CLOUD DETECTION BY FUSING MULTI-SCALE CONVOLUTIONAL FEATURES

Zhiwei Li [1], Huanfeng Shen [1, 5, *], Yancong Wei [2], Qing Cheng [3], Qiangqiang Yuan [4, 5]

[1] School of Resource and Environmental Sciences, Wuhan University, Wuhan, P. R. China - (lizw, shenhf)@whu.edu.cn
[2] State Key Laboratory of Information Engineering in Surveying, Mapping and Remote Sensing, Wuhan University, Wuhan, P. R. China - ycwei@whu.edu.cn
[3] School of Urban Design of Wuhan University, Wuhan University, Wuhan, P. R. China - qingcheng@whu.edu.cn
[4] School of Geodesy and Geomatics, Wuhan University, Wuhan, P. R. China - yqiang86@gmail.com
[5] Collaborative Innovation Center of Geospatial Technology, Wuhan, P. R. China

KEY WORDS: Cloud detection, Deep learning, Convolutional feature fusion, Multi-scale, MSCN

ABSTRACT:

Clouds detection is an important pre-processing step for accurate application of optical satellite imagery. Recent studies indicate that deep learning achieves best performance in image segmentation tasks. Aiming at boosting the accuracy of cloud detection for multispectral imagery, especially for those that contain only visible and near infrared bands, in this paper, we proposed a deep learning based cloud detection method termed MSCN (multi-scale cloud net), which segments cloud by fusing multi-scale convolutional features. MSCN was trained on a global cloud cover validation collection, and was tested in more than ten types of optical images with different resolution. Experiment results show that MSCN has obvious advantages over the traditional multi-feature combined cloud detection method in accuracy, especially when in snow and other areas covered by bright non-cloud objects. Besides, MSCN produced more detailed cloud masks than the compared deep cloud detection convolution network. The effectiveness of MSCN make it promising for practical application in multiple kinds of optical imagery.

1. INTRODUCTION

Cloud cover impedes optical satellites from obtaining clear views of the Earth's surface, and thus the existence of clouds influences the availability of useful satellite data. Accurately extracting clouds from cloud-contaminated imagery can help to reduce the negative influences that cloud coverage brings to the application of the imagery. Therefore, cloud detection in optical imagery is of great significance.

In recent years, scholars have undertaken a great deal of research into cloud detection for different types of remote sensing data (Fisher, 2014; Li et al., 2017; Luo et al., 2008; Zhu and Woodcock, 2012). The traditional threshold-based cloud detection methods suffer from the problems of thin cloud omission and bright non-cloud object commission, especially for those images which have limited spectral information. To further improve the accuracy of cloud detection from single image, more spatial features such as geometric and texture features are combined with spectral features to enhance the diversity of clouds (Li et al., 2017). However, since most of cloud detection methods proposed in previous studies only used low-level spectral and spatial features, there are still rooms to promote cloud detection accuracy with the use of features at higher vision levels.

Recent advances have proven deep learning a very successful set of tools (Zhu et al., 2017). Benefiting from the application of the deep convolutional features, deep learning based methods such as ResNet (He et al., 2016) achieves high accuracy for image analysis tasks, and the accuracy is continuously promoted with appearances of new techniques. Deep learning is taking off in remote sensing as well. Cloud detection methods based on deep learning gradually appeared in recent studies, which can be divided into three categories according to input and output of the network: The first category is patch-label based approach, as for this approach, an image patch is used as the input of network, and the output is a label which denotes whether the image patch is cloudy (Gómez-Chova et al., 2017); The second category is region-label based approach, where the cloudy images are first segmented, then patches of different regions are labeled by the pre-trained network (Xie et al., 2017); The third category is pixel-label based approach. This kind of approach trained an end-to-end network, in which the input is an image patch of arbitrary size, while the output is a pixel-level labels which has same size of height and width as the input (Zhan et al., 2017).

