CNN Cloud Detection Algorithm Based on Channel and Spatial Attention and Probabilistic Upsampling for Remote Sensing Image

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

***Abstract*— In the field of remote sensing image, how to transmit image information more efficiently with limited bandwidth has always been a research hotspot. Compared with other ground objects, cloud pixels in remote sensing image are invalid information, so it is a meaningful research work to remove cloud before transmitting image and reduce the waste of useless information. In remote sensing image, due to the existence of thin clouds and the complexity of the underlying surface, most of the cloud detection algorithms struggle to achieve effective separation of clouds and ground objects. A deep learning (DL) cloud detection algorithm based on attention mechanism and probability upsampling has been proposed in this article. In order to enhance the information of the key areas, in the channel attention module, crucial information is highlighted in the channel dimension of the encoder, and the useless information is weakened. The spatial attention module is in the spatial dimension. The information fusion between each point in the image is strengthened. To reduce the information loss caused by the down-sampling module, a probabilistic upsampling block (PUB) is proposed to restore the image. Eventually, experiments are performed on Gaofen-1WFV data, and the results indicate that the algorithm proposed in this article has better detection results than other cloud detection algorithms in different scenarios.**

***Index Terms*— Channel attention, cloud detection, probabilistic upsampling, spatial attention.**

I. Introduction

With the development of remote sensing technology, satellite images are increasingly being used in various research [1]–[4]. An increasing amount of remote sensing data is being used in agriculture, environmental protection, urban development, military, the monitoring of land changes, hydrology, and so on [5]–[8]. However, nearly 70% of the earth's surface is covered by clouds [9], including thick clouds and thin clouds. The underlying surface of thick clouds cannot be known from the images, undoubtedly, the areas covered by clouds in remote sensing images are invalid information [10]. Although thin clouds do not completely cover the ground features, they still cannot fully know the ground features under the thin clouds when mixed with the underlying surface, therefore, thin clouds and thick clouds are the same invalid information. Because the transmission bandwidth on remote sensing satellites is limited, when these cloud regions are detected and identified as invalid information, the Region of Interest compression method will greatly improve the transmission efficiency between satellite and ground. Therefore, accurate and efficient detection of cloud regions is a hot spot in remote sensing image preprocessing.

Traditionally, the research methods of cloud detection are the combination of multi-band threshold, texture analysis, pattern recognition, and so on. Li *et al.* [11] presented the MFC algorithm, which combined band threshold with texture analysis method, firstly, threshold segmentation based on spectral features was realized, followed by which, a preliminary cloud mask was generated through mask thinning guided filtering, finally, the geometric features and texture features were combined to improve the cloud detection results. Li *et al.* [12] adjusted the segmentation threshold value by analyzing the physical properties of clouds and made it more suitable for mitigating the effects of clouds. The textural feature difference between clouds and the underlying surface is strengthened before texture identification. Using some morphology operations were eventually used to further refine the coarse cloud regions and extract the thin clouds. Chen *et al.* [13] aimed at the problem of complicated types of cloud and land. Firstly, the image to be detected was enhanced, and then the texture features of the image were analyzed in multi-scale space to distinguish between cloud and ground. Liu *et al.* [14] proposed a thin cloud removal method based on the cloud physical model, which uses the correction method and adaptive brightness factor to decrease the effect of transmission and obtain the final image. The results show that the method can more effectively remove thin clouds, improve the contrast of the image, and retain more details. These traditional methods usually require a lot of time for adjusting the parameters and tuning threshold. Also, this multistage process usually has poor detection accuracy.

With the development of new technologies, artificial neural networks (ANNs) have achieved impressive development. ANN is basically a mathematical model that simulates the human brain processing information. It is composed of a large number of processing units and can independently process multiple sets of information. The advantage of ANN is that it

Manuscript received April 28, 2021; revised June 7, 2021 and July 17, 2021; accepted August 10, 2021. Date of publication August 26, 2021; date of current version January 26, 2022. This work was supported in part by the Natural Science Foundation of China under Grant 61801359 and in part by the Pre-Research of the "Thirteenth Five-Year-Plan" of China Grant 305020903. *(Corresponding author: Jing Zhang.)*

The authors are with the State Key Laboratory of Integrated Services Networks, Xidian University, Xi'an 710071, China (e-mail: jingzhang@xidian.edu.cn).

Digital Object Identifier 10.1109/TGRS.2021.3105424

can deal with a lot of nonlinear problems and find the effective and optimal solution of the model by some constraints. In the 1990s, Key and Barry [15] first used this network to cloud detection of remote sensing images, and then more and more researchers began to use it in the field of cloud detection. Shi [16] based on the five-channel data of NOAA-AVHRR, used a simple neural network model to classify the images, including cumulonimbus, cumulus, cumulus, cirrus, medium cloud, low cloud and land, water, and unknown pixels. In 2006, Hinton put forward the concept of deep learning (DL). In the DL network, single-layer neurons are firstly constructed layer by layer so that a single-layer network is trained every time. After all the layers are trained, the optimization process begins. After multi-layer nonlinear feature extraction, high-level features with strong expression ability can be derived, and DL of data features can be realized without human participation. With the improvement of DL systems and the contribution of many scientific workers, in recent years, the deep convolutional neural network has achieved substantial success in the field of computer vision [17]–[21]. End to end does not require any human intervention, and through its powerful feature expression ability, it has become the prime research method in many fields of image processing [24]. Zhang *et al.* [25] integrated wavelet features into the DL network and achieved the task of speckle removal for SAR image. Yolo series [26] uses an end-to-end DL algorithm to accomplish the task of high-speed target detection.

In the exploration of DL methods for cloud detection, some new methods are proposed, which significantly improve the performance of cloud detection. By comparing the traditional cloud detection methods [28], it can be concluded that the feature learned by the convolutional neural network is better than the traditional manual feature. Owing to the complexity and diversity of the underlying surface, it is usually difficult to identify thin clouds compared to thick clouds, the multi-scale feature convolution neural network proposed by Shao *et al.* [29] can detect thin clouds as well as thick clouds at the same time, and the detection results are impressive. Li *et al.* [10] proposed a multi-scale convolutional feature fusion method based on DL. Dense connection groups are added in the symmetric encoder-decoder module to seek local and global information; the algorithm exhibits good detection performance in bright regions. Segal-Rozenhaimer *et al.* [30] proposed a convolutional neural network that can be adaptively applied to a variety of datasets. The robustness of this algorithm is verified by experiments. Aiming at the complicated underlying surfaces and the variety of cloud types, Liu *et al.* [14] proposed an innovative model named fuzzy auto encode model (FAEM), which combines the coding network and fuzzy function to achieve high-precision cloud detection of remote sensing image in complex environments.

In this article, we propose a cloud detection algorithm based on a convolutional neural network. Starting from the particularity of cloud detection task, we mainly focus on the relationship between the spectral segments of multispectral image and the points of spatial dimension. At the same time, we focus on the texture complexity of cloud images and a large number of thin clouds.

II. Background of U-Net

U-Net [32] is an image segmentation algorithm based on FCN-network architecture [33], which has been widely used in various image segmentation fields, such as medical image segmentation, industrial detection, and satellite image segmentation. U-Net network consists of two parts: encoder and decoder. The encoder is similar to VGG-Net [34] for feature extraction. Through convolution and downsampling, the input image size is compressed gradually, and the number of channels becomes more. In the decoder, the image is recovered by linking the convolution layer and the upsampling layer, and fused with the corresponding size feature extraction layer of the coding segment after each upsampling. At the final layer, a convolution is used to map each feature vector to the desired number of classes.

As an efficient and lightweight convolutional neural network, U-Net has attracted the attention of many scholars [35]–[37]. U-Net was first applied to medicine image segmentation. With the later expansion, U-Net is introduced into the field of remote sensing image segmentation. For example, Jeppesen *et al.* [38] and others have made improvements based on U-Net and proposed RS-Net remote sensing cloud detection algorithm. Guo *et al.* [39] added a channel attention module to U-Net, and realized efficient cloud and non-cloud segmentation algorithm.

