This article has been accepted for publication in a future proceedings of this conference, but has not been fully edited. Content may change prior to final publication.
Citation information: DOI: 10.1109/ICASSP49357.2023.10095819, ICASSP 2023 - 2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)

LOCAL-GLOBAL SIAMESE NETWORK WITH EFFICIENT INTER-SCALE FEATURE LEARNING FOR CHANGE DETECTION IN VHR REMOTE SENSING IMAGES

Yue Zhang[1], *Tao Lei*[1*], *Shaoxiong Han*[2], *Yetong Xu*[1], *Asoke K. Nandi*[3]

[1]Shaanxi Joint Laboratory of Artificial Intelligence, Shaanxi University of Science and Technology, Xi'an 710021, P. R. China
[2]Norinco Group Testing And Research institute, Weinan 714200, P. R. China
[3]Department of Electronic and Computer Engineering, Brunel University London, Uxbridge, Middlesex, UB8 3PH, United Kingdom

ABSTRACT

The popular networks for change detection (CD) in very-high-resolution (VHR) remote sensing (RS) images usually suffer from two problems. First, it is difficult for these networks to model simultaneously the local and global features of changed targets, which leads to the limited feature representation ability of popular CD networks. Second, these networks often have a large number of parameters and high computational costs due to complex network architecture. To address the above issues, we propose a local-global siamese network (LGS-Net) for CD in VHR RS images. First, we design an encoder with a parallel dual-branch structure consisting of convolutional neural networks (CNNs) and Transformer to extract rich features from bi-temporal images. Furthermore, we design a local-global feature enhancement (LGFE) module to help our encoder improve its feature representation ability. Second, we design a compact and efficient convolution module called inter-scale separable convolution (ISSConv). This module first divides feature maps into multiple groups, and then performs depthwise separable convolution in each group using atrous convolution with different dilation rates, which can not only capture changed targets across scales but also effectively reduce the number of model parameters. Experiments demonstrate that the proposed LGS-Net is superior to the state-of-the-art CD networks in terms of parameters, computational costs, and detection accuracy.

Index Terms— deep learning, change detection, local-global siamese network, feature enhancement, lightweight network

1. INTRODUCTION

Change detection (CD) in remote sensing (RS) images aims to identify the difference between two images from different periods but in the same area [1], it is of great significance in many fields, including disaster monitoring [2] and urban expansion [3]. The early methods for CD in VHR RS images mainly rely on manual feature extraction, these methods are usually sensitive to noise and therefore have low recognition accuracy and poor robustness. In recent years, with the rapid development of deep learning technology, a large amount of CD in VHR RS images methods based on deep learning have been proposed. These methods can be roughly divided into three categories: convolutional neural networks (CNNs)-based methods, Transformer-based methods, and methods based on the hybrid architecture of CNNs and Transformer.

CNNs have been widely used in various computer vision tasks due to their powerful feature extraction ability [4], but these methods still have some limitations. VHR RS images are usually very complex, and the changed targets with the same semantic information may be far away from each other. Due to the characteristics of the local receptive field of convolution kernels, it is difficult to employ CNNs to establish long-range dependency on different targets, which leads to the lack of long-range correlation information in feature maps and greatly limits the detection performance of networks. Recently, the transformer has been widely used in image classification and other tasks because it can effectively capture the long-range dependency of different targets [5] [6]. However, the output of the transformer at different stages is uniform and globally consistent, resulting in poor local-information extraction and high feature redundancy between shallow and deep layers. Therefore, the hybrid architecture based on CNNs and Transformer is proposed [7] [8]. Though the hybrid architecture is better than CNNs and Transformer, most of these hybrid architectures just sequentially combine CNNs and Transformer, or simply introduce the Transformer into CNNs, resulting in a lack of sufficient information interaction between CNNs and Transformer, which limits the feature representation ability of hybrid architecture networks. Besides, the existing hybrid architecture networks for CD tasks often employ vanilla convolution to extract image features,

*Corrsponding author: Tao Lei. Thanks to NSFC (62271296) and the Natural Science Basic Research Program of Shaanxi (No.2021JC-47).

