sensors

MDPI

Article

MMST: A Multi-Modal Ground-Based Cloud Image Classification Method

Liang Wei †, Tingting Zhu †, Yiren Guo and Chao Ni *

College of Mechanical and Electronic Engineering, Nanjing Forestry University, Nanjing 210037, China; tingtingzhu@njfu.edu.cn (T.Z.)
* Correspondence: chaoni@njfu.edu.cn
† These authors contributed equally to this work.

Abstract: In recent years, convolutional neural networks have been in the leading position for ground-based cloud image classification tasks. However, this approach introduces too much inductive bias, fails to perform global modeling, and gradually tends to saturate the performance effect of convolutional neural network models as the amount of data increases. In this paper, we propose a novel method for ground-based cloud image recognition based on the multi-modal Swin Transformer (MMST), which discards the idea of using convolution to extract visual features and mainly consists of an attention mechanism module and linear layers. The Swin Transformer, the visual backbone network of MMST, enables the model to achieve better performance in downstream tasks through pre-trained weights obtained from the large-scale dataset ImageNet and can significantly shorten the transfer learning time. At the same time, the multi-modal information fusion network uses multiple linear layers and a residual structure to thoroughly learn multi-modal features, further improving the model's performance. MMST is evaluated on the multi-modal ground-based cloud public data set MGCD. Compared with the state-of-art methods, the classification accuracy rate reaches 91.30%, which verifies its validity in ground-based cloud image classification and proves that in ground-based cloud image recognition, models based on the Transformer architecture can also achieve better results.

Keywords: ground-based cloud image (GCI) classification; Swin Transformer; global features; feature fusion

Citation: Wei, L.; Zhu, T.; Guo, Y.; Ni, C. MMST: A Multi-Modal Ground-Based Cloud Image Classification Method. *Sensors* **2023**, *23*, 4222. https://doi.org/10.3390/s23094222

Academic Editor: Stefanos Kollias

Received: 18 February 2023
Revised: 19 April 2023
Accepted: 19 April 2023
Published: 23 April 2023

1. Introduction

As a clean new energy source, the advantage of solar power compared to ordinary thermal power systems is that sun light shines on the Earth and can be exploited directly without mining and transportation. Secondly, solar energy has the characteristics of high energy capacity, no pollution, and wide distribution [1]. In the process of large-scale grid-connected photovoltaic power generation, the impact of PV power fluctuations on the grid cannot be ignored. Among the many factors affecting photovoltaic power generation, the most important meteorological factor is solar irradiance, and the condition of cloud cover will directly affect solar irradiance. In cloud observation, there are three main elements: cloud amount, cloud base height, and cloud shape. Among these elements, the cloud shape can instantly reflect the local atmospheric conditions, so the study of the classification of cloud shape is an essential part of cloud observation research.

In the early period, most weather stations relied on manual visual inspection by weather observers for cloud recognition, and the classification effect would vary depending on the observers' experience. For this reason, researchers have used classical image features to establish traditional machine-learning classification models for ground-based cloud images. However, the cloud recognition effect is not desirable, especially since the recognition rate of clouds in complex backgrounds is difficult to guarantee. Liu et al. [2]

used threshold segmentation and morphological methods to detect the edges of clouds and extract structural features from them to build a supervised classifier to classify weather clouds. Heinle et al. [3] extracted cloud color and texture features and obtained a high classification accuracy based on the Leave-One-Out Cross Validation method. Oikonomou et al. [4] adopted Regional Local Binary Pattern (R-LBP) and Four Patch-Local Binary Pattern (FP-LBP) to describe the global and local features of the cloud image, and in the classification stage, they used Support Vector Machine (SVM) and Linear Discriminant Analysis (LDA) classifiers. From the result, the classification results achieved over 90% accuracy with both datasets, much higher than the methods proposed in other papers. Xiao et al. [5] first extracted the original features of ground-based clouds from the perspectives of color, texture, and structure; they obtained more descriptive representation vectors using a dense sampling method and finally encoded the features and put them into the support vector machine for classification.

In recent decades, with the generation of datasets, the improvement of computing capability, and the development of various machine learning algorithms, the utilization of deep learning to deal with specific problems has received growing attention. Zhang et al. [6] proposed a new convolutional neural network (CNN) model, namely, CloudNet. The author proposed that we could achieve better results by combining abundant information from different locations in the ground-based cloud image. However, he did not elaborate on the specific impact of information from different locations on the model. Liu et al. [7] considered each cloud image as a node in the graph; they used GCN to aggregate information from the cloud images themselves and their connected images in a weighted manner to establish strong and weak connections between different classes of cloud images and mine the inherent structural information of clouds. Although Transformer [8] architecture has become a fundamental model in the natural language processing field, its application in computer vision still needs to be improved. In computer vision, attention mechanisms are used in conjunction with or to replace certain parts of convolutional networks. Therefore, inspired by the Transformer, Alexey et al. [9] proposed the Vision Transformer (Vit), which mainly divides the image into multiple blocks, called Patches, and then embeds them into the linear layer of the Transformer. Li et al. [10] applied the Swin Transformer to cloud image classification but still added a convolutional layer to the overall model to extract local information without using meteorological multi-modal information to assist in the classification. The formation of clouds is affected by many natural factors [11]. Therefore, utilizing this multi-modal information is significant for comprehensively characterizing clouds. Liu et al. [12] proposed a new ground-based cloud classification method, namely, the multi-level modality fusion model (HMF), which fuses deep multi-modal features and deep visual features at different levels, namely, low-level fusion and high-level fusion. High-level fusion combines the output of low-level fusion with visual features and multi-modal features.

Therefore, this paper proposes a ground-based cloud image classification model based on the multi-modal Swin Transformer (MMST). Without using the convolution module, the visual backbone network can perform well, integrating multi-modal information to enhance the model's representation ability further. The approach outperforms the currently available methods in the publicly available multi-modal base cloud image dataset MGCD [12], demonstrating its feasibility.

The main contributions of this paper are as follows:

1. This paper proposes a novel method based on the Swin Transformer. The model fully relies on the attention mechanism and the linear layer to learn the features of cloud images and multi-modal information. This method solves the shortcomings of the traditional CNN model, namely, that it cannot conduct global modeling, and the performance ceiling of the model is restricted by too much inductive bias.
2. We address the deficiency of learning only the modeling of images in the cloud classification task. Residual blocks are added to the linear layer to learn more complex feature representations of meteorological multi-modal information.

3. An experimental evaluation is carried out on the multi-modal base cloud image dataset MGCD, showing that the method proposed in this paper has better classification results.

2. Methods

2.1. Overview of the Classification Process

This paper proposes the MMST model for the ground-based cloud image classification task. First, the original ground-based cloud image (the image size is 1024 × 1024) is taken by the all-sky camera, and four kinds of original meteorological information (temperature, humidity, pressure, and wind speed) are collected by the sensor. The original ground-based cloud image and original meteorological information are then pre-processed, and the processed ground-based cloud image and meteorological information are input, respectively, into the visual backbone network Swin Transformer and Multi-modal Information Network to, respectively, obtain the Vision Feature and Multi-Modal Feature. Finally, they are sent to the Feature Fusion Network, the Concat operation is conducted, and the classification result is output by the Linear layer. The overall classification process is shown in Figure 1.

