IGARSS 2020 - 2020 IEEE International Geoscience and Remote Sensing Symposium | 978-1-7281-6374-1/20/$31.00 ©2020 IEEE | DOI: 10.1109/IGARSS39084.2020.9324486

DEEP NETWORKS UNDER BLOCK-LEVEL SUPERVISION FOR PIXEL-LEVEL CLOUD DETECTION IN MULTI-SPECTRAL SATELLITE IMAGERY

Wei Chen [1], Student Member, IEEE, Yansheng Li [1], Yongjun Zhang [1], and Xiaolong Hao [2]

[1] School of Remote Sensing and Information Engineering, Wuhan University, China
[2] Beijing Tracking and Communication Technology Research Institute, China

ABSTRACT

Cloud cover hinders the usability of optical remote sensing imagery. Existing cloud detection methods either require hand-crafted features or utilize deep networks. Generally, deep networks perform better than hand-crafted features. However, deep networks for cloud detection need massive and expensive pixel-level annotation labels. To alleviate that, this paper proposes a weakly supervised deep learning-based cloud detection method using only block-level labels, with a new global convolutional pooling operation and a local pooling pruning strategy to improve the performance. For evaluating, we collect a training dataset containing over 160,000 image blocks with block-level labels and a testing dataset including ten large image scenes with pixel-level labels. Even under extremely weak supervision, our method performed well with the average overall accuracy reached 97.2%. Experiments demonstrate that our proposed method obviously outperforms the state-of-the-art methods.

***Index Terms*—** Cloud detection, weakly supervised deep learning, global convolutional pooling, local pooling pruning, high-resolution remote sensing imagery

1. INTRODUCTION

Optical remote sensing (RS) imagery often degenerates because of cloud cover. Driven by various applications, cloud detection in the RS imagery attracts extensive research interest. Although numerous methods have been proposed, the off-the-shelf cloud detection methods have limited performance and weak universality. Hence, cloud detection in the RS imagery is still facing challenges.

So far, cloud detection methods are mainly designed for the low or medium resolution RS imagery (e.g., MODIS [1], Landsat [2, 3]). These images generally consist of many spectral bands that benefit improving the accuracy. With high-resolution RS satellites launched, the multi-spectral RS imagery with four spectral bands have become increasingly prevalent. It is more difficult for cloud detection in the high-resolution RS imagery with only four spectral bands [4]. Accordingly, it becomes very urgent to exploit the cloud detection technique for the high-resolution RS imagery.

In the early years, cloud detection methods are mainly based on hand-crafted features, such as spectral, textural, geometrical features [3], man-made filters [5], hand-crafted indexes [4] and so on. Motivated by the tremendous success of deep learning [6], various variants of deep semantic segmentation networks [2,7] have been proposed to address the cloud detection. Although these methods outperform the methods based on hand-crafted features, their superior performance highly depends on the pixel-level cloud masks requiring lots of manual annotation labor. Therefore, it is quite significant to explore advanced deep learning-based method of saving annotation labor.

As is well known, block-level labels are much easier to collect than pixel-level annotations. With the global pooling operations like global average pooling (GAP), researchers [8] have shown that deep networks trained with only block-level labels are informative of object locations. However, because of the inherent defects of global pooling operations [9], there is a lack of the capability of obtaining the detail information of objects, which is quite important for accurately detecting the cloud boundary. The potential of weakly supervised deep learning has not been well exploited.

In this paper, we leverage only block-level supervision to train the deep networks for pixel-level cloud detection. We propose a global pooling operation called global convolutional pooling (GCP) in the training stage which learns channel-wise convolutional weights to enhance the representing ability of the feature map. Furthermore, we propose a local pooling pruning (LPP) strategy in the testing stage during generating the cloud activation map (CAM). By pruning the local pooling layers in the trained deep networks, the spatial resolution of CAM gets much better. After that, the final cloud mask of one RS image can be obtained through natively segmenting the CAM by an adaptively statistical threshold.

