

Multimodal Co-learning with VHR Multispectral Imagery and Photogrammetric Data

Yuxing Xie

Vollständiger Abdruck der von der TUM School of Engineering and Design der
Technischen Universität München zur Erlangung des akademischen Grades eines

Doktors der Ingenieurwissenschaften (Dr.-Ing.)

genehmigten Dissertation.

Vorsitz:

Prof. Dr.-Ing. Muhammad Shahzad

Prüfende der Dissertation:

1. Prof. Dr.-Ing. habil. Xiao Xiang Zhu
2. Hon.-Prof. Dr.-Ing. Peter Reinartz
3. Prof. Dr. techn. Friedrich Fraundorfer

Die Dissertation wurde am 20.03.2024 bei der Technischen Universität München eingereicht und durch die TUM School of Engineering and Design am 14.08.2024 angenommen.

Abstract

The rapid development of sensor technology has facilitated unprecedented access to diverse remote sensing data, intertwining numerous fields with remote sensing techniques more closely than ever before. Among these, urban remote sensing stands out as a crucial topic. Meanwhile, as the volume of available data far exceeds manual processing capacities, the development and implementation of automatic processing methodologies have become necessary. As a significant breakthrough for big data analysis, deep learning has brought revolutionary change in remote sensing tasks, achieving remarkable performance on various benchmarks.

In most cases, the success of deep learning techniques relies on supervised training with similar data. Therefore, single-modal networks' performance is restricted when meeting complex scenarios that cannot be fully described by a single modality alone. Benefiting from the potential to integrate complementary advantages of different data modalities, multimodal learning has been recognized as a solution to more challenging tasks. As the most widely applied multimodal approach, data fusion has made progress in several urban remote sensing applications. However, it has two inherent limitations. On the one hand, it necessitates the availability of complete multimodal data even during the testing phase, which poses a challenge for the vast volumes of single-modal data. On the other hand, it cannot utilize intrinsic information of raw heterogeneous data, potentially leading to diminished performance. To avoid the mentioned limitations and utilize the advantages of multimodal information in a more friendly manner, this dissertation conducts a comprehensive investigation on multimodal co-learning, the idea of transferring knowledge between different data modalities while maintaining their independence. Specifically speaking, this dissertation makes the following contributions:

- We conduct comprehensive studies on two essential yet crucial urban remote sensing tasks, building extraction and building change detection, with multispectral orthophotos and corresponding photogrammetric geometric data derived from the same source. We develop and employ proper networks for different modalities depending on the task, including unitemporal images, unitemporal point clouds, bitemporal images, and DSM-derived height difference maps.
- We develop three flexible multimodal co-learning frameworks adapted for different urban remote sensing tasks, which can outperform single-modal learning and circumvent the limitations of conventional data fusion approaches. The first framework enhances the building extraction networks in scenarios with limited labeled data. The second framework enhances the building extraction networks when using cross-domain labeled data. The third framework is for building change detection tasks, which also utilizes cross-domain labeled data. The cornerstone of these frameworks' success lies in their ability to effectively transfer beneficial knowledge between 2-D spectral modality and 2.5-D/3-D geometric modality, through labeled or unlabeled data pairs.

Abstract

- Considering the lack of public multimodal urban remote sensing data, we develop a large synthetic dataset that can be utilized to evaluate the algorithms for building extraction, semantic segmentation, and building change detection. We conduct a series of experiments between this synthetic dataset and real datasets. Promising results are achieved by utilizing our proposed multimodal co-learning frameworks. These results not only further demonstrate the capability of our methods, but also provide a potential to employ cost-efficient and annotation-friendly synthetic training data for real urban remote sensing applications.

Zusammenfassung

Die rapide Entwicklung der Sensortechnologie hat einen idealen Zugang zu verschiedenen Fernerkundungsdaten ermöglicht, wodurch zahlreiche Bereiche enger als je zuvor mit Fernerkundungstechniken verbunden sind. Darunter ist die städtische Fernerkundung ein wichtiges Thema. Da die Menge der verfügbaren Daten die manuellen Verarbeitungskapazitäten bei weitem übersteigt, sind die Entwicklung und Implementierung automatischer Verarbeitungsmethoden notwendig geworden. Als bedeutender Durchbruch für die Big-Data-Analyse hat Deep Learning einen revolutionären Wandel in der Fernerkundung bewirkt und eine bemerkenswerte Leistung bei verschiedenen Benchmarks erzielt.

In den meisten Fällen hängt der Erfolg von Deep-Learning-Techniken vom überwachten Training mit ähnlichen Daten ab. Daher ist die Leistung monomodaler Netze bei komplexen Szenarien, die mit einer einzigen Modalität allein nicht vollständig beschrieben werden können, eingeschränkt. Multimodales Lernen nutzt das Potenzial, komplementäre Vorteile verschiedener Datenmodalitäten zu integrieren, und wurde als Lösung für anspruchsvollere Aufgaben identifiziert. Als der am weitesten verbreitete multimodale Ansatz hat die Datenfusion in mehreren städtischen Fernerkundungsanwendungen Fortschritte gemacht. Sie hat jedoch zwei inhärente Einschränkungen. Zum einen müssen bereits in der Testphase vollständige multimodale Daten zur Verfügung stehen, was bei den riesigen Mengen an monomodalen Daten eine Herausforderung darstellt. Andererseits können die intrinsischen Informationen der heterogenen Rohdaten nicht genutzt werden, was zu einer verminderten Leistung führen kann. Um die genannten Einschränkungen zu vermeiden und die Vorteile multimodaler Informationen auf eine angenehme Art und Weise zu nutzen, führt diese Dissertation eine umfassende Untersuchung zum multimodalen Co-Learning durch, d.h. der Idee, Wissen zwischen verschiedenen Datenmodalitäten unter Beibehaltung ihrer Unabhängigkeit zu transferieren. Konkret leistet diese Dissertation die folgenden Beiträge:

- Wir führen umfassende Studien zu zwei wesentlichen, aber entscheidenden städtischen Fernerkundungsaufgaben durch, der Gebäudeextraktion und der Erkennung von Gebäudeveränderungen, mit multispektralen Orthofotos und entsprechenden photogrammetrischen geometrischen Daten, die aus derselben Quelle stammen. Wir nutzen und entwickeln geeignete Netzwerke für unterschiedliche Modalitäten je nach Aufgabe, einschließlich monotemporaler Bilder, monotemporaler Punktwolken, bitemporaler Bilder und DSM-abgeleiteter Höhendifferenzkarten.
- Wir entwickeln drei flexible multimodale Co-Learning-Frameworks, die für verschiedene Aufgaben in der städtischen Fernerkundung geeignet sind und welche die Grenzen herkömmlicher Datenfusionsansätze überwinden können. Das erste Framework verbessert die Gebäudeextraktionsnetzwerke in Szenarien mit begrenzten gelabelten Daten. Das zweite Framework verbessert die Gebäudeextraktionsnetzwerke bei der Verwendung bereichsübergreifender markierter Daten. Das dritte Framework ist für Aufgaben zur Erkennung von Gebäudeveränderungen gedacht, bei denen

Zusammenfassung

ebenfalls domänenübergreifende markierte Daten verwendet werden. Der Grundstein für den Erfolg dieser Methoden liegt in ihrer Fähigkeit, vorteilhaftes Wissen zwischen der 2-D-Spektralmodalität und der 2,5-D/3-D-Geometriemodalität durch beschriftete oder unbeschriftete Datenpaare zu übertragen.

- In Anbetracht des Mangels an öffentlichen multimodalen städtischen Fernerkundungsdaten entwickeln wir einen großen synthetischen Datensatz, der zur Bewertung der Algorithmen für die Gebäudeextraktion, semantische Segmentierung und Erkennung von Gebäudeveränderungen verwendet werden kann. Wir führen eine Reihe von Experimenten zwischen diesem synthetischen Datensatz und realen Datensätzen durch. Durch den Einsatz der von uns vorgelegten multimodalen Co-Learning-Frameworks werden vielversprechende Ergebnisse erzielt. Diese Ergebnisse sind nicht nur ein weiterer Beweis für die Leistungsfähigkeit unserer Methoden, sondern bieten auch die Möglichkeit, kosteneffiziente und kommentarfreundliche synthetische Trainingsdaten für reale städtische Fernerkundungsanwendungen einzusetzen.

Acknowledgements

Finally, this dissertation has reached its end. I would like to express my sincere gratitude to everyone who supported me along the journey.

First and foremost, my sincerest thanks go to my supervisors. I am deeply grateful to Dr. Jiaojiao Tian for her mentoring. Over the past years, she has been the person guiding me and discussing most of my work with me, supporting me throughout my entire Ph.D. journey. I would like to thank Prof. Xiaoxiang Zhu for her supervision. She connected me with advanced topics and excellent researchers. My gratitude extends to Prof. Peter Reinartz, my department leader in DLR. He offered the funding that covered the majority of my expenses, allowing me the opportunity to live and work in Munich for several years. In addition, I sincerely appreciate Prof. Friedrich Fraundorfer for being the reviewer of my dissertation and Prof. Muhammad Shahzad for serving as the chair of the examination committee. I am grateful for their valuable time and effort amidst their busy schedules.

I would like to express my gratitude to my colleagues and friends at DLR. During the past years, I spent most of my time with them. I would like to thank Xiangtian Yuan. I will always remember our adventures in Europe. I would like to thank Dr. Guichen Zhang and Dr. Seyed Majid Azimi, my kind office mates. The pleasant time shared with them is unforgettable. I sincerely thank Dr. Wei Yao, Mario Fuentes Reyes, Chengfa Benjamin Lee, Philipp Schuegraf, Corentin Henry, Dr. Nina Merkle, Shuangyi Liu, Jens Hellekes, Dr. Yuanxin Xia, Dr. Xiangyu Zhuo, Dr. Ksenia Bittner, Christian Kempf, Prof. Danfeng Hong, Dr. Yao Sun, Prof. Yuansheng Hua, Dr. Reza Bahmanyar, Prof. Lichao Mou, Dr. Jingliang Hu, Dr. Lanlan Rao, Prof. Rong Liu, Jianxiang Feng, Haolin Li, and Teo Beker. I enjoy every talk, every lunch, and every party with them, and every event in Christmas gatherings and summer schools. Especially, I will never forget the table games, hiking, and dinner events during the COVID pandemic. I would like to thank Sabine Knickl, Maximilian Langheinrich, and Peter Schwind for their support in secretarial and IT issues, smoothing the way for my research work. I am also grateful to Dr. Stefan Auer, Dr. Thomas Krauss, and Dr. Franz Kurz. As senior researchers, their sharing often inspired me. I would like to give my special thanks to Dr. Pablo d'Angelo, helping me a lot in generating high-quality DSM data, and Dr. Daniele Cerra, always generous in helping proofread English writings.

I would like to thank researchers who are working or used to work in TUM, Jingwei Zhu, Dr. Qingyu Li, Prof. Yusheng Xu, Prof. Rong Huang, Dr. Kun Qian, and Dr. Qian Song, for the support and help in their research fields. I also would like to convey my gratitude to professors, colleagues, and friends in other universities or institutes. I appreciate Prof. Konrad Schindler and Prof. Jan Dirk Wegner for their kind guidance during my exchange time in ETH Zurich. I thank Shengyu Huang, Dr. Binbin Xiang, and Dr. Yapin Lin. The discussions with them gave me many ideas on 3-D data. I also thank Prof. Gaoqiao Wu. His nice monitor significantly improved my comfort when writing this dissertation. Additionally, I would like to give my thanks to Prof. Xin Li, Prof. Zheng Ji, and Prof. Qin Yan, who supported or encouraged me at the beginning of this journey, when I was applying for this Ph.D. position.

Acknowledgements

Furthermore, I sincerely and deeply appreciate all my non-academic friends in Munich and Zurich, from whom I learned a great deal. My greatest gains in recent years came from the conversations with them. Additionally, I would like to thank every named and anonymous teacher who taught me meditation. It helped me build the foundation for pursuing the essence of things.

Last but not least, I would like to express my deepest gratitude to my beloved wife Xu Cheng, and my parents. This journey has always been accompanied by their care, support, understanding, and trust.

Being the fool, I am still on the way.

Contents

| | |
|---|-------------|
| Abstract | iii |
| Zusammenfassung | v |
| Acknowledgements | vii |
| Acronyms | xiii |
| 1 Introduction | 1 |
| 1.1 Motivations and Objectives | 1 |
| 1.2 Dissertation Outline | 3 |
| 2 Basics | 5 |
| 2.1 Varied Dimensions in Remote Sensing Data | 5 |
| 2.1.1 Multispectral Imagery | 5 |
| 2.1.2 Point Cloud | 6 |
| 2.1.3 DSM | 7 |
| 2.2 Deep Learning Methodology for Multidimensional Remote Sensing Data . . | 7 |
| 2.2.1 Image Networks | 8 |
| 2.2.1.1 Convolutional Neural Networks | 8 |
| 2.2.1.2 Vision Transformer | 9 |
| 2.2.1.3 Siamese Networks | 10 |
| 2.2.2 Point Cloud Networks | 10 |
| 2.2.2.1 Voxel-based Semantic Segmentation Networks | 10 |
| 2.2.2.2 Point-based Semantic Segmentation Networks | 11 |
| 3 State of the Art | 13 |
| 3.1 2-D Multispectral Imagery Analysis with Deep Learning | 13 |
| 3.1.1 Building Extraction | 13 |
| 3.1.2 Change Detection | 15 |
| 3.2 2.5-D/3-D Remote Sensing Data Analysis with Deep Learning | 16 |
| 3.3 Multimodal Learning with 2-D and 2.5-D/3-D Remote Sensing Data | 17 |
| 4 Summary of the Work | 21 |
| 4.1 What is Multimodal Co-learning? | 21 |
| 4.2 Multimodal Datasets Involved in This Dissertation | 22 |
| 4.2.1 Building Extraction Datasets | 22 |
| 4.2.1.1 ISPRS Potsdam Dataset | 22 |
| 4.2.1.2 Munich WorldView-2 Dataset | 22 |
| 4.2.1.3 Simulated Multimodal Aerial Remote Sensing (SMARS) Dataset | 23 |

Contents

| | | |
|----------|---|-----------|
| 4.2.2 | Change Detection Datasets | 24 |
| 4.2.2.1 | SMARS Dataset | 24 |
| 4.2.2.2 | Istanbul WorldView-2 Dataset | 24 |
| 4.3 | Case Studies with 2-D and 2.5-D/3-D Multimodal Learning | 25 |
| 4.3.1 | Case I: Multimodal Co-learning Enhances the Building Extraction Networks with Limited Labeled Data | 25 |
| 4.3.1.1 | Background | 25 |
| 4.3.1.2 | Methodology | 26 |
| 4.3.1.3 | Experiments | 27 |
| 4.3.1.4 | Summary | 34 |
| 4.3.2 | Case II: Multimodal Co-learning Enhances the Building Extraction Networks with Cross-domain Data | 34 |
| 4.3.2.1 | Background | 34 |
| 4.3.2.2 | Methodology | 34 |
| 4.3.2.3 | Experiments | 36 |
| 4.3.2.4 | Summary | 39 |
| 4.3.3 | Case III: Multimodal Co-learning Enhances the Building Change Detection Networks with Cross-domain Data | 40 |
| 4.3.3.1 | Background | 40 |
| 4.3.3.2 | Methodology | 41 |
| 4.3.3.3 | Experiments | 46 |
| 4.3.3.4 | Summary | 52 |
| 5 | Conclusion and Outlook | 53 |
| 5.1 | Conclusion | 53 |
| 5.2 | Outlook | 54 |
| | List of Figures | 57 |
| | List of Tables | 59 |
| | Bibliography | 61 |
| | Appendices | 73 |
| A | Yuxing Xie, Jiaojiao Tian, and Xiao Xiang Zhu. “Linking points with labels in 3D: A review of point cloud semantic segmentation.” <i>IEEE Geoscience and remote sensing magazine</i> 8.4 (2020): 38-59. | 75 |
| B | Yuxing Xie, Jiaojiao Tian, and Xiao Xiang Zhu. “A co-learning method to utilize optical images and photogrammetric point clouds for building extraction.” <i>International Journal of Applied Earth Ob- servation and Geoinformation</i> 116 (2023): 103165. | 97 |

