Multimodal Co-learning for Building Change Detection: A Domain Adaptation Framework Using VHR Images and Digital Surface Models

Yuxing Xie, Xiangtian Yuan, Xiao Xiang Zhu, *Fellow, IEEE* and Jiaojiao Tian, *Senior member, IEEE*

***Abstract*—In this article, we propose a multimodal co-learning framework for building change detection. This framework can be adopted to jointly train a Siamese bitemporal image network and a height difference map (HDiff) network with labeled source data and unlabeled target data pairs. Three co-learning combinations (vanilla co-learning, fusion co-learning, and detached fusion co-learning) are proposed and investigated with two types of co-learning loss functions within our framework. Our experimental results demonstrate that the proposed methods are able to take advantage of unlabeled target data pairs and therefore enhance the performance of single-modal neural networks on the target data. In addition, our synthetic-to-real experiments demonstrate that the recently published synthetic dataset SMARS is feasible to be used in real change detection scenarios, where the optimal result is with the F1 score of 79.29%.**

***Index Terms*—change detection, co-learning, multimodal learning, domain adaptation, digital surface models (DSMs)**

I. Introduction

BUILDING change detection is an essential yet challenging task in the remote sensing (RS) field. It aims to identify the differences in the condition of building objects within defined areas from multi-temporal 2D, 2.5D, or 3D data [1]. Detection of building changes is required in a wide range of real-world applications, such as urban monitoring [2], disaster assessment [3], and map updating [4]. Building change detection methods can be categorized into two kinds of pipelines: (1) change detection based on post-classification, which first predicts building masks for bitemporal data and then generates building change maps based on the difference of predicted building masks. (2) Direct change detection, which directly extracts change features and converts the features to building change maps. In this work, we concentrate on the latter. Unless specified otherwise in the text, the follow-up "building change detection" or "change detection" in this article refers to direct change detection. Direct change detection commonly consists of two steps: feature extraction and change detection [5].

Before utilizing machine learning methods for change detection, traditional transformation-based algorithms and image algebraic operations were mainstream approaches [5]–[8]. These methods usually first calculate the difference between bitemporal images and then apply threshold- or clustering-based classification algorithms on the image difference to generate change maps [9]. However, these pixel-based methods are limited to processing low- or medium-resolution images because they cannot analyze contextual relationships. Although some improved object-based methods are designed to deal with high- and very high-resolution images, they still have obvious limitations such as being sensitive to noise and computationally expensive [5], [9]–[11]. They typically achieve low accuracy when dealing with large-scale diversity-enriched data due to the poor generalization ability of handcrafted features.

As change detection can be regarded as a classification problem, machine learning approaches are naturally introduced. Similar to machine learning-based studies in other remote sensing fields, support vector machine (SVM) and random forest (RF) [12]–[15] are the two most popular models for change detection before the deep learning methods are commonplace. Additionally, graphical models such as Markov random field and conditional random field are widely employed for the purpose of better utilizing contextual relationships and generating fine-grained boundaries [16]–[19]. However, these machine learning methods are still difficult to effectively apply in large-scale datasets with obvious domain gaps. It is a huge challenge to design effective universal change features manually.

The rapid advancement of deep neural networks in recent years has set new standards in supervised 2D change detection [5], [20]–[23]. Specifically, the success of convolutional neural networks (CNNs) in other remote sensing and computer vision tasks [24]–[27] has established CNNs as the backbone for change detection in numerous studies. Few of them are based on single-stream architectures [21], [28], [29], which take as input image differencing, hand-crafted change features, or concatenation of bitemporal images. Due to the large variability between the pre- and post-event images, the single-stream methods often suffer from noise and loss of information

Manuscript received xx xx, xxxx. (*Corresponding authors: Yuxing Xie.*)

The work of Y. Xie was supported by a DLR-DAAD Research Fellowship (No. 57424731). The work of X. Zhu is jointly supported by the German Federal Ministry of Education and Research (BMBF) in the framework of the international future AI lab "AI4EO – Artificial Intelligence for Earth Observation: Reasoning, Uncertainties, Ethics and Beyond" (grant number: 01DD20001) and by the Munich Center for Machine Learning.

Yuxing Xie is with the Data Science in Earth Observation, Technical University of Munich, 80333 Munich, Germany, and was with the Remote Sensing Technology Institute (IMF), German Aerospace Center (DLR), 82234 Wessling, Germany (e-mail: yuxing.xie@outlook.com, yuxing.xie@dlr.de).

Xiangtian Yuan is with the Remote Sensing Technology Institute (IMF), German Aerospace Center (DLR), 82234 Wessling, Germany (e-mail: xiangtian.yuan@dlr.de).

Xiao Xiang Zhu is with the Data Science in Earth Observation, Technical University of Munich, 80333 Munich, Germany and also with the Munich Center for Machine Learning, Munich, Germany (e-mail: xiaoxiang.zhu@tum.de).

Jiaojiao Tian is with the Remote Sensing Technology Institute (IMF), German Aerospace Center (DLR), 82234 Wessling, Germany (e-mail: jiaojiao.tian@dlr.de).

from the input, inhibiting a wider application in remote sensing change detection. Consequently, the mainstream methods are based on the Siamese architecture [21], [30], which extracts features from bitemporal images via two parallel encoders with the same network structure. The Siamese approaches not only maintain the information lost in single-stream approaches but also exhibit more robust and distinctive representations for the object of interest [31]–[40]. More recently, vision transformer (ViT) [41], [42] has achieved further success in deep learning-based image processing topics and attracted attention from the remote sensing community. Transformer-based methods have also been introduced in several recent change detection studies [30], [43]–[45] and achieved stunning results. With the development of foundation models, recent studies like [46]–[48] have successfully incorporated them into the change detection task, paving a way for further improvement in this area.

Despite the remarkable performance of the 2D deep learning methods on benchmark datasets, their real-world applications are still constrained. Myriad state-of-the-art change detection methods are in a fully supervised fashion. While satisfactory results on the test sets of benchmark datasets can be achieved, the performance of the trained models on other datasets usually displays a steep decline as a result of domain gap [34], [49]. In remote sensing data, domain gaps can be attributed to the differences in image sensors, spatial resolutions, acquisition conditions, etc. In RS change detection specifically, change features can be very dissimilar depending on the location of the data, e.g., urban and rural, Europe and Asia. domain gap is widespread between different change detection datasets of optical images. Yet this challenge has not been overcome by 2D fully supervised methods [34], [50]. To make things worse, annotating RS change detection data is not only extremely time-consuming and requires specific knowledge of the regions, but is also error-prone as unchanged areas are dominant. Therefore, creating the annotation of an unseen area for fine-tuning is not practical. Another issue is the intrinsic limits of 2D data in identifying changes. The change in height can not be quantified with only 2D orthorectified images. Consequently, geometric information is receiving increasing attention.

Benefiting from the development of photogrammetric techniques, 3D sensors such as LiDAR, as well as TomoSAR techniques, 2.5D and 3D data have been becoming easier to obtain. As 2.5D and 3D data have rich geometric information, they can better describe regular man-made objects including buildings and their changes, and provide more discriminative features [1], [51]. As a result, several traditional change detection methods employed bitemopral DSMs as the input data for building change detection. The simplest approach is DSMs subtraction, which is computationally cheaper, and achieves good performance when using high-quality DSMs from LiDAR and airborne stereo data [52], [53]. To improve the change detection accuracy, various refinement approaches are introduced. For instance, building indicators from images [54], [55], shape information [56] or the existing GIS cadastral maps [57]. In our previous study, we notice that 2.5D data has better generalization performance than 2D images with appropriate deep neural networks [51], [58]. Naturally, a question (A) comes out for the building change detection task: Do neural networks designed for DSMs also demonstrate better generalization performance than those designed for 2D images?

Although DSMs are good at describing geometric features, they also have disadvantages such as inevitable outliers and unsharpened building boundaries, which could result in incorrect change detection [1]. Furthermore, due to the diversity of the data, it is impossible to ensure that the domains of the target and source data are always consistent. Domain gap is also one of the main problems constraining the effectiveness of deep learning algorithms in the representation of 2.5D/3D data [58], [59]. Desiring beyond homogeneous data, a few learning-based studies have shifted the focus from single-modal methods to multimodal data fusion, enriching the features or probabilities via a fusion operation (e.g., summation, average, concatenation, etc.). 2D-2.5D/3D data fusion utilizing multimodal data as inputs for a fusion framework may increase the accuracy of change detection [60]. Recently, multimodal knowledge transfer semi-supervised learning architecture represented by co-learning utilizing multimodal data pairs only for the training phase [61], [62] has attracted the attention in remote sensing tasks such as building extraction [51], [58] and semantic segmentation [63]. These methods can further enhance the generalization performance of image networks and DSM/point cloud-based networks, breaking the constraints of domain gaps. Naturally, another question (B) comes to our minds: Are there any co-learning architectures suitable for building change detection when the source data and target data are with large domain gaps?

With the maturity of photogrammetry techniques like structure from motion and dense matching [64], [65], it is no longer a big challenge to obtain high-quality DSMs. Nowadays, UAV data are widely used in local and near real-time surveillance applications [66]. Almost any commercial UAV image data processing software can produce DSMs. For large-scale monitoring, more satellites like Pléiades-Neo [67], WorldView [68], and Gaofen [69] series are available to provide VHR optical images and stereo-/multi-view vision products including DSMs. At minimal cost, well-matched orthophotos and DSMs can even be derived from a single pair of high-resolution stereo images by photogrammetry algorithms. Such aligned orthophotos and DSMs require low acquisition costs and are therefore commonly used in real applications [67], [70], [71]. However, existing learning-based 2.5D change detection studies are very limited. Therefore, in this work, we investigate the advantage of utilizing 2.5D imagery-derived photogrammetric DSMs as the input for change detection, and an effective co-learning framework with corresponding 2D optical images, to answer the above-mentioned questions A and B. To sum up, the contributions of our work are as follows:

1) We propose a co-learning framework for bitemporal images and DSMs modalities, focusing on the building change detection task. Three well-designed co-learning combinations (vanilla co-learning, fusion co-learning, and detached fusion co-learning) are proposed, defined,

and investigated in this work. Furthermore, we present a way to determine whether these co-learning combinations are equivalent for different loss functions.

2) This work highlights the advantages of photogrammetric DSMs in the task of building change detection. Compared with 2D optical imagery, existing studies on photogrammetric DSMs are limited. We propose an end-to-end transformer-based network for change detection from HDiff maps and investigate the difference from 2D change detection in cross-domain scenarios.
3) This work also involves synthetic-to-real domain adaptation, a novel topic in remote sensing. To the best of our knowledge, this is the first study to address this topic specifically for the change detection task. We utilize co-learning as a domain adaptation method and explore the potential of using the recently published synthetic benchmark dataset SMARS [72] to train change detection deep neural networks for a real dataset.

Our experiments demonstrate that the proposed co-learning methods can effectively transfer mutual information across modalities and improve the performance of the Siamese network and the proposed HDiff map networks on cross-domain target data.

The remainder of this paper is organized as follows: section II reviews related works on multimodal deep learning with 2D images and 2.5D/3D Data, as well as multimodal change detection. Section III introduces the methodology employed in our work. Section IV describes the implementation of experiments and results comparisons of different methods. Section V presents the discussion on experiments and methodology. Last but not least, section VI concludes the paper.

II. Related works

A. Multimodal Deep Learning with 2D Multispectral Images and 2.5D/3D Data

Depending on how the information from both modalities is utilized, multimodal deep learning works with 2D images and 2.5D DSMs/3D point clouds in the remote sensing field can be generally classified into two categories: data fusion and knowledge transfer.

Data fusion refers to the techniques of combining multimodal data and related information during the process. It is based on the intuition that improved accuracy could be achieved with multimodal information compared with using single-modal data alone [73]. Depending on the locations where the fusion operations take place, data fusion approaches can be categorized into early fusion (observation-level fusion), middle fusion (feature-level fusion), late fusion (decision-level fusion) [51], [73], and their combinations.

Early fusion is carried out at the data input stage. In remote sensing tasks, 2D multispectral images are concatenated with the height values of DSMs or normalized DSMs (nDSMs) as the input channels to a single-modal network. For example, [74] proposes the gated residual refinement network (GRRNet) using multispectral images and LiDAR-derived nDSMs as the input. A gated feature labeling (GFL) unit is designed in the decoder to refine the semantic segmentation results. In a few early fusion studies, spectral information from images is added directly to 3D point clouds as per-point values, and colorized point clouds are processed in a three-dimensional domain with point cloud deep neural networks. However, till now no consensus has been reached on whether coloring the 3D point clouds brings advantages [75]. Some earlier studies found such fusion operations can even lead to a decline in the performance of point cloud networks [51], [76], [77].

Middle fusion is carried out at feature embedding levels in the middle of the model, aiming at fusing deeper features of different modalities into a composite one. The subsequent operations such as convolution are based on the fused features. For instance, [70] adopts a FuseNet-like [78] semantic segmentation architecture with feature fusion modules. Multispectral images and nDSMs are processed by two individual encoders. In addition, a third encoder, namely the virtual encoder for fused feature maps of two modalities is introduced. The virtual encoder takes its previous activations concatenated with the activations from the other two encoders as the input. A single-stream decoder is utilized to upsample the encoded fused representation afterward. This symmetrical design can alleviate the need to select the main modality source. [79] proposes a CNN architecture with a fusion operation combining features from three parallel networks for building extraction. Each parallel network processes one data modality. The input data to this architecture contain RGB images, panchromatic images, and nDSMs. Experimental results demonstrate that the fusion of several networks has superior generalization performance on unseen data. [80] proposes a dual-channel scale-aware semantic segmentation network with position and channel attentions (DSPCANet), which uses two branches to process multispectral images and DSM rasters individually. Multimodal features are concatenated and further refined by a channel attention module and an improved position attention module. [71] presents an end-to-end cross-modal gated fusion network (CMGFNet) for building extraction, which introduces a gated fusion module (GFM) for fusing features from separate multispectral image encoder and DSM encoder. Experiments on three datasets demonstrate that GFM can produce features that contain more discriminative information about building objects and backgrounds than traditional summation and concatenation feature fusion methods.

Late fusion is carried out at the decision stage of the model, which fuses probability maps output from deep learning models of different modalities. For instance, [70] designs a late fusion semantic segmentation architecture for multispectral images and nDSMs. This method first averages predictions from two modalities to generate a smooth fused prediction. Then a residual correction module is applied to refine the probability with a small offset. This architecture is tested with SegNet and ResNet as the backbone and is suited to combine different strong deep learning models that are confident in the predictions. To further exploit the advantages of each fusion strategy, some works adopted multiple fusion strategies and conducted more complex multimodal networks [81], [82].

Knowledge transfer does not directly operate on the data or extracted features. There are two principles of knowledge transfer methods: (1) employing different network branches

for different data modalities. (2) Bridging the relationships between different modalities by soft connections (usually loss functions). Each network only influences others in the training phase and can be utilized alone for testing single-modal data. Compared with the data fusion strategies, knowledge transfer is more flexible and therefore is more applicable in various scenarios, such as in the case of missing modalities during the testing time. In addition, another limitation of data fusion is the inefficient utilization of the complete information of the raw heterogeneous data and the complementary nature of multimodalities, which may result in incorrect and irrelevant feature representations [63], [71]. In contrast, knowledge transfer always uses different network branches to process different data modalities, effectively maintaining the completeness of heterogeneous information and reducing noisy information from the other modalities. In recent years, 2D/3D co-learning-based approaches belonging to the knowledge transfer category have been introduced in the remote sensing field. As a pioneer, our previous work [51] presents a co-learning framework for 2D and 3D building extraction networks with multispectral images and photogrammetric point clouds, which significantly improves the performance of both image and point cloud networks with very few labeled data pairs and a large quantity of unlabeled data pairs. In [58], we extend the co-learning framework proposed in [51] for the cross-domain building extraction task and the spaceborne-to-airborne experiment demonstrates the power of such methodology on an unlabeled target dataset. Recently, [63] proposes an imbalance knowledge-driven multimodal network (IKD-Net) for the semantic segmentation task, combining conventional data fusion and co-learning. In its network architecture, IKD-Net adopted a feature fusion module and a class knowledge-guided module to refine the image feature maps with the features from the strong LiDAR point cloud modality. A similarity constraint is enforced as the co-learning loss function to guide the weak image modality with mutual knowledge from the strong LiDAR point cloud modality.