This article has been accepted for publication in a future proceedings of this conference, but has not been fully edited. Content may change prior to final publication.
Citation information: DOI: 10.1109/ICASSP49357.2023.10095819, ICASSP 2023 - 2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)

which leads to a large number of parameters and computational costs [9]. Therefore, it is difficult to deploy a hybrid architecture network for CD tasks on low-resource devices.

To address the above problems, we design a local-global siamese network (LGS-Net) for CD tasks in VHR RS images. The main contributions of this work can be summarized as follows:

(1) we design a siamese encoder with a parallel dual-branch structure consisting of CNNs and Transformer, it can extract and interact with local-global information for bi-temporal images at the same time. Furthermore, we use an local-global feature enhancement (LGFE) module to enhance feature information and improve the feature expression ability. The encoder can fully integrate the local modeling advantages of CNNs and the global modeling advantages of Transformer, compared with mainstream CNNs and Transformer, hybrid structure have obvious advantages.

(2) We design an inter-scale separable convolution (ISSConv) module by utilizing the advantages of depthwise separable convolution, grouping convolution, and atrous convolution. It performs cross-scale feature extraction for bi-temporal images by grouping and introducing atrous convolution with a variable dilation rate, which minimizes feature redundancy and reduces computational costs.

(3) We design an LGS-Net for CD in VHR RS images. The experimental results show that the proposed network not only provides higher detection accuracy, but also requires less storage and computational costs compared with state-of-the-art networks.

2. THE PROPOSED NETWORK

2.1. Overall network structure

Different from normal image segmentation tasks, the input of CD is a pair of bi-temporal images. As shown in Fig.1, to fuse effectively bi-temporal image features, we build a local-global siamese network (LGS-Net) for CD tasks in VHR RS images. LGS-Net consists of an encoder and a decoder, and the encoder adopts a parallel dual-branch structure composed of CNNs and Transformer. It aims to extract local and global features of bi-temporal image simultaneously. In the CNNs branch, we replace the vanilla convolution with inter-scale separable convolution (ISSConv). In the Transformer branch, we employ PVT-v2-B1 [10] with a pyramid structure to capture multi-scale features with long-distance dependency.

Specifically, first, a pair of bi-temporal images are sent to the siamese encoder composed of Transformer block and ISSConv for local-global information extraction. Then the outputs from the Transformer block and ISSConv are efficiently integrated by a local-global feature enhancement (LGFE) module. The decoder consists of a deconvolution upsampling layer and ISSConv. At each stage of the encoder, we acquire a different image and connect it to the corresponding position of the decoder to obtain richer feature maps of changed targets. Finally, the network performs dimensionality reduction and normalization operations by using point convolution to output the final CD results.

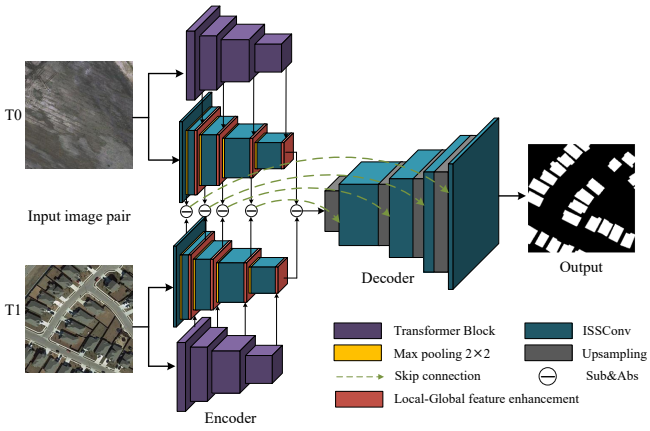

Fig. 1. The overall structure of LGS-Net.