- C Mario Fuentes Reyes*, Yuxing Xie*, Xiangtian Yuan*, Pablo d'Angelo, Franz Kurz, Daniele Cerra, and Jiaojiao Tian. "A 2D/3D multimodal data simulation approach with applications on urban semantic segmentation, building extraction and change detection." *ISPRS Journal of Photogrammetry and Remote Sensing* 205 (2023): 74-97. (* equal contribution) 115
- D Yuxing Xie, Xiangtian Yuan, Xiao Xiang Zhu, and Jiaojiao Tian. "Multimodal Co-learning for Building Change Detection: A Domain Adaptation Framework Using VHR Images and Digital Surface Models." *IEEE Transactions on Geoscience and Remote Sensing* (2024) 62 (2024): 5402520. 141

Acronyms

| | |
|-------|--|
| 2-D | two-dimensional. |
| 2.5-D | two-and-a-half-dimensional. |
| 3-D | three-dimensional. |
| AI | artificial intelligence. |
| ALS | airborne laser scanning. |
| BIT | bitemporal image transformer. |
| CNN | convolutional neural network. |
| DoG | difference of Gaussian. |
| DSM | digital surface model. |
| FN | false negative. |
| FNR | false negative rate. |
| FP | false positive. |
| FPR | false positive rate. |
| GAN | generative adversarial network. |
| GB | gigabytes. |
| GCN | graph convolutional network. |
| GIS | geographic information system. |
| GPU | graphical processing unit. |
| GSD | ground sampling distance. |
| HDiff | height difference. |
| HR | high resolution. |
| InSAR | interferometric synthetic aperture radar. |
| IoU | intersection over union. |
| ISPRS | international society for photogrammetry and remote sensing. |
| KL | Kullback-Leibler. |
| LiDAR | light detection and ranging. |
| MLP | multilayer perceptron. |

Acronyms

| | |
|---------|---|
| nDSM | normalized digital surface model. |
| OA | overall accuracy. |
| PSI | persistent scatterers interferometry. |
| RGB | red green blue. |
| SAR | synthetic aperture radar. |
| SMARS | simulated multimodal aerial remote sensing. |
| SOTA | state of the art. |
| TB | terabytes. |
| TN | true negative. |
| TomoSAR | synthetic aperture radar tomography. |
| TP | true positive. |
| UAV | unmanned aerial vehicle. |
| VHR | very high resolution. |
| ViT | vision transformer. |

1 Introduction

1.1 Motivations and Objectives

Benefiting from the diversification of sensors and the ever-growing data volume, the remote sensing field has entered the era of big data. Every day, just commercial satellite imaging companies alone can collect hundreds of terabytes (TB) of data per day [1]. If airborne and unmanned aerial vehicle (UAV)-based data are also included and counted, the amount of available data volume would be enormous and far beyond what could be handled manually. Therefore, investigating and developing automatic processing methodologies for large-scale remote sensing data has become increasingly crucial.

As a breakthrough for big data analysis in past years, deep learning has brought unprecedented performance to a variety of high resolution (HR) and very high resolution (VHR) optical remote sensing tasks [2], such as semantic segmentation [3, 4], building extraction [5, 6], and change detection [7, 8, 9]. Among these studies, two-dimensional (2-D) convolutional networks (CNNs) are first proven to be effective and have been the most widely used [2, 10]. More recently, vision transformer (ViT) methods have been becoming exceptional and eye-catching, due to their ability to better capture global information by the self-attention mechanism [11, 12, 13, 14]. Although advanced 2-D networks have demonstrated strong capabilities on VHR optical imagery-based tasks, they encounter challenges when processing complex scenarios with diverse objects and less-than-ideal imaging conditions in real applications. Sometimes 2-D neural networks find it hard to distinguish objects in different categories with similar colors [6]. For instance, Figure 1.1 is an example from the ISPRS Potsdam benchmark ¹. The cycled object is a building of which the color is close to the color of nearby vegetation. It is wrongly classified as a non-building object by the U-Net in [6]. Three-dimensional (3-D) point clouds and two-and-a-half-dimensional (2.5-D) digital surface models (DSMs) with rich geometric information can compensate for the shortcomings of 2-D optical images, as they are good at describing objects with regular geometric shapes. Taking advantage of the development in photogrammetric techniques, light detection and ranging (LiDAR) sensors, as well as synthetic aperture radar (SAR) tomography (TomoSAR) techniques, 3-D and 2.5-D data are becoming easier and cheaper to obtain. In addition, deep learning also demonstrates remarkable performance on these dimensions. Aiming at processing 3-D point clouds, numerous meticulously designed point cloud backbones have been proposed and made considerable achievements in point cloud semantic segmentation, superior to conventional methods in terms of accuracy [15, 16, 17, 18, 19].

As an attempt to combine the advantages of 2-D and 2.5-D/3-D remote sensing data, multimodal data fusion has attracted many researchers' attention [20, 21]. Data fusion methods try to enhance the features or probabilities via a fusion operation (e.g., summation, average, concatenation, etc.). They have achieved better results than single-modal learning in many cases. However, data fusion has two inherent limitations. One issue

¹<https://www.isprs.org/education/benchmarks/UrbanSemLab/2d-sem-label-potsdam.aspx>

1 Introduction



Figure 1.1: An example of the spectral confusion between buildings and vegetation in the ISPRS Potsdam dataset.

is that data fusion approaches rely on strict data composition. Full modalities should exist not only in the training phase but also in the testing phase. The other limitation is that data fusion may not utilize complete information of the raw heterogeneous data and the complementary nature of multimodalities [22, 23]. As multimodal information is processed with only one single network stream, incorrect and irrelevant representations may be calculated from the fused data. For example, processing colored point clouds (a kind of early fusion operation for point clouds and optical images) with point cloud networks might result in a decline in semantic segmentation and building extraction performance compared to processing single-modal raw point clouds [24, 25, 26, 6]. Based on these above-mentioned situations, several questions come out for the multimodal remote sensing data analysis:

- Can we reduce the data requirements of conventional data fusion and handle the case in which one modality is missing during the testing phase?
- Can we utilize two completely independent neural networks rather than a single network stream to process the 2-D and 2.5-D/3-D data, respectively? If so, the completeness of heterogeneous information from each modality can be maintained.
- Can we utilize limited labeled data in the training phase to train reasonable networks for the test data?

Inspired by the aforementioned facts, this dissertation aims to develop novel 2-D and 2.5-D/3-D multimodal deep learning algorithms and frameworks for remote sensing data analysis. Specifically speaking, it has accomplished the following objectives:

- Develop effective 2-D and 2.5-D/3-D multimodal learning frameworks to address the problem caused by imperfect (e.g., insufficient or cross-domain) training data.
- Develop flexible frameworks that can overcome the limitations of conventional data fusion. To keep the completeness of natural information from each modality, our

frameworks adopt two completely independent neural networks to process different modalities. In the testing phase, only single-modal data are needed.

- Design a training mode that can exploit unlabeled multimodal data to compensate for the limitation of labeled training data.
- Extend the availability of 2-D and 2.5-D/3-D multimodal datasets for urban remote sensing tasks.

The studies associated with this dissertation focus on two essential and critical tasks in the remote sensing field: **building extraction** and **building change detection**. In the following text, unless specified otherwise, 2-D, 2.5-D, and 3-D data refer to optical images, DSMs, and point clouds, respectively.

1.2 Dissertation Outline

This cumulative dissertation is based on the Ph.D’s research works in multimodal learning with VHR images and corresponding photogrammetric point clouds or DSMs, which are tested with tasks including building extraction and change detection. The contributions are mainly derived from four peer-reviewed journal papers, which are attached in the appendix:

- Yuxing Xie, Jiaojiao Tian, and Xiao Xiang Zhu. “Linking points with labels in 3D: A review of point cloud semantic segmentation.” *IEEE Geoscience and remote sensing magazine* 8.4 (2020): 38-59.
- Yuxing Xie, Jiaojiao Tian, and Xiao Xiang Zhu. “A co-learning method to utilize optical images and photogrammetric point clouds for building extraction.” *International Journal of Applied Earth Observation and Geoinformation* 116 (2023): 103165.
- Mario Fuentes Reyes*, Yuxing Xie*, Xiangtian Yuan*, Pablo d’Angelo, Franz Kurz, Daniele Cerra, and Jiaojiao Tian. “A 2D/3D multimodal data simulation approach with applications on urban semantic segmentation, building extraction and change detection.” *ISPRS Journal of Photogrammetry and Remote Sensing* 205 (2023): 74-97. (* equal contribution)
- Yuxing Xie, Xiangtian Yuan, Xiao Xiang Zhu, and Jiaojiao Tian. “Multimodal co-learning for building change detection: a domain adaptation framework using VHR images and digital surface models.” *IEEE Transactions on Geoscience and Remote Sensing* 62 (2024): 5402520.

In addition, parts of the content in this dissertation are based on the following two conference papers:

- Yuxing Xie, Konrad Schindler, Jiaojiao Tian, and Xiao Xiang Zhu. “Exploring Cross-City Semantic Segmentation of ALS Point Clouds.” *ISPRS-International Archives of the Photogrammetry, Remote Sensing and Spatial Information Sciences* 43 (2021): 247-254.

1 Introduction

- Yuxing Xie, and Jiaojiao Tian. “Multimodal Co-learning: A Domain Adaptation Method for Building Extraction from Optical Remote Sensing Imagery.” In *2023 Joint Urban Remote Sensing Event (JURSE)*, IEEE, (2023).

The remainder of this dissertation is organized as follows. Chapter 2 introduces the basics of remote sensing data of different dimensions, as well as deep learning methodologies to process them. Chapter 3 reviews related works on the topics of remote sensing data analysis with single-modal learning and 2-D and 2.5-D/3-D multimodal learning. Chapter 4 summarizes the Ph.D’s contributions with support from the aforementioned journal papers. Parts of the presented experiments are derived from another conference paper [27]. Chapter 5 concludes the dissertation and gives an outlook of potential future works.

2 Basics

2.1 Varied Dimensions in Remote Sensing Data

This section gives an overview of the optical remote sensing data with different dimensions, including 2-D multispectral imagery, 2.5-D DSMs, and 3-D point clouds. Figure 2.1 illustrates the comparisons in an example of these three modalities, which are derived from the same spaceborne WorldView-2 stereo image pair by the improved semi-global matching approach [28, 29].

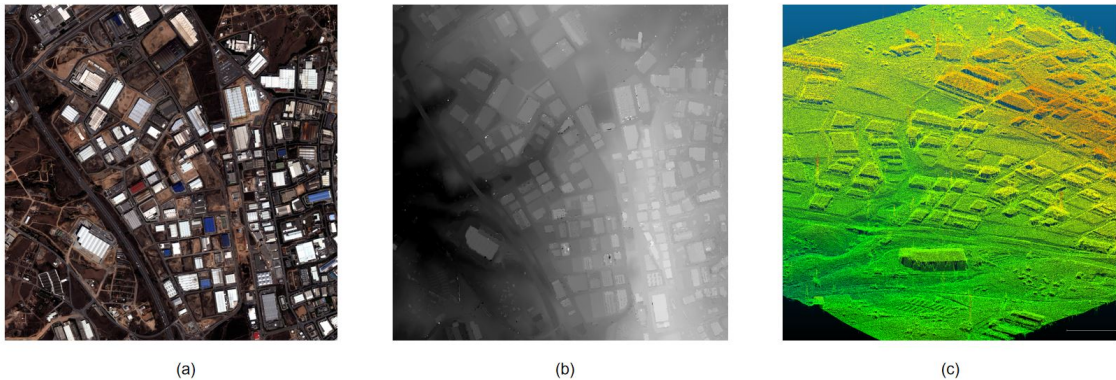


Figure 2.1: Three kinds of modalities derived from the same WorldView-2 stereo images. (a) RGB image. (b) DSM. (c) Point clouds.

2.1.1 Multispectral Imagery

Multispectral imagery is the most commonly utilized remote sensing data. Most multispectral images are acquired by the optical sensors measuring electromagnetic waves between 400 to 15,000 nanometers wavelength [30, 31], including visible light (e.g., red, green, and blue) and different infrared waves (e.g., near-, mid-, far-, and thermal infrared). Each multispectral image only consists of a small number (typically 3 - 15) of spectral bands. In remote sensing and earth observation tasks, multispectral sensors can be mounted at different mainstream platforms, including spaceborne, airborne, drone-based, as well as ground-based.

Raw images captured by the sensors are usually based on the perspective projection. Such images are distorted and therefore cannot be used directly in some real applications like geographic mapping. In such applications geometrically corrected true orthophotos are expected. Figure 2.2 compares perspective projection and orthographic projection. The perspective projection often leads to incorrect tilt in tall structures and inconsistent scales. In contrast, in an orthographic projection the nadir view for every pixel is calculated and the scale is corrected to uniform. Therefore, true distances can be measured directly in

the true orthophotos. To generate true orthophotos, an orthorectification process based on stereo-/multi-view images is necessary [32, 33].

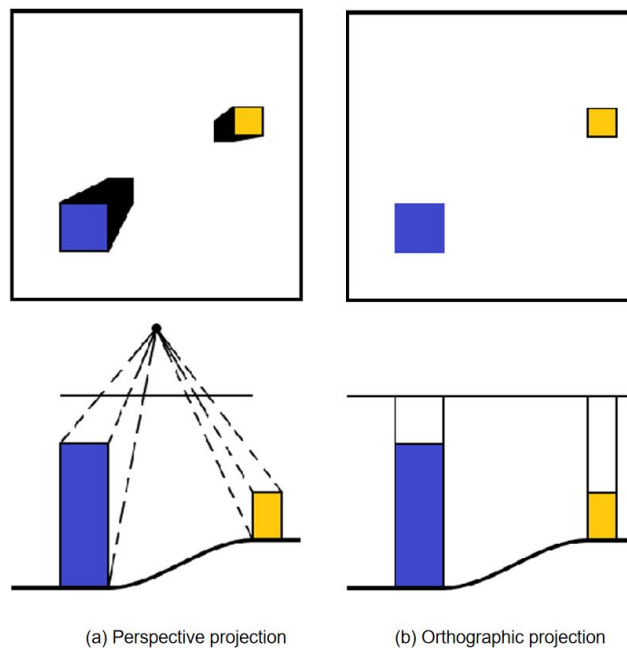


Figure 2.2: A comparison between perspective projection and orthographic projection.

Nowadays, benefiting from the development of multispectral sensors especially satellites like Pléiades-Neo [34], WorldView [35], and Gaofen [36] series, rapid-revisit HR and VHR multispectral images have become easily accessible. Multispectral data play a pivotal role in nearly every mainstream remote sensing task, such as building extraction [5, 37, 38, 39], semantic segmentation [40, 41], object detection [42, 43], vegetation monitoring [44], and so on. Moreover, when multitemporal images of the same location are available, recent intelligent algorithms have made it possible to realize the objective of change detection [7, 45].

2.1.2 Point Cloud

The point cloud is a kind of 3-D representation of an object or a space, which usually consists of at least millions of individual measurement points with x , y , and z coordinates. Some point clouds may also contain intensity or color information [15, 46]. Point clouds can be generated in several different ways. The most well-known way is to employ LiDAR sensors, actively measuring a huge amount of points in a scene. Another mainstream way is to generate point clouds from stereo-/multi-view images through photogrammetry techniques. Such point clouds are usually named imagery-derived point clouds [47, 27, 48] or photogrammetric point clouds [49, 50, 6]. In addition, as a developing way to generate point clouds, InSAR point clouds by TomoSAR or persistent scatterer interferometry (PSI) techniques provide a potential to obtain global data with meter-level accuracy [51, 52, 53, 54].

Point clouds can provide fundamental structures, so they are popularly used as the skeleton for 3-D modeling [55]. As 3-D geometric features derived from point clouds are good at distinguishing the shapes of diverse object types, they are also widely utilized with related intelligent algorithms in segmentation [56, 15], semantic segmentation (scene understanding) [15], and object detection [57] tasks.

