Multimodal Information Fusion for Weather Systems and Clouds Identification From Satellite Images

Cong Bai, *Member, IEEE*, Dongxiaoyuan Zhao, Minjing Zhang, and Jinglin Zhang

***Abstract*—Seeing the cloud and then understanding the weather is one of the important means for people to forecast weather. There has been a certain progress in the use of deep learning technology for weather forecasting, especially in the automatic understanding of disaster weather from satellite image, which can be seen as the image classification problem. Publicly available satellite image benchmark database tries to link weather directly with satellite images. However, single image modal is far from enough to correctly identify weather systems and clouds. Thus, we integrate images with meteorological elements, in which five kinds of meteorological elements, such as season, month, date stamp, and geographic longitude, and latitude, are labeled. To effectively use such various modalities for clouds and weather systems identification through satellite image classification tasks, we propose a new satellite image classification framework: multimodal auxiliary network (MANET). MANET consists of three parts: image feature extraction module based on convolutional neural network, meteorological information feature extraction module based on perceptron, and layer-level multimodal fusion. MANET successfully integrates the multimodal information, including meteorological elements and satellite images. The experimental results show that MANET can achieve better weather systems and clouds and land cover classification results based on satellite images.**

***Index Terms*—Clouds, image classification, meteorology, multimodal.**

I. Introduction

About 75% of global economic losses are due to disastrous weather, and more than 10 000 people die every year due to severe weather [1]. Disastrous weather, including tropical cyclone [2], [3], [4], severe convection [5], [6], [7], and sand storm [8], [9], seriously threaten people's lives and property. Monitoring the formation and development of disastrous weather is the basis for weather forecasting. Cloud plays an important role in weather systems since cloud type, cloud phase, and cloud height [10] profoundly affect the generation and development of weather systems. Remote sensing (RS) image is one of powerful tools to monitor clouds and weather systems. As one kind of RS image, which can get top–down observations of cloud cover and earth surface, satellite images can be used to understand different weather conditions, evaluate their strength and future development trends, and provide all-weather basis for weather forecasts and disaster weather predictions. This article tries to perform monitoring of clouds and weather systems, such as tropical cyclones, extratropical cyclones [11], [12], [13], and other possible disastrous weather [14], through satellite image classification tasks. There are different kinds of classification tasks. From the perspective of different forms of outputs, classification tasks can be divided into single-label classification and multilabel classification. The former task is aiming at finding the most significant label of images, and the latter allows to output multiple correct labels. In terms of describing complex images with multiple objects, multilabel classification is more suitable. Not only label information but also semantic and spatial relationships will be learned by multilabel classification models. On the other side, when we talk about inputs, single-modal and multimodal are two different forms. The former contains only one form of data, image for example, while the latter are data with different ones.

Multimodal classification [15], [16], [17], [18], [19] has became a hot topic recently. Various sensors, such as radar, infrared, and camera, can collect various kinds of data. And each of the above-mentioned kind of data can be seen as a modal. Single-modal learning is aiming at finding a mapping from data to its low-dimensional representation, while multimodal learning can further utilize the complementarity of diversified data and extract more powerful joint features. However, most of the existing research works on RS image classification are still focused on single-modal image classification of ground-base images. Different from the satellite images, the ground-based image is captured by a vision sensor located on the ground. And the related ground-based image classification mostly focuses on single image modal classification. For example, Li et al. [20] propose a cloud image detection method based on SVM to remove thick cloud data for reducing the amount of data to improve the efficiency of the data. But without taking other modal information into consideration, it just focus sub-block cloud image that is used as learning samples of SVM classifier. Zhang et al. [21] propose a ground-based cloud image dataset, consisting of 11 categories under meteorological standards as well as CloudNet for ground-based cloud image classification. Haut et al. [22] present a cloud implementation of a successful

Manuscript received 20 May 2022; revised 23 July 2022; accepted 19 August 2022. Date of publication 29 August 2022; date of current version 8 September 2022. This work was supported in part by National Key Research and Development Program of China under Grant 2018YFE0126100, in part by the Zhejiang Provincial Natural Science Foundation of China under Grant LR21F020002, and in part by the Key Research and Development Program of Jiangsu Province under Grant BE2021093. *(Corresponding author: Jinglin Zhang.)*

Cong Bai, Dongxiaoyuan Zhao, and Minjing Zhang are with the College of Computer Science and Technology, Zhejiang University of Technology, Hangzhou 310023, China, and also with the Key Laboratory of Visual Media Intelligent Processing Technology of Zhejiang Province, Hangzhou 310023, China (e-mail: congbai@zjut.edu.cn; zdxy@zjut.edu.cn; 2111912159@zjut.edu.cn).

Jinglin Zhang is with the School of Control Science and Engineering, Shandong University, Jinan 250061, China (e-mail: jinglin.zhang37@gmail.com).

Digital Object Identifier 10.1109/JSTARS.2022.3202246

technique for hyperspectral image classification: the multinominal logistic regression probabilistic classifier.

In summary, most of research works focus on the single modal ground cloud image classification. Hence, how to understand the weather systems and cloud from satellite image using its multimodal information will be an interesting topic. In this study, large-scale satellite cloud image database for meteorological research (LSCIDMR) [23] is upgraded to a multimodal database named as LSCIDMR database with meteorological element (LSCIDMRME). Different from LSCIDMR, LSCIDMRME not only has the image modal label information but also has season, month, data stamp, geographic longitude, and geographic latitude information. The LSCIDMRME contains 521 950 multimodal label tags that can provide a more complete description of weather information from multiple angles. At the same time, we design a network framework to fuse the characteristic information of multimodalities. The results of comprehensive comparison experiments show that multimodal image classification can achieve better performance than single-modal image classification. The main contributions of this article can be summarized as follow.

1) We upgrade original single-modal dataset LSCIDMR into a multimodal dataset LSCIDMRME, which will be uploaded to the IEEE Dataport for attracting more researchers involving deep learning based meteorological research.
2) LSCIDMRME has total six kind of information: image, season, month, data stamp, geographic longitude, and geographic latitude. The total 104 390 images consist of 414 211 multilabels and 40 625 unique labels. And the label of modal season, month, data stamp, geographic longitude, and geographic latitude is one modal corresponds to one image. That is to say, one image has five multimodal labels. In other words, LSCIDMRME consists 521 950 multimodal labels.
3) Multimodal auxiliary network (MANET) for satellite image classification is proposed to fuse multimodal information. MANET consists of three parts: image feature extraction module (IFEM) based on convolutional neural network (CNN), meteorological information feature extraction module (MIFEM) based on perceptron, and layer-level multimodal fusion. Experimental results show that the proposed MANET can achieve better classification performance than single image modal classification.

The reminder of this article is organized as follows. Section II presents related work on single-modal image classification, multimodal image classification, and RS image classification. Multimodal database LSCIDMRME is detailed in Section III. The proposed MANET is shown in Section IV followed by experimental evaluation in Section V. Finally, Section VI gives conclusion and perspectives.

II. Related Work

A. Single-Modal Image Classification

Image classification is a basic task in computer vision. From the 10-class gray-scale image handwritten digit recognition task performed on MNIST to 10-class cifar10 and 100-class cifar100 tasks, then to the later ImageNet [24], image classification is accompanied by the growth of the dataset. Nowadays, thanks to datasets containing more than 10 million images and more than 20 000 categories, such as ImageNet, the accuracy of image classification has surpassed that of humans. Classical convolutional networks, such as LeNet, AlexNet, GoogleNet, ResNet, and EfficientNet, utilize deep learning to investigate the problems of single-modal image classification. LeNet [25] is a multilayer neural network trained with backpropagation algorithm that is marked as the emergence of CNN. AlexNet increases the depth of the network and adopts dropout algorithm that is well avoids overfitting and significantly improves the accuracy of image classification. GoogleNet [26] successfully increases the depth of model without increasing the complexity of computation. ResNet [27] gets the highest accuracy of image classification by increasing the depth of neural network. EfficientNet [28] systematically study model compression and confirms that careful balance of network depth, width, and resolution can bring better results. Through this observation, they propose a new zoom method: use simple and efficient composite coefficients to uniformly zoom all dimensions, including depth, width, and resolution. However, the success of those models mentioned above has just improve the accuracy of image classification, none of them takes the advantages of the mutual enhancing between different modalities.

B. Multimodal Image Classification

Single-modal learning was aimed at learning a high-level representation of images, while multimodal learning attempts to extract complementary information of diversified forms of data. According to the classification tasks of different multimodal datasets, different multimodal fusion classification algorithms are designed. Camps-Valls et al. [29] contrive to use a cross-kernel function to map two modalities datasets into the same feature space. The versatility of classification has been improved after using this method. Couprie et al. [30] treat the multimodal data as multichannel input data into the CNN. The multichannel input method probably interferes with the classification process. Wang et al. [31] propose a train structure that can train two modalities, respectively, and input the result to two fully-connected layers. Wang et al. [32] concatenate the activation in the joint loss function to establish the correlation between the two different modals. In summary, there are many different multimodal tasks for classification, detection, segmentation, etc. Thus, we also propose a classification framework for this task focusing on multimodal classification task in meteorology research.