B. Change Detection with Multimodal Data

Compared with the single-modal image or DSM data, multimodal data provide more stable and accurate change features. Therefore, several studies have introduced multimodal strategies for change detection. For example, in our previous works the decision fusion method belief functions have been proven to be an efficient fusion module for multimodal change detection [60], [83], [84], which can effectively improve the building change detection results compared with single-modal change indicators. The paper [83] proposes a change detection pipeline based on the robust height differences between DSMs and the similarity measurement between corresponding optical image pairs. A fusion module based on the Dempster-Shafer theory is adopted to fuse these two change indicators, which significantly improves the change accuracy compared with the results of either single modality. Additionally, vegetation and shadow classification results are introduced as extra information to refine the initial change detection results, and a building extraction method based on shape features is performed to get more accurate building change maps. [84] proposes another multimodal change detection framework. First, it uses a refined basic belief assignments (BBAs) model to calculate the BBAs of the change indicators from optical images and DSMs. Then a building change detection decision fusion approach is applied to fuse these BBAs. Finally, four decision-making criteria are employed to convert the fused global BBAs to building change maps. [60] extends the framework in [84] and employs initial building probabilities extracted by the deep neural network Deeplabv3+ for the change decision, which shows better generalization ability than the previous version. Also based on the Dempster–Shafer theory, [85] introduces a complementary evidence fusion framework. In this framework, the image change indicator is calculated with the subtraction of the normalized difference vegetation index (NDVI) of bitemporal images. A complementary evidence combination rule is employed for the decision fusion to alleviate the conflicts between the change evidence from optical images and DSMs. Recently, [86] utilizes the morphological building index (MBI) as the image change feature and robust height difference proposed in [83] as the height change feature and proposes a co-segmentation framework for building change detection. The changed areas and unchanged areas are distinguished by a graph-cut-based energy minimization method.

Nevertheless, end-to-end deep learning-based multimodal change detection methods have not been widely investigated, which is partly due to the lack of sufficient public datasets [72], [87]. Although [60] involves deep learning, it only uses the network for building extraction rather than change detection. The lack of sufficient multimodal change detection data impedes the development of robust end-to-end methods with strong cross-domain generalizability. The flexible requirement for data of the co-learning framework could have huge implications for multi-modal change detection research.

III. Methodology

A. Overview

We aim to develop a generic image-DSM co-learning framework for the building change detection task. This framework is based on two individual CNN-transformer-fused networks for the modalities images and DSMs, respectively. Fig. 1 illustrates the overview of the framework. In this framework, two networks can be trained jointly with labeled training data and partially unlabeled multimodal data pairs. The DSMs are processed in the format of height difference. This is because height difference can play a better generalization ability with explicit geometric features, while bitemporal DSMs can not be well utilized by the Siamese image network. Related comparisons are presented in section IV. To generate HDiff maps, different methods can be used. In our framework, two height difference operations are designed: direct height difference and robust height difference [83].

The following subsections give detailed introductions and descriptions of the methods used in this framework.

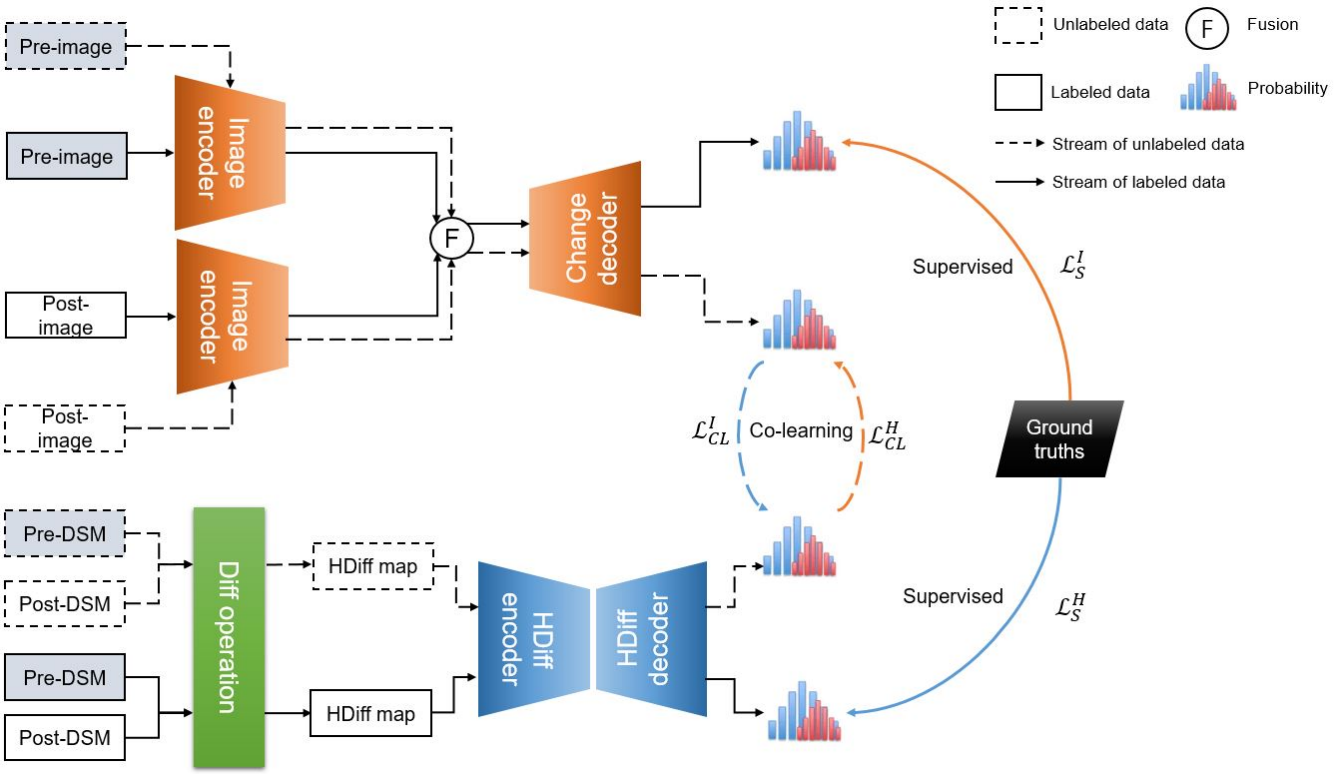


Fig. 1. Our proposed co-learning change detection framework.

B. Problem Statement: Co-learning for Cross-Domain Change Detection

Assume that there are two datasets in a cross-domain scenario, the source dataset $\mathbb{D}_s$ and the target dataset $\mathbb{D}_t$. Each dataset includes bitemporal data. In the following text, we use subscripts 1 and 2 to denote pre- and post-event data, respectively. $\mathbb{D}_s$ consists of labeled source samples $\{\{I_1^s, I_2^s\}, \{H_1^s, H_2^s\}, G^s\}$, including pre-images I_1^s, post-images I_2^s, pre-DSMs H_1^s, post-DSMs H_2^s, and the change detection ground truths G^s. $\mathbb{D}_t$ consists of unlabeled target samples $\{\{I_1^t, I_2^t\}, \{H_1^t, H_2^t\}\}$, including pre-images I_1^t, post-images I_2^t, pre-DSMs H_1^t, and post-DSMs H_2^t.

f_I is the image branch operation (i.e., the image change detection network) for pre-/post-image pairs $\{I_1^s, I_2^s\}$ and $\{I_1^t, I_2^t\}$. The building change probabilities P_I^s and P_I^t predicted by the image branch operation are calculated as follows:

$$P_I^s = f_I(I_1^s, I_2^s) \,, \tag{1}$$

$$P_I^t = f_I(I_1^t, I_2^t) \,, \tag{2}$$

f_H is the DSM branch operation (including a height difference preprocessing operation and HDiff map network) for pre-/post-DSM pairs. The probabilities P_H^s and P_H^t for DSM pairs$\{H_1^s, H_2^s\}$ and $\{H_1^t, H_2^t\}$ predicted by the DSM branch are calculated as the follows:

$$P_H^s = f_H(H_1^s, H_2^s) \,, \tag{3}$$

$$P_H^t = f_H(H_1^t, H_2^t) \,, \tag{4}$$

1) Supervised Change Detection with Labeled Source Data: To supervise the pixel-wise change detection, a generic loss function L_S measuring the difference between the source building change probability P_I^s/P_H^s and ground truth G_s is needed:

$$\mathcal{L}_S^I = L_S(G^s || P_I^s) \,, \tag{5}$$

$$\mathcal{L}_S^H = L_S(G^s || P_H^s) \,, \tag{6}$$

where $\mathcal{L}_S^I$ and $\mathcal{L}_S^H$ denote the supervised change detection loss function for image modality and DSM modality, respectively.

2) Co-learning with Unlabeled Target Data: In this subsection, we propose three co-learning combinations: vanilla co-learning, fusion co-learning, and detached fusion co-learning.

Vanilla Co-learning: This is the co-learning implementation following the idea presented in [51], which is based on the intuition that if both the image branch and DSM branch can produce good predictions, their building change probabilities P_I^t and P_H^t should be consistent with each other. Hence, the target co-learning problem is formulated as a generic consistency loss function L_C to minimize the distributions of P_I^t and P_H^t. The vanilla co-learning loss functions for image modality $\mathcal{L}_{CL-V}^I$ and DSM modality $\mathcal{L}_{CL-V}^H$ are calculated as follows:

$$\mathcal{L}_{CL-V}^I = L_C(P_{H,d}^t || P_I^t) \,, \tag{7}$$

$$\mathcal{L}^H_{CL-V} = L_C(P^t_{I,d}||P^t_H) \,, \tag{8}$$

where $P^t_{H,d}$ and $P^t_{I,d}$ refer to detached P^t_H and P^t_I, respectively. Detached probabilities mean they are variables removed from the gradient computational graph so they do not affect the update of the weights for the corresponding networks. They can be named shadow reference probability, utilized by the main modality network as the reference in the co-learning loss function [51].

Fusion Co-learning: This co-learning method is based on the intuition that if both the image branch and DSM branch can produce good predictions, their building change probabilities P^t_I and P^t_H should be consistent with the average fusion probability $\frac{P^t_I+P^t_H}{2}$. Hence, the target co-learning problem is formulated as a generic consistency loss function L_C to minimize the predicted probability distributions of P^t_I/P^t_H and shadow reference probability $\frac{P^t_I+P^t_H}{2}$. The fusion co-learning loss functions for image modality $\mathcal{L}^I_{CL-F}$ and DSM modality $\mathcal{L}^H_{CL-F}$ are calculated as follows:

$$\mathcal{L}^I_{CL-F} = L_C(\frac{P^t_I + P^t_{H,d}}{2}||P^t_I) \,, \tag{9}$$

$$\mathcal{L}^H_{CL-F} = L_C(\frac{P^t_{I,d} + P^t_H}{2}||P^t_H) \,, \tag{10}$$

where $P^t_{H,d}$ and $P^t_{I,d}$ refer to detached P^t_H and P^t_I, respectively.

Detached Fusion Co-learning: If the average probability $\frac{P^t_I+P^t_H}{2}$ is fully detached from the computational graph and as a constant, another co-learning format is obtained. We name it detached fusion co-learning. The detached fusion co-learning loss functions for image modality $\mathcal{L}^I_{CL-DF}$ and DSM modality $\mathcal{L}^H_{CL-DF}$ are calculated as follows:

$$\mathcal{L}^I_{CL-DF} = L_C(\frac{P^t_{I,d} + P^t_{H,d}}{2}||P^t_I) \,, \tag{11}$$

$$\mathcal{L}^H_{CL-DF} = L_C(\frac{P^t_{I,d} + P^t_{H,d}}{2}||P^t_H) \,, \tag{12}$$

where L_C denotes a generic consistency loss function. $P^t_{H,d}$ and $P^t_{I,d}$ refer to detached P^t_H and P^t_I, respectively.

In some cases, L_C may result in the situation that two or even all of $\mathcal{L}_{CL-V}$, $\mathcal{L}_{CL-F}$, and $\mathcal{L}_{CL-DF}$ are equivalent. Appendix A gives a way to evaluate whether three co-learning combinations are inequivalent.

3) Total loss function: The total loss function is a weighted sum of the above-mentioned individual losses calculated during the training iteration. In our framework, combining the supervised change detection loss function $\mathcal{L}^I_S$/$\mathcal{L}^H_S$ and the co-learning loss function $\mathcal{L}^I_{CL}$/$\mathcal{L}^H_{CL}$, the total loss function of the training phase can be obtained:

$$\mathcal{L}^I_{total} = \lambda_1\mathcal{L}^I_S + \lambda_2\mathcal{L}^I_{CL} \,, \tag{13}$$

$$\mathcal{L}^H_{total} = \lambda_1\mathcal{L}^H_S + \lambda_2\mathcal{L}^H_{CL} \,, \tag{14}$$

where $\mathcal{L}^I_{CL} \in \{\mathcal{L}^I_{CL-V}, \mathcal{L}^I_{CL-F}, \mathcal{L}^I_{CL-DF}\}$ and $\mathcal{L}^H_{CL} \in \{\mathcal{L}^H_{CL-V}, \mathcal{L}^H_{CL-F}, \mathcal{L}^H_{CL-DF}\}$. $\mathcal{L}^I_{total}$, $\mathcal{L}^I_S$, and $\mathcal{L}^I_{CL}$ are the total loss function, the supervised loss function, and the co-learning loss function for the image modality, respectively. $\mathcal{L}^H_{total}$, $\mathcal{L}^H_S$, and $\mathcal{L}^H_{CL}$ are the total loss function, the supervised loss function, and the co-learning loss function for the DSM modality, respectively. λ_1 and λ_2 are the hyperparameters to weigh the supervised loss function and the co-learning loss function.

Algorithm 1 Training Phase of the Proposed Change Detection Co-learning Method

Input: $\mathbb{D}_s$, $\mathbb{D}_t$
Output: W_I, W_H
Initialize W_I, W_H
while $n < N$ **do**
  Part 1: Learning with labeled source samples
  (1) Randomly sample B labeled source data samples $\{\{I^s_1, I^s_2\}, \{H^s_1, H^s_2\}, G^s\}$ from the source dataset $\mathbb{D}_s$.
  (2) Forward pass:
    $P^s_I \leftarrow f_I(I^s_1, I^s_2)$
    $P^s_H \leftarrow f_H(H^s_1, H^s_2)$
  (3) Calculate supervised loss:
    $\mathcal{L}^I_S \leftarrow L_S(G^s||P^s_I)$
    $\mathcal{L}^H_S \leftarrow L_S(G^s||P^s_H)$

  Part 2: Learning with unlabeled target samples
  (1) Randomly sample B unlabeled target data samples $\{\{I^t_1, I^t_2\}, \{H^t_1, H^t_2\}\}$ from the target dataset $\mathbb{D}_t$.
  (2) Forward pass:
    $P^t_I \leftarrow f_I(I^t_1, I^t_2)$
    $P^t_H \leftarrow f_H(H^t_1, H^t_2)$
  (3) Calculate co-learning loss:
    $\mathcal{L}^I_{CL} \leftarrow L_C(P^t_I, P^t_{H,d})$
    $\mathcal{L}^H_{CL} \leftarrow L_C(P^t_H, P^t_{I,d})$

  Part 3: Backward propagation and updating network parameters
  (1) Calculate total loss:
    $\mathcal{L}^I_{total} \leftarrow \lambda_1\mathcal{L}^I_S + \lambda_2\mathcal{L}^I_{CL}$
    $\mathcal{L}^H_{total} \leftarrow \lambda_1\mathcal{L}^H_S + \lambda_2\mathcal{L}^H_{CL}$
  (2) Backward pass:
    Calculate the backward pass for the image change detection network.
    Calculate the backward pass for the DSM change detection network.
  (3) Update: W_I, W_H
end while
Return W_I, W_H

Algorithm 1 presents how the proposed framework is implemented. During the training phase, each iteration consists of two groups of forward pass operations, with separate operations for the image and DSM networks. The first group of forward pass uses the labeled source samples, contributing to the supervised loss functions. The second group employs unlabeled target samples and contributes to the co-learning loss functions. The backward pass operations employ the total loss functions. At the end of each iteration, the parameters of

the image network W_I and the DSM network W_H are updated with the help of the optimizer.