In this article, our proposed algorithm based on encode–decode DL image segmentation network U-Net. We added a channel attention module between the decode and the encode of each layer, A spatial attention module is added in the last layer of the network, to improve the information loss caused by downsampling, probabilistic upsampling block (PUB) is added, in order to verify the effectiveness of the algorithm, and high score data is used to carry out the experiments.

III. Method

In this article, our proposed algorithm consists of three parts: channel attention module, spatial attention module, and probability upsampling module. To adaptively adjust the characteristic response value of each channel, the channel attention module fuses the information of the encoder-decoder, and models the dependency relationship between the channels. The obtained information is added to the decoder for image restoration. By modeling different position relationships, the spatial attention module adjusts the feature response value of the spatial dimension. The probabilistic upsampling module fuses the downsampling information from encode to decode, which improves the image edge problem caused by the roughness of downsampling as demonstrated in Fig. 1.

A. Background of Attention Mechanism

The visual attention mechanism is a special brain signal processing mechanism of human vision. By quickly scanning the global image, human vision can identify the target area that needs to be focused, that is, the focus of attention, and then invest more attention resources in this area to obtain more detailed information of the concerned target, so as to

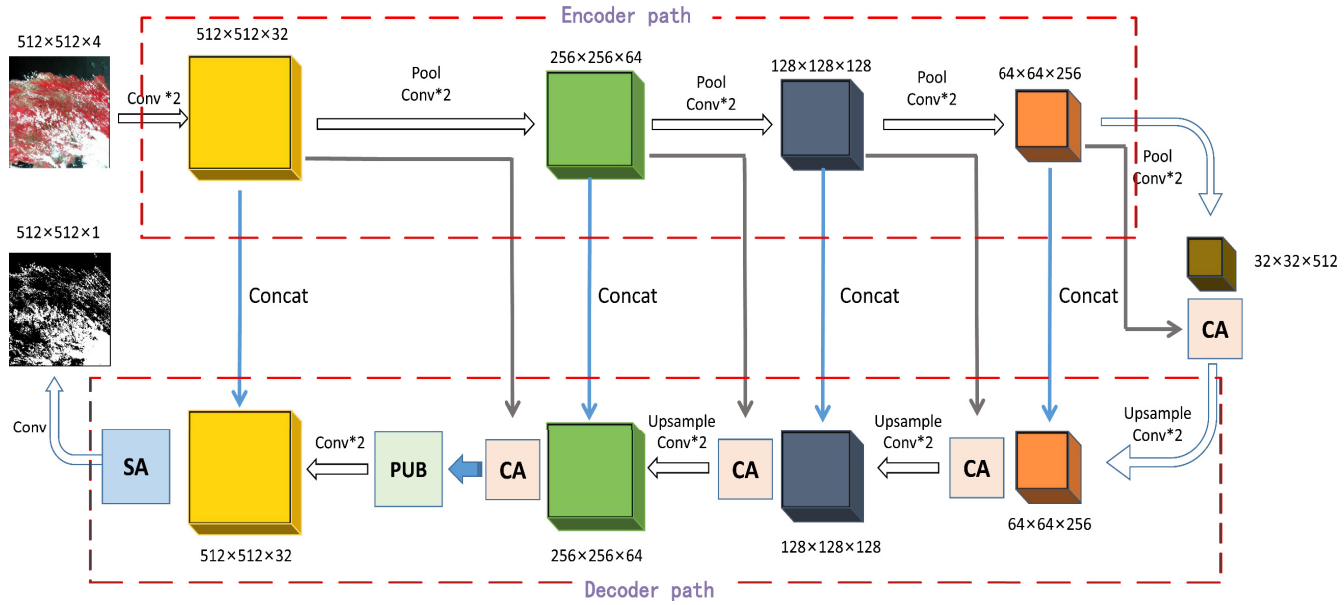


Fig. 1. Graphic model of proposed network structure. The input of the network is multi-spectral image with four spectral segments, and the output is a single channel binary image.

suppress other useless information. This is a means of rapidly screening the high-value information from a large volume of information by using limited attention resources. Also, this is a survival mechanism formed in the long-term evolution of human beings. The human visual attention mechanism significantly improves the efficiency and accuracy of visual information processing. Inspired by the process of human visual attention, the visual attention mechanism has been introduced into DL and is widely used in natural language processing, speech recognition, and image processing. In the year 2014, the Google team drew attention toward the RNN model for image classification [40] and achieved impressive performance. In 2017, researchers introduced the attention mechanism into the CNN network [41]. Subsequently, the attention mechanism based on CNN has been widely used. CNN's attention module can be categorized into two parts: the channel attention module and the spatial attention module. Channel attention emphasizes the correlation among the dimensions of the channel, which focuses the network attention on the useful channel information and suppresses the useless channel information. Hu *et al.* [42] proposed a new architectural unit, which has been termed as the squeeze-and-excitation (SE) block. The characteristic response value of each channel can be adjusted adaptively by modeling the relationship between the channels. Compared to SE-Net, the CBAM proposed by Woo *et al.* [43] not only adds spatial attention but also introduces the parallel structure of maximum pooling and average pooling in channel attention, and its effectiveness is verified by experiments. Spatial attention focuses on the region of interest in the spatial domain. In recent years, Wang *et al.* [44] proposed the spatial attention structure of non-local, which is widely used in various tasks. This structure improves the expression ability of the network by capturing long-distance dependence and expanding the receptive field to the whole picture.

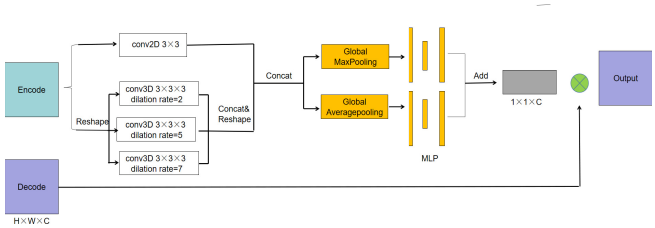


Fig. 2. Block diagram of the CA module.

B. Channel Attention Module

In this article, the proposed channel attention module includes two parts: a multi-scale sampling module and a spatial compression module, as shown in Fig. 2.

In order to increase the expression ability of the module, we choose to perform a multi-scale parallel convolution operation on the input before entering the spatial compression module. Szegedy *et al.* [45] indicate that width is a key factor in improving the performance of the model. More filter parallel structures with different dimensions can obtain features of different sizes of receptive fields, and the fusion of these features can improve the network performance. In the research of multi-scale modules, the prominent focus is put on the size and number of filters. The final structure has been chosen as a four-channel parallel structure, and 2-D convolution with the size of (3, 3) and 3-D dilated convolution with a three-way expansion rate of {2, 5, 7} have been used. Dilated convolution [46]–[48] is widely used as it can improve the network performance while keeping the parameters intact. Dilated convolution increases the receptive field without increasing the parameters and enables images to establish long-distance information association. Considering the channel attention module, we use the 3-D dilated convolution. Hence, long-distance information can be acquired in the channel dimension.

The fusion of this information can enhance the expression ability of the network. Experiments also verify this concept of using 3-D dilated convolution to get more expressive features.

The space compression module transforms the information in the original image into another space and retains the key information. The key part of the image is enhanced, and the useless information of other parts is suppressed. The information obtained by the multi-scale sampling module is taken as the input of the space compression module. Further, through the parallel structure of global maximum pooling and global average pooling, the spatial information aggregation is carried out, and the context information of $1 \times 1 \times C$ size is generated. The corresponding channel attention graph is generated through two channels of MLP. Eventually, these two channels of information are added to obtain the final channel attention graph.