In the experiment, we train deep networks under the supervision of RS image blocks with coarse labels, which only indicate whether an image block contains cloud or not, but purse the pixel-level cloud detection. Even under this extreme setting, our proposed method still yields promising re-

This work was supported by the National Natural Science Foundation of China under grants 41971284 and 41601352; the China Postdoctoral Science Foundation under grants 2016M590716 and 2017T100581; the Hubei Provincial Natural Science Foundation of China under grant 2018CFB501. (Corresponding author: Yansheng Li)

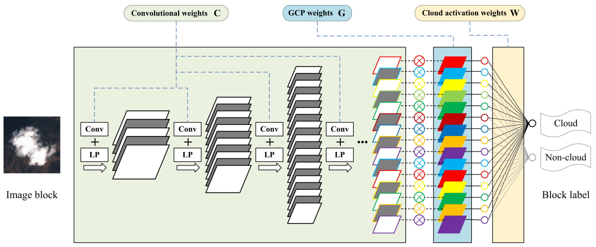

Fig. 1. The architecture of our adopted deep network.

sults, and outperforms the existing methods [8–10]. Considering that there are not any qualified datasets to evaluate the weakly supervised deep learning-based cloud detection (WDCD) method, we collect a dataset based on the GaoFen-1 multi-spectral imagery, which is one kind of typical high-resolution RS imagery.

The collected dataset and our proposed method own good generality. The rest of this paper is organized as follows. Section 2 introduces our proposed WDCD method. Section 3 describes the dataset and reports the experimental results. Finally, Section 4 gives the conclusion of this paper.

2. METHODOLOGY

In this section, we give the details of our proposed WDCD method. Section 2.1 gives the structure of our deep networks and the learning process in the training stage. Furthermore, how to perform the pixel-level cloud detection using the trained deep network will be introduced in Section 2.2.

2.1. Learning deep networks under block-level supervision

With block-level labels, it is easy to build a discriminative deep network (e.g., VGG [10]) only to classify the image blocks as cloud or non-cloud. GAP showed that block-level supervision can be used for object localization [8] but the accuracy needs to be improved. Paper [9] improved the GAP method with utilizing a two-stage-learning (TSL) method while the networks cannot be learned in an end-to-end way. To overcome aforementioned limitation, we proposed the WDCD framework.

As depicted in Fig.1, the architecture of our deep networks is quite similar to the common convolutional neural network (CNN) where the CNN is composed of local convolutional (Conv) operations and local pooling (LP) operations. The difference is that normal CNNs are designed for classification tasks whereas we employ it under block-level supervision to perform the cloud detection. That's why we replaced the GAP or fully connected layer with our proposed GCP layer, in order to promote the representing ability of the feature map. As displayed in Fig.1, the feature map is performed a channel-wise convolution with the GCP layer, by which the spatial variance will be well represented after several iterations of network propagation. Let $\{(b_n, y_n)|n = 1, 2, \cdots, N\}$ denote the training cloud dataset. N is the number of image blocks, b_n stands for the n-th image block, and y_n denotes its binary label. Let $\Psi = \{C, G, W\}$ denote the weights of the deep networks where C stands for the convolutional weights, G denotes the GCP weight, and W stands for the cloud activation weights. For a given image block b_n it is sent to the deep network and outputs the feature map f_n^k as Eq. (1).

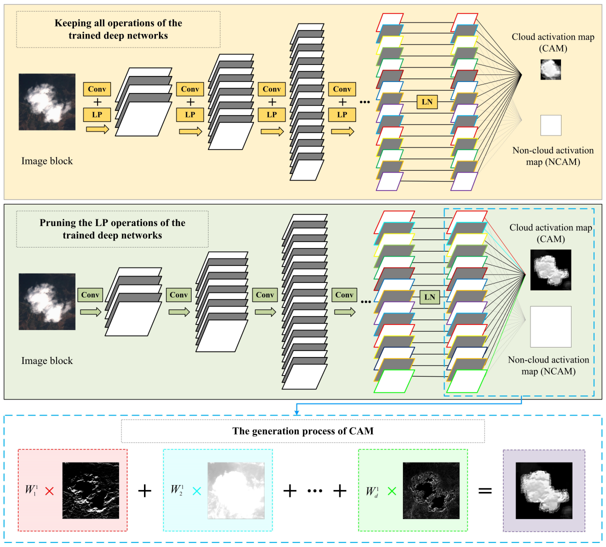
Keeping all operations of the trained deep networks
Image block
Conv
+
LP
LN
Cloud activation map (CAM)
Non-cloud activation map (NCAM)
Pruning the LP operations of the trained deep networks
Image block
Conv
LN
Cloud activation map (CAM)
Non-cloud activation map (NCAM)
The generation process of CAM
W_1^1 ×
+
W_2^1 ×
+ ··· +
W_d^1 ×
=

