Cloud Detection Method Using CNN Based on Cascaded Feature Attention and Channel Attention

Jing Zhang, Jun Wu, Hui Wang, Yuchen Wang, and Yunsong Li, *Member, IEEE*

***Abstract*— Cloud detection is of great significance for the subsequent analysis and application of remote-sensing images, and it is a critical part of remote-sensing image preprocessing. In this article, we propose a cloud detection method using convolutional neural networks based on cascaded feature attention and channel attention (CFCA-Net). The CFCA-Net uses cascaded feature attention module (CFAM) to enhance the attention of the network toward important color feature and texture feature. The CFAM cascaded the color feature attention and texture feature attention module in the encoder. The CFAN-Net also uses channel attention to highlight the important information in the channel dimensions. The attention module is based on multi-scale features and uses dilated convolution with different dilation rates to obtain information about multiple receptive fields. Moreover, a loss function combined quadtree and binary cross-entropy (BCE) was also introduced to make the network focus on the edge of cloud area. We validated our CFCA-Net on the Gaofen-1 wide field-of-view (WFV) imagery dataset. The experimental results show that the CFCA-Net performs well under different scenarios, and its overall accuracy reaches 97.55%. Moreover, subjective cloud detection results also prove the effectiveness of our algorithm.**

***Index Terms*— Attention mechanism, cascaded feature attention, channel attention, cloud detection, loss function, quadtree segmentation.**

I. Introduction

WITH the development of satellite remote-sensing technology, a large number of high-resolution remote-sensing images have been used widely in marine pollution monitoring, urban planning, agricultural monitoring, and other fields. However, the occlusion of clouds is inevitable in satellite images, as the clouds cover more than 60% of the earth's surface area [1]. Most satellite sensors cannot penetrate the clouds, and the cloud occlusion in remote-sensing images affects the extraction and application of remote-sensing image data significantly. Therefore, cloud detection, as an important step of remote-sensing image preprocessing, is crucial in various application fields of remote-sensing image.

Over the years, researchers have studied much about cloud detection methods. It is known that the traditional cloud detection method relies on the physical characteristics of the cloud and sets the threshold based on it. The cloud detection method based on physical characteristics studies mainly the reflectivity of clouds in different bands and the relationship between them (such as the ratio of reflectance between two bands, etc.). Using the difference between the physical characteristics of the cloud area and the non-cloud area, a better detection effect can be achieved by setting thresholds for the specific physical characteristics. In 1993, Rowssow and Garder [2] set thresholds in the near-infrared and visible light bands and proposed an International Satellite Cloud Climatology Project (ISCPP) cloud detection algorithm. Targeting the Landsat-7 remote-sensing data, Irish *et al.* [3] proposed an automatic cloud cover assessment (ACCA) algorithm. This method uses the multi-spectral and thermal infrared band reflection characteristics of the Landsat7 remote-sensing data to obtain cloud masks and non-cloud masks. This method is improved and also used in Gaofen-1 satellite imagery [4]. These methods use only a part of the band information about the remote-sensing data. The F-mask considers almost all the band information, conducts several physical tests, builds a probability model to calculate the cloud probability of each pixel, and can dynamically calculate the suitable threshold [5]–[7]. Chen *et al.* [8] used F-mask to integrate spectral information and contextual semantic information to improve the detection accuracy of Landsat images. The multi-feature combined (MFC) algorithm uses the relationship between the reflectivity and waveband of the GF-1 remote-sensing image and uses the aggregate and texture features to improve the inspection results to generate the final cloud mask [9].

Some remote-sensing images contain less band information, such as the Gaofen-1 satellite image which has only four bands of information. For such an image, the color and texture features are generally extracted to process the image. An and Shi [10] designed a cloud detection algorithm based on the least square method. This algorithm utilizes the color features, local statistical features, texture features, and structural features of the image. Liu *et al.* [11] applied a graphic model combined with color features for cloud segmentation. Li *et al.* [12] used support vector machines (SVMs) [13] to distinguish features, including brightness features, texture features, and average gray-level co-occurrence matrix (GLCM) [14], [15]. Shi *et al.* [16] used scale-invariant feature transform (SIFT) [17] and RGB features as the key features to evaluate whether a super-pixel [18] is a cloud. These methods extract the brightness, texture, and other variable features of image

Manuscript received July 1, 2021; revised September 7, 2021 and October 8, 2021; accepted October 12, 2021. Date of publication October 15, 2021; date of current version February 7, 2022. This work was supported by the Natural Science Foundation of China under Grant 61801359 and Grant 61571345. *(Corresponding author: Jing Zhang.)*

Jing Zhang is with the State Key Laboratory of Integrated Services Networks, and the School of Telecommunication Engineering, Xidian University, Xi'an 710071, China, and also with the Guangzhou Institute of Technology, Xidian University, Guangzhou 510555, China (e-mail: jingzhang@xidian.edu.cn).

Jun Wu, Hui Wang, and Yuchen Wang are with the School of Telecommunication Engineering, Xidian University, Xi'an 710071, China.

Yunsong Li is with the State Key Laboratory of Integrated Services Networks, and the School of Telecommunication Engineering, Xidian University, Xi'an 710071, China.

Digital Object Identifier 10.1109/TGRS.2021.3120752

pixels to obtain the cloud masks. However, these methods are not robust to images of extraordinary underlying surfaces (such as ice and snow).

In recent years, neural network methods have been used widely in the field of image processing and have achieved good results in object detection, classification, and segmentation. Remote-sensing image cloud detection tasks are categorized under semantic segmentation. The deep learning methods for cloud detection can avoid manually designing features and dig out more potential features. Key and Barry [19] took the lead in applying neural networks to cloud detection in remote-sensing images. Bankert [20] and Jianhua [21] used artificial neural networks and probabilistic neural networks, respectively, for Advanced Very High Resolution Radiometer (AVHRR) cloud detection. These two models have a great detection effect on thin clouds and thick clouds and have good stability in complex scenes. In deep learning methods, multi-scale features are widely used. Xie *et al.* [22] performed super-pixel segmentation on the remote-sensing image to be detected, used a convolutional neural network to extract multi-scale features from the super-pixel, and divided the pixels into cloud pixels and non-cloud pixels. Ji *et al.* [23] used cascaded convolutional neural networks to integrate cloud detection and cloud removal frameworks and used multi-scale aggregation to detect clouds and shows. Luotamo *et al.* [24] used multi-scale information and cascaded two CNN models to deal with undersampled and full-resolution images. Jeppesen *et al.* [25] suggested a cloud detection deep learning model for remote-sensing images based on the convolutional neural network model. Segal-Rozenhaimer *et al.* [26] proposed a domain-adaptive method based on CNN. This method can better adapt to different satellite platforms in the prediction step without the need to train each platform separately, which improves the robustness of multiple remote-sensing platform predictions. The deep learning methods can also handle situations such as missing information, no clouds labels, and so on. SAGAN used a semi-supervised algorithm to achieve cloud detection, requiring only a small number of image-level tags [27]. For thumbnails with missing resolution and spectral information, CDnet used feature pyramid module (FPM) and boundary refinement (BR) block to effectively extract cloud masks [28]. CDnetV2 had further improved the detection results of images with coexisting clouds and snow [29]. The main advantage of deep learning is the diversity of feature learning and the ability to learn in-deep features. The deep convolutional neural network can extract various features such as spatial features and spectral features.

However, most methods pay more attention to regional accuracy and less to boundary quality, which lead to the blurred boundary in the detection results [30]. In cloud boundaries and thin cloud areas, cloud information and underlying surface information are mixed. Due to the complexity and diversity of the underlying surface, it is very difficult to detect the boundaries and thin cloud areas accurately. In the face of this situation, it is unrealistic to only rely on increasing the width and depth of the network to solve it.

We have done a lot of research on cloud detection. We first consider using the multiple features of ground objects. We found that the texture difference between cloud and ground objects is very obvious, which is very effective for improving the accuracy of cloud detection. The multi-scale image decomposition based on the domain transform filter were used to extract the texture features of ground objects [31]. Then, we combined the color and texture features of remote-sensing images to design cloud detection methods [32]. Compared to the traditional algorithms, the deep learning method has significantly improved the detection performance. We noticed the development of deep learning and used convolutional neural networks for cloud detection. We designed a Gabor transform layer in the encoder–decoder network to extract texture features [33]. This network also combined with the attention module and achieved a good cloud detection effect. In the AUDI-Net [34], we proposed the Up-Down block and used wavelet transform, which significantly improves the density of thin cloud detection. However, the Up-Down block takes up a lot of parameters and calculations. In [35], we studied the lightweight network and achieved great performance with a smaller amount of parameters.

Through previous research, we found that color and texture features are very effective in cloud detection. The effective extraction and utilization of these features can often improve the performance of cloud detection. It has achieved good detection results on public dataset and also has a lighter network structure compared with addition input, up and down block implant network (ADUI-Net).

We have proposed a network for cloud detection, which contains cascaded feature attention module and channel attention module, named CFCA-Net. The CFCA-Net is built on the encoder–decoder structure. It has achieved good detection results on public dataset, it also achieved good detection results on thin cloud and the boundary of the cloud. And also has a lighter network structure compared with ADUI-Net. Our contributions include the following three parts.

1) We designed a cascaded feature attention module (CFAM) to enhance the useful spatial information of the multi-scale feature maps and suppress invalid information. This module extracts color features and texture features giving better results in remote-sensing images with fewer bands. We used dark channel prior to assisting the extract color feature, and nonsubsampled contourlet transform (NSCT) to assist the extract texture feature.
2) We sketched a channel attention module on the decoder to carry out the screening of characteristic channels. Our channel attention module uses dilated convolution with different dilation rates to obtain information about multiple receptive fields.
3) We designed a loss function based on quadtree segmentation. This loss function pays attention to the part of the detection results that has large edge changes and is difficult to distinguish. The similar points in the entire large area are finally replaced with single values, which reduce the proportion of the simple samples in the loss function compared to using all points to iterate the loss function.

II. Background

A. Encoder–Decoder Structure

In the field of semantic segmentation, the encoder–decoder structure is widely used and has achieved great results. The fully convolutional network (FCN) [36] introduced an end-to-end fully convolutional neural network structure for semantic segmentation. Unet [37] introduced skip connection and achieved good results. SegNet [38] applied the pooling layer result from the encoder to the decoder that introduced more encoding information.

DeepLab series proposed atrous spatial pyramid pooling (ASPP), which combines information at different scales [39]–[42]. DeepLabV3+ introduces a decoding module based on DeepLabV3, which further integrates the low-level features with the high-level features and improves the accuracy of the segmentation boundary.

B. Dilated Convolution

In order to expand the receptive field, there are usually two methods, one is to increase the size of the convolution kernel, and the other is to use a pooling operation. The pooling layer is an important structure in deep learning that can further extract abstract features and expand the receptive field. However, the large convolution kernel will increase the amount of calculation, and the pooling operation will inevitably reduce the resolution and cause the loss of detailed information. Dilated convolutions proposed to use dilated convolution to avoid the decrease in resolution and proposed a "context module" to aggregate multi-scale information [43]. Dilated convolution is realized by inserting spaces between the elements of the convolution kernel. This method of increasing the receptive field has a good effect while connecting multiple dilated convolutions [44]. The DeepLab series uses dilated convolutions, among which DeepLabv2 and DeepLabv3 study the effectiveness of dilated convolution in parallel and series for extracting multi-scale information [39]–[42].

C. Attention Mechanism

The attention mechanism resulted from human visual cognitive science. Scientists discovered that when humans perform visual tasks such as reading and observation, they pay more attention to the detailed information of the target area and suppress other useless data. The attention mechanism in deep learning is similar to this mechanism. The basic idea is to make the model focus on the important features and ignore those that are not important. The results of attention are generally displayed in the form of probability maps or probability feature vectors. Squeeze-and-excitation networks (SE-Nets), proposed by Hu *et al.* [45], use the SE module to realize the weight learning of feature maps of different channels. Woo *et al.* [46] proposed convolutional block attention module (CBAM) that combines spatial and channel attention. The attention mechanism is very effective in target detection [47], [48], image segmentation [49]–[51], super-resolution [52], [53], and other fields which can improve the effectiveness of the model.