Moreover, we explore the difference and relationship between the encoder segment and the decoder in U-Net network. The saliency map of low-level features contains many details, the saliency map of high-level features is only a rough result, and some basic areas may be weakened. Zhao and Wu [49] concluded through a series of experiments that deep features usually contain global context information, focus more on salient areas, and ignore some edge information. This will bring disastrous effects in the cloud detection tasks because edge regions are often thin clouds. The shallow feature contains more spatial information, so it also contains the thin cloud region. The proposed attention module uses the shallow feature-coding segment, generates the final weight vector through the multi-scale sampling module and spatial compression module, and then uses the decoder to multiply the weight vector to detect the final output feature. The features of the thin cloud region are enhanced to a certain extent by adding the elements with the original features of the decoder, while the features of the thick cloud area are not weakened.

C. Spatial Attention Module

The visual attention model of the human brain has been simulated in the spatial attention model of DL. In recent years, it has been widely used by scientific workers, especially in the field of DL [50], [51]. In order to extract the key information, the spatial attention module makes a corresponding spatial transformation of the spatial domain information in the picture, generates the mask of the space, score, and finally multiplies the elements with the original image to seek the desired result. The original image is usually fit by the compression channel and simple convolution. Through analysis, it has been found that for image segmentation or classification tasks, the computer is not as sensitive as the human vision to identify the category of a certain area. If the focus is only on the local area, the computer cannot complete the task of segmentation and classification. Usually, the convolution operation is limited by the size of the convolution kernel and can only fit the local information. For cloud feature extraction, long-distance information is often equally crucial because in the final analysis, cloud feature is only a classification problem for cloud and non-cloud regions, and most of the cloud pixels are identical. The effective use of the similarity between cloud pixels is the focus of this research work.

In order to make effective use of the long-distance information, one method is to enlarge the convolution kernel as much as possible, even to the whole image, or to expand the receptive field by accumulating the empty convolutions in the network, which will enlarge the receptive field and acquire the wider information distribution, However, such a continuous superposition method will substantially increase the amount of computation, and the deepening of the network will make it difficult for the network to converge, which boosts the difficulty of optimization. In 2011, Buades *et al.* [52] proposed a spatial filtering method non-local mean denoising. In this method, the pixels in the image are not considered to exist in isolation. There must be some association between the pixels of one point and the pixels of other places, which can be considered as gray correlation and geometric structure similarity. Meanwhile, it was also found that similar pixels are not limited to a certain local area, such as the long edge, structure, and texture in the image. Natural images contain abundant redundant information, so image blocks that can describe the structural features of images have to be used to seek similar blocks in the whole image. The basic idea of non-local mean denoising is that the gray value of the current pixel is obtained by a weighted average of all pixels in the image with a similar structure.

Inspired by the idea of non-local mean denoising, Wang *et al.* [44] introduced the idea of non-local mean denoising into DL, the maximum non-local information sharing is achieved by expanding the receptive field area to the size of the whole image. As an end-to-end module, non-local neural network modules can be added to any CNN network, and it will substantially improve the overall performance of the network. However, this module has a fatal disadvantage in that the matrix multiplication leads to a huge number of parameters leading to a high hardware requirement. Cao *et al.* [53] found that in the original non-local structure for each query location, the important areas are basically the same area, that is, the attention of each location is almost the same. Hence, by adding the characteristics of these important areas to each location, the accuracy of the network does not decline, but the amount of computation is reduced significantly. Based on this, the way of non-local neural network module has been modified to obtain the context information, and convolution has been done instead of a large number of matrix multiplication operations, In this way, the number of parameters is prominently reduced, and considering that the features learned by the non-local networks are location independent, the information of the whole graph has been integrated into one point. In the final structure, a 1×1 convolution has been used to fit the information.

Based on past experience, the convolutional neural network is very significant for visual tasks to represent features on multiple scales. While exploring the stronger expression ability of the network, it has been noticed that the better feature extraction ability can be obtained by grouping the features and fusing the results layer by layer. Hence, we proposed a multi-level convolution fusion-block (MFC-block) (Fig. 4).

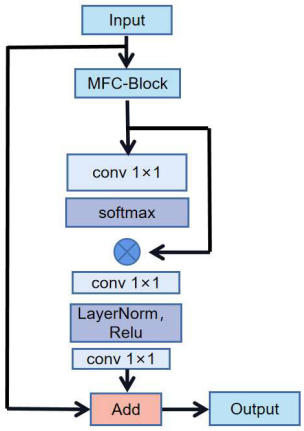


Fig. 3. Block diagram of the SA module, where MFC-block is multi-level convolution fusion-block.

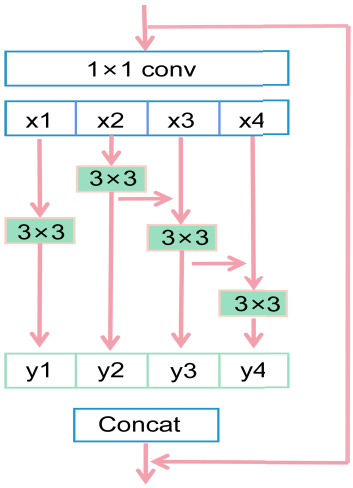


Fig. 4. MFC-block.

A packet convolution module is added to this spatial attention module to represent the multi-scale features in a more fine-grained way and increase the receptive field of each network layer, firstly, the input feature passes through a 1 $\times$ 1 convolution operation, then the features are divided into several groups based on the channel dimension, and each group is fused with the features of the previous group and then fit by a 3 $\times$ 3 convolution. The result of each group is spliced and then taken as output, at the same time, referring to the Res-Net structure, the input feature and output feature are added to obtain the final output.

The spatial attention module proposed here is demonstrated in Fig. 3. The module combines the multi-level convolution fusion module with the improved non-local neural network module. On the one hand, it boosts the relevance of each local region through grouping convolution, and on the other hand, it cross-fuses the long-distance information through the non-local idea. The module is designed as a residual structure and added to the last layer feature of the decoder.

D. Probabilistic Upsampling Block

In the image segmentation task, the boundary details play a significant role. If the spatial information of some key positions is ignored, the segmentation task turns out to be inefficient; hence, Seg-Net proposed by Kendall and Cipolla [54] stores the maxpooling index in the encoder feature map before downsampling. This suggests that the position of the maximum feature value of each pool window is used as the feature map from each encoder to decoder. This structure alleviates the loss of boundary information to a certain extent, and its effectiveness is manifested by experiments. In the cloud detection task, once the spatial location information is lost, the cloud or non-cloud texture will be incoherent; this will significantly affect the detection performance. In order to solve this problem, a PUB module has been designed, which will optimize the problem of common upsampling information loss.

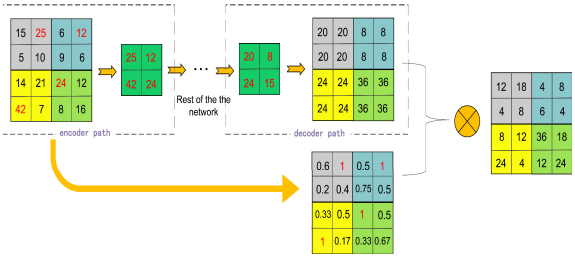


Fig. 5. Block diagram of probabilistic upsampling-block (PUB).

Compared to Seg-Net, the position information of the input image before downsampling has been achieved through this method, and the index of the maximum value is obtained. At the same time, the ratio of the other three positions is set to the maximum index value, that is, the maximum position is set to 1, and the other positions are tuned to the ratio of the maximum value. This information is stored in the weight graph when the decoder performs upsampling; it first performs upsampling, and then multiplies the result of the encoder. This structure not only optimizes the boundary missing issue but also maintains the continuity of space to a certain extent. The experimental results prove that this structure not only has better detection performance but also has a faster convergence speed; PUB is shown in Fig. 5.