2.1.3 DSM

The DSM is a 2.5-D digital product describing the ground surface and adjunct objects. It represents elevations in the regular grid cell (pixel) structure. DSMs can be created from airborne LiDAR point clouds by the direct rasterization operation or HR stereo images via dense image matching [29, 58]. Compared to 3-D point clouds, DSMs cannot provide façade information. As DSMs have the same structure as one-channel images, they can be processed by image-based algorithms. In real-world applications, it is commonly seen that fusing multispectral images and DSMs for building extraction, semantic segmentation, or change detection, as DSMs can better describe geometric features and increase the understanding of complex urban scenarios [59, 60, 61, 23]. If DSMs are converted to point cloud format with (x, y, z) structure, they can also be processed by point cloud algorithms such as 3-D convolutions. Figure 2.3 illustrates how a DSM is processed by a 2-D convolution and a 3-D convolution, respectively. In a 2-D convolutional operation, the elevation value z_i of the DSM is treated as a single-channel input and converted to a deep feature f_i . In contrast, during a 3-D convolutional operation, z is not directly used as input for the convolution. Instead, it serves as the index for the third dimension, analogous to how x and y index the first and second dimensions, respectively. The 3-D deep feature represents the geometric relationships between the voxel and its surrounding neighborhoods.

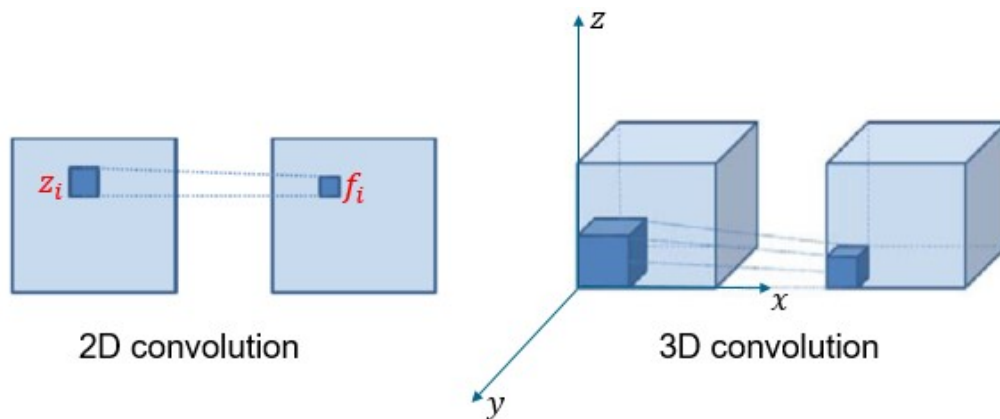


Figure 2.3: How 2-D convolution and 3-D convolution process the element of DSMs.

2.2 Deep Learning Methodology for Multidimensional Remote Sensing Data

Deep learning, originating from the artificial intelligence (AI) field, has revolutionized image processing including remote sensing data processing in past years [2, 15]. This

technique utilizes multi-layer neural networks with a huge amount of parameters to model the nonlinear relationships between input and output data. These parameters can be trained and optimized with annotated samples. When sufficient annotated training data are fed into neural networks, highly robust models can be obtained and achieve great performance on similar unlabeled target data. This section overviews mainstream deep neural networks applied to remote sensing data. According to the dimensions of the data, the networks introduced in this dissertation can be classified into two categories: image networks and point cloud networks.

2.2.1 Image Networks

As one of the most successful applications of deep learning, image networks significantly changed the field of computer vision and remote sensing. They have shifted the paradigm from relying on traditional handcrafted features to utilizing advanced neural architectures that automatically learn and extract features from image data, and achieved unprecedented breakthroughs on benchmark datasets of all kinds of tasks. By the time of writing this dissertation, mainstream image networks can be categorized into two groups: CNN-based and ViT-based, which are introduced in section 2.2.1.1 and section 2.2.1.2, respectively. Moreover, section 2.2.1.3 introduces the Siamese network architecture processing two groups of images with parallel branches, which is widely used in the task of change detection.

2.2.1.1 Convolutional Neural Networks

CNNs are characterized by their hierarchical stacks of convolutional layers and other essential layers such as pooling and normalization. In past years CNNs have made impressive achievements in image pattern recognition tasks [2, 62]. Many CNN-based backbones have been proposed for image feature extraction, such as VGG [63], DenseNet [64], ResNet [65], and MobileNet [66]. In this dissertation, ResNet is involved in several studies and is therefore subsequently introduced in detail.

ResNet, or residual network, is proposed in [65] to address the degradation problem commonly existing in deep learning methods at that time, which limits the successful training of deep neural networks. ResNet utilizes shortcut connections (also known as skip connections) to perform the identity mapping. Compared to plain networks, such operations can better stabilize the training and convergence and make it possible to optimize deep networks with over 1,000 layers. ResNet has different variants, based on their depth. Popular versions like ResNet-34, ResNet-50, ResNet-101, and ResNet-152 are extensively used due to their good balance between model size and accuracy. As a backbone, ResNet can be integrated into different deep learning architectures and has achieved state-of-the-art (SOTA) performance in multiple benchmark datasets across computer vision and remote sensing tasks.

Once deep features are extracted through backbones, it is crucial to organize and decode these features into the desired output format. Take semantic segmentation as an instance, where a semantic map of the same size as the original image is expected. To fulfill this goal, a specific architecture that bridges the features and the target output should be designed. Several significant works like FCN [67], U-Net [68], PSPNet [69], DeepLab series [70, 71], UPerNet [72] have introduced different architectures for this purpose. The following text introduces the U-Net to illustrate this concept. U-Net is an architecture

extensively employed in semantic segmentation tasks and also utilized in several studies presented in this dissertation.

U-Net is an architecture first developed for biomedical images [68], and soon gains popularity in other image domains, including natural imagery and remote sensing imagery. Its ability to retain multiscale features and capture contextual information makes it particularly effective for semantic segmentation. U-Net employs skip connections to concatenate the features in the encoder and decoder and therefore low-level and high-level features can be well preserved and utilized. The original U-Net encoder uses basic convolutional layers for feature extraction and refinement. However, with the development of more advanced backbones, many researchers have replaced these conventional convolutional layers with more powerful backbones like ResNet [65] and achieved enhanced performance [73, 3, 74].

2.2.1.2 Vision Transformer

Inspired by the success of transformer-based deep learning methods in the natural language processing field [13], recently some researchers have also introduced transformer-like architectures to vision tasks and achieved SOTA results on all kinds of benchmark datasets [11, 12]. A Transformer is a network based on the self-attention mechanism. In an attention layer, the input vector x is first transformed into three vectors: query, key, and value, denoted as q , k , and v , respectively. Different inputs are packed into a matrix X and then transformed into three matrices, Q , K , and V , respectively. The attention outputs of three matrices are calculated as:

$$Attention(Q, K, V) = softmax\left(\frac{QK^T}{\sqrt{d_k}}\right)V \quad (2.1)$$

where d_k is the dimension of the key vector k .

Instead of performing a vanilla single-head self-attention function, some works found it beneficial to utilize multi-head attention that concatenates the outputs from multiple parallel self-attention layers. This is because multi-head attention enables the model to simultaneously focus on information from various representation subspaces at different positions and produces more descriptive features [13].

In the original ViT architecture [11], the receptive field is of a fixed scale, which can be limiting to describing HR images in pixel-wise downstream tasks such as semantic segmentation. For an effective understanding of these images, a hierarchical feature representation is essential. To enhance the power of ViT, various studies have been conducted with hierarchical/multi-scale backbones [75, 12, 76]. Among them, Swin Transformer [12] has emerged as the most popular due to its greater accuracy and higher efficiency.

Swin Transformer adopts the design of shifted windows and patch merging to realize hierarchical feature computation. Generally speaking, in each stage of the network, a regular window partitioning operation is carried out, and self-attention is computed within each window.

As a backbone, Swin Transformer can be combined with various decoders for different tasks [77, 78, 79], even CNN decoders [9]. For example, in the open-source project mm-segmentation [80], Swin Transformer is employed with CNN-based UperNet for semantic segmentation. In the change detection network of [9], we integrate the Swin Transformer backbone into a U-Net architecture, of which the decoder is based on transposed convolutions and 1×1 convolutions.

However, two main issues are limiting the adoption of ViT [81]. One is that ViT-based models are generally huger than traditional CNNs, demanding much more computation resources. The other issue is that ViT-based models lack some inductive biases and require more training data to prevent overfitting. Therefore, ViT has not yet been able to replace CNNs in many light-weight applications.

2.2.1.3 Siamese Networks

A Siamese neural network is a type of architecture that employs two identical branches to process a pair of inputs. Here “identical” means that these branches have the same shape and share the same parameters. During the training phase, parameter updates are mirrored across both branches. In forward propagation, each input is handled by a separate branch, and the features extracted from two inputs are compared. This comparison could take the form of subtraction [7, 82] or concatenation [83, 82]. In the remote sensing field, Siamese network architecture is frequently used for analyzing bitemporal data and is particularly popular for the task of change detection. Siamese networks are not limited to specific backbones. Both CNN-based [82, 84] and ViT-based [78, 85] backbones can be well integrated into this architecture.

2.2.2 Point Cloud Networks

In past years, there also have been significant breakthroughs in point cloud analysis, with deep learning being introduced to this field. Depending on the data format ingested into neural networks, point cloud deep learning methods can be sorted into four categories: multiview image-based, point-based, voxel-based, and transformer-based. **Appendix A** involves a comprehensive introduction and comparison of the first three types of point cloud deep learning methods. Multiview image-based methods are with the idea that utilize 2-D image networks to process projected point clouds [86, 15]. However, these methods have notable limitations, including the need for complex preprocessing operations and the challenge of completely covering entire 3-D scenes. Consequently, their application has become infrequent. Similar to the trend in the image domain, transformer-based methods have also started to be used for point cloud processing recently [87, 88]. However, the studies involved in this dissertation do not cover them. Thus, the subsequent sections will briefly delve into the basic theory of point-based and voxel-based semantic segmentation networks, which have been mainstream and are pertinent to the research of this dissertation.

2.2.2.1 Voxel-based Semantic Segmentation Networks

Typical convolutions for raster images cannot directly be applied to disordered and unstructured point clouds. To adapt point clouds to the structured nature of convolutions, apart from multiview image-based methods, voxel-based methods emerged as another early solution in point cloud deep learning. These methods first convert raw point clouds into the voxel format and then employ 3-D convolutions for processing [15]. While early voxel-based approaches like SEGCloud [89], OctNet [90], and O-CNN [91] have been applied to point cloud semantic segmentation tasks, they exhibit a notable limitation. These methods lack an efficient way to handle a large number of voxels that represent empty spaces, leading to high computational and memory requirements. Subsequent advancements in

sparse convolution approaches effectively resolved this issue [92, 19]. Sparse convolutions utilize arbitrary shapes instead of dense shapes. Arbitrary kernels only take the voxels occupying kernel shapes into the convolution calculation. If the input vector is $\mathbf{x}_{\mathbf{u}}^{in} \in \mathbb{R}^{N^{in}}$ and the existing offset voxels covered by the sparse convolution are defined as $\mathbf{u} \in \mathbb{Z}^D$, the output feature vector by a sparse convolution calculation is calculated as [19]:

$$\mathbf{x}_{\mathbf{u}}^{out} = \sum_{\mathbf{i} \in \mathcal{N}^D(\mathbf{u}, C^{in})} W_{\mathbf{i}} \mathbf{x}_{\mathbf{u}+\mathbf{i}}^{in}, \quad (2.2)$$

where $\mathcal{N}^D(\mathbf{u}, C^{in}) = \{\mathbf{i} | \mathbf{u} + \mathbf{i} \in C^{in}, \mathbf{i} \in \mathcal{N}^D\}$. C^{in} is the predefined sparse tensors to be convolved. In the case of a point cloud, $D = 3$.

2.2.2.2 Point-based Semantic Segmentation Networks

Applying point-based networks is another mainstream solution to avoid information loss of point clouds in 2-D projected data formats. Point-based networks directly adopt raw points as the input. As a pioneering method, PointNet [93] uses a symmetric function to solve the ordering problem of point clouds:

$$f(\{x_1, \dots, x_n\}) \approx g(h(x_1), \dots, h(x_n)), \quad (2.3)$$

where $f: 2^{\mathbb{R}^N} \rightarrow \mathbb{R}$ and $h: \mathbb{R}^N \rightarrow \mathbb{R}^K$. $g: \underbrace{\mathbb{R}^K \times \dots \times \mathbb{R}^K}_n \rightarrow \mathbb{R}$ is a symmetric function.

In the implementation, the feature extraction operation h is realized by the multilayer perceptron (MLP). The symmetric function inside g can be max pooling, average pooling, or attention-based weighted sum [94]. Among them, max pooling is the most popular choice due to its best performance on the benchmark [93].

PointNet can capture global features but is weak at acquiring local features. To address this limitation, [16] proposed PointNet++, a hierarchical neural network extended from the basic PointNet model. PointNet++ is designed to capture multiscale local features within neighborhoods, enabling it to construct a more robust feature representation. This enhancement improves the model’s performance on semantic segmentation tasks across various scales. Inspired by the success of PointNet and PointNet++, many subsequent works adopted them as backbones and developed different modules to enhance them [95, 17, 96]. The summary of those mainstream PointNet-based methods can be found in **Appendix A**.

Leaning towards the application of convolutions, some researchers have designed continuous convolutions specifically to address the unordered and unstructured nature of point clouds. Contributions like PointCNN [97], KPConv [18], and PointConv [98] have provided alternative efficient backbones for point clouds, and achieved impressive performance on semantic segmentation tasks.

3 State of the Art

This dissertation is dedicated to developing novel multimodal co-learning frameworks that combine 2-D multispectral imagery and 2.5-D/3-D photogrammetric data, utilizing unlabeled data and mutual information to improve the performance of 2-D and 3-D neural networks for building extraction and building change detection. This chapter reviews related studies on these two remote sensing tasks, as well as on multimodal learning works involving 2-D and 2.5-D/3-D data.

3.1 2-D Multispectral Imagery Analysis with Deep Learning

3.1.1 Building Extraction

Building extraction is a binary segmentation task that separates pixels of remote sensing images into building areas and non-building backgrounds. This task is crucial and essential for many urban applications, such as urban monitoring [99], urban planning [100], urban energy modeling [101], and digital cadastral mapping [102]. In recent years, deep learning-based methods have taken the place of traditional algorithms and have become the most popular due to their superior performance [2]. In principle, every semantic segmentation network can also be utilized for building extraction. For example, encoder-decoder FCN [67], SegNet [103], U-Net [68], and multiscale HR-Net [104] are widely used in early building extraction works. These vanilla networks are easy to be restricted by their shallow local receptive field and limited training data. As a result, they could lose the spatial details especially building boundaries in the building extraction task, and might be hard to describe the features of discriminating buildings with different styles and varying sizes or scales [38, 39].

To overcome these issues, several enhanced methods are proposed in subsequent studies. For instance, considering the building scale problem, [105] develops a Siamese architecture that can utilize not only original training data but also their downsampled counterparts in the training phase, which enhances the diversity of the training samples and improves the generalization ability of the network. Attempting to address the scale issue as well, MHA-Net [106] utilizes a multipath hybrid dilated convolution module to fuse multiscale contextual information. This module merges multiscale contextual information, augmenting the network's ability to capture features from buildings of various sizes. Similarly, MAP-Net [38] adopts a channel-wise attention module to optimize the fused features from multiscale paths, which enhances the building representation by well combining the contextual information from different scales. BOMSC-Net [107] considers multiscale context by a parallel graph reasoning module that combines features of different scales. To improve the spatial details of buildings, [108] designed a U-Net architecture with Res2Net [108] as the backbone for building extraction, namely Res2-Unet. This network is trained with a boundary loss function to optimize the network in maintaining more details of building shapes. Trying to optimize building boundaries as well, [40] involves a holistically nested edge detection module for edge extraction and a boundary enhancement

module to combine the edge and building masks. These two modules can be flexibly inserted into different encoder-decoder architectures and effectively improve the networks' performance on building boundaries. [109] proposes a method conducting a dual-branch architecture for the training and introducing the difference of Gaussian (DoG) operator as a constraint, which augments the network's sensitivity to edge information. In [37], a feature pairwise conditional random field is employed to preserve sharp building boundaries, providing a flexible strategy for refining the results produced by CNN models. Moreover, [110] employs an attraction field representation method to optimize building boundaries. This boundary-aware attraction field can be inserted into CNN models and suppresses the influence of background. Inspired by the spatial correlation advantage of graph convolutional network (GCN), [5] proposed a method combining a GCN with structured feature embedding, which demonstrates superior building extraction performance compared with many CNN-only backbones.

