Ground-based Remote Sensing Cloud Detection using Dual Pyramid Network and Encoder-Decoder Constraint

Zhong Zhang, *Senior Member, IEEE,* Shuzhen Yang, Shuang Liu, *Senior Member, IEEE,* Xiaozhong Cao, and Tariq S. Durrani, *Fellow, IEEE*

***Abstract*—Many methods for ground-based remote sensing cloud detection learn representation features using the encoder-decoder structure. However, they only consider the information from single scale, which leads to incomplete feature extraction. In this paper, we propose a novel deep network named Dual Pyramid Network (DPNet) for ground-based remote sensing cloud detection, which possesses an encoder-decoder structure with Dual Pyramid Pooling Module (DPPM). Specifically, we process the feature maps of different scales in the encoder through dual pyramid pooling. Then, we fuse the outputs of the dual pyramid pooling in the same pyramid level using the attention fusion. Furthermore, we propose the Encoder-Decoder Constraint (EDC) to relieve information loss in the process of encoding and decoding. It constrains the values and the gradients of probability maps from the encoder and the decoder to be consistent. Since the number of cloud images in the publicly available databases for ground-based remote sensing cloud detection is limited, we release the TJNU Large-scale Cloud Detection Database (TLCDD) which is the largest database in this field. We conduct a series of experiments on TLCDD, and the experimental results verify the effectiveness of the proposed method.**

***Index Terms*—ground-based remote sensing cloud detection, Dual Pyramid Pooling Module, Encoder-Decoder Constraint.**

I. Introduction

CLOUD is an important weather phenomenon, and it has a great influence on the earth's radiation budget and climate change [1], [2]. Hence, cloud observation has drawn a lot of attention from both academia and industry due to its wide applications in weather forecasting and military operations [3], [4]. Cloud observation is mainly classified into the satellite cloud observation and the ground-based cloud observation [5]–[7]. The satellite cloud observation is more suitable for describing large-scale cloud information and changes, while the ground-based cloud observation is good at reflecting local cloud information [8]–[11]. In addition, the ground-based cloud observation has many advantages, such as low equipment cost, simple operation, and easy acquisition of data. The ground-based cloud observation mainly contains three tasks, i.e., cloud shape, cloud cover and cloud base height [12]. In this paper, we focus on ground-based remote sensing cloud detection which is the key technology for cloud cover estimation. There are two reasons for the urgent need to develop an automatic ground-based remote sensing cloud detection algorithm. Firstly, the detection results marked by different weather observers may be inconsistent due to their different skill levels. Secondly, manually labeling cloud images for ground-based remote sensing cloud detection is labour intensive and tedious, because this process is pixel-level labeled. When the ground-based cloud data is huge, the labeling process is very difficult.

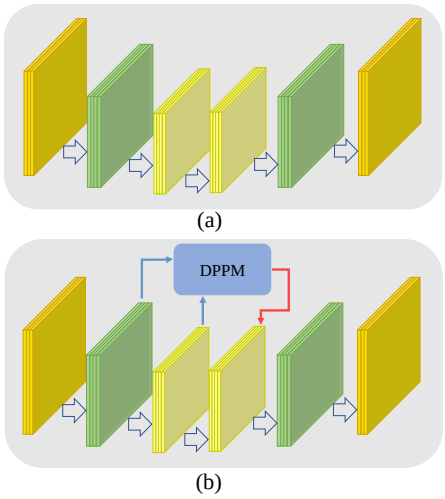

Fig. 1. The structures of (a) common-used deep network and (b) DPNet for ground-based remote sensing cloud detection.

Hence, many methods for ground-based remote sensing cloud detection have been proposed, and they are roughly divided into three categories, namely threshold-based methods, texture-based methods, and deep learning methods. Some threshold-based methods directly treat R and B values as

This work was supported by National Natural Science Foundation of China under Grant No. 62171321, Natural Science Foundation of Tianjin under Grant No. 20JCZDJC00180 and No. 19JCZDJC31500, the Open Projects Program of National Laboratory of Pattern Recognition under Grant No. 202000002, and the Tianjin Higher Education Creative Team Funds Program. (Corresponding author: Shuang Liu.)

Zhong Zhang, Shuzhen Yang, and Shuang Liu are with Tianjin Key Laboratory of Wireless Mobile Communications and Power Transmission, Tianjin Normal University, Tianjin 300387, China (e-mail: zhong.zhang8848@gmail.com, ysz2020zhen@gmail.com, shuangliu.tjnu@gmail.com).

Xiaozhong Cao is with the Meteorological Observation Centre, China Meteorological Administration, Beijing 100081, China (e-mail: xzhongcao@163.com).

Tariq S. Durrani is with the Department of Electronic and Electrical Engineering, University of Strathclyde, Glasgow Scotland, UK (e-mail: t.durrani@strath.ac.uk).

the threshold to distinguish cloud and sky, or employ adaptive thresholds, for example, Ostu algorithm and superpixel segmentation [13], [14]. The texture-based methods utilize the texture features to describe the local regions of cloud images [15], [16]. However, the performance of threshold-based methods and texture-based methods is unsatisfactory, because these methods are not learning-based, which is easily affected by the environmental changes.

Recently, Convolutional Neural Network (CNN) achieves extraordinary performance in many fields such as image recognition, speech analysis and object detection because of its deep network structure and changeable perception field [14], [17], [18]. CNN is also introduced in the field of remote sensing cloud detection, and many researchers [14], [18], [19] design the deep network as a structure with the encoder and the decoder as shown in Fig. 1(a). However, there are two limitations for these methods. Firstly, in the encoding process, they only consider the information from one scale, which leads to incomplete feature extraction. Secondly, the feature maps are conducted by max-pooling operations in the encoding process, which results in information loss. Meanwhile, the convolution operations have a negative impact on edge detail information.

To overcome the first limitation, we propose a novel deep network named Dual Pyramid Network (DPNet) for ground-based remote sensing cloud detection, which possesses an encoder-decoder structure with Dual Pyramid Pooling Module (DPPM) as shown in Fig. 1(b). The proposed DPPM combines the information from different scales of the encoder via fusing dual pyramids. Specifically, we process the feature maps of different scales in the encoder via spatial pyramid pooling. Then, we fuse the feature maps in the same pyramid level from different scales by learning attention weights which reflect the importance of different elements in the feature maps. As a result, we obtain completed features.

Furthermore, we propose the Encoder-Decoder Constraint (EDC) to relieve information loss. The quality of probability maps directly determines the performance of cloud detection, and meanwhile the information communication between the encoder and the decoder could fully utilize the information of them. Hence, the proposed EDC constrains the probability maps from the encoder and the decoder. In order to reflect the detail information and the local boundary, EDC expects the values and the gradients of probability maps from the encoder and the decoder to be consistent, simultaneously.

A large-scale database is necessary for the development of ground-based remote sensing cloud detection algorithms, especially for deep learning based algorithms [20]. The large-scale database could avoid model overfitting and improve the generalization ability of deep model. However, the publicly available databases on ground-based remote sensing cloud detection contain insufficient cloud images, which is difficult to meet actual demand. For example, Singapore All Weather Segmentation (SWIMSEG) database [21], CloudSegmentation database [13], Whole Sky Image SEGmentation (WSISEG) database [22] and BENCHMARK database [23] have 1013, 100, 400 and 32 cloud images, respectively. In this paper, we release the TJNU Large-scale Cloud Detection Database (TLCDD) consisting of 5000 cloud images. To the best of our knowledge, TLCDD is the largest database for ground-based remote sensing cloud detection.