Aiming at boosting the accuracy of cloud detection, in this paper, we proposed a deep learning based method to detect clouds by fusing multi-scale convolutional features. Compared to previous methods, the proposed method has made following improvements: 1) Better performance on distinguishing clouds and bright non-cloud objects; 2) More detailed cloud boundary information in output cloud mask. The rest contents of the paper include method introduction, experiment results, discussion and conclusions.

2. THE PROPOSED METHOD

Unlike the architectures of previous deep learning based cloud detection methods (Gómez-Chova et al., 2017; Xie et al., 2017) which uses full connected layers to output class labels, the architecture of our proposed network is designed based on fully convolutional network (FCN) (Shelhamer et al., 2017) and SegNet (Badrinarayanan et al., 2017). FCN is presented for end-to-end, pixels-to-pixels semantic segmentation, in which fully connected layers are replaced by convolutional layers to enable a classification net to output a spatial map, while SegNet is a deep convolutional encoder-decoder architecture which is proposed for semantic pixel-wise segmentation. MSCN shares a similar

[*] Corresponding author, Huanfeng Shen, Email: shenhf@whu.edu.cn

This contribution has been peer-reviewed. The double-blind peer-review was conducted on the basis of the full paper.
https://doi.org/10.5194/isprs-annals-IV-3-149-2018

architecture with FCN and SegNet, but with some improvements. To make full use of convolution features of different scales and further boost the cloud detection accuracy, multi-scale feature map fusion and residual network architecture (He et al., 2016) are applied in our network, respectively.

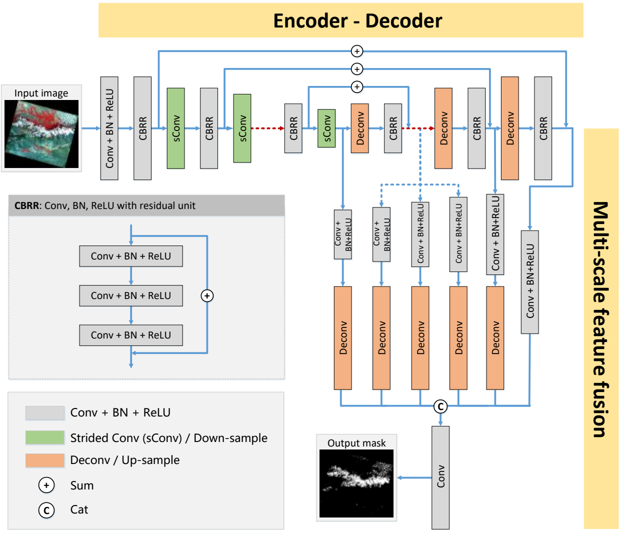


Figure 1. The architecture of the proposed MSCN. Feature maps with different scales are concatenated and fused. The output feature map of the final convolutional layer is regarded as cloud probability map and is fed to a binary classifier for pixel-wise binary cloud mask segmentation.

As shown in Figure 1, our network contains a symmetric architecture of convolutional encoders and corresponding decoders, then a feature fusion module follows to produce a final result. The encoder-decoder architecture consists of 41 layers, including basic convolutional/deconvolutional layers that do not change the scale of feature maps, and strided convolutional/deconvolutional layers with 2 pixels stride that down-sample and up-sample feature maps. The feature fusion module is employed to aggregate features at 6 scales from the decoder stage, in which feature maps are up-sampled by separated deconvolutional filters to the same size as the input image, then concatenated and fed to a simple convolutional layer for the final output. All the maps from convolutional layers except the last in the network are activated by Rectified Linear Unit (ReLU), and tricks for optimizing deep networks, such as Skip connection (Basement for Residual learning (He et al., 2016)) and Batch Normalization (Loffe and Szegedy, 2015), are also contained to boost its converging during training and improve its accuracy performance.

The training data was clipped from published GaoFen-1 WFV cloud and cloud shadow cover validation data, which was established in 108 global regions with a resolution of 16m, and was firstly released in (Li et al., 2017). To the best of our knowledge, it is the largest cloud and cloud shadow data set with manually labeled ground truth masks. In this paper, a total number of 73080 sample images with a size of 256x256x4 were selected with a fixed stride form the whole dataset. We randomly select 80% of samples as training data, and the remaining part as validation samples. For each training sample, a binary cloud mask was set as label and was artificially generated from the original image: 1 for cloud pixels and 0 for non-cloud pixels.

