






Article

Multi-Evidence and Multi-Modal Fusion Network for Ground-Based Cloud Recognition

Shuang Liu ¹, Mei Li ¹, Zhong Zhang ^{1,*}, Baihua Xiao ² and Tariq S. Durrani ³

¹ Tianjin Key Laboratory of Wireless Mobile Communications and Power Transmission, Tianjin Normal University, Tianjin 300387, China; s.liu@tjnu.edu.cn (S.L.); limeitjnu@gmail.com (M.L.)

² The State Key Laboratory of Management and Control for Complex Systems, Institute of Automation, Chinese Academy of Sciences, Beijing 100190, China; baihua.xiao@ia.ac.cn

³ Department of Electronic and Electrical Engineering, University of Strathclyde, Glasgow G11XQ, UK; t.durrani@strath.ac.uk

* Correspondence: zhangz@tjnu.edu.cn

Received: 27 December 2019; Accepted: 27 January 2020; Published: 2 February 2020



Abstract: In recent times, deep neural networks have drawn much attention in ground-based cloud recognition. Yet such kind of approaches simply center upon learning global features from visual information, which causes incomplete representations for ground-based clouds. In this paper, we propose a novel method named multi-evidence and multi-modal fusion network (MMFN) for ground-based cloud recognition, which could learn extended cloud information by fusing heterogeneous features in a unified framework. Namely, MMFN exploits multiple pieces of evidence, i.e., global and local visual features, from ground-based cloud images using the main network and the attentive network. In the attentive network, local visual features are extracted from attentive maps which are obtained by refining salient patterns from convolutional activation maps. Meanwhile, the multi-modal network in MMFN learns multi-modal features for ground-based cloud. To fully fuse the multi-modal and multi-evidence visual features, we design two fusion layers in MMFN to incorporate multi-modal features with global and local visual features, respectively. Furthermore, we release the first multi-modal ground-based cloud dataset named MGCD which not only contains the ground-based cloud images but also contains the multi-modal information corresponding to each cloud image. The MMFN is evaluated on MGCD and achieves a classification accuracy of 88.63% comparative to the state-of-the-art methods, which validates its effectiveness for ground-based cloud recognition.

Keywords: ground-based cloud recognition; convolution neural network; feature fusion

1. Introduction

Clouds are collections of very tiny water droplets or ice crystals floating in the air. They exert a considerable impact on the hydrological cycle, earth's energy balance and climate system [1–5]. Accurate cloud observation is crucial for climate prediction, air traffic control, and weather monitoring [6].

In general, space-based satellite, air-based radiosonde and ground-based remote sensing observations are three major ways for cloud observation [7]. Satellite observations are widely applied in large-scale surveys. However, they have deficiencies in providing sufficient temporal and spatial resolutions for localized and short-term cloud analysis over a particular area. Although air-based radiosonde observation is excellent in cloud vertical structure detection, its cost is considerably high. As a result, the equipment of ground-based remote sensing observations are rapidly developed, such as total-sky imager (TSI) [8,9] and all sky imager [10,11], which can provide high-resolution remote sensing images at a low cost so as to promote local cloud analysis.

Ground-based cloud recognition is an essential and challenging issue in automatically local sky observation. With the substantial amount of ground-based cloud images, ground-based cloud recognition has been extensively studied in the academic community in recent decades. Most traditional algorithms for ground-based cloud recognition utilize hand-crafted features, for example, brightness, texture, shape and color, to represent cloud images [12–18], but they are deficient in modeling complex data distribution. Recently, the convolutional neural network (CNN) [19–24] has achieved remarkable performance in various research fields due to the nature of learning highly nonlinear feature transformations. Inspired by this, some cloud-related researchers employ the CNNs to exploit visual features from ground-based cloud images and thrust the performance of ground-based cloud recognition to a new level. For example, Shi et al. [25] utilized deep features of cloud images from convolutional layers for ground-based cloud recognition, where the sum-pooling or max-pooling is applied to feature maps. Zhang et al. [26] presented a straightforward network named CloudNet with a couple of convolutional and fully connected layers for ground-based cloud recognition. Li et al. [27] propounded the dual guided loss which could integrate the knowledge of different CNNs in the learning process. These CNN-based methods only utilize the entire cloud images to learn global visual features. Nevertheless, different cloud categories may share similar patterns bringing confusion in the decision of the classifier. Ye et al. [28] gathered local visual features from multiple convolutional layers by the pattern mining and selection strategy, and then encoded them using the Fisher vector. However, it only considers the local visual features, and these features are obtained from a pre-trained CNN without a learning process. As ground-based cloud images are highly complex because of large intra-class and small inter-class variations (see Figure 1), the existing methods focusing on cloud visual information can hardly satisfy the requirement of accurate ground-based cloud recognition.

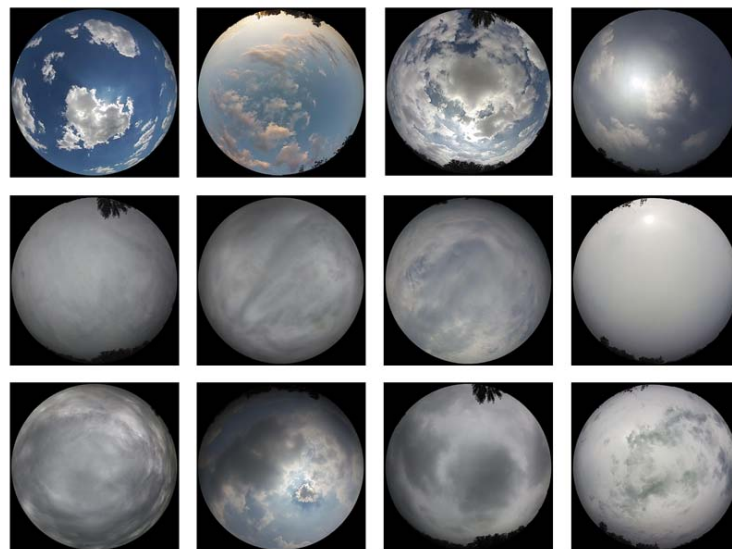


Figure 1. Some examples of ground-based cloud images. Each row indicates the cloud images from the same class.

The visual information contained in ground-based cloud image represents cloud only from the visual perspective, which cannot describe the cloud accurately due to the large variances in cloud appearance. It should be noted that cloud formation is a mutual process of multiple natural factors, including temperature, humidity, pressure and wind speed, which we name as *multi-modal information*. Clouds have a great correlation with multi-modal information [29,30]. For example, humidity influences cloud occurrence and cloud shape is affected by wind. Hence, instead of only focusing on cloud visual representations, it is more reasonable to enhance the recognition performance via combining ground-based cloud visual and multi-modal information. Liu and Li [31] extracted deep features by stretching the sum convolutional map obtained from pooling activation at the same

position of all the feature maps in deep convolutional layers. Then the deep features are integrated with multimodal features with weight. Liu et al. [32] propounded a two-stream network to learn ground-based cloud images and multi-modal information jointly, and then employed a weighted strategy to fuse these two kinds of information. In spite of these efforts, recognizing ground-based cloud using both cloud images and multi-modal information still remains an open issue.

Furthermore, existing public ground-based cloud datasets [26,33,34] lack data richness, which restricts the research of multi-modal ground-based cloud recognition. Specifically, none of these datasets contain both the ground-based cloud images and multi-modal information. Moreover, in practice, meteorological stations have been installed with equipment for collecting cloud images and multi-modal information, and therefore this information can easily be acquired.

In this paper, considering the above-mentioned issues, we propose the multi-evidence and multi-modal fusion network (MMFN) to fuse heterogeneous features, namely, global visual features, local visual features, and multi-modal information, for ground-based cloud recognition. To this end, the MMFN mainly consists of three components, i.e., main network, attentive network and multi-modal network. The main and attentive networks could mine the *multi-evidence*, i.e., global and local visual features, to provide discriminative cloud visual information. The multi-modal network is designed with fully connected layers to learn multi-modal features.

To optimize the existing public datasets, we release a new dataset named multi-modal ground-based cloud dataset (MGCD) which contains both the ground-based cloud images and the corresponding multi-modal information. Here, the multi-modal information refers to temperature, humidity, pressure and wind speed. It contains 8000 ground-based cloud samples constructed from a long-time period and larger than any of the existing public ones.

The contributions of this paper are summarized as follows:

- The proposed MMFN could fuse the multi-evidence and multi-modal features of ground-based cloud in an end-to-end fashion, which maximizes their complimentary benefits for ground-based cloud recognition.
- The attentive network refines the salient patterns from convolutional activation maps which could learn the reliable and discriminative local visual features for ground-based clouds.
- We release a new dataset MGCD which not only contains the ground-based cloud images but also contains the corresponding multi-modal information. To our knowledge, the MGCD is the first public cloud dataset containing multi-modal information.