C. Siamese ResNet with Bitemporal Image Transformer Layer

Considering the balance between the network depth and GPU memory, we employ the ResNet-50 convolutional network [88] in a Siamese structure as the encoder and a bitemporal image transformer (BIT) module [30] at the bottleneck to refine the original bitemporal image features. In general, this architecture consists of three steps. (1) Employ a ResNet-50 backbone as the encoder, extracting initial features from pre-event and post-event images. (2) Use a BIT module to refine the initial features. (3) Fuse refined features by the subtraction operation and utilize an elegant change classifier to convert fused features to change maps. Fig. 2 presents the architecture of the Siamese image network ResNet-50-BIT. We use the ResNet-50 encoder to replace the ResNet-18 implemented by [30], so the image encoder can extract more robust features with the help of deeper structure [88]. In addition, we apply a small change classifier to control the size of the model and make sure it can be successfully run on an 11 GB RTX 2080 Ti GPU.

D. Transformer-based UNet for HDiff Maps

In this method, f_H contains two steps: (1) generate HDiff maps and (2) apply the HDiff network to process HDiff maps. As HDiff rasters have 3D information of coordinates X, Y, and $\triangle Z$, there are two main approaches to processing them. One is to process them as point clouds [51], [58] with 3D neural networks, while the other is to process them as 2D rasters and the height difference values $\triangle Z$ are utilized as input channels to a 2D network. Considering that the height difference values in different cities typically fall within a certain range and 2D networks are usually more efficient than point cloud networks with the same scales [89], in this study we employed a 2D SwinTransformer-based [42] U-shape network (SwinTransUNet) as the processing branch for the HDiff maps. Fig. 3 presents the architecture of our HDiff map network SwinTransUNet. As it shows, the encoder is conducted with Swin Transformer and patch merging blocks, generating multiscale features with a hierarchical structure, which has a good capability to capture global features. A U-Net structure is utilized as the decoder, so different scales of features can be utilized more efficiently. To control the computational cost and GPU memory usage, the dimensionality reduction blocks and upsampling blocks of the decoder are based on convolution and transposed convolution operations, respectively. Therefore, our HDiff network can also be trained and tested on a relatively cheaper GPU with lower memory such as an 11 GB RTX 2080 Ti.

E. Robust Height Difference

Due to limited resolution, illumination distortion, and cloud cover, the matching quality of spaceborne images is often limited, resulting in unsatisfactory quality of DSMs [83], [90]. These DSMs, along with generated HDiff maps obtained through direct pixel-wise subtraction, tend to contain numerous unexpected outlier pixels. Such outliers can adversely affect the performance of classification algorithms, such as building extraction or change detection. To address the noise issue and improve the quality of the HDiff map, a robust difference method is proposed by [83].

The robust difference between bitemporal DSM H_1 and DSM H_2 for the pixel (i, j) is defined as the minimum of differences calculated with the pixel (i, j) in the post-DSM and a certain neighborhood (with windows size $2 \times w + 1$) of the pixel $H_1(i, j)$ in the pre-DSM. The robust positive and negative differences $Diff_P^H(i, j)$ and $Diff_N^H(i, j)$ with respect to the pixel (i, j) are defined in following equations:

$$Diff_P^H(i,j) = \begin{cases} \min\limits_{p,q}\{H_2(i,j) - H_1(p,q)\}, & x_2(i,j) - x_1(p,q) > 0 \\ 0, & x_2(i,j) - x_1(p,q) \leq 0 \end{cases} \tag{15}$$

$$Diff_N^H(i,j) = \begin{cases} 0, & x_2(i,j) - x_1(p,q) \geq 0 \\ \max\limits_{p,q}\{H_2(i,j) - H_1(p,q)\}, & x_2(i,j) - x_1(p,q) < 0 \end{cases} \tag{16}$$

where $p \in [i - w, i + w]$ and $q \in [j - w, j + w]$ in a squared window around the pixel (i, j). This operation only takes the minimum value (greater than zero) of the positive change, or the maximum value of the negative change within the defined window region. Noisy outliers can be effectively eliminated from the original height difference map.

In this work, we only consider building change or non-change. Therefore, we utilize a combined binary robust difference map $Diff_R^H(i, j)$ including both positive and negative differences, which is computed as follows:

$$Diff_R^H(i,j) = Diff_P^H(i,j) + Diff_N^H(i,j)\ , \tag{17}$$

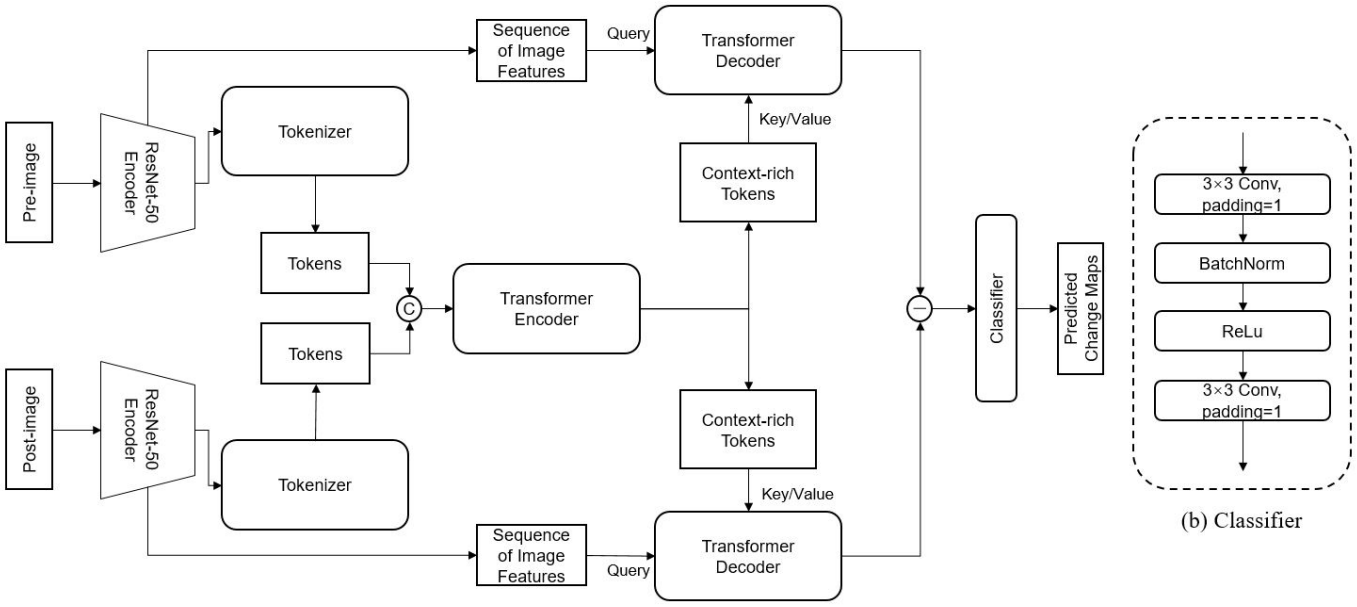
F. Loss Functions

Our framework employs two categories of loss functions in each training phase. First, a pixel-wise supervised loss function is used in the labeled source data for the purpose of change detection. Second, an unsupervised loss function is applied to the unlabeled target data.

1) The loss function for supervised change detection: Change detection is a pixel-wise classification task. Therefore, we employ cross-entropy as the supervised loss function, denoted as:

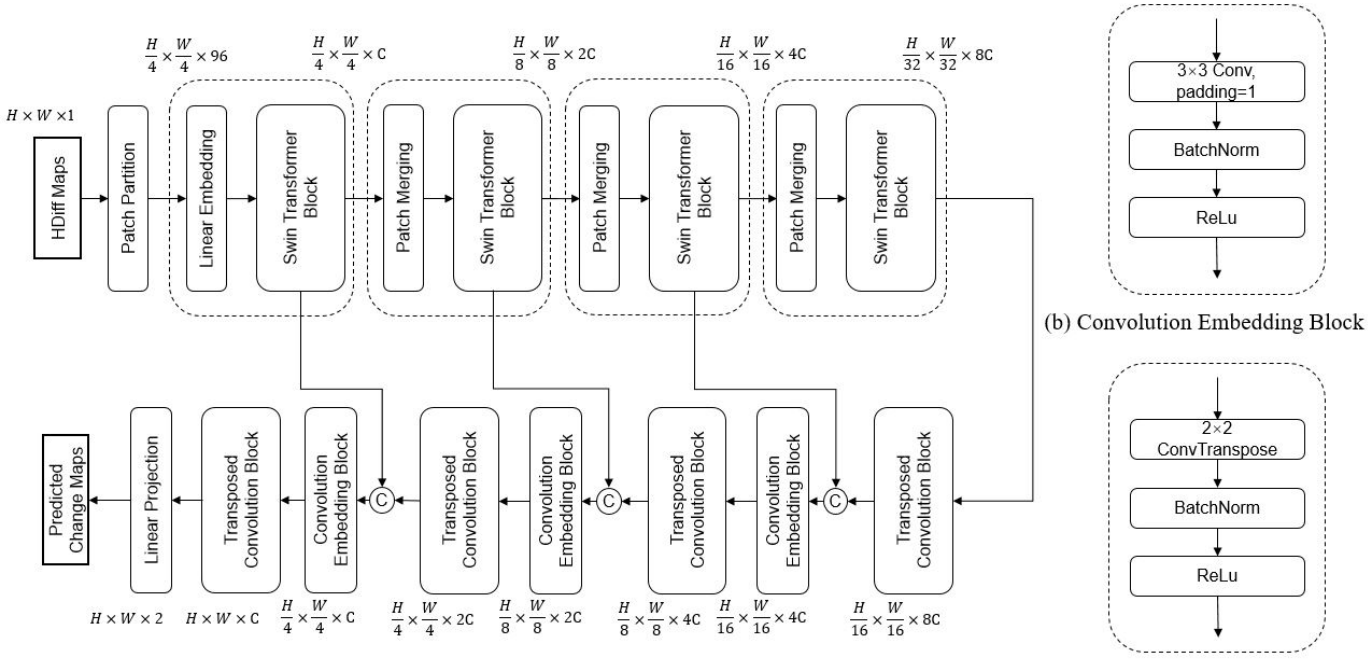
$$\begin{aligned} \mathcal{L}_S(G^s||P_I^s) &= CE(G^s||P_I^s) \\ &= \sum_{x \in \mathcal{X}} G^s(x) \log P_I^s(x)\ , \end{aligned} \tag{18}$$

where G^s and P_I^s are defined on the same probability space $\mathcal{X}$. G^s is the distribution of the source domain's ground truths. P_I^s is the predicted probability distribution of the image modality from the source domain. This is the supervised change detection loss applied for the image modality.



(a) The architecture of the Siamese image network ResNet-50-BIT

Fig. 2. (a) The architecture of the Siamese image network ResNet-50-BIT. (b) The classifier block. The modules of the tokenizer, transformer encoder, and transformer decoder are forked from the official implementation of [30] https://github.com/justchenhao/BIT_CD.



(a) The architecture of the HDiff network SwinTransUnet

(b) Convolution Embedding Block

(c) Transposed Convolution Block

Fig. 3. (a) The architecture of the proposed HDiff network SwinTransUNet. (b) Convolution Embedding Block. (c) Transposed Convolution Block. The swin transformer encoder modules are forked from the official implementation of [42] https://github.com/microsoft/Swin-Transformer.

In the same way, the supervised loss function for the DSM modality is

$$\begin{aligned}\mathcal{L}_S(G^s||P_H^s) &= CE(G^s||P_H^s) \\ &= \sum_{\hat{x}\in\mathcal{X}} G^s(\hat{x}) \log P_H^t(\hat{x}) \ , \end{aligned} \tag{19}$$

where P_H^s is the predicted probability distribution of the DSM modality from the source domain.

2) Loss functions for unsupervised multimodal co-learning: In this work, two kinds of loss functions, KL divergence and mean square error (MSE), are adopted as the co-learning loss function.

As presented in section III-B, each loss function can be integrated into our framework and generate three co-learning combinations. It is possible for certain loss functions to result in equivalent combinations, which have identical effects during backpropagation and updating parameters. Appendix A outlines a method for determining whether $\mathcal{L}_{CL-V}$, $\mathcal{L}_{CL-F}$, and $\mathcal{L}_{CL-DF}$ are equivalent. Appendix B presents the derivation for KL divergence and MSE loss functions. According to its conclusion, when KL divergence is employed as L_C, $\mathcal{L}_{CL-V}$, $\mathcal{L}_{CL-F}$, and $\mathcal{L}_{CL-DF}$ are inequivalent. So they are three different methods. When MSE is employed as L_C, $\mathcal{L}_{CL-V}$ and $\mathcal{L}_{CL-F}$ are equivalent. Therefore, only Vanilla co-learning and detached fusion co-learning are reported for the MSE-based experimental results in the following text.

IV. Experiments

A. Datasets

1) Simulated Multimodal Aerial Remote Sensing (SMARS) dataset: SMARS[1] is a recently published synthetic aerial remote sensing dataset by the German Aerospace Center (DLR) and the International Society for Photogrammetry and Remote Sensing (ISPRS) [72]. This dataset is designed for multimodal urban semantic segmentation, building extraction, and building change detection tasks. Its feasibility of being employed as a benchmark for algorithm training and evaluation has been proven [72]. It consists of two sub-datasets with distinct urban styles. One simulated city is named Synthetic Paris (SParis). The other is named Synthetic Venice (SVenice). Each sub-dataset includes bitemporal orthophotos, bitemporal photogrammetric DSMs, corresponding semantic maps, and corresponding building change maps. SMARS provides two versions, with resolutions of 30cm and 50cm, respectively. In this work, we employ the version of 50cm to evaluate the co-learning-based cross-domain change detection experiments. The training, validation, and testing raster sizes of the 50cm-SParis dataset are 1518×3560 pixels, 1008×3560 pixels, and 1974×3560 pixels, respectively. The training, validation, and testing raster sizes of the 50cm-SVenice dataset are 2800×5600, 2800×2128, and 2800×3472, respectively. Based on SParis and SVenice data, two groups of cross-domain experiments are conducted in this work: (1) SParis→SVenice: SParis used for training, SVenice for testing, and (2) SVenice→SParis: SVenice used for training, SParis for testing.

[1]https://www2.isprs.org/commissions/comm1/wg8/benchmark_smars/

2) Istanbul WorldView-2 dataset: The Istanbul WorldView-2 dataset is a building change detection dataset covering two areas of Istanbul, Türkiye with a GSD of 50 cm. This dataset consists of 100 pairs of bitemporal orthophotos with RGB channels and photogrammetric DSMs from 2011 and 2012 and the corresponding building change ground truth annotated by hand. The orthophotos and photogrammetric DSMs are generated from stereo WorldView-2 satellite images using the improved semi-global matching approach [90], [91]. Each patch has a pixel size of 400×400. In this work, the Istanbul WorldView-2 dataset is used as the testing data in a series of synthetic→real experiments, of which the training set is the SMARS dataset.

Fig. 4 presents samples of the SMARS dataset and Istanbul WorldView-2 dataset.

