Bi-Temporal Semantic Reasoning for the Semantic Change Detection in HR Remote Sensing Images

Lei Ding, Haitao Guo, Sicong Liu, *Member, IEEE,* Lichao Mou, Jing Zhang and Lorenzo Bruzzone, *Fellow, IEEE*

***Abstract*—Semantic change detection (SCD) extends the multi-class change detection (MCD) task to provide not only the change locations but also the detailed land-cover/land-use (LCLU) categories before and after the observation intervals. This fine-grained semantic change information is very useful in many applications. Recent studies indicate that the SCD can be modeled through a triple-branch Convolutional Neural Network (CNN), which contains two temporal branches and a change branch. However, in this architecture, the communications between the temporal branches and the change branch are insufficient. To overcome the limitations in existing methods, we propose a novel CNN architecture for the SCD, where the semantic temporal features are merged in a deep CD unit. Furthermore, we elaborate on this architecture to reason the bi-temporal semantic correlations. The resulting Bi-temporal Semantic Reasoning Network (Bi-SRNet) contains two types of semantic reasoning blocks to reason both single-temporal and cross-temporal semantic correlations, as well as a novel loss function to improve the semantic consistency of change detection results. Experimental results on a benchmark dataset show that the proposed architecture obtains significant accuracy improvements over the existing approaches, while the added designs in the Bi-SRNet further improves the segmentation of both semantic categories and the changed areas. The codes in this paper are accessible at: *https://github.com/ggsDing/Bi-SRNet*.**

***Index Terms*—Remote Sensing, Convolutional Neural Network, Semantic Segmentation, Change Detection, Semantic Change Detection**

I. Introduction

Change detection (CD) refers to the task of identifying the areas in remote sensing images (RSIs) where changes have occurred during the observation intervals [1]. CD is useful for various kinds of real-world applications, such as urban management, environment monitoring, crop monitoring and damage assessment. Although binary change detection (BCD) algorithms allow us to automatically monitor and analyze the region of interests in RSIs, the information provided is coarse-grained and does not describe the detailed change types. In many applications we are interested in not only 'where' the changes occurred, but also 'what' are the changes. To overcome this limitation, multi-class change detection (MCD) techniques [1]–[3] and approaches to the detection of LCLU transitions [4]–[7] have been presented in the literature. They have been referred as Semantic Change Detection (SCD) in recent literature [8], [9] and provide not only the change information, but also the detailed LCLU maps before and after the change events. This allows representation of richer and more complex semantic change information.

Recently with the development of Convolutional Neural Networks (CNNs) [10], great improvements have been achieved in terms of CD. Instead of extracting difference information (which is the common practice in statistical and image processing methods), CNNs learn to directly segment multi-temporal images [11]. CNNs typically have a hierarchy bottom-up design, where the bi-temporal features are embedded and down-scaled through stacked convolutional layers. The change information is modelled through weighted combination and transformation of the features. Compared with statistical and image processing methods, CNN-based methods have the advantages of: i) Improved robustness. The CNN-based CD methods extract numerous features while being free of hyper-parameters (such as weights and thresholds), thus can stably process large volumes of data; ii) Modelling more complex changes. CNNs can learn to model some complex change types which can not be well described by hand-crafted features.

However, the CNN-based SCD of RSIs has been rarely studied in existing works. In the perspective of image processing, BCD is essentially a binary segmentation task where a binary map is produced to represent the changed/ unchanged regions. However, the SCD is a complex task containing two underlying sub-tasks: i) Semantic segmentation (SS) of the LCLU classes. It is required to segment the bi-temporal semantic labels of either all the RSI or the changed areas; ii) binary CD of the changed areas. Therefore, the results of SCD should be either two temporal LCLU maps and a change map [9] or two semantic change maps [12]. Previous CNN-based CD methods may not be suitable for the SCD, since they typically contain only a single branch to embed the difference features [13]. Although several recent works have proposed task-specific methods for the SCD [9], [12], they are based on a triple-branch architecture where the sub-tasks are separately modelled without considering their intrinsic correlations. This often results in inconsistence between the detected changes

L. Ding and H. Guo are with the PLA Strategic Force Information Engineering University, ZhengZhou, China (E-mail: dinglei14@outlook.com, ghtgjp2002@163.com).

L. Ding, 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@studenti.unitn.it, lorenzo.bruzzone@unitn.it).

S. Liu is with the College of Surveying and Geoinformatics, Tongji University, No.1239 Siping Road, Shanghai, China (E-mail: sicongliu.rs@gmail.com).

L. Mou is with the Remote Sensing Technology Institute, German Aerospace Center, and Data Science in Earth Observation, Technical University of Munich, Germany (E-mail: lichao.mou@dlr.de).

This document is funded by the National Natural Science Foundation of China (No. 41876105, 41671410, 42071324). It is also funded by the scholarship from China Scholarship Council (grant NO.201703170123).

and the segmented LCLU maps. Moreover, the semantic features are separately extracted without considering their temporal coherence.

To fill the research gap in SCD, in this paper, we exploit the spatial and temporal semantic correlations in SCD to improve the accuracy. The major contributions in this paper are as follows:

1) Proposing a novel CNN-based architecture for the SCD. The sub-tasks in SCD (SS and CD) are disentangled, whereas their features are shared and are deeply fused. The loss functions are also disentangled to supervise respectively SS and CD in the SCD. The resulting architecture SSCD-l (SS and CD, late fusion) shows significant accuracy improvements over existing approaches;
2) Proposing a Bi-temporal Semantic Reasoning Network (Bi-SRNet) for the SCD. Building on top of the novel CNN architecture, the Bi-SRNet further integrates i) two Siamese Semantic Reasoning (Siam-SR) blocks to model the semantic information in each temporal branch; ii) a Cross-temporal SR (Cot-SR) block to model the temporal correlations, and iii) a Semantic Consistency Loss (SCLoss) function to align the semantic and change representations. These designs are verified in an ablation study, whereas the resulting Bi-SRNet is evaluated in comparisons with state-of-the-art (SOTA) methods.

The remainder of this paper is organized as follows. Section II introduces the literature works on CD of RSIs. Section III elaborates on the proposed CNN architecture, as well as the Bi-SRNet. Section IV describes the experimental settings and the evaluation metrics. Section V reports results of the ablation study and comparative experiments. Section VI summarizes this work and draws the conclusions.

II. Related Work

This section is organized following the development of CD methods. The pre-CNN and CNN-based methods are separately introduced. Recent methods for the SCD are also reviewed.

A. Expert knowledge-based Change Detection

In the past decades, CD techniques have experienced a rapid development due to the increasing availability of remote sensing images, the importance of CD applications, and the evolution of machine learning. There are many excellent review works in literature (e.g., [14]–[17]), focusing on the analyzing of typical CD problems and corresponding methods. Among the traditional statistical and image processing unsupervised CD techniques, in [5] an approach based on the expectation-maximization (EM) algorithm and Markov Random Fields (MRF) was proposed for automatically solving binary CD problems by analyzing a difference image generated by Change Vector Analysis (CVA). CVA [2], [18] and its different variation versions (e.g., [3], [19], [20]) starting from a theoretical definition of the CD problem, provided different successful solutions for CD in multispectral and hyperspectral images also considering the detection of multiple changes. Another popular transformation-based unsupervised CD approach is the Multivariate Alteration Detection (MAD) [21] and its iterative reweighted version (IR-MAD) [22], which exploits the nature of changes in multiple spectral bands. For the supervised CD, approaches are usually designed by taking advantages from robust supervised classifiers such as the Support Vector Machine (SVM), the Random Forest (RF) and the Extreme Learning Machine (ELM) in order to achieve better CD performance with a high accuracy [23]–[25]. Other attempts developed from the semi-supervised perspective combined the merits of both unsupervised and supervised methods in order to improve the automation and robustness of the CD process [26], [27]. On the other hand, CD performance can be also enhanced by considering different features, for example the spectral-spatial features [28], the kernel-based features [29], and the target-driven features [30]. However, the conventional hand-designed features usually represent the low-level descriptions of the change objects, whereas the use of deep learning based approaches provides the capability to learn more complex and effective high-level change features from the data.

B. CNN-based Change Detection

The development of convolutional neural networks (CNNs) in the filed of computer vision provides insights into CD. By modelling the CD in hyper/multi-spectral imagery as a classification task, [31] presents a 2D CNN to learn spectral-spatial feature representation and [32] proposes a recurrent CNN that is able to learn spectral-spatial-temporal features and produce accurate results. Furthermore, in [33], the authors introduce 3D CNN. For high spatial resolution remote sensing images, CD is usually deemed as a dense prediction problem and solved by semantic segmentation CNNs. For instance, in [11] an improved UNet++ is designed for the CD. [34] presents a network with a hierarchical supervision. In [35], the authors introduce a Siamese CNN to extract features of bi-temporal images and utilize a weighted contrastive loss to alleviate the influence of imbalanced data. In [34] and [36], Siamese CNNs are employed to extract bi-temporal features, while CNN decoders are designed to fuse the features and to learn change information. To tackle pseudo changes caused by seasonal transitions in CD tasks, [37] proposes a metric learning-based generative adversarial network (GAN), termed MeGAN, to learn seasonal invariant feature representations. In addition, some works recently focused on devising novel CNN architectures for unsupervised CD [38]–[44]. For example, [38] proposes an unsupervised deep learning-based change vector analysis model for MCD in very high resolution images. In [45] this method is also used for CD in SAR images, after converting them into optical-like features. The authors of [44] incorporate deep neural networks and low-rank decomposition for predicting saliency maps where high values indicate large change probabilities. Another important branch in CNN for CD is heterogeneous CD, also known as multimodal CD, which aims to detect changes between heterogeneous images. In [46], the authors propose two novel network architectures using an affinity-based change prior learnt from the input

data for heterogeneous CD. [47] introduces a CNN entitled symmetric convolutional coupling network (SCCN) for CD in heterogeneous optical and SAR images. In [48], the optical images are translated into the SAR domain to reduce differences, before performing CD with the SAR images.