With the rise of transformer networks, recently more and more studies have introduced transformer architecture into the building extraction task. For example, [111] applies Swin Transformer [12] and SegFormer [76] for cross-area building extraction. Avoiding the high computational costs and data demands of transformer networks, more researchers choose a balanced way that combines transformer and CNN. For instance, Easy-Net [112] adopts transformer blocks to fuse and refine raw features from a multiscale CNN, augmenting the building extraction performance in a lightweight way. Considering local spatial details may not be maintained by ViT, [113, 114, 115] adopt similar dual-path networks that simultaneously utilize a CNN-based and a transformer-based encoder for feature extraction. Output features from two paths are fused and refined by subsequent decoding operations.

More recently, several researchers have noticed the problem of large amounts of imperfect data in real-world applications and shifted the focus from fully supervised learning towards semi-supervised or weakly supervised learning. For example, as instances of semi-supervised learning, [116, 117] leverage unlabeled data in the training phase through consistency learning modules, which enhance the generalization capability of networks. To overcome the limitations of weak annotations, ALNet [118] employs a feature-to-image restoration branch as an auxiliary task to acquire additional pixel-wise supervisory information from available labels. [119] explores the utilization of scribble labels with a structure-aware scribble module and an edge-aware loss function.

Another trend in building extraction studies is the growing interest in cross-domain scenarios. In real-world applications, the diversity of most unlabeled regions cannot be fully represented by limited annotated data, leading to widespread domain gaps across different datasets. Domain adaptation methods, aimed at leveraging the available labeled data more effectively and training more generalizable networks, are being increasingly introduced into the building extraction task. For example, [120] explores cross-domain building extraction with a series of SOTA mainstream networks and concludes that fusing the probabilities achieved by different methods can improve the results on unseen target data. [121] designs a full-level domain adaptation framework consisting of an image alignment preprocessing method, an adversarial learning module, a mean-teacher model, as well as a self-training postprocessing step. This framework considers potential domain gaps in every stage and provides effective solutions. The effectiveness of these methods has been verified on three public datasets from different countries and with different resolutions. [122] utilizes a cross-geolocation attention module to improve the generalization ability of

the CNN for building extraction. In the proposed Siamese architecture, information from both source data and target data can be learned by the network.

3.1.2 Change Detection

Change detection is another fundamental yet challenging task in remote sensing applications, crucial for urban planning, disaster assessment, map updating, and more [123, 124, 125]. The goal of this task is to identify differences in a specific area across various time frames [45]. According to the type of labels, change detection can be classified into binary change detection [82, 7] and semantic change detection [126, 127]. In this dissertation, we focus on the former. Unless specified otherwise the follow-up “change detection” in this dissertation refers to binary change detection. The advent of deep learning in past years has brought unprecedented solutions to change detection using multispectral images, elevating performance on public benchmarks to new heights. Just as with building extraction, early deep learning-based 2-D change detection methods predominantly adopt CNNs [82, 84], but more recently, transformer networks have gained prominence [78, 85].

As illustrated in Figure 3.1, change detection architectures can be categorized into two main types: single-branch and dual-branch. The single-branch architecture embodies an early fusion approach, merging bitemporal inputs before introducing them into the network. For example, [82] presents a U-Net-based CNN architecture with an early fusion operation, which concatenates pre- and post-event images as different color channels. Similarly, CD-Net [128] employs the concept of stacking contraction and expansion blocks, and bitemporal images are concatenated to a single input along the channel dimension and then fed into a single-branch network for change detection. Also based on concatenated image pairs, [129] introduces an improved UNet++ architecture capable of learning multiscale features from concatenated channels for change detection tasks.

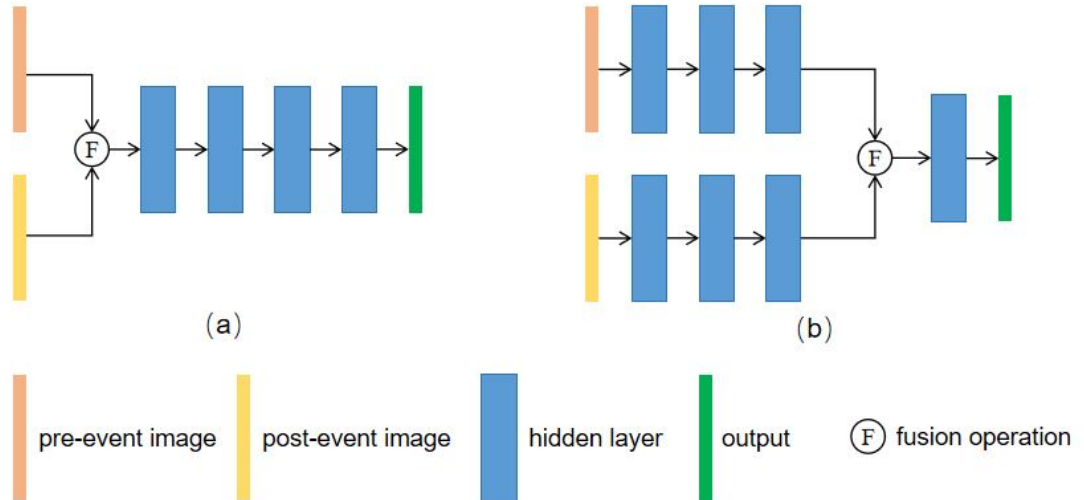


Figure 3.1: Change detection architectures. (a) Single-branch architecture. (b) Dual-branch architecture.

Dual-branch architectures can be classified into two categories: Siamese network introduced in section 2.2.1.3 and pseudo-Siamese network. The distinction between them hinges

on whether they share weights across branches or not [130]. Siamese networks, characterized by shared weights, boast fewer parameters and tend to converge more rapidly than equivalent architectures without weight sharing. Furthermore, several studies demonstrate that their performance on homogeneous data is comparable [45, 130]. Consequently, for research involving homogeneous data, the preference typically leans towards non-pseudo-Siamese networks. For example, [82] presents two end-to-end Siamese change detection networks, FC-Siam-conc and FC-Siam-diff, with different fusion strategies. Having proved the effectiveness of Siamese architecture, subsequent studies focus on enhancing the networks' feature discriminative ability with various strategies. These include multiscale feature fusion [131, 84, 132], optimizing invariant features using generative adversarial networks (GANs) [133, 134], and increasing the availability of contextual information [135, 83, 136]. For heterogeneous data, the pseudo-Siamese structure holds an advantage in that each branch can be specifically optimized to be more suitable for its respective type of data [45, 137].

Owing to their ability to capture global information, transformer-based networks are being increasingly employed in 2-D change detection tasks. As an early experiment utilizing transformer block, [7] introduced a bitemporal image transformer (BIT) module to model the long-range context of the bitemporal images. Subsequently, purely transformer-based models like ChangeFormer [85] emerged, delivering superior results on public benchmarks. Given the computational cost and local feature limitation of pure transformer networks, some of the latest studies have implemented hybrid architectures. These architectures combine both convolutional and transformer blocks, effectively leveraging the power of both network types [138, 8, 139].

3.2 2.5-D/3-D Remote Sensing Data Analysis with Deep Learning

In the remote sensing field, most scene-level point cloud analysis concentrates on semantic segmentation [15]. Early works like [140, 141] project point clouds to images and utilize 2-D CNNs for semantic segmentation. With the rise of more convenient and powerful 3-D point cloud networks, image-based methods are no longer the mainstream. Later on, inspired by the success of PointNet [93] and PointNet++ [16], a series of later remote sensing studies utilize them as feature descriptors. For example, [142] extends the base of PointNet to a multiscale approach and successfully applies it to large-scale airborne laser scanning (ALS) datasets. To optimize the features extracted from PointNet++, [143] proposes a framework utilizing a manifold learning-based algorithm for multiscale feature embedding and a graph-cut method to refine the semantic segmentation result. To capture more accurate local relationships, GraNet [144] introduces local spatial geometric learning modules that consider the orientation information, spatial distribution, and elevation information.

With the development of point convolution [18] and sparse convolution [19, 92] in computer vision, an increasing number of researchers have started to use them as backbones due to their efficiency and higher accuracy compared to PointNet-based methods. For example, [145] extends KPConv [18] with a recurrent residual dual attention mechanism, which can improve the diversity of local features and refine the semantic segmentation results on ALS point clouds. Aiming at enhancing local features as well, [146] designs a hybrid block combining KPConv [18] and 2-D point convolutions. [147] conducts a regres-

sion network [148] with sparse convolution of Minkowski Engine [19], achieving superior results compared with PointNet and KPconv.

As a novel topic, deep learning-based point cloud change detection has attracted attention recently [149, 150]. As a pioneer in this field, [151] brings the idea of Siamese architecture to 3-D space and proposes a Siamese KPConv network for building change detection from point clouds, achieving superior accuracy in comparison with conventional method random forest. Building upon this, [150] extends the scope from building change detection to multiclass change detection, showing remarkable performance on real datasets. Additionally, [152] presents a semantic-supported change detection framework. Initially, this framework acquires semantic segmentation results of bitemporal point clouds using an enhanced multiscale network based on PointNet++. Then these semantic results are processed by a semantic-supported change detection module to compute the final change detection results.

Similar to the dilemma in image modality, domain gaps also restrict the effective use of point cloud networks for broader applications across datasets from varying regions and with different urban styles [153, 154, 155]. Our previous study [153] on cross-city semantic segmentation explores the disparities between two ALS point cloud datasets from Germany and China, revealing a pronounced domain gap that remains unresolved by a point cloud network and an unsupervised alignment method. Especially, the minority class poses a considerable challenge. To better apply point cloud networks in large-scale real-world scenarios, there is a pressing need for effective methods that leverage limited labeled training data to boost the generalization capabilities of these networks [15].

As mentioned in section 2.1.3, the 2.5-D DSMs can be analyzed either as image rasters or point clouds within the realm of deep learning research. In most remote sensing studies, DSMs are processed through an image network branch, and output features or probabilities are fused with those derived from multispectral images. These fusion operations are aimed at utilizing multimodal information to improve the results of tasks such as semantic segmentation [156], building extraction [157, 23], and change detection [158]. They will be introduced with more details in section 3.3. When processing cross-domain DSMs with varying elevation ranges, 3-D point cloud networks are superior options [27]. This preference stems from the fact that directly using absolute elevations in 2-D networks may result in a significant domain shift, potentially leading to poor results. 2-D networks are better suited for processing nDSMs [159]. In contrast, 3-D point cloud networks can utilize relative, as opposed to absolute, coordinates. Such approaches enable a more natural representation of the ground objects' features and are employed in our studies [27, 6]. For the task of change detection, height difference maps are frequently employed [60, 59]. These maps are generated through straightforward subtraction or window-based averaging strategies and serve as indicators of potential changes. The process of change detection from height difference maps essentially constitutes a pixel-wise semantic segmentation task. In our latest study [9], a hybrid multiscale network combining Swin Transformer [12] and CNNs is proposed for detecting building changes from height difference maps.

3.3 Multimodal Learning with 2-D and 2.5-D/3-D Remote Sensing Data

Despite the development of numerous methods for 2-D, and 2.5-D/3-D remote sensing data, challenges inherent to each single-modal data type persist. Multispectral images are

susceptible to light reflection, shadows, and blockages [60, 156]. Additionally, networks reliant on spectral features may struggle to distinguish between objects of similar colors [6]. Spectral influence issues are particularly pronounced in change detection tasks, where pseudo-changes such as seasonal shadows, vegetation changes, and varying lighting conditions pose significant challenges for 2-D change detection networks [130, 60]. To better address these issues, seasonal-invariant features are expected [133, 134, 45]. Regarding 2.5-D/3-D data, while they are unaffected by changes in shadows or lighting, they encounter limitations such as the potential for missing structures, blurred boundaries, and, as a result, incomplete object extraction results [59]. Moreover, distinct objects sharing similar geometric characteristics may also confuse 3-D networks. Considering that multispectral images and point cloud/DSM data can provide information that is complementary to each other, some researchers have shifted their attention from single-modal methodologies to multimodal learning and achieved notable improvement on various remote sensing tasks. Related studies can be categorized into two classes: data fusion [20] and knowledge transfer [6, 27, 9]. Multimodal learning has contributed to various remote sensing tasks, including building extraction [23, 6], semantic segmentation [160, 22], and change detection [158, 9].

Data fusion combines multimodal data or embedded information during forward propagation, based on the intuition that multimodal information can outperform single-modal data [161, 20]. According to the stages where the fusion is carried out, data fusion methodologies can be classified into early fusion (or observation-level fusion), middle fusion (or feature-level fusion), late fusion (or decision-level fusion), and hybrid approaches [20, 162, 6]. Early fusion occurs at the data input phase. In remote sensing studies, 2-D multispectral images are commonly concatenated with height maps of DSMs or nDSMs as multi-channel inputs for 2-D networks. For example, [163] merges multispectral images and nDSMs as the input for building extraction. In another type of early fusion instance, spectral values from multispectral images are projected into 3-D space and appended to point clouds as per-point attributes, generating colorized point clouds for 3-D point cloud networks. However, the efficacy of such 3-D early fusion techniques remains uncertain [26]. Some studies even indicate colorized data bring potential negative impacts when compared to utilizing raw point clouds [24, 6, 25].

Middle fusion takes place during the intermediate phase of a deep learning model. It merges feature tensors of multiple modalities into a unified composite one. This composite feature is subsequently processed by later stages of the model. For example, [159] presents a CNN-based multimodal feature fusion architecture for RGB images, panchromatic images, and nDSMs. Experiments on building extraction demonstrate that multimodal features can improve the generalization performance on unseen datasets. CMGFNet [23] fuses image-branch and DSM-branch features using a gated fusion module, which produces discriminative information that can improve the building extraction results. Similarly, employing feature-level fusion, DSPCANet [160] develops a channel attention module and an improved position attention module to refine the fused features of multispectral images and DSMs, leading to improved semantic segmentation results.

Late fusion occurs at the decision-making phase of a deep learning model. It merges probability maps of distinct modalities. For example, [158] introduces a multimodal change detection architecture that utilizes two-stage decision fusion operations. The initial fusion step merges the change probabilities derived from both image difference and DSM difference. A subsequent fusion operation then integrates these merged probabilities with building extraction probabilities generated by a CNN. To maximize the benefits of dif-

ferent fusion strategies, some recent studies have combined multiple fusion techniques in their architectures, offering innovative solutions for building extraction [157] and semantic segmentation [164].

Knowledge transfer diverges from direct operations on data or extracted features. As summarized in our study **Appendix D**, knowledge transfer methods follow two main principles: (1) the use of distinct network branches for varying data modalities, and (2) the establishment of connections between different modalities through soft connections, typically implemented as loss functions. This allows single-modal networks to influence each other during the training phase while enabling them to operate independently for testing single-modal data. Unlike data fusion strategies, knowledge transfer offers greater flexibility, making it better suited for a variety of scenarios, including those where certain modalities may be absent during testing. Moreover, a notable drawback of data fusion is its inefficient harnessing of information present in raw heterogeneous data and the inherent complementary nature of multimodalities, potentially leading to inaccurate and irrelevant feature representations [23, 22]. In contrast, knowledge transfer approaches employ separate networks for processing distinct modalities, which effectively preserves the completeness of heterogeneous information and minimizes the interference of noise from other modalities. As a prominent example of knowledge transfer, co-learning approaches that integrate 2-D and 2.5-D/3-D data have been developed and verified on several remote sensing tasks. As a pioneer of multimodal co-learning in remote sensing, this dissertation involves our three case studies on this subject. They are published as **Appendix B**, **Appendix D**, and [27]. Our work **Appendix B** presents a multimodal building extraction framework for 2-D images and photogrammetric point clouds, which demonstrates the effectiveness of standard and enhanced co-learning methods that leverage a minimal amount of labeled data pairs alongside a substantial volume of unlabeled data pairs. In [27] this building extraction co-learning framework is extended for cross-domain tasks. Recently, our study **Appendix D** has developed a co-learning framework for building change detection with VHR multispectral images and DSMs, significantly enhancing the performance of single-modal networks on cross-domain datasets. A summary of these studies will be provided in Chapter 4.

