MaCon: A Generic Self-Supervised Framework for Unsupervised Multimodal Change Detection

Jian Wang, Li Yan, Jianbing Yang, Hong Xie, Qiangqiang Yuan, Pengcheng Wei, Zhao Gao, Ce Zhang, and Peter M. Atkinson

***Abstract*—Change detection (CD) is important for Earth observation, emergency response and time-series understanding. Recently, data availability in various modalities has increased rapidly, and multimodal change detection (MCD) is gaining prominence. Given the scarcity of datasets and labels for MCD, unsupervised approaches are more practical for MCD. However, previous methods typically either merely reduce the gap between multimodal data through transformation or feed the original multimodal data directly into the discriminant network for difference extraction. The former faces challenges in extracting precise difference features. The latter contains the pronounced intrinsic distinction between the original multimodal data; direct extraction and comparison of features usually introduce significant noise, thereby compromising the quality of the resultant difference image. In this article, we proposed the MaCon framework to synergistically distill the common and discrepancy representations. The MaCon framework unifies mask reconstruction (MR) and contrastive learning (CL) self-supervised paradigms, where the MR serves the purpose of transformation while CL focuses on discrimination. Moreover, we presented an optimal sampling strategy in the CL architecture, enabling the CL subnetwork to extract more distinguishable discrepancy representations. Furthermore, we developed an effective silent attention mechanism that not only enhances contrast in output representations but stabilizes the training. Experimental results on both multimodal and monomodal datasets demonstrate that the MaCon framework effectively distills the intrinsic common representations between varied modalities and manifests state-of-the-art performance across both multimodal and monomodal CD. Such findings imply that the MaCon possesses the potential to serve as a unified framework in the CD and relevant fields. Source code will be publicly available once the article is accepted.**

***Index Terms*—Self-supervised learning, mask reconstruction, contrastive learning, multimodal data, change detection, unsupervised learning, remote sensing, Earth observation**

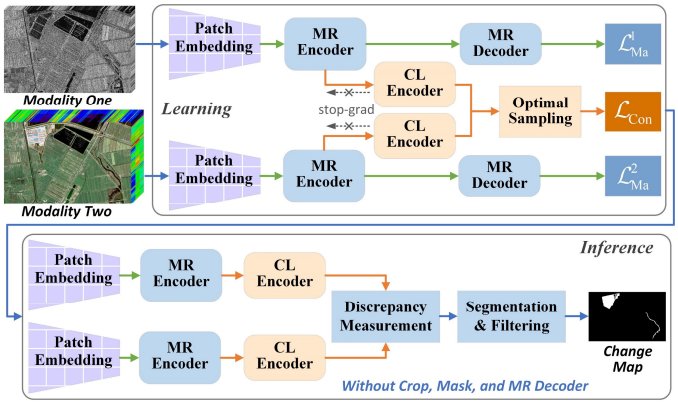


Fig. 1. Overview of the MaCon framework. The green and orange lines denote the flow of the MR and CL subnetworks, respectively. The gray dashed line represents stopping gradient backpropagation. The meaning of these symbols is the same in all figures. The CL encoder, MR encoder and decoder can be most mainstream architectures like CNN or Transformer.

I. Introduction

Change detection (CD) aims to characterize the information differences between multi-temporal images of the same area, and use these to identify the spatial changes[1], [2]. At present, CD is a common task in the field of image vision and perception [3], [4], as well as an important topic in remote sensing (RS) and Earth observation [5], [6]. Traditional CD is performed on monomodal images, namely, monomodal CD [7], [8], [9]. Recently, the amount of data in various modalities has increased rapidly with the development of different types and numbers of sensors and platforms [10], [11], [12], [13]. Researchers found that the availability and quality of monomodal data are often limited in specific scenarios. Additionally, many practical applications require fine temporal resolution, such as military reconnaissance, rescue and assessment of disasters, whereas acquiring multi-temporal

This research was supported in part by the National Natural Science Foundation of China under Grants 42394061 and 42371451 and the Science and Technology Major Project of Hubei Province under Grant 2021AAA010. The research was also supported in part by the Open Fund of Hubei Luojia Laboratory under Grant 220100053 and the State Scholarship Fund of China, which funded a one-year research visit of Jian Wang to Lancaster University through 2023-24. Jian Wang and Li Yan contributed equally. *(Corresponding author: Hong Xie).*

Jian Wang, Li Yan, Jianbing Yang, Hong Xie, Qiangqiang Yuan, and Pengcheng Wei are with the School of Geodesy and Geomatics, Hubei Luojia Laboratory, Wuhan University, Wuhan 430079, China. Jian Wang is also with the Faculty of Science and Technology, and Lancaster Environment Centre, Lancaster University, Lancaster, LA1 4YQ, UK (e-mail: wj_sgg@whu.edu.cn; lyan@sgg.whu.edu.cn; jbyang11@163.com; hxie@sgg.whu.edu.cn; qqyuan@sgg.whu.edu.cn; wei.pc@whu.edu.cn).

Zhao Gao is with the School of Computer Science, Wuhan University, Wuhan, 430079, China (e-mail: gaozzz@whu.edu.cn).

Ce Zhang is with the School of Geographical Sciences, University of Bristol, Bristol, BS8 1SS, UK. He is also with the UK Centre for Ecology & Hydrology, Lancaster, LA1 4AP, UK. (e-mail: ce.zhang@bristol.ac.uk).

Peter M. Atkinson is with the Faculty of Science and Technology, Lancaster University, Lancaster, LA1 4YQ, UK. He is also with Geography and Environmental Science, University of Southampton, Highfield, Southampton, SO17 1BJ, UK and the College of Surveying and Geo-Informatics, Tongji University, No.1239, Siping Road, Shanghai, PR China, 200092 (e-mail: pma@lancaster.ac.uk).

monomodal data requires a long period usually. In this context, multimodal change detection (MCD) brings obvious utility.

MCD identifies changes by comparing specific multitemporal images captured over the same geographical area at different times, but under varied conditions [14]. The MCD is an increasingly popular and challenging research topic, essentially representing a generalization of the monomodal CD problem [15], [16]. The input images could be acquired by different sensors, or recorded with different sensor parameters, or under different environmental conditions. The advantages of MCD are two-fold: first, it can increase the temporal resolution or extend the time span for time-series monitoring by inserting multimodal data; second, it is beneficial to shorten the response time of CD by relaxing the data acquisition requirements, which is imperative for emergency events [17].

Despite the above, MCD is a challenging task since multimodal data cannot be compared directly to obtain information about change differences as in monomodal CD. While numerous supervised methods exist for MCD [18], [19], [20], they reveal the following limitations. First, the expense of manual labeling is prohibitive, making it infeasible to label all scenes, particularly in the era of big data. Second, models trained on specific domains or geographical regions frequently lack generalization, making it hard to adaptively and accurately detect new targets in different environments or scenarios [21]. Third, both datasets and labels are scarce in the MCD task, and labels are usually subjective and may lack precision for scenes with intricate surface cover. As a result, the practical application of these supervised methods is constrained. Instead, unsupervised methods avoid these limitations as they do not require labels. For example, SFPPI [22] first generates a similarity-feature map and then fuses multiple binary segmentation results to output the final change map. PP [23], [24] computes differences between pixels in each image separately before generating the change map by comparing the difference scores.

At present, unsupervised MCD methods can be divided broadly into three classes: classification, transformation and discrimination. The classification methods first classify multimodal images. Subsequently, the derived classification outcomes can be compared directly to identify changes, such as the multidimensional evidential reasoning method, post-classification comparison method and compound classification method [25], [26], [27]. Since unsupervised classification models struggle to obtain accurate classification results, the classification methods are susceptible to the accumulation of classification errors.

In general, the core objective of transformation methods is to make the multimodal images comparable. Most transformation methods aim to either transfer “incomparable” images to a common domain or transform one image to the domain of another, thereby rendering them “comparable” [28]. In the transformation approach, the mappings of multimodal data typically demand training using unchanged pairs of multimodal data. Consequently, these transformation-based MCD methods either need pre-constructed pseudo labels [29], [30] or prior results [31] to guide the training. Alternatively, they may involve a complex iterative process to create the pseudo labels set concurrently with learning the mappings [32], [33]. Therefore, these methods can be deemed as pseudo or automatic supervised learning. Another limitation is that, after transformation, prevailing methods typically employ directly simple algorithms to extract the discrepancy information, such as difference [34], [35], [36], ratio [29], [37], distance function [30], [32], [34] or compound [14], [25], [38]. Regrettably, these methods are incompetent at deriving high-quality differences since neighbor information and further discrepancy enhancement are not utilized.

The discrimination approach is an emerging approach. It is intuitive and represented by self-supervised contrastive learning (CL) methods [39], which discriminate the characteristics between the dual stream outputs of the network by designing positive and negative pretext samples and loss function [9], [40]. However, there are two deficiencies in existing discrimination methods. First, the design of positive and negative samples is improper. They simply deem patch pairs at different locations as negative samples; positive samples are patch pairs at the same locations [9], [39], [40], [41]. However, patch pairs at different locations may be of the same class, and those at the same locations may be changed as well. Therefore, the obtained samples contain numerous exceptions. Second, existing methods input multimodal data directly into the CL network for learning [39], [40]. Since there are considerable distinctions between the original multimodal data, those methods are not conducive to accurate difference discrimination, thereby weakening the ability to extract differences.

In the context of the above shortcomings, we sought to couple the transformation and discrimination unsupervised methods and utilize both of their merits. The transformation module alleviates the domain deviation between multimodal data, and the discriminant module extracts refined differences. At the same time, we aimed to improve the sampling strategy to enable the network to distill better discrepancy representations. The main contribution of this paper can be summarized as follows:

1) A generic end-to-end self-supervised learning framework, namely MaCon. The MaCon innovatively coupled the mask reconstruction (MR) and CL architecture. Within this framework, the MR subnetwork distills the global information and transforms the multimodal data into a common domain, and the CL subnetwork extracts local information and discriminates the distinction between multimodal representations.

2) An optimal sampling strategy in the CL architecture. This strategy enhances the framework’s ability to learn more distinguishable object representations by utilizing more accurate samples, thereby enhancing the efficacy of subsequent change detection tasks.

3) A robust and plug-and-play attention mechanism. It can suppress features with low correlation, enhance contrast in output representations and stabilize training.

4) Experimental results on multiple multimodal and monomodal datasets show that the performance of the MaCon framework is better than the compared state-of-the-art (SOTA) methods, and even exceeds some supervised methods. The MaCon is expected to provide a unified framework for the task of CD.

The rest of this paper is structured as follows. Section II reviews the related work on self-supervised learning and its situation in CD task. Section III elaborates on the principle and algorithm of the MaCon framework. Section IV expounds on the experiments on multimodal and monomodal datasets. Section V analyzes the working mechanism of MaCon. Section VI draws the concluding remarks.

II. Related Work

Self-supervised learning focuses on various pretext tasks instead of the labels for pre-training, and they show a strong learning ability for representation and have seen significant interest in artificial intelligence [42], [43]. Currently, mainstream self-supervised learning can be broadly categorized into two paradigms: mask reconstruction (MR) and contrastive learning (CL) [44].

A. Mask Reconstruction

The MR self-supervised learning paradigm reserves a segment of the input sequence and trains models to forecast the masked content. Research has indicated their great performance and scalability, and evidence suggests that these pre-trained representations exhibit strong generalization across diverse downstream tasks [42], [45], [46].

In RS, all existing works using the MR paradigm employ MR as a foundation model for pre-training, followed by supervised fine-tuning for downstream CD. Wang et al. [47] made the first attempt to explore pre-training vision models tailored to RS tasks with large-scale RGB dataset. This work pre-trained multiple networks and tested transfer performance on CD task. After that, Sun et al. [48] developed an RS foundation model based on the MR self-supervised learning, RingMo, which is designed for dense and small objects in complicated RS scenes and training on massive monomodal datasets. The final change map can be generated by an appended CD head after fine-tuning on downstream CD dataset. To more effectively process RS spectral images, Hong et al. [49] created a large RS foundation model using a three-dimensional MR method with three-dimensional tokens to couple spatial-spectral information. This model was initially trained on one million RS optical images, and then, the CD task was performed by supervised retraining.

B. Contrastive Learning

The CL is an important paradigm in self-supervised learning. The core idea is to maximize the similarity between views augmented from the same image while minimizing the similarity between views augmented from distinct images [50], [51]. Numerous studies have shown that positive and negative samples are essential for CL, and the quantity and quality of negative samples generally determine the effectiveness of CL [50], [52], [53].

Some studies have adopted CL for CD. Akiva et al. [41] presented a material and texture-based method, which compares multi-temporal, spatially aligned large-scale multispectral images over unchanged regions to learn invariance to illumination and viewing angle as a mechanism to achieve consistency of material and texture representation. Then, the change map can be obtained by fine-tuning on a CD head. To leverage domain knowledge and characteristics of satellite images to learn better self-supervised features, Li et al. [54] proposed a geographical knowledge-driven CL method for multispectral RS images, which adopts global land cover (LC) products and geographical location associated with each RS image as geographical knowledge to provide supervision for network pre-training. Mall et al. [55] developed a new contrastive loss and used the temporal signal to contrast enormous RGB images with long- and short-term differences. Then, the CD task can be conducted with supervised retraining on specific datasets.

C. Limitations

Although the experimental results of the abovementioned methods show great performance on downstream CD task, the following limitations exist in these self-supervised methods.

1) It not only requires massive data and resources for pre-training but also necessitates retraining for downstream tasks, making it highly cost-intensive and difficult to develop extensively.

2) Only designed for monomodal datasets and are unsuitable for multimodal datasets.

