OPTICAL SATELLITE IMAGE CHANGE DETECTION VIA TRANSFORMER-BASED SIAMESE NETWORK

*Yang Wu, Yuyao Wang, Yanheng Li, Qizhi Xu**

College of Mechanical and Electrical Engineering, Beijing Institute of Technology

IGARSS 2022 - 2022 IEEE International Geoscience and Remote Sensing Symposium | 978-1-6654-2792-0/22/$31.00 ©2022 IEEE | DOI: 10.1109/IGARSS46834.2022.9884408

ABSTRACT

Optical satellite image change detection is essential to monitor the use of Earth's resources. Convolutional neural networks(CNN)-based methods exhibit excellent performance on change detection. As Transformers became the de-facto standard in the field of natural language processing(NLP), there were more and more methods based on it are proposed in computer vision, such as image classification, object detection, semantic segmentation and so on. Many proposed models based on vision Transformer(ViT) have surpassed the performance of CNN and show effectiveness and superiority. With the emergence of more and more applications of ViT in the field of image processing, it's advantages are gradually being explored. In terms of change detection, the CNN-based models have already shown great advantages over traditional methods. In view of current achievements of Transformer, we decided to apply Transformer to change detection in optical satellite image. Change detection of bi-temporal images, we need to take two images as inputs. So we proposed a Siamese extensions of ViT networks which achieve the best results in tests on two open change detection datasets. Experimental results on real datasets show the effectiveness and the superiority of the proposed network.

Index Terms— optical satellite image, vision Transformer, change detection, Siamese networks

1. INTRODUCTION

Considering the importance of optical satellite image for observing, understanding and using the Earth's resources, change detection is an indispensable technical tool for optical satellite image utilization. Optical satellite image change detection can help us dynamically monitor the use of the earth's resources, the progress of urban construction[1], the impact of natural disasters[2]. Change detection of optical satellite image can be generally described as using of satellites to acquire images of the same area at different periods, and the researcher marks each pixel of the acquired image pair whether there have changed or not through comprehensive analysis. The results obtained using manual processing have a high accuracy, precision and low errors. In the face of a massive data, however, this method is inefficient and the labour cost is too high to accept. We need to design an efficient algorithm for automatic change detection on optical satellite image images.

The development of optical satellite image change detection also has advanced alongside computer vision and image processing techniques[3, 4]. When using traditional methods for change detection, the researcher's experience and prior knowledge play a crucial role in design of the algorithm, and requires rigorous analysis and complex processing of images, sometimes manual intervention. The large-scale application of deep learning has turned things around. As deep learning can automatically obtain complex features from massive amounts of data to effectively perform change detection without human intervention. Deep learning can fully utilize the parallel computing power of computers, and it's more efficient compared to traditional methods. Therefore, the use of deep learning for change detection in optical satellite image has become a hotspot.

In this paper, we propose a Transformer-based Siamese network for change detection of bi-temporal optical satellite images. It's an end-to-end network that can training from scratch using change detection datasets. When the model training is finished, we input two images of different periods in the same place, then the model could directly output a binary image, which each pixel of the image is used to indicate changed or unchanged. Objects tend to have complex features in optical satellite image with high-resolution. Compared with CNN-based deep learning methods, Transformer make it better to obtain the global information of images, extract complex features of images. To achieve a such function, two-stream Transformer model with the Cross Entropy loss function is adopted.

The rest of this paper is organized as follows. Section 2 discusses some other work that were used for inspiration or comparison during the development of this work. We describes in detail the proposed Transformer architectures for change detection of optical satellite image in Section 3. Section 4 contains the results of experiments and the comparison with several previous change detection methods. This work's concluding remarks in Section 5.

This work was supported by the National Natural Science Foundation of China under Grant 61972021 . (Corresponding author: Qizhi Xu.)