C. Semantic Change Detection

Binary CD produces as outputs binary maps to represent the change information, which are often not enough informative in practical applications. In many applications there are multiple change types, such as seasonal changes, urbanization, damages, deforestation and pollution. To elaborate the change information, the MCD [7], [19] not only detects the changes, but also classifies the LCLU transition types. In [14] an intuitive method to the MCD, i.e., the post-classification comparison (PCC) is presented, which directly compares the LCLU classification maps to produce transition statistics. A major limitation of this approach is that it omits the temporal correlation between the two RSIs, as each LCLU classification map is produced independently. Multi-date direct classification (MDC) [14] overcomes this limitation by jointly training on the two RSIs. However, it requires multitemporal training data that model all possible LCLU transitions. To address this problem, in [4], [6], [7], [49] the compound classification (CC) technique is introduced for MCD, which computes and maximizes the posterior joint probabilities of LC transitions. This method models the temporal dependence between two RSIs through an iterative semi-supervised estimation of the probability of transitions without requiring training data for each possible transition. One of the representative CNN-based methods for MCD is the work in [32] which integrates RNN units into CNN: the CNNs are employed to extract spatial information, whereas the RNN units detect multi-class changes from temporal features.

Recently there are CNN-based methods developed for SCD. In [13] two SCD methods with triple embedding branches are introduced. Two of the branches segment temporal images into LCLU maps, while a CD branch detects the difference information. In [12], the triple-branch CNN is further extended by introducing gating and weighting designs in the decoders to improve the feature representations. In this work, a benchmark dataset for the SCD is also released together with task-specific evaluation metrics. In [50] a CNN framework for SCD is proposed, where a Siamese CNN is employed to extract semantic features and a decoder module is designed to detect changes. The work in [51] also performs the CD in the decoder and introduces attention designs to enhance the change features. However, two major problems remained: i) the bi-temporal LCLU features are separately embedded without considering their temporal coherence; and ii) frequent inconsistences occur between the detected changes and the segmented LCLU maps.

III. Proposed Bi-temporal Semantic Reasoning Network (Bi-SRNet) for SCD

In this section we introduce the Bi-SRNet for SCD. First, we summarize the existing CNN approaches for the SCD and introduce a novel task-specific architecture. Then, we introduce the Bi-SRNet built on top of this proposed SCD architecture. Finally, we introduce the semantic reasoning designs and the loss functions in the Bi-SRNet.

A. Task Formulation and Possible Approaches

Before introducing the proposed approach, let us first define the task of semantic change detection (SCD) and its connections with the semantic segmentation (SS) and the binary change detection (BCD). Given an input image I, the task of SS is to find a mapping function f_s that projects I into a semantic map P:

$$f_s(p_{i,j}) = c_{i,j} \tag{1}$$

where $p_{i,j}$ is a pixel on I, c_{ij} is the projected LCLU class. Meanwhile, a BCD function f_{bcd} projects two temporal images I_1, I_2 into a binary change map C. For two image pixels $p_{1_{i,j}}, p_{2_{i,j}}$ on I_1, I_2 that are related to the same spatial location, the calculation is:

$$f_{bcd}(p_{1_{i,j}}, p_{2_{i,j}}) = \begin{cases} 0, & c_{1_{i,j}} = c_{2_{i,j}} \\ 1, & c_{1_{i,j}} \neq c_{2_{i,j}} \end{cases} \tag{2}$$

where the projected signal (0 or 1) reports if there is change in LCLU classes ($c_{1_{i,j}}$ and $c_{2_{i,j}}$) or other change types (e.g., damage and status change). The SCD function f_{scd} is a combination of f_s and f_{bcd}:

$$f_{scd}(p_{1_{i,j}}, p_{2_{i,j}}) = \begin{cases} (0,0), & c_{1_{i,j}} = c_{2_{i,j}} \\ (c_{1_{i,j}}, c_{2_{i,j}}), & c_{1_{i,j}} \neq c_{2_{i,j}} \end{cases} \tag{3}$$

It produces two semantic change maps S_1 and S_2 which indicate both the change areas and the bi-temporal LCLU classes. The required training labels in the SCD can be either i) two ground truth (GT) semantic change maps L_1 and L_2 [12], or alternatively ii) one GT binary CD map L_c and two GT LCLU maps $L_{S_1}$and L_{S_2} [9]. In the later , L_1 and L_2 can be easily generated by masking L_{S_1} and L_{S_2} with L_c.

An intuitive approach to the SCD is the post-classification comparison (PCC) [14], which first classifies the LCLU maps and then compare the difference information. However, this approach has been proved sub-optimal since it neglects the temporal correlations and may cause accumulation of errors [7], [9]. Alternatively, we employ CNNs to learn directly the semantic changes, which is often referred as multidate direct classification (MDC) in MCD tasks [7]. The possible CNN-based approaches can be summarized as follows:

1) Direct SCD, early fusion (DSCD-e, Fig.1(a)). The temporal images I_1 and I_2 are concatenated as input data. A single CNN encoder $\mathcal{E}$ is employed to directly learn the f_{scd}:

$$S_1, S_2 = \mathcal{E}(I_1, I_2). \tag{4}$$

Many literature works that treat CD as a semantic segmentation task can be included into this category, including the FC-EF in [13] and the UNet++ in [11]. A major limitation of this architecture is that LCLU information in each temporal branch is not fully exploited [9]. Since the changed areas are minority, $\mathcal{E}$ is driven to pay more attention to the unchanged areas.

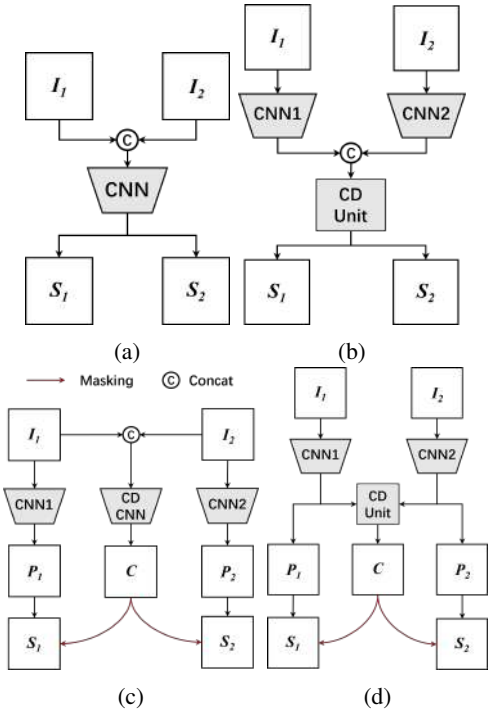


Fig. 1: Four possible CNN architectures for the SCD: (a) Direct SCD, early fusion (DSCD-e); (b) Direct SCD, late fusion (DSCD-l), (c) Disentangled SS and CD, early fusion (SSCD-e), and (d) Disentangled SS and CD, late fusion (SSCD-l).

2) Direct SCD, late fusion (DSCD-l, Fig.1(b)). I_1 and I_2 are separately fed as inputs into two CNN encoders $\mathcal{E}_1$ and $\mathcal{E}_2$ (which can be weight-sharing, i.e., siamese [13], if I_1 and I_2 belong to the same domain). The encoded features are then fused and modelled through a convolutional CD unit $\mathcal{D}$. $\mathcal{E}_1$ and $\mathcal{E}_2$ serve as feature extractors, whereas the f_{scd} is learned by $\mathcal{D}$ with the embedded semantic features:

$$S_1, S_2 = \mathcal{D}[\mathcal{E}(I_1), \mathcal{E}(I_2)]. \tag{5}$$

This architecture may also include decoder networks and skip connections. In the FC-Siam-conc and the FC-Siam-diff [13] the CD units are multi-scale concatenation blocks, whereas in the ReCNN [32] they are Recurrent Neural Networks (RNNs). However, in this architecture, the *no-change* class is still competing with other classes during the network inference, which does not correctly reflect their intrinsic correlations.

3) Disentangled SS and CD, early fusion (SSCD-e, Fig.1(c)). Three CNN encoders $\mathcal{E}_1$, $\mathcal{E}_2$ and $\mathcal{E}_c$ are separately trained with I_1, I_2 and (I_1, I_2) as the inputs. $\mathcal{E}_1$ and $\mathcal{E}_2$ can be siamese if I_1 and I_2 belong to the same domain. The semantic and change information are separately modelled. $\mathcal{E}_1$ and $\mathcal{E}_2$ produce the semantic maps P_1 and P_2, while $\mathcal{E}_c$ produces a binary change map C. S_1, S_2 are then generated by masking P_1, P_2 with C. The calculations are as follows:

$$P_1, P_2, C = \mathcal{E}_1(I_1), \mathcal{E}_2(I_2), \mathcal{E}_c(I_1, I_2) \tag{6}$$

$$S_1, S_2 = C \cdot (P_1, P_2). \tag{7}$$

A representative of this architecture is the HRSCD-str.3 in [9]. However, a disadvantage of this architecture is that it neglects the temporal correlations, since there are limited communications between the temporal branches and the CD branch. Although there can be connections between $\mathcal{E}_1$, $\mathcal{E}_2$ and $\mathcal{E}_c$ (e.g., the skip-connections in the HRSCD-str.4 [9] and the gating operations in the ASN [12]), they do not fully exploit the LCLU information in temporal features. Additionally, training the f_{cd} from scratch with input images is not computation-efficient.

4) Disentangled SS and CD, late fusion (SSCD-l, Fig.1(c)). This is a novel architecture proposed to address the limitations of above-mentioned approaches. In the SSCD-l, two CNN encoders $\mathcal{E}_1$ and $\mathcal{E}_2$ are employed to extract semantic information from I_1 and I_2. The extracted semantic features are further merged to train a CD unit $\mathcal{D}$, which exploits the difference information. The calculations can be represented as follows:

$$P_1, P_2 = \mathcal{E}_1(I_1), \mathcal{E}_2(I_2) \tag{8}$$

$$C = \mathcal{D}[\mathcal{E}(I_1), \mathcal{E}(I_2)] \tag{9}$$

$$S_1, S_2 = C \cdot (P_1, P_2). \tag{10}$$

The advantages of the SSCD-l architecture are three-fold: i) Both the LCLU information and the change information are explicitly modelled. This enables the SSCD-l to embed more oriented features for the underlying sub-tasks (SS and CD); ii) Instead of producing two separate change maps (such as in the DSCD-e and DSCD-l), where there may be discrepancies, the SSCD-l learns to produce a single change map. The change information is therefore more consistent in the two masked temporal predictions S_1 and S_2; iii) Instead of extracting change features from scratch in the two RSIs (such as in the SSCD-e), the SSCD-l learns more semantic changes from the extracted features of the two temporal branches.