Fig. 2. The difference between deep networks with and without local pooling pruning (LPP). It is noted that we don't generate NCAM in the figure.

$$f_n^k = \varphi^k(b_n; C) \tag{1}$$

where f_n^k denotes the k-th channel of the last convolutional layer's output feature map, φ denotes the representation of computation in the deep networks. By global convolutional pooling f_n^k per channel, we calculate the activation value of f_n^k at each channel as depicted in Eq. (2).

$$O_n^k = f_n^k \otimes G^k \tag{2}$$

where O_n^k denotes the activation value of f_n^k at the k-th channel, $G^k \in G$ stands for the weights of the GCP layer at the k-th channel, $\otimes$ denotes the channel-wise convolution.

Softmax-based cross-entropy loss function is taken to learn the network $\Psi = \{C, G, W\}$ specified by Eq.(3).

$$\min_{\Psi=\{C,G,W\}} J = -\sum_{n=1}^{N}\sum_{c=1}^{2} y_n^c \cdot log\left[\frac{\exp(\sum\limits_{k=1}^{d} W_k^c \cdot O_n^k + W_0^c)}{\sum\limits_{c=1}^{2}\exp(\sum\limits_{k=1}^{d} W_k^c \cdot O_n^k + W_0^c)}\right] \tag{3}$$

By optimizing the function in Eq.(3), the convolutional weights C, the GCP weights G, and the cloud activation weights W are learned simultaneously. Where $W_k^1 \in W$ indicates the contribution of the feature map for cloud.

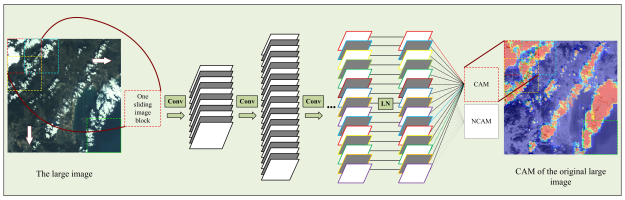

Fig. 3. The process of computing the cloud activation maps (CAM) of one large RS image.

2.2. Pixel-level cloud detection using the trained deep networks under block-level supervision

Based on the findings [11] and our task requirement of preserving precise spatial information, we prune the local pooling layers from our cloud detection network when generating the CAM and name this operation as LPP. Extensive experiments tend out that LPP enhances the spatial resolution of the output CAM. Fig.2 depicts the difference between deep networks with and without LPP when generating the CAMs. The spatial resolution of CAM increases significantly from 20 × 20 to 230 × 230 when we adopt the LPP strategy. It is noted that, the spatial resolution of generated CAMs will be resized to 250 × 250 so that the CAM will correspond to the size of the input image block. Given one image block b , the feature map f of the last convolutional layer can be calculated by Eq. (1), and f is used to compute the activation value at each channel with the GCP weights G. Then we perform the channel-wise multiplication to f with its activation values. After that we adjust the value of f to the appropriate range with a linear normalization (LN) operation by Eq. (4).

$$T^k = \frac{\delta(f^k)}{\tau(f^k)} \times f^k \tag{4}$$

where $T^{\,k}$ is the modified feature map of the k-th channel; $\delta(f^k) = f^k \otimes G^k$ denotes the activation value of the k-th channel of the last convolutional layer; $\tau(f^k)$ stands for a statistic value such as the average or median of f^k.

Furthermore, we calculate the CAM M^b of the block b by Eq. (5).

$$M^b = \sum_{k=1}^{d} W_k^1 \times T^k = \sum_{k=1}^{d} W_k^1 \times \frac{\delta(f^k)}{\tau(f^k)} \times f^k \tag{5}$$

where $W_k^1, k = 1, 2, \cdots, d$ stands for the cloud activation weights.

