Recurrent Semantic Change Detection in VHR Remote Sensing Images Using Visual Foundation Models

Jing Zhang, Lei Ding, Tingyuan Zhou, Jian Wang, Peter M. Atkinson and Lorenzo Bruzzone, *Fellow, IEEE*

***Abstract*—Semantic change detection (SCD) involves the simultaneous extraction of changed regions and their corresponding semantic classifications (pre- and post-change) in remote sensing imagery (RSI). Despite recent advancements in Vision Foundation Models (VFMs), the Fast Segment Anything Model has demonstrated insufficient performance in SCD. In this paper, we propose a novel Visual Foundation Model architecture for SCD, designated as VFM-ReSCD. This architecture integrates a side adapter (SA) into the VFM-ReSCD to fine-tune the Fast Segment Anything Model (FastSAM) network, enabling zero-shot transfer to novel image distributions and tasks. This enhancement facilitates the extraction of spatial features from Very High-Resolution (VHR) RSIs. Moreover, we introduce a Recurrent Neural Network (RNN) to model semantic correlation and capture feature changes. We evaluated the proposed methodology on two benchmark datasets. Extensive experiments show that our method achieves state-of-the-art performances over existing approaches and outperforms other CNN-based methods on two RSI datasets.**

***Index Terms*—Remote Sensing, Semantic Change Detection, Visual Foundation Model, Recurrent Neural Network**

I. Introduction

Semantic Change Detection (SCD), also known traditionally as ”detection of land-cover transitions” or ”multiclass change detection” [1] [2] [3], has been reinvigorated through the application of deep learning (DL) methodologies. This is relatively new task with Deep Learning. SCD is attracting growing attention with considerable interest due to Earth Observation (EO) applications. It is essential to identify both the location and the categories of change in the SCD task [4] [5] [6], which combines the two tasks semantic segmentation and change detection: acquire the semantic map and change map, and subsequently overlay and analyze these maps to derive the results map containing the semantic and change information. SCD is very useful and meaningful in EO, as it provides numerous practical benefits in many applications, including resource management, environmental monitoring, analysis of ecosystem change, urban planning, and land-cover and land-use (LCLU) monitoring [4] [7]. In this paper, we investigate employing the latest computer vision methods to exploit more accurate semantic information from bi-temporal images, with the aim of enhancing accuracy.

Previous studies have focused on modeling the multitemporal dependence and utilizing change vector analysis (CVA) [8] to classify changes in different land cover (LC) types. Markov chains were used in the study and prediction of changes using multi-temporal data [9] [10]. Recently, SCD has emerged as an innovative research area that utilizes the potential of deep learning methodologies. For example, in [11], the authors present the first large-scale, very high-resolution SCD dataset and used predicted land-cover information to predict changes through deep learning techniques.

Recent advances in deep learning applications for semantic change detection (SCD) have demonstrated significant potential and effectiveness. Notably, a novel convolutional neural network (CNN) approach introduced by Ding et al. [5] was designed to learn the relationship between bitemporal images using a change detection (CD) unit, which has achieved accurate results in SCD. This development highlights the importance of understanding the temporal differences within image pairs. Based on this, Zhang et al. [6] employed a recurrent neural network (RNN) method, which, unlike the CNN approach, did not use a dedicated CD unit. Instead, their method modeled the correlation between bi-temporal images while simultaneously learning the change features, further expanding the repertoire of deep learning techniques for SCD. To address the challenge of capturing long-distance dependencies, Chen et al. [12] introduced the State Space Model for Change Detection, which excels at capturing the dynamics of change and processing unbalanced datasets. This research laid the foundation for the mamba-based approach in SCD, emphasizing the model's adaptability to datasets with varying class distributions and its ability to track changes over time. Furthermore, in a related development, Ding et al. [13] utilized VFM as an encoder to extract visual representations in remote sensing scenes. They proposed a convolutional adapter to focus on specific ground objects and introduced a task-agnostic semantic learning branch to leverage semantic features in bitemporal remote sensing images (RSIs), thereby enhancing the model's ability to understand and interpret changes in the scene.

VFM can generalize to unfamiliar data distributions and

J. Zhang and L. Bruzzone are with the Department of Information Engineering and Computer Science, University of Trento, 38123 Trento, Italy (E-mail: jing.zhang-1@.unitn.it, lorenzo.bruzzone@unitn.it).

L. Ding is with the Information Engineering University, Zhengzhou, China (E-mail: dinglei14@outlook.com).

Peter M. Atkinson, Ting Y. Zhou, and Jian Wang are with the Lancaster Environment Centre, Lancaster University, Lancaster, LA1 4YQ, UK. Peter is also with the School of Geography and Environmental Science, University of Southampton, Highfield, Southampton, SO17 1BJ, UK (e-mail: rcdzhouty@gmail.com, wj_sgg@whu.edu.cn, pma@lancaster.ac.uk).

This document is funded by the China Scholarship Council under Grant No. 202006540012. It is also supported by the National Natural Science Foundation of China under Grant 42201443. (Corresponding author: Lorenzo Bruzzone.)

exhibit compatibility with various models for more complex tasks. VFM such as the Segment Anything Model [14] have achieved significant performance in segmentation tasks, leading to a new paradigm in various remote sensing technologies. In [15], the SAM pseudo-label optimizer refines the initial pseudo-labels proposed to increase their accuracy. In [16], authors employed SAM to segment agricultural land in the context of research on spatio-temporal fusion. In general, SAM has excellent identification capabilities in segmentation tasks. Thus, we consider to extend fastSAM [17], an efficient version of SAM, to extract visual representation in the SCD domain.

In SCD tasks, it is crucial to explore the relationship between semantic transitions and temporal dependencies. There are two main challenges in SCD tasks. The first challenge involves consideration of semantic relevance in detecting non-salient changes within specific regions, such as the identification of small objects. The challenge is exacerbated by the frequent occurrence of false and missed alarms, which can be attributed to the presence of external noise, as well as variations in the quality of the input images. The second challenge emerges from the discrepancies between the outcomes of SCD tasks and bi-temporal analyses, which directly impact the quality of the results. This highlights the need to properly consider semantic relevance.

This paper investigates the model of nuanced semantic change transition patterns to reduce false alarms and discrepancies. VFM has achieved significant performance in various scenarios for object recognition and segmentation. We introduce VFM as a context encoder for remote sensing SCD, utilizing the parameter-efficient fine-tuning technique known as the side adapter network [18]. The VFM and the side adapter network jointly extract semantic features that contain spatial details. This technique is designed to capture semantic transitions and temporal dependencies. Moreover, we introduce Recurrent Neural Network (RNN) to model the evolution of semantics over time, therefore precisely detecting the semantic changes. Our method exhibits excellent performance in our validation experiments on two benchmark datasets. Our contributions can be summarized as follows.

1) Introducing VFM to the task of SCD. The proposed VFM-ReSCD leverages the semantic embedding capability of VFM, thus can better identify the LCLU changes.
2) Proposing a side adapter for better adaptation of the VFM semantics. This allows better exploitation of both the general visual context as well as RS-specific knowledge.
3) Proposing a bi-directional RNN block to model the semantic correlations between dual temporal branches, thus achieving a better understanding to the LCLU transition.

This paper has been organized into the following parts. Section II introduces the literature work on CD in RSIs. Section III elaborates on the proposed VFM-ReSCD architecture. Section IV describes the experimental settings and the evaluation metrics. Section V reports the results of an ablation study and comparative experiments. Section VI summarizes this work and draws conclusions.

II. Related Work

SCD is a crucial task in remote sensing, which aims at identifying and analyzing semantic changes in two satellite images taken in the same geographical area. This section reviews previous related work in SCD, including traditional methods and deep learning-based methods, which are introduced separately.

A. Binary Change Detection

Most traditional CD methods are based on image processing techniques, which rely on simple pixel-level differences or spectral analysis to detect changes in images [19] [20]. However, these approaches often struggle with complex scenarios due to their reliance on basic quantitative comparisons of multi-temporal images. To overcome this limitation, deep learning (DL) technologies have emerged as a powerful alternative due to their ability to automatically extract complex features and significantly improve detection accuracy. Specifically, convolutional neural networks (CNNs) have demonstrated excellent performance in image classification and feature extraction, leading to their widespread application in CD for bitemporal images. For example, Daudt et al. [21] proposed a Siamese network based on a fully convolutional network, achieving accurate results in CD tasks. Moreover, Zhang et al. [22] introduced a deeply supervised image fusion network that enhances accuracy and efficiency in high-resolution bitemporal remote sensing images through advanced fusion techniques and deep supervised learning. However, the model is complex with high computational costs. Building on the strengths of deep learning, RNNs [23] have proven well-suited for processing time-series data, as they can model temporal relationships between images to identify change features more accurately. Nevertheless, RNNs have limitations for capturing long distance information. Therefore, transformer-based approaches have introduced additional advancements. For instance, Bandara et al. [24] incorporated self-attention mechanisms and the parallel processing capabilities of transformers to capture spatio-temporal relationships between bi-temporal images, thereby improving the detection accuracy of small objects. Likewise, Chen et al. [25] proposed an object-guided transformer architecture that integrates paired OpenStreetMap (OSM) data and high-resolution optical imagery, combining semantic object-level information with spatial visual features to detect LC changes effectively. In addition to these developments, recent research highlights the potential of visual foundation models [26] [27], which exhibit excellent generalization and model compatibility, enabling them to handle complex CD tasks and perform well under varying data distributions. Moreover, Mamba-based approaches have recently gained attention in CD. For instance, Chen et al. [12] introduced the Mamba model for CD, SCD, and semantic segmentation, offering new research directions for Mamba-based approaches in remote sensing.

