Self-Supervised Remote Sensing Image Change Detection and Data Fusion

Yuxing Chen

Advisor: Prof. Lorenzo Bruzzone

Co-Advisor: Prof. Stefano Vitale

Department of Information Engineering and Computer Science
University of Trento

This dissertation is submitted for the degree of
Doctor of Philosophy

ICT International Doctoral School November 2023

I would like to dedicate this thesis to my beloved family and my aunt Mrs Jiang, Yan and Mrs Guo, Hanwen.

Acknowledgements

This thesis is a result of all my friends and Professors who have aided me along the way. Most importantly, I am deeply thankful to Dr Liu, Qi for his unwavering support and encouragement during my studies abroad, especially during the challenging period of the COVID-19 pandemic. His empathy, compassion and encouragement gave me the strength to overcome my weaknesses and keep motivated in my research. His support was a source of strength, especially during the difficult times confronting the Asian hate from racist individuals. Without his support, I might have given up on this journey during these challenging times. Carry this courage and get the light from Professors, I finally get here. I express my sincere gratitude to Prof. Lorenzo Bruzzone for providing me with the opportunity to pursue a PhD in RSLab. Despite my shortcomings in paper writing, his patience, understanding, and continuous support gave me confidence in carrying on with the work. I am also grateful to Prof. Stefano Vitale for introducing me to the fascinating world of gravitational wave detection.

My appreciation also goes out to my colleagues who were a guiding light throughout this journey. In particular, I would like to thank Dr Sancharia for her invaluable help in resolving data collection problems, Elena for providing me with radar sounder data, and Elisa for encouraging me to explore new research topics and accompanying me on an unforgettable conference journey to KalaLump. I am grateful to my friend and brother, Lei, who introduced me to RSLab and supported me in the early stages of my research. We also shared a beautiful summer travelling experience in Europe. I must also express my gratitude to my RSLab peers, especially Lifeng, Abhishek, Jordy, Maofan, Jing, Jianming, Haohao, and Raktim, whose camaraderie made research less burdensome by providing moments of joy and hope amidst perceived failures. I will cherish the memories of my time spent pursuing my PhD in this beautiful mountain city, Trento. I extend my heartfelt thanks to Walter, Elia, and Aprillia for being wonderful friends and creating a memorable experience. In closing the Trento chapter, I must convey my deepest gratitude to Walter for his adventure with me in the Alps, which helped me heal and get through dark moments.

The final words of gratitude are to my father for his unwavering support during my twenty-year absence from home.

Abstract

Self-supervised learning models, which are called foundation models, have achieved great success in computer vision. Meanwhile, the limited access to labeled data has driven the development of self-supervised methods in remote sensing tasks. In remote sensing image change detection, the generative models are extensively utilized in unsupervised binary change detection tasks, while they overly focus on pixels rather than on abstract feature representations. In addition, the state-of-the-art satellite image time series change detection approaches fail to effectively leverage the spatial-temporal information of image time series or generalize well to unseen scenarios. Similarly, in the context of multimodal remote sensing data fusion, the recent successes of deep learning techniques mainly focus on specific tasks and complete data fusion paradigms. These task-specific models lack of generalizability to other remote sensing tasks and become overfitted to the dominant modalities. Moreover, they fail to handle incomplete modalities inputs and experience severe degradation in downstream tasks.

To address these challenges associated with individual supervised learning models, this thesis presents two novel contributions to self-supervised learning models on remote sensing image change detection and multimodal remote sensing data fusion. The first contribution proposes a bi-temporal / multi-temporal contrastive change detection framework, which employs contrastive loss on image patches or superpixels to get fine-grained change maps and incorporates an uncertainty method to enhance the temporal robustness. In the context of satellite image time series change detection, the proposed approach improves the consistency of pseudo labels through feature tracking and tackles the challenges posed by seasonal changes in long-term remote sensing image time series using supervised contrastive loss and the random walk loss in ConvLSTM. The second contribution develops a self-supervised multimodal RS data fusion framework, with a specific focus on addressing the incomplete multimodal RS data fusion challenges in downstream tasks. Within this framework, multimodal RS data are fused by applying a multi-view contrastive loss at the pixel level and reconstructing each modality using others in a generative way based on MultiMAE. In downstream tasks, the proposed approach leverages a random modality

combination training strategy and an attention block to enable fusion across modal-incomplete inputs.

The thesis assesses the effectiveness of the proposed self-supervised change detection approach on single-sensor and cross-sensor datasets of SAR and multispectral images, and evaluates the proposed self-supervised multimodal RS data fusion approach on the multimodal RS dataset with SAR, multispectral images, DEM and also LULC maps. The self-supervised change detection approach demonstrates improvements over state-of-the-art unsupervised change detection methods in challenging scenarios involving multi-temporal and multi-sensor RS image change detection. Similarly, the self-supervised multimodal remote sensing data fusion approach achieves the best performance by employing an intermediate fusion strategy on SAR and optical image pairs, outperforming existing unsupervised data fusion approaches. Notably, in incomplete multimodal fusion tasks, the proposed method exhibits impressive performance on all modal-incomplete and single modality inputs, surpassing the performance of vanilla MultiViT, which tends to overfit on dominant modality inputs and fails in tasks with single modality inputs.

Table of contents

List of figures

List of tables

Chapter 1

Introduction

The present chapter provides an introductory overview of the thesis. We highlight our motivation and briefly overview the related literature on remote sensing image change detection and data fusion tasks, along with an exposition of the problem that forms the core of the thesis. Further, the novel contributions of the thesis are highlighted. Lastly, an outline of the structure of the thesis is presented. This chapter lays the foundation for the subsequent content in the thesis.

1.1 Overview

Remote sensing is a collection of techniques that aims to retrieve and process information of the Earth's surface using reflected or emitted electromagnetic radiation. Hundreds of terabytes of Remote Sensing data are accumulated per day from various systems, which cover most bands of the electromagnetic spectrum and include both active and passive sensors [34]. However, the inherent value of raw remote sensing data in facilitating downstream tasks is limited. The remote sensing data processing and analysis techniques further drain the insights into spatial-temporal information on human activities, Earth's environment, and their mutual influences across our planet. These insights benefit many downstream applications, including but not limited to agriculture, climate change studies, and natural resource management. They are usually provided by Land-use Land-cover (LULC) mapping [38] and change detection [114] tasks.

The importance of LULC mapping lies in its provision of essential information for comprehending the intricate relationship between human activities and the environment. Land Cover retains the physical attributes of the Earth's surface, encompassing elements such as vegetation, water, and soil. In contrast, Land Use delineates the purposes for which humans exploit the Land Cover, reflecting changes induced by anthropogenic activities.

LULC changes refer to the dynamic interplay between human actions and environmental conditions. Nowadays, the importance of accurate LULC and change detection products extends beyond academic interest to the realm of policy implementation concerning the management of natural resources (e.g., crop mapping, forest resource management, mineral mapping) and environmental predicaments (e.g., flooding, wildfires, landslides, deforestation). In the domain of natural resource management, scholars like Hutt et al. [76] advocate the utilization of multitemporal Sentinel-2 images and external agricultural data to formulate comprehensive LULC maps, encompassing annually changing crop types. Similarly, Junaid et al. [80] employ a random forest classifier and Landsat image time series to analyze forest cover in Malam Iabba's forest land. Johnson et al., [79] on the other hand, leverage LULC methods on 16 bands of World View-3 SWIR and VNIR imagery to map seven geological materials. Conversely, for environmental monitoring, Long et al. [100] employed a change detection approach to demarcate the extent of flooding in the Chobe floodplain, located in the Caprivi region of Namibia. Zanetti [164] utilized a one-class classification change detection model to identify prominent wildfire expansions in Rhodes (Greece), Corfu (Greece), and Palermo (Italy) during the summer of 2023, utilizing Sentinel-2 images. Shi et al. [143] employed a Deep Neural Network (DNN)-based change detection method to successfully identify two landslides in Hong Kong, covering a total area exceeding 70 km2. Furthermore, Bem et al. [39] employed a ResUnet-based change detection method in conjunction with Landsat time-series images to detect instances of deforestation between 2017 and 2019 within the Brazilian Amazon. These diverse applications underscore the imperative for meticulous LULC mappings and change detections, serving as catalysts for sustainable development. Consequently, the process of LULC mapping and change detection emerges as an indispensable undertaking within the Remote Sensing (RS) community.

LULC mapping is a process of dividing a remote sensing image into multiple segments or regions, each of which corresponds to a homogeneous object or a feature in the scene. The purpose of LULC mapping is to simplify the analysis and interpretation of RS images by reducing the amount of data and making it easier to identify features of interest. They provide key spatial information for urban planning and natural resource management. In earlier work [19, 21], the LULC mapping methods were mainly based on the spectral information of each pixel, because the spectral information of each pixel can completely characterize various underlying materials (e.g. crops, urban) in coarse-resolution imagery. With the development of remote sensing satellite technology, the spatial and spectral resolutions of RS images have become higher and higher in the past decades. However, the spectral information alone often is not enough to distinguish neighbouring LULC classes. Hence, the joint use of spatial contextual [41] which is based on patches or the superpixel and spectral information

to determine the LULC classes became popular. Supervised and unsupervised methods [147, 3, 74, 53] were widely used to segment images into most discriminate classes given the labels or the class number. Moreover, Deep Neural Networks (DNNs) [89, 52] have been used for automatical feature learning for pixel-wise mapping. The supervised approaches are often limited by the availability of annotated datasets. It is expensive and often not possible to obtain a large amount of annotated samples for network training. In this context, unsupervised methods are preferred to supervised ones in many operational applications while supervised methods are always conditioned on specific tasks. Thus we need different LULC maps for different downstream tasks.

Compared with LULC maps, LULC change introduces additional temporal information showing variants in the Earth's surface. Change detection is the process of identifying and mapping changes in the Earth's surface between two or more remote sensing images acquired at different times. This is a crucial step in monitoring and understanding the dynamics of various environments, including urbanization [44], deforestation [39], LULC changes [1], and natural disasters [172]. Many irrelevant changes, such as radiometric and atmospheric variations, seasonal changes in vegetation, and building shadows, limit the accuracy of change maps. Early approaches to change detection in bitemporal RS images include image algebra, image transformation and image classification methods [15]. Image rationing and change vector analysis (CVA) [20] are early examples of such algebraic approaches. Starting from these algebra attempts in this field, many supervised and unsupervised techniques have been developed. Most of them are based on image transformation algorithms [25, 159, 168] where the important point is to obtain robust features from multi-temporal images. To get a good feature, deep learning methods have been shown to be widely used in this domain. One common approach is direct classification [36], a binary segmentation approach, where models are trained using annotated binary labels. The image transformation approach has also been improved using deep learning, where deep neural networks are utilized to extract discriminative features, such as Generative Adversarial Networks (GAN) [119], AutoEncoders (AEs) [104] and self-supervised learning [91]. The challenge of detecting changes in remote sensing image time series is compounded by the presence of seasonal noise, which can be difficult to distinguish true changes. Many supervised Recurrent Neural Networks (RNNs) [106] and unsupervised approaches are proposed to solve this problem. However, the lack of annotations and the complex change types makes self-supervised change detection much more needed.

In addition to the technical development of LULC mapping and change detection tasks, the development of sensors intrinsically improves the ability of LULC mapping and change detection from data itself, which in turn requires technical developments. Multimodal RS

data fusion further improves this ability by integrating the complementary information extracted from individual sensor data. The accumulation of EO data from different sensors and their increasing temporal and spatial resolutions give a new emphasis to data fusion and information extraction techniques. For example, optical and Synthetic Aperture Radar (SAR) remote sensing data characterize target features in different ways but contain complementary information. Multi-spectral and hyper-spectral images acquire spectral information and enable the interpretation of the land-cover categories on the basis of spectral signatures, while radar images provide dielectric properties and are not affected by cloud occlusions. If used in combination, they can enhance the accuracy and reliability of LULC mapping and change detection tasks. It is well known that the complementary use of multimodal remote sensing data offers more complete information on a scene and can result in better performance in downstream applications [57]. However, most multimodal remote sensing data fusion works [77, 2] are designed in a particular context and in a supervised way, thus they are conditioned at the specific tasks. Moreover, these methods often assume that all modalities are available during the training and inference time. This assumption greatly limits applications of multimodal remote sensing data analysis because in practice data collection process may result with missing modalities. In this situation, these approaches may fail to deal with incomplete image modalities and face severe degradation in downstream applications. Consequently, an incomplete multimodal learning method is highly desired for a flexible and practical remote sensing application with one or more missing modalities. The flexibility of Transfomer [154] makes it possible to train a model across different modalities. However, the severe degradation with modal-incomplete inputs is still present.

To sum up, the DNNs allow us to get effective and robust features in LULC mapping and change detection tasks, by achieving state-of-the-art results in the supervised approach. Nevertheless, the existing RS image change detection and data fusion approaches are still in need of improvement due to their limited generality and heavy reliance on annotated data as well as the limitation on specific downstream tasks.

1.2 Motivation

The aftermentioned introduction underscores the pivotal role of LULC mapping and change detection in the Remote Sensing domain, serving as foundational tasks that provide essential spatial and temporal information for subsequent RS downstream applications. These two tasks are developed by utilizing the advanced multi-temporal and multimodal RS data, while concurrent with the development of RS sensors. For example, low-resolution RGB image is only used for urban or non-urban area classification while multispectral image

enable more refined class distinctions. Furthermore, the integration of Synthetic Aperture Radar (SAR) and multispectral images, or the utilization of hyperspectral images, facilitates the identification of various crop types and crop growth status. However, the exhaustive categorization of Earth's surface into fine-grain classes, sometimes beyond semantic classes, underscores the effort of supervised learning on downstream tasks given task-specific labels. This also happened to change detection tasks, post-classification comparison and supervised change detection approaches are often limited to the given semantics in network training Given the different types of change classes, we usually need to train specific models.

This supervised paradigm is often limited by the availability of annotated datasets and tasks. It is expensive and often not possible to obtain a large amount of annotated data for modeling change maps or fusing multimodality remote sensing data for downstream applications. The scarcity of labels, not only for semantic labels, further renders the supervised paradigm unsuitable for these two fundamental tasks. Consequently, there has been a rising interest in the RS community to build generalist models that can perform a variety of tasks. Hence, We assert that LULC mapping and change detection, beyond the training on limited semantics, are inherently data-centric endeavors. A good feature representation, obtained from RS data itself, can obviate the need for laborious task-specific model training from scratch.

In contrast to manually defining classes, our approach advocates learning class patterns and anomalies directly from RS data, emphasizing feature representation. Specifically, for the LULC mapping task, our objective is to amalgamate multimodal RS data to derive a feature representation conducive to diverse downstream mapping tasks. For change detection, we advocate learning changes intrinsically from the data and subsequently selecting pertinent change types. Thus, within this context, unsupervised methods emerge as preferable over their supervised counterparts.

Confronted with limited access to labeled data, the development of unsupervised methods, including Generative Adversarial Networks (GAN) [61], Mask AutoEncoder (MAE) [108], and self-supervised learning [91], has gained prominence in data fusion and change detection tasks. However, existing research has demonstrated that CNN-based generative models overly emphasize pixels at the expense of abstract feature representations. Recent advancements in contrastive self-supervised learning [151, 63] and Transformer-based AutoEncoder[68] underscore the potential for more interpretable and meaningful feature representations, with applications extending to classification and segmentation tasks.

The focal point of this thesis is the formulation of a self-supervised RS image change detection and data fusion methodology, aiming to circumvent the repetitive task of training models from scratch for each downstream task based on task-specific labels. The motivation

for this research stems from the imperative need for a foundational model applicable across diverse RS downstream tasks. The proposed self-supervised approach, avoiding reliance on specific task-based labels, seeks to streamline model training efforts and enhance adaptability across various applications.

1.3 Problem Definition

According to the motivation of this thesis, we define the research problem of self-supervised remote sensing image change detection and data fusion in the following. In the unsupervised setting, change detection is the operation of distinguishing changed and unchanged pixels by comparing multi-temporal images acquired by different or same sensors at different dates. Let us consider two images I_1 and I_2 acquired at two different dates t_1 and t_2, respectively. The aim of change detection is to create a change intensity map that contains permanent changes, from multi-view images I_1 and I_2. As described in related works, the crucial point in this task is to align the features of unchanged pixels or patches from the different view data $T_1 = f_\gamma(p_1)$ and $T_2 = g_\delta(p_2)$. Here, p_1 and p_2 are unchanged patches or pixels in images I_1 and I_2, respectively. The f and g functions are used to extract the features from multi-temporal images, where γ and δ denote the corresponding parameters. The objective function of our task can be defined as:

$$\gamma, \delta = \arg\min_{\gamma,\delta} \{d[f_\gamma(p_1), g_\delta(p_2)]\} \tag{1.1}$$

where d is a measure of feature distance between T_1 and T_2.

The field of data fusion encompasses a wide range of methods and mathematical tools, including spectral analysis and plausibility theory to diverse themes and applications [155]. The tools employed in a data fusion process can be customized for a specific use case. While there are several early tries on establishing a precise definition of data fusion, it remains challenging. Klein [88] defined data fusion as a multilevel, multifaceted process that involves detecting, associating, correlating, estimating, and combining data and information from single or multiple sources. However, this definition does not account for the quality and reliability of the fusion result. In EO domain, for instance, one can use certain features extracted from multisource images to improve the accuracy and reliability of LULC mapping. Wald [156] refined this definition by stating that data fusion is a formal framework that employs means and tools to combine data from various sources to obtain higher-quality information. In this study, we aim to define self-supervised remote sensing data fusion, which

is a method of learning from multi-source data based on a generative and discriminative approach without explicit supervision.

Generative data fusion is achieved through reconstruction techniques. Recently, Transformer based generative models have gained popularity in learning universal representations that can be transferred to a wide range of downstream tasks. For instance, models such as BEiT [10] and MAE [68] have been proposed that predict discrete tokens and randomly mask patches of input images and then reconstruct them. Moreover, MultiMAE [8] proposes to aggregate multiple modalities and learn a shared representation that can reconstruct each modality by others. In this thesis, we follow a similar paradigm to generate a fusion representation of multimodality data by reconstructing the missing patches of different modalities from visible parts using the MAE framework. Given the observed modality x_1, in order to obtain the reconstruction x_2 of the missing modality, we optimize the following objective for the reconstruction network:

$$\varepsilon^* = \arg\max_{\varepsilon} \sum_{\{x_1, x_2\}} -\log p\left(x_2 \mid x_1; \varepsilon\right) \tag{1.2}$$

where ε is the parameters of the model.

In addition to reconstruction, self-supervised data fusion also involves the use of contrastive learning, which endeavours to capture similarities and differences between various data modalities. By utilizing independent per-modality encodings, this paradigm can help to learn invariant information from multi-view inputs. To this end, distinct models are trained for each modality to generate a final representation for respective inputs. For instance, for image i and text t inputs, separate models f_i and f_t are utilized to produce corresponding representations $z^i = f_i(i)$ and $z^t = f_t(t)$. Consequently, the resulting "two-tower" architecture can be employed to learn representations for a collection of n image and text pairs $\{(i_j, t_j)\}_{j=1}^n$. The representation $Z_n = \{(z^i{}_j, z^t{}_j)\}_{j=1}^n$ is the learned corresponding features for paired inputs, which are closer in feature space than those of unpaired inputs. This is achieved by forming the contrastive loss with the temperature τ:

$$L_j = -\log \frac{e^{sim(z_j^i, z_j^t)/\tau}}{\sum_{k=1}^n e^{sim(z_j^i, z_k^t)/\tau}} \tag{1.3}$$

In terms of fusion paradigms, there are two possibilities: the first treats distinct modalities as different views and aims to eliminate noise or bias while retaining common features. Meanwhile, the second approach stacks the dependent modalities to preserve complementary information. In practice, we can contrast single modality-specific representations to the fusion ones.

1.4 Novel Contributions

By developing new methods that can effectively handle the complexity and variability of multimodal remote sensing data, this thesis aims to contribute to the advancement of remote sensing data analysis technology and its applications. The main novel contributions of this PhD thesis in the field of remote sensing image change detection and data fusion are described as follows.

As far as we know, we are the first to apply contrastive learning on multi-view remote sensing image change detection tasks. To improve the performance of the CNNs-based generative models, we propose the patch-wise contrastive method to remote sensing change detection tasks and assess its performance on bi-temporal and bi-sensor datasets. We further propose a self-supervised change detection approach at the pixel level and introduce a simple but effective uncertainty approach in the change detection task to reduce the impact of seasonal changes. For remote sensing image time-series change detection, we propose to use feature tracking to extract reliable change pixels in image sequences that are insensitive to seasonal changes. To ensure the robustness and consistency of change maps, we propose to use supervised contrastive loss and contrastive random walk loss on change feature learning. These losses encourage the pixels in the same class to have a closer feature representation. To extend the approach to arbitrary long-time series, we jointly use Unet and ConvLSTM as the model architectures. All these works on self-supervised change detection further improve the accuracy of unsupervised binary change detection on multitemporal and multimodal remote sensing images.

For multimodal remote sensing data fusion, we first introduce and verify the effectiveness of multi-view contrastive loss in SAR-optical data fusion. In detail, we propose a self-supervised approach that can obtain pixels-wise feature representation from SAR and optical image pairs without using any annotation. This is achieved by using U-net and the contrastive loss, by preserving local information at the superpixel level. We also studied three different fusion strategies (i.e., early fusion, intermediate fusion and late fusion). To further use the fusion features on downstream applications, we propose a self-supervised land-cover segmentation approach based on contrastive learning and vector quantization in the proposed SAR-optical data fusion framework. However, the proposed approach works only under the availability of all modalities in the inference stage. This assumption greatly limits the application of the fusion features on downstream applications. In this situation, we further propose a unified model for incomplete multimodal learning of remote sensing data, which leverages a mask attention strategy, Bi-LSTM, the contrastive and reconstruction losses in a Multimodal Transformer framework to build the fusion across different modalities in pre-training and supervised training. The proposed approach allows the network learning and

inference on an incomplete modality input. The two proposed remote sensing data fusion approaches broaden the data fusion algorithms and open the door to more complex remote sensing downstream applications, further improving the feature representation ability with respect to end-to-end supervised learning.

These novel contributions fill significant gaps in the current state of the art and provide new insights into remote sensing image change detection and data fusion. The proposed methods can be applied on a wide range of remote sensing applications and will be valuable for researchers and practitioners in the field.

1.5 Structure of the Thesis

The thesis is structured into seven chapters.

Chapter 1 has provided an overview of the background, problem definition, motivation, and novel contributions of the thesis.

Chapter 2 presents a comprehensive review of the state of the art in remote sensing image change detection and data fusion, including the backgrounds of remote sensing data and the algorithm basics of deep learning approaches. This chapter provides an in-depth understanding of the current methods and techniques used in the field and identifies the challenges and limitations of these approaches.

Chapter 3 focuses on the proposed algorithms and techniques for bi-temporal remote sensing image change detection. This chapter details the novel contributions of the proposed approach, including the patch-wise and pixel-wise self-supervised algorithm for remote sensing image change detection. Additionally, it showcases the change maps obtained from the proposed patch-wise and pixel-wise algorithms using bi-temporal and bi-sensor images, along with a comprehensive analysis of the proposed methods against state-of-the-art approaches and also the discussion of the robustness of the uncertainty-enhanced approach on water areas.

Chapter 4 describes the proposed algorithms for remote sensing image time-series change detection. It describes the novel contributions, such as the feature tracking algorithm for pseudo label generation and the use of supervised contrastive loss with contrastive random walk loss on Unet-ConvLSTM, as well as the fine-tuning stage using supervised contrastive learning. The remote sensing image time-series change maps obtained from the proposed algorithms on Landsat-8 and Sentinel-2 multispectral images are also presented and analysed. It also ablates the effectiveness of each algorithm of the proposed approach.

Chapter 5 delves into the self-supervised SAR-optical fusion and segmentation approach. It introduces the novel contributions of the proposed approach, including the self-supervised

fusion algorithm and the self-supervised segmentation using contrastive learning and vector quantization. It showcases three different fusion strategies, which are the early, intermediate and late fusion approaches, at the image-level and superpixel level of SAR and optical image pairs. Experiments on the fine-tuning of learned features of the proposed approach against state-of-the-art approaches and the effectiveness of unsupervised land-cover segmentation on the fusion of SAR-optical image pairs.

Chapter 6 focuses on multimodal remote sensing data fusion methods that can handle modal-incomplete inputs in training and inference. It introduces the proposed Transformer framework with the masked attention strategy and Bi-LSTM as well as the contrastive and reconstruction losses in supervised training and pre-training. This chapter provides insight into a unified and general data fusion approach for diverse remote sensing downstream applications with modal-incomplete inputs. Moreover, it presents the results of two tasks: building instance / semantic segmentation and LULC mapping using optical, SAR and DEM data as well as remote sensing products. The chapter includes a comprehensive analysis of the performance of each component in the proposed methods and compares them with the vanilla Transformer approach. In addition, the comparison between generative and contrastive pre-training is also included in this chapter.

Chapter 7 concludes the thesis and discusses future work. It summarizes the main findings and contributions, emphasizing the potential impact and implications of the proposed methods for remote sensing image change detection and data fusion. Furthermore, this chapter outlines potential avenues for future research, highlighting areas that could benefit from further investigation.

Chapter 2

Background and Related Works

The previous chapter introduced remote sensing image change detection and data fusion, addressing their definition and importance. This chapter aims to present the remote sensing data used for change detection and data fusion, list the existing tools and algorithms and discuss the state of the art of remote sensing image change detection and data fusion techniques. The purpose is not to present an exhaustive review of a large number of published works, but to give an overview of the existing methods.

The chapter presents multispectral, SAR images, and some remote sensing products (i.e., DEM/DSM and Land-use Land-cover map) for image change detection and data fusion (Section 2.1-2.2). Section 2.3 presents a review of multimodal remote sensing data fusion methods, while Section 2.4 details a review of change detection methods based on multi-temporal and cross-sensor data. Section 2.5 summarizes existing neural networks and learning algorithms used in remote sensing image change detection and data fusion. Finally, Section 2.6 outlines the self-supervised generative and discriminative models used in this thesis.

2.1 Optical and SAR Remote Sensing Images

Remote sensing is a technology that enables the identification, measurement, and analysis of characteristics of objects of interest without direct contact. This technology has the advantages of cost-effective, large-scale and real-time retrieval of land surface information compared with in-situ observation data. Remote sensing relies on the measurement of Electromagnetic (EM) energy emitted or reflected by Earth's surface objects to observe and retrieve their features. This is because the EM radiation from Earth's surface can be detected and translated into important information on individual objects. Remote sensing systems can be grouped into two categories: passive and active remote sensing. Passive remote sensing

employs the sun's energy as its source and detects both reflected and emitted energy from Earth's surface objects. In contrast, active remote sensing generates an EM radiation and transmits it to Earth's objects, subsequently capturing the backscattered energy from these objects.

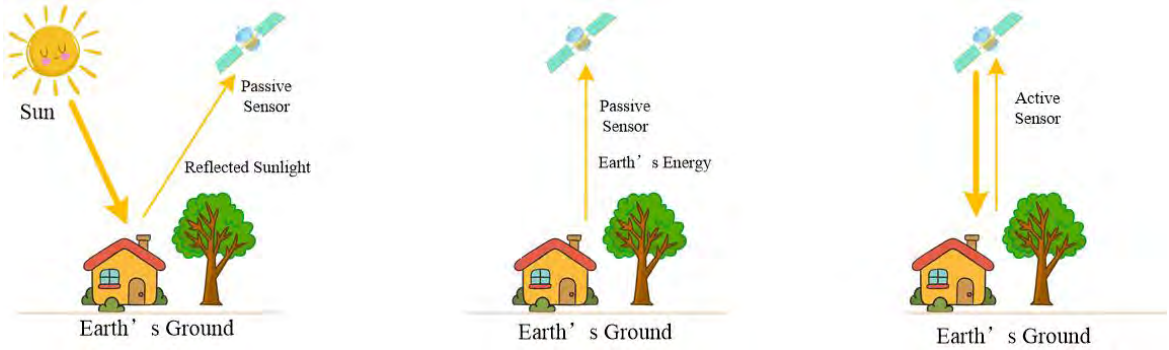

Fig. 2.1 Illustration of remote sensing system (active and passive)

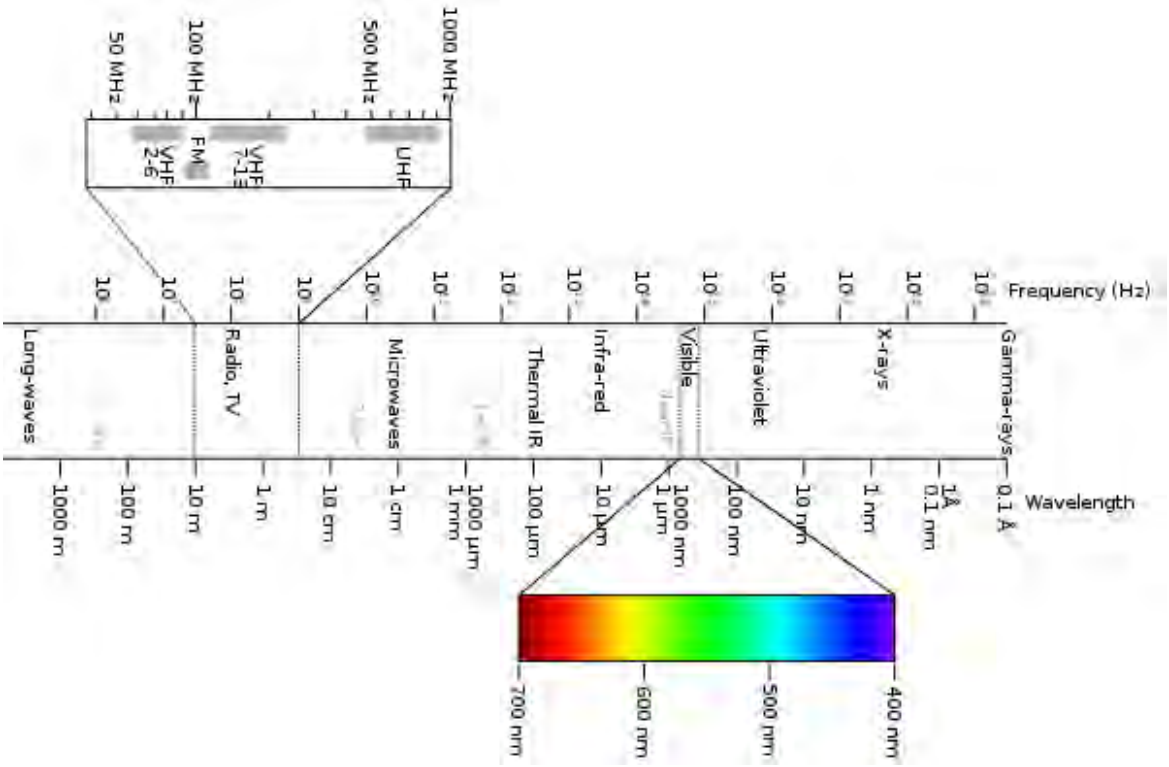

Fig. 2.2 Electromagnetic Spectrum.

2.1.1 Optical Remote Sensing Image Characteristics

Passive optical remote sensing systems work on sunlight reflection, which can only function during daylight hours. Optical sensors require dealing with atmospheric conditions, such as the effect of varying illumination conditions resulting from the position of the sun. Passive sensors detect electromagnetic energy from the optical parts of the spectrum, including the visible, near-infrared (IR), short-wave infrared, and thermal infrared domains (as shown in Fig. 2.2 by Victor Blacus). The visible spectrum band is within the wavelengths between about 400-700 nm [1], while infrared wavelengths occupy a range from approximately 700 nm to 1 mm [2]. The visible region represents only a small portion of the entire electromagnetic spectrum. Compared to the visible region, infrared light possesses longer wavelengths that are suitable for estimating surface temperature (3 - 14 um for thermal infrared) or vegetation conditions. In thermal infrared (TIR) remote sensing, cameras mainly gather energy directly emitted from the surface of the Earth, allowing passive sensors to operate during the day or night.

Optical remote sensing data come in various types, including panchromatic imagery, multispectral imagery, and hyperspectral imagery: (1) Panchromatic images are characterized by a single spectral band acquired on a broad spectrum, including wavelengths from various visible bands and a portion of the TIR. Panchromatic images are visualized as grayscale images and usually have a higher resolution than multispectral images. Notably, most satellites, including Landsat, Digital Globe's satellites and the SPOT constellations, produce panchromatic imagery along with multispectral imagery.

(2) Multispectral imaging is a technique that involves capturing images in multiple spectral bands. The majority of optical remote sensing systems operate in this mode (e.g., Landsat, Sentinel-2 and Sentinel-3). In comparison to the panchromatic mode, the multispectral mode is considered more beneficial due to the abundance of spectral information it provides.

(3) Hyperspectral images acquire data in the form of a sequence of narrow and contiguous wavelength bands, typically spaced at intervals ranging from 10 to 20 nm. Some hyperspectral systems, such as AVIRIS, EO-1 Hyperion and PRISMA, are capable of capturing hundreds of spectral bands. This vast amount of bands allows for the detection of subtle variations in reflected energy, rendering it highly sensitive and therefore capable of differentiating between LULC features with greater precision.

There are several satellites with different characteristics that acquire multispectral images of the Earth's surface, such as Landsat-8 and Sentinel-2. They have proven to be especially valuable for land cover monitoring because they offer cost-free images and their data has

[1] https://en.wikipedia.org/wiki/Visible_spectrum
[2] https://en.wikipedia.org/wiki/Infrared

been gathered for the preceding decades. The Landsat program, consisting of a suite of multispectral satellites, was developed by the National Aeronautics and Space Administration (NASA) of the United States in the early 1970s. One of the notable features of Landsat images is their utility for environmental research. The spatial resolutions of the sensors on the Landsat 8 platform are presented in Table 2.1. Differently, the Sentinel-2 satellite, which was developed by the European Space Agency (ESA) as a component of the Copernicus land monitoring initiative, acquires multispectral data across 12 spectral bands, with spatial resolutions of 10, 20, and 60 meters depending on the specific band, as detailed in Table 2.2.

Table 2.1 Landsat-8 Spectral Bands

Landsat 8 Bands	Wavelength [mm]	Resolution [m]
Band 1 - Coastal aerosol	0.43 - 0.45	30
Band 2 - Blue	0.45 - 0.51	30
Band 3 - Green	0.53 - 0.59	30
Band 4 - Red	0.64 - 0.67	30
Band 5 - Near Infrared (NIR)	0.85 - 0.88	30
Band 6 - SWIR 1	1.57 - 1.65	30
Band 7 - SWIR 2	2.11 - 2.29	30
Band 8 - Panchromatic	0.50 - 0.68	15
Band 9 - Cirrus	1.36 - 1.38	30
Band 10 - Thermal Infrared (TIRS) 1	10.60 - 11.19	100 (resampled to 30)
Band 11 - Thermal Infrared (TIRS) 2	11.50 - 12.51	100 (resampled to 30)

Table 2.2 Sentinel-2 Spectral Bands

Sentinel-2 Bands	Central Wavelength [mm]	Resolution [m]
Band 1 - Coastal aerosol	0.443	60
Band 2 - Blue	0.490	10
Band 3 - Green	0.560	10
Band 4 - Red	0.665	10
Band 5 - Vegetation Red Edge	0.705	20
Band 6 - Vegetation Red Edge	0.740	20
Band 7 - Vegetation Red Edge	0.783	20
Band 8 - NIR	0.842	10
Band 8A - Vegetation Red Edge	0.865	20
Band 9 - Water vapour	0.945	60
Band 10 - SWIR - Cirrus	1.375	60
Band 11 - SWIR	1.610	20
Band 12 - SWIR	2.190	20

2.1.2 SAR Remote Sensing Image Characteristics

There are several types of instruments in active remote sensing systems, where RADAR stands out as a particularly effective technology for mapping and monitoring land cover and land use. Its unique abilities include self-illumination, which allows to capture images during both day and night, and the capability to penetrate cloud cover, light rain and fog. This advantage allows it to overcome the optical limitations of clouds and other atmospheric obstructions. Additionally, RADAR is sensitive to the physical structures of different land cover objects, whose backscattering value is varying with the physical status of the land surface. RADAR technology works in the microwave range of the electromagnetic spectrum.

Synthetic Aperture Radar (SAR) is a fundamental technology in the accurate mapping of land cover. The principle of SAR is to synthesize a large antenna by exploiting the Doppler history from radar echoes generated by the forward motion of the satellite. Consequently, it can achieve a high azimuth resolution, despite having a physically relatively small antenna. SAR signals are transmitted by the radar in a side-looking direction towards the Earth's surface objects, with a given look angle and incidence angle, which differentiates SAR from optical imagery. Differently from optical imagery, spatial resolution in SAR images is defined by range and azimuth resolution. The SAR system's ability to distinguish between two object targets in the along-track direction of the sensor is the azimuth resolution. The range resolution (across-track) depends on the size of the SAR system pulse, while the azimuth resolution depends on antenna size and radar wavelength.

SAR imagery with varying wavelengths can slightly penetrate different types of materials, offering opportunities for diverse applications. The shorter frequency has a stronger penetration ability into Earth's surface. For example, the *L* band has a longer wavelength than *X* and *C* bands, thus penetrating more into the vegetation and, under dry conditions, to some extent, into the soil, such as dry snow or sand. The brightness variation in the SAR image follows a non-uniform pattern and is marked by a granular texture referred to as a speckle. Speckle results from the interference of EM waves due to elementary scatters present in a reduction cell. SAR signals can transmit and receive either horizontal (H) or vertical (V) electric field vectors, which is termed polarization. Typically, there are four types of polarization (HH, HV, VV and VH).

SAR satellite images of both C-band (e.g. ERS-1/2 AMI, Envisat ASAR, RadarSat-1, RISAT-1, Sentinel-1 A/B) and L-band (e.g. SeaSat-1, JERS-1, ALOS PALSAR-1/2) have been applied to perform land cover mapping and change detection tasks for nearly two decades due to their capability to capture changes in high spatial resolution. Recently, X-band TerraSAR-X / TanDEM-X and COSMO-SkyMed (CSK) constellations with very high spatial resolution have enhanced the ability to detect small objects. Moreover, Sentinel-1

is a C-band Copernicus mission of satellites that offers SAR images at medium resolution (approximately 10 m) with high revisit time (about 5 days), along with a wide swath (250 km), and different operational modes including the Interferometric Wide swath (IW) mode which applies the Terrain Observation with Progressive Scanning SAR (TOPSAR) imaging technique. The Level-1 data products provided by Sentinel-1 include Single Look Complex (SLC) and Ground Range Detected (GRD). The Sentinel-1 provides available horizontal (H) or vertical (V) polarization modes including VV and VH polarimetric channels for classifying and analysing land covers such as vegetation or built-up areas. In addition, the Sentinel-1 GRD image preprocessing, which includes applying the orbit file, removing GRD border noise and thermal noise to reduce sub-swath discontinuity, calculating backscatter intensity using radiometric calibration, performing orthorectification (terrain correction) using SRTM 30 m DEM, and converting backscatter coefficient to dB, is performed through the SNAP Graph Processing Tool (GPT).

2.2 Other Modality Remote Sensing Data

2.2.1 DEM and DSM

The digital elevation model (DEM) and digital surface model (DSM) are the main remote sensing product nowadays, which are used to represent three-dimensional earth surfaces. DEMs represent the earth's topography, including things like mountains, hills, and valleys, while DSMs are able to represent more than just topography. DSMs represent above-ground features like buildings and vegetation, as well as topography. There are various sources that are used to generate these models, such as Lidar, InSAR, and stereo satellite images. Lidar (Light Detection and Ranging) uses laser pulses to measure the distance between the sensor and the ground. The technology works by firing rapid pulses of laser light at the terrain, which then bounce back to the sensor, allowing it to determine the distance between the object and the sensor. By scanning an area with a laser, it is possible to create a very accurate 3D point cloud and extract detailed information about above-ground features like vegetation and buildings which can be used to create a DEM or DSM. There are several public DEM or DSM created by Lidar in GEE, such as France, Netherlands and Australia. InSAR (Interferometric Synthetic Aperture Radar) uses radar waves to measure the distance between the satellite and the ground. By using interferogram technology on two radar SLC images, it is possible to create a DEM or DSM. InSAR is considered to be more useful than Lidar when it comes to generating a global Teerain map, as satellites can visit the Earth periodically. There is two famous global DEM product generated by InSAR, SRTM [153]

and TanDEM [174]. Stereo satellite images are a third source of data used to generate DEMs and DSMs. These images are created by taking two satellite images of the same location but from different angles or viewing directions. By overlaying the images and processing them together, it's possible to create three-dimensional models of the terrain. While stereo satellite images are not as accurate as Lidar or InSAR, they are still useful in generating data for larger areas and can be used to fill in gaps in datasets created by other sources.

2.2.2 Land-Use Land-Cover Maps

Land cover and land use maps are two important remote sensing products which can be used for many different downstream applications, such as resource management and change detection. Land cover refers to the physical and biological cover of the Earth's surface, while land use refers to the human activities that take place on the land. Here we introduce two main land-use land-cover products in remote sensing: the dynamic world dataset [18] and European urban altas [113]. The Dynamic World (DNW) dataset is a continuously updating image collection of globally consistent, 10 m resolution, near real-time LULC predictions created from Sentinel-2 imagery. Images in this dataset include ten bands: nine bands with estimated probabilities for each of the nine LULC classes (water, trees, grass, crops, shrub and scrub, flooded vegetation, built-up area, bare ground, and snow & ice) as well as a class "label" band indicating the class with the largest estimated probability. These unique properties enable users to do multi-temporal analysis as well as create custom products suited to their needs. European Urban Atlas provides reliable, inter-comparable, high-resolution LULC maps for over 300 large urban zones and their surroundings for the 2006 reference year in EU member states and for about 800 functional urban areas and their surroundings for the 2012 and 2018 reference year. European Urban Atlas includes 17 urban classes and 10 rural classes. The urban classes include categories such as industrial, commercial, residential, green urban areas, and water bodies. The European Urban Atlas is useful for a variety of applications such as urban planning, environmental monitoring, and climate change research. It can also be used to monitor changes in land use over time and to assess the impact of human activities on the environment. There are also countless LULC maps accumulated in the past decades, which not only can be used as labels for network training but also can be a kind of data source in multimodal learning.

2.3 Multimodal Remote Sensing Data Fusion

The most common approach is based on deep learning techniques applied to single modality data, e.g., multispectral, hyperspectral, LiDAR, or SAR. The fusion of various RS data from different sensors has not received sufficient attention yet. However, it is well known that the complementary use of multimodal RS data offers more complete information on a scene and can result in better performance in many applications [57]. By integrating the complementary information provided by different modality data, such as SAR and multispectral images, traditional methods have been intensively studied by designing handcrafted features based on domain-specific knowledge and exploiting rough fusion strategies. Various feature fusion methods, including supervised learning and unsupervised learning techniques, have been investigated to improve the performance of combining complementary information from SAR and optical images. Early works already proved the effectiveness of combining SAR and optical data with the multi-layer perception (MLP) classifier [19, 21]. Recent Sentinel-1 and Sentinel-2 images are combined to improve the LULC classification accuracy on monsoon regions using a random forest model in [147]. However, these fusion algorithms are a simple concatenation of SAR and optical images and have no capability to learn high-level features.

Thanks to the growth of deep learning, DNNs show great potential in modelling the complicated relationship between different modality data and different downstream applications. Kussul *et al.*[89] first explore the deep CNNs in SAR-optical fusion for LULC classification and demonstrate their superiority with respect to traditional MLP classifiers. In [52], Feng *et al.* propose a multi-branch CNN to improve the classification accuracy in coastal areas by fusing Sentinel-1 and Sentinel-2 images. A multi-temporal W-Net is proposed to integrate Sentinel-1 and Sentinel-2 images in land-cover mapping [54]. Recently, Dino *et al.* [77] propose a deep learning architecture, namely TWINNS, to fuse Sentinel-1 and Sentinel-2 time series data in land-cover mapping. Adrian *et al.* [2] use the 3-dimensional deep learning network to fuse multi-temporal Sentinel-1 and Sentinel-2 data for mapping ten different crop types, as well as water, soil and urban area. In addition to SAR-optical fusion, there are also many other tries in the integration in Lidar-optical. Paisitkriangkrai et al. [124] proposed fusing optical and Lidar data through concatenating deep and expert features as inputs to random forests. Several advanced techniques have subsequently been developed, with the aim of enhancing feature extraction ability. Audebert et al. [6] suggest the use of deep fully convolutional networks to investigate the early and late fusion of LiDAR and multispectral data. Similarly, Chen et al [31], employ a two-branch network to separately extract spectral-spatial-elevation features, followed by utilizing a fully connected layer to integrate these heterogeneous features for final classification. Other novel fusion strategies are also introduced, such as cross-attention module [112], a reconstruction-based network [73],

and a graph fusion network [50]. Additionally, recent studies by Roy et al. [134] propose a multimodal Transformer network to fuse Lidar and hyperspectral images for classification. Similar to Lidar-optical fusion, many researchers also developed the DSM-optical fusion methods, where the DSM was acquired from stereo-optical images.

Their results have shown that deep learning techniques play a significant role in multimodal RS data fusion. However, these techniques in multimodal RS data fusion mainly focused on supervised methods, which are often limited by the availability of annotated data. Labelled remote sensing data are often scarce. The limited access to such labelled data has driven the development of the unsupervised method. These techniques can learn feature representations from unlabeled multimodal data. In [3], Amarsaikhan *et al.* use PCA to enhance the features extracted from SAR-optical images and improve the urban land-cover maps. Fernandez-Beltran *et al.* [53] propose a hierarchical multi-modal probabilistic latent semantic analysis (HMpLSA) model to fuse SAR and multispectral imaging (MSI) data for unsupervised land cover categorization tasks. Similarly, multi-view learning methods also provide a solution to the unsupervised combination of complementary information from SAR-optical images. In [55], Jie *et al.* propose a deep bimodal autoencoder (BDAE) to fuse SAR and multispectral images for classification. In [118], Nielsen *et al.* jointly analyze Sentinel SAR and optical data for change detection using CCA. Based on CCA, Andrew *et al.* [4] propose the deep canonical correlation analysis (DCCA), which learns separate representations for each modality from a shared latent subspace using CNNs.

All these works we mentioned above assume that all modalities are available in inference time. This assumption can greatly limit applications of multi-modal analysis because in practice data collection process may with missing modalities. In this situation, most existing multimodal data fusion methods may fail to deal with incomplete imaging modalities and face severe degradation in downstream tasks. Consequently, a robust multimodal method is highly desired for a flexible and practical remote sensing application with one or more missing modalities. The algorithm used in this situation is called incomplete multimodal learning, which aims at learning methods that are robust with any subset of available modalities at inference. A straightforward strategy for incomplete multimodal learning of remote sensing tasks is synthesizing the missing modalities by generative models [13]. Another stream of methods explores knowledge distillation from complete modalities to incomplete ones [83]. Although promising results are obtained, such methods have to train and deploy a specific model for each subset of missing modalities, which is complicated and burdensome in downstream tasks. Meanwhile, all these methods require complete modalities during the training process. Recent incomplete multimodal learning methods focused on learning a unified model, instead of a bunch of distilled networks, for downstream tasks. In this context,

the modality-invariant fusion embedding across different modalities may contribute to more robust performance, especially when one or more modalities are missing. Transformer is widely used in this task for its flexibility and multimodality modelling abilities. Current works [115, 128] exploited the Transformer for contrastive learning using audio and video data. However, the dedicated Transformer for incomplete multimodal learning of remote sensing tasks has not been carefully tapped yet and cannot allow missing data in the training process.

2.4 Multitemporal and Multimodal Remote Sensing Image Change Detection

Detection of changes in multi-temporal remote sensing (RS) images has been extensively studied in the past decades [95]. Early approaches to change detection in bi-temporal RS images include image algebra, image transformation and image classification methods [15]. These methods have limitations, such as relying on empirical feature extraction algorithms or being sensitive to classification results, which limit their application in change detection. Image algebra methods directly compare image values, such as in the case of change vector analysis (CVA)-based methods [15, 16, 94, 165] that provide spectral change information in terms of magnitude and direction of the spectral change vectors. CVA [20] and its object-based variants are one of the most popular unsupervised single-sensor change detection methods. They calculate the change intensity maps and the change direction for change detection and related classification.

On the other hand, image transformation methods map images into the same feature space for comparison. The most common transformation methods include principal component analysis (PCA) [25], slow feature analysis (SFA) [159], and canonical correlation analysis (CCA) [168]. Another popular method is the combination of PCA and K-means (PCA-KM)[40], which transforms and compares the bi-temporal images in the feature space, and then determines the binary change map using k-means. In [117], Nilsen *et al.* treat the bi-temporal images as multi-view data and proposed the multivariate alteration detection (MAD) based on canonical correlations analysis (CCA), which maximizes the correlation between the transformed features of bi-temporal images for change detection. Wu *et al.* [160] propose a novel change detection method to project the bi-temporal images into a common feature space and detected the changed pixels by extracting the invariant components based on the theory of slow feature analysis (SFA). Unlike single-sensor-based transformation methods, the greatest challenge in cross-sensor change detection is to align the inconsistent

feature representation of different modality images. This requires transforming heterogeneous representation into a common feature space where performing change detection. There are a few traditional methods that focus on this transformation of different modalities. Gong *et al.* [59] propose an iterative coupled dictionary learning method that learns two couple dictionaries for encoding bi-temporal images. Luppino *et al.* [102] propose to perform image regression by transforming images to the domain of each other and to measure the affinity matrice distance, which indicates the change possibility of each pixel. Sun *et al.* [149] develope a nonlocal patch similarity-based method by constructing a graph for each patch and establishing a connection between heterogeneous images.