B. Experiment Setup

Our experiments are carried out based on the PyTorch framework [92]. Single-modal baseline models are trained and tested on a Geforce RTX 2080 Ti GPU with 11 GB RAM. The co-learning experiments are performed on two Geforce RTX 2080 Ti GPUs, one of which is used for training the Siamese network for bitemporal images, while the other is used for training the HDiff map network. In implementing the ResNet-50-BIT network, the token length, decoder depth, and dimension of heads are set to 4, 8, and 16, respectively. In the settings of HDiff SwinTransUNet, the depths of 4 layers in the encoder are 2, 2, 18, and 2, and the number of attention heads of each layer is 3, 6, 12, and 24 respectively. The token size of each patch is 4. The size of the windows is set to 12. In the training phase, we adopt the Adam optimizer with a learning rate of 0.001. The training batch size is 3. All models are trained for 30 epochs, which indicates a complete pass through the labeled source training dataset. Considering different methods may rely on different weights for the co-learning functions, we report the best results from cases with experience values $\lambda_2 = 0.1$ and 1.0. λ_1 remains equal to 1.0. Considering the 400×400 size of the Istanbul WorldView-2 patches, the training data of SMARS dataset are cropped to the patches with the same size and an overlap of 200 pixels. SParis and SVenice training sets consist of 96 and 351 training patches, respectively.

We test two co-learning loss functions and three types of co-learning combinations in our experiments. To quantitatively evaluate the performance of different methods, we employed $F1$ and intersection over union (IoU) scores as the primary evaluation metrics. In order to better demonstrate the confusion between changed and unchanged pixels, *precision* and *recall* are also reported in our work. These metrics are calculated according to the following equations:

$$F1 = \frac{2TP}{2TP + FP + FN} \ , \tag{20}$$

$$IoU = \frac{TP}{TP + FP + FN} \ , \tag{21}$$

$$Precision = \frac{TP}{TP + FP} \ , \tag{22}$$

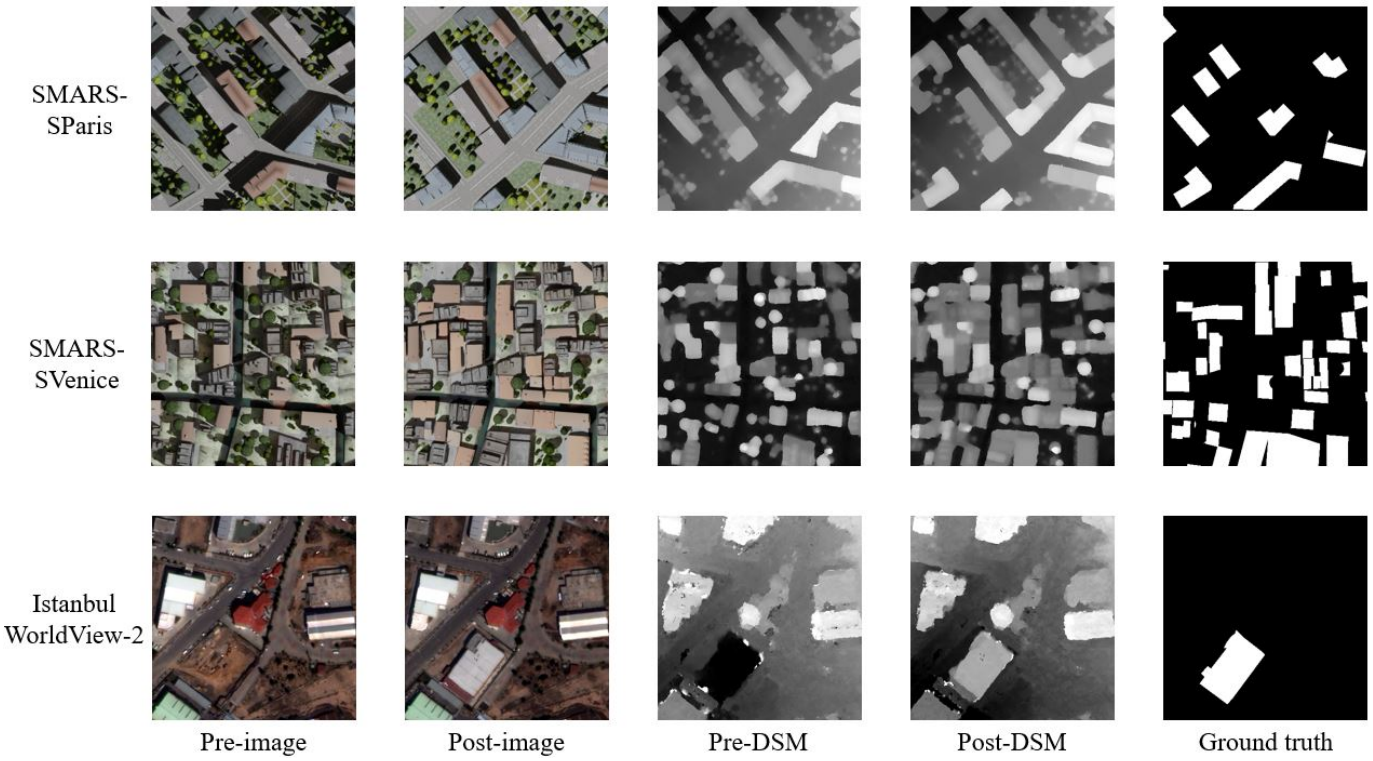


Fig. 4. Samples of SParis dataset and SVenice dataset of SMARS, as well as the Istanbul WorldView-2 dataset.

$$Recall = \frac{TP}{TP + FN}\ , \tag{23}$$

where TP denotes the number of true positives, TN the true negatives, FP the false positives, and FN the false negatives.

C. Experiment I: Domain Adaptation with Synthetic Data

As mentioned in section IV-A1, this experiment includes two parts: SParis→SVenice and SVenice→SParis.

Table I presents the qualitative results of SParis→SVenice. All co-learning combinations with either KL divergence or MSE loss functions can achieve significant improvement in the image network compared with the baseline. The network trained using co-learning with detached fusion strategy and the MSE loss function achieves the highest IoU and F1 scores, with an improvement of 62.19% on IoU and 63.97% on F1, compared with the baseline method by single-modal learning. In the results achieved by the HDiff network, the best quantitative results are obtained by the co-learning-enhanced network optimized by the MSE-based CL-V loss, of which the IoU is 71.71% and the F1 score is 83.52%.

Among the results of SVenice→SParis experiments in Table II, the single-modal image network with bitemporal images has the poorest performance, with the IoU of 38.08% and the F1 of 55.16%. All reported co-learning combinations with two types of loss functions are able to improve the results. The best image modality result is achieved when applying detached fusion co-learning and using the MSE as the co-learning loss, leading to an IoU of 88.04% and F1 of 93.64%. The HDiff network SwinTransUNet can also benefit from co-learning in this case. The method detached fusion co-learning (KL divergence as the loss) achieves an increase of 2.71% on IoU and 1.52% on F1.

Fig. 5 shows the qualitative results of SParis→SVenice. From the given examples, the baseline bitemporal method employing ResNet-50-BIT struggles to effectively identify building changes in both images and DSMs. In fact, no single changed building is fully detected. When using the baseline method to process HDiff maps, reasonable results can be achieved. However, numerous false positive pixels still exist as highlighted with green color. With the help of co-learning, the performance of the bitemporal network ResNet-50-BIT is significantly better on the target domain images. At the same time, the performance of the HDiff network is also enhanced on the HDiff maps. Compared with the baseline single-modal method, the HDiff network trained with co-learning approaches generates fewer false negatives.

The results of SVenice→SParis shown in Fig. 6 are similar to what happens in SParis→SVenice, which also demonstrates the effectiveness of the proposed co-learning approaches. All methods (co-learning or single-modal learning) in the case of SVenice→SParis yield better results than in the SParis→SVenice case. It can be explained by the higher building diversities (sizes and shapes) of SVenice, which are conducive to the robustness and generalizability of models [72].

Single-modal HDiff baseline method achieves much better results than the single-modal bitemporal image baseline method. Yet, image networks possess greater improvement potential when co-learning is applied. HDiff network is more

prone to generate more false positive pixels, as shown in example A in Fig. 5 and example B in Fig. 6. Since the HDiff network is designed to detect the shapes with certain height differences in the HDiff map, some non-man-made object changes have similar geometric features with changes in buildings and are therefore wrongly recognized. In Fig. 5 A, a noticeable false positive object of round shape at the left border of all results by the HDiff network is the change of a tree rather than a building. In the results of image-based methods, however, only the network trained with the KL-CL-F strategy makes the same mistake.

TABLE I
QUANTITATIVE RESULTS OF DIFFERENT METHODS ON THE EXPERIMENT SPARIS→SVENICE. THE BEST SCORE IS SHOWN IN BOLD.

Modality	Methods		Precision	Recall	F1	IoU
Image	Baseline		40.92	14.19	21.07	11.78
	KL	CL-V	**91.97**	65.17	76.29	61.66
		CL-F	83.49	66.11	73.79	58.47
		CL-DF	86.59	76.49	81.23	68.39
	MSE	CL-V	86.46	83.14	84.77	73.56
		CL-DF	86.15	**83.96**	**85.04**	**73.97**
DSM	Baseline (Siamese)		55.95	30.51	24.60	39.48
	Baseline (HDiff)		84.37	75.44	79.66	66.20
	KL	CL-V	78.88	88.15	83.26	71.32
		CL-F	81.35	85.02	83.15	71.16
		CL-DF	74.90	**90.96**	82.16	69.71
	MSE	CL-V	81.04	86.17	**83.52**	**71.71**
		CL-DF	**84.11**	82.07	83.08	71.05

TABLE II
QUANTITATIVE RESULTS OF DIFFERENT METHODS ON THE EXPERIMENT SVENICE→SPARIS. THE BEST SCORE IS SHOWN IN BOLD.

Modality	Methods		Precision	Recall	F1	IoU
Image	Baseline		82.97	41.31	55.16	38.08
	KL	CL-V	95.99	89.35	92.55	86.13
		CL-F	**99.19**	83.21	90.50	82.64
		CL-DF	97.91	**89.64**	93.59	87.96
	MSE	CL-V	97.35	89.41	93.21	87.29
		CL-DF	98.63	89.13	**93.64**	**88.04**
DSM	Baseline (Siamese)		54.64	35.25	42.85	27.27
	Baseline (HDiff)		94.10	92.26	93.17	87.21
	KL	CL-V	93.53	92.70	94.10	88.85
		CL-F	97.28	90.32	93.66	87.86
		CL-DF	95.55	**93.85**	**94.69**	**89.92**
	MSE	CL-V	96.03	93.08	94.53	89.62
		CL-DF	**97.65**	91.79	94.63	89.81

D. Experiment II: SMARS→stanbul WorldView-2

In this experimental case, we adopt the full 50cm-SMARS training data (including both SParis and SVenice) as the source data and Istanbul WorldView-2 patches as the target data. Additionally, to verify whether robust height difference can improve building change detection results, two groups of comparison experiments are presented. One group utilizes the direct height difference operation to generate the HDiff maps for Istanbul data, marked with a red D in Table III and the following text. The other employs the robust height difference method to calculate optimized HDiff maps for Istanbul data, marked with a blue R in Table III and the following text.

TABLE III
QUANTITATIVE RESULTS OF DIFFERENT METHODS ON THE EXPERIMENT SMARS→ISTANBUL.

Modality	Methods		Precision	Recall	F1	IoU
Image	Baseline		7.95	3.21	4.57	2.34
	KL	CL-V (D)	89.67	65.09	75.43	60.55
		CL-F (D)	83.80	57.33	68.08	51.61
		CL-DF (D)	85.03	64.04	73.06	57.55
		CL-V (R)	92.27	63.03	74.90	59.87
		CL-F (R)	80.43	59.79	68.59	52.20
		CL-DF (R)	86.61	70.11	77.49	63.25
	MSE	CL-V (D)	86.89	69.08	76.97	62.56
		CL-DF (D)	87.32	68.27	76.63	62.11
		CL-V (R)	84.71	74.51	79.29	65.68
		CL-DF (R)	87.34	72.32	79.12	65.46
DSM	Baseline (Siamese)		40.33	27.83	32.94	19.71
	Baseline (D)		66.12	78.43	71.76	55.95
	Baseline (R)		74.41	72.93	73.67	58.31
	KL	CL-V (D)	81.09	70.93	75.67	60.86
		CL-F (D)	79.86	70.81	75.06	60.08
		CL-DF (D)	80.97	70.07	75.13	60.16
		CL-V (R)	77.11	76.55	76.83	62.38
		CL-F (R)	75.76	76.55	76.16	61.49
		CL-DF (R)	80.37	73.60	76.84	62.38
	MSE	CL-V (D)	78.64	74.59	76.56	62.02
		CL-DF (D)	78.37	76.64	77.49	63.26
		CL-V (R)	81.42	73.33	77.17	62.82
		CL-DF (R)	82.93	72.94	77.62	63.42

1) Co-learning with direct HDiff maps: As presented in Table III, the Siamese image baseline network ResNet-50-BIT trained with SMARS has abysmal performance on the unseen Istanbul dataset, in which only 4.57% of the F1 score and 2.34% of the IoU score are obtained. This performance can be attributed to the significant spectral domain gap between the synthetic images and real WorldView-2 images. The Siamese DSM baseline method also produces poor results, again demonstrating that the Siamese DSM approach has a poor generalization ability. By contrast, the baseline HDiff network can achieve reasonable results with either R or D.

With the help of co-learning, the performance of the image network is greatly improved. The best result by the Siamese image network is achieved with the MSE-CL-V co-learning variety, bringing up the F1 to 76.97% and the Iou to 62.56%. The HDiff network SwinTransUNet can also be enhanced by co-learning methods. All the results from different co-learning combinations are superior to the baseline change detection result of the HDiff map. Among them, the best result is achieved with the co-learning variety MSE-CL-DF, leading to a 12.25% higher precision, a 5.73% higher F1, and a 7.31% higher IoU compared with the baseline method.

2) Co-learning with robust HDiff maps: According to our past experience processing spaceborne DSMs [83], the window size for robust height difference is set to 5 (i.e. $w = 2$). The baseline results of R in Table III demonstrate the advantage of robust height difference in single-modal learning. In comparison to the baseline (D) using direct HDiff maps, baseline (R) employing robust HDiff maps achieves an increase of 1.91% and 2.36% on F1 and IoU, respectively.