B. Semantic Change Detection

Binary change detection generates a change map without category and location information, which limits its practical

applications. With advances in remote sensing technology and deep learning methods, researchers have begun to focus on higher-level representation changes in images, known as SCD. SCD identifies changes in land-cover categories and includes understanding the semantic context underlying these changes. For example, in the context of urban expansion, SCD not only identifies the emergence of new buildings but also infers changes in building types or uses [28]. SCD has attracted much attention for enhancing the represented information. Teppei [29] conducted a study on how to identify differences between two images in a scene and represent them with semantic information. The authors proposed the concept of SCD by combining semantic segmentation and CD. Subsequently, unsupervised CD methods have been developed. For instance, in [30], the authors present Kernel Principal Component Analysis (KPCA) convolution, which extracts features from multi-temporal high-resolution remote sensing images and identifies changes without requiring labeled data by utilizing deep twin networks and polar domain mapping. Similarly, Saha et al. [31] propose Deep Change Vector Analysis (DCVA), a novel framework leveraging CNN features to effectively model spatial context and analyze multitemporal VHR satellite images. While these unsupervised methods have shown promise, they still face challenges in providing sufficient accuracy in complex scenarios. Many researchers have used CNN-based networks to achieve significant outcomes. For example, in [32], the author proposed a deep object-based SCD framework for building damage assessment that addresses semantic inconsistency issues. In [33], a Siamese UNet architecture was used for large-scale SCD, where semantic change maps were generated with only coarse boundary or scarce category information. The network mainly used two encoders and two decoders to share weights and then used a multi-scale atrous convolution unit to enlarge the receptive field and capture multi-scale information. Finally, the authors proposed an attention mechanism and a deep supervision strategy to improve network performance.

Recently, CNN-based methods have been developed for SCD. In [5], two SCD methods with triple embedding branches were introduced. Two branches segment temporal images into LCLU maps, while a CD branch detects the change information. In [34], the triple branch was further extended by introducing gating and weighting designs into the decoders to improve the representations of the features. These works also released benchmark datasets for SCD and task-specific evaluation metrics. In [35], a CNN framework for SCD was proposed, where a Siamese CNN was employed to extract semantic features, and a decoder module was designed to detect changes. The study in [36] focuses on spatial-temporal dependency to enhance the learning of semantic features. It is worth nothing that exploring spatial-temporal dependencies and temporal dependencies for SCD are significant research topics. This research work provides a foundation for the mamba-based approach in SCD.

C. Visual Foundation Models

Incorporating VFM into the SCD domain constitutes a profound advance in technological innovation. These advanced models are capable of capturing a more comprehensive set of features and are significantly more efficient at processing multimodal data and image information at different times, leading to increased accuracy and robustness in SCD. This Challenge is that we rely on the tack of well-annotated training data in remote sensing images for SCD and CD tasks. To address this issue, we found that one of the VFM, called SAM [14], can segment natural images without annotations. Although SAM performs excellently in natural image segmentation, the results obtained by directly extending it to remote sensing are not enough accurate. For example, Chen et al. [37] propose a SAM-based method, taking advantage of SAM excellent zero-shot transfer capability, which enables high-quality optical image segmentation maps to be obtained. In [16] SAM was employed to segment agricultural land in research on spatiotemporal fusion. However, its ability to detect small objects is limited in these tasks. Therefore, In [13], authors consider using the adaptor to fine-tune the VFM to learn semantics in RSIs that a CNN-based fastSAM as the backbone network for CD. In [38] the author presented a dual encoder that combines MobileSAM and CNN, which extracts asymptotic and local features in parallel. However, in these studies, the adapter and CNN are merely employed to fine-tune the backbone network for feature extraction. Differently, in our research, we not only employ an adapter to fine-tune the fastSAM, but also design a RNN to correlate semantic features in bi-temporal images. There are differences between VFM-ReSCD and SAM-CD. The SAM-CD [13] architecture utilizes FastSAM as a frozen encoder, introduces a trainable adapter to improving generalization in RSIs, and incorporates multi-scale features in a UNet-like decoder with a change branch and a semantic learning branch, thus enhancing its semantic awareness for increased object change detection in VHR RSIs. Although an adapter is introduced in SAM-CD, SAM-CD is hardly effective for SCD. There are two reasons: 1) SCD provides full supervision of the LCLU, so more trainable layers lead to higher accuracy. The SAM-CD is generally light-weight and includes only a few trainable parameters. Although its visual encoder, to some extent, extracts the semantic representations, it is not fully trained on VHR RSI, so it does not lead to higher semantic learning accuracy; 2) SCD is a more complicated task than only extracting temporal semantics, as it requires modeling of the correlations of semantic change. Thus, the proposed VFM-ReSCD introduces a side adapter designed specifically for semantic information and an RNN to model semantic correlation and change features.

III. Proposed VFM-ReSCD Architecture for SCD

In this section, we introduce the VFM-ReSCD architecture for SCD. Firstly, we present a summary of the VFM-ReSCD approaches to SCD and introduce a novel task-specific architecture that precisely models the correlation of semantic changes in SCD. Secondly, we propose a bidirectional recurrent neural network (Bi-RNN) approach to learn the features of change from SCD data. Finally, we introduce loss functions in the VFM-ReSCD architecture which serve as the basis for optimizing the performance of our proposed method.

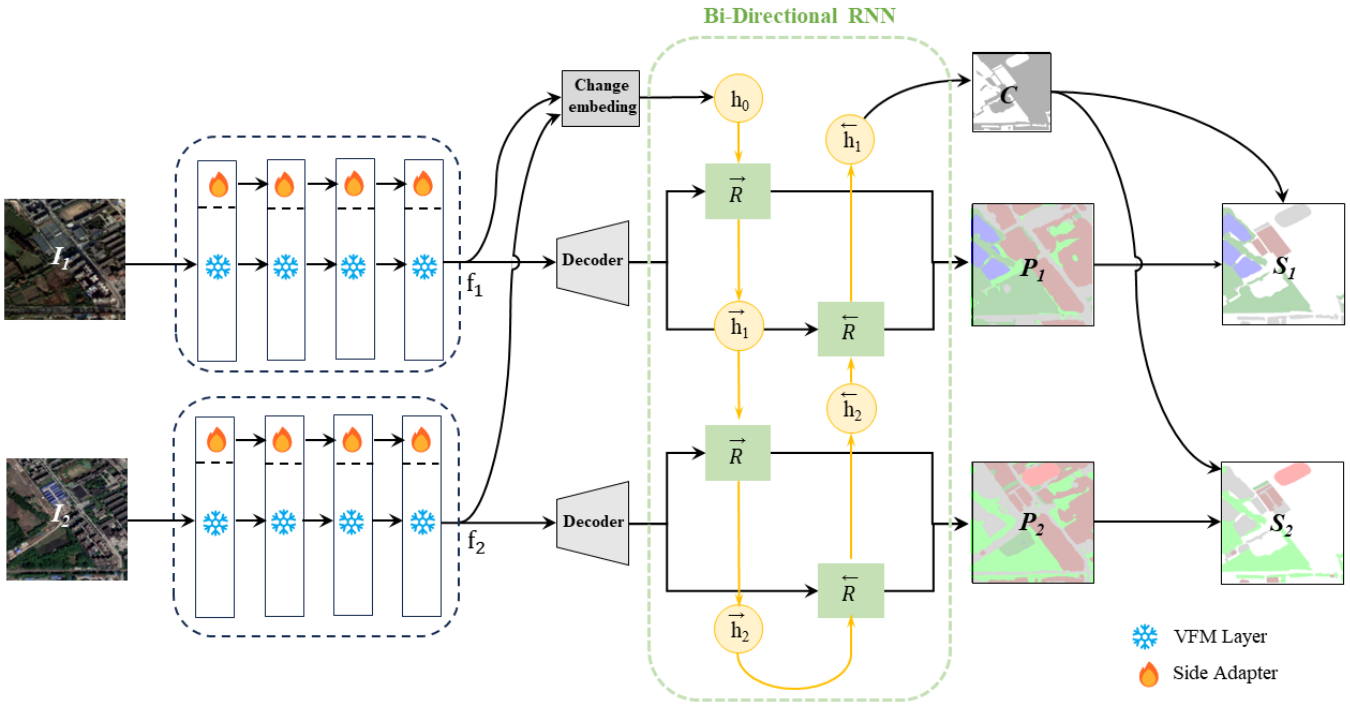

Fig. 1: Architecture of the proposed **VFM**-driven **Re**current **S**emantic **C**hange **D**etection (VFM-ReSCD).