Supervised image classification methods project image values into different classes at each date and compare directly class labels. This approach, known as post-classification [159] change detection, is widely used in large-scale land-cover change detection. In general, image algebra and transformation methods heavily rely on empirical feature extraction algorithms, while post-classification methods are sensitive to the classification results of each image and to error propagation. These limitations hinder the application of conventional change detection methods.

Deep learning methods have been shown to significantly improve the performance of conventional change detection methods by using DNNs [60] and stochastic gradient descent [14]. One common approach is direct classification, where models are trained using pre-defined labels and then used to classify change and unchanged pixels. For example, Rodrigo et al. [36] present three Unet-based convolutional neural network (CNN) architectures for detecting binary changes between pairs of registered RGB images. In the absence of ground truth, pseudo labels from conventional change detection methods can be used to train models in a self-training paradigm. Zhou et al. [173] propose a self-training algorithm based on pseudo labels for change detection, where the pseudo labels are generated by the traditional CVA approach and used to train a new network end-to-end.

The image transformation approach has also been improved using deep learning, where DNNs are utilized to extract discriminative features. In [49], Du *et al.* developed the slow feature analysis into deep learning methods to calculate the change intensity maps and highlight the changed components in the transformed feature space. Instead of pixel-based analysis, Saha *et al.* [135] use pre-trained CNNs to extract deep spatial-spectral features from multi-temporal images and analyze the features using traditional CVA. Many new techniques also have been developed for extracting discriminative features from bi-temporal RS images, such as generative [122] and discriminative [96] models. Generative models [105, 11, 82, 129, 47] have been adopted to generate features of multi-temporal or multi-sensor image pairs and detect the changes by an explicit comparison of the generated features.

Liu *et al.* propose a stacked autoencoder to extract the temporal change features of multi-temporal SAR images based on superpixels. Bergamasco *et al.* [11] further propose a multilayer convolutional autoencoder (CAE) for multi-temporal Sentinel-1 images change detection in an unsupervised way. Besides autoencoders, generative adversarial networks are also used for change detection tasks. Gong *et al.* [58], for example, treats change detection as a generative learning procedure that connects bi-temporal images and generates the desired change maps. Due to the misregistration between multi-temporal very high resolution (VHR) images, Ren *et al.* [129] use the generative adversarial network (GAN) to generate better-coregistered images, and then generate binary change maps by comparing these generated images explicitly. Dong *et al.* [47] utilize the GAN's discriminator to differentiate samples from bi-temporal images and transform bi-temporal images into more consistent feature representations for direct comparison. Generative models are used not only for homogeneous image change detection but also for heterogeneous image change detection. In [103], Luppino *et al.* combine domain-specific affinity matrices and autoencoders to align the related pixels from multimodal images. Niu *et al.* [119] propose the conditional generative adversarial network (cGAN) to translate two heterogeneous images into a single domain for comparison. Liu *et al.* [98] further use the cycle-consistent adversarial networks (CycleGANs) to learn the mapping relation between heterogeneous image pairs. Discriminative models used in self-supervised change detection include "pretext" tasks and contrastive methods. In [91], Leenstra *et al.* define two pretext tasks for feature representation learning. They further pre-train a discriminative model to extract features from bi-temporal images on these pretext tasks for change detection. Although pretext tasks are widely used in self-supervised learning, they are not a direct way for the change detection task.

The application of deep learning in post-classification change detection can follow two main directions. One is to use a deep learning-based segmentation approach to classify the object of interest on bi-temporal images and then compare them. For example, Nemoto et al. [116] first segment buildings in an urban area and then compare the building maps at two different times to detect changes. Another approach is to perform binary change detection and segmentation of both images simultaneously. Ding et al. [46] propose combining post-classification and direct classification methods using a bi-temporal semantic reasoning network, where the network produces both a change map and two classification maps. These approaches demonstrate the ability of deep learning in deriving changes from image pairs.

The challenge of detecting changes in remote sensing images time-series is compounded by the presence of seasonal noise, which can be difficult to distinguish from true changes. One approach to addressing this challenge is to use graph-based methods [65], which present detected spatiotemporal phenomena as evolution graphs composed of spatiotemporal entities

belonging to the same geographical location in multiple timestamps. Image time-series change detection is often associated with sequential data, making it necessary to evaluate temporal dynamics. The computer vision community has addressed the modelling of temporal relationships among features using recurrent neural networks, which have proven effective for a wide range of applications such as object tracking and action recognition. Long short-term memory networks (LSTM) are particularly effective for such problems, as they mitigate the vanishing gradient problem when dealing with long-term dependencies. The combination of recurrent neural networks and deep learning architectures has also been used for time-series tasks, aiming to produce more useful feature representations by extracting both spatial and temporal information during the learning process. Recent RS image time-series change detection tasks have extensively integrated LSTM techniques. In [114], an LSTM is integrated into a CNN to consider both spatial and temporal features in an end-to-end framework. Sefrin et al. [140] propose combining FCN and LSTM to study land-cover changes using Sentinel-2 images. For high-resolution image change detection, Sun et al. [148] propose using atrous Unet-ConvLSTM to better model multiscale spatial information.

However, supervised methods often require a large number of labelled training samples, which can be difficult to obtain for long image time series. For unsupervised approaches, Saha et al. [136] treat change detection as an anomaly detection problem, using an LSTM network to learn a representation of the image time series. In this method, they used a pretext task of reordering the image sequence. However, the predefined task cannot resist the influence of seasonal noise, which leads to many pseudo-changes in the results. Some researchers have shown that pseudo-labels can help solve this problem. Kalinicheva et al. [81] propose a new framework that combines a graph model and pseudo-labels, using a gated recurrent unit (GRU) AE-based model to associate the changes of consecutive images with different spatial objects. Yang et al. [162] propose an unsupervised time-distance-guided convolutional recurrent neural network (UTRnet) for change detection in irregularly collected images, using a weighted pre-change detection to obtain reliable training samples. However, pseudo labels often have a high level of noise and do not consider temporal information, and the pre-trained model can not adapt to various changes in the image time series.

2.5 Deep Neural Networks and Loss Functions

2.5.1 Introduction of Classic Deep Neural Networks

Multi-layer Perception (MLP) [131] is treated as the fundamental architecture of deep neural networks (DNNs). It consists of a set of simple neurons called perceptrons. The

perceptron computes a single output from multiple real-valued inputs by linearly combining the inputs with the input weights, followed by a non-linear activation function like the hyperbolic tangent $tanh(x)$ or the logistic sigmoid $1/(1+e^{-x})$. Mathematically this can be written as:

$$y = \varphi\left(\sum_{i=1}^{n} w_i x_i + b\right) = \varphi\left(\mathbf{w}^T \mathbf{x} + b\right) \tag{2.1}$$

where $\mathbf{w}$ denotes the vector of weights, $\mathbf{x}$ is the vector of inputs, b is the bias and φ is the activation function. A single perceptron, because of its limited mapping abilities, has

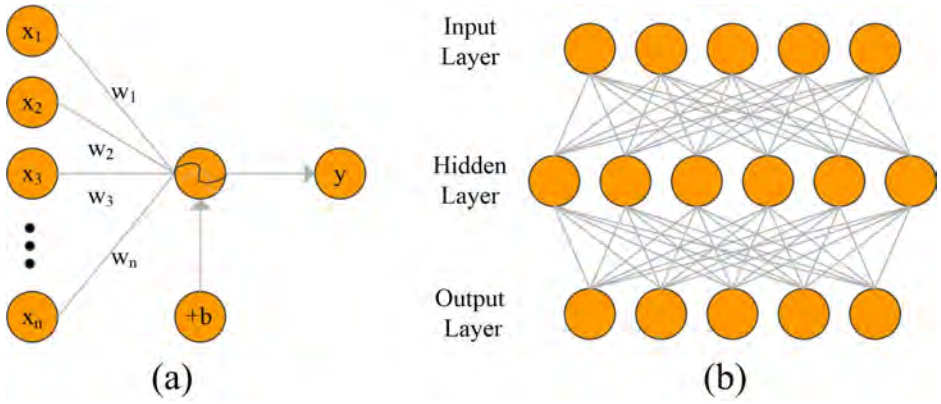

Fig. 2.3 Signal-flow graph of the perceptron and MLP

limited performance. However, stacking perceptrons by layers creates a multilayer perceptron network, which propagates the input signal through each layer. This network can model a wide range of mappings, including strongly and mildly nonlinear mappings. Its signal-flow is shown in Fig. 2.3. Such feedforward networks with a single hidden layer, composed of nonlinear activation functions and a linear output layer, can be expressed mathematically as:

$$y = f(x) = B\varphi(Ax + a) + b \tag{2.2}$$

where x represents the vector of inputs and y represents the vector of outputs. A is the first layer weight matrix, a is the bias vector of the first layer, B is the weight matrix of the second layer, and b is the bias vector of the second layer. Moreover, the function φ denotes an elementwise nonlinearity. The MLP network with only one hidden layer has an astonishingly powerful ability to approximate any continuous function $\mathbf{f} : \mathbb{R}^n \to \mathbb{R}^m$ to any given precision, provided it has enough hidden units.

Convolutional neural networks (ConvNets) [90] are a specific type of artificial neural network (ANN) which have proven remarkably successful in a wide range of computer vision applications. The unique connectivity patterns between the neurons of a ConvNet are

inspired by the structure and function of the visual cortex of animals, which is characterized by neurons that respond to overlapping regions of the visual field. ConvNets are a variant of multilayer perceptron (MLP). A key distinguishing feature of a ConvNet is replacing matrix multiplication with convolution to compute neuron activations. A convolutional neural network comprises several critical components, including convolutional and pooling layers, an activation function, residual connections, and fully connected layers. The convolutional layer is designed to process 2D inputs signals, such as images, using convolution operation with 2D kernel:

$$(K * I)(i, j) = \sum_{m,n} K(m,n) I(i+n, j+m) \tag{2.3}$$

where K is a 2d-kernel. During the computation, the convolution is performed over a 2D grid of pixels in the corresponding layer, with the kernel moving with a predefined stride size s. Padding may be added to the input image to control the size of the output feature. The width of the output can be computed using the equation $W_o = \frac{W_i - k + 2p}{s} + 1$, where W_i is the input image width, k is the kernel size, p is the padding size, and s is the stride size. When working with images having multiple channels, the convolution operation is performed individually on each channel, and the results are combined to form the final output.

The convolution operation employs an activation function φ such as ReLU to induce non-linearity. The activation function applies a point-wise non-linearity to each element in the feature maps after a convolution layer, similar to its role in MLP. Notably, an activation layer has no trainable parameters. Given a kernel K size $k \times k$ and an input image x, the activation can be obtained from the convolution operation by sliding K and computing $z(x) = \varphi(K * x + b)$, where b represents the bias. A pooling layer also called a subsampling layer, is deployed to decrease dimensionality by computing the mean or maximum on image patches. Like convolutional layers, the pooling layers operate on small image patches while following a stride. The output shape is computed by dividing the input shape by the stride. Alternatively, dimensionality can be reduced using a convolutional layer by selecting a stride size over one and zero padding. Pooling has the added benefit of making the network less sensitive to minor translations of the input images.

The depth of CNNs has a significant impact on their performance. More layers are generally better as they allow the network to extract richer features. However, as the deeper and deeper networks, the gradient in backpropagation sometimes vanishes, which results in a degradation problem. In this context, He *et al.* [70] propose the deep residual network using skip connections in each residual unit. Each residual unit can be expressed in the following general form:

$$\begin{aligned} \mathbf{x}_{i+1} &= \mathbf{x}_i + F(\mathbf{x}_i, W_i) \\ \mathbf{x}_{i+1} &= f(\mathbf{x}_{i+1}) \end{aligned} \tag{2.4}$$

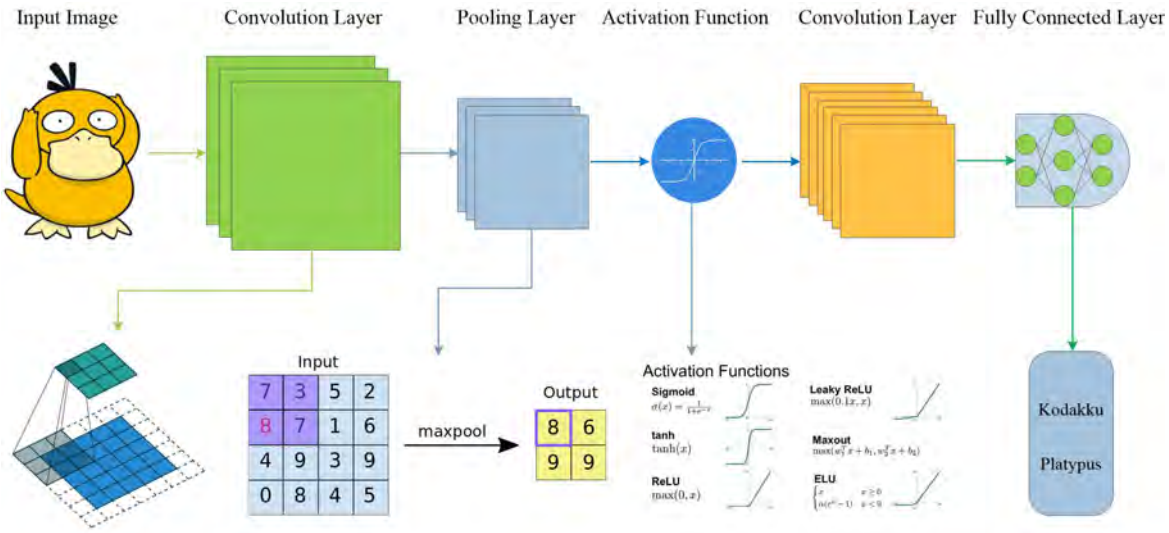

Fig. 2.4 The CNN Components.

where $\mathbf{x}_i$ and $\mathbf{x}_{i+1}$ are the input and output of the $i-th$ residual unit, respectively, $F(\cdot)$ is the residual function, W_i is the weight matrix, $f(\cdot)$ is the activation function.

The fully connected layer is a stack of MLP. Each input is connected to every output node, hence the fully connected layers can associate different combinations of complex features obtained from previous layers with the multiple classes. As more layers are added to the network, the number of trainable parameters and network complexity increases, but this common practice often leads to overfitting issues. To overcome this, the dropout technique is frequently applied. Dropout, which occurs during every epoch, randomly selects neurons based on a defined rate and disables them during training. Since the active neurons change with every epoch, this forces the layer to generalize better, making it more robust. During prediction, all neurons are typically active. Network normalization, often achieved through normalization such as batch normalization, is another popular method. Batch normalization scales the values within a range, improving the performance of gradient calculations during backpropagation, and helping to reduce the training time. It also allows for higher learning rates and has regularizing effects.

Long Short-Term Memory (LSTM). Recurrent Neural Network (RNN) computes the output vector y_t of each tokens x_t in the sequence embeddings by iterating the following equations from $t = 1$ to n:

$$\begin{aligned} h_t &= H\left(W_{xh}x_t + W_{hh}h_{t-1} + b_h\right) \\ y_t &= W_{hy}h_t + b_y \end{aligned} \tag{2.5}$$

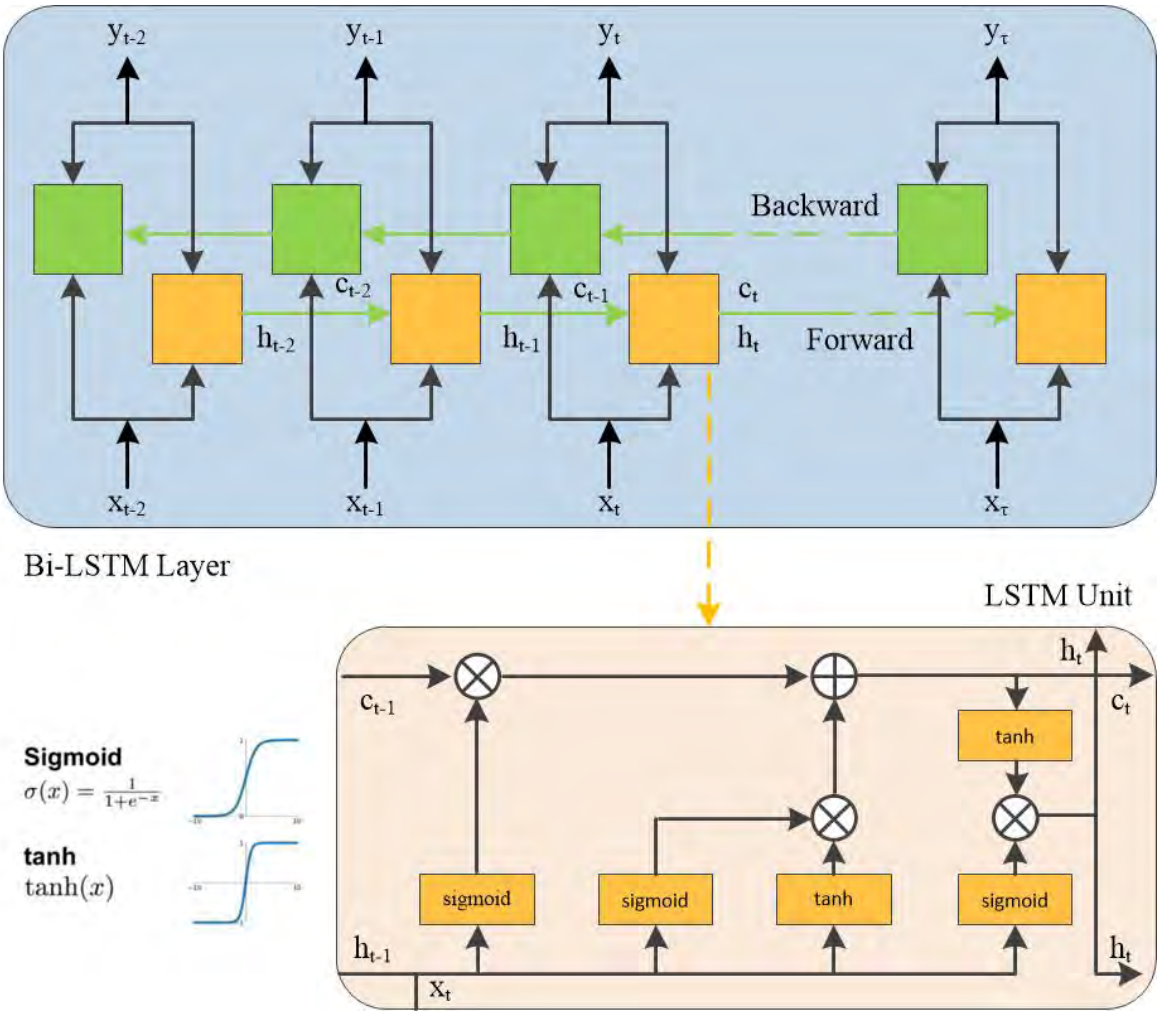

Fig. 2.5 The Bi-LSTM Components.

where h_t is the hidden vector sequence, W denotes weight matrices (w_{xh} is the matrix of the weights connecting the input layer and the hidden layer), b denotes bias vector, and H is activation function of the hidden layer. This equation represents the connection between the previous and the currently hidden states, thus RNNs make use of the previous context in sequence. However, the RNN is not able to use effectively all inputs in sequence due to the vanishing gradient problem. Hence, an improved Long Short-Term Memory [72] architecture was proposed. The LSTM is conceptually defined as an RNN but replaced the hidden layer as memory cells. An LSTM model consists of three gates: forgot f_t, input i_t and output gates

o_t and a cell activation vector c_t.

$$\begin{aligned} i_t &= \sigma\left(W_{xi}x_t + W_{hi}h_{t-1} + W_{ci}c_{t-1} + b_i\right) \\ f_t &= \sigma\left(W_{xf}x_t + W_{hf}h_{t-1} + W_{cf}c_{t-1} + b_f\right) \\ c_t &= f_t c_{t-1} + i_t \tanh\left(W_{xc}x_t + W_{hc}h_{t-1} + b_c\right) \\ o_t &= \sigma\left(W_{xo}x_t + W_{ho}h_{t-1} + W_{co}c_t + b_o\right) \\ h_t &= o_t \tanh\left(c_t\right) \end{aligned} \tag{2.6}$$

where σ is the logistic function and all b are learned biases. The forget gate makes the decision of preserving or removing the existing information by using a sigmoid function. The output of this gate is a value between 0 and 1, where 0 indicates completely getting rid of the learned value and 1 implies preserving the whole value. The input gate makes the decision of whether or not the new information will be added to the LSTM memory. A cell activation vector indicates a vector of new candidate values that will be added to LSTM memory. The combination of the input gate and the forgot gate provides an update for the LSTM memory in which the current value is forgotten and updated. The output gate uses a sigmoid layer to make the decision of what part of the LSTM memory contributes to the output. Then, we put the cell state through tanh (to push the values to be between -1 and 1) and multiply it by the output of the sigmoid gate, so that we only output the parts we decided to.

The Bi-LSTM is an extension of the described LSTM model in which we can use both the past context and future context of an input sequence. It consists of two separate hidden layers. It first computes the forward hidden sequence $\overrightarrow{h_i}$; then it computes the backward hidden sequence $\overleftarrow{h_i}$; finally, it combines both forwards to generate the final output y_t. Let the hidden states h be LSTM blocks, a Bi-LSTM is implemented by the following functions:

$$\begin{aligned} \overrightarrow{h_t} &= H\left(W_{x\overrightarrow{h}}x_t + W_{\overrightarrow{h}\overrightarrow{h}}\overrightarrow{h}_{t-1} + b_{\overrightarrow{h}}\right) \\ \overleftarrow{h_t} &= H\left(W_{x\overleftarrow{h}}x_t + W_{\overleftarrow{h}\overleftarrow{h}}\overleftarrow{h}_{t-1} + b_{\overleftarrow{h}}\right) \\ y_t &= W_{\overrightarrow{h}y}\overrightarrow{h}_t + W_{\overleftarrow{h}y}\overleftarrow{h}_t + b_y \end{aligned} \tag{2.7}$$

Applying the LSTM twice leads to improving learning long-term dependencies and thus consequently will improve the accuracy of the model.

Transformer and Vision Transformer (ViT). Although CNNs have been successful in various tasks, one of their limitations is the receptive field, which restricts their ability to capture dependencies between distant positions. To address this issue in the Natural Language Processing (NLP) domain, the Transformer architecture was proposed as a solution. Unlike CNNs, Transformers [154] rely solely on self-attention and do not use convolution. The

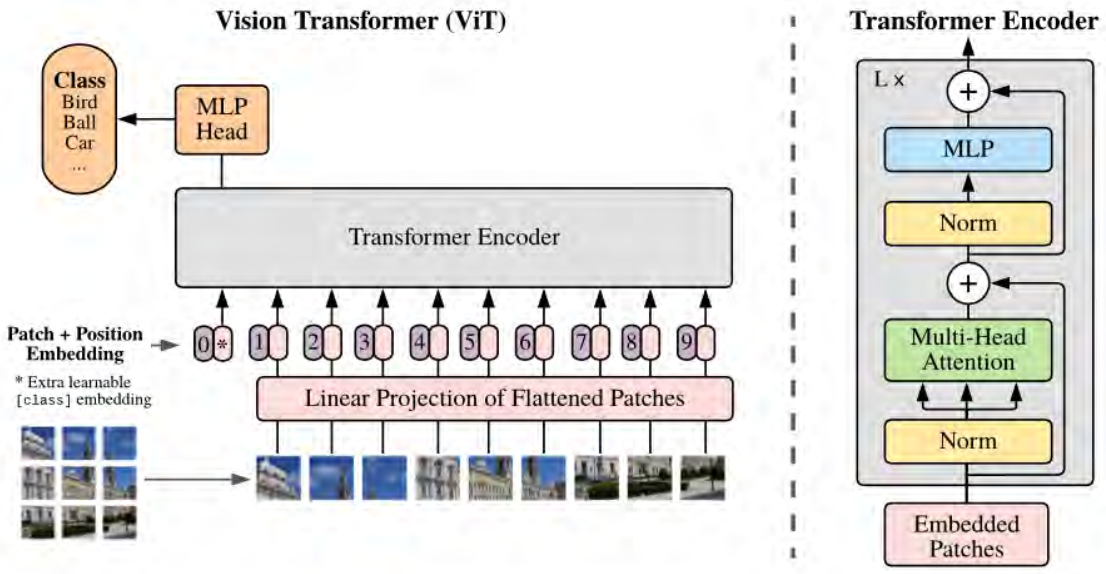

Fig. 2.6 The ViT Components by [48].

attention mechanism is based on a query, key, and value concept, where queries are matched against keys, which are assigned values. The more a query matches a key, the higher the weighting of the corresponding value. This matching is performed using scaled dot-product attention, where the input is represented by matrices Q, K and V, corresponding to queries, keys and values, respectively. By multiplying them together, an attention output is computed, which is a weighted sum of the values:

$$Attention(Q,K,V) = softmax(\frac{QK^T}{\sqrt{d_k}})V \tag{2.8}$$

The input consists of queries and keys of dimension d_k and values of dimension d_v The softmax function is applied to ensure that the weights sum up to one and the scaling factor $\sqrt{d_k}$ has been shown to empirically improve performance. Using multi-head attention, scaled dot-product attention is applied to multiple sets of queries, keys, and values that undergo linear transformations. These inputs are fed in parallel into the attention function, and the results are concatenated into a single matrix that is transformed linearly to the desired output dimensions. This can be expressed as $MultiHead(Q,K,V) = Concat(head_1,...,head_h)W^o$. Each $head_i$ is the result of running scaled dot-product attention on the i^{th} set of transformed queries, keys, and values ($head_i = Attention(QW_i^Q, KW_i^K, VW_i^V)$, where h indicates the number of attention heads). With the setting shown in the previous sections, the transformer would not be able to use position information. To incorporate positional information, a

position embedding is added to the input before being processed by the network, thereby enabling the Transformer to capture dependencies and relationships between positions.

In light of the success of Transformers in NLP, researchers have sought to extend this architecture to the field of vision. However, the self-attention mechanism's computational and memory complexities pose a challenge when dealing with longer image sequences, given that the number of pixels in an image greatly exceeds the number of words in a sentence. The Vision Transformer (ViT) [48] addresses this issue by dividing the image into smaller patches and transforming each patch into a vector embedding, similar to word embedding in NLP. The ViT is a pure Transformer architecture that takes a sequence of image patches as input, where the sequence length is proportional to the number of patches. The input image, denoted by $x \in R^{H\times W\times C}$, is split into $L = \frac{H\times W}{p^2}$ patches, each of dimension $P\times P\times C$. Following BERT [43], a learnable classification embedding x_{class} is prepended to the image sequence along with the added $1D$ positional embeddings E_{pos} to formulate the patch embedding h_0. $\mathbf{E}$ is the patch encoder. The architecture of ViT follows the Transformer architecture:

$$\begin{aligned}
\mathbf{h}_0 &= \left[\mathbf{x}_{\text{class}}\,; \mathbf{x}_p^1\mathbf{E}; \mathbf{x}_p^2\mathbf{E}; \cdots; \mathbf{x}_p^L\mathbf{E}\right] + \mathbf{E}_{pos}, && \mathbf{E} \in \mathbb{R}^{\left(P^2\cdot C\right)\times D}, \mathbf{E}_{pos} \in \mathbb{R}^{(L+1)\times D} \\
\mathbf{h}'_\ell &= \text{MSA}\left(\text{LN}\left(\mathbf{h}_{\ell-1}\right)\right) + \mathbf{h}_{\ell-1}, && \ell = 1,\ldots,L \\
\mathbf{h}_\ell &= \text{MLP}\left(\text{LN}\left(\mathbf{h}'_\ell\right)\right) + \mathbf{h}'_\ell, && \ell = 1,\ldots,L \\
\mathbf{y} &= \text{LN}\left(\mathbf{h}_L^0\right)
\end{aligned} \tag{2.9}$$

This equation applies multi-headed self-attention (MSA). Given learnable matrices W_q, W_k, W_v corresponding to query, key and value representations, a single self-attention head is computed by:

$$Attention_h(X) = softmax(\frac{QK^T}{\sqrt{d_h}})V \tag{2.10}$$

where $Q = XW_q$, $K = XW_k$ and $V = XW_v$. Multi-headed self-attention aggregates information from H self-attention heads by means of concatenation and linear projection: $MSA(X) = concat_{h=1}^{H}[Attention_h(X)]W + b$.

2.5.2 Loss Functions

The goal of a neural network (E) is to optimize a loss function J with respect to the parameters θ over a set of n network inputs $D = (x_1, y_1), \cdots, (x_n, y_n)$, where $x_j \in E_1$ is the j^{th} input data point with an associated response or target $y_j \in E_{L+1}$. Most optimization methods are gradient-based, meaning that we must calculate the gradient of J with respect to the parameters at each layer $i \in [L]$. Let's start to introduce the loss function for both the regression and classification settings. Then we take the derivatives of these loss functions for

a single data point $(x,y) \equiv (x_j, y_j)$ for some $j \in [n]$, and then present error backpropagation in a concise format. Finally, we present algorithms for performing gradient descent steps for both regression and classification [24].

Regression. In the case of regression, the target variable $y \in E_{L=1}$ can be any generic vector of real numbers. Thus, for a single data point, the most common loss function to consider is the squared loss, given by:

$$J_R(x,y;\theta) = \frac{1}{2}\|y - F(x;\theta)\|^2 = \frac{1}{2}\langle y - F(x;\theta), y - F(x;\theta)\rangle. \tag{2.11}$$

In this case, the network prediction $\hat{y}_R \in E_{L+1}$ is given by the network output $F(x;\theta)$. We can calculate the gradient of J_R with respect to the parameter θ_i.

Classification. For the case of classification, the target variable y is often a one-hot encoding, i.e., the component of y corresponding to the class of the data point is equal to 1, and the other components are 0. Therefore, we must constrain the output of the network to be a valid discrete probability distribution. We can enforce this by applying the softmax function σ to the network output $F(x;\theta)$. Then, we can compare this prediction, $\hat{y}_C = \sigma(F(x;\theta))$, to the target variable by using the cross-entropy loss function. For a single point (x,y), we can write the full expression for this loss but with an inner product instead of a sum:

$$J_C(x,y;\theta) = -\langle y, (\log \circ \sigma)(F(x;\theta))\rangle \tag{2.12}$$

We can calculate the gradient of J_C with respect to the parameter θ_i.

2.6 Self-Supervised Discriminative and Generative Models

Supervised learning has demonstrated remarkable accomplishments on a range of tasks such as natural language processing and image understanding. Generally, supervised learning models are trained on specific tasks utilizing a voluminous dataset that has been manually labeled. However, supervised learning is meeting its bottleneck because of the difficulty to have enough labeled samples. The significant reliance on expensive manual labeled data resulted in a generalization error. Therefore, self-supervised learning has emerged as an attractive alternative that has drawn massive attention for its potential to overcome the challenges of data inefficiency and its generalization ability. In this section, we take a look into two typical self-supervised learning methods, autoencoder and contrastive models, for representation learning in vision tasks. Furthermore, these two models belong to two distinct categories based on their objectives: generative and discriminative.

Autoencoder (AE). The concept of Autoencoder (AE) [9] was initially introduced as a method to pre-train artificial neural networks. It is regarded as a directed graphical model that is mainly employed for the purpose of dimensionality reduction. Being a feed-forward neural network, the autoencoder is trained to reconstruct its input at the output layer. It is composed of two networks - encoder and decoder - functionally represented as $h = f_{enc}(x)$ and $x^{'} = f_{dec}(h)$. The primary objective of AE is to minimize the difference between input x and output $x^{'}$ (usually evaluated by mean-square error). It is noteworthy that the linear autoencoder is equivalent to the PCA method.

Denoising AE Model. From an academic standpoint, the fundamental principle underlying denoising autoencoder (DAE) [9] models is the idea that representations should possess the resilience to the presence of noise. Specifically, the masked language model (MLM) can be interpreted as a type of DAE model, exemplified by popular models such as BERT. In the realm of computer vision, DAE models are often deployed to learn robust image representations via the restoration of corrupted images, and the concept has been notably applied to the Vision Transformer (ViT). A particularly impactful contribution to the field in this respect is the work on Masked AutoEncoder (MAE) [68] .

Variational AE Model. The variational autoencoder (VAE) [9] model posits an underlying latent representation that generates the observed data. To approximate the posterior distribution over unobserved variables $Z = z_1, z_2, \cdots, z_n$ given some data X, the VAE employs a variational distribution $q(z|x)$.

$$p(z|x) \approx q(z|x) \tag{2.13}$$

During training, variational inference and the evidence lower bound (ELBO) are used to maximize the log-likelihood of the observed data. The ELBO is defined as:

$$\log p(x) \geq -D_{KL}(q(z \mid x) \| p(z)) + \mathbb{E}_{\sim q(z|x)}[\log p(x \mid z)] \tag{2.14}$$

where $D_{KL}(q(z \mid x) \| p(z))$ denotes the the Kullback-Leibler (KL) divergence of $q(x)$ from $p(x)$, $p(x)$ represents evidence probability, $p(z)$ is prior and $p(x|z)$ is likelihood probability. Within the autoencoder framework, the first term of ELBO serves as a regularizer to enforce the posterior's ability to reconstruct the input data based on the latent variables. VAE assumes the prior $p(z)$ and the approximate posterior $q(z|x)$ both follow Gaussian distributions. Specifically, $p(z) \sim N(0,1)$. Furthermore, the reparameterization trick is utilized for modelling approximate posterior $q(z|x)$. Assume $z \sim N(\mu, \sigma^2)$, $z = \mu + \sigma\varepsilon$, where $\varepsilon \sim N(0,1)$. The decoder network is then used to reconstruct the input data based on the calculated latent variable z, with parameterized μ and σ.

Contrastive Learning. [64] Given the joint distribution $P(X,Y)$ of the input X and target Y, the generative model calculates the $p(X|Y=y)$ by:

$$p(X \mid Y=y) = \frac{p(X,Y)}{p(Y=y)} = \frac{p(X,Y)}{\int_x p(Y,X=x)} \tag{2.15}$$

while the discriminative model tries to model the $P(Y|X=x)$ by:

$$p(Y \mid X=x) = \frac{p(X,Y)}{p(X=x)} = \frac{p(X,Y)}{\int_y p(Y=y,X)} \tag{2.16}$$

Notice that most of the representation learning takes hope to model relationships between different views. Thus for a long time, people believed that the generative model is the only choice for representation learning. However, contrastive learning brings the breakthroughs, such as InfoMax [123], MoCo [69] and SimCLR [26]. Contrastive learning aims to "learn to compare" through a Noise Contrastive Estimation (NCE) [64] objective formatted as:

$$L = \mathbb{E}_{x,x^+,x^-}\left[-\log\left(\frac{e^{f(x)^T f\left(x^+\right)}}{e^{f(x)^T f(x^+)} + e^{f(x)^T f^- x^-)}}\right]\right. \tag{2.17}$$

where x^+ is similar to x, x^- is dissimilar to x and f is an encoder (representation function). The similarity measure and encoder may vary from task to task, but the framework remains the same. With more dissimilar pairs involved, we have the InfoNCE formulated as:

$$L = \mathbb{E}_{x,x^+,x^k}\left[-\log\left(\frac{e^{f(x)^T f\left(x^+\right)}}{e^{f(x)^T f(x^+)} + \sum_{k=1}^{K} e^{f(x)^T f\left(x^k\right)}}\right]\right. \tag{2.18}$$

2.7 Conclusion

This chapter presents a comprehensive overview of the remote sensing images and products utilized in this thesis, along with an extensive review of remote sensing image change detection and data fusion. Furthermore, the last two sections introduce neural networks and self-supervised generative and discriminative models.

Regarding remote sensing images, our focus lies on high-resolution multispectral images, specifically Landsat-8 and Sentinel-2 images. We also delve into the characteristics of SAR images, with particular emphasis on Sentinel-1 backscattering images and their pre-processing techniques. These SAR and multispectral images will be employed in multi-temporal remote sensing image change detection and data fusion tasks. Additionally, we

introduce remote-sensing products such as DEM/DSM and LULC maps, which play a pivotal role in the multimodal fusion of remote sensing data discussed in Chapter 6.

We conduct a thorough review of state-of-the-art methods in multimodal remote sensing data fusion and underscore the significance of self-supervised pre-training for multimodal remote sensing data fusion, as well as incomplete multimodal learning in downstream tasks. For remote sensing image change detection, we review unsupervised change detection methods in both multitemporal and multimodal settings, thereby highlighting the primary direction of our thesis work.

At last, we introduce three key neural networks, namely Convolutional Neural Networks (CNNs), Long Short-Term Memory (LSTM), and Transformers, along with two loss functions, regression and classification, which form the foundation for the approaches proposed in this thesis. Finally, we present two fundamental self-supervised methods, autoencoders and contrastive learning, which serve as the underlying paradigms for our thesis work.

Chapter 3

Self-Supervised Bi-temporal RS image Change Detection

In this chapter, a self-supervised change detection approach based on an unlabeled multi-view setting is proposed to overcome the limitation of CNN-based generative models on unsupervised binary change detection. We present a novel approach to perform unsupervised change detection in both single-sensor and cross-sensor scenarios based on a multi-view contrastive learning method. To overcome the limitation of the patch-based image processing algorithm, we further propose a pixel-wise contrastive framework which uses the contrastive loss on superpixels to get fine-grained change maps and exploits an uncertainty method to enhance the temporal robustness. Results demonstrate both improvements over state-of-the-art unsupervised methods and that the proposed approach narrows the gap between unsupervised and supervised change detection methods.

3.1 Self-supervised Change Detection in Multi-view Remote Sensing Images

In this section, we present the proposed approach to multi-temporal and multi-sensor remote sensing image change detection based on self-supervised learning and image patches.

3.1.1 Introduction

Change maps are one of the most important products of remote sensing and are widely used in many applications including damage assessment and environmental monitoring. The spatial and temporal resolutions play a crucial role in obtaining accurate and timely change

detection maps from multi-temporal images. In this context, irrelevant changes, such as radiometric and atmospheric variations, seasonal changes of vegetation, and changes in the building shadows, which are typical of multi-temporal images, limit the accuracy of change maps. In the past decades, many researchers developed techniques that directly compare pixels values of multi-temporal images to get the change maps from coarse resolution images [22, 20, 15], assuming that the spectral information of each pixel can completely characterize various underlying land-cover types. With the increase of spatial and spectral resolutions of remote sensing images, the use of spectral information only is often not enough to distinguish accurately land-cover changes. Many supervised and unsupervised techniques were developed to determine the land-cover changes by jointly using spatial context and spectral information. Recently, deep learning techniques and in particular Convolutional Neural Networks (CNNs) methods [135] have been widely used in this domain. CNNs allow one to model high-level features from images in terms of spatial and spectral information, achieving state-of-the-art results in a supervised way [166].

Most of the past works are limited to the use of single-modality images. Cross-domain change detection has not received sufficient attention yet. Current Earth Observation satellites provide a vast amount of multi-view observations from different sensors and at different times. On the one hand, images taken by different types of sensors can improve the time resolution thus satisfying the requirement of specific applications with tight constraints. A possible example of this is the joint use of Sentinel-2 and Landsat-8 images for regular and timely monitoring of burned areas [133]. However, the differences in acquisition modes and sensor parameters present a big challenge for traditional methods. On the other hand, multimodal data are complementary to the use of single modality images and their use becomes crucial especially when only images from different sensors are available in some specific scenarios. This could be the case of emergency management when, for example, optical and SAR images could be jointly exploited for flood change detection tasks [75]. In this scenario, methods capable of computing change maps from images of different sensors in the minimum possible time can be very useful. This has led to the development of multi-source change detection methods, which can process either multi-sensor or multi-modal images.

The recent success of deep learning techniques in change detection is mainly focused on supervised methods [36, 127, 126], which are often limited by the availability of annotated datasets. Especially in multi-temporal problems, it is expensive and often not possible to obtain a large amount of annotated samples for modeling change classes. Thus, unsupervised methods are preferred to supervised ones in many operational applications. The limited access to labelled data has driven the development of unsupervised methods, such as Generative Adversarial Network (GAN)[61] and Convolutional AutoEncoder (CAE)[108], which are

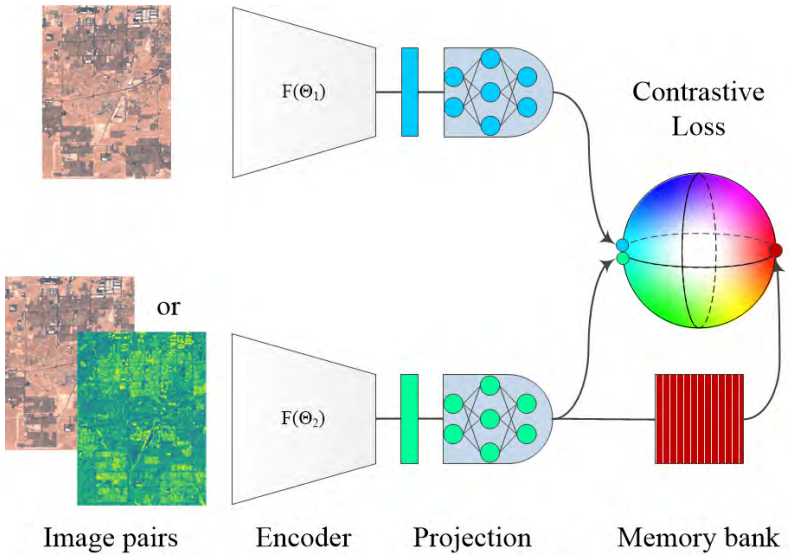

Fig. 3.1 The pre-training part of the proposed approach to change detection for bi-temporal remote sensing image pairs. In the cross-sensor setting, the image pair consists of two images acquired by different types of sensors and the architecture of the network is symmetric with each side consisting of an encoder and a projection layer. In the single-sensor setting, the image pair consists of bi-temporal images acquired by the same sensor and two symmetric subnetworks that share almost identical architectures.

currently among the most used deep learning methods in unsupervised change detection tasks. Recent research in self-supervised learning [151, 69] encourages the network to learn more interpretable and meaningful feature representations in CV tasks, where they outperformed the generative counterparts. To overcome the drawbacks of generative models, in this section, we exploit contrastive learning in multi-view remote sensing image change detection. We present a novel general approach to perform unsupervised change detection in both single-sensor and cross-sensor scenarios that are based on a multi-view contrastive learning method [151]. Rather than training generative models on a predefined task, the proposed approach is trained end-to-end on large unlabeled images by minimizing the distance between features directly from bi-temporal images. To this purpose, a pseudo-Siamese network (which exploits ResNet-34 as the backbone) is trained to regress the output between two branches that were pre-trained in a contrastive way on a large archived multi-view image. Then, we introduce a change score that can accurately model the feature distance between bi-temporal images. Changes are identified when there is a significant disagreement between the feature vectors of the two branches.

3.1.2 Methodology

Pseudo-Siamese Network

Siamese networks [17] is the most used model in entity comparison. However, the comparison of cross-sensor image pairs can not be performed by Siamese networks directly for their different imaging mechanism. Siamese networks share identical weights in two branches, while cross-sensor image pairs have dissimilar low-level features. Hence, the pseudo-Siamese network is used as the model architecture for cross-sensor image change detection. It has two branches that share the same architecture except for the input channel, but with different weights. For single-sensor images, we propose to use the mean teacher network [150] as the architecture of our model. The mean teacher is a common pseudo-Siamese network used in self-supervised learning, which uses an exponential moving average (EMA) weight to produce a more accurate model than using the same weights directly in the single-sensor images setting. In this way, the model has a better intermediate feature representation by aggregating the information of each step. Fig. 3.1 (a) shows the architecture used in this work, where two branches are designed to extract the features of bi-temporal image pairs. In this work, the ResNet-34 [70] is adopted as the backbone of the two branches and the input channels are changed for adapting to the image pairs, i.e., the polarization of SAR image patches and the spectral bands of optical image patches. In particular, we change the parameters of the strider from 2 to 1 in the third and fourth layers of the backbone for adapting the network to the relatively small input size. In greater detail, the bi-temporal image pairs are passed through the unshared branches and are then modelled in output from the related feature vectors. The output feature vectors of the two branches are normalized and then used to compute the similarity with each other and the negative samples of the batch. Finally, the model parameters are updated by minimizing a loss function.

For SAR-Optical fusion data, we implement BYOL by using the mean teacher network [150] as the architecture of our model (Fig. 3.1 (b)). It can be perceived as a kind of pseudo-Siamese network, which consists of two identical branches (online network and target network). However, the weight of one branch is the expositional moving average (EMA) weight of the other.

Self-supervised Learning Approach

Contrastive learning is a popular methodology for unsupervised feature representation in the machine learning community [151, 120]. The main idea behind the contrastive loss is to find a feature representation that attributes the feature distance between different samples. For change detection, let us consider each bi-temporal image pair $\{I_1^i, I_2^i\}_{i=1,2,\ldots,N}$ on a given

scene i, which is considered as a positive pair sampled from the joint distribution $p(I_1^i, I_2^i)$. Another image pair $\{I_1^i, I_2^j\}$ taken from a different scene is considered as a negative pair sampled from the product of marginals $p(I_1^i)p(I_2^j)$. The method introduces a similarity function, $h_\theta(\cdot)$, which is used to model the feature distance between positive and negative pairs. The pseudo-Siamese network is trained to minimize the $L_{contrast}^S$ defined as:

$$L_{contrast}^S = -\mathop{\mathbb{E}}_{S}\left[\log \frac{h_\theta(I_1^1, I_2^1)}{\sum_{j=1}^N h_\theta(I_1^1, I_2^j)}\right] \tag{3.1}$$

where (I_1^1, I_2^1) is a positive pair sample, $(I_1^1, I_2^j | j \geq 1)$ are negative pair samples and $S = \{I_1^1, I_2^1, I_2^2, \cdots, I_2^{N-1}\}$ is a set that contains $N-1$ negative samples and one positive sample. During the training, positive image pairs are assigned to a higher value whereas negative pairs to a lower value. Hence, the network represents positive pairs at a close distance whereas negative pairs at a high distance. The self-supervised method takes different augmentations of the same image as positive pairs and negative pairs sampled uniformly from different scenes. For change detection, we can construct bi-temporal image sets S_1 and S_2 by fixing one set and enumerating positives and negatives from the other set. This allows us to define a symmetric loss as:

$$L(S_1, S_2) = L_{\text{contrast}}^{S_1} + L_{\text{contrast}}^{S_2} \tag{3.2}$$

In practice, the Noise-Contrastive Estimation [64] method is used to make a tractable computation of (3.2) when N is extremely large. This multi-view contrastive learning approach makes the unsupervised change detection possible.

Change Detection

The change detection strategy described in this subsection is based on the feature learned by the contrastive learning method. Let $S = \{I_1, I_2, I_3, ..., I_n\}$ be a dataset of either single-sensor or cross-sensor multi-temporal remote sensing images. Our goal is to detect changes between satellite images from different dates. As mentioned before, most changes of interest are those relevant to human activities, while the results are easily affected by irrelevant changes, such as seasonal changes. Other relevant changes are usually rare, whereas irrelevant changes are common during a long period. This means that, under this assumption, the features of relevant changes can be derived from the unchanged features. To this purpose, the models are trained to regress the features of images acquired at different dates. As shown in Fig. 3.2, here we use the considered contrastive learning algorithm to get features of either single-sensor or cross-sensor multi-temporal images. After training, a change intensity map can be derived by assigning a score to each pixel indicating the probability of change.

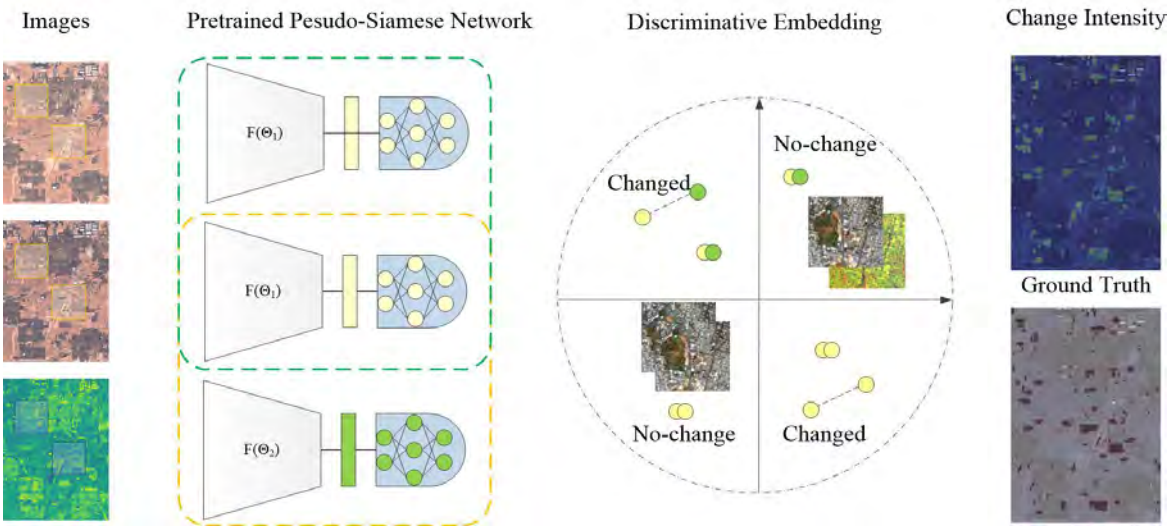

Fig. 3.2 Schematic overview of the proposed change detection approach (SSL). Input images are fed through the pre-trained pseudo-Siamese network that extracts feature vectors from single-sensor or cross-sensor bi-temporal image patches. Then, change intensity maps are generated by estimating regression errors for each pixel and the final binary change map is obtained by setting a threshold.

