IGARSS 2022 - 2022 IEEE International Geoscience and Remote Sensing Symposium | 978-1-6654-2792-0/22/$31.00 ©2022 IEEE | DOI: 10.1109/IGARSS46834.2022.9883360

SCTRANS: A TRANSFORMER NETWORK BASED ON THE SPATIAL AND CHANNEL ATTENTION FOR CLOUD DETECTION

Wenke Jiao [1], *Yongjun Zhang* [1], *Bin Zhang* [1], *Yi Wan* [1*]

[1] School of Remote Sensing and Information Engineering, Wuhan University, 430079, China
*Corresponding author: yi.wan@whu.edu.cn

ABSTRACT

Cloud detection is an important preprocessing step for remote sensing image processing and analysis. The current deep-learning-based cloud detection methods are mostly based on Convolutional Neural Network (CNN) which pay more attention to local information. To make more use of the global information, in this article, we propose a transformer-based cloud detection method (SCTrans) based on the spatial and channel attention mechanism. The experiment results show that when using only three-band images on the Landsat7 dataset, the mIoU of the validation set reaches 85.92% and the mIoU of the test set reaches 87.86%. The experimental results show that the proposed network has a higher mIoU and F1 score than Fmask and other networks.

Index Terms— Cloud detection, deep learning, transformer, attention mechanism, neural networks, remote sensing

1. INTRODUCTION

Remote sensing images are inevitably contaminated by clouds, which limit the subsequent use of remote sensing images. A large proportion of remote sensing data are destructed due to the existence of clouds, which affects the application of target detection, image fusion, and image registration [1]. Therefore, cloud detection is an important preprocessing step for remote sensing images. Cloud detection methods include threshold methods, methods based on the spatial and texture characteristics, and methods based on machine learning [2]. Fmask [3] is a classic cloud detection method based on the threshold, the threshold methods need to set the threshold according to the cloud and spectral characteristics to achieve better detection results. Because of seasons and geographic location, the threshold will also change. Since different seasons and land covers will need different thresholds, the threshold methods tend to have lower accuracy. The texture-based detection methods detect clouds according to the spatial and geometric characteristics of the cloud. Tian et al. [4] used the grey level cooccurrence matrix to get the spatial distribution of numeric counts. Compared with the threshold methods, the texture-based detection methods improve the accuracy and scalability. However, due to the diversity of cloud features, the effect is still not ideal. The cloud detection methods based on machine learning use models to extract features from the training set and adjust hyperparameters through multiple experiments to obtain the optimal model.

Deep learning is a subfield of machine learning. In recent years, with the proposal of FCN [5], semantic segmentation networks have shined in the field of remote sensing, such as cloud detection, building detection, and so on. Researchers begin to use semantic segmentation networks for cloud detection. Drönner et al. [6] proposed CS-CNN, which is a fast cloud detection method. Francis et al. [7] proposed CloudFCN based on an encoder-decoder structure to detect clouds. Cloud detection is essentially a pixel-by-pixel classification problem. The main goal is to distinguish clouds from ground objects in remote sensing images. Although people have explored many methods based on CNN to improve the accuracy of cloud detection, their ability to extract features is still limited. In recent years, the vision transformer (ViT) has shown the potential for global modeling to learn long-distance information in hyperspectral image classification [8]. Thus, we propose a semantic segmentation network based on Mix Transformer (MiT) [9] and incorporate the attention mechanism. Experiments on the landsat7 dataset show that our method significantly outperforms previous methods. The network structure is shown in the Fig. 1. This is an Encoder-Decoder architecture. Our main contributions are as follows:

1) A transformer-based network is proposed for cloud detection. The MiT is introduced to extract features, which has higher accuracy than the CNN-based architectures. At the same time, we take advantage of the U-shaped network in cloud detection to construct a U-shaped network based on the MiT, which helps to identify global features and improves the computing efficiency.

2) On the decoder side, we incorporate the attention mechanism named Convolutional Block Attention Module (CBAM) [10], which helps our model to highlight critical information and ignore unimportant information.