2. Methods

The proposed MMFN is used for the multi-modal ground-based cloud recognition by fusing ground-based cloud images and multi-modal information. As depicted in Figure 2, it comprises three networks, i.e., main network, attentive network and multi-modal network, two fusion layers (*concat1* and *concat2*) and two fully connected layers (*fc5* and *fc6*). In this section, we detail the main network, the attentive network and the multi-modal network, respectively. Then, the fusion strategy between visual and multi-modal features is elaborated.

2.1. Main Network

The main network is used to learn global visual features from the entire ground-based cloud images, and it evolves from the widely-used ResNet-50 [20]. Figure 3 summaries the framework of ResNet-50 which mainly consists of six components, i.e., *conv1*, *conv2_x* ~ *conv5_x* and a fully connected layer. Additionally, *conv2_x* ~ *conv5_x* are constituted by 3, 4, 6 and 3 residual building blocks, respectively. Taking *conv3_x* as an example, it contains 4 residual building blocks, each of which is made up of three convolutional layers, where the convolution kernels are with the size of 1×1 , 3×3 and 1×1 , respectively. Note that the final fully connected layer of ResNet-50 is discarded in the main network. The output of *conv5_x* is aggregated by the average pooling layer (*avgpool1*) resulting in a 2048-dimensional vector which is treated as the input of the fusion layer (*concat1*).

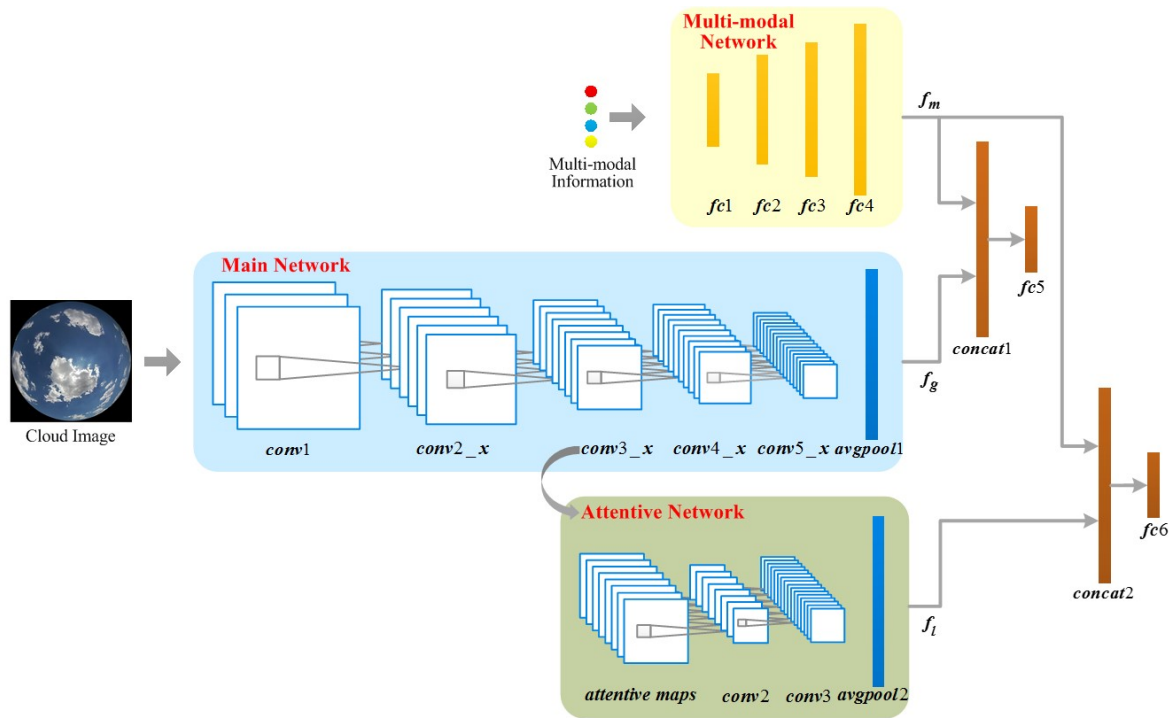


Figure 2. The architecture of the proposed multi-evidence and multi-modal fusion network (MMFN).

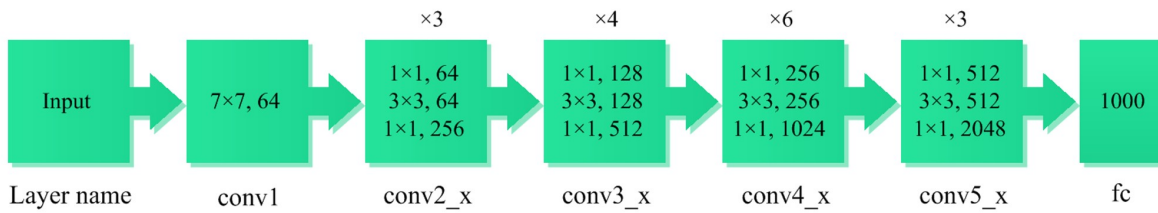


Figure 3. The overall framework of ResNet-50. Herein, the residual building blocks are expressed in green boxes and their number is displayed above the boxes.

2.2. Attentive Network

CNNs tend to pay more attention to local regions where the structure and texture information of clouds is reflected. Figure 4 visualizes the features of CNN implying that the salient parts or regions in ground-based cloud images play a decisive role in the recognition process. Hence, there is an inevitable need to extract local features from ground-based cloud images so as to complement global features. The attentive network is inspired by the great potential of the attention mechanism and used for exploiting local visual features for ground-based cloud recognition. Attention is the process of selecting and gating relevant information based on saliency [35] and it has been widely investigated in speech recognition [36,37], object detection [38], image captioning [39] and many other visual recognition works [40–43]. Meanwhile, the convolutional activation maps in shallow convolutional layers contain rich low-level patterns, such as structure and texture. Hence, we design the attentive network which consists of the attentive maps, two convolutional layers (*conv2* and *conv3*) and one average pooling layer (*avgpool2*) to extract local visual features from convolutional activation maps. Specifically, we first refine salient patterns from convolutional activation maps to obtain attentive maps that contain more semantic information. We then optimize the local visual features from attentive maps.

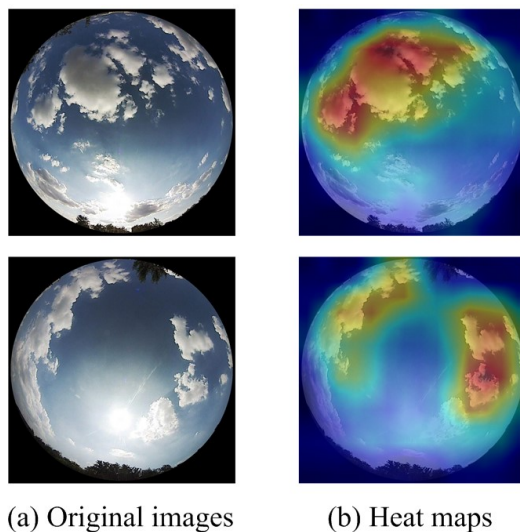


Figure 4. (a) represents the original ground-based cloud images and (b) represents the visualization features of convolutional neural networks (CNNs).

To learn reliable and discriminative local visual features, we propose attentive maps as the first part of the attentive network, which are generated by refining the salient patterns from the convolutional activation map. Explicitly, we treat the convolutional activation maps of the first residual building block in *conv3_x* as the input of attentive maps. Let $X_i = \{x_{i,j} \mid j = 1, 2, \dots, h \times w\}$ denote the i -th convolutional activation map, where $x_{i,j}$ is the response at location j , and h, w denote the height and width of convolutional activation map. Herein, there are 512 convolutional activation maps, and $h = w = 28$ in the first block of *conv3_x*. For the i -th convolutional activation map, we sort $x_{i,1} \sim x_{i,h \times w}$ in descending order and select the top $n \times n$ responses. Afterward, we reconstruct them into an $n \times n$ attentive map and maintain the descending order. Figure 5 illustrates the process of obtaining an attentive map, where n is set to 5. We exert the same strategy to all the convolutional activation maps, and therefore obtain 512 attentive maps. Hence, the attentive maps gather higher responses for the meaningful content and eliminate the negative effects caused by the non-salient responses.

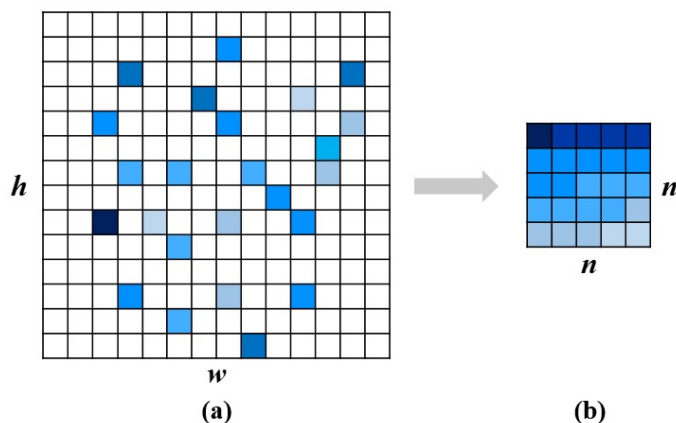


Figure 5. The process of obtaining an attentive map. (a) is the convolutional activation map with the highlighted top $n \times n$ responses, and (b) is the corresponding attentive map, where n is set to 5. Note that, the deeper the color, the larger the response.

