MSCD-Net: From Unimodal to Multimodal Semantic Change Detection

Jian Wang, Hong Xie, Li Yan, Tingyuan Zhou, Yanheng Wang, Jing Zhang, Lorenzo Bruzzone, and Peter M. Atkinson

***Abstract*—Semantic change detection (SCD) involves temporal changes and spatial semantics. Its working principle and processing flow usually include land semantic segmentation (LSS) and binary change detection (BCD). Due to its significant impact and practical value, SCD has received consistently wide attention in Earth observation. Nowadays, remote sensing data in various modalities are proliferating, calling for an urgent need to develop intelligent algorithms for multimodal remote sensing data. However, no efficient multimodal SCD methods exist currently. To address this limitation, this work proposes the first deep learning-based multimodal SCD method: MSCD-Net. MSCD-Net extracts multi-scale semantic and difference features after fusing multimodal features, and then aggregates and refines these features to output high-quality semantic segmentation and change maps. Additionally, a semantic difference decoder (SDD) module is designed to model semantic and difference features jointly. It can be integrated with existing methods to increase accuracy. Experimental results demonstrate that MSCD-Net achieves state-of-the-art performance on both multimodal and unimodal SCD datasets, and SDD has strong feature learning ability and compatibility. These findings imply that MSCD-Net is expected to promote the development and application of multimodal SCD.**

***Index Terms*—Land semantic segmentation, change detection, multimodal data, semantic change detection, remote sensing**

I. INTRODUCTION

SEMANTIC change detection (SCD) is a specialized task within change detection (CD) that provides detailed information on land-cover/land-use (LCLU) changes. SCD involves spatial semantics and temporal changes, with its principles and workflow typically including land semantic segmentation (LSS) and binary change detection (BCD). Consequently, SCD plays a significant role in, and fulfills the extensive requirements of, land resource surveys, emergency response, disaster monitoring, military reconnaissance, map updating and various other scenarios [1], [2], [3]. SCD has experienced significant growth in recent years, driven by rapid advances in computing power, artificial intelligence and contributions from the research community [4], [5], [6].

BCD has been consistently the most common and widely researched sub-task of CD. Early BCD methods included image difference, ratio and log operations [7]. During the same period, the intuitive and effective post-classification comparison method emerged. These methods leveraged the spectral properties of remote sensing (RS) data to identify changes by comparing images from different time points. Subsequently, more sophisticated algorithms, such as those based on decision trees, CVA [8], MAD [9] and MRF [10] were developed. These methods are primarily data-driven, often utilizing small datasets tailored to specific types or scenarios. They focus on analyzing the distribution patterns of changes in specific images based on physical characteristics, such as reflectance and spatial shape [11], [12]. As machine learning began to be incorporated into change detection, transform-based and object-based methods became gradually mainstream. Representative methods involve IR-MAD [13], RCVA [14], SFA and DSFA [15]. These methods, characterized by weak artificial intelligence, provide a more nuanced understanding of changes and enhance BCD accuracy through manual feature engineering. In the past decade, deep learning (DL) -based methods have been developed and applied widely, such as SST-Former [16], SSN-Siam-conc [17], STS-STAM-CMR [18] and FDCNN [19]. These methods excel at automatically learning features from large datasets, reducing the need for well-designed manual features and significantly improving BCD accuracy and efficiency.

The purpose of BCD is only to identify the location and area of changes, without concern for the property of those changes. However, in many situations, it is crucial not only to know where changes have occurred, but also to understand the detailed semantics before and after the change event. This is the focus of SCD. Due to the complexities and challenges of the SCD task compared to BCD, early SCD studies focused primarily on small areas and single LCLU categories, with very limited public datasets available [20], [21]. As a result, SCD developed more slowly and produced fewer research results than BCD. Remarkably, there has been no DL-based

This work was supported in part by the National Natural Science Foundation of China under Grants 42394061 and 42371451, in part by the Science and Technology Major Project of Hubei Province under Grant 2021AAA010, in part by the Open Fund of Hubei Luojia Laboratory under Grant 220100053, and in part by the State Scholarship Fund of China. Jian Wang and Hong Xie contributed equally. *(Corresponding author: Li Yan.)*

Jian Wang, Hong Xie, and Li Yan are with the School of Geodesy and Geomatics, Hubei Luojia Laboratory, Wuhan University, Wuhan 430079, China. Jian Wang is also with the Faculty of Science and Technology, Lancaster University, Lancaster, LA1 4YQ, UK (e-mail: wj_sgg@whu.edu.cn; hxie@sgg.whu.edu.cn; lyan@sgg.whu.edu.cn).

Tingyuan Zhou and Peter M. Atkinson are with the Faculty of Science and Technology, Lancaster University, Lancaster, LA1 4YQ, UK. Peter M. Atkinson is also with the School of Geography and Environmental Science, University of Southampton, Highfield, Southampton, SO17 1BJ, UK (e-mail: rcdzhouty@gmail.com; pma@lancaster.ac.uk).

YanHeng Wang is with the College of Intelligent Systems Science and Engineering, Harbin Engineering University, Harbin 150001, China (e-mail: b220040023@hrbeu.edu.cn).

Jing Zhang and Lorenzo 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).

SCD method for a long time. During this period, high-impact traditional SCD methods include ICC [22] PCA [23], C2VA[24] and S2CVA[25]. This situation changed after the HRSCD [26] and SECOND [27] datasets were introduced in 2019. Many DL-based SCD methods, such as TCRPN [28], Bi-SRNet [29], SAAN [30], ChangeMask [31] and CLAFA [32], have been proposed within the last few years. These methods produced outstanding performance compared to traditional SCD methods.

Modality refers to the way in which something expressed or perceived. Each source or form of information can be regarded as a modality. Multimodal involves multiple modalities, which typically manifests in three forms: 1) multimedia data describing the same object, such as images, audio, text, and videos; 2) the same type of media data from different sensors, such as image data captured by various imaging sensors; 3) symbolic representations and information with distinct structural or representational characteristics, such as different languages [33]. With the expansion of sensor types, platforms and data volumes, coupled with the extension of application scenarios, demand for the development of high-performance algorithms has become imperative [34], [35], [36], [37]. Additionally, the explosion of "big data" and the rise of strong artificial intelligence methodologies have established multimodality as a significant trend. Generally, unimodal data processing can be seen as a special case of multimodal processing. Therefore, constructing a high-performance SCD method suitable for both multimodal and unimodal data is crucial for enhancing the influence and applicability of SCD. However, after reviewing the current state of SCD development, we found that predominantly existing methods were designed to process merely unimodal RS data but neglect multimodal RS data, and no DL-based research has been conducted on multimodal SCD. As a result, the development of multimodal SCD has not kept pace with community demands.

In light of this motivation, we propose a pioneering multimodal SCD method, MSCD-Net, which unifies the SCD tasks for multimodal and unimodal RS data. The MSCD-Net can efficiently process multimodal or unimodal data and output accurate SCD maps in a fully end-to-end manner. Furthermore, we build a practical semantic difference decoder (SDD) that features strong intrinsic feature learning ability and high compatibility. The SDD can be integrated with existing methods to enhance accuracy significantly. Experimental results manifest that MSCD-Net achieves the highest accuracy across multimodal and unimodal SCD datasets, and it holds significant potential to advance the development and unification of SCD methods.

The remainder of this paper is organized as follows. Section II details related work. Section III elaborates on the principles of MSCD-Net. Section IV demonstrates the experiments and results, while Section V draws the conclusions.

II. Related Work

A. Binary Change Detection

1) Unimodal BCD: Recently, there has been much research on unimodal BCD. Typically, Zhang and Shi [19] pioneered a high-resolution RS image CD framework using a deep feature difference convolutional neural network (FDCNN), which learns the deep features and then generates multi-scale and multi-depth feature difference maps for CD. Liu et al. [38] presented the local restricted CNN (LRCNN) to detect changed areas in multi-temporal polarimetric synthetic aperture radar (SAR) images by imposing a local spatial constraint on the output layer of the CNN. However, high-resolution RS BCD remains challenging due to the complexity of objects in the scene. To this end, Li et al. [39] proposed a deep-supervised dual discriminative metric network (SDMNet) by combining a discriminative implicit metric module and multiple losses. The SDMNet can effectively distinguish changes of interest and pseudo-changes in high-resolution RS images. Cao et al. [40] proposed a multi-scale weakly supervised learning method, which utilizes a large number of single-temporal high-resolution images and image-level labels to detect changes in built-up area.

Because BCD pipelines based on CNNs fail to adequately capture long-range concepts in space-time, Chen et al. [41] proposed a bitemporal image transformer (BIT) to model contexts within the spatial-temporal domain. Furthermore, CNN methods often focus on the extraction of spatial information, but ignore important spectral and temporal sequences. To deal with this limitation, Wang et al. [16] proposed a joint spectral, spatial and temporal transformer for hyperspectral image change detection, named SST-Former. Considering that the CD task commonly has the problem of class imbalance (i.e., unchanged samples far outnumber changed samples), Mou et al. [42] explored the one-class CD and proposed a data-enclosing-ball minimizing autoencoder (DebM-AE) that is trained with reconstruction error and a minimum volume criterion.

To alleviate the labeling cost, numerous unsupervised BCD methods have been developed, especially for SAR data, as obtaining a substantial number of labeled samples for SAR data is challenging. Zhang et al. [43] proposed an unsupervised approach to small area BCD using multi-scale superpixel reconstruction and a two-stage centre-constrained fuzzy *c*-means clustering algorithm. Subsequently, Zhang et al. [44] proposed an unsupervised BCD method called adaptive contourlet fusion clustering based on adaptive contourlet fusion and fast non-local clustering for multi-temporal SAR images. Nevertheless, unsupervised methods usually exhibit low accuracy due to lacking constraints or guidance during training. To tackle this limitation, Ji et al. [17] proposed an end-to-end unsupervised BCD network based on self-adaptive superpixel segmentation. Yan et al. [45] proposed a domain knowledge-guided self-supervised learning BCD by associating the domain knowledge of RS.

2) Multimodal BCD: Currently, the amount of data available in various modalities has increased rapidly with the development of new types, and additional numbers, of sensors and platforms [46], [47], [48], [49], [50]. The availability and quality of unimodal data are often constrained in specific scenarios. Additionally, many practical applications require fine temporal resolution, such as military reconnaissance and disaster assessment, whereas acquiring multi-temporal unimodal data usually requires a long period. In this context,

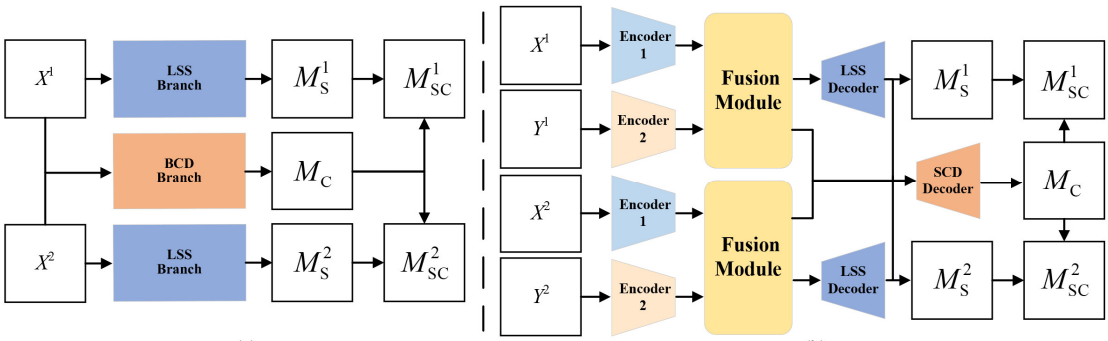


Fig. 1. The SCD paradigms of unimodal and multimodal data. (a) unimodal data and (b) multimodal data.

multimodal BCD brings obvious utility.

Some multimodal BCD research focused on supervised methods. For example, Lv et al. [51] proposed a hierarchical attention feature fusion (HAFF) -based network by integrating multi-scale convolution fusion filters to explore the global semantic features of the targets of interest from multiple perspectives. Because pairing and annotating multimodal RS images is both expensive and time-consuming, much research has aimed to develop unsupervised methods for multimodal BCD. These unsupervised multimodal BCD methods can be divided broadly into three classes: classification, transformation and discrimination. The classification methods first classify multimodal images. Subsequently, the derived classification outcomes can be compared directly to identify changes, such as the multidimensional evidential reasoning method, post-classification comparison method and compound classification method [52], [53]. Since unsupervised classification models struggle to obtain accurate classification results, the classification methods are susceptible to the accumulation of classification errors [54].

