A TRANSFORMER-BASED SIAMESE NETWORK FOR CHANGE DETECTION

Wele Gedara Chaminda Bandara, Vishal M. Patel

Johns Hopkins University, Baltimore, Maryland, USA.
{wbandar1, vpatel36}@jhu.edu

ABSTRACT

This paper presents a transformer-based Siamese network architecture (abbreviated by *ChangeFormer*) for Change Detection (CD) from a pair of co-registered remote sensing images. Different from recent CD frameworks, which are based on fully convolutional networks (ConvNets), the proposed method unifies hierarchically structured transformer encoder with Multi-Layer Perception (MLP) decoder in a Siamese network architecture to efficiently render multi-scale long-range details required for accurate CD. Experiments on two CD datasets show that the proposed end-to-end trainable *ChangeFormer* architecture achieves better CD performance than previous counterparts. Our code and pre-trained models are available at github.com/wgcban/ChangeFormer.

Index Terms— Change detection, transformer Siamese network, attention mechanism, multilayer perceptron, remote sensing.

1. INTRODUCTION

Change Detection (CD) aims to detect relevant changes from a pair of co-registered images acquired at distinct times [1]. The definition of *change* may usually vary depending on the application. The changes in man-made facilities (e.g., buildings, vehicles, etc.), vegetation changes, and environmental changes (e.g., polar ice cap melting, deforestation, damages caused by disasters) are usually regarded as relevant changes. A better CD model is the one that can recognize these relevant changes while avoiding complex *irrelevant changes* caused by seasonal variations, building shadows, atmospheric variations, and changes in illumination conditions.

The existing state-of-the-art (SOTA) CD methods are mainly based on deep convolutional networks (ConvNets) due to their ability to extract powerful discriminative features. Since it is essential to capture *long-range contextual information* within the spatial and temporal scope to identify relevant changes in multi-temporal images, the latest CD studies have been focused on increasing the *receptive field* of the CD model. As a result, CD models with stacked convolution layers, dilated convolutions, and attention mechanisms [2] (channel and spatial attention) have been proposed [3]. Even though the attention-based methods are effective in capturing global details, they struggle to relate long-range

details in space-time because they use attention to re-weight the bi-temporal features obtained through ConvNets in the channel and spatial dimension.

The recent success of *Transformers* (i.e., non-local self-attention) in Natural Language Processing (NLP) has led researchers in applying transformers in various computer vision tasks. Following the transformer design in NLP, different architectures have been proposed for various computer vision tasks, including image classification and image segmentation such as Vision Transformer (ViT), SEgmentation TRansformer (SETR), Vision Transformer using Shifted Windows (Swin), Twins [4] and SegFormer [5]. These Transformer networks have comparatively larger *effective receptive field (ERF)* than deep ConvNets - providing much stronger context modeling ability between any pair of pixels in images than ConvNets.

Although Transformer networks have a larger receptive field and stronger context shaping ability, very few works have been done on transformers for CD. In a more recent work [6], a transformer architecture is applied in conjunction with a ConvNet encoder (ResNet18) to enhance the feature representation while keeping the overall ConvNet-based feature extraction process in place. *In this paper, we show that this dependency on ConvNets is not necessary, and a hierarchical transformer encoder with a lightweight MLP decoder can work very well for CD tasks.*

2. METHOD

The proposed *ChangeFormer* network consists of three main modules as shown in Fig. 1: a hierarchical transformer encoder in a Siamese network to extract coarse and fine features of bi-temporal image, four feature difference modules to compute feature differences at multiple scales, and a lightweight MLP decoder to fuse these multi-level feature differences and predict the CD mask.