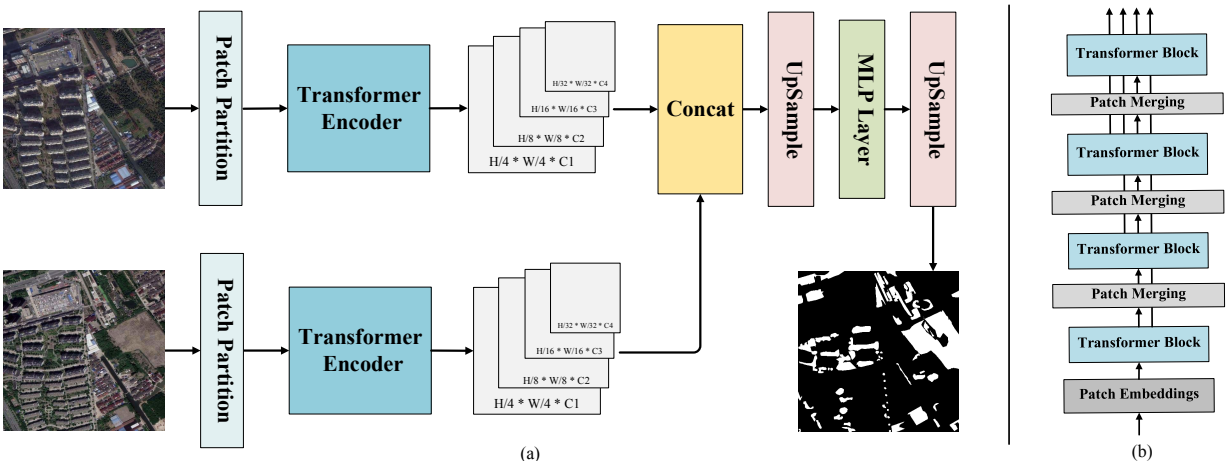


Fig. 1: (a) Overview of proposed framework. (b) structure of Transformer Encoder

2. RELATED WORK

The history of change detection technology move forward alongside with the development of computer technology. The development of change detection methods can be divided into two stages: traditional methods stage and deep learning stage. Traditional methods such as OTSU thresholding, PCA-K-means and feature fusion[5]. When researchers design algorithms for change detection using optical satellite image, they usually need to manually design features for extraction, which relies heavily on the researcher's experience and technical ability. Thus, this situation conduct low accuracy and high false detection rate. The emergence of deep learning has greatly improved this situation. When we using deep learning to detect changes in optical satellite image, we can extract complex and abstract features from massive data for learning without manual intervention, so as to obtain better results. We can also use an end-to-end structure when designing deep neural network, which reduces the complexity of operations and labour costs, improves efficiency.

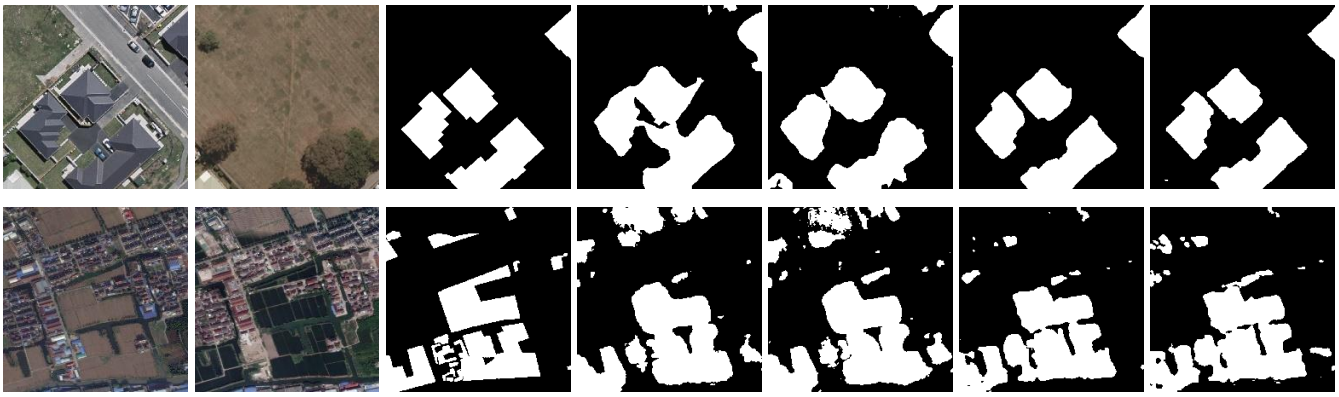
In terms of image analysis and processing, CNN-based architectures dominate in deep learning previously, and change detection via deep learning is mostly done using CNN-based architectures. Wang Q et al. proposed a high-resolution optical satellite image change detection framework based on Faster R-CNN[6]. DTIS, ONERA et al. proposed two Siamese extensions of fully convolutional networks as FC-Siam-conc and FC-Siam-diff[7]. Nevertheless, it is difficult to use global information for change detection using CNN alone. Hao Chen et al. proposed spatial-temporal attention-based method(STANet) for change detection[8], They introduced a self-attention mechanism to enhance the features extracted by the CNN.

As Transformer occupies a dominant position in the field of NLP, it has also been widely used in the computer vision and image processing. ViT[9] have already achieved outstanding achievements in image classification, object detection, semantic segmentation and many other research directions. ViT provides a new through of detecting changes in optical satellite image. Transformer can obtain better global information compared with the CNN-based architecture, but it has the disadvantage of single scale feature extraction. Ze Liu et al. proposed swin-Transformer to solve the shortcoming[10]. A lightweight decoder was proposed by Enze Xie et al to reduce the number of parameters on [11]. Based on that, we proposed Transformer-based Siamese network for change detection. This network requires two inputs, and outputs a binary image, and described in section 3.