In general, the core objective of transformation methods is to make the multimodal images comparable. Most transformation methods aim to either transfer "incomparable" images to a common domain or transform one image to the domain of another, thereby rendering them "comparable" [55]. The former transformation approaches can be categorized into: 1) feature space-based methods [56], [57], [58], and 2) DL-based methods [59], [60], [61]. The latter transformation approaches can be viewed as image regression or image translation, and can be categorized into: 1) classical signal-processing methods [62], [63], [64] and 2) DL-based methods [65], [66], [67].

The discrimination methods are an emerging approach. The methods are intuitive and represented by self-supervised contrastive learning methods [2], [68], [69], which discriminate the characteristics between the dual stream outputs of the network by designing appropriate positive and negative samples and a loss function. When the distance between the positive and negative samples is maximized while the loss is minimized, one obtains the best model, thereby inferring the difference image and change map [45], [2].

The above BCD works have greatly expanded the CD method library and enhanced CD applicability in various scenarios. However, they do not provide the semantic properties of the changed targets, which are in great demand in engineering applications.

B. Semantic Change Detection

Nowadays, representative methods exist for unimodal SCD. Daudt et al. [26] proposed four SCD strategies, among which the two most effective strategies are HRSCD-str3 and -str4. The str3 fuses bitemporal data early and inputs them into the CD branch, while str4 is based on str3 and appends the features extracted by the LSS encoder into the CD decoder. Mou et al. [5] proposed a recurrent convolutional neural network architecture, which is trained to learn a joint spectral-spatial-temporal feature representation in a unified framework for CD in multispectral images. These two works have sparked enthusiasm for subsequent SCD research. Zheng et al. [31] proposed the ChangeMask by exploring two inductive biases: sematic-change causal relationship and temporal symmetry. Xia et al. [70] proposed a deep Siamese post-classification fusion network to alleviate the accumulation of misclassification errors in post-classification method. Ding et al. [29] summarized four feasible CNN architectures for the SCD. These derivative researches exhibit higher accuracy.

In addition, many researchers believe that multi-level feature interaction benefits the performance of SCD task. Wang et al. [32] proposed an approach characterized by two attentive feature aggregation schemes that handle cross-level features in different processes. Guo et al. [30] designed a similarity-aware attention flow network (SAAN). The SAAN incorporates a similarity-guided attention flow module with deeply supervised similarity optimization to achieve effective change detection. Zhang et al. [71] presented the VFM-ReSCD architecture, which combines vision foundation models and multi-level decoder for SCD on RS images to learn sufficient LCLU transition information.

By summarizing the related works for the SCD task, we found that no DL-based methods related to multimodal SCD have been proposed. However, given the massive amount of multimodal data in current engineering applications, there is a need for methods capable of efficiently obtaining high-precision semantic changes for LCLU. Therefore, constructing a high-performance SCD method suitable for both multimodal and unimodal data is of great significance for increasing the influence and applicability of SCD.

To deal with the above limitations, we aimed to develop a method suitable for both multimodal and unimodal SCD. Rethinking related SCD methods, we found that the architecture commonly adopted in the SCD task is a three-branch structure, as shown in Fig. 1 (a). The bitemporal

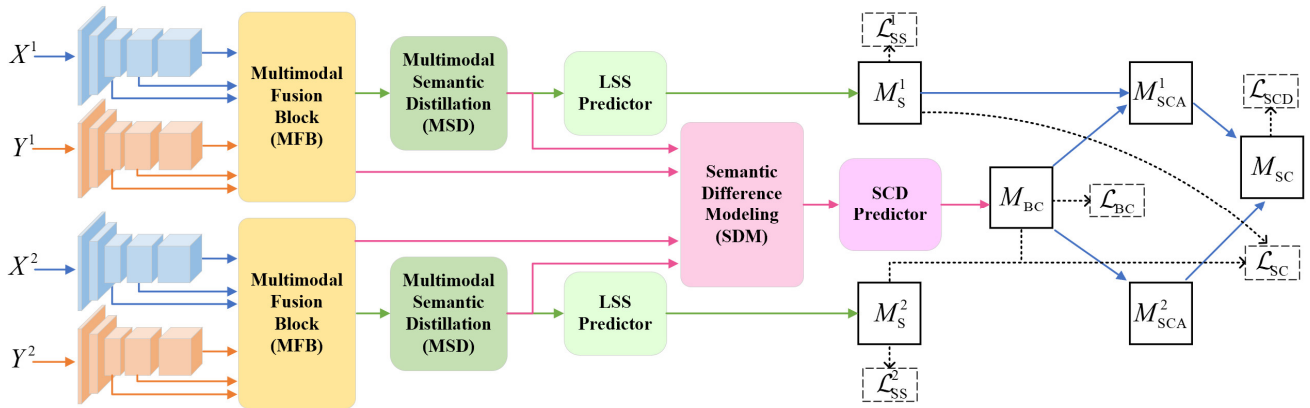


Fig. 2. Overview of the MSCD-Net.

unimodal data passes through two LSS branches and one BCD branch, outputting the following five products: bitemporal semantic segmentation maps M_S^1 and M_S^2, binary change map M_{BC}, bitemporal semantic change maps M_{SC}^1 and M_{SC}^2.

Although some existing methods have achieved a good performance, most of them do not focus on mining deep difference features in the CD branch and, rather, rely too heavily on the features extracted by the LSS branch. This is reflected in the CD decoder, which typically processes the semantic features extracted by the LSS branch through simple convolution blocks [28], [72], [26], [30], [32], concatenation or mapping [31], [29], [5], before outputting the binary changes. In addition, most existing SCD methods use a patch-based pure convolutional architecture, which cannot model long-range correlations [26], [31], [28], [29], [71], [5], [30], [32], [72]. This leads to an excessive focus on specific local details and difficulty in capturing global context. Consequently, the generated change maps often contain many missed detections and false alarms.

III. Methodology

First, we conceived a paradigm for multimodal SCD by considering the key components and challenges of multimodal data processing, as shown in Fig. 1 (b). Its main characteristics include: 1) multiple Siamese encoders that extract features from multimodal data; 2) a module dedicated to the fusion of multimodal features; 3) a SCD decoder responsible for extracting global context and modeling the intrinsic association between semantic features and change features. In this way, the model not only receives the semantic features extracted by the LSS branch but also incorporates the fused features, thereby avoiding over-reliance on semantic features and increasing the quality of difference features and the robustness of the change maps.

Subsequently, we proposed the MSCD-Net based on the conceived paradigm for multimodal SCD. An overview of the MSCD-Net is shown in Fig. 2. It contains two Siamese encoders corresponding to the two input data with different modalities, which extract multi-scale features from the multimodal data. Two multimodal fusion blocks (MFBs) are built to fuse the multimodal features. The multimodal semantic distillation (MSD) module and LSS predictor are responsible for distilling the high-quality multi-scale land semantic features and predicting semantic map, respectively. A semantic difference modeling (SDM) module is designed to aggregate and refine multi-scale semantic difference features. The SCD predictor outputs the binary change map. Finally, the SCD map is derived by the binary change map and bitemporal land semantic maps.

During the optimization process, we designed four different loss functions to provide extensive supervision signals, involving the semantic map, the BCD map, the semantic consistency and the SCD map. All modules collaborate to distill, fuse and enhance various features embedded in the multimodal data and guide the model to learn numerous task-specific patterns, thereby enabling the model to acquire profound knowledge for SCD.

A. Encoder

The purpose of the encoder is to extract multimodal features preliminarily. We employ two weight-shared networks (i.e., Siamese network) for each modal data to extract features of each modality independently at different times. Let X and Y be the data in different modalities, we have $F_s^t = \mathrm{E}_X(X^t)$ and $G_s^t = \mathrm{E}_Y(Y^t)$, where X^t and Y^t are the input modalities X and Y at time t, respectively; $\mathrm{E}_{(\cdot)}(\cdot)$ denotes the encode function, which can be implemented with most mainstream architectures like the CNN or Transformer; F_s^t and G_s^t are the output features with respect to modalities X and Y at scale s and at time t, respectively.

Noteworthy, traditional encoders in image processing with DL adopt downsampling at every stage to reduce computational load and extract multi-scale features. Nevertheless, this also causes a loss of local details, so a skip connection from encoder to decoder is used to alleviate this problem. Since the features extracted from multimodal data differ greatly before and after fusion, retaining directly the pre-fusion features for the post-fusion decoder will impair performance. Therefore, we discard the downsampling operation and skip connection in the encoder after s reaching 2, to preserve details and more accurately identify boundaries and small objects.

B. MFB

The function of MFB is to fuse the features extracted from the multimodal data. For clarity of presentation, we illustrate

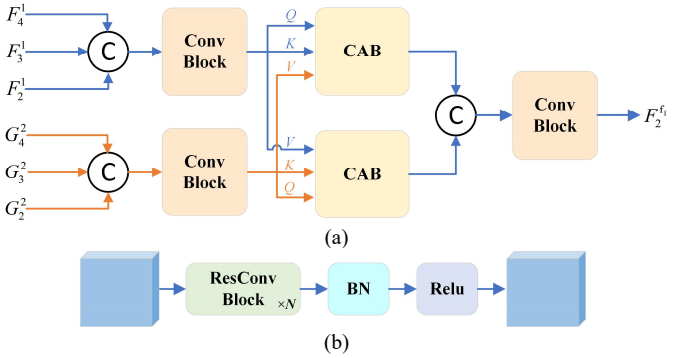


Fig. 3. The structure of the (a) MFB module and (b) conv block.

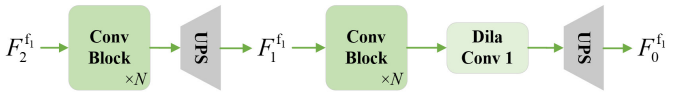


Fig. 4. The pipeline of the MSD module.

the MFB pipeline at the first time (t=1), as shown in Fig. 3 (a).

The MFB contains three conv blocks and two cross-attention blocks (CABs) to interact and fuse multimodal features, with its operation defined as:

$$F_2^{f_1} = \mathrm{CB}(\mathrm{CAT}(\mathrm{CAB}(\mathrm{CB}(\mathrm{CAT}(F_4^1, F_3^1, F_2^1), \mathrm{CAT}(F_4^2, F_3^2, F_2^2))))) \quad (1)$$

where CB(·), CAT(·), and CAB(·) represent the conv block, the concatenation and the cross-attention block, respectively. The features of any modality at the last three stages of the encoder are initially merged through the conv block after concatenating. Subsequently, cross-attention interactions are performed to associate multimodal features globally. Then, the multimodal features are concatenated into one feature, and the fused features $F_2^{f_t}$ are output after being refined by a conv block. The conv block is illustrated in Fig. 3 (b), which comprises several ResConv blocks, a batch normalization and a rectified linear unit (ReLU) activation function.

It should be noted that the attention mechanism in CAB is not a vanilla cross-attention. In the CAB, the query Q and key K come from the same modality, whereas the value V comes from another modality. By contrast, in the vanilla cross-attention, the Q comes from one modality, but the K and V from another. Although this seems to be a minor difference, the adopted attention mechanism retrieves information from another modality based on its own correlation rather than the correlation between itself and the other modality. Given that the correlation between multimodal data or features is often weak, but their spatial relationships are similar, this strategy is more suitable for multimodal feature fusion.

C. MSD and LSS Predictor

The MSD module can be considered a semantic segmentation sub-task, providing high-quality multi-scale land cover semantic features for CD tasks. The pipeline of a single MSD branch is illustrated in Fig. 4. The MSD module consists of two stages, each corresponding to a different feature scale, and both stages include a conv block and an upsampling operation.

Due to the large feature size, a dilated convolution block is utilized in the second stage to expand the convolutional receptive field and enhance multi-scale context information. Lastly, the semantic features $F_0^{f_t}$ are processed by LSS predictor to generate the semantic segmentation maps M_S^t.

D. SDM and SCD Predictor

The structure of the SDM module is shown in Fig. 5 (a). SDM takes as input the bitemporal features and multi-scale fused semantic features generated by the MSD and MFB modules. These features are merged through differentiation and concatenation operations to form a unified feature, which is then fed into the Semantic Difference Decoder (SDD) for joint modeling of both semantic and change features.

The SDD (see Fig. 5 (b)) we built is similar to the Transformer architecture, but the main difference is that SDD uses multilayer embedding, semantic-aware attention (SAA), shortcut with linear projection (LP) and upsampling. The multilayer embedding outputs feature embeddings suitable for SAA processing through multiple convolutional layers, while the upsampling converts feature scales.

In SAA, long-range associations and global information can be perceived in the entire image by performing attention operations in overlapped windows, which means that more samples are used for calculation and the associations between all targets can be obtained. This mechanism produces semantically strong and contextually rich difference representations and suppresses pseudo-change. The distinction between different categories is further expanded through multiplication and convolution operations.