E. Loss Function

While performing the binary classification task, binary cross-entropy [55] can be selected as the loss function. When the output of the network is activated through the sigmoid function, the probability value of the output ranges between 0 and 1; if the value exceeds 0.5, it can be classified as a positive sample, if it is less than 0.5, it is a negative sample. Cloud detection is a binary image segmentation task, positive samples in cloud detection are cloud pixels, and the negative samples are non-cloud pixels, the binary cross-entropy can be used as the loss function; through the experiment, good detection results can be achieved. But through the analysis, it can be analyzed that the distribution of cloud and non-cloud is not balanced in a scene cloud image, for the thick cloud and non-cloud areas, since these regions are easy to distinguish for the network when the prediction value of an area is close to 1, it is surely a cloud area. On the contrary, when the prediction value of an area is close to 0, it is certainly a non-cloud area.

However, if it is the boundary between cloud and non-cloud or the thin cloud area, the prediction value is usually close to 0.5, which makes it difficult for the network to identify, to which area, the point belongs to. This makes the cloud detection task strenuous. Often, whether the detection of these regions is correct or not is the key to judge the performance of an algorithm, the use of the binary cross-entropy loss function will make the iterative process of the loss function slow and unable to converge to the optimal in a large number of simple samples. In view of this, the focus of the presented research is to identify the means of improving the binary cross-entropy loss function to make the loss function more suitable for cloud detection tasks, the binary cross-entropy loss function can be computed as follows:

$$L = -y\log y' - (1-y)\log\left(1-y'\right). \tag{1}$$

In the above formula, y' is the output through the activation function, and the size is between 0 and 1.We notice that in the Focal loss function [56], add a constraint term to the binary cross-entropy loss function to achieve better detection performance, such as adding a balance factor to balance positive and negative samples, and adding a constraint term $(1-y')$ to make the network pay more attention to difficult and misclassified samples. Compared with thick clouds and other ground objects, thin clouds are easy to be misclassified, so if the relevant constraints are added to the loss function, the network will pay more attention to the detection of thin cloud areas. In the cloud detection task, because the number of cloud pixels and non-cloud pixels in a scene is uncertain, the balance factor is not necessary for us. To avoid a lot of meaningless optimization in simple samples, relevant constraints are added to the binary cross-entropy function to improve the loss function. We use focal loss with only constraints. The formula is as follows:

$$L_b = -y\left(1-y'\right)\log y' - (1-y)y'\log\left(1-y'\right). \tag{2}$$

When the positive samples (cloud pixels), when they are misclassified due to their dimension, the modulation factor $(1-y')$ is close to 1, and the loss will not be greatly affected. When they are correctly classified, it is close to 1, so the modulation factor is a value close to 0. The same is true for the negative samples. For the samples with correct classification and the samples with the wrong classification, the loss will be reduced, but the reduction degree of the samples with correct classification is greater than that of the samples with false classification. Hence, more attention will be transferred to the samples with false classification, and the proportion of the samples with better classification will be reduced. For the cloud detection task, for example, a certain region is very similar to the underlying surface of the cloud, which will lead the cloud area to be assigned most likely to the negative samples. Therefore, more attention is paid to these areas, which makes the presented network more robust.

IV. Dataset and Experimental Environment

A. Dataset

In 2013, China launched the first high-resolution earth observation satellite called GF-1 WFV, which was equipped with two 2 m/8 m panchromatic cameras and four 16 m multispectral cameras. Gaofen-1 satellite breaks through the optical remote sensing technology of high spatial resolution, multispectral and high temporal resolution, multi-payload image mosaic and fusion technology, high precision, and high stability attitude control technology. The wide field of view camera consists of four bands of visible light and near infrared, namely R, G, B, and NIR. The spatial resolution is 16 m, and the observation range is 800 km. Owing to high precision and wide observation range, it has been applied in many fields, such as environmental disaster reduction, ocean, security, and remote sensing. It is a challenging task to do any research activity on cloud detection algorithms based on high score data. Because there are only four bands of information, and there are no bands such as thermal infrared bands, which are supreme in the cloud detection task, it is a challenging and meaningful research work to use the limited spectral information to better segment the cloud and underlying surface.

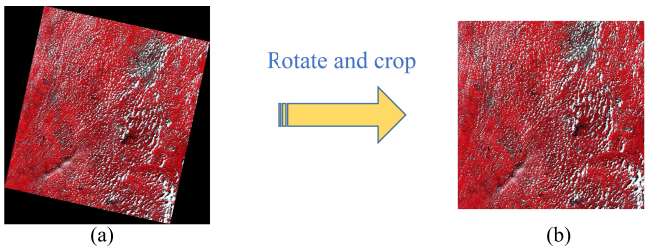


Fig. 6. (b) Resultant image that was obtained by processing (a) original image.

The dataset that is used in the experiment is the open-access GF-1 WFV imagery [10]. This set of data comprises 108 data collected from various locations around the world. For scenes with different cloud cover distributions, all data have corresponding cloud masks. A variety of geomorphic environments, including urban areas, barren areas, areas covered by snow, areas covered by a lot of vegetation, and oceans or lakes, are covered in this dataset. The resolution of the image is 16 m, covering the visible and near-infrared bands. The image size is approximately 17 000 × 16 000 × 4. Among the 108 scenes, 86 were selected as training data. The rest are test data. As shown in Fig .6, to remove the black area around each scene, all the images are rotated and cut to 11 264 × 11 264, and each scene is cut to 512 × 512. In this way, 52 272 × 512 × 512 images are obtained for training and testing including 41 624 images for training and 10 648 images for testing the size of 11 264 × 11 264 × 4. Finally, the pixel values of these images are divided by 1023 to normalize between 0 and 1.

As shown in Fig. 7, it can be seen that the ground features are quite different, and the colors of various ground objects are very different, especially the snow and water images, which are very similar to clouds.

B. Experimental Environment

In this article, all the experiments are programmed and implemented with Keras framework on Ubuntu 18.04 and trained with NVIDIA RTX 2080 Ti GPU. The training batch

TABLE I
EVALUATION RESULTS WITH U-NET MODELS AND PU-NET MODELS ON THE TEST DATASET

Class	OA(%)	FAR(%)	Pre(%)	Recall(%)	Kappa(%)	F1(%)
U-Net	96.73	4.46	89.86	89.66	82.31	89.74
PU-Net	96.89	3.60	92.32	91.24	86.46	91.28

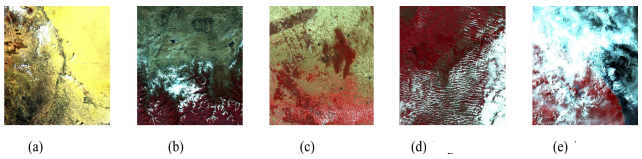


Fig. 7. NIR-R-G images of five experimental data, including (a) barren, (b) snow, (c) urban, (d) vegetation, and (e) water.

size is set to 4 and the maximum training epoch to 50. Adam optimizer is used to optimize [57], and the learning rate is set to 10^{-6}.

V. EXPERIMENTS

A. Evaluation Metrics

The overall accuracy (OA) and false alarm rate (FAR) [30], [38], [60] are chosen as the main experimental verification indicators. In addition, we also selected precision, Recall, Kappa, and F1-Score as auxiliary indicators. In the field of DL, OA refers to the ratio of the number of samples correctly classified by the classifier to the total number of samples in a given test dataset. FAR indicates the ratio of the negative samples misclassified by the classifier to the total number of all the negative samples. Recall represents the ratio between the correct number of detected cloud pixels to the actual number of cloud pixels in the ground truth. Kappa is a coefficient used to evaluate the consistency in image segmentation. The higher the value of Kappa, the better the model performance of the network. F1 (F1-Score) is a measure of classification problem, which is the harmonic average of Precision and Recall. Generally, the higher the score of F1, the better the quality of the model [58].