To find out the optimal CNN architecture for the SCD, experimental comparisons have been conducted in Sec.V-A and Sec.V-C.

B. Bi-temporal Semantic Reasoning Network

In the proposed SSCD-l architecture, the semantic features are separately extracted through two temporal branches and are merged through a deep CD unit. However, one of the remaining problems is to model the temporal correlation and coherence between the two feature extraction branches in many applications. The majority of image regions remain unchanged through the observation intervals, thus they exhibit similar visual patterns. To better exploit this information, we propose the Bi-temporal Semantic Reasoning Network (Bi-SRNet) as illustrated in Fig.2.

The Bi-SRNet is built on top of the SSCD-l architecture by introducing two extra Semantic Reasoning (SR) blocks [52], [53] and a semantic consistency loss. Given 2 input temporal images I_1 and I_2, the Bi-SRNet first employs 2

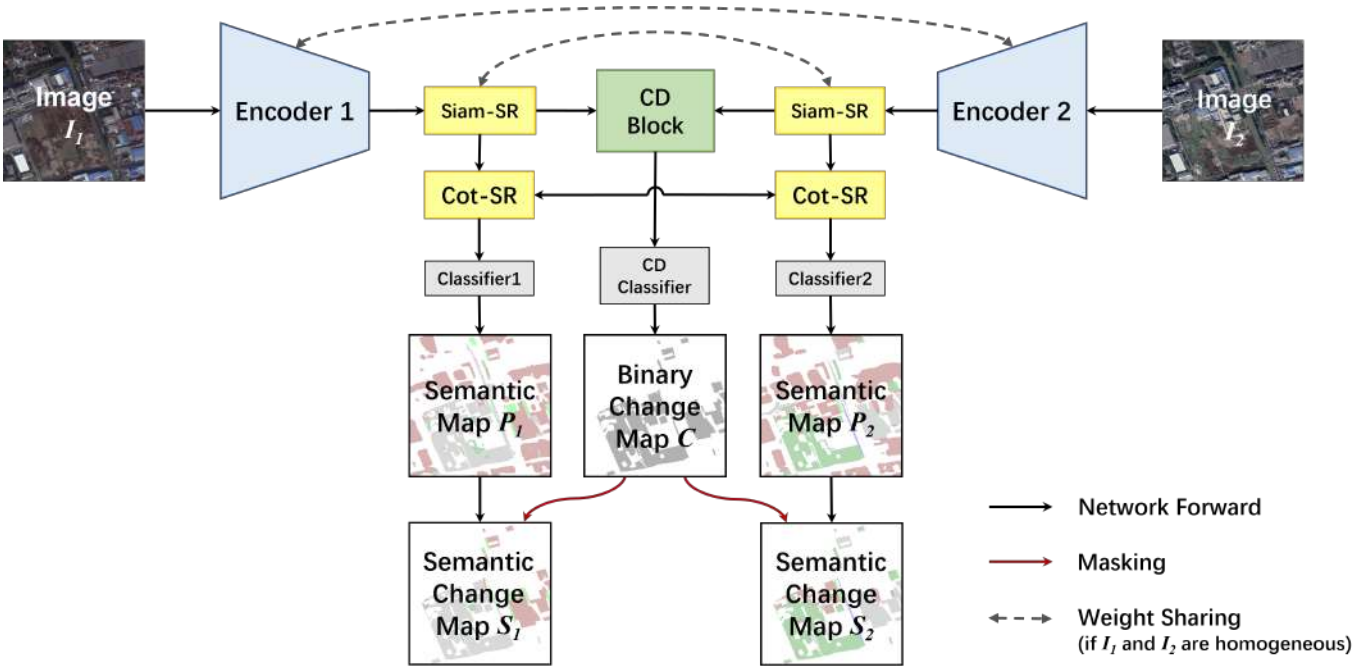


Fig. 2: Architecture of the proposed **Bi**-temporal **S**emantic **R**easoning **Net**work (Bi-SRNet) for SCD. The modelling of semantic and changed information is disentangled to enhance the feature exploitation, whereas the semantic representations in the two temporal branches are aligned through the SR blocks.

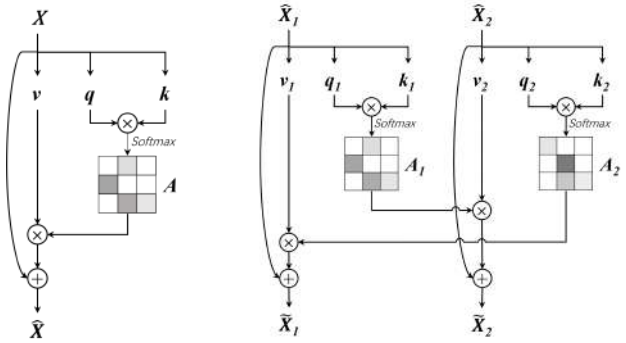
CNN encoders $\mathcal{E}_1$ and $\mathcal{E}_2$ to extract the semantic features X_1 and X_2. Differently from the SSCD-l, X_1 and X_2 are then processed by two Siamese SR (Siam-SR) blocks to improve their semantic representations. Under the circumstance that there is no significant domain difference, the weights of both $\mathcal{E}_1, \mathcal{E}_2$ and the Siam-SRs are shared to reduce over-fitting risks. The enhanced features $\hat{X}_1$ and $\hat{X}_2$ are further sent to a Cross-temporal SR (Cot-SR) block to model their semantic correlations. These correlations between the two temporal branches are also supervised by the semantic consistency loss (introduced in Sec.III-D). The temporally aligned features (denoted as $\tilde{X}_1$ and $\tilde{X}_2$) are then projected into two semantic maps P_1 and P_2. Meanwhile, the CD block models change information through the unaligned features $\hat{X}_1$ and $\hat{X}_2$, before them being projected into a binary change map C. All the projections are made through 1×1 convolutional layers whose weights are not shared. Same as the SSCD-l, the Bi-SRNet produces as outputs 3 direct maps: the semantic maps P_1, P_2 and a binary CD map C. Finally, the semantic change maps S_1 and S_2 are generated by masking P_1 and P_2 with C. $\mathcal{E}_1$, $\mathcal{E}_2$ and $\mathcal{D}$ in the Bi-SRNet are identical to those in the SSCD-l. The simplified calculations (omitting the convolutional classifiers) are:

$$\hat{X}_1 = \text{SiamSR}[\mathcal{E}_1(I_1)], \hat{X}_2 = \text{SiamSR}[\mathcal{E}_2(I_2)] \quad (11)$$

$$P_1, P_2 = \text{CotSR}(\hat{X}_1, \hat{X}_2) \quad (12)$$

$$C = \mathcal{D}(\hat{X}_1, \hat{X}_2)\ \ S_1, S_2 = C \cdot (P_1, P_2). \quad (13)$$

Since the focus of this work is to investigate architecture designs and to model the semantic correlations for SCD, no decoder structures are adopted. The SR blocks operate on a spatial scale of $1/8$ of the input resolution. This spatial resolution is suggested for enhancing the semantic features in the literature [54] [55], as it achieves a balance between spatial accuracy and context modelling distance. The network outputs are directly enlarged as the results.



(a) (b)

Fig. 3: Structures of the (a) Siamese Semantic Reasoning (Siam-SR) and (b) Cross-temporal Semantic Reasoning (Cot-SR) blocks.

C. Semantic Reasoning Blocks

Non-local units [56] have been proved effective to model long-range dependencies in images and have been widely used in semantic segmentation tasks [52], [53], [57]. In the SCD task it is beneficial to take into account both i) the spatial correlations within each temporal image and ii) the semantic correlation and consistency between the bi-temporal images.

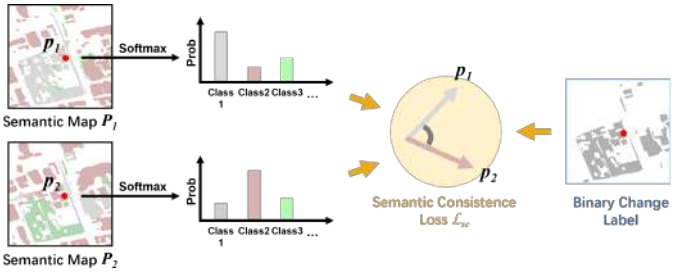


Fig. 4: Illustration of the calculation of the Semantic consistency Loss (SCLoss).

In the Bi-SRNet these information are learned through the Siam-SR blocks and the Cot-SR block, respectively.