Subsequently, the attentive maps are followed by a dropout layer. *conv2*, *conv3* and *avgpool2* are used to transform the attentive maps to a high dimensional vector by non-linear transformations. The convolution kernels of *conv2* and *conv3* are with the size of 3×3 and 1×1 , and with the stride of 2 and 1, respectively. Notice that, the target of using the convolution kernel with the size of 1×1

in *conv3* is to increase the output dimension. In addition, the numbers of convolution kernels in *conv2* and *conv3* are 512 and 2048, respectively. Both *conv2* and *conv3* are normalized by the batch normalization followed by the Leaky rectified linear unit (Leaky ReLU). The output of *avgpool2* is a 2048-dimensional vector and is fed into *concat2*.

Hence, with the help of the main and attentive networks, the proposed MMFN could mine the multi-evidence to represent ground-based cloud images in a unified framework. This provides discriminative cloud visual information.

2.3. Multi-Modal Network

The multi-modal information indicates the procedure of cloud formation, and thus we apply the multi-modal network to learn multi-modal features for completed cloud representation. As the input of the multi-modal network is the multi-modal information, which is represented as a vector, it is designed with four fully connected layers, i.e., $fc1 \sim fc4$, where the neuron numbers of them are 64, 256, 512 and 2048, respectively. Additionally, the batch normalization and the Leaky ReLU activation are connected to each of the first three. The output of *fc4* is passed through the Leaky ReLU activation, and then it is treated as the input of *concat1* and *concat2* at the same time. Herein, we denote the input of *concat1* and *concat2*, namely the output of the multi-modal network, as f_m .

2.4. Heterogeneous Feature Fusion

Feature fusion has been proved to be a robust and effective strategy to learn rich information in various areas, such as scene classification [44], facial expression recognition [45], action recognition [46] and so on. Especially, feature fusion methods based on deep neural networks are deemed to be extremely powerful and they are roughly divided into homogeneous feature fusion [47–50] and heterogeneous feature fusion [51–53]. For the former, many efforts concentrate on fusing the homogeneous features extracted from different components of CNN for recognition tasks. Compared with the homogeneous feature fusion, the heterogeneous feature fusion is rather tough and complex, because heterogeneous features possess significant different distributions and data structures. Herein, we focus the emphasis on the heterogeneous feature fusion.

The outputs of the main network, the attentive network and the multi-modal network, i.e., f_g , f_l and f_m , are treated as global visual features, local visual features and multi-modal features respectively, each of which is a 2048-dimensional vector. f_g is learned from the entire cloud images, and it contains more semantic information because of being extracted from the deeper layer of the main network. While f_l is learned from salient patterns in the shallow convolutional activation maps and it contains more texture information. Different from the visual features, f_m describes the clouds from the aspect of multi-modal information. Hence, these features describe the ground-based clouds from different aspects, and they contain some complementary information. To take full advantage of the complementary strengths among them, we combine the multi-modal features with the global and local visual features, respectively.

In this work, we propose two fusion layers *concat1* and *concat2* to fuse f_m with f_g and f_l respectively. In *concat1*, the integration algorithm for f_g and f_m can be formulated as

$$F_{gm} = g(f_g, f_m), \quad (1)$$

where $g(\cdot)$ denotes the fusion operation. In this work, $g(\cdot)$ is represented as

$$g(f_g, f_m) = [\lambda_1 f_g^T, \lambda_2 f_m^T]^T, \quad (2)$$

where $[\cdot, \cdot]$ means to concatenate two vectors, and λ_1 and λ_2 are the coefficients to trade-off the importance of f_g and f_m .

Similarly, the fusion features of f_l and f_m for *concta2* can be expressed as

$$F_{lm} = g(f_l, f_m) = [\lambda_3 f_l^T, \lambda_4 f_m^T]^T, \quad (3)$$

where λ_3 and λ_4 are the coefficients to balance the importance of f_l and f_m .

The final fully connected layers *fc5* and *fc6* are used for the recognition task and are connected to *concat1* and *concat2*, respectively. Each of them has K neurons, where K refers to the number of ground-based cloud categories. The output of *fc5* is fed into the softmax activation, and a series of label predictions over K categories are obtained to represent the probability of each category. The softmax activation is defined as

$$y_k = \frac{e^{x_k}}{\sum_{t=1}^K e^{x_t}}, \quad (4)$$

where x_k and $y_k \in [0, 1]$ are the value of the k -th neuron of *fc5* and the predicted probability of the k -th category, respectively. The cross-entropy loss is employed to calculate the loss value

$$L_1 = - \sum_{k=1}^K q_k \log y_k, \quad (5)$$

where q_k denotes the ground-truth probability, and it is assigned with 1 when k is the ground-truth label, otherwise, it is assigned with 0.

As for *fc6*, it is similar to *fc5*. Namely, its output is activated by the softmax and then evaluated by the cross-entropy loss L_2 . The total cost of the proposed MMFN is computed as

$$L = \alpha L_1 + \beta L_2, \quad (6)$$

where α and β are the weights to balance L_1 and L_2 . Hence, the optimization target of MMFN is to minimize Equation (6), and the training of MMFN is an end-to-end process, which is beneficial to the fusion of the multi-modal features with the global and local visual features under a single network. After training the MMFN, we extract the fused features F_{gm} and F_{lm} from ground-based cloud samples according to Equation (2) and Equation (3). Finally, F_{gm} and F_{lm} are directly concatenated as the final representation for ground-based cloud samples.

In short, the proposed MMFN has the following three properties. Firstly, the attentive network is utilized to refine the salient patterns from convolutional activation maps so as to learn the reliable and discriminative local visual features for ground-based clouds. Secondly, the MMFN could process the heterogeneous data. Specifically, the three networks in MMFN transform the corresponding heterogeneous data into the uniform format, which provides conditions for the subsequent feature fusion and discriminative feature learning. Thirdly, the MMFN could learn more extended fusion features for the ground-based clouds. It is because the heterogeneous features, i.e., global (local) visual features and the multi-modal features, are fused by two fusion layers which are optimized under one unified framework.

2.5. Comparison Methods

In this subsection, we describe comparison methods that are conducted in the experiments including variants of MMFN, and hand-crafted and learning-based methods.

2.5.1. Variants of MMFN

A unique advantage of the proposed MMFN is the capability of learning supplementary and complementary features, i.e., global visual features, local visual features and multi-modal features, from ground-based cloud data. Successful extraction of the expected features is guaranteed by several pivotal components, i.e., three networks, two fusion layers and the cooperative supervision of two

losses. For the purpose of demonstrating their effectiveness on MGCD, we list several variants of the proposed MMFN as follows.

variant1. The *variant1* is designed to only learn the global visual features of the ground-based cloud. It directly connects a fully connected layer with 7 neurons to the main network and the output of *avgpool1*, a 2048-dimensional vector, is used to represent the ground-based cloud images. Furthermore, the concatenation of the global visual features with the multi-modal information is denoted as *variant1* + MI.

variant2. The *variant2* is utilized to simply learn the local visual features. It maintains the architecture of the main network before the second residual building block in *conv_3* and the attentive network. Then, it adds a fully connected layer with 7 neurons after the attentive network. The output of *avgpool2*, a 2048-dimensional vector as well, is regarded as the representation of ground-based cloud. Similarly, the concatenation of the local visual features and the multi-modal information is denoted as *variant2* + MI.

variant3. The *variant3* is designed for integrating global and local visual features. To this end, the multi-modal network, as well as the two fusion layers of MMFN, are removed. Furthermore, the *variant3* adds a fully connected layer with 7 neurons after the main network and the attentive network, respectively. The outputs of *avgpool1* and *avgpool2* of the *variant3* are concatenated resulting in a 4096-dimensional vector as the final representation of ground-based cloud images. The feature representation of the *variant3* is integrated with the multi-modal information, which is referred to as *variant3* + MI.

variant4. The *variant4* fuses the global visual features and the multi-modal feature under a unified framework. Hence, it removes the attentive network and the fusion layer *concat2* of MMFN, and only uses L_1 to jointly learn cloud visual and multi-modal information. The output of *concat1*, a 4096-dimensional vector, is treated as the fusion of global visual features and the multi-modal features.

variant5. The *variant5* integrates local visual and multi-modal features under a unified framework. It discards the layers after the first residual building block in *conv3_x* of the main network and the fusion layer *concat1* of MMFN. The output of *concat2*, a 4096-dimensional vector, is regarded as the fusion result of local visual and multi-modal features.

variant6. For the purpose of demonstrating the effectiveness of two fusion layers, the *variant6* integrates the outputs of three networks using one fusion layer which is followed by one fully connected layer with 7 neurons. The output of the fusion layer is treated as the ground-based cloud representation which is a 6144-dimensional vector.

variant7. To learn discriminative local visual features, the MMFN reconstructs the salient responses to form the attentive maps. The counterpart *variant7* is employed to highlight the advantage of this innovative strategy. Instead of aggregating the top $n \times n$ responses, *variant7* randomly selects $n \times n$ responses from each convolutional activation map.

