TIME TRAVELLING PIXELS: BITEMPORAL FEATURES INTEGRATION WITH FOUNDATION MODEL FOR REMOTE SENSING IMAGE CHANGE DETECTION

Keyan Chen[1], *Chengyang Liu*[1], *Wenyuan Li*[2], *Zili Liu*[1,3], *Hao Chen*[3], *Haotian Zhang*[1], *Zhengxia Zou*[1], *Zhenwei Shi*[1,*]

[1]Beihang University, [2]University of Hong Kong, [3]Shanghai AI Laboratory

ABSTRACT

Change detection, a prominent research area in remote sensing, is pivotal in observing and analyzing surface transformations. Despite significant advancements achieved through deep learning-based methods, executing high-precision change detection in spatio-temporally complex remote sensing scenarios still presents a substantial challenge. The recent emergence of foundation models, with their powerful universality and generalization capabilities, offers potential solutions. However, bridging the gap of data and tasks remains a significant obstacle. In this paper, we introduce Time Travelling Pixels (TTP), a novel approach that integrates the latent knowledge of the SAM foundation model into change detection. This method effectively addresses the domain shift in general knowledge transfer and the challenge of expressing homogeneous and heterogeneous characteristics of multi-temporal images. The state-of-the-art results obtained on the LEVIR-CD underscore the efficacy of the TTP. The Code is available at `https://kychen.me/TTP`.

Index Terms— Remote sensing, change detection, foundation model, efficient tuning, bitemporal modeling

1. INTRODUCTION

As remote sensing technology for earth observation continues to evolve, remote sensing image change detection has surged to the forefront of research in this field. The primary objective is to analyze the changes of interest within multi-temporal remote sensing products. These changes are typically expressed as pixel-level binary classifications (changed/unchanged). The dynamic attributes of remote sensing surfaces are influenced not only by natural elements but also by human activities. The precise perception of these changes holds immense significance for the quantitative analysis of land cover alterations. This serves as a potent tool for illustrating macroeconomic trends, human activities, and climate changes. The invaluable application is apparent across various domains, encompassing urban expansion, glacier melting, and the evaluation of economic crop yields [1–5].

High-resolution remote sensing images have emerged as a potent tool for intricate change detection. However, executing robust change detection in complex scenarios remains a formidable challenge [6, 7]. Change detection primarily concentrates on “effective changes” amidst “non-semantic changes” [1, 2]. Specifically, non-semantic changes instigated by atmospheric conditions, remote sensors, registration, and semantic changes that are irrelevant to downstream applications (“invalid changes”) should be disregarded. This presents considerable obstacles to precise change detection. Deep learning technology has made significant strides in the realm of change detection. For example, algorithms based on CNN can unveil robust features in changing areas with their strong feature extraction capabilities, achieving impressive performance in a variety of complex scenarios [5, 8]. Recently, methods anchored on Transformers have further accelerated the advancement of this field. Transformers can capture long-distance dependencies across the entire image, endowing the model with a global receptive field, and opening up new avenues for tasks like change detection that necessitate high-level semantic knowledge [1, 3]. Despite the remarkable success of these methods, their adaptability in complex and evolving spatiotemporal environments is still a considerable distance from practical application. Furthermore, as the model scale expands, the limited annotated data for change detection significantly curtails the potential of these models. While some strides have been made in self-supervised representation learning and simulated data generation, they still fall short in covering the diversity of remote sensing image scenarios caused by spatiotemporal variability. Nor can they propel the performance of large-parameter models across different scenes [7, 9].