D. Nonsubsampled Contourlet Transform

Da Cunha *et al.* [54] proposed the NSCT. NSCT not only has the multi-resolution and time-frequency local characteristics of the wavelet transform, but also has multi-directivity and anisotropy, which can well represent the texture, edge direction, and other information in the image. The NSCT is a transformation based on the non-subsampled pyramid (NSP) and the non-subsampled direction filter bank (NSDFB). First, the NSP decomposes the input image in a tower shape and decomposes it into two parts, high-pass and low-pass. Then, the NSPFB decomposes the high-frequency sub-band into multiple directional sub-bands, and the low-frequency part continues to decompose as above. The NSP uses a translation-invariant filter structure to achieve the filter function.

Using NSCT to extract the texture information of the image for segmentation is conducive to improving the performance of image segmentation [55], [56].

III. Method

A. Overview

We use the encoder–decoder structure as the framework of the cloud detection network model and introduce the attention mechanism. The backbone of the CFCA-Net is similar to the existing encoder–decoder network models. The overall network framework of the CFCA-Net, in this article, is composed of two parts: encoder and decoder. The encoder encodes the entire input image, expands the number of feature map channels of the image gradually, and obtains features of different scales through the pooling structure. Each step of the encoding end comprises two consecutive convolutional layers and a maximum pooling with size of 2 × 2. Each convolutional layer uses a convolution operation with a kernel size of 3 × 3 and ReLU linear correction unit. The maximum pooling is used to down-sample the feature map; in each down-sampling step, the number of feature channels will be doubled. Contrary to the encoding side, the decoding side is needed to restore the feature map to the size of the input image. Hence, each step of the decoder includes up-sampling and convolution with kernel sizes of 3 × 3. To make up for the loss of information in the sampling process, the feature map of the corresponding scale at the encoder is connected to the decoder through skip connection, and the feature information is shared with the decoder. Finally, we used 1 × 1 convolution and Sigmoid activation function to get the final prediction result.

Table I shows the structure of basic encoder–decoder network of CFCA-Net. In the table, (×2) means that there are two layers with the same structure.

The ground objects in the remote-sensing image are complicated, and too many invalid features affect the performance of the cloud detection model. The introduction of the attention module can enable the network to learn the compelling features of the cloud region, reduce the learning of invalid features such as ground objects, and improve the effectiveness of feature extraction and the accuracy of the cloud detection model. We used the CFAM in the encoder to emphasize the color, texture, and other related features of the cloud area while

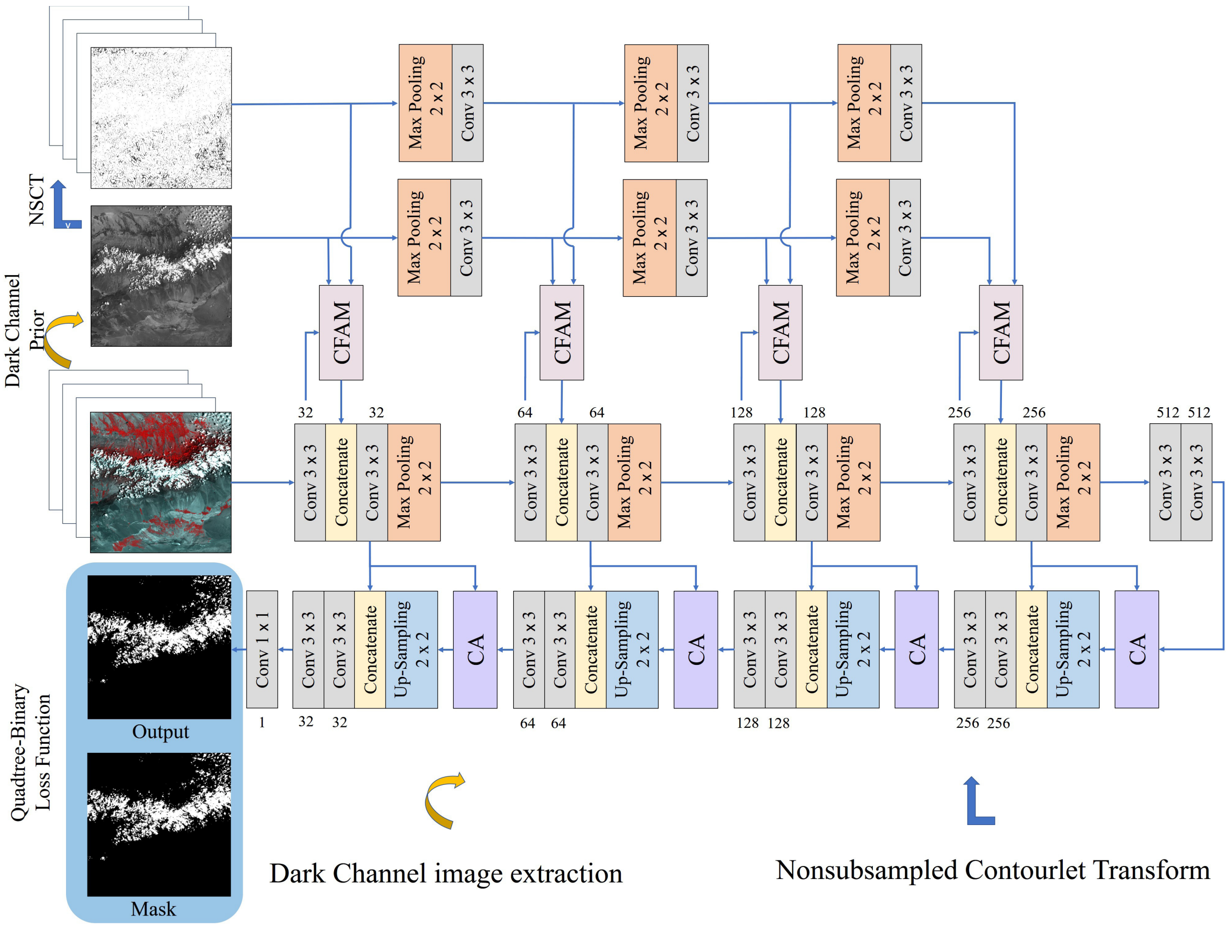

Fig. 1. Structure of CFCA-Net.

TABLE I
STRUCTURE OF BASIC ENCODER–DECODER NETWORK

Encoder		Decoder	
Input	$(512 \times 512 \times 4)$	**Output** $1 \times 1 \times 1$	$(512 \times 512 \times 1)$
Conv1$(\times 2)$ $3 \times 3 \times 32$	$(512 \times 512 \times 32)$	**Conv9**$(\times 2)$ $3 \times 3 \times 32$	$(512 \times 512 \times 32)$
Pooling 2×2	$(256 \times 256 \times 32)$	**UpSampling** 2×2	$(512 \times 512 \times 64)$
Conv2$(\times 2)$ $3 \times 3 \times 64$	$(256 \times 256 \times 64)$	**Conv8**$(\times 2)$ $3 \times 3 \times 64$	$(256 \times 256 \times 64)$
Pooling 2×2	$(128 \times 128 \times 64)$	**UpSampling** 2×2	$(256 \times 256 \times 128)$
Conv3$(\times 2)$ $3 \times 3 \times 128$	$(128 \times 128 \times 128)$	**Conv7**$(\times 2)$ $3 \times 3 \times 128$	$(128 \times 128 \times 128)$
Pooling 2×2	$(64 \times 64 \times 128)$	**UpSampling** 2×2	$(128 \times 128 \times 256)$
Conv4$(\times 2)$ $3 \times 3 \times 256$	$(64 \times 64 \times 256)$	**Conv6**$(\times 2)$ $3 \times 3 \times 256$	$(64 \times 64 \times 256)$
Pooling 2×2	$(32 \times 32 \times 256)$	**UpSampling** 2×2	$(64 \times 64 \times 512)$
	Conv5$(\times 2)$ $3 \times 3 \times 512$	$(32 \times 32 \times 512)$	

ignoring the invalid features of the non-cloud area. As shown in Fig. 1, we used the cascaded attention module at each scale of the encoder to form continuous multi-scale cascaded feature attention. In this way, the information loss caused by down-sampling can be effectively compensated, and the feature map of the next level can be guided to make it pay more attention to the color features and texture features to preserve the features of the cloud area.

On the decoder, after multiple convolutions and pooling operations, a multi-channel feature map containing complex information is generated. The feature map of each channel is a component extracted from the original image that contains different feature information. Some channels contain more features that can highlight the cloud area, while some do not. The channel attention mechanism is a good feature map screening mechanism. The channel attention mechanism is often to mine the correlation of data from itself [45], [46], [57]. However, according to the characteristics of the cloud detection encoder–decoder network in this article, we sketched a channel attention module that uses the feature map of the encoder to guide the feature map of the decoder. As shown in Fig. 1, because of the symmetrical structure of the encoder–decoder network, it is necessary to perform multiple up-sampling operations on the network at the decoder. In the process of up-sampling, the number of feature map channels decreases gradually. Hence, we used the channel attention module before the up-sampling process of the decoder to retain the channels that can highlight the features of the cloud area in the feature map.

The overall performance of the model is affected by both the network structure and the design of loss function [58], [59]. A proper loss function can make the model converge faster during the training process, and the obtained model also has a more reliable prediction performance. Therefore, choosing a suitable loss function is also extremely important for the development of the model. In the semantic segmentation of the ordinary images, the cross-entropy loss function and Adam optimization algorithm are used to train the model to achieve better results [60], [61].

In remote-sensing images, cloud areas and non-cloud areas often occupy a large portion which is easier to identify. However, the boundary between the two is mostly thin clouds, extremely difficult to detect. Therefore, for cloud detection networks, we hope that the network can be more accurate in the edge detection of cloud and non-cloud areas. Quadtree image segmentation is used widely in image processing applications to locate regions of interest [62]. The cloud mask of the cloud detection network is a result of pixel-level binarization. This article designs a selective guided loss function for the quadtree classification. Through quadtree segmentation of the cloud mask, we determine which parts are more heterogeneous than other parts, and let the loss function focus on the edges that are difficult to distinguish. In the network training process, we adopt a combination of the cross-entropy loss function and quadtree loss, so that the loss function can focus on the indistinguishable parts of the edge, while also considering the overall prediction results.

B. Cascaded Feature Attention

The traditional cloud detection methods extract several texture features and color features to improve the performance of cloud detection. Compared to the conventional cloud detection methods, deep-learning-based cloud detection methods generate functions that map the input data to predicted cloud masks by using statistical analysis of the training set. The cloud detection method based on deep learning does not rely on prior knowledge but autonomously learns relevant features through the network. This process relies on a large number of training datasets and enormous computing power support. If we can guide and add prior to the network training, we can make the network's feature learning ability stronger. We used traditional methods to extract the color and detail texture features of the cloud layer and generated attention weights to add to the cloud detection network. It would help the network to pay more attention to these features and enhance the ability of feature learning.

In this section, we explain the cascaded attention module in detail. As shown in Fig. 2, the cascaded attention module contains two sub-modules, the color feature map attention module and the detail texture feature map attention module.

1) Color Feature Attention: The color feature is one of the most significant visual features of an image, and the color feature has a strong correlation with the scene displayed. In addition, the color feature has a small effect on the size, direction, and viewing angle of the image itself; hence, it is more robust. He *et al.* [63] found that for most distant images, there will always be some pixels (called dark channel pixels) that contain a very low pixel value in the three-color channel components. It can be seen from Fig. 3 that the cloud area of the dark channel image is still bright, while the non-cloud area is very dark, similar to the ground truth. Upon extraction of the original image from the dark channel feature, the attention feature map is extracted through the color feature map attention module. This is because the clouds generally have a higher reflectivity in the visible light band. Therefore, we extracted the dark channel features of the image and constructed the color feature map attention module. The dark channel extraction method is shown as follows:

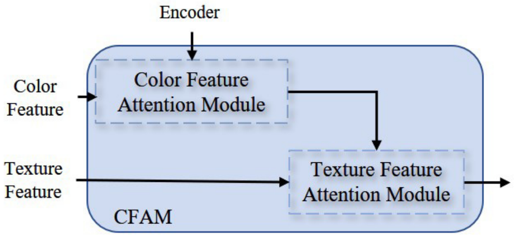
Encoder
Color Feature
Texture Feature
Color Feature Attention Module
Texture Feature Attention Module
CFAM

Fig. 2. Structure of cascaded feature attention module (CFAM).

$$f_{\text{dark}}(x,y) = \min_{c\in[r,g,b]} f(x,y,c). \tag{1}$$

In the above formula, $f_{\text{dark}}(x,y)$ is the dark channel image. And $f(x,y,c)$ is the original remote-sensing image, which has three visible bands.