The above-mentioned metrics are defined, as follows:

$$\mathrm{OA} = \frac{\mathrm{TP}+\mathrm{TN}}{\mathrm{TP}+\mathrm{TN}+\mathrm{FP}+\mathrm{FN}} \tag{3}$$

$$\mathrm{FAR} = \frac{\mathrm{FP}}{\mathrm{TN}+\mathrm{FP}} \tag{4}$$

$$\mathrm{Precision} = \frac{\mathrm{TP}}{\mathrm{TP}+\mathrm{FP}} \tag{5}$$

$$\mathrm{Recall} = \frac{\mathrm{TP}}{\mathrm{TP}+\mathrm{FN}} \tag{6}$$

$$\mathrm{Kappa} = \frac{P_a - P_e}{1 - P_e} \tag{7}$$

$$P_a = \frac{\mathrm{TP}+\mathrm{TN}}{\mathrm{TP}+\mathrm{TN}+\mathrm{FP}+\mathrm{FN}} \tag{8}$$

$$P_e = \frac{P(\mathrm{TP}+\mathrm{FP}) + N(\mathrm{FN}+\mathrm{TN})}{(P+N)^2} \tag{9}$$

$$F1 = \frac{2 \times \mathrm{precision} \times \mathrm{Recall}}{\mathrm{Precision} + \mathrm{Recall}} \tag{10}$$

where TP indicates the true positive outcomes, i.e., the number of cloud pixels that are correctly identified as cloud pixels in the generated mask; TN represents the true negative outcomes, i.e., the number of non-cloud pixels correctly identified as the non-cloud pixels in the generated mask; FP indicates the false-positive outcomes, i.e., the number of non-cloud pixels wrongly identified as cloud pixels in the generated mask; while FN is the false-negative outcomes, i.e., the number of cloud pixels falsely identified as non-cloud pixels in the generated mask. P denotes the number of cloud pixels in the ground truth and N denotes the number of non-cloud pixels in the ground truth.

When calculating far, if all the pixels in the result are non-cloud pixels, TN and FP are both 0, then the denominator of far is 0. We add an infinitely small number $\varepsilon = e^{-10}$ to avoid the situation where the denominator is 0.

B. Evaluation of PUB

The purpose of PUB is to ensure the continuity of spatial information while restoring the image at the decoder and to optimize the problem of boundary loss caused by the upsampling. The spatial attention and channel attention modules are removed, the network is identified as PU-Net. Compared to U-Net, PU-Net only adds the PUB, therefore, the performance of PU-Net is compared with U-Net. In the process of training the network, it is observed that PU-Net not only has better detection results than U-Net but also has a faster convergence speed. The test results are mentioned in Table I. From the above test results, it can be analyzed that PU-Net has achieved higher OA, and far has also decreased by 0.86%.

As highlighted in Fig. 8, it is a test picture of water classification. The results obtained by the U-Net algorithm demonstrate obvious discontinuities in the texture, while the results of PU-Net do not have this situation, the subjective and objective results are combined, the proposed PUB thus seems effective. As shown in Fig. 9, the convergence speed of PU-Net in the first few epochs is significantly higher than that of U-Net, and the final OA is also higher than that of U-Net, so we can draw a conclusion, PUB can effectively improve the convergence speed and network performance.

C. Evaluation of CA Module

In order to verify the CA module, a few random experiments have been carried out. Firstly, for the demonstration of the multi-scale sampling module structure of the CA module, the comparative experiments of 2-D dilated convolution and 3-D dilated convolution are performed. The CA module is then

TABLE II
EVALUATION RESULTS WITH CA MODELS USING 3-D CONVOLUTION AND 2-D CONVOLUTION ON THE TEST DATASET

Class	OA(%)	FAR(%)	Pre(%)	Recall(%)	Kappa(%)	F1(%)
CA-2D	97.05	4.41	90.41	90.12	83.49	90.26
CA-3D	97.16	3.41	91.52	90.55	85.88	91.03

TABLE III
EVALUATION RESULTS WITH U-NET MODELS AND U-NET-SA ON THE TEST DATASET

Class	OA(%)	FAR(%)	Pre(%)	Recall(%)	Kappa(%)	F1(%)
U-Net	96.73	4.46	89.86	89.66	82.31	89.74
U-Net add SA	97.15	3.59	92.02	90.85	86.32	91.43

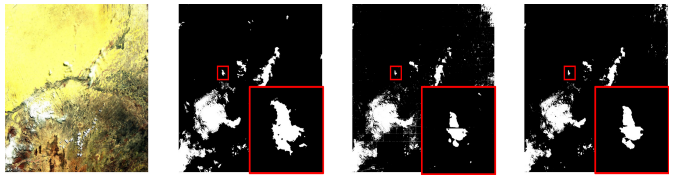


Fig. 8. (a) NIR-R-G image, (b) ground truth, and masks generated by (c) U-Net and (d) PU-Net.

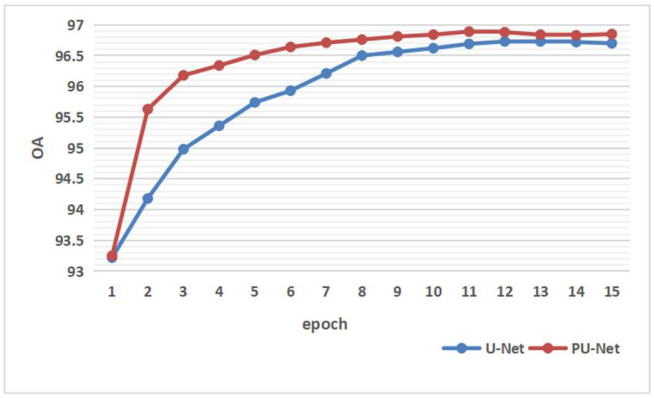


Fig. 9. Comparison of the convergence speed of OA between U-Net and PU-Net.

added to the deepest layer of the U-Net network. Through many experiments, it has been concluded that one convolution in parallel and three convolution groups with expansion rates {2, 5, 7} have the best detection results. Therefore, 2-D and 3-D convolutions are chosen for the three-way expansion convolution. From the table, it can be noticed that the accuracy and FAR of the structure with 3-D dilated convolution are optimized compared to that of 2-D dilated convolution. 3-D dilated convolution can not only fit the information in spatial dimension but also operate in channel dimension. Therefore, 3-D dilated convolution is believed to be better than 2-D dilated convolution, and the detection results also confirm this conjecture, as listed in Table II.

To further verify the effectiveness of the CA module, experiments are conducted by controlling the number of CA modules. Since U-Net has a total of five feature layers,

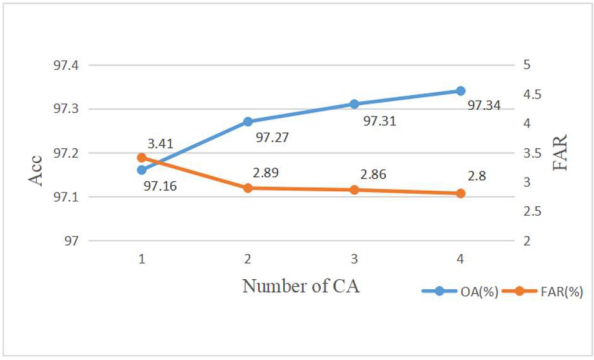


Fig. 10. Influence of the number of CA modules on the test results.

as demonstrated in the figure, from the deepest layer of the network up, add 1, 2, 3, and 4 CA module units, respectively. Through the detection results of 22 scenes, it can be analyzed that more CA modules can improve the performance of the network. As demonstrated in Fig. 10, while OA increased, FAR also decreased.

D. Evaluation of SA Module

For the SA module, we choose to add the SA module to the last layer of the U-Net network because we added the CA module in the deepest four layers of the U-Net network. To avoid the mutual influence caused by the mixing of the CA module and the SA module in the final network, the only discussion is done for adding SA to the last layer. At the same time, another reason is that the proposed SA is doing information fusion in a spatial dimension, and the last layer is the largest compared to other layers. It will make this SA module play a maximum role, making it more effective. The SA module added by U-Net is later compared with the original U-Net to verify the effectiveness of the SA module. The results are mentioned in Table III. As shown in Fig. 11, the network with SA shows better detection performance for snow-covered areas.