With robust HDiff maps, all co-learning methods can also improve the performance of both the ResNet-50-BIT image network and the SwinTransUNet HDiff network. The MSE-CL-V co-learning variety achieves the best image modality

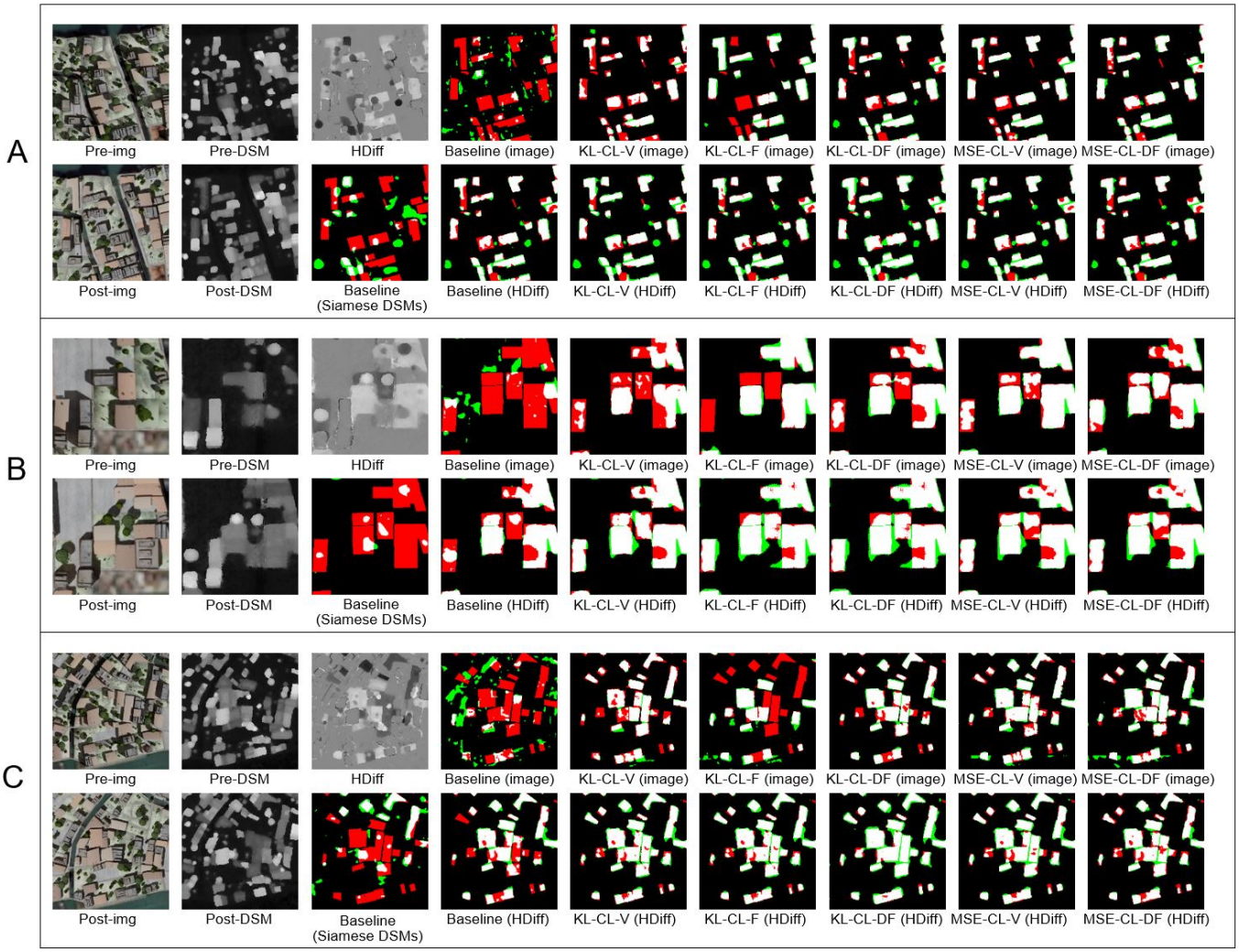


Fig. 5. Building change detection results of SParis→SVenice. Color legend: TP TN FN FP.

result, with an F1 score of 79.29% and an IoU score of 65.68%. For the DSM modality, the best result is achieved by MSE-CL-DF, with the F1 score of 77.62% and the IoU of 63.42%. In addition, each co-learning-enhanced HDiff network with robust HDiff maps yields better results compared with the same method utilizing direct HDiff maps. For the image modality, the best result achieved by MSE-CL-V (R) has a 2.32% higher F1 and a 3.12% higher IoU compared with MSE-CL-V (D). For the DSM modality the best result achieved by MSE-CL-DF (R) has a 0.13% higher F1 and a 0.16% higher IoU compared with MSE-CL-DF (D).

According to the visualization results presented in Fig. 7, Baseline (D) is more prone to generate obvious false positives due to the outlier values in direct HDiff maps, especially as exemplified by the green clusters in A and B. As the robust height difference approach can filter out a portion of such outliers, Baseline (R) (using the same model with Baseline (D)) results contain fewer false positive pixels. Whether using Direct HDiff maps or Robust HDiff maps, the co-learning training approaches lead to significant improvements in image results by ResNet-50-BIT and HDiff results by SwinTransUNet. In Fig. 7 A, the results of robust HDiff maps with co-learning varieties are superior to those of direct HDiff maps with the same approach. In the results of direct HDiff maps, more building change pixels are wrongly recognized as unchanged pixels. In the image results, similar phenomena can be observed. MSE-CL-V (D/image), which achieves the highest score among all co-learning varieties with direct HDiff maps, cannot recognize the change of a small building at the bottom border of A, while MSE-CL-V (R/image) is capable.

Nevertheless, applying robust HDiff maps may have negative effects in a few cases. For instance, in Fig. 7 C, the left building is an extension and only the extended part is defined as a building change in the ground truth. In the robust HDiff map, the height difference values of the narrow rectangular area are processed to the same values of its connected extended part. Therefore, the narrow rectangular area is completely recognized as a building change by SwinTransUNet. Even co-learning cannot correct this error. In this case, the image network trained with co-learning performs better, and MSE-

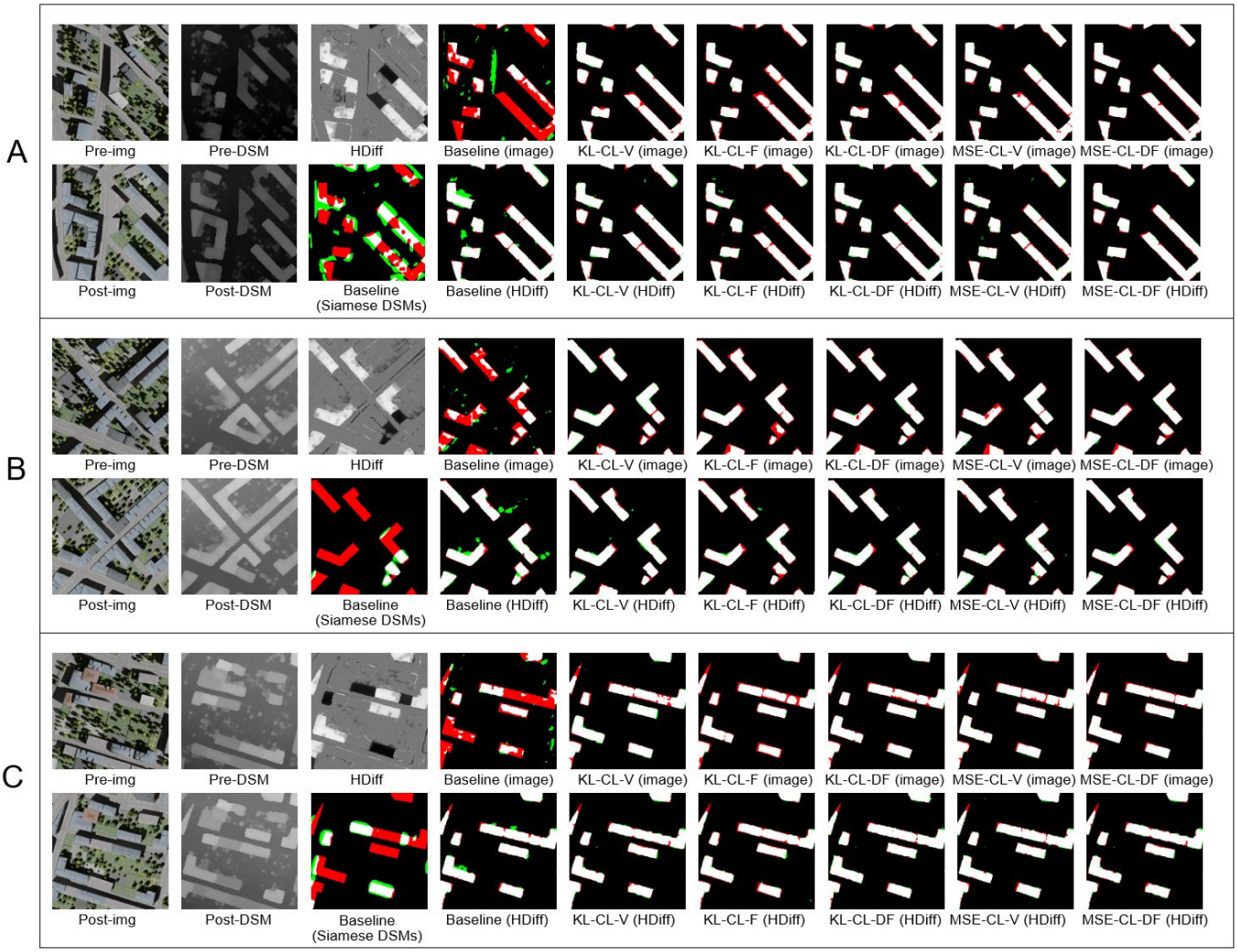


Fig. 6. Building change detection results of SVenice→SParis. Color legend: TP TN FN FP.

CL-V(R/image) correctly recognizes this area as a non-change area.

V. Discussion

A. Domain Gaps in Different Modalities

Due to the differences in imaging sensors, capturing conditions, and preprocessing operations for the raw data, the domain gaps of spectral distribution widely exist between different source and target datasets in remote sensing tasks [93]. Therefore, domain adaptation is becoming an essential topic.

This study presented the building change detection results of three baseline networks across three variants of two modalities: Siamese optical images, Siamese DSMs, and HDiff maps. Among the three paradigms, the HDiff maps demonstrate the most remarkable generalization ability in cross-domain scenarios. On the contrary, Siamese images and Siamese DSMs fail to produce reasonable results in our experiments. This phenomenon underscores the domain gap issues in these Siamese modalities, including synthetic→synthetic and synthetic→real cases, which is less pronounced in the HDiff maps for building change detection tasks. The superior cross-domain generalizability of HDiff maps can be attributed to its explicit geometric features, which excel in representing building changes. As a result, SwinTransUNet can learn robust knowledge and yield reasonable results in HDiff map single-modal learning mode. Nevertheless, the domain gaps of HDiff maps between different synthetic data and those between synthetic and real data are different. Since the two sub-datasets of SMARS focus on urban scenes and have similar building geometry, the domain gaps in HDiff maps between them are not significant. The baseline method for HDiff maps can yield commendable results, with the F1 score of 79.66% in SParis→SVenice and 93.17% in SVenice→SParis. As mentioned in section IV-C SVenice has a higher building diversity than SParis, which causes the main difference in building changes between these two sub-datasets. Larger domain gaps exist between SMARS and Istanbul datasets. First, Istanbul data are derived from space borne WorldView-2 data that are under the influence of real-world capturing conditions, which

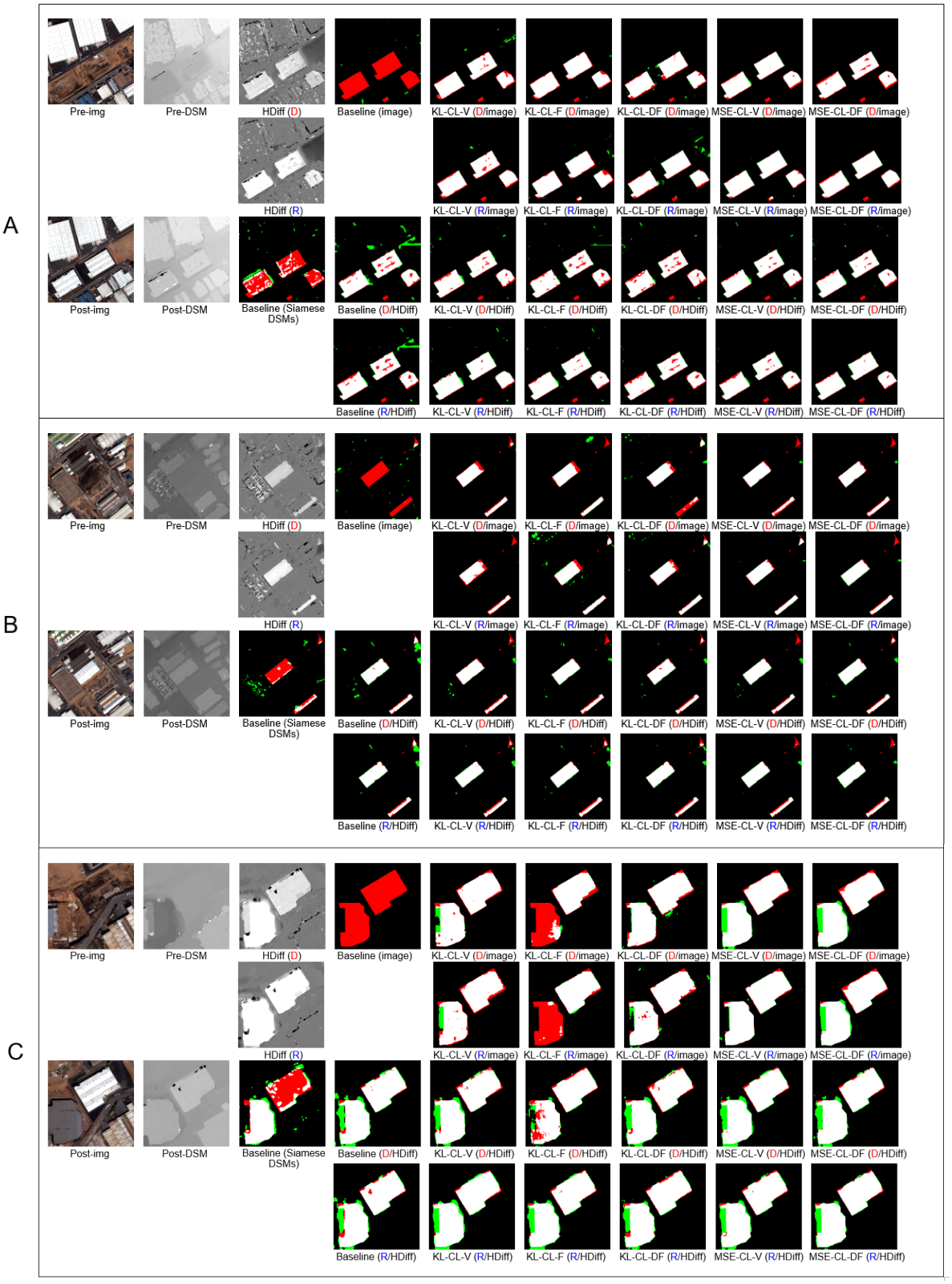


Fig. 7. Building change detection results of SMARS→Istanbul. Color legend: TP TN FN FP.

could also lead to variation in the quality of DSMs. Second, the Istanbul dataset encompasses not only urban scenes but also suburban industrial areas, where the building and the building change characteristics differ from those in the urban scenes of SMARS. The aforementioned points present the challenge for cross-domain experiments as exemplified by the case C in Fig. 7.

B. Co-learning-Enhanced Siamese Image Modality and HDiff Map Modality

As HDiff maps demonstrate superior generalization ability to Siamese DSMs, our co-learning experiments are conducted with the strong modality HDiff map and weaker modality Siamese images. An intuition is that the strong modality can assist the weaker modality's hidden feature map refinement with cross-modal learning [63]. Our experiments demonstrate that the performance of the Siamese image change detection branch ResNet-50-BIT can be significantly improved on the target data. In addition, the performance of the HDiff map network can be further boosted with the help of hidden knowledge from the Siamese image modality, which has very poor performance with single-modal learning. However, the Siamese image modality sometimes outperforms the HDiff map modality in the co-learning mode. As described in section IV, the co-learning-enhanced Siamese image network can accurately differentiate building changes and tree changes. It can even achieve higher evaluation metrics in experiments SParis→SVenice and SMARS→Istanbul. These promising results demonstrate that the proposed co-learning building change detection framework can boost the performance of each modality.

C. Multimodal Co-learning

Co-learning is a concept first proposed in the multimodal learning field [61], [62]. We follow the definition in papers [61] and [62]. Its main idea is to transfer mutual information/knowledge between different modalities with a consistency constraint, based on the intuition that the predictions from different modalities should be consistent when they are correct. In other words, the co-learning concept is based on maximizing the mutual information between the representations of the networks of different modalities.