C. RS Image Classification

From the aspect of classification granularity, pixel-level classification (PLC) and image-level classification (ILC) are required by different applications in RS field. The target of PLC is to generate a classification map of the given images. In other words, PLC task is designed to find the corresponding category for every pixel in given images. Some PLC benchmarks, such as Houston2013,[1] Houston2018,[2] and CWI [33], are proposed for various purposes. For ILC tasks, labels are annotated at image level, and ILC can further be subdivided into single-label classification and multilabel classification problems. The former is aiming at finding the most significant categories of the given images, while the latter allows multiple correct labels for a single image. Most of existing classification datasets are developed in single-labeled form [34], [35], [36]. However, due to the requirement of describing a complex image with multiple objects, some multilabeled datasets [37], [38] have also been proposed. In terms of the modal of the data, except single-modal image datasets, multimodal benchmarks are also available, such as above-mentioned Houston2013, Houston2018, and BigEarthNet-MM [39] which is the extended version of BigEarthNet [38].

CNN is mainstream solution in RS image classification. Furthermore, many specific methods are developed in consideration of the nature of RS images. To balance performance and efficiency, a lightweight discriminative model [40] and LCNN-BEF [41] have been proposed. What is more, LCNN-BEF considers the validity of both deep and shallow CNNs, and for the same reason, best representation branch model [42] and SCCov [43] have also been brought forward. Inconsistencies in scales of RS images motivated the appearance of SEMSD-Net [44] and SF-CNN [45]. To fully utilize the spectralwise information of hyper-spectral images, HybridSN [46] and mixed-convolution [47] design 3D–2D hybrid CNNs. For narrowing the gap between the amount of annotated data and raw RS data, semisupervised and unsupervised methods, such as GAN-based method MARTA-GAN [48], and Attention-GAN [49], similarity-based auxiliary training method Siamese-CNN [50], and kernel collaborative representation [51] are proposed. To take multimodal inputs, FUSION-FCN [52], deep-shallow [53], and two-branch network [54] have been developed, and modal fusion techniques are studied in detail in [55].

Related works discussed above are for single-label RS image classification, as for multilabel RS image classification, not only the most significant semantic representation but also semantic and spatial relationships between different labels should be learned by the model. Specifically in RS field, some approaches are directly transplanted from general CV filed, using off-the-shelf deep learning tools, such as CNN [56]. However, there are also some methods especially designed for RS images. Two-branch network [57], [58], attention mechanism [59], [60], [61], and GCNs [62] are introduced in RS field to model above-mentioned semantic and spatial relationships between objects in images.

In the formation and development of Weather systems, cloud plays an important role. Cloud and Weather classification via images is of great significance. [63] uses traditional method to extract the feature of satellite cloud imagery. Modern deep learning methods are powerful tools for solving cloud and weather image classification tasks. Li et al. [64] detect and classify clouds with Deep neural networks from the perspective of radiance. Except RS images, ground-based cloud images are also explored in cloud classification [21].

In conclusion, there are limited researches on cloud and weather system classification especially based on satellite images and deep learning. And this article, to some extent, is filling such a research gap.

TABLE I
NUMBER OF IMAGE IN EACH CLASS AND CORRESPONDING RATIO IN LSCIDWS-S AND LSCIDWS-M

Type	LSCIDWS-S	Ratio	LSCIDWS-M	Ratio
Tropical Cyclone	3305	8.14%	3305	3.17%
Extratropical Cyclone	4984	12.27%	4984	4.77%
Frontal Surface	634	1.56%	634	0.61%
Westerly Jet	628	1.55%	628	0.60%
Snow	7631	18.78%	8700	8.33%
Low Water Cloud	1774	4.37%	99312	95.14%
High Ice Cloud	5278	12.99%	96033	91.99%
Vegetation	7831	18.28%	47421	42.43%
Desert	4518	11.12%	59448	56.95%
Ocean	4042	9.95%	93756	89.81%
LabelLess	63765	–	–	–
Total Number	104390	–	414221	–

III. LSCIDMRME: LSCIDMR Database With Meteorological Element Label

LSCIDMR [23] is the first public available large-scale cloud image database for meteorological research. This database has 104 390 images with sizes of 1000*1000 pixels. Two forms of database are available, single-labeled LSCIDMR-S and multilabeled LSCIDMR-M. Table I lists detailed information of LSCIDMR. The ratio of a specific label in LSCIDMR-S equals the number of that label divided by the total number of labels in the database without Labeless. The ratio of a specific label in LSCIDMR-M equals the number of that label divided by the total number of images in the database. Fig. 1 gives two image examples of each categories of LSCIDMR-S. One image is annotated with one label in LSCIDMR-S that could not show the rich information. Thus, LSCIDMR-M have a total 414 211 multiple labels that could provide more information in an image. The second and third columns of Table II give five examples of two different annotation methods.

However, all the labels of LSCIDMR are only from the perspective of image information, which is not enough in recognizing clouds and weather systems. In fact, weather conditions have the essential connection with the geographic location and seasonal information. Hence, such information elements labels are added as follows: Season, Month, Date, Longitude, and Latitude. Adding these geographic information elements will enrich the LSCIDMR from the image to seasonal and geographical information. The motivations of choosing such five elements are as follows.

Season [65]: There are several different stages of climate change in a year that can be generally divided into—spring, summer, autumn, and winter. Fig. 2 shows the statistical analysis of typical weather systems in different seasons. From this figure, we can see tropical cyclones and extratropical cyclones in all seasons, but we observe tropical cyclones mostly in summer

[1] [Online]. Available: https://hyperspectral.ee.uh.edu/?page_id=459

[2] [Online]. Available: https://hyperspectral.ee.uh.edu/?page_id=1075

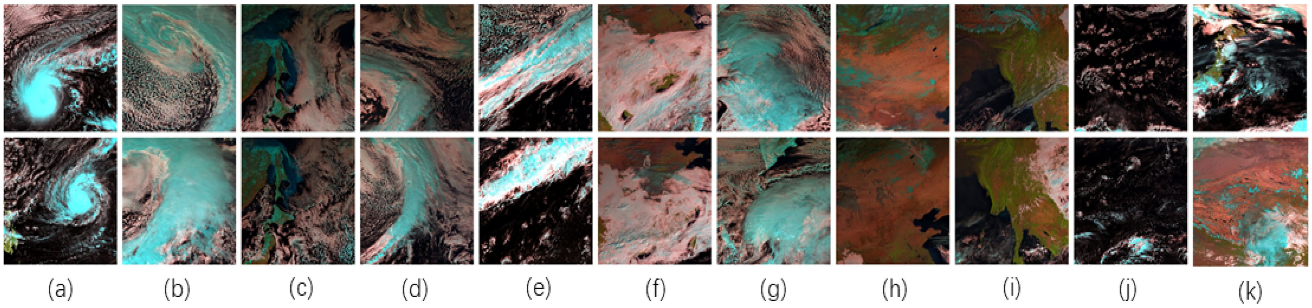


Fig. 1. Images in LSCIDMR-S: selected examples of 11 classes are shown. (a) Tropical cyclone. (b) Extratropical cyclone. (c) Snow. (d) Frontal surface. (e) Westerly jet. (f) Low water cloud. (g) High ice cloud. (h) Desert. (i) Vegetation. (j) Ocean. (k) LabelLess.

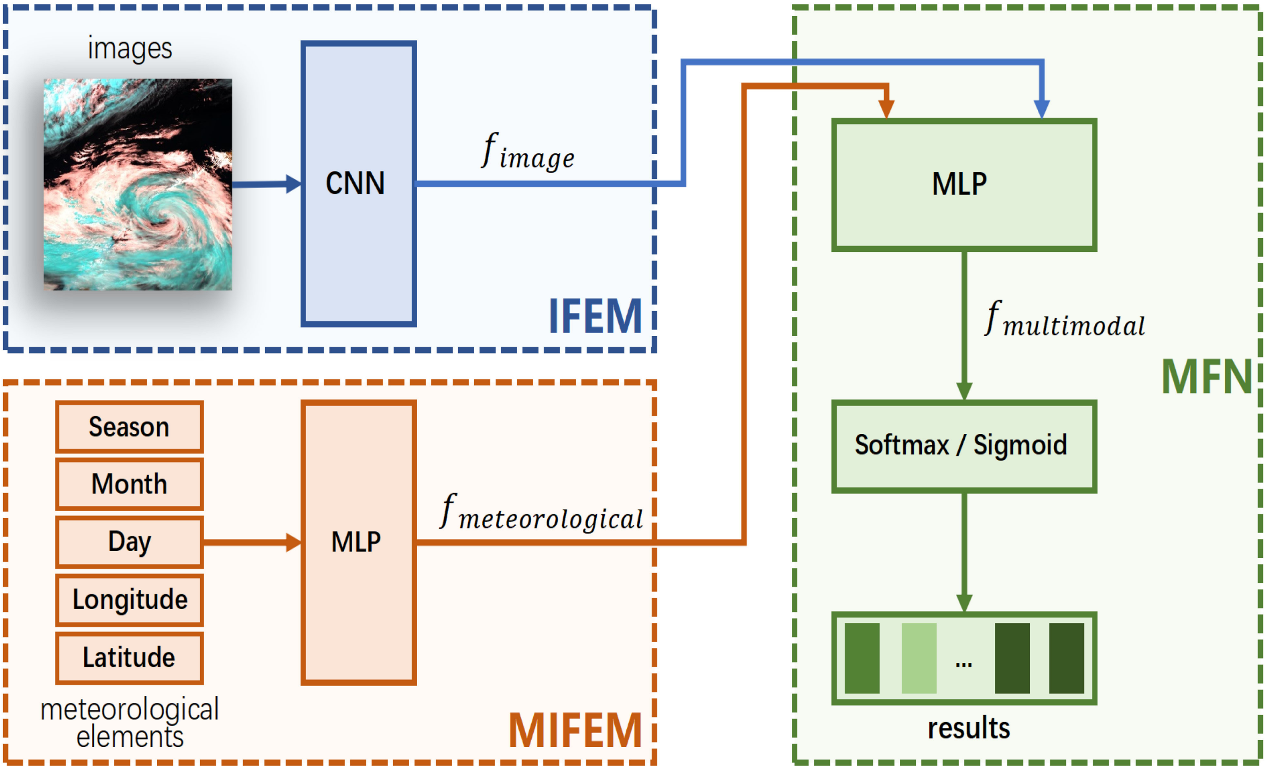


Fig. 2. Statistical analysis of typical weather systems in different seasons.

and autumn. Meanwhile, extratropical cyclones are normally observed in spring and winter. Thus, the season is an important factor for meteorological research.