E. Evaluation of Proposed Method

The final network is based on U-Net, adding CA module, SA module, and PUB module. The CA module strengthens

TABLE IV

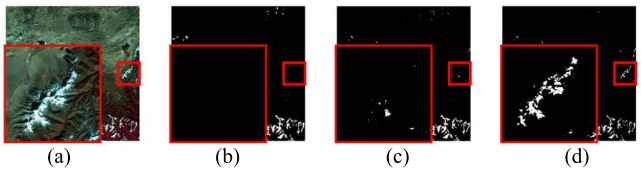
EVALUATION RESULTS WITH PROPOSED MODELS USING BINARY CROSS-ENTROPY AND OUR LOSS FUNCTION ON THE TEST DATASET

Class	OA(%)	FAR(%)	Pre(%)	Recall(%)	Kappa(%)	F1(%)
Binary cross-entropy	97.42	3.38	94.14	91.33	86.28	92.72
Focal loss	97.45	2.65	94.45	91.24	87.00	92.82

TABLE V

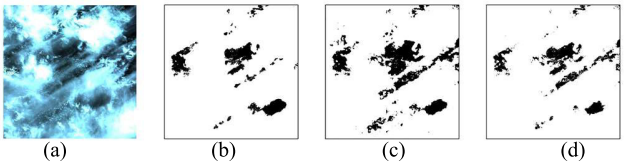
EVALUATION RESULTS WITH DEEPLABV3+, RS-NET, NGAD, AND OUR METHOD MODELS ON THE TEST DATASET

Class	method	RS-Net	Deeplabv3+	NGAD	Proposed method
Barren	OA(%)	97.74	97.83	**98.42**	98.24
	FAR(%)	1.71	0.88	**0.12**	0.20
	Precision(%)	95.21	86.94	94.54	**95.50**
	Recall(%)	72.83	76.61	**79.14**	79.00
	Kappa(%)	80.46	80.04	**86.55**	85.41
	F1(%)	81.60	81.22	84.69	**86.47**
Vegetation	OA(%)	98.21	97.54	98.54	**98.56**
	FAR(%)	**0.47**	1.16	0.87	0.83
	Precision(%)	95.22	91.50	94.47	**95.04**
	Recall(%)	90.49	87.85	**94.19**	93.97
	Kappa(%)	91.42	87.31	**93.34**	90.56
	F1(%)	92.60	88.93	94.33	**94.50**
Snow	OA(%)	94.85	94.32	96.21	**96.72**
	FAR(%)	4.33	4.60	**2.63**	5.80
	Precision(%)	87.55	87.50	**93.88**	91.69
	Recall(%)	87.43	84.00	86.20	**88.16**
	Kappa(%)	78.34	75.77	83.03	**83.67**
	F1(%)	87.42	85.70	**89.88**	**89.89**
Water	OA(%)	91.64	91.07	93.38	**93.40**
	FAR(%)	18.10	11.34	**8.25**	8.86
	Precision(%)	95.76	**98.02**	97.01	95.53
	Recall(%)	90.63	87.39	92.02	**93.22**
	Kappa(%)	69.69	71.67	78.01	**78.35**
	F1(%)	92.70	92.03	**94.45**	94.36
Urban	OA(%)	99.53	99.30	99.56	**99.58**
	FAR(%)	0.31	0.47	0.31	**0.29**
	Precision(%)	89.10	82.98	**88.59**	87.73
	Recall(%)	92.36	90.27	91.56	**93.48**
	Kappa(%)	90.44	86.09	**91.25**	90.24
	F1(%)	90.70	86.47	90.05	**90.51**
All	OA(%)	96.71	96.18	97.42	**97.45**
	FAR(%)	3.97	3.25	**2.22**	2.65
	Precision(%)	94.34	91.31	94.39	**94.45**
	Recall(%)	87.92	85.99	90.57	**91.24**
	Kappa(%)	84.74	82.37	**88.12**	87.00
	F1(%)	91.02	88.57	92.44	**92.82**



(a) (b) (c) (d)

Fig. 11. (a) NIR-R-G image, (b) ground truth, and masks generated by (c) U-Net add SA and (d) U-Net.



(a) (b) (c) (d)

Fig. 12. (a) NIR-R-G image, (b) ground truth, and masks the loss function uses generated by (c) binary cross-entropy and (d) our loss function.

the fusion of channel dimension information, the SA module strengthens the fusion of the spatial dimension information, and the PUB module guides the acquisition module of the code terminal by supervising the location information of the acquisition module. Also, we compare focal loss with binary cross-entropy loss function. Focal loss focuses more attention on the samples that are difficult to classify, through the detection results as Table IV, it can be observed that the focal loss function has better detection performance.

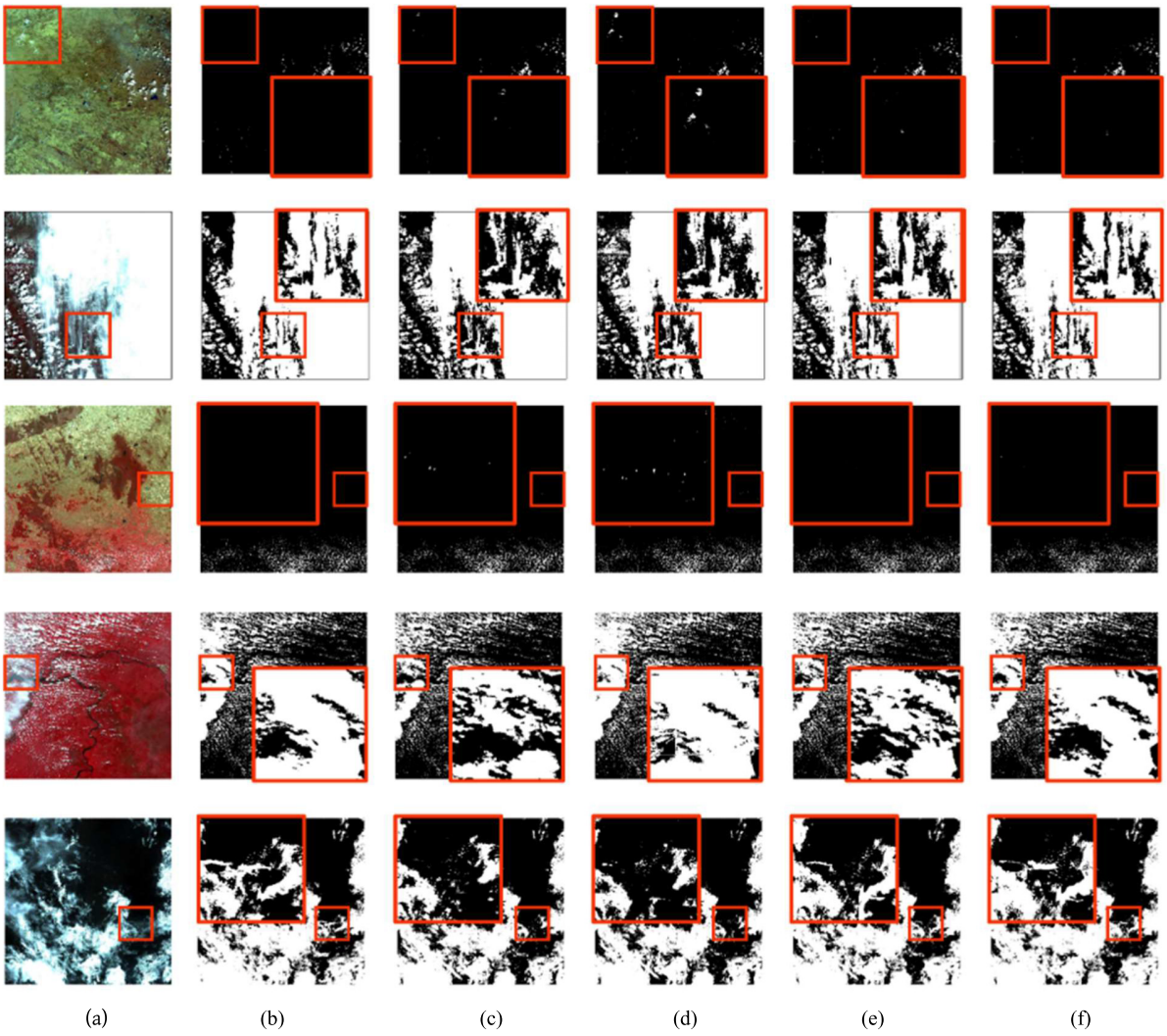


Fig. 13. (a) NIR-R-G image, (b) ground truth, and masks generated by (c) RS-Net, (d) DeeplabV3+, (e) NGAD, and (f) our proposed method.