The structure of the attention module of the color feature map is given in Fig. 4, which can be expressed as follows:

$$f_{C-\text{Dilated}} = \underset{\text{rate}\in\{1,3,5\}}{D}\{C[A(f_C), M(f_C)]\} \tag{2}$$

$$f_{C-\text{Att}} = \text{Conv}\{C[f_{C-\text{Dilated}}]\} \tag{3}$$

$$f_{C-\text{OUT}} = \text{Conv}\{\text{Conv}\{\text{Sigmoid}\{f_{C-\text{Att}}\} \times f_E\} + f_E\}. \tag{4}$$

First, the dark channel feature map f_C is compressed in the channel dimension, and the average pooling and maximum pooling are performed, respectively, in the channel dimension, and the maximum and average values on the channel are extracted. Next, three dilated convolutions with different dilated rates are connected in parallel to further extract color features. By using dilated convolution, the receptive field can be increased without reducing the image resolution and increasing the amount of calculation. Different receptive fields are concatenated in the channel dimension, and after the convolution and Sigmoid activation, feature fusion is realized, and the attention weight of the color feature map $f_{C-\text{Att}}$ is obtained. We used the attention weight obtained from the color feature map to guide the encoder. Multiplying the feature map pixel by pixel with the feature map at the encoder f_E and $f_{C-\text{Att}}$, we obtained a feature map with assigned weights. For the module to maintain the original encoding end information, the feature map after the attention weight assigned is subjected to convolution learning and then added to the original encoding end feature map and convolved to obtain the output of the attention module $f_{C-\text{OUT}}$.

2) Texture Feature Attention: The texture feature describes the surface properties of a scene corresponding to the image. It is expressed by the gray-scale spatial distribution of the

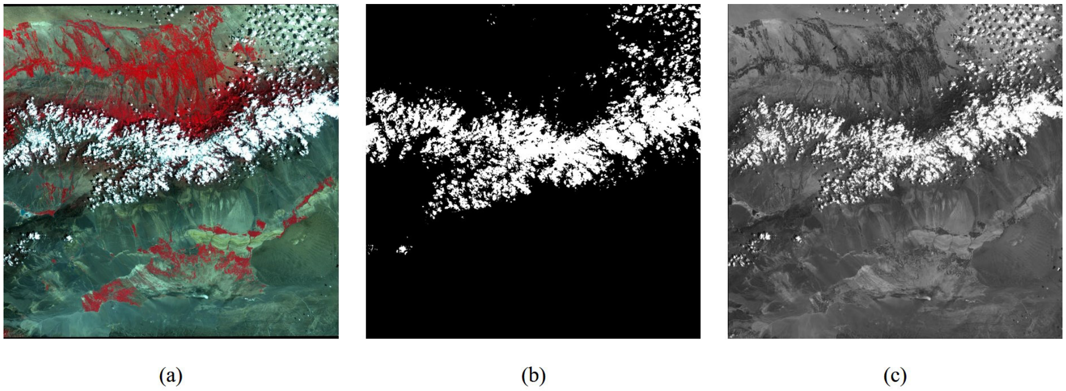

Fig. 3. Comparison of (a) original image, (b) ground truth, and (c) dark channel image.

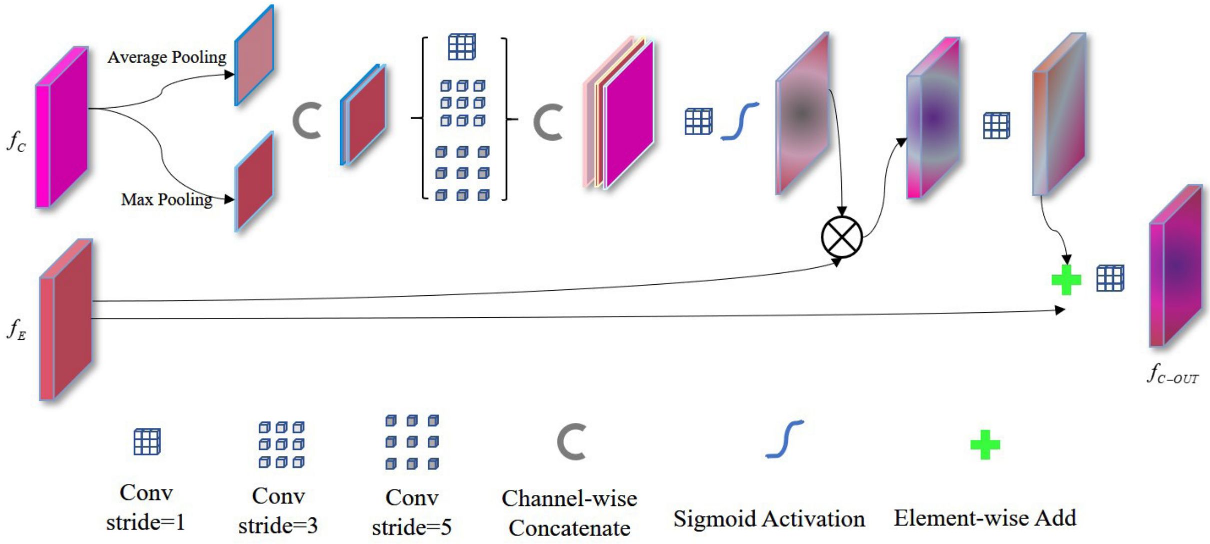

Fig. 4. Structure of color feature attention module. The feature map with color gradient in the figure indicates that it is weighted by attention, and the deeper color indicates the part that needs pay more attention.

pixels and their surrounding spatial neighborhoods, reflecting the slowly changing or periodically changing surface structure organization and arrangement properties of the surface of the object. High-resolution remote-sensing satellite images show rich and detailed information due to their high resolution. As most of the particles that make up the cloud layer are similar and have uniform radiation characteristics, the cloud area in the remote-sensing image is generally smooth, with small gray value changes, strong continuity, and similar texture characteristics. However, the texture details are more obvious because of the complex distribution of the ground features. In remote-sensing images, texture information is an important feature for identifying the cloud and non-cloud areas. Therefore, the effective extraction of the remote-sensing image texture features is conducive to the distinction between cloud and non-cloud areas in cloud detection.

The NSCT helps to maintain the edge information and contour structure of the image. In the cloud detection of remote-sensing image, the NSCT can extract the edge contour of the cloud area. The attention mechanism using the NSCT extracted texture features that can help the network identify cloud areas and non-cloud areas. Simultaneously, it can enhance the detection accuracy of the edge of the cloud area and improve the detection performance. In this article, we use NSCT to perform a two-level decomposition, and set the sub-band decomposition coefficients of each level as 2 and 4, respectively, as shown in Fig. 5.

It can be seen from Fig. 6 that the cloud area is very smooth, while the texture characteristics of the non-cloud area are obvious. The texture characteristics of the cloud and non-cloud areas are very different. After a detailed texture feature is extracted by the NSCT, the attention feature map is extracted through the attention module of the detailed texture feature map. The structure of the detailed texture feature attention module is shown in Fig. 7, which can be expressed by the following equations:

$$f_{C-\text{Dilated}} = \underset{\text{rate}\in\{1,3,5\}}{D}\{C[A(f_T), M(f_T)]\} \tag{5}$$

$$f_{T-\text{Att}} = \text{Sigmoid}\{\text{Conv}\{C[f_{C-\text{Dilated}}]\}\} \tag{6}$$

$$f_{T-\text{OUT}} = \text{Conv}\{\text{Conv}\{f_{T-\text{Att}} \times f_{C-\text{OUT}}\} - f_{C-\text{OUT}}\}. \tag{7}$$

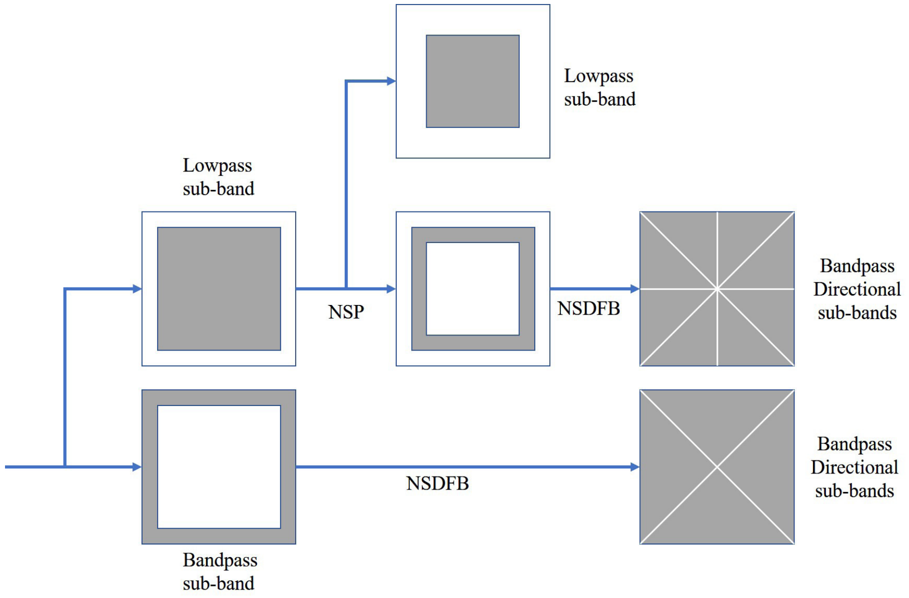

Fig. 5. Structure of NSCT. The sub-band decomposition coefficients of level one is 2, and of level two is 4.

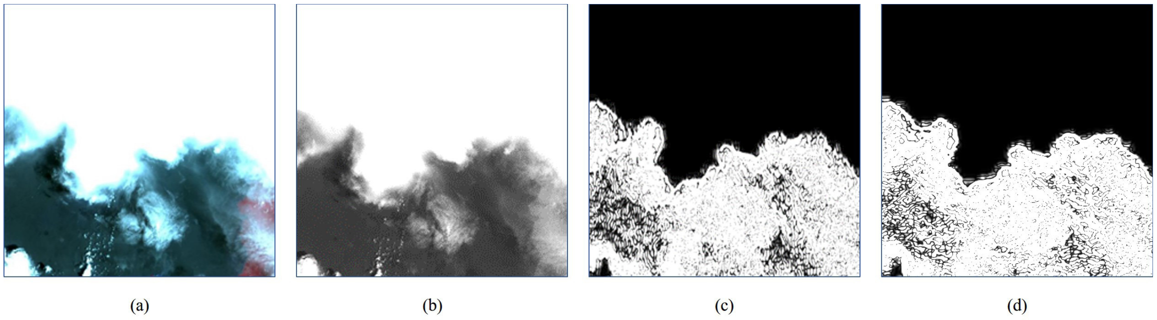

Fig. 6. Result of an image after dark channel extract and NSCT. (a) Original image. (b) Extracted dark channel image and dark channel image after NSCT. (c) Level-1 direction-1. (d) Level-1 direction-2.

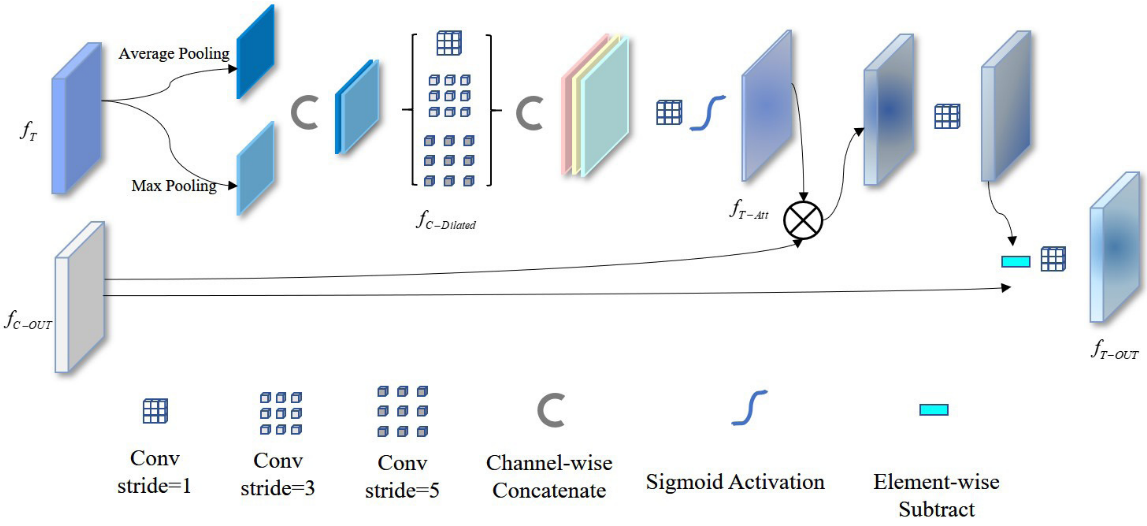

Fig. 7. Structure of texture feature attention module. The feature map with color gradient in the figure indicates that it is weighted by attention, and the deeper color indicates the part that needs pay more attention.