2. METHOD

U-Net [11] is an image segmentation algorithm, originally used for medical image segmentation. As an efficient

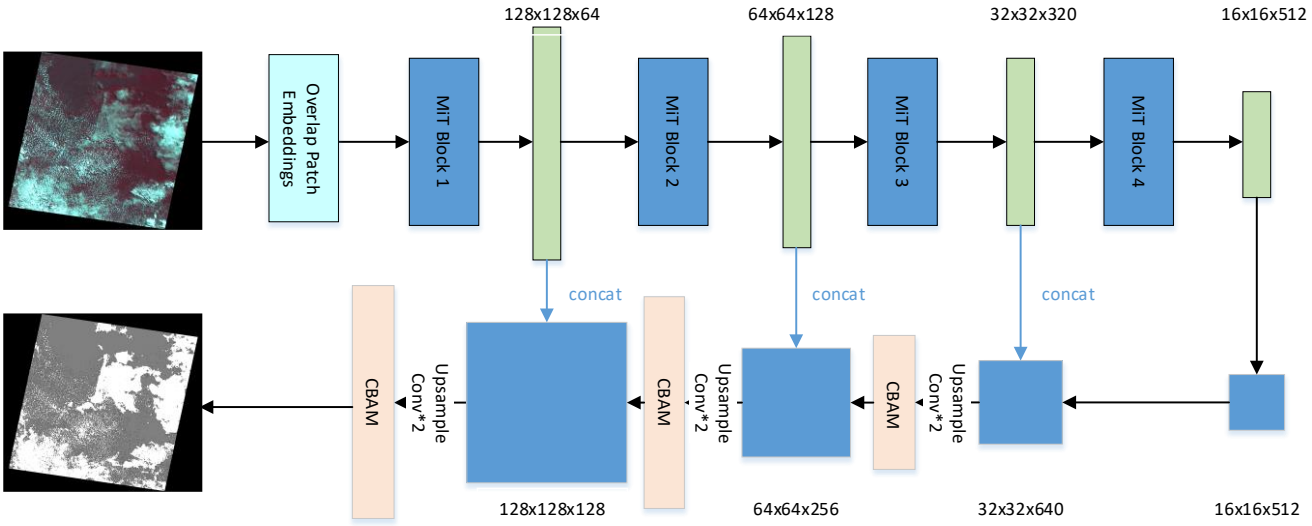

Fig. 1. Framework of the proposed SCTrans method.

semantic segmentation network. Researchers introduce U-Net into the field of remote sensing. U-Net has shown excellent performance on cloud detection tasks, but there is still room for improvement. ViT exhibits very competitive modeling capabilities, proving that the transformer-based architecture in the field of natural language processing performs better than CNN in image classification tasks. Convolution is a local operation and usually models the relationship between neighboring pixels, while the transformer is a global operation, which can model the relationship between all pixels. We combine the advantages of U-Net and the MiT and introduce CBAM [10]. U-Net combines multi-scale features through jump connections, and focuses on the structured information of the image. The transformer focuses more on the semantic information of the image. Next, we will introduce our model from the encoder and the decoder.

2.1. Transformer encoder

ViT is the first work to prove that the transformer is effective in image classification, ViT reshapes the image into a sequence of patches, which also serves as the input sequence for the transformer.

In semantic segmentation, the encoder is usually a pre-trained classification network, such as VGG, Resnet. It generally outputs high-resolution coarse-grained features and low-resolution fine-grained features. The feature maps extracted by the encoder can be input into the decoder to obtain the result of pixel-by-pixel classification. A suitable encoder is very important. Compared with the commonly used CNN in semantic segmentation, ViT has obvious advantages in accuracy performance on public dataset, but the disadvantages are also obvious. The parameters and calculations are large and require large video memory. Due to the position embedding in ViT, images of different resolutions need to be interpolated during the test process, which will lead to a decrease in accuracy.

Inspired by ViT, Xie et al. [9] design a series of the MiT encoders, MiT-B0 to MiT-B5, with the same architecture but different sizes, our model uses MiT-B2 as the encoder. ViT can only generate a single-resolution feature map, which results in limited information used by the decoder. The MiT can output CNN-like multi-resolution feature maps.