Besides, to reduce the negative influences of few incorrect labels in training data, in the end of our network we set a convolutional layer followed by mean square error loss to drive the network output cloud probability map, instead of binary cloud mask. Given a training dataset $\{x_i, y_i\}_{i=1}^{N}$ including multispectral images x_i and corresponding cloud masks y_i, our goal is to learn a model f that predicts cloud probability $f(x)$, optimal parameters in f can be learned by minimizing the mean square error loss which is averaged over the training set and defined as follow:

$$Loss = \frac{1}{N}\sum_{i=1}^{i=N}||y_i - f(x_i)||^2 \qquad (1)$$

For an output of cloud probability map, users can select appropriate threshold to segment the cloud probability map to a binary cloud mask, according to certain needs of different tasks. In this paper, a default segment parameter of 0.5 is set for better balancing the errors of commission and omission in segmented binary cloud mask. The training process relied on stochastic gradient descent (SGD) with a batch size of 128. The learning rate descended from 10^{-1} to 10^{-4} with an interval of every 30 epochs, and the momentum was fixed to 0.9. Finally, the proposed network was trained for 120 epochs (36560 iterations) in about 3 days on MatConvNet (Vedaldi and Lenc, 2015) with support from a Titan Xp GPU.

3. EXPERIMENT RESULTS AND ANALYSIS

To demonstrate the benefits of adding residual unit and multi-scale feature fusion, we compared the performances of the proposed network in situations with and without feature fusion or without residual unit. They are tested on a test dataset including another 73080 samples that are not overlapped with training data. As shown in Figure 2, from which it can be found that both feature fusion and residual unit are helpful to boosting the accuracy on cloud detection tasks.

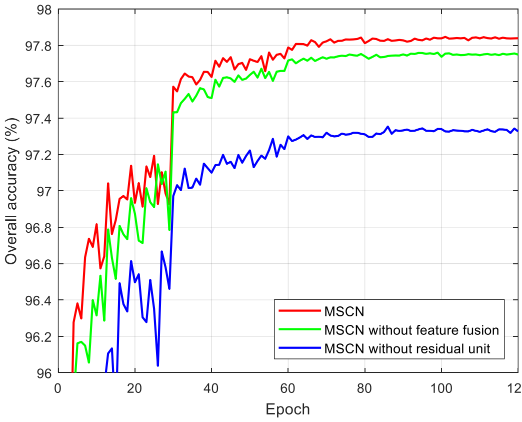


Figure 2. Accuracy curves for MSCN, MSCN without feature fusion, and MSCN without residual unit.

Additionally, the proposed MSCN method is compared with two types of cloud detection methods in quantitative and visual manners. One compared cloud detection method is our previously proposed multi-feature combined method (MFC) [1] which uses multiple spectral features and single scale spatial features to implement threshold-based cloud segmentation. Another compared method (termed as DCN in this paper) proposed in [9] designs a deep convolutional network with multi-scale prediction module for cloud and snow detection task. Both MFC and DCN method can be used for cloud detection in Gaofen-1 WFV imagery.

The accuracy curves of the proposed MSCN method and its comparisons with MFC are shown in Figure 3, in which we can see that the overall cloud accuracy is promoted with the number of training epoches increases and finally converged. The cloud accuracy of MSCN surpass MFC in terms of average overall accuracy (97.85% VS 96.80%), producer’s accuracy and user’s accuracy. The Figure 4 shows cloud detection examples of MSCN and MFC, in which we can see that MSCN has a superior advantage over MFC in areas of bright non-cloud objects covered including snow and bright water. In this paper, a global MODIS land cover product of 0.05 degree resolution is used for classifying the validation dataset into six categories. We observe that MSCN and MFC both acquired high accuracies in vegetation, urban and wetlands areas, and it should be noted that the pixel-level average overall cloud accuracy in snow covered areas is significantly boosted from 39.94% of MFC to 86.51% of MSCN, the overall cloud accuracies of MSCN in water and barren areas are also higher than MFC.