The sketches of *variant1* ~ *variant6* are shown in Figure 6.

2.5.2. Hand-Crafted and Learning-Based Methods

In this part, we provide descriptions of a series of methods for ground-based cloud classification involving hand-crafted methods, i.e., local binary patterns (LBP) [54] and completed LBP (CLBP) [55] and the learning-based method, i.e., bag-of-visual-words (BoVW) [56].

(a) The BoVW designs the ground-based cloud image representation with the idea of a bag of features framework. It densely samples SIFT features [57], which are then clustered by the K -means to generate a dictionary with 300 visual words. Based on the dictionary, each ground-based cloud image is transformed into the histogram of visual word frequency. To exploit the spatial information, the spatial pyramid matching scheme [58] is employed to partition each ground-based cloud image into an additional two levels with 4 and 16 sub-regions. Therefore, each ground-based cloud image with BoVW is represented by a 6300-dimensional histogram. This method is denoted as PBoVW.

(b) The LBP operator assigns binary labels to the circular neighborhoods of a center pixel according to their different signs. In this work, the uniform invariant LBP descriptor $LBP_{P,R}^{riu2}$ is used as the texture descriptor of ground-based cloud images, where (P, R) means P sampling points on a circle with the radius of R . We evaluate the cases with (P, R) being set to $(8, 1)$, $(16, 2)$ and $(24, 3)$, resulting into the descriptor being a feature vector with the dimensionality of 10, 18 and 26, respectively.

(c) The CLBP is proposed to improve LBP, and it decomposes local differences into signs and magnitudes. Besides, the local central information is considered as an operator. These three operators are jointly combined to describe each ground-based cloud image. The (P, R) is set to $(8, 1)$, $(16, 2)$ and $(24, 3)$, resulting in a 200-dimensional, a 648-dimensional and a 1352-dimensional feature vector, respectively.

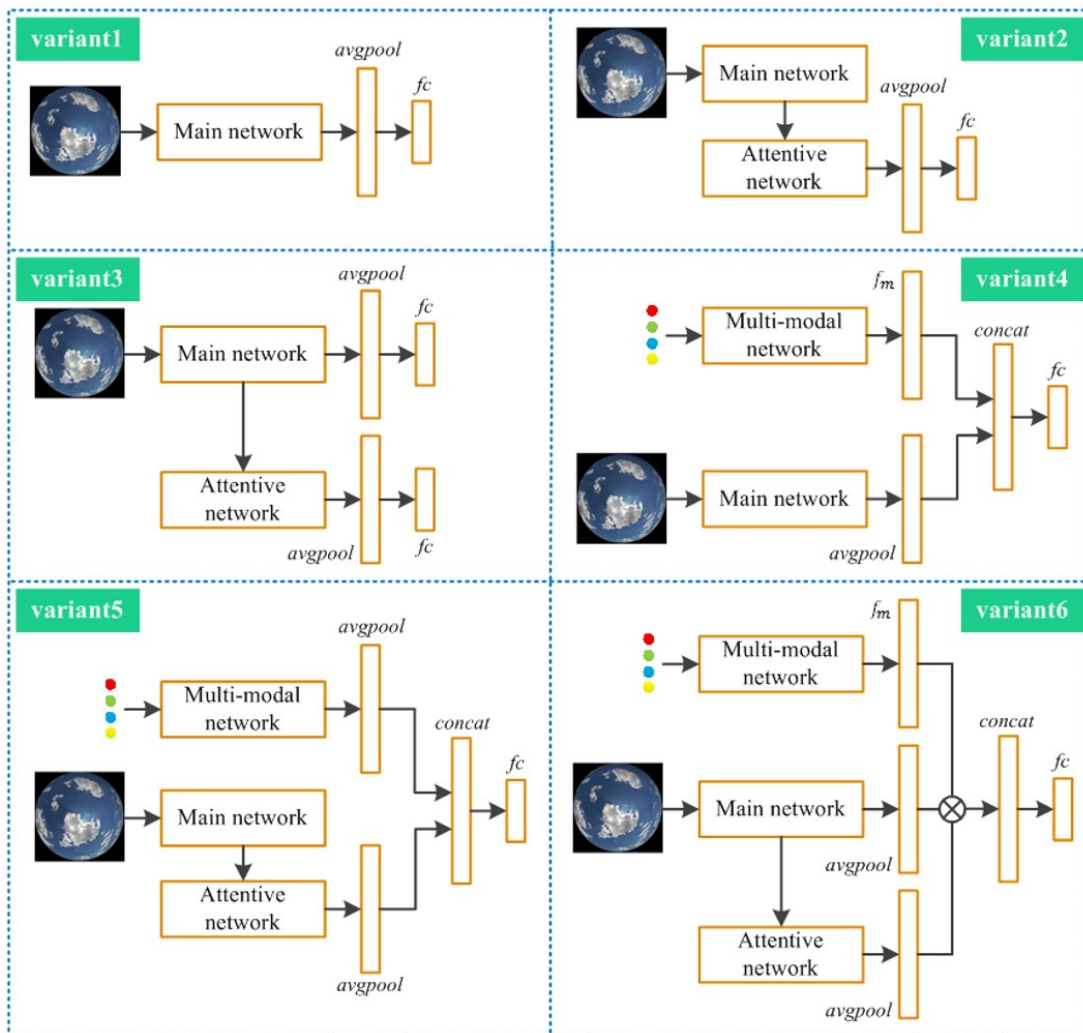


Figure 6. The sketches of variant1 ~ variant6.

2.6. Implementation Details

We first resize the ground-based cloud images into 252×252 and then randomly crop them to the fixed size of 224×224 . The ground-based cloud images are also augmented by the random horizontal flip. Thereafter, each of them is normalized by the mean subtraction. To ensure the compatibility, the values of multi-modal information, i.e., temperature, humidity, pressure and wind speed, are scaled to $[0, 1]$.

The main network is initialized by the ResNet-50 pre-trained on the ImageNet dataset. We adopt the weight initialization method in [59] for the convolutional layers (*conv2* and *conv3*) and the

fully connected layers. The weights of the batch normalization layers are initialized by the normal distribution with the mean and the standard deviation of 1 and 0.02, respectively. All the biases in the convolutional layers, fully connected layers and the batch normalization layers are initialized to zero. For the attentive maps, we discard the top 5 responses in each convolutional activation map to alleviate the negative effects of noise or outliers. We then form the attentive maps.

During the training phase, the SGD [60] optimizer is employed to update the parameters of the MMFN. The total training epochs are 50 with a batch size of 32. The weight decay is set to 2×10^{-4} with a momentum of 0.9. The learning rate is initialized to 3×10^{-4} and reduced by a factor of 0.2 at epoch 15 and 35, respectively. The slope of Leaky ReLU is fixed to 0.1, and the drop rate in the dropout layer of the attentive network is a constant of 0.5. Furthermore, the parameters in the multi-modal network are restricted to the box of $[-0.01, 0.01]$. After training the MMFN, each cloud sample is represented as an 8192-dimensional vector by concatenating the fusion features F_{gm} and F_{lm} . Then, the final representations of training samples are utilized to train the SVM classifier [61].

Furthermore, several parameters introduced in this paper, i.e., the parameter n in the attentive network, the parameters $\lambda_1 \sim \lambda_4$ in Equations (2) and (3) and the parameters α and β in Equation (6) are set as: $n = 7$, $(\lambda_1, \lambda_2) = (0.3, 0.7)$, $(\lambda_3, \lambda_4) = (0.3, 0.7)$ and $\alpha = \beta = 1$. Their influences on ground-based cloud classification performance with different settings are analyzed in Section 4.3.

3. Data

The multi-modal ground-based cloud dataset (MGCD) was the first one composed of ground-based cloud images and the multi-modal information. It was collected in Tianjin, China from March 2017 to December 2018 over a period of 22 months, at different locations and day times in all seasons, which ensured the diversity of cloud data. The MGCD contains 8000 ground-based cloud samples, which is the largest cloud dataset. Each sample was composed of one ground-based cloud image and a set of multi-modal information which had a one-to-one correspondence. The cloud images were collected by a sky camera with a fisheye lens, with a resolution of 1024×1024 pixels in JPEG format. The multi-modal information was collected by a weather station and stored in a vector with four elements, namely, temperature, humidity, pressure and wind speed.

We divided the sky conditions into seven sky types, as listed in Figure 7, including (1) cumulus, (2) altocumulus and cirrocumulus, (3) cirrus and cirrostratus, (4) clear sky, (5) stratocumulus, stratus and altostratus, (6) cumulonimbus and nimbostratus and (7) mixed cloud, under the genera-based classification recommendation of the World Meteorological Organization (WMO) as well as the cloud visual similarities in practice. The sky is often covered by no less than two cloud types, and this sky type is regarded as mixed cloud. Additionally, cloud images with cloudiness no more than 10% are categorized as clear sky. It should be noticed that all cloud images are labeled by meteorological experts and ground-based cloud-related researchers after much deliberation.