Month: Different catastrophic weather has different probabilities in the early, middle, and late stages of the same season. Thus, we add the Month as a factor for meteorological research.

Date stamp: A date is a specific time that can be named, for example, a particular day or a particular year. The probability of same type of disastrous weather occurring in the first ten days of the same month, the middle ten days, and the second ten days of the same month is also different. Hence, we take the Data stamp into the consideration.

Longitude and latitude: The probability of different severe weather occurring in different geographical locations is also different as different geographic locations have different geographic characteristics, such as oceans, deserts, vegetation, and etc. All of them have a certain impact on the formation of weather systems. Thus, the geographic information including longitude and latitude is also added in the database.

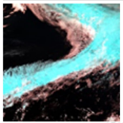
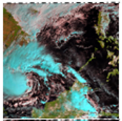
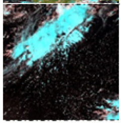
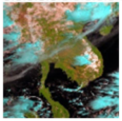
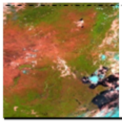
The weather system [66] is a very complex system and many factors must be considered. However, information mentioned above is added as it can be extracted from Himawari-8 satellite directly.

IV. MANET: Multimodal Auxiliary Network for Satellite Image Classification

A. Overview

The structure of our MANET is shown in Fig. 3. And the Algorithm 1 shows the pipeline for training MANET. Our framework

TABLE II
FIVE EXAMPLES OF LSCIDMRME IMAGES

Image	LSCIDWS-S	LSCIDWS-M	Season	Month	Date	Longitude	Latitude
	Frontal Surface	Frontal Surface Low Water Cloud High Ice Cloud Ocean	Summer	May	21	21	05
	Tropical Cyclone	Tropical Cyclone Low Water Cloud High Ice Cloud Ocean, Desert, Vegetation	Summer	May	01	07	11
	LableLess	Low Water Cloud High Ice Cloud Ocean	Autumn	July	22	19	07
	LableLess	Low Water Cloud High Ice Cloud Ocean, Desert, Vegetation	Autumn	July	14	04	10
	LableLess	Low Water Cloud High Ice Cloud Ocean, Desert, Vegetation	Autumn	July	13	07	03

LSCIDMRME includes LSCIDMR-S, LSCIDMR-M, season, month, date, longitude, and latitude. LSCIDMR-S corresponds to the single-label information of the picture, while LSCIDMR-M to the multilabel information. Season, month, date, longitude, and latitude are the corresponding representation of modal information is added to the dataset.

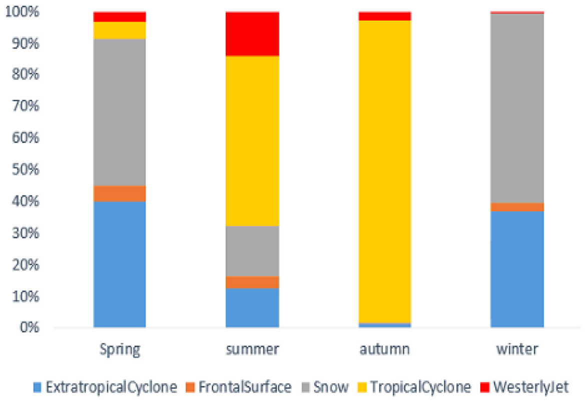


Fig. 3. *Framework of MANET.* There are three main modules in MANET—IFEM, MIFEM, and MFN. In IFEM, images are encoded as semantic representations f_{image} through a CNN; in MIFEM, meteorological features $f_{\text{meteorological}}$ are extracted from meteorological elements through an MLP; then f_{images} and $f_{\text{meteorological}}$ are fused in MFN through another MLP to get joint feature $f_{\text{multimodal}}$; finally, depending on whether the task is multilabeled or single-labeled, $f_{\text{multimodal}}$ are fed into Softmax or Sigmoid function to get final result.

contains three main modules: IFEM, MIFEM, and multimodal fusion based on neural network (MFN). The main purpose of IFEM is to use deep learning method to extract the image representation from images. And MIFEM module contains two main steps. First, it performs nondimensional data processing on each geographic information element and then performs feature extraction on the processed data with a multilayer perceptron. MFCN is a self-designed neural network to fuse the image and meteorological feature information extracted from IFEM and MIFEM, respectively. Following the feature extraction of

Algorithm 1: Pipeline of training MANET: Multimodal auxiliary network for satellite image classification.

```
Input: The training set of image modal, $IX_{\text{train}}$; The
       training set of geographic information modal,
       $GX_{\text{train}}$; The number of model training, $m$;
Output: The classification result of testing set of image
        modal, $IT_{\text{train}}$; Model weights of the entire
        MANET after training $M_{\text{MANET}}$;
1: Initialize CNN with parameters pretrained on
   ImageNet, and initialize other parts with random value;
2: Calculate the loss $L$ based on the network prediction
   result and the real label;
3: Backpropagation updates the model parameter W of
   the entire network framework $M_{\text{MANET}}$;
4: Repeat steps 1–3 of the algorithm until the model
   converges or reaches the maximum number of training
   $m$; return $IT_{\text{train}}$; $M_{\text{MANET}}$;
```

the fused multimodal representation information, multifeature information is classified. These three modules are detailed in the following three sections.

B. Image Feature Extraction Module

The main task of IFEM is to use CNNs to extract high-dimensional features from images. The four kinds of layers in CNNs are: convolution layer, pooling layer, fully connected layer, and activation layer. Convolutional layer is used for image feature extraction. The pooling layer compresses the input feature map. On the one hand, it makes the feature map smaller and

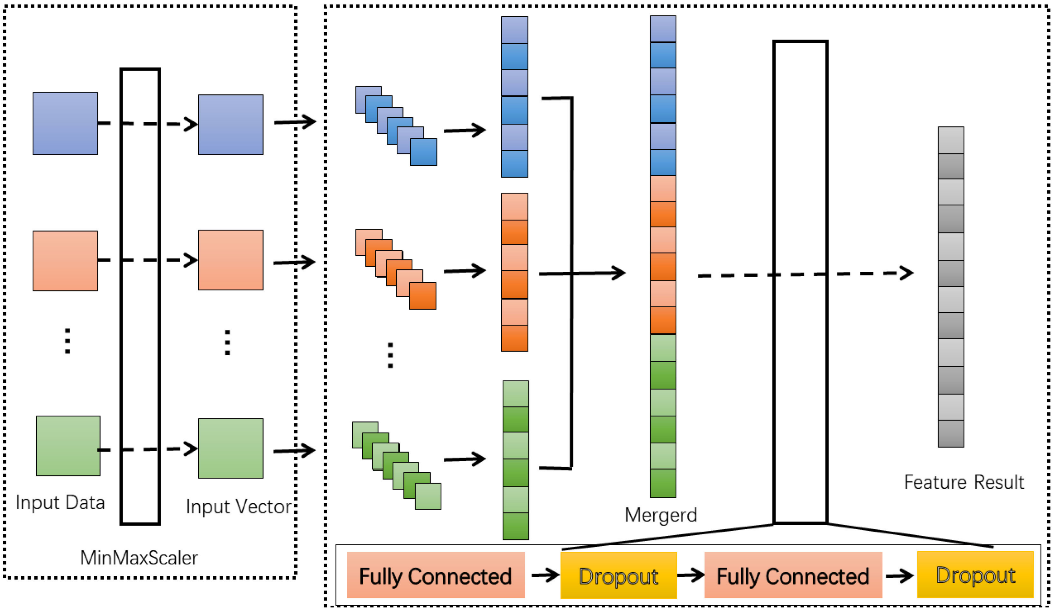


Fig. 4. *Framework of meteorological information process model.* Dimensionless processing is firstly performed on meteorological elements through *min–max scaler*, and processed information is than concatenated and fed into an MLP to get meteorological feature.

simplifies the network calculation complexity; on the other hand, it performs feature compression and extracts main features. The fully connected layer connects all the features and sends the output value to the classifier. The activation function is used to add nonlinear factors, as the expressive power of the linear model is not enough. Most of CNNs are the stacking and improvement of layers mentioned above. Therefore, in our module, we use mainstream image classification networks to complete the feature extraction f_{image} of image modalities.

C. Meteorological Information Feature Extraction Module

MIFEM is used to extract the feature of meteorological information other than images. Fig. 4 is the flowchart of overall processing of this module that mainly includes two parts: data processing and meteorological feature extraction.