3) Mounting a simple CD head in the downstream CD task makes extracting the great discrepancy representations challenging for high-performance change detection.

4) Some studies have shown that fine-tuning with labels can lead to catastrophic forgetting, and results are unsatisfactory when applied to new data with substantial changes in domain and distribution [56].

To deal with these limitations, we creatively couple MR and CL into a framework to leverage their strengths synergistically. Specifically, MR is designed to serve as a transformation mechanism for multimodal data, while CL is employed to effectively distinguish differences. Furthermore, instead of adopting a two-stage strategy with pre-training and fine-tuning, we trained and inferred end-to-end on the target multimodal CD dataset directly.

III. MaCon Framework

A. Framework Overview

The overview of the proposed framework is shown in Fig. 1. The framework comprises two pseudo-Siamese subnetworks, that is, MR and CL subnetworks, and all of them have two branches for multimodal images. Given two images from any modality, MaCon distills common representations and highlights discrepancy representations in the learning phase first, and then predicts the changed area in the inference phase.

The learning phase is the core of the MaCon framework. In the learning phase, the MaCon learns how to cross the gap between different modalities and differentiate the LC discrepancy from multi-temporal images in a self-supervised way. Next, we first expound the components and principles of the learning phase and then extend them to the inference phase. The dimension variation during the forward phase is annotated above all operation blocks to facilitate understanding.

B. Common Representations Learning with MR

The architecture of the MR subnetwork is shown in Fig. 2. The detailed forward pipeline of the MR subnetwork is as follows. Firstly, to extend and take full advantage of the data,

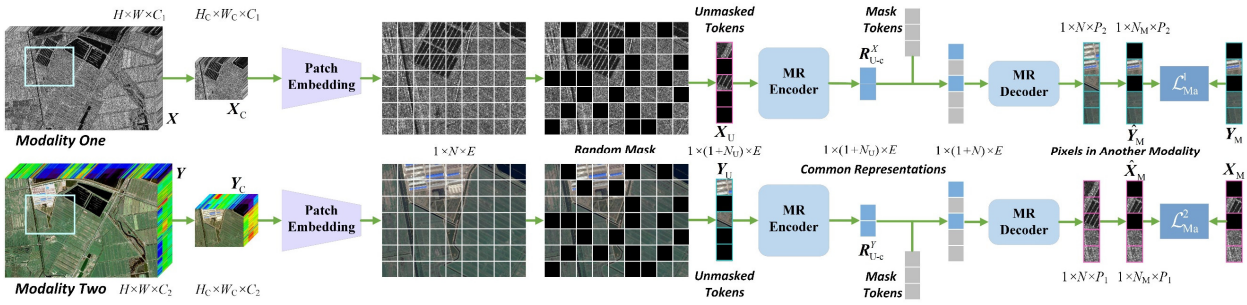


Fig. 2. The architecture of the MR subnetwork.

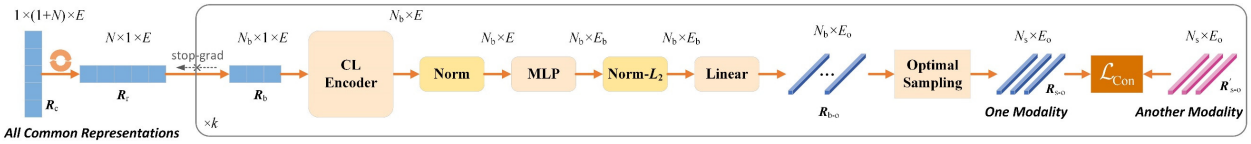


Fig. 3. The architecture of the CL subnetwork (single branch).

one adopts several data augmentation strategies to the normalized multimodal data $\boldsymbol{X}$ and $\boldsymbol{Y}$, including random crop with an $H_C \times W_C$ fixed-size window, and random horizontal or vertical flip. Next, the augmented data $\boldsymbol{X}_C$ and $\boldsymbol{Y}_C$ are tokenized to overlapping patches via a convolutional layer with a small stride, then flattened to a batch of embedded vectors with a dimension of E; this differs from other MR networks [45], [48], [57], [58] and is important for the proposed MaCon, especially for RS images, because it provides the universally fundamental embeddings for both MR and CL subnetworks and induces the patch tokens to assimilate more neighbor information. Notably, for clarity and to aid visual interpretation, the patches in Fig. 2 are illustrated in non-overlapping form.

Subsequently, to force the network to learn global information better and extract high-level semantic features, instead of just storing the pixel mapping relationship between multimodal data, the patch tokens are masked randomly with a fixed mask ratio β after prepending a learnable class token to the patch tokens. We recommend setting β in the range of [0.4, 0.7] based on numerical experiments to achieve optimal performance. Then, the unmasked patch tokens $\boldsymbol{X}_U$ and $\boldsymbol{Y}_U$ are fed into the MR encoder to distill the unmasked common representations $\boldsymbol{R}^X_{U\text{-}c}$ and $\boldsymbol{R}^Y_{U\text{-}c}$ of the multimodal data. Here, we fill the MR encoder with a series of Transformer blocks.

Next, the $\boldsymbol{R}^X_{U\text{-}c}$ and $\boldsymbol{R}^Y_{U\text{-}c}$ are concatenated with the masked tokens, and they are restored to their original patch positions. Then, they are decoded to reconstruct the masked pixels of another modality ($\hat{\boldsymbol{Y}}_M$ and $\hat{\boldsymbol{X}}_M$) through the MR decoder. Here, the MR decoder is composed of Transformer blocks and a linear predictor.

Finally, the loss can be evaluated by the mean squared L_2 distance between the reconstructed ($\hat{\boldsymbol{Y}}_M$ and $\hat{\boldsymbol{X}}_M$) and actual normalized pixels in another modality ($\boldsymbol{Y}_M$ and $\boldsymbol{X}_M$). Additionally, we attach a regularization term to reduce overfitting by penalizing the magnitude of the network parameters. Thus, the MR loss $\mathcal{L}_{Ma}(\theta_1,\theta_2)$ is defined as

$$\mathcal{L}_{Ma}(\theta_1,\theta_2) = \mathcal{L}^1_{Ma}(\theta_1) + \mathcal{L}^2_{Ma}(\theta_2), \tag{1}$$

where θ_1 and θ_2 are the parameters of the two MR branches, $\mathcal{L}^1_{Ma}(\theta_1)$ and $\mathcal{L}^2_{Ma}(\theta_2)$ are the MR losses of two MR branches and derived as

$$\begin{cases} \mathcal{L}^1_{Ma}(\theta_1) = \mathbb{E}\left\|\hat{\boldsymbol{Y}}_M(\theta_1) - \boldsymbol{Y}_M(\theta_1)\right\|_2^2 + \lambda_1\|\theta_1\|_2 \\ \mathcal{L}^2_{Ma}(\theta_2) = \mathbb{E}\left\|\hat{\boldsymbol{X}}_M(\theta_2) - \boldsymbol{X}_M(\theta_2)\right\|_2^2 + \lambda_2\|\theta_2\|_2 \end{cases}, \tag{2}$$

where the λ_1 and λ_2 are the regularization coefficients, determined automatically by the optimization algorithm. With this pipeline, the MR encoder will learn the common representations with reduced domain bias in multimodal data.

C. Discrepancy Representations Learning with Optimized CL

The architecture of the CL subnetwork is illustrated in Fig. 3. Due to the architecture being general for different modalities and space limitations, Fig. 3 takes a single branch as an example. The CL subnetwork is responsible for learning local information and high-level semantics, and generates discrepancy representations.

As shown in Fig. 3, after the MR encoder outputs the common representations $\boldsymbol{R}_c$ for all patch tokens, the $\boldsymbol{R}_c$ are shuffled and permuted into $\boldsymbol{R}_r$ with shape $N \times B \times E$ (*length* × *batch size* × *embeddings*), and where B is 1, that is, $\boldsymbol{R}_r = \mathrm{Pm}(\mathrm{Sf}(\boldsymbol{R}_c))$, where $\mathrm{Sf}(\cdot)$ and $\mathrm{Pm}(\cdot)$ denote shuffle and permutation operations, respectively. Then, a loop is run over the dimension N of $\boldsymbol{R}_r$ with a mini-batch step of N_b, which means the loop will iterate the following operations for k times.

1) Sample mini-batch representations $\boldsymbol{R}_b$ from the $\boldsymbol{R}_r$ and prepend a class token to the $\boldsymbol{R}_b$, i.e., $\boldsymbol{R}_b = \mathrm{Pp}(\mathrm{Sp}(\boldsymbol{R}_r))$, where $\mathrm{Sp}(\cdot)$ and $\mathrm{Pp}(\cdot)$ denote sampling and prepend operations, respectively.

2) Flow through the CL encoder to extract high-level semantic distinctions. Here, we use multiple Transformer blocks as the CL encoder.

3) Normalize the representations with layer normalization to reduce the impacts of internal covariate shift, leading to faster convergence.

4) Feed into a deeper and larger MLP than that in the Transformer blocks and project the embeddings to dimension

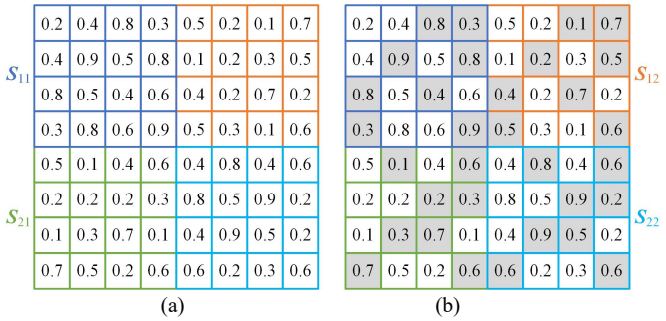


Fig. 4. Illustration of the similarity matrix and preliminary negative sampling. (a) Original similarity matrix. (b) Optimal sampled similarity matrix. The length N_b is 4, and k is 0.75. The cells filled with gray and white colors represent masked and reserved samples, respectively.

E_b. Deeper and larger hidden layers help to learn and capture more complex patterns, and more GELUs increase stronger nonlinear modeling capabilities.

5) Conduct the L_2 normalization to encourage the CL subnetwork to distribute the importance of embeddings more evenly and prevent any single embedding from dominating the learning process by constraining the weights to stay small, resulting in the model being more robust and better generalized.

6) Map the representations to sampling space $\boldsymbol{R}_{\text{b-o}} \in \mathbb{R}^{N_b \times E_o}$ to further extract better discrepancy representations of different LC types.

7) Optimize sampling for the representations with the detailed algorithm presented in Section III-D and output optimized sampling space $\boldsymbol{R}_{\text{s-o}} \in \mathbb{R}^{N_s \times E_o}$.

8) Estimate the contrastive loss $\mathcal{L}_{\text{Con}}$ with Eq. (13) in terms of optimized sampling space between the current modality $\boldsymbol{R}_{\text{s-o}}$ and another modality $\boldsymbol{R}'_{\text{s-o}}$.

9) Backpropagate gradients and update the parameters of the CL subnetwork.

D. Optimal Sampling and Contrastive Loss

After the linear heads of the CL subnetwork output the representations in two branches, the previous methods use these representations directly to calculate contrastive loss and update the network [9], [39], [40], [59], which means that all patch pairs corresponding to different locations are deemed as negative samples. Obviously, the traditional sampling methods are defective, given that not all patch pairs cover different land classes. Therefore, we designed a better strategy for sampling.

First, L_2 normalization Norm_{L2} is performed on the multimodal representations $\boldsymbol{R}_{\text{b-o}}$ and $\boldsymbol{R}'_{\text{b-o}}$, followed by concatenation along the dimension N_b

$$\boldsymbol{R}^{12}_{\text{b-o}}(\eta_1,\eta_2) = \text{Concat}_{N_b}\left(\text{Norm}_{L2}\left(\boldsymbol{R}_{\text{b-o}}(\eta_1,\eta_2)\right), \text{Norm}_{L2}\left(\boldsymbol{R}'_{\text{b-o}}(\eta_1,\eta_2)\right)\right), \tag{3}$$

where η_1 and η_2 are the parameters of the two CL branches, $\text{Concat}_{N_b}(\cdot)$ means the operation to concatenate the input along dimension N_b.