The ground-based cloud samples from the first 11 months constitutes the training set and these from the second 11 months are the test set, where each of the sets contains 4000 samples. Figure 7 presents the samples and the number from each cloud category in MGCD, where the multi-modal information is embedded in the corresponding cloud image. The MGCD is available at <https://github.com/shuangliutjnu/Multimodal-Ground-based-Cloud-Database>.

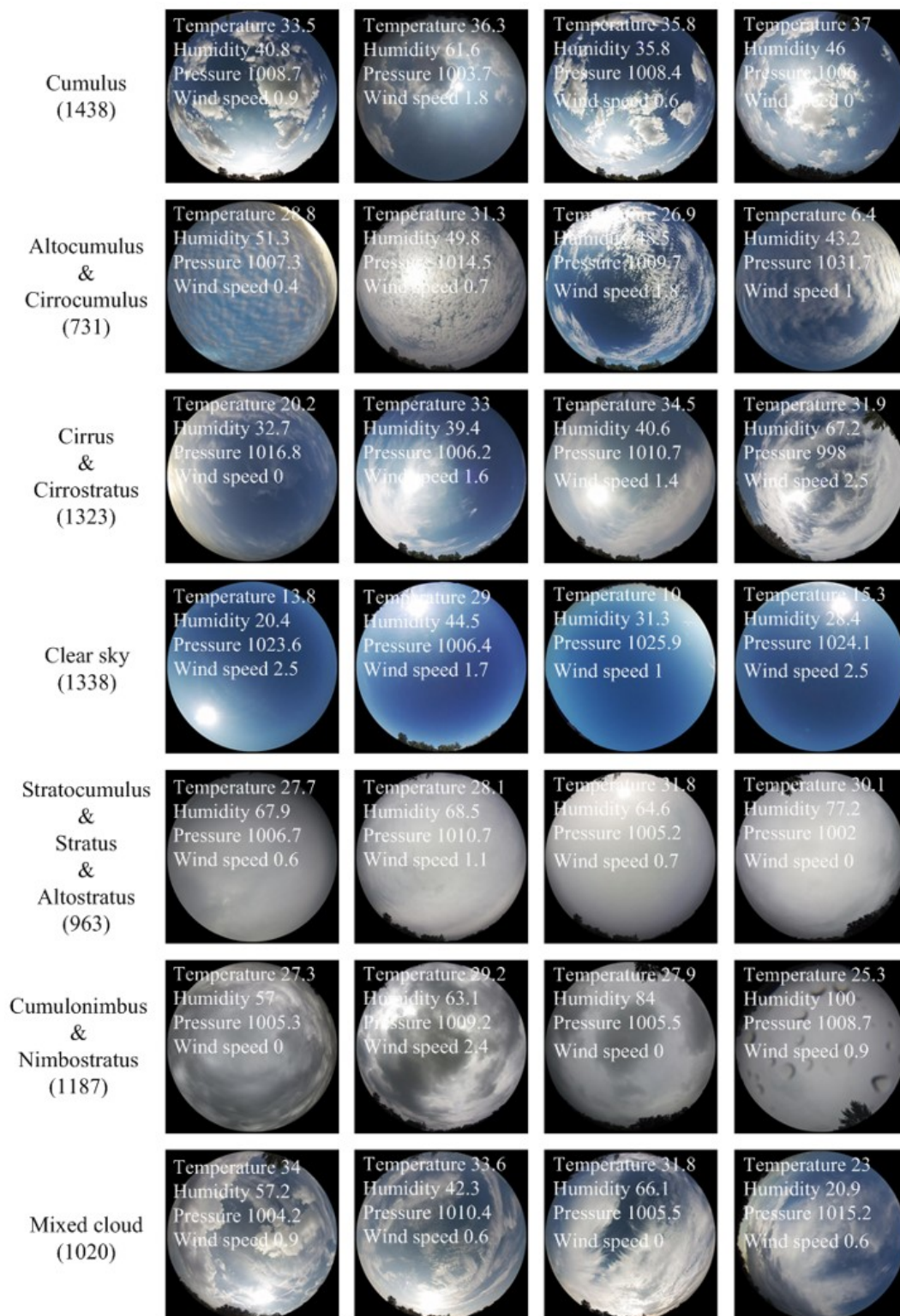


Figure 7. Some samples from each category in multi-modal ground-based cloud dataset (MGCD), where the multi-modal information is embedded in the corresponding ground-based cloud image.

4. Results

In this section, we present the comparisons of the proposed MMFN with the variants of MMFN and other methods on MGCD followed by the analysis of classification results with different parameters

4.1. Comparison with Variants of MMFN

The recognition accuracy of MMFN and its variants based on MGCD are presented in Table 1 where several conclusions can be drawn. Firstly, both variant1 and the variant2 achieve promising recognition accuracy. It indicates that both global visual features and local visual features are essential for cloud recognition, in which global visual features contain more semantic and coarse information while local visual features contain more texture and fine information. The variant3 achieves a recognition accuracy of 86.25% which exceeds 3.1% and 4.02% over the variant1 and the variant2, respectively. It is because the variant3 combines the benefits of global visual features and local visual features.

Table 1. The recognition accuracy (%) of the proposed multi-evidence and multi-modal fusion network (MMFN) as well as its variants. The notation “+” indicates the concatenation operation.

Methods	Accuracy (%)
variant1	83.15
variant1 + MI	84.48
variant2	82.23
variant2 + MI	83.70
variant3	86.25
variant3 + MI	87.10
variant4	85.90
variant5	83.70
variant6	87.38
variant7	87.60
MMFN	88.63

Secondly, the methods (variant1 + MI, variant2 + MI, variant3 + MI, variant4 and variant5) which employ the multi-modal information are more competitive than those (variant1, variant2 and variant3) that do not. Specifically, compared with variant1, variant1 + MI has an improvement of 1.33%, and so for variant2 + MI and variant3 + MI which have improvements of 1.47% and 0.85%, respectively. More importantly, the improvements of the variant4, variant5 and MMFN are 2.75%, 1.47% and 2.38% over the variant1, variant2 and variant3, respectively. We therefore conclude that jointly learning the cloud visual features and the multi-modal features under a unified framework can further improve the recognition accuracy.

Thirdly, from the comparison between variant6 and the proposed MMFN, we can discover that the recognition performance of the latter is superior to the former even though both of them learn the global visual features, local visual features and the multi-modal features under a unified framework. It indicates that fusing the multi-modal features with the global and local visual features, respectively, can sufficiently mine the complementary information among them and exploit more discriminative cloud features.

Finally, the proposed MMFN achieves better results than variant7 because the attentive maps of MMFN learn local visual features from the salient patterns in the convolutional activation maps. While variant7 randomly selects responses from convolutional activation maps. Hence, the proposed attentive map could learn effective local visual features.

4.2. Comparison with Other Methods

The comparison results between the proposed MMFN and other methods, such as [32,62,63], are summarized in Table 2. Firstly, most results in the right part of the table are more competitive than those in the left part, which indicates that the multi-modal information contains useful information for ground-based cloud recognition. The visual features and the multi-modal information are supplementary to each other, and therefore the integration of them could obtain the extended information for ground-based cloud representation. Secondly, the CNN-based methods, such as

CloudNet, JFCNN, DTFN and so on, are much better than the hand-crafted methods (LBP and CLBP) and the learning-based methods (BoVW and PBoVW). It is because the CNNs are with the nature of highly nonlinear transformations which enables them to extract effective features from highly complex cloud data. Thirdly, the proposed MMFN has an improvement over CNN-based methods, which verifies the effectiveness of the multi-evidence and multi-modal fusion strategy. Such a strategy thoroughly investigates the correlations between the visual features and the multi-modal information and takes into consideration the complementary and supplementary information between them as well as their relative importance for the recognition task.

Table 2. The recognition accuracies (%) of the proposed MMFN and other methods.

Methods	Accuracy (%)	Methods	Accuracy (%)
BoVW	66.15	BoVW + MI	67.20
PBoVW	66.13	PBoVW + MI	67.15
LBP ^{riu2} _{8,1}	45.38	LBP ^{riu2} _{8,1} + MI	45.25
LBP ^{riu2} _{16,2}	49.00	LBP ^{riu2} _{16,2} + MI	47.25
LBP ^{riu2} _{24,3}	50.20	LBP ^{riu2} _{24,3} + MI	50.53
CLBP ^{riu2} _{8,1}	65.10	CLBP ^{riu2} _{8,1} + MI	65.40
CLBP ^{riu2} _{16,2}	68.20	CLBP ^{riu2} _{16,2} + MI	68.48
CLBP ^{riu2} _{24,3}	69.18	CLBP ^{riu2} _{24,3} + MI	69.68
VGG-16	77.95	DMF [31]	79.05
DCAFs [25]	82.67	DCAFs + MI	82.97
CloutNet [26]	79.92	CloutNet + MI	80.37
		JFCNN [32]	84.13
		DTFN [62]	86.48
		HMF [63]	87.90
		MMFN	88.63

4.3. Parameter Analysis

In this subsection, we analyse the parameter n in the attentive network, the parameters $\lambda_1 \sim \lambda_4$ in Equations (2) and (3), and the parameters α and β in Equation (6).