According to Fig. 1, we can see that our proposed LGS-Net differs from other popular CD networks in two aspects. First, our proposed LGS-Net adopts the latest hybrid architecture of CNNs and Transformer. It fuses the local features from CNNs branch and the global features from Transformer branch at each layer. Furthermore, it uses a feature enhancement module to improve feature representation. Thus, our encoder achieves better feature representation than popular network architecture for CD in VHR RS images. Second, we employ a novel ISSConv that is superior to vanilla convolution and improved convolutions such as depthwise separable convolution and its variants. Thus, our proposed LGS-Net requires fewer parameters and computational costs than the popular hybrid architecture of CNNs and Transformer.

2.2. Local and global feature representation

In CD of VHR RS images, as the receptive field of the convolution kernel is usually limited, the CNNs are unable to obtain the global feature of bi-temporal images efficiently. Although Transformer has a strong ability to capture global information [11], it is weak at obtaining local information. To effectively correlate local and global features simultaneously, we design a local-global siamese encoder for VHR RS images feature extraction. At the same time, in the encoder, we also design an LGFE module, which aims to efficiently integrate the output of ISSConv and Transformer. The specific structure of LGFE module is shown in Fig.2.

Channel attention operation is performed on the obtained local and global feature maps [12]. Let T_i and C_{i+1} be the output of the i-th layer of Transformer branch and CNNs branch respectively, where $i = 1, 2, 3, 4$. We first aggregate spatial information of a feature map by using average-pooling

This article has been accepted for publication in a future proceedings of this conference, but has not been fully edited. Content may change prior to final publication.
Citation information: DOI: 10.1109/ICASSP49357.2023.10095819, ICASSP 2023 - 2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)

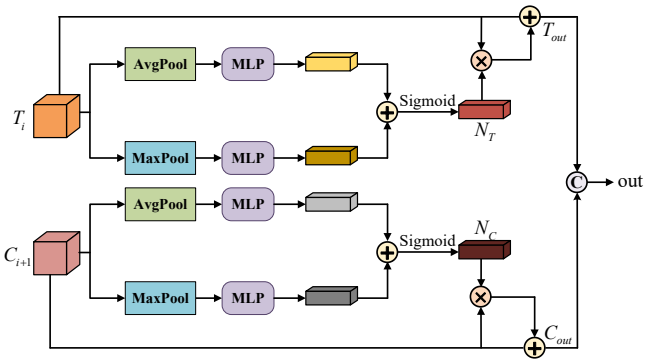

Fig. 2. The structure of LGFE module.

and max-pooling operations, then we use multi-layer perceptron (MLP) to obtain attention weights, and finally, use the element-wise summation to merge the output feature vectors. The specific channel attention operation is defined as:

$$N_T = \delta\left(MLP\left(AvgPool\left(T_i\right)\right) + MLP\left(MaxPool\left(T_i\right)\right)\right), \tag{1}$$

$$N_C = \delta\left(MLP\left(AvgPool\left(C_{i+1}\right)\right) + MLP\left(MaxPool\left(C_{i+1}\right)\right)\right), \tag{2}$$

where N_T and N_C denote the obtained attention map after channel attention operation, MLP represents the multi-layer perceptron network, $AvgPool$ and $MaxPool$ denote the average-pooling and max-pooling operations, respectively, and δ stands for the Sigmoid function.

To obtain the enhanced feature map T_{out} and C_{out}, the attention map is multiplied to the original feature map and then the obtained results is added to the original feature map. The specific operation is defined as:

$$T_{out} = N_T \times T_i + T_i, \tag{3}$$

$$C_{out} = N_C \times C_{i+1} + C_{i+1}. \tag{4}$$

Finally, the enhanced maps obtained from the two branches are concatenated to get the final feature map as follows:

$$out = Concat\left(T_{out}, C_{out}\right), \tag{5}$$

where $Concat$ represents the splicing of feature maps in the channel dimension.

Different from most of encoders used for CD in VHR RS images, our encoder does not simply connect CNNs and Transformer in series or parallel, but fully fuses the local features from CNNs branch and the global features from Transformer in each stage to realize feature complementation. Therefore, our encoder can achieve better feature representation than other encoders for CD in VHR RS images.