Then, considering spatial distance and topological relation have a slight association with the positive and negative samples in the CD task, compared with similarity, we adopt the inner product similarity with efficient and intuitive characteristics as the sampling metric. To accelerate calculation in parallel, we estimate the similarity matrix $\boldsymbol{S}$ as

$$\boldsymbol{S}(\eta_1,\eta_2) = \boldsymbol{R}^{12}_{\text{b-o}} * \text{Trans}(\boldsymbol{R}^{12}_{\text{b-o}}), \tag{4}$$

where $\text{Trans}(\cdot)$ denotes the transpose operation and * is matrix multiplication. In essence, the similarity matrix $\boldsymbol{S}$ contains four subblocks, that is, $\boldsymbol{S}_{11}$ (upper left), $\boldsymbol{S}_{12}$ (upper right), $\boldsymbol{S}_{21}$ (low left), and $\boldsymbol{S}_{22}$ (low right), and implies self-similarities with $\boldsymbol{S}_{11}$ and $\boldsymbol{S}_{22}$ and cross-similarities with $\boldsymbol{S}_{12}$ and $\boldsymbol{S}_{21}$, as shown in Fig. 4. The four blocks of similarities signify four sampling spaces, so we can impose constraints on them to optimize sampling. For this purpose, we obtain the preliminary negative samples $\boldsymbol{S}_{\text{Ng}}$ as follows:

$$\boldsymbol{S}_{\text{Ng}}(\eta_1,\eta_2) = \text{Samp}_2\left(\text{Samp}_1\left(\exp(\boldsymbol{S}/\tau)\right)\right), \tag{5}$$

where τ is the temperature to scale the similarity and set as 0.5. Since there is a logarithmic operation in the objective loss to be adopted later, we resort to an exponential operator here to ensure non-negativity. $\text{Samp}_1(\cdot)$ and $\text{Samp}_2(\cdot)$ are the sampling functions and can be expressed as

$$\begin{cases} \text{Samp}_1(\cdot) = topk_{11\text{-}12} \oplus topk_{12\text{-}11} \\ \text{Samp}_2(\cdot) = topk_{22\text{-}21} \oplus topk_{21\text{-}22} \end{cases}, \tag{6}$$

where $topk_{11\text{-}12} \oplus topk_{12\text{-}11}$ means sampling in $\boldsymbol{S}_{11}$ followed by in $\boldsymbol{S}_{12}$ with the constraint of similarity smaller than k, then sampling in $\boldsymbol{S}_{12}$ followed by in $\boldsymbol{S}_{11}$ with the same constraint; $topk_{22\text{-}21} \oplus topk_{21\text{-}22}$ is analogous. Such a sampling strategy ensures that negative samples are sufficiently clean in all four similarity blocks. The threshold k represents k times of N_b, $k \in (0,1]$. An illustration of the preliminary negative sampling is shown in Fig. 4, where the N_b and k are 4 and 0.75, respectively. Notably, the negative samples are essential to contrastive learning, and their quantity and quality jointly determine the performance of contrastive learning [50], [52], [61], so the number of negative samples should be of adequate size. Additionally, because the final equivalent threshold is the square of k in the two-fold sampling, after experiment (in Section V) and trade-off, we recommend setting k in the range of [0.7, 0.95], generally.

In addition, we generate the preliminary positive samples $\boldsymbol{S}_{\text{Ps}}$ by evaluating the cross-similarity of corresponding locations between the multimodal representations $\boldsymbol{R}_{\text{b-o}}$ and $\boldsymbol{R}'_{\text{b-o}}$, that is

$$\boldsymbol{S}_{\text{Ps}}(\eta_1,\eta_2) = \exp\left(\tau^{-1}\left\langle \text{Norm}_{L2}\left(\boldsymbol{R}_{\text{b-o}}(\eta_1,\eta_2)\right), \text{Norm}_{L2}\left(\boldsymbol{R}'_{\text{b-o}}(\eta_1,\eta_2)\right)\right\rangle_{b.}\right), \tag{7}$$

where $\langle\cdot,\cdot\rangle_{b.}$ represents the inner product in rows between input tensors. Then, we purify positive samples with

$$\boldsymbol{S}'_{\text{Pos}}(\eta_1,\eta_2) = \text{Concat}\left(\coprod_0^{kN_b} topk(\boldsymbol{S}_{\text{Ps}}), \coprod_{b_{ls}}^{b_{le}} topk(\boldsymbol{S}_{\text{Ps}})\right), \tag{8}$$

where $\coprod_{b_{ls}}^{b_{le}} topk(\boldsymbol{S}_{\text{Ps}})$ means retaining the elements from b_{ls} to b_{le} in descending order of $\boldsymbol{S}_{\text{Ps}}$, and $\coprod_0^{kN_b} topk(\boldsymbol{S}_{\text{Ps}})$ is analogous; $b_{le} = kN_b/2 + (1-k)N_b = (2-k)N_b/2$ and

$b_{\mathrm{ls}} = kN_{\mathrm{b}} / 2$. Recall that the similarity matrix $\boldsymbol{S}$ contains two symmetric blocks of cross-similarities, $\boldsymbol{S}_{12}$ and $\boldsymbol{S}_{21}$, and the rows of $\boldsymbol{S}_{\mathrm{Ng}}$ is $2N_{\mathrm{b}}$; whereas only a single cross-similarity block is considered in $\boldsymbol{S}_{\mathrm{Ps}}$ and $\boldsymbol{S}'_{\mathrm{Pos}}$ with N_{b} rows. We obtain the optimized positive samples $\boldsymbol{S}^{\mathrm{O}}_{\mathrm{Pos}}$ by concatenating $\boldsymbol{S}'_{\mathrm{Pos}}$ sequentially, ensuring that positive and negative samples have equal sizes for tensor computation and that their numbers are balanced to a certain degree, that is

$$\boldsymbol{S}^{\mathrm{O}}_{\mathrm{Pos}}(\eta_1, \eta_2) = \mathrm{Concat}(\boldsymbol{S}'_{\mathrm{Pos}}, \boldsymbol{S}'_{\mathrm{Pos}}). \tag{9}$$

After the $\boldsymbol{S}_{\mathrm{Ng}}$ and $\boldsymbol{S}^{\mathrm{O}}_{\mathrm{Pos}}$ are obtained, we derive the reweighted negative samples $\boldsymbol{S}_{\mathrm{Neg}}$ as

$$\boldsymbol{S}_{\mathrm{Neg}}(\eta_1, \eta_2) = \frac{-p \cdot N_{\mathrm{S}} \cdot \boldsymbol{S}^{\mathrm{O}}_{\mathrm{Pos}} + \boldsymbol{S}'_{\mathrm{Ng}}}{1-p}, \tag{10}$$

where N_{S} denotes the number of preliminary negative samples, p is the class probability and generally set to 0.1 [60], $\boldsymbol{S}'_{\mathrm{Ng}}$ represents the reprojected negative samples and can be deduced as

$$\boldsymbol{S}'_{\mathrm{Ng}}(\eta_1, \eta_2) = \sum_{i=1}^{N_{\mathrm{S}}} \frac{\exp(\sigma \log \boldsymbol{S}^i_{\mathrm{Ng}}) \cdot \boldsymbol{S}^i_{\mathrm{Ng}}}{\mathbb{E}_{i=1}^{N_{\mathrm{S}}} \exp(\sigma \log \boldsymbol{S}^i_{\mathrm{Ng}})}, \tag{11}$$

where σ is the concentration parameter, scheduled linearly with an initial value of 1.

Finally, the optimized negative samples $\boldsymbol{S}^{\mathrm{O}}_{\mathrm{Neg}}$ can be determined by rectifying the outliers, that is

$$\boldsymbol{S}^{\mathrm{O}}_{\mathrm{Neg}}(\eta_1, \eta_2) = \mathrm{Max}\left(N_{\mathrm{S}} \exp(-1/\tau),\ \boldsymbol{S}_{\mathrm{Neg}}\right). \tag{12}$$

The contrastive loss $\mathcal{L}_{\mathrm{Con}}$ between the optimized negative and positive samples can be evaluated as

$$\mathcal{L}_{\mathrm{Con}}(\eta_1, \eta_2) = -\mathbb{E}\left(\log \frac{\boldsymbol{S}^{\mathrm{O}}_{\mathrm{Pos}}}{\boldsymbol{S}^{\mathrm{O}}_{\mathrm{Pos}} + \boldsymbol{S}^{\mathrm{O}}_{\mathrm{Neg}}}\right) + \mu \left\| (\eta_1, \eta_2) \right\|_2, \tag{13}$$

where μ is the regularization coefficient and is determined automatically by the optimization algorithm. The computational complexity of optimal sampling and contrastive loss is $O\left((N_{\mathrm{b}})^2 E_{\mathrm{o}}\right)$. In implementation, we run Algorithm 1.

Algorithm 1. Optimal sampling and contrastive loss

Input: multimodal representations $\boldsymbol{R}_{\mathrm{b\text{-}o}}$ and $\boldsymbol{R}'_{\mathrm{b\text{-}o}}$, mini-batch N_{b}, similarity threshold k

Operation:

1: Calculate the similarity matrix $\boldsymbol{S}$ with Eqs. (3) and (4)

2: Obtain preliminary samples $\boldsymbol{S}_{\mathrm{Ng}}$ and $\boldsymbol{S}_{\mathrm{Ps}}$ with Eqs. (5) to (7)

3: Determine the optimized samples $\boldsymbol{S}^{\mathrm{O}}_{\mathrm{Pos}}$ and $\boldsymbol{S}^{\mathrm{O}}_{\mathrm{Neg}}$ with Eqs. (9) and (12)

4: Estimate the contrastive loss $\mathcal{L}_{\mathrm{Con}}$ with Eq. (13)

Output: The contrastive loss $\mathcal{L}_{\mathrm{Con}}$

E. Silent Attention

The vanilla attention mechanism in the Transformer can be expressed as

$$\mathrm{Attention}(Q, K, V) = \mathrm{softmax}\left(\frac{QK^{\mathrm{T}}}{\sqrt{d}}\right)V, \tag{14}$$

where Q, K, and V denote queries, keys and values with dimension of $n \times d$, respectively; the softmax is defined as

$$s_{\mathrm{a}} = \mathrm{softmax}(x) = \frac{\exp(x_i)}{\sum_{j=1}^{n} \exp(x_j)}, \tag{15}$$

where x is a vector with dimension n. Then the negative limit of the softmax is deduced as

$$\lim_{x_i \to -\infty} \cdots \lim_{x_n \to -\infty} \left(\mathrm{softmax}(x)\right)_i = \frac{1}{n}, \tag{16}$$

which means the negative limit of the softmax is a positive constant. This raises a problem: the softmax restrains the attention from outputting zero values, even if there is no correlation between two tokens. Consequently, these irrelevant tokens are assigned weights to extract information with V, and learning is unstable and prone to collapse in our experiments.

To address this problem, we proposed an improved softmax function as

$$s_{\mathrm{s}} = \mathrm{softmax}_{\mathrm{s}}(x) = \frac{\exp(x_i)}{C_{\mathrm{s}} + \sum_{j=1}^{n} \exp(x_j)}, \tag{17}$$

where C_{s} is a positive constant. The key difference from the original softmax is the negative limit

$$\lim_{x_i \to -\infty} \cdots \lim_{x_n \to -\infty} \left(\mathrm{softmax}(x)\right)_i = 0. \tag{18}$$

When the input x contains significantly negative correlations, the proposed $\mathrm{softmax}_{\mathrm{s}}$ (s_{s}) tries to avoid scoring. Moreover, the derivative of the s_{s} is positive, so we always have a non-zero gradient; its sum is in the range of zero to one, and the relative ratio in the output vector is the same as that in the original softmax, which means the output is under control.

In addition, we add a dropout operation after s_{s} to balance the representational and generalized ability, that is

$$A_{\mathrm{s}}(Q, K, V) = \mathrm{dropout}\left(\mathrm{softmax}_{\mathrm{s}}\left(\frac{QK^{\mathrm{T}}}{\sqrt{d}}\right)\right)V. \tag{19}$$

The $\mathrm{softmax}_{\mathrm{s}}$ can drive the scores of irrelevant tokens toward zero while slightly reducing the scores of others. This reduction is compensated during the subsequent normalization. As a result, it suppresses features with low correlation, increases the discrepancy of output features, and stabilizes the training.

Since the attention mechanism costs enormous time and space resources, we implement silent attention with a specially optimized algorithm to be faster and simpler. Specifically, for running in parallel on CUDA, we transform Eq. (17) as

$$s_{\mathrm{s}} = \mathrm{softmax}_{\mathrm{s}}(x) = \frac{\exp(x_i)}{\exp(\log C_{\mathrm{s}}) + \sum_{j=1}^{n} \exp(x_j)}, \tag{20}$$

then, we execute Algorithm 2 on CUDA. Note that we recommend setting C_{s} to 1, so we just prepend zero to the scaled correlation in the last dimension and avoid precision loss during the ferrying between exponentiation and logarithm. In Algorithm 2, it can be observed that the silent attention receives the same input as the vanilla attention and can be swapped into network flexibility. The computational complexity of silent

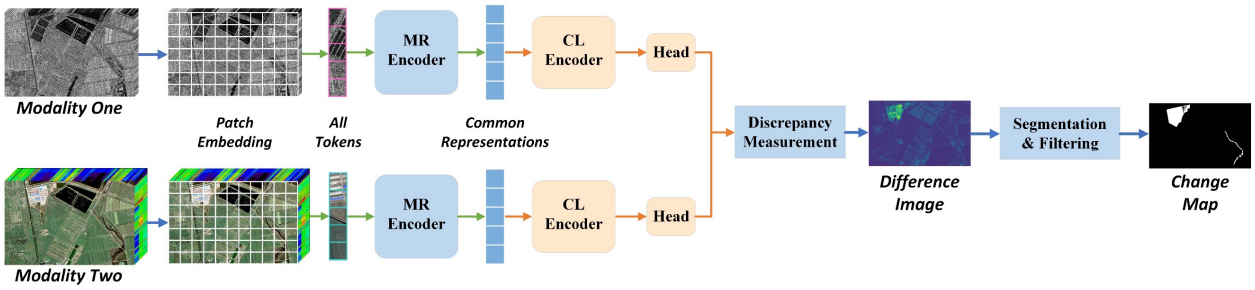


Fig. 5. Inference pipeline for multimodal change detection.

TABLE I
DESCRIPTION OF THE FIVE MULTIMODAL DATASETS

Dataset	Sensor (& modality)	Size	Location	Event (& Spatial resolution)
OSCD-S2S1	Sentinel-1/Sentinel-2 (SAR-Multispectral)	$H \times W \times 13(2)$	Scattered around the world	Mixed (10 m)
Shuguang	Radarsat-2/Google Earth (SAR-RGB)	593 × 921 × 1(3)	Shuguang Village, China	Construction (8 m)
Sardinia	Landsat-5/Google Earth (NIR-RGB)	300 × 412 × 1(3)	Sardinia, Italy	Lake expansion (30 m)
Toulouse	Pleiades/WorldView-2 (RGB-Pseudo RGB)	2000 ×2000 × 3(3)	Toulouse, France	Construction (0.52 m)
Sutter	Landsat-8/Sentinel-1A (Multispectral-SAR)	875 × 500 × 11(3)	Sutter County, USA	Flooding (≈15m)

attention is the same as that of vanilla attention, which is $O(n^2 d)$. Therefore, we replaced vanilla attention with silent attention in all Transformer blocks of MaCon.