4 Summary of the Work

4.1 What is Multimodal Co-learning?

Multimodal co-learning refers to strategies that transfer knowledge between two or more modalities, their representation, and their predictive probabilities, which enhance the modeling of a modality by exploiting the hidden information from another modality [161, 165]. These approaches can be implemented in either a supervised or semi-supervised manner. In our study **Appendix B** [6], we name the fully supervised version as *standard co-learning*, which exclusively employs the labeled data for training. The semi-supervised variant is named *enhanced co-learning*, which leverages not only labeled data but also unlabeled multimodal pairs during the training phase.

Figure 4.1 illustrates the difference between conventional data fusion techniques and a naïve standard co-learning structure transferring knowledge via a loss function between probabilities. Such a co-learning method addresses the two challenges mentioned in section 1.1. Unlike the dual-input-and-single-output architectures of data fusion approaches, co-learning independently trains two single-modal networks. This operation is particularly advantageous for real-world scenarios where a modality might be missing or multimodal data are not aligned during the testing phase. In addition, by utilizing single-modal networks separately, the unique characteristics of each modality are preserved. Therefore, the negative impacts of data fusion, particularly concerning colorized point clouds, can be eliminated [6].

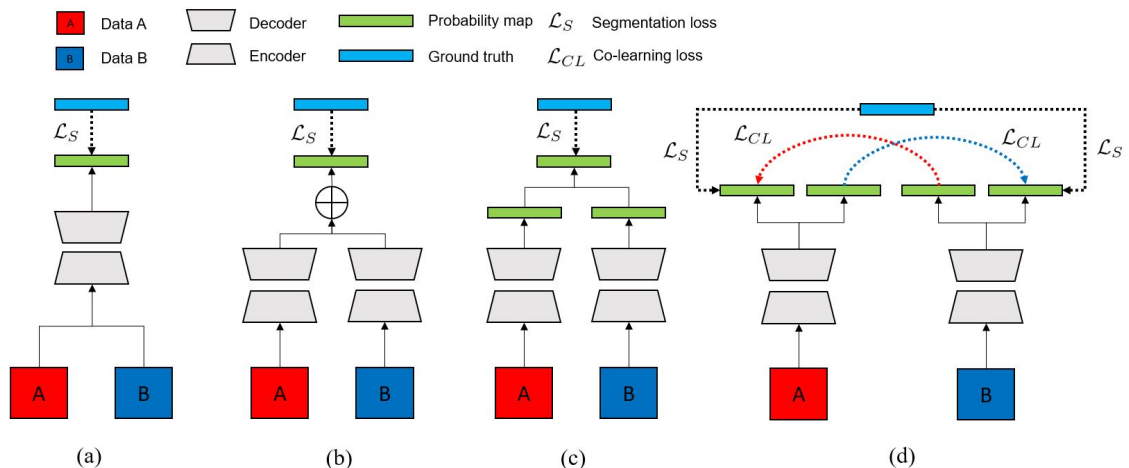


Figure 4.1: The difference between conventional data fusion and co-learning. (a) Early fusion. (b) Middle fusion. (c) Late fusion. (d) Multimodal co-learning.

In remote sensing applications, orthophotos along with their corresponding photogrammetric point clouds or DSMs are ideally suited for use within multimodal learning frameworks due to their perfect alignment. Through our comprehensive investigation of various

multimodal co-learning approaches utilizing this data combination, we have demonstrated their efficacy in tasks including building extraction and building change detection from VHR multispectral images and corresponding photogrammetric point clouds or DSMs [6, 27, 9]. Drawing on insights from our recent studies, the following content of this chapter is structured into two sections. Section 4.2 introduces the multimodal datasets utilized in our experiments. Section 4.3 presents three case studies with multimodal co-learning:

- The first case study focuses on the building extraction task, addressing the challenge of training with limited labeled orthophotos and photogrammetric point clouds.
- The second case study also targets the building extraction task, but it tackles the domain gap issue that the source data and the target data are from different datasets.
- The last case study centers on the building change detection task, similarly dealing with the challenge of domain gaps between diverse datasets.

4.2 Multimodal Datasets Involved in This Dissertation

This section summarizes the multimodal datasets employed in our works [6, 27, 9, 155], including ISPRS Potsdam benchmark¹ and Munich Munich WorldView-2 dataset for building extraction, Istanbul WorldView-2 dataset for building change detection, as well as the newly released synthetic benchmark SMARS [155] for both building extraction and building change detection. SAMRS benchmark dataset has been published as **Appendix C** to support this dissertation.

4.2.1 Building Extraction Datasets

4.2.1.1 ISPRS Potsdam Dataset

ISPRS Potsdam dataset¹ is a benchmark published by the International Society for Photogrammetry and Remote Sensing (ISPRS). It is designed for the semantic segmentation task but is also widely used as a benchmark to evaluate building extraction methodologies [5, 37]. The original ISPRS Potsdam dataset consists of 38 airborne multispectral orthophotos with red, green, blue, and infrared channels and matched dense-image-matching DSMs. Each orthophoto and DSM has a pixel size of 6000×6000 and a ground sampling distance (GSD) of 5 cm.

4.2.1.2 Munich WorldView-2 Dataset

The Munich WorldView-2 is a private dataset collecting WorldView-2 satellite imagery and matched point clouds over the city center of Munich, Germany. The image part of this dataset is orthophotos with RGB channels. The point cloud part of this dataset is unrasterized colorless point clouds generated from a pair of stereo WorldView-2 panchromatic images with an improved semi-global matching approach [28, 29]. The GSD of the RGB orthophotos is 0.5 m. As illustrated in Figure 4.2, the whole region of this dataset is divided into 6 areas. A1, A2, and A3 are set as the training data, each with a size of 6000×6000 pixels. A5 and A6, with a size of 6000×3200 pixels, are used as the validation set. The test dataset is A4, also with a size of 6000×6000 pixels. The building masks

¹<https://www.isprs.org/education/benchmarks/UrbanSemLab/2d-sem-label-potsdam.aspx>

(i.e., ground truths) of orthophotos are annotated manually, while the ground truths of point clouds are projected from the building masks via an affine transformation.

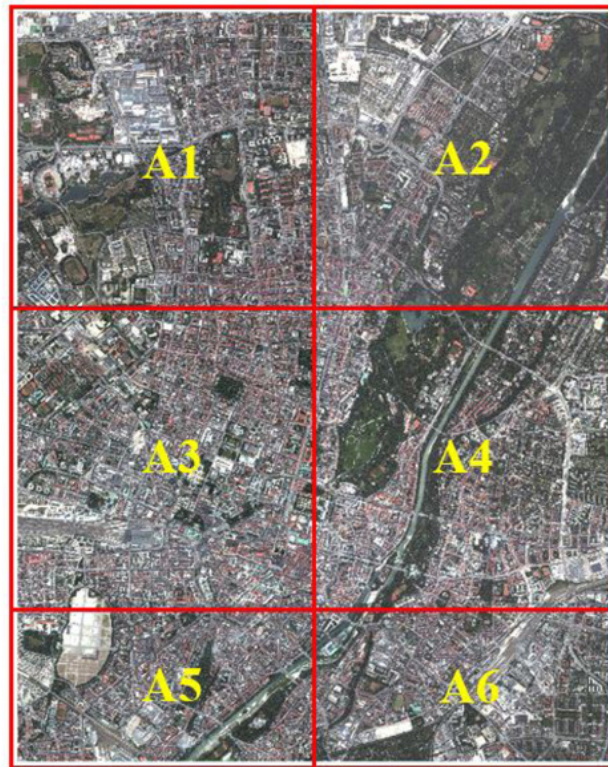


Figure 4.2: The coverage of the Munich WorldView-2 dataset used in this dissertation.

4.2.1.3 Simulated Multimodal Aerial Remote Sensing (SMARS) Dataset

The SMARS dataset² serves as a public benchmark created by the German Aerospace Center (DLR) and ISPRS. It is a synthetic dataset crafted for multimodal urban remote sensing tasks, including semantic segmentation, building extraction, and building change detection. We created this dataset due to the concern that real-world multimodal urban remote sensing datasets are difficult to obtain and current benchmarks cannot satisfy the demands of evaluating increasing 2-D and 2.5-D/3-D multimodal algorithms. For the details of generating and rendering this dataset with Blender³ and CityEngine⁴, please refer to **Appendix C**.

SMARS features two urban sub-datasets, each styled differently. One, named Synthetic Paris (SParis), simulates the architectural essence of Paris, France, while the other, named Synthetic Venice (SVenice), simulates the style of Venice, Italy. Each sub-dataset is comprised of bitemporal orthophotos, DSMs, semantic maps, building masks, and building change maps, as shown in Figure 4.3. Their data have been rendered at two different GSDs of 30 cm and 50 cm and both have been validated in [155] as a reliable benchmark for the training and evaluation of algorithms. In the building extraction experiments

²https://www2.isprs.org/commissions/comm1/wg8/benchmark_smars/

³<https://www.blender.org/>

⁴<https://www.esri.com/en-us/arcgis/products/arcgis-cityengine/overview/>

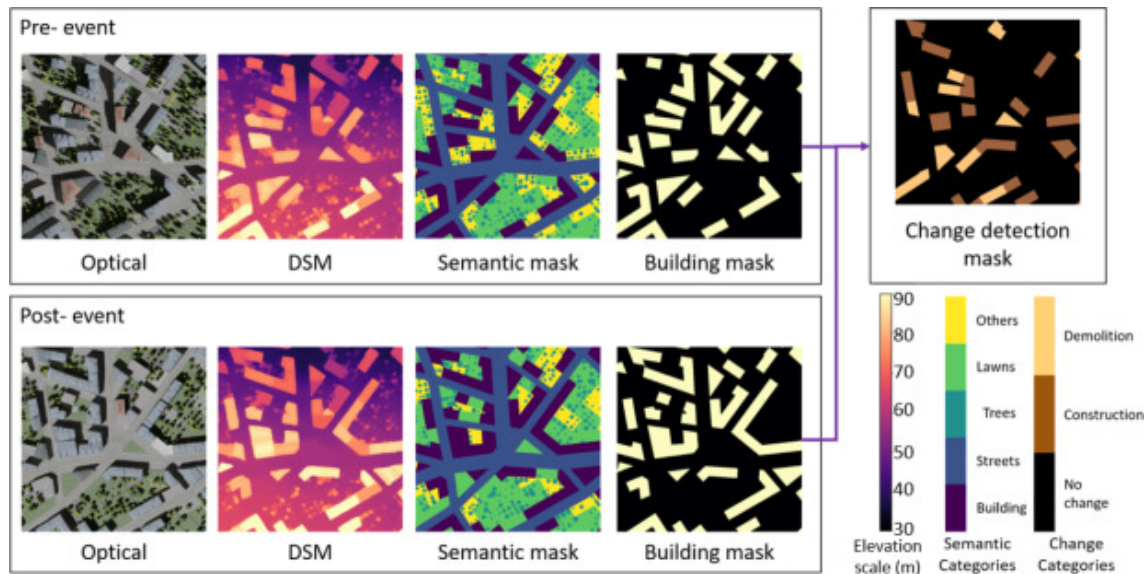


Figure 4.3: Available data types in SMARS dataset. Scales are given as a reference for the displayed information.

presented in section 4.3.2, the 30 cm resolution version is utilized as the experimental data.

4.2.2 Change Detection Datasets

4.2.2.1 SMARS Dataset

The essential information of SMARS dataset has been introduced in 4.2.1.3. The building change masks in this dataset are generated via the difference between pre-event and post-event building masks. Therefore, two versions of change masks are available. One is 3-class change masks with the labels of construction, demolition, or no change. The other is binary change masks with labels of change or no change. In the related study involved in this dissertation, the focus is binary change detection.

4.2.2.2 Istanbul WorldView-2 Dataset

The Istanbul WorldView-2 is a private dataset for building change detection between the years 2011 and 2012. It contains bitemporal RGB orthophotos, bitemporal photogrammetric DSMs, and corresponding building change masks. The orthophotos and DSMs have a GSD of 50 cm. They are generated from stereo WorldView-2 satellite images by the improved semi-global matching approach [29, 28]. This dataset covers two regions of Istanbul, Türkiye, as shown in Figure 4.4. Images and DSMs in Region I have a pixel size of 4692×3435 , while those in Region II have a pixel size of 1964×1245 .

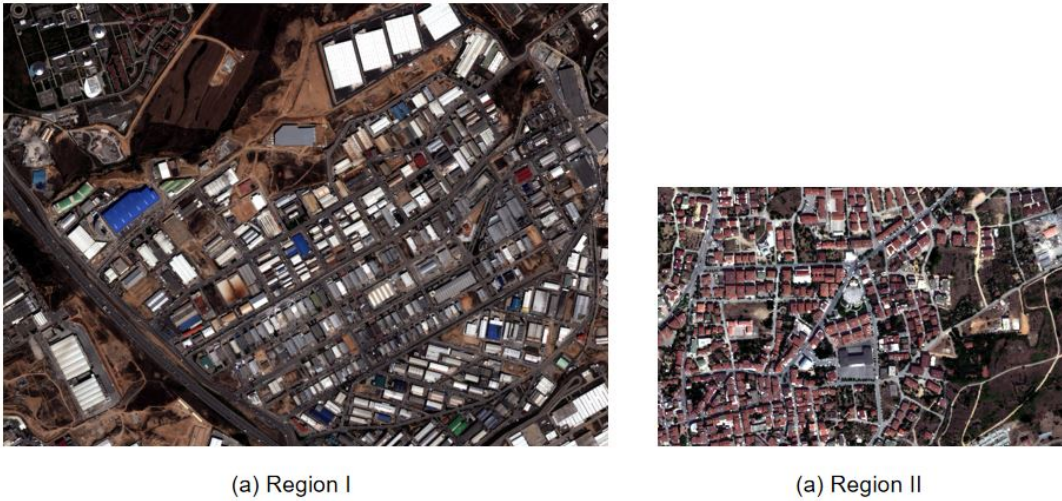


Figure 4.4: The coverage of the Istanbul WorldView-2 dataset used in this dissertation.

4.3 Case Studies with 2-D and 2.5-D/3-D Multimodal Learning

4.3.1 Case I: Multimodal Co-learning Enhances the Building Extraction Networks with Limited Labeled Data

This section gives a brief overview of the published journal paper **Appendix B**, presenting how multimodal co-learning enhances the image and point cloud building extraction networks with limited labeled data.

4.3.1.1 Background

Despite the big success that deep learning techniques have brought to automatic building extraction, several challenges still hinder its effective and practical deployment in real-world applications. The industry is eager for higher accuracy and more flexible data utilization. A common issue arises when deep neural networks are trained with limited samples, leading to overfitting and poor generalization to new, unseen data. To align with more practical scenarios, there is a pressing need for enhanced accuracy with reduced reliance on extensive annotated datasets.

In this section, we explore the capabilities of multimodal co-learning for the task of building extraction, utilizing multispectral orthophotos and corresponding photogrammetric point clouds. This case study operates with a minimal set of labeled training data, consisting of only 10 pairs of samples. The improvement of single-modal networks is achieved through the exchange of mutual information between 2-D and 3-D modalities. We introduce a flexible co-learning framework capable of concurrently training both an image network and a point cloud network. This approach is particularly designed to accommodate scenarios where one modality may be absent during testing, demonstrating the framework’s adaptability to varying data availability.