The Siam-SR blocks are two standard non-local units that share the same weights. Fig.3(a) illustrates the operations inside a Siam-SR block. Given an input feature $\mathbf{X} \in \mathbb{R}^{c \times h \times w}$ where c is the number of channels and h, w are the spatial sizes, a Siam-SR block first projects it into three vectors $\mathbf{q} \in \mathbb{R}^{H \times c'}, \mathbf{k} \in \mathbb{R}^{c' \times H}$ and $\mathbf{v} \in \mathbb{R}^{c \times H}$, where $H = hw$ and $c' = c/r$, r is a channel reduction factor (normally set to 2). An attention matrix $\mathbf{A} \in \mathbb{R}^{H \times H}$ is then calculated as:

$$\mathbf{A} = \phi(\mathbf{q} \times \mathbf{k}) \tag{14}$$

where ϕ is a *softmax* normalization function along the row dimension. $\mathbf{A}$ records the correlations between each pair of spatial positions. An enhanced feature $\hat{\mathbf{X}}$ is then obtained with:

$$\hat{\mathbf{X}} = \mathbf{X} + \mathbf{v} \times \mathbf{A} \tag{15}$$

Intuitively, the bi-temporal information provides more clues of the image context. Therefore, we propose the Cot-SR block, which is an extension of the non-local unit to model cross-temporal information. As illustrated in Fig.3(b), the Cot-SR block simultaneously enhances two temporal features $\hat{\mathbf{X}}_1, \hat{\mathbf{X}}_2 \in \mathbb{R}^{c \times h \times w}$. First, 6 vectors $\mathbf{q}_1, \mathbf{q}_2 \in \mathbb{R}^{H \times c'}$, $\mathbf{k}_1, \mathbf{k}_2 \in \mathbb{R}^{c' \times H}$ and $\mathbf{v}_1, \mathbf{v}_2 \in \mathbb{R}^{c \times H}$ are projected from $\hat{\mathbf{X}}_1$ and $\hat{\mathbf{X}}_2$. Second, two attention matrices $\mathbf{A}_1, \mathbf{A}_2 \in \mathbb{R}^{H \times H}$ are generated to record the semantic focuses in each temporal branch:

$$\mathbf{A}_1 = \phi(\mathbf{q}_1 \times \mathbf{k}_1) \tag{16}$$

$$\mathbf{A}_2 = \phi(\mathbf{q}_2 \times \mathbf{k}_2) \tag{17}$$

Finally, each attention map operates on its opposite temporal branch to project cross-temporal correlations:

$$\tilde{\mathbf{X}}_1 = \hat{\mathbf{X}}_1 + \mathbf{v}_1 \times \mathbf{A}_2 \tag{18}$$

$$\tilde{\mathbf{X}}_2 = \hat{\mathbf{X}}_2 + \mathbf{v}_2 \times \mathbf{A}_1 \tag{19}$$

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

The Siam-SR blocks aggregate spatial information to embed semantic focuses into each temporal branch, whereas the Cot-SR learns cross-temporal semantic consistency to enhance the features in unchanged areas.

D. Loss Functions

We use three loss functions to train the Bi-SRNet: the semantic class loss $\mathcal{L}_{sem}$, the binary change loss $\mathcal{L}_{change}$, and a proposed semantic consistency loss $\mathcal{L}_{sc}$.

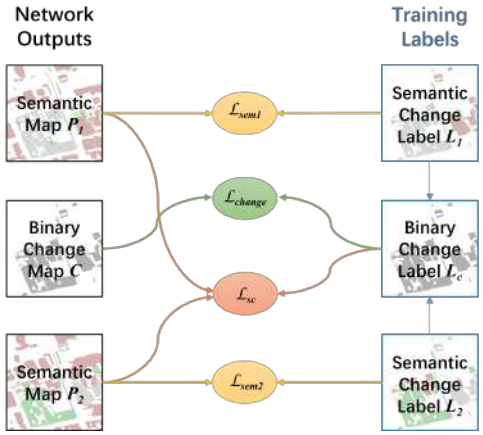


Fig. 5: The calculation of supervision losses in the Bi-SRNet.

The semantic loss $\mathcal{L}_{sem}$ is the mutli-class cross entropy loss between the semantic segmentation results P_1, P_2 and the GT semantic change maps L_1, L_2. The calculation of $\mathcal{L}_{sem}$ on each pixel is:

$$\mathcal{L}_{sem} = -\frac{1}{N}\sum_{i=1}^{N} y_i log(p_i) \tag{20}$$

where N is the number of semantic classes, y_i and p_i denote the GT label and the predicted probability of the i-th class, respectively. N is set according to the number of LCLU classes in the dataset. The *no-change* class (indexed as '0') is excluded from loss calculation to encourage the temporal branches to focus on extracting the semantic features.

The change loss $\mathcal{L}_{change}$ is the binary cross entropy loss between the predicted binary change map C and the reference change map L_c. The L_c is generated with L_1 or L_2 by replacing all their non-zero labels with a *changed* label (indexed as '1'). The $\mathcal{L}_{change}$ for each pixel is calculated as:

$$\mathcal{L}_{change} = -y_c log(p_c) - (1 - y_c) log(1 - p_c) \tag{21}$$

where y_c and p_c denote the GT label and the predicted probability of change, respectively.

$\mathcal{L}_{sem}$ and $\mathcal{L}_{change}$ are designed to drive the learning of the semantic information and of the CD, respectively. We further propose a task-specific semantic consistency loss (SCLoss) to link SS with CD. It aligns the semantic representations between the two temporal branches, as well as guides training of the Cot-SR block. As illustrated in Fig.4, the SCLoss awards predictions with similar probability distributions in the *no-change* areas, whereas punishing those in the changed areas. This aligns the bi-temporal semantic and change information in the SCD task. The SCLoss $\mathcal{L}_{sc}$ is calculated between the predicted semantic maps P_1, P_2 and the change label L_c using the Cosine loss function:

$$\mathcal{L}_{sc} = \begin{cases} 1 - cos(x_1, x_2), & y_c = 1 \\ cos(x_1, x_2), & y_c = 0 \end{cases} \tag{22}$$

where x_1, x_2 are feature vectors of a pixel on P_1 and P_2, respectively. y_c is the value at the same position on L_c.

Training of the two feature embedding branches is directly supervised by L_1 and L_1 and is assisted by L_c (through the

$\mathcal{L}_{sc}$), while training of the CD block is directly supervised by L_c. The relationships between the 3 outputs P_1, P_2, C and the GT maps L_1, L_2 and L_c are illustrated in Fig.5. The total loss $\mathcal{L}_{scd}$ is calculated as:

$$\mathcal{L}_{scd} = (\mathcal{L}_{sem_1} + \mathcal{L}_{sem_2})/2 + \mathcal{L}_{change} + \mathcal{L}_{sc} \tag{23}$$

where $\mathcal{L}_{sem_1}$ and $\mathcal{L}_{sem_2}$ are the semantic loss of each temporal branch. They are added and averaged to represent the $\mathcal{L}_{sem}$. Since the learning of LCLU information and change information is disentangled in the SSCD-l architecture, no hyper-parameters are used to balance the loss functions.

IV. Dataset Description and Experimental Settings

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

A. Dataset

The experiments in this study are conducted on the SEmantic Change detectiON Dataset (SECOND) [12], a benchmark dataset for the SCD. The SECOND is constructed with bi-temporal HR optical images (containing RGB channels) collected by several aerial platforms and sensors. The observed regions include several cities in China, including Hangzhou, Chengdu and Shanghai. Each image has the same size of 512×512 pixels. The spatial resolution varies from 0.5m to 3m (per pixel).

The LC categories before and after the change events are provided. In each GT semantic change map, one change class and six LC classes are annotated, including: *no-change, non-vegetated ground surface, tree, low vegetation, water, buildings* and *playgrounds*. These LC classes are selected considering the commonly interesting LC classes and the frequent geographical changes [58]. The bi-temporal LC transitions raise 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 ones are openly available. We further split them into a training set and a test set with the numeric ratio of $4:1$ (i.e., 2375 image pairs for training, 593 ones for testing).

B. Evaluation Metrics

In this study, 3 evaluation metrics are adopted to evaluate the SCD accuracy, including: overall accuracy (OA), mean Intersection over Union (mIoU) and Separated Kappa (SeK) coefficient. OA has been commonly adopted in both semantic segmentation [55], [59] and CD [9] tasks. 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 ($i, j \in \{0, 1, ..., N\}$ (0 represents *no-change*). OA is calculated as:

$$OA = \sum_{i=0}^{N} q_{ii} / \sum_{i=0}^{N} \sum_{j=0}^{N} q_{ij}. \tag{24}$$

Since OA is mostly determined by the identification of *no-change* pixels, it cannot well evaluate the segmentation of LCLU classes. Additionally, it does not count the pixels that are identified as *changed* but are predicted into wrong LCLU classes. Alternatively, mIoU and SeK are suggested in the SECOND [12] to 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):

$$mIoU = (IoU_{nc} + IoU_c)/2, \tag{25}$$

$$IoU_{nc} = q_{00} / (\sum_{i=0}^{N} q_{i0} + \sum_{j=0}^{N} q_{0j} - q_{00}), \tag{26}$$

$$IoU_c = \sum_{i=1}^{N} \sum_{j=1}^{N} q_{ij} / (\sum_{i=0}^{N} \sum_{j=0}^{N} q_{ij} - q_{00}), \tag{27}$$

The SeK coefficient is calculated based on the confusion matrix $\hat{Q} = \{\hat{q}_{ij}\}$ where $\hat{q}_{ij} = q_{ij}$ except that $\hat{q}_{00} = 0$. This is to exclude the true positive *no-change* pixels, whose number is dominant. The calculations are as follows:

$$\rho = \sum_{i=0}^{N} \hat{q}_{ii} / \sum_{i=0}^{N} \sum_{j=0}^{N} \hat{q}_{ij}, \tag{28}$$

$$\eta = \sum_{i=0}^{N} (\sum_{j=0}^{N} \hat{q}_{ij} * \sum_{j=0}^{N} \hat{q}_{ji}) / (\sum_{i=0}^{N} \sum_{j=0}^{N} \hat{q}_{ij})^2, \tag{29}$$

$$SeK = e^{IoU_c - 1} \cdot (\rho - \eta)/(1 - \eta). \tag{30}$$

The mIoU and SeK directly evaluate the sub-tasks in SCD, i.e., the CD and the SS of LCLU classes, respectively. Additionally, to evaluate more intuitively the segmentation of LCLU classes in changed areas, we introduce a new metric F_{scd} (derived from the F_1 score in segmentation and CD tasks [13] [60]):

$$P_{scd} = \sum_{i=1}^{N} q_{ii} / \sum_{i=1}^{N} \sum_{j=0}^{N} q_{ij}, \tag{31}$$

$$R_{scd} = \sum_{i=1}^{N} q_{ii} / \sum_{i=0}^{N} \sum_{j=1}^{N} q_{ij}, \tag{32}$$

$$F_{scd} = \frac{2 * P_{scd} * R_{scd}}{P_{scd} + R_{scd}} \tag{33}$$

where P_{scd} and R_{scd} are variants of the $Precision$ and $Recall$ [55] which focus only on the changed areas. F_{scd} describes the segmentation accuracy of LCLU classes in the changed areas.