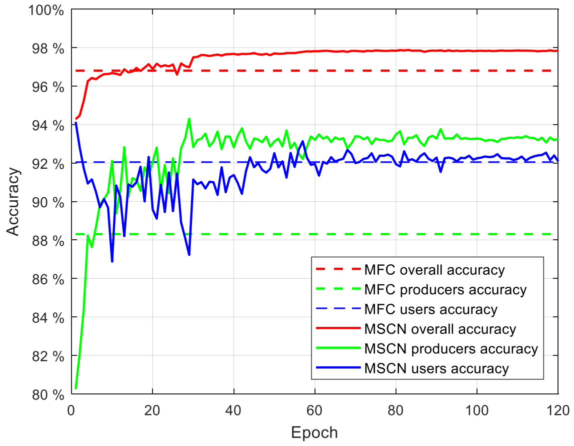


Figure 3. Accuracy tests of the proposed MSCN method.

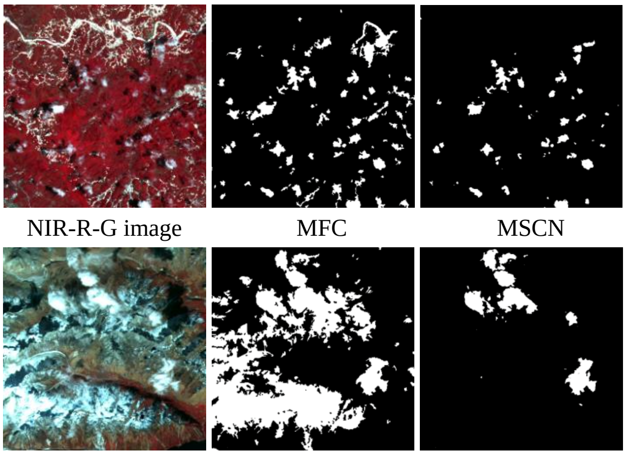

NIR-R-G image MFC MSCN

Figure 4. Cloud detection results of MSCN and MFC.

The DCN method was trained on large cloud and snow data set which is based on 50 Gaofen-1 WFV images. Considering that the providing experiment results of DCN are geometrically mismatched with the established reference cloud masks, we only compare MSCN with DCN in visual inspection manner. The Figure 5 shows cloud detection examples of MSCN and DCN. It can be observed that cloud detection results of DCN are rough and lose details in cloud boundaries, while MSCN provides more refined cloud masks because of the use of deconvolutional layers and multi-scale convolutional feature fusion. We will discuss the reasons of these differences in the following section.

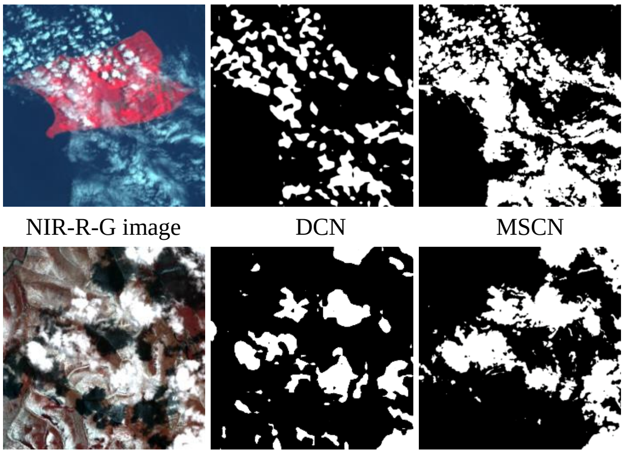

NIR-R-G image DCN MSCN

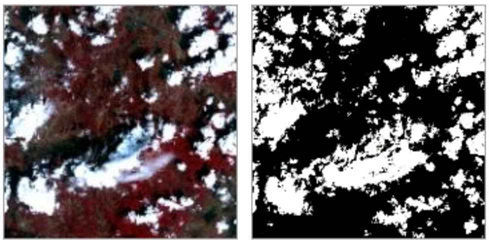
Figure 5. Cloud detection results of MSCN and DCN.