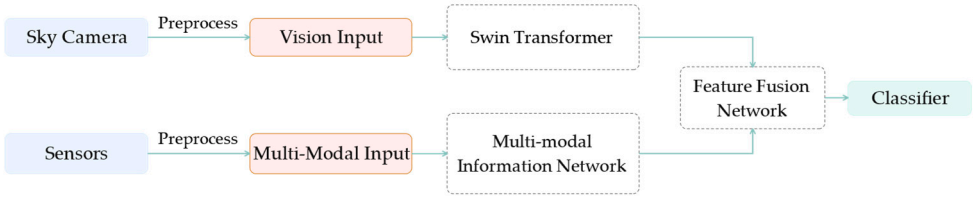


Figure 1. The classification process of ground-based cloud images.

2.2. Introduction of the Proposed Method

Figure 2 shows the overall structure of MMST. The visual backbone network Swin Transformer learns the feature relationship between image patch sequences by calculating the attention that the Patch can achieve for the purpose of global modeling. Shifting the windows can achieve the effect of fusing local information by interacting with the windows, similar to the convolution in CNN. With patch merging, the feature map size is downsized to achieve cascading multi-scale features and extract higher-level information, such as pooling in CNN. Moreover, the Multi-modal Information Network and Feature Fusion Network consist of blocks composed of linear layers to learn better feature representation using the residual structure, and the presence of the residual unit also ensures the stability of the gradients in model training.

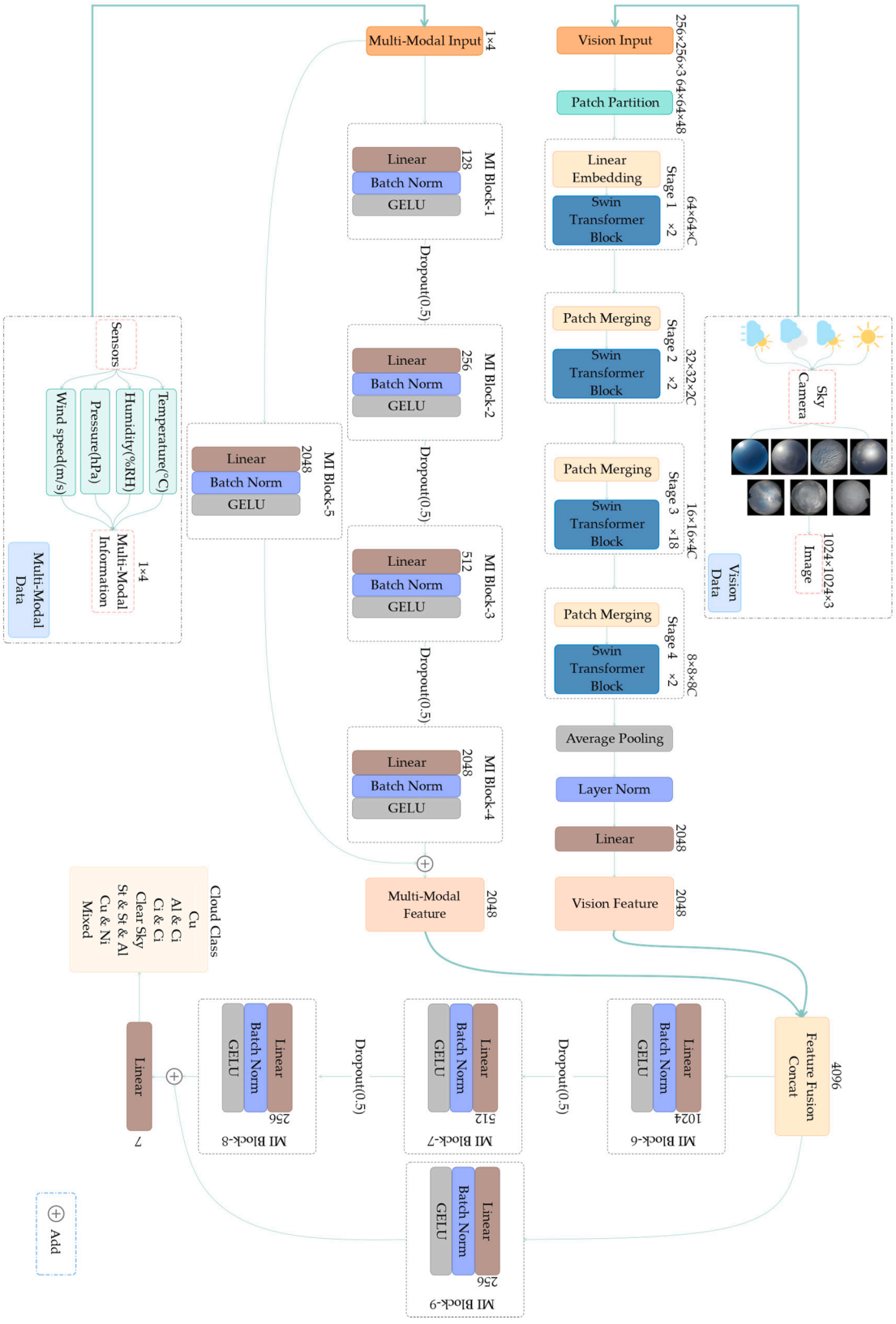


Figure 2. The structure of the proposed classification network (MMST).

2.3. Visual Backbone Network

2.3.1. Swin Transformer V2

In the MMST model, Swin Transformer V2 [13] is used as the vision backbone network. Swin Transformer V2 is an improved version of the Swin Transformer [14] (i.e., the

V1 version), which is also a model based on the Vit architecture. Due to the similarity of the main structure, in this article, the V1 and V2 versions of the Swin Transformer are not distinguished by name. The improvements proposed by the V2 version relative to the V1 version are roughly divided into two parts as follows:

(1) Improvements to self-attention

For the original Vit model, the two-dimensional (2D) image is first segmented, and the image of $X_p \in R^{H\times W\times C}$ is divided into several non-overlapping patches, namely, $X_p \in R^{P\times P\times C}$, where $H \times W$ is the size of the original image, C is the number of channels in the image, and $P \times P$ is the size of each patch. Since the spatial position of each patch has some influence on the later classification, position encoding is added to the vector of the patch projections to preserve their spatial information. The core components of the Vit encoder are a multi-head self-attention (MSA) module and a feed-forward multilayer perceptron (MLP) [8]. Figure 3 illustrates the structure of the MSA module, which is composed of multiple self-attention. The structure of self-attention is shown in Figure 3. The input vector is transformed by three trainable matrices to obtain three different homologous matrices: query matrix, Q; key matrix, K; and value matrix, V. The weight of each element (i.e., the importance of each element to the context) is obtained by computing the dot product of the query matrix and the key matrix and then applying $Softmax$ to scale the weight values to the interval of (0, 1). The transformed weight is multiplied by the value matrix to obtain the element value carrying the global importance information, which is different from local modeling in CNN [15]. The calculation of self-attention is as follows:

$$Attention(Q, K, V) = Softmax\left(\frac{QK^T}{\sqrt{d_k}}\right) \cdot V, \quad (1)$$

where $\sqrt{d_k}$ is the square root of the key matrix dimension, the purpose is to make the weight distribution smoother and more reasonable, and also to make the gradient more stable. MSA divides the input into n parts, i.e., n-head self-attention, performs the attention calculation of Equation (2) on the n parts of the input, and obtains n attention output results. Finally, the n output is concatenated and restored into the original dimension through the linear layer. The MSA calculation process is shown as follows:

$$head_i = Attention\left(QW_i^Q, KW_i^K, VW_i^V\right), \quad (2)$$

$$MSA(Q, K, V) = Concat(head_1, \ldots, head_i)W^0, \quad (3)$$

where i is the number of input vectors divided, $head_i$ is the attention output of the ith head, and the W_i^Q, W_i^K, W_i^V, and W^0 are all learnable parameter matrices.