The potent universality and adaptability of recent foundational models have been firmly established. These models are trained on vast quantities of data, thereby acquiring generalized knowledge and representations [10]. Foundational models in the visual domain, such as CLIP [11] and SAM [12], have been extensively investigated and utilized by researchers. These models are repositories of a wealth of general knowledge, enabling cross-domain transfer and sharing. This significantly diminishes the need for annotated data for specific tasks. However, current visual foundational models are primarily designed for natural images, which creates a domain gap when these models are employed for change detection tasks in remote sensing images [10]. Moreover, while most visual foundational models excel at comprehending single images, they often fall short in extracting homogeneity

[*] Corresponding author

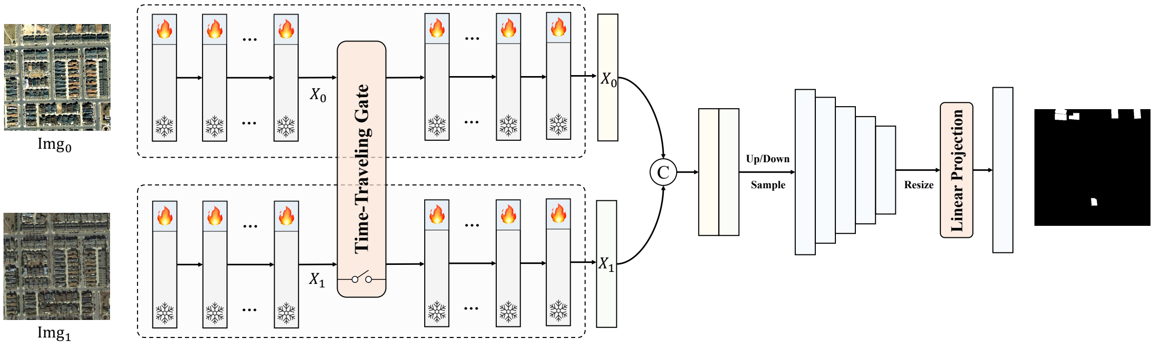

Fig. 1. The overview of the proposed TTP. The snowflake icon symbolizes that the model parameters are frozen, while the fire signifies training.

and heterogeneity from multiple images, particularly when significant changes occur in the images. This capability is crucial for change detection as it necessitates the model to concentrate solely on “effective changes”.

In this paper, we amalgamate the general knowledge of visual foundational models into the task of change detection. This approach overcomes the domain shift encountered during the knowledge transfer and the challenge of expressing the homogeneity and heterogeneity characteristics of multi-temporal images. We introduce Time Travelling Pixels, or TTP, a method that seamlessly integrates temporal information into the pixel semantic feature space. Specifically, TTP leverages the general segmentation knowledge based on the SAM (Segment Anything) model [12]. It introduces low-rank fine-tuning parameters into the SAM backbone to mitigate the domain shift of spatial semantics. Furthermore, TTP proposes a time-traveling activation gate that allows temporal features to permeate the pixel semantic space, thereby equipping the foundational model with the capacity to comprehend homogeneity and heterogeneity features between bitemporal images. Lastly, we devise a lightweight and efficient multi-level change prediction head to decode the dense high-level change semantic features. This innovative approach paves the way for more accurate and efficient change detection in remote sensing images.

The primary contributions of this paper can be encapsulated as follows: 1) We address the issue of insufficient annotated data by transferring the generalized latent knowledge of foundational models to the task of change detection. We introduce the Time Travelling Pixels (TTP) to bridge the time-space domain gap in the knowledge transfer process. 2) More specifically, we incorporate low-rank fine-tuning to mitigate the domain shift of spatial semantics, propose a time-traveling activation gate to augment the foundational model’s capacity to discern inter-image correlations and design a lightweight and efficient multi-level prediction head to decode the dense semantic information encapsulated in the foundational model. 3) We compare the proposed method with various advanced methods on the LEVIR-CD dataset. The results demonstrate that our method achieves state-of-the-art performance, underscoring its effectiveness and potential for further applications.

2. METHODOLOGY

2.1. Overview

To mitigate the annotation requirements of change detection, we leverage the general knowledge transferred from the foundational model. In this paper, we exploit the general segmentation capabilities of the SAM [12] to construct a change detection network, TTP. TTP is primarily composed of three components: a foundational model backbone based on low-rank fine-tuning; a time-traveling activation gate interposed between dual-temporal features; and an efficient multi-level decoding head. The structure is depicted in Fig. 1.