Since deep features usually lack details, retaining the pre-fusion features in the post-fusion decoder weakens the feature quality and the final accuracy. To deal with this deficiency, we designed a shortcut operation with LP in SDD. This operation not only preserves high-quality features from the previous step, but also offers greater flexibility in reconstructing details. In particular, the linear projected embedding in the SAA is a weighted sum with globally long-range semantic association, and the semantic embedding is further strengthened through the conv block so that the SDD can reconstruct fine semantic difference features.

Algorithm 1. Forward of MSCD-Net

Input: multimodal data X^t and Y^t

Operation:

1: F_s^t and G_s^t are extracted from $\mathrm{E}_X(X^t)$ and $\mathrm{E}_Y(Y^t)$, respectively.

2: $F_2^{f_t}$ is derived by fusing F_s^t and G_s^t with MFB

3: $F_1^{f_t}$ and $F_0^{f_t}$ are generated via MSD, and then $F_0^{f_t}$ is transformed to M_S^t by the LSS predictor

4: $F_2^{f_t}$, $F_1^{f_t}$ and $F_0^{f_t}$ are integrated via SDM and SCD predictor to produce M_{BC}

5: M_{SCA}^t is derived by masking M_S^t with M_{BC}, and then M_{SC} is determined by comparing M_{SCA}^1 and M_{SCA}^2

Output: M_{SC}, M_{SCA}^t, M_{BC}, M_S^t

In this way, the semantic features and difference features in the spatiotemporal domain are jointly modeled by the

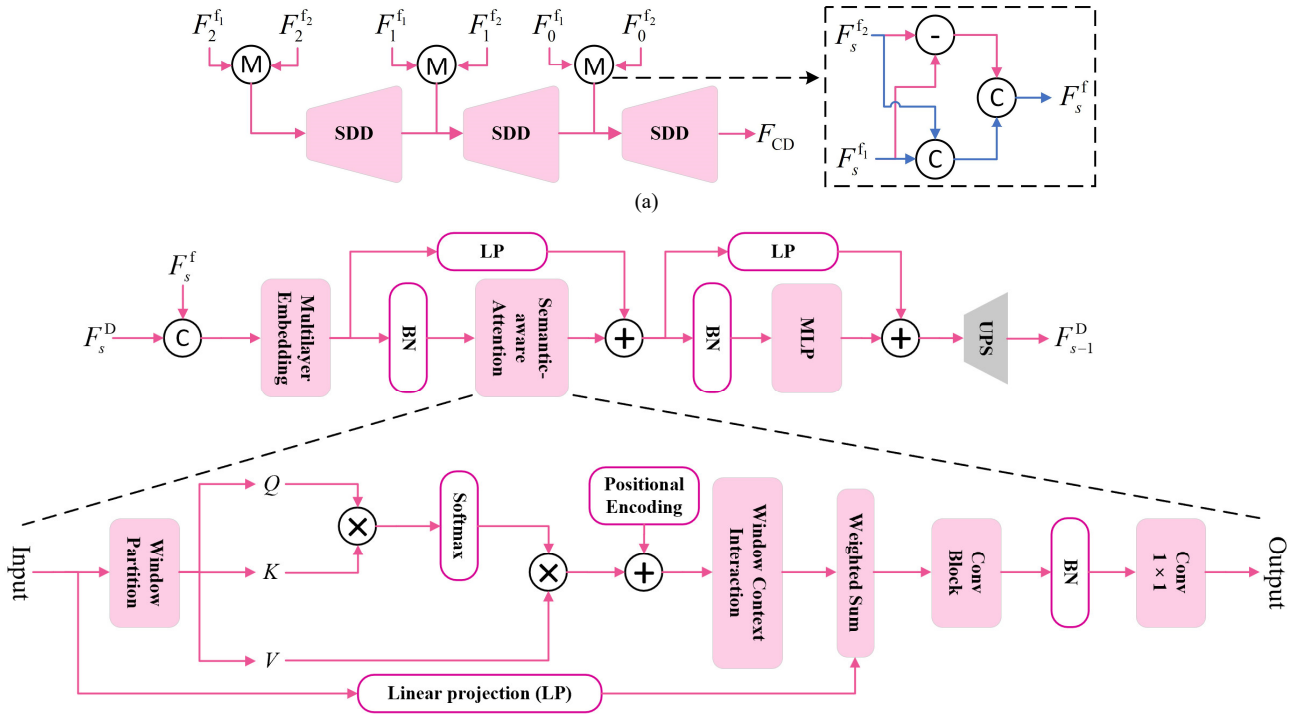


Fig. 5. The structure of the (a) SDM and (b) SDD modules.

interaction, differentiation and fusion of multi-temporal spatial features, and the dependency between the semantic features and the difference features is explored thoroughly. Multi-level recurrent aggregation can integrate the feature representations of various scales and dimensions, refine the semantic differences and enhance the CD robustness. Finally, the BCD map M_{BC} is produced by processing the ultimate difference feature F_{CD} through the SCD predictor. After we obtain M_{S}^{t} and M_{BC}, M_{SCA}^{t} and M_{SC} can be readily derived. The forward of MSCD-Net can be summarized as Algorithm 1.

E. Loss Function

We adopted four types of loss functions for MSCD-Net: semantic segmentation loss $\mathcal{L}_{\mathrm{SS}}$, binary change detection loss $\mathcal{L}_{\mathrm{BC}}$, semantic consistency loss $\mathcal{L}_{\mathrm{SC}}$ and semantic change detection loss $\mathcal{L}_{\mathrm{SCD}}$. The semantic segmentation loss $\mathcal{L}_{\mathrm{SS}}$ can be estimated by the cross-entropy between the bitemporal predicted land cover semantic maps and target labels, i.e.,

$$\begin{aligned}\mathcal{L}_{\mathrm{SS}} &= \mathbb{E}\left(\mathcal{L}_{\mathrm{SS}}^{1}+\mathcal{L}_{\mathrm{SS}}^{2}\right)\\ &= -\mathbb{E}\left(M_{\mathrm{S}}^{1}\log(\mathrm{softmax}(\hat{M}_{\mathrm{S}}^{1}))+M_{\mathrm{S}}^{2}\log(\mathrm{softmax}(\hat{M}_{\mathrm{S}}^{2}))\right),\end{aligned} \tag{2}$$

where M_{S}^{t} and $\hat{M}_{\mathrm{S}}^{t}$ denote the predicted and target semantic map at time *t*, respectively.

The binary change detection loss $\mathcal{L}_{\mathrm{BC}}$ can be calculated by the binary cross-entropy with the predicted and target binary change map M_{BC} and $\hat{M}_{\mathrm{BC}}$ as follows:

$$\mathcal{L}_{\mathrm{BC}} = M_{\mathrm{BC}}\log(1-\hat{M}_{\mathrm{BC}})+(1-M_{\mathrm{BC}})\log(\hat{M}_{\mathrm{BC}}). \tag{3}$$

The semantic consistency loss $\mathcal{L}_{\mathrm{SC}}$ can be estimated with the M_{S}^{1}, M_{S}^{2} and M_{BC}, i.e.,

$$\mathcal{L}_{\mathrm{SC}} = \begin{cases} 1-\cos(M_{\mathrm{S}}^{1},M_{\mathrm{S}}^{2}), & M_{\mathrm{BC}}=0 \\ \max(0,\cos(M_{\mathrm{S}}^{1},M_{\mathrm{S}}^{2})), & M_{\mathrm{BC}}=1 \end{cases}. \tag{4}$$

The $\mathcal{L}_{\mathrm{SC}}$ synergistically increases the accuracy of LSS and BCD by ensuring that the bitemporal semantic predictions in unchanged regions are as consistent as possible, while minimizing the similarity of bitemporal predictions in changed regions.

Additionally, the semantic change detection loss $\mathcal{L}_{\mathrm{SCD}}$ is derived by the predicted semantic change map M_{SC} and target $\hat{M}_{\mathrm{SC}}$, i.e.,

$$\mathcal{L}_{\mathrm{SCD}} = \mathbb{E}(l_{\mathrm{SCD}}) \tag{5}$$

where

$$l_{\mathrm{SCD}} = \begin{cases} 0.5(M_{\mathrm{SC}}-\hat{M}_{\mathrm{SC}})^{2}/\gamma, & \text{for } |M_{\mathrm{SC}}-\hat{M}_{\mathrm{SC}}|<\gamma \\ |M_{\mathrm{SC}}-\hat{M}_{\mathrm{SC}}|-0.5\gamma, & \text{otherwise} \end{cases}, \tag{6}$$

and γ is a balance coefficient that ranges in the interval (0, +∞).

Finally, we obtain the total loss $\mathcal{L}_{\mathrm{T}}$ as follows:

$$\mathcal{L}_{\mathrm{T}} = \mathcal{L}_{\mathrm{SS}}+\mathcal{L}_{\mathrm{BC}}+\mathcal{L}_{\mathrm{SC}}+\mathcal{L}_{\mathrm{SCD}}. \tag{7}$$

These four loss functions provide comprehensive supervision signals, and their collaboration fosters a closer relationship between semantic and difference features. Consequently, the model can achieve accurate results in all LSS, BCD and SCD outputs.

IV. Experiments and Results

In this section, we conducted experiments on both multimodal and unimodal datasets to evaluate the performance of the proposed MSCD-Net. All experiments were performed

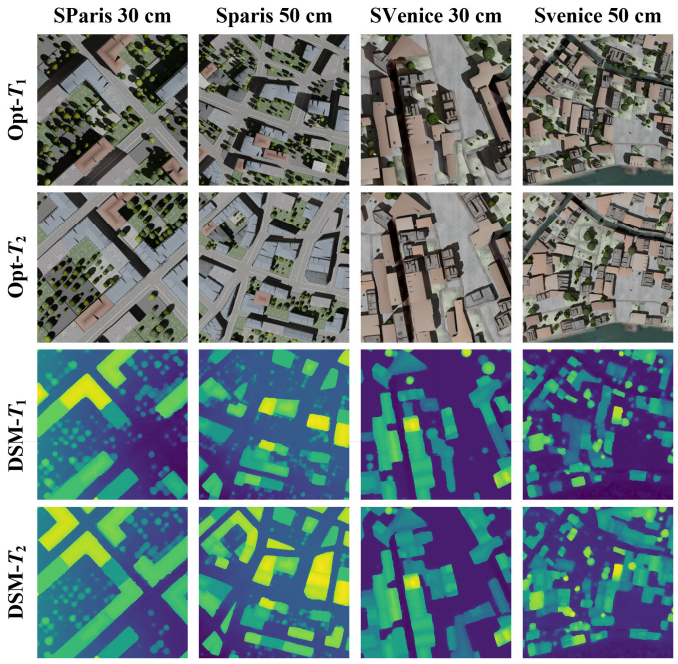


Fig. 6. Visualization of representative input data on four scenes.

on a single PC equipped with an Intel Core i9-10850K CPU operating at a clock rate of 3.6 GHz and two NVIDIA GeForce RTX 3090 GPUs.

A. Description of Datasets

1) SMARS-SCD: The SMARS-SCD multimodal dataset contains pairs of scenes with urban changes, and it was built based on the SMARS dataset [73]. The SMARS dataset is simulated based on the topographies of two European cities, Paris and Venice, and includes two pairs of scenes named SParis and SVenice, respectively. These scenes come with associated orthoimages and DSMs. The dataset features two different spatial resolutions of 30 cm and 50 cm, in which urban land cover is classified into five categories: building, streets, trees, lawns and others. The sizes of both the SParis and SVenice rasters with 30 cm spatial resolution are 5600 × 5600 pixels, and their sizes with 50 cm resolution are 4500 × 3560 pixels and 5600 × 5600 pixels, respectively.

The labels for the BCD and SCD maps were derived by calculating the difference between the bitemporal labels of the land semantic maps. The division methods and parameters for training, validation and test sets in all scenes are consistent with those recommended by the original SMARS dataset. The window size is 512 × 512 and the stride is 256 × 256. The SMARS-SCD dataset includes rasters in GeoTIFF format for all maps and contains six types of files: optical images, DSM, semantic label, semantic change area (SCA) label, BCD label and SCD label. The optical images are rendered in 24-bit RGB format, the DSMs are stored with float precision, and reference labels as discrete integers.

2) SECOND: The semantic change detection (SECOND) dataset [27] employs semantic labeling and targets six distinct object types for annotation: ground, trees, low vegetation, water, buildings and playgrounds. This dataset sources its optical imagery from multiple platforms and sensors, featuring 4,662 pairs of aerial images with spatial resolutions ranging from 0.5 to 3 m. All images have a size of 512 × 512 pixels. The dataset was divided randomly into training, validation and test sets in a ratio of 7:1:2.