As shown in Fig. 12, from the ground truth lesson, we can see that most areas of the test map are covered by clouds, and many thin cloud areas can be seen from the corresponding NIR-R-G image. Compared with the binary cross-entropy loss function, the focal loss function shows less missed detection.

To further verify the performance of the algorithm, the algorithm with some existing image segmentation networks, including DeeplabV3+ [19], RS-Net [38], and NGAD [59] are compared. To ensure the fairness and accuracy of the experiment, the code used in the experiment is given by the algorithm author, and the same data is used for training and testing. The final results can be analyzed from the subjective image and objective data that the proposed algorithm has many advantages over other algorithms in cloud detection tasks.

As demonstrated in Fig. 13, the first row and the third row are barren and urban classified images, respectively. In the upper left corner of the image, some ground objects similar to clouds are shown. The deeplabV3+ and RS-Net algorithms identify these areas as clouds, and the proposed algorithm correctly detects these non-cloud areas. Therefore, it can be considered that our algorithm has better detection ability in local high reflection areas.

The second row and the fourth row are snow and vegetation classification images, as shown in the red frame selection area. From the JPEG image, it can be seen that there are a lot of thin clouds in these areas. DeeplabV3+, Rs-Net, and NGAD are the three contrast algorithms having the phenomenon of missing thin clouds. Since attention is paid to these difficult areas, the presented algorithm shows a better detection effect, when the texture is more similar to the ground truth.

The fifth row is the water classification image, as shown in the red frame area. Since it is difficult to distinguish water from the cloud under the illumination of light, the low detection performance of the algorithm is observed in these areas; nevertheless, the proposed method is better than the other three methods on the whole.

Table V lists the average values of objective indexes of each category of 22 test images of several algorithms, including

TABLE VI
COMPARISON OF DETECTION PARAM SIZE AND FLOPS OF DIFFERENT ALGORITHMS

Model	Param Size	Flops
RS-Net	29.93 MB	15.69 MFLOPs
DeeplabV3+	159.10 MB	83.90 MFLOPs
NGAD	58.48 MB	30.65 MFLOPs
Proposed method	42.98 MB	22.52 MFLOPs

DeeplabV3+, RS-Net, and NGAD. Among them, NGAD is the algorithm of other students in the research group; it can be observed from the table that the FAR of NGAD and the presented method is lower than that of the other algorithms. This proves that in the face of cloud-like areas, these two algorithms divide them correctly into non-cloud areas. From the perspective of OA, the proposed algorithm has greater improvements than other algorithms. On the whole, it has a similar performance to the NGAD, the algorithm proposed in this article has higher OA than NGAD, but also higher FAR than NGAD. In general, our algorithm has greater advantages compared with RS-Net and DeeplabV3+. Compared with NGAD, although FAR performance is not good, other indicators are improved compared with NGAD. F1-Score, in particular, reflects the quality of a network. We have achieved the best results in this area, which also reflects the effectiveness of our work at the same time. NGAD uses complex Gabor features in the network, which greatly increases the network computation, while our network is relatively light and easier to deploy.

Table VI presents a comparison of the parameters of different algorithms. By comparing the parameters of several algorithms, we can see that RS-Net has the smallest parameters, but because of the simple model, the detection performance is not enough. While DeeplabV3+ has the largest parameters, and the detection performance is not high, and the detection performance of NGAD is similar to that of our algorithm, but its parameters are 36% more than our model. We can conclude that our algorithm achieves the best detection performance, also well controls the model parameters; therefore, we can think that our algorithm is advanced compared with other cloud detection algorithms.

After analysis, we can find that our FAR is slightly higher than NGAD because we pay more attention to the thin cloud area. Although our OA and Precision detection results have achieved better results, in the edge of thin cloud region, the probability of non-cloud region recognition for cloud region is greatly improved, how to better identify the edge of cloud. This is what we need to further improve in our future work.

In summary, from a subjective and objective point of view, the proposed algorithm exhibits excellent cloud detection performance. This justifies the meaningfulness and effectiveness of this algorithm, at the same time; it also has the characteristics of lightweight, which makes our network have fewer limitations in application and better deployable.

VI. CONCLUSION

With the enhancement of DL theory systems, more and more researchers choose to use the convolutional neural network based on DL for cloud detection and other related research. Nevertheless, when the convolutional neural network is used to get effective cloud information, a large number of redundant information will be fed into the network at the same time, which will lead to subsequent false classification. Cloud detection, a special segmentation task, is very sensitive to the distribution of texture, once an area is classified incorrectly, it will lead to the final result image texture disorder. There are also some common problems: the binary cross-entropy loss function cannot take into account the regions that are more difficult to classify, resulting in low detection performance. In view of these issues, the proposed attention module automatically adjusts the weight of the region to retain more useful information and suppress the useless information. In order to solve the problem that the convolutional neural network does not take into account the boundary texture information, the PUB module has been proposed. The upsampling module has been thus optimized; finally, the binary cross-entropy loss function has been optimized to pay more attention to the critical regions such as thin clouds. Experiments prove the effectiveness of this algorithm, as the OA is 97.45% and far is 2.65%. Compared to other algorithms, the proposed algorithm achieves better detection results.

REFERENCES

[1] C. J. Van Westen, "Remote sensing and GIS for natural hazards assessment and disaster risk management," *Treatise Geomorphol.*, vol. 3, pp. 259–298, Mar. 2013.
[2] H. Adab, K. D. Kanniah, and K. Solaimani, "Modeling forest fire risk in the northeast of Iran using remote sensing and GIS techniques," *Natural Hazards*, vol. 65, no. 3, pp. 1723–1743, Feb. 2013.
[3] T. Shi, Q. Xu, Z. Zou, and Z. Shi, "Automatic raft labeling for remote sensing images via dual-scale homogeneous convolutional neural network," *Remote Sens.*, vol. 10, no. 7, p. 1130, Jul. 2018.
[4] Y. Shi, Z. Qi, X. Liu, N. Niu, and H. Zhang, "Urban land use and land cover classification using multisource remote sensing images and social media data," *Remote Sens.*, vol. 11, no. 22, p. 2719, Nov. 2019.
[5] J. Li, M. Khodadadzadeh, A. Plaza, X. Jia, and J. M. Bioucas-Dias, "A discontinuity preserving relaxation scheme for spectral–spatial hyperspectral image classification," *IEEE J. Sel. Topics Appl. Earth Observ. Remote Sens.*, vol. 9, no. 2, pp. 625–639, Feb. 2016.
[6] Y. Gu, T. Liu, X. Jia, and J. A. Benediktsson, "Nonlinear multiple kernel learning with multiple-structure-element extended morphological profiles for hyperspectral image classification," *IEEE Trans. Geosci. Remote Sens.*, vol. 54, no. 6, pp. 3235–3247, Jun. 2016.
[7] P. Ghamisi, M. D. Mura, and J. A. Benediktsson, "A survey on spectral–spatial classification techniques based on attribute profiles," *IEEE Trans. Geosci. Remote Sens.*, vol. 53, no. 5, pp. 2335–2353, May 2015.
[8] P. Ghamisi, J. A. Benediktsson, and J. R. Sveinsson, "Automatic spectral–spatial classification framework based on attribute profiles and supervised feature extraction," *IEEE Trans. Geosci. Remote Sens.*, vol. 52, no. 9, pp. 5771–5782, Sep. 2014.
[9] 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. Meteorological Soc. Jpn.*, vol. 82, no. 2, pp. 761–773, 2004.
[10] 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.
[11] 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.