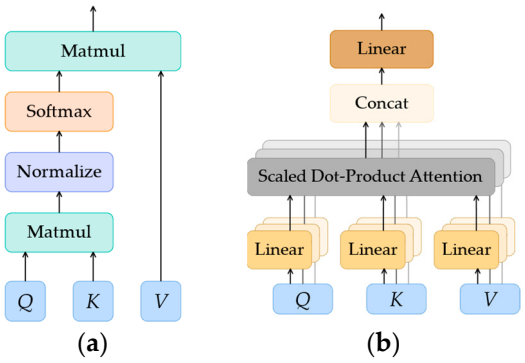


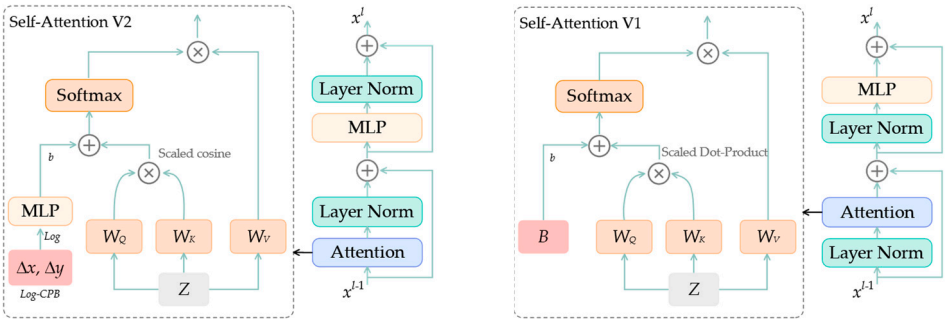
Figure 3. (**a**) The self-attention and (**b**) the multi-head self-attention.

Scaled cosine attention [13] is used in Swin Transformer V2 instead of Scaled Dot-Product attention, as shown in Figure 4. A comparison of self-attention V1 and V2 is shown in Figure 4, where Figure 4 shows the improved structure of self-attention in V2,

and Figure 4 shows the original structure. Scaled cosine attention is proposed to solve a problem that arises in the res-post-norm configuration, namely, that the learned attention maps of some blocks and heads are frequently dominated by a few pixel pairs. Scaled cosine attention can compare the similarity of each element in two vectors, and its calculation is as follows:

$$Scaled\ cosine\ attention(q_i, k_j) = cos(q_i, k_j)/\tau + B_{ij}, \quad (4)$$

where i and j are different pixel coordinate indices, q and k are query and key vectors, respectively, B_{ij} is the relative position bias of pixels at i and j coordinates, and τ is a globally shared learnable scalar. The cosine function is naturally normalized and thus can have milder attention values.



(**a**) (**b**)

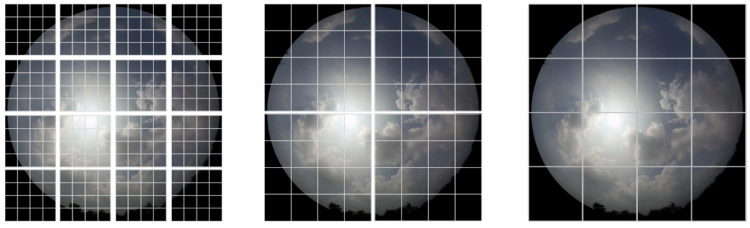
Figure 4. (**a**) The self-attention V2 and (**b**) the self-attention V1.

(2) Improvements to shifted windows multi-head self-attention

The Swin Transformer is proposed to build hierarchical feature maps by merging image patches [16], as shown in Figure 5. Compared with the method of keeping the feature map size invariant in Vit, hierarchical feature maps not only use multi-scale features for modeling, but they can also greatly reduce the complexity of self-attention operations [14]. The model based on Transformer architecture itself cannot implicitly learn the position information of the sequence, and in the process of dividing the picture into multiple patches, the border between the patches has a specific meaning; therefore, the Swin Transformer adds relative position encoding (RPE) to each patch [17], which can compensate for the relative position information between the two elements that is lost when computing self-attention with absolute position encoding. Therefore, the self-attention calculation formula using relative position encoding is as follows:

$$Attention - RPE(Q, K, V) = Softmax\left(\frac{QK^T}{\sqrt{d_k}} + B\right) \cdot V, \quad (5)$$

where $B \in R^{M^2 \times M^2}$, the value in B comes from a smaller-sized bias matrix $\hat{B} \in R^{(2M-1)\times(2M-1)}$, M^2 is the number of patches in a window, and Q, K, V are the query matrix, key matrix, and value matrix, respectively.



×4 Downsample ×8 Downsample ×16 Downsample

Figure 5. Hierarchical feature maps in Swin Transformer.

The Swin Transformer uses the concept of Windows Multi-head Self-Attention (WMSA) to partition the image into multiple non-intersecting windows, and Multi-head Self-Attention is only computed in each individual Window. However, in Vit, Multi-head Self-Attention is directly performed on the global Window. The purpose is to reduce the amount of calculation, but this occurs at the expense of information transmission between different Windows. Therefore, in this paper, we propose using Shifted Windows Multi-head Self-Attention (SW-MSA) to address the shortcomings of WMSA. Through this method, information can be transmitted in adjacent windows. The block based on Windows Multi-head Self-Attention (W-MSA) and the block based on Shifted Windows Multi-head Self-Attention (SW-MSA) constitute the core components of the Swin Transformer. W-MSA refers to the mutual calculation of self-attention between patches in a Window with a specified size. Compared with MSA in Vit, W-MSA can effectively save computing resources and improve model computing efficiency. Equations (6) and (7) represent the computational complexity of MSA and W-MSA, respectively:

$$\Omega(MSA) = 4hwC^2 + 2(hw)^2C, \tag{6}$$

$$\Omega(W-MSA) = 4hwC^2 + 2M^2hwC, \tag{7}$$

where h, w, and C represent the height, width, and channel of an image, respectively, M denotes that each window contains $M \times M$ patches. The former is quadratic to patch number $h \times w$, and the latter is linear when M is fixed. The difference between the two computational complexities increases as the input image size increases. SW-MSA implements communication between windows through the Shift window, Masked MSA, and Reverse shift, as detailed in Figure 6.

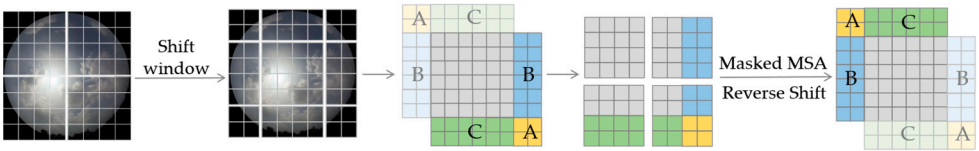


Figure 6. Illustration of the self-attention in shifted window partitioning.