3) *OSCD-S1S2*: The OSCD-S1S2 dataset contains Sentinel-1 SAR and Sentinel-2 multispectral data collected from 24 cities around the world from 2015 to 2018. The SAR data include VV and VH polarization with a spatial resolution of 10 m. The multispectral data include 13 bands with the multiple spatial resolutions of Sentinel-2. The labels indicate binary change.

TABLE I
EVALUATION METRICS FOR THE LSS, BCD AND SCD OUTPUTS

Task	Metric	Equation
LSS	mIoU	$\mathrm{mIoU}=\mathbb{E}(\mathrm{IoU})$ $\mathrm{IoU}=\mathrm{TP}/(\mathrm{TP}+\mathrm{FP}+\mathrm{TN})$
	Kappa	$\mathrm{Kappa}=(\mathrm{OA}-\mathrm{PE})/(1-\mathrm{PE})$ $\mathrm{OA}=\frac{\mathrm{TP}+\mathrm{TN}}{\mathrm{TP}+\mathrm{TN}+\mathrm{FP}+\mathrm{FN}}$ $\mathrm{PE}=\frac{(\mathrm{TP}+\mathrm{FN})(\mathrm{TP}+\mathrm{FP})+(\mathrm{TN}+\mathrm{FP})(\mathrm{TN}+\mathrm{FN})}{(\mathrm{TP}+\mathrm{TN}+\mathrm{FN}+\mathrm{FP})^2}$
BCD	cIoU	The equation of cIoU is the same as the IoU, but only for the change class.
	bmIoU	$\mathrm{bmIoU}=(\mathrm{IoU}_{\mathrm{uc}}+\mathrm{IoU}_{\mathrm{c}})/2$ $\mathrm{IoU}_{\mathrm{uc}}=q_{00}/\left(\sum_{i=0}^{N}q_{i0}+\sum_{j=0}^{N}q_{0j}-q_{00}\right)$ $\mathrm{IoU}_{\mathrm{c}}=\sum_{i=0}^{N}\sum_{j=0}^{N}q_{ij}/\left(\sum_{i=0}^{N}\sum_{j=0}^{N}q_{ij}-q_{00}\right)$
	KC	The equation same as the Kappa, but the calculated classes of Kappa is N whereas that of KC is 2.
	F1	$\mathrm{F1}=\frac{2\times\mathrm{Pre}\times\mathrm{Rec}}{\mathrm{Pre}+\mathrm{Rec}}$ $\mathrm{Pre}=\frac{\mathrm{TP}}{\mathrm{TP}+\mathrm{FP}}$ $\mathrm{Rec}=\frac{\mathrm{TP}}{\mathrm{TP}+\mathrm{FN}}$
SCD	SeK	$\mathrm{SeK}=e^{(IoU_c-1)}(\rho-\eta)/(1-\eta)$ $\rho=\sum_{i=0}^{N}\hat{q}_{ii}/\left(\sum_{i=0}^{N}\sum_{j=0}^{N}\hat{q}_{ij}\right)$ $\eta=\sum_{i=0}^{N}\left(\sum_{j=0}^{N}\hat{q}_{ij}\cdot\sum_{j=0}^{N}\hat{q}_{ji}\right)/\left(\sum_{i=0}^{N}\sum_{j=0}^{N}\hat{q}_{ij}\right)^2$
	Fscd	$\mathrm{Fscd}=2P_{\mathrm{scd}}R_{\mathrm{scd}}/(P_{\mathrm{scd}}+R_{\mathrm{scd}})$ $P_{\mathrm{scd}}=\sum_{i=1}^{N}q_{ii}/\sum_{i=1}^{N}\sum_{j=0}^{N}q_{ij}$ $R_{\mathrm{scd}}=\sum_{i=1}^{N}q_{ii}/\sum_{i=0}^{N}\sum_{j=1}^{N}q_{ij}$

B. Comparison Methods

We selected leading existing unimodal SCD and BCD methods for comparison. For unimodal SCD methods, we concatenated the multimodal data along the channel dimension before inputting them into these methods. For unimodal BCD methods, we not only performed the same concatenation operation, but also replaced all the loss functions with those used in MSCD-Net to adapt them to the multimodal SCD task. The selected unimodal SCD methods contain SAAN [30], CLAFA [32], ChangeMask [31], HRSCD-str4 and -str3 [26], and ResNet-LSTM [5]. The selected unimodal BCD methods

include FC-Siam-conv and FC-EF [74]. On the SECOND dataset, more advanced methods—including MTSCD [75], SMNet [76] and EGMS-Net [77]—were adopted to ensure a fairer and more comprehensive comparison.

On the OSCD-S1S2 dataset, we selected the unimodal BCD methods of STADE [78], DKSSCD-EF [45] and SiamUnet-diff-EF [74]. Additionally, the multimodal BCD methods including SiamU-conc [79], HFA-PANet [80] and Semi-MCD [81] were also evaluated.

C. Implementation Details and Evaluation Metrics

1) Implementation details: We constructed the encoders using the classic ResNet-34 model to balance efficiency and accuracy. For both the SMARS-SCD and SECOND datasets, we used a small batch size of 2 for training and 4 for testing due to the large input image size (512 × 512 pixels). Our data augmentation strategy included random rotations and flipping while loading image pairs. Performance was measured on all datasets by calculating evaluation metrics on the test set. Moreover, the balance coefficient γ in $\mathcal{L}_{\mathrm{SCD}}$ was set to 1. We employed an optimization algorithm based on stochastic gradient descent to train MSCD-Net. The learning rate, initially set at 0.1, was decayed by 10% every 2 epochs. The weight decay and momentum coefficients were set to 5e-4 and 0.9, respectively.

2) Evaluation Metrics: In the experiments, we not only compared the accuracy of the final SCD results, but also collected the accuracy of the intermediate LSS and BCD outputs (M_{S}^{t} and M_{BC}). Specifically, two metrics, i.e., mean intersection over union (mIoU) and Cohen's Kappa coefficient for LSS (Kappa), were used to evaluate the performance of LSS quantitatively. Four metrics, i.e., IoU for change class (cIoU), balanced mIoU (bmIoU), Cohen's Kappa coefficient for CD (KC) and F1 score (F1), were used for the BCD. For SCD, we used two comprehensive quality metrics, i.e., separated kappa (SeK) and F1 score for SCD (Fscd), to measure the performance quantitatively.

Let $Q=\{q_{ij}\}$ be the confusion matrix, where q_{ij} denotes the number of pixels predicted as class i while the true label is class j $(i,j)\in\{0,1,\cdots,N\}$ (0 indicates unchanged). To exclude the true positive no-change pixels, whose number is dominant, we let $\hat{q}_{ij}=q_{ij}$ but without q_{00}. Then, we can derive all the above metrics, with the equations presented in Table I. A larger value means better performance for all metrics.

D. Results and Analysis

1) SMARS-SCD dataset: The metric statistics of all methods on the SMARS-SCD dataset are listed in Tables II to V. Due to space limitations, we present the visualized outputs of representative methods on four input pairs in Fig. 6. The visualized outputs are shown in Figs. 7 to 10. From Tables II to V and Figs. 7 to 10, we obtain the following insights:

1) The proposed MSCD-Net shows optimal performance across all scenes, outperforming other methods in SCD, BCD and LSS outputs. Overall, small or linear objects are more accurately identified, the boundaries of changed areas are more precise, and semantic changes are identified with improved accuracy.

TABLE II
METRIC STATISTICS ON THE SPARIS 30 CM OF THE SMARS-SCD DATASET

Method	LSS		BCD				SCD	
	mIoU	Kappa	cIoU	bmIoU	KC	F1	SeK	Fscd
MSCD-Net	**88.53**	**93.25**	**94.63**	**96.72**	**96.64**	**97.24**	**82.21**	**93.64**
CLAFA	85.83	91.42	93.13	95.65	95.68	96.44	78.24	92.25
SAAN	78.59	87.01	94.45	96.61	96.52	97.14	76.03	89.28
HRSCD-str4	83.24	89.84	93.54	96.04	95.93	96.66	76.09	90.12
ChangeMask	75.72	84.42	93.92	96.28	96.18	96.86	76.22	89.94
HRSCD-str3	41.64	50.12	91.96	95.04	94.86	95.80	56.26	77.33
FC-Siam-conv	79.67	87.23	92.16	95.24	95.06	95.93	73.65	89.67
ResNet-LSTM	79.86	87.14	90.96	94.48	94.25	95.27	71.56	89.28
FC-EF	71.75	80.55	87.26	92.23	91.68	93.20	62.79	85.58

TABLE III
METRIC STATISTICS ON THE SPARIS 50 CM OF THE SMARS-SCD DATASET. * SIGNIFIES USING THEIR ORIGINAL LOSS FUNCTIONS

Method	LSS		BCD				SCD	
	mIoU	Kappa	cIoU	bmIoU	KC	F1	SeK	Fscd
MSCD-Net	**82.44**	**89.77**	**91.92**	**95.01**	**94.83**	**95.79**	**74.06**	**90.19**
CLAFA	77.76	85.95	90.45	94.07	93.81	94.94	69.12	87.64
SAAN	75.55	84.26	89.33	93.40	93.08	94.36	67.64	87.57
HRSCD-str4	73.20	83.41	91.28	94.56	94.31	95.35	66.86	85.01
ChangeMask	71.33	81.99	90.53	94.11	93.86	95.03	66.18	85.15
HRSCD-str3	61.07	72.28	81.85	88.49	87.35	90.14	51.41	78.09
ResNet-LSTM	68.75	81.14	86.14	91.36	90.81	92.55	57.05	81.33
FC-Siam-conv	73.06	83.89	87.08	91.92	91.48	93.13	61.03	83.96
FC-EF	61.81	74.15	84.35	90.25	89.58	91.51	52.60	79.13
FC-Siam-conv*			86.39	91.57	91.04	92.69		
FC-EF*			81.98	88.85	87.94	90.23		

TABLE IV
METRIC STATISTICS ON THE SVENICE 30 CM OF THE SMARS-SCD DATASET. * SIGNIFIES USING THEIR ORIGINAL LOSS FUNCTIONS

Method	LSS		BCD				SCD	
	mIoU	Kappa	cIoU	bmIoU	KC	F1	SeK	Fscd
MSCD-Net	**86.97**	**92.96**	**93.28**	**95.43**	**95.3**	**96.52**	**79.48**	**92.93**
CLAFA	81.76	89.78	91.96	94.48	94.28	95.81	74.92	90.66
SAAN	79.73	87.97	92.42	94.79	94.63	96.06	73.74	89.28
HRSCD-str4	79.34	88.27	91.34	93.99	93.77	95.47	72.76	89.31
ChangeMask	81.68	89.16	91.92	94.45	94.28	95.8	73.16	89.28
HRSCD-str3	63.33	77.07	82.50	87.81	86.18	90.41	53.26	80.81
ResNet-LSTM	68.42	82.05	85.02	89.85	89.34	91.90	61.24	85.52
FC-Siam-conv	79.86	88.35	88.90	92.47	92.10	94.12	68.91	88.89
FC-EF	60.08	75.42	85.43	89.83	89.18	92.14	55.61	79.57
FC-Siam-conv*			89.04	92.53	92.17	94.20		
FC-EF*			80.85	86.88	85.69	89.28		

TABLE V
METRIC STATISTICS ON THE SVENICE 50 CM OF THE SMARS-SCD DATASET

Method	LSS		BCD				SCD	
	mIoU	Kappa	cIoU	bmIoU	KC	F1	SeK	Fscd
MSCD-Net	**72.06**	**84.51**	**90.78**	**93.94**	**93.69**	**95.16**	**67.25**	**85.55**
CLAFA	69.58	84.36	87.09	91.64	91.16	93.10	63.59	86.44
SAAN	71.18	84.06	87.33	91.48	90.98	93.17	62.24	84.88
HRSCD-str4	67.08	81.22	87.09	91.67	91.02	93.11	62.07	84.83
ChangeMask	64.81	81.80	86.95	91.49	90.96	93.02	61.68	84.96
HRSCD-str3	42.96	61.36	71.81	80.92	78.32	83.59	28.76	62.30
ResNet-LSTM	56.99	74.90	79.73	86.47	84.98	88.72	46.53	77.75
FC-Siam-conv	60.43	78.36	83.77	89.32	88.52	91.17	53.74	80.46
FC-EF	46.38	63.31	71.42	80.52	77.88	83.33	33.85	69.69

2) For almost all methods, the accuracy at 30 cm spatial resolution is greater than that at 50 cm resolution within the same city of the SMARS-SCD dataset. The primary reason for this is that object edges at 50 cm resolution are more blurred, and there are more small objects