The contributions of this paper are summarized into three aspects. Firstly, we propose DPNet to construct dual pyramids in the encoder in order to fuse the information from different scales. Secondly, we propose EDC to constrain the feature maps of the encoder and the decoder so as to relieve information loss. Finally, we release the largest cloud detection database, i.e., TLCDD, and the proposed method achieves better performance than other state-of-the-art methods on TLCDD.

II. Related Work

A. Ground-based Remote Sensing Cloud Detection

At present, more and more researchers are devoted to the ground-based remote sensing cloud detection. These studies are mainly composed of threshold-based methods, texture-based methods, and deep learning methods. The threshold-based methods usually adopt RGB color values as criteria to distinguish cloud and sky. For example, Long *et al.* [2] and Kreuter *et al.* [24] proposed to utilize the thresholds of 0.6 and 0.77 on R/B for cloud detection. When the ratio is over 0.6 or 0.77, the pixel is identified as cloud. Souzaecher *et al.* [25] recommended to employ B-R for identifying cloud, and the pixels with B-R$>$30 are classified as sky. The above methods directly utilize the fixed threshold to detect cloud, and easily affected by environmental changes. To overcome the drawback, some researchers present adaptive threshold algorithms. Yang *et al.* [26] calculated the adaptive threshold based on the B-R feature image using the Otsu algorithm. The superpixel segmentation algorithm [13], [27] was utilized to divide the cloud image into a series of subregions, and further detect cloud in each subregion. Furthermore, texture feature extraction as a better kind of methods is used in cloud detection. For example, Başeski *et al.* [15] applied the Homogenous Texture Descriptor (HTD) as the complement of color features for cloud detection. The HTD could describe the regularity, directionality and coarseness of texture. Tulpan *et al.* [16] proposed to utilize six kinds of image moments for cloud detection, where the image moments include the area of the image, two edge detectors, a cross detector, and the elongation and direction of the image. The threshold-based methods and the texture-based methods solve the difficulty of manually labeling cloud pixels to a certain extent, but the performance is unsatisfactory. Thus, the ground-based remote sensing cloud detection algorithms still need to be improved.

CNN possesses the strong capability of feature representation, so it has been widely used in many research fields with excellent performance [17], [19], [28], [29]. Inspired by this, many researchers design different network structures under the framework of CNN to improve the performance of ground-based remote sensing cloud detection. For example, Dev *et al.* [18] proposed the CloudSegNet where the basic structure is designed as the encoder-decoder structure. In the training stage, CloudSegNet is optimized by daytime and nighttime cloud images. Xie *et al.* [14] presented the SegCloud

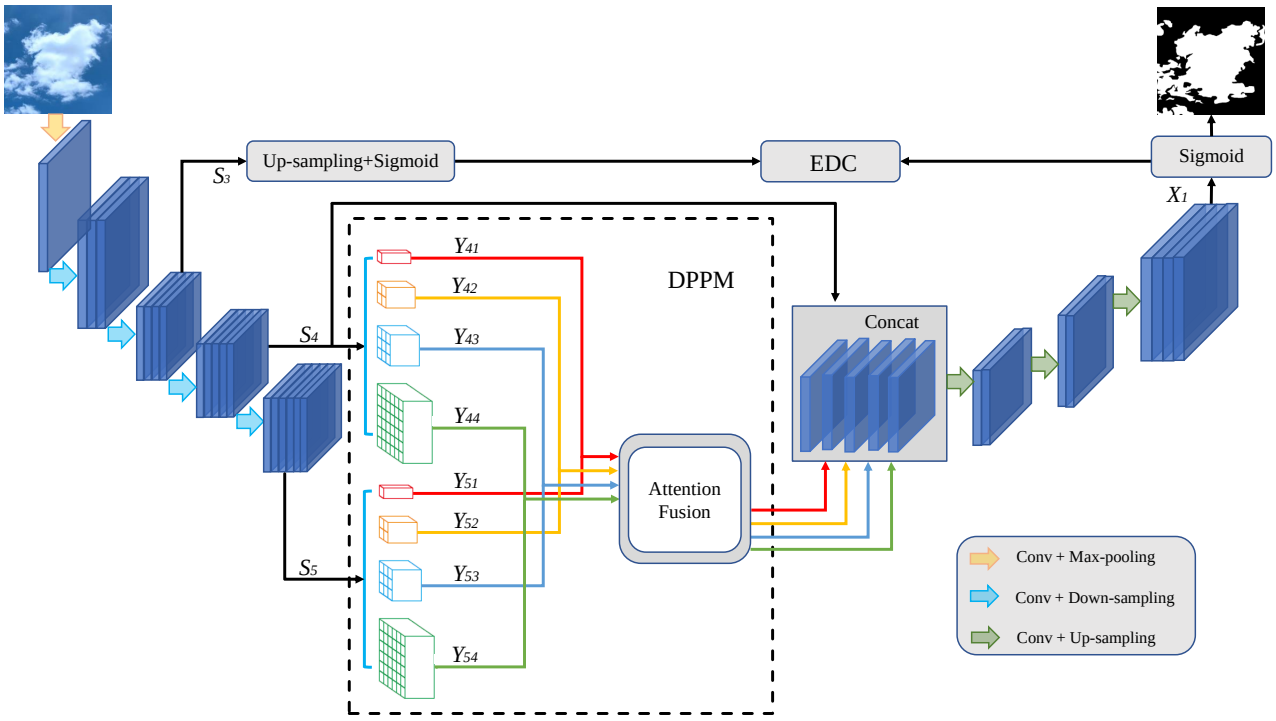

Fig. 2. The framework of DPNet. S_3 S_4 and S_5 are the feature maps from $Scale3$, $Scale4$ and $Scale5$ respectively, Y_{4i} and Y_{5i} are the outputs of pyramid pooling and X_1 is the feature maps of the decoder.

model which possesses a symmetric encoder-decoder structure followed by a softmax classifier. The softmax classifier realizes the pixel classification and outputs the segmentation results. Zhang *et al.* [30] proposed the Multi-scale Attention Convolutional Neural Network (MACNN) for cloud detection by exploiting multi-scale information and attention connection between the encoder and the decoder.

B. Encoder-Decoder Structure for Semantic Segmentation

The task of cloud detection is to classify each pixel of cloud image into cloud or sky, which is regarded as a two-category segmentation problem. Hence, we introduce semantic segmentation [31]–[34] in this subsection. The encoder-decoder structure which mainly includes an encoder and a decoder dominates the semantic segmentation task [18]. The encoder maps an image to a specific high-dimensional feature to capture semantic information. The decoder gradually transforms the high-dimensional feature into the score map for the sequence segmentation, and it restores object detail and spatial information. The high-dimensional feature is treated as the bridge between the original image and the score map. Furthermore, the skip connection is usually inserted into the encoder-decoder structure, and it realizes the feature fusion between the encoder and the decoder [35], which is beneficial to preserve the detail information from the encoder [23], [36].