In the practice of machine learning algorithms, we often need to convert data of different specifications to the same specification or to convert data from different distributions to a specific distribution. This requirement is collectively referred to as "dimensionless" data. Linear dimensionless [67] includes centering (Zero-centered or Mean-subtraction) processing and scaling processing (Scale). The essence of centralization is to subtract a fixed value from all records, that is, to move the data sample data to a certain position. The essence of scaling is to fix the data in a certain range by dividing by a fixed value. Taking the logarithm is also a kind of scaling process. As for the characteristics of meteorological information, we choose the Min–Max scaling method to process it. When the data are centered according to the minimum value and then scaled by the range (maximum–minimum), the data move by the minimum unit and will be converged to between [0,1], and this process is called data as Min–Max Scaling. x_i in (1) represent the ith data in the modal. $\min(x_i)$ and $\max(x_i)$ represent the smallest value and the largest value in this modal, respectively. ${x_i}^*$ is the ith data after Min–Max Scaling processing.

$$ {x_i}^* = \frac{x_i - \min(x_i)}{\max(x_i) - \min(x_i)}. \tag{1} $$

After processing each modal data, we input the processed data into the multilayer perceptron built by ourselves. Our self-built multilayer perceptron mainly includes two fully connected layers. In (2), let i as the subscript of the previous layer of neurons or the input layer node, j as the subscript of the current layer of neurons or hidden layer of neurons, and w_{ij} represents the weight of each neuron in the previous layer to the current neuron, that is, the weight of neuron j. h_j represents the weighted sum of all inputs of the current node.

$$ h_j = \sum_{i=0}^{M} w_{ij} x_j. \tag{2} $$

In (3), a_j represents the output value of the hidden layer neuron.

$$ a_j = g(h_j) = g\left(\sum_{i=0}^{M} w_{ij} x_{ij}\right). \tag{3} $$

$f_{\text{meteorological}}$ in (4) represents the output value of the output layer, it is also meteorological feature extracted from the network structure. h_k represents the input weighted sum of neurons k in the output layer.

$$ f_{\text{meteorological}} = a_k = g(h_k) = g\left(\sum_{i=0}^{M} w_{jk} x_{jk}\right). \tag{4} $$

We add dropout [68] to the multilayer perceptron we established to prevent possible overfitting of the model. Briefly speaking, dropout is to let the activation value of a certain neuron stop working with a certain probability p when the network is propagating forward, which can make the model more generalized because it will not rely too much on certain parts characteristics.

D. Multimodal Fusion Network

Multimodal fusion network is used to get joint feature $f_{\text{multimodal}}$ after extracting features from image modal and meteorological modal. What is more, MFN further extract the fused features for final classification. f_{image} and $f_{\text{meteorological}}$ features extracted from IFME and MIFEM, respectively, are the input of our self-built MFN. The input of this module is the features extracted from two front modules. Unlike the IFME, using mainstream CNN as the feature extractor, MFCN is designed on the basis of multilayer perceptron. And the purpose of MFCN is to extract the multimodal feature $f_{\text{multimodal}}$. MFCN mainly includes three fully connected layers and choose ReLU as activation function. And we also add two dropout layers after first and second fully connected layers in order to avoid overfitting. The activation function selected in the last fully connected layer is different according to the specific classification task. We choose the softmax activation function for the single-label classification. In (5), k represents the total number of outputs in MFN, Z_j represents the jth original output value. $\sum_k^{k=1} e^{Z_k}$ represent all the factors of the original output value, which means that the different probabilities obtained by the Softmax function are related to each other.

$$\text{softmax}(Z_j) = \frac{e^{Z_j}}{\sum_k^{k=1} e^{Z_k}} \text{for } j = 1, 2, ..., k. \tag{5}$$

Different from single-label classification, we choose sigmoid function [69] for multilabel classification to project output logits to probability domain because in multilabel classification problem, for one image, multiple correct answers exist and the separate processing of different logits is needed. In (6), Z_j represents the jth original output value.

$$\text{sigmoid}(Z_j) = \frac{e^{Z_j}}{1 + e^{Z_j}}. \tag{6}$$

We get the final classification result based on the output value of the final activation function.

V. Experiments and Analysis

Experiments are conducted on a NVIDIA Quadro RTX A6000 GPU with 48 G memory. LSCIDMRME is composed of LSCIDMRME-S and LSCIDMRME-M, dealing with single-label and multilabel. Hence, two groups of experiments are carried in this section.

A. Experiments on LSCIDMRME-S

1) Baseline Modal Used in the Image Feature Extraction Module: We utilize three classical CNNs as the base model in IFEM in the experiments: AlexNet [24], ResNet-101 [70], and EfficientNet-B5 [28]. These base models and corresponding CNN part of MANET are pretrained on ImageNet, and other parts of MANET are initialized randomly. Feature vectors of 4096 dimensions, 2048 dimensions, and 2048 dimensions are constructed by the features extracted from the last fully connected layer of AlexNet, ResNet-101, and EfficientNet, respectively.

TABLE III
LSCIDMR With LabelLess: Overall Accuracies (%) of Four Kinds of Deep Learning Methods Under Different Testing Ratios

CNN	Testing ratios	
	10%	20%
AlexNet-image	78.92±0.71	77.32±0.20
MANET-AlexNet	**84.94± 0.21**	**84.40±0.13**
ResNet-101-image	84.18± 0.10	83.96±0.15
MANET-ResNet-101	**86.64± 0.23**	**86.31±0.09**
EfficientNet-image	85.25± 0.39	84.51±0.58
MANET-EfficientNet	**86.19± 0.23**	**85.13±0.09**

2) Parameter Setting: Two different training and testing ratios are taken into consideration in order to get a more comprehensive evaluation: 10% and 20%. For the former, 10% of data in each category is used for testing and the rest for training. For the latter, 20% of the data in each category is used for testing, while the rest is served as the training set. During training, the input to the CNN model is a batch of RGB images, whose sizes are fixed at 256 × 256 pixels. Simple data augmentation such as vertical flipping at random with a certain probability and proportional cropping is performed. We choose cross-entropy loss as the loss function, and stochastic gradient descent (SGD) is selected as the optimizer. The momentum rate of SGD is set as 0.9 and the learning rate is initialized as 1×10^{-5}. Training will last 100 epochs and the learning rate will be decreased by a factor of 5 every 20 epochs.

3) Evaluation Metrics: Overall accuracy [71] and confusion matrix [72] are used to evaluate the performance of image classification models. We perform training of these networks with 10 epochs, and the means and standard variances of the overall accuracy for each epoch would be calculated. During the training process, the model that can get the highest means and standard variances of the overall accuracy is saved as best models for the computation of confusion matrix. And then, through each of these best models, the correct and incorrect classification of each category would be calculated and put to corresponding position of the confusion matrix.

4) Results and Analyses: Table III presents the means and standard variance of the overall accuracy of each model with different testing ratio. From this table, it can be observed that the proposed MANET is more effective compared to the classification accuracy of a single modal. The overall classification accuracy of the proposed method has been improved after fusing the information of other meteorological modalities. Compared with the test ratio 20%, the classification accuracy got when the test ratio is 10%, reaches a better result. As for 10% testing ratio, there is more data that can be used in training and get more information during this process. And MANET-ResNet-101 achieves the highest classification accuracy compared with others.

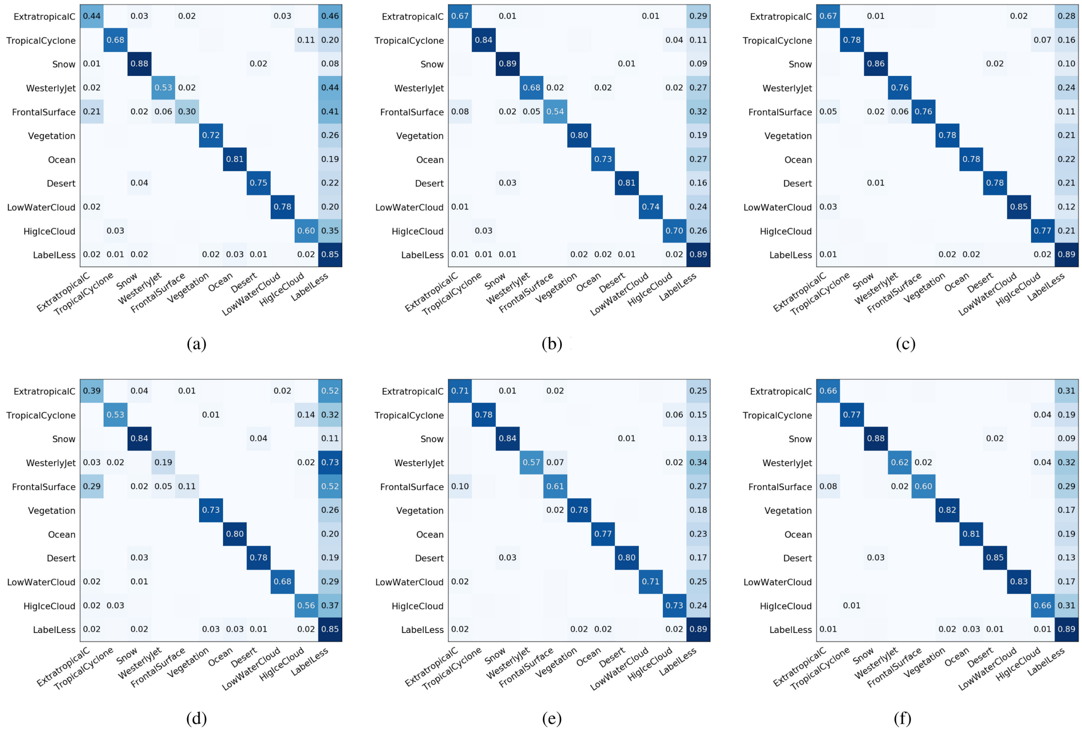


Fig. 5. LSCIDMR-S: Confusion matrices of baseline models under the different testing ratios. That of 10% are shown in (a)–(c) and that of under 20% are shown in (d)–(f). ExtratropicalC refers to Extratropical Cyclone. (a), (d) AlexNet. (b), (e) ResNet-101. (c), (f) EfficientNet-B5.