A. Overview

Prior research on SCD tasks has predominantly relied on deep learning methodologies that require labeled data. In contrast, VFMs exhibit superior performance without labels in image segmentation [14], obviating the need for labeled data in SCD tasks. To mitigate excessive computational demands, we propose substituting SAM with fastSAM, a variant of SAM that operates as an encoder for the extraction of semantic features. An overview of the VFM-driven recurrent semantic change detection (VFM-ReSCD) framework is illustrated in Fig.1. The first step involves using a frozen encoder, fastSAM, to extract features from the bi-temporal images. To enhance the extraction of both global and local features, we employ side adapters to fine-tune the fastSAM network. Subsequently, the VFM-ReSCD network is fine-tuned to recover generalized image features through the decoder. Finally, the RNN component leverages modeling to capture the semantic relevance and change features between the bi-temporal branches. This methodology is meticulously designed to effectively discern semantic alterations within SCD tasks. Let (I_1, I_2) indicate bi-temporal images and $\xi(\cdot)$ represent the fastSAM encoder, which is a VFM layer and a side adapter. $\zeta(\cdot)$ represents the decoder network. The relationship between these components can be expressed as follows:

$$f_1 = f_{v_1} + f_{s_1} \tag{1}$$

$$f_2 = f_{v_2} + f_{s_2} \tag{2}$$

$$\mathcal{F}'_C = M_{change} = \zeta[\xi(I_1, I_2)] \tag{3}$$

$$\mathcal{F}'_a = \zeta(f_1), \mathcal{F}'_b = \zeta(f_2) \tag{4}$$

We propose a novel architecture incorporating these components to perform SCD tasks effectively. Here, f_1 and f_2 are the features of the encoder. f_{v_1}, f_{s_1}, f_{v_2}, and f_{s_2} are the features extracted from the VFM layer and the side adapter. M_{change} is change embeding, and $\mathcal{F}'_C$ is the feature of the change branch. After $\xi(\cdot)$ and $\zeta(\cdot)$, we obtain the features $\mathcal{F}'_C$,$\mathcal{F}'_a$ and $\mathcal{F}'_b$, which are the head of the bidirectional RNN network. This is illustrated in Sec.III-D.

B. VFM Encoder

FastSAM, as a foundational model for segmentation tasks, demonstrates robust efficacy in processing natural images [17], but exhibits various limitations in different research domains [39]. A particular challenge lies in the extraction of small and irregular objects. To mitigate these limitations, we introduce a side adapter [18], which has been demonstrated to be adept at fine-tuning region recognition tasks, thus enhancing the performance of SCD.

Firstly, we present FastSAM as an encoder that includes the VFM layer and the side adapter. The VFM layer extracts features at the spatial scales of $1/32, 1/16, 1/8$, and $1/4$, denoted as ν_1,ν_2,ν_3, and ν_4, respectively. While the side adapter adapts the extracted semantic features. Each feature ν_i is processed by a corresponding side adapter san, denoted as:

$$san(\upsilon_i) = \gamma\{bn[conv(\upsilon_i)]\} \tag{5}$$

$$f_i = (\nu_i, san(\upsilon_i)) \tag{6}$$

where $conv$ denotes a 1×1 convolutional layer, bn denotes a batch normalization function, and $\gamma()$ is a RELU function.

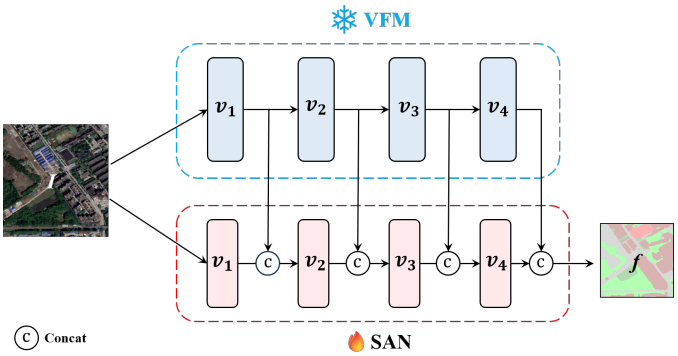

Fig. 2: Architecture of the Side Adapter Network (SAN).

We consider that there are fewer object classes in RSI relative to natural images. Thus, to reduce the feature redundancy, we employ a strategy that reduces the number of f_i channels. The outputs f_1 and f_2 of this encoder provide a rich set of features essential for SCD tasks.

C. Recurrent Neural Network for Semantic Learning

Recurrent neural networks (RNN) have demonstrated superior performance in sharing learning results and learning efficiency. To further increase the model's ability to capture temporal correlation, we propose a bi-directional RNN architecture that models the correlation between bi-temporal images and learns the change features simultaneously. The proposed BiRNN module inputs bitemporal semantic features and memory features into all neural units R, which consequently output augmented semantic and memory features. Let h_0, $\overrightarrow{h_1}$, and $\overrightarrow{h_2}$ be the hidden states in the RNN (h_0, $\overrightarrow{h_1}$, and $\overrightarrow{h_2}$ are learning parameters), which represent memory information. The bidirectional memory information contains both pre-to-post and post-to-pre temporal datasets, ensuring a comprehensive temporal understanding.
The first direction of change is computed as:

$$\overrightarrow{h_1} = f(\overrightarrow{W} * f_1 + \overrightarrow{V} * h_0 + \overrightarrow{b}) \tag{7}$$

$$\overrightarrow{h_2} = f(\overrightarrow{W} * f_2 + \overrightarrow{V} * \overrightarrow{h_1} + \overrightarrow{b}) \tag{8}$$

The second change direction is calculated as:

$$\overleftarrow{h_2} = f(\overleftarrow{W} * f_1 + \overleftarrow{V} * \overrightarrow{h_2} + \overleftarrow{b}) \tag{9}$$

$$\overleftarrow{h_1} = f(\overleftarrow{W} * f_2 + \overleftarrow{V} * \overleftarrow{h_2} + \overleftarrow{b}) \tag{10}$$

Where h_0, $\overrightarrow{h_1}$, and $\overrightarrow{h_2}$ represent the memory (change) information in the time dimension, and the initial value of h_0 is 0. $\overrightarrow{R}$ and $\overleftarrow{R}$ are the calculation units, $\overrightarrow{h_1}$, $\overrightarrow{h_2}$, $\overleftarrow{h_2}$ and $\overleftarrow{h_1}$ are the results of the output from the computing units. $\overrightarrow{W}$, $\overrightarrow{V}$ and b are model parameters. After calculations in both directions, $\overrightarrow{R}$ and $\overleftarrow{R}$ represent the information obtained on the changes in the forward and backward directions, respectively. The last step is to utilize the weight matrix $\overrightarrow{V}$ to fuse $\overrightarrow{R}$ to generate the output feature f_1' and f_2' at the times a and b as:

$$f_1' = g(U[\overrightarrow{h1}; \overleftarrow{h1}] + c) \tag{11}$$

$$f_2' = g(U[\overrightarrow{h2}; \overleftarrow{h2}] + c) \tag{12}$$

where f, g, and U are activation functions, and c is a model parameter (see (3)). Here, $\overrightarrow{h_1}$ and $\overrightarrow{h_2}$ are connected together. Therefore, the proposed architecture can better correlate temporal semantic information and change information. As a result, two different training phases, high-level and low-level, are used to obtain information on the low-level and high-level semantic features that we expect to obtain. Low-level semantic features with relatively lower spatial dependencies are learned from cropped local patches. In contrast, high-level semantic features that do not require accurate localization are learned from a larger perspective. Thus, additional information can be collected to make predictions for large-size images.

Finally, each attention map is analyzed for its relative temporal branching to predict cross-temporal correlation:

$$\tilde{\mathbf{X}}_1 = \hat{\mathbf{X}}_1 + (\mathbf{v}_1 \times \mathbf{A}_2) \tag{13}$$

$$\tilde{\mathbf{X}}_2 = \hat{\mathbf{X}}_2 + (\mathbf{v}_2 \times \mathbf{A}_1) \tag{14}$$

where $\tilde{\mathbf{X}}_1$, $\tilde{\mathbf{X}}_2$ are the enhanced features.

D. Loss Functions

In the training of the VFM-ReSCD, we employ three distinct loss functions: the semantic class loss $\mathcal{L}_{sem}$, the binary change loss $\mathcal{L}_{change}$, and the semantic consistency loss $\mathcal{L}_{sc}$ [5], which is introduced in this study. The semantic loss $\mathcal{L}_{sem}$ is calculated as the multiclass cross-entropy loss between the semantic segmentation results P_1, P_2 and the corresponding ground truth (GT) semantic change maps L_1, L_2. It is calculated on a per-pixel basis as follows:

$$\mathcal{L}_{sem} = -\frac{1}{\eta}\sum_{i=1}^{\eta} \phi_i log(\psi_i) \tag{15}$$

In our model, η represents the total number of semantic classes. Here, ϕ_i and $\psi_i$refer to the GT label and the predicted probability associated with the i class, respectively. The value of η is determined based on the number of LCLU classes. The 'no-change' class (indexed as '0') is deliberately excluded from the loss calculation to encourage the temporal branches to focus on extracting semantic features. The change loss, denoted as $\mathcal{L}_{change}$, quantifies the binary cross-entropy divergence between the predicted binary change map C and a reference change map L_c. L_c is constructed using either L_1 or L_2 by substituting all non-zero labels with a *changed* label (indexed as '1'). The $\mathcal{L}_{change}$ for each pixel is calculated as:

$$\mathcal{L}_{change} = -y_c log(p_c) - (1 - y_c) log(1 - p_c) \tag{16}$$

where y_c and p_c indicate the GT label and the predicted probability of change, respectively. $\mathcal{L}_{sem}$ and $\mathcal{L}_{change}$ are designed to drive the learning of semantic information and CD,

respectively. We extend our proposal by introducing a task-specific semantic consistency loss (SCLoss) to link SS to CD. SCLoss encourages consistent predictions in the *no change* regions and penalizes discrepancies in changed regions. This consistency is useful for integrating bi-temporal semantics and change information in the SCD task. The SCLoss $\mathcal{L}_{sc}$ is calculated between the predicted semantic maps T_1, T_2 and the change map l_c using the Cosine loss function:

$$\mathcal{L}_{sc} = \begin{cases} 1 - cos(T_{x1}, T_{x2}), & l_c = 1 \\ cos(T_{x1}, T_{x2}), & l_c = 0 \end{cases} \tag{17}$$

where T_{x1}, T_{x2} are the feature vectors of a pixel in P_1 and P_2, respectively. The training of the two branches of feature embedding is supervised directly by L_1 and L_2, with further assistance provided by L_c through $\mathcal{L}_{sc}$. Meanwhile, the CD block is directly supervised by L_c. The relationships between the 3 outputs T_1, T_2, C, and the GT maps L_1, L_2, and L_c are crosswise and parallel. The total loss $\mathcal{L}_{scd}$ is calculated as:

$$\mathcal{L}_{scd} = (\mathcal{L}_{sem_{t1}} + \mathcal{L}_{sem_{t2}})/2 + \mathcal{L}_{change} + \mathcal{L}_{sc} \tag{18}$$

Each temporal branch has its semantic loss $\mathcal{L}sem_{t1}$ and $\mathcal{L}sem_{t2}$ indicated in (14). Their summation is computed and averaged to represent $\mathcal{L}_{sem}$. Using L_{sc}, the joint consideration of temporal semantic information from two images can enhance the discrimination of critical areas.