Long *et al.* [37] presented the Fully Convolutional Network (FCN) for semantic segmentation, which could accept the image with any size. It utilizes the deconvolution layer to upsample the feature maps in the last layer in order to restore to the size of input image. The widely used U-Net [23] is a typical encoder-decoder network, in which the encoder contains convolution and max-pooling operations and the decoder restores the feature maps to the original resolution through convolution and up-sampling operations. Some methods [28], [34], [38] presented the spatial pyramid structure under the framework of the encoder-decoder structure to aggregate the context information based on different regions. However, most existing encoder-decoder methods extract incomplete features in the encoding processing, and meanwhile they suffer from the information loss in the encoding and decoding process. Therefore, we propose DPNet and EDC to overcome these limitations.

III. Approach

In this section, we first present an overview of the proposed DPNet as shown in Fig. 2. We then describe the major parts of DPNet, i.e., encoder-decoder structure and DPPM. Finally, we introduce how to implement EDC.

A. Overview of DPNet

Encoder-Decoder Structure. We apply the encoder-decoder structure to conduct the pixel labeling in the cloud image. The encoder is designed as the common used ResNet-50 [39] which utilizes the max-pooling operations and the convolution operations to continuously reduce the size of feature maps and increase the number of channels. The decoder employs the up-sampling operations to increase the size of feature maps, continuously.

Dual Pyramid Pooling Module. The proposed DPPM aims to extract completed information from cloud images during

TABLE I
THE STRUCTURE OF ENCODER.

Name	Input Size	Filters	Output Size
Scale1	512×512	$\begin{bmatrix} 3\times3, 64, s=2 \\ 3\times3, 64, s=1 \\ 3\times3, 128, s=1 \end{bmatrix} \times 1$	256×256
Max-pooling	256×256	$3 \times 3,\ s = 2$	128×128
Scale2	128×128	$\begin{bmatrix} 1\times1, 64, s=1 \\ 3\times3, 64, s=1 \\ 1\times1, 256, s=1 \end{bmatrix} \times 3$	128×128
Scale3	128×128	$\begin{bmatrix} 1\times1, 128, s=1 \\ 3\times3, 128, s=1 \\ 1\times1, 512, s=1 \end{bmatrix} \times 4$	128×128
Scale4	64×64	$\begin{bmatrix} 1\times1, 256, s=1 \\ 3\times3, 256, s=1 \\ 1\times1, 1024, s=1 \end{bmatrix} \times 6$	64×64
Scale5	64×64	$\begin{bmatrix} 1\times1, 512, s=1 \\ 3\times3, 512, s=1 \\ 1\times1, 2048, s=1 \end{bmatrix} \times 3$	64×64

the encoding process. The feature maps from two scales are conducted by the pyramid pooling to obtain different pyramid levels. Afterwards, we apply attention mechanism to fuse the feature maps from different scales under the same pyramid level. In this way, each pixel is assigned to different attention weight, which is beneficial to the subsequent decoding process.

Encoder-Decoder Constraint. The information loss occurs in the process of encoding and decoding, and therefore we exchange the information between them to overcome this drawback. The proposed EDC contains two constraints which expect the probability maps from the encoder and the decoder to be consistent from different aspects.

B. Encoder-Decoder Structure

The proposed DPNet utilizes an asymmetric encoder-decoder structure. The encoder consists of five scales, and each scale contains several blocks. The detailed information of the encoder is listed in Table I. Here, s represents the size of stride. $Scale2$-$Scale5$ include 3, 4, 6, and 3 blocks respectively, and each block contains three convolutional layers. Taking $Scale5$ as an example, it contains 3 blocks where each block consists of three convolutional layers with the sizes of 1×1, 3×3 and 1×1, and the number of filters are 512, 512 and 2048, respectively.

The decoder is composed of three up-sampling layers and three convolutional blocks, and the structure is shown in Table II. After each up-sampling layer, the size of feature maps is doubled, and the decoder outputs the feature maps which have the same size of the input image.

TABLE II
THE STRUCTURE OF DECODER.

Name	Input Size	Filters	Output Size
Up-sampling	64×64	$2 \times 2,\ s = 2$	128×128
Conv	128×128	$\left[3 \times 3, 1024, s = 1\right] \times 2$	128×128
Up-sampling	128×128	$2 \times 2,\ s = 2$	256×256
Conv	256×256	$\left[3 \times 3, 256, s = 1\right] \times 2$	256×256
Up-sampling	256×256	$2 \times 2,\ s = 2$	512×512
Conv	512×512	$\begin{bmatrix} 3\times3, 64, s=1 \\ 3\times3, 64, s=1 \\ 3\times3, 16, s=1 \\ 3\times3, 1, s=1 \end{bmatrix}$	512×512

C. Dual Pyramid Pooling Module

The pyramid pooling [40] is usually inserted into the segmentation networks, such as PSPNet to exploit contextual information by pooling feature maps at different pyramid levels. The pyramid pooling is formulated as:

$$Y_i = P(S, k_i), i = 1, 2, 3, 4 \tag{1}$$

where P refers to the average pooling, S represents the feature maps, and k_i indicates the i-th pyramid level. Normally, there are four pyramid levels ($i = 1, 2, 3, 4$), and the bin sizes of four pyramid levels are 1×1, 2×2, 3×3, and 6×6, respectively. However, the pyramid pooling only considers the feature maps from one scale, and ignores the information from different scales. Hence, we propose the dual pyramid pooling which conducts the pyramid pooling on the feature maps from different scales and fuse them via the attention fusion. It is formulated as:

$$Z_i = A(P(S_4, k_i), P(S_5, k_i)), i = 1, 2, 3, 4 \tag{2}$$

where S_4 and S_5 are the feature maps from $Scale4$ and $Scale5$, and A indicates the attention fusion.

From Eq. 2, we can see that the feature maps from different scales should be fused together. Some segmentation networks, such as U-Net [23] and FCN [37] are usually direct addition or concatenation to fuse the feature maps. They treat all elements in the feature maps equally, and ignore the importance of different elements. Hence, we propose the attention fusion to assign different weights to the elements in order to fuse the feature maps from different scales after pyramid pooling. The attention fusion is formulated as:

$$A(Y_{4i}, Y_{5i}) = W_i Y_{4i} + Y_{5i} \tag{3}$$

where W_i is the attetion coefficient of the i-th pyramid level.

Fig. 3 shows the flowchart of attention fusion, where the feature maps Y_{4i} and Y_{5i} are the outputs of pyramid pooling from $Scale4$ and $Scale5$ respectively. Y_{4i} and Y_{5i} are processed by the up-sampling operation, the convolutional layer with the kernel size of 1×1, and the rectified linear unit (ReLU) activation, respectively. Afterwards, the obtained feature maps

are added in the element-wise manner, and then fed into the convolutional layer with the kernel size of 1×1 and the ReLU activation to obtain the attention coefficient W_i.

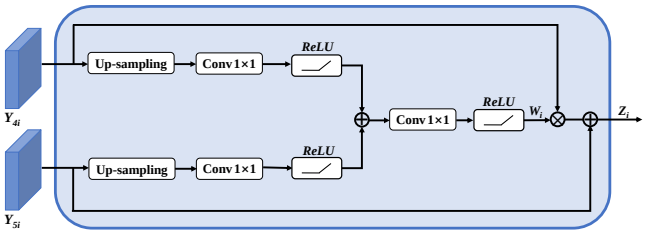

Fig. 3. The flowchart of the attention fusion.