Figs. 5 and 6 show confusion matrices of baseline models and MANET, respectively, under different test ratios. Comparing these two figures, it can be told that with the same baseline model, the addition of meteorological information would improve the classification accuracy of each category, especially for weather systems and clouds. Take Frontal surface as an example, in Fig 5(a), the accuracy of it through AlexNet is 0.30, and if we add meteorological information into model as the form of MANET, the accuracy will improve to 0.43, as shown in Fig. 6(a). This kind of improvement of specific categories can be seen on different baseline models and test ratios. Above all, the accuracy of weather systems and cloud types are improved to varying degrees with the help of meteorological information. Thus, it is proved that the purpose of improving the performance of identifying weather systems and clouds is achieved by MANET. This improvement comes from the introduction of prior knowledge about what weather systems or cloud would emerge at what time and location. What is more, because 60% of samples are in the Labeless category, a shortcut solution for models is simply predicting unrecognized images as Labeless, and this cheating way ensures 60% correct probability of guessing. But we can tell from confusion matrices that this harmful situation will be mitigated by MANET since the number of false Labeless of MANET is smaller than baseline models.

B. Experiments on LSCIDMRME-M

1) Parameter Setting: For LSCIDMRME-M, the setting of training and testing ratio is the same as LSCIDMRME-S: 10% and 20%, respectively. The network structures mentioned above are utilized, but some slight modifications in the structure are applied since there are some differences in the processes of multi-label classification and single-label classification. The input image size of networks is 256×256 pixels, same as LSCIDMRME-S. The activation function is changed to Sigmoid [69] in the added fully connected layer of these networks; the loss function, further more, is replaced by binary cross entropy [73], [74]. Sigmoid is utilized as activation function to map the vector of each category's prediction score into the probability domain, which is in the range of 0–1. The threshold is set to 0.5. If the prediction score of a sample for a category is greater than this threshold, this sample is then categorized as that class. We still chose SGD as the optimizer. The initial learning rate and adjustment strategy of learning rate are the same as the above-mentioned experiments on LSCIDMRME-S.

2) Method Standard: Precision, recall, accuracy, and AbsoluteTrue are introduced as four indicators to evaluate our multi-label classification models; the specific formulas and principles of these four metrics are as follows. For a better illustration of the relationship between the ground truth and the predicted labels in a same patch, we draw a chart in Fig. 7. Given N as the number of patches in the dataset and L_k as the subset that contains every label for the kth patch, while L_k^* denotes the subset which contains every predicted label for the kth patch. In (9), the formula $L_k \cup L_k^*$ represents that every element in this set is a member of L_k or L_k^*. And an element in $L_k \cap L_k^*$ in (7)–(9) denotes a sample that belongs to both L_k and L_k^*. $||* ||$ denotes the number of elements of a specific set.

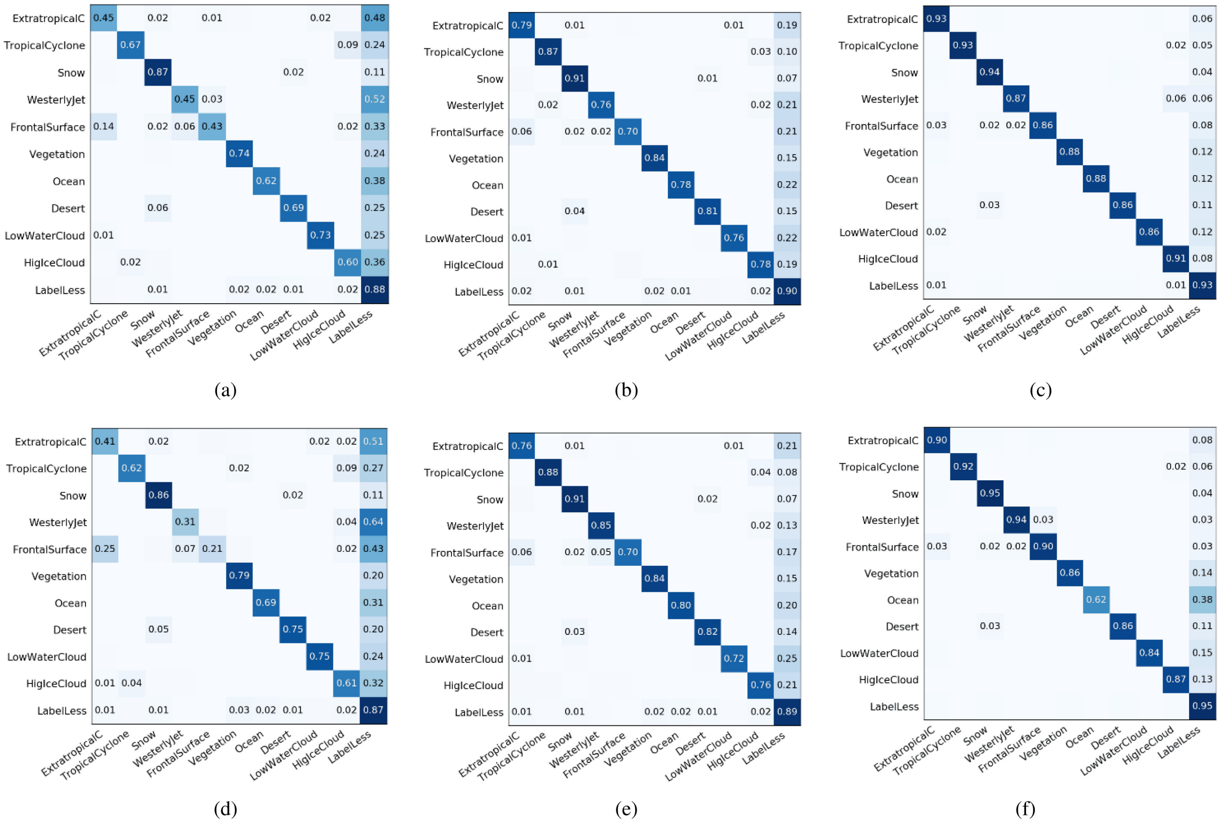


Fig. 6. LSCIDMRME-S: Confusion matrices of proposed MANET under different testing ratios. That of 10% are shown in (a)–(c) and that of under 20% are shown in (d)–(f). *ExtratropicalC* refers to *Extratropical Cyclone*. (a), (d) MANET-AlexNe. (b), (e) MANET-ResNet-101. (c), (f) MANET-EfficientNet.

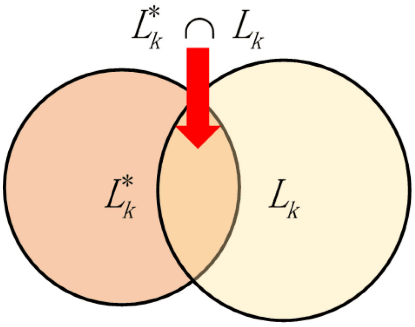


Fig. 7. Schematic drawing to show the meanings of some set theory symbols used in evaluation protocols in multilabel image classification. L_k represents the subset that contains all the label(s) for the kth sample; L_k^* represents the subset that contains all the predicted label(s) for the kth sample.

Precision [75]: *Precision* is the ratio of the number of properly predicted labels to all predicted labels.

$$\text{Precision} = \frac{1}{N}\sum_{k=1}^{N}\left(\frac{||L_k \cap L_k^*||}{||L_k^*||}\right). \tag{7}$$

Recall [75], [76]: *Recall* is the ratio of the number of correctly predicted labels to the real labels.

$$\text{Recall} = \frac{1}{N}\sum_{k=1}^{N}\left(\frac{||L_k \cap L_k^*||}{||L_k||}\right). \tag{8}$$

Accuracy [71]: *Accuracy* is the ratio of correctly predicted labels to the total labels including correctly and incorrectly predicted labels, those real labels missed in the prediction are also included.

$$\text{Accuracy} = \frac{1}{N}\sum_{k=1}^{N}\left(\frac{||L_k \cap L_k^*||}{||L_k \cup L_k^*||}\right). \tag{9}$$

AbsoluteTrue [76]: In the kth sample, when and only when all its label(s) predicted are identical to its true label(s) can be scored with 1; otherwise, 0.

$$\text{AbsoluteTrue} = \frac{1}{N}\sum_{k=1}^{N}\Delta\left(||L_k, L_k^*||\right) \tag{10}$$

$$\Delta(||L_k, L_k^*||) = \begin{cases} 1, & \text{all labels in } \mathrm{L_k} \text{ are same as in } \mathrm{L_k^*} \\ 0, & \text{otherwise} \end{cases}. \tag{11}$$

To calculate the four metrics mentioned-above, during the training process, the best model of each deep learning network in different training and testing ratios is saved.