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

C. Experimental settings

The experiments in this study are conducted on a workstation with a NVIDIA Quadro P6000 GPU. All the CNN

models are implemented with the PyTorch library. The same experimental parameters are used in all the experiments, including batch size (8), running epochs (50) and initial learning (0.1). The gradient descent optimization method is Stochastic Gradient Descent (SGD) with Nesterov momentum. The augmentation strategy include random flipping and rotating while loading the image pairs. For simplicity, no test-time augmentation operation is applied.

V. Experimental Results

In this section, a series of experiments are conducted to verify the effectiveness of the proposed architecture for SCD (the SSCD-l) and the components in the Bi-SRNet. Finally the proposed methods are compared with several recent methods in both SCD and binary CD.

A. Comparison of SCD Architectures

To find the optimal CNN architecture for the SCD, we compare the 4 approaches discussed in Sec.III-A: the DSCD-e, DSCD-l, SSCD-e and SSCD-l. For simplicity and fairness, all the tested architectures are implemented with only the basic convolutional designs (i.e., sophisticated designs such as encoder-decoder structures, dilated convolutions and attention units are not adopted). They employ the same CNN encoder (the ResNet34 [61]), adopt the same down-sampling stride ($\times 1/8$) and have same number of inner channels at each convolutional stage. The late-fusion approaches (DSCD-e and SSCD-l) employ an identical CD block, i.e., a CNN block with 6 stacked residual units.

The quantitative results are reported in Table.I. The DSCD-e architecture obtains the lowest accuracy, although it requires the lowest computations. Meanwhile, DSCD-l achieves much higher metric values, showing that the separate embedding of temporal features and CD features is essential in SCD. Compared to the DSCD-l, the SSCD-e slightly increases the mIoU values at the cost of taking much heavier computations. This shows that the embedding of CD features from scratch is in-efficient. On the contrary, the proposed SSCD-l re-uses and merges the temporal features for CD, which leads to significant performance improvements. It surpasses the SSCD-e by 4.81% in Sek and 4.26% in F_{scd}, whereas its computational cost is only slightly higher than that of the DSCD-l.

Fig.6 qualitatively compares the results obtained by the SSCD-e and SSCD-l architectures. The SCD maps are generated by masking the LC maps with the change maps, as illustrated in Sec.III-A. The intermediate results reveal one of the major limitations of the SSCD-e architecture. Since the change information is separately embedded in the SSCD-e architecture, the segmented changes are often inconsistent with those in the bi-temporal LC maps. For example, in Fig.6(a) the LC change of 'ground' turning into 'low vegetation' is indicated in the LC maps but it is not represented in the change map. Also a LC change from 'building' to 'building' is indicated in the SCD map, which is self—contradictory. In Fig.6(c) the emergence of a building is also omitted. However, these LC changes are easily captured by the SSCD-l architecture. Through direct modelling of the LC features, the CD unit is aware of the semantic changes. Additionally, some of the non-salient differences are adaptively omitted in the results of SSCD-l, such as the transitions between between 'ground' and 'vegetation' in Fig.6(b).

These experimental comparisons indicate that the proposed SSCD-l provides the most accurate results in SCD among the compared architectures. Its advantages in embedding the semantic information (indicated by the SeK) is particularly dominant. However, there are also many variants of these architectures exploited in previous studies. In Sec.V-C these methods are further compared to summarize common features of the different SCD architectures.

B. Ablation Study

After verifying the effectiveness of the SSCD-l, we further perform an ablation study to evaluate the auxiliary components in the proposed Bi-SRNet. The quantitative results are presented in Table II. First we test the effectiveness of the SCLoss by adding it as an auxiliary loss to train the SSCD-l. This boosts the accuracy by around 0.82% in SeK and 0.76% in F_{scd}, indicating that the semantic embedding of features is improved. Taking this method (SSCD-l with SCLoss) as the baseline, we further assess the performance of each SR blcok. The Siam-SR blocks, which are placed on each temporal branch, lead to noticeable accuracy improvements (0.4% in mIoU and 0.43 in F_{scd}). Meanwhile, the Cot-SR block that models the temporal coherence improves the SeK (by over 0.41%) and the F_{scd} (by over 0.46%). This indicates that both the Siam-SR and the Cot-SR improve the semantic embedding of temporal features, while the former also improves the detection of change information. Finally, we evaluate the Bi-SRNet which contains all these auxiliary designs. Compared to the standard SSCD-l, its improvements are around 0.81% in mIoU, 1.36% in SeK and 1.39 % in F_{scd}. These results indicate that the Bi-SRNet integrating all the designs brings an increase in accuracy.

The qualitative results in some testing areas are presented in Fig.7. The prediction maps from left to right are provided for the proposed methods in Table II, which are organized in the sequence of number of contained components. Compared to the results of the standard SSCD-l, the predicted LC categories after introducing the SCLoss and SR blocks are gradually improved. The Bi-SRNet exhibits advantages in discriminating the critical areas. For example, in Fig.7(a2),(b1),(c1) and (c2), identification of the *ground*, *low vegetation* and *tree* classes is greatly improved.

Through this ablation study we find that: i) as indicated by increases in SeK values, all the tested auxiliary components result in improvements in semantic embedding; ii) the semantic reasoning designs in the Bi-SRNet improve not only the discrimination of LC categories, but also the detection of changes.

C. Comparative Experiments

To comprehensively evaluate the performance of the proposed SSCD-l architecture and the Bi-SRNet, we further

TABLE I: Comparison of the results provided by different CNN architectures for SCD.

Method	Computational Costs			Accuracy			
	Params (Mb)	FLOPS (Gbps)	Infer. Time (s/100e)	OA(%)	mIoU(%)	Sek(%)	F_{scd}(%)
DSCD-e	21.36	91.39	1.32	86.46	68.55	16.01	56.22
DSCD-l	23.31	189.54	2.74	86.58	68.86	16.43	56.67
SSCD-e	42.72	272.95	3.89	85.66	69.55	17.05	56.96
proposed SSCD-l	23.31	189.57	2.75	**87.19**	**72.60**	**21.86**	**61.22**

TABLE II: Quantitative results of the ablation study.

Proposed Method	Components				Accuracy			
	CD block	siam-SR block	Cot-SR block	SCLoss	OA(%)	mIoU(%)	Sek(%)	F_{scd}(%)
SSCD-l	√				87.19	72.60	21.86	61.22
SCLoss-SSCDl	√			√	87.48	73.06	22.68	61.98
SiamSR-SSCDl	√	√		√	87.73	73.45	23.15	62.41
CotSR-SSCDl	√		√	√	87.67	73.38	23.09	62.44
Bi-SRNet	√	√	√	√	**87.84**	**73.41**	**23.22**	**62.61**

TABLE III: Comparison of the proposed methods with literature methods for the SCD.

Method	Arch. Type	Computational Costs			Accuracy			
		Params (Mb)	FLOPs (Gbps))	Infer. Time (s/100e)	OA(%)	mIoU(%)	Sek(%)	F_{scd}(%)
FC-EF [13]	DSCD-e	1.66	17.75	0.73	85.18	64.25	9.98	48.45
UNet++ [11]	DSCD-e	9.16	139.82	3.46	85.18	63.83	9.90	48.04
HRSCD-str.2 [9]	DSCD-e	6.39	14.29	2.33	85.49	64.43	10.69	49.22
ResNet-GRU [32]	DSCD-l	21.45	182.36	2.73	85.09	60.64	8.99	45.89
ResNet-LSTM [32]	DSCD-l	21.48	182.63	2.73	86.77	67.16	15.96	56.90
FC-Siam-conv. [13]	DSCD-l	2.74	35.01	1.21	86.92	68.86	16.36	56.41
FC-Siam-diff [13]	DSCD-l	1.66	21.41	1.02	86.86	68.96	16.25	56.20
IFN [34]	DSCD-l	35.73	329.10	9.41	86.47	68.45	14.25	53.54
HRSCD-str.3 [9]	SSCD-e	12.77	42.67	6.26	84.62	66.33	11.97	51.62
HRSCD-str.4 [9]	SSCD-e	13.71	43.69	6.37	86.62	71.15	18.80	58.21
SSCD-l (proposed)	SSCD-l	23.31	189.57	2.75	87.19	72.60	21.86	61.22
Bi-SRNet (proposed)	SSCD-l	23.39	189.91	3.42	**87.84**	**73.41**	**23.22**	**62.61**

compare them with several SOTA methods in both CD and SCD tasks. The compared methods include:

1) The FC-EF, FC-Siam-conv and FC-Siam-diff [13]. These are three UNet-like [62] CNNs for binary CD. The FC-EF concatenates temporal images as inputs, whose architecture can be divided into the DSCD-e. The FC-Siam-conv and FC-Siam-diff both contain siamese encoders while their decoders also serve as CD blocks, which can be divided into the DSCD-l.

2) The UNet++ [11]. This model is a variant of the UNet and can be divided into the DSCD-e. Note that the deep supervision function introduced in [11] is not suitable for the SCD task, thus it is removed.

3) The HRSCD-str.2, HRSCD-str.3 and HRSCD-str.4 [9]. These are three methods introduced for the SCD, both containing residual blocks [61] and encoder-decoder structures. The HRSCD-str.2 directly produces semantic change maps, thus belongs to the type DSCD-e. The HRSCD-str.3 and HRSCD-str.4 both contain triple encoding branches, thus belong to the type SSCD-e.

4) The ResNet-GRU and ResNet-LSTM [32]. These methods are derived from the methods in [32] that combines CNN and RNNs for CD. Since the original methods are designed to classify low-resolution RSIs and contain only few convolutional layers, which are not suitable for processing HR RSIs, we further change their encoders into the ResNet34 [61]. The RNN units serve as CD blocks, thus these methods belong to the DSCD-l.

5) The IFN [34]. This method contains a VGGNet [63] encoder and an attention-based decoder. The features are merged in the decoder, thus it belongs to the DSCD-l.