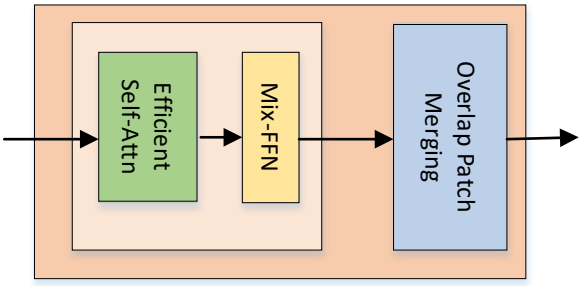

Fig. 2. Structure of the MiT block.

The structure of the MiT block is shown in the Fig. 2, the MiT block includes efficient self-attention, Mix-FFN and overlap patch merging. In self-attention, the attention map is obtained by multiplying the input vector **X** by the weight matrix **Q**, **K**, **V**, which have the same dimensions $H \times W \times C$. The multi-head self-attention generates multiple attention maps, then cascades the results of multiple attention maps, which is similar to the multi-channel mechanism in CNN. CNN obtains feature maps of different dimensions, concatenates them, and then maps back to the original dimensions through a parameter matrix. The d_{head} is dimension of the head. The calculation process of a conventional multi-head self-attention is as follows:

$$\text{Attention}(\mathbf{Q}, \mathbf{K}, \mathbf{V}) = \text{Softmax}\left(\frac{\mathbf{Q}\mathbf{K}^{\mathbf{T}}}{\sqrt{d_{head}}}\right)\mathbf{V} \qquad (1)$$

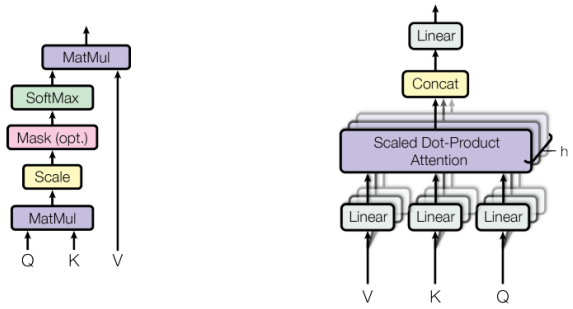

Fig. 3. Self-attention(left) and multi-head self-attention(right)

The structure of self-attention and multi-head self-attention is shown in the Fig. 3. To reduce the computational complexity of multi-head self-attention, efficient self-attention controls the size of the **K**, **V** matrix in multi-head self-attention by increasing the reduction ratio R. Efficient self-attention changes the shape of the **K** by controlling R, reducing the complexity from $O((H*W)^2)$ to $O(\frac{(HW)^2}{R})$. The formula is as follows:

$$\widehat{\mathbf{K}} = \text{Reshape}\left(\frac{HW}{R}, C \cdot R\right)(\mathbf{K}) \quad (2)$$

$$\mathbf{K} = \text{Linear}(C \cdot R, C)(\widehat{\mathbf{K}}) \quad (3)$$

To solve the problem of inconsistent resolution in the test process caused by position embedding in ViT, the MiT uses 3×3 convolution in Mix-FFN to transmit position information. Overlap Patch Merging reduces the size of the feature map and increases the number of channels.

2.2. Attention decoder

The decoder's input comes from two parts. The first is feature maps of different scales through jump connection after passing through the transformer encoder. The second is the feature maps from the upper level. We concatenate the two obtained feature maps. Then the height and width of the feature map are doubled by up sampling layer and the convolutional layer, and the number of channels is halved. Combining the feature maps obtained by these two methods can help the model obtain multi-scale information, thereby better identifying the cloud boundary. However, some redundant information may also affect the performance of the model. To avoid this effect and make the model more focused on cloud-related information, we introduced CBAM, which is a kind of attention mechanism, which combines Channel attention and spatial attention are combined. After obtaining the fusion feature map of each layer, we passed CBAM once to increase the focus on the cloud area and suppress unimportant features, the final mask image is obtained until the original resolution is restored.