3. PROPOSED METHOD

The proposed Transformer-based model consists of two part, the modified Transformer encoder for extracting features and the lightweight decoder for producing result. The concrete implementations are described in details in this section. An overview of the network architecture is presented in Fig.1. These architectures are able to be trained end-to-end, and we can test a pair of optical satellite image using the trained model. The images which fed into the model could be extracted multi-scale features by proposed Transformer encoder, the decoder directly fuse these multi-scale features and the change detection results would be produced.

The encoder contains two identical networks sharing weight, which is built by a patch partition module, the modified Transformer blocks with multi-head self-attention module and patch merge layers. The Transformer only accepts a 1-Dimension sequence of feature embeddings as input, so we reshape the image $\mathbf{x} \in \mathbb{R}^{H\times W\times C}$ into a sequence by flattening 2D patches $\mathbf{x}_p \in \mathbb{R}^{N\times(P^2\cdot C)}$, where *(H,W)* is the resolution of the original image, *C* is the number of chan-



(a) Image 1 (b) Image 2 (c) Ground Truth (d) FC-EF (e) FC-Siam-Conc (f) STANet (g) Ours

Fig. 2: Comparison between the results obtained, first row represents WHU-CD, second represents SE-CD.

nels,*(P,P)* is the resolution of each image patch, we use a patch size of 4×4 and thus the feature dimension of each patch is $4 \times 4 \times 3 = 48$, and $N = HW/P^2$ is the resulting number of patches. Each patch is treated as a "token". A embedding layer is applied on this feature to project it to an arbitrary dimension denoted as *C*. Several Transformer blocks are applied on these patch tokens to extract features. To produce a hierarchical representation, the number of tokens is reduced by patch merging layers as the network gets deeper. As a result, the proposed architecture can conveniently get multi-scale features.

The proposed lightweight decoder consisting only of MLP layers and up-sampling layers. It consists of four main steps. First, two group of multi-scale features from the Siamese Transformer encoder are concatenated together at same scale. Then the concatenated low-resolution features are up-sampled to same size of high-resolution features. Third, a MLP layer is adopted to fuse the concatenated features predict the mask *M* with a $\frac{H}{4} \times \frac{W}{4}$ resolution. Last, up-sampling mask to original size as change detection result of two input images.

On a pair of bi-temporal images, the changed pixels represent only a small fraction of all pixels. We use the Balanced Cross Entropy loss function to balance the ratio of positive and negative samples. The loss function is defined as follows:

$$\begin{aligned} Loss = -\sum (\alpha * p * log q \\ + (1-\alpha)(1-p) log(1-q)) \end{aligned} \tag{1}$$

In the above, $\alpha \in [0, 1]$ is a weighting factor, α for class 1 and $1 - \alpha$ for class -1. $q \in \pm 1$ specifies the ground-truth class and $p \in [0, 1]$ is the model's estimated probability for the class with label *q=1* .

4. EXPERIMENTS

We use the WuHan University change detection datasets(WHU-CD)[12] and SenseEarth change detection datasets to validate the effectiveness of the proposed method. WHU-CD is a public change detection dataset for buildings, it contains one pair of images of size 32507 × 15354 pixels. But the dataset has no data split solution provided, we crop the images into small patches of size 256×256 and split it into train, validation, test part, 6245 patches for train, 446 patches for validation, and 743 patches for test. SenseEarth change detection datasets(SE-CD) is SenseTime's change detection dataset, it contains 2968 pair of optical satellite images of 512 × 512 pixels. Also, SE-CD has no data split solution provided, we split it into 2494/178/296 for training/validation/test.

We trained the model using the Adam optimizer for 200 epochs on WHU-CD and SE-CD datasets, and the learning rate is set to 10e-3 at the beginning of training, we keep the learning rate at the first 100 epochs, which is then divided by 6 every 20 epochs. We use horizontal and vertical image flipping, 45° and 90° rotation form of data augmentation. Validation set is tested after each training epoch, and the best model on the validation set is used for evaluating the model on test set. The proposed model is implemented using the PyTorch framework on two Nvidia GeForce GTX 1080Ti.