Algorithm 2. Silent attention

Input: queries Q, keys K, values head V, head dim d
Calculate correlation:
1: Estimate the correlation by matrix multiplication between the Q and transpose of K
2: Divide by the square root of d to scale the correlation
Estimate softmax$_s$:
1: Prepend $\log C_s$ in column to scaled correlation in the last dimension
2: Calculate s_s by performing softmax operation with Eq. (15)
3: Remove the prepended column in the s_s
Determine the Silent attention:
1: Conduct dropout on s_s
2: Derive the A_s by matrix multiplication between the s_s and V
Output: The silent attention A_s

F. Optimization and Inference

After completing the forward pass and deriving the MR loss $\mathcal{L}_{\text{Ma}}$ and contrastive loss $\mathcal{L}_{\text{Con}}$, the optimization process begins. We designed stop-gradient operation in the CL subnetwork and asynchronous backpropagation for MR and CL subnetworks. That means the gradients from the $\mathcal{L}_{\text{Ma}}$ and $\mathcal{L}_{\text{Con}}$ are asynchronously backpropagated to update the parameters of the respective MR and CL subnetworks. In this way, the two subnetworks not only collaborated in forward modelling, but were also independent in backpropagation without conflicting gradient updates.

When the relative changes in both losses $\mathcal{L}_{\text{Ma}}$ and $\mathcal{L}_{\text{Con}}$ remain below 1e-3 for 20 epochs, or the maximum training epochs is reached, learning is stopped. Then, the change map can be inferred by utilizing the common and discriminative representations of the multimodal data. The inference pipeline is shown in Fig. 5. Before the block of discrepancy measurement, the main difference from the learning phase is that the inference phase is without crop and mask operations, MR decoder and reconstruction modules. The difference image can be derived by measuring the distance between discrepancy representations output from the head in the CL subnetwork. After that, the preliminary binary change is output through a segmentation algorithm, and finally, the refined change map is obtained through morphological filtering.

Note that we developed a trick for the MaCon framework to hack the parameters of patch embedding so as to change the patch stride in the inference phase. In this way, we can set a smaller stride size to model finer boundaries and representations or a larger one to accelerate inference.

IV. EXPERIMENTS AND ANALYSIS

To evaluate the performance of the proposed MaCon framework, we experimented on both multimodal and monomodal datasets, as well as analyzed the computational cost and impact of key submodules and hyperparameters on performance.

A. Datasets Description

1) Multimodal datasets: We experimented on five multimodal datasets, including 14 multimodal image pairs distributed worldwide, as listed in Table I. The Shuguang dataset contains SAR and RGB modalities, with a size of 593 × 921 × 1(3). The Sardinia dataset includes NIR and RGB data before and after the lake expansion. The Toulouse dataset consists of large-scale images with a size of 2000 × 2000 × 3(3),

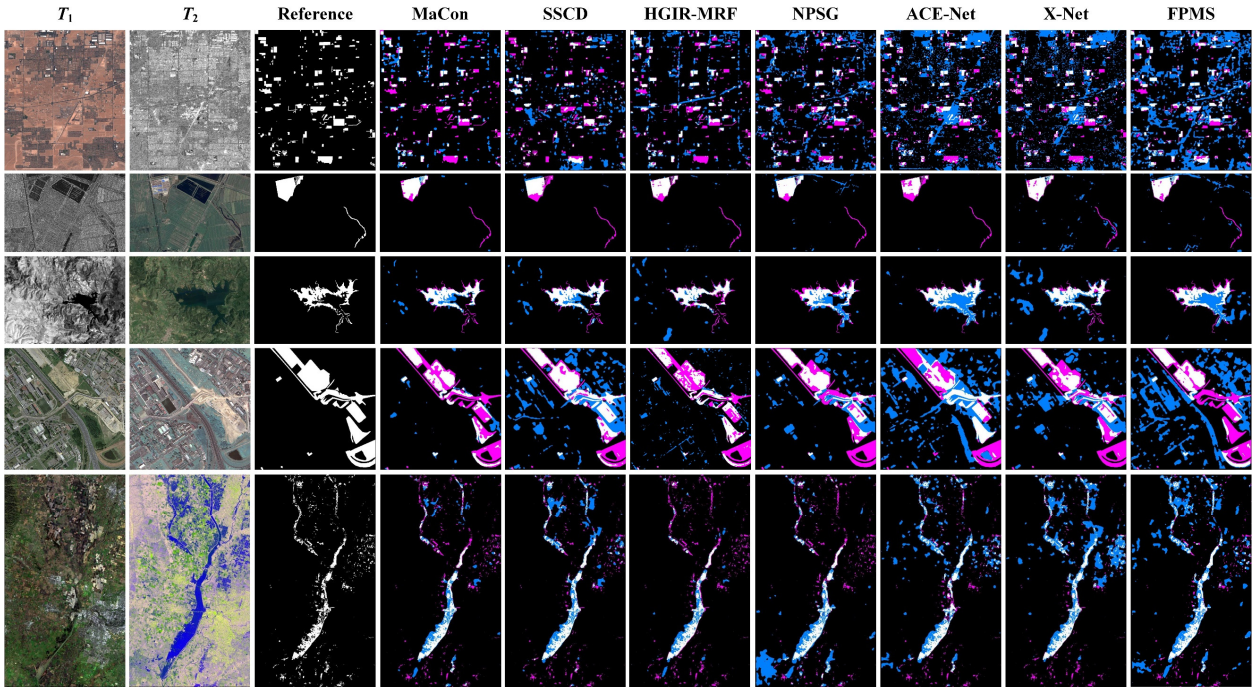


Fig. 6. Rendered change maps obtained by the proposed framework and representative comparison methods on the multimodal datasets. From top to bottom, they correspond to the Lasvegas case, Shuguang, Sardinia, Toulouse and Sutter datasets, respectively. In the rendered change maps, white: true positives (TP); black: true negatives (TN); azure: false positives (FP); magenta: false negatives(FN). The meaning of these symbols is the same in all figures.

TABLE II
PERFORMANCE COMPARISON ON THE MULTIMODAL DATASETS

Method	Metric	OSCD-S2S1	Shuguang	Sardinia	Toulouse	Sutter
MaCon (ours)	F1	**0.217**	**0.815**	**0.736**	**0.559**	**0.531**
	KC	**0.175**	**0.806**	**0.717**	**0.503**	**0.507**
SSCD	F1	0.202	0.713	0.714	0.519	0.511
	KC	0.165	0.697	0.692	0.414	0.486
HGIR-MRF	F1	0.186	0.791	0.704	0.542	0.514
	KC	0.141	0.779	0.683	0.487	0.491
NPSG	F1	0.165	0.737	0.651	0.472	0.443
	KC	0.112	0.724	0.624	0.380	0.406
M3CD-EMAP	F1	0.163	0.768	0.591	0.457	0.425
	KC	0.111	0.754	0.514	0.348	0.393
ACE-Net	F1	0.142	0.690	0.632	0.398	0.465
	KC	0.091	0.680	0.601	0.258	0.428
X-Net	F1	0.138	0.735	0.587	0.434	0.426
	KC	0.088	0.710	0.548	0.321	0.385
M3CD	F1	0.146	0.622	0.603	0.480	0.187
	KC	0.097	0.602	0.526	0.405	0.124
FPMS	F1	0.143	0.662	0.562	0.311	0.429
	KC	0.086	0.645	0.533	0.145	0.388

including RGB and pseudo-RGB modalities. The Sutter dataset contains multispectral and SAR images before and after the flood, with a size of 875 × 500 × 11(3).

It should be noted that the OSCD-S2S1 dataset [18] we used is its test set, which contains 10 image pairs. The OSCD-S2S1 dataset is more challenging and practical than the other datasets because it possesses the following characteristics: mixed changing events, complex LC, very few changed pixels in several cases, more small changing elements and discrete distribution, fine temporal resolution and accessibility for the data of Sentinel-1 and 2. Regrettably, the recently published OSCD-S2S1 dataset comprises many wrong labels, likely due to its complexity, and unsupervised CD methods were rarely tested on the OSCD-S2S1 dataset. Therefore, we want to provide a benchmark for testing on the OSCD-S2S1 dataset.

2) Monomodal datasets: To evaluate the performance of the proposed method on the monomodal dataset, we tested it on two optical datasets, Montpellier and ZY3, and one SAR dataset, San Francisco. The Montpellier dataset, contained in the OSCD-S2S2 dataset [62], consists of a pair of multispectral images of size 426× 451 × 13(13) pixels and a spatial resolution of 10 m. The ZY3 dataset [63] contains two RGB images of size 458 × 559 × 3(3) pixels and a spatial resolution of 5.8 m [63]. The San Francisco dataset [64] consists of a pair of SAR images of size 256 × 256 × 1(1) pixels, acquired by ERS-2.

B. Implementation Details

The proposed MaCon was implemented based on the Pytorch with a single NVIDIA GeForce RTX 3090 (24-GB RAM). We adopted the Transformer architecture with silent attention for all the CL encoder, MR encoder and decoder.

For the MR subnetwork, we set the crop size to 200 × 200 in learning. We used the convolutional layer with kernel and stride size of 8 × 8 and 4 × 4 in learning, and that of 8 × 8 and 2 × 2 in inference, to tokenize the multimodal data to the embeddings. The ViT-Small [65] was set as the encoder with 384 embedding dimensions; the decoder comprises 8 Transformer blocks with 16 heads and 512 embedding dimensions and contains a linear predictor additionally. The random mask and dropout ratio were set to 0.5 and 0.1, respectively. During learning, the number of input tokens in the MR subnetwork was 2401.

For the CL subnetwork, we set the layers of the Transformer block to 8 with 6 heads and 384 embeddings, the similarity threshold k to 0.9, mini-batch N_b to 256, dropout ratio to 0.1, temperature τ to 0.5, class probability p to 0.1, 3 linear layers in MLP with hidden embeddings to 2048, bottleneck embeddings E_b to 256, and output embeddings E_o to 384.

Moreover, the optimization algorithm based on adaptive estimates of low-order moments with decoupled weight decay

TABLE III
REPORTED PERFORMANCE COMPARISON WITH RECENTLY PUBLISHED SOTA METHODS ON THE MULTIMODAL DATASETS. NOTABLY, * MEANS WE RAN THEIR CODE AND TRIED OUR BEST TO TUNE HYPERPARAMETERS TO ACHIEVE OPTIMAL; + SIGNIFIES SUPERVISED METHOD

OSCD-S2S1	F1	KC	Shuguang	F1	KC	Sardinia	F1	KC	Toulouse	F1	KC	Sutter	F1	KC
MaCon (ours)	**0.217**	**0.175**	**MaCon (ours)**	**0.815**	**0.806**	**MaCon (ours)**	**0.736**	**0.717**	**MaCon (ours)**	**0.559**	**0.503**	**MaCon (ours)**	**0.531**	**0.507**
FC-EF+ [19]	0.171		IRGMcS [28]	0.804	0.794	FD-MCD [74]	0.732	0.714	HGIR-MRF [38]	0.549	0.501	IRGMcS [28]	0.512	0.490
SSCD* [39]	0.202	0.165	HGIR-MRF [38]	0.790	0.779	ALSC-P [75]		0.713	AGSCC [17]	0.540	0.490	HGIR-MRF [38]	0.511	0.489
HGIR-MRF* [38]	0.186	0.141	PSGM [76]		0.744	NACCL [77]	0.700		CAAE [78]	0.520	0.451	SSCD [39]	0.510	0.460
NPSG* [66]	0.165	0.112	NPSG [66]		0.729	PSGM [76]		0.682	IRGMcS [28]	0.481	0.421	SCCN [34]	0.500	0.454
M3CD-EMAP* [67]	0.163	0.111	X-Net [31]	0.731	0.696	AGSCC [17]	0.680	0.658	FPMS [35]	0.296		ALSC-P [75]		0.420
ACE-Net* [31]	0.142	0.091	ACE-Net [31]	0.726	0.689	CAAE [78]	0.628	0.598	NACCL [77]	0.290		ACE-Net [31]	0.459	0.415

TABLE IV
PERFORMANCE COMPARISON ON THE MONOMODAL DATASETS. + SIGNIFIES SUPERVISED METHOD

Montpellier	F1	KC	ZY3	F1	KC	San Francisco	F1	KC
MaCon (ours)	**0.553**	**0.520**	**MaCon (ours)**	**0.571**	**0.523**	**MaCon (ours)**	**0.905**	**0.897**
FDCNN+ [63]	0.440	0.390	FDCNN+ [63]	0.548	0.500	FDCNN+ [63]	0.882	0.873
FCD-GN [68]	0.544	0.503	FCD-GN [68]	0.566	0.519	LR-CNN [64]	0.893	0.883
DSFA [69]	0.427	0.371	DSFA [69]	0.543	0.487	PCAKM [71]	0.878	0.870
RCVA [70]	0.435	0.378	RCVA [70]	0.539	0.485	SSN-Siam-diff [8]	0.869	0.859
ISFA [69]	0.388	0.327	ISFA [69]	0.506	0.421	FC-Siam-conc [72]	0.761	0.743
SSN-Siam-diff [8]	0.486	0.445	SSN-Siam-diff [8]	0.471	0.385	Ms-CapsNet+ [73]	0.903	0.894

(AdamW) was employed to train MaCon with the decayed learning rate set from 5e-3 to 1e-5 and weight decay from 0.04 to 0.4 in the cosine schedulers; the mixed-precision computing technique was adopted to decrease memory cost during learning. In inference, the difference map was generated by mean squared L_2 distance, i.e., $\mathbb{E}\left\|\boldsymbol{R}_{\text{b-o}} - \boldsymbol{R}'_{\text{b-o}}\right\|_2^2$, and Otsu was applied as the segmentation algorithm to obtain the binary change map. Otsu is one of the most widely used segmentation methods for automatic thresholding, and its criterion is to maximize the inter-class variance between classes.