2.2. Efficient Fine-tuning of Foundation Model

The backbone of the SAM is comprised of transformer encoders, which can be categorized into base, large, and huge versions, corresponding to 12, 24, and 32 layers, respectively. To bolster computational efficiency, the majority of transformer layers in the backbone employ local attention, with only four layers utilizing global attention. In this study, we leverage the pre-trained, robust visual backbone, maintaining its parameters in a frozen state to expedite adaptation to downstream tasks. To bridge the gap between the domains of natural images and remote sensing images, we introduce low-rank trainable parameters into the multi-head attention layers, as demonstrated in the subsequent equation,

$$\begin{aligned} W^* &= W_0 + W_a W_b^T \\ Q = W_q^* X, K &= W_k^* X, V = W_v^* X \\ H &= \text{Softmax}(\frac{QK^T}{\sqrt{d}})V \end{aligned} \tag{1}$$

where $W_0 \in \mathbb{R}^{d \times d}$ signifies the original frozen model parameters, while $W_a \in \mathbb{R}^{d \times r}$ and $W_b \in \mathbb{R}^{d \times r}$, $r \ll d$ represent the additional fine-tuning parameters introduced. We incorporate low-rank fine-tuning in the linear projection layer of the self-attention matrix Q, K, V in each layer of the encoder. $X \in \mathbb{R}^{b \times n \times d}$ denotes the input features, and $H \in \mathbb{R}^{b \times n \times d}$ is the output following the self-attention operation.

2.3. Time-traveling Activation Gate

Current visual foundational models excel at interpreting the content of single images, yet they fall short in extracting homogenous and heterogeneous features from multiple images. However, in change detection, it is crucial for the model to concentrate on the “effective differences” in bi-temporal images while disregarding “irrelevant differences”. To tackle this, we introduce the time-traveling activation gate, which facilitates the flow of bi-temporal features into the pixel feature semantic space. This empowers the foundational model to comprehend the changes in bi-temporal images and focus on “effective changes”. For efficiency, we only incorporate the activation gate after the global attention layer in the backbone, *i.e.*, we only employ four bi-temporal time-traveling activation gates. Let’s consider $X_0 \in \mathbb{R}^{b\times c\times h\times w}$ and $X_1 \in \mathbb{R}^{b\times c\times h\times w}$ as the features of the previous and subsequent temporal phases, respectively. We follow the formula below to integrate bi-temporal information,

$$\begin{aligned} M &= \delta(\Phi^1_{\text{proj}}(\Phi_{\text{cat}}(X_0, X_1))) \\ X_0 &= X_0 + \Phi^2_{\text{proj}}(M \circ X_1) \\ X_1 &= X_1 + \Phi^2_{\text{proj}}(M \circ X_0) \end{aligned} \tag{2}$$

where Φ_{cat} symbolizes vector concatenation along the channel dimension, Φ^1_{proj} denotes linear channel compression, δ is a sigmoid activation function, and $\circ$ signifies pixel-wise multiplication. Φ^2_{proj} indicates linear mapping.

2.4. Multi-level Decoding Head

Remote sensing image scenes are diverse, and the scale of surfaces can vary significantly. However, visual encoders based on ViT typically generate feature map of a single scale. Despite the map containing high-level global semantic information, their performance advantages can be challenging to demonstrate without multi-level decoding heads. To address this, we introduce a lightweight and efficient multi-level change prediction head. This head constructs multi-level features through transposed convolution upsampling and max pooling downsampling. It then employs a lightweight MLP mapping layer to output the final change probability map,

$$\begin{aligned} \{F_i\} &= \Phi_{\text{sampling}}(\Phi_{\text{cat}}(X_0, X_1)) \\ F_i &= \Phi_{\text{resize}}(\Phi^1_{\text{proj}}(F_i)) \\ M &= \Phi^2_{\text{proj}}(\Phi_{\text{cat}}(\{F_i\})) \end{aligned} \tag{3}$$

where Φ_{sampling} signifies the feature maps of various levels generated by upsampling/downsampling, Φ^1_{proj} and Φ^2_{proj} represent the MLP mapping layer, and Φ_{resize} refers to applying bilinear interpolation to the features to unify the scale for concatenation.