To improve the stability of the model and the performance loss caused by the SW-MSA process, Swin Transformer V2 uses log-spaced continuous position bias to solve the above problem. The introduction of a log-spaced continuous position bias approach guarantees that the relative position bias can be smoothly transferred across window resolutions. The continuous position bias method uses a small meta-network (e.g., an MLP network) in relative coordinates to optimize the bias parameters, which is different than directly optimizing the parameters in a traditional network:

$$B(\Delta x, \Delta y) = \mathcal{G}(\Delta x, \Delta y), \tag{8}$$

where $\mathcal{G}$ is a small network that can be designed artificially. Then, to alleviate the computational problem of needing to calculate relative position coordinates when converting between windows and the problem of inconsistent relative position coordinate ranges for windows of different sizes,

$$\begin{aligned} \widehat{\Delta x} &= sign(x) \cdot log(1 + |\Delta x|), \\ \widehat{\Delta y} &= sign(y) \cdot log(1 + |\Delta y|), \end{aligned} \tag{9}$$

where Δx, Δy and $\widehat{\Delta x}$, $\widehat{\Delta y}$ are the linear-scaled and log-spaced coordinates, respectively.

2.3.2. Feature Map Visualization

To illustrate the feature extraction ability of the proposed method more intuitively, we used the Gradient-weighted Class Activation Mapping++ (Grad-CAM++) [18] method for feature visualization. The method shows the crucial regions of the image predicted by generating a rough attention map from the chosen layer of the model. For our work, we used the last norm layer of the last block of the Swin Transformer as the chosen layer. The redder the color of the attention map, the higher the importance of the corresponding region of the image. The visualization results are shown in Figure 7. The proposed method can focus on the critical parts of the cloud image, and for images with fewer clouds, the model expands the area of focus (e.g., the third image from left to right). The model will focus on the overall features of images with dense clouds and no apparent features.

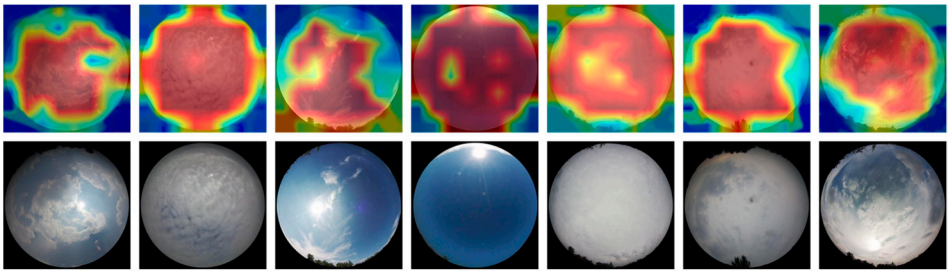


Figure 7. Grad-CAM++ visualization results in the MGCD dataset.

2.4. Multi-Modal Information Network and Feature Fusion Network

In the Multi-modal Information Network and the Feature Fusion Network, we utilize the Multi Information Block (MI Block) as the main component. The MI Block consists of Linear, Batch Norm, and GELU, and the Dropout layer is used between MI Blocks to prevent overfitting; the default setting of the Dropout ratio is 0.5. Regarding the choice of activation function, we use the Gaussian Error Linear Unit (GELU) [19]. In the modeling process of neural networks, the fundamental property of the model is non-linearity. At the same time, it is necessary to include the stochastic regular for the model generalization ability. The non-linear activation and the stochastic regular determine the model's input. In the activation, the *GELU* introduces the idea of the stochastic regular, a probabilistic description of the input of the neuron, which is calculated as follows:

$$GELU(x) = 0.5x\left(1 + tanh\left(\sqrt{2/\pi}(x + 0.044715x^3)\right)\right), \quad (10)$$

In addition to the regularization operations employed above, MMST uses residual units between certain MI Blocks. This idea is borrowed from the Deep crossing model [20], where the use of residual units in the designed network structure is likely to perform regularization, ensuring the stability of the model implicitly.

2.5. Implementation Details

We downscaled the ground-based cloud image from 1024 × 1024 to 256 × 256 and imported it into MMST at this size. Then, to perform data augmentation on the image, we used random horizontal and vertical flips with a probability of 50%. Subsequently, each image was normalized according to the mean and variance of ImageNet. On the other hand, to ensure data matching, we normalized the values of multi-modal information with normal distribution.

The experimental platform comprised a server containing an Intel(R) Core(TM) i9-9900 K 3.60 GHz CPU, an NVIDIA GTX 2080 Ti, and a memory of 32 GB. The operating system was Windows 10 Professional Edition. The software environment was Python3.10 in Pytorch 1.11.

In terms of models, the main network was initialized by the pre-trained Swin Transformer V2 on the ImageNet dataset, and we fine-tuned it on MGCD. For the parameters

of the linear layer in the Multi-modal Information Network and Feature Fusion Network, we used Kaiming Initialization [21], which can solve the problem that Xavier Initialization [22] is only applicable to linear activation functions and guarantees the stability of the gradient to a certain extent. Adding Batch Norm after each linear layer and using Dropout between MI Blocks can effectively prevent the model from overfitting. During the training phase, we used the NAdam optimizer to update the parameters of the network. NAdam adds the accumulation of Nesterov momentum on the basis of Adam. NAdam has more substantial constraints on the learning rate and has a more direct impact on the updating of the gradient. The initial learning rate was set to 5×10^{-6}. Limited by the experimental equipment, the total number of training iterations was set to 30 and the batch size was set to 16. The computation complexity of MMST was 15.246 GFLOPs and the number of MMST's parameters was about 74.104 M. The total training time was about 2.5 h. The loss function was Cross Entropy Loss, which is calculated as follows:

$$Loss = -\sum_{c=1}^{C} w_c \log \frac{\exp(x_{n,c})}{\sum_{i=1}^{C} \exp(x_{n,i})} y_{n,c}, \tag{11}$$

where x is the input, y is the target, w is the weight, and C is the number of classes. *Accuracy* is calculated as follows:

$$Accuracy = \frac{n_{\text{correct}}}{n_{\text{total}}}, \tag{12}$$

where n_{correct} is the number of correctly classified samples, and n_{total} is the number of all samples in the test dataset.

3. Data Collection

The Multi-modal Ground-based Cloud Image Dataset (MGCD) combines meteorological cloud images and corresponding multi-modal information. MGCD contains 8000 ground-based cloud samples, and each sample includes a cloud image with a resolution of 1024 × 1024 and a set of multi-modal information. The cloud image is collected by a fisheye lens sky camera, which can provide observations of a wide range of sky conditions at 180° horizontal and vertical angles. Multi-modal information includes temperature (°C), humidity (%RH), pressure (hPa), and wind speed (m/s). According to the genus-based classification recommendations of the World Meteorological Organization (WMO), the collected ground-based cloud images are divided into seven categories: (1) Cumulus, denoted as Cu; (2) Altocumulus and Cirrocumulus, denoted as Al-Ci; (3) Cirrus and Cirrostratus, denoted as Ci-Ci; (4) Clear sky, denoted as Cs; (5) Stratocumulus and Stratus and Altostratus, denoted as St-St-Al; (6) cumulonimbus and nimbostratus, denoted as Cu-Ni; and (7) mixed cloud, denoted as Mc. Note that mixed clouds are meteorological clouds in which the sky is usually covered by no less than two types of clouds, and clear skies are cloud images with no more than 10% cloud volume. Figure 8 shows an example of each type of cloud and the corresponding multi-modal information in the MGCD.