C. Comparison Methods and Evaluation Metrics

1) Comparison methods on the multimodal datasets: To demonstrate the superiority of the MaCon for unsupervised multimodal change detection, we compared eight recently proposed methods with MaCon since they are representative and their code is open-source, including SSCD [39], HGIR-MRF [38], NPSG [66], M3CD-EMAP [67],ACE-Net [31], X-Net [31], M3CD [14] and FPMS [35]. For these methods, we ran their code to test and tried our best to tune hyperparameters to achieve optimal.

Note that on the OSCD-S2S1 dataset, we chose the best metric for all the compared methods on each image pair after repeatedly adjusting their hyperparameters. In contrast, the MaCon directly adopted the mean metric on all 10 image pairs with the same hyperparameters to evaluate its adaptive ability on different images more objectively. We also compared the reported metrics from their original papers to reflect the SOTA accuracy achieved on each dataset, including many methods without open-source code.

2) Comparison methods on the monomodal datasets: We compared the MaCon with advanced monomodal CD methods. Specifically, the FCD-GN [68], DSFA [69], RCVA [70], ISFA [69], FDCNN [63] and SSN-Siam-diff [8] were compared on the optical datasets of Montpellier and ZY3. Similarly, the PCAKM [71], FC-Siam-conc [72], LR-CNN [64], FDCNN [63], SSN-Siam-diff [8] and Ms-CapsNet [73] were compared on the SAR dataset of San Francisco. For these methods, if reported metrics existed for tested datasets in the associated papers, we adopted them directly; if not, we ran the code and tried our best to adjust the hyperparameters to train them optimally.

Notably, we did not compare the MaCon with the same methods that were compared on the multimodal datasets, because traditional methods for MCD are usually not tested on monomodal datasets, and the metrics are generally worse than those customized for monomodal datasets.

3) Evaluation Metrics: In the experiments, two comprehensive quality metrics, F1 score (F1) and Cohen's Kappa coefficient (KC), were used to evaluate the performance of all methods quantitatively. A larger value signals better performance for all of these metrics

$$\begin{cases} \mathrm{F1} = \dfrac{2 \times Pre \times Rec}{Pre + Rec}, \\ \mathrm{KC} = \dfrac{OA - PE}{1 - PE} \end{cases}, \tag{21}$$

where the subitems are defined as

$$\begin{cases} Pre = \dfrac{TP}{TP + FP}, \\ Rec = \dfrac{TP}{TP + FN}, \\ OA = \dfrac{TP + TN}{TP + TN + FP + FN}, \\ PE = \dfrac{(TP + FN)(TP + FP) + (TN + FP)(TN + FN)}{(TP + TN + FN + FP)^2} \end{cases}, \tag{22}$$

where TP is the true positive, denoting the number of pixels correctly classified as changed; TN is true negative, which means the number of pixels correctly classified as unchanged; FP is false positive, which denotes means the number of pixels misclassified as changed; FN is false negative, representing the number of pixels misclassified as unchanged; OA denotes overall accuracy; PE means expected agreement between the ground reference and predictions given the class distributions.

TABLE V
ABLATION PERFORMANCE ON THE MULTIMODAL DATASETS. THE MEANING OF THE ABBREVIATIONS ON THE HEAD ROW ARE THE FOLLOWING. DA: DATA AUGMENTATION; MR-CL: THE STAGED SERIES COMBINATION OF MR AND CL; OS: OPTIMAL SAMPLING; SA: SILENT ATTENTION; CYCL: CYCLE LOSS; EMA: EXPONENTIAL MOVING AVERAGE; PE: POSITIONAL EMBEDDING; MF: MORPHOLOGICAL FILTERING

MR	CL	DA	MR-CL	OS	SA	CycL	ema	PE	MF	OSCD-S2S1	Shuguang	Sardinia	Toulouse	Sutter	Params	FLOPs	Memory
√		√								0.122	0.378	0.410	0.319	0.328	96 M	253 G	12.2 GB
	√	√								0.163	0.561	0.502	0.422	0.374	41 M	157 G	3.5 GB
√	√	√								0.186	0.703	0.624	0.471	0.425	134 M	405 G	14.6 GB
√	√	√	√							0.161	0.589	0.607	0.429	0.417	134 M	405 G	14.6 GB
√	√	√		√						0.207	0.763	0.704	0.535	0.491	134 M	406 G	14.6 GB
√	√	√			√					0.198	0.726	0.676	0.507	0.463	134 M	405 G	14.6 GB
√	√	√		√	√					0.214	0.791	0.714	0.537	0.518	134 M	406 G	14.6 GB
√	√	√		√	√				√	**0.217**	**0.815**	**0.736**	**0.559**	**0.531**	**134 M**	**407 G**	**14.6 GB**
√	√	√		√	√	√				0.202	0.711	0.623	0.504	0.472	134 M	656 G	25.7 GB
√	√	√		√	√		√			0.107	0.290	0.309	0.233	0.198	134 M	406 G	14.3 GB
√	√	√		√	√			√		0.203	0.731	0.692	0.525	0.485	134 M	406 G	14.6 GB
√	√			√	√					0.208	0.762	0.718	0.536	0.503	134 M	406 G	14.6 GB
√	√			√	√			√		0.191	0.660	0.647	0.474	0.462	134 M	406 G	14.6 GB

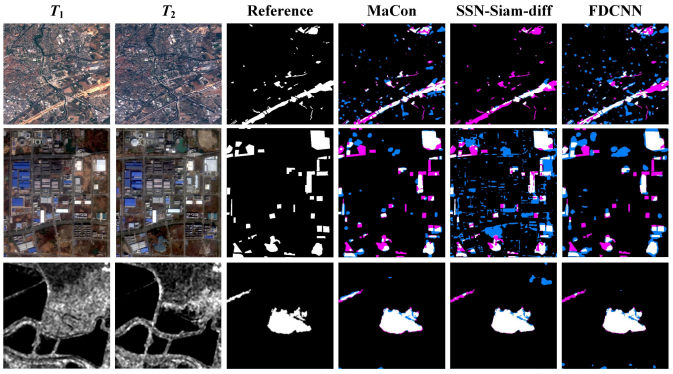


Fig. 7. Rendered change maps obtained by MaCon and representative comparison methods on the monomodal datasets. From top to bottom, they correspond to the Montpellier, ZY3 and San Francisco datasets, respectively.

D. Results

1) On multimodal datasets: The visualized CD results on the multimodal datasets are shown in Fig. 6, and the statistical metrics are listed in Table II. From Fig. 6, we can see that the proposed MaCon is the best at suppressing spurious changes, detecting small change elements and determining the boundary between the changed and unchanged classes. In Table II, the proposed MaCon method outperforms all the other benchmark methods on the multimodal datasets, with preponderant leading in metrics. These findings indicate that the MaCon framework achieves great accuracy on different multimodal datasets.

To evaluate the accuracy of the proposed method in detail, we also compared the reported metrics in the original papers with recently published SOTA methods on each multimodal dataset, as listed in Table III. These methods cover the three classes of approaches mentioned in Section I.

Note that, as mentioned before, almost no unsupervised methods have been tested on the OSCD-S2S1 dataset. Therefore, we include the results of the comparison methods above and compare them with the reported metrics of the supervised FC-EF methods. It can be seen that our MaCon framework outperforms the SOTA methods on all datasets. Noteworthy, the MaCon even significantly surpasses the supervised method of FC-EF. These results demonstrate that MaCon framework has excellent accuracy for MCD.

TABLE VI
COMPLEXITY AND ACCURACY (F1) COMPARISON

(A) ON THE MULTIMODAL DATASETS

Method	Params	FLOPs	Memory	Shuguang	Sutter
MaCon (ours)	**134 M**	**407 G**	**14.6 GB**	**0.815**	**0.531**
SSCD	57 M	389 G	6.9 GB	0.713	0.511
ACE-Net	45 M	684 G	4.7 GB	0.690	0.465
X-Net	37 M	553 G	3.4 GB	0.735	0.426

(B) ON THE MONOMODAL DATASETS

Method	Params	FLOPs	Memory	Montpellier	ZY3
MaCon(ours)	**67 M**	**203 G**	**7.5 GB**	**0.553**	**0.571**
DSFA	43 M	89 G	4.0 GB	0.427	0.543
FDCNN	146 M	284 G	6.4 GB	0.440	0.548

2) On monomodal datasets: The visualized CD results on monomodal datasets are shown in Fig. 7, and the statistical metrics are listed in Table IV. We can see that the proposed framework outperforms all the other comparison methods on the monomodal datasets. Notably, the MaCon exceeds the supervised methods of FDCNN and Ms-CapsNet.

The results on multimodal and monomodal datasets indicate that the proposed framework can extract the common representations in essential between different modalities, so it has remarkable performance and generalization for both multimodal and monomodal CD. This also implies that the proposed framework promises to provide a unified model for the field of CD.

E. Ablation Study

To further quantitatively explore the contribution of the main modules in MaCon on MCD, we conducted sufficient ablations. The complexity involved was calculated based on data sizes of 200 × 200 × 3. The results are shown in Table V, from which we can obtain the following insights.

TABLE VII
COMPARISON OF COMPUTATIONAL TIME (IN SECONDS) ON FOUR MULTIMODAL DATASETS

Datasets	Image size	MaCon (ours)		SSCD		HGIR-MRF	NPSG	ACE-Net	X-Net	FPMS
		Learning	Inference	Learning	Inference					
Lasvegas	824×716	168.2	5.2	261.5	5.3	141.9	351.4	872.8	699.4	20.3
Shuguang	593×921	168.5	4.8	266.4	4.9	129.1	295.6	745.2	646.1	18.8
Montpellier	426×451	160.7	1.6	260.3	1.7	148.3	119.7	494.5	412.7	12.6
Sardinia	300×412	159.3	1.1	256.7	1.1	128.4	67.2	319.3	282.6	8.2

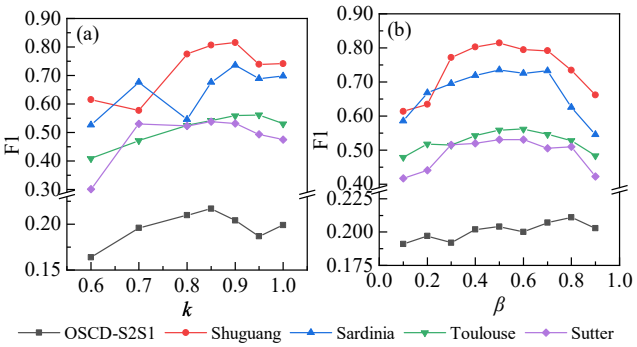


Fig. 8. Sensitivity of (a) similarity threshold k, and (b) mask ratio β.

1) Only a single MR or CL network does not work well, especially only MR, because the learning objective of MR is not intuitively consistent with the task of CD. However, we find that the MR subnetwork holds the property of accelerating representation distilling. Specifically, only a single CL network or the staged series combination of MR and CL requires more training epochs to achieve the best performance.

2) The effect of the staged series combination of MR and CL is not as accurate as the parallel coupling.

3) Quantitative evaluation confirms again that both optimal sampling and silent attention have a significant positive correlation with the performance of MCD.

4) Without flipping augmentation, the accuracy does not drop by much. That is because random masking with shuffling in MaCon is equivalent to a strong enhancement.

5) The coupling of MR with CL is complex, and the output of the first reconstruction is not good enough to provide high-quality input for the second learning, so the performance is unsatisfactory after adding cycle loss [17]. Moreover, the computational cost and training difficulty surge, and anomalies such as gradient vanishing and loss NaN are apt to occur.

6) The ema strategy is highly unsuitable for updating parameters on branches with different modalities. Because the distinctions in multimodal data are considerable, the ema induces the inability to learn how to extract representations in another modality effectively, and the parameters of another modality are probably even misled far away from its truth.

F. Computational Cost

Another aspect to be analyzed is the computational cost of the proposed framework. We compared the complexity and accuracy of MaCon and open-source deep learning-based methods on both multimodal and unimodal datasets. The results are presented in Table VI. Additionally, we tested the practical runtime of MaCon and six compared methods on the Lasvegas, Shuguang, Montpellier and Sardinia multimodal datasets. The results are reported in Table VII. Note that the source codes for these methods were implemented in different programming environments: FPMS in C++, HGIR-MRF and NPSG in MATLAB, and others are deep learning-based methods, implemented in Python.

Although MaCon has more parameters and requires additional memory, it achieves the highest accuracy with acceptable computational complexity compared to the other methods. Because MaCon adopts several parallel computation strategies and large batch size in inference, its runtime is the shortest among all the deep learning-based methods. ACE-Net and X-Net rely heavily on prior computation and, as a result, have large FLOPs and long runtime.