D. Encoder-Decoder Constraint

The encoder-decoder structure dominates the field of cloud detection, but this structure easily causes the information loss due to the following two reasons. Firstly, the max-pooling in the encoder reduces the size of feature maps, which results in the information loss. Secondly, the convolutional layers have a negative impact on edge detail information. Furthermore, after a series of convolutional layers, it is hard to find the corresponding position in the cloud image.

In order to solve the above-mentioned issues, we propose EDC to constrain the probability maps, which consists of two constraint terms. The first constraint term of EDC focuses on constraining the probability maps from the encoder and the decoder. Specifically, the feature maps S_3 in the encoder are fed into the up-sampling layer and the Sigmoid function, and then we obtain the probability map $S_3^{'}$ which is the same size as the input image. The feature maps X_1 in the decoder are input into the Sigmoid function to obtain the probability map $X_1^{'}$. Then, this constraint expects that the probability maps from the encoder and the decoder are consistent:

$$L_1 = \frac{1}{H \times W} ||S_3^{'} - X_1^{'}||_1 \tag{4}$$

where H and W are the height and the width of the probability maps respectively, and $|| \cdot ||_1$ is the l_1 norm of matrix.

The edge information is vital to the ground-based remote sensing cloud detection, and therefore the second term of EDC utilizes the gradients of probability maps of the encoder and the decoder. It is formulated as:

$$L_2 = \frac{1}{H \times W} ||G(S_3^{'}) - G(X_1^{'})||_2 \tag{5}$$

where $|| \cdot ||_2$ is the l_2 norm of matrix, and G is the Prewitt operator [41] which is utilized to compute the gradient. G consists of two templates G_x and G_y, where G_x detects horizontal edges and G_y detects vertical edges. They are defined as:

$$G_x = \begin{pmatrix} 1, & 0, & -1 \\ 1, & 0, & -1 \\ 1, & 0, & -1 \end{pmatrix} \tag{6}$$

$$G_y = \begin{pmatrix} 1, & 1, & 1 \\ 0, & 0, & 0 \\ -1, & -1, & -1 \end{pmatrix} \tag{7}$$

The expressions of the probability maps $S_3^{'}$ and $X_1^{'}$ after going through G are:

$$G(S_3^{'}) = G_x * S_3^{'} + G_y * S_3^{'} \tag{8}$$

$$G(X_1^{'}) = G_x * X_1^{'} + G_y * X_1^{'} \tag{9}$$

where $*$ represents the convolution operation.

The loss of EDC is expressed as:

$$L_E = L_1 + \alpha L_2 \tag{10}$$

where α is the parameter to balance the two constraints.

Furthermore, we apply the binary cross-entropy loss after the probability map $X_1^{'}$ to optimize the network:

$$L_G = -\frac{1}{H \times W} \sum_{i=1}^{H \times W} y_i \ln x_i + (1 - y_i) \ln(1 - x_i) \tag{11}$$

where y_i is the ground-truth label, and x_i is the element value of $X_1^{'}$. In a word, the total loss of the proposed method is formulated as:

$$L = L_G + \beta L_E \tag{12}$$

where β is the parameter to balance the importance of different components.

IV. Experiments

In this section, we evaluate the performance of the proposed method on TLCDD. We first introduce the TLCDD and the implementation details of our experiments. Afterwards, we show the experimental results to verify the superiority of the proposed method.

A. TLCDD

The TLCDD which consists of 5000 cloud images is utilized to study the ground-based remote sensing cloud detection. There are 4208 images for training and 792 images for testing. It has no cloud image overlap between the training set and the test set. Each cloud image in the database corresponds to a ground-truth cloud mask which is jointly annotated by meteorologists and cloud-related researchers. The cloud image is stored in PNG format with a pixel resolution of 512×512. The collection of all the images in the database lasted for two years and came from nine provinces of China including Tianjin, Anhui, Sichuan, Gansu, Shandong, Hebei, Liaoning, Jiangsu, and Hainan. As a result, the TLCDD guarantees the diversity of cloud images, which makes the experimental results convincing. Fig. 4 illustrates the cloud images and the corresponding ground-truth cloud masks.

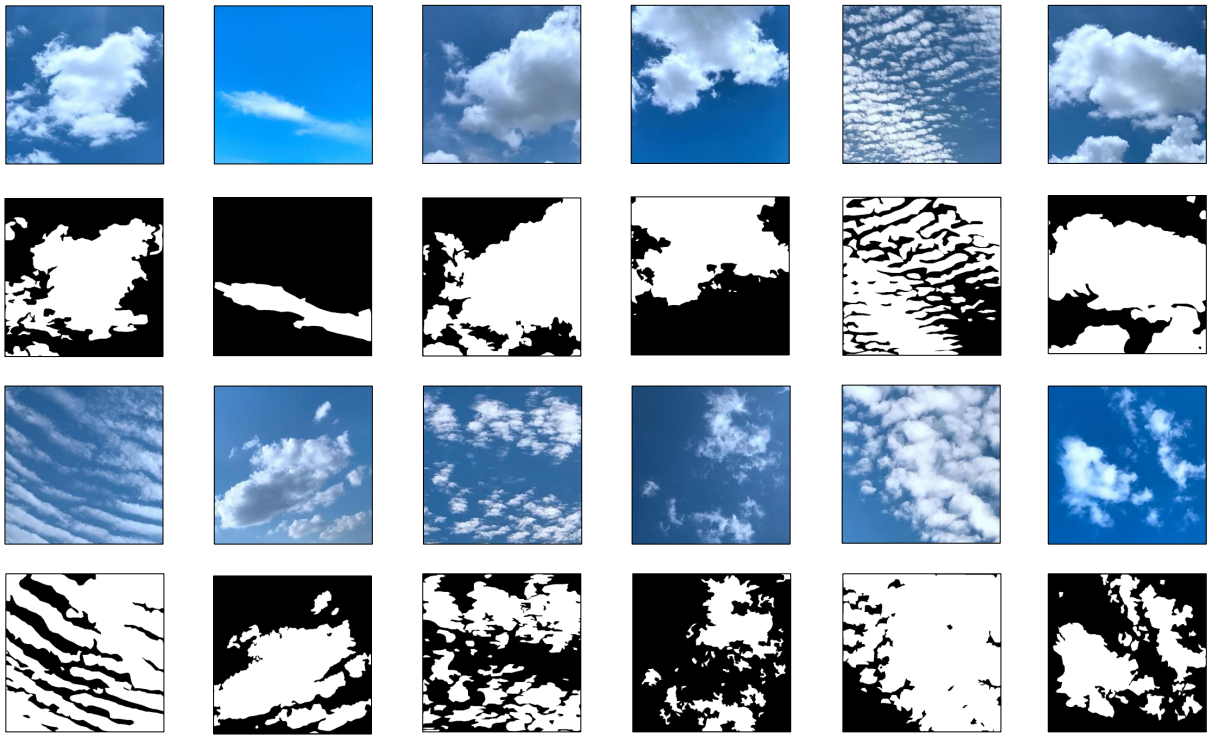

Fig. 4. Several cloud samples in TLCDD.