During the network training, images acquired by the different sensors or at different dates are treated as two-views in our approach. Image patches centred at each pixel are fed in input to the network, and the output is a single feature vector for each patch-sized input. In detail, given an input image $\mathbf{I} \in \mathbb{R}^{w \times h}$ of width w, height h, we can get a feature vector $T(r,c)$ of a square local image region with a side length p for each image pixel at row r and column c. During the inference, the model provides as output a feature map that is generated by a given size of input image patches. Let $T_1(r,c)$ and $T_2(r,c)$ denote the feature vectors at the row r and column c for the considered bi-temporal images. The change intensity map is defined as the pair-wise regression error $e(r,c)$ between the feature vectors of bi-temporal images:

$$e(r,c) = \|T_1(r,c) - T_2(r,c)\|_2^2 \tag{3.3}$$

It is worth noting that the proposed model allows the use of different input sizes. To prevent the possible degeneration of the detection accuracy at a given input size, we recommend using a larger input size when the trial on a smaller input size fails.

One can see from Fig. 3.2 that pixels can be classified as changed and unchanged by thresholding the feature distance in the change intensity map. In this case, two strategies are considered. The simplest strategy is to choose the opposite minimum value of standardized intensity maps as the threshold value. An alternative strategy is the Robin thresholding

method [132], which is robust and suitable for long-tailed distribution curves. In this method, the threshold value is the "corner" on the distribution curve of the intensity map and the maximum deviation from the straight line drawn between the endpoints of the curve. In our technique, the threshold value is determined by the first strategy if the absolute difference of these two threshold values is smaller than half of their average value. Otherwise, the threshold value is determined by the Robin thresholding method.

3.1.3 Experimental Results

In this section, we first present the considered datasets, then the state-of-the-art change detection methods used in the comparison, and finally conduct a thorough analysis of the performance of different approaches and the analysis of robustness and efficiency.

Description of Datasets

We developed our experiments on five different datasets including two single-sensor datasets and three cross-sensor datasets. All remote sensing images in this work are raw images from the google earth engine (GEE) and without any specific pre-processing.

OSCD_S2S2/_S1S1/_S1S2/_L8S2: The Onera Satellite Change Detection (OSCD) dataset [37] was created for bi-temporal change detection using Sentinel-2 images acquired between 2015 and 2018. These images have a total of 13 bands with a relatively high resolution (10 m) for Visible (VIS) and near-infrared (NIR) band images and 60 m resolution for other spectral channels. The images of this dataset include urban areas and present the change type of urban growth and changes. The dataset consists of 24 pairs of multispectral images and the corresponding pixel-wise ground truth acquired in different cities and including different landscapes. The pixel-wise ground truth labels, which were manually annotated, focus on urban growth and built-up changes and contain some errors on the identification of bare lands. At the original supervised setting, 14 pairs were selected for the training set and the rest 10 pairs were used to evaluate the performance of methods. To use this dataset in self-supervised training, we downloaded additional Sentinel-2 images in the same location as the original bi-temporal images between 2016 and 2020. We considered images from each month to augment existing image pairs. Similarly, Landsat-8 multi-temporal images and Sentinel-1 ground range detected (GRD) image products are also provided in this dataset corresponding to the given Sentinel-2 scenes. The Landsat-8 images have nine channels covering the spectrum from deep blue to shortwave infrared and two long-wave infrared channels and their resolution range from 15 m to 100 m. The Sentinel-1 GRD products have been terrain corrected, multi-looked, and transformed to the ground range

and geographical coordinates. They consist of two channels including Vertical-Horizontal (VH) and Vertical-Vertical (VV) polarization as well as of additional information on the incidence angle. To use this dataset for multi-view change detection, we separate it into four sub-datasets: OSCD_S2S2 (225 pairs), OSCD_S1S1 (577 pairs), OSCD_S1S2 (334 pairs) and OSCD_L8S2 (156 pairs). These datasets are composed of single-sensor images (OSCD_S2S2, OSCD_S1S1) and cross-sensor images (OSCD_L8S2, OSCD_S1S2). To keep consistency with previous research, 10 image pairs of these four datasets corresponding to the OSCD test image pairs are treated as the test set to evaluate the performance of different methods, and image pairs acquired on other scenes and on each month of four years are used for the self-supervised pre-training. In practice, it is impossible to acquire the test image pairs of OSCD_S1S1, OSCD_L8S2, and OSCD_S1S2 at the same time as the OSCD_S2S2. Hence, we only obtained these image pairs at the closest time to OSCD_S2S2 test image pairs.

Flood in California: The California dataset is also a cross-sensor data set that includes a Landsat-8 (multi-spectral) and a Sentinel-1 GRD (SAR) image. The multispectral and SAR images are acquired on 5 January 2017 and 18 February 2017, respectively. The dataset represents a flood occurred in Sacramento County, Yuba County, and Sutter County, California. The ground truth was extracted from a Sentinel-1 SAR image pair where the pre-event image is acquired approximately at the same time as the Landsat-8 image. However, we realized that the ground truth in [102] contains many mistakes. Hence, we updated the reference data with the PCC method according to bi-temporal Sentinel-1 images. The other three image pairs of Sentinel-1 and Landsat-8 images of the same scene acquired in 2017 and 2018, respectively, were used for the self-supervised pre-training of the proposed approach.

S1-2 fusion: We considered the OSCD dataset and use Sentinel-2 as well as the corresponding Sentinel-1 GRD images, where the two polarization images (vertical-horizontal and vertical-vertical) of Sentinel-1 GRD products were used to complement the Sentinel-2 images. The dataset is split into the training part (14 pairs) used for the self-supervised pre-training and the test part (10 pairs) used for evaluation. Each part consists of Sentinel-1 and Sentinel-2 images of the given scene.

Experimental Settings

Literature Methods for Comparison: We considered different state-of-the-art methods for comparisons with the proposed approach on the five datasets mentioned above. On the first two optical data sets (OSCD_S2S2 and OSCD_L8S2), the proposed SSL approach was compared with two unsupervised deep learning approaches (DSFA [49] and CAA [103]) and two deep supervised methods (FC-EF [36] and FC-EF-Res [35]). Code-Aligned

Autoencoders (CAA) is a deep unsupervised methodology to align the code spaces of two autoencoders based on affinity information extracted from the multi-modal input data. It allows for achieving a latent space entanglement even when the input images contain changes by decreasing the interference of changed pixels. However, it degrades its performance when only one input channel is considered. It is also well suited for single-sensor change detection, as it does not depend on any prior knowledge of the data. Fully convolutional-early fusion (FC-EF) is considered for the supervised change detection method on the OSCD dataset. In this method, the bi-temporal image pair are stacked together as the input. The architecture of FC-EF is based on U-Net [130], where the skip connections between the encoder and decoder help to localize the spatial information more precisely and get clear change boundaries. FC-EF-Res is an extension of FC-EF with residual blocks to improve the accuracy of change results. In addition, it is worth noting that the first dataset (OSCD_S2S2) has previously been extensively used in other works. Hence, we also compare our results with those of some conventional methods [37] (Log-ratio, GLRT and Image difference), an unsupervised deep learning method (ACGAN [137]) and supervised deep learning techniques (FC-Siam-conc and FC-Siam-diff [37]) reported in previous papers. On the Sentinel-1 SAR images dataset, only unsupervised methods (DSFA, SCCN, and CAA) are used for comparison. Note that some change information present in multi-spectral images is not detectable in SAR images, hence we did not use supervised methods on them. On the two cross-sensor remote sensing image datasets (OSCD_S1S2 and California), two state-of-the-art methods are used for comparisons, including the symmetric convolutional coupling network (SCCN) and CAA. Considering that only significant changes in the backscattering of SAR images can be detected, we only consider the LasVegas site in the OSCD_S1S2 data set.

Implementation details: During the training on a single-sensor data set, we randomly composed all images acquired at adjacent months into pairs as the input. While SAR/multi-spectral image pairs acquired in the same month have been used as the input of multi-sensor fusion pairs. After finishing the training process, the test image pairs are fed into the pre-trained network and then the related change intensity maps are derived. For the supervised method (FC-EF and FC-EF-Res), we used the 14 bi-temporal training images considered in the previous work [35]. In the self-supervised and supervised method, we use four channels (VIS and NIR) in Landsat-8 and Sentinel-2 images, while two polarizations (VH and VV) in Sentinel-1 images. CAA and SCCN methods require cross-sensor image pairs having the same number of input channels. To keep consistency with the four input channels of multi-spectral images, we augmented Sentinel-1 images with the plus and minus operation between the two polarizations as the other two channels.

Evaluation Criteria: To appraise the different methods presented above, five evaluation metrics (precision (Pre), recall (Rec), overall accuracy (OA), F1 score and Cohen's kappa score (Kap)) are used in this paper. We simply classify the image pixels into two classes by setting an appropriate threshold value according to the presented strategy and analysing them with reference to the ground truth map. Then, the number of unchanged pixels incorrectly flagged as change is denoted by FP (false positive) and the number of changed pixels incorrectly flagged as unchanged is denoted by FN (false negative). In addition, the number of changed pixels correctly detected as change is denoted by TP (true positive) and the number of unchanged pixels correctly detected as unchanged is denoted by TN (true negative). From these four quantities, the five evaluation metrics can be defined as :

$$Pre = \frac{TP}{TP+FP} \tag{3.4}$$

$$Rec = \frac{TP}{TP+FN} \tag{3.5}$$

$$F_1 = \frac{2Pre \cdot Rec}{Pre+Rec} \tag{3.6}$$

$$OA = \frac{TP+TN}{TP+TN+FP+FN} \tag{3.7}$$

$$Kap = \frac{OA-PE}{1-PE} \tag{3.8}$$

$$\begin{aligned} \text{PE} = &\frac{(TP+FP)\cdot(TP+FN)}{(TP+TN+FP+FN)^2} \\ &+ \frac{(FN+TN)\cdot(FP+TN)}{(TP+TN+FP+FN)^2} \end{aligned} \tag{3.9}$$

Obviously, a higher value of Pre results in fewer false alarms, and a higher value of Rec represents a smaller rate of incorrect detections. The overall accuracy OA is the ratio between correctly detected pixels and all pixels of the image. However, these three metrics will give a misleading overestimate of the result when the amount of changed pixels is a small fraction of the image. $F1$ score and Kap can overcome the problem of Pre and Rec and better reveal the overall performance. Note that large $F1$ and Kap values represent better overall performance.

Results on Single-sensor Data Sets: We first evaluate the change detection performance of the proposed approach and state-of-the-art methods (DSFA, CAA and supervised methods) applied to the single-sensor change detection scenario. This includes bi-temporal Sentinel-2 images (OSCD_S2S2 test dataset) and bi-temporal Sentinel-1 images (OSCD_S1S1 test dataset). The performance metrics obtained on the OSCD_S2S2 test dataset are reported in

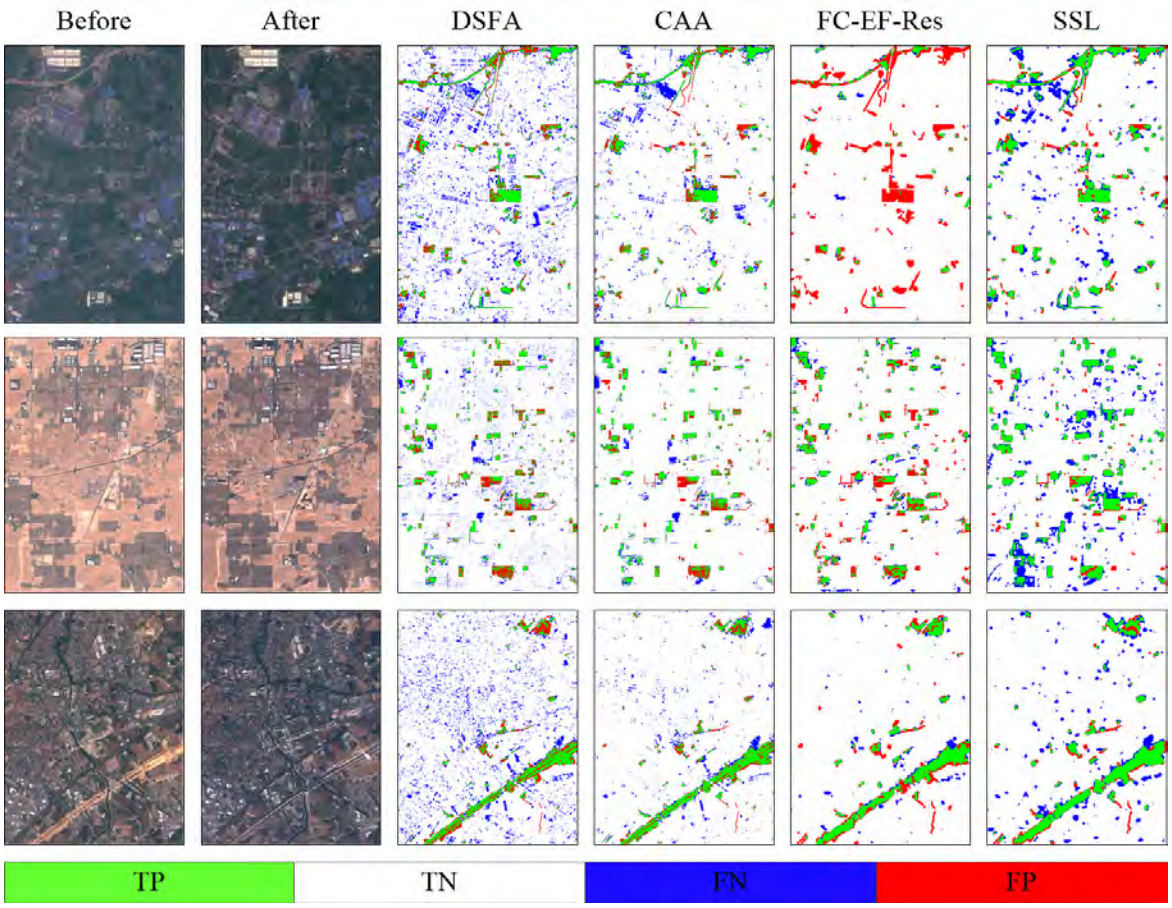

Fig. 3.3 Examples of change detection results on OSCD_S2S2, organized in one row for each location. Col. 1: pre-event image (Sentinel-2); Col. 2: post-event image (Sentinel-2). Change maps obtained by: DSFA (Col. 3), CAA (Col. 4), FC-EF-Res (Col. 5), and the proposed SSL (Col. 6).

Table 3.1. As expected the FC-EF and FC-EF-Res supervised methods applied to raw images achieved a better performance than most unsupervised methods except the proposed SSL approach on most metrics. Among all unsupervised methods, the proposed SSL approach (with the input size of 8 pixels) with an OA of 93 % and a Kappa coefficient of 0.48, obtained the best performance on all five metrics and the best performance among all methods (including the supervised ones) implemented in this work. Although the two supervised methods performed better than other methods on most metrics, they have worse performance on Recall than the proposed SSL approach. In addition, the results of other unsupervised methods (i.e., ACGAN, Image difference, GLRT, and Log-ratio) and supervised methods (i.e., Siamese and EF) on VIS and NIR channels in [37] are reported in the table. They are all worse than those of the proposed SSL approach. The results of other supervised methods (i.e., FC-EF*, FC-EF-Res*, FC-Siamese-Con* and FC-Siamese-Diff*) applied to carefully

processed RGB channel images are reported in the last rows of Table 3.1. Their accuracies on most metrics are slightly better than those of the proposed SSL approach, but they can not be achieved when working on raw images as a high registration precision is required. Indeed, in the related papers, multi-temporal images are carefully co-registered using GEFolki toolbox to improve the accuracy of change maps [37]. On the contrary, the proposed SSL approach is based on image patches where the registration precision of the Sentinel system is enough for obtaining a good change map.

Besides the quantitative analysis, we also provide a visual qualitative comparison in Fig. 3.3, where the TP, TN, FN and FP pixels are colored in green, white, blue and red, respectively. One can see that change maps provided by DSFA and CAA are affected by a significant salt-and-pepper noise where plenty of unchanged buildings are misclassified as changed ones. This is due to the lack of use of spatial context information in these methods. This issue is well addressed by the proposed SSL approach and the FC-EF-Res supervised method, which provides better maps. Most of the changed pixels are correctly detected in the proposed SSL approach but with more false alarms than in the supervised FC-EF-Res method. Note that this is probably due to some small changes that are ignored in the ground truth. Nonetheless, since these results are processed in patches, some small objects are not classified correctly and false alarms on boundaries of buildings are provided by the proposed SSL approach. Instead, the change maps obtained by the FC-EF-Res method are in general more accurate and less noisy due to the use of spatial-spectral information in U-Net and

Table 3.1 Quantitative evaluations of different approaches applied to the OSCD_S2S2 dataset.

Type	Method	Pre(%)	Rec(%)	OA(%)	F1	Kap
Unsupervised	Prop. SSL	**40.44**	**69.10**	**93.00**	**0.51**	**0.48**
	DSFA	26.77	54.24	92.63	0.36	0.32
	CAA	23.49	52.96	91.66	0.33	0.29
	ACGAN[44]	-	64.63	77.67	-	-
	Img. Diff[41]	-	63.42	76.12	-	-
	GLRT[41]	-	60.48	76.25	-	-
	Log-ratio[41]	-	59.68	76.93	-	-
Supervised	FC-EF	**55.34**	39.48	95.13	0.46	**0.44**
	FC-EF-res	54.97	38.39	95.10	0.45	0.43
	Siamese[41]	21.57	79.40	76.76	0.34	-
	EF[41]	21.56	**82.14**	83.63	0.34	-
	FC-EF*[42]	44.72	53.92	94.23	0.49	-
	FC-EF-Res*[42]	52.27	68.24	**95.34**	**0.59**	-
	FC-Siamese-Con*[42]	42.89	47.77	94.07	0.45	-
	FC-Siamese-Diff*[42]	49.81	47.94	94.86	0.49	-

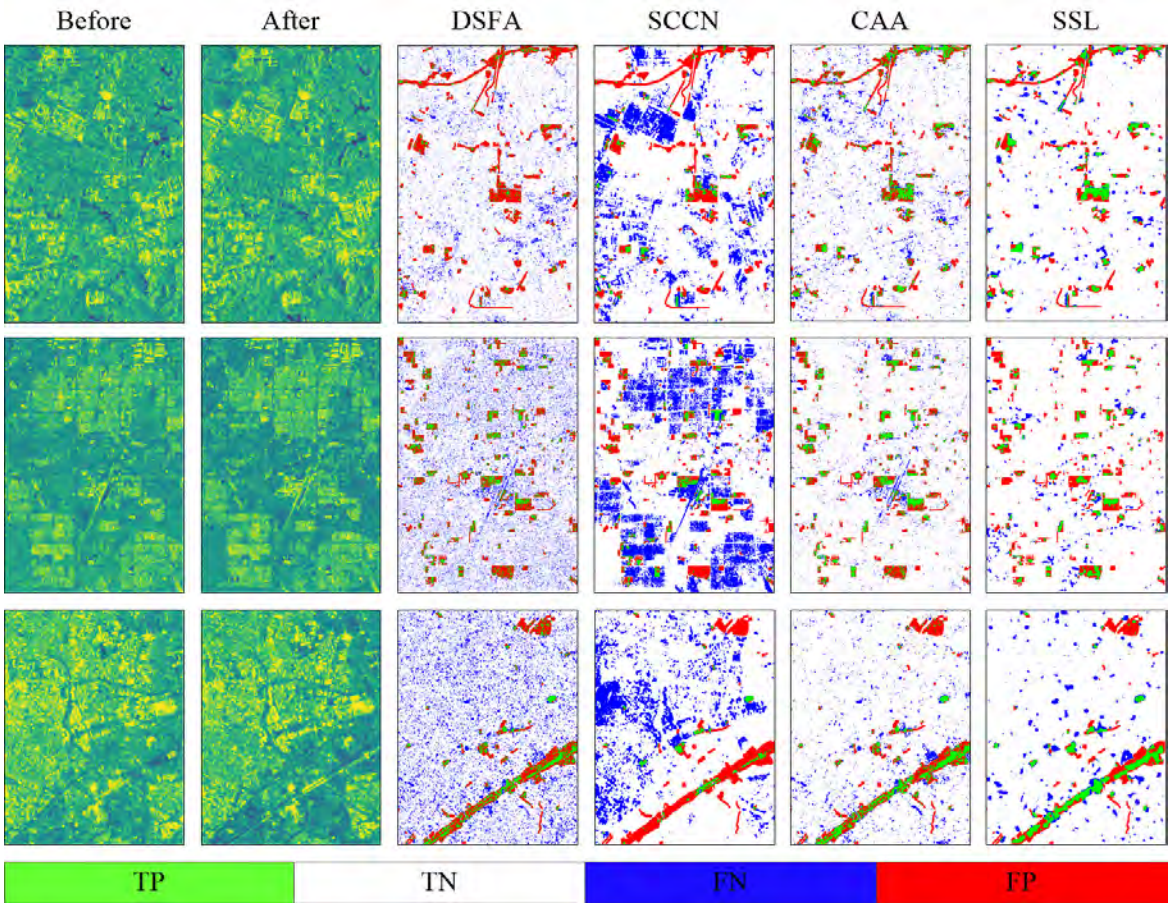

Fig. 3.4 Examples of change detection results on OSCD_S1S1, organized in one row for each location. Col. 1: pre-event image (Sentine-1); Col. 2: post-event image (Sentine-1). Change maps obtained by: DSFA (Col. 3), SCCN (Col. 4), CAA (Col. 5), and the proposed SSL (Col. 6).

Table 3.2 Quantitative evaluations of different unsupervised approaches applied to the OSCD_S1S1 datasets.

Method	Pre(%)	Rec(%)	OA(%)	F1	Kap
Prop. SSL	**27.52**	27.72	92.33	**0.28**	**0.24**
SCCN	7.48	27.80	78.04	0.12	0.04
CAA	19.80	**34.81**	89.12	0.25	0.20
DSFA	10.96	22.78	**92.63**	0.15	0.08

the supervised learning algorithm. However, the FC-EF-Res method failed to detect most of changed pixels in the first scenario. This confirms that the change detection results of supervised methods heavily rely on the change type distribution and the quality of training samples. This is not an issue for the proposed SSL approach.

Table 3.3 Quantitative evaluations of different approaches applied to the OSCD_L8S2 dataset.

Type	Method	Pre(%)	Rec(%)	OA(%)	F1	Kap
Unsup.	Prop. SSL	**37.31**	32.22	**93.57**	**0.35**	**0.31**
	CAA	18.45	**45.80**	90.25	0.26	0.22
	DSFA	8.08	24.29	86.64	0.12	0.07
Sup.	FC-EF	29.75	**34.08**	92.27	**0.32**	0.28
	FC-EF-res	**39.14**	27.14	**93.93**	**0.32**	**0.29**

To complete the evaluation on single-sensor datasets, the performance of all unsupervised methods is validated on the OSCD_S1S1 test dataset. The quantitative results are reported in Table 3.2, which shows that the proposed SSL approach (with the input size of 8 pixels) produces better accuracy than other methods. The binary change maps obtained by each unsupervised method are shown in Fig. 3.4. One can see that all results appear much noisier due to the influence of speckle in SAR images. It is worth noting that only a new building that appeared in the post-event SAR image can be detected because minor growth of the building does not cause significant backscatter change. Apart from this, the boundaries of the detected objects are not accurate as those in the optical dataset due to the side-looking imaging mechanism. In general, the above two experiments based on single-sensor images demonstrate that the proposed SSL approach obtained the best quantitative and qualitative performance with respect to all the other considered unsupervised change detection techniques.

Results on Cross-sensor Data Sets: In the second change detection scenario, we consider three cross-sensor data sets which consist of a Landsat-8/Sentinel-2 images set (OSCD_L8S2 test dataset), a Sentinel-1/Sentinel-2 image pair (OSCD_S1S2) and a Sentinel-1 / Landsat-8 image pair (California). The performance of each model applied to OSCD_S2S2 is also validated on the OSCD_L8S2 test dataset, which was obtained by different optical sensors having different spatial resolutions, and the quantitive evaluation is reported in Table 3.3. In general, the supervised methods outperform DSFA and CAA considering all five metrics. However, the performance of FC-EF-res on Recall is much worse than those of CAA and the proposed SSL approach. Meanwhile, the proposed SSL approach (with an input size of 16 pixels) with an overall accuracy of 93.57% and a Kappa coefficient of 0.31, obtained the best accuracy among the methods. Fig. 3.5 presents the binary change maps obtained by all methods on the OSCD_L8S2. One can see that the change maps contain a larger number of false alarms for all methods compared with the maps obtained on the OSCD_S2S2. This is probably due to the relatively lower resolution of Landsat-8 VIS and NIR channel images with respect to the counterparts in Sentinel-2 images. Consistently with the results obtained on OSCD_S2S2 (see Fig. 3.3), the proposed SSL approach has a better segmentation result

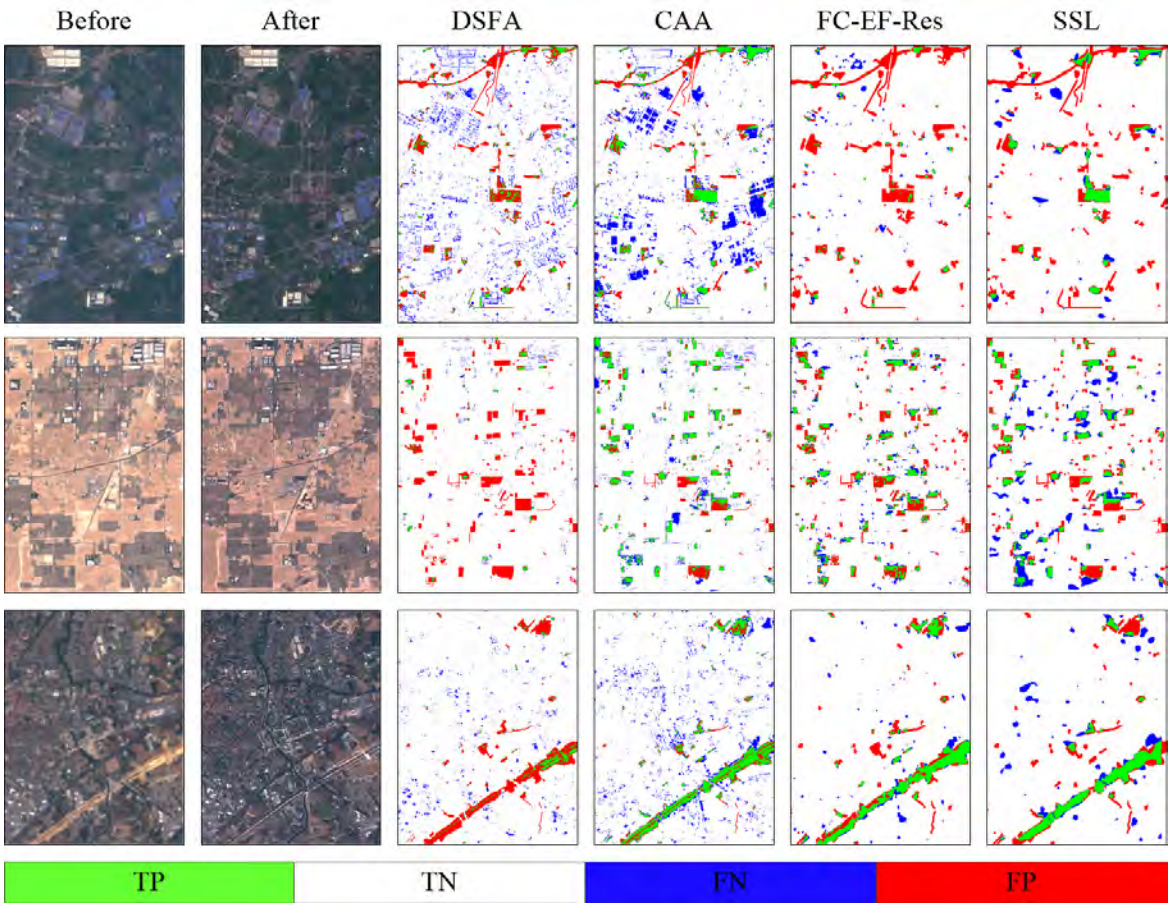

Fig. 3.5 Examples of change detection results on OSCD_L8S2, organized in one row for each location. Col. 1: pre-event image (Landsat-8); Col. 2: post-event image (Sentinel-2). Change maps obtained by: DSFA (Col. 3), CAA (Col. 4), FC-EF-Res (Col. 5), and the proposed SSL (Col. 6).

but with lower accuracy on all metrics, which indicates that the different resolution images increase the difficulty of change detection tasks.

The performance of three unsupervised methods (SCCN, CAA and SSL) on OSCD_S1S2 is reported in Table 3.4. One can see that the proposed SSL approach (with an input size of 16 pixels) performs much better than the other two unsupervised methods on most metrics due to the separated training on the archived images. In contrast, SCCN and CAA are both trained on the test image only and the complicated background in the scene makes them hard to separate the unchanged pixels for the network training causing too many false alarms in change detection maps. Compared with the results obtained in the optical image experiments, the results presented here are much worse. This demonstrates the difficulty of SAR-optical change detection in complicated backgrounds, such as urban area. Fig. 3.6 presents the qualitative visual results in terms of binary change maps. One can observe that the results

Table 3.4 Quantitative evaluations of different approaches applied to the heterogeneous images OSCD_S1S2 and the California datasets.

Dataset	Method	Prec(%)	Rec(%)	OA(%)	F1	Kap
S1S2	SCCN	7.38	22.45	68.54	0.11	-
	CAA	21.91	**28.71**	84.79	0.25	0.17
	Prop. SSL	**62.82**	24.19	**92.10**	**0.35**	**0.32**
California	SCCN	51.42	64.44	92.88	**0.57**	**0.53**
	CAA	**76.49**	40.38	**94.68**	0.53	0.50
	Prop. SSL	40.43	**68.14**	90.24	0.51	0.46

provided by SCCN and CAA are affected by many more missed detections and false alarms than in the single-sensor case. The result of the proposed SSL approach has fewer false alarms but with more missed detections with respect to the single-sensor setting owing to the larger domain discrepancy.

Differently from the previous dataset, the California dataset is related to a simpler background and to more significant changes resulting from the flood. Table 3.4 presents

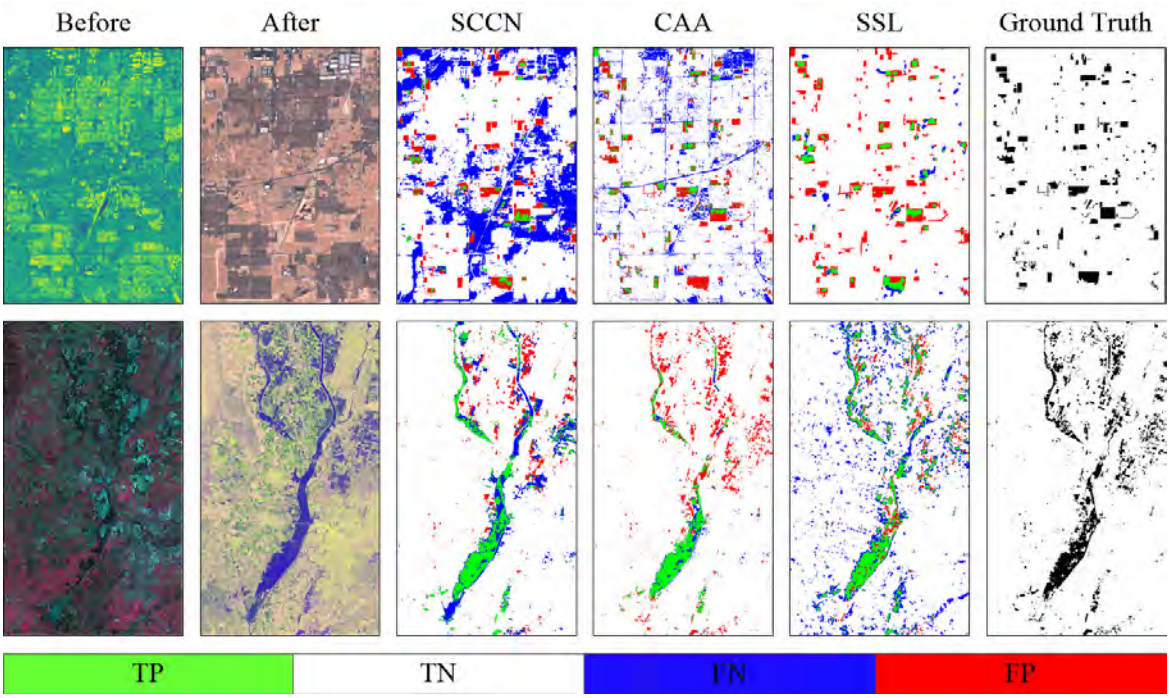

Fig. 3.6 Change detection results on OSCD_S1S2 and California flood, organized in one row for each location. Col. 1: pre-event image (Sentine-1 for OSCD_S1S2 and Landsat-8 for the California flood); Col. 2: post-event image (Sentine-2 for OSCD_S1S2 and Sentine-1 for the California flood). Change maps obtained by: SCCN (Col. 3), CAA (Col. 4), and the proposed SSL (Col. 5). Col. 6: the ground truth.

the results of all methods on this dataset. The three unsupervised methods (SCCN, CAA and SSL) have similar performance on overall evaluation metrics (OA, F1 and Kappa). The SCCN achieves the best F1 score, Kappa and the second-best values on Precision, Recall and OA, while the CAA achieved the highest Recall. The proposed SSL approach (with an input size of 8 pixels) gets the third-best values on four of five metrics, thus it does not show obvious superiority. Fig. 3.6 illustrates the Landsat 8 and Sentinel-1 images and the change maps from the compared methods. Maps provided by SCCN and ACC show a clear boundary of change areas, whereas one of the proposed SSL approaches is less precise. The map of the proposed SSL approach contains more false alarms, while the map of the CAA has more missed detections. In general, considering the results on the two SAR-optical test datasets, the proposed SSL approach achieved the best performance in urban areas whereas a slightly worse performance on flood detection.

Analysis of Robustness and Efficiency: To better analyze the robustness of the proposed SSL approach, we further evaluated the performance in terms of the five metrics considered under different input sizes. Here we provide an example of results considering the OSCD_L8S2 dataset. We consider the effects of varying input sizes from small to large, where each input size is a multiple of 8 pixels. Table 3.5 shows that the optimal input size for the contrastive method is 16; too small (8) or too large input size (24) sharply degenerates the accuracy. The architecture of the proposed approach allows for generalization to an arbitrary input size, which can prevent failure under a given input size.

We also compared the efficiency between the proposed approach and the other selected methods in terms of inference time (Table 3.6). From a general perspective, it is not possible to provide a fair efficiency comparison between them. This because they are not end-to-end deep learning models. Because the supervised method does not allow arbitrary input size and memory limitation, we used a small patch size ($1\times8\times512\times512$) that fits in memory to

Table 3.5 Quantitative evaluations of contrastive method applied to OSCD_L8S2 under different input sizes.

Input Size(pixels)	Pre(%)	Rec(%)	OA(%)	F1	Kappa
8	17	10.73	92.52	0.13	0.09
16	**37.31**	**32.22**	**93.57**	**0.35**	**0.31**
24	11.8	11.19	90.9	0.11	0.07

Table 3.6 Efficiency comparisons between different methods.

Models	CAA	DSFA	SCCN	FC-EF	FC-EF-res	Prop. SSL
Kappa	0.29	0.32	0.06	0.44	0.43	**0.48**
Time (s)	**0.003**	0.09	0.001	0.015	0.015	7.849

test inference time for each model. The times reported for all models are based on the use of PyTorch on a 7.8 GB RTX 2070ti GPU.

From the analysis of Table 3.6, one can see that the unsupervised models (CAA, DSFA and SCCN) need much less inference time but result in far low accuracy. The proposed approach needs more inference time. However, this is acceptable at an operational level when working with a proper GPU setting. Note that a key performance indicator to consider is that, when compared with the supervised methods, the proposed approach achieved much higher accuracy and without any label during the network training. We expect that the time cost of annotating labels in supervised methods can be higher than the inference time in the proposed approach. Moreover, the proposed approach also has a parameter redundancy and the tradeoff between accuracy and inference time depends on the particular task considered.

3.1.4 Conclusion

We have presented a self-supervised approach to unsupervised change detection in multi-view remote sensing images, which can be used with both single-sensor and cross-sensor images. The main idea of the presented framework is to extract good feature representations from multi-view images using contrastive learning. Images from satellite mission archives are used to train the pseudo-Siamese network without using any label. Under the reasonable assumption that the change event is rare in long-time archived images, the network can properly align the features learned from images obtained at different times even when they contain changes. After completing the pre-training process, the regression error of image patches captured from bi-temporal images can be used as a change score to indicate the change probability. If required, a binary change map can be directly calculated from change intensity maps by using a thresholding method. Experimental results on both single-sensor and cross-sensor remote sensing image data sets proved that the proposed SSL approach can be applicable in practice, and demonstrated its superiority over several state-of-the-art unsupervised methods. Results also show that the performance declines when the resolution of the two sensors is different in a homogeneous setting.

Moreover, in the SAR-optical change detection setting, the change detection results are affected by the complexity of the background. Experimental results show that the fusion of SAR and optical images can improve the change detection results and the considerable potential of the proposed method in unsupervised change detection.

3.2 A Self-supervised Approach to Pixel-level Change Detection in Bi-temporal RS images

In this section, we present the proposed approach to multi-view remote sensing image change detection based on pixel-wise feature representation. It includes the network architecture of the proposed method, the contrastive learning and the uncertainty aware feature learning approach. We first get the pixel-wise feature representation of bi-temporal images from the proposed network, which is trained by using the averaged feature values over superpixels in the contrastive loss rather than using the pixel-level features. Then, the uncertainty aware feature learning approach is used to reduce the impact of seasonal changes in pixel-wise feature representation. Afterward, the binary change map is generated by comparing the cosine similarity between the feature vectors of each pixel within bi-temporal images given a threshold value.

3.2.1 Introduction

The basic idea of self-supervised change detection in remote sensing is to align the shared information between multi-view images and reduce the impact of the sensor- and time-related noise. In this context, self-supervised learning can play a major role in multi-view remote sensing image change detection. The use of self-supervised learning in change detection is possible with both multi-sensor and multi-resolution image pairs. Recently, self-supervised learning [61, 108, 151, 55, 119] has been recognized as a promising technique for obtaining meaningful representations and overcoming both season-related and sensor-related noise in the image processing domain. The intuitive idea is to start from the two temporal views and reconstruct their counterpart using generative models. Nevertheless, some studies have shown that such generative models overly focus on pixels rather than on abstract feature representations [97]. Research in contrastive learning [151, 69, 26] has encouraged the network to learn more interpretable and meaningful feature representations from multi-view images, where they outperformed the generative counterparts. However, these works focus on image-level tasks. The patch-based algorithm in the image-level processing also makes these methods very computationally expensive. How to perform contrastive learning in pixel-level change detection is a problem that until now has been relatively unexplored.

From a different perspective, few approaches have considered the aleatoric uncertainty of seasonal changes for binary change detection tasks. For example, the feature map of cropland shows a high uncertainty, whereas that of the forest is relatively stable. This is because the cropland changes significantly with the seasons. Traditional change detection methods

simply treat the multi-temporal croplands in the same way, which is not sufficient to alleviate season-related noise. Modelling the uncertainty of multi-temporal images can reduce the impact of such seasonal changes, thus resulting in superior and robust performance. In the computer vision (CV) community, some novel approaches [111, 85] have been proposed to estimate aleatoric uncertainty during the training and inference of models. For the regression task, models can predict the uncertainty in one forward pass. Most of them require training labels to perform uncertainty estimation during training. Unfortunately, there are no labels that can be used in such an unsupervised change detection task.

For these reasons, in the second work, we propose a pixel-wise self-supervised change detection approach based on contrastive learning. The proposed approach consists of two branches and is trained on shift-augmented images. Instead of adopting contrastive loss on each pixel feature, this work exploits the averaged feature over superpixels, where each averaged feature is treated as a single instance during the training. Superpixels obtained from the same location of multi-view image pairs are called positive pairs. Negative pairs are obtained from different locations or different batches of multi-view image pairs. In the training process, the features of positive pairs can be pulled close and those of negative pairs can be pushed apart. In addition to the contrastive approach, we also introduce an uncertainty approach to reduce the impact of seasonal changes at the second step of the network training.

In summary, our contributions are as follows:

- As far as we know, we are the first to apply the pixel-wise contrastive method to unsupervised remote sensing change detection tasks and assess its performance on bi-temporal and bi-sensor datasets.

- We propose a self-supervised change detection approach at the pixel level and introduce a simple but effective uncertainty approach in the change detection task to reduce the impact of seasonal changes.

- We provide a comparison with the state-of-the-art approaches on various types of datasets. Experimental results show that our method obtains comparative results with state-of-the-art methods and a sharp improvement compared to the traditional unsupervised approaches. Moreover, the pixel-wise approach outperforms the patch-based contrastive method in the efficiency and robustness of water areas. We also showed the further improvements obtained by the uncertainty approach.

3.2.2 Methodology

Network Architecture

The proposed approach has two branches (Fig. 3.7). Two temporal views of images are used as inputs to each branch. Each branch contains a ResUnet block (U) [170] and an additional three-layered perceptron (MLP) projector (P) in the online branch. In addition, the same shift transformation is applied to both input and output following the shift equivariance principle. This is used as a kind of geometric data augmentation. We adopt a ResUnet architecture similar to the one presented in the [170] but replace all padding types with the same padding. Like U-net [130], ResUnet consists of encoder, decoder, bridge, and skip connections while using residual units instead of plain neural units. The encoder is used to get compact features

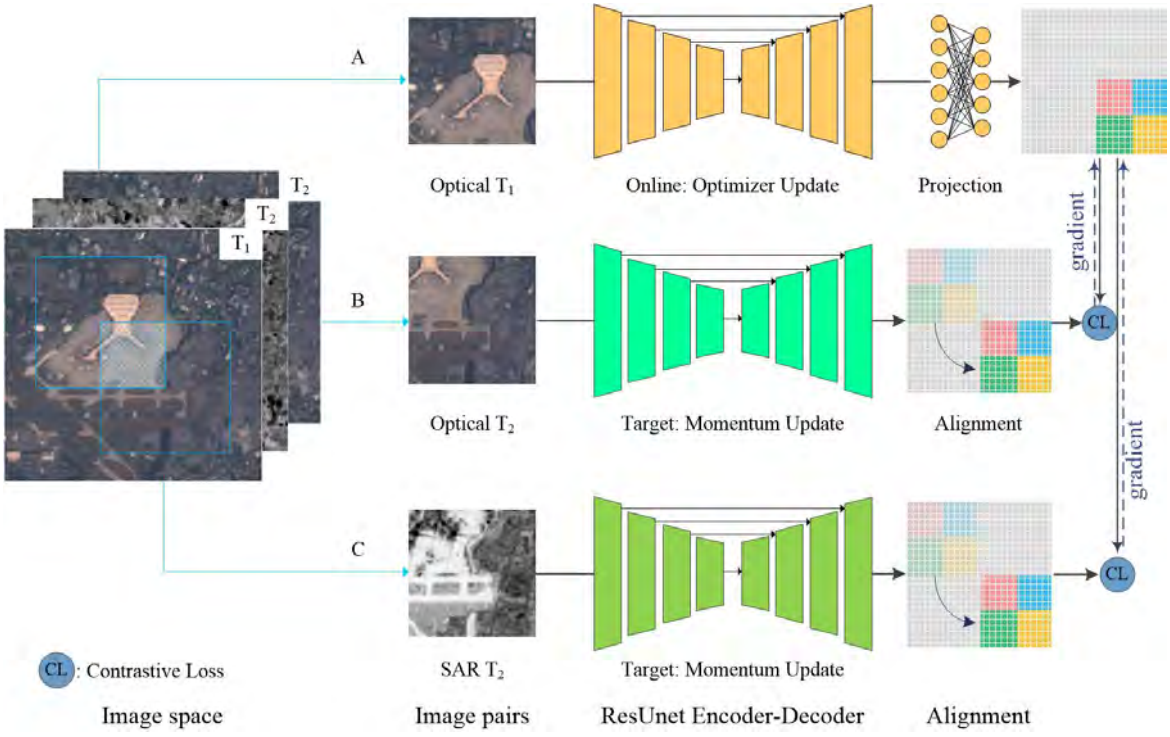

Fig. 3.7 Overview of the proposed pixel-wise self-supervised change detection approach. We perform a shift operation between two input views (T_1 and T_2) but still keep an overlap. The approach is based on a pseudo-Siamese architecture with two branches both consisting of a ResUnet block and an additional projector in the online branch (A). At the end of the network, the output features of two branches are used as the inputs to the contrastive loss. The weights of the target branch (B or C) are then updated by a momentum update of the online branch. Note that the branches A and B denote the homogeneous image change detection scenario and the branches A and C denote the heterogeneous image change detection scenario. T_1 and T_2 denote that the images are acquired at two different times.

with convolution layers, while the decoder reconstructs the compact features at the pixel level. Multi-level features from the encoder are aggregated by using skip connections, which reduces the number of parameters of the network achieving better performance.

As the deeper and deeper networks, the gradient in backpropagation sometimes vanishes, which results in a degradation problem. In this context, He *et al.* [70] propose the deep residual network using skip connections in each residual unit. Each residual unit can be expressed in the following general form:

$$\begin{aligned} \mathbf{x}_{i+1} &= \mathbf{x}_i + F(\mathbf{x}_i, W_i) \\ \mathbf{x}_{i+1} &= f(\mathbf{x}_{i+1}) \end{aligned} \tag{3.10}$$

where $\mathbf{x}_i$ and $\mathbf{x}_{i+1}$ are the input and output of the $i-th$ residual unit, respectively, $F(\cdot)$ is the residual function, W_i is the weight matrix, $f(\cdot)$ is the activation function.

In this work, each residual unit of the encoder consists of two Conv-BN-ReLU blocks and two identity mappings. There are three residual units in the encoder and a Conv-BN-ReLU-Pool block before the residual units. In the last two blocks, instead of using a pooling operation to downsample the feature maps, a stride of two is applied to the convolution block to reduce the feature map by half. The bridge part consists of a convolutional layer, a BN layer and a ReLU activation layer and followed by an up-sampling operation. The decoder part has three blocks but without residual connections and uses a stride of one in all convolution. In each decoder block, there is a concatenation with the feature maps from the corresponding encoding path and then an up-sampling operation for concatenated feature maps. At the last of the decoding path, a linear layer is used to reconstruct the learned representations. At last, the projector consists of a 1×1 Conv of 192 channels and a ReLU, and then of a 1×1 Conv with 128 channels for each pixel. The parameters and channel size of each unit of the online branch are presented in Table 3.7. Each unit ([]) consists of a convolutional layer, a BN layer and a ReLU activation layer. The parameters of the target branch ϕ are updated in a momentum update controlled by τ and the parameters of the online branch θ, i.e.,

$$\phi \leftarrow \tau\phi + (1-\tau)\theta \tag{3.11}$$

Shift equivariance is achieved by using shift operation on the input image and output features of the two branches, respectively. Specifically, given an image pair, we randomly crop the areas with the same size in both images and keep the overlap between the two cropped areas. During the training, the cropped image pair is fed into the two branches, respectively, to obtain two feature maps. To align the feature map of the two branches, the

same transformation is applied to the counterpart output. During the inference, the model provides two feature maps for the considered bi-temporal images. The change intensity map is defined as the cosine similarity between the feature vector of each pixel within bi-temporal images. To get binary change maps, the Rosin thresholding [132] method is used in this work.

Loss Function

The training objective function is a contrastive loss. The contrastive loss is used to distinguish the representations of each superpixel from others. The loss is sampled over the corresponding superpixel features between two input views (Fig. 3.8). This aims to keep the consistency of the normalized pixel-wise representations between the two branches. Each superpixel-wise feature pair (z_1^i, v_2^i), where z_1^i is the output of ResUnet and v_2^i is the output of the projector, is sampled from the same location i that is called positive. Let v_2^j be taken from another location that is called negative. The contrastive loss can be written as L_{contrast}:

$$L_{\text{contrast}} = -\mathop{\mathbb{E}}_{S}\left[\log \frac{h_\theta(z_1^i, v_2^i)}{\sum_{j=1}^{N} h_\theta(z_1^i, z_2^j)}\right] \tag{3.12}$$

Table 3.7 Structure of the network of the proposed online branch.