2.3. ISSConv for efficient feature learning

The existing CD methods usually have complex network structures leading to massive parameters and excessive computational costs. To solve this problem, we design a compact and efficient convolution module called ISSConv. Based on the depthwise separable convolution [13], ISSConv realizes the cross-scale feature expression for changed targets in the bi-temporal images. The ISSConv has lower computational costs and fewer parameters than vanilla convolution and improved convolutions because it groups the depthwise convolution and employs atrous convolution with variable dilation rates. The structure of ISSConv is shown in Fig. 3.

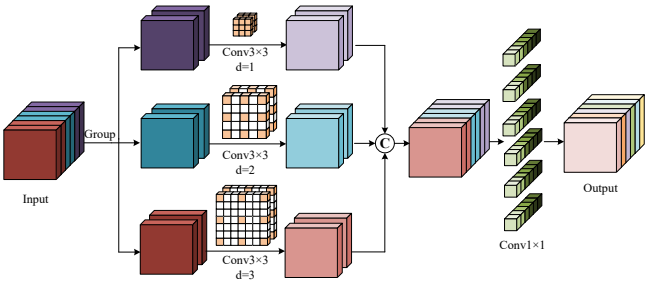

Fig. 3. The structure of ISSConv.

Specifically, first, the input feature map $X \in R^{C\times H\times W}$ is divided into g groups. Second, the depthwise convolution is performed on feature maps at each group, where the convolutional kernels are 3×3 with different dilation rates at each group. Third, feature maps at each group are concatenated to output a completed feature map. Finally, the point convolution is performed on the feature map to achieve cross-scale channel information exchange. Compared to other convolution operations used for CD of VHR RS images, such as depthwise separable convolution, atrous convolution, and grouping convolution, ISSConv not only realizes the cross-scale feature extraction but also enhances the compactness of feature maps.

Let $X \in R^{C\times H\times W}$ be an input feature map, where C, H and W are the number of channels, the height, and width of the feature map, and $C^{'}$ is the number of channels of the output feature map. For vanilla convolution, the number of parameters is denoted by P and the computational cost is denoted by Q, then

$$P = K \times K \times C \times C^{'}, \tag{6}$$

$$Q = K \times K \times C \times C^{'} \times H \times W. \tag{7}$$

For ISSConv, the number of parameters is denoted by P_m and the computational cost is denoted by Q_m, then

$$P_m = K \times K \times C + C \times C^{'}, \tag{8}$$

$$Q_m = K \times K \times C \times H \times W + C \times C^{'} \times H \times W. \tag{9}$$

Compared with the vanilla convolution, the proposed ISSConv can effectively reduce the number of parameters and

This article has been accepted for publication in a future proceedings of this conference, but has not been fully edited. Content may change prior to final publication.
Citation information: DOI: 10.1109/ICASSP49357.2023.10095819, ICASSP 2023 - 2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)

computational cost as follows:

$$r = \frac{K \times K \times C + C \times C^{'}}{K \times K \times C \times C^{'}} = \frac{1}{K^2} + \frac{1}{C^{'}}. \quad (10)$$

From (10), the parameters and computations of the proposed ISSConv are only $\left(1/K^2 + 1/C^{'}\right)$ of vanilla convolution. Meanwhile, the ISSConv enables the exchange of cross-scale features between different targets while effectively reduces the number of network parameters and computational costs.

3. EXPERIMENTS

To evaluate the performance of LGS-Net for CD task in VHR RS images, we select the LEVIR-CD dataset [14] and the CDD dataset [15] as experimental data. The LEVIR-CD dataset contains 637 VHR Google Earth image pairs with a resolution of 0.5m and a size of 1024×1024 pixels. In our experiments, we used 70% of the data as the training set, 10% as the validation set, and 20% as the testing set. They are cropped into 256×256 image pairs by random cropping. The CDD dataset includes RS images with seasonal changes in the same region obtained by Google Earth. A total of 16,000 image pairs of size 256×256 are obtained through random cropping and data enhancement, where 10,000 pairs are used for training, 3,000 pairs are for verifying, and the rest 3,000 pairs are for testing.