2.1. Hierarchical Transformer Encoder

Given an input bi-temporal image, the hierarchical transformer encoder generates ConvNet-like multi-level features with high-resolution coarse features and low-resolution fine-grained features required for the CD. Concretely, given a pre-change or post-change images of resolution $H \times W \times 3$,
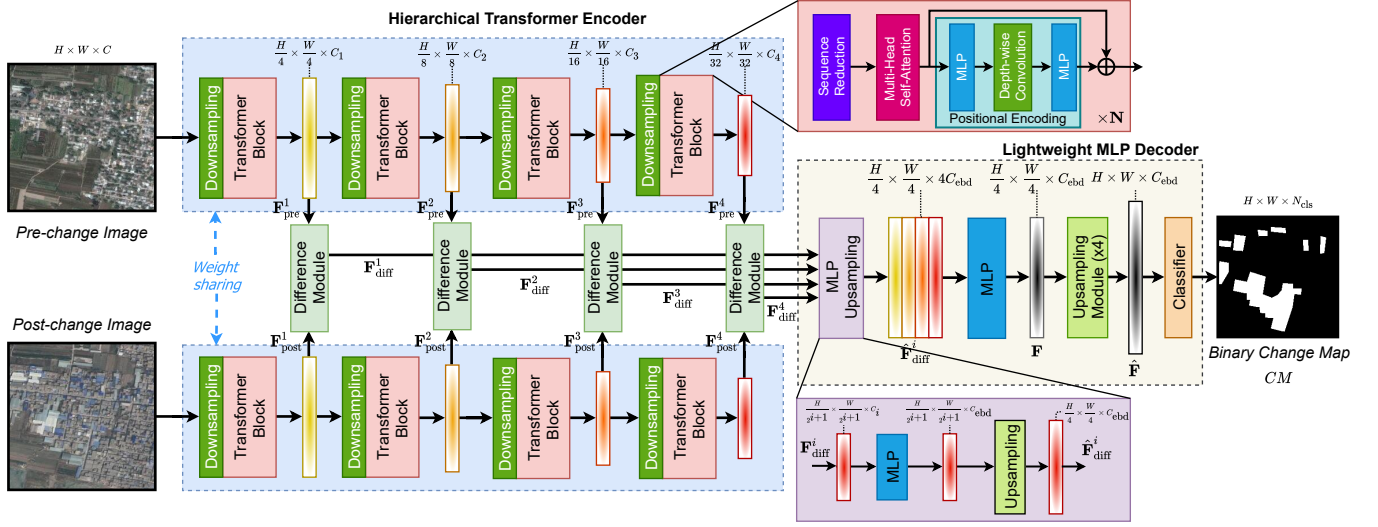
Fig. 1. The proposed *ChangeFormer* network for CD.

the transformer encoder outputs feature maps $\mathbf{F}_i$ with a resolution $\frac{H}{2^{i+1}} \times \frac{W}{2^{i+1}} \times C_i$, where $i = \{1, 2, 3, 4\}$ and $C_{i+1} > C_i$ which will be further processed through the difference modules followed by MLP decoder to obtain the change map.

Transformer Block. The main building block of the transformer encoder is *self-attention* module. In the original work [7], self-attention is estimated as:

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

where $\mathbf{Q}$, $\mathbf{K}$, and $\mathbf{V}$ denote Query, Key, and Value, respectively, and have the same dimensions of $HW \times C$. However, the computational complexity of eqn. (1) is $O((HW)^2)$ which prohibits its application on high-resolution images. To reduce the computational complexity of eqn. (1), we adopt the *Sequence Reduction* process introduced in [8] which utilizes reduction ratio R to reduce the length of the sequence HW as follows:

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

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

where $\mathbf{S}$ denotes the sequence to be reduced i.e., $\mathbf{Q}$, $\mathbf{K}$, and $\mathbf{V}$, $\text{Reshape}(h, w)$ denotes tensor reshaping operation to the one with shape of (h, w), and $\text{Linear}(C_{\text{in}}, C_{\text{out}})$ denotes a linear-layer with C_{in} input channels and C_{out} output channels. This results in a new set of $\mathbf{Q}$, $\mathbf{K}$, and $\mathbf{V}$ of size $\left(\frac{HW}{R}, C\right)$, hence reduces the computational complexity of eqn. (1) to $O((HW)^2/R)$.

To provide *positional information* for transformers, we utilize two MLP layers along with a 3×3 depth-wise convolutions as follows:

$$\mathbf{F}_{\text{out}} = \text{MLP}(\text{GELU}(\text{Conv2D}_{3\times3}(\text{MLP}(\mathbf{F}_{\text{in}})))) + \mathbf{F}_{\text{in}}, \quad (4)$$

where $\mathbf{F}_{\text{in}}$ are the features from self-attention, and GELU denotes Gaussian Error Linear Unit activation. Our positional encoding scheme differs from the fixed positional encoding utilized in previous transformer networks like ViT [9] which allows our *ChangeFormer* to take test images that are different in resolution from the ones used during training.

Downsampling Block. Given an input patch $\mathbf{F}_i$ from the i-th transformer layer of resolution $\frac{H}{2^{i+1}} \times \frac{W}{2^{i+1}} \times C_i$, downsampling layer shrink it to obtain $\mathbf{F}_{i+1}$ of resolution $\frac{H}{2^{i+2}} \times \frac{W}{2^{i+2}} \times C_{i+1}$ which will be the input to the $(i + 1)$-th Transformer layer. To achieve this, we utilize a 3×3 Conv2D layer with kernel size $K = 7$, stride $S = 4$, and padding $P = 3$ for the initial downsampling, and $K = 3$, $S = 2$, and $P = 1$ for the rest.