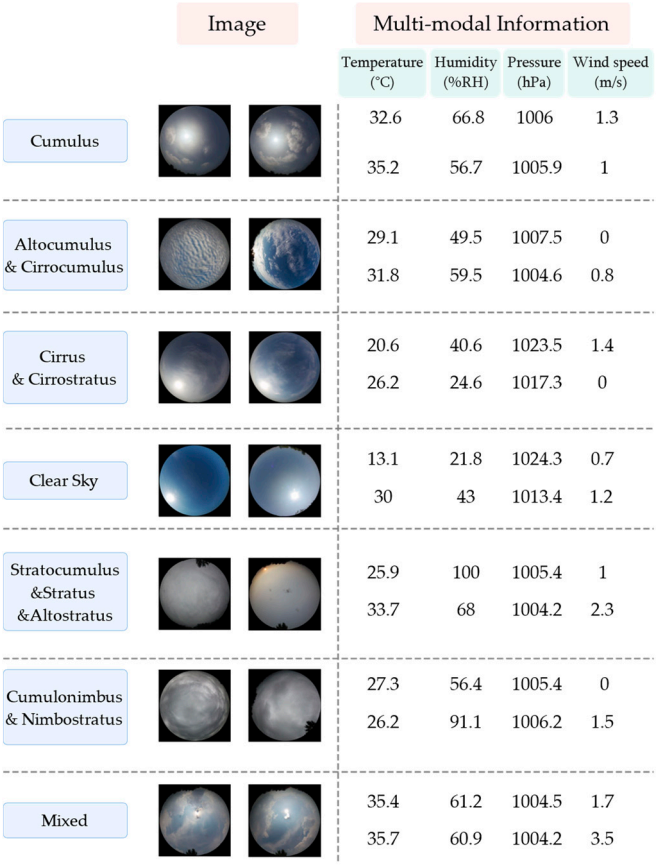


Figure 8. Some samples from the MGCD dataset.

4. Results

In this section, we compared the classification performance of variants of MMST, hand-crafted, and learning-based methods, and other classical classification methods on MGCD, verifying the effectiveness of the proposed MMST classification.

4.1. Comparison with Variants of MMST

The advantage of the proposed MMST model is that it can combine ground-based cloud images and corresponding meteorological information to improve the classification accuracy further. To demonstrate their effectiveness on the MGCD dataset, we listed several variants of MMST. The structures of variant 1 through variant 4 are shown in Figure 9.

Variant 1. Variant 1 is a version of MMST, called MMST-small, with a smaller number of parameters (MMST is also called MMST-base). Compared with the C of the Swin Transformer Block in MMST-base, which is 128, MMST-small sets C to 96, the output layer uses fewer neurons (1024), and the feature fusion method uses **Add**. To further reduce the number of MMST parameters, we reduced the number of MI Blocks from 9 to 7.

Variant 2. The structure of variant 2 is basically the same as that of MMST-base. The difference is that the **Concat** operation of Feature Fusion is changed to the **Add** operation to verify that for MMST, **Add** or **Concat** is more effective for feature fusion.

Variant 3. Variant 3 is based on Variant 2, using **Add** to fuse ground-based cloud images and multi-modal meteorological features. The difference is that the MI Block residual connection is removed, and the rest of the structure remains unchanged to verify the impact of the residual connection on deep fusion.

Variant 4. Variant 4 is based on MMST-base and uses **Concat** to fuse ground-based cloud images and multi-modal meteorological features. The difference is that the MI Block residual connection is removed, and the rest of the structure remains unchanged to verify the impact of the residual connection on deep fusion.

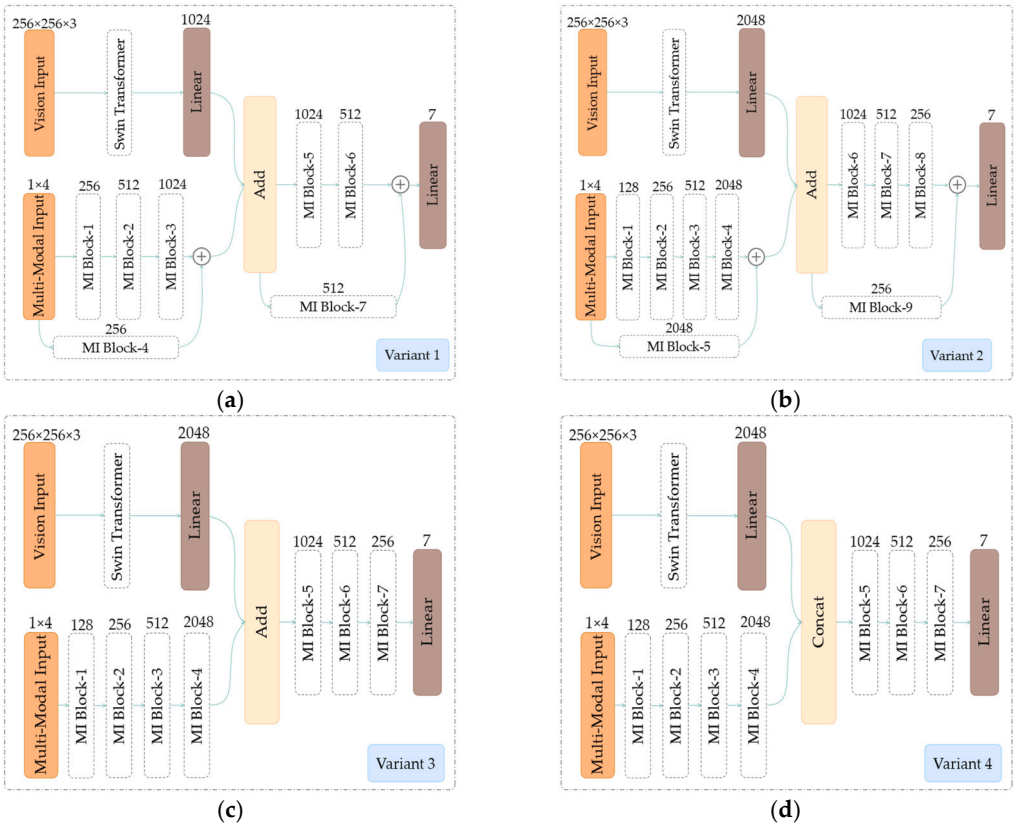


Figure 9. The four different variants of MMST. (**a**) Variant 1. (**b**) Variant 2. (**c**) Variant 3. (**d**) Variant 4.

Figure 10 gives the classification accuracy of MMST and its different variants on the MGCD dataset with both inputs of the ground-based cloud image and multi-modal information. The MMST model achieved the highest accuracy at 91.30%.

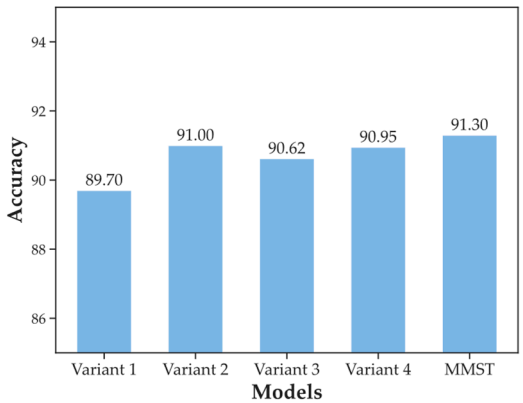


Figure 10. Comparison of accuracy of variants of MMST classification results.

To further analyze the performance of the proposed MMST model, its confusion matrix is shown in Figure 11, and the Recalls and F1-scores of the proposed MMST for different classes of clouds are listed in Table 1.