The attention of the detailed texture feature map is similar to the attention of the color feature map that extracts attention from space. We implement a structure similar to that of the attention module of the color feature map. The extraction of the attention weight of the detail texture is consistent with the attention module of the color feature map, which

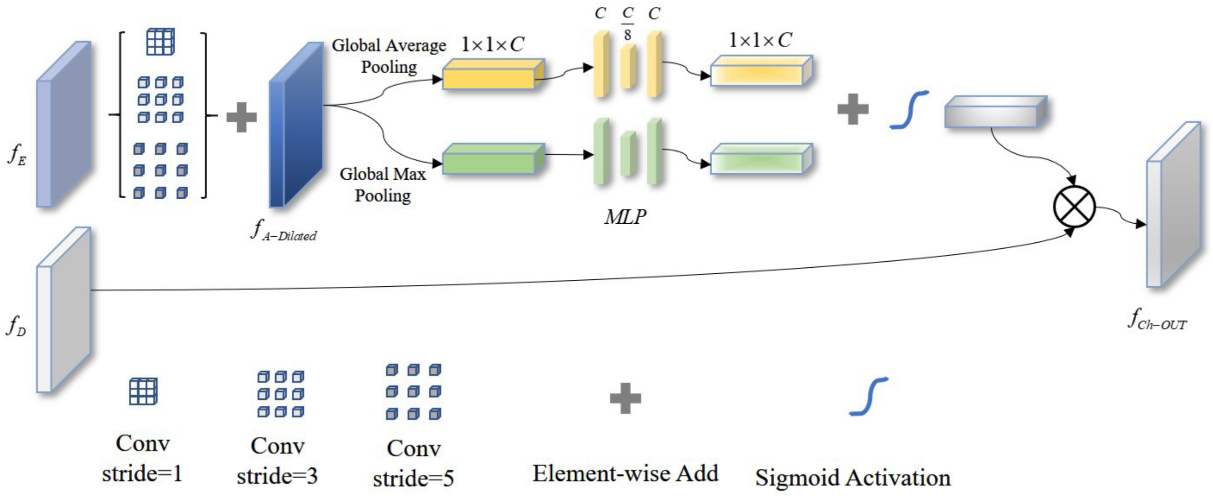

Fig. 8. Structure of channel attention module.

uses dilated convolution with different dilated rates and also uses the Sigmoid as activation function. We use the attention weight obtained from the detailed texture feature map to guide the encoder. We then perform a pixel-by-pixel multiplication operation with $f_{T-\text{Att}}$ and the feature map $f_{C-\text{OUT}}$ at the encoder to obtain the feature map with attention weights. Since the cloud area is relatively smooth and the texture features are similar, the texture features of the non-cloud area are richer, and hence, the texture attention is more focused on the texture feature of the non-cloud. To make the network pay more attention to the characteristics of the cloud region, the feature map, after the attention weight is assigned, is subjected to convolutional learning and then subtracted from the feature map of the original encoding end to obtain detailed texture difference information. The attention module output $f_{T-\text{OUT}}$ is obtained after the convolution output.

C. Channel Attention Module

In the encoder, the original image is subjected to operations such as convolution and pooling to generate a multi-channel feature map containing a variety of complex information. Each channel is a component extracted from the original image and contains variety feature information. Some channels contain more information, which highlights the characteristics of the cloud area. This information is helpful for the network to segment the cloud area from the image and is the key information for the network to complete the segmentation task. However, in the decoding process, at the decoding end, the feature maps of these channels are regarded as equally important, causing a certain degree of useless information interference. We employ the channel attention mechanism to filter these irrelevant feature channels. For clouds with different sizes, different receptive fields are required. Large cloud areas require larger receptive fields to obtain richer semantic information, while small cloud areas should use smaller receptive fields. In order to deal with the cloud areas with different size, we use parallel dilated convolutions to obtain different receptive fields and capture multiscale information.

The structure of the channel attention module is presented in Fig. 8, which can be expressed by the following formulas:

$$f_{A-\text{Dilated}} = \underset{\text{rate}=1}{D}(f_E) + \underset{\text{rate}=3}{D}(f_E) + \underset{\text{rate}=5}{D}(f_E) \tag{8}$$

$$f_{\text{Ch}-\text{Att}} = \text{MLP}\{\text{Mp}(f_{A-\text{Dilated}}), \text{Ap}(f_{A-\text{Dilated}})\} \tag{9}$$

$$f_{\text{Ch}-\text{OUT}} = \text{Sigmoid}\{f_{\text{Ch}-\text{Att}}\} \times f_D. \tag{10}$$

First, the feature maps of the encoding end are, respectively, subject to the dilated convolution with the dilated rate of $\{1, 3, 5\}$ and the feature maps of different receptive fields are obtained. The corresponding elements of the feature maps of different receptive fields are added together for feature fusion. Then, global average pooling and global maximum pooling are performed on the fused feature maps to obtain global information on each channel. The vector generated after using global maximum pooling and global average pooling has the extracted high-level features. Using these two pooling methods, models can obtain relatively rich information. The information of these two vectors is transformed, and feature is extracted using a fully connected layer, and after the addition, the Sigmoid function is used for normalization to obtain the channel attention weight. We use the attention weights generated by the feature map of the code segment, containing the shallow features, to guide the feature map of the decoder. The weight extracted by the encoder is multiplied with the feature map of the decoding end to obtain the reconstructed feature map.

D. Quadtree-Binary (QTB) Loss Function

Binary cross-entropy (BCE) is usually used as the loss function in binary classification tasks. The formula of BCE is as follows:

$$L_{\text{BCE}} = -\frac{1}{wh}\sum_{i=1}^{w}\sum_{j=1}^{h}\left(y_{i,j}\log y'_{i,j} + \left(1 - y_{i,j}\right)\log\left(1 - y'_{i,j}\right)\right). \tag{11}$$

The output of the network is normalized to 0–1 by the sigmoid function. The pixels can be regarded as positive samples if the probability value exceeds 0.5. In cloud detection, the cloud pixels can be regarded as positive samples and non-cloud pixels as negative samples.

In the cloud detection task, it is found that the large cloud or large non-cloud areas are simple samples and are easier to detect. For cloud detection, we hope that the network focuses more on the edge of the cloud and non-cloud areas because these areas are difficult to detect and often have a greater impact on detection performance. The prediction value of cloud pixels in these areas is about 0.5, which is the challenge of the cloud detection task. Using BCE function cannot converge to the optimal in a large number of simple samples.

Based on this, we design quadtree loss. We introduce the quadtree structure into the loss function and refine the segmentation sub-region on the real cloud mask. The same eigenvalues are classified into the same category after the quadtree segmentation of the whole image is completed. Similarly, the probability value of the prediction image is divided into sub-regions according to the quadtree segmentation result of the cloud mask. The formula of quadtree loss is given as

$$L_{\mathrm{QT}_k} = -\frac{1}{w_k h_k}\sum_{i=1}^{w_k}\sum_{j=1}^{h_k}\left(y_{i,j}\log y'_{i,j} + \left(1-y_{i,j}\right)\log\left(1-y'_{i,j}\right)\right) \tag{12}$$

$$L_{\mathrm{QT}} = \frac{1}{M}\sum_{k=1}^{M} L_{\mathrm{QT}_k}. \tag{13}$$

This formula of the L_{QT_k} suggests that after the quadtree segmentation of the cloud mask, BCE is done for each region, which represents the local detection accuracy between prediction result and ground truth. These local regions are obtained by the quadtree segmentation, and the k denotes kth sub-regions. The L_{QT} is the quadtree loss. M is the number of sub-regions divided by quadtree, that is, the size of the set is obtained by quadtree. This means that after the BCE of each region is finished, the average of all regions is calculated.

Fig. 9 is a simple example of calculating the quadtree loss. We first perform the quadtree segment on the mask to obtain segmented blocks. Then calculate the cross-entropy for each segmented block. Finally calculate the average of all blocks. There are 16 pixels in the original image, calculated according to the BCE. After using the quadtree loss, the final result only needs to average the value of ten points.

The advantage of the quadtree loss is to focus the attention of loss function on the parts with large edge changes and is difficult to distinguish. Thus, compared with iterating the loss function with all points, the proportion of simple samples in the loss function is reduced. The application of quadtree classification selective guidance loss function can enhance the network performance effectively by drawing the network attention to the samples that are difficult to detect.

Fig. 10 shows quadtree segmentation results of ground truth. It can be seen from the results that the large cloud and non-cloud areas are divided into large blocks, while the edge area is densely distributed with many small blocks. Therefore, when calculating the quadtree loss, the proportion of these edge region samples in the loss function will increase.

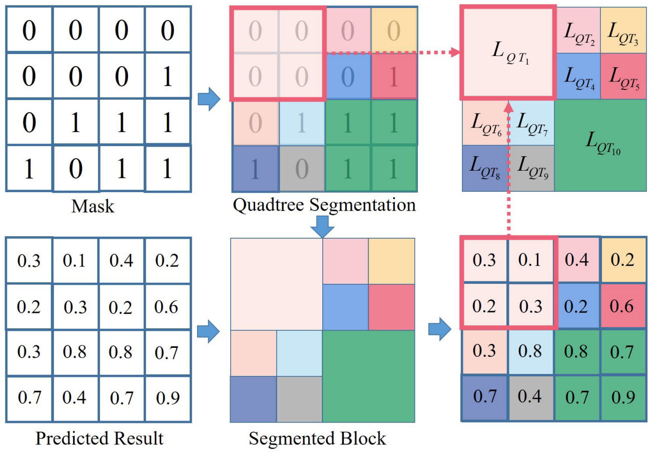

Fig. 9. Simple example of calculating the quadtree loss.

TABLE II
GAOFEN-1 SPECTRAL BANDS

Spectral Band	Wavelength	Resolution
Blue	0.45-0.52μm	16m
Green	0.52-0.59μm	16m
Red	0.63-0.69μm	16m
Near Infrared(NIR)	0.77-0.89μm	16m

We use BCE and quadtree loss at the same time to make the network have better convergence performance and improve the effect of edge detection at the same time. The final loss function is named quadtree-binary (QTB) loss; the formula is as follows:

$$L_{\mathrm{quadtree-binary}} = \gamma_1 * L_{\mathrm{BCE}} + \gamma_2 * L_{\mathrm{QT}} \tag{14}$$

where γ_1 and γ_2, respectively, represent the weight of L_{BCE} and L_{QT} and can be adjusted for different data. The final loss function will automatically adjust the influence of the samples with different degrees of difficulty. At the same time, the integration of the entire region is equivalent to adjusting the proportion of this type of samples, and it optimizes the problem of inter-class competition caused by the uneven proportion among samples.

IV. DATASET AND EVALUATION METRICS

A. Dataset

The dataset used in the experiment is obtained from the Gaofen-1 satellite. The satellite was launched in 2013 and is equipped with two panchromatic cameras and four spectroscopic cameras. It can achieve an imaging width of more than 800 km with a resolution of 16 m. The cloud detection algorithm based on Gaofen-1 data is challenging as the wide-field of a camera carried by Gaofen-1 consists of three visible light bands and near-infrared bands. Employing limited spectral information to achieve better segmentation results is very challenging and meaningful research.