Difference Module. We utilize four Difference Modules to compute the difference of multi-level features of pre-change and post-change images from the hierarchical transformer encoder as shown in Fig. 1. More precisely, our Difference Module consists of Conv2D, ReLU, BatchNorm2d (BN) as follows:

$$\mathbf{F}^i_{\text{diff}} = \text{BN}(\text{ReLU}(\text{Conv2D}_{3\times3}(\text{Cat}(\mathbf{F}^i_{\text{pre}}, \mathbf{F}^i_{\text{post}})))), \quad (5)$$

where $\mathbf{F}^i_{\text{pre}}$ and $\mathbf{F}^i_{\text{post}}$ denote the feature maps of pre-change and post-change images from the i-th hierarchical layer, and Cat denotes the tensor concatenation. Instead of computing the absolute difference of $\mathbf{F}^i_{\text{pre}}$ and $\mathbf{F}^i_{\text{post}}$ as in [6], the proposed difference module learn the optimal distance metric at each scale during training - resulting in better CD performance.

2.2. MLP Decoder

We utilize a simple decoder with MLP layers that aggregates the multi-level feature difference maps to predict the change map. The proposed MLP decoder consists of three main steps.

MLP & Upsampling. We first process each multi-scale

208

feature difference map through an MLP layer to unify the channel dimension and then upsample each one to the size of $H/4 \times W/4$ as follows:

$$\tilde{\mathbf{F}}^i_{\text{diff}} = \text{Linear}(C_i, C_{\text{ebd}})(\mathbf{F}^i_{\text{diff}}) \forall i, \qquad (6)$$

$$\hat{\mathbf{F}}^i_{\text{diff}} = \text{Upsample}((H/4, W/4), \text{``bilinear''})(\tilde{\mathbf{F}}^i_{\text{diff}}), \quad (7)$$

where C_{ebd} denotes the embedding dimension.

Concatenation & Fusion. The upsampled feature difference maps are then concatenated and fused through an MLP layer as follows:

$$\mathbf{F} = \text{Linear}(4C_{\text{ebd}}, C_{\text{ebd}})(\text{Cat}(\hat{\mathbf{F}}^1_{\text{diff}}, \hat{\mathbf{F}}^2_{\text{diff}}, \hat{\mathbf{F}}^3_{\text{diff}}, \hat{\mathbf{F}}^4_{\text{diff}})). \quad (8)$$

Upsampling & Classification. We upsample the fused feature map $\mathbf{F}$ to the size of $H \times W$ by utilizing a 2D transposed convolution layer with $S = 4$ and $K = 3$. Finally, the upsampled fused feature map is processed through another MLP layer to predict the change mask **CM** with a resolution of $H \times W \times N_{\text{cls}}$, where N_{cls} (=2) is the number of classes i.e., *change* and *no-change*. This process can be formulated as follows:

$$\hat{\mathbf{F}} = \text{ConvTranspose2D}(S = 4, K = 3)(\mathbf{F}), \quad (9)$$

$$\mathbf{CM} = \text{Linear}(C_{\text{ebd}}, N_{\text{cls}})(\hat{\mathbf{F}}). \qquad (10)$$

3. EXPERIMENTAL SETUP

Datasets. We use two publically available CD datasets for our experiments, namely LEVIR-CD [10] and DSIFN-CD [11]. The LEVIR-CD is a building CD dataset that contains RS image pairs of resolution 1024×1024. From these images, we crop non-overlapping patches of size 256×256 and randomly split them into three parts to make train/val/test sets of samples 7120/1024/2048. The DSIFN dataset is an general CD dataset that contains the changes in different land-cover objects. For experiments, we create non-overlapping patches of size 256×256 from the 512×512 images while utilizing the authors' default train/val/test sets. This results in 14400/1360/192 samples for training/val/test, respectively, for the DSIFN dataset.

Implementation Details. We implemented our model in PyTorch and trained using an NVIDIA Quadro RTX 8000 GPU. We randomly initialize the network. During training, we applied data augmentation through random flip, random re-scale (0.8-1.2), random crop, Gaussian blur, and random color jittering. We trained the models using the Cross-Entropy (CE) Loss and `AdamW` optimizer with weight decay equal to 0.01 and beta values equal to (0.9, 0.999). The learning rate is initially set to 0.0001 and linearly decays to 0 until trained for 200 epochs. We use a batch size of 16 to train the model.

