected by the College of Electronic and Communication Engineering of Tianjin Normal University and the Meteorological Observation Center of Beijing Meteorological Administration of China from 2017 to 2018. The total number of ground-based cloud images in GRSCD is consistent with MGCD, with a training set and a testing set each accounting for 50%, including 7 types of clouds: altostratus (Ac), cirrus(Ci), clear sky(Cl), cumulonimbus(Cb), cumulus(Cu), stratocumulus(Sc), and mix(Mx). Among them, the features of cumulonimbus and stratocumulus in MGCD are not distinct and easy to confuse; the features of altostratus and cumulus in GRSCD are not distinct and easy to confuse. In addition, each sample contains a ground-based cloud image and a set of multi-modal cloud information, and cloud images with cloud cover not exceeding 10% are classified as clear sky. Figure 11 depicts a partial sample of the GRSCD dataset. The specific data are listed in Table 3.

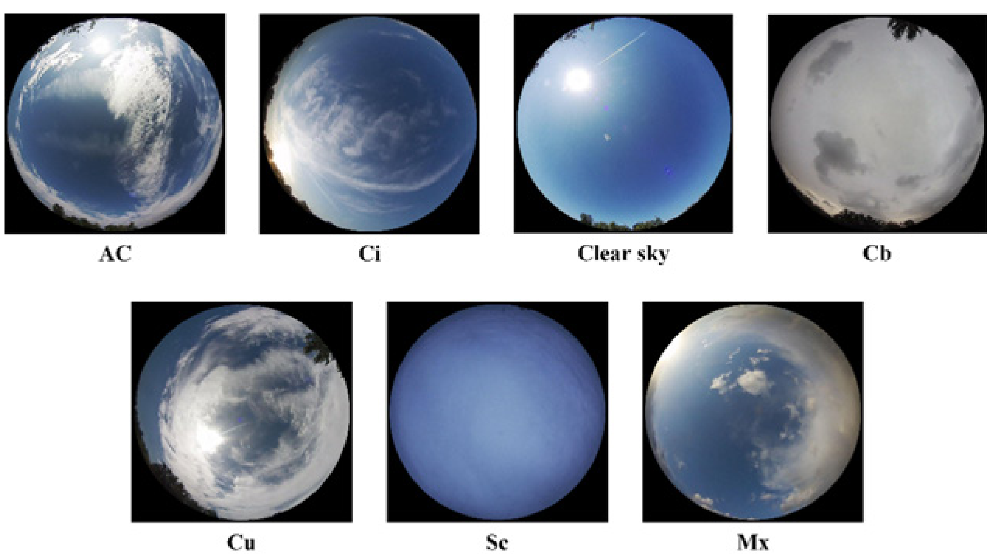


Figure 11: Sample legend of GRSCD dataset (Liu et al., 2020b).

Table 3. GRSCD dataset-specific information

No	Class	Training	Testing	Total
1	Ac	400	331	731
2	Ci	650	673	1323
3	Cl	650	688	1338
4	Cb	600	587	1187
5	Cu	690	748	1438
6	Sc	500	463	963
7	Mx	510	510	1020
	Total	4000	4000	8000

3.2 Experimental Setting

3.2.1 Implementation Details

All experiments in this paper adopt Python programming language and run on Intel(R) Core (TM) i9-12700K CPU @ 3.60GHz. NVIDIA GeForce RTX 3090 24G Graphical Processing Unit (GPU) platform and uses Pytorch as a deep learning framework. The CNN experiment is trained on the ground-based cloud image classification datasets MGCD and GRSCD respectively. The number of training data accounts for 50%, the initial learning rate is set to 0.0002, Batchsize is set to 32, and Adam optimizer (Kingma and Ba, 2015) is used to optimize all available parameters in the network. In addition, to improve the generalization ability of the CNN model and the convergence speed of the experiment, the transfer learning method is adopted in the training stage, and model parameters are obtained by training RepVGG with the ground-based cloud image classification dataset made by the team and used as the weight of pre-training. CNN experiment directly trains based on pre-training weight, which can accelerate the model convergence speed and shorten the training time, avoid the problem of parameter overfitting, and promote the rapid gradient decline.

3.2.2 Evaluation Index

To objectively evaluate the ground-based cloud image classification performance of CloudRVE and other CNN models, the accuracy rate, recall rate, and the average values of different indices of 7 types of clouds in MGCD and GRSCD datasets are calculated in the experiment, which is used as evaluation indices of CNN model. The accuracy rate and average accuracy rate is derived based on positive and negative samples, n represents the number of cloud types and the calculation process is as follows:

$$Accuracy(Acc) = \frac{TP+TN}{TP+TN+FP+FN}\ ,\ \overline{Accuracy}(\overline{Acc}) = \frac{1}{n}\sum_{i=1}^{n}\frac{TP_i+TN_i}{TP_i+TN_i+FP_i+FN_i} \tag{15}$$

TP (True Positive) parameter is the number of correctly classified samples for a specific genus, TN (True Negative) parameter is the number of correctly classified samples for the remaining genus, and FN (False Negative) parameter is the number of misclassified samples for a specific class genus. FP (False Positive) parameter is the number of misclassified samples for the remaining classes genera. The precision rate, average precision rate, recall rate and average recall rate can be expressed as:

$$Precison(Pr) = \frac{TP}{TP+FP}\ ,\ \overline{Precison}(\overline{Pr}) = \frac{1}{n}\sum_{i=1}^{n}\frac{TP_i}{TP_i+FP_i} \tag{16}$$

$$Recall(Re) = \frac{TP}{TP+FN}\ ,\overline{Recall}(\overline{Re}) = \frac{1}{n}\sum_{i=1}^{n}\frac{TP_i}{TP_i+FN_i} \tag{17}$$

In addition, the specificity, average specificity, F1_score and average F1_score are also used as evaluation indices of the CNN model in the experiment, and their expressions are shown as follows:

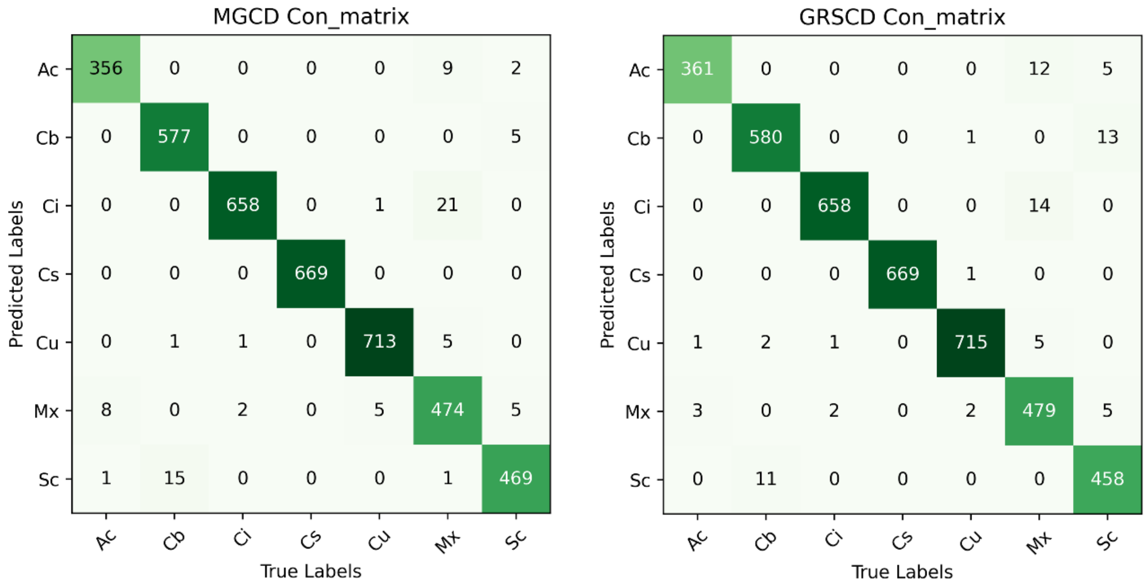
$$Specificity(TNR) = \frac{TN}{FP+TN}\ ,\overline{Specificity}(\overline{TNR}) = \frac{1}{n}\sum_{i=1}^{n}\frac{TN_i}{FP_i+TN_i} \tag{18}$$

$$F1\_score(F1) = \frac{2\times Pr\times Re}{Pr+Re}\ ,\ \overline{F1\_score}(\overline{F1}) = \frac{1}{n}\sum_{i=1}^{n}\frac{2\times Pr_i\times Re_i}{Pr_i+Re_i} \tag{19}$$

4. Experimental Results and Discussion

4.1 Classification Results of Ground-Based Cloud Images

Figure 12 shows the confusion matrix of MGCD and GRSCD datasets, showing CloudRVE prediction results on MGCD and GRSCD datasets. The horizontal axis represents the true cloud image classification, while the vertical axis represents the predicted cloud image classification, where the value of the diagonal element represents the correct number of cloud image classifications and the value of the off-diagonal element represents the number of cloud image classification errors. As can be seen from Figure 12(a), in the MGCD dataset, the correct classification of the Cu is the largest, while the misclassification of the cloud images mainly comes from Sc and Mx. The reason is that the cloud base of Sc is blackened by illumination, making it easily confused with Cb. In addition, the dynamic change of cloud will lead to a change in the viewpoint of the whole sky camera, thus increasing the difficulty of cloud genus identification. As can be seen in Figure 12(b), in the GRSCD dataset, the correctly classified cloud images of the same Cu had the largest number, while the incorrectly classified ones mainly came from Mx and Sc. The Mx cloud is a hybrid cloud, containing a variety of different cloud genera, with large shares of Ac, Ci, and Cu, which could be erroneously classified as Mx. Similarly, Sc could be taken for Cb, due to their similar features, impeding the correct identification.



(a) (b)

Figure 12: Confusion matrix images. (a)MGCD confusion matrix image. (b) GRSCD confusion matrix image.

The overall classification accuracy of the CloudRVE method proposed in this paper in MGCD and GRSCD datasets and the classification results of each cloud genus are listed in Tables 4 and 5. It can be seen that the accuracy of CloudRVE in MGCD and GRSCD datasets reached 98.15 and 98.07%, respectively. The characteristics of the Cl in MGCD and GRSCD datasets were easy to identify, resulting in the accuracy rate, recall rate, specificity, and F1 value reaching 100%. In the MGCD dataset, the accuracy rate, recall rate, and F1 value of the other six cloud genera all exceeded 95.00%, and the specificity was above 99.50%. The accuracy and specificity of the Ci were the highest, reaching 98.64 and 99.73%, respectively. Cu had the highest recall rate and F1 value, reaching 99.17 and 98.89%, respectively. In addition, the recall rate and F1 value of Sc and Mx were about 2.00% lower than other cloud genera, mainly their characteristics in the MGCD dataset were similar to those of Cb and Ci, respectively, reducing CloudRVE's ability to classify them.

Table 4. Classification results for the MGCD dataset.

Genus	$\overline{Acc}$ (%)	Pr (%)	Re (%)	TNR (%)	F1 (%)
Cu		98.62	99.17	99.70	98.89
Ac		97.02	98.08	99.70	97.55
Ci		98.64	98.94	99.73	98.79
Cl	98.15	100.0	100.0	100.0	100.0
Sc		97.26	95.84	99.63	96.54
Cb		97.13	97.13	99.51	97.13
Mx		97.24	96.67	99.60	96.95

Table 5. Classification results for the GRSCD dataset.

Genus	$\overline{Acc}$ (%)	Pr (%)	Re (%)	TNR (%)	F1 (%)
Cu		99.30	99.03	99.85	99.16
Ac		94.24	98.63	99.39	96.39
Ci		97.91	99.24	99.58	98.57
Cl	98.07	100.0	100.0	100.0	100.0
Sc		98.10	96.47	99.74	97.27
Cb		97.33	98.48	99.53	97.90
Mx		97.74	93.33	99.68	95.49

In the GRSCD dataset, the accuracy rate, recall rate, and F1 value of the other six cloud genera exceeded 94.00%, and the specificity as over 99.30%. Cu had the highest accuracy, specificity, and F1 value, reaching 99.30, 99.85, and 99.16%. The recall rate of Ci was the highest, reaching 99.17%. In addition, the Ac accuracy was only 94.24%, mainly because Ac contained a small amount of Sc, and CloudRVE could easily to misjudge Ac as Sc or Mx. Mx contained a variety of other clouds, and the images composition was complex. Cloud clusters of different can genera varied in size and shape, resulting in lower recall rate and F1 values.

4.2. Ablation Experiment

Table 6. Results of the ablation experiment.