	Encoder	Decoder	
Conv1 Maxpool	[3×3, 64], stride 2 3×3, stride 2	Cat. ResBlk3 [3×3, 128] upsampling 2	DecBlk1 stride 1
ResBlk1 stried 1	$\begin{bmatrix} 3\times3,\ 64 \\ 3\times3,\ 64 \end{bmatrix}\times 2$	Cat. ResBlk2 [3×3, 128] upsampling 2	DecBlk2 stride 1
ResBlk2 stride 2	$\begin{bmatrix} 3\times3,128 \\ 3\times3,128 \end{bmatrix}\times 2$	Cat. ResBlk1 [3×3, 128] upsampling 2 Conv 1×1, 128	DecBlk3 stride 1
ResBlk3 stride 2	$\begin{bmatrix} 3\times3,256 \\ 3\times3,256 \end{bmatrix}\times 2$	1×1, 192 ReLU 1×1, 128	Projector
Bridge stride 1	[3×3, 256] upsampling 2		

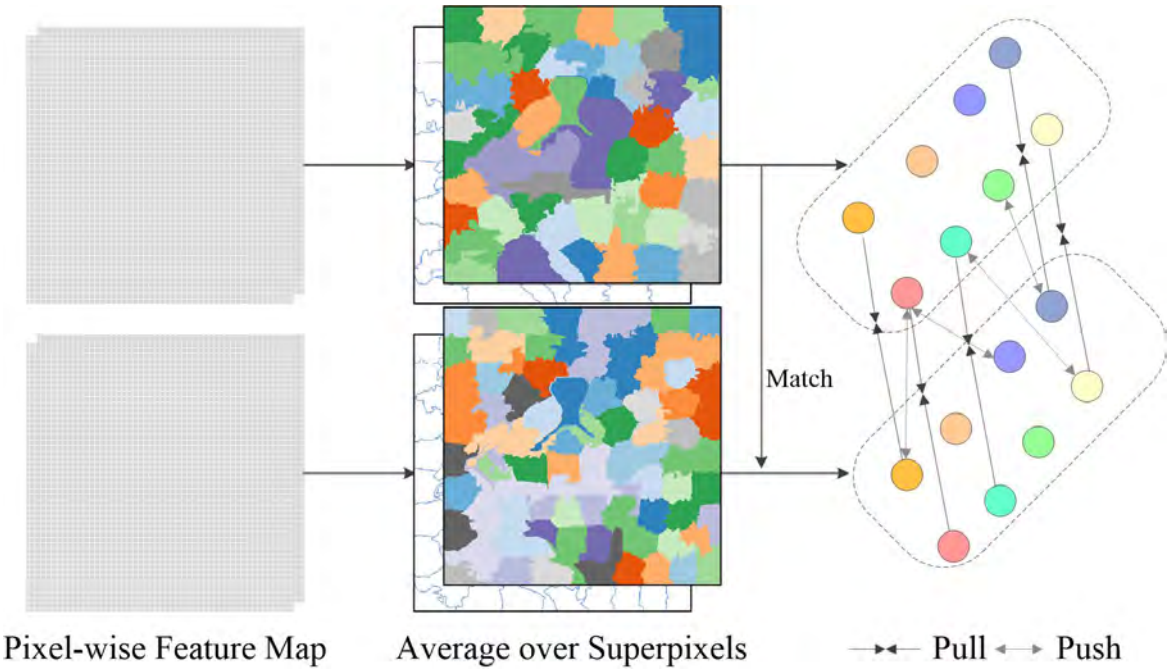

Fig. 3.8 Overview of the contrastive loss performed on superpixel features. F_1 and F_2 denote the feature from bi-temporal images.

where $h_\theta(\cdot)$ is a similarity function (i.e., cosine similarity), $\{(z_1^i, v_2^i)\}$ is a normalized latent representation pair i of N scenes, $\{(z_1^i, v_2^j | j \geq i)\}$ is a negative feature pair and $S = \{z_1^1, v_2^1, v_2^2, \cdots, v_2^N\}$ is a set that contains $N-1$ negative feature pairs and one positive feature pair by anchoring at z_1^1. In the training process, the network is trained to increase the value of positive pairs and decrease the value of negative pairs. This results in a feature representation that is close for positive pairs whereas it is not for negative pairs. Compared with the instance-level contrastive learning, this loss function is able to make the model get more detailed representations and more suitable for dense prediction downstream tasks.

Uncertainty-aware Feature Learning

In this section, we propose a deterministic model to approximate the feature representations that are invariant to the seasonal changes (Fig. 3.9). Specifically, the model learns to directly infer both feature representation and its uncertainty in a forward pass. The network architecture is based on a teacher-student paradigm, where the parameters of the teacher model are fixed during the network training. The teacher network is pre-trained by the proposed approach. Then, bi-temporal predictive samples are generated from the teacher model to train the student network. Like the works in regression tasks, we use the KL loss to approximate the variational predictive distribution and estimate the log variance (s) from the

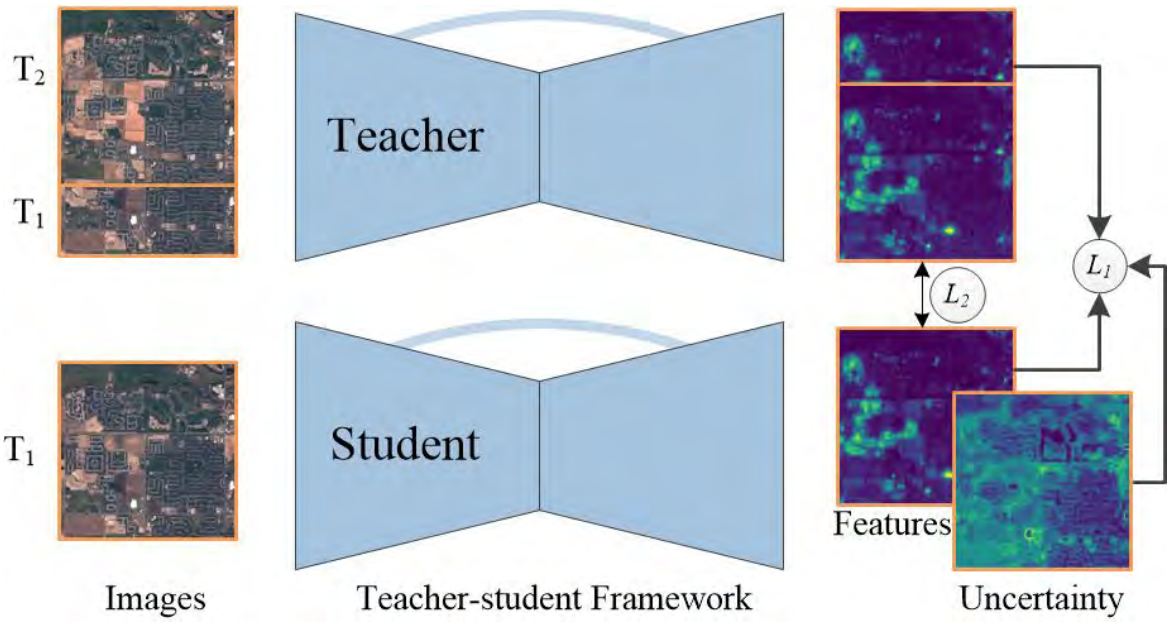

Fig. 3.9 Overview of the teacher-student paradigm for uncertainty-aware feature learning. T_1 and T_2 denote that the images are acquired at two different times. L_1 and L_2 are the two components of the uncertainty loss.

network directly avoiding the gradient explosion. The loss function can be written as:

$$L_1 = \frac{1}{H}\frac{1}{W}\sum_i \frac{1}{2}\exp(-s_i)\, d(y_i^1 - \mu_i^2) + \frac{1}{2}s_i \tag{3.13}$$

where i corresponds to each pixel within an image; H and W are the height and width of the image; y^1 and μ^2 are the predictions from teacher and student network at time T_1 and T_2, respectively. In most works, d is the l_2 distance between the prediction of teacher and student networks, whereas we use the cosine distance substituted for l_2 distance. Empirically, we found that training solely with the above loss function sometimes leads to sub-optimal predictive performance. This may be due to too large amount of noise between different temporal images. Thus we leverage the feature of images generated by the teacher network at the same time to stabilize the training process. The teacher-student model is trained with the cosine distance between bi-temporal features, leading to the total loss:

$$L_{\text{un}} = L_1 + \lambda \frac{1}{H}\frac{1}{W}\sum_i d(y_i^2 - \mu_i^2) \tag{3.14}$$

where the λ is a hyper-parameter to be tuned. We found that $\lambda = 1$ generally performs well in our experiments.

3.2.3 Experimental Results

In this section, we present the considered datasets, the experiment settings and then evaluate the performance obtained using the proposed approach for binary change detection tasks.

Description of Datasets

We developed our experiments on three multi-view data sets, which consist of two homogeneous data sets and one heterogeneous data set.

OSCD_S2S2: The Onera Satellite Change Detection (OSCD) dataset [37] was created for bi-temporal change detection using Sentinel-2 images acquired between 2015 and 2018. The dataset was acquired in 24 cities and includes different landscapes. The pixel-wise ground truth labels, which were manually annotated, focus on urban growth and built-up changes while containing some errors on the identification of bare lands. To use this dataset in self-supervised training, we downloaded additional Sentinel-2 images between 2016 and 2020 in the same location as the original bi-temporal images. We considered images from each month to augment existing image pairs. To keep consistency with previous research, 10 image pairs obtained from 10 different cities are treated as the test set for evaluation.

MUDS_S2S2: Multi-temporal Urban Development (MUDS) dataset [152] is an open-source dataset of the native Planet 4 *m* resolution imagery acquired between 2017 and 2020. The imagery comprises 24 consecutive monthly mosaic images of 101 locations over 6 continents. To use this dataset with OSCD_S2S2, we downloaded additional Sentinel-2 images between 2017 and 2020 in the same location as the original images and resized each image to 512×512 pixels. We chose 33 of 110 locations as the test set, where the first image was defined as pre-image and the last image was defined as post-image. In addition, we manually labeled the three types of changes, such as built-up, bare land and water. Note that only Sentinel-2 images of this dataset are used in this work.

Flood in California: The California dataset is a cross-sensor data set that includes a Landsat-8 (multi-spectral) and a Sentinel-1 GRD (SAR) image. The multispectral and SAR images are acquired on 5 January 2017 and 18 February 2017, respectively. The dataset represents a flood that occurred in Sacramento County, Yuba County, and Sutter County, California. The ground truth was extracted from a Sentinel-1 SAR image pair where the pre-event image is acquired approximately at the same time as the Landsat-8 image [28]. The other three image pairs of Sentinel-1 and Landsat-8 images of the same scene acquired in 2017 and 2018, are used for the self-supervised pre-training of the proposed approach.

Experiment Settings

Evaluation Metrics: To appraise the different methods in binary change detection tasks, five evaluation metrics (precision (Pre), recall (Rec), overall accuracy (OA), F1 score and Cohen's kappa score (Kappa)) are used in this paper. We simply classify the image pixels into two classes by setting an appropriate threshold value according to the Rosin thresholding method and then analyze them with reference to the ground truth.

Implementation Details: Concerning the geometric data augmentation in this work, we applied shift transformations with a crop size of 128 $\times$ 128 pixels and an overlap between bi-temporal images ranging between 64% and 100%. In addition, we also applied random flip to further improve the performance of the proposed approach in the network training. The photometric augmentation was not considered for capturing the seasonal change better. We also introduce the multi-crop strategy on superpixel features to improve the performance of the proposed approach. It consists in segmenting each bi-temporal image using different superpixel algorithms or superpixel parameters. During the training, we first sample superpixel indices of one temporal input and then select the corresponding ones on the other temporal input. Because of the different sizes of superpixels in the two inputs, the corresponding superpixels are decided by the maximum overlap criterion (Fig. 3.8). Once corresponding superpixels are decided, corresponding features between bi-temporal images are averaged over the selected superpixels, respectively. To select the appropriate samples for calculating contrastive loss, we segment the image into superpixels and randomly select one superpixel from each image patch. We used the felzenszwalb approach [51] to generate superpixels.

For the self-supervised training of the teacher network, we adopt Adam with an initial learning rate of $3e^{-4}$ and decay the learning rate with the step scheduling without restarts and set the batch size as 100. Models are run for 200 epochs. We used bi-temporal images to train the student network of the teacher-student paradigm for uncertainty-aware feature representation. In order to capture the teacher predictive distribution, the image used to train the student model should not be the same as the one for the teacher model. To alleviate this problem, both temporal images were given as input to the teacher model and one of them was given as input to the student model during the training of the teacher-student network. This extra image is crucial for the enhanced quality of feature maps in the student model. We emphasize that the uncertainty estimation comes from the bi-temporal images. To achieve faster convergence, we initialize the student network using the weights of the teacher network. To this end, a smaller initial learning rate of $1e^{-4}$ is used to train the student network for 200 epochs. We employ a step learning rate policy on the student network only and a batch size of 10.

There are two baseline approaches, patch-based self-supervised approach (PatchSSL) [28] and Code-Aligned Autoencoders (CAA) [103], that we categorize to make a comparison. Besides these two approaches, we also used SCCN [171] as a comparison in heterogeneous change detection. Meanwhile, we also include the results of the teacher network in the proposed approach as a baseline comparison.

Experimental Results

Experimental Results on Bi-temporal Sentinel-2 Images: As mentioned before, two bi-temporal Sentinel-2 image data sets are proposed to evaluate the effectiveness of the proposed approach. They are the OSCD_S2S2 and the MUDS_S2S2 data sets. The performance of the proposed approach (PixSSLs) is compared with Code-Aligned Autoencoders (CAA), patch-based self-supervised methods (PatchSSL) and the result of the teacher network (PixSSLt) in the proposed approach. Two supervised approaches (FC-EF and FC-EF-res) from previous research [36, 35] are also considered.

The performance metrics obtained on the OSCD test set are reported in Table 3.8. As one can see, the PixSSLt obtained an OA of 94.08% and a Kappa coefficient of 0.49, which outperforms the results obtained by PatchSSL and CAA as well as supervised methods. One can also observe that there is an improvement in OA and Kappa with about 1% and 0.02 after applying the uncertainty approach. The results obtained by the proposed PixSSLs that exploit the uncertainty approach are better than those obtained using the pixel-wise self-supervised approach only in almost all metrics except the Recall. Nevertheless, the patch-based self-supervised approach with an OA of 93.00% and a Kappa coefficient of 0.48, obtained the best performance on Recall. The table also presents the results of the supervised approach as presented in [28]. Note that three self-supervised approaches (PatchSSL, PixSSLt and PixSSLs) all outperform the supervised approaches on this dataset. The proposed PixSSLs not only outperforms the literature self-supervised approaches but also is more efficient during the inference phase.

Table 3.8 Quantitative evaluations of different approaches applied to the OSCD_S2S2 dataset.

Type	Method	Pre(%)	Rec(%)	OA(%)	F1	Kappa
Unsup.	Proposed PixSSLs	**62.46**	46.59	**95.70**	**0.53**	**0.51**
	PixSSLt	45.42	60.64	94.08	0.52	0.49
	PatchSSL	40.44	**69.10**	93.00	0.51	0.48
	CAA	23.49	52.96	91.66	0.33	0.29
Sup.	FC-EF	55.34	39.48	95.13	0.46	0.44
	FC-EF-res	54.97	38.39	95.10	0.45	0.43

Table 3.9 Quantitative evaluations of different approaches applied to the MUDS dataset.

Methods	Pre(%)	Rec(%)	OA(%)	F1	Kappa
Proposed PixSSLs	**43.81**	**70.54**	**94.45**	**0.54**	**0.51**
PixSSLt	32.30	69.15	91.87	0.44	0.40
PatchSSL	34.43	65.41	92.39	0.45	0.41
CAA	32.68	48.26	93.01	0.39	0.35

Similar performance can be found on the MUDS test set (Table 3.9). The PixSSLt obtained results similar to those of PatchSSL, by outperforming the results of CAA. The results obtained by the proposed PixSSLs with an OA of 94.45% and a Kappa coefficient of 0.51, which are better than those obtained by PixSSLt. Compared with the results on OSCD,

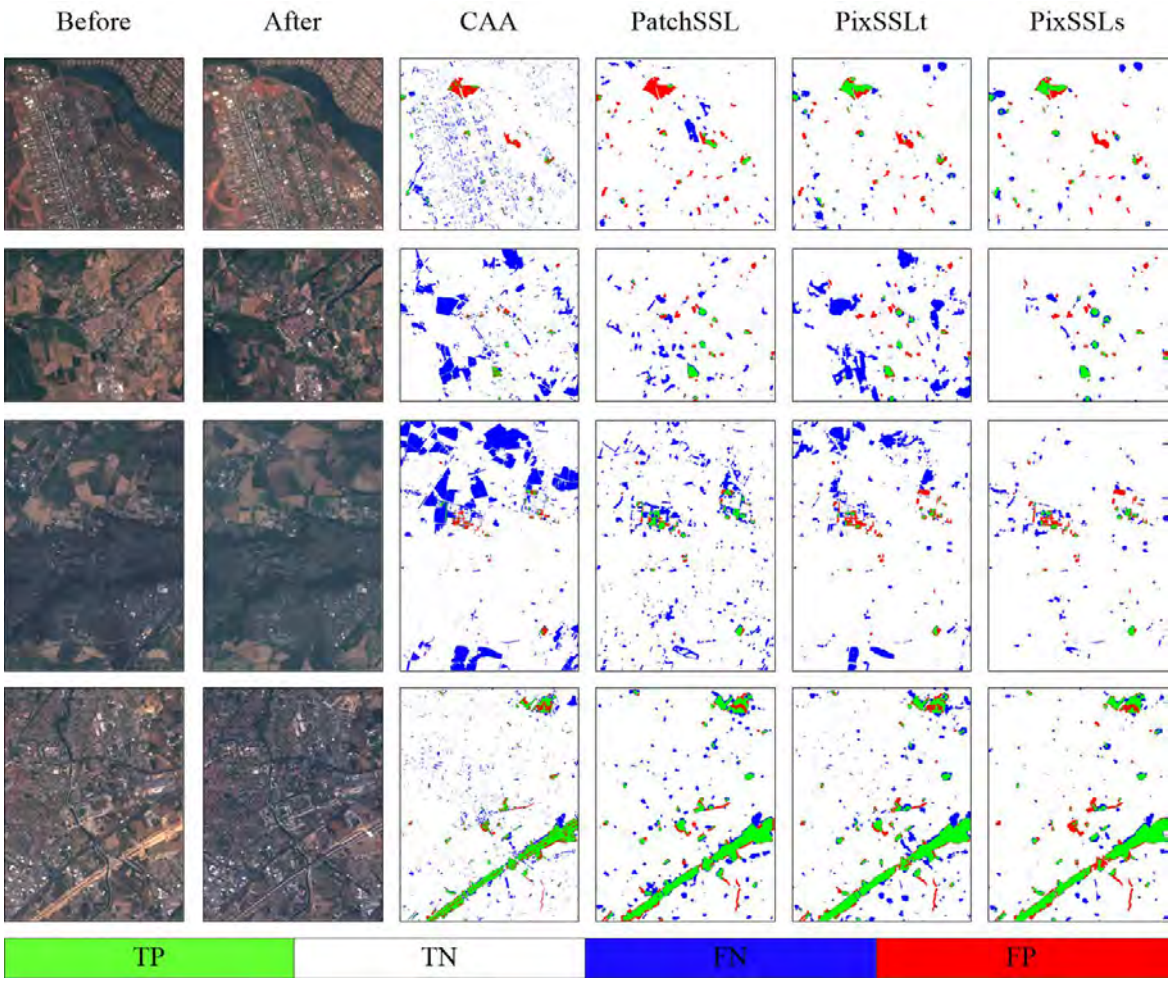

Fig. 3.10 Examples of change detection results on OSCD_S2S2, organized in one row for each location. Col. 1: pre-event image; Col. 2: post-event image. Change maps obtained by: CAA (Col. 3), PatchSSL (Col. 4), PixSSLt (Col. 4) and the proposed PixSSLs (Col. 6).

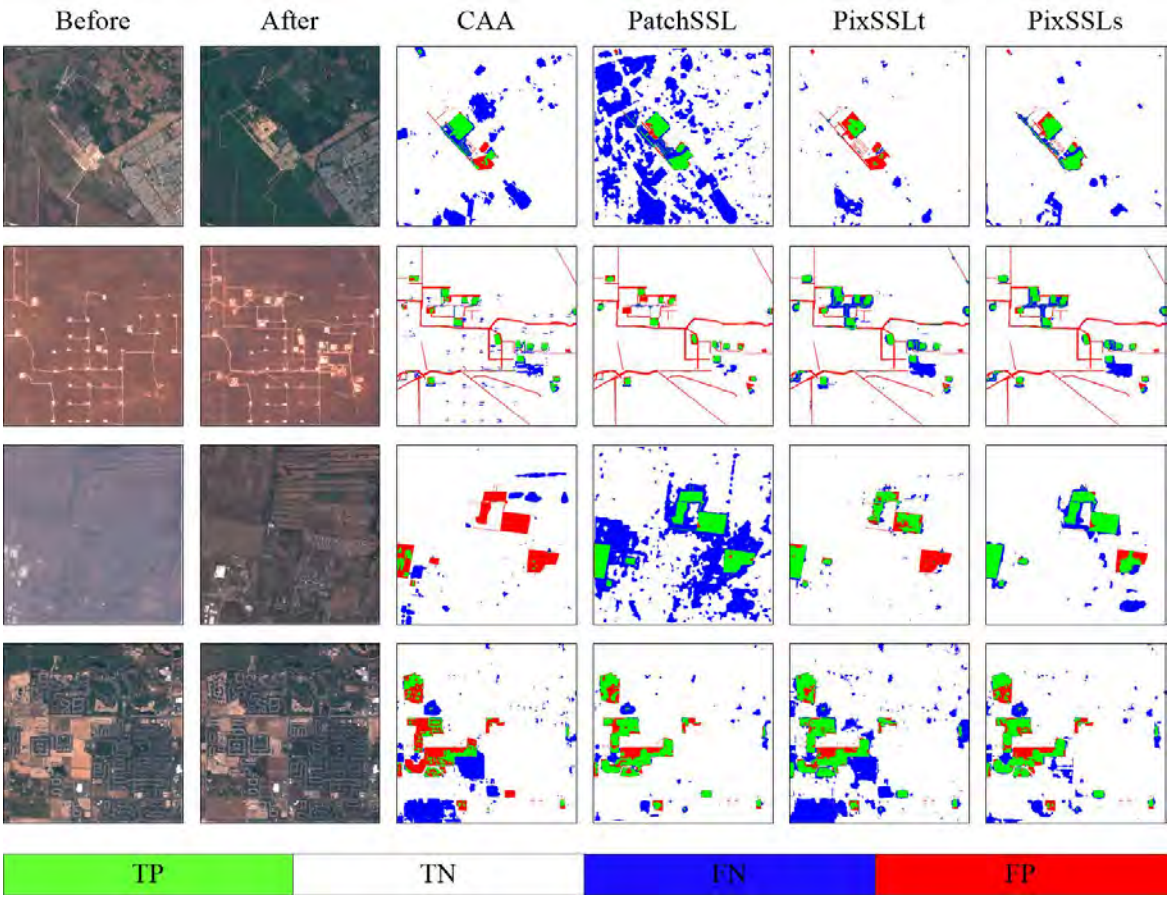

Fig. 3.11 Examples of change detection results on MUDS_S2S2, organized in one row for each location. Col. 1: pre-event image; Col. 2: post-event image. Change maps obtained by: CAA (Col. 3), PatchSSL (Col. 4), PixSSLt (Col. 4) and the proposed PixSSLs (Col. 6).

the improvement in OA and Kappa with about 2.6% and 0.11 is more prominent when using the uncertainty approach. Moreover, it also outperformed the PatchSSL, which obtained an OA of 92.39% and a Kappa coefficient of 0.41. The PatchSSL performs relatively poorly, which is mostly due to the fact that the MUDS dataset contains more seasonal changes. On the contrary, more built-up changes are presented in the OSCD dataset.

Besides the quantitative analysis, we also provide a visual qualitative comparison, where the TP, TN, FN and FP pixels are colored in green, white, blue and red, respectively. In Fig. 3.10, we show a comparison between all methods on the OSCD test set. One can see that the change maps obtained by CAA are noisy and contains more false alarms. Instead, the change maps obtained by other methods are in general more accurate and less noisy. Change maps provided by the PixSSLt contain more false alarms, as many unchanged pixels are wrongly classified as changed ones. The proposed uncertainty enhanced PixSSLs suppresses most unchanged regions but also fails to highlight some clearly changed regions. Among all

considered methods, PixSSLs successfully detects most of the changed pixels and achieve the best change maps.

Fig. 3.11 shows a comparison between all methods on the MUDS test set. One can see that the PatchSSL fail to suppress most seasonal changes. The results obtained by CAA and the proposed PixSSLt contain fewer false alarms whereas including a large number of missed detection. This issue is well addressed in the proposed uncertainty-enhanced PixSSLs. As shown in the first and third row, the proposed PixSSLs successfully suppresses most seasonal changes and detects the changed areas. In addition, most of the changed pixels are correctly detected in all contrastive approaches. The experiments on these two datasets demonstrate that self-supervised methods obtained the best quantitative and qualitative performance with respect to the considered autoencoder approach and the uncertainty-enhanced PixSSLs shows a sharp improvement in suppressing seasonal changes. Although the PatchSSL still achieved the comparable results, the proposed approach is more efficient.

Experimental Results on the Cross-sensor Image Pair: In the second change detection scenario, we consider one cross-sensor data set which consists of a Sentinel-1/Landsat-8 image pair (California flood). Table 3.10 lists the quantitative statistics on the changed maps obtained by four unsupervised methods (SCCN, CAA, PatchSSL and the proposed PixSSLt). As one can see, SCCN achieves the best F1 score and Kappa and the second-best values on Precision, Recall and OA, while it is trained on the test image itself. The proposed PixSSLt has a comparable performance to SCCN obtaining the second-best value. Although PixSSLt does not get the best performance, its results are superior to those of the PatchSSL and CAA approaches.

Fig. 3.12 illustrates the Landsat-8 and Sentinel-1 images and the change maps from the compared methods. SCCN achieves the best change maps with a clear boundary of flood areas, while the CAA just detects the main flood area and misses the small areas. PtachSSL highlights most of the flood areas, while its map is noisier than those of other comparison methods and contains more false alarms. Compared with the PatchSSL, the proposed PixSSLt produced a more clear change map, which is very close to the result of SCCN. Overall, the proposed PixSSLt improves the performance and effectiveness of multi-sensor change detection scenarios compared with the PatchSSL and CAA.

3.2.4 Discussion and Conclusion

Discussion

To better analyze the robustness of the proposed approach, we further evaluated the performance in terms of the five metrics under more challenging water areas. Here we provide two

Table 3.10 Quantitative evaluations of different approaches applied to the Flood dataset.

Methods	Pre(%)	Rec(%)	OA(%)	F1	Kappa
Proposed PixSSLt	50.64	59.11	92.73	0.55	0.51
PatchSSL	40.43	**68.14**	90.24	0.51	0.46
CAA	**76.49**	40.38	**94.68**	0.53	0.50
SCCN	51.42	64.44	92.88	**0.57**	**0.53**

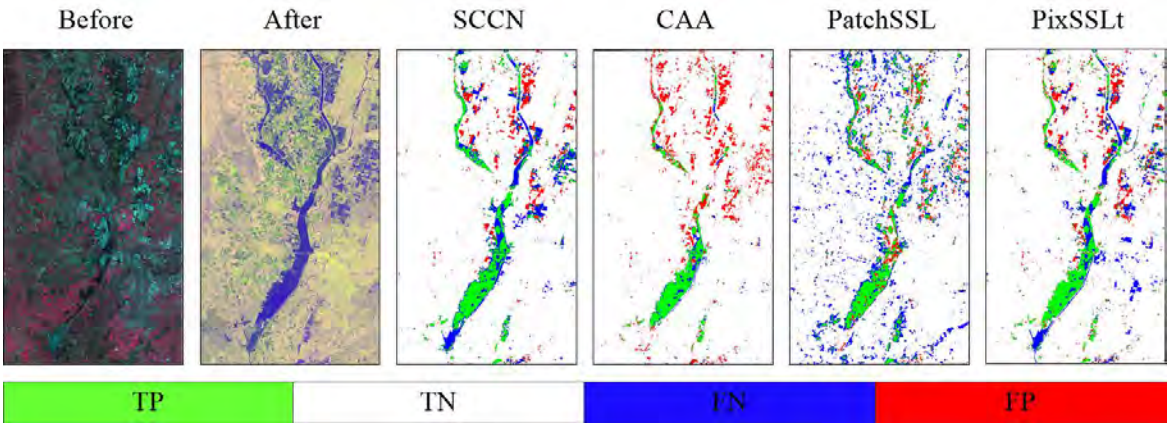

Fig. 3.12 Change detection results on the Califorlia flood dataset, organized in one row for each location. Col. 1: pre-event image (Landsat-8); Col. 2: post-event image (Sentine-1). Change maps obtained by: SCCN (Col. 3), CAA (Col. 4), and PatchSSL (Col. 5) and the proposed PixSSLt (Col. 6).

examples: one comes from OSCD and the other comes from MUDS. Fig. 3.13 shows the change intensity maps and change maps obtained by PatchSSL and PixSSLt. As we can see, both scenarios present many false alarms of water areas using PatchSSL, whereas the results of PixSSLt correctly identified the water areas as non-change in change maps. Similarly, water areas show a relatively high value in the change intensity maps of PatchSSL, whereas they are suppressed in the results of PixSSLt. In a quantitative way, the proposed PixSSLt obtains an OA of 97.2% and a Kappa coefficient of 0.47, whereas the PatchSSL obtained an OA of 89.6% and a Kappa coefficient of 0.23. Thus the proposed PixSSLt achieves the best performance. This demonstrates again that the proposed PixSSLt is more robust than PatchSSL.

In order to have an intuitive understanding of the efficiency between different methods, Table 3.11 presents a detailed comparison. The number of multiply-accumulate (MAC) and the number of parameters are considered as two relevant metrics of the model efficiency. The MAC operations is used to measure the computational cost. Following the common practice, we use them to measure the network efficiency in terms of computational cost and memory consumption. The metrics reported for all models are based on the use of PyTorch on a 7.8

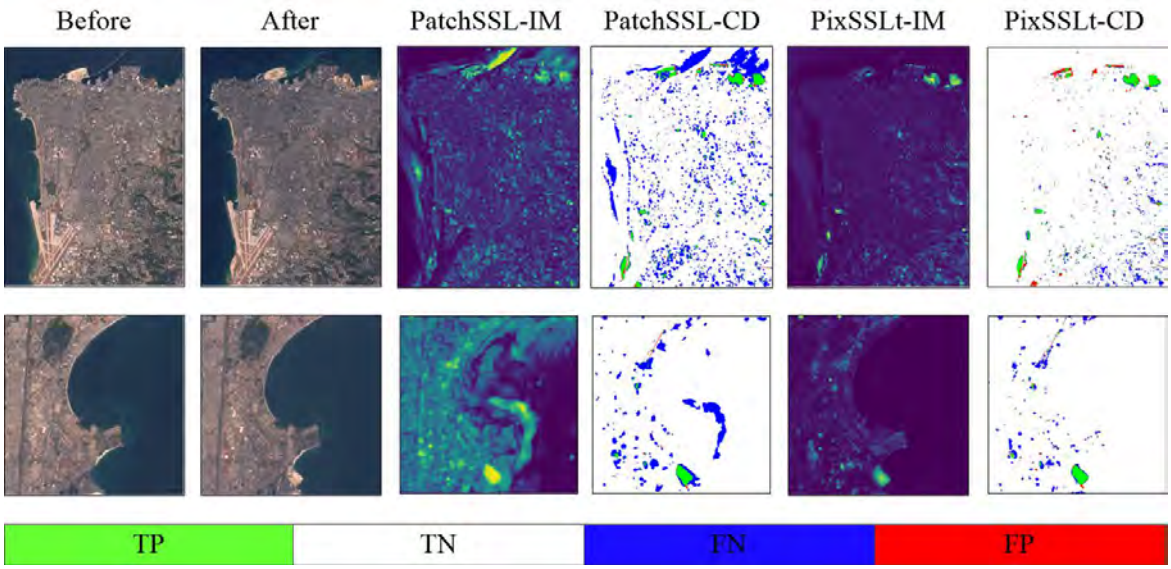

Fig. 3.13 Examples of change intensity maps and change maps obtained by PatchSSL and the proposed PixSSLt on water areas, organized in one row for each location. Col. 1: pre-event image; Col. 2: post-event image. Change maps obtained by: PatchSSL (Col. 4) and the proposed PixSSLt (Col. 6), and change intensity maps obtained by PatchSSL (Col. 3) and the proposed PixSSLt (Col. 5).

GB RTX 2070ti GPU. Table 3.11 shows the Kappa, MACs and the number of parameters of each model. The performance of each model on Kappa is obtained on the OSCD test set. From the analysis of Table 3.11, one can see that CAA needs much lower MACs but results in far low accuracy. Unlike CAA, the two self-supervised methods are heavy-weight networks and provide an high accuracy. Compared with PatchSSL, the proposed PixSSL achieved competitive results but with much lower computational costs.

We then derived change maps with different thresholding methods (OTSU and Rosin) using two self-supervised methods: PixSSLs and PatchSSL. In Table 3.12, we present the change detection results obtained on the MUDS dataset utilizing the Rosin and OTSU thresholding methods considering two self-supervised approaches. For both methods, the binary change results obtained using the Rosin thresholding approach are much better than those obtained by using the OTSU method. This indicates that the Rosin method is more robust than the OSTU method in the presented self-supervised change detection scenario.

Conclusion

We have presented a pixel-wise contrastive learning approach to multi-view remote sensing image change detection. It uses the ResUnet as the architecture of the network and exploits an uncertainty approach during the network training. The main idea of the presented approach is

Table 3.11 Efficiency comparisons between different approaches.

Models	Kappa	MACs (G)	Params (M)
Proposed PixSSL	**0.51**	84.9	4.216
PatchSSL	0.48	8026.1	21.353
CAA	0.29	40.7	**0.103**
FC-EF	0.44	14.4	1.351
FC-EF-res	0.43	**8.1**	1.104

Table 3.12 Change detection results of PixSSLs on the MUDS dataset using Rosin and otsu threholding methods.

Thresholding	Method	Pre(%)	Rec(%)	OA(%)	F1	kappa
Rosin	PatchSSL	34.43	65.41	92.39	0.45	0.41
	Prop. PixSSLs	43.81	70.54	94.45	0.54	0.51
OTSU	PatchSSL	18.05	86.44	80.60	0.30	0.24
	Prop. PixSSLs	33.63	74.17	92.04	0.46	0.43

the use of both the contrastive loss in pixel-wise feature learning and the uncertainty approach in suppressing seasonal changes.

Experimental results on multi-view remote sensing image data sets demonstrated the superiority and efficiency of the proposed approach over other state-of-the-art methods. Among the methods used in the comparison, the results produced by CAA contain more false alarms and missing detections. The results obtained by PatchSSL are similar to those of the proposed approach but with less suppression of seasonal changes. Moreover, PatchSSL is working on the patch level and is computationally more expensive. Compared with the PatchSSL approach, the proposed PixSSL is more effective in the inference phase and obtains better change maps, especially in vegetation and water areas. Results also show that the use of the uncertainty approach further suppresses the seasonal changes with respect to the only use of the contrastive learning method.

3.3 Conclusion

In this chapter, we have presented the proposed self-supervised change detection framework based on image patches and image pixels. Meanwhile, we have also considered the images from different sensors, different resolutions and also the fusion of different sensors. The main idea of the presented approach is the use of contrastive loss between different temporal and sensor images in suppressing seasonal and sensor noises. However, we found that contrastive learning solely cannot suppress the seasonal noise accurately. We further proposed to use the

uncertainty approach to distil more discriminant features in multi-temporal images. Experimental results on multi-view remote sensing image data sets demonstrated the superiority and efficiency of the proposed approach over other state-of-the-art methods. The proposed approach is yet unable to both handle long image time series and classify change types. In the future, we will explore how to track changes among time-series images and map the corresponding change types based on the spectral property of the image itself.

Chapter 4

Self-Supervised Change Detection in Satellite Image Time Series

In this chapter, we propose a two-stage approach to unsupervised change detection in satellite image time-series using contrastive learning with feature tracking. By deriving pseudo labels from pre-trained models and using feature tracking to propagate them within the image time-series, we improve the consistency of our pseudo labels and address the challenges of seasonal changes in long-term remote sensing image time-series. We adopt the self-training algorithm with ConvLSTM on the obtained pseudo labels, where we first use supervised contrastive loss and contrastive random walks to further improve the feature correspondence in space-time. Then a fully connected layer is fine-tuned on the pre-trained multi-temporal features for generating the final change maps. Through comprehensive experiments on two datasets, we demonstrate consistent improvements in accuracy on fitting and inference scenarios.

4.1 Introduction

The challenge of detecting changes in RS images time-series is compounded by the presence of seasonal noise, which can be difficult to distinguish from true changes. One approach to address this challenge is to use graph-based methods [65], which present detected spatio-temporal phenomena as evolution graphs composed of spatio-temporal entities belonging to the same geographical location in multiple timestamps. Deep learning methods have also been applied to RS image time-series change detection, using techniques such as recurrent neural networks (RNNs) [106] to extract discriminative features from image sequences. However, supervised methods often require a large number of labelled training samples,

which can be difficult to obtain for long image time-series. In this context, self-training approaches such as self-supervised and pseudo-label learning have become popular, where networks are trained on a pretext task such as image restoration using 3D CNN [109, 110] and predict the correct order of shuffled image sequences [136]. For example, Kalincheva et al. [81] proposed a framework combining a graph model and pseudo labels, which associates changes in consecutive images with different spatial objects using a gated recurrent unit (GRU) AE-based model. Meshkini et al. [109] further proposed the use of a pre-trained 3D CNN to extract spatial-temporal information from long satellite image time-series, where they can detect the times and locations of changes in image sequences. However, pseudo labels often have a high level of noise and do not consider temporal information, and the pre-trained model can not adapt to various changes.

In this work, we propose the use of contrastive learning [86] and feature tracking [5] to address these challenges and improve the performance of change detection in RS image time-series. We leverage contrastive learning methods both to get good pre-trained features for pseudo label generation and to reduce the overfitting that results in incorrect pseudo labels when considering supervised contrastive learning [86] and contrastive random walks [78]. Additionally, by incorporating a feature tracking-based pseudo label generation task and a convolutional long short-term memory network (ConvLSTM) [144], we are able to extract time-series change maps from image time-series and further train a new model from scratch. In detail, the pseudo-label generation is based on the pre-trained model using contrastive learning. The change detection model is trained from pseudo labels by the joint use of Unet [170] and ConvLSTM networks. We first extract pseudo labels from change pair time-series and then use them with images to train the proposed network, which outputs change maps relative to the first image in the sequence. During the training, supervised contrastive loss, contrastive random walk loss and logistic regression are used to optimize the parameters of the feature encoder and the last classifier, respectively. The supervised contrastive loss is used to mitigate the noise in pseudo labels, while the contrastive random walk loss improves the quality of the consecutive change results. Finally, we demonstrate the effectiveness of our approach on two data sets.

In this chapter, we propose the following main novel contributions:

- To generate time-related pseudo labels for network training, we propose to use feature tracking to extract reliable change pixels in image sequences that are insensitive to seasonal noise.

- To ensure the robustness and consistency of change maps, we propose to use supervised contrastive loss and contrastive random walk loss on change feature learning. These

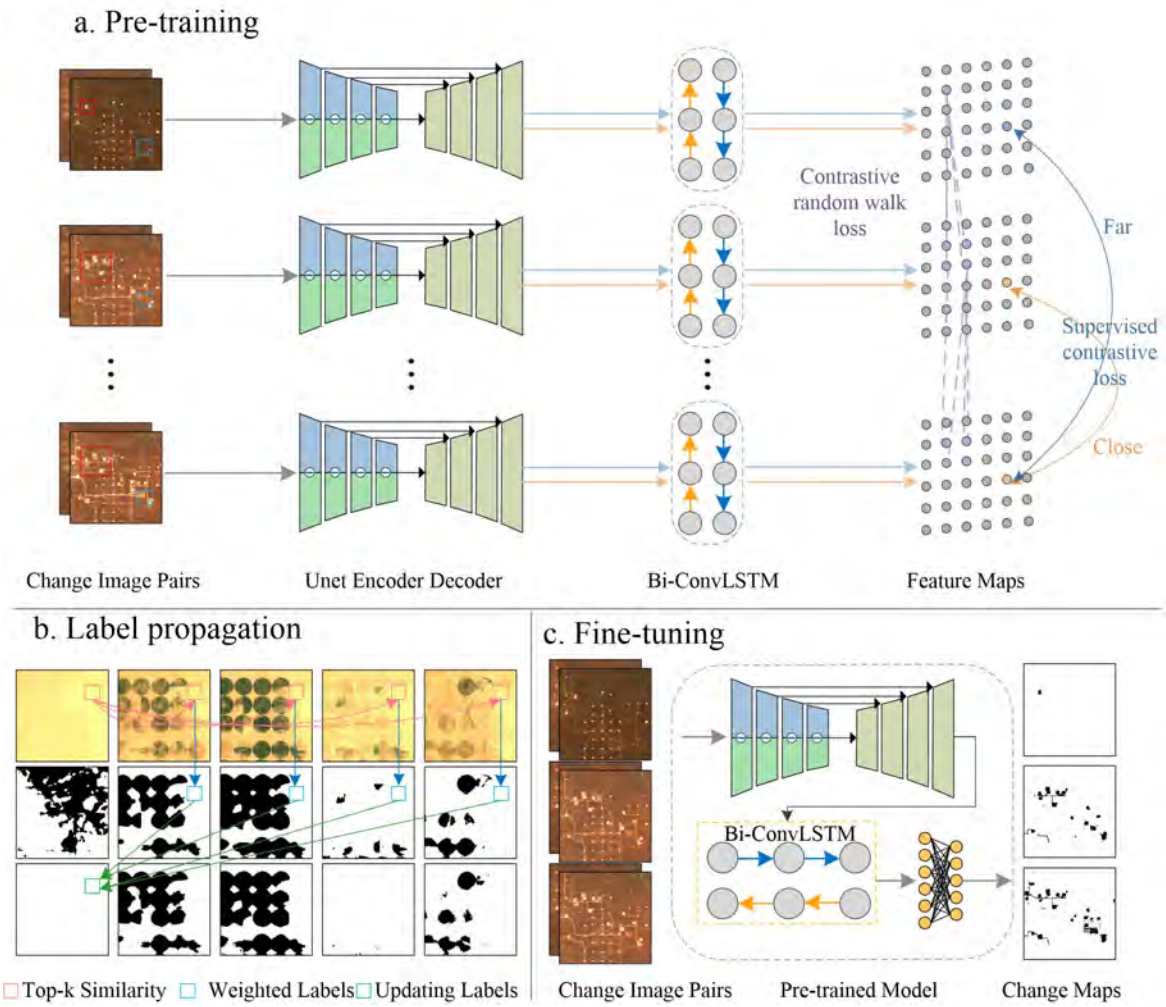

Fig. 4.1 Overview of the proposed approach for RS image time-series change detection, where the proposed network is based on the Unet and Bi-ConvLSTM. a. The pre-training step uses supervised contrastive learning on spatial feature representation and uses contrastive random walk loss on temporal features for temporal feature modelling. b. The label propagation step uses k-NN for noise reduction among change map time-series. c. The fine-tuning step uses an MLP and logistic regression to predict the final change maps.

losses encourage the pixels in the same class to have a closer feature representation among image time-series.

- To extend the approach to arbitrary long time-series, we jointly use Unet and ConvLSTM as the model architectures. To verify the performance of the proposed approach, we provide a comparison with state-of-the-art methods and an ablation study. Our experiments show that our method obtains competitive results on the datasets.

4.2 Methodology

In this section, we present the proposed two-stage RS image time-series change detection framework. It includes a feature tracking-based pseudo label generation task and a self-training change-detection module that follows the training setting of supervised contrastive learning [86]. We first get the pixel-wise feature representation of each image in the image sequence using the pre-trained model [30] and then get the pseudo change maps using the thresholding approach. Then, the feature tracking approach is used to update the threshold-based pseudo change maps. Afterwards, the pseudo change labels are used to learn the representation of change maps using the supervised contrastive loss and the contrastive random walk loss. Finally, a fully connected layer is fine-tuned on the learned change map representation using logistic regression. In the following subsections, we will describe the network architecture of the proposed framework, the supervised contrastive loss, the contrastive random walk loss and the feature tracking-based pseudo label updating.

4.2.1 Network Architecture

The proposed approach uses an Unet-ConvLSTM network architecture, which consists of two components: ResUnet and Bi-ConvLSTM 4.1. For the Unet, we adopt a similar architecture as the FC-Siam-diff [36]. Instead of concatenating features from two encoders, it instead concatenates the absolute value of their difference. It consists of two encoders, one bridge, one decoder, and skip connections between the downsampling and upsampling paths. The decoder part has three blocks, each of which consists of a convolution layer (Conv), batch normalization (BN), ReLU, and upsampling. A 1×1 Conv is used after the last block to reconstruct the learned representations. We changed the padding type of all blocks to "same" padding. The parameters and channel size of each unit are presented in Table 4.1. Each convolution unit ($[\cdot]$) includes a convolutional layer, a BN layer, and a ReLU activation layer. Each residual block (ResBlk) in the encoding path has two residual units, each of which consists of two convolution units and an identity mapping.

The output features of time-series change pairs are given in the input to the Bi-ConvLSTM layer. Different from the standard LSTM, ConvLSTM uses convolution operations in the input-to-state and state-to-state transitions to improve the modelling of the spatial correlation among sequence images. It consists of an input gate i_t, an output gate o_t, a forget gate f_t, and a memory cell C_t. The input, output and forget gates act as controlling gates to access, update, and clear memory cell. ConvLSTM can be formulated as follows (for convenience

Table 4.1 Structure of the proposed network.

	Encoder 1 & 2	Decoder	
Conv1 Maxpool	[3×3, 32], stride 2 3×3, stride 2	Cat. Diff. ResBlk2 [3×3, 64] upsampling 2	DecBlk1 stride 1
ResBlk1 stride 1	$\begin{bmatrix} 3\times3,\ 32 \\ 3\times3,\ 32 \end{bmatrix}\times2$	Cat. Diff. ResBlk1 [3×3, 32] upsampling 2	DecBlk2 stride 1
ResBlk2 stride 2	$\begin{bmatrix} 3\times3,64 \\ 3\times3,64 \end{bmatrix}\times2$	Cat. Diff. Conv1 [3×3, 32] upsampling 2	DecBlk3 stride 1
ResBlk3 stride 2	$\begin{bmatrix} 3\times3,128 \\ 3\times3,128 \end{bmatrix}\times2$	[3×3, 16] Bi-ConvLSTM	LSTM Blk
Bridge stride 1	[3×3, 128] upsampling 2	[1×1, 32/1]	Logistic Regression

we remove the subscript and subscript from the parameters):

$$\begin{aligned} i_t &= \sigma\left(W_{xi}X_t + W_{hi}H_{t-1} + W_{ci}C_{t-1} + b_i\right) \\ f_t &= \sigma\left(W_{xf}X_t + W_{hf}H_{t-1} + W_{cf}C_{t-1} + b_f\right) \\ C_t &= f_t \circ C_{t-1} + i_t \tanh\left(W_{xc}X_t + W_{hc}H_{t-1} + b_c\right) \\ o_t &= \sigma\left(W_{xo}X_t + W_{ho}H_{t-1} + W_{co} \circ C_t + b_c\right) \\ H_t &= o_t \circ \tanh\left(C_t\right) \end{aligned} \tag{4.1}$$

where o denotes the Hadamard functions. X_t is the input tensor and H_t is the hidden state tensor.W_{x*} and W_{h*} are 2D convolution kernels corresponding to the input and hidden state, respectively, and b_i, b_f, b_o and b_c are the bias terms. In this study, we employ Bi-ConvLSTM [62, 146] to encode the features of time-series change pairs. It was proposed to use both past and future information to model sequential data. Bi-ConvLSTM uses two ConvLSTMs to process the input data in both forward and backward directions and then makes a decision for the current input by taking into account the data dependencies in both directions. It has been shown that analyzing both forward and backward temporal perspectives improves predictive performance. Each forward and backward ConvLSTM can be considered as a standard one, with two sets of parameters for backward and forward states. The output of Bi-ConvLSTM

is calculated as follows:

$$\mathbf{Y}_t = \tanh\left(\mathbf{W}_y^{\overrightarrow{H}}\overrightarrow{H}_t + \mathbf{W}_y^{\overleftarrow{H}}\overleftarrow{H}_t + b\right) \tag{4.2}$$

where $\overrightarrow{H}_t$ and $\overleftarrow{H}_t$ denote the hidden state tensors for forward and backward states, respectively, b is the bias term, and $\mathbf{Y}_t$ indicates the final output considering bidirectional spatio-temporal information. The hyperbolic tangent (tanh) is used to combine the output of both forward and backward states in a non-linear manner. After the last layer of Bi-ConvLSTM, an MLP block is used to reconstruct output features at the feature learning stage and predict the binary change maps at the finetuning stage.