4.3.1.2 Methodology

The comprehensive flowchart of our proposed co-learning-based network architecture is depicted in Figure 4.5. As it illustrates, we apply the knowledge transfer between two modalities through the use of probability maps. The intuition behind this strategy is that improved outputs will lead to a reduced dissimilarity in predictions between the two modalities. To harness this, we introduce a co-learning loss function designed to minimize the similarity in the predictions made by the 2-D image network and the 3-D point cloud network. Consequently, during the training phase, two loss functions are optimized: a supervised loss function aimed at building semantic extraction, and an unsupervised co-learning loss function that quantifies the discrepancies between the 2-D and 3-D predictions. In our framework, we adopt a U-Net [68] with ResNet-34 backbones [65] as the image network, and a sparse convolutional network (SparseConvNet) [92] as the point cloud network. Each network generates two kinds of probability maps. The first, referred to as the *predicted probability*, is engaged in the loss functions to the same modality network, affecting its backward propagation. The second type, *termed the shadow reference probability*, is distinct from actual ground truth data and is employed by the network of the other modality as the reference in the co-learning loss function.

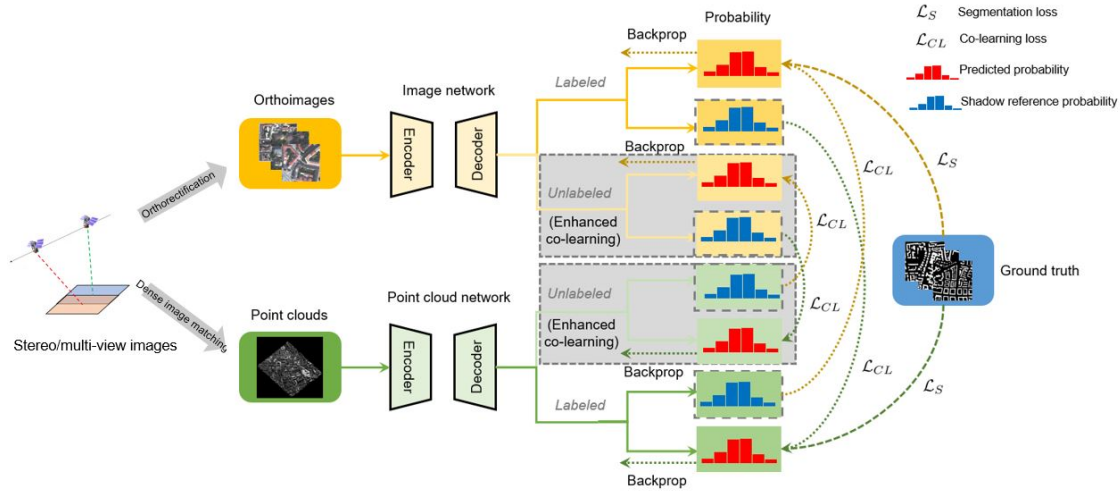


Figure 4.5: The training phase of the proposed co-learning framework for building extraction.

In the multimodal co-learning framework, both labeled and unlabeled training data can be employed during the training phase. Labeled training data engage the optimization of both the supervised and co-learning loss functions, while unlabeled training data pairs can benefit another modality via co-learning loss. We name the setting with only labeled pairs as *standard co-learning*, and the situation trained partly with additional unlabeled pairs as *enhanced co-learning*. Figure 4.5 contains the training procedures of both standard and enhanced versions.

In this work, the cross-entropy is used as the supervised loss function:

$$\mathcal{L}_S(P||Q) = H(P||Q) \quad (4.1)$$

$$= \sum_{x \in \mathcal{X}} P(x) \log(Q(x)), \quad (4.2)$$

where P and Q are defined on the same probability space \mathcal{X} . The term P denotes the distribution of the ground truth, while Q is the probability distribution of the predicted output.

Co-learning can be realized by a similarity loss function. Referring to [166], we adopt Kullback-Leibler (KL) divergence:

$$\mathcal{L}_{CL}(P||Q) = \mathcal{D}_{KL}(P||Q) \quad (4.3)$$

$$= \sum_{x \in \mathcal{X}} P(x) \log\left(\frac{P(x)}{Q(x)}\right), \quad (4.4)$$

where P and Q are defined on the same probability space \mathcal{X} . The item P denotes the probability distribution of the target data, while Q is the probability distribution of the predicted output. In our co-learning framework, P and Q are from two different modalities. P is the shadow reference probability, while Q is the predicted probability.

The integration of a co-learning loss function \mathcal{L}_{CL} with a supervised loss function \mathcal{L}_S forms the basis for the total standard co-learning loss function \mathcal{L}_{total} for each single-modal network. For a 2-D image network, the total loss function is defined as:

$$\mathcal{L}_{total} = \mathcal{L}_S + \lambda_1 \mathcal{L}_{CL}(P_{3D}||P_{2D}), \quad (4.5)$$

where λ_1 acts as a hyperparameter to weight the co-learning loss function. In this equation, the probability map of point clouds P_{3D} serves as the shadow reference for the image network within the co-learning loss function, treated as a constant coefficient.

Similarly, for a 3-D point cloud network, the total loss function is given by:

$$\mathcal{L}_{total} = \mathcal{L}_S + \lambda_1 \mathcal{L}_{CL}(P_{2D}||P_{3D}), \quad (4.6)$$

with λ_1 similarly weighting the co-learning loss. Here, the image network's probability map P_{2D} is utilized as the shadow reference in the co-learning loss function for the point cloud network.

In the scenario of enhanced co-learning, the framework also incorporates unlabeled data pairs into the co-learning loss to enrich the learning process. The total loss function for the image network loss function, integrating the enhanced co-learning loss $\mathcal{L}_{CL}^{unlabeled}$, is defined as:

$$\mathcal{L}_{total} = \mathcal{L}_S + \lambda_1 \mathcal{L}_{CL}^{labeled}(P_{3D}||P_{2D}) + \lambda_2 \mathcal{L}_{CL}^{unlabeled}(P_{3D}||P_{2D}), \quad (4.7)$$

Similarly, for the point cloud network trained in enhanced co-learning mode, the total loss function is:

$$\mathcal{L}_{total} = \mathcal{L}_S + \lambda_1 \mathcal{L}_{CL}^{labeled}(P_{2D}||P_{3D}) + \lambda_2 \mathcal{L}_{CL}^{unlabeled}(P_{2D}||P_{3D}), \quad (4.8)$$

4.3.1.3 Experiments

This section reports three experiments designed to evaluate the effectiveness of the proposed co-learning methods in enhancing building extraction networks. Experiment I and Experiment II utilize only 10 labeled training samples from the Munich WorldView-2 dataset and the ISPRS Potsdam dataset, respectively. These experiments aim to demonstrate the capacity of co-learning approaches to improve performance with limited labeled

4 Summary of the Work

data. Experiment III, on the other hand, employs the fully labeled training set from the ISPRS Potsdam dataset, facilitating a direct comparison of the results with SOTA methods.

To evaluate and compare the results, the intersection over union (IoU) and the F1 score of the building class are utilized as the metrics. To better evaluate the confusion between the background and buildings, overall accuracy (OA), false negative rate (FNR), and false positive rate (FPR) are also reported. They are computed as follows:

$$OA = \sum_{i=1}^n \left(\frac{TP_i + TN_i}{TP_i + TN_i + FP_i + FN_i} \right), \quad (4.9)$$

$$F1 = \frac{2TP}{2TP + FP + FN}, \quad (4.10)$$

$$IoU = \frac{TP}{TP + FP + FN}, \quad (4.11)$$

$$FNR = \frac{FN}{TP + FN}, \quad (4.12)$$

$$FPR = \frac{FP}{TN + FP}, \quad (4.13)$$

where i is the class index and n is the total number of classes; in the building extraction task $n = 2$. TP refers to the number of true positives, FP the false positives, TN the true negatives, and FN the false negatives.

Experiment I: 10-shot Munich WorldView-2 Dataset

(1) *Dataset*: In this experiment, we randomly selected 10 image samples with 512×512 pixels (as illustrated in Figure 4.6) and corresponding photogrammetric point clouds in the same range from the entire training set as the labeled training data. Enhanced co-learning utilizes all remaining patches of original training data as unlabeled pairs.

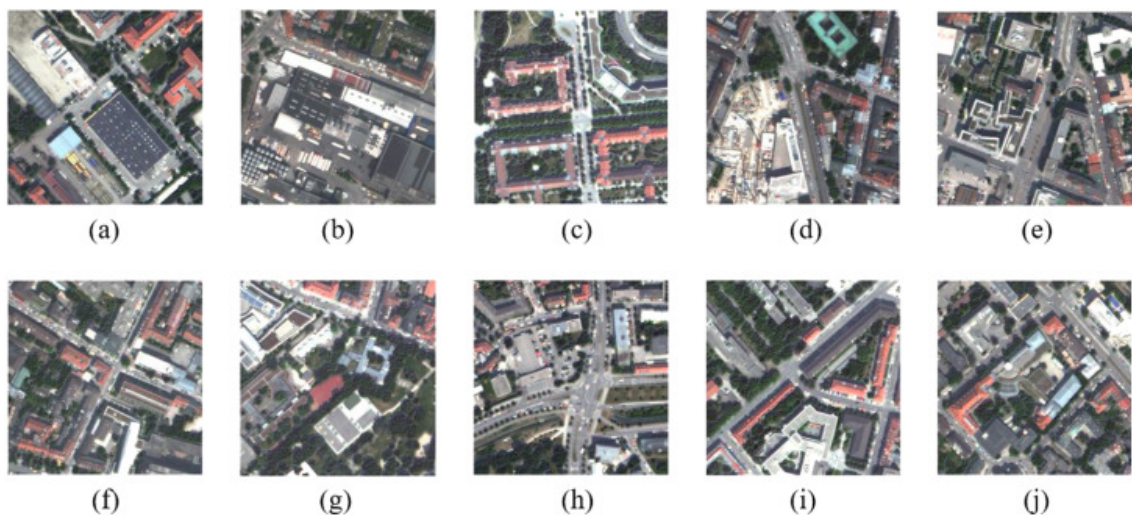


Figure 4.6: 10 labeled training samples of the Munich WorldView-2 dataset.

(2) *Results*: As illustrated in Figure 4.7, the single-modal image baseline network incorrectly classifies numerous non-building objects as buildings, such as areas of low vegetation

and water bodies that possess light colors and regular boundaries, as indicated within the red oval. This misclassification stems from the fact that only 10 labeled image samples cannot provide adequate spectral and textual information to the baseline model. Through the incorporation of geometric knowledge from the point cloud modality via both standard and enhanced co-learning methods, such false positives are substantially reduced.

Regarding point clouds, Figure 4.8 gives three instances where co-learning proves beneficial. The point cloud network, augmented by standard co-learning, correctly identifies more building points compared to its single-modal learning-trained counterpart. Enhanced co-learning demonstrates superior precision in identifying complete building structures over standard co-learning. However, it sometimes introduces more false positives, as exemplified in (a) with red and (c) with yellow annotations. In these scenarios, implementing an additional probability fusion operation can rectify many errors in the results by enhanced co-learning, such as those circled in (a) and (c).

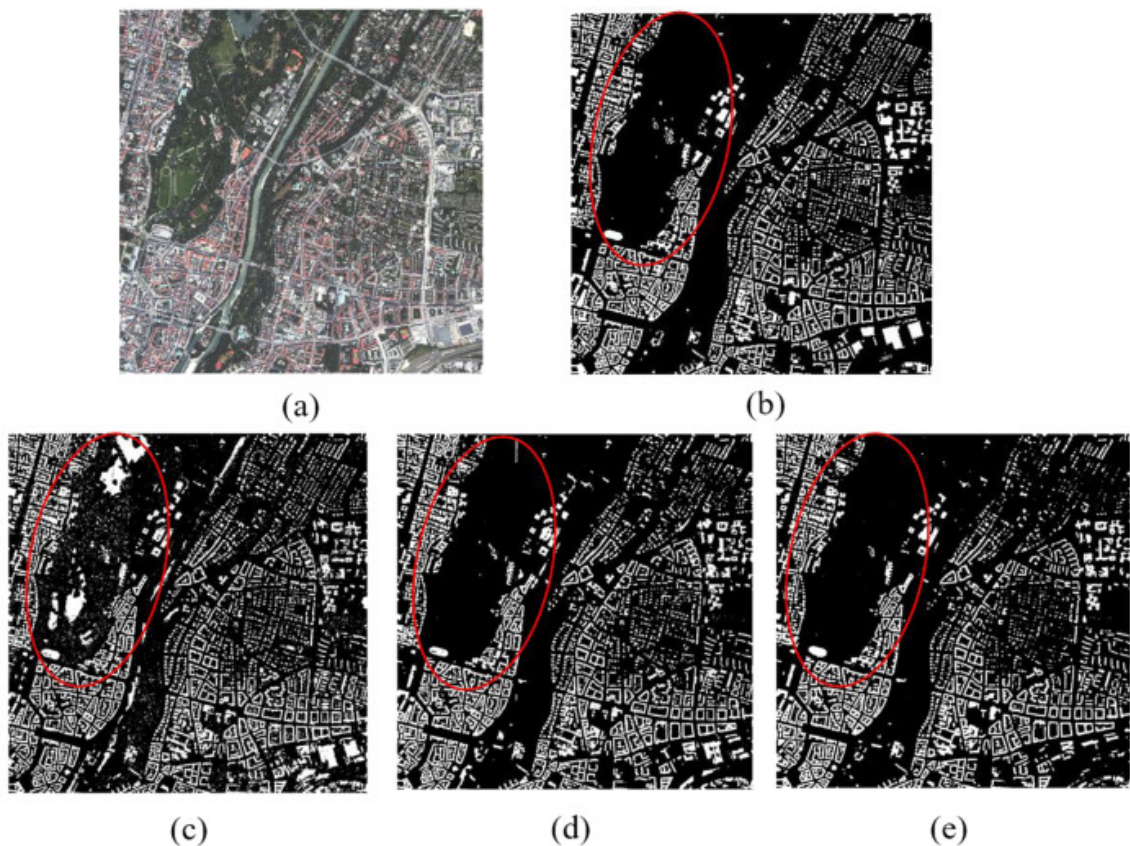


Figure 4.7: The overview of image results obtained from 10-shot Munich WorldView-2 dataset using 10 labeled training samples and various training strategies. (a) Original image. (b) Ground truth. (c) Single-modal. (d) Standard co-learning. (e) Enhanced co-learning.

Quantitative results presented in Table 4.1 also demonstrate the capability of co-learning methodologies. For the image modality, when compared with the baseline approach, the standard co-learning method archives a 3.42% improvement in OA, an 8.68% increase in IoU, and a 6.45% enhancement in F1 score. Additionally, FNR and FPR are diminished by 0.41% and 4.18%, respectively. For the point cloud modality, the

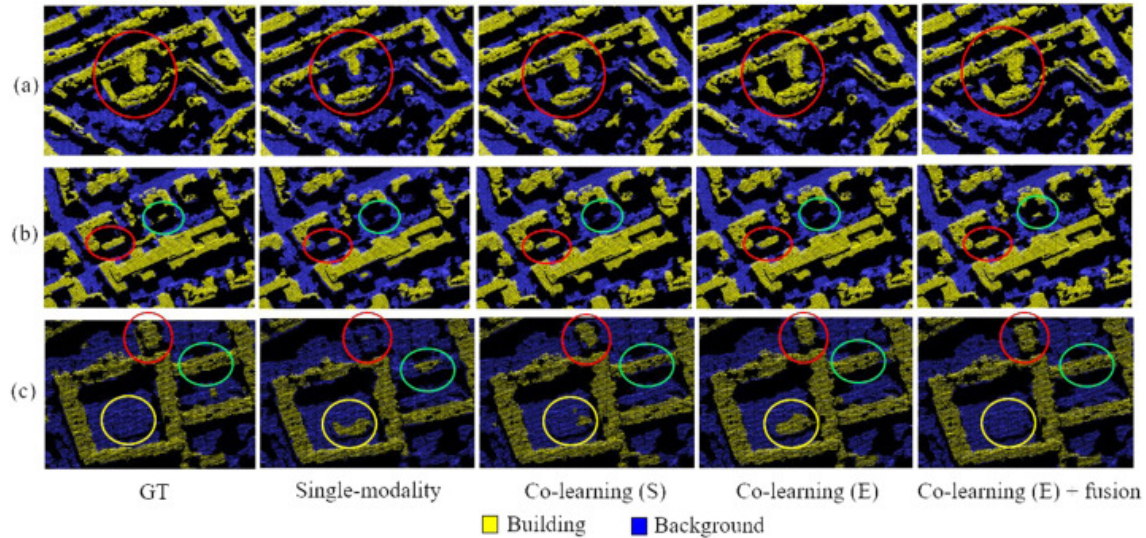


Figure 4.8: Close-up views of point cloud results obtained from 10-shot Munich WorldView-2 dataset using various training strategies. GT: Ground truth. Co-learning (S): standard co-learning. Co-learning (E): enhanced co-learning. Co-learning (E) fusion: enhanced co-learning and probability fusion.