We first analyse the parameter n which determines the size of the attentive map. The recognition accuracy with different n are displayed in Figure 8. We can see that the best recognition accuracy is achieved when n is equal to 7, namely 25% salient responses are selected from the convolutional activation map. While n is set to away from 7, the corresponding recognition accuracy is below the peak value of 88.63%.

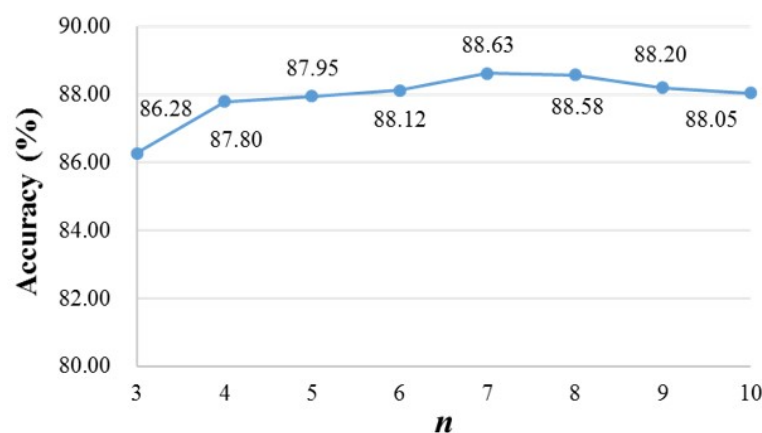


Figure 8. The recognition accuracy of MMFN with different n .

Then we analyse two pairs of parameters, i.e., λ_1 and λ_2 , λ_3 and λ_4 , in Equation (2) and Equation (3). λ_1 and λ_2 balance the significance of the global visual features and the multi-modal features respectively. Similarly, λ_3 and λ_4 balance the significance of the local visual features and the multi-modal features. Appropriate $\lambda_1 \sim \lambda_4$ settings can optimize the recognition result. The recognition accuracies with different (λ_1, λ_2) and (λ_3, λ_4) settings are illustrated in Table 3 and Table 4. From Table 3 we can see that when (λ_1, λ_2) is equal to (0.3, 0.7), the best recognition accuracy is obtained. Similarly, Table 4 shows that when (λ_3, λ_4) is with the setting of (0.3, 0.7), the best recognition accuracy is achieved.

Table 3. The recognition accuracy (%) with different (λ_1, λ_2) settings.

(λ_1, λ_2)	Accuracy (%)
(0.2, 0.8)	75.02
(0.3, 0.7)	88.63
(0.4, 0.6)	88.33
(0.5, 0.5)	88.53
(0.6, 0.4)	88.30
(0.7, 0.3)	88.10
(0.8, 0.2)	87.85

Table 4. The recognition accuracies (%) with different (λ_3, λ_4) settings.

(λ_3, λ_4)	Accuracy (%)
(0.2, 0.8)	87.90
(0.3, 0.7)	88.63
(0.4, 0.6)	87.75
(0.5, 0.5)	87.85
(0.6, 0.4)	87.90
(0.7, 0.3)	87.85
(0.8, 0.2)	87.80

Afterwards, we evaluate the parameters α and β which are a tradeoff between the losses L_1 and L_2 in Equation (6). The recognition performances with different α and β settings are summarized in Table 5. It is observed when $\alpha = \beta = 1$, the best recognition accuracy is obtained.

Table 5. The recognition accuracy (%) with different (α, β) settings.

(α, β)	Accuracy (%)
(0.6, 0.4)	87.15
(0.7, 0.3)	87.30
(0.8, 0.2)	87.85
(1, 1)	88.63
(1, 1.5)	87.93
(1, 2)	87.80
(1.5, 1)	87.38
(2, 1)	87.00

Besides, as more training data means better training of the network, we implement the experiment with the dataset being divided into different ratios, i.e., 60/40, 70/30 and 80/20, to evaluate the influences on recognition accuracies caused by the training data numbers, and the results are illustrated in Figure 9. As shown, more training samples lead to higher recognition performance.

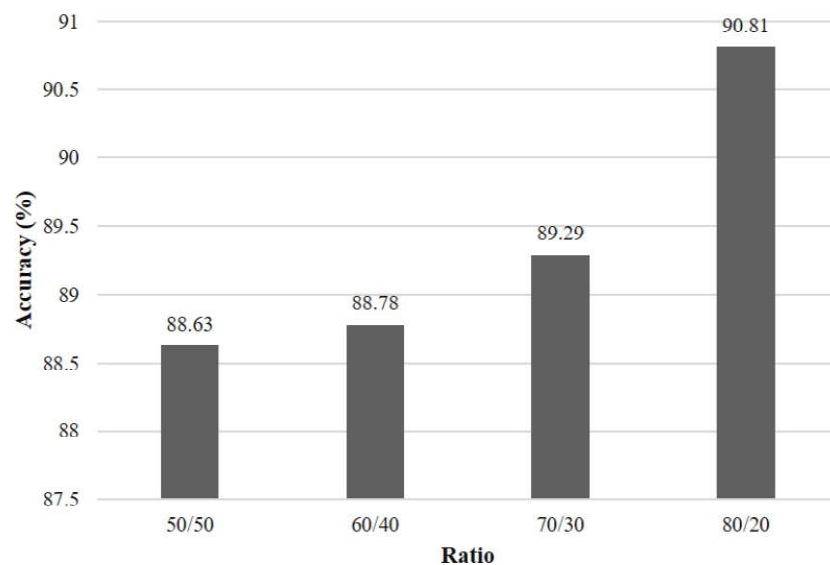


Figure 9. The recognition accuracies (%) of MMFN with the training and test samples under different ratios.

5. Discussion

5.1. Overall Discussion

Cloud classification is both a basic and necessary part of climatological and weather research and provides indicative knowledge about both short-term weather conditions and long-term climate change [64]. There are many algorithms developed for automatic cloud classification. Xiao et al. [65] aggregated texture, structure and color features which are extracted simultaneously from ground-based cloud images and encoded by the Fisher vector as a cloud image descriptor. Afterwards, a linear SVM was employed to group 1231 ground-based cloud images into six classes with an accuracy of 87.5%. Wang et al. [7] applied the selection criterion which is based on the Kullback–Leibler divergence between LBP histograms of the original and resized ground-based cloud images to select the optimal resolution of the resized cloud image. The criterion was evaluated on three datasets with a total of ground-based cloud images 550 and 5000 from five classes, and 1500 from seven classes, respectively. The overall classification results of these three datasets are around 65.5%, 45.8% and 44% respectively. Zhang et al. [26] employed the CloudNet composed of five convolutional layers and two fully connected layers to divide 2543 ground-based cloud images into 10 categories with an average accuracy of 88%. Li et al. [62] presented a deep tensor fusion network which fuses cloud visual features and multimodal features at the tensor level so that the spatial information of ground-based cloud images can be maintained. They obtained the classification result of 86.48% over 8000 ground-based cloud samples. Liu et al. [63] fused deep multimodal and deep visual features in a two-level fashion, i.e., low-level and high-level. The low-level fused the heterogeneous features directly and its output was regarded as a part of the input of the high-level which also integrates deep visual and deep multimodal features. The classification accuracy of the hierarchical feature fusion method was 87.9% over 8000 ground-based cloud samples.

Of the above-mentioned studies, most of them have achieved high accuracy, but the datasets used in these studies are either with a small number of ground-based images or not public. The availability of sufficient ground-based cloud samples is a fundamental factor to allow CNNs to work effectively. In addition, since cloud types change over time, appropriate fusion of multi-modal information and cloud visual information could improve the classification performance. The JFCNN [32] achieved excellent performance with the accuracy of 93.37% by learning ground-based cloud images and multi-modal information jointly. However, the dataset used in [32] only contains 3711 labeled cloud

samples, and it is randomly split into the training set and the test set with the ratio of 2:1, which means there may exist high dependence between training and test samples. In this study, we create a more extensive dataset MGCD containing 8000 ground-based cloud samples with both cloud images and the corresponding multimodal information. All the samples are classified into the training set and the test set and both of them are with 4000 samples, where the training set is grouped from the first 11 months and the test set is grouped from the second 11 months. Hence, the training and test sets in the MGCD are independent. As salient local patterns play a decisive role in the decision-making procedure, we devise MMFN with three networks, i.e., main network, attentive network and multi-modal network, to generate global visual features, local visual features and multi-modal features, and fuse them at two fusion layers. The proposed MMFN obtains the best result of 88.63%. We first assess the rationality of each component of MMFN by comparing with its variants. Then, we exert comparisons between MMFN and other methods, including the hand-crafted methods (LBP and CLBP), the learning-based methods (BoVW and PBoVW) and the CNN-based methods (DMF, JFCNN, HMF and so on), where the accuracy gaps between the proposed MMFN and the hand-crafted and learning-based methods are over 18 percentage points, and the gap between the proposed MMFN and the second-best CNN-based method, i.e., HMF is 0.73 percentage points.