3.1. Training details

Experiments were performed on a server with NVIDIA GeForce RTX 3090 24GB and PyTorch 1.7. The number of training epoch is set to 200, and the Adam optimizer is used to optimize the model. In the training process, the learning rate is set to 0.0001, and the batch size is set to 32. We used binary cross-entropy loss to optimize the network weights.

3.2. Evaluation and results

In this experiment, we used Precision (*Pre*), Recall (*Rec*), and F1-Score (*F1*) to evaluate the experimental results. Table 1 shows the quantitative analysis of the results of our proposed method and other methods [2] [8] [9] [16] [17] [18] on the test set. It can be seen that our method performs better than other popular methods and *F1* achieves 90.36% and 90.25%. In particular, compared with BIT that directly uses CNNs and Transformer in series, our method is more excellent, and *F1* is improved by about 1% on both datasets.

To verify the effectiveness of different modules in our network, we conducted a series of ablation experiments on the LEVIR-CD dataset, as shown in Table 2, where we used Siam-UNet as our baseline network, and Transformer means that we introduced the local-global siamese encoder built by Transformer Block. In Table 2, it can be seen that the introduction of Transformer, LGFE, and ISSConv have improved the detection accuracy to a certain extent, and the combination of them can achieve higher detection accuracy with lower parameters.

Table 1. Quantitative analysis of different networks on the LEVIR-CD and the CDD, The best values are in bold.

Methods	LEVIR-CD			CDD		
	Pre(%)	Rec(%)	F1(%)	Pre(%)	Rec(%)	F1(%)
FC-Siam-diff [16]	84.44	86.38	85.40	61.85	76.69	68.48
FCN-PP [2]	82.09	84.48	83.27	88.14	84.22	86.14
DSIFN [9]	86.00	89.73	87.03	90.72	86.50	88.56
FDCNN [17]	83.87	87.56	85.68	87.90	86.99	87.44
MSTDSNet [18]	85.52	90.84	88.10	86.95	89.54	88.23
BIT [8]	89.24	89.37	89.31	88.91	89.68	89.29
LGS-Net(ours)	**89.41**	**91.32**	**90.36**	**90.73**	**89.78**	**90.25**

Table 2. Ablation experiments.

Backbone	Transformer	LGFE	ISSConv	Params(M)	F1(%)
Base(Siam-UNet)+				31.07	87.11
Base+	✓			51.62	89.51
Base+	✓	✓		51.84	89.91
Base+	✓	✓	✓	21.94	**90.36**

4. CONCLUTION

In this work, we propose an LGS-Net for CD in VHR RS images. First, LGS-Net uses a parallel dual-branch structure consisting of ISSConv and Transformer to extract rich features from bi-temporal images, and afterward, it uses the LGFE module to enhance the local-global feature information, fully integrate the advantages of CNNs and Transformer, and enhance the network feature expression ability. Second, ISSConv is proposed to extract cross-scale features of targets with low memory usage and computational costs. The experiments are performed on two popular CD datasets, including LEVIR-CD and CDD, and the results show that our proposed LGS-Net is superior to the current popular CD networks.

At present, industrial deployments have become an important challenge for the practical applications of deep learning models. Therefore, in the subsequent research, we will further improve the detection accuracy of LGS-Net and continue to simplify it. We hope that in the future, LGS-Net can effectively address the challenge of deployment for CD in VHR RS images on low-resource devices. In addition, we will further explore the application of LGS-Net in multi-modal for CD tasks in VHR RS images, such as the application of CD in synthetic aperture radar and optical images.

5. REFERENCES

[1] Abu Sebastian, Tomas Tuma, Nikolaos Papandreou, Manuel Le Gallo, Lukas Kull, Thomas Parnell, and Evangelos Eleftheriou, "Temporal correlation detection using computational phase-change memory," *Nature Communications*, vol. 8, no. 1, pp. 1–10, 2017.

[2] Tao Lei, Yuxiao Zhang, Zhiyong Lv, Shuying Li, Shigang Liu, and Asoke K Nandi, "Landslide inventory mapping from bitemporal images using deep convolutional neural networks," *IEEE Geoscience and Remote Sensing Letters*, vol. 16, no. 6, pp. 982–986, 2019.