Dataset	Model	$\overline{Acc}$ (%)	$\overline{Pr}$ (%)	$\overline{Re}$ (%)	$\overline{TNR}$ (%)	$\overline{F1}$ (%)
MGCD	RepVGG	95.57	95.31	94.99	99.26	95.14
	RepVGG_M	95.97	95.65	95.67	99.33	95.56
	RepVGG_M+ECA	96.80	96.60	96.37	99.47	96.45
	CloudRVE	**98.15**	**97.99**	**97.98**	**99.68**	**97.83**
GRSCD	RepVGG	95.42	94.99	94.88	99.24	94.92
	RepVGG_M	95.70	95.46	95.30	99.29	95.36
	RepVGG_M+ECA	96.10	95.67	95.74	99.35	95.68
	CloudRVE	**98.07**	**97.80**	**97.88**	**99.68**	**97.82**

In this section, the ablation experiment is used to compare the original structure and different improvement stages of the proposed method on the MGCD and GRSCD datasets respectively, and the results are shown in Table 6. RepVGG_M is the main improved network, ECA is the attention module, CloudRVE is the combined improved network of RepVGG_M and NECA, and is the final version of the method proposed in this paper. It can be seen from the data in the table that the performance of each improvement stage of the network model is improved compared to the previous stage, which not only verifies the feasibility of extracting more cloud image detail features by adding 1×1 convolutional layer branches but also verifies that NECA can effectively improve the noise suppression ability and enhance the channel feature extraction ability. Compared with the original network structure, the accuracy of CloudRVE in the MGCD dataset increased by 2.58%, the average accuracy rate increased by 2.68%, the average recall rate increased by 2.99%, the average specificity increased by 0.42%, and the average F1 value increased by

2.69%. In the GRSCD dataset, the accuracy rate increased by 2.65%, the average accuracy rate increased by 2.81%, the average specificity increased by 0.44%, and the average F1 value increased by 2.69%. Therefore, it can be seen from the data display that the method proposed in this paper has the best performance.

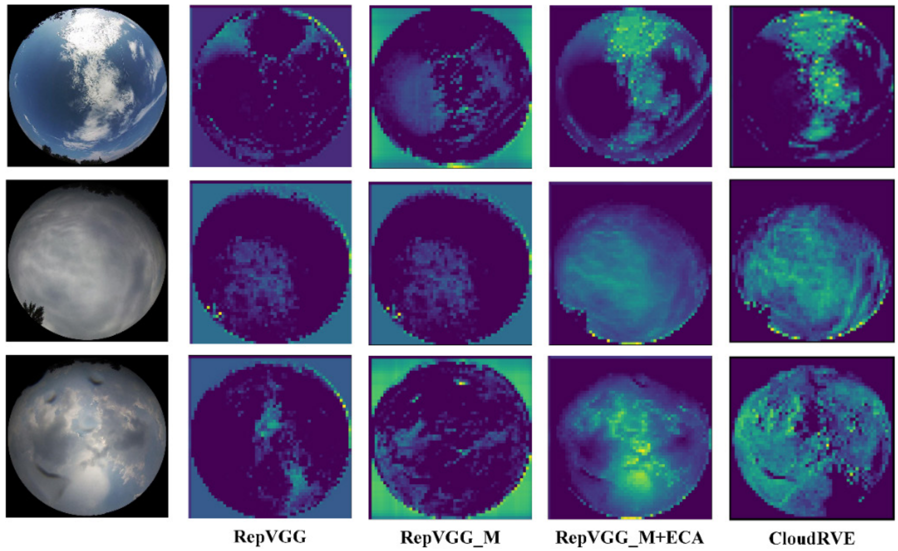


Figure 13: Feature extraction of different models based on MGCD (Liu et al., 2020a).

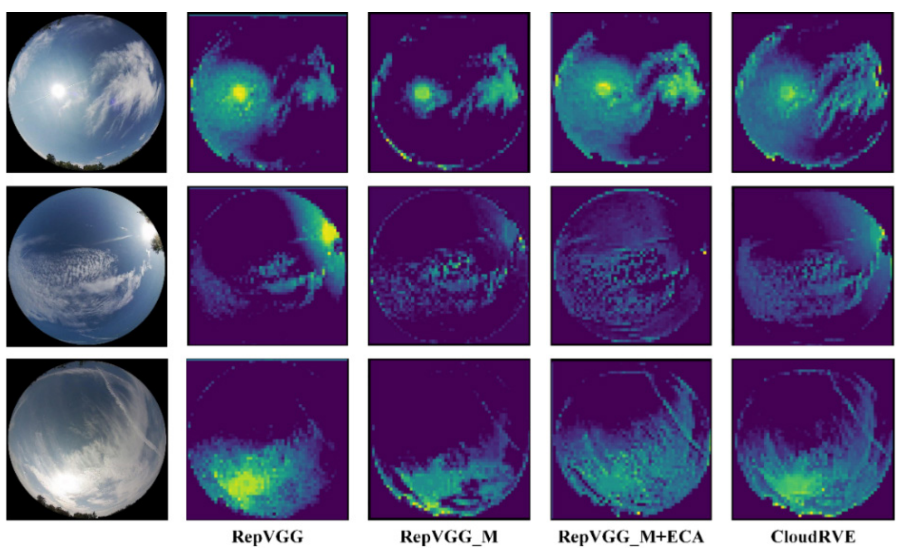
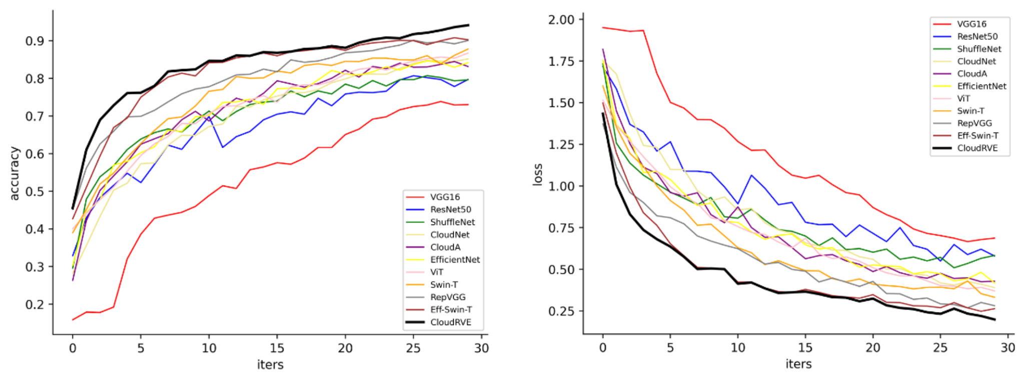


Figure 14: Feature extraction of different models based on GRSCD (Liu et al., 2020b).