It is quite reasonable that the proposed MMFN has a competitive edge over other methods. Affected by temperature, wind speed, illumination, noise, deformation and other environmental factors, the cloud images are with the characteristic of volatility leading to intractability of cloud recognition. A more effective strategy is imperative and required to obtain extended cloud information. Hence, the proposed MMFN jointly learns the cloud multi-evidence and the multi-modal information and extracts powerful and discriminative features from ground-based cloud samples. Accordingly, the proposed MMFN makes a significant improvement over other methods to the recognition accuracy.

5.2. Potential Applications and Future Work

Generally, this research points out a new method to promote the accuracy of cloud classification using cloud images and multi-modal information, which is beneficial to the regional weather prediction. Furthermore, this research may provide a novel solution to other studies related to heterogeneous information fusion, for example, image-text recognition.

In this work, we utilized four weather parameters to improve the cloud classification, and we will investigate how to employ other measurements such as cloud base height for cloud classification in future work. Additionally, we cannot guarantee that the MMFN trained with the MGCD can be generalized well to another dataset, for example the cloud samples gathered from more windy, colder, warmer or on lower ground. Thus, we will utilize unsupervised domain adaptation to enhance the model generalization ability in the future work.

6. Conclusions

In this paper, we have proposed a novel method named MMFN for ground-based cloud recognition. The proposed MMFN has the ability of learning and fusing heterogeneous features under a unified framework. Furthermore, the attentive map is proposed to extract local visual features from salient patterns. To discover the complementary benefit from heterogeneous features, the multi-modal features are integrated with global visual features and local visual features respectively by using two fusion layers. We have also released a new cloud dataset MGCD which includes the cloud images and the multi-modal information. To evaluate the effectiveness of the proposed MMFN, we have conducted a range of experiments and the results demonstrate that the proposed MMFN can stand comparison with the state-of-the-art methods.

Author Contributions: Methodology, S.L. and M.L.; software, M.L.; validation, Z.Z., B.X. and T.S.D.; formal analysis, Z.Z. and B.X.; data curation, S.L., M.L. and Z.Z.; writing—original draft preparation, S.L. and M.L.; writing—review and editing, Z.Z. and T.S.D.; supervision, S.L. and Z.Z. All authors have read and agreed to the published version of the manuscript.

Funding: This work was supported by National Natural Science Foundation of China under Grant No. 61711530240, Natural Science Foundation of Tianjin under Grant No. 19JCZDJC31500, the Fund of Tianjin Normal University under Grant No. 135202RC1703, the Open Projects Program of National Laboratory of Pattern Recognition under Grant No. 202000002, and the Tianjin Higher Education Creative Team Funds Program.

Conflicts of Interest: The authors declare no conflict of interest. The sponsors had no role in the design, execution, interpretation or writing of the study. This research was funded by National Natural Science Foundation of China, grant number 61711530240, Natural Science Foundation of Tianjin, grant number 19JCZDJC31500, the Fund of Tianjin Normal University, grant number 135202RC1703, the Open Projects Program of National Laboratory of Pattern Recognition, grant number 202000002, and the Tianjin Higher Education Creative Team Funds Program. The APC was funded by Natural Science Foundation of Tianjin, grant number 19JCZDJC31500, and the Open Projects Program of National Laboratory of Pattern Recognition, grant number 202000002.

Abbreviations

The following abbreviations are used in this manuscript:

MMFN	Multi-evidence and multi-modal fusion network
MGCD	Multimodal ground-based cloud dataset
TSI	Total-sky imager
CNN	Convolutional neural network
Leaky ReLU	Leaky rectified linear unite
SGD	Stochastic gradient descent
BoVW	Bag-of-visual-words
SIFT	Scale invariant feature transform
LBP	Local binary pattern
CLBP	Completed LBP
DMF	Deep multimodal fusion
JFCNN	Joint fusion convolutional neural network

References

1. Ceppi, P.; Hartmann, D.L. Clouds and the atmospheric circulation response to warming. *J. Clim.* **2016**, *29*, 783–799.
2. Zhou, C.; Zelinka, M.D.; Klein, S.A. Impact of decadal cloud variations on the Earth's energy budget. *Nat. Geosci.* **2016**, *9*, 871.
3. McNeill, V.F. Atmospheric aerosols: Clouds, chemistry, and climate. *Annu. Rev. Chem. Biomol.* **2005**, *8*, 258–271.
4. Huang, W.; Wang, Y.; Chen, X. Cloud detection for high-resolution remote-sensing images of urban areas using colour and edge features based on dual-colour models. *Int. J. Remote Sens.* **2018**, *39*, 6657–6675.
5. Liu, Y.; Tang, Y.; Hua, S.; Luo, R.; Zhu, Q. Features of the cloud base height and determining the threshold of relative humidity over southeast China. *Remote Sens.* **2019**, *11*, 2900.
6. Calbo, J.; Sabburg, J. Feature extraction from whole-sky ground-based images for cloud-type recognition. *J. Atmos. Ocean. Technol.* **2008**, *25*, 3–14.
7. Wang, Y.; Wang, C.; Shi, C.; Xiao, B. A Selection Criterion for the Optimal Resolution of Ground-Based Remote Sensing Cloud Images for Cloud Classification. *IEEE Trans. Geosci. Remote* **2018**, *57*, 1358–1367.
8. Shi, W.; Caballero, J.; Huszár, F.; Totz, J.; Aitken, A.P.; Bishop, R.; Rueckert, D.; Wang, Z. Real-time single image and video super-resolution using an efficient sub-pixel convolutional neural network. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Las Vegas, NV, USA, 27–30 June 2016; pp. 1874–1883.
9. Ryu, A.; Ito, M.; Ishii, H.; Hayashi, Y. Preliminary analysis of short-term solar irradiance forecasting by using total-sky Imager and convolutional neural network. In Proceedings of the IEEE PES GTD Grand International Conference and Exposition Asia, Bangkok, Thailand, 21–23 March 2019; pp. 627–631.
10. Nouri, B.; Wilbert, S.; Segura, L.; Kuhn, P.; Hanrieder, N.; Kazantzidis, A.; Schmidt, T.; Zarzalejo, L.; Blanc, P.; Pitz-Paal, R. Determination of cloud transmittance for all sky imager based solar nowcasting. *Sol. Energy* **2019**, *181*, 251–263.