IV. Dataset Description and Experimental Settings

In this section, we describe the dataset, the evaluation metrics, and the experimental settings.

A. Datasets

We perform experiments on two well-established SCD benchmark datasets, namely the SECOND [34] dataset and the Landsat [40] dataset.

SECOND Dataset.

The Semantic Change Detection Dataset (SECOND) is a benchmark dataset for SCD, and it is well-annotated. It is built using bi-temporal high-resolution optical images, which include RGB channels, acquired from various aerial platforms and sensors. These pairs of images are sourced from multiple urban areas, including Hangzhou, Chengdu, and Shanghai. Each image has the same size of 512×512 pixels. The spatial resolution varies from 0.5 to 3 m (per pixel) [41].

The annotation of the SECOND was conducted by a group of experts specialized in Earth vision applications, ensuring a high degree of labeling accuracy. In each GT semantic change map, one change class and six Land Cover (LC) classes are annotated, including *non-vegetated ground surface, tree, low vegetation, water, buildings* and *playgrounds*. These LC classes were selected considering them as common and interesting LC classes and their frequent geographical changes [42]. The bi-temporal LC transitions create a total of 30 LC change types. The changed pixels account for 19.87% of the total image pixels. Among the 4662 pairs of temporal images, 2968 are openly available. We further split them into a training set and a test set with a proportion of $4:1$ (i.e, 2375 image pairs for training, 593 for testing).

Landsat Dataset.

The Landsat-SCD dataset is made up of Landsat images collected between 1990 and 2020. The observation area is Tumshuk, Xinjiang, China. The dataset contains 8468 pairs of images, each of which has a fixed size of 416 × 416 pixels with a spatial resolution of 30 m. The dataset contains a no-change class and four land cover classes, including farmland, desert, buildings, and water. Figure 5 shows sample images from the Landsat-SCD dataset. The dataset contains many complex detection scenes, where the buildings are small and scattered. Changed pixels account for about 19 % of the total, which provides a realistic evaluation dataset for SCD methods.

B. Evaluation Metrics

In this research, three semantic change detection metrics were selected that commonly used to measure performance [11], [43]: overall accuracy (OA), mean intersection over union (mIoU), and Separated Kappa coefficient (SeK). OA has been commonly adopted in both semantic segmentation tasks [44], [45] and CD [11]. Let us denote $Q = \{q_{i,j}\}$ as the confusion matrix where $q_{i,j}$ represents the number of pixels that are classified into class i while their GT index is j ($i, j \in \{0, 1, ..., C\}$ (0 represents *no-change*). OA is calculated as:

$$OA = \sum_{i=0}^{C} q_{ii} / \sum_{i=0}^{C} \sum_{j=0}^{C} q_{ij}. \tag{19}$$

Since OA is mainly determined by the identification of *no-change* pixels, it cannot evaluate LCLU class segmentation well. Additionally, it does not count the pixels that are identified as *changed*, but are predicted into the wrong LCLU classes. Alternatively, mIoU and SeK can evaluate the discrimination of *changed/no-change* regions and the segmentation of LC classes, respectively.

mIoU is the mean value of the IoU of *no-change* regions (IoU_{nc}) and that of the changed regions (IoU_c), i.e.,

$$mIoU = (IoU_{nc} + IoU_c)/2 \tag{20}$$

where:

$$IoU_{nc} = q_{00} / (\sum_{i=0}^{N} q_{i0} + \sum_{j=0}^{N} q_{0j} - q_{00}) \tag{21}$$

$$IoU_c = \sum_{i=1}^{N} \sum_{j=1}^{N} q_{ij} / (\sum_{i=0}^{N} \sum_{j=0}^{N} q_{ij} - q_{00}) \tag{22}$$

Calculation of the SeK coefficient depends on the information provided by the confusion matrix $\hat{Q} = \{\hat{q}_{ij}\}$, where $\hat{q}_{ij} = q_{ij}$ except that $\hat{q}_{00} = 0$. This eliminates true positive *no-change* pixels, whose number is dominant. It is computed as follows:

$$SeK = e^{IoU_c - 1} \cdot (\rho - \eta)/(1 - \eta) \tag{23}$$

where:

$$\rho = \sum_{i=0}^{N} \hat{q}_{ii} / \sum_{i=0}^{N} \sum_{j=0}^{N} \hat{q}_{ij} \tag{24}$$

$$\eta = \sum_{i=0}^{N} (\sum_{j=0}^{N} \hat{q}_{ij} * \sum_{j=0}^{N} \hat{q}_{ji}) / (\sum_{i=0}^{N} \sum_{j=0}^{N} \hat{q}_{ij})^2 \tag{25}$$

The mIoU and SeK directly evaluate the subtasks in SCD, that is the CD and the SS of LCLU classes, respectively. In addition, to more intuitively assess the segmentation of LCLU classes in changed areas, we introduce a new metric F_{scd} (derived from the F_1 score on the segmentation and CD tasks [21] [46]) denoted as follows:

$$F_{scd} = \frac{2 * P_{scd} * R_{scd}}{P_{scd} + R_{scd}} \tag{26}$$

where:

$$P_{scd} = \sum_{i=1}^{N} q_{ii} / \sum_{i=1}^{N} \sum_{j=0}^{N} q_{ij} \tag{27}$$

$$R_{scd} = \sum_{i=1}^{N} q_{ii} / \sum_{i=0}^{N} \sum_{j=1}^{N} q_{ij} \tag{28}$$

Note that P_{scd} and R_{scd} are variants of the $Precision$ and $Recall$ [44] which focus only on the changed areas. F_{scd} describes the segmentation accuracy of the LCLU classes in the changed areas.

Finally, three metrics are provided to measure the computational costs, including the number of parameters (Params), the number of floating-point operations (FLOPs), and the inference (Infer) time for 100 epochs. The FLOPs and Infer time are measured by considering the calculations for a pair of input images, each with 512×512 pixels.

C. Implementation Details

The method we propose was implemented with PyTorch. The training process involves 50 epochs. We set the initial learning rate at 0.1 and updated it at each iteration to $0.1 * (1 - iterations/total\_iterations)^{1.5}$. The adapted gradient descent optimization method used was based on Stochastic Gradient Descent (SGD) with Nesterov momentum. The augmentation strategy included repetition and rotation while loading the image pairs. We applied only simple geometric augmentations to the input images, including repetition and random cropping. During inference, we apply a test-time augmentation operation which includes eight times flipping operations to produce more stable prediction results. For more implementation details, readers are encouraged to visit the associated codes at: https://github.com/Gaia0811/VFM-SCD.

V. Experimental Results

In this section, we conduct a series of experiments to evaluate the effectiveness of the proposed SCD approach and its components within the VFM-ReSCD architecture. First, ablation studies are performed to demonstrate quantitatively the contribution of each component in the proposed technique. Next, we present qualitative results obtained from sample test data. Finally, we compare the performance of the proposed methods with state-of-the-art (SOTA) SCD methods.

A. Ablation Study

1) **Quantitative Results**. To assess the effectiveness of the methods proposed in Section 3, we used SSCD-l [5] as a baseline. We perform an ablation study to evaluate the components in the proposed VFM-ReSCDD. The quantitative results are presented in Table I. First, we tested the effectiveness of the VFM-ReSCD by adding it as an auxiliary loss to train the SSCD-l. This increased the precision by around 0. 82% in SeK and 0.76% in F_{scd}, indicating that the semantic embedding of features improved. Taking this method (SSCD-l with SCLoss) as the baseline, we further assessed the performance of each SR block. The Siam-SR blocks on each temporal branch lead to noticeable increases in accuracy (0.4% in $mIoU$ and 0.43% in F_{scd}). Meanwhile, the Cot-SR block that models temporal coherence improves SeK by more than 0. 41% and F_{scd} by more than 0. 46%. This indicates that both the VFM-ReSCD (without SAN and RNN models) and the VFM-ReSCD (with RNN model) improved the semantic embedding of temporal features. At the same time, the former also increases the detection of change information. Then, we evaluate the VFM-ReSCD, which contains SAN and RNN. Compared to the standard SSCD-l, its increases are around 0.38% in OA, 0.78% in $mIoU$, 2.56% in SeK, and 2.65 % in F_{scd}. As demonstrated by these results, a significant increase in accuracy was achieved by incorporating all the proposed components into VFM-ReSCD.