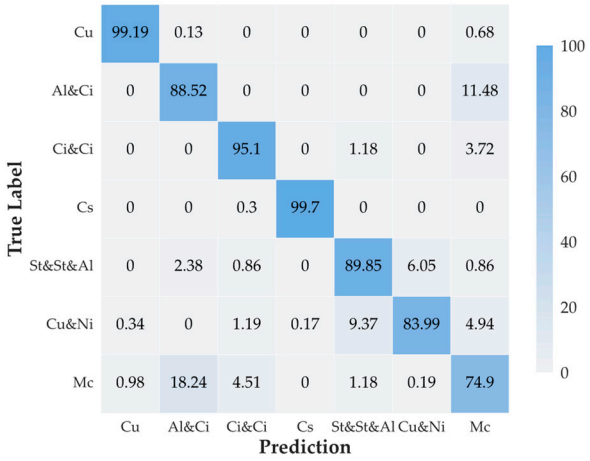


Figure 11. Confusion matrix for MMST classification results.

Table 1. Evaluation metrics for MMST classification results.

Class	Precision (%)	Recall (%)	F1-Score (%)
Cu	99.13	99.62	99.03
Al-Ci	73.95	89.23	80.27
Ci-Ci	95.78	95.49	95.91
Cs	100	100	100
St-St-Al	86.17	90.55	88.17
Cu-Ni	95.49	84.13	89.04
Mc	79.32	75.28	77.51

4.2. Comparison with Hand-Crafted Methods

In this section, we used local binary patterns (LBP) [23] and completed LBP (CLBP) [24], bag-of-visual-words (BoVW) [25], and pyramid BoVW (PBoVW) [26] to conduct experiments on the MGCD dataset to explore the performance of the hand-crafted methods. In the experiments of this subsection, the values of (P, R) of LBP were set to (8, 1), (16, 2), and (24, 3), respectively. Table 2 illustrates the classification results of these methods on the MGCD dataset. The proposed MMST achieved the greatest performance with vision inputs or vision+MI inputs among these classification models.

Table 2. Traditional machine learning classification results.

Class	Vision Input				Vision + MI Input			
	Accuracy (%)	Precision (%)	Recall (%)	F1-Score (%)	Accuracy (%)	Precision (%)	Recall (%)	F1-Score (%)
BoVW	66.15	62.80	66.95	63.94	67.20	66.19	67.91	66.60
PBoVW	66.13	63.53	65.51	64.54	67.15	67.00	65.85	65.23
$LBP^{riu2}_{8,1}$	45.38	44.33	45.94	44.99	45.25	46.22	45.07	45.65
$LBP^{riu2}_{16,2}$	49.00	49.27	51.34	49.85	47.25	49.53	51.58	50.13
$LBP^{riu2}_{24,3}$	50.20	49.55	52.96	50.08	50.53	46.94	49.31	47.11
$CLBP^{riu2}_{8,1}$	65.10	64.45	65.39	64.32	65.40	65.12	65.57	65.07
$CLBP^{riu2}_{16,2}$	68.20	67.88	67.47	67.78	68.48	69.19	68.18	68.68
$CLBP^{riu2}_{24,3}$	69.18	70.71	66.20	68.73	69.68	69.92	71.67	70.50
MMST	**88.22**	**86.87**	**87.48**	**86.79**	**91.30**	**89.86**	**90.17**	**89.17**

4.3. Comparison with Other Deep Learning Methods

Since MMST is a neural network model trained with an end-to-end architecture, we compare MMST with other deep-learning methods (e.g., VGG16 [27], ResNet50 [28], DMF

[29], DCAFs [30], CloudNet [6], JFCNN [31], DTFN [32], MMFN [12], HMF [33], Vit-base [9]) in both input cases. The choice of deep learning architecture will also affect the results [34], so the models for comparative experiments in this section include CNN and Transformer architectures. The results are shown in Figure 12.

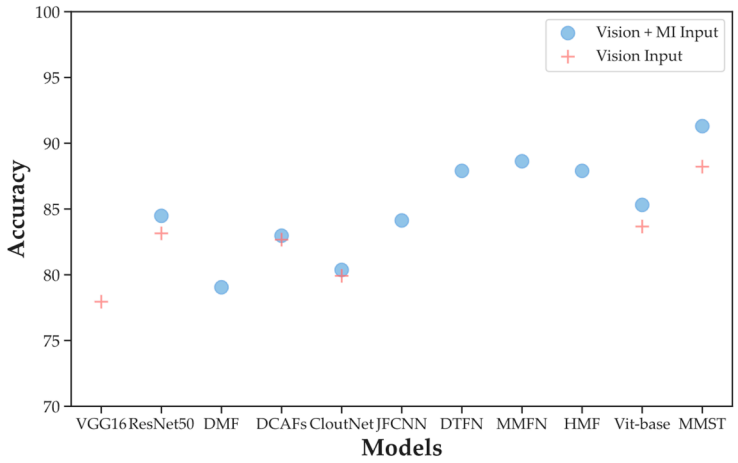


Figure 12. Comparison with other classification methods in two input cases.

5. Discussion

5.1. Analyses of the Experiments with Variants of MMST

Comparing the performance of the proposed MMST and variant 2, as shown in Figure 10, it shows that using the Swin Transformer as the visual backbone network and using the residual structure to fuse multi-modal features have positive effects. The difference between MMST and variant 2 is whether the feature fusion layer uses **Add** or **Concat**. The experimental results showed that using the **Concat** method can retain information to the greatest extent, but this is at the cost of increasing the number of calculations, while using the **Add** method is a special form of **Concat**. That is, the dimension of the image description itself does not increase, but the increasing amount of information under each dimension reduces the computational effort; however, the disadvantage is that the feature information will be lost. For the MGCD dataset, it is more advantageous to use **Concat** to fuse the modal information. Secondly, variant 1 has fewer parameters, which means a faster calculation speed. Compared with MMST and other variants, variant 1 has the lowest classification accuracy.

Finally, comparing variants 2, 3, and 4 and the proposed MMST, we discovered that one of the differences is in the use of the residual structure. As seen from the results, we found that the model using the residual structure was more effective than the model without the residual structure. Specifically, variant 2 classification accuracy was 0.38% higher than that of variant 3, and MMST was 0.35% higher than that of variant 4. In terms of the information fusion method, variant 2 and variant 3 used the **Add** method, and variant 4 and MMST used the **Concat** method. The results showed that for the same architecture, the model performance was improved by simply replacing **Add** with **Concat**, with MMST achieving an accuracy 0.30% higher than that of variant 2, and variant 4 achieving an accuracy 0.33% higher than that of variant 3.

Combining Figure 11 and Table 1, we can see that the accuracy of Cumulus (Cu), Cirrus and Cirrostratus (Ci-Ci), and Clear sky (Cs) reached 99.19%, 95.10%, and 99.7%, respectively. It is clear from combining the datasets that these three categories of cloud images had clear contours and distinct features relative to other categories. Recall and F1-score both reach above 95%. In contrast, Stratocumulus and Stratus and Altostratus (St-St-Al) and Cumulonimbus and Nimbostratus (Cu-Ni) had no obvious cloud features, and the overall color of the cloud image was grayish, which made it more challenging to distinguish the clouds from the sky. The worst classification effect of cloud type was of mixed

cloud (Mc), with Precision, Recall, and F1-score results of only 79.32%, 75.28%, and 77.51%, respectively. The existence of at least two types of clouds within the image increased the difficulty of model learning. Observing the cloud map, we could see that the distribution and shape of mixed cloud (Mc) and Altocumulus and Cirrostratus (Al-Ci) clouds are more similar, so the model can easily misclassify them as Altocumulus and Cirrostratus (Al-Ci).