3. EXPERIMENTS

3.1. Experimental Dataset and Settings

We carried out experiments on the LEVIR-CD to substantiate the efficacy of our method [5]. This dataset encompasses 637 pairs of bi-temporal images, each with a resolution of 1024×1024, and includes over 31,333 annotated instances of changes. We adhered to the official standards, partitioning the dataset into three subsets: training, validation, and testing, comprising 445, 64, and 128 image pairs, respectively.

3.2. Evaluation Protocol and Metrics

To assess the performance, we utilized widely recognized evaluation metrics, including Intersection over Union (IoU), F1 score, Precision, and Recall for the change category, as well as Overall Accuracy (OA) [1, 13].

3.3. Implementation Details

Architecture Details: TTP capitalizes on SAM’s visual backbone for the transfer of general knowledge. During the low-rank fine-tuning phase, we set $r = 16$. To guarantee efficiency in the decoding head, we limit upsampling to $\frac{1}{4}$ of the original image during supervised training.

Training Details: TTP employs a binary cross-entropy function for training. We set the model input size to 512×512 and utilize data augmentation techniques such as rotation, flipping, random cropping, and photometric distortion to enhance the sample size. During the training phase, the SAM backbone remains frozen. We utilize the AdamW optimizer with a learning rate of 0.0004 and a cosine annealing scheduler with a linear warmup to decay the learning rate. Our batch size is set to 16, and the maximum epoch is 300.

3.4. Comparison with the State-of-the-Art

We have compared the proposed TTP with a series of state-of-the-art change detection methods, including FC-Siam-Di [8], DTCDSCN [14], STANet [5], SNUNet [15], BIT [1], ChangeFormer [3], ddpm-CD [13], WNet [2], and CSTSUNet [4]. The comparative results are presented in Tab. 1. As illustrated in the table, the proposed TTP achieved the highest performance (92.1/85.6 F1/IoU), significantly surpassing the contemporary state-of-the-art methods, WNet (90.7/82.9) and CSTSUNet (90.7/83.0). This underscores that the transfer of general knowledge from the foundational model can bolster the effectiveness of change detection. It also validates the efficacy of the proposed transfer method.

Table 1. Comparative results on the LEVIR-CD dataset.

Method	P	R	F1	IoU	OA
FC-Siam-Di [8] (2018)	89.5	83.3	86.3	75.9	98.7
DTCDSCN [14] (2020)	88.5	86.8	87.7	78.1	98.8
STANet [5] (2020)	83.8	91.0	87.3	77.4	98.7
SNUNet [15] (2021)	89.2	87.2	88.2	78.8	98.8
BIT [1] (2021)	89.2	89.4	89.3	80.7	98.9
ChangeFormer [3] (2022)	92.1	88.8	90.4	82.5	99.0
ddpm-CD [13] (2022)	-	-	90.9	83.4	99.1
WNet [2] (2023)	91.2	90.2	90.7	82.9	99.1
CSTSUNet [4] (2023)	92.0	89.4	90.7	83.0	99.1
TTP (Ours)	**93.0**	**91.7**	**92.1**	**85.6**	**99.2**
TTP (w/o ttg)	92.2	90.3	91.1	84.2	99.1
TTP (w/o ttg, ml)	91.9	89.3	90.6	82.8	99.0
TTP (w/o ttg, ml, tuning)	80.9	69.3	74.6	59.5	97.6

3.5. Ablation Study

To thoroughly evaluate the effectiveness of each component, we conducted a series of ablation experiments on the LEVIR-CD dataset, adhering to the same training settings as TTP. As illustrated in Tab. 1, the performance experienced a decline when the time travel gate (ttg) and multi-level decoding head (ml) were removed. Moreover, the removal of the low-rank fine-tuning parameters in the foundational model led to a dramatic drop in performance. These observations underscore that the method proposed in this paper can effectively bridge the domain gap and enhance spatio-temporal understanding. They also validate the effectiveness of each component in the change detection task.