We utilize the GF-1 wide field of view (WFV) dataset provided by Li *et al.* [9]. This set of data includes 108 images

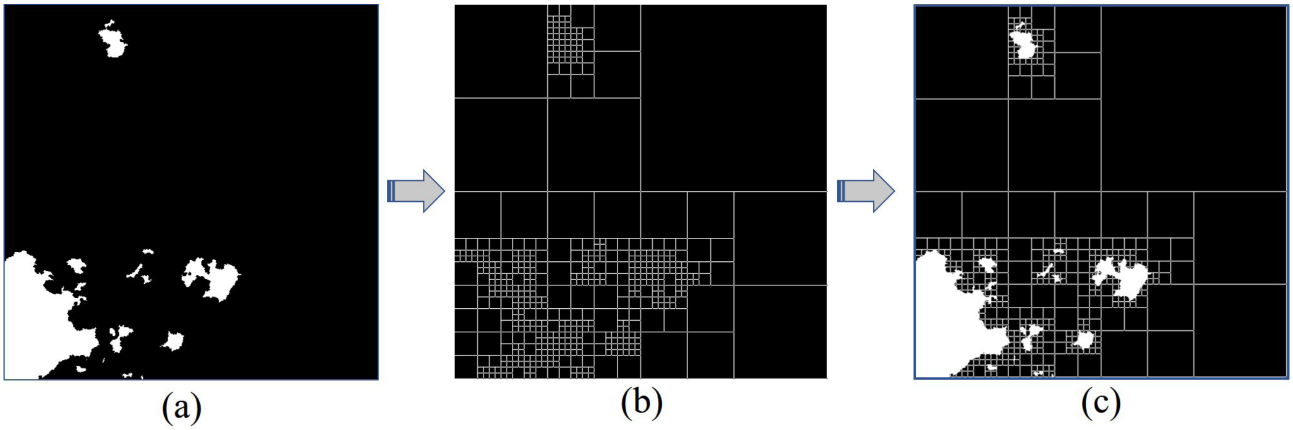

Fig. 10. Quadtree segmentation results of ground truth. (a) Ground truth, (b) blocks after quadtree segmentation, and (c) quadtree segmentation results (ground truth).

collected from all over the world. The dataset covers different geomorphic environments, including urban, barren, snow, vegetation, and water. The resolution of the image in the dataset is 16 m, and there are four bands of information of R, G, B, and NIR. Table II shows the bands and resolution of Gaofen-1. The dimensions of each image are approximately 17 000 × 16 000 × 4. We selected 86 scenes as the training set for the experiment and the rest as the test set.

To remove the black area around the scene, all images are rotated and cut to 11 264 × 11 264 as the black area does not contain any remote-sensing information and is not helpful for feature extraction. Due to the limitation of hardware resources, each scene is cut into 512 × 512 × 4 small pictures. In the end, 41 624 pictures were used for training, and 10 648 pictures were used for testing.

B. Evaluation Metrics

To evaluate the algorithm objectively, we use OA [64], Precision [65], Recall [66], F1-Score [25], Kappa [67], and FAR [64], [68] to evaluate the results. These metrics are calculated by

$$\mathrm{OA} = \frac{\mathrm{TP}+\mathrm{TN}}{\mathrm{TP}+\mathrm{TN}+\mathrm{FP}+\mathrm{FN}} \tag{15}$$

$$\mathrm{Precision} = \frac{\mathrm{TP}}{\mathrm{TP}+\mathrm{FP}} \tag{16}$$

$$\mathrm{Recall} = \frac{\mathrm{TP}}{\mathrm{TP}+\mathrm{FN}} \tag{17}$$

$$\mathrm{F1\_score} = \frac{2 \times \mathrm{Precision} \times \mathrm{Recall}}{\mathrm{Precision}+\mathrm{Recall}} \tag{18}$$

$$\mathrm{kappa} = \frac{p_a - p_e}{1 - p_e} \tag{19}$$

$$p_a = \frac{(\mathrm{TP}+\mathrm{TN})}{(\mathrm{TP}+\mathrm{TN}+\mathrm{FP}+\mathrm{FN})} \tag{20}$$

$$p_e = \frac{(P * (\mathrm{TP}+\mathrm{FP}) + N * (\mathrm{FN}+\mathrm{TN}))}{(P+N)^2} \tag{21}$$

$$\mathrm{FAR} = \frac{\mathrm{FP}}{\mathrm{TN}+\mathrm{FP}}. \tag{22}$$

Here, TP denotes the correct prediction of cloud pixels, TN denotes the number of non-cloud pixels correctly identified as non-cloud pixels, FN and FP represent the incorrect detecting results, FP denotes the false positive outcomes, and FN denotes the false negative outcomes. P and N denote the number of cloud pixels and non-cloud pixels, respectively. In order to avoid a situation where the denominator is 0, we add a very small number $\epsilon = e^{-10}$ to the term where the denominator may be 0.

These metrics are calculated based on each large picture of 11 246 × 11 246, and the final result is obtained by averaging all large pictures in the test sample.

C. Implementation Details

In this study, all the experiments were programmed and implemented on Ubuntu 16.04. The implementation of the models is based on Python 3.6 and employing Keras 2.2.4 and TensorFlow 1.12 deep learning framework. The models are trained and evaluated on NVIDIA GEFORCE RTX 2080 Ti. The network uses the Adam optimization algorithm with a learning set to 0.00001 in training stage. For the BGR-NIR images, the batch size is set to 2, and the number of iterations is 30. The value of pixels was normalized between 0 and 1. In QTB loss function, $\gamma_1 = 0.9$ and $\gamma_2 = 0.1$, and these values are obtained through experimental testing.

V. Experimental Result and Discussion

A. Evaluation of CFAM

We performed a series of experiments to verify the effectiveness of the CFAM. We used different combinations of the color feature attention module and the texture feature attention module for training and testing. In order to verify the performance of CFAM, we cascaded the color feature attention module and texture feature attention module to form the final cascaded feature attention network (CFAN). In order to verify the effectiveness of color feature extraction, we added color feature attention to the encoder and then the attention weight was added to the coding network through concatenate in channel dimension. This network is called the color feature network (CFN). Then, we used texture feature attention in the coding network to design texture feature network (TFN). The structure of TFN is similar to CFN.

TABLE III

QUANTITATIVE COMPARISON OF U-NET, CFN, TFN, AND CFAN

Method	OA	Precision	Recall	F1-Score	Kappa	FAR
U-Net	96.73	89.89	89.65	86.98	82.30	3.29
CFN	97.13	90.55	**91.46**	89.99	85.00	4.96
TFN	97.22	**94.13**	88.59	90.13	85.75	**2.95**
CFAN	**97.35**	93.94	90.10	**91.53**	**87.31**	3.26

The difference is that the texture feature attention module is used to replace the color feature attention module. We trained and tested three networks, CFN, TFN, and CFAN, and compared the results with the basic encoder–decoder network U-Net [37]. The performance of these networks is given in Table III.

As shown in Table III, the color feature attention module and the texture feature attention module improve the performance of the network significantly. Specifically, the texture feature attention module has improved the OA and Precision. The OA has reached 97.22%, and Precision has increased by 4.24% compared to U-Net. In addition, F1-Score, Kappa, and FAR have also been improved. In other words, the main function of the texture feature attention module is to improve the detection accuracy of cloud pixels.

As for color feature attention module, the results show that the Recall has been improved significantly. This means that CFN can better identify cloud pixels in cloud pixels and non-cloud pixels correctly. This is due to the color feature extraction based on the dark channel prior.

In the CFAN, we used both the color feature attention module and texture feature attention module. The CFAN shows better performance as compared to the CFN and TFN in general. Its OA has increased to 97.35%, and F1-Score has improved to 91.53%. Kappa also has reached to 87.31%. This shows that cascaded the color feature attention module and the texture attention module has a better effect, compared with using a single module. A large number of experiments have also proved that the proposed CFAN shows high stability.

Therefore, it can be considered that the CFAM has shown excellent performance in the cloud detection of remote-sensing images by extracting color and texture features.

B. Evaluation of CA

The channel attention module is used to classify useless channel information and useful information. We used channel attention to assist in the interpretation of information at the decoder, so we added the channel attention module before the up-sampling of the decoder. The network was marked as CA-Net. We used BCE as the loss function of the network. We compared the performance of CA-Net with U-Net.

Moreover, we compared the network performance under different CA modules. At the decoding end, we had a total of five feature layers, which needed to go through four times of up-sampling and channel reduction. We controlled the number of CA modules as {1, 2, 3, 4}. The results are mentioned in Table IV.

TABLE IV

QUANTITATIVE COMPARISON OF CA-NET WITH DIFFERENT NUMBERS OF CA MODULE

Method	OA	Precision	Recall	F1-Score	Kappa	FAR
U-Net	96.73	89.89	**89.65**	86.98	82.30	3.29
CA-Net($n = 1$)	97.15	93.14	87.31	88.05	83.83	**2.32**
CA-Net($n = 2$)	97.19	92.73	88.82	89.38	85.06	2.33
CA-Net($n = 3$)	97.26	**93.66**	87.74	89.41	85.15	2.83
CA-Net($n = 4$)	**97.31**	93.16	89.07	**90.17**	**85.94**	2.54

The results proved that the model performs better as the number of channel attention module increases. When the CA modules are used before all up-sampling, the OA of the network reaches 97.31%, and the F1-Score reaches 90.17%. At the same time, FAR remains at a low level. This indicates that our proposed CA module can assist the decoder to interpret the information.

C. Evaluation of CFCA-Net

The CFCA-Net is based on the encoder–decoder structure, constructing the Dark&NSCT subnet on the encoder, using the multi-scale CFAM, and adding the channel attention module on the decoding end. The Dark&NSCT subnet extracts the color and texture features and injects the CFAM to the encoder to pay more attention to color and texture features. The channel attention module strengthens the fusion of the channel dimension information at the decoder. The QTB loss function and the BCE loss are also compared.

To verify the performance of the proposed CFCA-Net, we compared our algorithm with other centralized algorithms, including SegNet [38], DeepLab v3+ [42], RS-Net [25], and our previously proposed ADUI-Net [34].

1) Analysis of CFCA-Net: Fig. 11 shows the detection results of this method and comparison algorithm in different land-cover scenarios. We selected five representative scenes: urban, water, barren, snow, and vegetation. From the area marked by the red box in the figure, it can be seen that our algorithm shows better thin cloud detection performance. In the scene in the first column, a thin cloud on the left side of the image is seen, which looks such as a mountain range. Our algorithm detected this area.

In the water scene in the second column, there is a thin cloud at the top of the picture, visually indistinguishable from the underlying water. It can be seen that the comparison algorithms have different degrees of missed detection; ADUI-Net is over-detected, and our algorithm has a better detection performance.

In the snow scene presented in the fourth column, there are a large number of thin cloud areas, and only ADUI-Net and the method in this article have detected those thin cloud areas. The thin clouds do not completely cover the background of the ground objects, so they are easily confused with the underlying surface, and the detection is extremely difficult. It thus reflects the strong, thin cloud and edge detection performance of our algorithm.