3. EXPERIMENT

3.1. Dataset

To verify the performance of the model, we conducted training, validation and testing on the Landsat-7 Irish dataset [12], which contains eight bands. Our study selected three of these bands: Band 1(Blue), Band 2(Green) and Band 3(Red) to compose three RGB channel images. Landsat-7 Irish dataset contains 206 images, which are uniformly distributed in nine latitude regions, and the ground truths (GTs) of these images are manually pixel-wise annotated. We selected 180 images, including 107 images in the training set, 36 images in the validation set, and 37 images in the test set. We divided these images into two categories, cloud and others. During the training process, it is difficult to input large images into the network to train the model, so we cropped the images in the training set into 24,504 images whose size is 512x512 with an overlap rate of 20%, and we used large images for evaluation during validation and testing.

3.2. Parameter setup

Our method is implemented with PyTorch on Ubuntu and two NVIDIA Tesla V100 (16GB) GPUs and optimized by the Adam with decoupled weight decay (AdamW [13]). We use poly as the learning rate policy with the initial learning rate of 0.00006. Besides, the total number of iterations is 80000, where the batch size is set to 8. We select mean intersection over union (mIoU), overall accuracy (OA) and F1 score as evaluation matrices to quantitatively evaluate the performance of cloud detection networks.

3.3. Results and analysis

We compared our SCTrans with other approaches, including Danet[14], DeeplabV3[15], Pspnet[16], FCN, U-Net. The results of the validation set and test set of Landsat 7 Irish are shown in Table 1 and Table 2.

Table 1. Results on val set of different methods

Method	Acc/%	mIoU/%	mF1/%
Fmask [3]	89.18	79.03	88.18
Danet [14]	91.88	82.83	90.49
DeeplabV3 [15]	92.53	84.01	91.2
Pspnet [16]	92.85	84.82	91.7
FCN [5]	93.01	85.02	91.81
U-Net [11]	93.06	85.18	91.91
SCTrans(ours)	**93.45**	**85.92**	**92.35**

Table 2. Results on test set of different methods

Method	Acc/%	mIoU/%	mF1/%
Fmask[3]	89.19	79.48	88.5
Danet[14]	92.18	85.8	92.29
DeeplabV3[15]	92.58	86.85	92.9
Pspnet[16]	93.15	87.1	93.05
FCN [5]	93.32	87.54	93.3
U-Net [11]	93.47	87.8	93.46
SCTrans(ours)	**93.61**	**87.86**	**93.49**

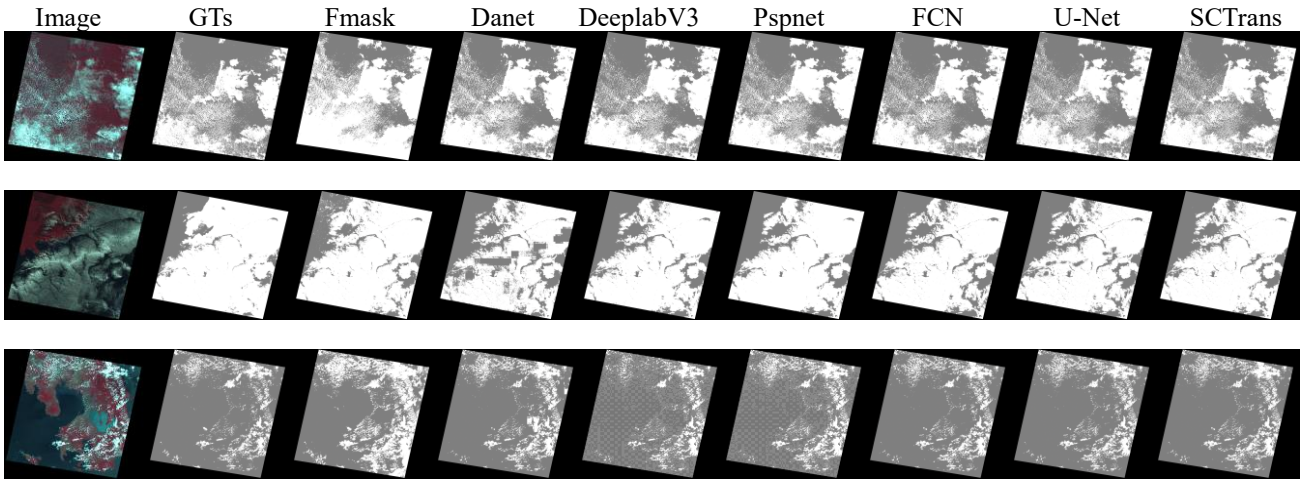

Fig. 4. Comparison between the results of different methods in Landsat7 dataset

It can be seen from Table 1 and Table 2 that SCTrans is better than Fmask and other mainstream semantic segmentation networks. On the validation set, mIoU reached 85.92%, and on the test set, mIoU reached 87.86%, which shows that SCTrans has higher accuracy and better robustness. Fig. 4 shows the comparison results of SCTrans and other methods. SCTrans handles broken clouds and thin clouds better than other methods.