4.2.2 Loss Function

During training, each image $I_i (i > 0)$ is used to construct a change pair anchored at the initial image ($I_{i=0}$). In this way, the proposed network captures the temporal changes related to the first image rather than the cumulated changes of the image sequence. The training process uses a teacher-student paradigm and the exponential moving average (EMA) algorithm [150]. The inputs of the student network are the original time-series image pairs, while the teacher network uses the same time-series image pairs with color jitter.

According to supervised contrastive learning, the training process consists of the feature learning and fine-tuning stages. In the feature learning stage, we use supervised contrastive loss [86] with the contrastive random walk loss [78]. The loss can be written as:

$$L_{\text{feat}} = L_{sc} + \lambda L_{crw} \tag{4.3}$$

where L_{sc} is the supervised contrastive loss, and L_{crw} is the contrastive random walk loss. In the finetuning stage, we use logistic regression to directly predict the change probability. The hyper-parameter λ is used to tune the loss. A value of $\lambda = 0.1$ generally performed well in our experiments.

Image time-series change detection is often treated as a simple extension of bi-temporal change detection in time. The supervised contrastive loss is used to differentiate representations between changed and unchanged pixels spatially in time-series change pairs. However, the incorporation of temporal information to mitigate seasonal noise poses a significant challenge because the change depicted in frame t might not have any relation to what we find at the same location in frame $t+k$. To overcome this limitation, the contrastive random walks leverage pathfinding on a space-time graph and associate features across space and

time. It establishes feature correspondence shared by neighboring frames. In the following, we provide details on the supervised contrastive loss and the contrastive random walk loss.

Contrastive loss

The supervised contrastive loss is calculated by sampling the pixel features in the constructed time-series pairs. The pixel feature pairs at the same location in the output of the teacher and student networks are called positive pairs, while pixel features from different locations are called negative pairs. Given a positive feature pair (v_1^i, v_2^i) and a pixel feature v_2^j taken from another location, the contrastive loss can be formulated as L_{contrast}:

$$L_{\text{contrast}} = -\mathop{\mathbb{E}}_{S}\left[\log \frac{e^{sim(v_1^i, v_2^i)/\tau}}{\sum_{j=1}^{N} e^{sim(v_1^i, v_2^j)/\tau}}\right] \tag{4.4}$$

where *sim* is a similarity function (i.e., cosine similarity), (v_1^i, v_2^i) is the normalized latent representation of pixel i, $(v_1^i, v_2^j | j \geq i)$ is the normalized latent representation of negative pair and $S = \{s_1^1, s_2^1, s_2^2, \cdots, s_2^{N-1}\}$ is a set that contains $N-1$ negative samples and one positive sample. One limitation of self-supervised contrastive learning is that, since the class labels of the inputs are ignored, samples from the same class may end up being treated as negative pairs, which can affect the training performance. To avoid this limitation and enable the contrastive loss to learn in a supervised fashion, Khosla et al. [86] extended the approach to account for input labels. Following the original supervised contrastive learning method, we randomly sample N pixel features in each change pair from the teacher-student network, generating two data views $\{(x_i, y_i)\}_{i=1}^{2N}$, where $i \in I = [2N]$ is the index of an arbitrary sample. Given $A = \{A_{i,j} | y_i = y_j, (x_i, y_i), (x_j, y_j)\}$, we perform supervised contrastive learning with sampled pixel features:

$$L_i^{\text{sup}} = \sum_{i \in I} \frac{-1}{|A(i)|} \cdot \sum_{a \in A(i)} \log \frac{e^{\text{sim}(x_i, y_a)/\tau}}{\sum_{b \in B(i)} e^{\text{sim}(x_i, y_b)/\tau}} \tag{4.5}$$

where $B(i)$ means the set of indices excluding i, i.e., $B(i) = I \backslash i$; $A(i) = \{A_{i,j} | j \in B(i)\}$ is the positive set distinct from sample i and $|\cdot|$ stands for cardinality. In this case, the labels are binary pseudo labels.

Contrastive Random Walk Loss

This work builds upon the contrastive random walk framework by Jabri *et al.* [78]. In the contrastive random walks, we are given an input image time-series with k frames. We select

N nodes q_t from the frame t, which serve as vertices of a graph. Pairwise similarities of nodes are converted into non-negative affinities by applying a softmax function (with temperature τ) over edges departing from each node. This process generates the stochastic affinity matrix between frames t and $t+1$ as a bipartite graph with each edge given as:

$$C_t^{t+1}(i,j) = \text{softmax}(q_t q_{t+1}^T)_{i,j} \tag{4.6}$$

The affinity matrix for the entire graph denotes the edge weights between all pairs of nodes in change image pairs. To model affinities over multiple change image pairs, we take the product of the sequential affinity matrices:

$$\overline{C}_t^{t+k} = \prod_{i=0}^{k-1} C_{t+i}^{t+i+1} \tag{4.7}$$

Ultimately, we train the model to maximize the likelihood of cycle consistency, i.e. the event that the walker returns to the node it started from:

$$L_{crw} = -tr(log(\overline{C}_t^{t+k}\overline{C}_{t+k}^{t})) \tag{4.8}$$

4.2.3 Pseudo Label Updating

Bi-temporal remote sensing image change detection usually gets the change map through thresholding methods, which inevitably introduce errors when there is no change between bi-temporal images. The pseudo-change maps can be made less noisy by propagating the threshold-based pseudo labels to each change pair using the label propagation algorithm [5]. This is because the false alarms in one frame can be mitigated by similar features with correct labels from other frames. This algorithm propagates labels in the feature space considering both spatial and temporal neighbours. In detail, the labels of target nodes are determined by computing the matrix of transitions between target nodes and source nodes, considering only the top$-k$ transitions, and multiplying it by the labels of the source nodes. Given the feature embedding of a frame I_t and a one-hot format label of the frame I_{t-1}, we compute its cosine similarity with the feature embedding of the frame I_{t-1}:

$$M_{t-1,t} = f(\phi(I_{t-1}), \phi(I_t)) \tag{4.9}$$

Then we compute the label y_i of the pixel i in I_t according to the label of pixel j in I_{t-1}:

$$y_i = \sum_j M_{t-1,t}(j,i) y_j \tag{4.10}$$

where $M_{t-1,t}(j,i)$ is the affinity between pixel i of I_t and pixel j of I_{t-1}, and ϕ is the encoder. For each pixel i, we propagate from the top N pixels with the greatest affinity $M_{t-1,t}(j,i)$ for each pixel i. For example, let us assume we get a false alarm f_{ij} at the pixel $(i,j)_t$ of the frame t. Then we compute its cosine similarity with the feature of pixel $(i,j)_{\setminus t}$ from the remaining frames in the time series and select the $N = 10$ feature embeddings with the highest similarity. We then use these embeddings to compute a weighted sum of the label predictions at the pixel $(i,j)_t$ of frame t. This process is repeated for all change pairs in the image sequence, updating the labels and embedding contexts using the top-N change pairs. Finally, the false alarms in the frame $(i,j)_t$ are corrected by the pseudo labels from other frames.

4.3 Experimental Description

In this section, we first describe the datasets used in our experiments and then introduce the related experiment setting on the network training and the pseudo-label updating. Finally, we present the results of the proposed approach and the comparison methods. We also present an ablation study of each component of the proposed approach and the hyperparameter in the loss function.

4.3.1 Description of Datasets

We conducted experiments on two multi-spectral datasets, one from the Sentinel-2 satellite constellation and the other from the Landsat-8 satellite.

Sentinel-2 dataset

The Multi-temporal Urban Development (MUDS) dataset [152] was designed to monitor urbanization by tracking changes in building construction from 2017 to 2020. It is an open-source dataset that includes native Planet 4-meter resolution imagery and Sentinel-2 multi-spectral images with irregular observation intervals across six continents. However, the original Sentinel-2 images often contain clouds and missing values. To improve the utility of this dataset, we selected only 74 locations with a minimum of 12 clean images and resized each image to 512×512 pixels. Of these 74 locations, we labelled the significant change pair of 53 scenes for future evaluation, with all change pairs referenced to the first image. We used only four bands in this work, all of them with a spatial resolution of 10 meters. Due to the unsupervised nature of this dataset, we only considered three types of changes: built-up, bare land, and water.

Landsat-8 dataset

The UTRnet dataset [162] was specifically designed for validation of the UTRnet model. The dataset consists of the satellite image time-series collected by Landsat-8 from 2013 to 2021, with a spatial resolution of 30 meters. Four spectral bands covering the visible to the shortwave infrared region are used, including blue, green, red, and near-infrared bands. The dataset includes nine typical scenes located in different cities in China, each with a different land-cover type. For each scene, ten cloud-free Landsat-8 images were selected to cover different seasons. The image size for each scene is 400×400 pixels. The ground truth includes three classes: changed pixels, unchanged pixels, and unlabeled pixels. The changed and unchanged pixels are labelled using Google Earth images. In this study, unlabeled pixels are treated as unchanged pixels to validate the influence of seasonal noise. Due to the temporal limitations of high-resolution image labelling, the labels only include the longest interval pairs. The change maps include city expansion, water change, and soil change.

4.3.2 Experiment Settings

Evaluation Metrics

In order to evaluate the effectiveness of different methods in binary change detection, this paper employs five evaluation metrics: precision (Pre), recall (Rec), overall accuracy (OA), F1 score (F1), and Cohen's kappa score (Kap).

Implementation Details

In the process of generating pseudo labels, we first derived pseudo labels of each change pair using a thresholding approach on pre-trained features [30]. Then propagate the threshold-based labels to each change pair using the feature tracking approach. In the setting of feature tracking parameters, the spatial neighbours P are set to 10, the temporal neighbours N_T are set to 3 most correlated change pairs, and the top_k pixel is set to 10. In the self-training algorithm, the proposed approach uses a two-layer Bi-ConvLSTM. We choose the Adam optimizer with an initial learning rate of $3e^{-4}$ at the feature learning stage, which is decreased using step scheduling without restarts. The batch size is set to 2 and the model is trained for 200 epochs. For the finetuning, we use an SGD optimizer with a learning rate of 0.01, a mini-batch size of 10 and a number of epochs equal to 10. To evaluate the proposed approach, it is compared with the state-of-the-art method UTRnet in fitting and unseen scenarios. UTRnet is an improved LSTM-based self-training approach that uses CVA to generate pseudo labels. Unlike the proposed approach, UTRnet is not designed to generalize to unseen scenarios and

requires fitting a separate model for each scene. In addition to the image time-series change detection approaches, we also considered three state-of-the-art unsupervised multi-temporal image change detection approaches [30, 103**?**]. The PixSSL change detection approach [30] exploits the contrastive loss and uncertainty approach to align multitemporal images. The CAA [103] is an autoencoder-based generative model and the GMCD [**?**] is based on graph convolutional network (GCN) and metric learning.

In the evaluation, we choose the fitting evaluation on the Landsat-8 dataset due to the lack of training data, whereas considered the inference evaluation on the Sentinel-2 dataset. For the fitting scenarios of the Landsat-8 dataset, we chose Scene 3, Scene 5 and Scene 7 as the evaluation set. For the unseen scenarios of the Sentinel-2 dataset, we chose scene $T1286\_2921\_13$, scene $T1736\_3318\_13$ and scene $T6730\_3430\_13$ as the evaluation set. In addition to the comparison with UTRnet, we also conduct extensive ablation experiments on the labelled 53 scenes of the Sentinel-2 dataset to evaluate the impact of different components of the proposed approach and using different pseudo labels. In particular, the proposed approach is compared with its versions that do not use the contrastive random walk loss or only use logistic regression directly. It should be noted that only the change pair with the most significant change is labelled as ground truth for evaluating the performance of different approaches.

4.4 Experimental Results

4.4.1 Experimental Results on Landsat-8 Image Time-series

In this chapter, the effectiveness of the proposed approach is evaluated using the Landsat-8 dataset. The performance of the proposed approach is compared with the state-of-the-art approach UTRnet, which has been validated by fitting on each scene in the dataset. In order to evaluate the generalization capability of the proposed approach, results are provided for fitting on all scenes, while UTRnet results are provided for both fitting on each scene and fitting on all scenes. Quantitative evaluation is performed on the most significant change map of the change map time-series, due to the challenges of differentiating changes in continuous change scenarios. The results of the proposed approach and comparison methods are presented in Table 4.2. Among all bi-temporal image change detection approaches, the PixSSL obtained a Kappa coefficient of 0.557, which outperforms the results obtained by GMCD and CAA. This is because the PixSSL used the uncertainty approach to further suppress the seasonal noise among multi-temporal images with respect to the only use of the self-supervised learning method. In the one-scene fitting setting, UTRnet achieves an OA

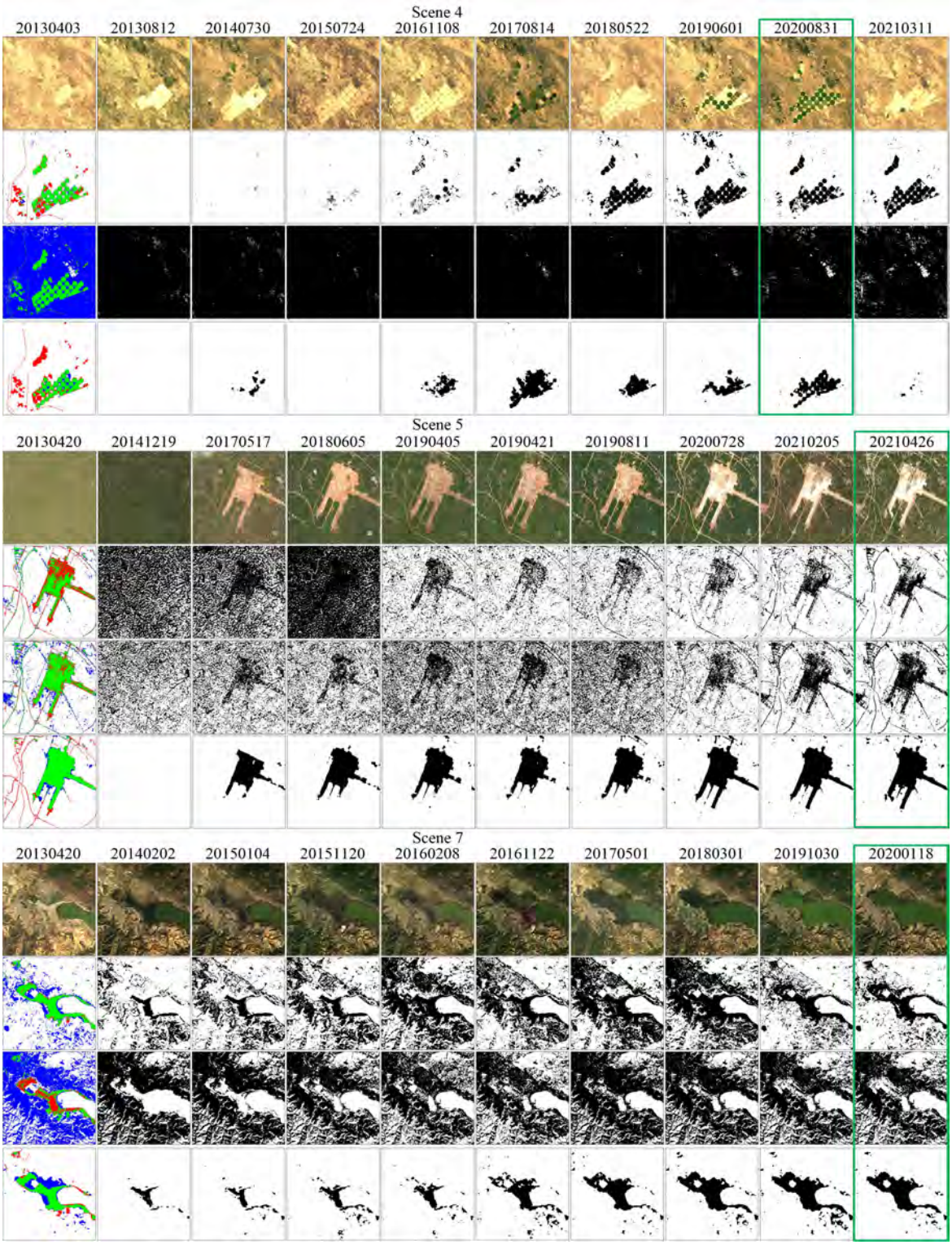

Fig. 4.2 Examples of change detection results on three scenes for the Landsat-8 dataset. Row 1: image time-series; Row 2: change maps of one-scene fitting obtained by UTRnet; Row 3: change maps of all-scene fitting obtained by UTRnet; Row 4: change maps of all-scene fitting obtained by the proposed approach. Col. 1 of Row 2, 3, 4 in each scene is the most significant change map versus the ground truth (Green: TP, White: TN, Blue: FN, Red: FP). The Green box indicates the most significant changed image pair.

Table 4.2 Quantitative evaluations of different approaches applied to the fitting test set on the Landsat-8 dataset.

Method	Pre(%)	Rec(%)	OA(%)	F1	Kap
GMCD	38.06	42.17	83.51	0.400	0.305
CAA	61.09	49.14	89.29	0.545	0.485
PixSSL	53.11	76.01	88.07	0.625	0.557
UTRnet (One-Scene)	60.86	62.57	89.66	0.617	0.557
UTRnet (All-Scene)	17.50	**77.19**	48.50	0.285	0.087
Proposed. (All-Scene)	**76.78**	69.72	**93.15**	**0.731**	**0.692**

of 89.66% and a Cohen's kappa score of 0.56, underperforming the results obtained from the pseudo labels. However, for the all-scene fitting setting, UTRnet fails to differentiate changed and unchanged pixels, achieving an OA of 48.50% and a Cohen's kappa score of 0.09. In contrast, the proposed approach achieves significantly better results than UTRnet in both settings, with an OA of 93.15% and a Cohen's kappa score of 0.69. Comparing the results of bi-temporal image change detection approaches, we can see that self-training approaches further improve the results of the bi-temporal pseudo labels using pseudo labels.

We also provide a visual comparison of the results obtained by the proposed approach and the UTRnet method. We present the results of UTRnet obtained by both fitting on each scene and on all scenes, as well as the results of the proposed approach. We also present the most significant change maps in each first column of the change maps in each scene in Fig 4.2, where true positives, true negatives, false negatives, and false positives are colored in green, white, blue, and red, respectively. From the visual comparison of the most significant change map, we can see that the change map obtained by UTRnet using all-scene fitting is noisy and contains a high number of false alarms. In contrast, the change maps obtained by the other two settings are more accurate and have less noise. In addition, the proposed approach is able to successfully detect most of the changed pixels and suppress the effects of seasonal changes. When comparing the change map time-series, we can see that the change maps obtained by UTRnet (one-scene fitting) have more false alarms that are affected by historical changes. In contrast, the change maps obtained by the proposed approach are robust to seasonal changes and only focus on real changes that happened at each time. While the one-scene fitting UTRnet still achieves good results on all test scenes, the all-scene fitting UTRnet can perform well rarely in a few scenes with less seasonal noise, but its results are heavily influenced by the imbalanced training samples.

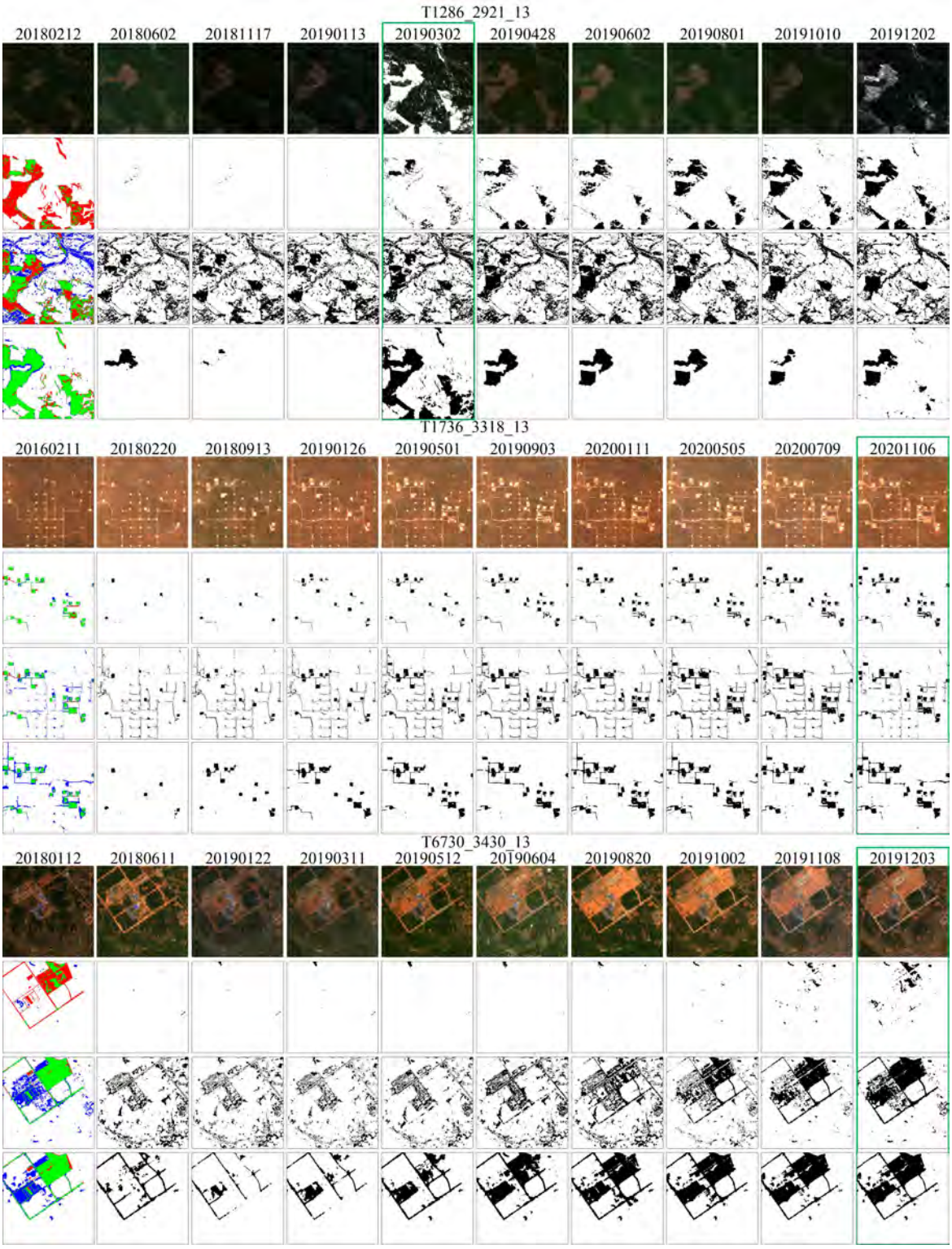

Fig. 4.3 Examples of change detection results on three scenes for the Sentinel-2 dataset. Row 1: image time-series; Row 2: change maps of one-scene fitting obtained by UTRnet; Row 3: change maps on inference setting obtained by UTRnet; Row 4: change maps on inference setting obtained by the proposed approach. Col. 1 of Row 2, 3, 4 in each scene is the most significant change map versus the ground truth (Green: TP, White: TN, Blue: FN, Red: FP). The Green box indicates the most significant changed image pair.

Table 4.3 Quantitative evaluations of different approaches applied to the inference test set on the Sentinel-2 dataset.

Method	Pre(%)	Rec(%)	OA(%)	F1	Kap
GMCD	35.78	31.52	82.45	0.335	0.235
CAA	63.85	37.16	88.23	0.470	0.409
PixSSL	55.18	**92.97**	88.02	0.693	0.624
UTRnet (One-Scene)	**85.25**	25.17	88.89	0.388	0.347
UTRnet (Inference)	46.64	64.07	84.67	0.540	0.451
Proposed. (Inference)	70.79	90.76	**93.22**	**0.795**	**0.755**

4.4.2 Experimental Results on the Sentinel-2 Image Time-series

The Sentinel-2 dataset is characterized by a diversity of land-cover scenes and a larger number of training samples. In contrast to the results obtained on the Landsat-8 dataset, we present the results of one-scene fitting UTRnet and inference on unseen scenarios based on models trained on all training samples. Similar to the evaluation on the Landsat-8 dataset, we only consider the most significant change map in each scene to assess its quantitative performance (Table 4.3). Among all bi-temporal image change detection approaches, the PixSSL obtained a Kappa coefficient of 0.624, which outperforms the results obtained by GMCD and CAA. In image time-series change detection results, as one can observe, the one-scene fitting UTRnet achieves worse results than those obtained on the Landsat-8 dataset, with an OA of 88.89% and a Cohen's kappa score of 0.35. The possible reason is that the Sentinel-2 dataset contains more seasonal changes such as snow. However, its performance is improved when inferred to unseen scenarios. Nevertheless, it still shows significant improvements compared to the all-scene fitting setting on the Landsat-8 dataset, which is largely due to the increased number and diversity of training samples. On the other hand, the inference results obtained by the proposed approach are significantly better than those obtained by UTRnet in both the one-scene fitting and inference on unseen scenarios settings. Across all five performance metrics, the proposed approach achieves the best performance in most cases, except for precision, achieving an OA of 93.22% and a Cohen's kappa score of 0.755. This indicates that the proposed approach not only outperforms the state-of-the-art method UTRnet on trained samples but also on unseen samples. In the proposed approach the improvement is more pronounced when using a larger and more diverse set of training samples. Similarly to the experiments on the Landset-8 dataset, UTRnet underperforms the results of the state-of-the-art bi-temporal image change detection approach (PixSSL) and the proposed image time-series change detection approach further improved the bi-temporal image change detection results.

In addition to the quantitative analysis, we also provide a visual comparison of the most significant change map and the change map time-series obtained by the proposed approach and the UTRnet method in each scene. Fig. 4.3 shows a comparison of all methods on the Sentinel-2 test set. The true positive, true negative, false negative, and false positive pixels of the significant change map are colored green, white, blue, and red, respectively. We first analyze the performance of the most significant change map in each scene. As shown in the figure (first column of change maps in each scene), the proposed approach successfully detects most changed pixels and suppresses seasonal noise whereas the results of UTRnet contain more false alarms and missing detections. For the change map time-series, one can see that the results obtained by one-scene fitting UTRnet contain many missed detection in particular related to the cultivated errors in the image sequence. As for the inference results, UTRnet fails to suppress most seasonal noise and presents more false alarms, while getting big improvements in noise reduction compared with its performance on the Landset-8 dataset. This issue is well addressed in the proposed approach, where abrupt changes and continuous changes are both well detected.

4.4.3 Discussion

In this section, we conduct extensive ablation studies on the proposed approach to analyze the contribution of different components. To better understand the proposed approach, we choose the scene T4780_3377_13 with significant vegetation changes over time for visualization. However, the quantitative evaluation was implemented on the selected 53 scenes of the Sentinel-2 dataset as the ablation test set.

Pseudo labels

Many unsupervised change detection approaches employ a thresholding approach for change detection. However, determining a reasonable threshold is often a challenging task. To demonstrate the effectiveness of the proposed pseudo-label generation approach, we present the pseudo-labels obtained by thresholding and feature tracking methods, individually. Then, we train the proposed approach using these two sets of pseudo-labels. Finally, we evaluate the performance of the trained models on the ablation test set. Fig. 4.4 shows the details of the pseudo-labels and the results obtained by the trained models. As one can see, the thresholding approach produces more false alarms in the pseudo change maps with the shorter time interval change pair. In contrast, the feature tracking approach can mitigate the effect of this type of seasonal change while maintaining the most significant changes in the change map time-series. Similarly, the model trained on threshold-based labels produces

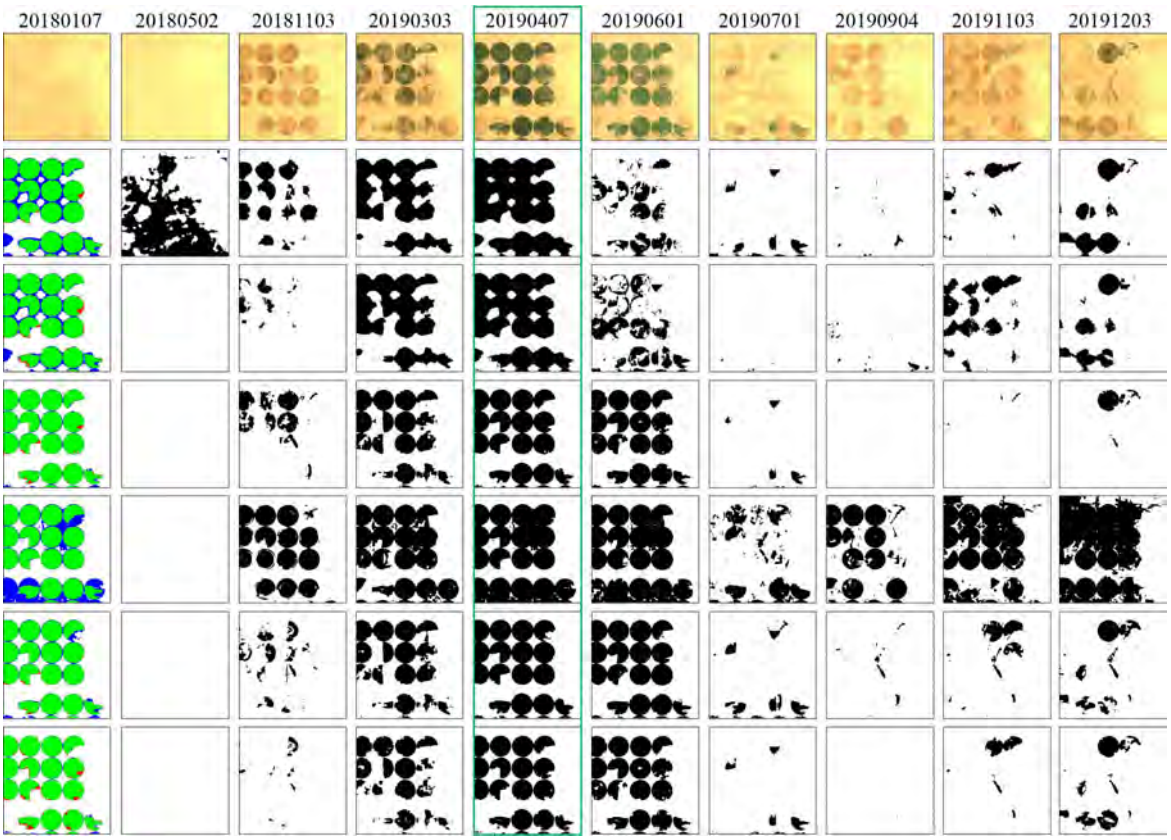

Fig. 4.4 Examples of change detection results on the Sentinel-2 ablation test set. Row 1: image time-series; Row 2: pseudo labels obtained by thresholding approach; Row 3: pseudo labels obtained by feature tracking; Row 4: change maps obtained by the proposed approach only using cross-entropy loss; Row 5: change maps obtained by the proposed approach only using contrastive loss; Row 6: change maps obtained by the proposed approach trained on threshold-based pseudo labels; Row 7: change maps obtained by the proposed approach trained on feature tracking-based pseudo labels. Col. 1 of Row 2-7 is the most significant change map versus the ground truth (Green: TP, White: TN, Blue: FN, Red: FP). The Green box indicates the most significant changed image pair.

more missed detections due to this type of noise, while the model trained on feature tracking refined labels further reduces the false alarms. Table 4.4 presents all five metrics on the ablation test set for the two trained models. Threshold-based labels and feature-tracking refined labels present similar performance on the ablation test set. This is because we only considered the most significant changed map in the evaluation. Among these results, the model trained on feature tracking refined labels provides the best result in almost all metrics, including the highest overall accuracy of 94.53% and the highest kappa coefficient of 0.669. Compared to the threshold-based labels, the OA and Kappa on the feature tracking refined labels are further improved by about 2% and 0.05. This indicates that the feature tracking

Table 4.4 Quantitative evaluations of different approaches applied to the Sentinel-2 test set in the ablation study.

Method	Pre(%)	Rec(%)	OA(%)	F1	Kap
threshold-based	49.59	77.22	91.02	0.604	0.556
feature-tracking	52.53	67.35	91.71	0.606	0.561
only logistic regression	66.46	60.12	93.77	0.631	0.597
only contrastive learning	59.27	78.23	93.30	0.674	0.638
Pro. on thres.	54.35	**82.49**	92.31	0.655	0.614
Pro. on feature.	**68.18**	71.76	**94.53**	**0.699**	**0.669**

method further improves the pseudo labels of remote sensing image time-series and thus benefits the self-training. In addition, the performance of the model trained on threshold-based labels even is worse than the one trained on feature-tracking refined labels only using contrastive loss. This demonstrates the significance of the refined pseudo labels within the proposed approach again.

Supervised contrastive loss and contrastive random walk loss

To verify the effectiveness of the contrastive loss and contrastive random walk loss, we set up experiments with training on the proposed pseudo-labels. These experiments encompassed training with both supervised contrastive loss and contrastive random walk loss, training solely with supervised contrastive loss, and training solely with logistic regression. The same ablation test set is used here. Fig. 4.4 and Table 4.4 present the results obtained under three different settings. Results show that the supervised contrastive loss and the contrastive random walk loss achieve significant improvements in noise reduction and maintain the consistency of the time-series change maps. The only use of the contrastive loss achieves the OA of 93.3% and the Kappa of 0.638, which are slightly lower than the values obtained by using both loss functions. In addition, the use of both loss functions increases by about 1% and 0.07 on the OA and the Kappa, respectively, with respect to the only use of the logistic regression. This demonstrates that the joint use of contrastive loss and contrastive random walk loss can further improve the performance of the self-training paradigm.

4.5 Discussion and Conclusion

In this chapter, we have proposed a new framework for detecting changes in RS image time-series without any manually annotated training data. Our framework jointly uses an architecture based on Unet and ConvLSTM and adopts a self-training algorithm. We

first extract pseudo labels using the feature-tracking method and then further improve the results by training a model from scratch. Feature-tracking approach detects most changes in the RS image time-series, while alleviating the presence of seasonal noise. The proposed self-training algorithm combines the use of supervised contrastive loss, contrastive random walk loss and logistic regression following the two-stage setting of supervised contrastive learning. This mitigates the effects of the noise in pseudo labels and keeps the consistency of the change map time-series. Our experiments on two different datasets demonstrate the effectiveness of the proposed approach compared to state-of-the-art methods. It is worth noting that the proposed approach can also generalize well to unseen scenarios. Although our method is demonstrated in the context of multi-spectral images, it can be applied to other sensors, such as synthetic aperture radar and RGB images.

In future work, we plan to extend our method to detect different types of changes using prior information from multi-spectral images.

Chapter 5

Self-Supervised SAR-Optical Data Fusion and Segmentation

In this chapter, we propose a self-supervised framework for SAR-optical data fusion and land-cover segmentation tasks. SAR and optical images are fused by using a multi-view contrastive loss at image-level and super-pixel level according to one of those possible strategies: in early fusion, intermediate fusion and late fusion. For the land-cover segmentation task, we further propose a self-supervised approach by jointly using the previous fusion framework and the vector quantization. Experimental results show that the proposed approach not only achieves comparable accuracy to the weakly-supervised approach but also reduces the dimension of features with respect to the image-level contrastive learning method. Among the three considered fusion strategies, the intermediate fusion strategy achieves the best performance. In addition, the further use of vector quantization brings improvements over the current state-of-the-art techniques of unsupervised land-cover segmentation on SAR-optical image pairs.

5.1 Self-supervised SAR-optical Data Fusion of Sentinel-1 and Sentinel-2 Images

In this section, we introduce a self-supervised framework for SAR-optical data fusion. SAR and optical images are fused by using a multi-view contrastive loss at image-level and super-pixel level according to one of those possible strategies: in the early, intermediate and late strategies.

5.1.1 Introduction

Every year a large number of Earth Observation Satellites are operated to monitor human activities, Earth's environment, and their mutual influences across our planet. Hundreds of terabytes of remote sensing data are accumulated per day from various systems, which cover most bands of the electromagnetic spectrum and include both active and passive sensors [34]. In this context, deep learning methods, especially supervised deep learning approaches, have been developed to process and analyze such massive amounts of multimodal RS data for specific applications, such as land-cover mapping, target recognition, and change detection. However, these applications are mostly limited to the use of a single type of image and require a large amount of labeled data for the training of the algorithm. The most common approach is based on deep learning techniques applied to single modality data, e.g., multispectral, hyperspectral, LiDAR, or Synthetic Aperture Radar (SAR). The fusion of various RS data from different sensors has not received sufficient attention yet. However, it is well known that the complementary use of multimodal RS data offers more complete information on a scene and can result in better performance in many applications [57]. For example, multispectral/hyperspectral images acquire information that characteristics land-cover categories on the basis of their spectral signatures, while radar images provide dielectric properties and are not affected by cloud occlusions.

Inspired by the success of deep learning in computer vision (CV), some works [6, 12, 92, 77] have investigated the fusion of multimodal RS data using deep learning methods. Their results have shown that deep learning techniques play a significant role in multimodal RS data fusion. However, the recent success of deep learning techniques in multimodal RS data fusion mainly focused on supervised methods, which are often limited from the availability of annotated data. Labeled remote sensing data are often scarce. The limited access to such labeled data has driven the development of unsupervised methods, such as generative models (e.g., GAN[61], CAE[108], VAE[87]). Nevertheless, recent research [96] has shown that such CNNs-based generative models overly focus on pixels rather than on abstract feature representations. In this context, unsupervised approaches are an interesting alternative to address land-cover mapping tasks [28]. Most unsupervised approaches rely on the prior information of spectral indices derived from SAR and optical images, such as normalized difference water index (NDWI), normalized difference vegetation index (NDVI), bare soil index (BI), and backscattering values (BS). These indices can be used to select training samples for network training and then segment images using the well-trained network. Even these indices can identify different land-cover classes. They are not able to extract all the semantic available in the data.

To address these limitations, we propose a new self-supervised approach to fuse the complementary information presented in SAR and optical images at the pixel level. The proposed data fuison approach can be implemented according to three pixel-wise fusion strategies: i) early fusion (PixEF), ii) intermediate fusion (PixIF) and iii) late fusion (PixLF). The proposed SAR-optical fusion approach is compared with the instance-level contrastive method under the common linear protocol in the context of the land-cover segmentation task.

The main contributions of this chapter are as follows.

1) We first introduce and verify the effectiveness of multi-view contrastive loss in SAR-optical data fusion. Then, we propose a self-supervised approach, which can obtain pixels-wise feature representations from SAR and optical image pairs without using any annotation. This is achieved by using U-Net [130] and the contrastive loss, by preserving local information at the superpixel level.

2) We compare different fusion strategies (i.e., early fusion, intermediate fusion and late fusion) in the proposed approach. Concretely, late and intermediate fusion strategies learn feature representations by comparing SAR and optical images directly, whereas the early fusion strategy distills the complementary information from a concatenation of image pairs. In addition, the efficiency of SAR-optical fusion with respect to the use of a single modality in the land-cover mapping task is analyzed.

5.1.2 Methodology

This section presents the methodology of the proposed self-supervised approach to SAR-optical image fusion, which aims to learn pixel-wise representations from unlabeled SAR-optical image pairs. Like previous self-supervised works [69, 120, 26], the three key ingredients (i.e., instance discrimination, contrastive loss and aggressive augmentation) are included in our work, where the shift transformation is used as a data augmentation approach in the contrastive paradigm.

Network Architecture

The proposed approach has two branches (5.1 (a)), where the input image of each branch has a relative shift. Each branch contains a ResUnet [170] block followed by a linear layer projector. After the projection, the same shift operation is performed on the output for feature alignment between two branches. We adopt a similar ResUnet architecture as the [170] and only use residual blocks in the encoder part. Like U-net, ResUnet consists of an encoder, a bridge, a decoder and skip connections between the downsampling and upsampling path. In this work, ResNet-18 is used as the encoder of the ResUnet block but without the

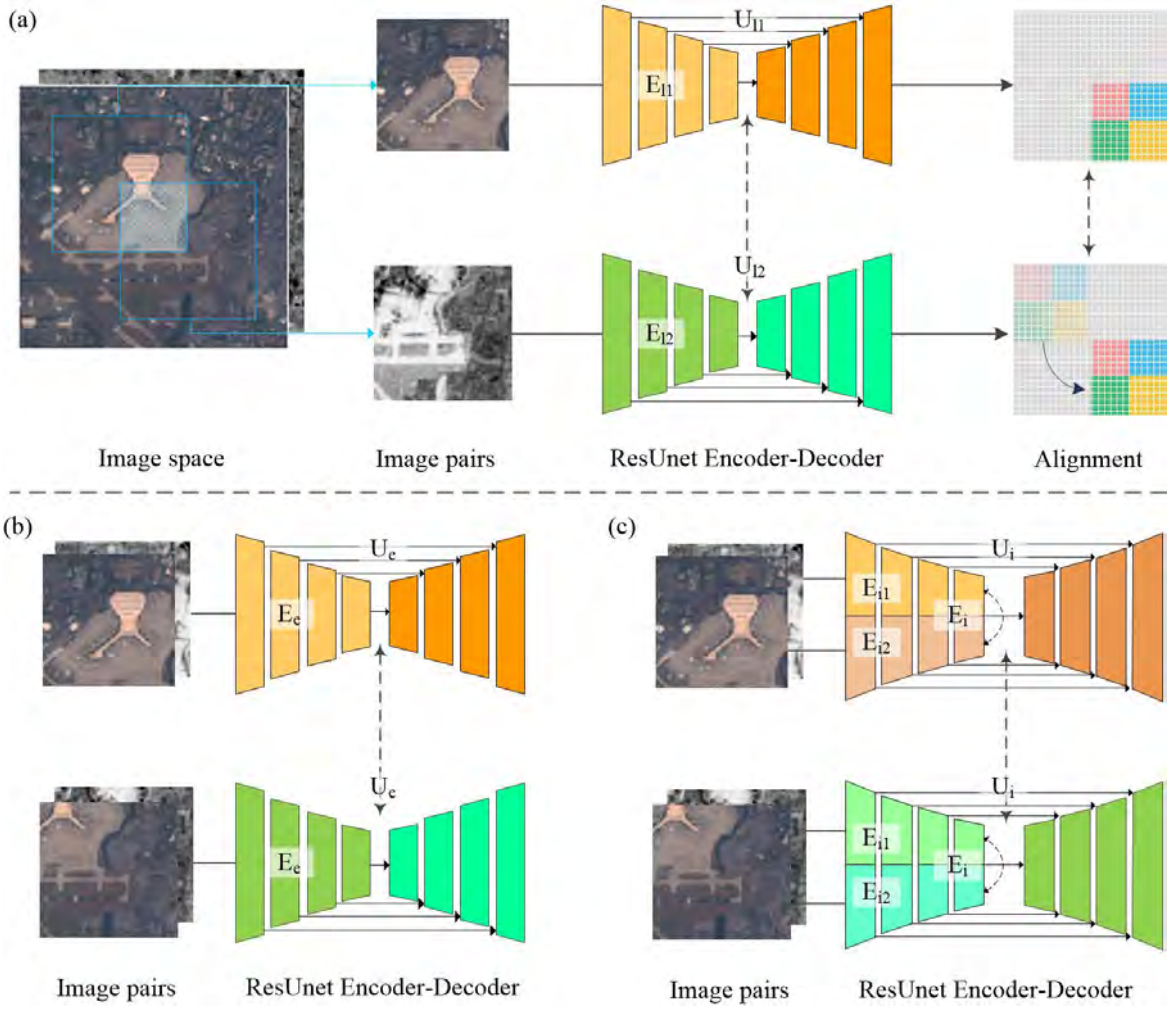

Fig. 5.1 Overview of the presented self-supervised SAR-optical fusion approach. The dash arrow line represents a contrastive loss. (a) An illustration of pixel-wise representation learning framework for the late fusion strategy. The two inputs have an offset but keep an overlap. The approach follows the common contrastive learning architecture where both branches consist of a ResUnet block and a projection. Then, a shift transformation is included in the one branch for aligning representations between two branches. (b) The ResUnet block follows the early fusion strategy. (c) The ResUnet block follows the intermediate fusion strategy where the encoder contains two parts used for encoding SAR and optical images independently.

fourth residual block. The decoder part has three blocks, where each block consists of a convolution layer (Conv) and a batch normalization (BN) layer, a ReLU activation layer, and an upsampling operation. A 1×1 Conv, following with the last block, is used to reconstruct the learned representations. All the padding types in the ResUnet block were changed to the "same" padding.

Feature alignment is achieved by using the same shift operation on both input images and output features of two branches. Specifically, given an input I_1 consisting of the SAR-optical image pair obtained from the same scene, we use a shift operation to make a random offset of I_1 along the x and y-axis directions. In this way, we can define the random shift transformation as T and obtain an augmented view $I_2 = T(I_1)$. During the training, the augmented view and the original input are fed into two branches, respectively, to obtain pixel-wise representations v_1 and v_2. To align pixel-wise representations of two branches, the same transformation is applied to the output of the other branch $v_1 = T(v_1)$.

In particular, two branches of PixEF and PixIF share the same parameters, but the encoder of PixIF is split into two groups where each group has half channels of the counterpart of PixEF in each layer. We denote the network of PixEF as U_e and its encoder as E_e. Similarly, the network of PixIF as U_i and its two independent encoders as E_{i1} and E_{i2}. Unlike these two models, the PixLF has two independent branches with half channels of PixEF in each layer, where the input channels were adjusted to the input images. We denote the network of PixLF as U_l and its two independent encoders as E_{l1} and E_{l2}. The parameters and channel size of each unit are presented in Table 5.1, where each convolution unit ([ ]) includes a convolutional layer, a BN layer and a ReLU activation layer. Each residual block (ResBlk) in the encoding path has two residual units. Each residual unit consists of two convolution units and an identity mapping.

Loss Function

The proposed approach consists of two types of contrastive loss based on images and superpixels individually. The main idea behind a contrastive loss is to find a feature representation that is invariant to augmentations. Given a dataset S that consists of a collection of image pairs $\{(s_1^i, s_2^i)\}_{i=1}^N$ across N different scenes, we consider each image pair (s_1^i, s_2^i) sampled from the joint distribution $p(s_1^i, s_2^i)$, which we call positives. Let s_2^j be taken from another scene $(j \neq i)$, then samples (s_1^i, s_2^j) sampled from the product of marginals $p(s_1^i)p(s_2^j)$, which we call negatives. The model $h(\cdot)$ is expected to know which pair is drawn from the joint distribution while the other is not exactly, by computing their cosine similarity with a hyperparameter τ. In the multi-view setting, the model $h(\cdot)$ is a neural network consiting of two branches with independent or same parameter f_{θ_1} and f_{θ_2}.

$$h(s_1, s_2) = \exp\left(\frac{f_{\theta_1}(s_1) \cdot f_{\theta_2}(s_2)}{\left\|f_{\theta_1}(s_1)\right\| \cdot \left\|f_{\theta_2}(s_2)\right\|} \cdot \frac{1}{\tau}\right) \tag{5.1}$$

Table 5.1 The network structure of the proposed PixEF, PixLF and PixIF.

Name	PixEF	PixLF	PixIF
	Encoding	Encoding	Encoding 1&2
Conv1	[3×3, 64], stride 2	[3×3, 32], stride 2	[3×3, 32], stride 2
MaxPool	3×3, stride 2	3×3, stride 2	3×3, stride 2
ResBlk1 stride 2	$\begin{bmatrix} 3\times3,\ 64 \\ 3\times3,\ 64 \end{bmatrix}\times2$	$\begin{bmatrix} 3\times3,\ 32 \\ 3\times3,\ 32 \end{bmatrix}\times2$	$\begin{bmatrix} 3\times3,\ 32 \\ 3\times3,\ 32 \end{bmatrix}\times2$
ResBlk2 stride 2	$\begin{bmatrix} 3\times3,128 \\ 3\times3,128 \end{bmatrix}\times2$	$\begin{bmatrix} 3\times3,\ 64 \\ 3\times3,\ 64 \end{bmatrix}\times2$	$\begin{bmatrix} 3\times3,\ 64 \\ 3\times3,\ 64 \end{bmatrix}\times2$
ResBlk3 stride 2	$\begin{bmatrix} 3\times3,256 \\ 3\times3,256 \end{bmatrix}\times2$	$\begin{bmatrix} 3\times3,128 \\ 3\times3,128 \end{bmatrix}\times2$	$\begin{bmatrix} 3\times3,128 \\ 3\times3,128 \end{bmatrix}\times2$
Bridge stride 1	[3×3, 256] upsampling 2	[3×3, 128] upsampling 2	[3×3,256] upsampling 2
	Decoding	Decoding	Decoding
Block4 stride 1	Cat. Block2 [3×3, 128] upsampling 2	Cat. Block2 [3×3, 64] upsampling 2	Cat. Block2 [3×3, 128] upsampling 2
Block5 stride 1	Cat. Block1 [3×3, 192] upsampling 2	Cat. Block1 [3×3, 96] upsampling 2	Cat. Block1 [3×3, 192] upsampling 2
Block6 stride 1	Cat. Conv1 [3×3, 256] upsampling 2 Conv 1×1	Cat. Conv1 [3×3, 128] upsampling 2 Conv 1×1	Cat. Conv1 [3×3, 256] upsampling 2 Conv 1×1

where s_1 and s_2 are the inputs in two branches of the network. The final loss function can be writen as $L(f_{\theta_1}, f_{\theta_2}, S)$ given the dataset S:

$$L(f_{\theta_1}, f_{\theta_2}, S) = -\mathop{\mathbb{E}}_{S}\left[\log \frac{h(s_1^i, s_2^i)}{\sum_{j=1}^{N} h(s_1^i, s_2^j)}\right] \tag{5.2}$$

where (s_1^i, s_2^i) is a positive pair sample, $(s_1^i, s_2^j | j \neq i)$ is a negative pair sample and $\{s_1^1, s_2^1, s_2^2, \cdots, s_2^N\}$ is a set that contains $N-1$ negative samples and one positive sample by anchoring at s_1^1. In the training process, the network is trained to increase the value of positive pairs and

decrease the value of negative pairs. This results in a feature representation that is close to positive pairs whereas it is not appropriate for negative pairs.