Table 4.1: Performance of different methods for building extraction in the experiment conducted on Munich WorldView-2 dataset utilizing only 10 labeled training sample pairs.

| | Methods | OA | IoU | F1 | FNR | FPR |
|--------------|---|--------|--------|--------|--------|--------|
| Image | Single-modal U-Net (baseline) | 0.8903 | 0.5979 | 0.7484 | 0.1940 | 0.0883 |
| | Co-learning U-Net (standard) | 0.9245 | 0.6847 | 0.8129 | 0.1899 | 0.0465 |
| | Co-learning U-Net (enhanced) | 0.9224 | 0.6682 | 0.8011 | 0.2282 | 0.0393 |
| Point clouds | Single-modal SparseConvNet (baseline) | 0.8465 | 0.4753 | 0.6443 | 0.3958 | 0.0811 |
| | Early fusion SparseConvNet (colorized point clouds) | 0.7938 | 0.4756 | 0.6446 | 0.1874 | 0.2118 |
| | Co-learning SparseConvNet (standard) | 0.8492 | 0.5024 | 0.6688 | 0.3388 | 0.0946 |
| | Co-learning SparseConvNet (enhanced) | 0.8790 | 0.5746 | 0.7298 | 0.2902 | 0.0703 |
| | Enhanced co-learning + late fusion | 0.9371 | 0.7456 | 0.8543 | 0.1984 | 0.0224 |

enhanced co-learning strategy boosts IoU and F1 by 9.93% and 8.55%, respectively, while reducing FNR and FPR by 10.56% and 1.08% compared to the results from the baseline of single-modal learning.

Experiment II: 10-shot ISPRS Potsdam Dataset

(1) *Dataset:* In this experiment, point clouds are converted from DSMs. We randomly selected 10 image samples with 512×512 pixels (as illustrated in Figure 4.9) and corresponding DSM-derived point clouds in the same range from the entire training set as the labeled training data. Enhanced co-learning utilizes all remaining samples of original training data as unlabeled pairs.

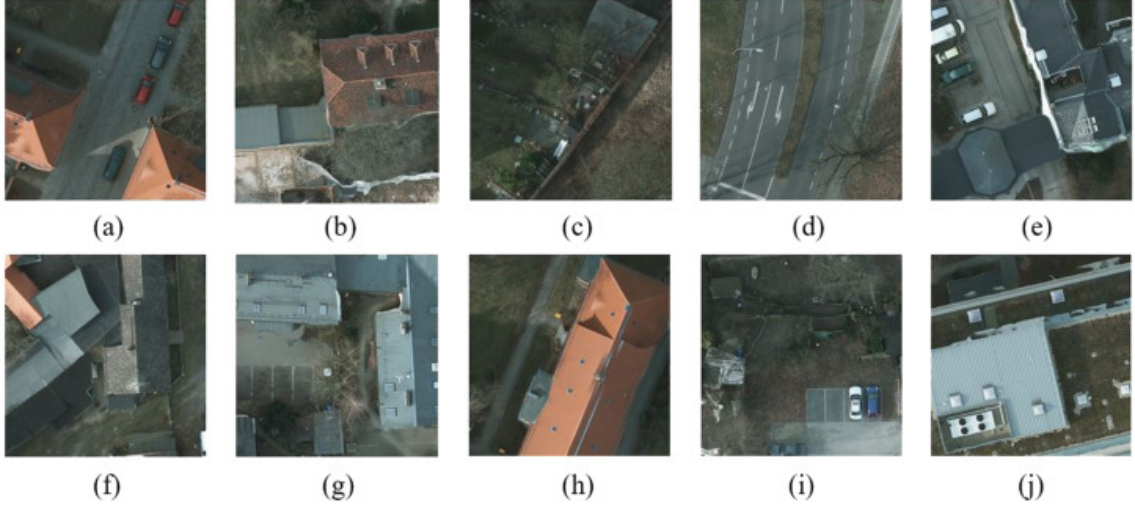


Figure 4.9: 10 labeled training samples of the Potsdam dataset.

(2) *Results:* Figure 4.10 presents five examples of building extraction results from RGB images, illustrating the challenges faced by single-modal learning methods when limited to only 10 labeled samples. In these cases, almost every building suffers from defects, due to the poor features learned from only 10 labeled samples. Standard co-learning brings some improvement to buildings. However, several background pixels are still erroneously identified as part of buildings. By contrast, the enhanced co-learning approach, which leverages a substantial volume of unlabeled training data, delivers significantly better results. Here, the primary issues are confined to building boundaries, small structures, and auxiliary features. Furthermore, applying the probability fusion operation refines the building masks by enhanced co-learning, particularly in detecting small-sized buildings as demonstrated in examples (d) and (e).

Figure 4.11 presents a close-up view of the results by the point cloud networks. A common limitation across all three methods is the tendency to mistakenly classify points from tall objects as buildings, due to the absence of spectral textural information to serve as a distinguishing factor. Fortunately, the knowledge transferred from the image modality can eliminate such errors to some extent. As highlighted in the circled area, the outputs generated by both co-learning strategies exhibit a reduced number of false positives compared to those produced by single-modal learning.

Table 4.2 lists the quantitative results. For both image and point cloud modalities, enhanced co-learning is superior to standard co-learning and baseline networks. When comparing these results to those trained via single-modal learning, the improvements in image modality achieved through enhanced co-learning are with increases of 5.75%, 18.06%, and 13.30% in OA, IoU, and F1, respectively. In the case of point clouds, the performance differences among the three models are relatively minor. Compared to the baseline performance, the top results from the enhanced co-learning strategy exhibit improvements of 0.95% in OA, 2.86% in IoU, and 1.78% in F1 score, respectively.

Experiment III: Full ISPRS Potsdam Dataset

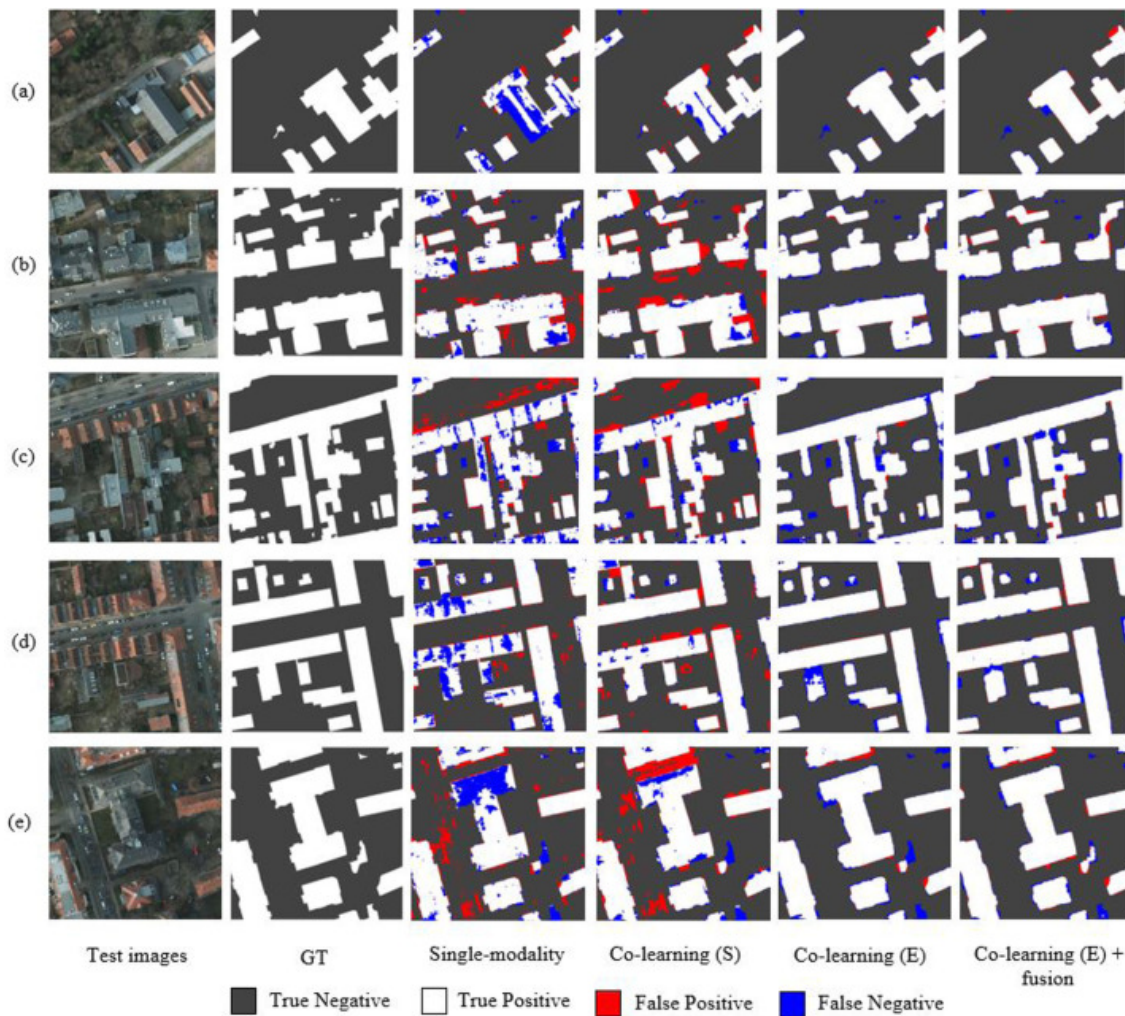


Figure 4.10: Close-up views of building extraction (image) results obtained from 10-shot ISPRS Potsdam dataset using various training strategies. GT: Ground truth. Co-learning (S): standard co-learning. Co-learning (E): enhanced co-learning. Co-learning (E) fusion: enhanced co-learning and probability fusion.

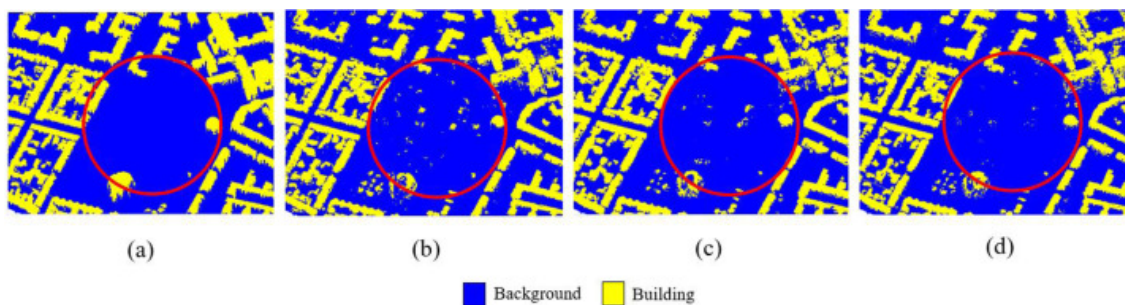


Figure 4.11: Close-up views of building extraction (point clouds) results obtained from 10-shot ISPRS Potsdam data set using various training strategies. (a) Ground truth. (b) Single-modality. (c) Standard co-learning. (d) Enhanced co-learning.

4.3 Case Studies with 2-D and 2.5-D/3-D Multimodal Learning

Table 4.2: Performance of different methods for building extraction in the experiment conducted on ISPRS Potsdam dataset utilizing only 10 labeled training sample pairs.

| | Methods | OA | IoU | F1 | FNR | FPR |
|--------------|---|--------|--------|--------|--------|--------|
| Image | Single-modal U-Net (baseline) | 0.8795 | 0.5633 | 0.7202 | 0.3502 | 0.0483 |
| | Early fusion U-Net (RGB + elevation) | 0.9004 | 0.6471 | 0.7857 | 0.2364 | 0.0566 |
| | Co-learning U-Net (standard) | 0.8850 | 0.6018 | 0.7514 | 0.2734 | 0.0652 |
| | Co-learning U-Net (enhanced) | 0.9370 | 0.7439 | 0.8532 | 0.2349 | 0.0089 |
| | Enhanced co-learning + late fusion | 0.9581 | 0.8291 | 0.9066 | 0.1509 | 0.0076 |
| Point clouds | Single-modal SparseConvNet (baseline) | 0.9409 | 0.7773 | 0.8747 | 0.1379 | 0.0343 |
| | Early fusion SparseConvNet (colorized point clouds) | 0.9167 | 0.6958 | 0.8206 | 0.2034 | 0.0455 |
| | Co-learning SparseConvNet(standard) | 0.9450 | 0.7906 | 0.8831 | 0.1321 | 0.0307 |
| | Co-learning SparseConvNet (enhanced) | 0.9504 | 0.8059 | 0.8925 | 0.1390 | 0.0215 |

To delve deeper into the capabilities of the co-learning framework and to compare it with the SOTA single-modal networks, we conduct experiments utilizing fully labeled training data. This part reports the experiment on the public benchmark ISPRS Potsdam dataset.

(1) *Dataset:* In this experiment, the entire labeled training set is employed as the labeled training pairs. We follow the data splitting settings of [110, 167]. For the enhanced co-learning approach, the testing data are utilized as unlabeled training pairs to facilitate the transfer of more knowledge between the image and point cloud modalities.

(2) *Results:* Table 4.3 outlines the results achieved by our co-learning framework and SOTA methods [110, 167, 168, 169]. Compared with the result achieved by our single-modal learning, the standard co-learning method exhibits a 1.37% increase in OA, a 5.56% in IoU, and a 1.63% in F1. Notably, our 2D U-Net model, when trained using the standard co-learning strategy, outperforms the SOTA results achieved by the mentioned single-modal networks.

Table 4.3: Performance of single-modal learning and co-learning results in the ISPRS Potsdam data set with full labels. The results of EPUNet and ESFNet are from [167].

| Methods | OA | IoU | F1 | FNR | FPR |
|-------------------------------------|--------|--------|--------|--------|--------|
| EPUNet [168] | - | 0.7941 | 0.8852 | - | - |
| ESFNet [169] | - | 0.8023 | 0.8865 | - | - |
| RegGAN [167] | - | 0.8248 | 0.9040 | - | - |
| SegNet-8s-AFM [110] | - | 0.8275 | 0.9056 | - | - |
| Single-modal U-Net | 0.9486 | 0.7928 | 0.8844 | 0.1770 | 0.0120 |
| Co-learning U-Net (standard) | 0.9623 | 0.8484 | 0.9180 | 0.1183 | 0.0123 |
| Co-learning U-Net (enhanced + test) | 0.9673 | 0.8676 | 0.9291 | 0.1048 | 0.0100 |

4.3.1.4 Summary

This section presents a multimodal co-learning framework suitable for building extraction from multispectral images and photogrammetric point clouds with limited labeled training data. The proposed co-learning methods can simultaneously train an enhanced image network and an enhanced point cloud network. The experiments have demonstrated the effectiveness of both standard co-learning and enhanced co-learning in the building extraction tasks with spaceborne and airborne data. For more technical details, experimental results, and comprehensive analysis, please refer to **Appendix B**.

4.3.2 Case II: Multimodal Co-learning Enhances the Building Extraction Networks with Cross-domain Data

This section presents the case study of how multimodal co-learning enhances the image and point cloud building extraction networks with cross-domain training data. It consists of the spaceborne \rightarrow airborne study in [27] and synthetic \rightarrow real experiment utilizing the SMARS dataset proposed in **Appendix C**.