B. Implementation Details and Evaluation Criterions

Before feeding the cloud images into the deep model, we conduct the preprocessing operations. Specifically, the cloud images are normalized by the mean values and the standard deviation values. The horizontal flip is conducted with the probability of 0.5. The size of images is 512×512. The encoder of DPNet is initialized by the pre-trained ResNet-50. Specifically, $Scale2 - Scale5$ correspond to $conv2x$-$conv5x$ in the ResNet-50 model, respectively. The proposed deep network is optimized by the SGD algorithm with the weight decay of 10^{-9} and the momentum of 0.9. In the training phase, the initial learning rate is set to 0.001 , and the number of training epochs is set to 45. In addition, the hyper-parameter α in Eq. 10 is equal to 1.1 , and β in Eq. 12 is equal to 0.4.

In order to verify the effectiveness of the proposed method, five quantitative evaluation criterions, i.e., Precision (Pre), Recall (Rec), F-score (F-s), Accuracy (Acc) and IoU are applied. The Precision refers to the pixels that are correctly predicted as the cloud accounting for the pixels that are predicted as the cloud in the image. The Recall refers to the proportion of pixels correctly predicted as cloud to all ground-truth cloud pixels in the image. The F-score considers both Recall and Precision, and it is interpreted as the harmonic mean of Precision and Recall. The Accuracy refers to the proportion of pixels that are correctly predicted as cloud and sky to all pixels in the image. We also consider IoU in the evaluation criterions for the cloud detection task. It quantifies a ratio of overlap between the intersection and the union of two sets. The two sets indicate the set of predicted cloud pixels and the set of ground-truth cloud pixels. The ratio can also be interpreted as the number of true positives over the sum of true positives, false positives, and false negatives. In a word, the five evaluation criterions are defined as:

$$Pre = \frac{TP}{TP + FP} \tag{13}$$

$$Rec = \frac{TP}{TP + FN} \tag{14}$$

$$F - s = \frac{2 \times Pre \times Rec}{(Pre + Rec)} \tag{15}$$

$$Acc = \frac{TP + TN}{TP + TN + FP + FN} \tag{16}$$

$$IoU = \frac{TP}{TP + FP + FN} \tag{17}$$

where TP, FP, TN and FN denote true positives, false positives, true negatives and false negatives, respectively.

C. Experimental Results

1) Ablation Studies: The advantage of the proposed DPNet is to learn rich and accurate features for cloud detection. We conduct ablation studies to verify the role of different components in DPNet, namely DPPM and EDC.

Framework1 only utilizes the encoder-decoder structure to detect clouds, that is it does not apply DPPM and EDC.

TABLE III
COMPARISONS WITH DIFFERENT ABLATION METHODS.

Methods	Pre (%)	Rec (%)	F-s (%)	Acc (%)	IoU (%)
$Framework1$	67.21	76.24	66.21	76.83	56.82
$Framework2$	68.35	78.19	67.46	78.61	57.27
$Framework3$	69.18	79.93	69.09	79.57	59.82
$Framework4$	70.12	81.07	71.12	81.43	62.07
$Framework5$	69.18	80.13	71.05	79.81	60.37
$Framework6$	69.02	79.07	70.12	79.43	60.66
$Framework7$	70.16	80.33	71.77	80.20	62.47
Ours	**72.09**	**82.18**	**72.96**	**85.70**	**64.38**

Framework2 employs the encoder-decoder structure with one pyramid pooling to extract the information from one scale. This structure is similar to PSPNet.

Framework3 applies dual pyramid pooling to extract the features from different scales and directly concatenate them without the attention fusion.

Framework4 uses the proposed DPPM to learn the features in the encoder, where the proposed DPPM contains the dual pyramid pooling and the attention fusion.

Framework5 conducts the cloud detection using the encoder-decoder structure with the first constraint term of EDC.

Framework6 inserts the second constraint term of EDC into the encoder-decoder network.

Framework7 employs the proposed EDC to constrain the encoder and the decoder.

The results of ablation studies are listed in Table III from which we can draw four conclusions. Firstly, our method achieves the best results because we combine the encoder-decoder structure with DPPM and EDC. Secondly, the performance of Framework2 and Framework3 is better than that of Framework1, which demonstrates the pyramid pooling strategy is effective to the cloud detection task. Meanwhile, the performance of Framework3 is better than that of Framework2, because the dual pyramid pooling could extract the features from different scales while the pyramid pooling learns features only from one scale. As a result, the dual pyramid pooling obtains richer and more complete information which is beneficial to ground-based remote sensing cloud detection. Thirdly, Framework4 obtains better results than Framework3 due to the attention fusion which assigns different weights to the elements of feature maps.

Finally, Framework5 and Framework6 are obtained by adding the first and second constraints of EDC on the basis of Framework1, respectively. They obtain better performance than Framework1, which verifies the effectiveness of the constraints between the encoder and the decoder. Furthermore, the results of Framework7 are superior to Framework5 and Framework6, which proves that the combination of the two constraints, i.e., EDC has a further performance improvement. Furthermore, we also study the influence of cloud image preprocessing. The experimental results are listed in Table IV where we can see that the results with preprocessing is better than without preprocessing

TABLE IV
THE RESULTS OF THE INFLUENCE OF PREPROCESSING. "WITH PRE" AND "WITHOUT PRE" INDICATE THE CLOUD IMAGE WITH PREPROCESSING AND WITHOUT PROCESSING, RESPECTIVELY.

Methods	Pre (%)	Rec (%)	F-s (%)	Acc (%)	IoU (%)	Time (Hours)
$With\ Pre$	**72.09**	**82.18**	**72.96**	**85.70**	**64.38**	**18.84**
$Without\ Pre$	71.34	81.65	71.64	84.21	63.98	20.46

2) Comparisons with State-of-the-Art Methods: We compare the proposed method with other methods and the results of the evaluation criterions are listed in Table V. These compared methods contain threshold-based methods and deep learning methods. The threshold-based methods usually include R/B (0.77) [24], B-R (30) [25], and Otsu [26]. The first two methods belong to the fixed threshold algorithms which perform different operations on R channel and B channel. Otsu is an adaptive threshold algorithm, which performs the segmentation task on the grayscale image, such as B-R by maximizing the inter-class variance.

We also compare the proposed method with the deep learning methods, for example, FCN [37], U-Net [23], CloudSegNet [18] and SegCloud [14]. FCN is the first network with fully convolutional layers for pixel-wise prediction. It utilizes five down-sampling blocks to extract the feature maps, and three deconvolution layers to restore the feature maps. It defines the skip architecture to combine deep-semantic information and shallow-appearance information. U-Net is a symmetrical encoder-decoder network which has four max-pooling blocks and four up-sampling blocks. It also utilizes the skip architecture on each corresponding convolutional block. CloudSegNet is composed of the encoder including three convolutional layers and three max-pooling layers, and the decoder including four deconvolution layers and three up-sampling layers. SegCloud consists of 10 convolutional layers and 5 max-pooling layers in the encoder, and 5 up-sampling layers and 10 convolutional layers in the decoder. Then, the outputs of decoder are fed into a softmax classifier.

From Table V, we can draw the following conclusions. Firstly, the proposed method achieves the best results. Specifically, it outperforms the second highest results by 3.29%, 1.68%, 5.64%, 7.06% and 6.22% in Pre, Rec, F-s, Acc and IoU, respectively. Secondly, the adaptive threshold method achieves better performance than the fixed threshold methods, because the adaptive threshold could vary with different cloud images. Thirdly, the detection results of the deep learning methods are better than those of the threshold-based methods. It is because the deep learning methods automatically learn features from cloud images through multiple layers. Meanwhile, the threshold-based methods directly apply the thresholds on the cloud images without feature learning, which is difficult to adapt to the environmental changes.