3) Results and Analyses: Different metrics of baseline models and MANET are listed in Tables IV and V, respectively. The addition of meteorological information can improve the performance of multilabel classification on most metrics of all three baseline models. What is more, we can draw similar conclusions with single-label classification experiments that smaller

TABLE IV
EXPERIMENTAL RESULTS MEASURED BY PRECISION, RECALL, ACCURACY, AND ABSOLUTETRUE OF SINGLE-MODAL IMAGE CLASSIFICATION ON LSCIDMR-M IN DIFFERENT TEST RATIOS

CNN	AlexNet-image		ResNet-101-image		EfficientNet-image	
Test Ratio	10%	20%	10%	20%	10%	20%
Precision	96.73±0.03	96.88±0.11	97.87±0.05	97.87±0.06	96.24±2.52	95.6±0.02
Recall	94.21±0.08	94.53±0.01	**97.78±0.04**	97.66±0.04	94.80±0.33	93.51±0.19
Accuracy	96.35±0.06	96.58±0.05	98.28±0.03	98.23±0.02	96.46±1.09	95.75± 0.08
AbsoluteTrue	71.29±0.08	71.79±0.21	84.77±0.25	84.53±0.16	71.85±7.29	65.75±0.86

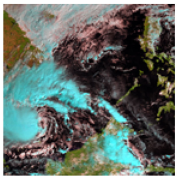
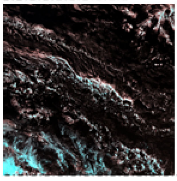
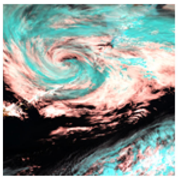
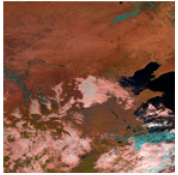
Better performances compared with multimodal models are in **Bold**.

TABLE V
EXPERIMENTAL RESULTS MEASURED BY PRECISION, RECALL, ACCURACY, AND ABSOLUTETRUE OF MULTIMODAL IMAGE CLASSIFICATION ON LSCIDMRME-M IN DIFFERENT TEST RATIOS

CNN	MANET-AlexNet		MANET-ResNet-101		MANET-EfficientNet	
Test Ratio	10%	20%	10%	20%	10%	20%
Precision	**97.44±0.08**	**97.60±0.07**	**98.07±0.04**	**98.09±0.03**	**98.21±0.04**	**98.20±0.02**
Recall	**97.25±0.11**	**97.19±0.04**	97.76±0.07	**97.73±0.02**	**97.93±0.04**	**97.90±0.03**
Accuracy	**97.91±0.02**	**97.94±0.01**	**98.34±0.01**	**98.34±0.01**	**98.47±0.02**	**98.46±0.01**
AbsoluteTrue	**82.03±0.21**	**82.36±0.10**	**85.45±0.14**	**85.79±0.10**	**86.53±0.19**	**86.50±0.09**

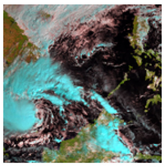
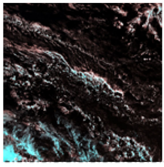
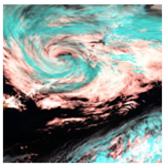
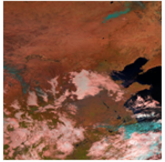
Better performances compared with single-modal models are in **Bold**.

TABLE VI
SINGLE-MODAL IMAGE CLASSIFICATION ON LSCIDMR-M: EXAMPLE OF LSCIDMR-M IMAGES WITH THE TRUE MULTILABELS AND THE MULTILABELS ASSIGNED BY DIFFERENT METHODS IN DIFFERENT TEST RATIOS

TestImage	TrueLabel	AlexNet		ResNet-19		EfficientNet-B5	
		10%	20%	10%	20%	10%	20%
	Tropical Cyclone	0	0	0	0	0	0
	Low Water Cloud	1	1	1	1	1	1
	High Ice Cloud	1	1	1	1	1	1
	Ocean	1	1	1	1	1	1
	Desert	1	1	0	1	1	1
	Vegetation	1	1	0	1	1	1
	Low Water Cloud	1	1	1	1	1	1
	High Ice Cloud	1	1	1	1	1	1
	Ocean	1	1	1	1	1	1
	Desert	0	0	1	1	0	0
	Vegetation	0	0	1	1	0	0
	Extratropical Cyclone	1	1	0	0	1	1
	Low Water Cloud	1	1	1	1	1	1
	High Ice Cloud	1	1	1	1	1	1
	Ocean	1	1	1	1	1	1
	Desert	0	0	1	1	0	0
	Vegetation	0	0	1	1	0	0
	Low Water Cloud	1	1	1	1	1	1
	High Ice Cloud	1	1	1	1	1	1
	Ocean	1	1	1	1	1	1
	Desert	1	1	1	1	1	1
	Vegetation	1	0	1	1	1	1

1 means the model predicts the label, 0 means not. The red mark is an indication of the wrong prediction of model.

TABLE VII
MULTIMODAL IMAGE CLASSIFICATION ON LSCIDMRME-M: EXAMPLE OF LSCIDMR-M IMAGES WITH THE TRUE MULTILABELS AND THE MULTILABELS ASSIGNED BY DIFFERENT METHODS IN DIFFERENT TEST RATIOS

TestImage	TrueLabel	MANET-AlexNet		MANET-ResNet-19		MANET-EfficientNet-B5	
		10%	20%	10%	20%	10%	20%
	Tropical Cyclone	**1**	**1**	0	0	**1**	**1**
	Low Water Cloud	1	1	1	1	1	1
	High Ice Cloud	1	1	1	1	1	1
	Ocean	1	1	1	1	1	1
	Desert	1	1	**1**	1	1	1
	Vegetation	1	1	0	1	1	1
	Low Water Cloud	1	1	1	1	1	1
	High Ice Cloud	1	1	1	1	1	1
	Ocean	1	1	1	1	1	1
	Desert	0	0	**0**	**0**	0	0
	Vegetation	0	0	**0**	**0**	0	0
	Extratropical Cyclone	1	1	0	0	1	1
	Low Water Cloud	1	1	1	1	1	1
	High Ice Cloud	1	1	1	1	1	1
	Ocean	1	1	1	1	1	1
	Desert	0	0	**0**	**0**	0	0
	Vegetation	0	0	**0**	**0**	0	0
	Low Water Cloud	1	1	1	1	1	1
	High Ice Cloud	1	1	1	1	1	1
	Ocean	1	1	1	1	1	1
	Desert	1	1	1	1	1	1
	Vegetation	1	0	1	1	1	1

1 means the model predicts the label, 0 means not. The red mark is an indication of the wrong prediction of model. The part in **Bold** is the part predicted correctly compared to the previous module.

test ratio can bring better performance due to a bigger training set. Generally speaking, consistent with the experiment results on LSCIDMRME-S, Tables IV and V proved that MANET has the ability to improve overall performance of classification.

From the perspective of model implementation, the multilabel classification task is done by changing the last Softmax function of single-label classification model to Sigmoid function. In other words, we canceled the incompatible assumptions of single-label classification to make model has the ability to predict multiple labels. This make our models actually groups of binary classifiers that can model the different labels separately. As given in Tables VI and VII, some examples of prediction results are listed. We can see that the first example image in Table VI is annotated with Tropical Cyclone and all baseline models cannot identify that label, but with the help of meteorological information, MANET-AlexNet and MANET-EifficientNet-B5 are able to detect such labels. From these two tables, we can also see that MANET is helpful for identifying land cover labels such as *Desert* and *Vegetation*. These proved that the proposed MANET has better performance on modeling weather systems, clouds, as well as other land cover labels.

VI. CONCLUSION

In this article, we upgrade the original single-modal dataset LSCIDMR into a multimodal dataset LSCIDMRME. Compared with LSCIDMR, LSCIDMRME has 521 950 multimodal tags. The new multimodal dataset will be uploaded to the GitHub and IEEE Data port. And we design MANET for satellite image classification by fusing multimodal information of LSCIDMRME. And the experimental results reflect that the proposed framework is able to achieve better classification performance than a single image modal classification. The purpose of identifying specific cloud types and weather systems also benefits from MANET through RS image classification tasks.

ACKNOWLEDGMENT

LSCIDMRME is processed from the original data supplied by P-Tree System, Japan Aerospace Exploration Agency. The authors would like to thank the researchers who kindly shared the codes utilized in this article. LSCIDMRME have been released at Github: https://github.com/Zjut-MultimediaPlus

REFERENCES

[1] C. Parmesan, T. L. Root, and M. R. Willig, "Impacts of extreme weather and climate on terrestrial biota*," *Bull. Amer. Meteorological Soc.*, vol. 81, no. 3, pp. 443–450, 2000.

[2] K. Emanuel and D. S. Nolan, "Tropical cyclone activity and the global climate system," in *Proc. 26th Conf. Hurricanes Trop. Meteorol.*, Miami, FL, USA, 2004, pp. 240–241.

[3] W. M. Gray, "Tropical cyclone genesis," *Atmospheric Science Paper*, Dept. Atmosph. Sci., Colorado State Univ., Fort Collins, CO, USA, 1975, Paper 234.

[4] H. Yang, L. Wu, and T. Xie, “Comparisons of four methods for tropical cyclone center detection in a high-resolution simulation,” *J. Meteorological Soc. Jpn. II*, vol. 98, no. 2, pp. 379–393, Jan. 2020.

[5] C. A. Doswell, “The distinction between large-scale and mesoscale contribution to severe convection: A case study example,” *Weather Forecast.*, vol. 2, no. 1, pp. 3–16, 1987.