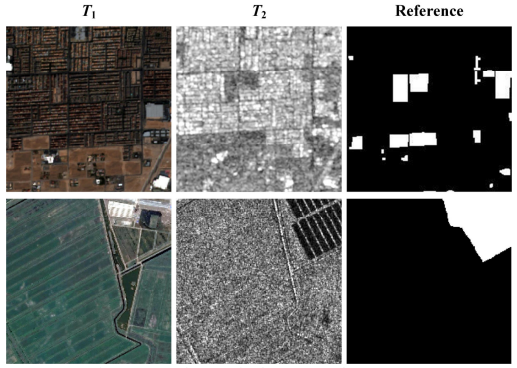


Fig. 9. Representative cropping window on the Lasvegas case (top) and Shuguang dataset (bottom).

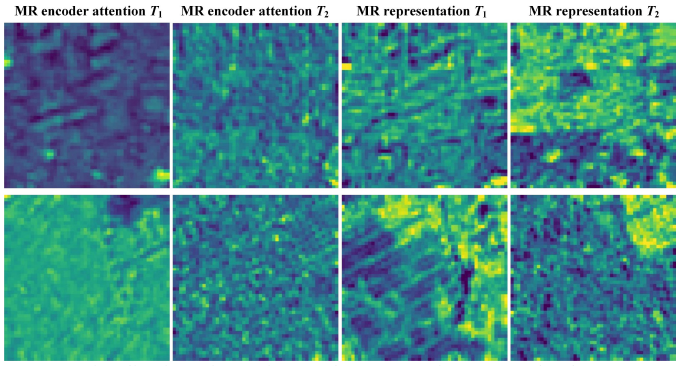


Fig. 10. Visualization of attentions and representations output by the MR encoders on the Lasvegas case (top) and Shuguang dataset (bottom).

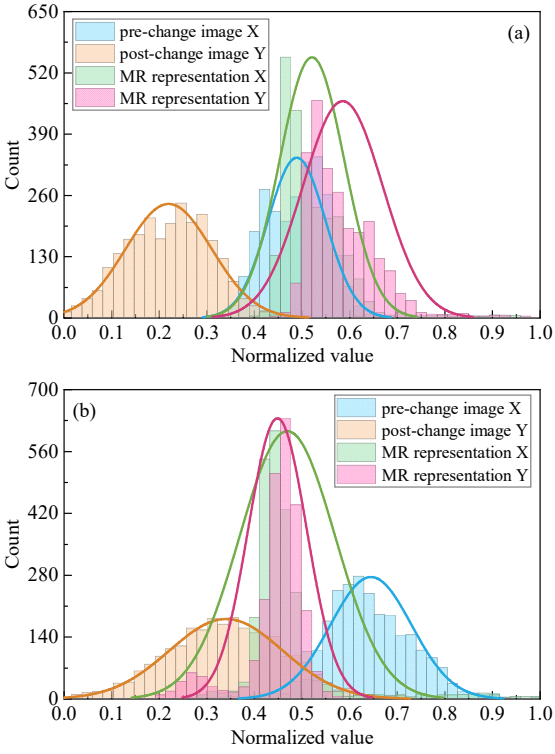


Fig. 11. The distributions between representations output by the MR encoders and original images on the (a) Lasvegas case and (b) Shuguang dataset.

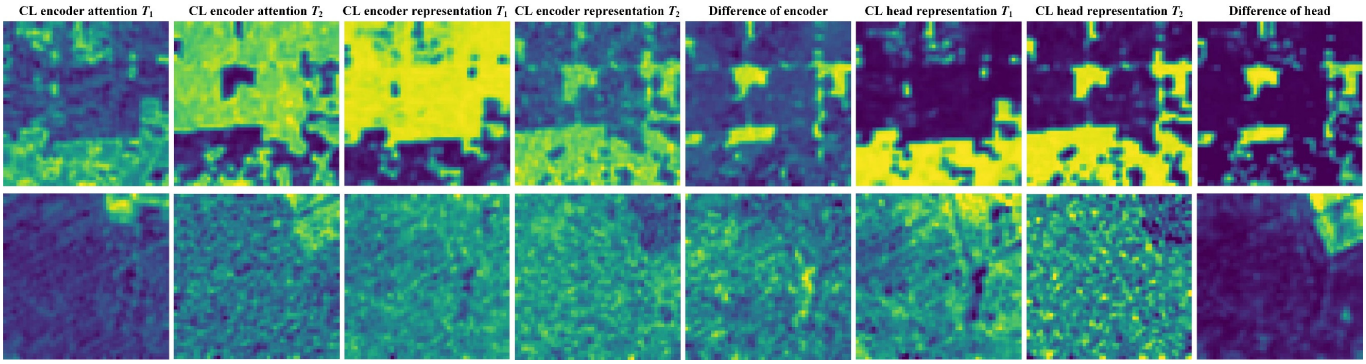


Fig. 12. Visualization of the attentions, representations and difference image output by the CL encoders, and that of the representations and difference image output by the CL heads. The first and second rows are the Lasvegas case and Shuguang dataset, respectively.

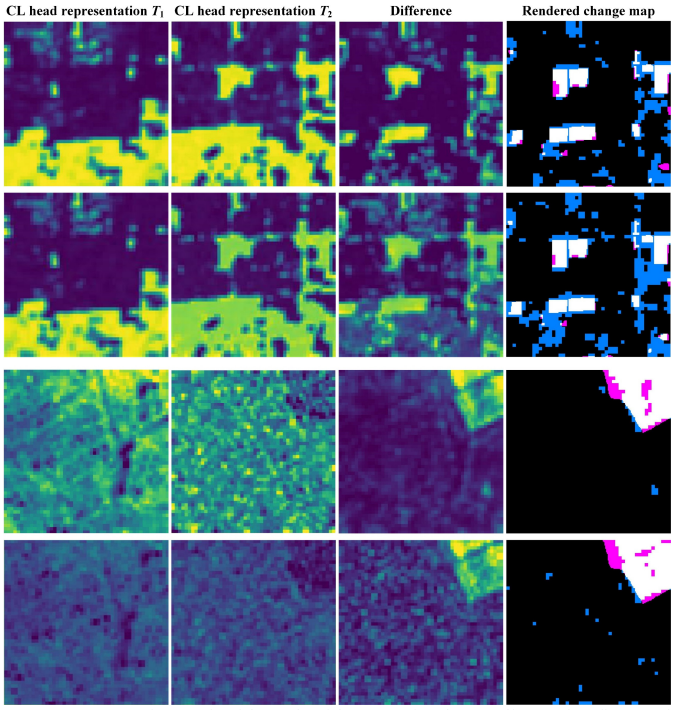


Fig. 13. Comparison of with and without optimal sampling. The first and second rows use optimal sampling and vanilla sampling on the Lasvegas case, respectively; the third and fourth rows use optimal sampling and vanilla sampling on the Shuguang dataset, respectively.

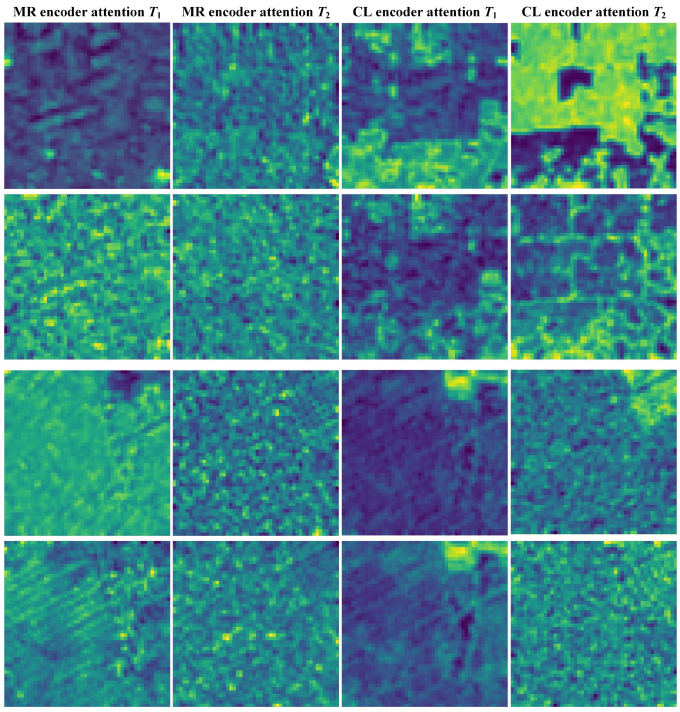


Fig. 14. Comparison of with and without silent attention. The first and second rows use silent and vanilla attention on the Lasvegas case, respectively; the third and fourth rows use silent and vanilla attention on the Shuguang dataset, respectively.

G. Sensitivity of Hyperparameters

1) Similarity threshold k: The impact of the similarity threshold k on performance is shown in Fig. 8 (a). We find that the accuracy is greatest when k is 0.85 or 0.9; specifically, when k is 0.85 in the OSCD-S2S1 and Sutter datasets, and k is 0.9 in the other datasets, to achieve the best performance. This is because the OSCD-S2S1 and Sutter datasets possess a considerably complex LC distribution, so they need more strict sampling constraints to help discriminate the discrepancy.

2) Mask ratio β: The impact of mask ratio β on performance is shown in Fig. 8 (b). It can be seen that the accuracy ascends with β until β reaches 0.5 on all datasets; when the mask ratio is larger than 0.5, the accuracy still increases until β reaches approximately 0.8 on the OSCD-S2S1 dataset, but the accuracy decreases generally on the other datasets. To avoid the high cost of parameter adjustment and improve the versatility of the proposed framework, we uniformly used the results of β at 0.5 on all datasets.

V. Discussion

To understand thoroughly how MaCon works and the role of each key module in the entire framework, we analyzed the mechanisms of MR and CL subnetworks, optimal sampling, and silent attention from the perspectives of interpretability. To exhibit the details clearly, we use representative cropping windows during the learning phase on the Lasvegas case in the OSCD-S2S1 dataset and the Shuguang dataset for illustration. The cropped dual temporal images and binary ground reference of the Lasvegas and Shuguang are shown in Fig. 9.

A. Common representation extraction in the MR branch

First, we visualized the attentions and representations output by the MR encoders, and compared the distributions between the representations and original images to investigate the role of the MR subnetwork. The visualization of attentions and

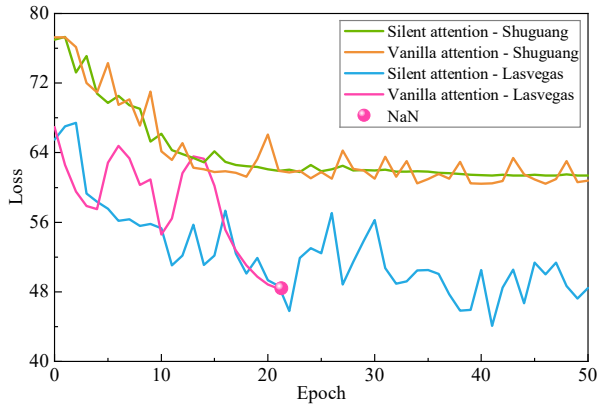


Fig. 15. Loss curves during learning on the Lasvegas case and Shuguang dataset.

representations on the Lasvegas and Shuguang datasets are shown in Fig. 10.

We can see that the dual temporal attentions do not apparently focus on local or specific objects; the representations abstract and generalize the surface object information and are more similar than the original images.

Additionally, the distributions of the representations output by the MR encoders and original images are shown in the Fig. 11. Since the range of representations output by the MR encoders is quite different from that of the original images, we normalized them to the range of 0 to 1. We find that the distributions of the MR representations are more similar and closer than those of the original images. To quantify this effect, we estimated the root mean square distance d_r and the relative root mean square distance r_d. On the Lasvegas case, the d_r of the original images and MR representations are 0.286 and 0.067, and their r_d are 0.808 and 0.121, respectively. On the Shuguang dataset, the d_r of the original images and MR representations are 0.336 and 0.137, and their r_d are 0.681 and 0.298, respectively. These findings demonstrate that the MR encoders can distill the common representations from multimodal data and shrink their domain bias.

B. Discrepancy representation extraction in the CL branch

We studied the role of the CL subnetwork. We visualized the attentions, representations and difference image output by the CL encoders, as well as the representations and difference image output by the CL heads (Fig. 12). Examining all the attentions and representations in Fig. 12 (especially lasvegas), we obtained the following insights.

The attentions and the representations output by CL encoders pay more attention to local information and high-level semantics; object-level and edge features are prominent. This differs from the MR subnetwork, which pays more attention to global information and low-level details. Certain specific land classes relevant to changes (such as buildings, cultivated land, and bare land) and changed regional information are emphasized within the representations. Distinctions among changed objects are conspicuously highlighted in the difference image, particularly when generated by the CL heads. Evidently, these intermediate outputs play a pivotal role in enhancing the detection of changed information.

C. Effectiveness of optimal sampling

We compared the visualizations and MCD results with or without optimal sampling to evaluate the effectiveness of optimal sampling. The comparisons are shown in Fig. 13, and we can get the following findings.

1) After using the optimal sampling strategy, the generated representations have fewer misclassifications for the foreground and background, higher quality for object segmentation and consistency, and more prominent edge information; 2) the difference image not only highlights the changed objects but also significantly suppresses the unchanged ones; 3) for the Lasvegas case, the F1 with and without optimal sampling are 0.872 and 0.818, respectively, and for the Shuguang dataset, the F1 with and without optimal sampling are 0.530 and 0.453, respectively. These findings manifest that optimal sampling is of great gain for the task of MCD.

D. Effectiveness of silent attention

We tested the effectiveness of silent attention in the MR encoders and CL encoders, as shown in Fig. 14. We find that silent attention focuses on high-correlation objects but suppresses low-correlation ones, thereby expanding the contrast in output representations and making it easier to distinguish changed objects, compared to vanilla attention.