[12] C. Li, Z. Lin, and X. Deng, "An efficient cloud detection method for high resolution remote sensing panchromatic imagery," in *Proc. SPIE*, vol. 10615, Apr. 2018, Art. no. 106154V.

[13] C. Zhenwei, Z. Guo, N. Jinsheng, and T. Xinming, "An automatic cloud detection method for ZY-3 Satellite," *Acta Geodaetica et Cartographica Sinica*, vol. 44, no. 3, pp. 292–300, 2015.

[14] J. Liu *et al.*, "Thin cloud removal from single satellite images," *Opt. Exp.*, vol. 22, no. 1, p. 618, Jan. 2014.

[15] J. Key and R. G. Barry, "Adaptation of the ISCCP cloud detection algorithm to combined AVHRR and SMMR arctic data," in *Proc. 12th Can. Symp. Remote Sens. Geosci. Remote Sens. Symp.*, 1989, pp. 188–191, doi: 10.1109/IGARSS.1989.567199.

[16] C. Shi, "Cloud classification for NOAA-AVHRR data by using a neural network," *Acta Meteorologica Sinica*, Feb. 2002, pp. 250–256.

[17] S. Ren, K. He, R. Girshick, and J. Sun, "Faster R-CNN: Towards real-time object detection with region proposal networks," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 39, no. 6, pp. 1137–1149, Jun. 2017.

[18] 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. 2018.

[19] 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.

[20] J. Ding, B. Chen, H. Liu, and M. Huang, "Convolutional neural network with data augmentation for SAR target recognition," *IEEE Geosci. Remote Sens. Lett.*, vol. 13, no. 3, pp. 364–368, Mar. 2016.

[21] K. He, X. Zhang, S. Ren, and J. Sun, "Spatial pyramid pooling in deep convolutional networks for visual recognition," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 37, no. 9, pp. 1904–1916, Sep. 2015.

[22] R. Girshick, J. Donahue, T. Darrell, and J. Malik, "Region-based convolutional networks for accurate object detection and segmentation," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 38, no. 1, pp. 142–158, Jan. 2016.

[23] K. Fragkiadaki, S. Levine, P. Felsen, and J. Malik, "Recurrent network models for human dynamics," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, Dec. 2015, pp. 4346–4354.

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

[25] J. Zhang, W. Li, and Y. Li, "SAR image despeckling using multiconnection network incorporating wavelet features," *IEEE Geosci. Remote Sens. Lett.*, vol. 17, no. 8, pp. 1363–1367, Oct. 2019.

[26] J. Redmon, S. Divvala, R. Girshick, and A. Farhadi, "You only look once: Unified, real-time object detection," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2016, pp. 779–788.

[27] J. Redmon and A. Farhadi, "YOLO9000: Better, faster, stronger," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jul. 2017, pp. 7263–7271.

[28] M. L. Goff, J. Y. Tourneret, H. Wendt, M. Ortner, and M. Spigai, "Deep learning for cloud detection," in *Proc. Int. Conf. Pattern Recognit. Syst.*, 2018, p. 10.

[29] S. Zhenfeng *et al.*, "Cloud detection in remote sensing images based on multiscale features-convolutional neural network," *IEEE Trans. Geosci. Remote Sens.* vol. 57, no. 6, pp. 4062–4076, Jun. 2019.

[30] 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.

[31] Z. Shao, J. Deng, L. Wang, Y. Fan, N. S. Sumari, and Q. Cheng, "Fuzzy AutoEncode based cloud detection for remote sensing imagery," *Remote Sens.*, vol. 9, no. 4, p. 311, 2017.

[32] O. Ronneberger, P. Fischer, and T. Brox, "U-Net: Convolutional networks for biomedical image segmentation," in *Medical Image Computing and Computer-Assisted Intervention (MICCAI)*. Springer, 2015, pp. 234–241.

[33] J. Long, E. Shelhamer, and T. Darrell, "Fully convolutional networks for semantic segmentation," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 39, no. 4, pp. 640–651, Apr. 2017.

[34] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," in *Proc. Int. Conf. Learn. Represent. (ICLR)*, May 2015, pp. 1–14.

[35] T.-Y. Lin, P. Dollar, R. Girshick, K. He, B. Hariharan, and S. Belongie, "Feature pyramid networks for object detection," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jul. 2017, pp. 2117–2125.

[36] Ö. Çiçek, A. Abdulkadir, S. S. Lienkamp, and T. Brox, "3D U-Net: Learning dense volumetric segmentation from sparse annotation," in *Medical Image Computing and Computer-Assisted Intervention (MICCAI)*, vol. 9901, 2016, pp. 424–432.

[37] B. Lim, S. Son, H. Kim, S. Nah, and K. M. Lee, "Enhanced deep residual networks for single image super-resolution," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. Workshops (CVPRW)*, Jul. 2017, pp. 136–144.

[38] 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, Aug. 2019.

[39] 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.

[40] V. Mnih, N. Heess, and A. Graves, "Recurrent models of visual attention," in *Proc. Adv. Neural Inf. Process. Syst.*, vol. 3, 2014, pp. 2204–2212.

[41] A. Vaswani *et al.*, "Attention is all you need," in *Proc. Conf. Neural Inf. Process. Syst.*, vol. 30, 2017, pp. 5998–6008.

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

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

[44] X. Wang, R. Girshick, A. Gupta, and K. He, "Non-local neural networks," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2018, pp. 7794–7803.

[45] C. Szegedy *et al.*, "Going deeper with convolutions," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2015, pp. 1–9.

[46] P. Wang *et al.*, "Understanding convolution for semantic segmentation," 2017, *arXiv:1702.08502*. [Online]. Available: http://arxiv.org/abs/1702.08502

[47] F. Yu, V. Koltun, and T. Funkhouser, "Dilated residual networks," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jul. 2017, pp. 472–480.

[48] S. Watanabe *et al.*, "ESPnet: End-to-end speech processing toolkit," in *Proc. Interspeech*, 2018, pp. 2207–2211.

[49] T. Zhao and X. Wu, "Pyramid feature attention network for saliency detection," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2019, pp. 3085–3094.

[50] L.-C. Chen, Y. Yang, J. Wang, W. Xu, and A. L. Yuille, "Attention to scale: Scale-aware semantic image segmentation," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2016, pp. 3640–3649.

[51] J. Fu *et al.*, "Dual attention network for scene segmentation," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2019, pp. 3146–3154.

[52] A. Buades, B. Coll, and J.-M. Morel, "Non-local means denoising," *Image Process. Line*, vol. 1, pp. 208–212, Sep. 2011.

[53] Y. Cao, J. Xu, S. Lin, F. Wei, and H. Hu, "GCNet: Non-local networks meet squeeze-excitation networks and beyond," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. Workshop (ICCVW)*, Oct. 2019, pp. 1971–1980.

[54] 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.

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

[56] T.-Y. Lin, P. Goyal, R. Girshick, K. He, and P. Dollar, "Focal loss for dense object detection," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 42, no. 2, pp. 318–327, Feb. 2020.

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

[58] M. Luotamo, S. Metsämäki, and A. Klami, "Multi-scale cloud detection in remote sensing images using a dual convolutional neural network," 2020, *arXiv:2006.00836*. [Online]. Available: http://arxiv.org/abs/2006.00836

[59] 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.

[60] 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.

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 a 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.

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

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

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

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

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.

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