In the pixel-level contrastive loss, we sample and average features from two branches over superpixels that are located on the overlap between the two branches. This aims to keep the consistency of the normalized pixel-wise representations between two branches. Here, we construct a set of pixel-wise feature pairs $P_{i=1}^{N}$ where the positive feature pair (p_1^i, p_2^i) is sampled from the same location, while $p_2^j | j \neq i$ in negative pairs is taken from another location. Compared with the instance-level contrastive learning, this loss function can make the model get more detailed representations and thus more suitable for dense prediction downstream tasks. To overcome the noise when using the single pixel, we adopt the contrastive loss at the superpixel level. Together with the pixel-level contrastive loss, an instance-level contrastive loss is used to improve the performance. The instance-level loss help to discriminate the similarity between the shifted views. Like pixel-level loss, we can construct a set $M_{i=1}^{N}$ of concatenated image pairs, where (m_1^i, m_2^i) is sampled from the same scene i while $m_2^j | j \neq i$ is taken from another scene. Finally, we use the pixel-wise contrastive loss in conjunction with the instance-level contrastive loss, leading to the total loss of three fusion approches:

$$\begin{aligned} L_{\mathrm{e}} &= L(U_e, U_e, P) + L(E_e, E_e, M) \\ L_{\mathrm{l}} &= L(U_{l1}, U_{l2}, P) + L(E_{l1}, E_{l2}, M) \\ L_{\mathrm{i}} &= L(U_i, U_i, P) + L(E_{i1}, E_{i2}, M) + L(E_i, E_i, M) \end{aligned} \tag{5.3}$$

where L_{l},L_{e} and L_{i} are the loss functions of PixLF, PixEF and PixIF, respectively.

5.1.3 Experimental Results

In this section, we present the dataset for the training and validation of the proposed self-supervised SAR-optical fusion. Besides, the details of network setup and evaluation experiments are introduced.

Description of the Dataset

DFC2020: We developed our experiments on the DFC2020 dataset. The DFC2020, which has been issued by the IEEE-GRSS 2020 Data Fusion Contest [163], is used as the training set and the evaluation set for comparison between different methods. This dataset consists of a total of 6114 quadruple samples, which are SAR-optical image pairs, MODIS-derived labels, and more accurate semi-manually derived high resolution (10 m) land-cover maps [139]. SAR images were acquired by Sentinel-1 and consist of dual-polarized (VV and

VH) components, and the optical images were taken by multi-spectral Sentinel-2. Each SAR-optical image pair was obtained within the same season. Each pixel in the DFC2020 was assigned to a land-cover class manually, which has eight fine-grained classes (i.e., Forest, Shrub-land, Grassland, Wetlands, Croplands, Urban/built-up, Barren and Water). We also provide the image-level label for each image, which is derived by the majority class of the related pixel-level land-cover maps. Previous research [157] pointed out the effectiveness of training a CNN on image-level labels that can guide the weakly supervised model (WSL) to learn a powerful representation of images.

Training and Test Sets: A random split of the DFC2020 dataset into a training set (1000) and a test set (5114) was applied in this work. Here, the training set is used to tune the parameters of the linear protocol in the evaluating phase. The test set is used to validate the effectiveness of the features learned using different methods. To assess the effectiveness of these methods with limited labels, we randomly split the training set into five groups with 10, 50, 100, 200, and 1000 samples. Each small number group is a sub-sampled version of the corresponding full training set. Note that all self-supervised and unsupervised models were trained on unlabeled SAR-optical image pairs.

Network Setup

The training process of the self-supervised approach includes three parts for PixEF, PixLF and PixIF. For PixEF, SAR-optical image pairs were concatenated as one input. For PixIF and PixLF, SAR and optical images are in input to two branches independently. The Adam with a learning rate of $3e^{-4}$, a weight decay of $4e^{-4}$ and a momentum of 0.9 was adapted as our optimizer. We use a mini-batch size of 1000 with an input size of 16×16; models are run for 700 epochs. We deploy a step scheduling learning rate policy in the training process. We use shift transformation to augment different inputs, where the vertical and horizontal range of the pixel shift is one-fifth of the input width. Apart from the shift transformation, we further apply a random flip transformation to improve the performance of the proposed approach.

Experiment Settings

To evaluate the learned feature representation of different methods, we provide an evaluation with a linear classifier followed by the frozen features on the test set. In particular, the feature representation in the proposed approach has 256 channels, while that of the comparison methods (i.e., DCCA and MCL) is a concatenation of multi-level features with 512 channels. Note that we decided to use a linear classifier as our main evaluation metric for the quality

of representations since it is simple and has a small number of extra parameters. This is an effective way to focus on the intrinsic discrimination capability of the classifier. The learning rate is set to 0.05 and the SGD with a mini-batch size of eight was adopted as the optimizer for the linear protocol as well as the maximum number of epochs is 50.

In this experiment, we also assess the performance of the proposed approach applied to the Sentinel-1 image, the Sentinel-2 image, and both of them. This is done to validate whether the SAR-optical fusion can obtain more discriminative representations than single modality for the downstream land-cover mapping task. To this purpose, we trained the proposed approach only on single modality images in the proposed early fusion strategy. Then the linear protocol on the frozen pre-trained models was used to evaluate the effectiveness of learned representations with limited training labels. However, the instance-level self-supervised method does not have an decoder and can not perform the same strategy. To provide a fair comparison, we also adopt the same decoder and the classifier of the proposed approach as a readout. The decoder is followed by encoders pre-trained by instance-level self-supervised methods and used to reconstruct the concatenation features for downstream tasks.

Linear Evaluation on Pre-trained Features

The performance of the proposed self-supervised approaches (PixEF, PixIF and PixLF) were evaluated on the test set in comparison to the two instance-level self-supervised methods (DCCA and MCL) and the weakly supervised method (WSL). Also in this case we considered different amounts of labeled data for the training of the linear classifier (see Fig. 5.2). The

Table 5.2 Class-wise and overall accuracies achieved on the test set by a linear classifier used with the different methods considering 1000 SAR-optical training samples.

Class	Average Accuracies (%)					
	WSL	DCCA	MCL	PixIF	PixLF	PixEF
Forest	90.2	92.7	90.9	**92.6**	92.3	91.3
Shrub-land	**70.4**	26.3	40.4	50.8	46.3	53.2
Grassland	57.6	63.0	68.7	73.0	64.5	**74.5**
Wetlands	**70.4**	56.9	64.9	62.7	57.2	59.9
Croplands	81.2	72.8	82.8	84.7	**85.8**	80.0
Urban	86.6	86.9	87.7	**87.8**	84.4	87.1
Barren	34.5	6.3	**46.3**	30.9	29.5	37.0
Water	99.2	99.0	99.3	99.3	**99.4**	99.2
AA	**73.8**	63.0	72.7	72.7	69.9	72.8
mIoU	0.490	0.411	0.487	**0.498**	0.476	0.490

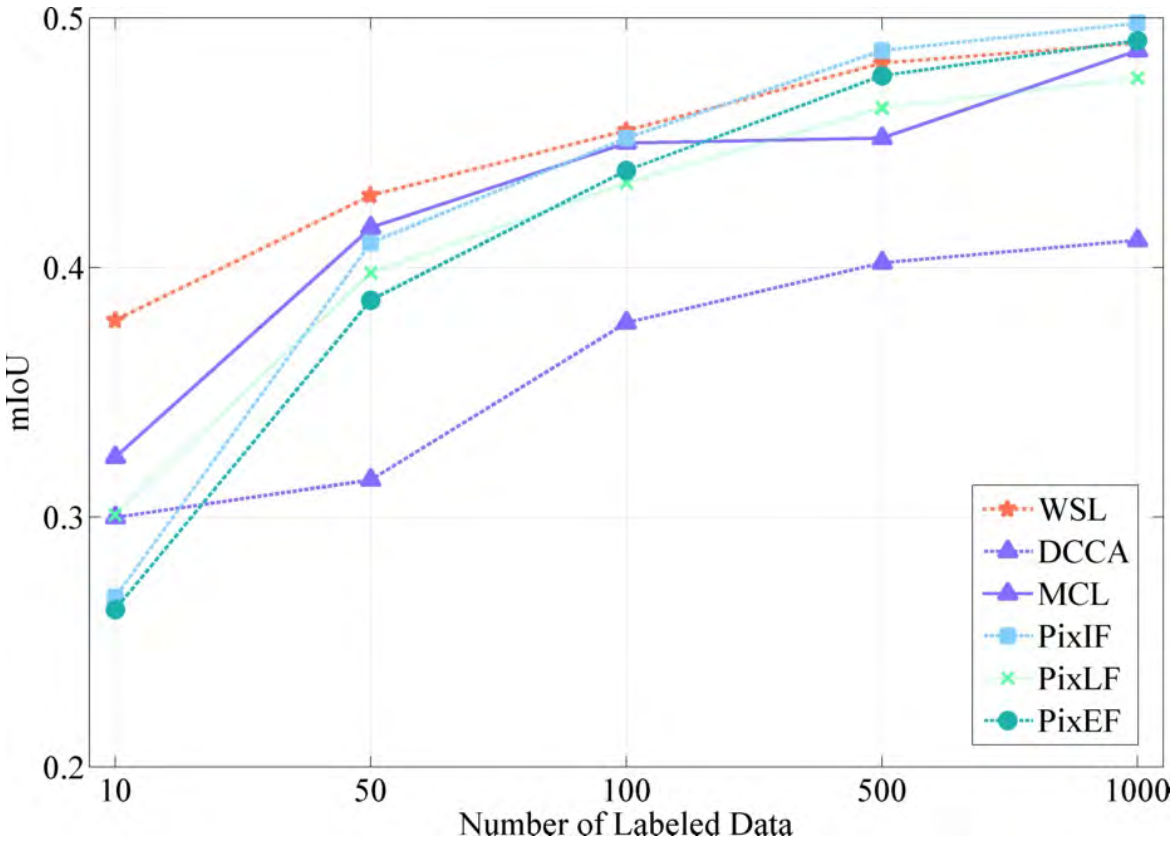

Fig. 5.2 The mean intersection over union metric (mIoU) achieved by different methods on test set versus the number of samples used for the training of the linear classifier on frozen encoders.

average class accuracy (AA) and mean intersection over union (mIoU) are common metrics used to assess the performance in land-cover mapping and are used to evaluate the overall precision of all land-cover classes in our work.

Fig. 5.2 shows the linear protocol results on mIoU. As one can see the proposed PixIF outperforms all other methods, whereas the DCCA performs significantly worse against any other methods when the number of training samples increases. However, the WSL method, which is weakly supervised, outperforms all other methods with few labels. In general, the proposed PixEF and PixLF as well as the MCL have a similar performance. The gap between WSL and all contrastive approaches is reduced when the number of labeled samples increases. In particular, the proposed PixIF outperforms WSL in the case of 1000 training samples. Among contrastive approaches, the performance of PixIF and PixEF is slightly better than that of MCL. Moreover, it is worth noting that MCL obtained representations with 512 channels, while the proposed approaches with only 256 channels. This means that

Table 5.3 Class-wise and overall accuracies achieved by PixEF on Sentinel-1 images alone (S1), Sentinel-2 images alone (S2) and Sentinel-1/-2 image fusion (S1S2) with the linear protocol and the fine-tuning evaluation.

class	Linear Evaluation (%)			Fine-tuning Evaluation (%)		
	S1	S2	S1S2	S1	S2	S1S2
Forest	84.2	90.3	**91.3**	86.2	**92.2**	92.0
Shrubland	27.5	48.9	**53.2**	31.2	44.2	**62.3**
Grassland	61.1	67.2	**74.5**	61.9	68.0	**78.0**
Wetlands	35.0	58.8	**59.9**	51.5	62.3	**62.3**
Croplands	66.0	**81.4**	80.0	71.1	79.6	**85.7**
Urban	78.8	86.5	**87.1**	84.7	**89.7**	86.1
Barren	3.5	30.0	**37.0**	7.8	30.7	**38.6**
Water	98.8	99.2	**99.2**	99.2	**99.5**	99.4
AA	56.9	70.0	**72.8**	61.7	70.8	**75.6**
mIoU	0.362	0.470	**0.490**	0.395	0.474	**0.521**

the proposed approach significantly reduces the dimensionality of features while keeping or even improving the feature representation ability.

Table 5.3 presents a detailed comparison of the class-by-class accuracy obtained by the linear classifier on the test set for each approach when trained with 1000 labeled samples. According to the results, the proposed PixIF and PixEF as well as the weakly-supervised WSL achieve an AA higher than 72% and an mIoU over 0.49 on the test set. They sharply outperform DCCA, which obtains an AA smaller than 65% and a mIoU smaller than 0.42. Among the proposed approaches, PixIF obtained the highest mIoU, whereas PixLF got the lowest value. In addition to the quantitative evaluations, we also provide a qualitative visual comparison of the land-cover maps predicted by different methods. Fig. 5.3 illustrates five examples of the results. Each example includes the land-cover maps predicted by different approaches as well as the ground truth in DFC2020. As one can observe, the proposed self-supervised PixIF, PixEF and the weakly-supervised WSL show better results than the rest of the methods in all cases, and the proposed PixIF confirms to be more effective than the other two fusion strategies (PixEF and PixLF). For each land-cover class, similar conclusions to those derived by Table 5.2 can be given.

In general, the results obtained in all comparisons confirm that the contrastive approach is superior to DCCA in this land-cover mapping task. The effectiveness of the proposed PixIF against other methods is due to its ability to use a three-level contrastive loss. It is interesting to note that despite the annotations are used in WSL, the proposed approach achieves comparable performance without any use of labels. This confirms the effectiveness of the self-supervised methods in feature representation learning.

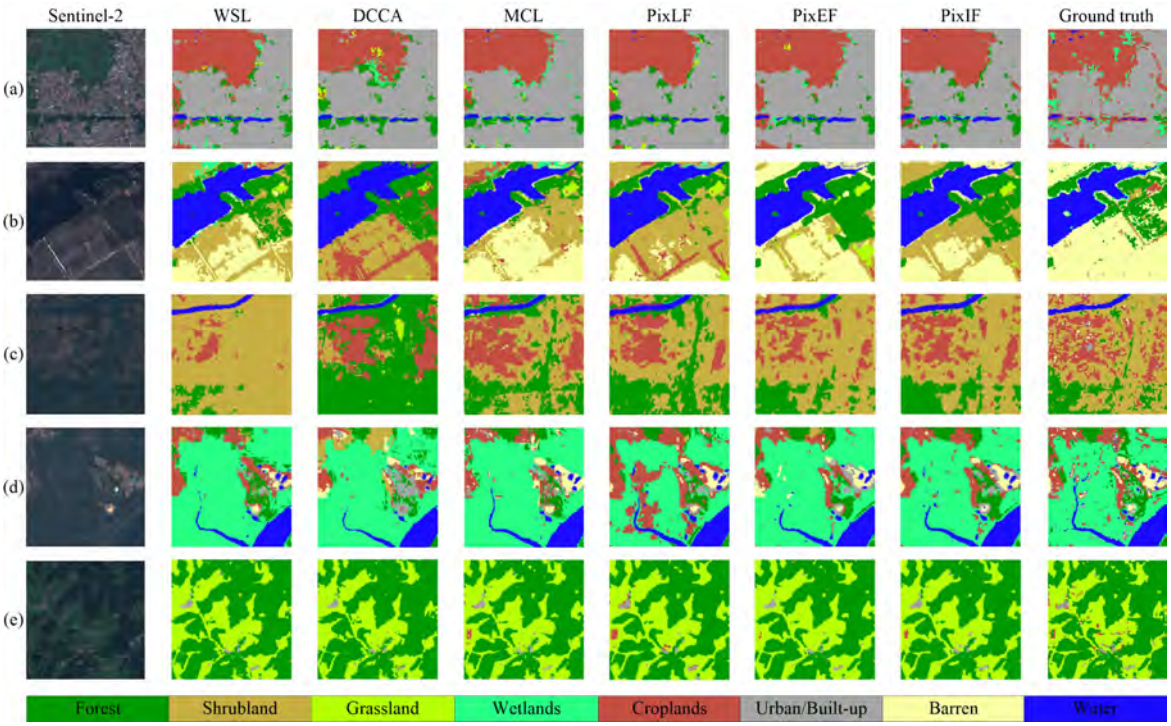

Fig. 5.3 Land-cover maps achieved on five different images by different considered methods with a linear classifier (see Table 5.2 for quantitative results).

We further investigate the performance of images from a single sensor in the land-cover mapping task, by training the proposed PixEF approach on Sentinel-1 (S1) and Sentinel-2 (S2) image as well as on the concatenation of Sentinel-1/-2 (S1S2) images. Table 5.3 shows a quantitative evaluation of the accuracy of each class by considering S1, S2 and S1S2 with the 1000 training samples and both linear and fine-tuning evaluations. As one can see, the SAR-optical fusion outperforms the use of any single modality data in both linear protocol and fine-tuning evaluations. The performance of PixEF on S2 is very close to the performance on S1S2, while the performance on S1S2 has an increase of more than 10% of AA with respect to the performance on S1 in both types of evaluation. In addition, the results of sentinel-2 images in higher classification accuracy on all classes than the use of only sentinel-1 images for both evaluation methods. For individual classes, water achieved the highest accuracy in all conditions, whereas the barren has the lowest accuracy. The classification accuracy of water does not show an obvious improvement after fusion, because there is already enough information in each single modality data. However, the SAR-optical fusion improved the performance of shrubland, grassland and barren.

Apart from quantitative assessment, we also made a visual comparison of the results obtained with both evaluations on S1, S2 and S1S2 images. The performances of each

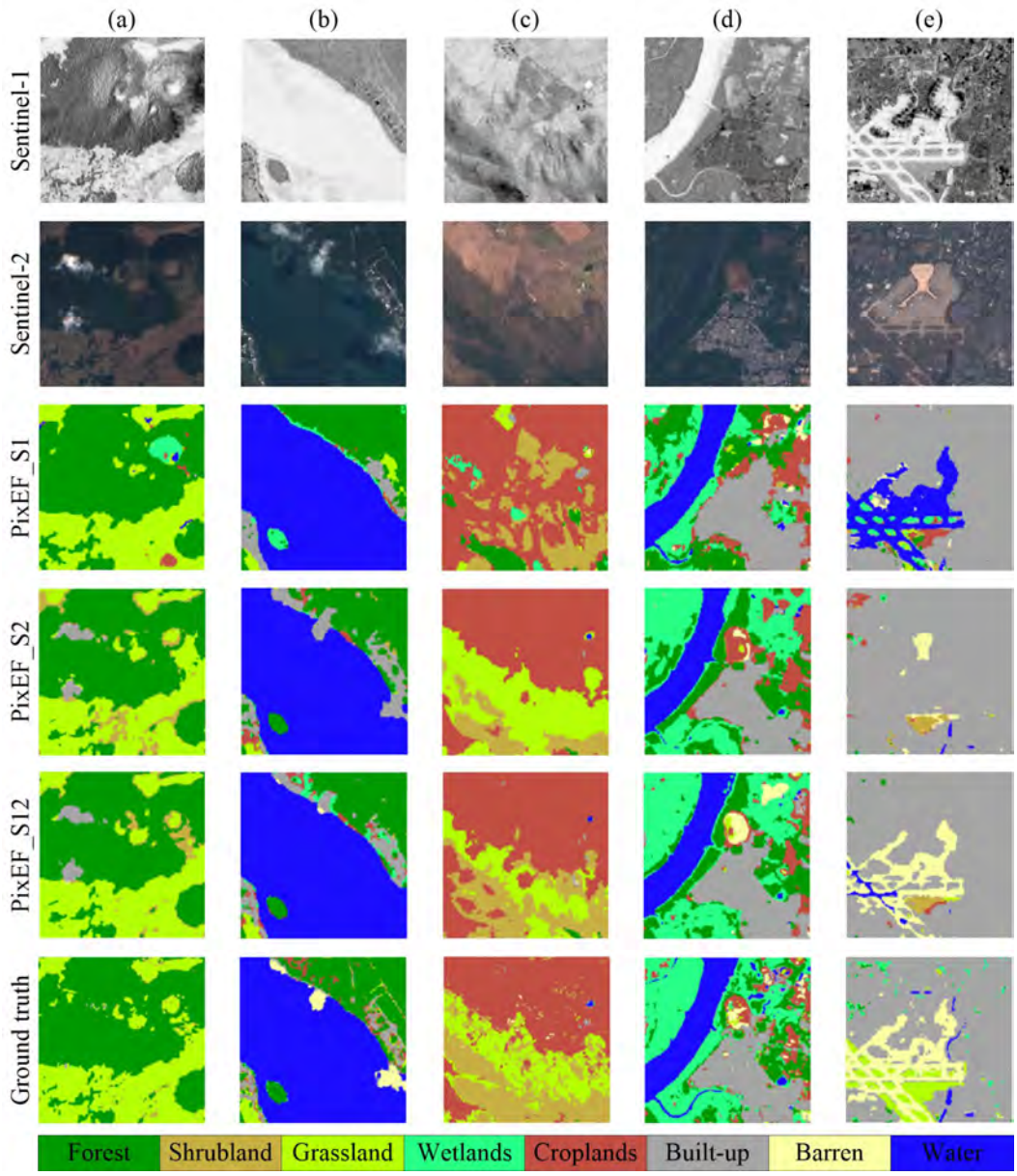

Fig. 5.4 Land-cover maps obtained by PixEF on Sentinel-1 images alone (S1), Sentinel-2 images alone (S2) and Sentinel-1/-2 image fusion with the linear classifier and fine-tuning evaluation for five different images (see Table IV for quantitative results).

modality and SAR-optical fusion keep consistent with both linear protocol and fine-tuning evaluations. As shown in Fig. 5.4, the SAR-optical fusion classifies various classes in a more accurate way, especially in barren, which obtains a significant improvement with respect to the use of single modality images. Besides, a trend can be figured out, that is, the methods with the input of S1S2 data achieve more smooth parsing results compared with the input of single modality data. Moreover, Fig. 5.4 also shows the advantages of each single

modality and other improvements after SAR-optical fusion. Both Fig. 5.4 (a) and Fig. 5.4 (b) present clouds on Sentinel-2 (but of course not in Sentinel-1) and the obvious influence on the corresponding classification maps. The presence of clouds leads to misclassifications in Sentinel-2, but not in Sentinel-1. Although the misclassification induced by clouds is still present in the classification maps of SAR-optical fusion, it is significantly mitigated compared to the results of Sentinel-2 alone. A similar phenomenon is also presented in Fig.5.4 (e), where the airport was distinguished as water in the Sentinel-1 result. This is the result of the similar backscatter between a flat runway and water. Conversely, this was correctly distinguished as built-up in the Sentinel-2 result. After SAR-optical fusion, the classification errors in each modality were obviously reduced. Overall, the visual comparison is coherent with the quantitative results presented in Table 5.3 and confirms again the effectiveness of the presented self-supervised SAR-optical fusion approach.

5.1.4 Discussion and Conclusion

Discussion

In this section, we discuss the effects of different components of PixIF that contribute to its performance. All results shown in Table 5.4 are trained and tested on the same setup of the linear protocol with 1000 training samples. The first row refers to the proposed approach using only shift operation and additional global loss. The second row refers to the proposed approach using geometric transformation (shift, rotate, resize and sheer) instead of shift operation and additional global loss. The third row refers to the proposed approach using photometric transformation (gaussian blur and noise) and additional global loss. And the fourth row refers to the proposed approach only using shift operation.

As one can see, the combination of shift operation and the use of global contrastive loss achieves the highest accuracy. In contrast, the lack of global loss makes the performance slightly decayed. This demonstrates the benefits of the use of instance-level contrastive loss. We also investigate a universal geometric transformation instead of shift operation in the

Table 5.4 The effect of the use of geometric, photometric, shift augmentation and global loss in the proposed approach.

Only Shift	Geometric	Photometric	Global loss	AA	mIoU
✓			✓	**72.7**	**0.498**
	✓		✓	68.4	0.460
		✓	✓	68.8	0.464
✓				70.6	0.479

network training. In this case, the performance drops about 0.04 on mIoU. The proposed approach can work with the geometric (affine) transformation but with little performance drops. Similarly in photometric transformation, it leads to 0.03 drops on mIoU. In general, in this work, we found that the shift operation is the most useful and simple data augmentation approach for the proposed approach.

Conclusion

In this chapter, we proposed a new self-supervised SAR-optical data fusion approach by jointly using the instance-level and pixel-level contrastive loss and the shift data augmentation. The proposed approach explores three fusion strategies to distill related representations from different modalities data. We additionally investigate the efficiency of SAR-optical fusion with respect to the single modality in the use of the proposed approach.

To evaluate the performance of the proposed approach, we compared it with two instance-level self-supervised methods (i.e., CML and DCCA) and also with a weakly supervised method (WSL) considering the linear protocol evaluation with different numbers of training samples. The results show that the proposed PixIF achieves the best performance among all self-supervised methods and a comparable performance to that of a weakly supervised method. The effectiveness of the proposed PixIF can be explained by the use of different levels of contrastive loss for a dense prediction task. Comparisons between the performance of proposed PixEF on SAR-optical fusion and single modality data were also considered. The experiment confirmed again the benefit of SAR-optical fusion in the land-cover mapping task.

5.2 Unsupervised Land-Cover Segmentation Based on Contrastive Learning and Vector Quantization

This section proposes a new unsupervised land-cover segmentation approach based on contrastive learning and vector quantization that jointly uses SAR and optical images. This approach exploits a pseudo-Siamese network to extract and discriminate features of different categories, where one branch is a ResUnet and the other branch is a gumble-softmax vector quantizer. The core idea is to minimize the contrastive loss between the learned features of the two branches. To segment images, for each pixel the output of gumble-softmax is discretized as a one-hot vector and its proxy label is chosen as the corresponding class.

5.2.1 Introduction

Land-cover maps provide spatial and categories information on land-cover classes. They are widely used for policy decisions, environmental monitoring, resource management, disaster discovery, etc. Land-cover maps are often generated by classifying pixels of images acquired by remote sensing systems. In this context, the complementary use of multimodal remote sensing data offers more complete information for land-cover segmentation tasks than the exploitation of single sensor data. For example, multi-spectral images collect rich spectral information on the land-cover categories on a wide range of the electromagnetic spectrum, while synthetic aperture radar (SAR) images provide measures on the dielectric and backscattering properties that are subjected to geometric features. Many land-cover mapping approaches have been developed to combine complementary information from SAR and optical images. Early works performed land-cover mapping tasks with machine learning approaches and proved the effectiveness of combining SAR and optical data in this task. Nevertheless, their performance is limited to the feature learning ability.

More recent works have instead used Convolutional Neural Networks (CNNs) to fuse SAR and optical images for performing land-cover maps, demonstrating the superiority of these deep learning architectures in SAR-optical fusion. However, most of them focus on supervised learning methods, which are often limited by the availability of annotated data. In this context, unsupervised approaches are an interesting alternative to address land-cover mapping tasks [28]. Most unsupervised approaches rely on the prior information of spectral indices derived from SAR and optical images, such as normalized difference water index (NDWI), normalized difference vegetation index (NDVI), bare soil index (BI), and backscattering values (BS). These indices can be used to select training samples for network training and then segment images using the well-trained network. Even these indices can identify different land-cover classes. They are not able to extract all the semantic available in the data.

Recent research [7, 120, 151, 71, 27] in contrastive learning demonstrates how these methodologies can encourage the network to learn more interpretable and meaningful feature representations. This resulted in improvements in classification and segmentation tasks, where contrastive methods outperformed the generative counterparts. Recent research on unsupervised image segmentation in the computer vision domain demonstrates that maximizing mutual information between different augmentations can encourage a deep network to learn and discriminate the features of different classes. However, existing methods, such as InfoSeg [67], rarely consider land-cover tasks and do not show robust performance on scenarios with more than two classes.

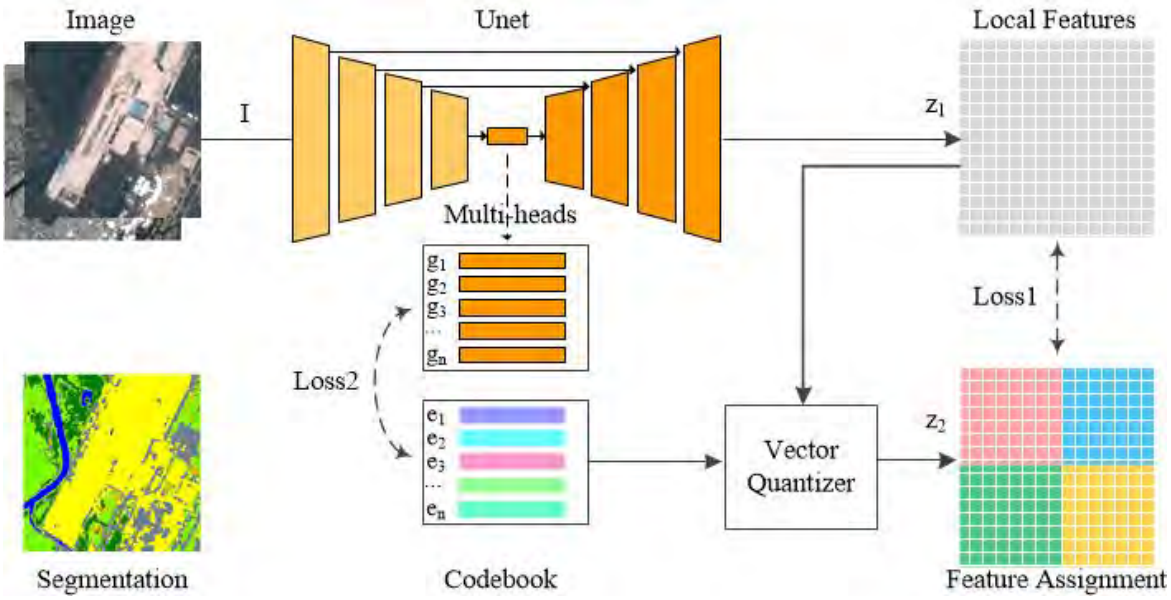

Fig. 5.5 Overview of the proposed unsupervised segmentation approach. The framework is a pseudo-Siamese architecture, where one branch is a ResUnet and the other branch is the gumbel-softmax vector quantizer. During the training, an input image is fed into ResUnet to get pixel-wise representation. We then reconstruct this feature representation from limited vectors using vector quantization. During the inference, the segmentation is obtained using hard selection in the gumbel-softmax operation.

To address this limitation, we further propose a new unsupervised land-cover segmentation approach, using contrastive learning and vector quantization, which can obtain the features for discriminating different land-cover categories based on the information provided by SAR and optical images. The evaluation of the proposed land-cover segmentation approach compared with the Infoseg approach was performed on a subset of DFC2020 dataset [163] including six land-cover classes.

5.2.2 Methodology

This section presents the methodology of the proposed unsupervised land-cover segmentation based on contrastive learning and vector quantization, where the network is learnt to reconstruct images using learned class vectors.

Network Architecture

We propose a pseudo-Siamese network with two branches Fig.5.5. One branch is a ResUnet and the other branch is a gumbel-softmax quantizer. For the ResUnet, we adopt the ResNet-18 as the encoder but without using the fourth layer. The decoder part has the same blocks

as the encoder and each block consists of a convolutional layer, a batch normalization, a ReLU activation function, and an upsampling operation. All padding type in the ResUnet is changed to the same padding type. An MLP projection followed by the last block is used to reconstruct the learned representations. It consists of the 1×1 Conv of 256 channels and a ReLU, and then a 1×1 Conv with 128 channels for each pixel. In addition, several classification heads are considered followed by the bottleneck block to encode the global features.

In the other branch, we introduce a gumbel-softmax vector quantizer to reconstruct the local feature representation from the limited vectors in the codebook. The quantization module takes the pixel-wise representation from the ResUnet and maps it into a new representation. This is done by selecting one entry from a fixed codebook using pixel-wise scaled softmax. However, this hard selecting process results in suboptimal performance. To alleviate this problem, we compute a soft feature assignment for each pixel using vectors and their class probabilities. The class probabilities are obtained by performing a pixel-wise scaled softmax between the fixed codebook and local features. In this way, we can obtain an augmented feature representation. During the inference, we chose a hard selection to assign each pixel a class, i.e. the proxy label of the corresponding vector. To make the vectors represent the image-level features, we force the multi-global features close to vectors and make the vectors distributed in feature space uniformly. It is noted that the number of global features and the number of vectors are the same as the number of classes. In the post-processing phase, the spectral indices are used to determine the landscape of each class and remove the isolated classified pixels.

Loss Functions

The proposed approach consists of two types of contrastive loss individually based on images and superpixels. For superpixel level loss, each positive feature pair (z_1^i, z_2^i) is sampled from the same location i, whereas each negative sample z_2^j is taken from another location. This loss L_{spix} can be written as:

$$L_{\text{spix}} = -\mathop{\mathbb{E}}_{S}\left[\log \frac{h_\theta(z_1^i, z_2^i)}{\sum_{j=1}^{N} h_\theta(z_1^i, z_2^j)}\right] \tag{5.4}$$

where $h_\theta(\cdot)$ is a similarity function (i.e., cosine similarity), (z_1^i, z_2^i) is the normalized latent representation of superpixel i, $(z_1^i, z_2^j | j \geq i)$ is the normalized latent representation of negative pairs and $S = \{z_1^1, z_2^1, z_2^2, \cdots, z_2^{N-1}\}$ is a set that contains $N-1$ negative samples and one positive sample.

Together with the superpixel level contrastive loss, a global level contrastive loss between global features g and vectors e follows the alignment and uniformity terms. The alignment loss L_{align} is strightforwardly defined with the cosine distance between positive pairs:

$$L_{\text{align}}(g,e) \triangleq \underset{(g,e)\sim p_{\text{pos}}}{\mathbb{E}} \left[\|g-e\|_2^{\alpha}\right], \quad \alpha > 0 \tag{5.5}$$

where $\|\cdot\|_2$ is l_2-norm, g and e are copied as the same numbers. This is equivalent to the mean squared error of l_2-normlized vectors. The uniformity loss L_{uniform} is defined as the logarithm of the average pairwise Gaussian potential:

$$L_{\text{uniform}}(e) \triangleq \underset{(e_i,e_j)\sim p_{\text{neg}}}{\mathbb{E}} \left[e^{-t\|e_i-e_j\|_2^2}\right], \quad t > 0 \tag{5.6}$$

This term decorrelates the different vectors in the codebook and prevents them from obtaining the same information. The overall loss function L is a sum of a superpixel, alignment and uniformity terms:

$$L = L_{\text{spix}}(z_1,z_2) + L_{\text{align}}(g,e) + L_{\text{uniform}}(e) \tag{5.7}$$

5.2.3 Experimental Results

Description of Dataset

We developed our experiments on a subset of DFC2020 dataset [163]. This subset consists of 2000 triple samples, which are SAR-optical image pairs and more accurate semi-manually derived high resolution (10 m) land-cover maps. SAR images with dual-polarized (VV and VH) components were acquired by the Sentinel-1 satellite. The optical images with 12 bands were taken by the multi-spectral sensor of the Sentinel-2 satellite. In this subset, the cropland, wetland, and grassland are reclassified as grassland; shrubland and barren are reclassified as bared land. Finally, only six classes (i.e., forest, grassland, urban, bare land, water, and sparse vegetation) are included in this data set according to the land surface properties.

Table 5.5 Class-wise and overall accuracies of different approaches achieved on the subset of DFC2020.

Meth.	For.	Spa.	Gra.	Bui.	Bar.	Wat.	AA	mIoU
Info.	48.3	38.0	72.5	61.1	0.80	82.2	50.5	0.33
fInfo.	49.9	41.4	75.0	61.2	0.00	70.8	50.0	0.32
Prop.	88.0	35.5	57.1	78.9	60.2	97.4	69.5	0.44
fProp.	88.2	36.1	58.0	81.8	57.3	97.5	69.8	0.45

Results

Empirically, we found that the six and seven class settings in the network training lead to the best performance. The final result combines the two settings with six land-cover classes. The evaluation is performed on the results after the post-processing using spectral indices. We also provide a fine-tuning evaluation on initial results due to the missing data in the post-processing. Table 5.5 presents a quantitative evaluation of the accuracy on each class obtained by the InfoSeg (Info.) and the proposed approach (Prop.) as well as their fine-tuning results (fInfo. and fProp.). As one can see, the proposed approach achieves an AA of 69.5% sharply outperforming InfoSeg, which obtains an AA of 50.5%. The fine-tuning result of the proposed approach obtains a higher AA and mIoU, with an improvement of 0.3 % and 0.01 with respect to the initial results. However, the fine-tuning result of InfoSeg is worse than the initial result. The possible reason for this is that the initial results of InfoSeg contain too many misclassifications. By analyzing performance on individual classes, forest, water and built-up achieve higher accuracy. In contrast, the accuracy of sparse vegetation is below 50%. This is due to the fact that there is no clear boundary among sparse vegetation, barren and grassland.

Apart from quantitative assessment, we also made a visual comparison of the results. As shown in Fig. 5.6, the proposed approach classifies various classes in a more accurate way, especially in small areas. There is an interesting trend visible from the results: InfoSeg only captures the spatial patterns, whereas the proposed approach can identify the clear boundaries of different land-cover classes. Moreover, InfoSeg showed higher difficulty to separate forest, grassland, and sparse vegetation accurately. Finally, the fine-tuning results of the proposed approach lead to a further performance improvement, while it also induces the loss of some details.

5.2.4 Conclusion

In this section, we have investigated the unsupervised land-cover segmentation based on the SAR-optical fusion framework and vector quantization. The core of the presented approach is to minimize the contrastive loss between the local features output from ResUnet and the reconstructed features from limited vectors. The land-cover maps can be obtained by assigning each pixel with its proxy label of the most contributed vector. This approach is assessed quantitatively and qualitatively on the selected subset of DFC2020 and achieves an average accuracy of 68% considering six land-cover classes. Experimental results show that the proposed approach can learn semantically meaningful representation and discriminate different land-cover categories.

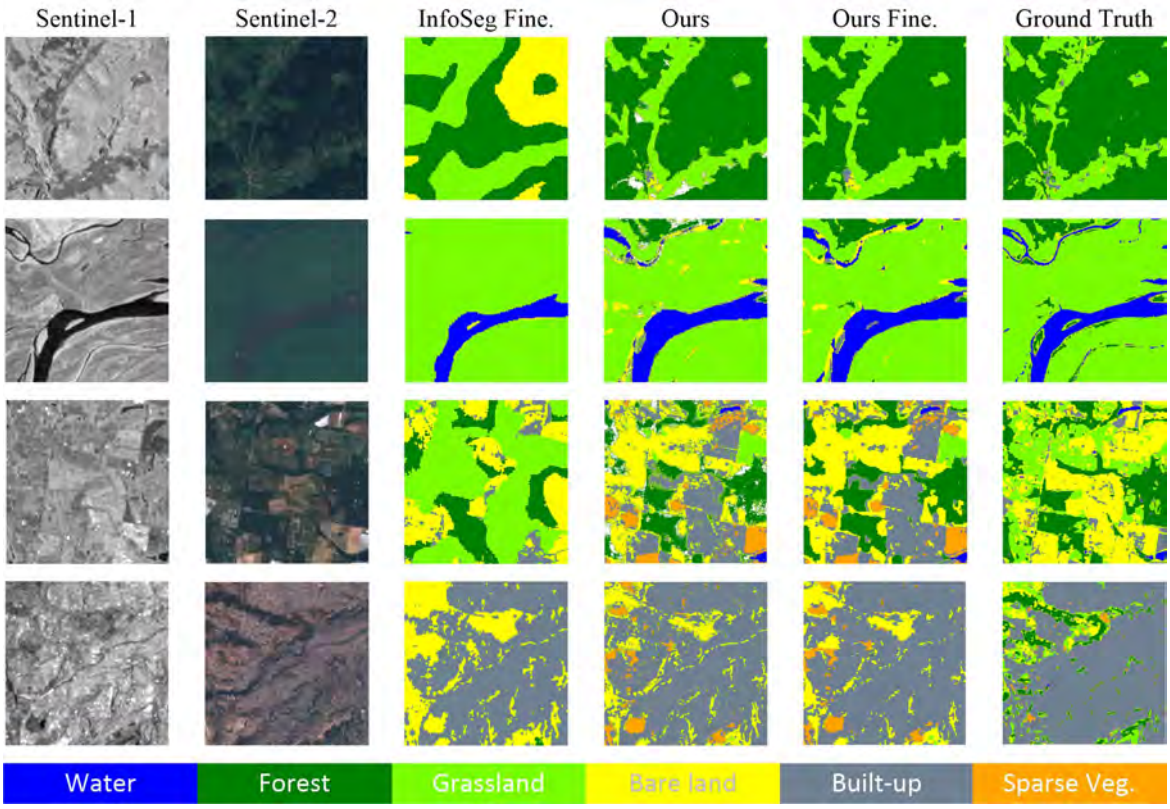

Fig. 5.6 Unsupervised land-cover maps obtained by InfoSeg and the proposed approach as well as their fine-tuning results.

5.3 Conclusion

In this chapter, we have investigated the self-supervised SAR-optical data fusion approach under three strategies: early fusion (PixEF), intermediate fusion (pixIF), and late fusion (PixLF). The results show that early fusion and intermediate fusion are better than late fusion. This is because the late fusion strategy discards the modal-specific task-relevant information but only keeps the shared information between the two modalities. This once again reminds us that contrastive learning between different modalities is not the best strategy to fuse multimodality information.

After that, we explored how to do unsupervised LULC mapping task in the self-supervised learning framework. We found that the complementary information on different land-cover provided by SAR and optical images is the key to accurate performance. Maximizing the mutual information between the global vectors and local information can segment images into semantic parts. On the basis of the freely accessed Sentinel-1/-2 data, the proposed approach demonstrates a promising potential for automatic large-scale land-cover mapping. In addition, the proposed approach can also be used to fuse other raster data. However, it

has also some limitations in the classes assignment. The considered training strategy just gives six classes and focuses on backscatter values of Sentinel-1 images while ignoring the polarization information. Accordingly, as future development, we plan to explore the possibility of including more specific classes in the presented approach for the land-cover mapping task.

Chapter 6

Incomplete Multimodal Learning for Remote Sensing Data Fusion

To address the limitation of the severe degradation with modal-incomplete inputs, in this chapter, we propose an approach that introduces a novel model for incomplete multimodal learning in the context of remote sensing data fusion. This approach can be used in both supervised and self-supervised pretraining paradigms and leverages the additional learned fusion tokens in combination with Bi-LSTM attention and masked self-attention mechanisms to collect multimodal signals. The proposed approach employs reconstruction and contrastive loss to facilitate fusion in pre-training, while allowing for random modality combinations as inputs in network training. Our approach delivers state-of-the-art performance on two multimodal datasets for tasks such as building instance / semantic segmentation and land-cover mapping tasks when dealing with incomplete inputs during inference.

6.1 Introduction

Remote sensing becomes more and more important in various Earth Observation (EO) tasks. With the increasing availability of multimodal RS data, researchers now can develop more diverse downstream applications. Despite the abundance of multimodal remote sensing data, each modality captures only certain specific properties and, therefore, cannot thoroughly describe the observed scenes. Thus the use of single-mode data results in limitations in many applications. Multimodal RS data fusion addresses these limitations [56]. For instance, synthetic aperture radar (SAR) provides physical structure information, while LiDAR collects both structure and depth information [125]. Meanwhile, multispectral (MS) and hyperspectral (HS) scenarios measure radiation reflectance across different wavelengths

of the electromagnetic spectrum. By integrating the complementary information confined in multimodal data, it is possible to improve the accuracy and reliability in many data analysis tasks, such as change detection [28] and land-cover mapping [29]. To integrate the complementary information provided by different sensors and remote sensing products (e.g., Land Cover Land Use Maps), traditional methods [42] exploit handcrafted features based on domain-specific knowledge and fusion strategies that often are not able to capture all the information present in the data.

Thanks to the growth of artificial intelligence, deep learning shows great potential in modelling the complex relationships between different modality data and is widely used in remote sensing data fusion tasks. There are three main multimodal RS data fusion applications, SAR-optical [138, 2, 2, 77, 89] and LiDAR-optical [125, 141, 169, 45] as well as image-map [84, 161], where the deep CNNs and Transformer networks are widely used. Nevertheless, deep Convolutional Neural Networks (CNNs) methods assume that all modalities are available during training and inference, which can be a limiting factor in practical applications, as data collection processes may miss some data sources for some instances. In such cases, existing multimodal data fusion methods may fail to deal with incomplete modalities, leading to severe degradation in performance. The approach used in this situation is called incomplete multimodal learning and aims at learning methods that perform inference which is robust to any subset of available modalities. A simple strategy for incomplete multimodal learning using CNNs is to synthesize the missing modalities using generative models. For instance, Generative Adversarial Networks (GANs) can effectively overcome the problems arising from missing or incomplete modalities in building footprint segmentation [13]. Another set of methods explores knowledge distillation from complete to incomplete modalities. In this approach, Kampffmeyer et al. [83] proposed to use an additional network, the hallucination network, for mitigating missing data modalities in the testing of urban land cover classification tasks. The network takes a modality as input that is assumed to be available during both training and testing, trying to learn a mapping function from this modality to the missing one.

Although promising results are obtained, such methods have to train and deploy a specific model for each subset of missing modalities, which is complicated and often unreliable in downstream tasks. Moreover, all these methods require complete modalities during the training process. Recent incomplete multimodal learning methods focus on learning a unified model, instead of a bunch of distilled networks, for downstream tasks. In this context, the modality-invariant fusion embedding across different modalities may contribute to more robust performance, especially when one or more modalities are missing. As a competitive multimodal data fusion model, Transformer does not need to access all modalities in the

network training and inference as its flexibility and sequence modelling strategy, which can be effective in both scenarios: with and without missing modalities. Current works exploited Transformers for multimodal RS data fusion in a complete fusion scenario, such as lidar and hyperspectral data fusion [134]. For incomplete multimodal data fusion, MBT [115] and Zorro [128] propose to fuse audio and video data using learnable tokens in the Transformer network. However, the definition of a dedicated Transformer for incomplete multimodal learning in remote sensing tasks has not been addressed yet and the existing multimodal RS data fusion methods do not allow missing data in the training process. Moreover, Ma et al. [107] point out that the vanilla Transformer tends to be overfitted on one modality data.

In addition, most multimodal data fusion methods are based on the supervised learning paradigm. Supervised approaches are task-specific and have limitations to be generalized to other tasks. Moreover, training on a large amount of multimodal data is cost expensive and collecting an adequate labeled data for each task is challenging for end-users. Thus, the research community usually relies on a few fine-tuning steps on a pre-trained model to adapt a network to a specific task. Pre-training without supervision has gained a lot of attention as it is more general and does not require labeled data. The self-supervised learning method for SAR-optical feature fusion [28] is an example of such an approach. However, this pre-training approach also needs to access all modalities during network training.

Hence, this paper proposes to exploit Transformer to build a unified model for incomplete multimodal learning for remote sensing tasks, which can be used in both the supervised and self-supervised pre-training paradigms. This is achieved by using additional learned fusion tokens for multimodal signal collection in the network. However, only using the additional learned fusion token cannot capture enough information from other modality tokens. In this context, we use a Bi-LSTM attention block to further distil different modality information to fusion tokens. Using this technique, the proposed approach can leverage MultiMAE and contrastive loss to build fusion across the different modalities in pre-training. Moreover, it can use a random modality combination training strategy in downstream task fine-tuning. This makes the learning and inference feasible also when incomplete modality data are given as input.

The three main contributions of this chapter consist in: (1) we propose to use Bi-LSTM and masked self-attention in multimodal Transformer to build additional fusion tokens across different modalities, which enable both contrastive and generative self-supervised pre-training for incomplete multimodal inputs; (2) based on the proposed approaches, we use the random modality combination training strategy in downstream tasks, which ensures task performance with incomplete inputs on inference. (3) we benchmark our approach on two datasets: the public DFC2023 track2 and the created quadruplet dataset, obtaining results that show the

proposed approach can be pre-trained on a large-scale remote sensing multimodal dataset in a self-supervised manner. The proposed approach achieves state-of-the-art performance when compared with the vanilla multimodal Transformer [115] on RS.

6.2 Related Work

6.2.1 Masked Autoencoder

The MAE (masked autoencoder) [68] is a novel self-supervised learning algorithm that demonstrates state-of-the-art performance on various vision benchmarks. Instead of relying on a contrastive objective, the MAE utilizes a pretext task that involves reconstructing masked patches of the input.