5.2. Analyses of the Experiments with Hand-Crafted Methods

By comparing the classification effects of Vision Input and Vision + MI Input, as shown in Table 2, the BoVW and PBoVW models based on the bag-of-words model were limited by the size of the lexicon as well as the dimensionality; even though the features of the ground-based cloud map were extracted, they could not describe the complete information of the cloud image well. After adding multi-modal information, it had richer features, so the accuracy increased to 67.20%. PBoVW was similar to BoVW, except that it incorporated a pyramidal hierarchical feature extraction technique, which was even less effective than BoVW. LBP had the highest accuracy of only 50.53% (with the combination of parameters (P, R) of (24, 3)) because of its features, such as rotation invariance and gray-scale invariance, but it was mainly used to describe local texture features. According to the experimental results, CLBP had a considerable improvement compared to LBP, and the classification accuracy reached 69.68% under the setting of $P = 24$, $R = 3$ with the image and multi-modal information jointly input into CBLP, which was 2.53% higher than PBoVW (Vision + MI Input) and 19.15% higher than $\mathrm{LBP}_{24,3}^{riu2}$ (Vision + MI Input); however, the accuracy of this classification method never exceeded 70%. It is evident that the classification method based on manual feature extraction has certain constraints, which are not only limited by the difficulty of designing features but also depend on the effectiveness of the extracted features. Compared with the MMST model proposed in this paper, even the best performing $\mathrm{CLBP}_{24,3}^{riu2}$ was 19.04% and 21.62% lower than MMST in Vision Input and Vision + MI Input cases, respectively, which shows that the performance ceiling of the machine learning-based classification algorithm is far from the ceiling of the deep learning model.

5.3. Analyses of the Experiments with Other Deep Learning Methods

From Figure 12, we can outline the points as follows. Firstly, multi-modal features are complementary to visual features, and their combination can improve the performance of single visual features as input. Secondly, the CNN-based methods, such as CloudNet, JFCNN, DTFN, and so on, are much better than the hand-crafted methods, and the classification accuracies are all higher than 75%. This is attributed to the highly nonlinear transformation nature of CNNs, which enables them to extract effective features from highly complex cloud data. Thirdly, although better than the classical CNN model ResNet50 (0.83% higher), the Vit-base model is less effective than the one designed for MGCD. Owing to the lack of local modeling capability, the results are not as good as the well-designed CNN models for MGCD, such as DTFN, MMFN, and HMF. Fourthly, the proposed MMST improves the accuracy from 90% to 91.30%. Moreover, it works well in both Vision Input and Vision + MI Input, and the classification accuracy in the case of Vision Input is 88.22%, which verifies the effectiveness of the MMST model.

In addition to the classification-model-based supervised learning discussed in this section, in the development process of deep learning, relevant researchers also proposed multi-task learning and weak supervised learning. Multi-task learning is usually used to deal with different tasks through multiple different models and different loss functions for input data. However, for ground-based cloud image classification tasks, rich feature information can already be obtained from cloud images and multimodal data, so obtaining auxiliary information from multi-task learning is very limited. On the other hand, weak supervised learning only labels part of the data, and the rest of the data depends on the model for reasoning. However, the annotation of ground-based cloud maps requires

a lot of professional knowledge and the complexity of cloud movement increases the difficulty of weak supervised learning.

6. Conclusions

In this paper, we introduced a novel multi-modal ground-based cloud map recognition method called MMST. The proposed MMST uses only the attention mechanism and linear layers to extract cloud maps and multi-modal features. Since the Transformer architecture lacks CNN-like prior knowledge, the improved Swin Transformer was used as the visual backbone network, where W-MSA and SW-MSA replace the local modeling in the CNN and use improved Scaled cosine attention, residual post normalization, and log-spaced continuous position bias to further promote the model representation. In addition, we incorporated the residual structure into the visual and multi-modal information fusion to ensure the stability of the training process and the adequacy of information fusion. We performed validation on the MGCD and the results show that the proposed MMST is comparable to state-of-the-art methods.

In future work, the following four processes can be considered to improve the proposed model:

1. Collect a ground-based cloud image dataset with a considerably larger amount of data.
2. Obtain more multi-modal information combined with image information to improve classification accuracy.
3. Improve the image coding methods and local modeling capabilities to enable the model of the Transformer architecture to gradually surpass or even replace CNN in the visual field. At the same time, the ability of the model to distinguish mixed cloud layers should be improved.
4. Scaling images before entering the network results in a loss of information. In theory, the Swin Transformer can process input data of any length (that is, images of any size). Future work can be directed toward processing high-resolution ground-based cloud images.

Author Contributions: Conceptualization, L.W. and T.Z.; methodology, L.W. and Y.G.; software, L.W.; validation, L.W., T.Z. and Y.G.; formal analysis, T.Z.; writing—original draft preparation, L.W.; writing—review and editing, T.Z. and C.N.; funding acquisition, T.Z. and Y.G. All authors have read and agreed to the published version of the manuscript.

Funding: This research was funded by National Natural Science Foundation of China, grant number 62006120, and it was funded by the Graduate Research Practice Innovation Plan of Jiangsu in 2021, grant number KYCX21-0878.

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: The MGCD dataset is available at https://github.com/shuangliut-jnu/Multimodal-Ground-based-Cloud-Database, 17 February 2023.

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

References

1. Zhu, T.; Wei, L.; Guo, Y. Cloud Classification of Ground-Based Cloud Images Based on Convolutional Neural Network. *J. Phys. Conf. Ser.* **2021**, *2035*, 012020. https://doi.org/10.1088/1742-6596/2035/1/012020.
2. Liu, L.; Sun, X.; Chen, F.; Zhao, S.; Gao, T. Cloud Classification Based on Structure Features of Infrared Images. *J. Atmos. Ocean. Technol.* **2011**, *28*, 410–417. https://doi.org/10.1175/2010JTECHA1385.1.
3. Heinle, A.; Macke, A.; Srivastav, A. Automatic Cloud Classification of Whole Sky Images. *Atmos. Meas. Tech.* **2010**, *3*, 557–567. https://doi.org/10.5194/amt-3-557-2010.
4. Oikonomou, S.; Kazantzidis, A.; Economou, G.; Fotopoulos, S. A Local Binary Pattern Classification Approach for Cloud Types Derived from All-Sky Imagers. *Int. J. Remote Sens.* **2019**, *40*, 2667–2682. https://doi.org/10.1080/01431161.2018.1530807.