Especially in the third scene and the last scene, we can see a thin cloud in the area marked by the blue frame in the RGB image. Due to the contrast of colors on the underlying

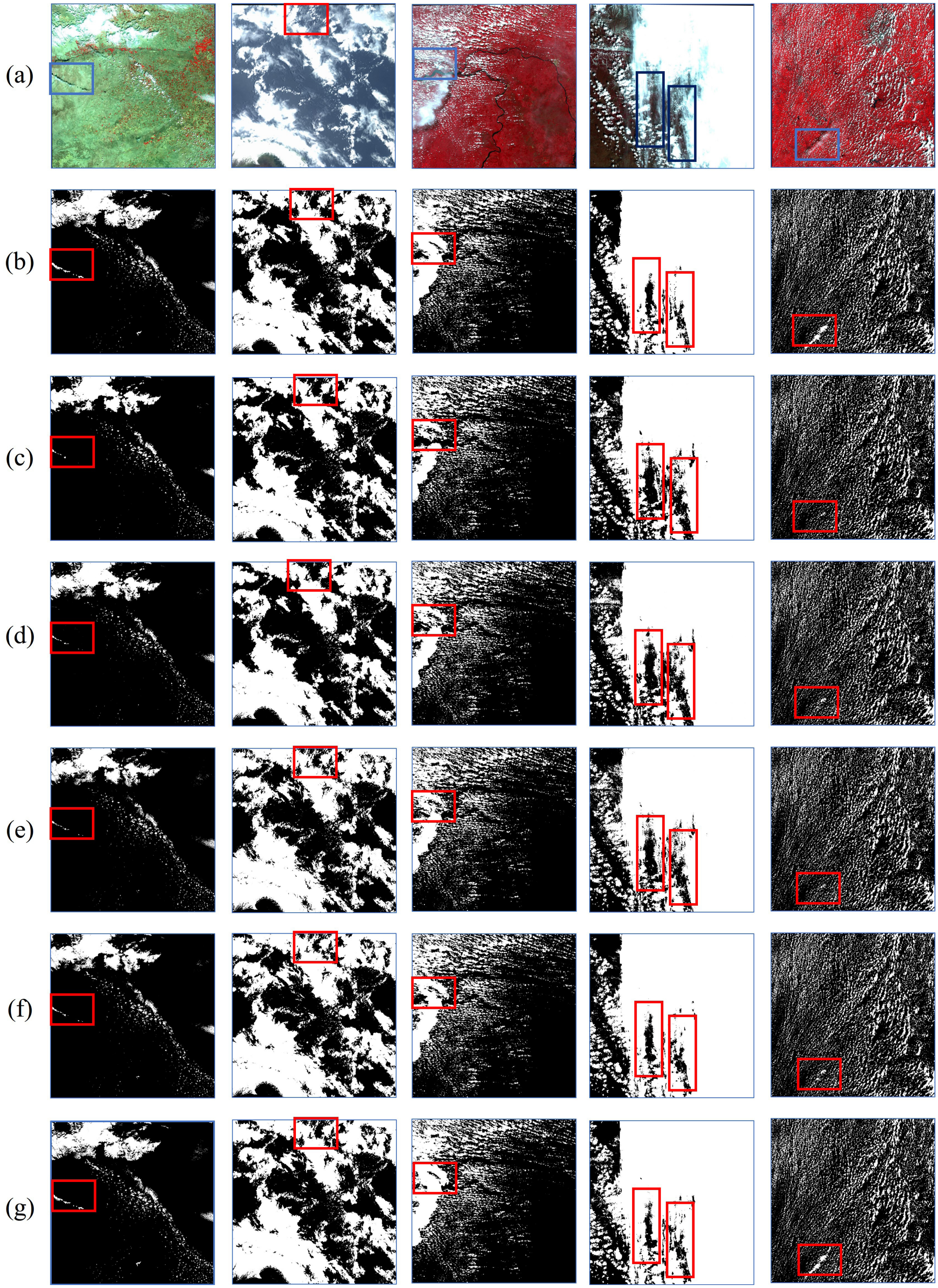

Fig. 11. Visual comparisons of the detection results. (a) Input images, (b) ground truth, and detection results obtained by (c) SegNet, (d) DeepLabv3+, (e) RS-Net, (f) ADUI-Net, and (g) our CFCA-Net.

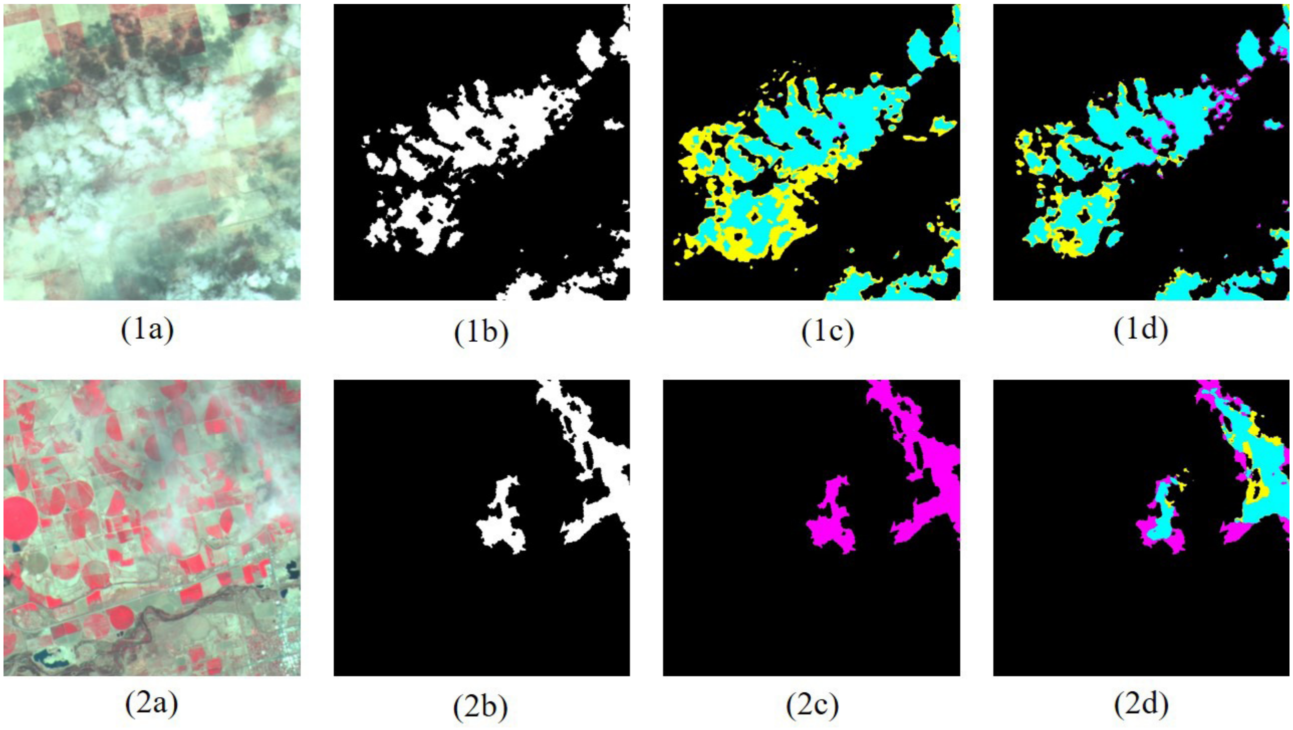

Fig. 12. Comparison of detection result obtained by BCE loss and QTB loss. (a) Original image, (b) ground truth, and results from (c) BCE loss, and (d) QTB loss. In (c) and (d), the cyan pixel indicates the pixel (TP) that is correctly detected as a cloud, the black pixel indicates the pixel that correctly detects the non-cloud (TN), and the yellow pixel indicates that the non-cloud pixel is incorrectly detected as a cloud pixel (FP), the purple pixel indicates that the cloud pixel is incorrectly detected as a non-cloud pixel (FN).

TABLE V
QUANTITATIVE COMPARISON WITH OTHER METHODS ON THE TESTING SET

Method	OA	Precision	Recall	F1-Score	Kappa	FAR
SegNet	96.02	94.25	84.04	86.78	82.72	**1.37**
DeepLabv3+	96.18	91.31	85.99	88.03	82.37	3.25
RS-Net	96.71	94.34	87.92	90.56	84.74	3.97
ADUI-Net	97.51	**95.98**	90.35	92.76	88.44	3.88
CFCA-Net (BCE)	97.45	94.54	**90.59**	92.31	88.17	2.22
CFCA-Net (QTB)	**97.55**	95.15	90.56	**92.79**	**88.48**	3.43

surface, it is difficult to distinguish the boundary of the cloud, even with the naked eye. Our algorithm detected the thin cloud area correctly. Compared with ADUI-Net, the detection performance of thin clouds has been further improved owing to the use of NSCT and quadtree binary loss in this method.

The metrics performance of these networks are described in Table V. It can be found that after using QTB loss, the OA is increased by 0.1% and the Precision is increased by 0.61% compared with BCE. This shows that QTB loss can effectively improve the accuracy of cloud detection models.

2) Analysis of QTB Loss: Fig. 12 shows the detection results obtained by QTB loss and BCE loss. As shown in Fig. 12(1c), it can be seen that BCE loss has a lot of false detections in the edge area. Compared with Fig. 12(1a), it can be seen that the edge of the cloud is blurry. Especially in the lower left corner, the boundary between the edge and the underlying surface is not clear. After training with the QTB loss, it can be seen from Fig. 12(1d) that the false detection of the edge has decreased. Most of the edges have good detection results, although there still a small range of false detections at some edges. This means that QTB loss can improve the cloud detection effect in the edge area. In the case of thin cloud, as shown in Fig. 12(2a), the underlying surface information is mixed with the cloud, and it is difficult to distinguish. The model trained with BCE loss failed to detect this cloud area, as shown in Fig. 12(2c). The model trained with QTB loss detected this thin cloud. As shown in Fig. 12(2d), although there are still false detection at the edge of the thin cloud, the main part of the cloud is correctly detected.

Fig. 13 shows the convergence during the training. As shown in Fig. 13, the convergence speed of QTB loss in the first few epochs is significantly higher than that of BCE loss, and the final OA is also higher than that of BCE loss. Therefore, we can conclude that QTB loss can improve the detection effect of thin clouds and edges and improve the convergence speed.

We also compared the cloud detection results in various scenarios, as mentioned in Table VI. From these results, we can see that the proposed algorithm has a better performance in

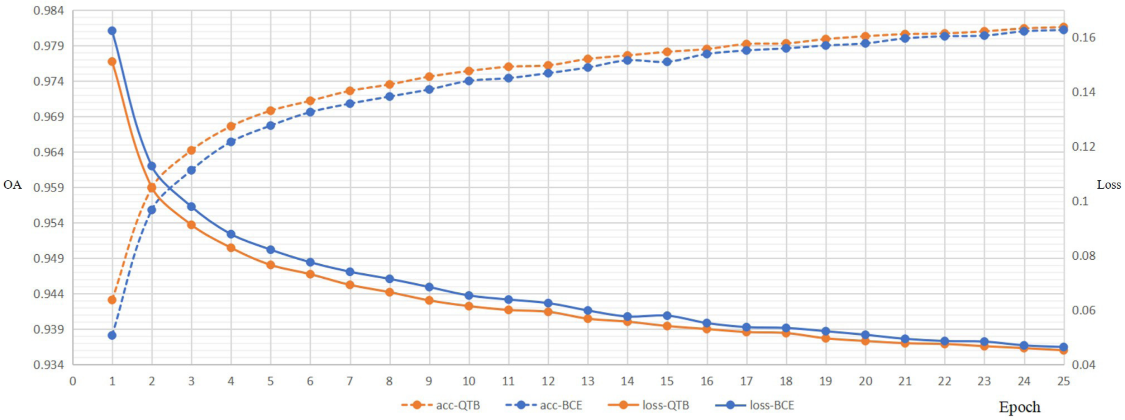

Fig. 13. Comparison of the convergence speed of loss between BCE loss and QTB loss. The basic network is CFCA-Net.

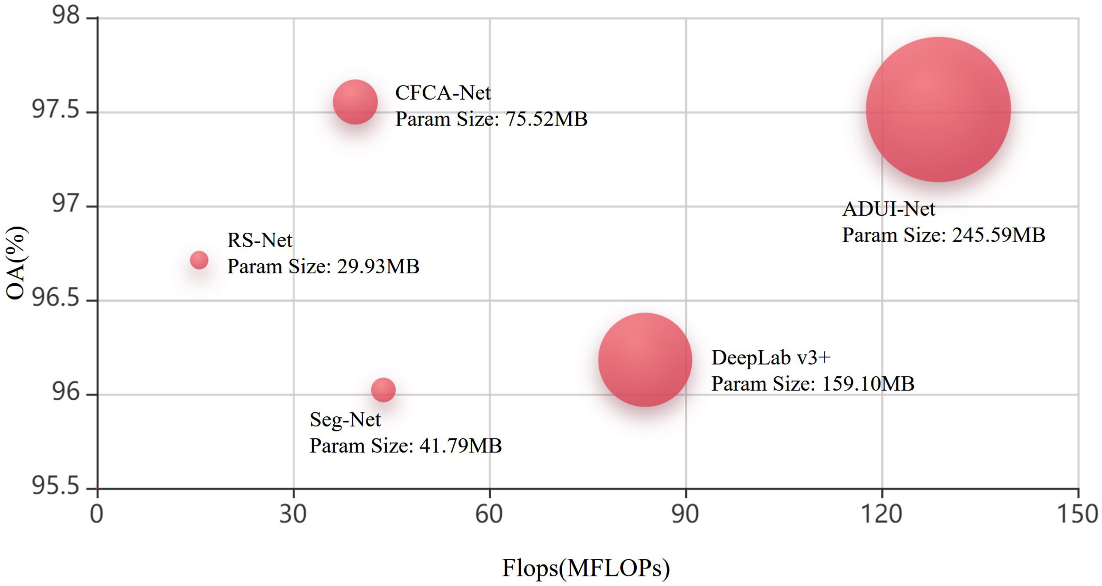

Fig. 14. Comparison of parameters with other methods.

most scenarios, and the performance of some scenarios is slightly worse than ADUI-Net. However, the overall performance on the entire test set is better than ADUI-Net.