The MAE network follows an asymmetric encoding and decoding scheme. Suppose the input image is a tensor of dimensions $I \in R^{C\times H\times W}$, where H,W are the height and width of the image, respectively, and C is the number of channels. The image is initially divided into non-overlapping patches $S \in R^{L\times P^2C}$, where P is the height and width of the patch, and $L = (H/P)\times(W/P)$ is the number of patches. These patches are then transformed into a sequence of embedded patch tokens $S' \in R^{L\times D}$, using a patch embedding function $f_p : R^{P^2C} \to R^D$. A fraction p_m of the sequence tokens is randomly masked, and the remaining visible tokens are fed into an encoder, which is a Vision Transformer (ViT). Due to the lack of positional information, additional positional embeddings are then added to patch embeddings to capture the spatial location of the patch in the image. The decoder is composed of multiple transformer blocks that are trained for all tokens, where the masked tokens are replaced as the initialized learnable tokens. The decoder produces a reconstructed image, which is compared to the original image using mean-squared error (MSE) loss, computed only on masked patches. Positional encoding allows the transformer to encode positional information. In MAE the positional encoding is:

$$\text{Encode}(k,2i) = \sin\frac{k}{\Omega^{\frac{2i}{d}}}, \text{Encode}(k,2i+1) = \cos\frac{k}{\Omega^{\frac{2i}{d}}} \tag{6.1}$$

Here, k is the position, i is the index of feature dimension in the encoding, d is the number of possible positions, and Ω is a large constant. In MAE, the position is defined as the index of the patch along the x or y axis. Therefore, k ranges from 0 to H/P or W/P. This encoding provides two unique dimensions, one for x and one for y coordinates, which are concatenated for the final encoding representation.

The Multimodal Masked Autoencoder (MultiMAE) [8] is based on a standard single-modal ViT and the modality-specific encoders. The encoder is equipped with 2-D sine-cosine positional embeddings following the linear projection. MultiMAE does not make use of modality-specific embeddings, as the bias term in each linear projection is sufficient. MultiMAE employs a separate decoder for each task that is responsible for reconstructing the masked-out tokens from the visible tokens. The input to each decoder is a full set of visible tokens from all different modalities, including the learnable modality embeddings with 2-D sine-cosine positional embeddings. The input is then followed by MLPs and Transformer blocks. Only the masked tokens are considered in the loss calculation. The mask sampling strategy employed in MultiMAE plays a crucial role in achieving predictive coding across different modalities. This sampling strategy ensures that most modalities are represented to similar degrees. MultiMAE adopts a symmetric Dirichlet distribution to select the proportion of tokens per modality λ ($\lambda_i \sim Dir(\alpha)$), where $\sum \lambda_i = 1, \lambda > 0$. The concentration parameter $\alpha > 0$ controls the sampling. For simplicity and better representation parameter $\alpha = 1$ in MultiMAE.

6.2.2 Multimodal Transformer

The self-attention blocks of Transformers build a natural bridge among multimodal signals in a unified architecture. Differently from the CNNs that use one network for each modality, the Transformer only use the same main architecture for all modalities with a modal-specific projector. Transformers integrate input tokens from all modalities into a single representation, while CNNs fuse features of each modality through concatenation or tensor fusion. However, such explicit integration requires the presence of all modalities during training, which undermines the pipeline in case of a missing modality. In contrast, Transformers use self-attention to embed a holistic multimodal representation and handle the absence of modalities by applying a mask on the attention matrix. Thus, multimodal Transformers are more adaptable to deal with modal-incomplete inputs. In addition, an easy-to-train model is vital for multimodal learning. The training load of a conventional multimodal backbone grows as the number of modalities increases since the backbone usually consists of modality-specific sub-models that need to be trained independently for each modality. Instead, Transformers process modalities altogether in a single model, significantly reducing the training load.

However, Transformer models exhibit significant deterioration in performance with model-incomplete inputs, especially in the context of multimodal inference where Transformer models tend to overfit the dominating modalities. To overcome this challenge, MBT [115] builds a multimodal architecture for video and audio, by using an additional fusion token to force information among different modalities to pass through by using cross-attention.

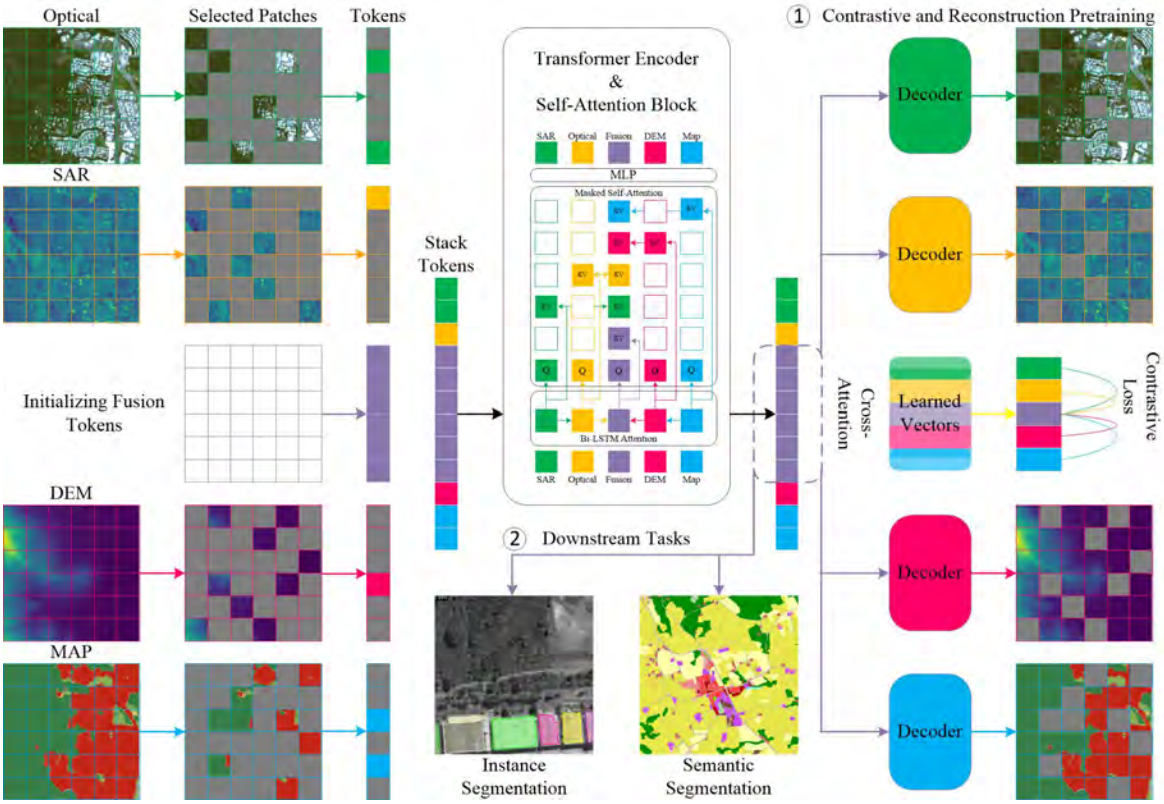

Fig. 6.1 Overview of the proposed framework. The inputs to our model are optical images, SAR images, DEM and Maps. Each of those inputs is patched using a 2D convolution and projected to feature vectors. All inputs are concatenated with a set of learnable fusion tokens and added to the position embedding. Next, we process these inputs through the Transformer Encoder, where the Bi-LSTM Attention and the masked Self-Attention strategy are applied. (1) In pre-training, task-specific decoders reconstruct the masked patches by using the output fusion tokens. Meanwhile, the global vectors of each modality and fusion tokens are output using cross-attention, which allows using contrastive loss between fusion tokens and each modality. (2) In the supervised training, the proposed framework can be trained on a specific downstream task by using a random modality combination strategy.

However, the representation of each modality can also access to the others in MBT, which means they are not independent. In [128], a modality-aware masking mechanism is used in all attention operations to isolate the allocation of latent representations of individual modalities, which leads to a resultant representation that is partially unimodal (i.e., part of the representation attends to a single modality) and partially multimodal (i.e., part of the representation attends to all modalities), thereby allowing for the use of contrastive learning.

6.3 Methodology

In this section, we describe the proposed incomplete multi-modal fusion architecture with additional learned fusion tokens, Bi-LSTM and masked self-attention. This is done using as an illustration case, an optical-SAR-DEM-MAP data fusion example. Then, we introduce the details of both the pre-training using MultiMAE and contrastive loss, as well as those of training using random modality combination on downstream tasks (see Fig. 6.1).

6.3.1 Network Architecture

The main architecture of the proposed approach is a ViT with modality-specific patch projection layers for each input modality. In detail, patches of each modality are projected to tokens using a specific linear projection for each modality. In this work, we use a 2D convolution to extract 16×16 patches and project them to the input dimension D. Next, position embeddings are added to the projected vectors so that the model is able to localize and distinguish each embedded patch. In addition to the multimodal input data, the learnable fusion tokens are introduced as one of the inputs. Differently to the bottleneck fusion tokens in MBT [115] and Zorro [128], we use the spatial tokens for dense downstream tasks, which have the same number of tokens of full input patches. In order to get local features, we add 2D sine-cosine positional embeddings on the spatial fusion tokens and use Bi-LSTM to aggregate all modality information to fusion tokens. Then the projected patches together with the learnable tokens are concatenated into a sequence of tokens and given as input to the same Transformer encoder with masked attention. Since all our input data have a 2D structure, we add 2D sine-cosine positional embeddings after linear projection. Following the setting of MultiMAE, we do not consider any modality-specific positional embedding.

Bi-LSTM Attention. We use a Bi-LSTM with an attention mechanism to integrate different modality input embeddings into learned fusion tokens for improving the feature learning ability. Consider one direction of the LSTM network: let $\overrightarrow{h_i}$ be the output of the LSTM for the multimodal inputs (in our example, Optical, SAR, DEM and MAP) and the learned fusion tokens. Bi-LSTM performs forward training and backward training separately for each training sequence and then combines the results of forward training and backward training together as the output of each modality, which is denoted as $h_i = [\overrightarrow{h_i}, \overleftarrow{h_i}]$. We use h_f (fusion tokens) to represent all multimodal inputs h_o (optical tokens), h_s (SAR tokens), h_d (DEM tokens), h_m (map tokens) and measure the importance of each modality through the similarity with a learning parameter u. Then we get a normalized importance weight β_i

through a softmax function.

$$\beta_i = \frac{exp(u^\top \tanh\left(W\left[h_f; h_i\right] + b\right))}{\sum_{i=1}^{t-1} exp(u^\top \tanh\left(W\left[h_f; h_i\right] + b\right))} \tag{6.2}$$

where u and h have the same dimension as the cell state of the LSTM, and [] is the concatenate operation. W is a weight matrix and b is a bias vector of the MLP. The final new fusion token is thus:

$$a = \sum_{i=1}^{t-1} \beta_i \cdot h_i \tag{6.3}$$

Masked Self-Attention. Masked self-attention is the key block of multimodal Transformer in contrastive pre-training. Using masked attention, we force part of the representation to attend only to itself, while other parts can attend to the whole representation. In the considered illustration case, the main goal of this approach is to split the representation into five parts: a part which only focuses on Optical tokens, a part which focuses on SAR tokens, a part which focuses on DEM tokens, a part which focuses on MAP tokens, and the fusion tokens which consider the whole representation. In this architecture, the self-attention in each layer and the cross-attention in the last layer both used this masking strategy. Here we introduce the masking binary tensor m that specifies which vectors can access each other. Entries of the masking matrix are $m_{i,j} = 1$ if information can flow from latent j to latent i. Versus, we set $m_{i,j} = 0$. The mask is applied to the standard attention output operation, which performs on keys k, values v and queries q, can be expressed as:

$$o_i = \sum_j \frac{m_{ij} \exp\left(\frac{q_i^\top k_j}{\sqrt{d_k}}\right)}{\sum_{\{j', m_{ij'}=1\}} \exp\left(\frac{q_i^\top k_{j'}}{\sqrt{d_k}}\right)} \cdot v_j \tag{6.4}$$

where the d_k is the dimension of k vector. In order to keep the performance of a single modality when other modalities are absent, the modality-specific representation can not access the fusion representation or other modalities. This explicitly prevents the information of the fusion stream from leaking into the unimodal representation. This is the key to preserving pure streams that correspond to single modalities. Thus, after applying this mask, the specific output o_s, o_o, o_d, o_m only contains information coming from the SAR, optical, DEM, MAP inputs, respectively. The fusion output o_f access all outputs in the model.

Reconstruction Pre-training In order to train our network in an MAE way, we use a separate decoder for each generation task. The input to each decoder is the spatial tokens output from the cross attention. Following the same setting of MAE, we use shallow decoders

with a low dimensionality, which consists of two Transformer blocks. MultiMAE mask across different modalities ensures the model develops predictive coding across different modalities besides different spatial patches. According to MultiMAE, we set a constant number of visible tokens at 256, which corresponds to 1/4 of all tokens in our experiment (learned fusion tokens and four modality inputs with 256×256 image size and 16×16 patch size). The proportion of tokens per modality λ are sampled from a symmetric Dirichlet distribution $(\lambda_{Optical}, \lambda_{SAR}, \lambda_{DEM}, \lambda_{MAP}) \sim Dir(\alpha)$, where $\lambda_{Optical} + \lambda_{SAR} + \lambda_{DEM} + \lambda_{MAP} = 1, \lambda \geq 0$. For simplicity and better representation of any possible sampled task, we use a concentration parameter $\alpha = 1$. As shown in Fig. 6.1, we adopt reconstruction loss (l_1 distance Mean Squared Error) to recover the pixel color and height information following MultiMAE and using cross-entropy loss (l_{ce}) on land-cover map reconstruction:

$$\begin{aligned} L_{DEM} &= l_1(Dec(o_f), DEM) \\ L_{SAR\_RGB} &= l_2(Dec(o_f), SAR) + l_2(Dec(o_f), RGB) \\ L_{MAP} &= l_{ce}(Dec(o_f), MAP) \end{aligned} \tag{6.5}$$

Contrastive Pretraining. We also add the class token for each modality input data and an additional global class token for the learned fusion tokens. To integrate information from the encoded visible tokens of other modalities, we add a single cross-attention layer using these tokens as queries that cross-attend to the encoded tokens of the last self-attention layer. We utilize the standard cross-attention operation and produce five different outputs: the vector outputs for each modality and a fusion vector output. This design opens the possibility to use contrastive learning among different modalities and fusion tokens. For a better multimodality alignment, we propose to use extra contrastive loss between each modality-specific output and the fusion vector. Specifically, given the optical vector output $z_o = g_o(o_o)$ and the fusion output $z_f = g_f(o_f)$, where g_o and g_f are the linear projection for each modality, the contrastive loss can be formulated as:

$$L_c(z_o, z_f) = -\mathop{\mathbb{E}}_{S}\left[\log \frac{e^{sim(z_o^i, z_f^i)/\tau}}{\sum_{j=1}^{N} e^{sim(z_o^i, z_f^j)/\tau}}\right] \tag{6.6}$$

where *sim* is a similarity function (i.e., cosine similarity), S is a set that contains $N-1$ negative samples and one positive sample. This equation introduces the loss for RGB-FUSION contrastive training. In order to contrast the output of all outputs, we define a contrastive loss between unimodal representations and fusion representations. Thus, we can

write the full loss as:

$$\begin{aligned} L = & L_{DEM} + L_{SAR\_RGB} + \lambda_2 * (L_c(z_f, z_o) \\ & + L_c(z_f, z_s) + L_c(z_f, z_d) + L_c(z_f, z_m)) \end{aligned} \tag{6.7}$$

Random Modalities Combination. Besides the network design, the training strategy is vital to the performance of modal-incomplete inputs. The research in [107] finds that the Transformer models tend to overfit the dominating modalities in a task. To improve the robustness of the proposed approach against modal-incomplete data, we propose to leverage a random modality combination training strategy. Thanks to the proposed approach, we can randomly choose the different modality combinations or unimodal data in pre-training or supervised training on downstream tasks. The proposed approach fuses all modalities using additional learned tokens, thus it greatly reduces the effects of modal-incomplete inputs.

6.4 Experiments

In this section, we evaluate the proposed approach in multiple settings. We first introduce the multimodal dataset used in this work. Then, we present the details of both pre-training and training on downstream tasks, as well as the evaluation procedures. Finally, we ablate the performance of the complete and the incomplete multimodal inputs to show the proposed approach's flexibility.

6.4.1 Experimental Details

In order to showcase the proposed approach across the different modalities, we train the proposed approach in both a completely supervised paradigm and a fine-tuning paradigm with pre-trained weights. Many works have pointed out that the pre-training of a big model on multimodal data can be beneficial on downstream tasks [145]. The pre-trained model can be then used for arbitrary downstream tasks with the fine-tuning of the task-specific decoder. Hence we can train a giant model on a large multimodal data set with as many modalities as possible. The pre-trained model can strengthen the ability to extract features that are only trained on a few or single modality data. In this section, we provide the details of the self-supervised pre-training and the supervised training on downstream tasks as well as the multimodal datasets.

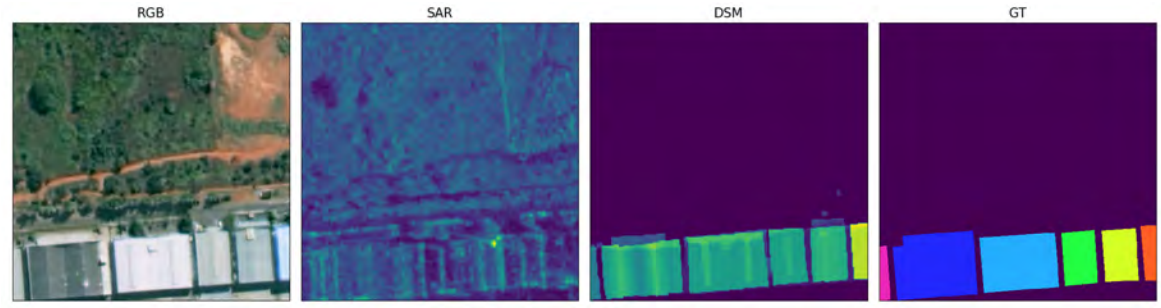

Fig. 6.2 Example of DFC2023 track2 data sample containing RGB and SAR images, DSM and ground truth.

Description of Datasets

We train and test the performance of the proposed approach on two multimodal datasets for two downstream tasks, namely building instance / semantic segmentation and LULC mapping.

DFC2023 track2 - Building instance / semantic segmentation. The first data set is the track 2 dataset of DFC2023, which comprises a combination of RGB images, SAR images, and Digital Surface Model (DSM) data. While the objective of the original task is building height estimation, this study simplifies it as building instance / semantic segmentation. The dataset consists of images obtained from GaoJing-1, GaoFen-2 and GaoFen-3 satellites, with spatial resolutions of 0.5 m, 0.8 m and 1 m, respectively. Normalized Digital Surface Models (nDSMs) are used as a reference in Track2 and are created from stereo images captured by GaoFen-7 and WorldView-1 and -2 with approximately 2 m ground sampling distance (GSD). The dataset was collected from seventeen cities across six continents and hence is highly diverse in terms of landforms, building types and architecture. The labels of building instance segmentation adopt the MS COCO format and are provided in a JSON file. A sample of the labels is shown in Fig. 6.2 for illustration.

Quadruplet Dataset - Land-Use Land-Cover (LULC) mapping The second dataset considers diverse data sources obtained from Google Earth Engine (GEE) platform, encompassing Sentinel-1, Sentinel-2, LiDAR DEMs and Dynamic World LULC maps, as shown in Fig. 6.3 and Fig. 6.4. The dataset comprises 37 regions across various landscapes and LULC classes in France and Australia. The Sentinel-1 mission provides data from a dual-polarization C-band SAR instrument and produces the calibrated and ortho-corrected S1 GRD products. We download the data from the COPERNICUS/S1_GRD category on GEE, resampling it into 10 m resolution and using dual-band VV+VH. Similarly, we download the Sentinel-2 data from the COPERNICUS/S2_SR_HARMONIZED category, which provides multispectral imaging with 13 spectral bands suitable for large-scale LULC mapping. We

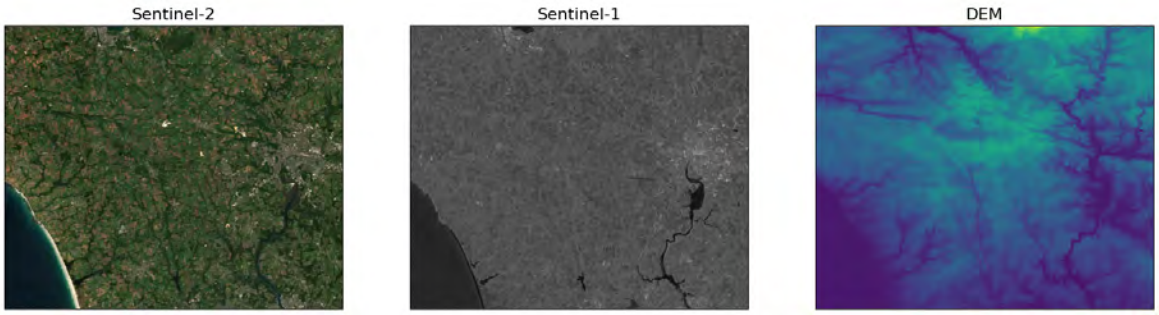

Fig. 6.3 Example of Quadruplets Data Set containing Sentinel1, Sentinel-2 and DEM data.

resample the Sentinel-2 data into 10 m resolution, and use the RGBN bands in this work. Two types of LiDAR DEMs are provided in this research. In France, we utilize the RGE ALTI dataset, which is a digital elevation model created using airborne lidar, with a pixel size of 1 m. We resample this dataset to 10 meters, with a vertical accuracy that ranges from 0.2 m to 0.5 m and an average accuracy of 7 m in steep slope areas. In Australia, we use a digital elevation model 5 m grid derived from 236 individual LiDAR surveys conducted between 2001 and 2015. We compile and resample the available 5 m resolution LiDAR-derived DEMs using a neighbourhood-mean method to create 10 m resolution datasets for each survey area, which we used in this work. The Dynamic World MAP (DNW) dataset comprises globally consistent, 10 m resolution, near real-time land-use and land cover predictions derived from Sentinel-2 imagery. It features ten bands that include estimated probabilities for each of the nine LULC classes (water, trees, grass, crops, shrub and scrub, flooded vegetation, built-up area, bare ground, and snow & ice). It also has a class "label" band indicating the class with the highest estimated probability, which makes it suitable for multi-temporal analysis and custom product creation. Lastly, we utilize the labeled class-reference from the UrbanAtlas 2018 database containing 27 LULC classes as the label of this dataset. The dataset provides integer rasters with index labels. We create raster maps with 10 m resolution that geographically match the Sentinel-1/-2 images using the open-data vector images freely available on the European Copernicus program website.

Downstream Tasks

We evaluate the proposed approach against state-of-the-art methods on two downstream tasks: building instance / semantic segmentation, and LULC mapping. In particular, the evaluation is performed on the supervised learning and the fine-tuning paradigms. For these two downstream tasks, we replace the pre-trained decoders with randomly initialized Mask2Former. In the following, we give an overview of the two tasks.

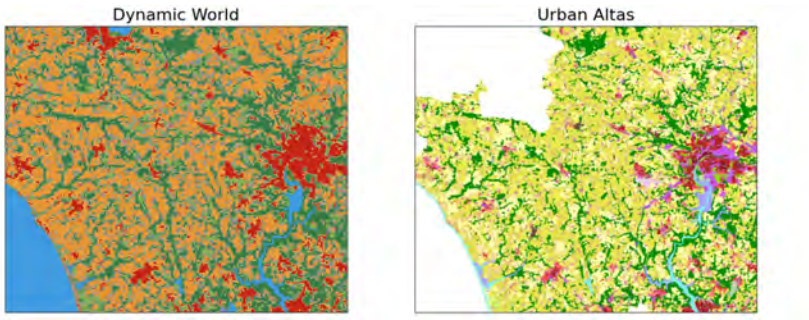

Fig. 6.4 Example of Dynamic World Map and European Urban Atlas data.

Building Instance / Semantic Segmentation: We follow the Mask2Former but replace the backbone with the proposed network. In the supervised experiments, we train the whole network from scratch using a random modality combination strategy. In the fine-tuning experiments, we consider two strategies, one is only to update the network on the pre-trained ViT-T backbones using a generative way, and the other is to update the whole network on the pre-trained ViT-T backbones using reconstruction and contrastive losses. We train our model on DFC2023 track2 train split and report the validation accuracy on the validation split. Along with the results of building instance segmentation, we also provide the binary building semantic segmentation results.

Land-Use Land-Cover Mapping: We still use the Mask2Former with the proposed backbone on the quadruplet dataset to generate LULC maps. However, we consider seven classes merged from the semantic hierarchy defined by UrbanAtlas. For that, we extract 7 semantic classes by taking the argmax of the prediction head. The same training strategy as that of the building instance segmentation is used in this task. We train our model on 10 (5340 samples) cities and report the validation accuracy on the other 2 (783 samples) cities.

Architectural Details

The proposed approach uses a ViT-T as the main structure and consists of 4 and 5 input adapters with a patch size of 16×16 pixels for the pre-training in the two different tasks. Differently from the standard MultiMAE, we add the learnable fusion tokens as input by using an additional input adapter to add 2D sine-cosine position encoding. The fusion tokens are as many as the number of all patched inputs.

After adding the position encodings, the fusion tokens with all modality inputs are given as input to a one-layer Bi-LSTM attention block. In self-attention, we use the masked algorithm to avoid the fusion information leak to a single modality. In order to get the global features of each modality and the fusion output, we use an additional cross-attention layer to

map the patch embeddings into the vector output. Then an auxiliary contrastive loss is added between each modality output vector and the fusion output vector.

For reconstruction learning, we follow the same setting of the MultiMAE decoder but without positional embeddings and cross-attention layer. The fusion tokens are projected into the decoder dimension by using a linear projection layer and then added to a learned modality embedding. After this, two Transformer blocks and a linear projector are used to project and reshape it to form an image or a map.

For the two downstream tasks, we adopt the same settings from Mask2Former. For the pixel decoder, we use 6 MSDeformAttn layers applied to feature maps with resolution 1/8, 1/16 and 1/32, and use a simple upsampling layer with lateral connection on the final 1/8 feature map to generate the feature map of resolution 1/4 as the per-pixel embedding. We use the Transformer decoder with 9 layers and 100 queries for instance segmentation, 9 queries for binary building semantic segmentation and 9 queries for LULC mapping. We use the binary cross-entropy loss and the dice loss for the mask loss. The final loss is a combination of mask loss and classification loss. For instance segmentation, we use the standard AP@50 (average precision with a fixed IoU of 0.5) metric. For semantic segmentation, we use the mIoU (mean Intersection-over-Union) metric.

Training Details

For pre-training, we train our model for 1600 epochs on 5700 triplet data on the DFC2023 track2 data set and 6123 quadruplet data on the quadruplet data set, individually. We use the AdamW optimizer with a base learning rate of 1e-4 and weight decay of 0.05. We warm up training for 40 epochs, starting from using cosine decay. We set the batch to 40 using a single Nvidia RTX 3090. All data are resized to 256×256. The number of non-masked tokens given to the encoder is set to 256 on the two data sets. For the second dataset, where we use the land-cover map as an additional modality input with 64-dimensional class embeddings.

For instance segmentation and semantic segmentation using Mask2Former, we use AdamW optimizer and the step learning rate schedule. We use an initial learning rate of 0.0001 and a weight decay of 0.05. A learning rate multiplier of 0.1 is applied to the backbone with the pre-training and not in the supervised learning. We decay the learning rate at 0.9 and 0.95 fractions of the total number of training steps by a factor of 10. We train our models for 50 epochs with a batch size of 10 in the semantic segmentation task and 300 epochs in the instance segmentation task.

6.4.2 Experimental Results

Multimodal Comparison

We evaluate the proposed approach with the two paradigms, one is supervised from scratch, and the other is fine-tuning with pre-trained weights. Considering no dedicated Transformer for incomplete multimodal remote sensing data fusion, we compare the proposed approach against a technique that uses origin self-attention and learned fusion tokens on the audio and video fusion task [115], termed MultiViT, on modal-complete and modal-incomplete inputs for building instance/semantic segmentation and LULC mapping tasks. The results reported in Tables 6.1 and 6.2 reveal that the proposed approach outperforms MultiViT in building instance/semantic segmentation tasks when evaluated with modal-complete inputs. However, for the LULC mapping task, the performance of the proposed approach and MultiViT are comparable. With regards to model-incomplete inputs, the proposed approach performs impressively well on all modality incomplete inputs and single modality inputs for both tasks due to the proposed attention block and random modality combination training strategy. For building instance/semantic segmentation, there is a visible dominance of RGB images over all other modalities, followed by DSM, while SAR images make the slightest contribution to the task, even causing noise. In this situation, MultiViT completely overfits on dominant modality inputs and fails on the task with single modality inputs when evaluated with model-incomplete inputs. Similarly, for LULC mapping, Sentinel-2 images along with the dynamic world map have a significant influence on the task, followed by Sentinel-1 and DEM images. The proposed approach achieves the best performance with a mIoU of

Table 6.1 Quantitative evaluations of proposed approach versus MultiViT with complete and incomplete multimodality inputs on the DFC2023 track2 dataset. Results are reported on AP@50 for instance segmentation and mIoU for semantic segmentation and consider the supervised result (sup.) and the fine-tuning result with the generative pre-trained weights (Fine. w/G) as well as the fine-tuning results with both the generative and contrastive pre-trained weights (Fine. w/G&C).

Multimodal Input	Sup. MultiViT		Sup. Propsed		Fine. w/ G.		Fine. w/ G. & C.	
	ins.	sem.	ins.	sem.	ins.	sem.	ins.	sem.
SAR, RGB, DSM	0.147	0.820	0.333	0.851	0.298	0.852	0.300	0.849
SAR, RGB	0.002	0.523	0.296	0.809	0.257	0.797	0.260	0.798
SAR, DSM	0.064	0.700	0.233	0.779	0.217	0.776	0.202	0.780
RGB, DSM	0.105	0.736	0.332	0.847	0.298	0.848	0.300	0.844
SAR	0.001	0.392	0.040	0.552	0.036	0.532	0.037	0.566
RGB	0.003	0.457	0.291	0.799	0.252	0.788	0.254	0.784
DSM	0.036	0.683	0.211	0.753	0.200	0.754	0.187	0.754

0.244 with modal-complete inputs, whereas MultiViT overfits on dynamic world maps, and performs slightly better when a dynamic world map is present but fails when it is not present in the inputs.

In the context of the fine-tuning paradigm, the proposed approach is assessed through two distinct pre-training methods: one that employs generative pre-training and another that combines generative and contrastive pre-training. The outcomes of the evaluation for both tasks are presented in Table 6.1 and Table 6.2. As one can see, two tasks show controversial results. Specifically, in the case of building instance/semantic segmentation tasks, the training-from-scratch model outperforms all other models. However, the model that leverages both generative and contrastive pre-training methods is closely ranked as the second-best. In contrast, for the land-cover mapping task, the fully finetuned model is the top-performing model among all the models listed in the tables, demonstrating the potential of pre-training in augmenting downstream LULC tasks.

For the single modality input, our goal is not to show state-of-the-art performance in this setting, as we are trying to solve the dramatic degradation of unimodal inference with a multimodal backbone. Here we show the ability of the proposed approach to produce meaningful unimodal outputs when fed with unimodal data. To do this, we only input one modality and neglect other modality inputs. As we can see on both datasets (Table 6.1 and Table 6.2), the MultiViT suffers significant degradation from missing of modalities and completely fails to work on the non-dominated modalities. In contrast, the proposed approach using the random modality combination strategy achieves high performance also when only one modality is available. This is due to the fact that in the proposed models, some capacity is allocated to each modality specifically and the model is able to produce unimodal outputs. Besides the quantitative analysis, we also provide a visual qualitative comparison. Fig. 6.2 and Fig. 6.3 show the results of building instance / semantic segmentation and LULC mapping, respectively. For building instance / semantic segmentation, similarly to Table 6.1, the proposed approach with supervised paradigm achieved the best performance followed by the results of fine-tuning. The MultiViT achieves the worst performance, especially with the modal-incomplete inputs. Our experimental results reveal that the SAR modality produced inferior results compared to other modalities. For the LULC mapping task, the fine-tuning with contrastive and generative pre-trained weights outperformed other approaches, while MultiViT exhibited reliable performance only with DNW input. For different modalities, we conclude that the Sentinel-1/2 images and DNW maps contributed equally as effective modalities, while the DEM input was determined to be a single-class predictor, indicating its inability to extract useful information.

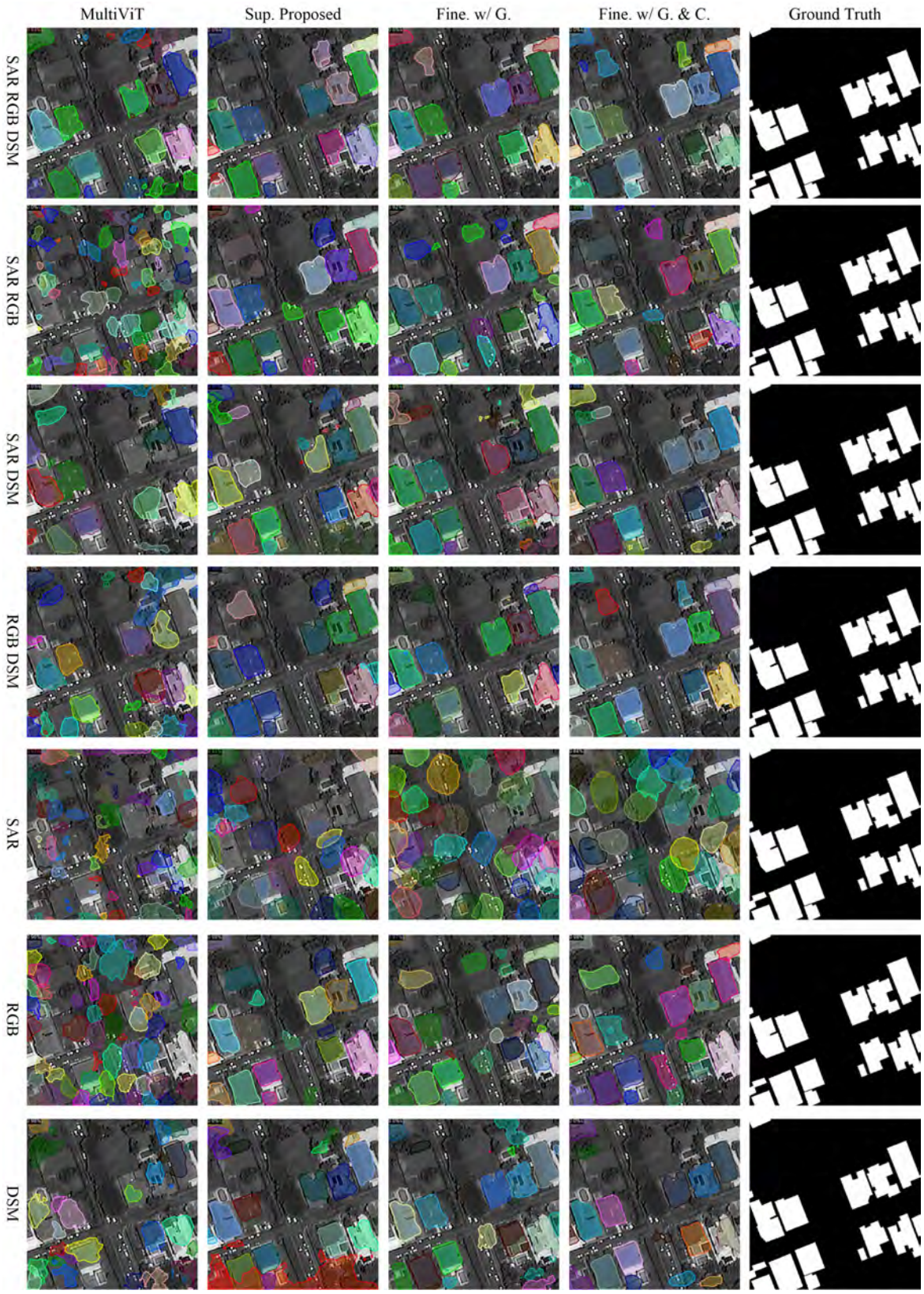

Fig. 6.5 Results of proposed approaches in the supervised and the two fine-tuning paradigms versus MultiViT on DFC2023 track2 dataset and consider the supervised result (sup.) and the fine-tuning result with the generative pre-trained weights (Fine. w/G) as well as the fine-tuning results with both the generative and contrastive pre-trained weights (Fine. w/G&C).

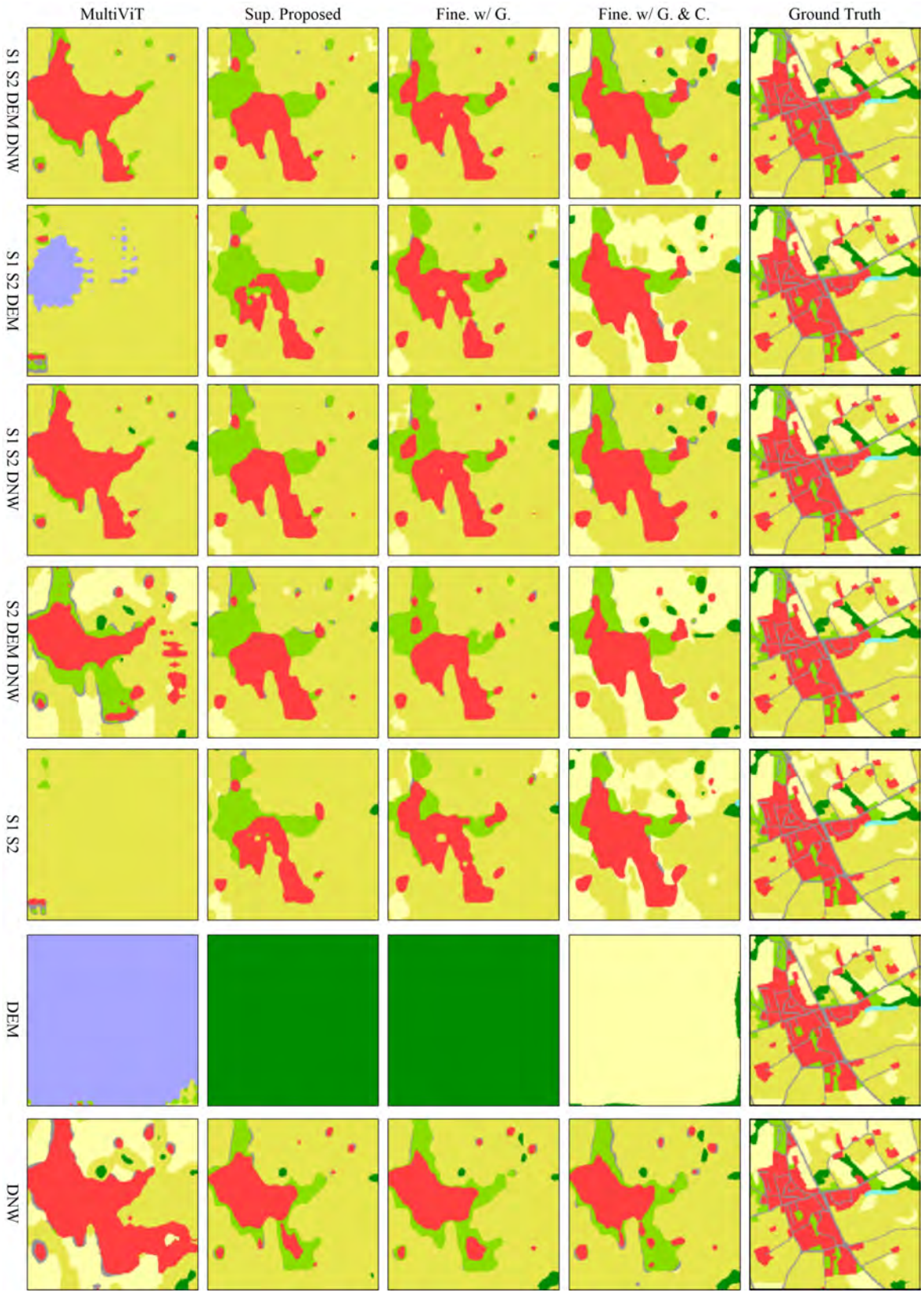

Fig. 6.6 Results of proposed approaches in the supervised and the two fine-tuning paradigms versus MultiViT on the quadruplets dataset and consider the supervised result (sup.) and the fine-tuning result with the generative pre-trained weights (Fine. w/G) as well as the fine-tuning results with both the generative and contrastive pre-trained weights (Fine. w/G&C).

Table 6.2 Quantitative evaluations of proposed approach versus MultiViT with complete and incomplete multimodality inputs on the quadruplets dataset. The results are reported in terms of mIoU values and consider the supervised result (sup.) and the fine-tuning result with the generative pre-trained weights (Fine. w/G) as well as the fine-tuning results with both the generative and contrastive pre-trained weights (Fine. w/G&C).

Multimodal Input	Sup. MultiViT	Sup. Proposed	Fine. w/ G.	Fine. w/ G. & C.
S1, S2, DEM, DNW	0.222	0.244	0.243	0.246
S1, S2, DEM	0.070	0.229	0.235	0.238
S1, S2, DNW	0.219	0.244	0.243	0.246
S1, DEM, DNW	0.219	0.235	0.235	0.235
S2, DEM, DNW	0.223	0.237	0.230	0.240
S1, S2	0.069	0.232	0.235	0.240
S1, DEM	0.074	0.208	0.221	0.216
S1, DNW	0.219	0.239	0.236	0.235
S2, DEM	0.054	0.210	0.204	0.227
S2, DNW	0.217	0.239	0.232	0.239
DEM, DNW	0.209	0.234	0.227	0.238
S1	0.079	0.208	0.222	0.214
S2	0.062	0.215	0.210	0.228
DEM	0.015	0.010	0.013	0.051
DNW	0.207	0.234	0.226	0.237

Albation

We now analyze the proposed approach through a series of ablation studies on both fine-tuning and supervised paradigms. To evaluate the generalizability of the proposed components, all ablations were performed on both tasks: the building instance / semantic segmentation and LULC mapping.

Random Modality Combination & Bi-LSTM Attention. We first validate the importance of the modality random combination training strategy on downstream tasks in a supervised paradigm. As shown in Tables 6.3 and 6.4, the model without the modality random combination training strategy experiences severe degradation with modal-incomplete inputs and even failed with a single modality on both tasks. In addition, we test the effect of the Bi-LSTM attention by removing it from the proposed network. The corresponding results show a significant drop in performance, indicating that the Bi-LSTM enables superior interaction of the fusion token with each modality and facilitates learning more discriminative features for downstream tasks.

Partial Fine-tuning and Non-masked Attention. In addition to the fine-tuning of the whole model, partial fine-tuning is also used to evaluate the quality of the learned repre-

Table 6.3 Quantitative evaluations of the proposed approach on the different settings of Bi-LSTM and random modality combination training strategy with complete and incomplete multimodality inputs on the DFC2023 track2 dataset. Results are reported in terms of AP@50 for instance segmentation and mIoU for semantic segmentation.

Multimodal Input	Sup. w/o LSTM		Sup. w/o Random		Sup. w/ all	
	ins.	seg.	ins.	seg.	ins.	seg.
SAR, RGB, DSM	0.265	0.809	0.301	0.854	0.333	0.851
SAR, RGB	0.213	0.728	0.083	0.660	0.296	0.809
SAR, DSM	0.173	0.763	0.061	0.696	0.233	0.779
RGB, DSM	0.165	0.807	0.224	0.782	0.332	0.847
SAR	0.028	0.509	0.000	0.372	0.040	0.552
RGB	0.210	0.722	0.061	0.577	0.291	0.799
DSM	0.168	0.749	0.040	0.664	0.211	0.753

sentation in a self-supervised approach. Partial fine-tuning involves freezing the backbone and updating only the task-specific decoder on the two tasks. It is important to note that contrastive pre-training relies on masked attention to keep each modality independent, especially when working with different data formats such as text and images. The use of masked attention in contrastive pre-training helps in avoiding information flow from one modality to the other, thereby keeping modality-specific information through the network. This is more beneficial for downstream tasks that involve only a single modality. However, when using generative pre-training, masked self-attention is not mandatory. Here, we show the fine-tuning results based on the combination of the pre-training (the use of reconstruction loss and contrastive loss), the generative pre-training (the only use of reconstruction loss), and the fine-tuning results without masked self-attention for both tasks (see Table 6.5 and Table 6.6). In the first row, we remove the masked Self-Attention blocks while keeping the random modality combination training strategy in fine-tuning, which results in a significant improvement in performance. This is probably because masked self-attention hinders the interaction between different modalities. Compared with the generative pre-training, the use of masked attention in the combination pre-training helps to avoid the information flow from one modality to the other. As one can see, the unimodal inference performs close to the modal-incomplete inputs as the modality streams are more independently treated. In contrast, the results without contrastive pre-training tend to overfit on dominant modalities and are relatively poor on other modalities. Moreover, lower performances are observed on one single modality.

Table 6.4 Quantitative evaluations of the proposed approach on the different settings of Bi-LSTM and random modality combination training strategy with complete and incomplete multimodality inputs on the quadruplets dataset. The results are reported in terms of mIoU.

Multimodal Input	Sup. w/o LSTM	Sup. w/o Random	Sup. w/ all
S1, S2, DEM, DNW	0.242	0.244	0.244
S1, S2, DEM	0.227	0.175	0.229
S1, S2, DNW	0.244	0.247	0.244
S1, DEM, DNW	0.237	0.198	0.235
S2, DEM, DNW	0.240	0.239	0.237
S1, S2	0.228	0.174	0.232
S1, DEM	0.201	0.058	0.208
S1, DNW	0.239	0.197	0.239
S2, DEM	0.211	0.139	0.210
S2, DNW	0.241	0.239	0.239
DEM, DNW	0.231	0.179	0.234
S1	0.203	0.051	0.208
S2	0.212	0.136	0.215
DEM	0.013	0.053	0.010
DNW	0.233	0.163	0.234

Table 6.5 Quantitative evaluations of the proposed approach in fine-tuning paradigm with different settings with complete and incomplete multimodality inputs on DFC2023 track2 dataset. Results are reported in terms of AP@50 for instance segmentation and mIoU for semantic segmentation.

Multimodal Input	Fine. w/o Mask		Partial Fine.		Full Fine.	
	ins.	seg.	ins.	seg.	ins.	seg.
SAR, RGB, DSM	0.317	0.850	0.215	0.807	0.300	0.849
SAR, RGB	0.276	0.799	0.136	0.711	0.260	0.798
SAR, DSM	0.220	0.783	0.173	0.767	0.202	0.780
RGB, DSM	0.318	0.845	0.206	0.800	0.300	0.844
SAR	0.034	0.562	0.022	0.499	0.037	0.566
RGB	0.276	0.789	0.132	0.694	0.254	0.784
DSM	0.205	0.752	0.152	0.747	0.187	0.754

6.5 Conclusion

In this chapter, we have introduced an incomplete multimodal learning framework for multimodal remote sensing data fusion which can be used in both supervised training and self-supervised pre-training paradigms. Unlike previous multimodal remote sensing data fusion approaches, the proposed approach enables the training and inference of models with modal-incomplete inputs. By using the Bi-LSTM attention mechanism and masked

Table 6.6 Quantitative evaluations of the proposed approach in fine-tuning paradigm with different settings with complete and incomplete multimodality inputs on the quadruplets dataset. The results are reported in terms of mIoU.

Multimodal Input	Fine. w/o Mask	Partial Fine.	Full Fine.
S1, S2, DEM, DNW	0.250	0.233	0.246
S1, S2, DEM	0.243	0.223	0.238
S1, S2, DNW	0.248	0.222	0.246
S1, DEM, DNW	0.238	0.217	0.235
S2, DEM, DNW	0.242	0.221	0.240
S1, S2	0.238	0.223	0.240
S1, DEM	0.221	0.212	0.216
S1, DNW	0.240	0.224	0.235
S2, DEM	0.231	0.198	0.227
S2, DNW	0.242	0.221	0.239
DEM, DNW	0.245	0.216	0.238
S1	0.226	0.212	0.214
S2	0.239	0.203	0.228
DEM	0.023	0.011	0.051
DNW	0.241	0.214	0.237

self-attention, we are able to pre-train the network using contrastive and reconstruction losses in the MultiMAE framework, and also to train the network from scratch or finetune the model on downstream tasks using a random modality combination strategy. This strategy allows the network to maintain high performance even when dealing with modal-incomplete inputs or a single modality in the inference stage.

We evaluated our model on two multimodal remote sensing datasets, demonstrating flexibility in network training and inference, and state-of-the-art performance when presented with modal-incomplete inputs. It is worth noting that this study focused solely on different modality raster data. In future work, we plan to incorporate diverse modalities data, such as text and vector data, into the proposed framework.

Chapter 7

Conclusions

This chapter concludes the dissertation by presenting an overall discussion of the thesis, a brief overview of the novel contributions, and the related critical analysis. Moreover, we propose possible future developments of the works.