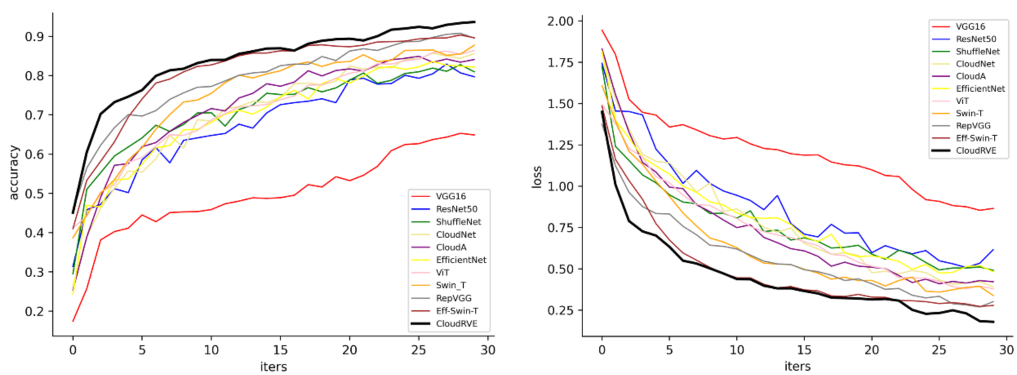
To visually compare the performance of the original structure and the method proposed in this paper in different improvement stages, we visualize the features by extracting the feature map of the middle layer of the network and then explain the feature extraction ability of the original structure and the method proposed in this paper in different improvement stages, as shown in Figures 13 and 14. The method generates a rough feature map to display the important region of the predicted images through the parameter weights generated by network training, in which the brighter the region indicates the higher its importance, and the darker the region represents the sky or those that cannot be extracted. Figure 13 shows that CloudRVE has the best feature location and extraction ability by showing the feature maps of three different cloud images in the MGCD dataset. Figure 14 shows that the three cloud images of the GRSCD dataset include not only clouds and sky but also strong sunlight, which affects the classification accuracy of the model. However, it can be seen from the feature maps that CloudRVE not only has the best feature extraction ability but also has a strong ability to suppress noise such as sunlight.

4.3 Comparison of Experimental Results



(a) **(b)**

Figure 15: Training accuracy (a) and training loss (b) curves of the MGCD dataset.



(a) **(b)**

Figure 16: Training accuracy (a) and training loss (b) curves of the GRSCD dataset.

To verify the feasibility of the proposed CloudRVE method, we compared it with other advanced methods, including CloudNet (Zhang et al., 2018), CloudA (Wang et al., 2020), Eff-Swin-T (Li et al., 2022), and other ground-based cloud image classification methods. These included such classic CNN models as VGG16 (Szegedy et al., 2015), ResNet50 (He et al., 2016), ShuffleNet (Zhang et al., 2018) and EfficientNet (Tan and Le, 2019). In addition, we compared it with other Transformer-based classification models such as ViT-L (Dosovitskiy et al., 2022), Swin-T(Liu et al., 2021), etc. Figures 15 and 16 illustrate the performances of different methods by displaying the training accuracy and training loss curves of MGCD and GRSCD datasets. Here the black bold curve represents the CloudRVE method, which has the largest accuracy value, the fastest convergence rate, the smallest loss rate, and the fastest decline rate in the training stage. This strongly indicates that the CloudRVE method has the best classification performance of ground-based cloud images.

Table 7. Comparison of experimental results.

Method	**MGCD**					**GRSCD**				
	$\overline{Acc}$ (%)	$\overline{Pr}$(%)	$\overline{Re}$ (%)	$\overline{TNR}$ (%)	$\overline{F1}$(%)	$\overline{Acc}$ (%)	$\overline{Pr}$(%)	$\overline{Re}$ (%)	$\overline{TNR}$ (%)	$\overline{F1}$ (%)
VGG-16	78.25	77.04	75.52	96.36	75.55	73.50	73.88	70.29	95.53	70.87
ResNet-50	85.98	85.24	84.55	97.67	84.82	86.51	85.56	85.38	97.75	85.34
ShuffleNet	86.95	86.08	85.68	97.83	85.71	86.99	86.85	85.18	97.82	85.71
CloudNet	90.01	89.24	89.08	98.34	89.13	89.60	89.06	88.60	98.27	88.79
CloudA	89.62	88.78	88.50	98.28	88.61	90.03	89.54	88.71	98.34	89.03
EfficientNet	91.17	90.66	90.22	98.53	90.27	90.10	89.68	88.92	98.35	89.13
ViT-L	91.11	90.91	90.21	98.55	90.40	90.98	90.49	90.33	98.50	90.39
Swin-T	92.87	92.44	91.63	98.63	91.76	93.55	93.22	92.87	98.93	92.71
RepVGG	95.57	95.31	94.99	99.26	95.14	95.42	94.99	94.88	99.24	94.92
Eff-Swin-T	96.93	96.73	96.44	99.49	96.56	95.62	95.41	95.11	99.27	95.21
CloudRVE	**98.15**	**97.99**	**97.98**	**99.68**	**97.83**	**98.07**	**97.80**	**97.88**	**99.68**	**97.82**

The comparative analysis results of the above methods are summarized in Table 7. It can be seen from the experimental results that RepVGG had the best performance among the CNN-based methods. Among them, the accuracy rate has the most significant improvement, and the precision and recall rates also have good improvement. The accuracy rate, precison rate, recall rate for the MGCD dataset reached 95.57, 95.31, and 94.99, respectively, while those for the GRSCD dataset were 95.42, 94.99, and 94.88, respectively. Ground-based cloud images have more texture features and deep semantic features than other images, and more image features need to be obtained to meet the classification requirements of such images. In recent years, Transformer has been widely used for image processing tasks due to its strong feature extraction capability. Several scholars have improved the Transformer derivative model through continuous exploration. Among them, Eff-Swin-T was an improvement based on Swin-T, and its performance on MGCD and GRSCD datasets was better than that of the classic CNN model. Its accuracy rate, precison rate, and recall rate reached 96.93, 96.73, 96.44, and 95.62, 95.41, 95.11, respectively. Compared with

Transformer and classical networks, the proposed method had much better classification performance of ground-based cloud images. For different cloud image classification datasets, it exhibited excellent generalization ability and strong robustness, which is instrumental in photovoltaic power generation prediction.