Paper [51] classifies multimodal co-learning methods into standard and enhanced versions, depending on whether unlabeled training data are employed. Since the enhanced co-learning utilizes the mutual information of unlabeled multimodal target data, it is suitable for cross-domain tasks. In this work, the proposed co-learning framework is an enhanced variant. Due to its ability to mutually enhance the feature representation of the other modality, we do not employ any co-learning loss function between the two modalities of the labeled source data. Instead, the co-learning loss functions are only applied between the unlabeled target modalities. By doing so, overfitting on the source data is avoided and the performance on the target data is prioritized, which is conducive to cross-domain results. Self-training is another common method used for domain adaptation that exploits the pseudo-label of the unlabeled data, which is produced by the model trained with the labeled source data. Compared to one-off enhanced co-learning, self-training relies on extra operations [51]. Specifically, extra algorithms are needed to select proper samples with pseudo labels, and repeating training procedures is required [94], [95].

The co-learning framework is versatile and easily extendable, allowing for integration with other multimodal learning methodologies. Two recent studies have blended traditional data fusion with co-learning, specifically for multimodal semantic segmentation [63] and building extraction [96], respectively. Augmenting the co-learning framework with a variety of modules may well be a future trend.

D. Efficiency and Computational Complexity

Co-learning requires training the networks of two modalities in parallel. Compared with single-modal learning, it introduces more loss functions and corresponding data transfer (e.g., detached probabilities when calculating the co-learning loss functions) operations, increasing the time for training two networks. Table IV records the training time, the number of trainable parameters (#Params), and the floating point operations (FLOPs) of each variant for the experiment SMARS→Istanbul with robust HDiff operation. All models are trained for 30 epochs. The total time of training two baseline networks is 39 min 47 s. The training time for the co-learning method is between 55 min and 57 min, which is about 1.4× of the baseline training. According to Algorithm 1, the image network and HDiff (DSM) network are trained individually without adding extra layers and introducing more computational complexity. Consequently, the total number of trainable parameters and FLOPs in our proposed co-learning framework remains unchanged and is equivalent to the sum of those when training the individual networks.

In this work, we adopt a 2D rather than a 3D network to process HDiff maps, which is also due to efficiency considerations. 3D networks calculate deep features in a way that traverses in 3D space, which incurs more computing costs and longer training time than the corresponding 2D version. Furthermore, more 2D image networks are available compared to point cloud networks. The framework based on 2D networks has better extensibility for further applications.

TABLE IV
TRAINING TIME, THE NUMBER OF TRAINABLE PARAMETERS, AND GFLOPS OF DIFFERENT METHODS IN THE EXPERIMENT SMARS→ISTANBUL (R).

Methods		Training time	#Params	FLOPs
Baseline (image)		20 min	43.22M	61.86G
Baseline (HDiff)		19 min 47 s	57.85M	57.93G
KL	CL-V	55 min 41 s	43.22M + 57.9M	61.86G + 57.93G
	CL-F	55 min 49 s	43.22M + 57.9M	61.86G + 57.93G
	CL-DF	56 min 37 s	43.22M + 57.9M	61.86G + 57.93G
MSE	CL-V	55 min 21 s	43.22M + 57.9M	61.86G + 57.93G
	CL-DF	56 min 15 s	43.22M + 57.9M	61.86G + 57.93G

E. The Potential of Co-learning Framework in Real-world Applications

Utilizing the co-learning framework with bitemporal image and HDiff map modalities, four distinct models can be acquired: a single-modal Siamese image network, a single-modal HDiff map network, a co-learning-enhanced Siamese image network, and a co-learning-enhanced HDiff map network. This is especially useful when the training data and test data do not have the same modalities, which poses great constraints for the Siamese methods. In addition, the co-learning change detection framework is flexible to extend. As depicted in Fig. 1, besides the change detection backbones for images and HDiff maps (comprising encoders and decoders), modules like the fusion operation in the Siamese network, the height difference operation for DSMs, and co-learning loss functions can be tailored for specific scenarios.

Nowadays, multisource and crowdsourced data from other fields, like social media [97] and web-retrieved images [98], can provide additional information not available in remote sensing data. The co-learning framework also holds the potential for utilizing such data and enhancing the performance beyond the limitations of 2D/2.5D/3D remote sensing data. However, a main issue with this concept lies in the accurate alignment of these varied data sources [62].

Our proposed co-learning framework can be considered a form of semi-supervised learning. Semi-supervised learning is a branch of machine learning methods involving both labeled and unlabeled data [99], which is suitable for real scenarios of the remote sensing field with a large amount of unlabeled data. A key challenge existing in semi-supervised learning methods is that not all unlabeled data can achieve improvement in the neural network models. Unlabeled data is only useful if it provides information benefiting label prediction that is not contained in the labeled data alone [99]. As another way to employ unlabeled data, self-supervised learning pre-trains a model on a pretext task using unlabeled data, thereby providing a foundation for subsequent fine-tuning on downstream tasks [100]. This could be a strategy to enhance the utilization efficiency of unlabeled data and offer a contribution different from semi-supervised learning. Integrating a self-supervised learning phase is another potential direction to improve our framework, making it more applicable to real-world scenarios.

VI. Conclusion

In this paper, we proposed a multimodal co-learning framework for building change detection with cross-domain data. This framework effectively utilizes the labeled source data and unlabeled target data pairs, presenting a promising solution to improve the Siamese image and HDiff map building change detection networks when bitemporal orthophotos and corresponding DSMs are available. We designed three co-learning combinations within the framework: vanilla co-learning, fusion co-learning, and detached fusion co-learning. They all present improved performance compared with single-modal baselines with two loss functions: KL divergence and MSE. The experiments demonstrate that the proposed co-learning method can enhance the ability of a single-modal change detection network on target data, with the help of mutual knowledge from another modality. We also explore the potential of the newly published synthetic benchmark dataset SMARS by conducting two groups of experiments. Our investigations indicate that SMARS data especially DSMs can be adapted to train deep learning models for realistic datasets. Compared with direct height difference, robust height difference can reduce the gap between synthetic data and realistic WorldView-2 data and improve the cross-domain results.

In the future, we would like to investigate more multimodal learning methods for remote sensing tasks. Specifically speaking, we will make efforts in the following aspects: (1) explore more co-learning variants and more knowledge transfer approaches employing unlabeled data such as self-supervised learning [100], [101]. As a huge amount of existing remote sensing data are unlabeled, they are currently far from being effectively utilized [102]. (2) Involve more types of multimodal combinations with co-learning methods, e.g., hyperspectral images and DSMs. Hyperspectral data are popular in multimodal applications [49] but suffer from spectral variability [93], which could be alleviated by the geometric information from DSMs [103]. (3) Investigate more complex and specific types of domain gaps. For instance, resolution gaps widely exist in remote sensing tasks, limiting the interactions between lower- and higher-resolution data. To address this problem, we would like to integrate additional modules such as super resolution [104] into the co-learning framework.

Appendix A

Assume P_I^t is the change probability of the target image modality, P_H^t is the change probability of the target DSM modality. P_I^t and $P_H^t$are calculated by the forward propagation of the image network and DSM network, respectively:

$$P_I^t = W_I^T X_I^t + b_I \ , \tag{24}$$

$$P_H^t = W_H^T X_H^t + b_H \ , \tag{25}$$

Where X_I^t and X_H^t are the original input target data of images and DSMs, respectively. W_I^T and W_H^T are the weights. b_I and b_H are the bias.

Here we take the image modality as an example. As introduced in III-B, there are three types of co-learning combinations for modality image $\mathcal{L}_{CL-V}^I$, $\mathcal{L}_{CL-F}^I$, and $\mathcal{L}_{CL-DF}^I$. If L_C is a generic co-learning loss function, three co-learning combinations for modality image are calculated as follows.

(1) Vanilla co-learning, which is calculated as:

$$\mathcal{L}_{CL-V}^I = L_C(P_{H,d}^t || P_I^t) \ , \tag{26}$$

(2) Fusion co-learning, which is calculated as:

$$\mathcal{L}_{CL-F}^I = L_C(\frac{P_I^t + P_{H,d}^t}{2} || P_I^t) \ , \tag{27}$$

(3) Detached fusion co-learning, which is calculated as:

$$\mathcal{L}_{CL-DF}^I = L_C(\frac{P_{I,d}^t + P_{H,d}^t}{2} || P_I^t) \ , \tag{28}$$

The derivatives of $\mathcal{L}^I_{CL-V}$, $\mathcal{L}^I_{CL-F}$, and $\mathcal{L}^I_{CL-DF}$ with respect to X_I are:

$$\frac{\partial \mathcal{L}^I_{CL-V}}{\partial X^t_I} = \frac{\partial \mathcal{L}^I_{CL-V}}{\partial P^t_I}\frac{\partial P^t_I}{\partial X^t_I}\,, \tag{29}$$

$$\frac{\partial \mathcal{L}^I_{CL-F}}{\partial X^t_I} = \frac{\partial \mathcal{L}^I_{CL-F}}{\partial P^t_I}\frac{\partial P^t_I}{\partial X^t_I}\,, \tag{30}$$

$$\frac{\partial \mathcal{L}^I_{CL-DF}}{\partial X^t_I} = \frac{\partial \mathcal{L}^I_{CL-DF}}{\partial P^t_I}\frac{\partial P^t_I}{\partial X^t_I}\,, \tag{31}$$

If $\frac{\partial \mathcal{L}^I_{CL-V}}{\partial P^t_I} \neq \alpha\frac{\partial \mathcal{L}^I_{CL-F}}{\partial P^t_I}$, $\frac{\partial \mathcal{L}^I_{CL-V}}{\partial P^t_I} \neq \beta\frac{\partial \mathcal{L}^I_{CL-DF}}{\partial P^t_I}$, and $\frac{\partial \mathcal{L}^I_{CL-F}}{\partial P^t_I} \neq \gamma\frac{\partial \mathcal{L}^{DI}_{CL-F}}{\partial P^t_I}$ $(\alpha,\beta,\gamma \neq 0)$, above three co-learning loss combinations are different respect to X^t_I. They can be regarded as three inequivalent methods. The co-learning loss combinations of DSM modality can be evaluated in the same way.

APPENDIX B

We use the case of image modality as an example. The situation of DSM modality can be calculated in the same way.'

A. KL-divergence

When KL-divergence is employed as the co-learning loss function, L^I_C for image modality is as follows:

$$L^I_C = P^I_S \ln \frac{P^I_S}{P^t_I}\,, \tag{32}$$

where P^I_S is the shadow reference probability of the image modality. $P^I_S \in \{P^t_{H,d}, \frac{P^t_I+P^t_{H,d}}{2}, \frac{P^t_{I,d}+P^t_{H,d}}{2}\}$, depending on which co-learning combination is employed.

We use the rule in A to evaluate the equivalence of three co-learning combinations, $\mathcal{L}^I_{CL-V}$, $\mathcal{L}^I_{CL-F}$, and $\mathcal{L}^I_{CL-DF}$:

(1) Vanilla co-learning:

$$\mathcal{L}^I_{CL-V} = P^t_{H,d} \ln \frac{P^t_{H,d}}{P^t_I}\,, \tag{33}$$

so,

$$\begin{aligned}\frac{\partial \mathcal{L}^I_{CL-V}}{\partial X^t_I} &= \frac{\partial \mathcal{L}^I_{CL-V}}{\partial P_I}\frac{\partial P^t_I}{\partial X^t_I} \\ &= \frac{\partial P^t_{H,d}\ln\frac{P^t_{H,d}}{P^t_I}}{\partial P^t_I}\frac{\partial P^t_I}{\partial X^t_I} \\ &= -\frac{P^t_{H,d}}{P^t_I}\frac{\partial P^t_I}{\partial X^t_I}\,,\end{aligned} \tag{34}$$

(2) Fusion co-learning

$$\mathcal{L}^I_{CL-F} = \frac{P^t_I + P^t_{H,d}}{2}\ln\frac{P^t_I + P^t_{H,d}}{2P^t_I}\,, \tag{35}$$

so,

$$\begin{aligned}\frac{\partial \mathcal{L}^I_{CL-F}}{\partial X^t_I} &= \frac{\partial \mathcal{L}^I_{CL-F}}{\partial P^t_I}\frac{\partial P^t_I}{\partial X^t_I} \\ &= \frac{\partial \frac{P^t_I+P^t_{H,d}}{2}\ln\frac{P^t_I+P^t_{H,d}}{2P^t_I}}{\partial P^t_I}\frac{\partial P^t_I}{\partial X^t_I} \\ &= \frac{1}{2}[\ln(P^t_{H,d}+P^t_I) - \ln P^t_{H,d} \\ &\quad - \frac{P^t_{H,d}}{P^t_I} - \ln 2]\frac{\partial P^t_I}{\partial X^t_I}\,,\end{aligned} \tag{36}$$

(3) Detached fusion co-learning

$$\mathcal{L}^I_{CL-DF} = \frac{P^t_{I,d}+P^t_{H,d}}{2}\ln\frac{P^t_{I,d}+P^t_{H,d}}{2P^t_I}\,, \tag{37}$$

so,

$$\begin{aligned}\frac{\partial \mathcal{L}^I_{CL-DF}}{\partial X_I} &= \frac{\partial \mathcal{L}^I_{CL-DF}}{\partial P_I}\frac{\partial P_I}{\partial X_I} \\ &= \frac{\partial \frac{P^t_{I,d}+P^t_{H,d}}{2}\ln\frac{P^t_{I,d}+P^t_{H,d}}{2P^t_I}}{\partial P^t_I}\frac{\partial P^t_I}{\partial X^t_I} \\ &= -\frac{P^t_{I,d}+P^t_{H,d}}{2P^t_I}\frac{\partial P^t_I}{\partial X^t_I}\,,\end{aligned} \tag{38}$$

As $\frac{\partial \mathcal{L}^I_{CL-V}}{\partial P^t_I} \neq \alpha\frac{\partial \mathcal{L}^I_{CL-F}}{\partial P^t_I}$, $\frac{\partial \mathcal{L}^I_{CL-V}}{\partial P^t_I} \neq \beta\frac{\partial \mathcal{L}^I_{CL-DF}}{\partial P^t_I}$, and $\frac{\partial \mathcal{L}^I_{CL-F}}{\partial P^t_I} \neq \gamma\frac{\partial \mathcal{L}^I_{CL-DF}}{\partial P^t_I}$ $(\alpha,\beta,\gamma \neq 0)$, KL divergence-based $\mathcal{L}^I_{CL-V}$, $\mathcal{L}^I_{CL-F}$, and $\mathcal{L}^I_{CL-DF}$ are inequivalent and they are three different co-learning methods.

B. MSE

When MSE is employed as the co-learning loss function, L^I_C for image modality is as follows:

$$L^I_C = |P^t_I - P^I_S|^2\,, \tag{39}$$

where P^I_S is the shadow reference probability of the image modality. $P^I_S \in \{P^t_{H,d}, \frac{P^t_I+P^t_{H,d}}{2}, \frac{P^t_{I,d}+P^t_{H,d}}{2}\}$, depending on which co-learning combination is employed.