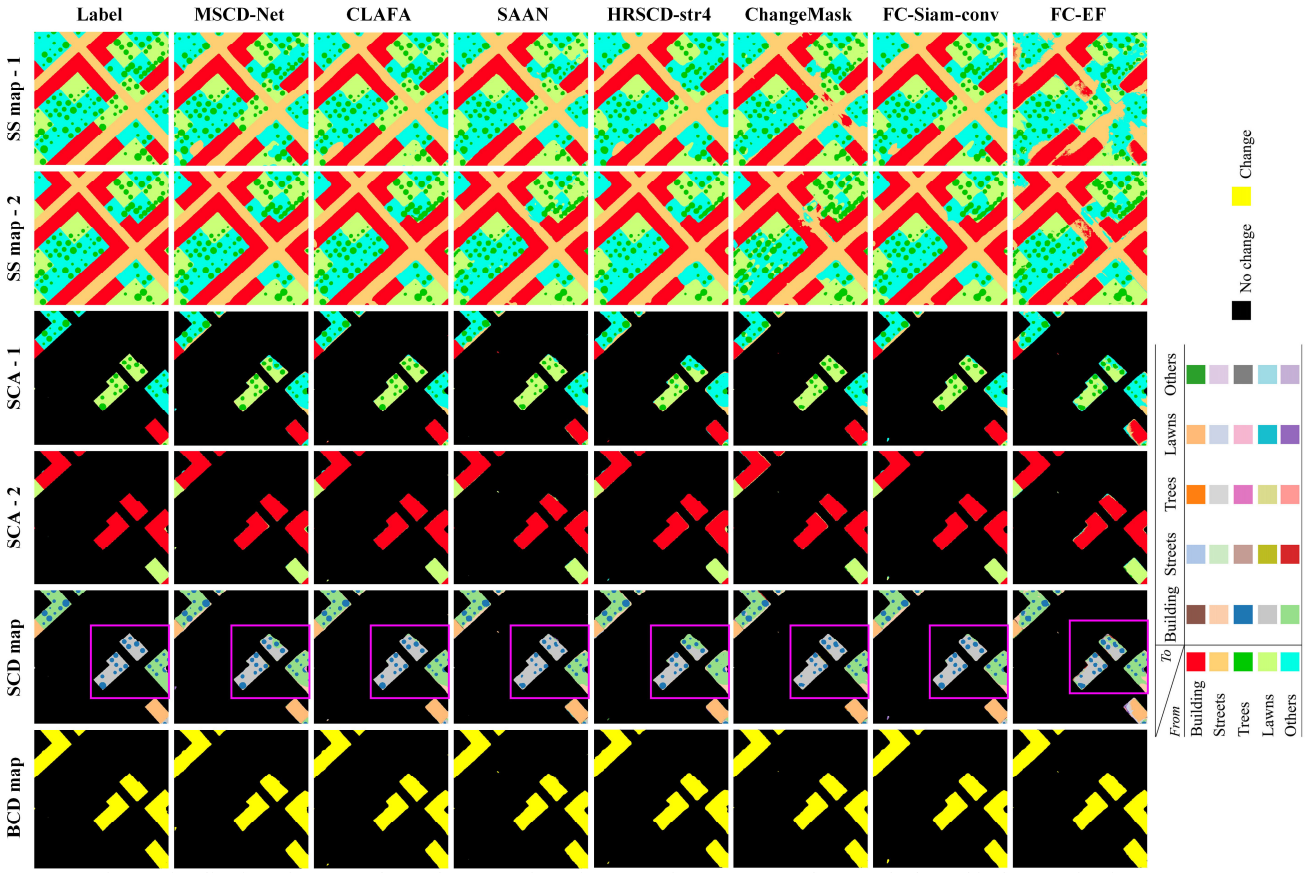


Fig. 7. Visualization of SCD results on the scene of SParis 30 cm of SMARS-SCD dataset. The legend is the same in Figs. 7 to 10.

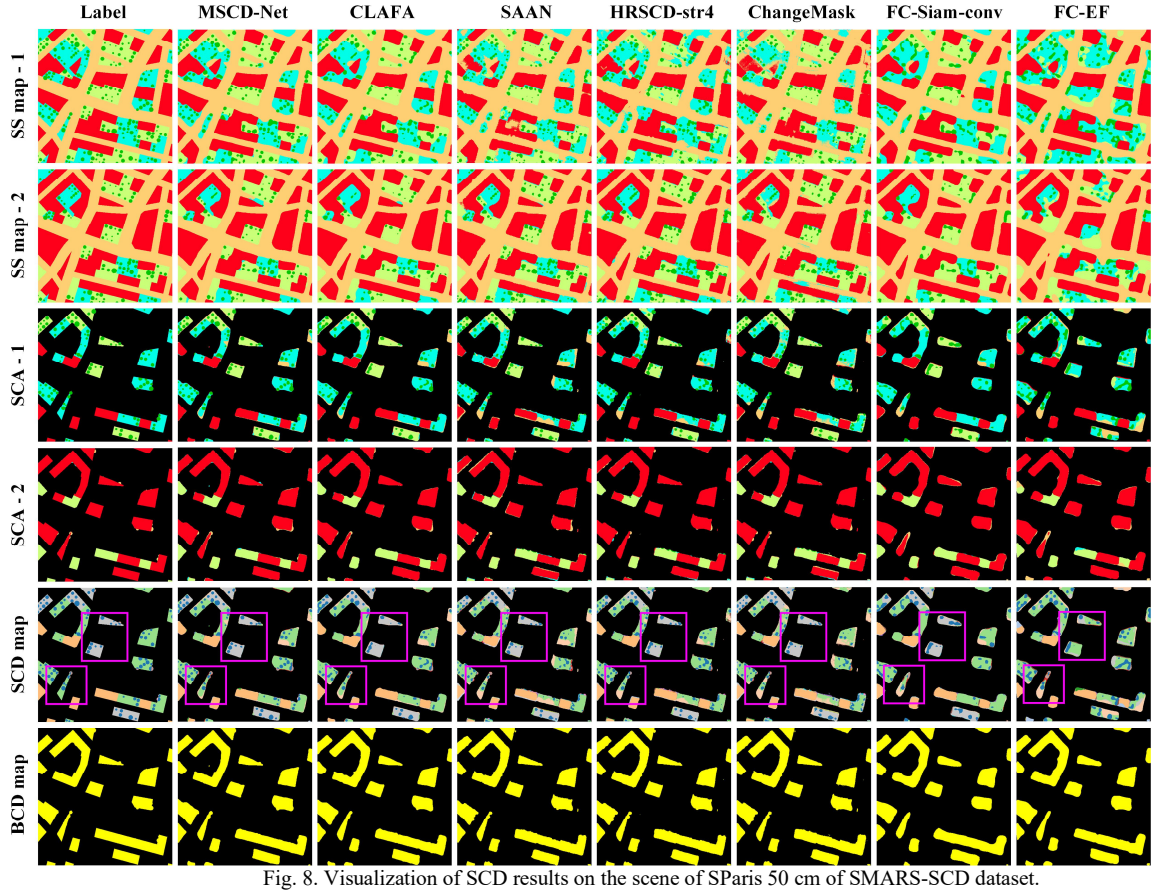


Fig. 8. Visualization of SCD results on the scene of SParis 50 cm of SMARS-SCD dataset.

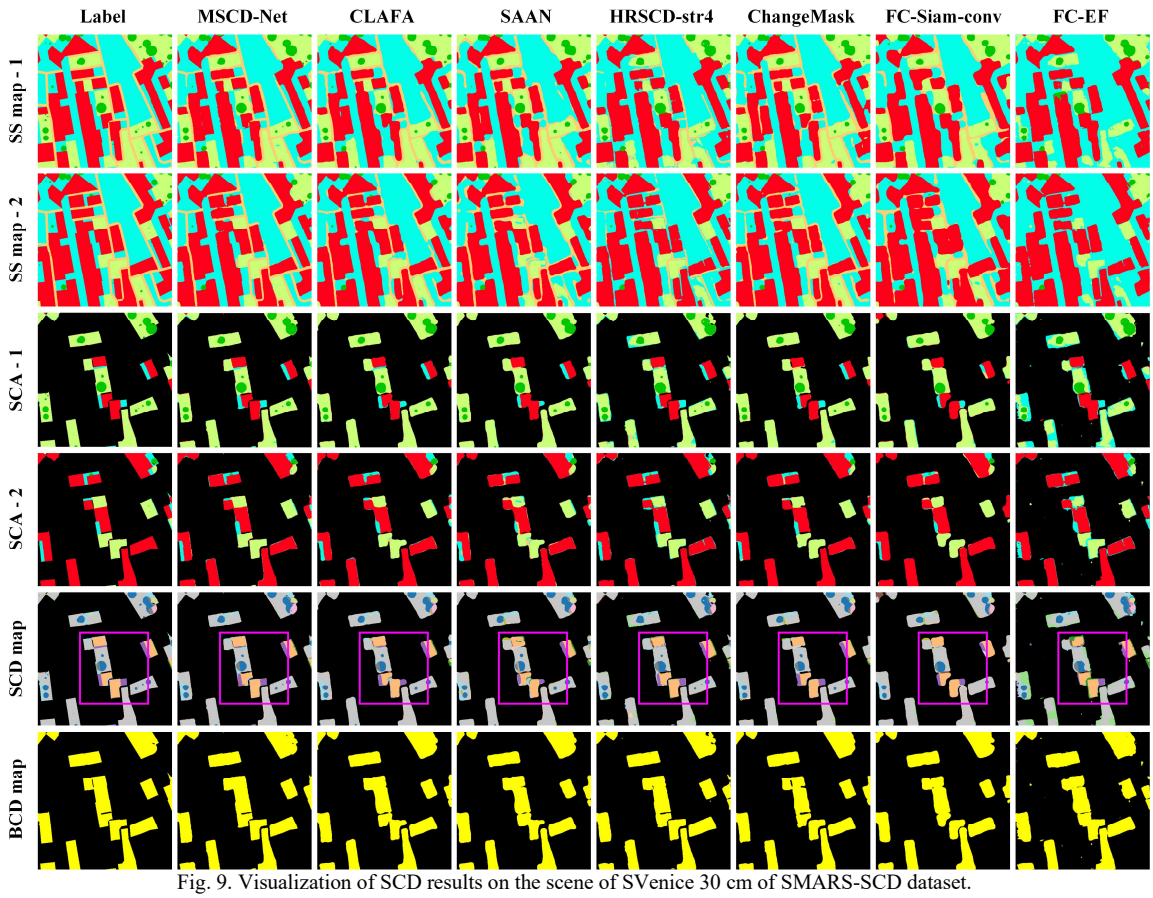


Fig. 9. Visualization of SCD results on the scene of SVenice 30 cm of SMARS-SCD dataset.

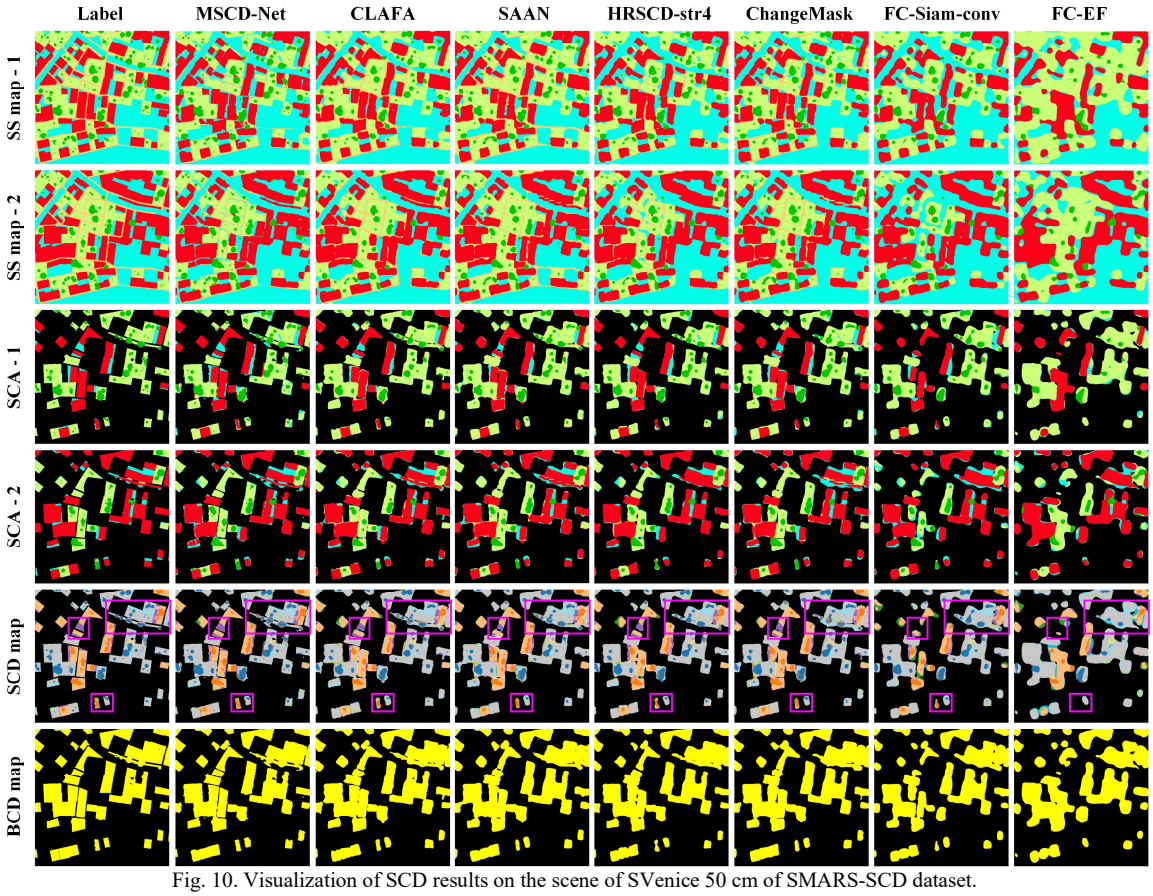


Fig. 10. Visualization of SCD results on the scene of SVenice 50 cm of SMARS-SCD dataset.