In order to intuitively observe the effectiveness of the proposed method, we show the predicted cloud masks of different methods in Fig. 5. From the figure, it can be seen that the detection results of the deep learning methods (column

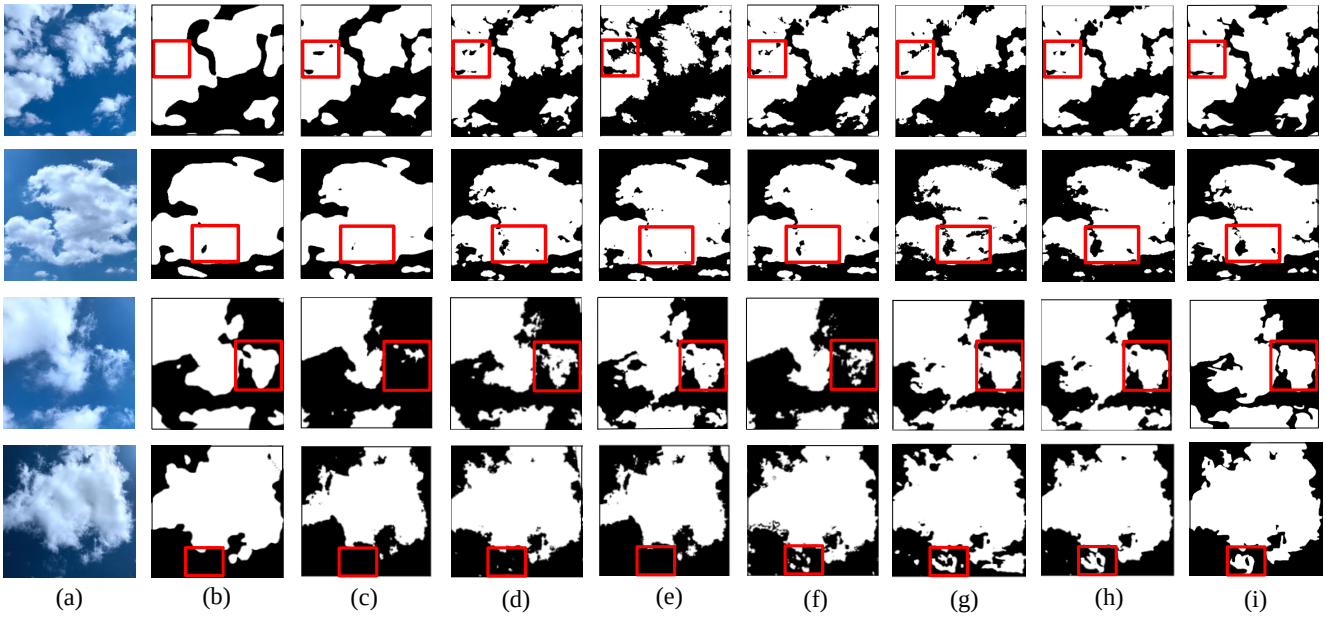

Fig. 5. The predicted cloud masks of different methods. (a) input images, (b) R/B (0.77) , (c) Otsu, (d) FCN, (e) U-Net, (f) SegCloud, (g) PSPNet, (h) Ours, (i) ground-truth cloud masks.

TABLE V
THE EVALUATION RESULTS OF DIFFERENT METHODS ON TCLDD.

Methods	Pre (%)	Rec (%)	F-s (%)	Acc (%)	IoU (%)
R/B (0.77) ([24])	65.88	22.55	25.54	69.11	18.95
B-R (30) ([25])	50.08	13.75	15.08	66.41	11.49
Otsu (B-R) ([26])	57.91	61.47	50.80	66.92	38.34
FCN ([37])	63.20	73.77	57.00	66.49	46.75
CloudSegNet ([18])	64.46	77.61	57.79	64.59	47.78
U-Net ([23])	68.80	80.43	67.32	74.13	58.16
SegCloud ([14])	68.35	80.50	66.95	73.06	57.76
PSPNet ([26])	68.74	77.75	67.00	78.64	57.43
Ours	**72.09**	**82.18**	**72.96**	**85.70**	**64.38**

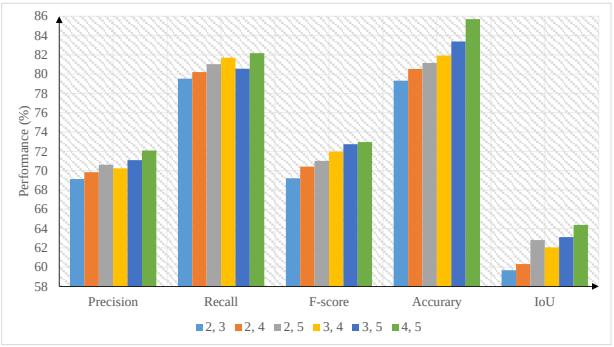

Fig. 6. The detection results with different inputs of DPPM.

(d) - (h)) are better than those of the threshold-based methods (column (b) - (c)). The proposed method shows promising performance in the difficult regions, for example the red rectangles in column (b) - (i) .

3) Parameter Analysis: In this subsection, we evaluate the the input of DPPM and the influence of the hyper-parameters including the number of pyramid levels in DPPM, the coefficients α in Eq. 10 and β in Eq. 12.

The input of DPPM. In this paper, we propose the dual pyramid pooling on the feature maps from two different scales. However, which two scales are selected is important for the detection results. We conduct the experiments with different two scales and the results are illustrated in Fig. 6. From the figure we can see that it is reasonable to choose $Scale4$ and $Scale5$ as the input of DPPM.

The number of pyramid levels in DPPM. The number of pyramid levels in DPPM is related to extract and aggregate the feature maps, and therefore we conduct experiments with different number of pyramid levels. The results are shown in Fig. 7 where we can see that when the number of pyramid levels in DPPM is set to 4 and the bin sizes are 1×1, 2×2, 3×3 and 6×6 respectively, the performance is the best.

The coefficient α in Eq. 10. The coefficient α is used to balance the two constraints in EDC, and the detection results with different α are listed in Fig. 8. The detection results increase when α gets larger, while the detection results decrease after 1.1. Hence, we set α to 1.1 in the experiments.

The coefficient β in Eq. 12. The detection results with different β are illustrated in Fig. 9. It can be seen that when β is equal to 0.4, the proposed method achieves the highest detection results.

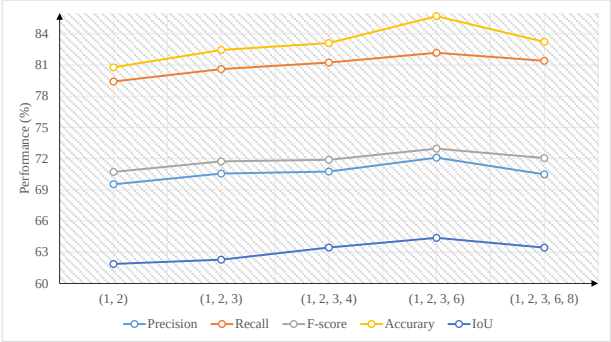

Fig. 7. The detection results with different number of pyramid levels in DPPM. The numbers in the bracket indicates the number of pyramid levels and the bin sizes of pyramid levels. For example, (1, 2, 3) represents that there are three pyramid levels and the bin sizes of the three pyramid levels are 1×1, 2×2 and 3×3, respectively.