We use the rule in A to evaluate the equivalence of three co-learning combinations, $\mathcal{L}^I_{CL-V}$, $\mathcal{L}^I_{CL-F}$, and $\mathcal{L}^I_{CL-DF}$:

(1) Vanilla co-learning:

$$\mathcal{L}^I_{CL-V} = |P^t_I - P^t_{H,d}|^2\,, \tag{40}$$

so,

$$\begin{aligned}\frac{\partial \mathcal{L}^I_{CL-V}}{\partial X^t_I} &= \frac{\partial \mathcal{L}^I_{CL-V}}{\partial P_I}\frac{\partial P^t_I}{\partial X^t_I} \\ &= \frac{\partial |P^t_I - P^t_{H,d}|^2}{\partial P^t_I}\frac{\partial P^t_I}{\partial X^t_I} \\ &= 2(P^t_I - P^t_{H,d})\frac{\partial P^t_I}{\partial X^t_I}\,,\end{aligned} \tag{41}$$

(2) Fusion co-learning

$$\begin{aligned}\mathcal{L}^I_{CL-F} &= |P^t_I - \frac{P^t_I + P^t_{H,d}}{2}|^2 \\ &= \frac{|P^t_I - P^t_{H,d}|^2}{4}\,,\end{aligned} \tag{42}$$

so,

$$\begin{aligned}\frac{\partial \mathcal{L}^I_{CL-F}}{\partial X^t_I} &= \frac{\partial \mathcal{L}^I_{CL-F}}{\partial P^t_I}\frac{\partial P^t_I}{\partial X^t_I} \\ &= \frac{\partial \frac{|P^t_I - P^t_{H,d}|^2}{4}}{\partial P^t_I}\frac{\partial P^t_I}{\partial X^t_I} \\ &= \frac{P^t_I - P^t_{H,d}}{4}\frac{\partial P^t_I}{\partial X^t_I}\ ,\end{aligned} \tag{43}$$

(3) Detached fusion co-learning

$$\mathcal{L}^I_{CL-DF} = |P^t_I - \frac{P^t_{I,d} + P^t_{H,d}}{2}|^2\ , \tag{44}$$

so,

$$\begin{aligned}\frac{\partial \mathcal{L}^I_{CL-DF}}{\partial X_I} &= \frac{\partial \mathcal{L}^I_{CL-DF}}{\partial P_I}\frac{\partial P_I}{\partial X_I} \\ &= \frac{\partial |P^t_I - \frac{P^t_{I,d}+P^t_{H,d}}{2}|^2}{\partial P^t_I}\frac{\partial P^t_I}{\partial X^t_I} \\ &= (2P^t_I - P^t_{I,d} - P^t_{H,d})\frac{\partial P^t_I}{\partial X^t_I}\ ,\end{aligned} \tag{45}$$

As $\frac{\partial \mathcal{L}^I_{CL-V}}{\partial P^t_I} = 4 \cdot \frac{\partial \mathcal{L}^I_{CL-F}}{\partial P^t_I}$, $\frac{\partial \mathcal{L}^I_{CL-V}}{\partial P^t_I} \neq \beta\frac{\partial \mathcal{L}^I_{CL-DF}}{\partial P^t_I}$, and $\frac{\partial \mathcal{L}^I_{CL-F}}{\partial P^t_I} \neq \gamma\frac{\partial \mathcal{L}^I_{CL-DF}}{\partial P^t_I}$ $(\beta, \gamma \neq 0)$, MSE-based $\mathcal{L}^I_{CL-V}$ and $\mathcal{L}^I_{CL-F}$ are equivalent. MSE-based $\mathcal{L}^I_{CL-V}$ and $\mathcal{L}^I_{CL-DF}$, as well as $\mathcal{L}^I_{CL-F}$ and $\mathcal{L}^I_{CL-DF}$ are inequivalent.

ACKNOWLEDGMENT

The authors thank Prof. Dr. Peter Reinartz for providing the necessary data and hardware.

REFERENCES

[1] R. Qin, J. Tian, and P. Reinartz, "3d change detection–approaches and applications," *ISPRS Journal of Photogrammetry and Remote Sensing*, vol. 122, pp. 41–56, 2016.

[2] I. R. Hegazy and M. R. Kaloop, "Monitoring urban growth and land use change detection with gis and remote sensing techniques in daqahlia governorate egypt," *International Journal of Sustainable Built Environment*, vol. 4, no. 1, pp. 117–124, 2015.

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

[4] Z. Ali, A. Tuladhar, and J. Zevenbergen, "An integrated approach for updating cadastral maps in pakistan using satellite remote sensing data," *International Journal of Applied Earth Observation and Geoinformation*, vol. 18, pp. 386–398, 2012.

[5] D. Wen, X. Huang, F. Bovolo, J. Li, X. Ke, A. Zhang, and J. A. Benediktsson, "Change detection from very-high-spatial-resolution optical remote sensing images: Methods, applications, and future directions," *IEEE Geoscience and Remote Sensing Magazine*, vol. 9, no. 4, pp. 68–101, 2021.

[6] 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.

[7] J. Deng, K. Wang, Y. Deng, and G. Qi, "Pca-based land-use change detection and analysis using multitemporal and multisensor satellite data," *International Journal of Remote Sensing*, vol. 29, no. 16, pp. 4823–4838, 2008.

[8] L. Bruzzone and D. F. Prieto, "An adaptive semiparametric and context-based approach to unsupervised change detection in multitemporal remote-sensing images," *IEEE Transactions on image processing*, vol. 11, no. 4, pp. 452–466, 2002.

[9] T. Lei, J. Wang, H. Ning, X. Wang, D. Xue, Q. Wang, and A. K. Nandi, "Difference enhancement and spatial–spectral nonlocal network for change detection in vhr remote sensing images," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 60, pp. 1–13, 2021.

[10] L. Bruzzone and F. Bovolo, "A novel framework for the design of change-detection systems for very-high-resolution remote sensing images," *Proceedings of the IEEE*, vol. 101, no. 3, pp. 609–630, 2012.

[11] Z. Lei, T. Fang, H. Huo, and D. Li, "Bi-temporal texton forest for land cover transition detection on remotely sensed imagery," *IEEE Transactions on Geoscience and remote sensing*, vol. 52, no. 2, pp. 1227–1237, 2013.

[12] F. Bovolo, L. Bruzzone, and M. Marconcini, "A novel approach to unsupervised change detection based on a semisupervised svm and a similarity measure," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 46, no. 7, pp. 2070–2082, 2008.

[13] C. Wu, L. Zhang, and L. Zhang, "A scene change detection framework for multi-temporal very high resolution remote sensing images," *Signal Processing*, vol. 124, pp. 184–197, 2016.

[14] K. J. Wessels, F. Van den Bergh, D. P. Roy, B. P. Salmon, K. C. Steenkamp, B. MacAlister, D. Swanepoel, and D. Jewitt, "Rapid land cover map updates using change detection and robust random forest classifiers," *Remote sensing*, vol. 8, no. 11, p. 888, 2016.

[15] H. Nemmour and Y. Chibani, "Multiple support vector machines for land cover change detection: An application for mapping urban extensions," *ISPRS Journal of Photogrammetry and Remote Sensing*, vol. 61, no. 2, pp. 125–133, 2006.

[16] L. Zhou, G. Cao, Y. Li, and Y. Shang, "Change detection based on conditional random field with region connection constraints in high-resolution remote sensing images," *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, vol. 9, no. 8, pp. 3478–3488, 2016.

[17] T. Kasetkasem and P. K. Varshney, "An image change detection algorithm based on markov random field models," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 40, no. 8, pp. 1815–1823, 2002.

[18] W. Gu, Z. Lv, and M. Hao, "Change detection method for remote sensing images based on an improved markov random field," *Multimedia Tools and Applications*, vol. 76, pp. 17 719–17 734, 2017.

[19] G. Cao, L. Zhou, and Y. Li, "A new change-detection method in high-resolution remote sensing images based on a conditional random field model," *International Journal of Remote Sensing*, vol. 37, no. 5, pp. 1173–1189, 2016.

[20] X. X. Zhu, D. Tuia, L. Mou, G.-S. Xia, L. Zhang, F. Xu, and F. Fraundorfer, "Deep learning in remote sensing: A comprehensive review and list of resources," *IEEE geoscience and remote sensing magazine*, vol. 5, no. 4, pp. 8–36, 2017.

[21] W. Shi, M. Zhang, R. Zhang, S. Chen, and Z. Zhan, "Change detection based on artificial intelligence: State-of-the-art and challenges," *Remote Sensing*, vol. 12, no. 10, p. 1688, 2020.

[22] A. Shafique, G. Cao, Z. Khan, M. Asad, and M. Aslam, "Deep learning-based change detection in remote sensing images: A review," *Remote Sensing*, vol. 14, no. 4, p. 871, 2022.

[23] H. Jiang, M. Peng, Y. Zhong, H. Xie, Z. Hao, J. Lin, X. Ma, and X. Hu, "A survey on deep learning-based change detection from high-resolution remote sensing images," *Remote Sensing*, vol. 14, no. 7, p. 1552, 2022.

[24] X. Wu, D. Hong, and J. Chanussot, "Uiu-net: U-net in u-net for infrared small object detection," *IEEE Transactions on Image Processing*, vol. 32, pp. 364–376, 2022.

[25] C. Li, B. Zhang, D. Hong, J. Yao, and J. Chanussot, "Lrr-net: An interpretable deep unfolding network for hyperspectral anomaly detection," *IEEE Transactions on Geoscience and Remote Sensing*, 2023.

[26] L. Ma, Y. Liu, X. Zhang, Y. Ye, G. Yin, and B. A. Johnson, "Deep learning in remote sensing applications: A meta-analysis and review," *ISPRS journal of photogrammetry and remote sensing*, vol. 152, pp. 166–177, 2019.

[27] L. Alzubaidi, J. Zhang, A. J. Humaidi, A. Al-Dujaili, Y. Duan, O. Al-Shamma, J. Santamaría, M. A. Fadhel, M. Al-Amidie, and L. Farhan, "Review of deep learning: Concepts, cnn architectures, challenges, applications, future directions," *Journal of big Data*, vol. 8, pp. 1–74, 2021.

[28] J. Zhao, M. Gong, J. Liu, and L. Jiao, "Deep learning to classify difference image for image change detection," in *2014 International Joint Conference on Neural Networks (IJCNN)*. IEEE, 2014, pp. 411–417.

[29] N. Lv, C. Chen, T. Qiu, and A. K. Sangaiah, "Deep learning and superpixel feature extraction based on contractive autoencoder for change detection in sar images," *IEEE transactions on industrial informatics*, vol. 14, no. 12, pp. 5530–5538, 2018.

[30] H. Chen, Z. Qi, and Z. Shi, "Remote sensing image change detection with transformers," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 60, pp. 1–14, 2021.

[31] 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.

[32] 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.

[33] 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.

[34] H. Chen, C. Wu, B. Du, and L. Zhang, "Dsdanet: Deep siamese domain adaptation convolutional neural network for cross-domain change detection," *arXiv preprint arXiv:2006.09225*, 2020.

[35] Q. Shi, M. Liu, S. Li, X. Liu, F. Wang, and L. Zhang, "A deeply supervised attention metric-based network and an open aerial image dataset for remote sensing change detection," *IEEE transactions on geoscience and remote sensing*, vol. 60, pp. 1–16, 2021.

[36] S. Fang, K. Li, J. Shao, and Z. Li, "Snunet-cd: A densely connected siamese network for change detection of vhr images," *IEEE Geoscience and Remote Sensing Letters*, vol. 19, pp. 1–5, 2021.

[37] G. Cheng, G. Wang, and J. Han, "Isnet: Towards improving separability for remote sensing image change detection," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 60, pp. 1–11, 2022.

[38] Z. Chen, Y. Zhou, B. Wang, X. Xu, N. He, S. Jin, and S. Jin, "Egde-net: A building change detection method for high-resolution remote sensing imagery based on edge guidance and differential enhancement," *ISPRS Journal of Photogrammetry and Remote Sensing*, vol. 191, pp. 203–222, 2022.

[39] Z. Li, C. Yan, Y. Sun, and Q. Xin, "A densely attentive refinement network for change detection based on very-high-resolution bitemporal remote sensing images," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 60, pp. 1–18, 2022.

[40] R. Zhang, H. Zhang, X. Ning, X. Huang, J. Wang, and W. Cui, "Global-aware siamese network for change detection on remote sensing images," *ISPRS Journal of Photogrammetry and Remote Sensing*, vol. 199, pp. 61–72, 2023.

[41] A. Dosovitskiy, L. Beyer, A. Kolesnikov, D. Weissenborn, X. Zhai, T. Unterthiner, M. Dehghani, M. Minderer, G. Heigold, S. Gelly *et al.*, "An image is worth 16x16 words: Transformers for image recognition at scale," *arXiv preprint arXiv:2010.11929*, 2020.

[42] Z. Liu, Y. Lin, Y. Cao, H. Hu, Y. Wei, Z. Zhang, S. Lin, and B. Guo, "Swin transformer: Hierarchical vision transformer using shifted windows," in *Proceedings of the IEEE/CVF international conference on computer vision*, 2021, pp. 10 012–10 022.

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

[44] C. Zhang, L. Wang, S. Cheng, and Y. Li, "Swinsunet: Pure transformer network for remote sensing image change detection," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 60, pp. 1–13, 2022.

[45] Q. Li, R. Zhong, X. Du, and Y. Du, "Transunetcd: A hybrid transformer network for change detection in optical remote-sensing images," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 60, pp. 1–19, 2022.

[46] D. Hong, B. Zhang, X. Li, Y. Li, C. Li, J. Yao, N. Yokoya, H. Li, X. Jia, A. Plaza *et al.*, "Spectralgpt: Spectral foundation model," *arXiv preprint arXiv:2311.07113*, 2023.

[47] K. Li, X. Cao, and D. Meng, "A new learning paradigm for foundation model-based remote sensing change detection," *arXiv preprint arXiv:2312.01163*, 2023.

[48] X. Guo, J. Lao, B. Dang, Y. Zhang, L. Yu, L. Ru, L. Zhong, Z. Huang, K. Wu, D. Hu *et al.*, "Skysense: A multi-modal remote sensing foundation model towards universal interpretation for earth observation imagery," *arXiv preprint arXiv:2312.10115*, 2023.

[49] D. Hong, B. Zhang, H. Li, Y. Li, J. Yao, C. Li, M. Werner, J. Chanussot, A. Zipf, and X. X. Zhu, "Cross-city matters: A multimodal remote sensing benchmark dataset for cross-city semantic segmentation using high-resolution domain adaptation networks," *Remote Sensing of Environment*, vol. 299, p. 113856, 2023.

[50] P. J. S. Vega, G. A. O. P. da Costa, R. Q. Feitosa, M. X. O. Adarme, C. A. de Almeida, C. Heipke, and F. Rottensteiner, "An unsupervised domain adaptation approach for change detection and its application to deforestation mapping in tropical biomes," *ISPRS Journal of Photogrammetry and Remote Sensing*, vol. 181, pp. 113–128, 2021.

[51] Y. Xie, J. Tian, and X. X. Zhu, "A co-learning method to utilize optical images and photogrammetric point clouds for building extraction," *International Journal of Applied Earth Observation and Geoinformation*, vol. 116, p. 103165, 2023.

[52] M. Turker and B. Cetinkaya, "Automatic detection of earthquake-damaged buildings using dems created from pre-and post-earthquake stereo aerial photographs," *International Journal of Remote Sensing*, vol. 26, no. 4, pp. 823–832, 2005.

[53] L. Zhu, H. Shimamura, K. Tachibana, Y. Li, and P. Gong, "Building change detection based on object extraction in dense urban areas," *International Archives of the Photogrammetry, Remote Sensing and Spatial Information Sciences*, 2008.

[54] F. Jung, "Detecting building changes from multitemporal aerial stereopairs," *ISPRS Journal of Photogrammetry and Remote Sensing*, vol. 58, no. 3-4, pp. 187–201, 2004.

[55] A. Sasagawa, E. Baltsavias, S. Kocaman-Aksakal, and J. D. Wegner, "Investigation on automatic change detection using pixel-changes and dsm-changes with alos-prism triplet images," *International archives of the photogrammetry, remote sensing and spatial information sciences*, vol. 40, no. 7/W2, pp. 213–217, 2013.

[56] J. Tian, H. Chaabouni-Chouayakh, and P. Reinartz, "3d building change detection from high resolution spaceborne stereo imagery," in *2011 International Workshop on Multi-Platform/Multi-Sensor Remote Sensing and Mapping*. IEEE, 2011, pp. 1–7.

[57] G. Dini, K. Jacobsen, F. Rottensteiner, M. Al Rajhi, and C. Heipke, "3d building change detection using high resolution stereo images and a gis database," *The International Archives of the Photogrammetry, Remote Sensing and Spatial Information Sciences; XXXIX-B7*, vol. 39, pp. 299–304, 2012.

[58] Y. Xie and J. Tian, "Multimodal co-learning: A domain adaptation method for building extraction from optical remote sensing imagery," in *2023 Joint Urban Remote Sensing Event (JURSE)*. IEEE, 2023, pp. 1–4.