2) **Qualitative Results**. The qualitative results obtained in some testing areas are presented in Fig.3. The prediction maps from left to right are provided for the methods proposed in Table I, which are organized in the sequence of the number of contained components. Compared with the results of the standard SSCD-l, the predicted LC categories after the introduction of the SCLoss and SR blocks are gradually increased. Bi-SRNet exhibits advantages in the discrimination in critical areas. For example, in Fig.3(a2),(b1),(c1) and (c2), identification of the *ground*, *low vegetation* and *tree* classes is greatly increased.

Through this ablation study, we find that: i) all the tested auxiliary components demonstrate increases in semantic embedding, as evidenced by the increases in SeK values; and ii) the semantic reasoning designs in the VFM-ReSCD increase not only the discrimination of LCLU categories but also the detection of changes.

Computational Costs Table III reports the cost of computational resources of different methods. The size of model parameters (Params) and the nomber of floating point operations (FLOPs) are used to measure the computational efficiency of various SCD methods. Amoung the compared methods, the SCDNet [33] has the highest Params. The FC-Siam-Conc and FC-Siam-diff have the lowest costs due to their simplified architectures. The proposed VFM-ReSCD utilizing VFM increases the Params and FLOPs compared to the baseline method (SSCD-l). However, since we adopt an efficient VFM (FastSAM) as the backbone encoder, the increases in computational cost is only marginal.

3) **SAM and fastSAM** In this section, we examine the

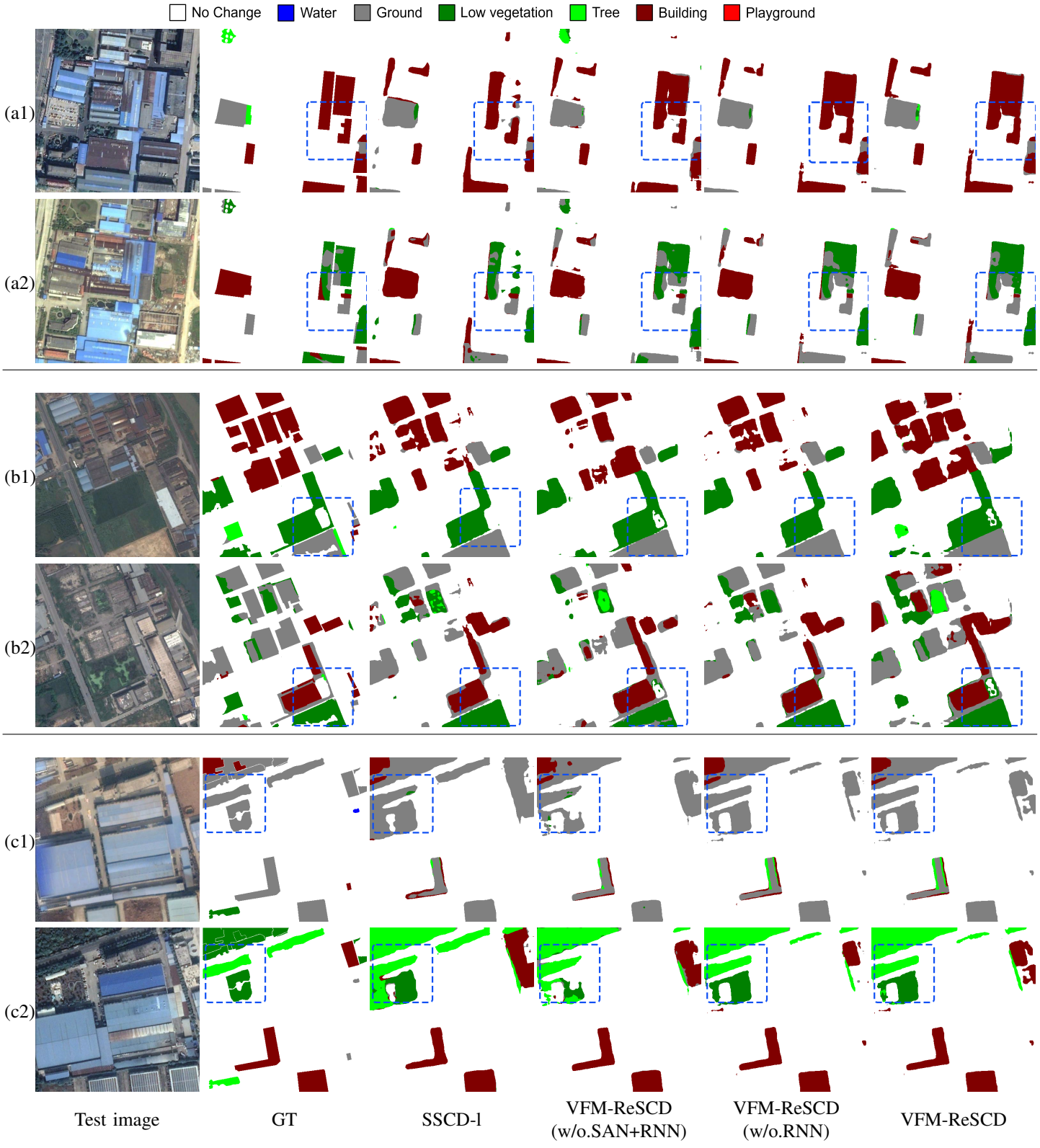

Fig. 3: Example of results provided by different proposed methods in the ablation study on the SECOND dataset. The major differences are highlighted in blue rectangles.

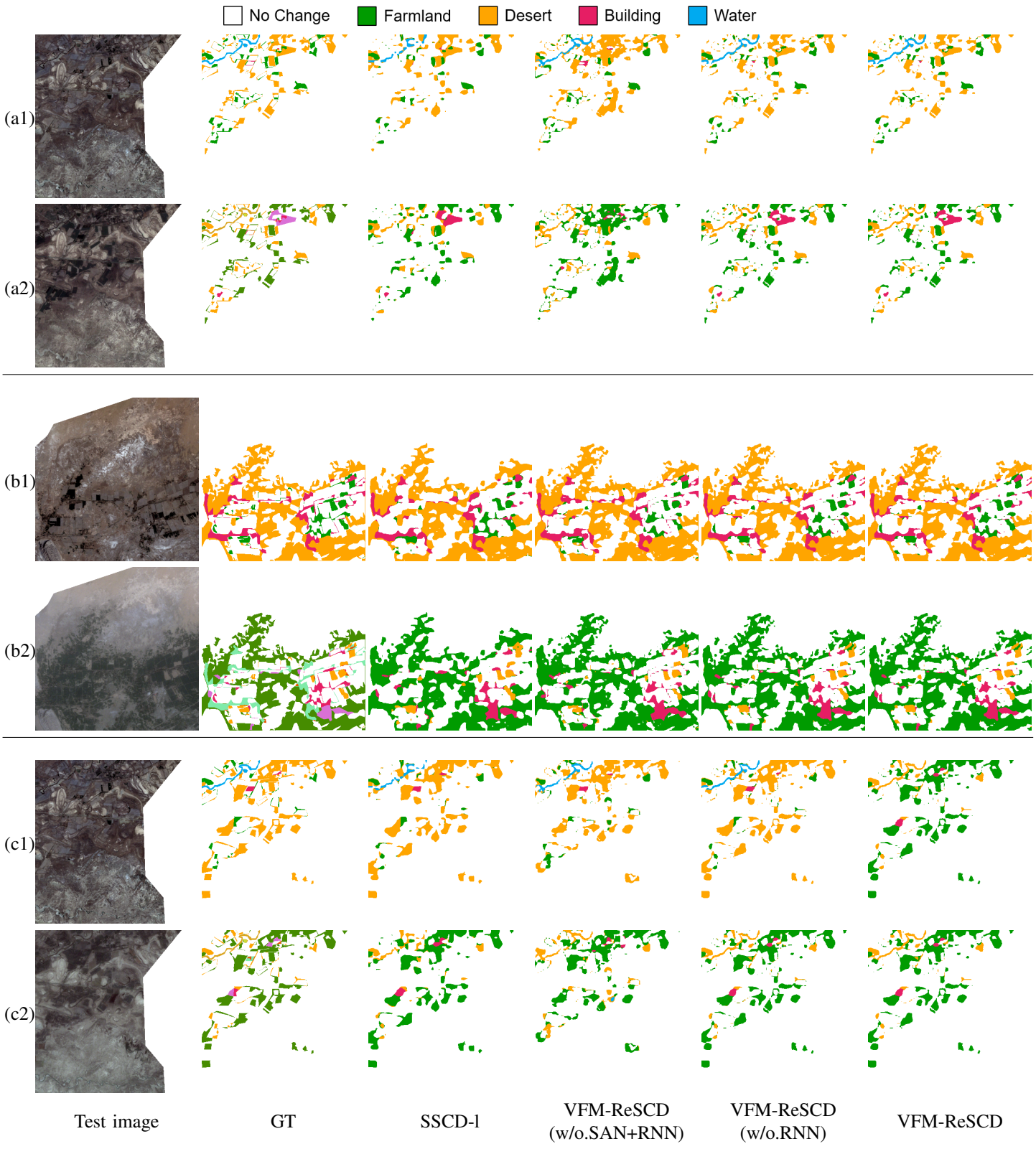

Fig. 4: Example of results provided by different proposed methods in the ablation study on the Landsat-SCD dataset.

TABLE I: Quantitative results obtained in the ablation study.