Given one large RS image, it is cropped to a set of overlapped blocks by sliding windows. By calculating the CAM of each block and mosaicking them together, the CAM of image is obtained as depicted in Fig.3. Due to the high-quality CAM, the binary cloud mask can be determined by an adaptive threshold segmentation algorithm. The visual segmenting results are shown as Fig.4.

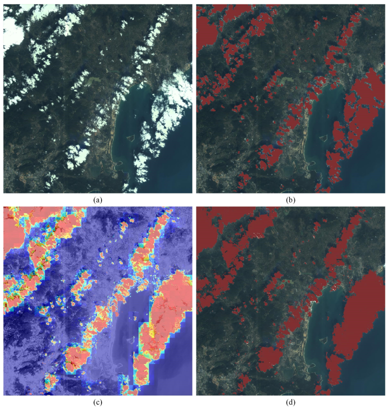

Fig. 4. The segmentation of CAM. (a) The original image, (b) The corresponding ground truth map of (a), (c) The computed CAM of (a), (d) The final cloud mask computed via segmenting the CAM (c).

3. EXPERIMENTS

In this section, we report the experiments. Section 3.1 introduces the detail description of the dataset, which is specifically collected for evaluating cloud detection via weakly supervised deep learning. Section 3.2 reports the experimental results.

3.1. The collected datasets

3.1.1. The training dataset with block-level labels

Given that there are no existing datasets for weakly supervised cloud detection like our WDCD method, we created a large-scale block-level dataset for it using the GaoFen-1 imagery. The dataset includes a training part and a validation part. The training part consists of 166,764 image blocks with their binary label which denotes the block containing cloud or not. Each image block is with the size of 250×250 and 166,764 image blocks are randomly cropped from 597 large image scenes from the GaoFen-1 satellite in various regions across China. More specifically, there are in total 79,316 image blocks containing the cloud while the rest 87,448 image blocks contain no cloud at all. With regard to training, we randomly select 90% of the training part to train the deep network while the rest 10% are used to adjust the hyper-parameters of the deep network.

3.1.2. The testing dataset with pixel-level labels

To evaluate the cloud detection performance, we build a testing dataset consisting of ten large image scenes and their pixel-level labeled cloud masks. Specifically, three images are from ZiYuan-3 satellite while the others are from GaoFen-1 satellite. The three large ZiYuan-3 image scenes are specially added to verify the generalization ability of our WDCD approach. The testing dataset is qualified to evaluate the cloud detection performance across multi-source RS data.

Each image in the testing dataset contains several typical land cover types such as cities, mountains, snow and ice, seas, lakes. And these images with good universality observed various locations from northeastern China to southwest China and even areas in southeast Asia.

3.2. The experimental results

In this section, we report quantitative detection results of our method as well as some baselines for evaluation of the cloud masks. As analyzed before, this is the first time that the weakly supervised deep learning idea is applied to cloud detection in RS imagery. To verify the superiority of our method, we design some baselines based on several methods designed for object detection using the weakly supervised deep learning idea in computer vision and RS domain.

Classify and assign (CAA) [10] uses DCNN to classify the image blocks as containing cloud or not. Considering the block as a whole, it roughly detects the cloud by assigning the value to all pixels inside the block based on the classification results. The network structure of CAM with GAP [8] is quite similar to CAA [10], while CAM with GAP replaces the fully connected layer with a global average pooling layer. After the training stage, CAM with GAP utilizes the activation weights to combine the feature map and generate the CAM, which is the same with our method. Paper [9] designs a method based on two-stage-learning (TSL) called CAM with TSL, which trains the convolutional weights and the cloud activation weights as two different stage and compute the CAM the same as CAM with GAP. We further evaluate the cloud detection performance of the cloud masks generated by segmenting these methods including: CAA [10], CAM with GAP [8], CAM with TSL [9], our proposed CAM with GCP and our proposed CAM with GCP+LPP. More specifically, several comprehensive metrics are used including overall accuracy (OA), Kappa, Intersection over Union (IOU), and F1. Table 1 shows that our proposed WDCD method obviously outperforms the baselines.