11. Nouri, B.; Kuhn, P.; Wilbert, S.; Hanrieder, N.; Prah, C.; Zarzalejo, L.; Kazantzidis, A.; Blanc, P.; Pitz-Paal, R. Cloud height and tracking accuracy of three all sky imager systems for individual clouds. *Sol. Energy* **2019**, *177*, 213–228.
12. Liu, S.; Wang, C.; Xiao, B.; Zhang, Z.; Shao, Y. Salient local binary pattern for ground-based cloud classification. *Acta Meteorol. Sin.* **2013**, *27*, 211–220.
13. Cheng H.Y.; Yu, C.C. Multi-model solar irradiance prediction based on automatic cloud classification. *Energy* **2015**, *91*, 579–587.
14. Kliangsuwan, T.; Heednacram, A. Feature extraction techniques for ground-based cloud type classification. *Expert Syst. Appl.* **2015**, *42*, 8294–8303.
15. Cheng, H.Y.; Yu, C.C. Block-based cloud classification with statistical features and distribution of local texture features. *Atmos. Meas. Tech.* **2015**, *8*, 1173–1182.
16. Gan, J.; Lu, W.; Li, Q.; Zhang, Z.; Yang, J.; Ma, Y.; Yao, W. Cloud type classification of total-sky images using duplex norm-bounded sparse coding. *IEEE J.-STARS* **2017**, *10*, 3360–3372.
17. Kliangsuwan, T.; Heednacram, A. A FFT features and hierarchical classification algorithms for cloud images. *Eng. Appl. Artif. Intel.* **2018**, *76*, 40–54.
18. 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.
19. Girshick, R.; Donahue, J.; Darrell, T.; Malik, J. Rich feature hierarchies for accurate object detection and semantic segmentation. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Columbus, OH, USA, 24–27 June 2014; pp. 580–587.
20. He, K.; Zhang, X.; Ren, S.; Sun, J. Deep residual learning for image recognition. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Las Vegas, NV, USA, 27–30 June 2016; pp. 770–778.
21. Liu, C.; Chen, L.C.; Schroff, F.; Adam, H.; Hua, W.; Yuille, A.L.; Fei-Fei, L. Auto-deeplab: Hierarchical neural architecture search for semantic image segmentation. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Long Beach, CA, USA, 16–20 June 2019; pp. 82–92.
22. Choi, J.; Kwon, J.; Lee, K.W. Deep meta learning for real-time target-aware visual tracking. In Proceedings of the IEEE International Conference on Computer Vision, Long Beach, CA, USA, 16–20 June 2019; pp. 911–920.
23. DeLancey, E.R.; Simms, J.F.; Mahdianpari, M.; Brisco, B.; Mahoney, C.; Kariyeva, J. Comparing deep learning and shallow learning for large-scale wetland classification in Alberta, Canada. *Remote Sens.* **2019**, *12*, 2.
24. Wang, Y.; Chen, C.; Ding, M.; Li, J. Real-time dense semantic labeling with dual-Path framework for high-resolution remote sensing image. *Remote Sens.* **2019**, *11*, 3020.
25. Shi, C.; Wang, C.; Wang Y.; Xiao, B. Deep convolutional activations-based features for ground-based cloud classification. *IEEE Geosci. Remote Sens.* **2017**, *14*, 816–820.
26. 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.
27. Li, M.; Liu, S.; Zhang, Z. Dual guided loss for ground-based cloud classification in weather station networks. *IEEE Access* **2019**, *7*, 63081–63088.
28. Ye, L.; Cao, Z.; Xiao, Y. DeepCloud: Ground-based cloud image categorization using deep convolutional features. *IEEE Trans. Geosci. Remote* **2017**, *55*, 5729–5740.
29. Baker, M.B.; Peter, T. Small-scale cloud processes and climate. *Nature* **2008**, *451*, 299.
30. 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.
31. Liu, S.; Li, M. Deep multimodal fusion for ground-based cloud classification in weather station networks. *EURASIP J. Wirel. Comm.* **2018**, *2018*, 48.
32. S. Liu, M. Li, Z. Zhang, B. Xiao and X. Cao, Multimodal ground-based cloud classification using joint fusion convolutional neural network. *Remote Sens.* **2018**, *10*, 822.
33. Li, Q.; Zhang, Z.; Lu, W.; Yang, J.; Ma, Y.; Yao, W. From pixels to patches: A cloud classification method based on a bag of micro-structures. *Atmos. Meas. Technol.* **2016**, *9*, 753–764.
34. Dev, S.; Lee, Y.H.; Winkler, S. Categorization of cloud image patches using an improved texton-based approach. In Proceedings of the IEEE International Conference on Image Processing, Quebec, QC, Canada, 27–30 September 2015; pp. 422–426.

35. Walther, D.; Rutishauser, U.; Koch, C.; Perona, P. On the usefulness of attention for object recognition. In Proceedings of the European Conference on Computer Vision Workshop on Attention and Performance in Computational Vision, Prague, Czech Republic, 15 May 2004; pp. 96–103.
36. Chang, X.; Qian, Y.; Yu, D. Monaural multi-talker speech recognition with attention mechanism and gated convolutional networks. In Proceedings of the Interspeech, Hyderabad, India, 2–6 September 2018; pp. 1586–1590.
37. Chorowski, J.K.; Bahdanau, D.; Serdyuk, D.; Cho, K.; Bengio, Y. Attention-based models for speech recognition. In Proceedings of the Advances in Neural Information Processing Systems, Montreal, QC, Canada, 7–12 December 2015; PP. 577–585.
38. Zhu, Y.; Zhao, C.; Guo, H.; Wang, J.; Zhao, X.; Lu, H. Attention couplenet: Fully convolutional attention coupling network for object detection. *IEEE Trans. Image Process.* **2019**, *28*, 113–126.
39. Chen, L.; Zhang, H.; Xiao, J.; Nie, L.; Shao, J.; Liu, W.; Chua, T.S. Sca-cnn: Spatial and channel-wise attention in convolutional networks for image captioning. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Honolulu, HI, USA, 21–26 July 2017; pp. 5659–5667.
40. Fu, J.; Zheng, H.; Mei, T. Look closer to see better: Recurrent attention convolutional neural network for fine-grained image recognition. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Honolulu, HI, USA, 21–26 July 2017; pp. 4438–4446.
41. Wang, F.; Jiang, M.; Qian, C.; Yang, S.; Li, C.; Zhang, H.; Wang, X.; Tang, X. Residual attention network for image classification. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition; Honolulu, HI, USA, 21–26 July 2017; pp. 3156–3164.
42. Peng, Y.; He, X.; Zhao, J. Object-part attention model for fine-grained image classification. *IEEE Trans. Image Process.* **2018**, *27*, 1487–1500.
43. Fu, J.; Liu, J.; Tian, H.; Li, Y.; Bao, Y.; Fang, Z.; Lu, H. Dual attention network for scene segmentation. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Long Beach, CA, USA, 16–20 June 2019; pp. 3146–3154.
44. Liu, Y.; Liu, Y.; Ding, L. Scene classification based on two-stage deep feature fusion. *IEEE Geosci. Remote Sens.* **2018**, *15*, 183–186.
45. Chen, J.; Chen, Z.; Chi, Z.; Fu, H. Facial expression recognition in video with multiple feature fusion. *IEEE Trans. Affect. Comput.* **2018**, *9*, 38–50.
46. Uddin, M.A.; Lee, Y. Feature fusion of deep spatial features and handcrafted spatiotemporal features for human action recognition. *Sensors* **2019**, *19*, 1599.
47. Chaib, S.; Liu, H.; Gu, Y.; Yao, H. Deep feature fusion for VHR remote sensing scene classification. *IEEE Trans. Geosci. Remote* **2017**, *55*, 4775–4784.
48. Song, W.; Li, S.; Fang, L.; Lu, T. Hyperspectral image classification with deep feature fusion network. *IEEE Trans. Geosci. Remote* **2018**, *56*, 3173–3184.
49. Tang, P.; Wang, H.; Kwong, S. G-MS2F: GoogLeNet based multi-stage feature fusion of deep CNN for scene recognition. *Neurocomputing* **2017**, *225*, 188–197.
50. Szegedy, C.; Liu, W.; Jia, Y.; Sermanet, P.; Reed, S.; Anguelov, D.; Erhan, D.; Vanhoucke, V.; Rabinovich, A. Going deeper with convolutions. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Montreal, QC, Canada, 7–12 June 2015; pp. 1–9.
51. Bodla, N.; Zheng, J.; Xu, H.; Chen, J.C.; Castillo, C.; Chellappa, R. Deep heterogeneous feature fusion for template-based face recognition. In Proceedings of the IEEE Winter Conference on Applications of Computer Vision, Santa Rosa, CA, USA, 27–29 March 2017; pp. 586–595.
52. Chen, Y.; Li, C.; Ghamisi, P.; Jia, X.; Gu, Y. Deep fusion of remote sensing data for accurate classification. *IEEE Geosci. Remote Sens.* **2017**, *14*, 1253–1257.
53. Guo, J.; Song, B.; Zhang, P.; Ma, M.; Luo, W. Affective video content analysis based on multimodal data fusion in heterogeneous networks. *Inform. Fusion* **2019**, *51*, 224–232.
54. Ojala, T.; Pietikainen, M.; Maenpaa, T. Multiresolution gray-scale and rotation invariant texture classification with local binary patterns. *IEEE Trans. Pattern Anal.* **2002**, *24*, 971–987.
55. 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.

56. Csurka, G.; Dance, C.; Fan, L.; Willamowski, J.; Bray, C. Visual categorization with bags of keypoints. In Proceedings of the European Conference on Computer Vision Workshop on Statistical Learning in Computer Vision, Prague, Czech Republic, 11–14 May 2004; pp. 1–16.
57. Lowe, D.G. Distinctive image features from scale-invariant keypoints. *Int. J. Comput. Vis.* **2004**, *60*, 91–110.
58. Lazebnik, S.; Schmid, C.; Ponce, J. Beyond bags of features: Spatial pyramid matching for recognizing natural scene categories. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Las Vegas, NV, USA, 17–22 June 2006; pp. 2169–2178.
59. He, K.; Zhang, X.; Ren, S.; Sun, J. Delving deep into rectifiers: Surpassing human-level performance on imagenet classification. In Proceedings of the IEEE International Conference on Computer Vision; Santiago, Chile, 7–13 December 2015; pp. 1026–1034.
60. Krizhevsky, A.; Sutskever, I.; Hinton, G.E. Imagenet classification with deep convolutional neural networks. In Proceedings of the Advances in Neural Information Processing Systems, Lake Tahoe, NV, USA, 3–6 December 2012; pp. 1097–1105.
61. Chang, C.C.; Lin, C.J. LIBSVM: A library for support vector machines. *ACM Trans. Intel. Syst. Technol.* **2011**, *2*, 27:1–27:27.
62. 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.
63. 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.
64. Huo, J.; Bi, Y.; Lü, D.; Duan, S. Cloud classification and distribution of cloud types in Beijing using Ka-band radar data. *Adv. Atmos. Sci.* **2019**, *36*, 793–803.
65. 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.



© 2020 by the authors. Licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license (<http://creativecommons.org/licenses/by/4.0/>).