Proposed Method	Components		SECOND				Landsat-SCD			
	SAN	RNN	$OA(\%)$	$mIoU(\%)$	$Sek(\%)$	$F_{scd}(\%)$	$OA(\%)$	$mIoU(\%)$	$Sek(\%)$	$F_{scd}(\%)$
SSCD-l (baseline)			87.28	72.55	21.45	61.62	94.75	85.25	50.17	84.91
VFM-ReSCD(w/o.SAN)		√	87.12	71.69	19.68	59.13	93.98	83.90	45.94	82.41
VFM-ReSCD(w/o.RNN)	√		87.50	73.28	23.23	63.21	95.80	88.03	57.32	87.74
VFM-ReSCD(w/o.SAN+RNN)	√	√	**87.66**	**73.33**	**24.01**	**64.27**	**95.92**	**88.15**	**58.12**	**88.29**

TABLE II: Comparison between the proposed VFM-ReSCD and some SOTA methods for SCD.

Method	SECOND				Landsat-SCD			
	OA(%)	mIoU(%)	Sek(%)	$F_{scd}(\%)$	OA(%)	mIoU(%)	Sek(%)	$F_{scd}(\%)$
ResNet-GRU [43]	80.79	64.20	8.58	46.47	90.55	74.16	26.51	71.87
FC-Siam-conc [21]	84.65	68.33	16.32	55.28	92.89	79.86	36.94	78.29
FC-Siam-diff [21]	84.34	68.33	16.08	55.16	91.95	76.44	30.23	73.97
HRSCD-str.3 [11]	82.40	64.68	10.24	50.85	91.10	78.33	31.43	73.17
HRSCD-str.4 [11]	85.84	71.16	18.62	58.60	91.27	79.10	32.29	73.34
SCDNet [33]	87.43	70.91	19.79	60.03	94.94	85.23	50.05	85.00
SSCD-l [5]	87.28	72.55	21.45	61.62	94.75	85.25	50.17	84.91
Bi-SRNet [5]	87.60	73.23	23.04	63.12	94.91	85.53	51.01	85.35
SSTNet [47]	-	72.87	22.84	63.18	-	82.94	44.54	81.90
EGMS-Net [48]	86.88	72.89	23.03	62.92	94.53	85.65	51.14	85.47
VFM-ReSCD (proposed)	**87.66**	**73.33**	**24.01**	**64.27**	**95.92**	**88.15**	**58.12**	**88.29**

TABLE III: Comparison of the computation costs of different methods.

Method	Params (Mb)	FLOPs (Gbps)
ResNet-GRU [43]	21.45	182.53
FC-Siam-conv. [21]	1.55	21.78
FC-Siam-diff [21]	1.35	19.36
HRSCD-str.3 [11]	12.77	42.94
HRSCD-str.4 [11]	13.71	43.69
SCDNet [33]	37.09	145.94
SSCD-l [5]	23.31	189.76
Bi-SRNet [5]	23.38	190.30
SSTNet [47]	-	-
EGMS-Net [48]	23.92	216.08
VFM-ReSCD (proposed)	26.09	235.26

TABLE IV: Performance of the VFM-ReSCD using different visual encoders (SECOND dataset).

VFM-ReSCD	Accuracy			
	$OA(\%)$	$mIoU(\%)$	$SeK(\%)$	$F_{csd}(\%)$
SAM-b	87.16	71.92	21.39	60.42
SAM-l	87.24	71.85	21.28	61.05
fastSAM	**87.66**	**73.33**	**24.01**	**64.27**

effectiveness of using different Visual Foundation models (VFMs) for SCD. Since SAM has excellent performance in the semantic segmentation domain, when we use the VFM for SCD tasks, the fastSAM exhibits better results than SAM both in performance and speed. Therefore, we select fastSAM as an encoder in our SCD architecture. The performance of the SAM-CD, equipped with various versions of the SAM and the FastSAM, is detailed in Table IV. The key component is the image encoder in SAM, which uses VIT (Vision Transformer) as the backbone. SAM is characterized by many parameters, large models, and high equipment requirements. To address this issue, we propose a CNN-based fastSAM for SCD.

Since SAM-h is much more resource intensive than SAM-b and SAM-l, our evaluation only tested the accuracy of SAM-b, SAM-l, and fastSAM (fastSAM here is the fastSAM encoder we use in this paper.). It can be observed that the fastSAM that we propose has a significant increase in terms of accuracy, with increase of SeK and F_{scd} of 2.72% and 2.23%, respectively.

In summary, this ablation study elucidates that: i) the proposed SAM framework and associated learning paradigm enhance change detection and facilitate the extraction of semantic information; ii) the VFM-ReSCD sharply advances the exploration of temporal semantic data; and iii) utilizing fastSAM as a foundational network confers a distinct advantage to the proposed approach.

B. Comparative Experiments

To comprehensively evaluate the performance of the proposed VFM-ReSCD architecture, we extended our comparative analysis to include several state-of-the-art (SOTA) methodologies in both CD and SCD tasks.

1) **Quantitative Results**. To quantitatively assess the effectiveness of the proposed methodology, we compare its performances with those of literature SOTA CD and SCD methods.

- ResNet-GRU combines CNN and RNN for CD as derived in [43]. As the methods for low-resolution RSIs with few convolutional layers are unsuitable for HR RSIs, we updated their encoders to ResNet34 [49].
- FC-Siam-conc and FC-siam-diff [21] are Siamese extensions of the FC-EF [21] model, which is based on the UNet network.
- HRSCD-str.3 and HRSCD-str.4 [11], both represent sophisticated methodologies introduced for SCD, incorpo-

This article has been accepted for publication in IEEE Transactions on Geoscience and Remote Sensing. This is the author's version which has not been fully edited and content may change prior to final publication. Citation information: DOI 10.1109/TGRS.2025.3546808

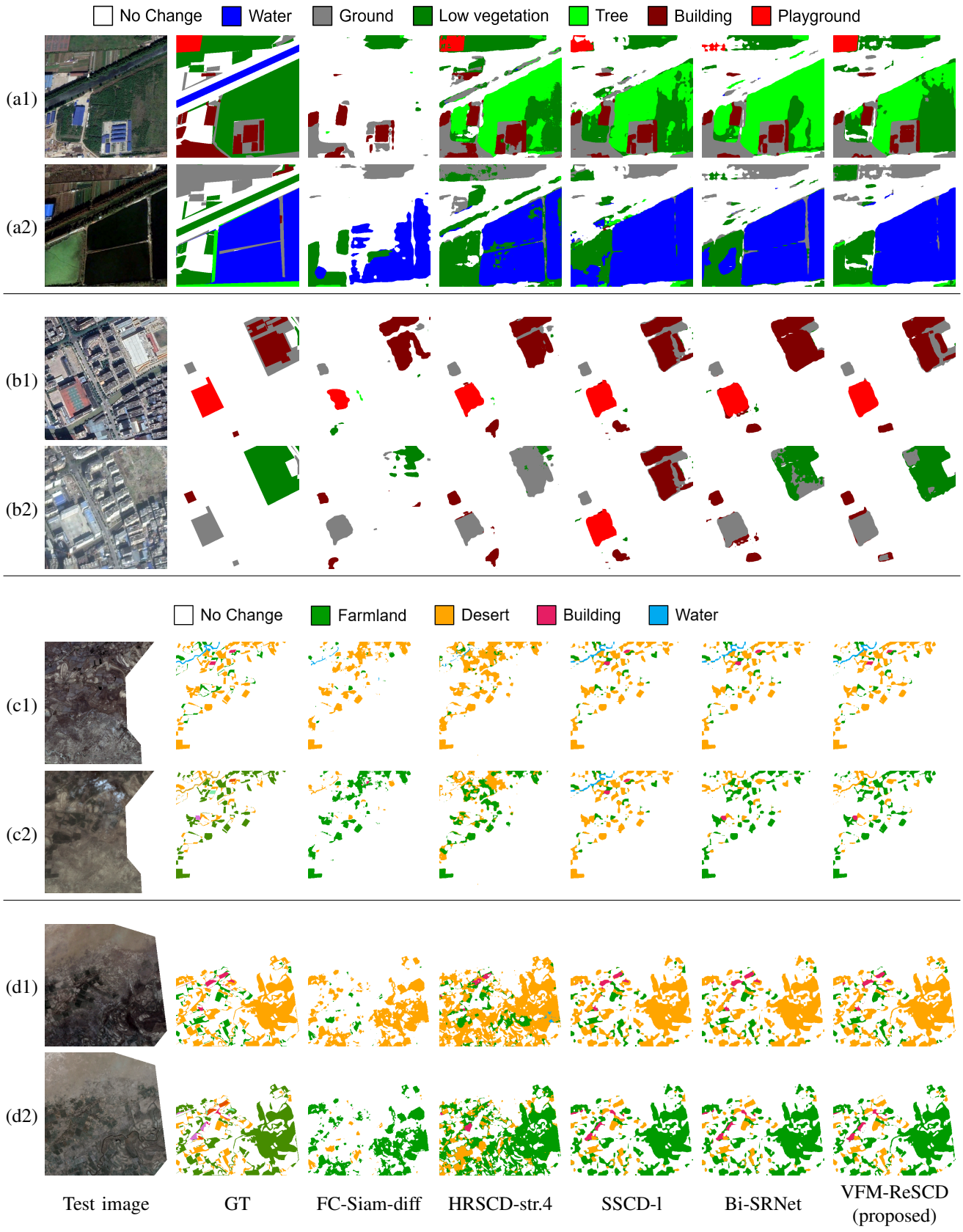

Fig. 5: Example of results provided by different methods in the comparative experiments. (a1)-(b2) results selected from the SECOND dataset, (c1)-(d2) results selected from the Landsat-SCD dataset.

rating residual blocks [49] and complex encoder-decoder architectures. Both methodologies feature triple-encoding branches.