We use precision, recall, intersection over union (IoU),overall accuracy (OA) and the F1-score with regard to the change category as the evaluation metrics.

$$\begin{aligned} Precision &= TP/(TP+FP) \\ Recall &= TP/(TP+FN) \\ IOU &= TP/(TP+FN+FP) \\ F1 &= (2*TP)/(2*TP+FP+FN) \end{aligned} \tag{2}$$

Where TP represents the number of true positive pixels, FP represents the number of false positive pixels, TN is the num-

Model	Precision	Recall	IOU	F1
FC-EF	0.732	0.872	0.781	0.796
FC-Siam-Conc	0.700	0.878	0.763	0.779
STANet	0.887	0.922	0.890	0.904
Ours	**0.903**	**0.932**	**0.904**	**0.917**

Table 1: quantitative results on WHU-CD

Model	Precision	Recall	IOU	F1
FC-EF	0.587	0.649	0.634	0.617
FC-Siam-Conc	0.493	0.738	0.590	0.591
STANet	0.600	**0.783**	0.669	0.679
Ours	**0.611**	0.778	**0.673**	**0.684**

Table 2: quantitative results on SE-CD

ber of true negative pixels, and FN is the number of false negative pixels.

We evaluate the performance of our proposed method on the WHU-CD and SE-CD datasets and make a comparison to several previous optical satellite image change detection methods: fully convolutional early fusion (FC-EF), FC-Siam-Conc and STANet.The results obtained as shown in Fig.2. And Table 1 is quantitative results of the different methods on WHU-CD. Table 2 is the quantitative results of the different methods on SE-CD. According to the table we can see, our proposed method is ahead of other comparison methods in all metrics on WHU-CD dataset. In SE-CD dataset, our method is slightly lower than STANet on recall. These results show that our method achieves better performance in quantitative evaluation.

5. CONCLUSION

In this paper, we propose a novel Siamese network based on Transformer for optical satellite image change detection. This model can detect the change of a whole image directly. Transformer encoder can exploits the global context information, and the patch merging operation can extract multi-scale features which is helpful for change decision. Then, the lightweight decoder is designed to improving efficiency. The experimental results is performed in two open datasets. However, the number of parameters is still large compare to CNN-based models. In the future, how to reduce the number of parameters and improve effectiveness of the model in detecting is under consideration.

6. REFERENCES

[1] X. Huang, L. Zhang, T. Zhu, "Building change detection from multitemporal high-resolution remotely sensed images based on a morphological building index". *IEEE J. Sel. Top. Appl. Earth Obs. Remote. Sens.*, pp.105–115, 2013.

[2] , M. Gong, J. Zhao, J. Liu, Q. Miao, L. Jiao, "Change detection in synthetic aperture radar images based on deep neural networks". *IEEE Trans. Neural Netw. Learn. Syst.*, pp.125–138, 2015.

[3] Ashbindu Singh, "Review article digital change detection techniques using remotely-sensed data," *International journal of remote sensing*, vol. 10, no. 6, pp.989–1003, 1989.

[4] Masroor Hussain, Dongmei Chen, Angela Cheng, Hui Wei, and David Stanley, "Change detection from remotely sensed images: From pixel-based to objectbased approaches," *ISPRS Journal of Photogrammetry and Remote Sensing*, vol. 80, pp.91–106, 2013.

[5] W. Zhang, X. Lu, "The Spectral-Spatial Joint Learning for Change Detection in Multispectral Imagery," *Remote Sens*, 11(3), 2019 .

[6] Q. Wang, X. Zhang, G. Chen, "Change detection based on Faster R-CNN for high-resolution remote sensing images," *Remote Sensing Letters*, pp.923-932, 2018.

[7] R. C. Daudt, B. Le. Saux, A. Boulch,"Fully convolutional siamese networks for change detection" *Proc. 25th IEEE Int. Conf. Image Process. (ICIP)*, pp. 4063–4067, 2018.

[8] H. Chen and Z. Shi, "A spatial-temporal attention-based method and a new dataset for remote sensing image change detection" *Remote Sens*, vol. 12, no. 10, pp.1662, 2020.

[9] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, et al, "An image is worth 16x16 words: Transformers for image recognition at scale," arXiv, 2020.

[10] Z. Liu, Y. Lin, Y. Cao, H. Hu, Y. Wei, Z. Zhang, S. Lin, B. Guo, "Swin transformer: Hierarchical vision transformer using shifted windows," arXiv, 2021.

[11] E. Xie, W. Wang, Z. Yu, A. Anandkumar, J. Alvarez, P. Luo, "SegFormer: Simple and Efficient Design for Semantic Segmentation with Transformers" arXiv, 2021.

[12] S. Ji, S. Wei, M. Lu, "Fully Convolutional Networks for Multi-Source Building Extraction from An Open Aerial and Satellite Imagery Dataset," *IEEE Transactions on geoscience and remote sensing*, 2018.