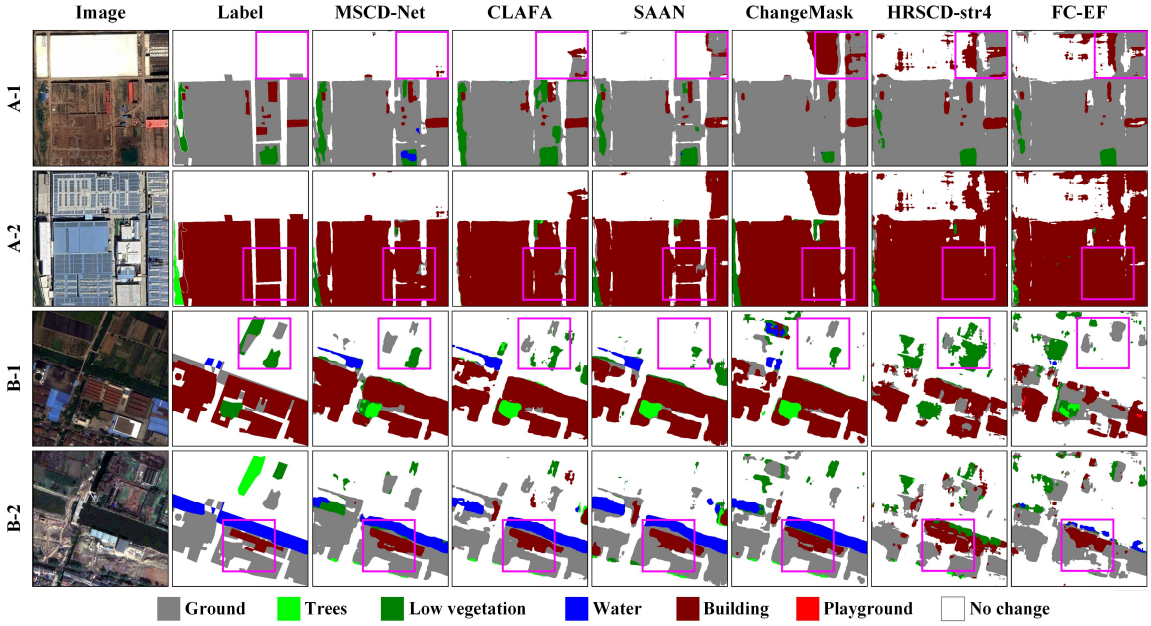


Fig. 11. Visualization of SCD results on the SECOND dataset across representative methods.

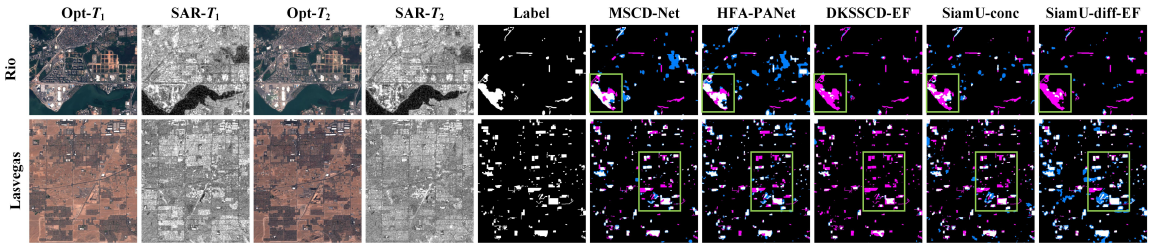


Fig. 12. Visualization of BCD results on the OSCD-S1S2 dataset across representative methods. White: true positives (TP); black: true negatives (TN); azure: false positives (FP); magenta: false negatives(FN).

TABLE VI
METRIC STATISTICS ON THE SECOND DATASET

Method	SeK	Fscd	bmIoU
MSCD-Net	**22.49**	**61.94**	**72.71**
EGMS-Net	21.51	60.65	72.10
MTSCD	20.57	60.55	71.68
CLAFA	21.55	60.72	72.16
SMNet	20.29	60.34	71.95
SAAN	18.03	57.91	70.48
HRSCD-str4	16.20	55.66	69.80
ChangeMask	17.89	57.60	70.41
HRSCD-str3	7.73	45.83	62.29
FC-Siam-conv	15.84	54.75	69.32
ResNet-LSTM	13.25	52.66	68.04
FC-EF	9.98	47.77	64.25

TABLE VII
BCD RESULTS ON THE OSCD-S1S2 DATASET

Method	cIoU	KC	F1
MSCD-Net	**32.66**	**46.17**	**49.23**
HFA-PANet	28.16	41.26	43.91
STADE	19.10	27.65	32.07
DKSSCD-EF	17.84	23.74	30.28
Semi-MCD	26.32	38.96	41.69
SiamU-conc	30.1	42.97	46.27
SiamUnet-diff-EF	22.25	34.01	36.34

(especially trees). In particular, the surface coverage of SVenice 50 cm is more crowded and complex, with more shadow coverage areas, resulting in the lowest accuracy for SVenice 50 cm.

3) The intermediate output of BCD is more accurate than that of LSS for most methods, indicating that the binary classification task is easier than the multi-class classification task on the SMARS-SCD dataset. Most of the compared methods achieve high segmentation and CD accuracy for buildings, but the accuracy is less effective for trees, others and lawns due to the presence of shadows and overlaps.
4) Because SMARS-SCD is a simulated dataset, it maintains high data quality and minor intra-class variance despite incorporating various illumination conditions and effects. Consequently, most methods achieve high accuracy, and the performance differences in SCD among the unimodal methods are minimal. Additionally, because most of the change areas in the SMARS-SCD dataset are related to buildings and the surface objects are relatively regular, there is little difference in BCD accuracy among the various methods.
5) The accuracy of unimodal BCD methods (FC-Siam-conc and FC-EF) was evaluated on the SParis 50 cm and SVenice 30 cm datasets using their original loss functions, without substitution with those employed in MSCD-Net. The results reveal that MSCD-Net's loss

TABLE VIII
INPUT DATA IN DIFFERENT MODALITIES AND THEIR COMBINATION

Input	SVenice 30 cm				SParis 50 cm			
	SeK	Fscd	mIoU	F1	SeK	Fscd	mIoU	F1
Opt	70.86	89.64	78.84	94.70	69.13	86.96	73.85	95.46
DSM	51.27	79.33	59.79	89.96	49.64	74.87	58.48	92.25
Opt & DSM	**79.48**	**92.93**	**86.97**	**96.52**	**74.06**	**90.19**	**82.44**	**95.79**

TABLE IX
THE PERFORMANCE OF DIFFERENT FUSION STRATEGIES IN MFB

Fusion strategy	SVenice 30 cm				SParis 50 cm			
	SeK	Fscd	mIoU	F1	SeK	Fscd	mIoU	F1
Siam-EF	73.83	89.41	71.21	96.01	68.16	87.82	77.15	94.48
Without CAB	69.98	88.85	77.64	94.64	65.21	86.17	73.24	94.06
Conv block rep. CAB	71.39	89.64	74.72	95.03	67.66	87.46	76.01	94.43
Without conv block	72.09	89.43	71.40	95.44	68.89	87.22	76.09	95.13
Vanilla cross-attention	75.58	91.49	84.57	95.69	69.46	87.14	74.21	95.48
Fusion layers-1	68.13	87.72	74.22	94.35	66.44	86.39	74.15	94.47
Fusion layers-2	75.52	90.34	81.51	96.24	70.07	88.07	82.37	95.15
Fusion layers-3	**79.48**	**92.93**	**86.97**	**96.52**	**74.06**	**90.19**	**82.44**	**95.79**
Fusion layers -4	77.45	92.36	85.12	96.05	72.07	88.98	79.38	95.58

TABLE X
THE PERFORMANCE OF DIFFERENT STRUCTURES IN THE SDM

Structure	SVenice 30 cm				SParis 50 cm			
	SeK	Fscd	mIoU	F1	SeK	Fscd	mIoU	F1
Only the features in scale 0	76.84	91.38	83.75	96.24	70.38	87.56	76.53	95.63
Contains the features in scale 0 and 1	78.56	92.50	84.63	96.41	72.74	89.3	80.59	95.68
Contains the features from scale 0 to 2	**79.48**	**92.93**	**86.97**	**96.52**	**74.06**	**90.19**	**82.44**	**95.79**
Without difference in merge block	73.23	89.29	81.93	95.81	66.76	87.25	75.72	94.13
Without concatenation in merge block	77.21	92.06	85.29	96.07	71.49	89.19	80.94	95.23
Conv block rep. SAA	74.64	91.09	79.01	95.56	67.88	87.9	80.14	94.28
SDD using vanilla attention	76.49	91.27	82.88	96.17	70.15	88.77	79.81	94.84
Without window context interaction	78.10	91.57	83.02	96.36	72.24	89.99	81.88	95.12
Shortcut without LP	78.53	91.72	86.36	96.71	72.87	89.48	80.16	95.66

function achieves superior accuracy compared to the original loss functions, while simultaneously providing LSS and SCD outputs.

2) SECOND dataset: We performed experiments on the SECOND dataset to evaluate the generalization performance of the MSCD-Net on the unimodal SCD dataset. The results are shown in Fig. 11, and accuracy metrics are listed in Table VI. From these results, it is evident that the MSCD-Net achieves the highest accuracy. MSCD-Net is more precise in dividing the boundaries of different classes and excels in identifying details, especially in scenes with complex spatial coverage. Owing to the insufficient exploration of long-range associations and global context, CLAFA and EGMS-Net still lag behind MSCD-Net.

3) OSCD-S1S2 dataset: The results on OSCD-S1S2 dataset are shown in Fig. 12, and metric statistics are listed in Table VII. It can be seen that the *cIoU*, *KC* and *F1* of MSCD-Net exceed the optimal comparison method by 2.88, 3.13 and 3.37, respectively, indicating that our method is the most accurate on the OSCD-S1S2 dataset. This result also shows that MSCD-Net has outstanding performance across different modalities and maintains proficiency in BCD tasks.

TABLE XI
THE IMPACT OF DIFFERENT LOSS FUNCTIONS ON PERFORMANCE

Loss function	SVenice 30 cm				SParis 50 cm			
	SeK	Fscd	mIoU	F1	SeK	Fscd	mIoU	F1
$\mathcal{L}_{SS}$ & $\mathcal{L}_{BCD}$	76.18	90.92	84.45	96.20	69.82	87.02	76.11	95.66
+ $\mathcal{L}_{SC}$	78.69	91.81	85.90	96.75	72.94	89.29	80.33	95.78
+ $\mathcal{L}_{SCD}$	78.25	91.51	84.01	96.71	72.27	89.10	81.09	95.61
Four Losses	**79.48**	**92.93**	**86.97**	**96.52**	**74.06**	**90.19**	**82.44**	**95.79**

E. Ablation Study

To explore the role and working mechanism of important modules within MSCD-Net, we conducted a variety of ablation studies. We present typical results for the scenes of SVenice 30 cm and SParis 50 cm from the SMARS-SCD dataset.