Among the above-mentioned methods, the ones in 3) are aimed to SCD, whereas those in 2), 4) and 5) are originally designed for binary CD. To apply these CD methods for the SCD, we slightly modified their last convolutional layers to meet the required numbers of output maps and channels. The quantitative results are reported in Table III. One can observe that the methods using the DSCD-e architecture generally produce unsatisfactory results. In these methods, the modelling of semantic information and change information is entangled, which leads to low SeK values. The FC-Siam-conv and the FC-Siam-diff obtain higher accuracy, which concatenate and merge the semantic features through their decoders. The ResNet-LSTM obtains the highest accuracy among the DSCD-l based methods due to its temporal modelling design. The HRSCD-str.4 (based on the SSCD-e architecture) obtains the highest accuracy amoung literature methods. It contains skip-connections between the temporal branches and the CD branch, which alleviates the drawbacks in the standard SSCD-e. The proposed methods based on the SSCD-l architectures

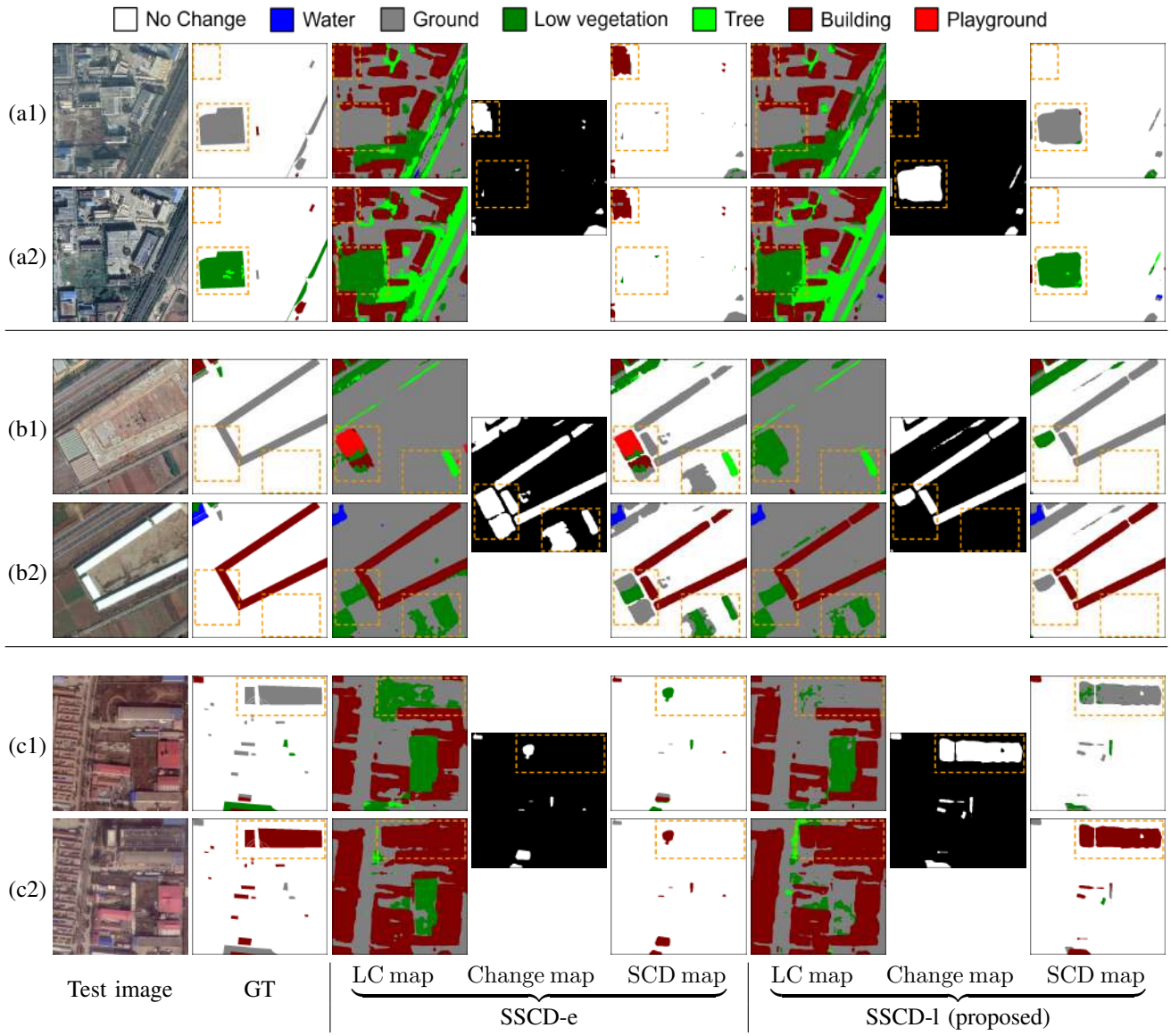


Fig. 6: Comparisons of the results provided by the SSCD-e and SSCD-l architectures.

obtain the best accuracy. Without using any CNN decoder or specialized encoder, the proposed Bi-SRNet outperform SOTA by a large margin in all the metrics (1.45%, 3.06% and 3% in mIoU, SeK and F_{scd}, respectively). This confirms that the SSCD-l is a better CNN architecture for the SCD.

The computational costs of different methods are reported in Table III. Generally, the models with ResNet34 encoders (ResNet-GRU, ResNet-LSTM, SSCD-l and Bi-SRNet) have more parameters and require more calculations. However, their inference time is close to some UNet-based methods (e.g., UNet++ and HRSCD-str.2), since most of the calculations are performed on down-scaled feature maps (at Stage 3 and 4 in the ResNet). The HRSCD-str.3 and HRSCD-str.4 that contain intense calculations in the decoder even require more inference time. The IFN that contains also cascaded attention blocks in the decoder requires highest computation and inference time. Inference time of the proposed methods (SSCD-l and Bi-SRNet) are at the middle level among the compared methods.

To visually assess the results, in Fig.8 we present comparisons of the results provided by different methods in several sample areas. One can observe that most of these methods are sensitive to the semantic changes between *building* and *low vegetation*. However, the DSCD-based methods commonly omit some other LC classes. Specifically, the UNet++, the ResNet-LSTM and the IFN omitted many *tree* and *water* areas. The detection of changes is much improved in SSCD-based results. However, the SSCD-e based approaches (HRSCD-str.3 and HRSCD-str.4) mis-classified the LC classes in some critical areas (e.g., confusion between *ground* and *low vegetation* in Fig.8(d1)) and omitted some minor changes (e.g., the *low vegetation* to *tree* changes in Fig.8(a) and the *tree* to *ground* changes in Fig.8(c)). Meanwhile, these change types are all captured by the proposed Bi-SRNet. This further confirms its advantages in both CD and semantic exploitation.

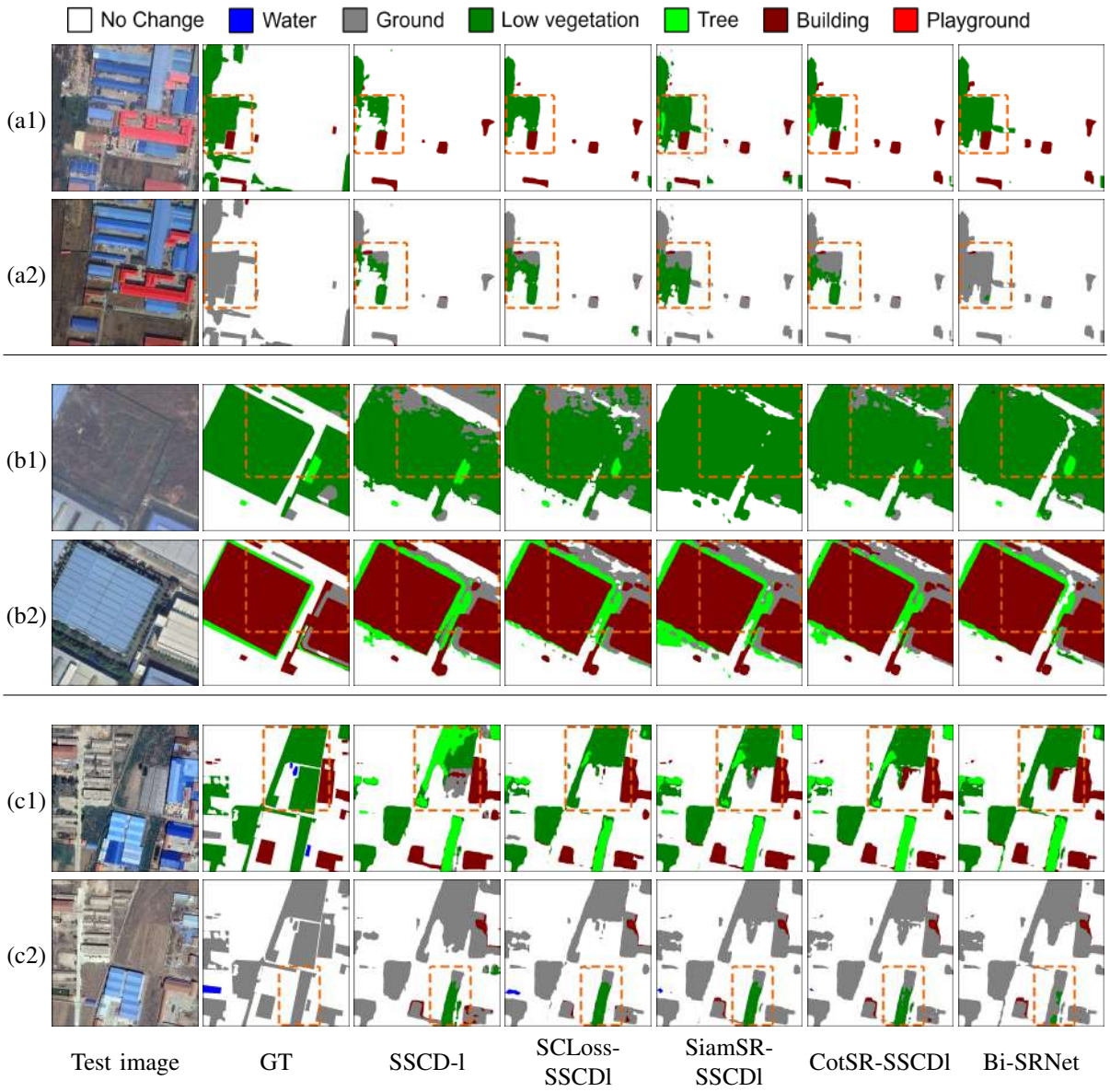


Fig. 7: Example of results provided by different proposed methods in the ablation study. The major differences are highlighted in orange rectangles.