5. Xiao, Y.; Cao, Z.; Zhuo, W.; Ye, L.; Zhu, L. MCLOUD: A Multiview Visual Feature Extraction Mechanism for Ground-Based Cloud Image Categorization. *J. Atmos. Ocean. Technol.* **2016**, *33*, 789–801. https://doi.org/10.1175/JTECH-D-15-0015.1.
6. Zhang, J.; Liu, P.; Zhang, F.; Song, Q. CloudNet: Ground-Based Cloud Classification with Deep Convolutional Neural Network. *Geophys. Res. Lett.* **2018**, *45*, 8665–8672. https://doi.org/10.1029/2018GL077787.
7. Liu, S.; Li, M.; Zhang, Z.; Cao, X.; Durrani, T.S. Ground-Based Cloud Classification Using Task-Based Graph Convolutional Network. *Geophys. Res. Lett.* **2020**, *47*, e2020GL087338. https://doi.org/10.1029/2020GL087338.
8. Vaswani, A.; Shazeer, N.; Parmar, N.; Uszkoreit, J.; Jones, L.; Gomez, A.N.; Kaiser, Ł.; Polosukhin, I. Attention Is All You Need. In *Proceedings of the Advances in Neural Information Processing Systems*; Curran Associates, Inc.: Red Hook, NY, USA, 2017; Volume 30.
9. Dosovitskiy, A.; Beyer, L.; Kolesnikov, A.; Weissenborn, D.; Zhai, X.; Unterthiner, T.; Dehghani, M.; Minderer, M.; Heigold, G.; Gelly, S.; et al. An Image Is Worth 16x16 Words: Transformers for Image Recognition at Scale. *arXiv* **2020**, arXiv:2010.11929.
10. Li, X.; Qiu, B.; Cao, G.; Wu, C.; Zhang, L. A Novel Method for Ground-Based Cloud Image Classification Using Transformer. *Remote Sens.* **2022**, *14*, 3978. https://doi.org/10.3390/rs14163978.
11. Farmer, D.K.; Cappa, C.D.; Kreidenweis, S.M. Atmospheric Processes and Their Controlling Influence on Cloud Condensation Nuclei Activity. *Chem. Rev.* **2015**, *115*, 4199–4217. https://doi.org/10.1021/cr5006292.
12. Liu, S.; Li, M.; Zhang, Z.; Xiao, B.; Durrani, T.S. Multi-Evidence and Multi-Modal Fusion Network for Ground-Based Cloud Recognition. *Remote Sens.* **2020**, *12*, 464. https://doi.org/10.3390/rs12030464.
13. Liu, Z.; Hu, H.; Lin, Y.; Yao, Z.; Xie, Z.; Wei, Y.; Ning, J.; Cao, Y.; Zhang, Z.; Dong, L.; et al. Swin Transformer V2: Scaling Up Capacity and Resolution. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, New Orleans, LA, USA, 18–24 June 2022; pp. 12009–12019.
14. Liu, Z.; Lin, Y.; Cao, Y.; Hu, H.; Wei, Y.; Zhang, Z.; Lin, S.; Guo, B. Swin Transformer: Hierarchical Vision Transformer Using Shifted Windows. In Proceedings of the 2021 IEEE/CVF International Conference on Computer Vision (ICCV), Montreal, QC, Canada, 10–17 October 2021; pp. 10012–10022.
15. Zheng, Z.; Zhao, Y.; Li, A.; Yu, Q. Wild Terrestrial Animal Re-Identification Based on an Improved Locally Aware Transformer with a Cross-Attention Mechanism. *Animals* **2022**, *12*, 3503. https://doi.org/10.3390/ani12243503.
16. Li, A.; Zhao, Y.; Zheng, Z. Novel Recursive BiFPN Combining with Swin Transformer for Wildland Fire Smoke Detection. *Forests* **2022**, *13*, 2032. https://doi.org/10.3390/f13122032.
17. Shaw, P.; Uszkoreit, J.; Vaswani, A. Self-Attention with Relative Position Representations. *arXiv* **2018**, arXiv:1803.02155.
18. Chattopadhay, A.; Sarkar, A.; Howlader, P.; Balasubramanian, V.N. Grad-CAM++: Generalized Gradient-Based Visual Explanations for Deep Convolutional Networks. In Proceedings of the 2018 IEEE Winter Conference on Applications of Computer Vision (WACV), Lake Tahoe, NV, USA, 12–15 March 2018; pp. 839–847.
19. Hendrycks, D.; Gimpel, K. Gaussian Error Linear Units (GELUs). *arXiv* **2020**, arXiv:1606.08415.
20. Shan, Y.; Hoens, T.R.; Jiao, J.; Wang, H.; Yu, D.; Mao, J. Deep Crossing: Web-Scale Modeling without Manually Crafted Combinatorial Features. In Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, San Francisco, CA, USA, 13 August 2016; Association for Computing Machinery: New York, NY, USA, 2016; pp. 255–262.
21. He, K.; Zhang, X.; Ren, S.; Sun, J. Delving Deep into Rectifiers: Surpassing Human-Level Performance on ImageNet Classification.*arXiv* **2015**, arXiv:1502.01852; pp. 1026–1034.
22. Glorot, X.; Bengio, Y. Understanding the Difficulty of Training Deep Feedforward Neural Networks. In Proceedings of the Thirteenth International Conference on Artificial Intelligence and Statistics; JMLR Workshop and Conference Proceedings, Sardinia, Italy, 31 March 2010; pp. 249–256.
23. Ojala, T.; Pietikainen, M.; Maenpaa, T. Multiresolution Gray-Scale and Rotation Invariant Texture Classification with Local Binary Patterns. *IEEE Trans. Pattern Anal. Mach. Intell.* **2002**, *24*, 971–987. https://doi.org/10.1109/TPAMI.2002.1017623.
24. Guo, Z.; Zhang, L.; Zhang, D. A Completed Modeling of Local Binary Pattern Operator for Texture Classification. *IEEE Trans. Image Process.* **2010**, *19*, 1657–1663. https://doi.org/10.1109/TIP.2010.2044957.
25. Csurka, G.; Dance, C.R.; Fan, L.; Willamowski, J.; Bray, C. Visual Categorization with Bags of Keypoints. In Proceedings of the Workshop on Statistical Learning in Computer Vision, ECCV, Prague, Czech Republic, 10−14 May 2004; Volume 1, pp. 1–2.
26. Lazebnik, S.; Schmid, C.; Ponce, J. Beyond Bags of Features: Spatial Pyramid Matching for Recognizing Natural Scene Categories. In *IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'06)*; IEEE: Piscataway, NJ, USA, 2006; Volume 2, pp. 2169–2178.
27. Simonyan, K.; Zisserman, A. Very Deep Convolutional Networks for Large-Scale Image Recognition. *arXiv* **2015**, arXiv:1409.1556.
28. He, K.; Zhang, X.; Ren, S.; Sun, J. Deep Residual Learning for Image Recognition. *arXiv* **2016**, arXiv:1512.03385.
29. Liu, S.; Li, M. Deep Multimodal Fusion for Ground-Based Cloud Classification in Weather Station Networks. *J. Wirel. Com. Netw.* **2018**, *2018*, 48. https://doi.org/10.1186/s13638-018-1062-0.
30. Shi, C.; Wang, C.; Wang, Y.; Xiao, B. Deep Convolutional Activations-Based Features for Ground-Based Cloud Classification. *IEEE Geosci. Remote Sens. Lett.* **2017**, *14*, 816–820. https://doi.org/10.1109/LGRS.2017.2681658.
31. Liu, S.; Li, M.; Zhang, Z.; Xiao, B.; Cao, X. Multimodal Ground-Based Cloud Classification Using Joint Fusion Convolutional Neural Network. *Remote Sens.* **2018**, *10*, 822. https://doi.org/10.3390/rs10060822.
32. Li, M.; Liu, S.; Zhang, Z. Deep Tensor Fusion Network for Multimodal Ground-Based Cloud Classification in Weather Station Networks. *Ad. Hoc Netw.* **2020**, *96*, 101991. https://doi.org/10.1016/j.adhoc.2019.101991.

33. Liu, S.; Duan, L.; Zhang, Z.; Cao, X. Hierarchical Multimodal Fusion for Ground-Based Cloud Classification in Weather Station Networks. *IEEE Access.* **2019**, *7*, 85688–85695. https://doi.org/10.1109/ACCESS.2019.2926092.
34. Amyar, A.; Guo, R.; Cai, X.; Assana, S.; Chow, K.; Rodriguez, J.; Yankama, T.; Cirillo, J.; Pierce, P.; Goddu, B.; et al. Impact of Deep Learning Architectures on Accelerated Cardiac T1 Mapping Using MyoMapNet. *NMR Biomed.* **2022**, *35*, e4794. https://doi.org/10.1002/nbm.4794.