The space complexities of CloudRVE and ten alternative methods are summarized and compared in Table 8. It can be seen from the table that CloudRVE had a spatial complexity of 105.17 Mb, which is in line with the spatial complexity of Swin-T and Eff-Swin-T, and far less than the spatial complexity of ViT-L. The spatial complexity of CloudRVE exceeded that of RepVGG by three times, achieving the best ground cloud image classification performance. Thus, CloudRVE achieved excellent ground cloud image classification performance at the expense of higher spatial complexity.

Table 8. Space complexity of the proposed and ten alternative methods.

Method	**Space complexity (Mb)**
VGG-16	512.28
ResNet-50	90.03
ShuffleNet	4.93
CloudNet	153.36
CloudA	87.57
EfficientNet	15.61
ViT-L	327.37
Swin-T	105.28
RepVGG	30.10
Eff-Swin-T	105.24
CloudRVE	**105.17**

In order to provide a more intuitive display of the advantages of CloudRVE over other advanced methods, we extracted the features of the intermediate layers of different methods to generate the ground cloud feature maps for the building foundation, demonstrating the strong feature extraction capabilities of CloudRVE and proving its superiority, as shown in Figures 17 and 18. Feature extraction was achieved by generating rough feature maps through network training with parameter weights to highlight the important regions of predicted images. The light colored regions represent the important features, while the dark colored regions represent the sky or unsuccessfully extracted features. Figure 17(b-i) shows the feature maps of different ground cloud classification methods based on MGCD dataset to demonstrate the CloudRVE capability to extract more extensive and comprehensive cloud features and suppress the black regions and sunlight, further illustrating the best feature localization and extraction capability of CloudRVE. Figure 18(b-i) shows the feature maps of different ground cloud classification methods based on GRSCD dataset to demonstrate that the cloud feature extracted by CloudRVE covers the effective area in Figure 18(a) with the best coverage and the best suppression of the sunlight, further proving that CloudRVE has the best feature localization and extraction capabilities.

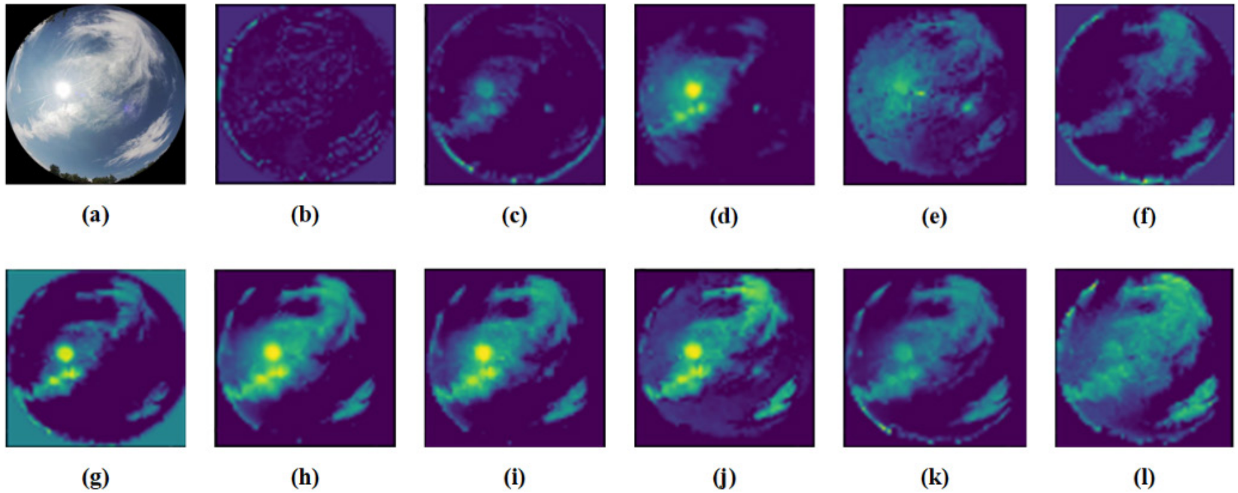


Figure 17: Feature extraction of different methods based on MGCD, (a) Original (Liu et al., 2020a); (b)VGG-16; (c) ResNet-50; (d) ShuffleNet; (e) CloudNet; (f) CloudA; (g) EfficientNet; (h) ViT-L; (i) Swin-T; (j) RepVGG; (k) Eff-Swin-T; (l) CloudRVE

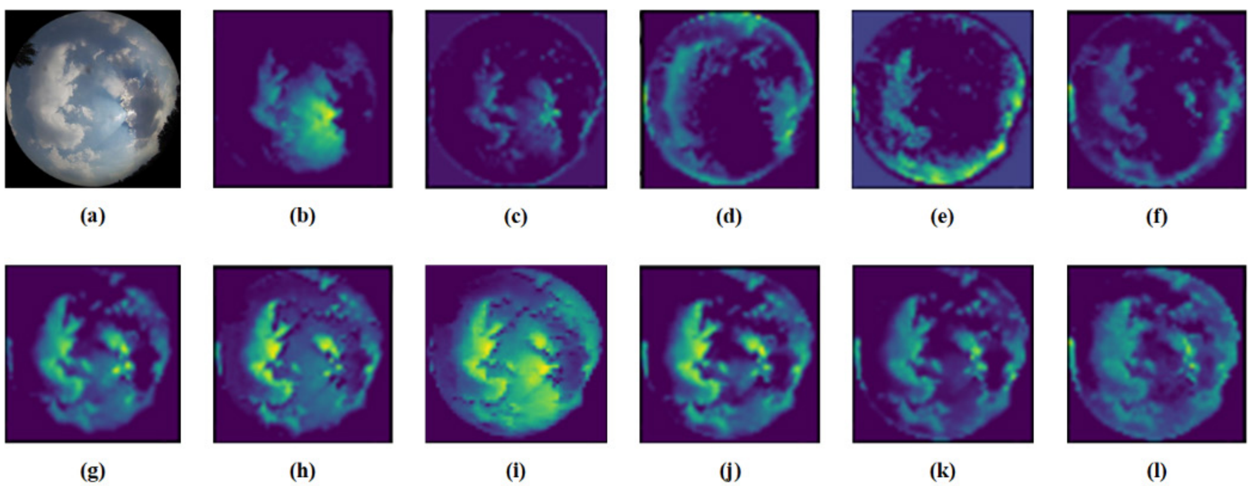


Figure 18: Feature extraction of different methods based on GRSCD: (a) Original (Liu et al., 2020b); (b)VGG-16; (c)ResNet-50; (d) ShuffleNet; (e) CloudNet; (f) CloudA; (g) EfficientNet; (h) ViT-L; (i) Swin-T; (j) RepVGG; (k) Eff-Swin-T; (l) CloudRVE