4. CONCLUSION

In this paper, we tackle the challenge of model generalization in complex spatiotemporal remote sensing scenarios by infusing the generic knowledge of foundational models into the task of change detection. Specifically, we introduce low-rank fine-tuning to bridge the spatial semantic chasm between natural and remote sensing images, thereby mitigating the limitations of the foundational model. We propose a time-travel activation gate to endow the foundational model with the capacity for temporal modeling. Additionally, we design a multi-level change prediction head to decode dense features. Experimental results on the LEVIR-CD dataset underscore the effectiveness of our proposed modules, with the proposed TTP achieving the best performance. This innovative approach paves the way for more accurate and efficient change detection in remote sensing images.

References

[1] 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.

[2] Xu Tang, Tianxiang Zhang, Jingjing Ma, Xiangrong Zhang, Fang Liu, and Licheng Jiao, “Wnet: W-shaped hierarchical network for remote sensing image change detection,” *IEEE Transactions on Geoscience and Remote Sensing*, 2023.

[3] Wele Gedara Chaminda Bandara and Vishal M Patel, “A transformer-based siamese network for change detection,” in *IGARSS 2022-2022 IEEE International Geoscience and Remote Sensing Symposium*. IEEE, 2022, pp. 207–210.

[4] Yaping Wu, Lu Li, Nan Wang, Wei Li, Junfang Fan, Ran Tao, Xin Wen, and Yanfeng Wang, “Cstsunet: A cross swin transformer based siamese u-shape network for change detection in remote sensing images,” *IEEE Transactions on Geoscience and Remote Sensing*, 2023.

[5] 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.

[6] Keyan Chen, Wenyuan Li, Sen Lei, Jianqi Chen, Xiaolong Jiang, Zhengxia Zou, and Zhenwei Shi, “Continuous remote sensing image super-resolution based on context interaction in implicit function space,” *IEEE Transactions on Geoscience and Remote Sensing*, 2023.

[7] Keyan Chen, Wenyuan Li, Jianqi Chen, Zhengxia Zou, and Zhenwei Shi, “Resolution-agnostic remote sensing scene classification with implicit neural representations,” *IEEE Geoscience and Remote Sensing Letters*, vol. 20, pp. 1–5, 2022.

[8] 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.

[9] Hao Chen, Haotian Zhang, Keyan Chen, Chenyao Zhou, Song Chen, Zhengxia Zou, and Zhenwei Shi, “Continuous cross-resolution remote sensing image change detection,” *IEEE Transactions on Geoscience and Remote Sensing*, 2023.

[10] Keyan Chen, Chenyang Liu, Hao Chen, Haotian Zhang, Wenyuan Li, Zhengxia Zou, and Zhenwei Shi, “Rsprompter: Learning to prompt for remote sensing instance segmentation based on visual foundation model,” *arXiv preprint arXiv:2306.16269*, 2023.

[11] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al., “Learning transferable visual models from natural language supervision,” in *International conference on machine learning*. PMLR, 2021, pp. 8748–8763.

[12] Alexander Kirillov, Eric Mintun, Nikhila Ravi, Hanzi Mao, Chloe Rolland, Laura Gustafson, Tete Xiao, Spencer Whitehead, Alexander C Berg, Wan-Yen Lo, et al., “Segment anything,” *arXiv preprint arXiv:2304.02643*, 2023.

[13] Wele Gedara Chaminda Bandara, Nithin Gopalakrishnan Nair, and Vishal M Patel, “Ddpm-cd: Remote sensing change detection using denoising diffusion probabilistic models,” *arXiv preprint arXiv:2206.11892*, 2022.

[14] 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.

[15] 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*, vol. 19, pp. 1–5, 2021.