Performance Metrics. To compare the performance of our model with SOTA methods, we report F1 and Intersection over Union (IoU) scores with regard to the *change-class* as the primary quantitative indices. Additionally, we report precision and recall of the change category and overall accuracy (OA).

4. RESULTS AND DISCUSSION

In this section, we compare the CD performance of our *ChangeFormer* with existing SOTA methods:

- **FC-EF** [12]: concatenates bi-temporal images and processes them through a ConvNet to detect changes.
- **FC-Siam-Di** [12]: is a feature-difference method, which extracts multi-level features of bi-temporal images from a Siamese ConvNet, and their difference is used to detect changes.
- **FC-Siam-Conc** [12]: is a feature-concatenation method, which extracts multi-level features of bi-temporal images from a Siamese ConvNet, and feature concatenation is used to detect changes.
- **DTCDSCN** [13]: is an attention-based method, which utilizes a dual attention module (DAM) to exploit the inter-dependencies between channels and spatial positions of ConvNet features to detect changes.
- **STANet** [14]: is an another Siamese-based spatial-temporal attention network for CD.
- **IFNet** [15]: is a multi-scale feature concatenation method, which fuses multi-level deep features of bi-temporal images with image difference features by means of attention modules for change map reconstruction.
- **SNUNet** [16]: is a multi-level feature concatenation method, in which a densely connected (NestedUNet) Siamese network is used for change detection.
- **BIT** [6]: is a transformer-based method, which uses a transformer encoder-decoder network to enhance the context-information of ConvNet features via semantic tokens followed by feature differencing to obtain the change map.

Table 1 presents the results of different CD methods on the test-sets of LEVIR-CD [10] and DSIFN-CD [11]. As can be seen from the table, the proposed *ChangeFormer* network achieves better CD performance in terms of F1, IoU, and OA metrics. In particular, our *ChangeFormer* improves previous SOTA in F1/IoU/OA by 1.2/2.2/0.1% and 20.0/44.3/6.4% for LEVIR-CD and DSIFN-CD, respectively. In addition, Fig. 2 compares the visual quality of different SOTA methods on test images from LEVIR-CD and DSIFN-CD. As highlighted in red, our ChangeFormer captures much finer details compared to the other SOTA methods. These quantitative and qualitative comparisons show the superiority of our proposed CD method over the existing SOTA methods.

5. CONCLUSION

In this paper, we proposed a transformer-based Siamese network for CD. By utilizing a hierarchical transformer encoder in a Siamese architecture with a simple MLP decoder, our method outperforms several other recent CD methods that employ very large ConvNets like ResNet18

Table 1. The average quantitative results of different CD methods on LEVIR-CD [10] and DSIFN-CD [11].*

Method	LEVIR-CD [10]					DSIFN-CD [11]				
	Precision	Recall	F1	IoU	OA	Precision	Recall	F1	IoU	OA
FC-EF [12]	86.91	80.17	83.40	71.53	98.39	72.61	52.73	61.09	43.98	88.59
FC-Siam-Di [12]	89.53	83.31	86.31	75.92	98.67	59.67	65.71	62.54	45.50	86.63
FC-Siam-Conc [12]	91.99	76.77	83.69	71.96	98.49	66.45	54.21	59.71	42.56	87.57
DTCDSCN [13]	88.53	86.83	87.67	78.05	98.77	53.87	77.99	63.72	46.76	84.91
STANet [14]	83.81	91.00	87.26	77.40	98.66	67.71	61.68	64.56	47.66	88.49
IFNet [15]	94.02	82.93	88.13	78.77	98.87	67.86	53.94	60.10	42.96	87.83
SNUNet [16]	89.18	87.17	88.16	78.83	98.82	60.60	72.89	66.18	49.45	87.34
BIT [6]	89.24	89.37	89.31	80.68	98.92	68.36	70.18	69.26	52.97	89.41
ChangeFormer (ours)	92.05	88.80	90.40	82.48	99.04	88.48	84.94	86.67	76.48	95.56