1) Input data: First, we studied the impact of different input modalities and multimodal combinations on model performance. The results are shown in Table VIII. Note that, for inputs of either Opt or DSM, the cross-attention block in MFB is replaced by a self-attention block, and MFB operates without concatenation at the end, outputting bitemporal features. We observed that the SeKs of using only DSM are 51.27 and 49.64 on SVenice 30 cm and SParis 50 cm, respectively; while the SeKs of using only optical data are 70.86 and 69.13 on corresponding scenes, showing suboptimal accuracy. By contrast, the combination of optical data with DSM yields superior accuracy, with all indicators reaching their highest values and the SeKs showing a significant improvement. This finding underscores the effectiveness of multimodal data fusion for common Earth observation tasks.

2) Fusion strategy in MFB: Next, we evaluated the performance of different fusion strategies in the MFB module; the results are shown in Table IX. Note that the Siam-EF means that after the multimodal data from the same temporal phase are concatenated, they are inputted into a Siamese encoder; the subsequent structure remains consistent with the unimodal input. “Conv block rep. CAB” means that the CAB is replaced by the conv block. From Table IX, we can derive the following insights:

1) Among all fusion strategies, those employing vanilla cross-attention or Siam-EF demonstrate intermediate accuracy, performing neither at the lowest nor highest levels observed. This indicates that these intuitive strategies can also achieve competitive performance. Compared with the conv block, CAB outputs higher accuracy.
2) The number of fusion layers positively correlates with performance up to the fourth layer. The reason is that a large gap exists among the multimodal original data or shallow features, making it challenging to integrate them into high-quality features. However, when the number of fusion layers reaches 4, the subsequent MSD and SDM modules need to start from the feature in scale 1. This results in only 1 and 2 sub-blocks in the MSD

TABLE XII
THE PERFORMANCE OF REPRESENTATIVE METHODS WITH SDD

Method	SVenice 30 cm				SParis 50 cm				SECOND		
	SeK	Fscd	mIoU	F1	SeK	Fscd	mIoU	F1	SeK	Fscd	bmIoU
CLAFA	74.92	90.66	81.76	95.81	69.12	87.64	77.76	94.94	21.55	60.72	72.16
CLAFA + SDD	76.85	91.36	83.67	96.25	71.17	88.98	80.37	95.2	22.14	61.33	72.58
HRSCD-str4	72.76	89.31	79.34	95.47	66.86	85.01	73.2	95.35	16.2	55.66	69.80
HRSCD-str4 + SDD	75.93	91.16	82.73	96.02	70.59	88.44	79.25	95.21	18.03	57.51	70.35
ChangeMask	73.16	89.28	81.68	95.8	66.18	85.15	71.33	95.03	17.89	57.60	70.41
ChangeMask + SDD	74.44	89.94	81.89	96.18	68.41	86.48	77.16	95.29	18.29	57.88	70.44

and SDM, respectively, leading to insufficient capacity to refine semantic and difference features and thereby reducing overall accuracy.

3) Structure of SDM: The impact of different structures of the SDM module on performance is shown in Table X. From this table, one can observe that aggregating more scales of semantic and difference features is more beneficial for enhancing performance. In multimodal feature merging, the difference operation contributes more to the final SCD accuracy than concatenation. The convolutional block does not increase accuracy as much as the proposed SAA, which outperforms vanilla attention. Additionally, window context interaction and shortcuts with LP in the SDD slightly improve performance.

4) Loss function: We studied the impact of different loss functions on performance; the obtained statistical metrics are listed in Table XI. The results show that both $\mathcal{L}_{SC}$ and $\mathcal{L}_{SCD}$ can increase overall performance when added to the two basic losses of the SCD task, $\mathcal{L}_{SS}$ and $\mathcal{L}_{BCD}$. The collaboration of all four losses results in the highest accuracy.

5) Feature learning ability and compatibility of SDD: Finally, we transferred the SDD module to other methods to evaluate its intrinsic feature learning ability and compatibility. We replaced the original CD decoder of these target methods with the SDD. The accuracy metrics on the SVenice 30 cm, SParis 50 cm, and SECOND datasets are shown in Table XII. We observe that after incorporating the SDD, the SeKs of CLAFA, HRSCD-str4 and ChangeMask increased by at least 1.93, 3.17 and 1.28, respectively, on the SVenice 30 cm and SParis 50 cm datasets. On the SECOND dataset, the SeKs of these methods increased by 0.59, 1.83 and 0.4, respectively. The HRSCD-str4 method, which is entirely based on the UNet network with a simple structure and feature extraction process, showed the most significant accuracy increase after using SDD. The results demonstrate that SDD possesses strong intrinsic feature learning ability and compatibility. It can leverage long-range associations and global information, thoroughly explore the dependence between semantic features and difference features, and effectively improve the quality of semantic difference features and SCD results across multiple methods.

6) Sensitivity of Hyperparameters: We conducted sensitivity analysis of batch size and balance coefficient γ, as illustrated in Fig. 13. The experimental results reveal a positive correlation between SeK and batch size, with this relationship being pronounced on the lower-resolution SParis 50 cm dataset. Interestingly, even-numbered batch sizes consistently demonstrate higher accuracy compared to their adjacent odd-numbered counterparts. Due to hardware memory constraints, our evaluation was limited to batch sizes $\leqslant 5$. Regarding the balance coefficient, our analysis indicates minimal accuracy variation within the range of (0,10), but a significant performance degradation occurs when balance coefficient exceeds 10. Based on these findings, we recommend implementing the balance coefficient within the range of [0.25, 8] and selecting the maximum feasible batch size according to available hardware memory capacity.

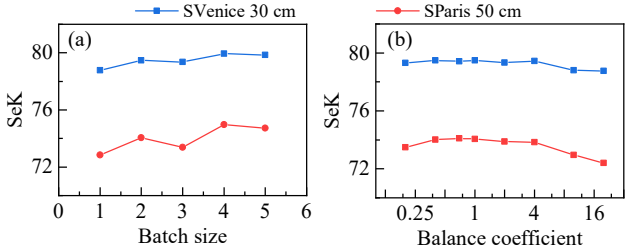


Fig. 13. Sensitivity of (a) batch size and (b) balance coefficient (the horizontal axis uses log2 scale).

V. CONCLUSION

SCD is an important, but challenging task in Earth observation. With the proliferation of sensor types, platforms and data volumes, as well as the continuous expansion of application scenarios, the demand for multimodal SCD methods is increasingly urgent. Thus, we proposed the first DL-based multimodal SCD method, named MSCD-Net. The MSCD-Net contains two Siamese encoders to extract multi-scale features from multimodal data. Two MFBs are built to fuse the multimodal features. The MSD module and LSS predictor are responsible for distilling high-quality multi-scale land semantic features and predicting the semantic map, respectively. A SDM module is designed to aggregate and refine multi-scale semantic difference features, while a SDD module jointly models semantic and difference features. Moreover, a loss function incorporating four distinct contributions is developed to provide extensive guidance. All modules collaborate to distill, fuse and enhance various features embedded in the multimodal data, guiding the model to learn delicate SCD patterns.

Experimental results confirm that the MSCD-Net achieves highest accuracy on both multimodal and unimodal SCD datasets. The SDD has strong feature learning ability and compatibility, can be used in multiple existing methods and significantly increases accuracy. These findings reveal the potential of MSCD-Net to advance the development and unification of SCD methods.

This research serves as a pioneering effort in multimodal SCD. However, constrained by the scarcity of available datasets, the multimodal data types employed in this work lack diversity. In the future, we will create a multimodal dataset incorporating authentic point clouds, optical, SAR and other modalities, aiming to provide the community with a high-quality benchmark for evaluating remote sensing multimodal tasks.

ACKNOWLEDGMENT

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

REFERENCES

[1] X. Gu, P. P. Angelov, C. Zhang, and P. M. Atkinson, "A Semi-Supervised Deep Rule-Based Approach for Complex Satellite Sensor Image Analysis," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 44, no. 5, pp. 2281–2292, May 2022.

[2] J. Wang *et al.*, "MaCon: A generic self-supervised framework for unsupervised multimodal change detection," *IEEE Trans. Image Process.*, vol. 34, pp. 1485–1500, 2025.

[3] Y. Zheng, S. Liu, H. Chen, and L. Bruzzone, "Hybrid FusionNet: A Hybrid Feature Fusion Framework for Multisource High-Resolution Remote Sensing Image Classification," *IEEE Trans. Geosci. Remote Sens.*, vol. 62, pp. 1–14, 2024.

[4] L. Ding *et al.*, "A survey of sample-efficient deep learning for change detection in remote sensing: Tasks, strategies, and challenges," *IEEE Geosci. Remote Sens. Mag.*, pp. 2–27, 2025.

[5] 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 Trans. Geosci. Remote Sens.*, vol. 57, no. 2, pp. 924–935, Feb. 2019.

[6] W. Zhang *et al.*, "A Novel Knowledge-Driven Automated Solution for High-Resolution Cropland Extraction by Cross-Scale Sample Transfer," *IEEE Trans. Geosci. Remote Sens.*, vol. 61, pp. 1–16, 2023.

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

[8] L. Bruzzone and R. Cossu, "An adaptive approach to reducing registration noise effects in unsupervised change detection," *IEEE Trans. Geosci. Remote Sens.*, vol. 41, no. 11, pp. 2455–2465, Nov. 2003.

[9] 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 Sens. Environ.*, vol. 64, no. 1, pp. 1–19, Apr. 1998.

[10] L. Bruzzone and D. F. Prieto, "Automatic analysis of the difference image for unsupervised change detection," *IEEE Trans. Geosci. Remote Sens.*, vol. 38, no. 3, pp. 1171–1182, May 2000.

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

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

[13] A. A. Nielsen, "The Regularized Iteratively Reweighted MAD Method for Change Detection in Multi- and Hyperspectral Data," *IEEE Trans. Image Process.*, vol. 16, no. 2, pp. 463–478, Feb. 2007.

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

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

[16] Y. Wang *et al.*, "Spectral–Spatial–Temporal Transformers for Hyperspectral Image Change Detection," *IEEE Trans. Geosci. Remote Sens.*, vol. 60, p. 5536814, 2022.

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

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

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

[20] C. Huang *et al.*, "Use of a dark object concept and support vector machines to automate forest cover change analysis," *Remote Sens. Environ.*, vol. 112, no. 3, pp. 970–985, Mar. 2008.

[21] G. Xian and C. Homer, "Updating the 2001 National Land Cover Database Impervious Surface Products to 2006 using Landsat Imagery Change Detection Methods," *Remote Sens. Environ.*, vol. 114, no. 8, pp. 1676–1686, Aug. 2010.

[22] L. Bruzzone and S. B. Serpico, "An iterative technique for the detection of land-cover transitions in multitemporal remote-sensing images," *IEEE Trans. Geosci. Remote Sens.*, vol. 35, no. 4, pp. 858–867, Jul. 1997.

[23] J. S. Deng, K. Wang, Y. H. Deng, and G. J. Qi, "PCA-based land-use change detection and analysis using multitemporal and multisensor satellite data," *Int. J. Remote Sens.*, vol. 29, no. 16, pp. 4823–4838, Aug. 2008.

[24] F. Bovolo, S. Marchesi, and L. Bruzzone, "A Framework for Automatic and Unsupervised Detection of Multiple Changes in Multitemporal Images," *IEEE Trans. Geosci. Remote Sens.*, vol. 50, no. 6, pp. 2196–2212, Jun. 2012.

[25] 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 Trans. Geosci. Remote Sens.*, vol. 53, no. 8, pp. 4363–4378, Aug. 2015.

[26] R. Caye Daudt, B. Le Saux, A. Boulch, and Y. Gousseau, "Multitask learning for large-scale semantic change detection," *Comput. Vis. Image Underst.*, vol. 187, p. 102783, Oct. 2019.

[27] K. Yang *et al.*, "Asymmetric Siamese Networks for Semantic Change Detection in Aerial Images," *IEEE Trans. Geosci. Remote Sens.*, vol. 60, pp. 1–18, 2022.

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

[29] 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 Trans. Geosci. Remote Sens.*, vol. 60, pp. 1–14, 2022.

[30] H. Guo, X. Su, C. Wu, B. Du, and L. Zhang, "SAAN: Similarity-aware attention flow network for change detection with VHR remote sensing images," *IEEE Trans. Image Process.*, pp. 1–1, 2024.

[31] Z. Zheng, Y. Zhong, S. Tian, A. Ma, and L. Zhang, "ChangeMask: Deep multi-task encoder-transformer-decoder architecture for semantic change detection," *ISPRS J. Photogramm. Remote Sens.*, vol. 183, pp. 228–239, Jan. 2022.

[32] G. Wang, G. Cheng, P. Zhou, and J. Han, "Cross-Level Attentive Feature Aggregation for Change Detection," *IEEE Trans. Circuits Syst. Video Technol.*, pp. 1–1, 2023.