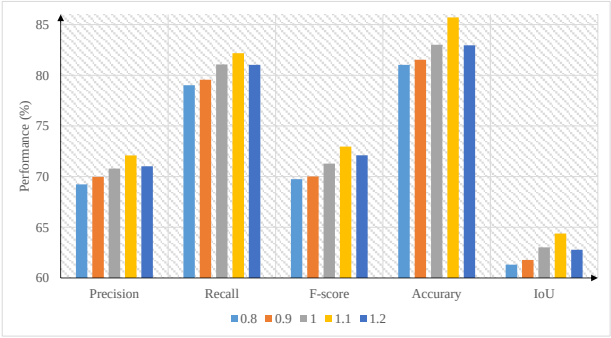

Fig. 8. The detection results with different α in Eq. 10.

V. Conclusion

In this paper, we have proposed DPNet for ground-based remote sensing cloud detection. Specifically, we first learn the feature maps from different scales using the encoder network, and then we feed the feature maps of two scales into DPPM which is composed of the dual pyramid pooling and the attention fusion to obtain complete and discriminative features. In order to solve the problem of information loss, we propose EDC to constraint the information of probability maps from the encoder and the decoder. In addition, we release the largest ground-based cloud database TLCDD, which is necessary to promote the research of ground-based remote sensing cloud detection. The experiments on TLCDD have demonstrated the effectiveness of the proposed method.

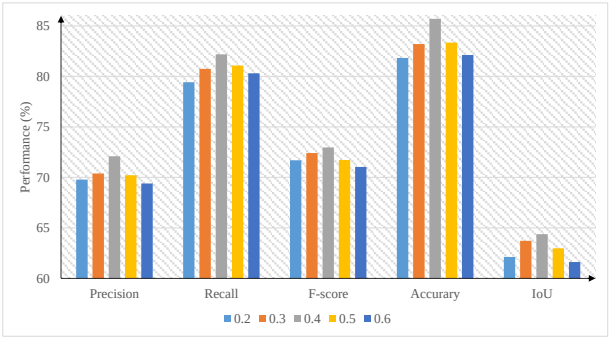

Fig. 9. The detection results with different β in Eq. 12.

References

[1] Y. Wang, C. Wang, C. Shi, and B. Xiao, "A selection criterion for the optimal resolution of ground-based remote sensing cloud images for cloud classification," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 57, no. 3, pp. 4062-4076, 2019.

[2] C. N. Long, J. M. Sabburg, J. Calb, and D. Pages, "Retrieving cloud characteristics from ground-based daytime color all-sky images," *Journal of Atmospheric and Oceanic Technology*, vol. 23, no. 5, pp. 633-652, 2006.

[3] Z. Kundzewicz, "Climate change impacts on the hydrological cycle," *Ecohydrology & Hydrobiology*, vol. 8, no. 2, pp. 195-203, 2008.

[4] G. Horváth, A. Barta, J. Gál, B. Suhai, and O. Haiman, "Ground-based full-sky imaging polarimetry of rapidly changing skies and its use for polarimetric cloud detection," *Applied Optics*, vol. 41, no. 3, pp. 543-559, 2002.

[5] G. Pfister, R. L. McKenzie, J. B. Liley, A. Thomas, B. W. Forgan, and C. N. Long, "Cloud coverage based on all-sky imaging and its impact on surface solar irradiance," *Journal of Applied Meteorology*, vol. 42, no. 10, pp. 1421-1434, 2003.

[6] J. Kalisch, and A. Macke, "Estimation of the total cloud cover with high temporal resolution and parametrization of short-term fluctuations of sea surface insolation," *Meteorologische Zeitschrift*, vol. 17, no. 5, pp. 603-611, 2008.

[7] Z. Shao, Y. Pan, C. Diao, and J. Cai, "Cloud detection in remote sensing images based on multiscale features-convolutional neural network," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 57, no. 6, pp. 4062-4076, 2019.

[8] J. R. Norris, R. J. Allen, A. T. Evan, M. D. Zelinka, C. W. O'Dell, and S. A. Klein, "Evidence for climate change in the satellite cloud record," *Nature*, vol. 536, no. 7614, pp. 72-75, 2016.

[9] B. Zhong, W. Chen, S. Wu, L. Hu, X. Luo, and Q. Liu, "A cloud detection method based on relationship between objects of cloud and cloud-shadow for chinese moderate to high resolution satellite imagery," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 10, no. 11, pp. 4898-4908, 2017.

[10] A. H. Young, K. R. Knapp, A. Inamdar, W. Hankins, and W. B. Rossow, "The international satellite cloud climatology project H-Series climate data record product," *Earth System Science Data*, vol. 10, no. 1, pp. 583-593, 2018.

[11] B. Nouri, S. Wilbert, L. Segura, P. Kuhn, N. Hanrieder, A. Kazantzidis, T. Schmidt, L. Zarzalejo, P. Blanc, and R. Pitz-Paal, "Determination of cloud transmittance for all sky imager based solar nowcasting," *Solar Energy*, vol. 181, pp. 251-263, 2019.

[12] L. Ye, Z. Cao, and Y. Xiao, "DeepCloud: Ground-based cloud image categorization using deep convolutional features," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 55, no. 10, pp. 5729–5740, 2017.

[13] C. Shi, Y. Wang, C. Wang, and B. Xiao, "Ground-based cloud detection using graph model built upon superpixels," *IEEE Geoscience and Remote Sensing Letters*, vol. 14, no. 5, pp. 557–567, 2017.

[14] W. Xie, D. Liu, M. Yang, S. Chen, B. Wang, and Z. Wang, "SegCloud: A novel cloud image segmentation model using a deep convolutional neural network for ground-based all-sky-view camera observation," *Atmospheric Measurement Techniques*, vol. 3, no. 13, pp. 1953-1961, 2020.

[15] E. Başeski and Ç. Cenaras, "Texture and color based cloud detection," in *Proceedings of International Conference on Recent Advances in Space Technologies*, 2015, pp. 311-315.

[16] D. Tulpan, C. Bouchard, K. Ellis and C. Minwalla, "Detection of clouds in sky/cloud and aerial images using moment based texture segmentation," in *Proceedings of International Conference on Unmanned Aircraft Systems*, 2017, pp. 1124-1133.

[17] J. Dai, K. He, and J. Sun, "Convolutional feature masking for joint object and stuff segmentation," in *Proceedings of IEEE Conference on Computer Vision and Pattern Recognition*, 2015, pp. 3992-4000.

[18] S. Dev, A. Nautiyal, Y. Lee, and S. Winkler, "CloudSegNet: A deep network for nychthemeron cloud image segmentation," *IEEE Geoscience and Remote Sensing Letters*, vol. 16, no. 12, pp. 1814-1818, 2019.

[19] J. Yang, J. Guo, H. Yue, Z. Liu, H. Hu and K. Li, "CDnet: CNN-based cloud detection for remote sensing imagery," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 57, no. 8, pp. 6195-6211, 2019.