5. Conclusion

This study proposed a new classification method called CloudRVE for ground-based cloud images based on the improved RepVGG network. In particular, its training stage structure was improved, the residual structure was broadened, and 1×1 convolutional layer branches were added to each block, extending the gradient information of the topology structure and enhancing the network ability to represent boundary features of cloud images. In addition, the NECA module was embedded after multi-branch fusion to learn the feature relationship between sequences, improve the network cross-channel interaction ability, and extract the best global features of cloud images. We validated the excellent performance of the proposed method on MGCD and GRSCD ground-based cloud image datasets, achieving the classification accuracy values of 98.15 and 98.07%, respectively, which outperformed ten other advanced methods. In addition, the MGCD and GRSCD ground-based cloud image datasets contain 7 types of cloud categories, which is more than the ground-based cloud image datasets used in other papers. This further demonstrates the excellent performance of the proposed method. The particular contributions of this paper were summarized in Section 1. However, this study shares some limitations with other methods of classifying ground-based cloud images via convolutional neural networks, which have reached a bottleneck due to continuous expansion of the capacity of ground-based cloud image datasets. A lucrative alternative is Transformer, which got a high reputation of a powerful deep neural network for processing sequences but has received little attention in ground-based cloud image classification. On the other hand, cloud classification is only based on ground-based cloud image features, while many physical features, such as height, thickness, etc., may be also used. Our follow-up study envisages combining CNN and Transformer models and using cloud height, cloud thickness, and other parameters in ground-based cloud image classification to improve the model's performance.

Author Contributions. LH performed the experiments and wrote the paper. CS, KZ, and HX analyzed the data and designed the experiments. CS conceived the method and reviewed the paper. XL, ZS, and XZ reviewed the paper and gave constructive suggestions.

Financial support. This research was funded by the National Science Foundation of China (NSFC) under Grant No. 62076093 and No. 62206095 and through the Fundamental Research Funds for the Central Universities of China under Grant No. 2022MS078 and No. 2020MS099.

Data Availability Statement. The MGCD dataset was accessed from https://github.com/shuangliutjnu/Multimodal-Ground-based-Cloud-Database. The GRSCD dataset was accessed from https://github.com/shuangliutjnu/TJNU-Ground-based-Remote-Sensing-Cloud-Database.

Acknowledgments. We would like to thank Professor Liu Shuang of Tianjin Normal University for providing the support of ground-based cloud image classification datasets and Student Meng Ru-oxuan from Guangxi Normal University for her contribution to this paper.

Declaration of Competing Interests. The authors declare that they have no conflict of interest.

References

Alonso-Montesinos, J., Martinez-Durban, M., del Sagrado, J., del Aguila, I. M., and Batlles, F. J.: The application of Bayesian network classifiers to cloud classification in satellite images, Renew. Energy, 97, 155–161, https://doi.org/10.1016/j.renene.2016.05.066, 2016.

Calbó, J. and Sabburg, J.: Feature Extraction from Whole-Sky Ground-Based Images for Cloud-Type Recognition, J. Atmospheric Ocean. Technol., 25, 3–14, https://doi.org/10.1175/2007JTECHA959.1, 2008.

Cazorla, A., Olmo, F. J., and Alados-Arboledas, L.: Development of a sky imager for cloud cover assessment, JOSA A, 25, 29–39, https://doi.org/10.1364/JOSAA.25.000029, 2008.

Ding, X., Zhang, X., Ma, N., Han, J., Ding, G., and Sun, J.: RepVGG: Making VGG-style ConvNets Great Again, in: 2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 13728–13737, https://doi.org/10.1109/CVPR46437.2021.01352, 2021.

Dosovitskiy, A., Beyer, L., Kolesnikov, A., Weissenborn, D., Zhai, X., Unterthiner, T., Dehghani, M., Minderer, M., Heigold, G., Gelly, S., Uszkoreit, J., and Houlsby, N.: An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale, International Conference on Learning Representations, 1909, 2022.

Fabel, Y., Nouri, B., Wilbert, S., Blum, N., Triebel, R., Hasenbalg, M., Kuhn, P., Zarzalejo, L. F., and Pitz-Paal, R.: Applying self-supervised learning for semantic cloud segmentation of all-sky images, Atmospheric Meas. Tech., 15, 797–809, https://doi.org/10.5194/amt-15-797-2022, 2022.

Goren, T., Rosenfeld, D., Sourdeval, O., and Quaas, J.: Satellite Observations of Precipitating Marine Stratocumulus Show Greater Cloud Fraction for Decoupled Clouds in Comparison to Coupled Clouds, Geophys. Res. Lett., 45, 5126–5134, https://doi.org/10.1029/2018GL078122, 2018.

Gorodetskaya, I. V., Kneifel, S., Maahn, M., Van Tricht, K., Thiery, W., Schween, J. H., Mangold, A., Crewell, S., and Van Lipzig, N. P. M.: Cloud and precipitation properties from ground-based remote-sensing instruments in East Antarctica, The Cryosphere, 9, 285–304, https://doi.org/10.5194/tc-9-285-2015, 2015.

Gyasi, E. K. and Swarnalatha, P.: Cloud-MobiNet: An Abridged Mobile-Net Convolutional Neural Network Model for Ground-Based Cloud Classification, Atmosphere, 14, 280, https://doi.org/10.3390/atmos14020280, 2023.

He, K., Zhang, X., Ren, S., and Sun, J.: Deep Residual Learning for Image Recognition, in: 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 770–778, https://doi.org/10.1109/CVPR.2016.90, 2016.