- SCDNet [33] is a Siamese UNet-based architecture that captures multi-scale information and generates detailed change maps to enhance large-scale semantic change detection.
- SSCD-l and Bi-SRNet [5]. Bi-SRNet is based on the foundational SSCD-l architecture and incorporates cross-temporal attention mechanisms.
- SSTNet [47] integrates semantic, spatial and temporal information and improves the detection accuracy of changes in narrow-length objects. It contains a multi-layer feature fusion module and a semantically changing feature interaction module to enhance the exchange and fusion of bi-temporal features.
- EGMS-Net [48] is a multitask Siamese network proposed to enhance SCD in RSIs. It integrates a coarse-to-fine multitask approach, an adaptive change information enhancement method, and a change information guidance module.

Our approach enables profound and intrinsic modeling of spatio-temporal dependencies within the SCD task. Thus, it significantly outperforms compared methods in all metrics. The proposed approach demonstrates an enhancement over the second-best results, with an improvement of approximately 0.98% in SeK and approximately 1.35% in F_{scd} on the SECOND dataset. Its improvements are around 2.5% in $mIoU$, 6.98% in SeK, and 2.94% in F_{scd} on the Landsat dataset.

2) **Qualitative Results**. In Fig.5, we present segmentation maps produced by various methods for visual comparison. The initial four rows in Fig.5(a1)-(b2) depict the results derived on the SECOND data set. One can observe that existing methods struggle to detect non-salient changes, e.g., the emergence of a playground in Fig.5(a) and the removal of small buildings in Fig.5(b). In addition, there are notable inconsistencies within the results. For example, in the SCD results of HRSCD-str.4 and Bi-SRNet, some areas are segmented as *Low vegetation* on both bitemporal segmentation maps, which is contradictory with respect to the represented change information. The results of the proposed methods mainly address these concerns. Through the integration of semantic learning objectives, there is an enhancement in detecting non-salient changes, leading to considerably fewer discrepancies in the bi-temporal results. The VFM-ReSCD further performs in recognizing critical areas, e.g., discrimination between *low vegetation* and *playgroud* in Fig.5(a).

Fig.5(c1)-(d2) present SCD results obtained on the Landsat-SCD dataset. Due to the relatively low GSD of the dataset, the utilization of the SCD model becomes essential for enhancing spatial detail preservation. The proposed methods demonstrate a notable capability to accurately capture nuanced changes within LU types, such as river drying in Fig.5(c), and the emergence of small *farms* in Fig.5(d). The VFM-ReSCD method exhibits advantages in effectively discerning semantic categories attributed to smaller objects.

3) **Change Analysis**. Fig. 3 and 4 present the VFM-ReSCD results on the SECOND and Landsat-SCD data sets. As anticipated, fastSAM performs poorly in detecting semantic changes. To enhance fastSAM for this task, we fine-tune it with a side adapter. This provides significant advantages in the SECOND dataset for buildings, ground, and low vegetation close, and in the Landsat dataset for all objects.

VI. Conclusions

Semantic Change Detection represents a pivotal task in Earth observation. Extracting spatio-temporal changes, analyzing pre- and post-change semantics, and modeling semantic-change correlations are fundamental to the SCD task. We have devised an architecture dedicated to the intricate modeling of spatial-temporal dependencies, thereby, significantly improving semantic-change representations. First, we proposed VFM-ReSCD, which outperforms mamba-based SCD architectures and uses VFM-ReSCD, which is an advanced version of SAM, to capture spatial features from ground objects and temporal constraints in RSIs for task-agnostic semantic representation learning. Thus, we used RNN to model spatiotemporal correlations in terms of semantic representations.

Extensive experiments have been conducted to evaluate the performance of the proposed method. Experiments on two SCD datasets have shown better detection results than other SOTA techniques. Future research will consider the potential effects of new computer vision approaches, such as Vision Mamba and the denoising diffusion model, to improve the accuracy of semantic-change relationships in SCD.

References

[1] L. Bruzzone and S. B. Serpico, "An iterative technique for the detection of land-cover transitions in multitemporal remote-sensing images," *IEEE transactions on geoscience and remote sensing*, vol. 35, no. 4, pp. 858–867, 1997.

[2] L. Bruzzone, R. Cossu, and G. Vernazza, "Detection of land-cover transitions by combining multidate classifiers," *Pattern Recognition Letters*, vol. 25, no. 13, pp. 1491–1500, 2004.

[3] B. Demir, F. Bovolo, and L. Bruzzone, "Detection of land-cover transitions in multitemporal remote sensing images with active-learning-based compound classification," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 50, no. 5, pp. 1930–1941, 2011.

[4] F. Bovolo and L. Bruzzone, "The time variable in data fusion: A change detection perspective," *IEEE Geoscience and Remote Sensing Magazine*, vol. 3, no. 3, pp. 8–26, 2015.

[5] L. Ding, H. Guo, S. Liu, L. Mou, J. Zhang, and L. Bruzzone, "Bi-temporal semantic reasoning for the semantic change detection in hr remote sensing images," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 60, pp. 1–14, 2022.

[6] J. Zhang, L. Ding, and L. Bruzzone, "Bi-directional temporal modelling for semantic change detection in remote sensing images," in *IGARSS 2023-2023 IEEE International Geoscience and Remote Sensing Symposium*. IEEE, 2023, pp. 5503–5506.

[7] R. Caye Daudt, B. Le Saux, A. Boulch, and Y. Gousseau, "Multitask learning for large-scale semantic change detection," *Computer Vision and Image Understanding*, vol. 187, p. 102783, 2019.

[8] D. Lu, P. Mausel, E. Brondizio, and E. Moran, "Change detection techniques," *International journal of remote sensing*, vol. 25, no. 12, pp. 2365–2401, 2004.

[9] H. Nguyen, T. Pham, M. Doan, and P. Tran, "Land use/land cover change prediction using multi-temporal satellite imagery and multi-layer perceptron markov model," *The International Archives of the Photogrammetry, Remote Sensing and Spatial Information Sciences*, vol. 44, pp. 99–105, 2020.

[10] S. A. Mohamed and M. E. El-Raey, "Land cover classification and change detection analysis of qaroun and wadi el-rayyan lakes using multi-temporal remotely sensed imagery," *Environmental monitoring and assessment*, vol. 191, no. 4, p. 229, 2019.

[11] R. C. Daudt, B. Le Saux, A. Boulch, and Y. Gousseau, "Multitask learning for large-scale semantic change detection," *Computer Vision and Image Understanding*, vol. 187, p. 102783, 2019.

[12] H. Chen, J. Song, C. Han, J. Xia, and N. Yokoya, "Changemamba: Remote sensing change detection with spatiotemporal state space model," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 62, pp. 1–20, 2024.

[13] L. Ding, K. Zhu, D. Peng, H. Tang, K. Yang, and L. Bruzzone, "Adapting segment anything model for change detection in vhr remote sensing images," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 62, pp. 1–11, 2024.

[14] A. Kirillov, E. Mintun, N. Ravi, H. Mao, C. Rolland, L. Gustafson, T. Xiao, S. Whitehead, A. C. Berg, W.-Y. Lo *et al.*, "Segment anything," *arXiv preprint arXiv:2304.02643*, 2023.

[15] L. Wang, M. Zhang, and W. Shi, "Cs-wscdnet: Class activation mapping and segment anything model-based framework for weakly supervised change detection," *IEEE Transactions on Geoscience and Remote Sensing*, 2023.

[16] H. Guo, D. Ye, and L. Bruzzone, "Obsum: An object-based spatial unmixing model for spatiotemporal fusion of remote sensing images," *arXiv preprint arXiv:2310.09517*, 2023.

[17] X. Zhao, W. Ding, Y. An, Y. Du, T. Yu, M. Li, M. Tang, and J. Wang, "Fast segment anything," *arXiv preprint arXiv:2306.12156*, 2023.

[18] M. Xu, Z. Zhang, F. Wei, H. Hu, and X. Bai, "Side adapter network for open-vocabulary semantic segmentation," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2023, pp. 2945–2954.

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

[20] L. Bruzzone and D. F. Prieto, "Automatic analysis of the difference image for unsupervised change detection," *IEEE Transactions on Geoscience and Remote sensing*, vol. 38, no. 3, pp. 1171–1182, 2000.

[21] R. C. Daudt, B. Le Saux, and A. Boulch, "Fully convolutional siamese networks for change detection," in *2018 25th IEEE International Conference on Image Processing (ICIP)*. IEEE, 2018, pp. 4063–4067.

[22] C. Zhang, P. Yue, D. Tapete, L. Jiang, B. Shangguan, L. Huang, and G. Liu, "A deeply supervised image fusion network for change detection in high resolution bi-temporal remote sensing images," *ISPRS Journal of Photogrammetry and Remote Sensing*, vol. 166, pp. 183–200, 2020.

[23] H. Chen, C. Wu, B. Du, L. Zhang, and L. Wang, "Change detection in multisource vhr images via deep siamese convolutional multiple-layers recurrent neural network," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 58, no. 4, pp. 2848–2864, 2019.

[24] W. G. C. Bandara and V. M. Patel, "A transformer-based siamese network for change detection," in *IGARSS 2022-2022 IEEE International Geoscience and Remote Sensing Symposium*. IEEE, 2022, pp. 207–210.

[25] H. Chen, C. Lan, J. Song, C. Broni-Bediako, J. Xia, and N. Yokoya, "Objformer: Learning land-cover changes from paired osm data and optical high-resolution imagery via object-guided transformer," *IEEE Transactions on Geoscience and Remote Sensing*, 2024.