[20] S. Liu, L. Duan, Z. Zhang, X. Cao, and T. S. Durrani, "Ground-based remote sensing cloud classification via context graph attention network," *IEEE Transactions on Geoscience and Remote Sensing*, 2021, doi: 10.1109/TGRS.2021.3063255.

[21] S. Dev, Y. Lee, and S. Winkler, "Color-based segmentation of sky/cloud images from ground-based cameras," *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, vol. 10, no. 1, pp. 231–242, 2017.
[22] F. Tao, W. Xie, Y. Wang, and Y. Xia, "Development of an all-sky imaging system for cloud cover assessment," *Applied Optics*, vol. 58, no. 20, pp. 5516–5524, 2019.
[23] R. Olaf, F. Philipp, and B. Thomas, "U-Net: Convolutional networks for biomedical image segmentation," in *Proceedings of International Conference on Medical Image Computing and Computer Assisted Intervention*, 2015, pp. 234–241.
[24] A. Kreuter, M. Zangerl, M. Schwarzmann, and M. Blumthaler, "All-sky imaging: A simple, versatile system for atmospheric research," *Applied Optics*, vol. 48, no. 6, pp. 1017–1091, 2009.
[25] M. Souzaecher, E. Pereira, L. Bins, and M. Andrade, "A simple method for the assessment of the cloud cover state in highlatitude regions by a ground-based digital camera," *Journal of Atmospheric and Oceanic Technology*, vol. 23, no. 3, pp. 427–447, 2006.
[26] J. Yang, W. Lu, Y. Ma, and W. Yao, "An automatic ground-based cloud detection method based on adaptive threshold," *Journal of Applied Meteorology and Climatology*, vol. 20, no. 6, pp. 713–721, 2009.
[27] S. Liu, L. Zhang, Z. Zhang, C. Wang, and B. Xiao, "Automatic cloud detection for all-sky images using super-pixel segmentation," *IEEE Geoscience and Remote Sensing Letters*, vol. 12, no. 2, pp. 354-358, 2015.
[28] H. Zhao, J. Shi, and X. Qi, "Pyramid scene parsing network," in *Proceedings of IEEE Conference on Computer Vision and Pattern Recognition*, 2017, pp. 2881-2890.
[29] M. Mostajabi, P. Yadollahpour, and G. Shakhnarovich, "Feedforward semantic segmentation with zoom-out features," in *Proceedings of IEEE Conference on Computer Vision and Pattern Recognition*, 2015, pp. 3376-3385.
[30] Z. Zhang, S. Yang, S. Liu, B. Xiao and X. Cao, "Ground-based cloud detection using multiscale attention convolutional neural network," *IEEE Geoscience and Remote Sensing Letters*, 2021, doi: 10.1109/LGRS.2021.3106337.
[31] Z. Qiao, Y. Zhou, and D. Yang, "Seed: Semantics enhanced encoder-decoder framework for scene text recognition," in *Proceedings of IEEE Conference on Computer Vision and Pattern Recognition*, 2020, pp. 13528-13537.
[32] J. Yao, S. Fidler, and R. Urtasun, "Describing the scene as a whole: Joint object detection, scene classification and semantic segmentation," in *Proceedings of IEEE Conference on Computer Vision and Pattern Recognition*, 2012, pp. 702-709.
[33] K. Yuan, G. Meng, D. Cheng, J. Bai, S. Xiang, and C. Pan, "Efficient cloud detection in remote sensing images using edge-aware segmentation network and easy-to-hard training strategy," in *Proceedings of IEEE International Conference on Image Processing*, 2017, pp. 61-65.
[34] X. Hou, J. Liu, B. Xu, B. Liu, X. Chen, M. Ilyas, et al. "Dual adaptive pyramid network for cross-stain histopathology image segmentation," in *Proceedings of International Conference on Medical Image Computing and Computer-Assisted Intervention*, 2019, pp. 101-109.
[35] D. O'Neill, B. Xue, and M. Zhang, "Evolutionary neural architecture search for high-dimensional skip-connection structures on densenet style networks," *IEEE Transactions on Evolutionary Computation*, 2021, doi: 10.1109/TEVC.2021.3083315.
[36] X. J. Mao, C. Shen, and Y. B. Yang, "Image restoration using very deep convolutional encoder-decoder networks with symmetric skip connections," in *Proceedings of Advances in Neural Information Processing Systems*, 2016, pp. 2802-2810.
[37] J. Long, E. Shelhamer, and T. Darrell, "Fully convolutional networks for semantic segmentation," in *Proceedings of IEEE Conference on Computer Vision and Pattern Recognition*, 2015, pp. 3431–3440.
[38] M. Zhai, X. Xiang, R. Zhang, N. Lv and A. El Saddik, "Optical flow estimation using dual self-attention pyramid networks," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 30, no. 10, pp. 3663-3674, 2020.
[39] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proceedings of IEEE Conference on Computer Vision and Pattern Recognition*, 2016, pp. 770–778.
[40] K. He, X. Zhang, S. Ren, and J. Sun, "Spatial pyramid pooling in deep convolutional networks for visual recognition," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 37, no. 9, pp. 1904-1916, 2015.
[41] R. Zhou, H. Yu, Y Cheng, "Quantum image edge extraction based on improved Prewitt operator," *Quantum Information Processing*, vol. 18, no. 9, pp. 1-24, 2019.

Zhong Zhang (M' 14 - SM' 19) is a Professor at Tianjin Normal University, Tianjin, China. He received the Ph.D. degree from Institute of Automation, Chinese Academy of Sciences, Beijing, China. He has published about 110 papers in international journals and conferences such as the IEEE Transactions on Geoscience and Remote Sensing, IEEE Transactions on Fuzzy Systems, Pattern Recognition, IEEE Transactions on Circuits Systems Video Technology, IEEE Transactions on Information Forensics and Security, Signal Processing (Elsevier), IEEE Conference on Computer Vision and Pattern Recognition (CVPR), AAAI Conference on Artificial Intelligence (AAAI), International Conference on Pattern Recognition (ICPR), and International Conference on Image Processing (ICIP). His research interests include remote sensing, computer vision, and deep learning. He is a senior member of IEEE.

Shuzhen Yang is a master student at Tianjin Normal University, Tianjin, China. Her research interests include ground-based cloud analysis and deep learning.

Shuang Liu (M' 18 - SM' 19) is a Professor at Tianjin Normal University, Tianjin, China. She received the Ph.D. degree from Institute of Automation, Chinese Academy of Sciences, Beijing, China. She has published over 60 papers in major international journals and conferences. Her research interests include computer vision, remote sensing and deep learning. She is a senior member of IEEE.

Xiaozhong Cao is a Professor at Meteorological Observation Centre in China Meteorological Administration. He received the Ph.D. degree in automatic control theory and application from Institute of Automation, Chinese Academy of Sciences in 1996. His current research interests include the theory of meteorological observation and climate change, and the automatic meteorological observation.

Tariq S. Durrani is Research Professor at University of Strathclyde, Glasgow Scotland. His research covers AI, Signal Processing and Technology Management. He has authored 350 publications; supervised 45 PhDs. He is a Fellow of the: IEEE, UK Royal Academy of Engineering, Royal Society of Edinburgh, IET, and the Third World Academy of Sciences. He was elected Foreign Member of the Chinese Academy of Sciences and the US National Academy of Engineering in 2021 and 2018, respectively.