At the same time, because the ADUI-Net uses Up-Down blocks with numerous parameters, the network has a huge number of parameters and calculations. To evaluate the complexity of the network, we counted the parameters and floating-point calculations of the model, as shown in Table VII and Fig. 14. The parameters of RS-Net and SegNet are small, but due to the simple model, the detection performance is not outstanding. DeepLab v3+ used ASPP, and therefore the parameters are relatively large. However, the detection performance is not good, which shows that merely increasing the complexity of the network cannot improve the performance of cloud detection. ADUI-Net designed the Up-Down block and wavelet transform to extract the texture characteristics of the cloud according to the characteristics of the cloud, showing high detection performance. Because the Up block and the Down block perform a large number of convolution operations on the feature map to extract features, this makes the amount of network parameters increase rapidly. Our CFCA-Net also use texture features, but our cascaded attention module uses less convolution to effectively extract color and texture features. Therefore, while our method achieves higher detection performance, the amount of parameters and complexity are still at a relatively low level. It can be seen that when our algorithm achieves the same or even better performance than ADUI-Net, while the parameter amount is only 30% of the latter. Therefore, we can conclude that our method is advanced compared with other cloud detection methods.

VI. Conclusion

With the development of deep learning theory, the convolution neural network, based on deep learning, has been used in remote-sensing image cloud detection research and achieved great results. Especially for the remote-sensing images with

TABLE VI
EVALUATION METRICS COMPARISON IN AREAS OF BARREN, VEGETATION, SNOW, WATER, AND URBAN FOR OUR METHODS WITH OTHER METHODS

Method	Land-cover types	OA	Precision	Recall	F1-Score	Kappa	FAR
SegNet	All	96.02	94.25	84.04	86.78	82.72	**1.37**
	Barren	97.64	93.99	90.13	79.28	78.08	0.20
	Vegetation	97.52	93.92	86.64	90.00	88.38	0.84
	Snow	94.81	92.70	84.64	88.47	78.94	3.00
	Water	90.09	98.55	85.75	91.43	70.41	3.30
	Urban	99.29	85.27	86.39	85.83	85.45	0.38
DeepLab v3+	All	96.18	91.31	85.99	88.03	82.37	3.25
	Barren	97.83	86.94	96.62	81.22	80.04	0.88
	Vegetation	97.54	91.50	87.85	88.93	87.31	1.16
	Snow	94.32	87.50	84.00	85.70	75.77	4.60
	Water	91.07	98.03	87.39	92.03	71.67	11.34
	Urban	99.30	89.10	92.36	90.70	86.09	0.47
RS-Net	All	96.71	94.34	87.92	90.56	84.74	3.97
	Barren	97.74	95.21	72.84	81.60	80.45	0.17
	Vegetation	98.21	95.22	90.49	92.61	91.42	0.47
	Snow	94.85	87.55	87.43	87.42	78.34	4.33
	Water	91.64	95.76	90.63	92.70	69.68	1.81
	Urban	99.53	89.10	92.36	90.70	90.44	2.87
ADUI-Net	All	97.51	**95.98**	90.35	92.76	88.44	3.88
	Barren	98.02	**96.43**	75.44	83.83	82.82	**0.16**
	Vegetation	98.55	**96.17**	**93.26**	**94.66**	**93.70**	0.61
	Snow	**97.49**	**97.54**	88.39	**92.52**	**87.94**	3.61
	Water	93.46	95.52	**93.34**	94.17	76.40	1.76
	Urban	**99.58**	91.49	**91.83**	**91.66**	**91.44**	0.22
CFCA-Net	All	**97.55**	95.15	**90.56**	**92.79**	**88.48**	3.43
	Barren	**98.15**	95.48	**77.75**	**85.55**	**84.57**	0.21
	Vegetation	**98.59**	95.60	92.87	94.13	93.19	**0.48**
	Snow	96.91	90.21	**91.39**	90.79	85.32	**1.72**
	Water	**93.79**	**96.65**	93.02	**94.60**	**78.27**	**1.95**
	Urban	99.57	**92.62**	89.73	91.15	90.92	**0.18**

TABLE VII
COMPARISON OF PARAMETERS WITH OTHER METHODS

Model	OA	Param Size(MB)	Flops($MFLOPs$)
SegNet	96.02	41.79	43.86
DeepLabv3+	96.18	159.10	83.90
RS-Net	96.71	29.93	15.69
ADUI-Net	97.51	245.59	128.75
CFCA-Net	97.55	75.52	39.57

few spectral segments, the cloud detection method based on deep learning can extract more useful information from limited spectral segments with more advantages than traditional methods. However, feature extraction using convolutional neural network carries redundant information. This information does not help in the detection of cloud region and leads to false detection affecting the performance of the network. In view of the large difference of color and texture features between the cloud region and underlying surface, this article proposes a cascade feature attention module to extract the color and texture features of cloud region. This article also designs a channel attention module to remove the redundant information and retain useful information. Moreover, this article optimizes the loss function to improve the performance of edge detection.

For GF-1 WFV image, the multi-scale cascade feature attention module and multi-scale channel attention model, proposed in this article, significantly improve the detection accuracy of thin cloud. To further evaluate the effectiveness of the proposed algorithm, we compared SegNet, DeepLabV3+, RS-Net, and ADUI-Net. Our algorithm shows better performance. Experimental results show the excellent performance of the proposed algorithm. On the Gaofen-1 WFV dataset, the overall accuracy of our method reached 97.55%. Subjective cloud detection results also proved the effectiveness of our algorithm.

References

[1] Y. Li, R. Yu, Y. Xu, and X. Zhang, "Spatial distribution and seasonal variation of cloud over China based on ISCCP data and surface observations," *J. Meteorolog. Soc. Jpn. Ser. II*, vol. 82, no. 2, pp. 761–773, 2004.

[2] W. B. Rossow and L. C. Garder, "Cloud detection using satellite measurements of infrared and visible radiances for ISCCP," *J. Climate*, vol. 6, no. 12, pp. 2341–2369, Dec. 1993.

[3] R. R. Irish, J. L. Barker, S. N. Goward, and T. Arvidson, "Characterization of the Landsat-7 ETM+ automated cloud-cover assessment (ACCA) algorithm," *Photogramm. Eng. Remote Sens.*, vol. 72, no. 10, pp. 1179–1188, 2006.

[4] 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 J. Sel. Topics Appl. Earth Observ. Remote Sens.*, vol. 10, no. 11, pp. 4898–4908, Nov. 2017.

[5] Z. Zhu and C. E. Woodcock, "Object-based cloud and cloud shadow detection in Landsat imagery," *Remote Sens. Environ.*, vol. 118, pp. 83–94, Mar. 2012.

[6] Z. Zhu and C. E. Woodcock, "Automated cloud, cloud shadow, and snow detection in multitemporal Landsat data: An algorithm designed specifically for monitoring land cover change," *Remote Sens. Environ.*, vol. 152, pp. 217–234, Sep. 2014.

[7] S. Qiu, B. He, Z. Zhu, Z. Liao, and X. Quan, "Improving Fmask cloud and cloud shadow detection in mountainous area for Landsats 4–8 images," *Remote Sens. Environ.*, vol. 199, pp. 107–119, Sep. 2017.

[8] Y. Chen, Q. Weng, L. Tang, X. Zhang, M. Bilal, and Q. Li, "Thick clouds removing from multitemporal Landsat images using spatiotemporal neural networks," *IEEE Trans. Geosci. Remote Sens.*, early access, Dec. 24, 2021, doi: 10.1109/TGRS.2020.3043980.

[9] Z. Li, H. Shen, H. Li, G. Xia, and L. Zhang, "Multi-feature combined cloud and cloud shadow detection in GaoFen-1 wide field of view imagery," *Remote Sens. Environ.*, vol. 191, pp. 342–358, Mar. 2017.

[10] Z. An and Z. Shi, "Scene learning for cloud detection on remote-sensing images," *IEEE J. Sel. Topics Appl. Earth Observ. Remote Sens.*, vol. 8, no. 8, pp. 4206–4222, Aug. 2015.

[11] S. Liu, Z. Zhang, B. Xiao, and X. Cao, "Ground-based cloud detection using automatic graph cut," *IEEE Geosci. Remote Sens. Lett.*, vol. 12, no. 6, pp. 1342–1346, Jun. 2015.

[12] P. Li, L. Dong, H. Xiao, and M. Xu, "A cloud image detection method based on SVM vector machine," *Neurocomputing*, vol. 169, no. 2, pp. 34–42, Dec. 2015.

[13] C. Cortes and V. Vapnik, "Support-vector networks," *Mach. Learn.*, vol. 20, no. 3, pp. 273–297, 1995.

[14] J. Munkres, "Algorithms for the assignment and transportation problems," *J. Soc. Ind. Appl. Math.*, vol. 5, no. 1, pp. 32–38, 1957.

[15] R. Osada, T. Funkhouser, B. Chazelle, and D. Dobkin, "Shape distributions," *ACM Trans. Graph.*, vol. 21, no. 4, pp. 807–832, 2002.

[16] C. Shi, Y. Wang, C. Wang, and B. Xiao, "Ground-based cloud detection using graph model built upon superpixels," *IEEE Geosci. Remote Sens. Lett.*, vol. 14, no. 5, pp. 719–723, May 2017.

[17] D. G. Lowe, "Distinctive image features from scale-invariant keypoints," *Int. J. Comput. Vis.*, vol. 60, no. 2, pp. 91–110, 2004.

[18] R. Achanta, A. Shaji, K. Smith, A. Lucchi, P. Fua, and S. Süsstrunk, "SLIC superpixels compared to state-of-the-art superpixel methods," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 34, no. 11, pp. 2274–2282, Nov. 2012.

[19] J. Key and R. G. Barry, "Adaptation of the ISCCP cloud detection algorithm to combined AVHRR and SMMR Arctic data," in *Proc. 12th Canadian Symp. Remote Sensing Geosci. Remote Sens. Symp.*, vol. 1, 1989, pp. 188–191.

[20] R. L. Bankert, "Cloud classification of AVHRR imagery in maritime regions using a probabilistic neural network," *J. Appl. Meteorol. Climatol.*, vol. 33, no. 8, pp. 909–918, 1994.

[21] S. C. Q. Jianhua, "Cloud classification for NOAA-AVHRR data by using a neural network," *Acta Meteorolog. Sinica*, pp. 250–256, 2002.

[22] F. Xie, M. Shi, Z. Shi, J. Yin, and D. Zhao, "Multilevel cloud detection in remote sensing images based on deep learning," *IEEE J. Sel. Topics Appl. Earth Observ. Remote Sens.*, vol. 10, no. 8, pp. 3631–3640, Aug. 2017.

[23] S. Ji, P. Dai, M. Lu, and Y. Zhang, "Simultaneous cloud detection and removal from bitemporal remote sensing images using cascade convolutional neural networks," *IEEE Trans. Geosci. Remote Sens.*, vol. 59, no. 1, pp. 732–748, Jan. 2021.

[24] M. Luotamo, S. Metsämäki, and A. Klami, "Multiscale cloud detection in remote sensing images using a dual convolutional neural network," *IEEE Trans. Geosci. Remote Sens.*, vol. 59, no. 6, pp. 4972–4983, Jun. 2021.

[25] J. H. Jeppesen, R. H. Jacobsen, F. Inceoglu, and T. S. Toftegaard, "A cloud detection algorithm for satellite imagery based on deep learning," *Remote Sens. Environ.*, vol. 229, pp. 247–259, May 2019.