VI. Conclusions

In this study, we investigated to improve the SCD task. First, we summarized the existing CNN architectures for SCD and identified the main limitations in semantic embedding and CD. Accordingly, we proposed a novel SCD architecture, the SSCD-l, where the semantic temporal features are merged in a deep convolutional CD block. Then, we further extend the SSCD-l architecture into the Bi-SRNet by introducing several semantic modelling designs. Three auxiliary designs are introduced, including two Siam-SR blocks to augment temporal information, a Cot-SR block to model temporal correlations, as well as a SCLoss to enhance temporal coherence. Finally, a set of experiments have been conducted to evaluate the effectiveness of the proposed methods.

Through experiments we found that: i) The proposed SSCD-l architecture outperforms other standard SCD architectures by a large margin; ii) The proposed Bi-SRNet containing semantic reasoning designs further improves the SSCD-l not only in the segmentation of LC classes, but also in the detection of changes; iii) The SSCD-l based methods (standard SSCD-l and the Bi-SRNet) outperform SOTA methods and obtain the highest accuracy metrics on the SECOND. The advantages of the Bi-SRNet are two-fold. First, in the SSCD-l architecture, the sub-tasks in SCD (SS and CD) are disentangled (with separate outputs and loss functions) but are deeply integrated (through re-use of the semantic features in the CD block). This leads to more accurate CD results. Second, the spatial and temporal correlations are modelled in the Bi-SRNet with both the SR blocks and the SCLoss. This improves the temporal consistence and enables the Bi-SRNet to discriminate better the LC classes in critical areas.

One of the remaining problems is to model the temporal

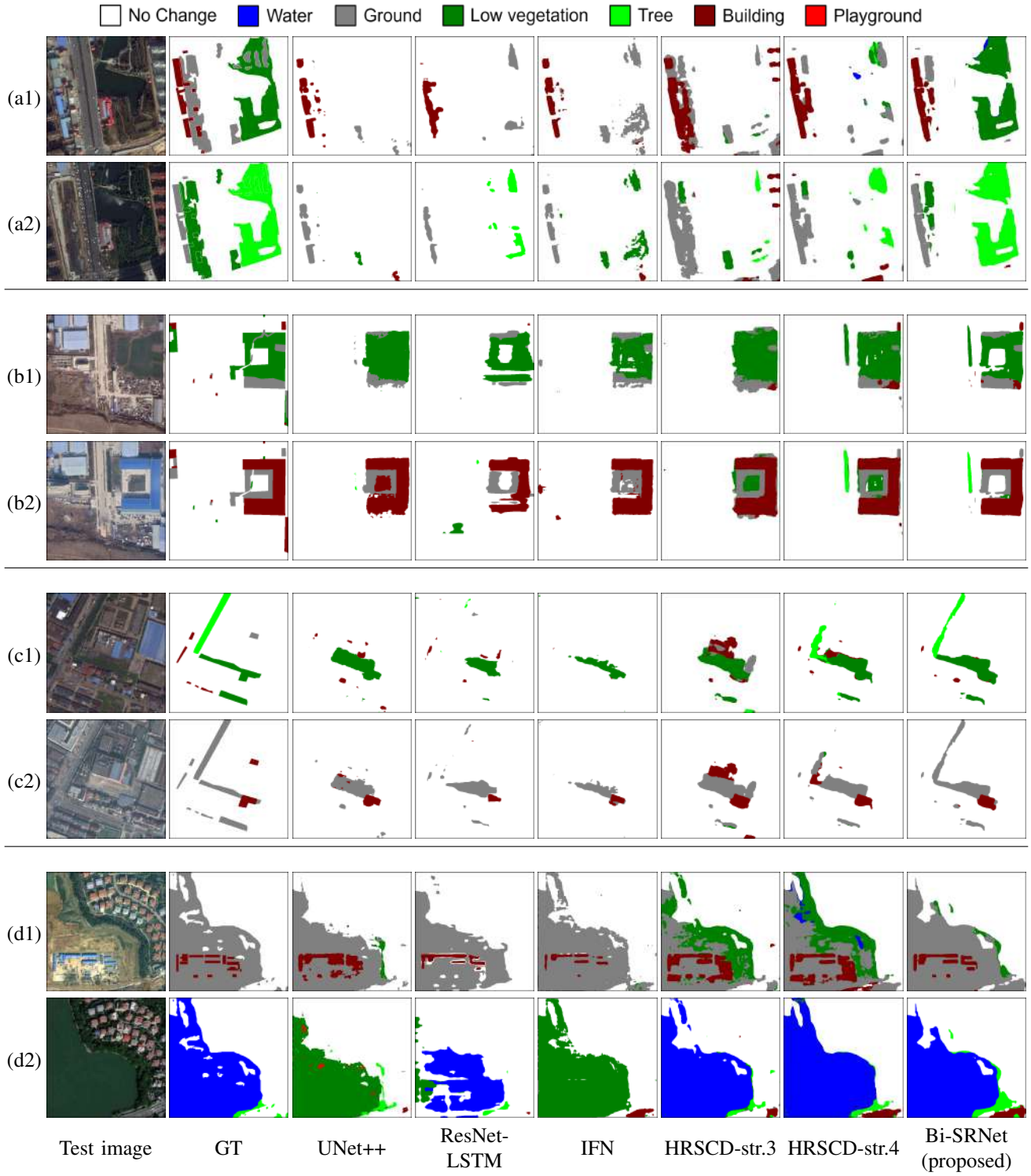


Fig. 8: Example of results provided by different methods in the comparative experiments.

correlation of LCLU classes especially in the changed areas, which is not fully exploited in the proposed method. The learning of LCLU transition types can potentially contribute to the recognition of semantic classes [7]. To model these transitions, more connections can be established between the CD unit and the temporal branches (in the DSCD-l architecture), which is left for future studies.

REFERENCES

[1] 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.

[2] F. Bovolo and L. Bruzzone, "A theoretical framework for unsupervised change detection based on change vector analysis in the polar domain," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 45, no. 1, pp. 218–236, 2007.

[3] F. Bovolo, S. Marchesi, and L. Bruzzone, "A framework for automatic and unsupervised detection of multiple changes in multitemporal images," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 50, no. 6, pp. 2196–2212, 2012.

[4] 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.

[5] 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.

[6] L. Bruzzone, D. F. Prieto, and S. B. Serpico, "A neural-statistical approach to multitemporal and multisource remote-sensing image classification," *IEEE Transactions on Geoscience and remote Sensing*, vol. 37, no. 3, pp. 1350–1359, 1999.

[7] 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.

[8] H. Kataoka, S. Shirakabe, Y. Miyashita, A. Nakamura, K. Iwata, and Y. Satoh, "Semantic change detection with hypermaps," *arXiv preprint arXiv:1604.07513*, vol. 2, no. 4, 2016.

[9] 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.

[10] J. Long, E. Shelhamer, and T. Darrell, "Fully convolutional networks for semantic segmentation," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2015, pp. 3431–3440.

[11] D. Peng, Y. Zhang, and H. Guan, "End-to-end change detection for high resolution satellite images using improved unet++," *Remote Sensing*, vol. 11, no. 11, p. 1382, 2019.

[12] 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.

[13] 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.

[14] A. Singh, "Digital change detection techniques using remotely-sensed data," *Int. J. Remote Sensing*, vol. 10, no. 6, pp. 989–1003, 1989.

[15] 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.

[16] Y. Ban and O. Yousif, *Change Detection Techniques: A Review*. Cham: Springer International Publishing, 2016, pp. 19–43.

[17] S. Liu, D. Marinelli, L. Bruzzone, and F. Bovolo, "A review of change detection in multitemporal hyperspectral images: Current techniques, applications, and challenges," *IEEE Geoscience and Remote Sensing Magazine*, vol. 7, no. 2, pp. 140–158, 2019.

[18] W. A. Malila, "Change vector analysis: an approach for detecting forest changes with landsat," in *LARS symposia*, 1980, p. 385.

[19] S. Liu, L. Bruzzone, F. Bovolo, M. Zanetti, and P. Du, "Sequential spectral change vector analysis for iteratively discovering and detecting multiple changes in hyperspectral images," *IEEE transactions on geoscience and remote sensing*, vol. 53, no. 8, pp. 4363–4378, 2015.

[20] S. Liu, Q. Du, X. Tong, A. Samat, L. Bruzzone, and F. Bovolo, "Multiscale morphological compressed change vector analysis for unsupervised multiple change detection," *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, vol. 10, no. 9, pp. 4124–4137, 2017.

[21] A. A. Nielsen, K. Conradsen, and J. J. Simpson, "Multivariate alteration detection (mad) and maf postprocessing in multispectral, bitemporal image data: New approaches to change detection studies," *Remote Sensing of Environment*, vol. 64, no. 1, pp. 1 – 19, 1998.

[22] A. A. Nielsen, "The regularized iteratively reweighted mad method for change detection in multi- and hyperspectral data," *IEEE Transactions on Image Processing*, vol. 16, no. 2, pp. 463–478, 2007.

[23] H. Nemmour and Y. Chibani, "Support vector machines for automatic multi-class change detection in algerian capital using landsat tm imagery," *Journal of the Indian Society of Remote Sensing*, vol. 38, no. 4, pp. 585–591, 2010.

[24] X. Wang, S. Liu, P. Du, H. Liang, J. Xia, and Y. Li, "Object-based change detection in urban areas from high spatial resolution images based on multiple features and ensemble learning," *Remote Sensing*, vol. 10, no. 2, 2018.