He, T., Zhang, Z., Zhang, H., Zhang, Z., Xie, J., and Li, M.: Bag of Tricks for Image Classification with Convolutional Neural Networks, 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Los Alamitos, 558–567, https://doi.org/10.1109/CVPR.2019.00065, 2019.

Heinle, A., Macke, A., and Srivastav, A.: Automatic cloud classification of whole sky images, Atmospheric Meas. Tech., 3, 557–567, https://doi.org/10.5194/amt-3-557-2010, 2010.

Hu, J., Shen, L., and Sun, G.: Squeeze-and-Excitation Networks, in: 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition, 7132–7141, https://doi.org/10.1109/CVPR.2018.00745, 2018.

Ioffe, S. and Szegedy, C.: Batch Normalization: Accelerating Deep Network Training by Reducing Internal Covariate Shift, in: International Conference on Machine Learning, Vol 37, San Diego, 448–456, 2015.

Kalisch, J. and Macke, A.: Estimation of the total cloud cover with high temporal resolution and parametrization of short-term fluctuations of sea surface insolation, Meteorol. Z., 603–611, https://doi.org/10.1127/0941-2948/2008/0321, 2008.

Kingma, D. P. and Ba, J.: Adam: A Method for Stochastic Optimization, 3rd International Conference for Learning Representations, San Diego, arXiv:1412.6980 [cs], https://doi.org/10.48550/arXiv.1412.6980, 2015.

Krizhevsky, A., Sutskever, I., and Hinton, G. E.: ImageNet classification with deep convolutional neural networks, Commun. ACM, 60, 84–90, https://doi.org/10.1145/3065386, 2017.

Li, X., Qiu, B., Cao, G., Wu, C., and Zhang, L.: A Novel Method for Ground-Based Cloud Image Classification Using Transformer, Remote Sens., 14, 3978, https://doi.org/10.3390/rs14163978, 2022.

Li, Z., Kong, H., and Wong, C.-S.: Neural Network-Based Identification of Cloud Types from Ground-Based Images of Cloud Layers, Appl. Sci., 13, 4470, https://doi.org/10.3390/app13074470, 2023.

Lin, F., Zhang, Y., and Wang, J.: Recent advances in intra-hour solar forecasting: A review of ground-based sky image methods, Int. J. Forecast., 39, 244–265, https://doi.org/10.1016/j.ijforecast.2021.11.002, 2023.

Liu, S., Li, M., Zhang, Z., Cao, X., and Durrani, T. S.: Ground-Based Cloud Classification Using Task-Based Graph Convolutional Network, Geophys. Res. Lett., 47, e2020GL087338, https://doi.org/10.1029/2020GL087338, 2020a.

Liu, S., Li, M., Zhang, Z., Xiao, B., and Durrani, T. S.: Multi-Evidence and Multi-Modal Fusion Network for Ground-Based Cloud Recognition, Remote Sens., 12, 464, https://doi.org/10.3390/rs12030464, 2020b.

Liu, S., Duan, L., Zhang, Z., Cao, X., and Durrani, T. S.: Multimodal Ground-Based Remote Sensing Cloud Classification via Learning Heterogeneous Deep Features, IEEE Trans. Geosci. Remote Sens., 58, 7790–7800, https://doi.org/10.1109/TGRS.2020.2984265, 2020c.

Liu, S., Duan, L., Zhang, Z., Cao, X., and Durrani, T. S.: Ground-Based Remote Sensing Cloud Classification via Context Graph Attention Network, IEEE Trans. Geosci. Remote Sens., 60, 1–11, https://doi.org/10.1109/TGRS.2021.3063255, 2022.

Liu, Z., Lin, Y., Cao, Y., Hu, H., Wei, Y., Zhang, Z., Lin, S., and Guo, B.: Swin Transformer: Hierarchical Vision Transformer using Shifted Windows, in: 2021 IEEE/CVF International Conference on Computer Vision (ICCV), 2021 IEEE/CVF International Conference on Computer Vision (ICCV), Montreal, QC, Canada, 9992–10002, https://doi.org/10.1109/ICCV48922.2021.00986, 2021.

Long, C., Li, X., Jing, Y., and Shen, H.: Bishift Networks for Thick Cloud Removal with Multitemporal Remote Sensing Images, Int. J. Intell. Syst., 2023, e9953198, https://doi.org/10.1155/2023/9953198, 2023.

Long, C. N., Sabburg, J. M., Calbó, J., and Pagès, D.: Retrieving Cloud Characteristics from Ground-Based Daytime Color All-Sky Images, J. Atmospheric Ocean. Technol., 23, 633–652, https://doi.org/10.1175/JTECH1875.1, 2006.

Meng, Q., Zhao, S., Huang, Z., and Zhou, F.: MagFace: A Universal Representation for Face Recognition and Quality Assessment, in: 2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 14220–14229, https://doi.org/10.1109/CVPR46437.2021.01400, 2021.

Norris, J. R., Allen, R. J., Evan, A. T., Zelinka, M. D., O’Dell, C. W., and Klein, S. A.: Evidence for climate change in the satellite cloud record, Nature, 536, 72–75, https://doi.org/10.1038/nature18273, 2016.

Nouri, B., Kuhn, P., Wilbert, S., Hanrieder, N., Prahl, C., Zarzalejo, L., Kazantzidis, A., Blanc, P., and Pitz-Paal, R.: Cloud height and tracking accuracy of three all sky imager systems for individual clouds, Sol. Energy, 177, 213–228, https://doi.org/10.1016/j.solener.2018.10.079, 2019.

Pfister, G., McKenzie, R. L., Liley, J. B., Thomas, A., Forgan, B. W., and Long, C. N.: Cloud Coverage Based on All-Sky Imaging and Its Impact on Surface Solar Irradiance, J. Appl. Meteorol. Climatol., 42, 1421–1434, https://doi.org/10.1175/1520-0450(2003)042<1421:CCBOAI>2.0.CO;2, 2003.

Qu, Y., Xu, J., Sun, Y., and Liu, D.: A temporal distributed hybrid deep learning model for day-ahead distributed PV power forecasting, Appl. Energy, 304, 117704, https://doi.org/10.1016/j.apenergy.2021.117704, 2021.

Sarukkai, V., Jain, A., Uzkent, B., and Ermon, S.: Cloud Removal in Satellite Images Using Spatiotemporal Generative Networks, in: 2020 IEEE Winter Conference on Applications of Computer Vision (WACV), 2020 IEEE Winter Conference on Applications of Computer Vision (WACV), 1785–1794, https://doi.org/10.1109/WACV45572.2020.9093564, 2020.