[26] M. Segal-Rozenhaimer, A. Li, K. Das, and V. Chirayath, "Cloud detection algorithm for multi-modal satellite imagery using convolutional neural-networks (CNN)," *Remote Sens. Environ.*, vol. 237, Feb. 2020, Art. no. 111446.

[27] Z. Wu, J. Li, Y. Wang, Z. Hu, and M. Molinier, "Self-attentive generative adversarial network for cloud detection in high resolution remote sensing images," *IEEE Geosci. Remote Sens. Lett.*, vol. 17, no. 10, pp. 1792–1796, Oct. 2020.

[28] J. Yang, J. Guo, H. Yue, Z. Liu, H. Hu, and K. Li, "CDNet: CNN-based cloud detection for remote sensing imagery," *IEEE Trans. Geosci. Remote Sens.*, vol. 57, no. 8, pp. 6195–6211, Aug. 2019.

[29] J. Guo, J. Yang, H. Yue, H. Tan, C. Hou, and K. Li, "CDnetV2: CNN-based cloud detection for remote sensing imagery with cloud-snow coexistence," *IEEE Trans. Geosci. Remote Sens.*, vol. 59, no. 1, pp. 700–713, Jan. 2020.

[30] Q. He, X. Sun, Z. Yan, and K. Fu, "DabNet: Deformable contextual and boundary-weighted network for cloud detection in remote sensing images," *IEEE Trans. Geosci. Remote Sens.*, early access, Jan. 5, 2021, doi: 10.1109/TGRS.2020.3045474.

[31] J. Zhang, Q. Zhou, X. Shen, and Y. Li, "Cloud detection in high-resolution remote sensing images using multi-features of ground objects," *J. Geovisualization Spatial Anal.*, vol. 3, no. 2, pp. 1–9, 2019.

[32] Y. Wang, J. Zhang, and G. shuang, "Hue–saturation–intensity and texture feature-based cloud detection algorithm for unmanned aerial vehicle images," *Int. J. Adv. Robotic Syst.*, vol. 17, no. 3, 2020, Art. no. 1729881420903532.

[33] J. Zhang, Q. Zhou, J. Wu, Y. Wang, and H. Wang, "A cloud detection method using convolutional neural network based on Gabor transform and attention mechanism with dark channel SubNet for remote sensing image," *Remote Sens.*, vol. 12, no. 19, p. 3261, 2020.

[34] J. Zhang, H. Wang, Y. Wang, Q. Zhou, and Y. Li, "Deep network based on up and down blocks using wavelet transform and successive multi-scale spatial attention for cloud detection," *Remote Sens. Environ.*, vol. 261, Aug. 2021, Art. no. 112483.

[35] J. Zhang, Y. Wang, H. Wang, J. Wu, and Y. Li, "CNN cloud detection algorithm based on channel and spatial attention and probabilistic upsampling for remote sensing image," *IEEE Trans. Geosci. Remote Sens.*, Aug. 26, 2021, doi: 10.1109/TGRS.2021.3105424.

[36] J. Long, E. Shelhamer, and T. Darrell, "Fully convolutional networks for semantic segmentation," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2015, pp. 3431–3440.

[37] O. Ronneberger, P. Fischer, and T. Brox, "U-Net: Convolutional networks for biomedical image segmentation," in *Proc. Int. Conf. Med. Image Comput. Comput.-Assist. Intervent*. Munich, Germany: Springer, Oct. 2015, pp. 234–241.

[38] V. Badrinarayanan, A. Kendall, and R. Cipolla, "SegNet: A deep convolutional encoder-decoder architecture for image segmentation," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 39, no. 12, pp. 2481–2495, Dec. 2017.

[39] L.-C. Chen, G. Papandreou, I. Kokkinos, K. Murphy, and A. L. Yuille, "Semantic image segmentation with deep convolutional nets and fully connected CRFs," 2014, *arXiv:1412.7062*. [Online]. Available: http://arxiv.org/abs/1412.7062

[40] L.-C. Chen, G. Papandreou, I. Kokkinos, K. Murphy, and A. L. Yuille, "DeepLab: Semantic image segmentation with deep convolutional nets, atrous convolution, and fully connected CRFs," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 40, no. 4, pp. 834–848, Apr. 2017.

[41] L.-C. Chen, G. Papandreou, F. Schroff, and H. Adam, "Rethinking atrous convolution for semantic image segmentation," 2017, *arXiv:1706.05587*. [Online]. Available: http://arxiv.org/abs/1706.05587

[42] L.-C. Chen, Y. Zhu, G. Papandreou, F. Schroff, and H. Adam, "Encoder-decoder with atrous separable convolution for semantic image segmentation," in *Proc. Eur. Conf. Comput. Vis. (ECCV)*, 2018, pp. 801–818.

[43] F. Yu and V. Koltun, "Multi-scale context aggregation by dilated convolutions," 2015, *arXiv:1511.07122*. [Online]. Available: http://arxiv.org/abs/1511.07122

[44] A. van den Oord *et al.*, "WaveNet: A generative model for raw audio," 2016, *arXiv:1609.03499*. [Online]. Available: http://arxiv.org/abs/1609.03499

[45] J. Hu, L. Shen, and G. Sun, "Squeeze-and-excitation networks," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2018, pp. 7132–7141.

[46] S. Woo, J. Park, J.-Y. Lee, and I. S. Kweon, "CBAM: Convolutional block attention module," in *Proc. Eur. Conf. Comput. Vis. (ECCV)*, Sep. 2018, pp. 3–19.

[47] C. Zhou, M. Wu, and S.-K. Lam, "SSA-CNN: Semantic self-attention CNN for pedestrian detection," 2019, *arXiv:1902.09080*. [Online]. Available: http://arxiv.org/abs/1902.09080

[48] J. Cao, Q. Chen, J. Guo, and R. Shi, "Attention-guided context feature pyramid network for object detection," 2020, *arXiv:2005.11475*. [Online]. Available: http://arxiv.org/abs/2005.11475

[49] H. Li, P. Xiong, J. An, and L. Wang, "Pyramid attention network for semantic segmentation," 2018, *arXiv:1805.10180*. [Online]. Available: http://arxiv.org/abs/1805.10180

[50] Z. Huang, X. Wang, L. Huang, C. Huang, Y. Wei, and W. Liu, "CCNet: Criss-cross attention for semantic segmentation," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Nov. 2019, pp. 603–612.

[51] Y. Liu, H. Shi, H. Shen, Y. Si, X. Wang, and T. Mei, "A new dataset and boundary-attention semantic segmentation for face parsing," in *Proc. AAAI Conf. Artif. Intell.*, vol. 34, no. 7, 2020, pp. 11637–11644.

[52] J. Li *et al.*, "Hyperspectral image super-resolution by band attention through adversarial learning," *IEEE Trans. Geosci. Remote Sens.*, vol. 58, no. 6, pp. 4304–4318, Jun. 2020.

[53] J. Li *et al.*, "Hybrid 2-D–3-D deep residual attentional network with structure tensor constraints for spectral super-resolution of RGB images," *IEEE Trans. Geosci. Remote Sens.*, vol. 59, no. 3, pp. 2321–2335, Mar. 2021.

[54] A. L. da Cunha, J. Zhou, and M. N. Do, "The nonsubsampled contourlet transform: Theory, design, and applications," *IEEE Trans. Image Process.*, vol. 15, no. 10, pp. 3089–3101, Oct. 2006.

[55] Y. Wu, P. Zhang, M. Li, Q. Zhang, F. Wang, and L. Jia, "SAR image multiclass segmentation using a multiscale and multidirection triplet Markov fields model in nonsubsampled contourlet transform domain," *Inf. Fusion*, vol. 14, pp. 441–449, Oct. 2013.

[56] A. Heshmati, M. Gholami, and A. Rashno, "Scheme for unsupervised colour–texture image segmentation using neutrosophic set and nonsubsampled contourlet transform," *IET Image Process.*, vol. 10, no. 6, pp. 464–473, Jun. 2016.

[57] F. Wang *et al.*, "Residual attention network for image classification," in *Proc. IEEE Conf. Comput. Vis. pattern Recognit.*, Jul. 2017, pp. 3156–3164.

[58] C. H. Sudre, W. Li, T. Vercauteren, S. Ourselin, and M. J. Cardoso, "Generalised dice overlap as a deep learning loss function for highly unbalanced segmentations," in *Deep Learning in Medical Image Analysis and Multimodal Learning for Clinical Decision Support*. Québec City, QC, Canada: Springer, Sep. 2017, pp. 240–248.

[59] Y. Guo, X. Cao, B. Liu, and M. Gao, "Cloud detection for satellite imagery using attention-based U-Net convolutional neural network," *Symmetry*, vol. 12, no. 6, p. 1056, Jun. 2020.

[60] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," 2014, *arXiv:1412.6980*. [Online]. Available: http://arxiv.org/abs/1412.6980

[61] S. Ruder, "An overview of gradient descent optimization algorithms," 2016, *arXiv:1609.04747*. [Online]. Available: http://arxiv.org/abs/1609.04747

[62] Z. F. Muhsin, A. Rehman, A. Altameem, T. Saba, and M. Uddin, "Improved quadtree image segmentation approach to region information," *Imag. Sci. J.*, vol. 62, no. 1, pp. 56–62, Jan. 2014.

[63] K. He, J. Sun, and X. Tang, "Single image haze removal using dark channel prior," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 33, no. 12, pp. 2341–2353, Dec. 2011.

[64] Y. Li, W. Chen, Y. Zhang, C. Tao, R. Xiao, and Y. Tan, "Accurate cloud detection in high-resolution remote sensing imagery by weakly supervised deep learning," *Remote Sens. Environ.*, vol. 250, Dec. 2020, Art. no. 112045.

[65] J. Yang *et al.*, "An automated cloud detection method based on the green channel of total-sky visible images," *Atmos. Meas. Techn.*, vol. 8, no. 11, pp. 4671–4679, Nov. 2015.

[66] Z. Li, H. Shen, Q. Cheng, Y. Liu, S. You, and Z. He, "Deep learning based cloud detection for medium and high resolution remote sensing images of different sensors," *ISPRS J. Photogramm. Remote Sens.*, vol. 150, pp. 197–212, Apr. 2019.

[67] M. Wieland, Y. Li, and S. Martinis, "Multi-sensor cloud and cloud shadow segmentation with a convolutional neural network," *Remote Sens. Environ.*, vol. 230, Sep. 2019, Art. no. 111203.

[68] Y. Benjamini and Y. Hochberg, "Controlling the false discovery rate: A practical and powerful approach to multiple testing," *J. Roy. Statist. Soc., Ser. B*, vol. 57, pp. 289–300, Jan. 1995.

Jing Zhang received the B.Sc. degree in information engineering and the Ph.D. degree in information and communication engineering from Xi'an Jiaotong University, Xi'an, China, in 2003 and 2009, respectively.

From September 2007 to September 2008, she was the Visiting Ph.D. Student with Mississippi State University, Starkville, MS, USA. She is currently with the State Key Laboratory of Integrated Services Networks, Xidian University, Xi'an. Her research interests include image processing and the design of unmanned systems.

Jun Wu is currently pursuing the master's degree with Xidian University, Xi'an, China.

His research interests include computer vision, remote sensing, and machine learning.

Hui Wang is currently pursuing the master's degree with Xidian University, Xi'an, China.

Her research interests include computer vision, remote sensing, and image compression algorithm.

Yuchen Wang is currently pursuing the master's degree with Xidian University, Xi'an, China.

His research interests include computer vision, remote sensing, and deep learning.

Yunsong Li (Member, IEEE) received the bachelor's degree in image transmission and processing, the master's degree in communication and information systems, and the Ph.D. degree in signal and information processing from Xidian University, Xi'an, China, in 1996, 1999, and December 2002, respectively.

In 2009, he was selected into the New Century Excellent Talents Program of the Ministry of Education. He is currently the Deputy Dean of the Aerospace Research Institute, Xidian University, the Academic Leader of communication and information systems, and a Ph.D. Supervisor of communication and information systems.

Dr. Li is also a member of the Scientific Application Expert Committee of lunar exploration engineering and the Deep Space Exploration Technology Professional Committee of the Chinese Society of Astronautics, and the Standing Director of the Shaanxi Provincial Graphics and Image Society.