[59] Y. Xie, K. Schindler, J. Tian, and X. X. Zhu, "Exploring cross-city semantic segmentation of als point clouds," *International Archives of the Photogrammetry, Remote Sensing and Spatial Information Sciences*, vol. 43, pp. 247–254, 2021.

[60] X. Yuan, J. Tian, and P. Reinartz, "Building change detection based on deep learning and belief function," in *2019 Joint Urban Remote Sensing Event (JURSE)*. IEEE, 2019, pp. 1–4.

[61] A. Rahate, R. Walambe, S. Ramanna, and K. Kotecha, "Multimodal co-learning: Challenges, applications with datasets, recent advances and future directions," *Information Fusion*, vol. 81, pp. 203–239, 2022.

[62] T. Baltrušaitis, C. Ahuja, and L.-P. Morency, "Multimodal machine learning: A survey and taxonomy," *IEEE transactions on pattern analysis and machine intelligence*, vol. 41, no. 2, pp. 423–443, 2018.

[63] Y. Wang, Y. Wan, Y. Zhang, B. Zhang, and Z. Gao, "Imbalance knowledge-driven multi-modal network for land-cover semantic segmentation using aerial images and lidar point clouds," *ISPRS Journal of Photogrammetry and Remote Sensing*, vol. 202, pp. 385–404, 2023.

[64] M. J. Westoby, J. Brasington, N. F. Glasser, M. J. Hambrey, and J. M. Reynolds, "'structure-from-motion'photogrammetry: A low-cost, effective tool for geoscience applications," *Geomorphology*, vol. 179, pp. 300–314, 2012.

[65] S. Gehrke, K. Morin, M. Downey, N. Boehrer, and T. Fuchs, "Semi-global matching: An alternative to lidar for dsm generation," in *Proceedings of the 2010 Canadian Geomatics Conference and Symposium of Commission I*, vol. 2, no. 6, 2010.

[66] R. Perko, H. Raggam, and P. M. Roth, "Mapping with pléiades—end-to-end workflow," *Remote Sensing*, vol. 11, no. 17, p. 2052, 2019.

[67] H. A. Al-Najjar, B. Kalantar, B. Pradhan, V. Saeidi, A. A. Halin, N. Ueda, and S. Mansor, "Land cover classification from fused dsm and uav images using convolutional neural networks," *Remote Sensing*, vol. 11, no. 12, p. 1461, 2019.

[68] M. Á. Aguilar, M. del Mar Saldaña, and F. J. Aguilar, "Generation and quality assessment of stereo-extracted dsm from geoeye-1 and worldview-2 imagery," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 52, no. 2, pp. 1259–1271, 2013.

[69] P. d'Angelo and J. Tian, "Geometric evaluation of gaofen-7 stereo data," *ISPRS Annals of the Photogrammetry, Remote Sensing and Spatial Information Sciences*, vol. 10, pp. 805–811, 2023.

[70] N. Audebert, B. Le Saux, and S. Lefèvre, "Beyond rgb: Very high resolution urban remote sensing with multimodal deep networks," *ISPRS journal of photogrammetry and remote sensing*, vol. 140, pp. 20–32, 2018.

[71] H. Hosseinpour, F. Samadzadegan, and F. D. Javan, "Cmgfnet: A deep cross-modal gated fusion network for building extraction from very high-resolution remote sensing images," *ISPRS journal of photogrammetry and remote sensing*, vol. 184, pp. 96–115, 2022.

[72] M. Fuentes Reyes, Y. Xie, X. Yuan, P. d'Angelo, F. Kurz, D. Cerra, and J. Tian, "A 2d/3d multimodal data simulation approach with applications on urban semantic segmentation, building extraction and change detection," *ISPRS Journal of Photogrammetry and Remote Sensing*, vol. 205, pp. 74–97, 2023.

[73] M. Schmitt and X. X. Zhu, "Data fusion and remote sensing: An ever-growing relationship," *IEEE Geoscience and Remote Sensing Magazine*, vol. 4, no. 4, pp. 6–23, 2016.

[74] J. Huang, X. Zhang, Q. Xin, Y. Sun, and P. Zhang, "Automatic building extraction from high-resolution aerial images and lidar data using gated residual refinement network," *ISPRS journal of photogrammetry and remote sensing*, vol. 151, pp. 91–105, 2019.

[75] T. Peters, C. Brenner, and K. Schindler, "Semantic segmentation of mobile mapping point clouds via multi-view label transfer," *ISPRS Journal of Photogrammetry and Remote Sensing*, vol. 202, pp. 30–39, 2023.

[76] S. Huang, M. Usvyatsov, and K. Schindler, "Indoor scene recognition in 3d," in *2020 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*. IEEE, 2020, pp. 8041–8048.

[77] S. Bachhofner, A.-M. Loghin, J. Otepka, N. Pfeifer, M. Hornacek, A. Siposova, N. Schmidinger, K. Hornik, N. Schiller, O. Kähler *et al.*, "Generalized sparse convolutional neural networks for semantic segmentation of point clouds derived from tri-stereo satellite imagery," *Remote Sensing*, vol. 12, no. 8, p. 1289, 2020.

[78] C. Hazirbas, L. Ma, C. Domokos, and D. Cremers, "Fusenet: Incorporating depth into semantic segmentation via fusion-based cnn architecture," in *Computer Vision–ACCV 2016: 13th Asian Conference on Computer Vision, Taipei, Taiwan, November 20-24, 2016, Revised Selected Papers, Part I 13*. Springer, 2017, pp. 213–228.

[79] K. Bittner, F. Adam, S. Cui, M. Körner, and P. Reinartz, "Building footprint extraction from vhr remote sensing images combined with normalized dsms using fused fully convolutional networks," *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, vol. 11, no. 8, pp. 2615–2629, 2018.

[80] Y.-C. Li, H.-C. Li, W.-S. Hu, and H.-L. Yu, "Dspcanet: Dual-channel scale-aware segmentation network with position and channel attentions for high-resolution aerial images," *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, vol. 14, pp. 8552–8565, 2021.

[81] P. Zhang, P. Du, C. Lin, X. Wang, E. Li, Z. Xue, and X. Bai, "A hybrid attention-aware fusion network (hafnet) for building extraction from high-resolution imagery and lidar data," *Remote Sensing*, vol. 12, no. 22, p. 3764, 2020.

[82] S. Zhou, Y. Feng, S. Li, D. Zheng, F. Fang, Y. Liu, and B. Wan, "Dsm-assisted unsupervised domain adaptive network for semantic segmentation of remote sensing imagery," *IEEE Transactions on Geoscience and Remote Sensing*, 2023.

[83] J. Tian, S. Cui, and P. Reinartz, "Building change detection based on satellite stereo imagery and digital surface models," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 52, no. 1, pp. 406–417, 2013.

[84] J. Tian and J. Dezert, "Fusion of multispectral imagery and dsms for building change detection using belief functions and reliabilities," *International Journal of Image and Data Fusion*, vol. 10, no. 1, pp. 1–27, 2019.

[85] S. Tian, Y. Zhong, A. Ma, and L. Zhang, "Three-dimensional change detection in urban areas based on complementary evidence fusion," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 60, pp. 1–13, 2021.

[86] H. Wang, X. Lv, K. Zhang, and B. Guo, "Building change detection based on 3d co-segmentation using satellite stereo imagery," *Remote Sensing*, vol. 14, no. 3, p. 628, 2022.

[87] Q. Li, Y. Shi, S. Auer, R. Roschlaub, K. Möst, M. Schmitt, C. Glock, and X. Zhu, "Detection of undocumented building constructions from official geodata using a convolutional neural network," *Remote Sensing*, vol. 12, no. 21, p. 3537, 2020.

[88] 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.

[89] Y. Cui, R. Chen, W. Chu, L. Chen, D. Tian, Y. Li, and D. Cao, "Deep learning for image and point cloud fusion in autonomous driving: A review," *IEEE Transactions on Intelligent Transportation Systems*, vol. 23, no. 2, pp. 722–739, 2021.

[90] J. Tian, P. Reinartz, P. d'Angelo, and M. Ehlers, "Region-based automatic building and forest change detection on cartosat-1 stereo imagery," *ISPRS Journal of Photogrammetry and Remote Sensing*, vol. 79, pp. 226–239, 2013.

[91] P. d'Angelo, "Improving semi-global matching: cost aggregation and confidence measure," *The International Archives of the Photogrammetry, Remote Sensing and Spatial Information Sciences*, vol. 41, pp. 299–304, 2016.

[92] A. Paszke, S. Gross, S. Chintala, G. Chanan, E. Yang, Z. DeVito, Z. Lin, A. Desmaison, L. Antiga, and A. Lerer, "Automatic differentiation in pytorch," in *NIPS-W*, 2017.

[93] D. Hong, N. Yokoya, J. Chanussot, and X. X. Zhu, "An augmented linear mixing model to address spectral variability for hyperspectral unmixing," *IEEE Transactions on Image Processing*, vol. 28, no. 4, pp. 1923–1938, 2018.

[94] L. Zhang, M. Lan, J. Zhang, and D. Tao, "Stagewise unsupervised domain adaptation with adversarial self-training for road segmentation of remote-sensing images," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 60, pp. 1–13, 2021.

[95] 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.

[96] Z. Huang, Q. Liu, H. Zhou, G. Gao, T. Xu, Q. Wen, and Y. Wang, "Building detection from panchromatic and multispectral images with dual-stream asymmetric fusion networks," *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, vol. 16, pp. 3364–3377, 2023.

[97] E. J. Hoffmann, K. Abdulahhad, and X. X. Zhu, "Using social media images for building function classification," *Cities*, vol. 133, p. 104107, 2023.

[98] G. Kyriakaki, A. Doulamis, N. Doulamis, M. Ioannides, K. Makantasis, E. Protopapadakis, A. Hadjiprocopis, K. Wenzel, D. Fritsch, M. Klein *et al.*, "4d reconstruction of tangible cultural heritage objects from web-retrieved images," *International Journal of Heritage in the Digital Era*, vol. 3, no. 2, pp. 431–451, 2014.

[99] J. E. Van Engelen and H. H. Hoos, "A survey on semi-supervised learning," *Machine learning*, vol. 109, no. 2, pp. 373–440, 2020.

[100] Y. Wang, C. M. Albrecht, N. A. A. Braham, L. Mou, and X. X. Zhu, "Self-supervised learning in remote sensing: A review," *IEEE Geoscience and Remote Sensing Magazine*, vol. 10, no. 4, pp. 213–247, 2022.

[101] V. Stojnic and V. Risojevic, "Self-supervised learning of remote sensing scene representations using contrastive multiview coding," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2021, pp. 1182–1191.

[102] E. Protopapadakis, A. Doulamis, N. Doulamis, and E. Maltezos, "Stacked autoencoders driven by semi-supervised learning for building extraction from near infrared remote sensing imagery," *Remote Sensing*, vol. 13, no. 3, p. 371, 2021.

[103] D. Hong, L. Gao, R. Hang, B. Zhang, and J. Chanussot, "Deep encoder–decoder networks for classification of hyperspectral and lidar data," *IEEE Geoscience and Remote Sensing Letters*, vol. 19, p. 5500205, 2020.

[104] Z. Qiu, H. Shen, L. Yue, and G. Zheng, "Cross-sensor remote sensing imagery super-resolution via an edge-guided attention-based network," *ISPRS Journal of Photogrammetry and Remote Sensing*, vol. 199, pp. 226–241, 2023.

Yuxing Xie received his B.Eng. degree in remote sensing science and technology from Wuhan University, Wuhan, China, in 2015, and the M.Eng. degree in photogrammetry and remote sensing from Wuhan University, Wuhan, China, in 2018. He is currently pursuing the Ph.D. degree at the Technical University of Munich (TUM), Munich, Germany. From 2018 to 2023, he worked at the Remote Sensing Technology Institute (IMF) of German Aerospace Center (DLR), Wessling, Germany. He was a guest scientist at the Department of Photogrammetry and Remote Sensing, ETH Zurich, Zurich, Switzerland in 2020. His research interests include digital image processing, point cloud processing, multimodal deep learning, urban remote sensing, and 3D simulation.

Xiangtian Yuan received his B.Eng. degree in civil and environmental engineering from East China Normal University, Shanghai, in 2015, and M.Sc. degree in Civil Engineering from the University of Washington, Seattle, in 2017. Since 2018, he has been working at the Photogrammetry and Image Analysis Department in the Remote Sensing Technology Institute, German Aerospace Center, Wessling, Germany, where he is also pursuing his Ph.D degree in geoinformatics.

His research interests are the application of deep learning in multimodal remote sensing, with focus on urban remote sensing, disaster monitoring, and 3D change detection.

Xiao Xiang Zhu (S'10–M'12–SM'14–F'21) received the Master (M.Sc.) degree, her doctor of engineering (Dr.-Ing.) degree and her "Habilitation" in the field of signal processing from Technical University of Munich (TUM), Munich, Germany, in 2008, 2011 and 2013, respectively.

She is the Chair Professor for Data Science in Earth Observation at Technical University of Munich (TUM) and was the founding Head of the Department "EO Data Science" at the Remote Sensing Technology Institute, German Aerospace Center (DLR). Since May 2020, she is the PI and director of the international future AI lab "AI4EO – Artificial Intelligence for Earth Observation: Reasoning, Uncertainties, Ethics and Beyond", Munich, Germany. Since October 2020, she also serves as a Director of the Munich Data Science Institute (MDSI), TUM. From 2019 to 2022, Zhu has been a co-coordinator of the Munich Data Science Research School (www.mu-ds.de) and the head of the Helmholtz Artificial Intelligence – Research Field "Aeronautics, Space and Transport". Prof. Zhu was a guest scientist or visiting professor at the Italian National Research Council (CNR-IREA), Naples, Italy, Fudan University, Shanghai, China, the University of Tokyo, Tokyo, Japan and University of California, Los Angeles, United States in 2009, 2014, 2015 and 2016, respectively. She is currently a visiting AI professor at ESA's Phi-lab, Frascati, Italy. Her main research interests are remote sensing and Earth observation, signal processing, machine learning and data science, with their applications in tackling societal grand challenges, e.g. Global Urbanization, UN's SDGs and Climate Change.

Dr. Zhu has been a member of young academy (Junge Akademie/Junges Kolleg) at the Berlin-Brandenburg Academy of Sciences and Humanities and the German National Academy of Sciences Leopoldina and the Bavarian Academy of Sciences and Humanities. She serves in the scientific advisory board in several research organizations, among others the German Research Center for Geosciences (GFZ, 2020-2023) and Potsdam Institute for Climate Impact Research (PIK). She is an associate Editor of IEEE Transactions on Geoscience and Remote Sensing, Pattern Recognition and serves as the area editor responsible for special issues of IEEE Signal Processing Magazine. She is a Fellow of IEEE.

Jiaojiao Tian (M'19–SM'21) received her B.S. degree in geoinformation systems from the China University of Geoscience, Beijing, in 2006, her M. Eng. degree in cartography and geoinformation at the Chinese Academy of Surveying and Mapping, Beijing, in 2009, and her Ph.D. degree in mathematics and computer science from Osnabrück University, Germany, in 2013. Since 2009, she has been with the Photogrammetry and Image Analysis Department, Remote Sensing Technology Institute, German Aerospace Center, Wessling, Germany, where she is currently head of the 3D and Modeling Group. In 2011, she was a guest scientist with the Institute of Photogrammetry and Remote Sensing, ETH Zürich, Switzerland. She serves as a co-chair of the ISPRS Commision WG I/8: Multi-sensor Modelling and Cross-modality Fusion. She is a member of the editorial board of the ISPRS Journal of Photogrammetry and Remote Sensing and of the International Journal of Image and Data Fusion.

Her research interests include 3D change detection, digital surface model (DSM) generation, 3D point cloud semantic segmentation, object extraction, and DSM-assisted building reconstruction, forest monitoring and classification.