7.1 Summary and Discussion

In this thesis, we have presented novel contributions to the field of self-supervised remote sensing image change detection and data fusion. The research highlights the importance of employing self-supervised learning methodologies for unsupervised change detection in high-resolution remote sensing images, particularly for Sentinel-1, Sentinel-2 and Landsat-8 images. The existing approaches in remote sensing image change detection predominantly rely on supervised learning algorithms, which encounter two primary challenges: the limitation of change semantics and the lack of generalizability. RSI semantic changes are especially limited due to the presence of non-semantic changes in high-resolution remote sensing images, such as variations in water quality and arid conditions. Consequently, the thesis emphasizes the alignment of pixel features in multitemporal and multisensor images, accounting for seasonal and sensor noises, which play a crucial role in unsupervised change detection in high-resolution remote sensing images. Seasonal and sensor noises are not only the natural augmentation of multitemporal and multisensor images but also restrict the detection performance. Furthermore, one crucial aspect of seasonal noise that is often related to the time scale and the change types we intend to detect. For example, if image pairs are captured with one-year intervals, phenomena like the snow may not be considered as changes. However, to detect subtle changes related to seasons, the interval between acquired images should not exceed one season. It is noteworthy that most research studies neglect this crucial fact in unsupervised change detection.

Similar to the supervised change detection approach, self-supervised change detection should also consider certain constraints. The conventional self-supervised change detection algorithms, based on the CNN-based generative models, have been found to focus primarily on pixel reconstruction rather than feature extraction. Meanwhile, this issue can be resolved by using Transformer-based generative models but patch-based algorithms make getting continuous feature representation difficult and require more distillation. Consequently, the thesis focuses more on the contrastive paradigm, treating multi-temporal and multi-sensor image pairs as distinct views.

In Chapter 3, the thesis explored image patch-based contrastive learning in multi-view remote sensing image change detection first, including both single-sensor and cross-sensor scenarios. Specifically, a pseudo-Siamese network utilizing ResNet-34 as the backbone is trained to regress the output between two branches, which is trained using contrastive loss on large archived multi-view image-patch pairs. Finally, changes are identified by a change score that can accurately model the feature distance between bi-temporal images. The experimental results on both single-sensor (e.g., Sentinel-1 SAR images, Sentinel-2 multispectral images and Landsat-8 multi-spectral images) and cross-sensor (e.g., Sentinel-1/2 image pairs and Landsat-8/Sentinel-2 image pairs) datasets demonstrate the superiority of the proposed approach over state-of-the-art unsupervised methods and narrow the gap with supervised approaches on performance. Additionally, the results reveal a decline in performance when using cross-sensor multispectral image pairs with different resolutions compared to multispectral image pairs from the same sensor. This case provides preliminary results of unsupervised change detection based on patch-level self-supervised learning. However, the patch-based approach (PatchSSL) is computationally expensive and neglects subtle changes. Consequently, the chapter further proposed a pixel-wise self-supervised change detection based on contrastive learning. The proposed approach incorporates two branches with input shift-augmented image pairs. Instead of applying contrastive loss on each pixel feature, the contrastive loss is employed on the averaged feature over superpixels. In addition, an uncertainty-based distillation process in the teacher-student paradigm is proposed to reduce the impact of seasonal changes. The experimental results on multi-view remote sensing image datasets demonstrate the superiority and efficiency of the proposed approach over state-of-the-art methods. Compared with the PatchSSL approach, the proposed PixSSL demonstrates better inference efficiency and yields improved change maps, particularly in vegetation and water areas. The results also indicate that the use of uncertainty-based approaches further suppresses seasonal changes compared to the sole use of contrastive learning.

Chapter 4 extends the unsupervised change detection from bi-temporal image pairs to satellite image time series using self-supervised learning. Unlike unsupervised bi-temporal change detection, satellite image time series analysis focuses on capturing the spatial-temporal information in image sequences. The unsupervised bi-temporal change detection, based on the pseudo-Siamese network, only learns the relationship between bi-temporal image pairs. To address this, this chapter adopted the self-training algorithm based on the ConvLSTM network and pseudo labels. Initially, pseudo labels are derived from pre-trained models, and feature tracking is employed to propagate the pseudo labels among the image time series. This strategy enhances the consistency of pseudo labels and enables the generation of change maps for long-term satellite image time series. To overcome the overfitting problem during self-training, supervised contrastive loss and contrastive random walk loss are utilized. The experimental results on the Landsat-8 and Sentinel-2 image time series demonstrate that the proposed approach suppresses most seasonal changes and achieves significant noise reduction compared to state-of-the-art models in both fitting and inference scenarios. Notably, the state-of-the-art model trains individual networks for each scene and tailors them to specific images during network training. The ablation studies on the proposed approach indicate that the feature tracking strategy mitigates seasonal changes in long change map time series, and the combined use of contrastive loss and contrastive random walk loss further improves the performance of the self-training paradigm compared to the use of cross-entropy loss alone.

The proposed self-supervised change detection approaches also have some limitations that should be properly understood for the correct use of them. To ensure reliable change detection, the time intervals used in the training data are critical. In this task, we utilized time-series images for the network training with interval spans that ranged from half a month to two years. However, the proposed approach does not account for specific sporadic weather phenomena, such as snow. In addition, seasonal changes in croplands pose another challenge for the method. Except for permanent croplands, the seasonal changes in croplands are often rapid and significant. For instance, after harvesting, most croplands exhibit shifts from grassy areas to bare land. To obtain built-up changes exclusively, one possible choice is to include more cropland data and push the network to fit on such data. However, we advocate for considering all changes and then identifying them using the spectral information of remote sensing data. Furthermore, the proposed image-time series change detection approach contains limitations in its use of contrastive random walk loss to resist noise in pseudo labels and preserve the consistency of changes in the image time series. The time window utilized in the contrastive random walk loss is crucial for generating stable change maps. When choosing the complete time series as a time window, the change maps will likely be constant

throughout the whole time series. Therefore, we recommend and provide evidence to include in the window the adjacent four images instead.

Most previous multimodal data fusion approaches are task-specific models, which need to train an individual model for each task. However, self-supervised learning models can generalize to downstream tasks with the fine-tuning of small labeled datasets. Here, we have presented another interesting application of self-supervised learning in RS multimodal data fusion. In multi-view contrastive learning, multimodal remote sensing images are treated as multiple augmented views. However, the network only captures the shared information among different views. Another approach involves stacking multimodal images together and learning the invariant representation between the stacked images and their augmentation. In multimodal autoencoders, the network learns to reconstruct each modality using the unmasked parts from the remaining modalities, thereby capturing the complementary information shared among multimodal remote sensing data.

In Chapter 5, the thesis proposed the fusion of SAR and optical images at the pixel level using contrastive learning. The proposed method encompasses three fusion strategies: early fusion (PixEF), intermediate fusion (pixIF) and late fusion (PixLF). It uses ResUnet as the backbone in a pseudo-Siamese network and the shift transformation to augment the inputs in two branches. These three fusion strategies, along with the state-of-the-art methods, are evaluated on linear protocol and fine-tuning settings. The experimental results on the DFC2020 dataset demonstrate that the proposed intermediate fusion (PixIF) outperforms state-of-the-art unsupervised data fusion approaches and the other two fusion strategies (PixEF and PixLF). PixIF achieves comparable performance with the weakly supervised method that utilizes image-level labels. In terms of land-cover maps, SAR-optical fusion outperforms the use of any single modality data. Among single modality data, the use of Sentinel-2 images yields similar results to SAR-optical fusion and outperforms the sole use of Sentinel-1 images. Furthermore, ablation studies on data augmentation reveal that shift augmentation alone achieves the best performance, while other geometric and photometric augmentations lead to a drop in performance. Building upon the proposed SAR-optical early-fusion framework, a new unsupervised land-cover segmentation approach using contrastive learning and vector quantization is proposed in this chapter. It employs a pseudo-Siamese architecture with one branch as a ResUnet and the other branch as the Gumbel-softmax vector quantizer. The experimental results on a subset of the DFC2020 dataset demonstrate that the proposed approach can learn semantically meaningful representations and effectively discriminate between different land-cover categories using SAR-optical data pairs.

Chapter 6 delves into the unsupervised multimodal remote sensing data fusion using both contrastive and generative models, as well as incomplete multimodal learning for

downstream tasks. A unified model is proposed in both supervised and self-supervised paradigms for incomplete multimodal learning in multimodal remote sensing data fusion. The proposed approach uses additional learned fusion tokens in the multimodal Transformer for multimodal information collection. A Bi-LSTM attention block is employed before the self-attention block to distil different modality information to the fusion tokens. During pre-training, the multimodal Transformer utilizes masked self-attention for contrastive learning and incorporates multimodal decoders for MultiMAE training. Experimental results on the building instance /semantic segmentation task (SAR, RGB and DSM modalities) and the LULC mapping task (Sentinel-1, Sentinel-2, DEM and Dynamic World maps) reveal that the proposed approach performs impressively well on all modality incomplete inputs and single modality inputs. On the contrary, the vanilla MultiViT model exhibits overfitting on dominant modality inputs and fails completely on tasks with single modality inputs. In the context of fine-tuning, the LULC mapping task indicates that the fine-tuning model achieves the best performance, where the pre-training approach combining contrastive learning and MultiMAE outperforms the use of MultiMAE alone. Ablation studies indicate that the random modality combination strategy is crucial for maintaining performance with modality-incomplete inputs, while the Bi-LSTM attention block enables superior interaction of the fusion token with each modality input. Additionally, the chapter includes LULC maps generated as a remote sensing product and as an additional data source in RS multimodal data fusion.

Among remote sensing data fusion methodologies, the proposed method based on contrastive learning has some inherent limitations. It can only obtain an invariant representation when dealing with varying noise and augmentations. While it works with two modalities that provide complementary information, such as SAR and optical images, it can only capture common information between the modalities rather than the complementary information they provide. On the other hand, contrastive learning can be an effective way of connecting two modalities, such as image and text. In the case of incomplete multimodal learning, this approach is used to enable fusion tokens that contain information from all modalities. However, this requires us to restrict each modality from accessing information from other modalities, which could potentially limit the model's ability to learn a robust representation of individual modalities.

7.2 Future Developments

This section presents some of the possible future developments of the works in this thesis. The current self-supervised change detection method primarily focuses on learning representative pixel features, which may not be sufficient for very high-resolution (VHR) remote sensing

images. Change detection from VHR remote sensing images is challenging due to the limited spectral information, the spectral variability, and the geometric distortions [158]. Geometric distortions, in particular, make it difficult to align pixels accurately. To address these challenges, an enhanced version of self-supervised learning can be developed by incorporating the DINO framework [23, 121]. DINO has the ability to learn the instance object in a self-supervised way, which can be utilized to align objects in bi-temporal VHR remote sensing image pairs and detect object-level changes. Another potential extension of this object-aware change detection approach is the development of change-type classification. The current self-supervised change detection only focuses on binary change detection, which may not be adequate for some downstream applications. Leveraging the feature distillation capabilities of pre-trained DINO models offers an opportunity for unsupervised object classification [66, 93]. A possible solution is to align the objects in bi-temporal VHR image pairs and thus detect the changed objects. Then an unsupervised classification approach, such as the simple K-means, can be used to further classify the different change objects. The spectral and physical information contained within the images can be utilized to assign different classes to various land-cover and land-use objects.

Multimodal RS data fusion involves a broad range of possible modalities, including the raster images, LULC maps, vector data and text data. In the current research, we only considered the raster data fusion. It is crucial to include diverse modalities for more advanced applications, such as text and vector data, in addition to remote sensing products. Text data are widely used as a prompt in remote sensing applications, such as question answering [99], text-image retrieval [33], image captioning[32], and referring segmentation [101, 167]. Incorporating vector data into remote sensing image processing can contribute to sound decision-making on downstream tasks, such as change detection [142]. A promising future extension of the current research is to develop a unified model that encompasses all these different modalities in both pre-training and downstream tasks. Unlike computer vision, remote sensing data exhibit diverse properties, including SAR, multispectral images, hyperspectral images, and LiDAR, in addition to RGB images. The proposed raster-based multimodal remote sensing data fusion approach can serve as an intermediary to connect data from different modalities.

References

[1] Abd El-Kawy, O., Rød, J., Ismail, H., and Suliman, A. (2011). Land use and land cover change detection in the western nile delta of egypt using remote sensing data. *Applied geography*, 31(2):483–494.

[2] Adrian, J., Sagan, V., and Maimaitijiang, M. (2021). Sentinel sar-optical fusion for crop type mapping using deep learning and google earth engine. *ISPRS Journal of Photogrammetry and Remote Sensing*, 175:215–235.

[3] Amarsaikhan, D., Blotevogel, H., Van Genderen, J., Ganzorig, M., Gantuya, R., and Nergui, B. (2010). Fusing high-resolution sar and optical imagery for improved urban land cover study and classification. *International Journal of Image and Data Fusion*, 1(1):83–97.

[4] Andrew, G., Arora, R., Bilmes, J., and Livescu, K. (2013). Deep canonical correlation analysis. In *International conference on machine learning*, pages 1247–1255. PMLR.

[5] Araslanov, N., Schaub-Meyer, S., and Roth, S. (2021). Dense unsupervised learning for video segmentation. *Advances in Neural Information Processing Systems*, 34:25308–25319.

[6] Audebert, N., Le Saux, B., and Lefèvre, S. (2018). Beyond rgb: Very high resolution urban remote sensing with multimodal deep networks. *ISPRS Journal of Photogrammetry and Remote Sensing*, 140:20–32.

[7] Bachman, P., Hjelm, R. D., and Buchwalter, W. (2019). Learning representations by maximizing mutual information across views. In *Advances in Neural Information Processing Systems*, pages 15535–15545.

[8] Bachmann, R., Mizrahi, D., Atanov, A., and Zamir, A. (2022). Multimae: Multi-modal multi-task masked autoencoders. In *Computer Vision–ECCV 2022: 17th European Conference, Tel Aviv, Israel, October 23–27, 2022, Proceedings, Part XXXVII*, pages 348–367. Springer.

[9] Bank, D., Koenigstein, N., and Giryes, R. (2020). Autoencoders. *arXiv preprint arXiv:2003.05991*.

[10] Bao, H., Dong, L., Piao, S., and Wei, F. (2021). Beit: Bert pre-training of image transformers. *arXiv preprint arXiv:2106.08254*.

[11] Bergamasco, L., Saha, S., Bovolo, F., and Bruzzone, L. (2019). Unsupervised change-detection based on convolutional-autoencoder feature extraction. In *Image and Signal Processing for Remote Sensing XXV*, volume 11155, page 1115510. International Society for Optics and Photonics.

[12] Bermudez, J., Happ, P., Oliveira, D., and Feitosa, R. (2018). Sar to optical image synthesis for cloud removal with generative adversarial networks. *ISPRS Annals of Photogrammetry, Remote Sensing & Spatial Information Sciences*, 4(1).

[13] Bischke, B., Helber, P., Koenig, F., Borth, D., and Dengel, A. (2018). Overcoming missing and incomplete modalities with generative adversarial networks for building footprint segmentation. In *2018 International Conference on Content-Based Multimedia Indexing (CBMI)*, pages 1–6. IEEE.

[14] Bottou, L. et al. (1991). Stochastic gradient learning in neural networks. *Proceedings of Neuro-Nımes*, 91(8):12.

[15] Bovolo, F. and Bruzzone, L. (2006). A theoretical framework for unsupervised change detection based on change vector analysis in the polar domain. *IEEE Transactions on Geoscience and Remote Sensing*, 45(1):218–236.

[16] Bovolo, F., Marchesi, S., and Bruzzone, L. (2011). A framework for automatic and unsupervised detection of multiple changes in multitemporal images. *IEEE Transactions on Geoscience and Remote Sensing*, 50(6):2196–2212.

[17] Bromley, J., Bentz, J. W., Bottou, L., Guyon, I., LeCun, Y., Moore, C., Säckinger, E., and Shah, R. (1993). Signature verification using a "siamese" time delay neural network. *International Journal of Pattern Recognition and Artificial Intelligence*, 7(04):669–688.

[18] Brown, C. F., Brumby, S. P., Guzder-Williams, B., Birch, T., Hyde, S. B., Mazzariello, J., Czerwinski, W., Pasquarella, V. J., Haertel, R., Ilyushchenko, S., et al. (2022). Dynamic world, near real-time global 10 m land use land cover mapping. *Scientific Data*, 9(1):251.

[19] Bruzzone, L., Conese, C., Maselli, F., and Roli, F. (1997). Multisource classification of complex rural areas by statistical and neural-network approaches. *Photogrammetric Engineering and Remote Sensing*, 63(5):523–532.

[20] Bruzzone, L. and Prieto, D. F. (2000). Automatic analysis of the difference image for unsupervised change detection. *IEEE Transactions on Geoscience and Remote sensing*, 38(3):1171–1182.

[21] Bruzzone, L., Prieto, D. F., and Serpico, S. B. (1999). A neural-statistical approach to multitemporal and multisource remote-sensing image classification. *IEEE Transactions on Geoscience and remote Sensing*, 37(3):1350–1359.

[22] Bruzzone, L. and Serpico, S. (1997). Detection of changes in remotely-sensed images by the selective use of multi-spectral information. *International Journal of Remote Sensing*, 18(18):3883–3888.

[23] Caron, M., Touvron, H., Misra, I., Jégou, H., Mairal, J., Bojanowski, P., and Joulin, A. (2021). Emerging properties in self-supervised vision transformers. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 9650–9660.

[24] Caterini, A. (2017). A novel mathematical framework for the analysis of neural networks. Master's thesis, University of Waterloo.

[25] Celik, T. (2009). Unsupervised change detection in satellite images using principal component analysis and *k*-means clustering. *IEEE geoscience and remote sensing letters*, 6(4):772–776.

[26] Chen, T., Kornblith, S., Norouzi, M., and Hinton, G. (2020). A simple framework for contrastive learning of visual representations. In *International conference on machine learning*, pages 1597–1607. PMLR.

[27] Chen, X., Duan, Y., Houthooft, R., Schulman, J., Sutskever, I., and Abbeel, P. (2016). Infogan: Interpretable representation learning by information maximizing generative adversarial nets. In *Advances in neural information processing systems*, pages 2172–2180.

[28] Chen, Y. and Bruzzone, L. (2021). Self-supervised change detection in multi-view remote sensing images. *arXiv preprint arXiv:2103.05969*.

[29] Chen, Y. and Bruzzone, L. (2022a). An approach based on contrastive learning and vector quantization to the unsupervised land-cover segmentation of multimodal images. In *IGARSS 2022-2022 IEEE International Geoscience and Remote Sensing Symposium*, pages 4811–4814. IEEE.

[30] Chen, Y. and Bruzzone, L. (2022b). A self-supervised approach to pixel-level change detection in bi-temporal rs images. *IEEE Transactions on Geoscience and Remote Sensing*, 60:1–11.

[31] Chen, Y., Li, C., Ghamisi, P., Jia, X., and Gu, Y. (2017). Deep fusion of remote sensing data for accurate classification. *IEEE Geoscience and Remote Sensing Letters*, 14(8):1253–1257.

[32] Cheng, Q., Huang, H., Xu, Y., Zhou, Y., Li, H., and Wang, Z. (2022). Nwpu-captions dataset and mlca-net for remote sensing image captioning. *IEEE Transactions on Geoscience and Remote Sensing*, 60:1–19.

[33] Cheng, Q., Zhou, Y., Fu, P., Xu, Y., and Zhang, L. (2021). A deep semantic alignment network for the cross-modal image-text retrieval in remote sensing. *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, 14:4284–4297.

[34] Chi, M., Plaza, A., Benediktsson, J. A., Sun, Z., Shen, J., and Zhu, Y. (2016). Big data for remote sensing: Challenges and opportunities. *Proceedings of the IEEE*, 104(11):2207–2219.

[35] Daudt, R., Le Saux, B., Boulch, A., and Gousseau, Y. (2019). Multitask learning for large-scale semantic change detection. *Computer Vision and Image Understanding*, 187:102783.

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

[37] Daudt, R. C., Le Saux, B., Boulch, A., and Gousseau, Y. (2018b). Urban change detection for multispectral earth observation using convolutional neural networks. In *IGARSS 2018-2018 IEEE International Geoscience and Remote Sensing Symposium*, pages 2115–2118. IEEE.

[38] De, S., Bruzzone, L., Bhattacharya, A., Bovolo, F., and Chaudhuri, S. (2017). A novel technique based on deep learning and a synthetic target database for classification of urban areas in polsar data. *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, 11(1):154–170.

[39] De Bem, P. P., de Carvalho Junior, O. A., Fontes Guimarães, R., and Trancoso Gomes, R. A. (2020). Change detection of deforestation in the brazilian amazon using landsat data and convolutional neural networks. *Remote Sensing*, 12(6):901.

[40] Deng, J., Wang, K., Deng, Y., and Qi, G. (2008). Pca-based land-use change detection and analysis using multitemporal and multisensor satellite data. *International Journal of Remote Sensing*, 29(16):4823–4838.

[41] Derksen, D., Inglada, J., and Michel, J. (2018). Spatially precise contextual features based on superpixel neighborhoods for land cover mapping with high resolution satellite image time series. In *IGARSS 2018-2018 IEEE International Geoscience and Remote Sensing Symposium*, pages 200–203. IEEE.

[42] Deus, D. (2016). Integration of alos palsar and landsat data for land cover and forest mapping in northern tanzania. *Land*, 5(4):43.

[43] Devlin, J., Chang, M.-W., Lee, K., and Toutanova, K. (2018). Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*.

[44] Dewan, A. M. and Yamaguchi, Y. (2009). Land use and land cover change in greater dhaka, bangladesh: Using remote sensing to promote sustainable urbanization. *Applied geography*, 29(3):390–401.

[45] Diab, A., Kashef, R., and Shaker, A. (2022). Deep learning for lidar point cloud classification in remote sensing. *Sensors*, 22(20):7868.

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

[47] Dong, H., Ma, W., Wu, Y., Zhang, J., and Jiao, L. (2020). Self-supervised representation learning for remote sensing image change detection based on temporal prediction. *Remote Sensing*, 12(11):1868.

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

[49] Du, B., Ru, L., Wu, C., and Zhang, L. (2019). Unsupervised deep slow feature analysis for change detection in multi-temporal remote sensing images. *IEEE Transactions on Geoscience and Remote Sensing*, 57(12):9976–9992.

[50] Du, X., Zheng, X., Lu, X., and Doudkin, A. A. (2021). Multisource remote sensing data classification with graph fusion network. *IEEE Transactions on Geoscience and Remote Sensing*, 59(12):10062–10072.

[51] Felzenszwalb, P. F. and Huttenlocher, D. P. (2004). Efficient graph-based image segmentation. *International journal of computer vision*, 59(2):167–181.

[52] Feng, Q., Yang, J., Zhu, D., Liu, J., Guo, H., Bayartungalag, B., and Li, B. (2019). Integrating multitemporal sentinel-1/2 data for coastal land cover classification using a multibranch convolutional neural network: A case of the yellow river delta. *Remote Sensing*, 11(9):1006.

[53] Fernandez-Beltran, R., Haut, J. M., Paoletti, M. E., Plaza, J., Plaza, A., and Pla, F. (2018). Remote sensing image fusion using hierarchical multimodal probabilistic latent semantic analysis. *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, 11(12):4982–4993.

[54] Gargiulo, M., Dell'Aglio, D. A., Iodice, A., Riccio, D., and Ruello, G. (2020). Integration of sentinel-1 and sentinel-2 data for land cover mapping using w-net. *Sensors*, 20(10):2969.

[55] Geng, J., Wang, H., Fan, J., and Ma, X. (2017). Classification of fusing sar and multispectral image via deep bimodal autoencoders. In *2017 IEEE International Geoscience and Remote Sensing Symposium (IGARSS)*, pages 823–826. IEEE.

[56] Ghamisi, P., Rasti, B., Yokoya, N., Wang, Q., Hofle, B., Bruzzone, L., Bovolo, F., Chi, M., Anders, K., Gloaguen, R., et al. (2019). Multisource and multitemporal data fusion in remote sensing: A comprehensive review of the state of the art. *IEEE Geoscience and Remote Sensing Magazine*, 7(1):6–39.

[57] Gómez-Chova, L., Tuia, D., Moser, G., and Camps-Valls, G. (2015). Multimodal classification of remote sensing images: A review and future directions. *Proceedings of the IEEE*, 103(9):1560–1584.

[58] Gong, M., Niu, X., Zhang, P., and Li, Z. (2017). Generative adversarial networks for change detection in multispectral imagery. *IEEE Geoscience and Remote Sensing Letters*, 14(12):2310–2314.

[59] Gong, M., Zhang, P., Su, L., and Liu, J. (2016). Coupled dictionary learning for change detection from multisource data. *IEEE Transactions on Geoscience and Remote sensing*, 54(12):7077–7091.

[60] Goodfellow, I., Bengio, Y., and Courville, A. (2016). *Deep learning*. MIT press.

[61] Goodfellow, I., Pouget-Abadie, J., Mirza, M., Xu, B., Warde-Farley, D., Ozair, S., Courville, A., and Bengio, Y. (2014). Generative adversarial nets. In *Advances in neural information processing systems*, pages 2672–2680.

[62] Graves, A., Fernández, S., and Schmidhuber, J. (2005). Bidirectional lstm networks for improved phoneme classification and recognition. In *International conference on artificial neural networks*, pages 799–804. Springer.

[63] Grill, J.-B., Strub, F., Altché, F., Tallec, C., Richemond, P. H., Buchatskaya, E., Doersch, C., Pires, B. A., Guo, Z. D., Azar, M. G., et al. (2020). Bootstrap your own latent: A new approach to self-supervised learning. *arXiv preprint arXiv:2006.07733*.

[64] Gutmann, M. and Hyvärinen, A. (2010). Noise-contrastive estimation: A new estimation principle for unnormalized statistical models. In *Proceedings of the Thirteenth International Conference on Artificial Intelligence and Statistics*, pages 297–304. JMLR Workshop and Conference Proceedings.

[65] Guttler, F., Ienco, D., Nin, J., Teisseire, M., and Poncelet, P. (2017). A graph-based approach to detect spatiotemporal dynamics in satellite image time series. *ISPRS Journal of Photogrammetry and Remote Sensing*, 130:92–107.

[66] Hamilton, M., Zhang, Z., Hariharan, B., Snavely, N., and Freeman, W. T. (2022). Unsupervised semantic segmentation by distilling feature correspondences. *arXiv preprint arXiv:2203.08414*.

[67] Harb, R. and Knöbelreiter, P. (2021). Infoseg: Unsupervised semantic image segmentation with mutual information maximization. In *German Conference for Pattern Recognition*.

[68] He, K., Chen, X., Xie, S., Li, Y., Dollár, P., and Girshick, R. (2022). Masked autoencoders are scalable vision learners. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 16000–16009.

[69] He, K., Fan, H., Wu, Y., Xie, S., and Girshick, R. (2020). Momentum contrast for unsupervised visual representation learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 9729–9738.

[70] He, K., Zhang, X., Ren, S., and Sun, J. (2016). Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778.

[71] Hjelm, R. D., Fedorov, A., Lavoie-Marchildon, S., Grewal, K., Bachman, P., Trischler, A., and Bengio, Y. (2018). Learning deep representations by mutual information estimation and maximization. *arXiv preprint arXiv:1808.06670*.

[72] Hochreiter, S. and Schmidhuber, J. (1997). Long short-term memory. *Neural computation*, 9(8):1735–1780.

[73] Hong, D., Gao, L., Hang, R., Zhang, B., and Chanussot, J. (2020). Deep encoder–decoder networks for classification of hyperspectral and lidar data. *IEEE Geoscience and Remote Sensing Letters*, 19:1–5.

[74] Hotelling, H. (1992). Relations between two sets of variates. In *Breakthroughs in statistics*, pages 162–190. Springer.

[75] Huang, M. and Jin, S. (2020). Rapid flood mapping and evaluation with a supervised classifier and change detection in shouguang using sentinel-1 sar and sentinel-2 optical data. *Remote Sensing*, 12(13):2073.

[76] Hütt, C., Waldhoff, G., and Bareth, G. (2020). Fusion of sentinel-1 with official topographic and cadastral geodata for crop-type enriched lulc mapping using foss and open data. *ISPRS International Journal of Geo-Information*, 9(2):120.

[77] Ienco, D., Interdonato, R., Gaetano, R., and Minh, D. H. T. (2019). Combining sentinel-1 and sentinel-2 satellite image time series for land cover mapping via a multi-source deep learning architecture. *ISPRS Journal of Photogrammetry and Remote Sensing*, 158:11–22.

[78] Jabri, A., Owens, A., and Efros, A. (2020). Space-time correspondence as a contrastive random walk. *Advances in neural information processing systems*, 33:19545–19560.

[79] Johnson, K. and Koperski, K. (2017). Worldview-3 swir land use-land cover mineral classification: Cuprite, nevada. *Remote Sens. GIS*.

[80] Junaid, M., Sun, J., Iqbal, A., Sohail, M., Zafar, S., and Khan, A. (2023). Mapping lulc dynamics and its potential implication on forest cover in malam jabba region with landsat time series imagery and random forest classification. *Sustainability*, 15(3):1858.

[81] Kalinicheva, E., Ienco, D., Sublime, J., and Trocan, M. (2020). Unsupervised change detection analysis in satellite image time series using deep learning combined with graph-based approaches. *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, 13:1450–1466.

[82] Kalinicheva, E., Sublime, J., and Trocan, M. (2019). Change detection in satellite images using reconstruction errors of joint autoencoders. In *International Conference on Artificial Neural Networks*, pages 637–648. Springer.

[83] Kampffmeyer, M., Salberg, A.-B., and Jenssen, R. (2018). Urban land cover classification with missing data modalities using deep convolutional neural networks. *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, 11(6):1758–1768.

[84] Kemper, H. and Kemper, G. (2020). Sensor fusion, gis and ai technologies for disaster management. *The International Archives of Photogrammetry, Remote Sensing and Spatial Information Sciences*, 43:1677–1683.

[85] Kendall, A. and Gal, Y. (2017). What uncertainties do we need in bayesian deep learning for computer vision? *arXiv preprint arXiv:1703.04977*.

[86] Khosla, P., Teterwak, P., Wang, C., Sarna, A., Tian, Y., Isola, P., Maschinot, A., Liu, C., and Krishnan, D. (2020). Supervised contrastive learning. In *Advances in Neural Information Processing Systems*.

[87] Kingma, D. P. and Welling, M. (2013). Auto-encoding variational bayes. *arXiv preprint arXiv:1312.6114*.

[88] Klein, L. A. (1999). Sensor and data fusion concepts and applications. Society of Photo-Optical Instrumentation Engineers (SPIE).

[89] Kussul, N., Lavreniuk, M., Skakun, S., and Shelestov, A. (2017). Deep learning classification of land cover and crop types using remote sensing data. *IEEE Geoscience and Remote Sensing Letters*, 14(5):778–782.

[90] LeCun, Y. et al. (2015). Lenet-5, convolutional neural networks. *URL: http://yann. lecun. com/exdb/lenet*, 20(5):14.

[91] Leenstra, M., Marcos, D., Bovolo, F., and Tuia, D. (2021). Self-supervised pre-training enhances change detection in sentinel-2 imagery. In *Pattern Recognition. ICPR International Workshops and Challenges*, pages 578–590. Springer International Publishing.

[92] Li, H., Ghamisi, P., Soergel, U., and Zhu, X. X. (2018). Hyperspectral and lidar fusion using deep three-stream convolutional neural networks. *Remote Sensing*, 10(10):1649.

[93] Li, K., Wang, Z., Cheng, Z., Yu, R., Zhao, Y., Song, G., Liu, C., Yuan, L., and Chen, J. (2023). Acseg: Adaptive conceptualization for unsupervised semantic segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 7162–7172.

[94] Liu, S., Bruzzone, L., Bovolo, F., Zanetti, M., and Du, P. (2015). Sequential spectral change vector analysis for iteratively discovering and detecting multiple changes in hyperspectral images. *IEEE transactions on geoscience and remote sensing*, 53(8):4363–4378.

[95] Liu, S., Marinelli, D., Bruzzone, L., and Bovolo, F. (2019). A review of change detection in multitemporal hyperspectral images: Current techniques, applications, and challenges. *IEEE Geoscience and Remote Sensing Magazine*, 7(2):140–158.

[96] Liu, X., Zhang, F., Hou, Z., Mian, L., Wang, Z., Zhang, J., and Tang, J. (2021a). Self-supervised learning: Generative or contrastive. *IEEE Transactions on Knowledge and Data Engineering*.

[97] Liu, X., Zhang, F., Hou, Z., Wang, Z., Mian, L., Zhang, J., and Tang, J. (2020). Self-supervised learning: Generative or contrastive. *arXiv preprint arXiv:2006.08218*, 1(2).

[98] Liu, Z.-G., Zhang, Z.-W., Pan, Q., and Ning, L.-B. (2021b). Unsupervised change detection from heterogeneous data based on image translation. *IEEE Transactions on Geoscience and Remote Sensing*.

[99] Lobry, S., Marcos, D., Murray, J., and Tuia, D. (2020). Rsvqa: Visual question answering for remote sensing data. *IEEE Transactions on Geoscience and Remote Sensing*, 58(12):8555–8566.

[100] Long, S., Fatoyinbo, T. E., and Policelli, F. (2014). Flood extent mapping for namibia using change detection and thresholding with sar. *Environmental Research Letters*, 9(3):035002.

[101] Lüddecke, T. and Ecker, A. (2022). Image segmentation using text and image prompts. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 7086–7096.

[102] Luppino, L. T., Bianchi, F. M., Moser, G., and Anfinsen, S. N. (2019). Unsupervised image regression for heterogeneous change detection. *IEEE Transactions on Geoence and Remote Sensing*, 57(12):9960–9975.

[103] Luppino, L. T., Hansen, M. A., Kampffmeyer, M., Bianchi, F. M., Moser, G., Jenssen, R., and Anfinsen, S. N. (2020). Code-aligned autoencoders for unsupervised change detection in multimodal remote sensing images. *arXiv preprint arXiv:2004.07011*.

[104] Luppino, L. T., Hansen, M. A., Kampffmeyer, M., Bianchi, F. M., Moser, G., Jenssen, R., and Anfinsen, S. N. (2022). Code-aligned autoencoders for unsupervised change detection in multimodal remote sensing images. *IEEE Transactions on Neural Networks and Learning Systems*.

[105] Lv, N., Chen, C., Qiu, T., and Sangaiah, A. K. (2018). Deep learning and superpixel feature extraction based on contractive autoencoder for change detection in sar images. *IEEE transactions on industrial informatics*, 14(12):5530–5538.

[106] Lyu, H., Lu, H., and Mou, L. (2016). Learning a transferable change rule from a recurrent neural network for land cover change detection. *Remote Sensing*, 8(6):506.

[107] Ma, M., Ren, J., Zhao, L., Testuggine, D., and Peng, X. (2022). Are multimodal transformers robust to missing modality? In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 18177–18186.

[108] Masci, J., Meier, U., Cireşan, D., and Schmidhuber, J. (2011). Stacked convolutional auto-encoders for hierarchical feature extraction. In *International conference on artificial neural networks*, pages 52–59. Springer.

[109] Meshkini, K., Bovolo, F., and Bruzzone, L. (2022). A 3d cnn approach for change detection in hr satellite image time series based on a pretrained 2d cnn. *The International Archives of Photogrammetry, Remote Sensing and Spatial Information Sciences*, 43:143–150.

[110] Meshkini, Khatereh and Bovolo, Francesca and Bruzzone, Lorenzo (2021). An unsupervised change detection approach for dense satellite image time series using 3d cnn. In *2021 IEEE International Geoscience and Remote Sensing Symposium IGARSS*, pages 4336–4339. IEEE.

[111] Mi, L., Wang, H., Tian, Y., and Shavit, N. (2019). Training-free uncertainty estimation for dense regression: Sensitivity as a surrogate. *arXiv preprint arXiv:1910.04858*.

[112] Mohla, S., Pande, S., Banerjee, B., and Chaudhuri, S. (2020). Fusatnet: Dual attention based spectrospatial multimodal fusion network for hyperspectral and lidar classification. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops*, pages 92–93.

[113] Montero, E., Van Wolvelaer, J., and Garzón, A. (2014). The european urban atlas. In *Land Use and Land Cover Mapping in Europe: Practices & Trends*, pages 115–124. Springer.

[114] Mou, L., Bruzzone, L., and Zhu, X. X. (2018). Learning spectral-spatial-temporal features via a recurrent convolutional neural network for change detection in multispectral imagery. *IEEE Transactions on Geoscience and Remote Sensing*, 57(2):924–935.

[115] Nagrani, A., Yang, S., Arnab, A., Jansen, A., Schmid, C., and Sun, C. (2021). Attention bottlenecks for multimodal fusion. *Advances in Neural Information Processing Systems*, 34:14200–14213.

[116] Nemoto, K., Hamaguchi, R., Sato, M., Fujita, A., Imaizumi, T., and Hikosaka, S. (2017). Building change detection via a combination of cnns using only rgb aerial imageries. In *Remote Sensing Technologies and Applications in Urban Environments II*, volume 10431, pages 107–118. SPIE.

[117] Nielsen, A. A., Conradsen, K., and Simpson, J. J. (1998). Multivariate alteration detection (mad) and maf postprocessing in multispectral, bitemporal image data: New approaches to change detection studies. *Remote Sensing of Environment*, 64(1):1–19.

[118] Nielsen, A. A. and Larsen, R. (2017). Canonical analysis of sentinel-1 radar and sentinel-2 optical data. In *Scandinavian Conference on Image Analysis*, pages 147–158. Springer.

[119] Niu, X., Gong, M., Zhan, T., and Yang, Y. (2018). A conditional adversarial network for change detection in heterogeneous images. *IEEE Geoscience and Remote Sensing Letters*, 16(1):45–49.

[120] Oord, A. v. d., Li, Y., and Vinyals, O. (2018). Representation learning with contrastive predictive coding. *arXiv preprint arXiv:1807.03748*.

[121] Oquab, M., Darcet, T., Moutakanni, T., Vo, H., Szafraniec, M., Khalidov, V., Fernandez, P., Haziza, D., Massa, F., El-Nouby, A., et al. (2023). Dinov2: Learning robust visual features without supervision. *arXiv preprint arXiv:2304.07193*.

[122] Oussidi, A. and Elhassouny, A. (2018). Deep generative models: Survey. In *2018 International Conference on Intelligent Systems and Computer Vision (ISCV)*, pages 1–8. IEEE.

[123] Ozsoy, S., Hamdan, S., Arik, S., Yuret, D., and Erdogan, A. (2022). Self-supervised learning with an information maximization criterion. *Advances in Neural Information Processing Systems*, 35:35240–35253.

[124] Paisitkriangkrai, S., Sherrah, J., Janney, P., Hengel, V.-D., et al. (2015). Effective semantic pixel labelling with convolutional networks and conditional random fields. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops*, pages 36–43.

[125] Paris, C. and Bruzzone, L. (2014). A three-dimensional model-based approach to the estimation of the tree top height by fusing low-density lidar data and very high resolution optical images. *IEEE Transactions on Geoscience and Remote Sensing*, 53(1):467–480.

[126] Peng, D., Zhang, Y., and Guan, H. (2019). End-to-end change detection for high resolution satellite images using improved unet++. *Remote Sensing*, 11(11):1382.

[127] Rahman, F., Vasu, B., Van Cor, J., Kerekes, J., and Savakis, A. (2018). Siamese network with multi-level features for patch-based change detection in satellite imagery. In *2018 IEEE Global Conference on Signal and Information Processing (GlobalSIP)*, pages 958–962. IEEE.

[128] Recasens, A., Lin, J., Carreira, J., Jaegle, D., Wang, L., Alayrac, J.-b., Luc, P., Miech, A., Smaira, L., Hemsley, R., et al. (2023). Zorro: the masked multimodal transformer. *arXiv preprint arXiv:2301.09595*.

[129] Ren, C., Wang, X., Gao, J., Zhou, X., and Chen, H. (2020). Unsupervised change detection in satellite images with generative adversarial network. *IEEE Transactions on Geoscience and Remote Sensing*, pages 1–15.

[130] Ronneberger, O., Fischer, P., and Brox, T. (2015). U-net: Convolutional networks for biomedical image segmentation. In *International Conference on Medical image computing and computer-assisted intervention*, pages 234–241. Springer.

[131] Rosenblatt, F. et al. (1962). *Principles of neurodynamics: Perceptrons and the theory of brain mechanisms*, volume 55. Spartan books Washington, DC.

[132] Rosin, P. L. and Hervás, J. (2005). Remote sensing image thresholding methods for determining landslide activity. *International Journal of Remote Sensing*, 26(6):1075–1092.

[133] Roy, D. P., Huang, H., Boschetti, L., Giglio, L., Yan, L., Zhang, H. H., and Li, Z. (2019). Landsat-8 and sentinel-2 burned area mapping-a combined sensor multi-temporal change detection approach. *Remote Sensing of Environment*, 231:111254.

[134] Roy, S. K., Deria, A., Hong, D., Rasti, B., Plaza, A., and Chanussot, J. (2022). Multimodal fusion transformer for remote sensing image classification. *arXiv preprint arXiv:2203.16952*.

[135] Saha, S., Bovolo, F., and Bruzzone, L. (2019a). Unsupervised deep change vector analysis for multiple-change detection in vhr images. *IEEE Transactions on Geoscience and Remote Sensing*, 57(6):3677–3693.

[136] Saha, S., Bovolo, F., and Bruzzone, L. (2020). Change detection in image time-series using unsupervised lstm. *IEEE Geoscience and Remote Sensing Letters*.

[137] Saha, S., Solano-Correa, Y. T., Bovolo, F., and Bruzzone, L. (2019b). Unsupervised deep learning based change detection in sentinel-2 images. In *2019 10th International Workshop on the Analysis of Multitemporal Remote Sensing Images (MultiTemp)*, pages 1–4. IEEE.

[138] Schmitt, M., Hughes, L. H., and Zhu, X. X. (2018). The sen1-2 dataset for deep learning in sar-optical data fusion. *arXiv preprint arXiv:1807.01569*.

[139] Schmitt, M., Prexl, J., Ebel, P., Liebel, L., and Zhu, X. X. (2020). Weakly supervised semantic segmentation of satellite images for land cover mapping–challenges and opportunities. *arXiv preprint arXiv:2002.08254*.

[140] Sefrin, O., Riese, F. M., and Keller, S. (2020). Deep learning for land cover change detection. *Remote Sensing*, 13(1):78.

[141] Seydi, S. T., Rastiveis, H., Kalantar, B., Halin, A. A., and Ueda, N. (2022). Bdd-net: An end-to-end multiscale residual cnn for earthquake-induced building damage detection. *Remote Sensing*, 14(9):2214.

[142] Shi, J., Liu, W., Zhu, Y., Wang, S., Hao, S., Zhu, C., Shan, H., Li, E., Li, X., and Zhang, L. (2022). Fine object change detection based on vector boundary and deep learning with high-resolution remote sensing images. *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, 15:4094–4103.

[143] Shi, W., Zhang, M., Ke, H., Fang, X., Zhan, Z., and Chen, S. (2020). Landslide recognition by deep convolutional neural network and change detection. *IEEE Transactions on Geoscience and Remote Sensing*, 59(6):4654–4672.

[144] Shi, X., Chen, Z., Wang, H., Yeung, D.-Y., Wong, W.-K., and Woo, W.-c. (2015). Convolutional lstm network: A machine learning approach for precipitation nowcasting. *Advances in neural information processing systems*, 28.

[145] Singh, M., Duval, Q., Alwala, K. V., Fan, H., Aggarwal, V., Adcock, A., Joulin, A., Dollár, P., Feichtenhofer, C., Girshick, R., et al. (2023). The effectiveness of mae pre-pretraining for billion-scale pretraining. *arXiv preprint arXiv:2303.13496*.

[146] Song, H., Wang, W., Zhao, S., Shen, J., and Lam, K.-M. (2018). Pyramid dilated deeper convlstm for video salient object detection. In *Proceedings of the European conference on computer vision (ECCV)*, pages 715–731.

[147] Steinhausen, M. J., Wagner, P. D., Narasimhan, B., and Waske, B. (2018). Combining sentinel-1 and sentinel-2 data for improved land use and land cover mapping of monsoon regions. *International journal of applied earth observation and geoinformation*, 73:595–604.

[148] Sun, S., Mu, L., Wang, L., and Liu, P. (2020). L-unet: An lstm network for remote sensing image change detection. *IEEE Geoscience and Remote Sensing Letters*.

[149] Sun, Y., Lei, L., Li, X., Sun, H., and Kuang, G. (2021). Nonlocal patch similarity based heterogeneous remote sensing change detection. *Pattern Recognition*, 109:107598.

[150] Tarvainen, A. and Valpola, H. (2017). Mean teachers are better role models: Weight-averaged consistency targets improve semi-supervised deep learning results. In *Advances in neural information processing systems*, pages 1195–1204.

[151] Tian, Y., Krishnan, D., and Isola, P. (2020). Contrastive multiview coding. In *Computer Vision – ECCV 2020*, pages 776–794.

[152] Van Etten, A., Hogan, D., Manso, J. M., Shermeyer, J., Weir, N., and Lewis, R. (2021). The multi-temporal urban development spacenet dataset. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 6398–6407.

[153] Van Zyl, J. J. (2001). The shuttle radar topography mission (srtm): a breakthrough in remote sensing of topography. *Acta astronautica*, 48(5-12):559–565.

[154] Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, Ł., and Polosukhin, I. (2017). Attention is all you need. *Advances in neural information processing systems*, 30.

[155] Wald, L. (1998a). Data fusion: a conceptual approach for an efficient exploitation of remote sensing images. In *2nd International Conference" Fusion of Earth Data: merging point measurements, raster maps and remotely sensed images"*, pages 17–24. SEE/URISCA.

[156] Wald, L. (1998b). A european proposal for terms of reference in data fusion. In *Commission VII Symposium" Resource and Environmental Monitoring"*, volume 32, pages 651–654.

[157] Wang, S., Chen, W., Xie, S. M., Azzari, G., and Lobell, D. B. (2020). Weakly supervised deep learning for segmentation of remote sensing imagery. *Remote Sensing*, 12(2):207.

[158] Wen, D., Huang, X., Bovolo, F., Li, J., Ke, X., Zhang, A., and Benediktsson, J. A. (2021). Change detection from very-high-spatial-resolution optical remote sensing images: Methods, applications, and future directions. *IEEE Geoscience and Remote Sensing Magazine*, 9(4):68–101.

[159] Wu, C., Du, B., Cui, X., and Zhang, L. (2017). A post-classification change detection method based on iterative slow feature analysis and bayesian soft fusion. *Remote Sensing of Environment*, 199:241–255.

[160] Wu, C., Du, B., and Zhang, L. (2013). Slow feature analysis for change detection in multispectral imagery. *IEEE Transactions on Geoscience and Remote Sensing*, 52(5):2858–2874.

[161] Xu, Q., Long, C., Yu, L., and Zhang, C. (2023). Road extraction with satellite images and partial road maps. *IEEE Transactions on Geoscience and Remote Sensing*.

[162] Yang, B., Qin, L., Liu, J., and Liu, X. (2022). Utrnet: An unsupervised time-distance-guided convolutional recurrent network for change detection in irregularly collected images. *IEEE Transactions on Geoscience and Remote Sensing*, 60:1–16.

[163] Yokoya, N., Ghamisi, P., Haensch, R., and Schmitt, M. (2020). 2020 ieee grss data fusion contest: Global land cover mapping with weak supervision [technical committees]. *IEEE Geoscience and Remote Sensing Magazine*, 8(1):154–157.

[164] Zanetti, M. (2023). A one-class classification model for burned-area detection based on mutual ordering of normalized differences. *IEEE Transactions on Geoscience and Remote Sensing*.

[165] Zanetti, M., Bovolo, F., and Bruzzone, L. (2015). Rayleigh-rice mixture parameter estimation via em algorithm for change detection in multispectral images. *IEEE Transactions on Image Processing*, 24(12):5004–5016.

[166] Zhan, Y., Fu, K., Yan, M., Sun, X., Wang, H., and Qiu, X. (2017). Change detection based on deep siamese convolutional network for optical aerial images. *IEEE Geoscience and Remote Sensing Letters*, 14(10):1845–1849.

[167] Zhan, Y., Xiong, Z., and Yuan, Y. (2022). Rsvg: Exploring data and models for visual grounding on remote sensing data. *arXiv preprint arXiv:2210.12634*.

[168] Zhang, L., Liao, M., Yang, L., and Lin, H. (2007). Remote sensing change detection based on canonical correlation analysis and contextual bayes decision. *Photogrammetric Engineering & Remote Sensing*, 73(3):311–318.

[169] Zhang, M., Li, W., Du, Q., Gao, L., and Zhang, B. (2018a). Feature extraction for classification of hyperspectral and lidar data using patch-to-patch cnn. *IEEE transactions on cybernetics*, 50(1):100–111.

[170] Zhang, Z., Liu, Q., and Wang, Y. (2018b). Road extraction by deep residual u-net. *IEEE Geoscience and Remote Sensing Letters*, 15(5):749–753.

[171] Zhao, W., Wang, Z., Gong, M., and Liu, J. (2017). Discriminative feature learning for unsupervised change detection in heterogeneous images based on a coupled neural network. *IEEE Transactions on Geoscience and Remote Sensing*, 55(12):7066–7080.

[172] Zheng, Z., Zhong, Y., Wang, J., Ma, A., and Zhang, L. (2021). 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*, 265:112636.

[173] Zhou, Y. and Li, X. (2020). Unsupervised self-training algorithm based on deep learning for optical aerial images change detection. *arXiv preprint arXiv:2010.07469*.

[174] Zink, M., Bachmann, M., Brautigam, B., Fritz, T., Hajnsek, I., Moreira, A., Wessel, B., and Krieger, G. (2014). Tandem-x: The new global dem takes shape. *IEEE Geoscience and Remote Sensing Magazine*, 2(2):8–23.