[6] R. P. McNulty, “Severe and convective weather: A central region forecasting challenge,” *Weather Forecast.*, vol. 10, no. 2, pp. 187–202, 1995.

[7] C. A. Doswell, *Severe Convective Storms—An Overview.* Boston, MA, USA: American Meteorological Society, 2001, pp. 1–26.

[8] D. S. Thomas, “Sandstorm in a teacup? Understanding desertification,” *Geograph. J.*, vol. 159, pp. 318–331, 1993.

[9] T. Ishihara and Y. Iwagaki, “On the effect of sand storm in controlling the mouth of the Kiku river,” *Bull.-Disaster Prevention Res. Inst., Kyoto Univ.*, vol. 2, pp. 1–32, 1952.

[10] H. Lin, Z. Li, J. Li, F. Zhang, M. Min, and W. P. Menzel, “Estimate of daytime single-layer cloud base height from advanced baseline imager measurements,” *Remote Sens. Environ.*, vol. 274, 2022, Art. no. 112970.

[11] X. L. Wang, V. R. Swail, and F. W. Zwiers, “Climatology and changes of extratropical cyclone activity: Comparison of ERA-40 with NCEP–NCAR reanalysis for 1958–2001,” *J. Climate*, vol. 19, no. 13, pp. 3145–3166, Jul. 2006.

[12] I. Simmonds and K. Keay, “Mean southern hemisphere extratropical cyclone behavior in the 40-year NCEP–NCAR reanalysis,” *J. Climate*, vol. 13, no. 5, pp. 873–885, Mar. 2000.

[13] H. Dacre, “A review of extratropical cyclones: Observations and conceptual models over the past 100 years,” *Weather*, vol. 75, no. 1, pp. 4–7, Jan. 2020.

[14] X. Xiangde, “The effects of sensitive region over tibetan plateau on disastrous weather and climate and its monitoring,” *Eng. Sci.*, vol. 11, pp. 96–107, 2009.

[15] T. Baltrušaitis, C. Ahuja, and L.-P. Morency, “Multimodal machine learning: A survey and taxonomy,” *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 41, no. 2, pp. 423–443, Feb. 2019.

[16] P. K. Atrey, M. A. Hossain, A. El Saddik, and M. S. Kankanhalli, “Multimodal fusion for multimedia analysis: A survey,” *Multimedia Syst.*, vol. 16, no. 6, pp. 345–379, Nov. 2010. [Online]. Available: https://doi.org/10.1007/s00530-010-0182-0

[17] D. Ramachandram and G. W. Taylor, “Deep multimodal learning: A survey on recent advances and trends,” *IEEE Signal Process. Mag.*, vol. 34, no. 6, pp. 96–108, Nov. 2017.

[18] L. Gómez-Chova, D. Tuia, G. Moser, and G. Camps-Valls, “Multimodal classification of remote sensing images: A review and future directions,” *Proc. IEEE*, vol. 103, no. 9, pp. 1560–1584, Sep. 2015.

[19] D. Zhang et al., “Multimodal classification of Alzheimer’s disease and mild cognitive impairment,” *Neuroimage*, vol. 55, no. 3, pp. 856–867, 2011.

[20] P. Li, L. Dong, H. Xiao, and M. Xu, “A cloud image detection method based on SVM vector machine,” *Neurocomputing*, vol. 169, pp. 34–42, 2015. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0925231215006864

[21] Z. Jinglin, P. Liu, Z. Feng, and S. Qianqian, “Cloudnet: Ground-based cloud classification with deep convolutional neural network,” *Geophys. Res. Lett.*, vol. 45, no. 16, pp. 8665–8672, Aug. 2018.

[22] J. Haut, M. Paoletti, A. Paz-Gallardo, J. Plaza, A. Plaza, and J. Vigo-Aguiar, “Cloud implementation of logistic regression for hyperspectral image classification,” in *Proc. 17th Int. Conf. Comput. Math. Methods Sci. Eng.*, 2017, vol. 3, pp. 1063–2321.

[23] C. Bai, M. Zhang, J. Zhang, J. Zheng, and S. Chen, “LSCIDMR: Large-scale satellite cloud image database for meteorological research,” *IEEE Trans. Cybern.*, early access, doi: 10.1109/TCYB.2021.3080121.

[24] J. Deng, W. Dong, R. Socher, L. J. Li, and F. F. Li, “ImageNet: A large-scale hierarchical image database,” in *Proc. IEEE Comput. Soc. Conf. Comput. Vis. Pattern Recognit.*, Miami, FL, USA, 2009, pp. 20–25.

[25] Y. Lecun and L. Bottou, “Gradient-based learning applied to document recognition,” in *Proc. IEEE Proc. IRE*, vol. 86, no. 11, pp. 2278–2324, Nov. 1998.

[26] C. Szegedy, W. Liu, Y. Jia, P. Sermanet, and A. Rabinovich, “Going deeper with convolutions,” in *Proc. IEEE Comput. Soc.*, 2014, pp. 1–9.

[27] K. He, X. Zhang, S. Ren, and J. Sun, “Deep residual learning for image recognition,” in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2016, pp. 770–778.

[28] M. Tan and Q. V. Le, “Efficientnet: Rethinking model scaling for convolutional neural networks,” in *Proc. Int. Conf. Mach. Learn.*, 2019, pp. 6105–6114.

[29] G. Camps-Valls, L. Gomez-Chova, J. Munoz-Mari, J. L. Rojo-Alvarez, and M. Martinez-Ramon, “Kernel-based framework for multitemporal and multisource remote sensing data classification and change detection,” *IEEE Trans. Geosci. Remote Sens.*, vol. 46, no. 6, pp. 1822–1835, Jun. 2008.

[30] C. Couprie, C. Farabet, L. Najman, and Y. Lecun, “Indoor semantic segmentation using depth information,” in *Proc. 1st Int. Conf. Learn. Representations*, 2013, pp. 1–8.

[31] W. Wang, X. Yang, B. C. Ooi, D. Zhang, and Y. Zhuang, “Effective deep learning-based multi-modal retrieval,” *Very Large Data Bases J.*, vol. 25, no. 1, pp. 79–101, 2016.

[32] A. Wang, J. Lu, J. Cai, T. J. Cham, and G. Wang, “Large-margin multimodal deep learning for RGB-D object recognition,” *IEEE Trans. Multimedia*, vol. 17, no. 11, pp. 1887–1898, Nov. 2015.

[33] M. Amani et al., “Evaluation of the Landsat-based canadian wetland inventory map using multiple sources: Challenges of large-scale wetland classification using remote sensing,” *IEEE J. Sel. Topics Appl. Earth Observ. Remote Sens.*, vol. 14, pp. 32–52, 2020.

[34] Y. Yang and S. Newsam, “Bag-of-visual-words and spatial extensions for land-use classification,” in *Proc. 18th SIGSPATIAL Int. Conf. Adv. Geographic Inf. Syst. Conf.*, 2010, pp. 270–279.

[35] W. Shao, W. Yang, and G.-S. Xia, “Extreme value theory-based calibration for the fusion of multiple features in high-resolution satellite scene classification,” *Int. J. Remote Sens.*, vol. 34, no. 23, pp. 8588–8602, 2013.

[36] P. Helber, B. Bischke, A. Dengel, and D. Borth, “EuroSAT: A novel dataset and deep learning benchmark for land use and land cover classification,” *IEEE J. Sel. Topics Appl. Earth Observ. Remote Sens.*, vol. 12, no. 7, pp. 2217–2226, Jul. 2019.

[37] B. Chaudhuri, B. Demir, S. Chaudhuri, and L. Bruzzone, “Multilabel remote sensing image retrieval using a semisupervised graph-theoretic method,” *IEEE Trans. Geosci. Remote Sens.*, vol. 56, no. 2, pp. 1144–1158, Feb. 2018.

[38] G. Sumbul, M. Charfuelan, B. Demir, and V. Markl, “Bigearthnet: A large-scale benchmark archive for remote sensing image understanding,” in *Proc. IEEE Int. Geosci. Remote Sens. Symp. Conf.*, 2019, pp. 5901–5904.

[39] G. Sumbul et al., “BigEarthNet-MM: A large-scale, multimodal, multilabel benchmark archive for remote sensing image classification and retrieval [software and data sets],” *IEEE Geosci. Remote Sens. Mag.*, vol. 9, no. 3, pp. 174–180, Sep. 2021.

[40] B. Zhang, Y. Zhang, and S. Wang, “A lightweight and discriminative model for remote sensing scene classification with multidilation pooling module,” *IEEE J. Sel. Topics Appl. Earth Observ. Remote Sens.*, vol. 12, no. 8, pp. 2636–2653, Aug. 2019.

[41] C. Shi, T. Wang, and L. Wang, “Branch feature fusion convolution network for remote sensing scene classification,” *IEEE J. Sel. Topics Appl. Earth Observ. Remote Sens.*, vol. 13, pp. 5194–5210, 2020.

[42] X. Zhang, W. An, J. Sun, H. Wu, W. Zhang, and Y. Du, “Best representation branch model for remote sensing image scene classification,” *IEEE J. Sel. Topics Appl. Earth Observ. Remote Sens.*, vol. 14, pp. 9768–9780, 2021.

[43] N. He, L. Fang, S. Li, J. Plaza, and A. Plaza, “Skip-connected covariance network for remote sensing scene classification,” *IEEE Trans. Neural Netw. Learn. Syst.*, vol. 31, no. 5, pp. 1461–1474, May 2020.