Shi, C., Wang, C., Wang, Y., and Xiao, B.: Deep Convolutional Activations-Based Features for Ground-Based Cloud Classification, IEEE Geosci. Remote Sens. Lett., 14, 816–820, https://doi.org/10.1109/LGRS.2017.2681658, 2017.

Shi, C., Zhou, Y., Qiu, B., He, J., Ding, M., and Wei, S.: Diurnal and nocturnal cloud segmentation of all-sky imager (ASI) images using enhancement fully convolutional networks, Atmospheric Meas. Tech., 12, 4713–4724, https://doi.org/10.5194/amt-12-4713-2019, 2019.

Simonyan, K. and Zisserman, A.: Very Deep Convolutional Networks for Large-Scale Image Recognition, https://doi.org/10.48550/arXiv.1409.1556, 10 April 2015.

Singh, M. and Glennen, M.: Automated ground-based cloud recognition, Pattern Anal. Appl., 8, 258–271, https://doi.org/10.1007/s10044-005-0007-5, 2005.

Szegedy, C., Liu, W., Jia, Y., Sermanet, P., Reed, S., Anguelov, D., Erhan, D., Vanhoucke, V., and Rabinovich, A.: Going deeper with convolutions, in: 2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 1–9, https://doi.org/10.1109/CVPR.2015.7298594, 2015.

Tan, M. and Le, Q. V.: EfficientNet: Rethinking Model Scaling for Convolutional Neural Networks, in: International Conference on Machine Learning, Vol 97, San Diego, 2019.

Tang, Y., Yang, P., Zhou, Z., Pan, D., Chen, J., and Zhao, X.: Improving cloud type classification of ground-based images using region covariance descriptors, Atmospheric Meas. Tech., 14, 737–747, https://doi.org/10.5194/amt-14-737-2021, 2021.

Taravat, A., Del Frate, F., Cornaro, C., and Vergari, S.: Neural Networks and Support Vector Machine Algorithms for Automatic Cloud Classification of Whole-Sky Ground-Based Images, IEEE Geosci. Remote Sens. Lett., 12, 666–670, https://doi.org/10.1109/LGRS.2014.2356616, 2015.

Wang, M., Zhou, S., Yang, Z., and Liu, Z.: CloudA: A Ground-Based Cloud Classification Method with a Convolutional Neural Network, J. Atmospheric Ocean. Technol., 37, 1661–1668, https://doi.org/10.1175/JTECH-D-19-0189.1, 2020.

Wang, M., Zhuang, Z., Wang, K., Zhou, S., Zhou, S., and Liu, Z.: Intelligent classification of ground-based visible cloud images using a transfer convolutional neural network and fine-tuning, Opt. Express, 29, 41176–41190, https://doi.org/10.1364/OE.442455, 2021.

Wu, X., Zhan, C., Lai, Y.-K., Cheng, M.-M., and Yang, J.: A Large-Scale Benchmark Dataset for Insect Pest Recognition, in: 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Conference Location: Long Beach, CA, USA, 8779–8788, https://doi.org/10.1109/CVPR.2019.00899, 2019.

Ye, L., Cao, Z., and Xiao, Y.: DeepCloud: Ground-Based Cloud Image Categorization Using Deep Convolutional Features, IEEE Trans. Geosci. Remote Sens., 55, 5729–5740, https://doi.org/10.1109/TGRS.2017.2712809, 2017.

Yu, A., Tang, M., Li, G., Hou, B., Xuan, Z., Zhu, B., and Chen, T.: A Novel Robust Classification Method for Ground-Based Clouds, Atmosphere, 12, 999, https://doi.org/10.3390/atmos12080999, 2021.

Zhang, J., Liu, P., Zhang, F., and Song, Q.: CloudNet: Ground-Based Cloud Classification With Deep Convolutional Neural Network, Geophys. Res. Lett., 45, 8665–8672, https://doi.org/10.1029/2018GL077787, 2018a.

Zhang, X., Zhou, X., Lin, M., and Sun, R.: ShuffleNet: An Extremely Efficient Convolutional Neural Network for Mobile Devices, in: 2018 Ieee/Cvf Conference on Computer Vision and Pattern Recognition (cvpr), New York, 6848–6856, https://doi.org/10.1109/CVPR.2018.00716, 2018b.

Zhang, Y., Liu, H., and Hu, Q.: TransFuse: Fusing Transformers and CNNs for Medical Image Segmentation, in: Medical Image Computing and Computer Assisted Intervention – MICCAI 2021, Cham, 14–24, https://doi.org/10.1007/978-3-030-87193-2_2, 2021.

Zhao, Z., Xu, G., Qi, Y., Liu, N., and Zhang, T.: Multi-patch deep features for power line insulator status classification from aerial images, in: 2016 International Joint Conference on Neural Networks (IJCNN), 2016 International Joint Conference on Neural Networks (IJCNN), 3187–3194, https://doi.org/10.1109/IJCNN.2016.7727606, 2016.

Zheng, Y., Rosenfeld, D., Zhu, Y., and Li, Z.: Satellite-Based Estimation of Cloud Top Radiative Cooling Rate for Marine Stratocumulus, Geophys. Res. Lett., 46, 4485–4494, https://doi.org/10.1029/2019GL082094, 2019.

Zhong, B., Chen, W., Wu, S., Hu, L., Luo, X., and Liu, Q.: A Cloud Detection Method Based on Relationship Between Objects of Cloud and Cloud-Shadow for Chinese Moderate to High Resolution Satellite Imagery, IEEE J. Sel. Top. Appl. Earth Obs. Remote Sens., 10, 4898–4908, https://doi.org/10.1109/JSTARS.2017.2734912, 2017.

Zhu, W., Chen, T., Hou, B., Bian, C., Yu, A., Chen, L., Tang, M., and Zhu, Y.: Classification of Ground-Based Cloud Images by Improved Combined Convolutional Network, Appl. Sci., 12, 1570, https://doi.org/10.3390/app12031570, 2022.

Zhuo, W., Cao, Z., and Xiao, Y.: Cloud Classification of Ground-Based Images Using Texture–Structure Features, J. Atmospheric Ocean. Technol., 31, 79–92, https://doi.org/10.1175/JTECH-D-13-00048.1, 2014.