[26] K. Li, X. Cao, and D. Meng, "A new learning paradigm for foundation model-based remote-sensing change detection," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 62, pp. 1–12, 2024.

[27] L. Ding, K. Zhu, D. Peng, H. Tang, K. Yang, and L. Bruzzone, "Adapting segment anything model for change detection in vhr remote sensing images," *IEEE Transactions on Geoscience and Remote Sensing*, 2024.

[28] L. Ding, J. Zhang, and L. Bruzzone, "Semantic segmentation of large-size vhr remote sensing images using a two-stage multiscale training architecture," *IEEE Transactions on Geoscience and Remote Sensing*, 2020.

[29] T. Suzuki, S. Shirakabe, Y. Miyashita, A. Nakamura, Y. Satoh, and H. Kataoka, "Semantic change detection with hypermaps," *arXiv preprint arXiv:1604.07513*, 2016.

[30] C. Wu, H. Chen, B. Du, and L. Zhang, "Unsupervised change detection in multitemporal vhr images based on deep kernel pca convolutional mapping network," *IEEE Transactions on Cybernetics*, vol. 52, no. 11, pp. 12 084–12 098, 2021.

[31] S. Saha, F. Bovolo, and L. Bruzzone, "Unsupervised deep change vector analysis for multiple-change detection in vhr images," *IEEE transactions on geoscience and remote sensing*, vol. 57, no. 6, pp. 3677–3693, 2019.

[32] Z. Zheng, Y. Zhong, J. Wang, A. Ma, and L. Zhang, "Building damage assessment for rapid disaster response with a deep object-based semantic change detection framework: From natural disasters to man-made disasters," *Remote Sensing of Environment*, vol. 265, p. 112636, 2021.

[33] D. Peng, L. Bruzzone, Y. Zhang, H. Guan, and P. He, "Scdnet: A novel convolutional network for semantic change detection in high resolution optical remote sensing imagery," *International Journal of Applied Earth Observation and Geoinformation*, vol. 103, p. 102465, 2021.

[34] K. Yang, G.-S. Xia, Z. Liu, B. Du, W. Yang, and M. Pelillo, "Asymmetric siamese networks for semantic change detection," *arXiv preprint arXiv:2010.05687*, 2020.

[35] Q. Zhu, X. Guo, W. Deng, Q. Guan, Y. Zhong, L. Zhang, and D. Li, "Land-use/land-cover change detection based on a siamese global learning framework for high spatial resolution remote sensing imagery," *ISPRS Journal of Photogrammetry and Remote Sensing*, vol. 184, pp. 63–78, 2022.

[36] L. Ding, J. Zhang, H. Guo, K. Zhang, B. Liu, and L. Bruzzone, "Joint spatio-temporal modeling for semantic change detection in remote sensing images," *IEEE Transactions on Geoscience and Remote Sensing*, 2024.

[37] K. Chen, C. Liu, H. Chen, H. Zhang, W. Li, Z. Zou, and Z. Shi, "Rsprompter: Learning to prompt for remote sensing instance segmentation based on visual foundation model," *IEEE Transactions on Geoscience and Remote Sensing*, 2024.

[38] L. Mei, Z. Ye, C. Xu, H. Wang, Y. Wang, C. Lei, W. Yang, and Y. Li, "Scd-sam: Adapting segment anything model for semantic change detection in remote sensing imagery," *IEEE Transactions on Geoscience and Remote Sensing*, 2024.

[39] W. Ji, J. Li, Q. Bi, W. Li, and L. Cheng, "Segment anything is not always perfect: An investigation of sam on different real-world applications," *arXiv preprint arXiv:2304.05750*, 2023.

[40] P. Yuan, Q. Zhao, X. Zhao, X. Wang, X. Long, and Y. Zheng, "A transformer-based siamese network and an open optical dataset for semantic change detection of remote sensing images," *International Journal of Digital Earth*, vol. 15, no. 1, pp. 1506–1525, 2022.

[41] K. Yang, G.-S. Xia, Z. Liu, B. Du, W. Yang, M. Pelillo, and L. Zhang, "Asymmetric siamese networks for semantic change detection in aerial images," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 60, pp. 1–18, 2021.

[42] X.-Y. Tong, G.-S. Xia, Q. Lu, H. Shen, S. Li, S. You, and L. Zhang, "Land-cover classification with high-resolution remote sensing images using transferable deep models," *Remote Sensing of Environment*, vol. 237, p. 111322, 2020.

[43] L. Mou, L. Bruzzone, and X. X. Zhu, "Learning spectral-spatial-temporal features via a recurrent convolutional neural network for change detection in multispectral imagery," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 57, no. 2, pp. 924–935, 2018.

[44] L. Ding, H. Tang, and L. Bruzzone, "Lanet: Local attention embedding to improve the semantic segmentation of remote sensing images," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 59, no. 1, pp. 426–435, 2020.

[45] L. Ding and L. Bruzzone, "Diresnet: Direction-aware residual network for road extraction in vhr remote sensing images," *IEEE Transactions on Geoscience and Remote Sensing*, 2020.

[46] L. Ding, H. Tang, Y. Liu, Y. Shi, X. X. Zhu, and L. Bruzzone, "Adversarial shape learning for building extraction in vhr remote sensing images," *IEEE Transactions on Image Processing*, 2021.

[47] Q. Xia, Y. Yan, W. Hou, B. Ge, N. Su, S. Feng, and C. Zhao, "Sstnet: A network based on modeling of semantic-spatio-temporal information for semantic change detection of remote sensing images," *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, 2024.

[48] X. Zuo, F. Jin, L. Ding, S. Wang, Y. Lin, B. Liu, and Y. Ding, "Multi-task siamese network guided by enhanced change information for semantic change detection in bi-temporal remote sensing images," *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, 2024.

[49] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2016, pp. 770–778.

Jing Zhang received a master's degree in software engineering from Beijing University of Technology. She is currently a Ph.D. student in the Department of Information Engineering and Computer Science at the University of Trento, Italy. Her current research interests include semantic change detection and remote sensing image processing.
She is a referee for many international journals, including the ISPRS Journal of Photogrammetry and Remote Sensing, IEEE Geoscience and Remote Sensing Letters, and Journal of Selected Topics in Applied Earth Observations and Remote Sensing.

Lei Ding received his MS's degree in Photogrammetry and Remote Sensing from the Information Engineering University (Zhengzhou, China), and his PhD (cum laude) in Communication and Information Technologies from the University of Trento (Trento, Italy). He is currently a Lecturer at the Information Engineering University. Since 2024, he has been a Post-doctoral Fellow at the Aerospace Information Research Institute, Chinese Academy of Sciences, Beijing, China. His research interests are related to the intelligent interpretation of remote sensing data.

Tingyuan Zhou received the M.S. degree in Geographic Information Systems (GIS) from the University of Manchester, Manchester, UK, in 2020. He is currently pursuing a Ph.D. in Geography at the Lancaster Environment Centre, Lancaster University, starting from 2023. His research interests include time-series analysis, deep learning, and urban remote sensing data analysis.

Jian Wang 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 Lancaster Environment Centre, Lancaster University, Lancaster, UK. His research interests include multimodal data fusion, semantic segmentation, time-series analysis, deep learning, and remote sensing data analysis.

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 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 nine journal special issues and eight books. The main focus of his research is in remote sensing, geographical information science, and spatial (and space–time) statistics applied to a range of environmental science and socio-economic problems. Prof. Atkinson was a recipient of the Peter Burrough Award of the International Spatial Accuracy Research Association and is a Fellow of the Learned Society of Wales. He is the 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*.

Lorenzo Bruzzone received the Laurea (M.S.) degree in electronic engineering (*summa cum laude*) and the Ph.D. degree in telecommunications from the University of Genoa, Italy, in 1993 and 1998, respectively.
He is currently a Full Professor of telecommunications at the University of Trento, Italy, where he teaches remote sensing, radar, and digital communications. Dr. Bruzzone is the founder and the director of the Remote Sensing Laboratory in the Department of Information Engineering and Computer Science, University of Trento. His current research interests are in the areas of remote sensing, radar and SAR, signal processing, machine learning, and pattern recognition. He promotes and supervises research on these topics within the frameworks of many national and international projects. He is the Principal Investigator of many research projects. Among others, he is the Principal Investigator of the *Radar for icy Moon exploration* (RIME) instrument in the framework of the *JUpiter ICy moons Explorer* (JUICE) mission of the European Space Agency. He is the author (or co-author) of 215 scientific publications in referred international journals (154 in IEEE journals), more than 290 papers in conference proceedings, and 21 book chapters. He is editor/co-editor of 18 books/conference proceedings and 1 scientific book. He was invited as a keynote speaker in more than 30 international conferences and workshops. Since 2009 he has been a member of the Administrative Committee of the IEEE Geoscience and Remote Sensing Society (GRSS).

Dr. Bruzzone was a Guest Co-Editor of many Special Issues of international journals. He is the co-founder of the IEEE International Workshop on the Analysis of Multi-Temporal Remote-Sensing Images (MultiTemp) series and is currently a member of the Permanent Steering Committee of this series of workshops. Since 2003 he has been the Chair of the SPIE Conference on Image and Signal Processing for Remote Sensing. He has been the founder of the IEEE Geoscience and Remote Sensing Magazine for which he has been Editor-in-Chief between 2013-2017. Currently, he is an Associate Editor for the IEEE Transactions on Geoscience and Remote Sensing. He has been Distinguished Speaker of the IEEE Geoscience and Remote Sensing Society between 2012-2016. His papers are highly cited, as proven form the total number of citations (more than 53000) and the value of the h-index (106) (source: Google Scholar).