*All values are reported in percentage (%). Color convention: **best**, **2nd-best**, and **3rd-best**.
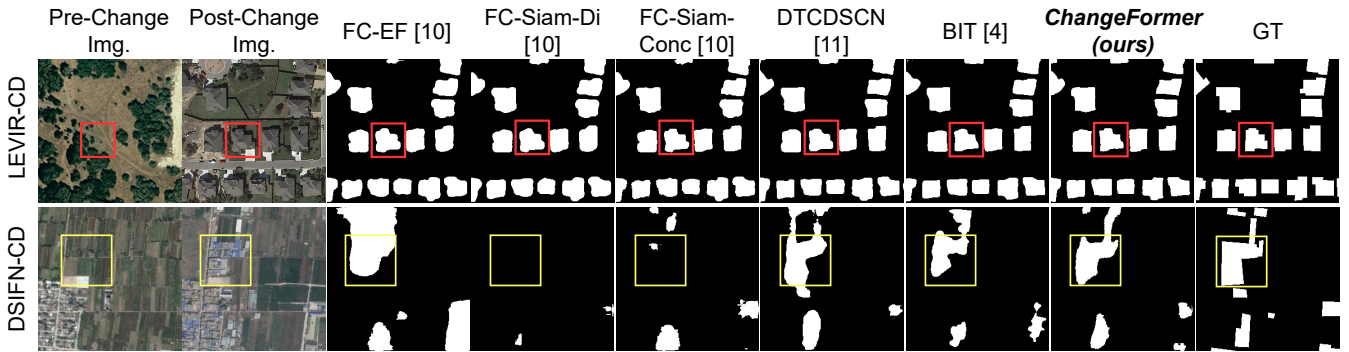
Fig. 2. Qualitative results of different CD methods on LEVIR-CD [10] and DSIFN-CD [11].

and U-Net as the backbone. We also show better performance in terms of IoU, F1 score, and overall accuracy than recent ConvNet-based (FC-EF, FC-Siam-DI, and FC-Siam-Conc), attention-based (DTCDSCN, STANet, and IFNet), and ConvNet+Transformer-based (BIT) methods. Hence, this study shows that it is unnecessary to depend on deep-ConvNets, and a hierarchical transformer in a Siamese network with a lightweight decoder can work very well for CD.

6. ACKNOWLEDGMENT

7. REFERENCES

[1] Wele Gedara Chaminda Bandara and Vishal M Patel, "Revisiting consistency regularization for semi-supervised change detection in remote sensing images," *arXiv preprint arXiv:2204.08454*, 2022.

[2] Wele Gedara Chaminda Bandara, Jeya Maria Jose Valanarasu, and Vishal M Patel, "Spin road mapper: Extracting roads from aerial images via spatial and interaction space graph reasoning for autonomous driving," *arXiv preprint arXiv:2109.07701*, 2021.

[3] Qian Shi, Mengxi Liu, Shengchen Li, Xiaoping Liu, Fei Wang, and Liangpei Zhang, "A deeply supervised attention metric-based network and an open aerial image dataset for remote sensing change detection," *IEEE Transactions on Geoscience and Remote Sensing*, 2021.

[4] Salman Khan, Muzammal Naseer, Munawar Hayat, Syed Waqas Zamir, Fahad Shahbaz Khan, and Mubarak Shah, "Transformers in vision: A survey," *arXiv preprint arXiv:2101.01169*, 2021.

[5] Enze Xie, Wenhai Wang, Zhiding Yu, Anima Anandkumar, Jose M Alvarez, and Ping Luo, "Segformer: Simple and efficient design for semantic segmentation with transformers," *arXiv preprint arXiv:2105.15203*, 2021.

[6] Hao Chen, Zipeng Qi, and Zhenwei Shi, "Remote sensing image change detection with transformers," *IEEE Transactions on Geoscience and Remote Sensing*, 2021.

[7] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Lukasz Kaiser, and Illia Polosukhin, "Attention is all you need," in *Advances in neural information processing systems*, 2017, pp. 5998–6008.

[8] Wenhai Wang, Enze Xie, Xiang Li, Deng-Ping Fan, Kaitao Song, Ding Liang, Tong Lu, Ping Luo, and Ling Shao, "Pyramid vision transformer: A versatile backbone for dense prediction without convolutions," *arXiv preprint arXiv:2102.12122*, 2021.

[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 preprint arXiv:2010.11929*, 2020.

[10] 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.

[11] 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.

[12] 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.

[13] Yi Liu, Chao Pang, Zongqian Zhan, Xiaomeng Zhang, and Xue Yang, "Building change detection for remote sensing images using a dual-task constrained deep siamese convolutional network model," *IEEE Geoscience and Remote Sensing Letters*, vol. 18, no. 5, pp. 811–815, 2020.

[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] 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.

[16] Sheng Fang, Kaiyu Li, Jinyuan Shao, and Zhe Li, "Snunet-cd: A densely connected siamese network for change detection of vhr images," *IEEE Geoscience and Remote Sensing Letters*, 2021.