4. DISCUSSION

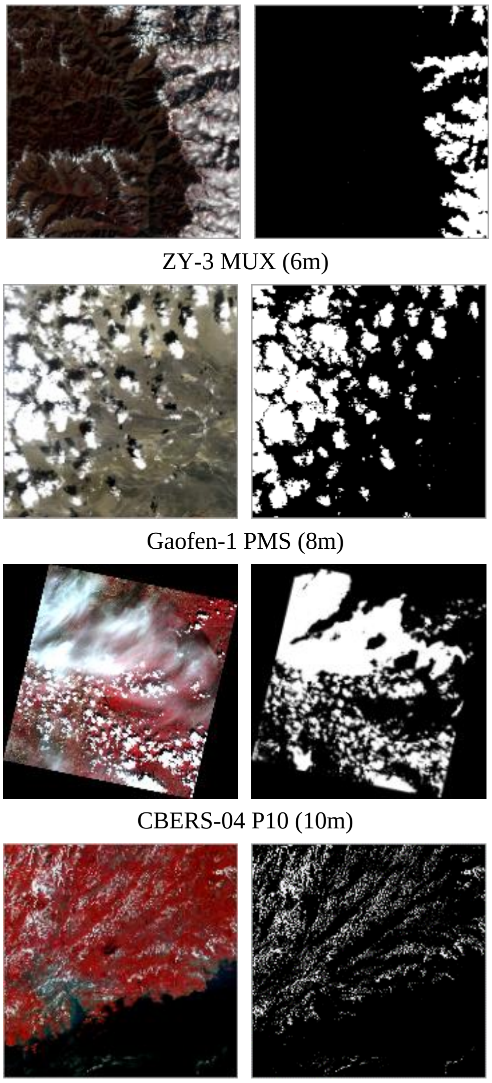
There are several reasons that MSCN produces more accurate masks than MFC and DCN. Firstly, the uses of deconvolutional layers and skip connection in MSCN help to fuse the feature hierarchy to combine shallow appearance information and deep sematic information, multi-scale convolution feature fusion additionally makes MSCN better ability to distinguish clouds and bright noncloud objects. Secondly, MSCN apply residual learning to find better optimal convergence and boost the accuracy. Thirdly, our network is trained on a global-scale data set which consists of many types of land cover, the diversity of training data makes MSCN a stronger capability to cope with different cases.

Without any parameters adjustment, the pre-trained MSCN network can also be applied to cloud detection for other types of multispectral imagery which has similar spectral setting. We have tested MSCN in multispectral images with different resolution ranges from 1m to 50m, such as Gaofen-2 PMS (4m), ZY-3 MUX (6m), Gaofen-1 PMS (8m), CBERS-04 P10 (10m), Gaofen-4 PMS (50m) etc., MSCN acquired not bad cloud detection results as show in Figure 6. To the best of our knowledge, it’s the first time that a single cloud detection method can directly process so many types of optical imagery with different resolution.

In our implementation, MSCN takes less than 10 seconds on GPU mode or 5 minutes on CPU mode to process a whole Gaofen-2 PMS image (4503x4548x4 pixels) in MatConvNet on a computer with a Titan Xp GPU and a Core i7-7700K CPU. In the future, in order to achieve better performance in a specific kind of imagery, it is essential to fine tune the pre-trained model with small learning rate using specific type of imagery.


Gaofen-2 PMS (4m)

https://doi.org/10.5194/isprs-annals-IV-3-149-2018 |



ZY-3 MUX (6m)

Gaofen-1 PMS (8m)

CBERS-04 P10 (10m)

Gaofen-4 PMS (50m)

Figure 6. Cloud detection examples of MSCN in other kinds of multispectral imagery in different land covers.