[44] T. Tian, L. Li, W. Chen, and H. Zhou, “SEMSDNet: A multiscale dense network with attention for remote sensing scene classification,” *IEEE J. Sel. Topics Appl. Earth Observ. Remote Sens.*, vol. 14, pp. 5501–5514, 2021.

[45] J. Xie, N. He, L. Fang, and A. Plaza, “Scale-free convolutional neural network for remote sensing scene classification,” *IEEE Trans. Geosci. Remote Sens.*, vol. 57, no. 9, pp. 6916–6928, Sep. 2019.

[46] S. K. Roy, G. Krishna, S. R. Dubey, and B. B. Chaudhuri, “HybridSN: Exploring 3-D–2-D CNN feature hierarchy for hyperspectral image classification,” *IEEE Geosci. Remote Sens. Lett.*, vol. 17, no. 2, pp. 277–281, Feb. 2020.

[47] J. Zheng, Y. Feng, C. Bai, and J. Zhang, “Hyperspectral image classification using mixed convolutions and covariance pooling,” *IEEE Trans. Geosci. Remote Sens.*, vol. 59, no. 1, pp. 522–534, Jan. 2021.

[48] D. Lin, K. Fu, Y. Wang, G. Xu, and X. Sun, “MARTA GANs: Unsupervised representation learning for remote sensing image classification,” *IEEE Geosci. Remote Sens. Lett.*, vol. 14, no. 11, pp. 2092–2096, Nov. 2017.

[49] Y. Yu, X. Li, and F. Liu, “Attention GANs: Unsupervised deep feature learning for aerial scene classification,” *IEEE Trans. Geosci. Remote Sens.*, vol. 58, no. 1, pp. 519–531, Jan. 2020.

[50] X. Liu, Y. Zhou, J. Zhao, R. Yao, B. Liu, and Y. Zheng, “Siamese convolutional neural networks for remote sensing scene classification,” *IEEE Geosci. Remote Sens. Lett.*, vol. 16, no. 8, pp. 1200–1204, Aug. 2019.

[51] X. Chen, M. Ma, Y. Li, and W. Cheng, "Fusing deep features by kernel collaborative representation for remote sensing scene classification," *IEEE J. Sel. Topics Appl. Earth Observ. Remote Sens.*, vol. 14, pp. 12429–12439, 2021.

[52] Y. Xu, B. Du, and L. Zhang, "Multi-source remote sensing data classification via fully convolutional networks and post-classification processing," in *Proc. IEEE Int. Geosci. Remote Sens. Symp.*, 2018, pp. 3852–3855.

[53] D. Cerra et al., "Combining deep and shallow neural networks with ad hoc detectors for the classification of complex multi-modal urban scenes," in *Proc. IEEE Int. Geosci. Remote Sens. Symp.*, 2018, pp. 3856–3859.

[54] S. Fang, D. Quan, S. Wang, L. Zhang, and L. Zhou, "A two-branch network with semi-supervised learning for hyperspectral classification," in *Proc. IEEE Int. Geosci. Remote Sens. Symp.*, 2018, pp. 3860–3863.

[55] D. Hong et al., "More diverse means better: Multimodal deep learning meets remote-sensing imagery classification," *IEEE Trans. Geosci. Remote Sens.*, vol. 59, no. 5, pp. 4340–4354, May 2021.

[56] I. Shendryk, Y. Rist, R. Lucas, P. Thorburn, and C. Ticehurst, "Deep learning-a new approach for multi-label scene classification in planetscope and Sentinel-2 imagery," in *Proc. IEEE Int. Geosci. Remote Sens. Symp. Conf.*, 2018, pp. 1116–1119.

[57] Y. Bazi, "Two-branch neural network for learning multi-label classification in UAV imagery," in *Proc. IEEE Int. Geosci. Remote Sens. Symp. Conf.*, 2019, pp. 2443–2446.

[58] Y. Hua, L. Mou, and X. X. Zhu, "Label relation inference for multi-label aerial image classification," in *Proc. IEEE Int. Geosci. Remote Sens. Symp. Conf.*, 2019, pp. 5244–5247.

[59] Y. Hua, L. Mou, and X. Zhu Xiang, "Multi-label aerial image classification using a bidirectional class-wise attention network," in *Proc. Joint Urban Remote Sens. Event Conf.*, 2019, pp. 1–4.

[60] G. Sumbul and B. Demir, "A novel multi-attention driven system for multi-label remote sensing image classification," in *Proc. IEEE Int. Geosci. Remote Sens. Symp. Conf.*, 2019, pp. 5726–5729.

[61] X. Tan, Z. Xiao, J. Zhu, Q. Wan, K. Wang, and D. Li, "Transformer-driven semantic relation inference for multilabel classification of high-resolution remote sensing images," *IEEE J. Sel. Topics Appl. Earth Observ. Remote Sens.*, vol. 15, pp. 1884–1901, 2022.

[62] Y. Li, R. Chen, Y. Zhang, and H. Li, "A CNN-GCN framework for multi-label aerial image scene classification," in *Proc. IEEE Int. Geosci. Remote Sens. Symp. Conf.*, 2020, pp. 1353–1356.

[63] D. Wu and B. Li, "Cloud feature extraction and classification of meteorological satellite cloud imagery," in *Proc. Int. Conf. Sensor Netw. Signal Process.*, 2018, pp. 292–296.

[64] W. Li, F. Zhang, H. Lin, X. Chen, J. Li, and W. Han, "Cloud detection and classification algorithms for Himawari-8 imager measurements based on deep learning," *IEEE Trans. Geosci. Remote Sens.*, vol. 60, 2022, Art. no. 4107117.

[65] B. Wang and L. Ho, "Rainy season of the Asian–Pacific summer monsoon," *J. Climate*, vol. 15, no. 4, pp. 386–398, 2002.

[66] M. R. Sinclair, J. A. Renwick, and J. W. Kidson, "Low-frequency variability of southern hemisphere sea level pressure and weather system activity," *Monthly Weather Rev.*, vol. 125, no. 10, pp. 2531–2543, 1997.

[67] M. Ruzicka, "On dimensionless numbers," *Chem. Eng. Res. Des.*, vol. 86, no. 8, pp. 835–868, 2008.

[68] G. E. Hinton, N. Srivastava, A. Krizhevsky, I. Sutskever, and R. R. Salakhutdinov, "Improving neural networks by preventing co-adaptation of feature detectors," *Comput. Sci.*, vol. 3, no. 4, pp. 212–223, 2012.

[69] S. Elfwing, E. Uchibe, and K. Doya, "Sigmoid-weighted linear units for neural network function approximation in reinforcement learning," *Neural Netw.*, vol. 107, pp. 3–11, Nov. 2018.

[70] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2016, pp. 770–778.

[71] *Accuracy (Trueness and Precision) of Measurement Methods and Results—Part II: Basic Method for the Determination of Repeatability and Reproducibility of a Standard Measurement Method*, ISO 5725-2:1994, 1994.

[72] T. Fawcett, "An introduction to ROC analysis," *Pattern Recognit. Lett.*, vol. 27, no. 8, pp. 861–874, Jun. 2006.

[73] D. Ramos, J. Franco-Pedroso, A. Lozano-Diez, and J. Gonzalez-Rodriguez, "Deconstructing cross-entropy for probabilistic binary classifiers," *Entropy*, vol. 20, no. 3, Mar. 2018, Art. no. 208.

[74] D. Chicco and G. Jurman, "The advantages of the Matthews correlation coefficient (MCC) over F1 score and accuracy in binary classification evaluation," *BMC Genom.*, vol. 21, no. 1, Jan. 2020, Art. no. 6.

[75] G. Tsoumakas, I. Katakis, and I. Vlahavas, "Mining multi-label data," in *Proc. Data Min. Knowl. Discov. Handbook*, 2010, vol. 7, pp. 667–685.

[76] K.-C. Chou, "Some remarks on predicting multi-label attributes in molecular biosystems," *Mol. Biosyst.*, vol. 9, no. 6, pp. 1092–1100, 2013.

Cong Bai (Member, IEEE) received the B.E. degree in electronic and information engineering from Shandong University, Jinan, China, in 2003, the M.E. degree in electronic circuit and system from Shanghai University, Shanghai, China, in 2009, and the Ph.D. degree in signal and image processing from the National Institute of Applied Sciences, Rennes, France, in 2013.

He is an Associate Professor with the College of Computer Science and Technology, Zhejiang University of Technology, Hangzhou, China. His research interests include computer vision and multimedia processing.

Dongxiaoyuan Zhao received the B.E. degree in computer science and technology from the Guilin University of Electronic Technology, Guilin, China. He is currently working toward the M.E. degree in computer science and technology with the College of Computer Sciences, Zhejiang University of Technology, Hangzhou, China.

His research interests include computer vision and remote sensing.

Minjing Zhang received the B.E. degree in computer science and technology from Jiangxi Normal University, Nanchang, China, in 2019 and the M.E. degree in computer technology from Zhejiang University of Technology, Hangzhou, China, in 2022.

Her research interests include multimedia information processing and remote sensing.

Jinglin Zhang received the M.E. degree in electronic circuit and system from Shanghai University, Shanghai, China, in 2010 and the Ph.D. degree in electronic and communication from the National Institute of Applied Sciences, Rennes, France, in 2014.

He is a Professor with the School of Control Science and Engineering, Shandong University, Jinan, China. His research interests include computer vision and remote sensing.