Additionally, we find that silent attention has more minor fluctuations during learning and overcomes the problem of collapse that appeared in vanilla attention on the Lasvegas case, as shown in Fig. 15. These results reveal that silent attention is preferable and more robust than vanilla attention.

VI. Conclusion

In this paper, we proposed a novel MaCon framework for unsupervised multimodal change detection. This framework ingeniously integrates the two self-supervised learning paradigms of MR and CL, harnessing their respective strengths synergistically. The MR subnetwork pays more attention to global information and low-level details, distilling common representations, while the CL subnetwork emphasizes local information and high-level semantics, extracting discrepancy representations. Additionally, we introduced an optimal sampling strategy to select more reasonable samples, thereby guiding the model to generate more distinguishable disparity. Moreover, we developed silent attention, a plug-and-play module that addresses the inability of traditional attention to assign negligible scores to irrelevant tokens. This advancement improves the differentiation in output representations.

Experimental evaluation indicated that the MaCon framework possesses strong generality on both multimodal and monomodal datasets; it outperforms existing SOTA methods and even exceeds the capabilities of certain supervised approaches. Interpretability experiments were conducted to understand the workings of the MaCon framework.

This study focused on two-dimensional multimodal Earth observation images, a domain with extremely wide application potential. Nevertheless, the need for change detection extends beyond Earth observation to include diverse applications such as high-definition maps, street view maintenance and medical imaging diagnostics. Investigating the applicability of the

MaCon framework to other fields and additional modalities presents an exciting avenue for future research.

ACKNOWLEDGMENT

The authors would like to thank the editors and anonymous reviewers for their in-depth reading and constructive comments and suggestions.

REFERENCES

[1] S. Tian, X. Tan, A. Ma, Z. Zheng, L. Zhang, and Y. Zhong, "Temporal-agnostic change region proposal for semantic change detection," *ISPRS J. Photogramm. Remote Sens.*, vol. 204, pp. 306–320, Oct. 2023.

[2] H. Su, X. Zhang, Y. Luo, C. Zhang, X. Zhou, and P. M. Atkinson, "Nonlocal feature learning based on a variational graph auto-encoder network for small area change detection using SAR imagery," *ISPRS J. Photogramm. Remote Sens.*, vol. 193, pp. 137–149, Nov. 2022.

[3] M. Hu, C. Wu, and L. Zhang, "HyperNet: Self-Supervised Hyperspectral Spatial–Spectral Feature Understanding Network for Hyperspectral Change Detection," *IEEE Trans. Geosci. Remote Sens.*, vol. 60, pp. 1–17, 2022.

[4] Z. Zheng, A. Ma, L. Zhang, and Y. Zhong, "Change is Everywhere: Single-Temporal Supervised Object Change Detection in Remote Sensing Imagery." arXiv, Aug. 11, 2022.

[5] Y. Wang *et al.*, "Spectral–Spatial–Temporal Transformers for Hyperspectral Image Change Detection," *IEEE Trans. Geosci. Remote Sens.*, vol. 60, pp. 1–14, 2022.

[6] X. Li, L. Yan, Y. Zhang, and N. Mo, "SDMNet: A Deep-Supervised Dual Discriminative Metric Network for Change Detection in High-Resolution Remote Sensing Images," *IEEE Geosci. Remote Sens. Lett.*, vol. 19, pp. 1–5, 2022.

[7] L. Yan, J. Yang, and Y. Zhang, "Building Instance Change Detection from High Spatial Resolution Remote Sensing Images Using Improved Instance Segmentation Architecture," *J. Indian Soc. Remote Sens.*, Sep. 2022.

[8] L. Ji, J. Zhao, and Z. Zhao, "A Novel End-to-End Unsupervised Change Detection Method with Self-Adaptive Superpixel Segmentation for SAR Images," *Remote Sens.*, vol. 15, no. 7, Art. no. 7, Jan. 2023.

[9] L. Yan, J. Yang, and J. Wang, "Domain Knowledge-Guided Self-Supervised Change Detection for Remote Sensing Images," *IEEE J. Sel. Top. Appl. Earth Obs. Remote Sens.*, vol. 16, pp. 4167–4179, 2023.

[10] L. Yan, J. Yang, Y. Zhang, A. Zhao, and X. Li, "Radiometric Normalization for Cross-Sensor Optical Gaofen Images with Change Detection and Chi-Square Test," *Remote Sens.*, vol. 13, no. 16, p. 3125, Aug. 2021.

[11] J. Wang *et al.*, "Deriving mining-induced 3-D deformations at any moment and assessing building damage by integrating single InSAR interferogram and gompertz probability integral model (SII-GPIM)," *IEEE Trans. Geosci. Remote Sens.*, vol. 60, pp. 1–17, 2022.

[12] F. Ye, Z. Wu, X. Jia, J. Chanussot, Y. Xu, and Z. Wei, "Bayesian Nonlocal Patch Tensor Factorization for Hyperspectral Image Super-Resolution," *IEEE Trans. Image Process.*, vol. 32, pp. 5877–5892, 2023.

[13] L. Yan, J. Huang, H. Xie, P. Wei, and Z. Gao, "Efficient Depth Fusion Transformer for Aerial Image Semantic Segmentation," *Remote Sens.*, vol. 14, no. 5, Art. no. 5, Jan. 2022.

[14] R. Touati, M. Mignotte, and M. Dahmane, "Multimodal Change Detection in Remote Sensing Images Using an Unsupervised Pixel Pairwise-Based Markov Random Field Model," *IEEE Trans. Image Process.*, vol. 29, pp. 757–767, 2020.

[15] Y. Wang *et al.*, "Mask DeepLab: End-to-end image segmentation for change detection in high-resolution remote sensing images," *Int. J. Appl. Earth Obs. Geoinformation*, vol. 104, p. 102582, Dec. 2021.

[16] L. Yan, and X. Li, "MAFNet: A Multi-Path Asymmetric Fusion Network for Metric-Based Change Detection in High-Resolution Remote Sensing Images," *Acta Electronica Sinica*., vol. 51, no. 7, pp. 1781–1790, 2023.

[17] Y. Sun, L. Lei, D. Guan, J. Wu, G. Kuang, and L. Liu, "Image Regression With Structure Cycle Consistency for Heterogeneous Change Detection," *IEEE Trans. Neural Netw. Learn. Syst.*, pp. 1–15, 2022.

[18] P. Ebel, S. Saha, and X. X. Zhu, "FUSING MULTI-MODAL DATA FOR SUPERVISED CHANGE DETECTION," *Int. Arch. Photogramm. Remote Sens. Spat. Inf. Sci.*, vol. XLIII-B3-2021, pp. 243–249, Jun. 2021.

[19] S. Saha, M. Shahzad, P. Ebel, and X. X. Zhu, "Supervised Change Detection Using Prechange Optical-SAR and Postchange SAR Data," *IEEE J. Sel. Top. Appl. Earth Obs. Remote Sens.*, vol. 15, pp. 8170–8178, 2022.

[20] M. Yang, L. Jiao, B. Hou, F. Liu, and S. Yang, "Selective Adversarial Adaptation-Based Cross-Scene Change Detection Framework in Remote Sensing Images," *IEEE Trans. Geosci. Remote Sens.*, vol. 59, no. 3, pp. 2188–2203, Mar. 2021.

[21] S. Hafner, Y. Ban, and A. Nascetti, "Unsupervised domain adaptation for global urban extraction using Sentinel-1 SAR and Sentinel-2 MSI data," *Remote Sens. Environ.*, vol. 280, p. 113192, Oct. 2022.

[22] R. Touati and M. Mignotte, "An Energy-Based Model Encoding Nonlocal Pairwise Pixel Interactions for Multisensor Change Detection," *IEEE Trans. Geosci. Remote Sens.*, vol. 56, no. 2, pp. 1046–1058, Feb. 2018.

[23] B. Ayhan and C. Kwan, "A New Approach to Change Detection Using Heterogeneous Images," in *2019 IEEE 10th Annual Ubiquitous Computing, Electronics & Mobile Communication Conference (UEMCON)*, Oct. 2019., pp. 0192–0197.

[24] C. Kwan, "Methods and Challenges Using Multispectral and Hyperspectral Images for Practical Change Detection Applications," *Information*, vol. 10, no. 11, Art. no. 11, Nov. 2019.

[25] L. Wan, Y. Xiang, and H. You, "An Object-Based Hierarchical Compound Classification Method for Change Detection in Heterogeneous Optical and SAR Images," *IEEE Trans. Geosci. Remote Sens.*, vol. 57, no. 12, pp. 9941–9959, Dec. 2019.

[26] T. Han, Y. Tang, X. Yang, Z. Lin, B. Zou, and H. Feng, "Change Detection for Heterogeneous Remote Sensing Images with Improved Training of Hierarchical Extreme Learning Machine (HELM)," *Remote Sens.*, vol. 13, no. 23, p. 4918, Dec. 2021.

[27] Z. Wu *et al.*, "Scheduling-Guided Automatic Processing of Massive Hyperspectral Image Classification on Cloud Computing Architectures," *IEEE Trans. Cybern.*, vol. 51, no. 7, pp. 3588–3601, Jul. 2021.

[28] Y. Sun, L. Lei, D. Guan, and G. Kuang, "Iterative Robust Graph for Unsupervised Change Detection of Heterogeneous Remote Sensing Images," *IEEE Trans. Image Process.*, vol. 30, pp. 6277–6291, 2021.

[29] L. T. Luppino, F. M. Bianchi, G. Moser, and S. N. Anfinsen, "Unsupervised Image Regression for Heterogeneous Change Detection," *IEEE Trans. Geosci. Remote Sens.*, vol. 57, no. 12, pp. 9960–9975, Dec. 2019.

[30] R. Touati, M. Mignotte, and M. Dahmane, "Anomaly Feature Learning for Unsupervised Change Detection in Heterogeneous Images: A Deep Sparse Residual Model," *IEEE J. Sel. Top. Appl. Earth Obs. Remote Sens.*, vol. 13, pp. 588–600, 2020.

[31] L. T. Luppino *et al.*, "Deep Image Translation With an Affinity-Based Change Prior for Unsupervised Multimodal Change Detection," *IEEE Trans. Geosci. Remote Sens.*, vol. 60, pp. 1–22, 2022.

[32] P. Zhang, M. Gong, L. Su, J. Liu, and Z. Li, "Change detection based on deep feature representation and mapping transformation for multi-spatial-resolution remote sensing images," *ISPRS J. Photogramm. Remote Sens.*, vol. 116, pp. 24–41, Jun. 2016.

[33] X. Jiang, G. Li, Y. Liu, X.-P. Zhang, and Y. He, "Change Detection in Heterogeneous Optical and SAR Remote Sensing Images Via Deep Homogeneous Feature Fusion," *IEEE J. Sel. Top. Appl. Earth Obs. Remote Sens.*, vol. 13, pp. 1551–1566, 2020.

[34] J. Liu, M. Gong, K. Qin, and P. Zhang, "A Deep Convolutional Coupling Network for Change Detection Based on Heterogeneous Optical and Radar Images," *IEEE Trans. Neural Netw. Learn. Syst.*, vol. 29, no. 3, pp. 545–559, Mar. 2018.

[35] M. Mignotte, "A Fractal Projection and Markovian Segmentation-Based Approach for Multimodal Change Detection," *IEEE Trans. Geosci. Remote Sens.*, vol. 58, no. 11, pp. 8046–8058, Nov. 2020.

[36] Z. Liu, G. Li, G. Mercier, Y. He, and Q. Pan, "Change Detection in Heterogenous Remote Sensing Images via Homogeneous Pixel Transformation," *IEEE Trans. Image Process.*, vol. 27, no. 4, pp. 1822–1834, Apr. 2018.

[37] T. Zhan, M. Gong, X. Jiang, and S. Li, "Log-Based Transformation Feature Learning for Change Detection in Heterogeneous Images," *IEEE Geosci. Remote Sens. Lett.*, vol. 15, no. 9, pp. 1352–1356, Sep. 2018.

[38] Y. Sun, L. Lei, X. Tan, D. Guan, J. Wu, and G. Kuang, "Structured graph based image regression for unsupervised multimodal change detection," *ISPRS J. Photogramm. Remote Sens.*, vol. 185, pp. 16–31, Mar. 2022.

[39] Y. Chen and L. Bruzzone, "Self-Supervised Change Detection in Multiview Remote Sensing Images," *IEEE Trans. Geosci. Remote Sens.*, vol. 60, pp. 1–12, 2022.

[40] S. Saha, P. Ebel, and X. X. Zhu, "Self-Supervised Multisensor Change Detection," *IEEE Trans. Geosci. Remote Sens.*, vol. 60, pp. 1–10, 2022.

[41] P. Akiva, M. Purri, and M. Leotta, "Self-Supervised Material and Texture Representation Learning for Remote Sensing Tasks," in *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, New Orleans, LA, USA: IEEE, Jun. 2022., pp. 8193–8205.

[42] J. Liang *et al.*, "Adapting Language-Audio Models as Few-Shot Audio Learners," in *INTERSPEECH 2023*, ISCA, Aug. 2023., pp. 276–280.

[43] K. Song, J. Xie, S. Zhang, and Z. Luo, "Multi-Mode Online Knowledge Distillation for Self-Supervised Visual Representation Learning," in *2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, Vancouver, BC, Canada: IEEE, Jun. 2023., pp. 11848–11857.

[44] M. Assran *et al.*, "Self-Supervised Learning from Images with a Joint-Embedding Predictive Architecture," in *2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, Vancouver, BC, Canada: IEEE, Jun. 2023., pp. 15619–15629.

[45] Z. Xie *et al.*, "SimMIM: A Simple Framework for Masked Image Modeling." arXiv, Apr. 17, 2022.