[25] X. Wang, P. Du, D. Chen, S. Liu, W. Zhang, and E. Li, "Change detection based on low-level to high-level features integration with limited samples," *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, vol. 13, pp. 6260–6276, 2020.

[26] W. Zhang, X. Lu, and X. Li, "A coarse-to-fine semi-supervised change detection for multispectral images," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 56, no. 6, pp. 3587–3599, 2018.

[27] S. Liu, X. Tong, L. Bruzzone, and P. Du, "A novel semisupervised framework for multiple change detection in hyperspectral images," in *2017 IEEE International Geoscience and Remote Sensing Symposium*. IEEE, 2017, pp. 173–176.

[28] S. Liu, Q. Du, X. Tong, A. Samat, and L. Bruzzone, "Unsupervised change detection in multispectral remote sensing images via spectral-spatial band expansion," *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, vol. 12, no. 9, pp. 3578–3587, 2019.

[29] M. Volpi, D. Tuia, G. Camps-Valls, and M. F. Kanevski, "Unsupervised change detection with kernels," *IEEE Geoscience Remote Sensing Letter*, vol. 9, no. 6, pp. 1026–1030, 2012.

[30] P. Du, S. Liu, L. Bruzzone, and F. Bovolo, "Target-driven change detection based on data transformation and similarity measures," in *2012 IEEE International Geoscience and Remote Sensing Symposium*, July 2012, pp. 2016–2019.

[31] W. Zhang and X. Lu, "The spectral-spatial joint learning for change detection in multispectral imagery," *Remote Sensing*, vol. 11, no. 3, p. 240, 2019.

[32] 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.

[33] A. Song, J. Choi, Y. Han, and Y. Kim, "Change detection in hyperspectral images using recurrent 3d fully convolutional networks," *Remote Sensing*, vol. 10, no. 11, p. 1827, 2018.

[34] 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.

[35] Y. Zhan, K. Fu, M. Yan, X. Sun, H. Wang, and X. Qiu, "Change detection based on deep siamese convolutional network for optical aerial images," *IEEE Geoscience and Remote Sensing Letters*, vol. 14, no. 10, pp. 1845–1849, 2017.

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

[37] W. Zhao, L. Mou, J. Chen, Y. Bo, and W. J. Emery, "Incorporating metric learning and adversarial network for seasonal invariant change detection," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 58, no. 4, pp. 2720–2731, 2020.

[38] 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.

[39] Y. Li, C. Peng, Y. Chen, L. Jiao, L. Zhou, and R. Shang, "A deep learning method for change detection in synthetic aperture radar images," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 57, no. 8, pp. 5751–5763, 2019.

[40] X. Li, Z. Yuan, and Q. Wang, "Unsupervised deep noise modeling for hyperspectral image change detection," *Remote Sensing*, vol. 11, no. 3, p. 258, 2019.

[41] M. Yang, L. Jiao, F. Liu, B. Hou, and S. Yang, "Transferred deep learning-based change detection in remote sensing images," *IEEE Trans-*

actions on Geoscience and Remote Sensing, vol. 57, no. 9, pp. 6960–6973, 2019.

[42] Q. Wang, Z. Yuan, Q. Du, and X. Li, "GETNET: A general end-to-end 2-D CNN framework for hyperspectral image change detection," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 57, no. 1, pp. 3–13, 2019.

[43] Z. Yuan, Q. Wang, and X. Li, "Robust pcanet for hyperspectral image change detection," in *2018 IEEE International Geoscience and Remote Sensing Symposium*, 2018, pp. 4931–4934.

[44] B. Hou, Y. Wang, and Q. Liu, "Change detection based on deep features and low rank," *IEEE Geoscience and Remote Sensing Letters*, vol. 14, no. 12, pp. 2418–2422, 2017.

[45] S. Saha, F. Bovolo, and L. Bruzzone, "Building change detection in vhr sar images via unsupervised deep transcoding," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 59, no. 3, pp. 1917–1929, 2020.

[46] L. T. Luppino, M. Kampffmeyer, F. M. Bianchi, G. Moser, S. B. Serpico, R. Jenssen, and S. N. Anfinsen, "Deep image translation with an affinity-based change prior for unsupervised multimodal change detection," *IEEE Transactions on Geoscience and Remote Sensing*, 2021.

[47] J. Liu, M. Gong, A. K. Qin, and P. Zhang, "A deep convolutional coupling network for change detection based on heterogeneous optical and radar images," *IEEE Trans. Neural Networks Learn. Syst.*, vol. 29, no. 3, pp. 545–559, 2018.

[48] X. Li, Z. Du, Y. Huang, and Z. Tan, "A deep translation (gan) based change detection network for optical and sar remote sensing images," *ISPRS Journal of Photogrammetry and Remote Sensing*, vol. 179, pp. 14–34, 2021.

[49] 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.

[50] 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.

[51] D. Wang, F. Zhao, C. Wang, H. Wang, F. Zheng, and X. Chen, "Y-net: A multiclass change detection network for bi-temporal remote sensing images," *International Journal of Remote Sensing*, vol. 43, no. 2, pp. 565–592, 2022.

[52] L. Mou, Y. Hua, and X. X. Zhu, "A relation-augmented fully convolutional network for semantic segmentation in aerial scenes," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2019, pp. 12 416–12 425.

[53] J. Zhang, S. Lin, L. Ding, and L. Bruzzone, "Multi-scale context aggregation for semantic segmentation of remote sensing images," *Remote Sensing*, vol. 12, no. 4, p. 701, 2020.

[54] L.-C. Chen, Y. Zhu, G. Papandreou, F. Schroff, and H. Adam, "Encoder-decoder with atrous separable convolution for semantic image segmentation," in *ECCV*, 2018.

[55] 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.

[56] X. Wang, R. Girshick, A. Gupta, and K. He, "Non-local neural networks," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2018, pp. 7794–7803.

[57] H. Nam, J.-W. Ha, and J. Kim, "Dual attention networks for multimodal reasoning and matching," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2017, pp. 299–307.

[58] 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.

[59] 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.

[60] 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.

[61] 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.

[62] O. Ronneberger, P. Fischer, and T. Brox, "U-net: Convolutional networks for biomedical image segmentation," in *International Conference on Medical image computing and computer-assisted intervention*. Springer, 2015.

[63] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," in *3rd International Conference on Learning Representations, ICLR 2015, San Diego, CA, USA, May 7-9, 2015, Conference Track Proceedings*, Y. Bengio and Y. LeCun, Eds., 2015.

Lei Ding received the MS's degree in Photogrammetry and Remote Sensing from the Information Engineering University (Zhengzhou, China), and the PhD (cum laude) in Communication and Information Technologies from the University of Trento (Trento, Italy). He is now a lecturer at the PLA Strategic Force Information Engineering University. His research interests are related to semantic segmentation, change detection and domain adaptation with Deep Learning techniques. He is a referee for many international journals, including IEEE TIP, TNNLS, TGRS, GRSL and JSTAR.

Haitao Guo received his M.S. degree and Ph.D. degree from Information Engineering University, China, in 2002 and 2008, respectively. He is currently an associate professor of photogrammetry and remote sensing at Information Engineering University, Zhengzhou, China, where he teaches digital photogrammetry and geopositioning for remote sensing imagery. His current research interests are in the areas of deep learning for image interpretation and change detection, geopositioning without ground control points for satellite imagery.

Sicong Liu (S'13–M'15-SM'21) received the B.Sc. degree in geographical information system and the M.E. degree in photogrammetry and remote sensing from the China University of Mining and Technology, Xuzhou, China, in 2009 and 2011, respectively, and the Ph.D. degree in information and communication technology from the University of Trento, Trento, Italy, 2015.

He is currently an Associate Professor with the College of Surveying and Geo-Informatics, Tongji University, Shanghai, China. His research interests include multitemporal data analysis, change detection, multispectral/hyperspectral remote sensing and planetary remote sensing. Dr. Liu was the winner (ranked as third place) of Paper Contest of the 2014 IEEE GRSS Data Fusion Contest. He is the Technical Co-Chair of the Tenth International Workshop on the Analysis of Multitemporal Remote Sensing Images (MultiTemp 2019). He serves as the Program Committee Member for SPIE Remote Sensing Symposium: Image and Signal Processing for Remote Sensing XXVI-XXVIII (2020-2022), and also served as the Session Chair for many international conferences such as International Geoscience and Remote Sensing Symposium (2017-2019). He is/was a Guest Editor for the IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing (JSTARS) and Remote Sensing.

Lichao Mou received the Bachelor's degree in automation from the Xi'an University of Posts and Telecommunications, Xi'an, China, in 2012, the Master's degree in signal and information processing from the University of Chinese Academy of Sciences (UCAS), China, in 2015, and the Dr.-Ing. degree from the Technical University of Munich (TUM), Munich, Germany, in 2020.

He is currently a Guest Professor at the Munich AI Future Lab AI4EO, TUM and the Head of Visual Learning and Reasoning team at the Department "EO Data Science", Remote Sensing Technology Institute (IMF), German Aerospace Center (DLR), Wessling, Germany. Since 2019, he is a Research Scientist at DLR-IMF and an AI Consultant for the Helmholtz Artificial Intelligence Cooperation Unit (HAICU). In 2015 he spent six months at the Computer Vision Group at the University of Freiburg in Germany. In 2019 he was a Visiting Researcher with the Cambridge Image Analysis Group (CIA), University of Cambridge, UK.

He was the recipient of the first place in the 2016 IEEE GRSS Data Fusion Contest and finalists for the Best Student Paper Award at the 2017 Joint Urban Remote Sensing Event and 2019 Joint Urban Remote Sensing Event.

Jing Zhang received a master's degree in software engineering from Beijing University of Technology. She is currently a Ph.D student at the department of Information Engineering and Computer Science, University of Trento, Italy. Her current research interests are related to the change detection and semantic segmentation of remote sensing image.

Lorenzo Bruzzone (S'95-M'98-SM'03-F'10) 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 the 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 coauthor) 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 keynote speaker in more than 30 international conferences and workshops. Since 2009 he is 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 27000) and the value of the h-index (78) (source: Google Scholar).