[3] Tao Lei, Jie Wang, Hailong Ning, Xingwu Wang, Dinghua Xue, Qi Wang, and Asoke K Nandi, "Difference enhancement and spatial–spectral nonlocal network for change detection in vhr remote sensing images," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 60, pp. 1–13, 2021.

[4] Youwei Pang, Xiaoqi Zhao, Tian-Zhu Xiang, Lihe Zhang, and Huchuan Lu, "Zoom in and out: A mixed-scale triplet network for camouflaged object detection," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2022, pp. 2160–2170.

[5] Sixiao Zheng, Jiachen Lu, Hengshuang Zhao, Xiatian Zhu, Zekun Luo, Yabiao Wang, Yanwei Fu, Jianfeng Feng, Tao Xiang, Philip HS Torr, et al., "Rethinking semantic segmentation from a sequence-to-sequence perspective with transformers," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2021, pp. 6881–6890.

[6] Anurag Arnab, Mostafa Dehghani, Georg Heigold, Chen Sun, Mario Lučić, and Cordelia Schmid, "Vivit: A video vision transformer," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2021, pp. 6836–6846.

[7] Jinsheng Fang, Hanjiang Lin, Xinyu Chen, and Kun Zeng, "A hybrid network of cnn and transformer for lightweight image super-resolution," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2022, pp. 1103–1112.

[8] Hao Chen, Zipeng Qi, and Zhenwei Shi, "Remote sensing image change detection with transformers," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 60, pp. 1–14, 2021.

[9] Chenxiao Zhang, Peng Yue, Deodato Tapete, Liangcun Jiang, Boyi Shangguan, Li Huang, and Guangchao Liu, "A deeply supervised image fusion network for change detection in high resolution bi-temporal remote sensing images," *ISPRS Journal of Photogrammetry and Remote Sensing*, vol. 166, pp. 183–200, 2020.

[10] Wenhai Wang, Enze Xie, Xiang Li, Deng-Ping Fan, Kaitao Song, Ding Liang, Tong Lu, Ping Luo, and Ling Shao, "Pvtv2: Improved baselines with pyramid vision transformer," *CoRR*, vol. abs/2106.13797, 2021.

[11] Xiaoyi Dong, Jianmin Bao, Dongdong Chen, Weiming Zhang, Nenghai Yu, Lu Yuan, Dong Chen, and Baining Guo, "Cswin transformer: A general vision transformer backbone with cross-shaped windows," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2022, pp. 12124–12134.

[12] Sanghyun Woo, Jongchan Park, Joon-Young Lee, and In So Kweon, "Cbam: Convolutional block attention module," in *Proceedings of the European conference on computer vision (ECCV)*, 2018, pp. 3–19.

[13] Mark Sandler, Andrew Howard, Menglong Zhu, Andrey Zhmoginov, and Liang-Chieh Chen, "Mobilenetv2: Inverted residuals and linear bottlenecks," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2018, pp. 4510–4520.

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

[15] MA Lebedev, Yu V Vizilter, OV Vygolov, VA Knyaz, and A Yu Rubis, "Change detection in remote sensing images using conditional adversarial networks.," *International Archives of the Photogrammetry, Remote Sensing & Spatial Information Sciences*, vol. 42, no. 2, 2018.

[16] Rodrigo Caye Daudt, Bertr Le Saux, and Alexandre Boulch, "Fully convolutional siamese networks for change detection," in *2018 25th IEEE International Conference on Image Processing (ICIP)*. IEEE, 2018, pp. 4063–4067.

[17] Min Zhang and Wenzhong Shi, "A feature difference convolutional neural network-based change detection method," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 58, no. 10, pp. 7232–7246, 2020.

[18] Fei Song, Sanxing Zhang, Tao Lei, Yixuan Song, and Zhenming Peng, "Mstdsnet-cd: Multiscale swin transformer and deeply supervised network for change detection of the fast-growing urban regions," *IEEE Geoscience and Remote Sensing Letters*, vol. 19, pp. 1–5, 2022.