[46] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, "BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding," in *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, Minneapolis, Minnesota: Association for Computational Linguistics, Jun. 2019., pp. 4171–4186.

[47] D. Wang, J. Zhang, B. Du, G.-S. Xia, and D. Tao, "An Empirical Study of Remote Sensing Pretraining," *IEEE Trans. Geosci. Remote Sens.*, pp. 1–1, 2022.

[48] X. Sun *et al.*, "RingMo: A Remote Sensing Foundation Model with Masked Image Modeling," *IEEE Trans. Geosci. Remote Sens.*, pp. 1–1, 2022.

[49] D. Hong *et al.*, "SpectralGPT: Spectral Remote Sensing Foundation Model," *IEEE Trans. Pattern Anal. Mach. Intell.*, pp. 1–18, 2024.

[50] D. Y. Fu, M. F. Chen, M. Zhang, K. Fatahalian, and C. Ré, "The Details Matter: Preventing Class Collapse in Supervised Contrastive Learning," in *AAAI Workshop on Artificial Intelligence with Biased or Scarce Data (AIBSD)*, MDPI, Apr. 2022., p. 4.

[51] X. Chen, H. Fan, R. Girshick, and K. He, "Improved Baselines with Momentum Contrastive Learning." arXiv, Mar. 09, 2020.

[52] Q. Hu, X. Wang, W. Hu, and G.-J. Qi, "AdCo: Adversarial Contrast for Efficient Learning of Unsupervised Representations from Self-Trained Negative Adversaries." arXiv, Mar. 05, 2021.

[53] T. Chen, S. Kornblith, M. Norouzi, and G. Hinton, "A Simple Framework for Contrastive Learning of Visual Representations." arXiv, Jun. 30, 2020.

[54] W. Li, K. Chen, H. Chen, and Z. Shi, "Geographical Knowledge-Driven Representation Learning for Remote Sensing Images," *IEEE Trans. Geosci. Remote Sens.*, vol. 60, pp. 1–16, 2022.

[55] U. Mall, B. Hariharan, and K. Bala, "Change-Aware Sampling and Contrastive Learning for Satellite Images," in *2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, Vancouver, BC, Canada: IEEE, Jun. 2023., pp. 5261–5270.

[56] Y. Zhai *et al.*, "Investigating the Catastrophic Forgetting in Multimodal Large Language Models." arXiv, Oct. 03, 2023.

[57] D. Wang *et al.*, "Advancing Plain Vision Transformer Towards Remote Sensing Foundation Model," *IEEE Trans. Geosci. Remote Sens.*, pp. 1–1, 2022.

[58] K. He, X. Chen, S. Xie, Y. Li, P. Dollár, and R. Girshick, "Masked Autoencoders Are Scalable Vision Learners." arXiv, Dec. 19, 2021.

[59] Y. Chen and L. Bruzzone, "Self-Supervised SAR-Optical Data Fusion of Sentinel-1/-2 Images," *IEEE Trans. Geosci. Remote Sens.*, vol. 60, pp. 1–11, 2022.

[60] J. Robinson, C.-Y. Chuang, S. Sra, and S. Jegelka, "Contrastive Learning with Hard Negative Samples." arXiv, Jan. 24, 2021.

[61] Y. Wang, C. Albrecht, N. Ait Ali Braham, L. Mou, and X. Zhu, "Self-Supervised Learning in Remote Sensing: A Review," *IEEE Geosci. Remote Sens. Mag.*, pp. 2–36, 2022.

[62] R. C. Daudt, B. Le Saux, A. Boulch, and Y. Gousseau, "Urban Change Detection for Multispectral Earth Observation Using Convolutional Neural Networks," in *IGARSS 2018 - 2018 IEEE International Geoscience and Remote Sensing Symposium*, 2018., pp. 2115–2118.

[63] M. Zhang and W. Shi, "A Feature Difference Convolutional Neural Network-Based Change Detection Method," *IEEE Trans. Geosci. Remote Sens.*, vol. 58, no. 10, pp. 7232–7246, 2020.

[64] F. Liu, L. Jiao, X. Tang, S. Yang, W. Ma, and B. Hou, "Local Restricted Convolutional Neural Network for Change Detection in Polarimetric SAR Images," *IEEE Trans. Neural Netw. Learn. Syst.*, vol. 30, no. 3, pp. 818–833, Mar. 2019.

[65] A. Dosovitskiy *et al.*, "An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale." arXiv, Jun. 03, 2021.

[66] Y. Sun, L. Lei, X. Li, H. Sun, and G. Kuang, "Nonlocal patch similarity based heterogeneous remote sensing change detection," *Pattern Recognit.*, vol. 109, p. 107598, Jan. 2021.

[67] C. Kwan, B. Ayhan, J. Larkin, L. Kwan, S. Bernabé, and A. Plaza, "Performance of Change Detection Algorithms Using Heterogeneous Images and Extended Multi-attribute Profiles (EMAPs)," *Remote Sens.*, vol. 11, no. 20, Art. no. 20, Jan. 2019.

[68] C. Wu, B. Du, and L. Zhang, "Fully Convolutional Change Detection Framework With Generative Adversarial Network for Unsupervised, Weakly Supervised and Regional Supervised Change Detection," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 45, no. 8, pp. 9774–9788, 2023.

[69] B. Du, L. Ru, C. Wu, and L. Zhang, "Unsupervised Deep Slow Feature Analysis for Change Detection in Multi-Temporal Remote Sensing Images," *IEEE Trans. Geosci. Remote Sens.*, vol. 57, no. 12, pp. 9976–9992, Dec. 2019.

[70] F. Thonfeld, H. Feilhauer, M. Braun, and G. Menz, "Robust Change Vector Analysis (RCVA) for multi-sensor very high resolution optical satellite data," *Int. J. Appl. Earth Obs. Geoinformation*, vol. 50, pp. 131–140, 2016.

[71] T. Celik, "Unsupervised Change Detection in Satellite Images Using Principal Component Analysis and k-Means Clustering," *IEEE Geosci. Remote Sens. Lett.*, vol. 6, no. 4, pp. 772–776, 2009.

[72] R. C. Daudt, B. L. Saux, and A. Boulch, "Fully Convolutional Siamese Networks for Change Detection." arXiv, Oct. 19, 2018.

[73] Y. Gao, F. Gao, J. Dong, and H.-C. Li, "SAR Image Change Detection Based on Multiscale Capsule Network," *IEEE Geosci. Remote Sens. Lett.*, vol. 18, no. 3, pp. 484–488, 2021.

[74] H. Chen, N. Yokoya, and M. Chini, "Fourier domain structural relationship analysis for unsupervised multimodal change detection," *ISPRS J. Photogramm. Remote Sens.*, vol. 198, pp. 99–114, Apr. 2023.

[75] L. Lei, Y. Sun, and G. Kuang, "Adaptive Local Structure Consistency-Based Heterogeneous Remote Sensing Change Detection," *IEEE Geosci. Remote Sens. Lett.*, vol. 19, pp. 1–5, 2022.

[76] Y. Sun, L. Lei, X. Li, X. Tan, and G. Kuang, "Patch Similarity Graph Matrix-Based Unsupervised Remote Sensing Change Detection With Homogeneous and Heterogeneous Sensors," *IEEE Trans. Geosci. Remote Sens.*, vol. 59, no. 6, pp. 4841–4861, Jun. 2021.

[77] M. Mignotte, "MRF Models Based on a Neighborhood Adaptive Class Conditional Likelihood For Multimodal Change Detection," *AI Comput. Sci. Robot. Technol.*, vol. 2022, pp. 1–20, Mar. 2022.

[78] L. T. Luppino *et al.*, "Code-Aligned Autoencoders for Unsupervised Change Detection in Multimodal Remote Sensing Images," *IEEE Trans. Neural Netw. Learn. Syst.*, pp. 1–13, 2022.

Jian Wang received the M.S. degree in surveying and mapping engineering from China University of Mining and Technology-Beijing, Beijing, China, in 2021.

He is currently pursuing the Ph.D. degree in photogrammetry and remote sensing at the Wuhan University, Wuhan, China. He is also a Visiting Researcher with the Faculty of Science and Technology, and Lancaster Environment Centre, Lancaster University, Lancaster, UK. His research interests include multimodal data fusion, semantic segmentation, spatiotemporal analysis, deep learning, and remote sensing data analysis.

Li Yan received the B.S., M.S., and Ph.D. degrees in photogrammetry and remote sensing from Wuhan University, Wuhan, China, in 1989, 1992, and 1999, respectively.

He is currently a Luojia Distinguished Professor with the School of Geodesy and Geomatics, Wuhan University, Wuhan, China. His research interests include real-time mobile mapping and surveying, multimodal data fusion, remote sensing, 3-D reconstruction, and precise image measurement.

Prof. Yan was a recipient of the high-level scientific and technological innovation talent from the Ministry of Natural Resources of China, the National Excellent Teacher Award, and the Model Teacher of Wuhan University. He has led and participated in dozens of national projects.

Jianbing Yang received the B.S. degree in surveying and mapping engineering from East China University of Technology, Nanchang, China, in 2017. He received the M.S. degree in geodesy and survey engineering from the Institute of Seismology, China Earthquake Administration, Wuwhan, China, in 2020. He received the Ph.D. degree in photogrammetry and remote sensing from Wuhan University, Wuhan, China, in 2023.

He is currently a lecturer with Air Force Early Warning Academy. His research interests include radiometric correction, deep learning, change detection, and image matching.

Hong Xie received the B.S., M.S., and Ph.D. degrees in photogrammetry and remote sensing from Wuhan University, Wuhan, China, in 2007, 2009, and 2013, respectively.

He is currently an Associate Professor with the School of Geodesy and Geomatics, Wuhan University. His research interests include object detection, deep learning, quality improvement for point cloud data, point cloud information extraction, and model reconstruction.

Qiangqiang Yuan received the B.S. degree in surveying and mapping engineering and the Ph.D. degree in photogrammetry and remote sensing from Wuhan University, Wuhan, China, in 2006 and 2012, respectively.

In 2012, he joined the School of Geodesy and Geomatics, Wuhan University, where he is currently a Professor. He has published more than 90 research articles, including more than 70 peer-reviewed articles in international journals, such as *Remote Sensing of Environment*, *ISPRS Journal of Photogrammetry and Remote Sensing*, *IEEE Transactions on Image Processing* and *IEEE Transactions on Geoscience and Remote Sensing*. His research interests include image reconstruction, remote sensing image processing and application, and data fusion.

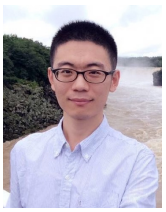
Prof. Yuan was a recipient of the Youth Talent Support Program of China in 2019, the Top-Ten Academic Star of Wuhan University in 2011, and the recognition of Best Reviewers of the IEEE GRSL in 2019. In 2014, he received the Hong Kong Scholar Award from the Society of Hong Kong Scholars and the China National Postdoctoral Council. He is an associate editor of 5 international journals and has frequently served as a referee for more than 40 international top journals, such as *Nature Climate Change, Nature Communications*, etc.

Pengcheng Wei received his Ph.D. degree from the School of Geodesy and Geomatics, Wuhan University, China, in 2023.

He is currently a Lecturer at the School of Civil Engineering and Architecture, China Three Gorges University. His research interests include spatiotemporal data processing and information extraction, computer vision and 3D model reconstruction, autonomous mobile mapping robot development, and intelligent interpretation of remote sensing imagery.

Zhao Gao received the M.S. degree in artificial intelligence from the School of Computer Science, Wuhan University, Wuhan, China, in 2024.

He is currently a Engineer at the Geely Auto Group, China. His research interests include multimodal data fusion, semantic segmentation, point cloud information extraction, computer vision, and remote sensing data analysis.



Ce Zhang received Ph.D. Degree in Geography from Lancaster Environment Centre, Lancaster University, U.K., in 2018.

He is currently a Lecturer in Environmental Data Science at the University of Bristol and a Fellow of UK Centre for Ecology & Hydrology. His major research interests include geospatial artificial intelligence, machine learning, deep learning, and remotely sensed image analysis.

Dr. Zhang was the recipient of a prestigious European Union (EU) Erasmus Mundus Scholarship for a European Joint MSc programme between the University of Twente (The Netherlands) and the University of Southampton (U.K.).

Peter M. Atkinson received the Ph.D. degree from the University of Sheffield (NERC CASE award with Rothamsted Experimental Station), Sheffield, U.K., in 1990, and the M.B.A. degree from the University of Southampton, Southampton, U.K., in 2012.

He is currently a Distinguished Professor of Spatial Data Science and the Executive Dean of the Faculty of Science and Technology, Lancaster University, Lancaster, U.K. He was previously a Professor of Geography at the University Southampton, where he is currently a Visiting Professor. He is also a Visiting Distinguished Professor with Tongji University, Shanghai, China. He previously held the Belle van Zuylen Chair at Utrecht University, Utrecht, The Netherlands. He has published over 400 peer-reviewed articles in international scientific journals and over 50 refereed book chapters. He has also edited 14 journal special issues and eight books. The main methodological *foci* of his research are remote sensing, spatial statistics, geostatistics, machine learning and AI applied to a range of environmental science and socio-economic problems.

Prof. Atkinson is the recipient of a range of awards including the Cuthbert Peek Award of the Royal Geographical Society-Institute of British Geographers and Peter Burrough Award of the International Spatial Accuracy Research Association, and he is a Fellow of the Learned Society of Wales. He is Editor-in-Chief of *Science of Remote Sensing*, a sister journal of *Remote Sensing of Environment*. He also sits on the editorial boards of several further journals, including *Environmetrics*, *Spatial Statistics*, and *Environmental Informatics.*