5. CONCLUSION

In this paper, a multi-scale cloud network of convolutional encoders and corresponding decoders architecture followed by a feature fusion module is proposed to implement cloud detection for multispectral images. Experimental results indicate that multi-scale convolutional feature fusion and residual network architecture are both helpful to boost the accuracy of cloud detection. Additionally, MSCN achieves higher accuracy than traditional MFC method, and has obvious advantage of keeping cloud boundary details in produced cloud mask over the deep convolutional network method. The effectiveness of MSCN make it promising for future practical application in more kinds of optical images.

In our future study, we will generalize the MSCN method to more images such as SPOT-6/7, WorldView-2/3, and investigate the possibility of cloud and cloud shadow detection for multiple kinds of imagery using single model.

ACKNOWLEDGEMENTS

We gratefully acknowledge the support of NVIDIA Corporation with the donation of GPU used for this research, and also the authors of (Zhan et al., 2017) and Lingxia Ma for providing experiment results of the DCN method.

REFERENCES

Badrinarayanan, V., Kendall, A., Cipolla, R., 2017. SegNet: A Deep Convolutional Encoder-Decoder Architecture for Image Segmentation. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 39(12), pp. 2481-2495.

Fisher, A., 2014. Cloud and cloud-shadow detection in SPOT5 HRG imagery with automated morphological feature extraction. *Remote Sensing*, 6(1), pp. 776-800.

G ómez-Chova, L., Mateo-Garc á, G., Camps-Valls, G., 2017. Convolutional neural networks for multispectral image cloud masking, In: *2017 IEEE International Geoscience and Remote Sensing Symposium (IGARSS)*, Texas, USA, pp. 2255-2258.

He, K., Zhang, X., Ren, S., Sun, J., 2016. Deep Residual Learning for Image Recognition, In: *the 29th IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, Las Vegas, USA, pp. 770-778.

Li, Z., Shen, H., Li, H., Xia, G., Gamba, P., Zhang, L., 2017. Multi-feature combined cloud and cloud shadow detection in GaoFen-1 wide field of view imagery. *Remote Sensing of Environment*, 191, pp. 342-358.

Loffe, S., Szegedy, C., 2015. Batch normalization: Accelerating deep network training by reducing internal covariate shift, In: *International Conference on Machine Learning (ICML)*, Lille, France, pp. 448-456.

Luo, Y., Trishchenko, A.P., Khlopenkov, K.V., 2008. Developing clear-sky, cloud and cloud shadow mask for producing clear-sky composites at 250-meter spatial resolution for the seven MODIS land bands over Canada and North America. *Remote Sensing of Environment*, 112(12), pp. 4167-4185.

Shelhamer, E., Long, J., Darrell, T., 2017. Fully convolutional networks for semantic segmentation. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 39(4), pp. 640-651.

Vedaldi, A., Lenc, K., 2015. Matconvnet: Convolutional neural networks for matlab, In: *Proceedings of the 23rd ACM International Conference on Multimedia*, Brisbane, Australia, pp. 689-692.

Xie, F., Shi, M., Shi, Z., Yin, J., Zhao, D., 2017. Multilevel Cloud Detection in Remote Sensing Images Based on Deep Learning. *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, 10(8), pp. 3631-3640.

Zhan, Y., Wang, J., Shi, J., Cheng, G., Yao, L., Sun, W., 2017. Distinguishing Cloud and Snow in Satellite Images via Deep Convolutional Network. *IEEE Geoscience and Remote Sensing Letters*, 14(10), pp. 1785-1789.

Zhu, X., Tuia, D., Mou, L., Xia, G.-S., Zhang, L., Xu, F., Fraundorfer, F., 2017. Deep Learning in Remote Sensing: A Review. *IEEE Geoscience and Remote Sensing Magazine*.

Zhu, Z., Woodcock, C.E., 2012. Object-based cloud and cloud shadow detection in Landsat imagery. *Remote Sensing of Environment*, 118, pp. 83-94.