4. CONCLUSION

In this paper, we proposed a new cloud detection method, SCTrans, which is based on transformer. SCTrans incorporates the spatial and channel attention mechanism which helps our model to highlight salient features and ignore irrelevant regions. The experiment results have proved the effectiveness of our network to extract cloud areas.

5. ACKNOWLEDGEMENT

This work was supported by the National Natural Science Foundation of China under project number 42030102.

6. REFERENCES

[1] J. Wei, Z. Li, Y. Peng, and L. Sun, "MODIS Collection 6.1 aerosol optical depth products over land and ocean: validation and comparison," *Atmos Environ*, vol. 201, pp. 428–440, 2019.

[2] Hou, S.W.; Sun, W.F.; Zheng, X.S., "Overview of cloud detection methods in remote sensing images," *Space Electron. Technol,* vol.11, pp. 68–76, 2014.

[3] Z Zhu and C E Woodcock, "Object-based cloud and cloud shadow detection in Landsat imagery," *Remote Sensing of Environment, vol,* 118, no. 6, pp. 83-94, 2012.

[4] Tian B et al., "A study of cloud classification with neural networks using spectral and textural features," *IEEE Transactions on Neural Networks, vol,* 10, no. 1, pp. 138–151, 1999.

[5] Long, J., Shelhamer., E, & Darrell, T. "Fully convolutional networks for semantic segmentation," *In Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 3431-3440, 2015.

[6] J. Drönner et al., "Fast cloud segmentation using convolutional neural networks," *Remote Sens.*, vol. 10, no. 11, pp. 1782, Nov. 2018.

[7] A. Francis, P. Sidiropoulos, and J.-P. Muller, "CloudFCN: Accurate and robust cloud detection for satellite imagery with deep learning," *Remote Sens.*, vol. 11, no. 19, pp. 2312, Oct. 2019.

[8] A. Dosovitskiy et al., "An image is worth 16 × 16 words: Transformers for image recognition at scale," *in Proc. Int. Conf. Learn. Represent. (ICLR)*, pp. 1–21, 2021.

[9] Enze Xie, Wenhai Wang, Zhiding Yu, Anima Anandkumar, Jose M Alvarez, and Ping Luo. "Segformer: Simple and efficient design for semantic segmentation with transformers." *In NeurIPS,* 2021. 8

[10] Sanghyun Woo et al., "CBAM: Convolutional Block Attention Module," *Proceedings of the European Conference on Computer Vision (ECCV),* pp. 3-19, 2018.

[11] O. Ronneberger et al., "U-net: Convolutional networks for biomedical image segmentation," *in Medical Image Computing and Computer-Assisted Intervention*. Springer, pp. 234–241,2015.

[12] Scaramuzza, P.L., Bouchard, M.A. & Dwyer, J.L. "Development of the Landsat data continuity mission cloud-cover assessment algorithms," *IEEE Transactions on Geoscience and Remote Sensing,* vol. 50, no. 4, pp, 1140-1154, 2012.

[13] Loshchilov I, Hutter F. "Decoupled weight decay regularization," *International Conference on Learning Representations*, 2019.

[14] J. Fu et al., "Dual Attention Network for Scene Segmentation," *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR),* pp. 3141-3149, 2019.

[15] Chen, L.C. el al., "Rethinking Atrous Convolution for Semantic Image Segmentation," *In Proceedings of the IEEE conference on computer vision and pattern recognition*, 2017.

[16] H. Zhao, J. Shi, X. Qi, X. Wang and J. Jia, "Pyramid Scene Parsing Network," *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR),* pp. 6230-6239, 2017.