[33] L. Pu, ZadehAmir, and MorencyLouis-Philippe, "Foundations & trends in multimodal machine learning: Principles, challenges, and open questions," *ACM Comput. Surv.*, Jun. 2024.

[34] X. Ma, X. Zhang, Z. Wang, and M.-O. Pun, "Unsupervised Domain Adaptation Augmented by Mutually Boosted Attention for Semantic Segmentation of VHR Remote Sensing Images," *IEEE Trans. Geosci. Remote Sens.*, vol. 61, pp. 1–15, 2023.

[35] S. Yao, K. Yang, X. Dong, X. Shi, and J. Wang, "Large-gradient deformation monitoring and parameter inversion in a mining area using a method combining a dynamic prediction model and InSAR," *Remote Sens. Lett.*, vol. 12, no. 9, pp. 838–847, Sep. 2021.

[36] J. Wang *et al.*, "Deriving mining-induced 3-D deformations at any moment and assessing building damage by integrating single InSAR interferogram and gompertz probability integral model (SII-GPIM)," *IEEE Trans. Geosci. Remote Sens.*, vol. 60, p. 4709817, 2022.

[37] Z. Gao, L. Yan, H. Xie, P. Wei, H. Wu, and J. Wang, "TSTD: A Cross-modal Two Stages Network with New Trans-decoder for Point Cloud Semantic Segmentation," in *Pattern Recognition and Computer Vision*, Q. Liu, H. Wang, Z. Ma, W. Zheng, H. Zha, X. Chen, L. Wang, and R. Ji,

Eds., in Lecture Notes in Computer Science. Singapore: Springer Nature, 2024., pp. 130–141.

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

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

[40] Y. Cao, X. Huang, and Q. Weng, "A multi-scale weakly supervised learning method with adaptive online noise correction for high-resolution change detection of built-up areas," *Remote Sens. Environ.*, vol. 297, p. 113779, Nov. 2023.

[41] H. Chen, Z. Qi, and Z. Shi, "Remote Sensing Image Change Detection With Transformers," *IEEE Trans. Geosci. Remote Sens.*, vol. 60, pp. 1–14, 2022.

[42] L. Mou, Y. Hua, S. Saha, F. Bovolo, L. Bruzzone, and X. X. Zhu, "Detecting Changes by Learning No Changes: Data-Enclosing-Ball Minimizing Autoencoders for One-Class Change Detection in Multispectral Imagery," *IEEE Trans. Geosci. Remote Sens.*, vol. 60, pp. 1–16, 2022.

[43] X. Zhang, H. Su, C. Zhang, X. Gu, X. Tan, and P. M. Atkinson, "Robust unsupervised small area change detection from SAR imagery using deep learning," *ISPRS J. Photogramm. Remote Sens.*, vol. 173, pp. 79–94, Mar. 2021.

[44] W. Zhang, L. Jiao, F. Liu, S. Yang, and J. Liu, "Adaptive Contourlet Fusion Clustering for SAR Image Change Detection," *IEEE Trans. Image Process.*, vol. 31, pp. 2295–2308, 2022.

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

[46] W. Shi, Q. Meng, L. Zhang, J. Wang, T. Zhou, and P. M. Atkinson, "Building Height Extraction from Monocular Off-Nadir Satellite Sensor Imagery," in *IGARSS 2024 - 2024 IEEE International Geoscience and Remote Sensing Symposium*, Jul. 2024., pp. 2537–2540.

[47] Y. Li, K. Yang, and H. Zhao, "Scale transfer learning of hyperspectral prediction model of heavy metal content in maize: From laboratory to satellite," *Int. J. Remote Sens.*, vol. 44, no. 8, pp. 2590–2610, Apr. 2023.

[48] J. Wang, J. Ma, K. Yang, S. Yao, and X. Shi, "Effects and laws analysis for the mining technique of grouting into the overburden bedding separation," *J. Clean. Prod.*, vol. 288, p. 125121, Mar. 2021.

[49] S. Yao and T. Balz, "Fringe Estimation in Distributed Scatterer Interferometry," *IEEE Trans. Geosci. Remote Sens.*, vol. 61, pp. 1–15, 2023.

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

[51] Z. Lv, J. Liu, W. Sun, T. Lei, J. A. Benediktsson, and X. Jia, "Hierarchical Attention Feature Fusion-Based Network for Land Cover Change Detection With Homogeneous and Heterogeneous Remote Sensing Images," *IEEE Trans. Geosci. Remote Sens.*, vol. 61, pp. 1–15, 2023.

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

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

[54] H. Li, M. Gong, M. Zhang, and Y. Wu, "Spatially Self-Paced Convolutional Networks for Change Detection in Heterogeneous Images," *IEEE J. Sel. Top. Appl. Earth Obs. Remote Sens.*, vol. 14, pp. 4966–4979, 2021.

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

[56] J. Prendes, M. Chabert, F. Pascal, A. Giros, and J.-Y. Tourneret, "A Bayesian Nonparametric Model Coupled with a Markov Random Field for Change Detection in Heterogeneous Remote Sensing Images," *SIAM J. Imaging Sci.*, vol. 9, no. 4, pp. 1889–1921, Jan. 2016.

[57] R. Touati, M. Mignotte, and M. Dahmane, "A Reliable Mixed-Norm-Based Multiresolution Change Detector in Heterogeneous Remote Sensing Images," *IEEE J. Sel. Top. Appl. Earth Obs. Remote Sens.*, vol. 12, no. 9, pp. 3588–3601, Sep. 2019.

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

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

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

[61] H. Chen, N. Yokoya, C. Wu, and B. Du, "Unsupervised Multimodal Change Detection Based on Structural Relationship Graph Representation Learning," *IEEE Trans. Geosci. Remote Sens.*, vol. 60, pp. 1–18, Dec. 2022.

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

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

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

[65] X. Niu, M. Gong, T. Zhan, and Y. Yang, "A Conditional Adversarial Network for Change Detection in Heterogeneous Images," *IEEE Geosci. Remote Sens. Lett.*, vol. 16, no. 1, pp. 45–49, Jan. 2019.

[66] L. T. Luppino *et al.*, "Deep Image Translation With an Affinity-Based Change Prior for Unsupervised Multimodal Change Detection," *IEEE Trans. Geosci. Remote Sens.*, vol. 60, p. 4700422, 2022.

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

[68] Y. Chen and L. Bruzzone, "Self-Supervised Change Detection in Multiview Remote Sensing Images," *IEEE Trans. Geosci. Remote Sens.*, vol. 60, p. 5402812, 2022.

[69] J. Wang, L. Yan, H. Xie, T. Zhou, W. Shi, and P. M. Atkinson, "Unsupervised Multimodal Change Detection by Distilling Common and Discrepant Representations," in *IGARSS 2024 - 2024 IEEE International Geoscience and Remote Sensing Symposium*, Jul. 2024., pp. 7817–7820.

[70] H. Xia, Y. Tian, L. Zhang, and S. Li, "A Deep Siamese Postclassification Fusion Network for Semantic Change Detection," *IEEE Trans. Geosci. Remote Sens.*, vol. 60, pp. 1–16, 2022.

[71] J. Zhang, L. Ding, T. Zhou, J. Wang, P. M. Atkinson, and L. Bruzzone, "Recurrent semantic change detection in VHR remote sensing images using visual foundation models," *IEEE Trans. Geosci. Remote Sens.*, vol. 63, pp. 1–14, 2025.

[72] P. Chen, B. Zhang, D. Hong, Z. Chen, X. Yang, and B. Li, "FCCDN: Feature constraint network for VHR image change detection," *ISPRS J. Photogramm. Remote Sens.*, vol. 187, pp. 101–119, May 2022.

[73] M. Fuentes Reyes *et al.*, "A 2D/3D multimodal data simulation approach with applications on urban semantic segmentation, building extraction and change detection," *ISPRS J. Photogramm. Remote Sens.*, vol. 205, pp. 74–97, Nov. 2023.

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

[75] F. Cui and J. Jiang, "MTSCD-net: A network based on multi-task learning for semantic change detection of bitemporal remote sensing images," *Int. J. Appl. Earth Obs. Geoinformation*, vol. 118, p. 103294, Apr. 2023.

[76] Y. Niu, H. Guo, J. Lu, L. Ding, and D. Yu, "SMNet: Symmetric multi-task network for semantic change detection in remote sensing images based on CNN and transformer," *Remote Sens.*, vol. 15, no. 4, Art. no. 4, Jan. 2023.

[77] X. Zuo *et al.*, "Multitask siamese network guided by enhanced change information for semantic change detection in bitemporal remote sensing

images," *IEEE J. Sel. Top. Appl. Earth Obs. Remote Sens.*, vol. 18, pp. 61–77, 2025.
[78] Z. Li *et al.*, "STADE-CDNet: Spatial–temporal attention with difference enhancement-based network for remote sensing image change detection," *IEEE Trans. Geosci. Remote Sens.*, vol. 62, pp. 1–17, 2024.
[79] P. Ebel, S. Saha, and X. X. Zhu, "FUSING MULTI-MODAL DATA FOR SUPERVISED CHANGE DETECTION," *Int. Arch. Photogramm. Remote Sens. Spat. Inf. Sci.*, vol. XLIII-B3-2021, pp. 243–249, Jun. 2021.
[80] T. Liu *et al.*, "Hierarchical feature alignment-based progressive addition network for multimodal change detection," *Pattern Recognit.*, vol. 162, p. 111355, Jun. 2025.
[81] S. Hafner, Y. Ban, and A. Nascetti, "Semi-supervised urban change detection using multi-modal sentinel-1 SAR and sentinel-2 MSI data," *Remote Sens.*, vol. 15, no. 21, Art. no. 21, Jan. 2023.

Jian Wang (Graduate Student Member, IEEE) received the M.S. degree in surveying and mapping engineering from China University of Mining and Technology-Beijing, Beijing, China, in 2021.

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

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

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

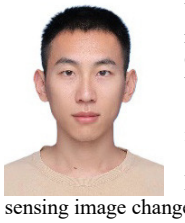
Li Yan received the B.S., M.S., and Ph.D. degrees in photogrammetry and remote sensing from Wuhan University, Wuhan, China, in 1989, 1992, and 1999, respectively.

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

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

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.



Yanheng Wang (Graduate Student Member, IEEE) received the M.S. degree in software engineering from the College of Information Technology, Shanghai Ocean University, Shanghai, China, in 2020.

He is currently pursuing the Ph.D. degree in artificial intelligence with the College of Intelligent Systems Science and Engineering, Harbin Engineering University, Harbin, China. His research interests include remote sensing image change detection, object detection, and deep learning.

Jing Zhang received a master's degree in software engineering from Beijing University of Technology, Beijing, China, in 2020. She is currently a Ph.D. student in the Department of Information Engineering and Computer Science at the University of Trento, Italy. She also was a visiting researcher at Lancaster University, UK. Her current research interests include semantic change detection and remote sensing mage 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*.

Lorenzo Bruzzone (Fellow, IEEE) received the Laurea (M.S.) degree in electronic engineering (*summa cum laude*) and the Ph.D. degree in telecommunications from the University of Genoa, 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. He is the Founder and the Director of the Remote Sensing Laboratory (https://rslab.disi.unitn.it/) with 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 currently the Principal Investigator of the Radar for Icy Moon Exploration (RIME) and of the Subsurface Radar Sounder (SRS) instruments in the framework of the JUpiter ICy moons Explorer (JUICE) and the Envision missions, respectively, of the European Space Agency (ESA) and is the for the High Resolution Land Cover project in the framework of the Climate Change Initiative of ESA. He is the author (or coauthor) of more than 400 scientific publications in referred international journals, more than 390 papers in conference proceedings, and 25 book chapters. He is editor/co-editor of 18 books/conference proceedings and two scientific books. His papers are highly cited, as proven from the total number of citations (more than 56500) and the value of the h-index (109) (source: Google Scholar). He was invited as keynote speaker in more than 40 international conferences and workshops.

Dr. Bruzzone was a member of the Administrative Committee of the *IEEE Geoscience and Remote Sensing Society* (GRSS) between 2009 and 2023. He ranked first place in the Student Prize Paper Competition of the 1998 IEEE International Geoscience and Remote Sensing Symposium (IGARSS), Seattle, in July 1998. Since that he was recipient of many international and national honors and awards, including among the most recent ones the IEEE GRSS 2015 Outstanding Service Award, the 2017 and 2018 IEEE IGARSS Symposium Prize Paper Awards, the 2019 WHISPER Outstanding Paper Award and the 2022 Letter Prize Paper Award for the best paper published on the *IEEE Geoscience and Remote Sensing Letters* in 2022. He 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.

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

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

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