4. CONCLUSION

This paper proposes a new framework that can train the deep networks with only block-level binary labels indicating the image block contains cloud or not. To improve the representing ability of the feature map, we propose a new global pooling operation called GCP which can learn channel-wise convolutional weights of each channel of the feature map. After the iterative backward propagations, the feature map owns the ability to represent the region of the cloud and will be used to compute the CAM. Furthermore, we propose the LPP to improve the spatial resolution of the computed CAM. Through adaptively segmenting the CAM, the pixel-level cloud mask is obtained. With GCP and LPP, the trained deep networks can detect the pixel-level cloud mask. Even under this extremely weak supervision, the proposed WDCD approach still achieves promising results and outperforms the state-of-the-art methods. In general, the cloud detection results can be utilized for many applications such as cloud removal and shadow detection [4] and further support the continuous cartography and wide-range environmental evaluation.

Table 1. Metrics(%) of results by our method and the baselines.

	F1	mIOU	Kappa	OA
CAA [10]	53.0	37.2	43.8	82.6
CAM with GAP [8]	73.8	59.6	70.5	94.7
CAM with TSL [9]	73.5	59.5	70.4	95.1
Our WDCD method via CAM+GCP	78.1	64.6	75.3	95.5
Our WDCD method via CAM+GCP+LPP	**81.2**	**69.2**	**79.5**	**97.2**

5. REFERENCES

[1] Haruma Ishida, Yu Oishi, Keitaro Morita, Keigo Moriwaki, and Takashi Y Nakajima, “Development of a support vector machine based cloud detection method for modis with the adjustability to various conditions,” *Remote sensing of environment*, vol. 205, pp. 390–407, 2018.

[2] Dengfeng Chai, Shawn Newsam, Hankui K Zhang, Yifan Qiu, and Jingfeng Huang, “Cloud and cloud shadow detection in landsat imagery based on deep convolutional neural networks,” *Remote sensing of environment*, vol. 225, pp. 307–316, 2019.

[3] Shi Qiu, Zhe Zhu, and Binbin He, “Fmask 4.0: Improved cloud and cloud shadow detection in landsats 4–8 and sentinel-2 imagery,” *Remote Sensing of Environment*, vol. 231, pp. 111205, 2019.

[4] Zhiwei Li, Huanfeng Shen, Huifang Li, Guisong Xia, Paolo Gamba, and Liangpei Zhang, “Multi-feature combined cloud and cloud shadow detection in gaofen-1 wide field of view imagery,” *Remote sensing of environment*, vol. 191, pp. 342–358, 2017.

[5] Yihua Tan, Ji Qi, and Feifei Ren, “Real-time cloud detection in high resolution images using maximum response filter and principle component analysis,” in *2016 IEEE International Geoscience and Remote Sensing Symposium (IGARSS)*. IEEE, 2016, pp. 6537–6540.

[6] Yann LeCun, Yoshua Bengio, and Geoffrey Hinton, “Deep learning,” *nature*, vol. 521, no. 7553, pp. 436, 2015.

[7] Zhenfeng Shao, Yin Pan, Chunyuan Diao, and Jiajun Cai, “Cloud detection in remote sensing images based on multiscale features-convolutional neural network,” *IEEE Transactions on Geoscience and Remote Sensing*, vol. 57, no. 6, pp. 4062–4076, 2019.

[8] Bolei Zhou, Aditya Khosla, Agata Lapedriza, Aude Oliva, and Antonio Torralba, “Learning deep features for discriminative localization,” in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2016, pp. 2921–2929.

[9] Yansheng Li, Yongjun Zhang, Xin Huang, and Alan L Yuille, “Deep networks under scene-level supervision for multi-class geospatial object detection from remote sensing images,” *ISPRS journal of photogrammetry and remote sensing*, vol. 146, pp. 182–196, 2018.

[10] Karen Simonyan and Andrew Zisserman, “Very deep convolutional networks for large-scale image recognition,” *arXiv preprint arXiv:1409.1556*, 2014.

[11] Ian Goodfellow, Yoshua Bengio, and Aaron Courville, *Deep learning*, MIT press, 2016.