4.3.2.1 Background

Deep learning-based algorithms have revolutionized remote sensing tasks with multispectral images, DSMs, and point clouds, showcasing significant advancements [2, 15]. However, the generalization ability of these algorithms can be significantly restricted when there is a substantial domain gap between the source datasets and target datasets. The performance of deep learning models trained on limited samples drops significantly when predicting unseen domains. In building extraction tasks, such domain gaps widely exist in multi-sensor and multi-seasonal datasets. Different urban/rural layouts and distinct building styles and distributions are other commonly seen factors causing these challenges.

Multimodal co-learning is a semi-supervised methodology that can utilize unlabeled data pairs for network training. This strategy transfers knowledge across different modalities, regardless of whether the data is labeled or unlabeled. Therefore, it can also employ multimodal target data to enhance the networks during the training phase. In this section, we extend the work introduced in section 4.3.1 to a framework suitable for cross-domain data, exploring how multimodal co-learning works on building extraction from stereo-/multi-view data consisting of multispectral images and corresponding DSMs.

4.3.2.2 Methodology

Figure 4.12 illustrates the workflow of the multimodal co-learning framework for cross-domain building extraction. This workflow is modified from the enhanced co-learning introduced in section 4.3.1 and **Appendix B**, so it can utilize unlabeled data pairs for the training. During the training phase, two networks operate in parallel: an image network that processes multispectral images and a sparse convolutional point cloud network that handles DSM data. Both networks are trained simultaneously yet independently. Similar to mainstream supervised building extraction, the labeled source data are trained with the semantic segmentation loss \mathcal{L}_S in this framework. In addition, the co-learning loss function \mathcal{L}_{CL} conducts a bridge between image and point cloud modalities, enabling them to leverage knowledge from each other adaptively. Therefore, unlabeled target data pairs can also be employed via this setting. Within this structure, there are four subsets of the

co-learning loss function, applicable either within the source domain or the target domain, with two subsets dedicated to optimizing the image network and the remaining two aimed at optimizing the point cloud network.

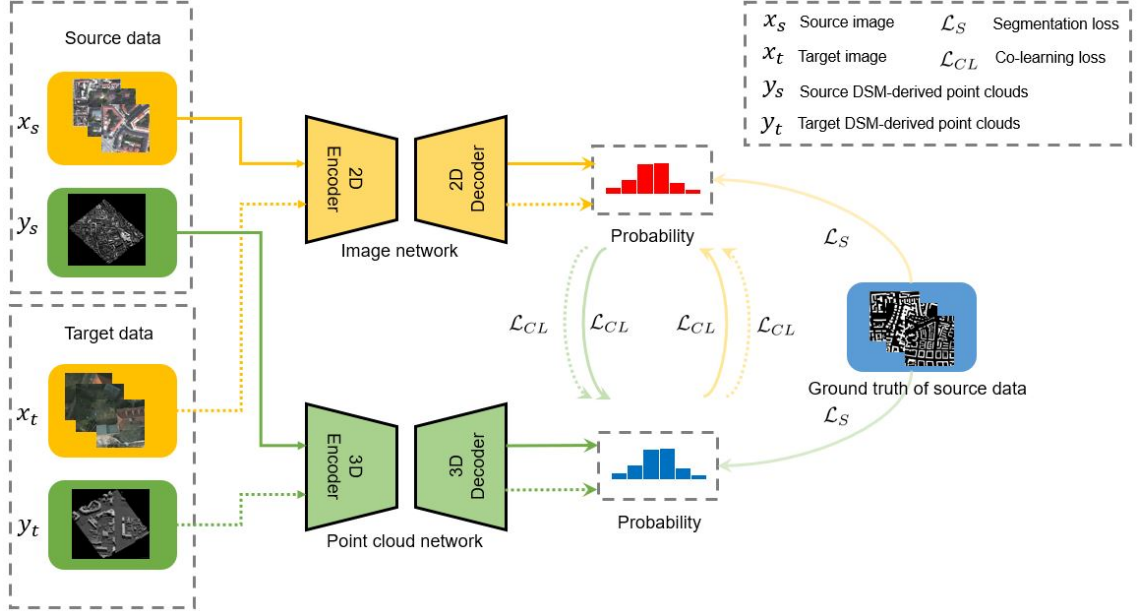


Figure 4.12: The workflow of the proposed multimodal and cross-domain co-learning framework for building extraction.

In this work, coss-entropy is adopted to measure building extraction loss for supervised learning:

$$\mathcal{L}_S(P||Q) = H(P||Q) \quad (4.14)$$

$$= \sum_{x \in \mathcal{X}} P(x) \log(Q(x)) , \quad (4.15)$$

where P and Q are defined within the identical probability space \mathcal{X} . P is the distribution of the ground truth, whereas Q is the probability distribution of the predicted output.

Following the success in **Appendix B**, KL divergence is used as the co-learning loss function:

$$\mathcal{L}_{CL}(P||Q) = \mathcal{D}_{KL}(P||Q) \quad (4.16)$$

$$= \sum_{x \in \mathcal{X}} P(x) \log\left(\frac{P(x)}{Q(x)}\right) , \quad (4.17)$$

where P represents the probability distribution of the target data, whereas Q is the probability distribution of the predicted output.

The total loss function of this framework is:

$$\mathcal{L}_{total} = \mathcal{L}_S + \lambda_1 \mathcal{L}_{CL}^{labeled}(M_1||M_2) + \lambda_2 \mathcal{L}_{CL}^{unlabeled}(M_1||M_2) , \quad (4.18)$$

4 Summary of the Work

where λ_i is the weighting coefficient. M_1 and M_2 represent two modalities, respectively.

To evaluate the effectiveness of the proposed framework, we select two kinds of mainstream 2-D CNN-based building extraction networks introduced in [39] as the baseline in the experiments. One is the encoder-decoder U-Net [68] enhanced with the ResNet-34 backbone [65]. The other is HR-Net [104] adopting multiscale subnetworks in parallel. Following **Appendix B**, SparseConvNet is employed as the point cloud network to process cross-domain DSMs.

4.3.2.3 Experiments

This section reports two domain adaptation experiments supporting the case of cross-domain building extraction. Experiment I is the domain adaptation between the spaceborne Munich WorldView-2 dataset and airborne ISPRS Potsdam dataset, which has been published in [27]. Experiment II is the domain adaptation between the SMARS dataset and the ISPRS Potsdam dataset via our proposed co-learning framework. Building class’s OA of the Equation 4.9, IoU of the Equation 4.11, and F1 of the Equation 4.10 are employed as the evaluation metrics.

Experiment I: Spaceborne Data \rightarrow Airborne Data

(1) *Datasets*: In this experiment, the source dataset is the spaceborne Munich WorldView-2 dataset introduced in section 4.2.1.2. The target dataset is the public ISPRS Potsdam dataset introduced in section 4.2.1.1. Different from the raw point cloud modality used in 4.2.2.1, in this experiment DSMs are used as the input of the Munich WorldView-2 dataset, maintaining consistency with the ISPRS Potsdam dataset. As a preprocessing operation of the target domain data, RGB images and DSMs in the ISPRS Potsdam dataset are downsampled to the GSD of 0.5m, consistent with WorldView-2 images and DSMs in the source domain. Original ISPRS Potsdam training data have 24 tiles. In this experiment, 20 of them (ID: 2-10, 2-12, 3-10, 3-11, 3-12, 4-11, 4-12, 5-10, 5-11, 6-7, 6-8, 6-9, 6-10, 6-11, 6-12, 7-7, 7-9, 7-10, 7-11, and 7-12) are utilized as the unlabeled training pairs and maintaining 4 tiles are employed as the validation set for the selection of optimal checkpoints. Figure 4.13 provides an example comparison between the abovementioned two datasets. They exhibit differences in both spectral style and the quality of DSMs.

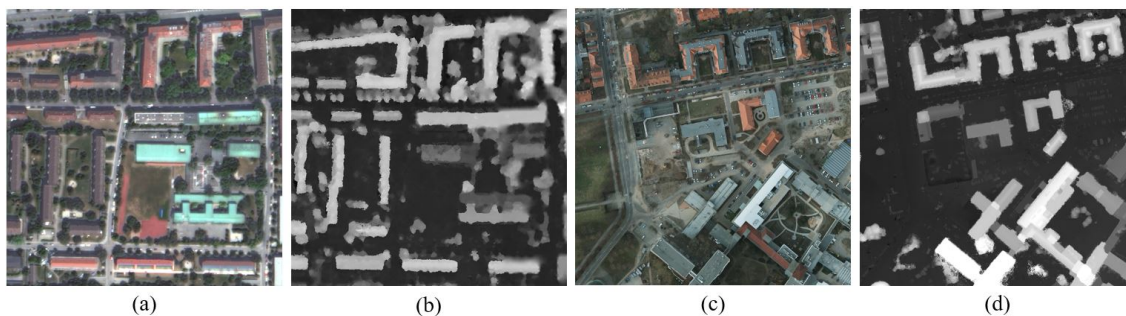


Figure 4.13: Examples of Munich WorldView-2 (source) and ISPRS Potsdam (target) datasets. (a) A WorldView-2 image patch with RGB channels. (b) A DSM patch derived from WorldView-2 images. (c) An RGB image patch from the ISPRS Potsdam airborne dataset. (d) A DSM patch from the ISPRS Potsdam airborne dataset.

(2) *Results:* Table 4.4 presents the quantitative results on the downsampled ISPRS Potsdam dataset. In comparison with the baseline U-Net and HR-Net trained only with source images, the application of enhanced co-learning mode with DSM-derived point clouds results in a significant performance boost. Specifically, when applied with HR-Net, enhanced co-learning achieves increases of 5.14% in OA, 9.36% in F1 score, and 9.72% in IoU. The benefits are even more pronounced with U-Net, where enhanced co-learning contributes to a remarkable improvement of 13.29% in OA, 17.97% in F1 score, and 21.52% in IoU. Regarding the point cloud modality, the results of the three methods are close. Co-learning with U-Net contributes a slight improvement to SparseConvNet, including a 0.85% in OA, a 1.96% in F1, and a 2.84% in IoU, compared with the baseline source-only method.

Table 4.4: Quantitative results by different methods in the building extraction experiment of spaceborne→airborne.

| | Methods | OA | F1 | IoU |
|-------------------|---|--------|--------|--------|
| Image | HR-Net (source only) | 0.8116 | 0.5648 | 0.3935 |
| | HR-Net (co-learning with SparseConvNet) | 0.8630 | 0.6584 | 0.4907 |
| | U-Net (source only) | 0.7710 | 0.6150 | 0.4441 |
| | U-Net (co-learning with SparseConvNet) | 0.9039 | 0.7947 | 0.6593 |
| DSM (Point Cloud) | SparseConvNet (source only) | 0.9241 | 0.8246 | 0.7015 |
| | SparseConvNet (co-learning with HR-Net) | 0.9187 | 0.8167 | 0.6902 |
| | SparseConvNet (co-learning with U-Net) | 0.9272 | 0.8363 | 0.7186 |

Figure 4.14 compares the qualitative results achieved by various methods. The performance of three point cloud networks is similar. In contrast, the co-learning strategy demonstrates a remarkable improvement in image networks, including both U-Net and HR-Net. As shown in Figure 4.14, the contrast in spectral style between the WorldView-2 image and the airborne Potsdam image may explain why the image networks trained with only source images fail to perform satisfactorily on the target data. A comparison between Figures 4.14 (c) and (d) reveals several critical limitations of source-only HR-Net. For example, marked by red circles, it misses a large portion of building structures. Conversely, when trained in co-learning mode with unlabeled target data pairs, the enhanced HR-Net demonstrates an improved ability to obtain more details of buildings. Figure 4.14 (e) illustrates that U-Net trained only with source images encounters significant challenges, incorrectly classifying numerous non-building pixels as buildings. In contrast, Figure 4.14 (f) achieved by co-learning-enhanced U-Net shows a great decrease in false positives.

Experiment II: Synthetic Data → Real Data

(1) *Datasets:* To further explore the potential of the proposed multimodal co-learning for domain adaptation framework, we conduct a synthetic → real experiment. In this experiment, the 30-cm SParis dataset of SMARS proposed in **Appendix C** is employed as the labeled source data. The real ISPRS Potsdam dataset is set as the target data. To reduce the influence of the resolution difference, we downsample the GSD of the RGB images and DSMs in ISPRS Potsdam data to 30 cm, keeping consistency with the source dataset.

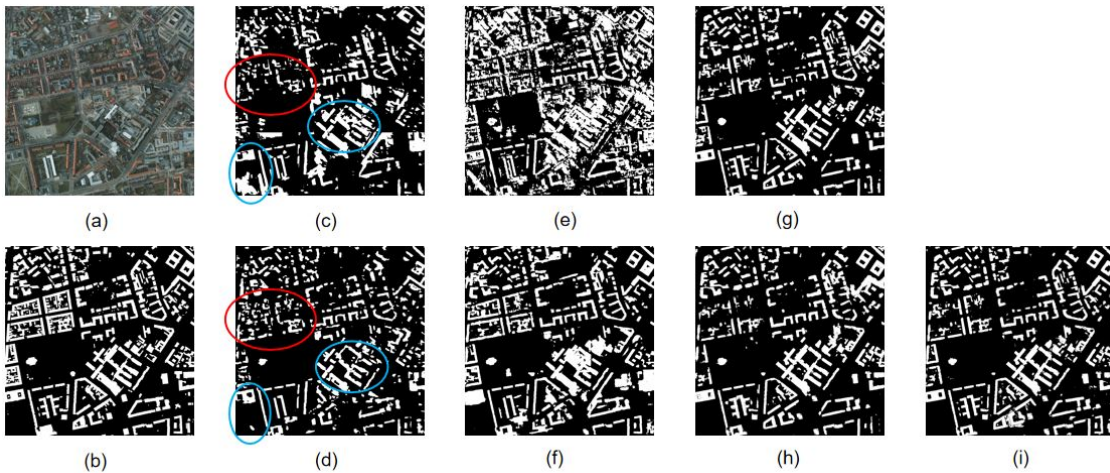


Figure 4.14: Building extraction results of ISPRS Potsdam target data. (a) RGB image. (b) Ground truth. Munich WorldView-2 dataset is employed as the source data. (c) HR-Net (source only). (d) HR-Net (co-learning with SparseConvNet). (e) U-Net (source only). (f) U-Net (co-learning with SparseConvNet). (g) SparseConvNet (source only). (h) SparseConvNet (co-learning with HR-Net). (i) SparseConvNet (co-learning with U-Net).

(2) *Results:* Table 4.5 lists the quantitative results achieved by each method for both image and DSM/point cloud modality. In comparing their OA, F1, and IoU scores, the proposed co-learning approach demonstrates a strong power that significantly improves the results of each network. For two image networks, HR-Net gains an improvement of 16.22% in OA, 12.35% in F1, and 13.23% in IoU, while U-Net experiences a greater increase of 22.51% in OA, 20.82% in F1, and 23.81% in IoU. For the point cloud network SparseConvNet, the mutual information of unlabeled data from both the HR-Net and U-Net can boost its performance. In comparison to the source-only baseline network, SparseConvNet achieves an increase of 17.14% in OA, 5.23% in F1, and 7.23% in IoU via the co-learning with U-Net, and a bigger boost of 18.56% in OA, 9.46% in F1, and 13.54% via the co-learning with HR-Net.

Table 4.5: Quantitative results by different methods in the building extraction experiment of SParis→Potsdam.

| | Methods | OA | F1 | IoU |
|-------------------|---|--------|--------|--------|
| Image | HR-Net (source only) | 0.5136 | 0.5709 | 0.3995 |
| | HR-Net (co-learning with SparseConvNet) | 0.6758 | 0.6944 | 0.5318 |
| | U-Net (source only) | 0.5322 | 0.5691 | 0.3977 |
| | U-Net (co-learning with SparseConvNet) | 0.7573 | 0.7773 | 0.6358 |
| DSM (Point Cloud) | SparseConvNet (source only) | 0.7556 | 0.7702 | 0.6263 |
| | SparseConvNet (co-learning with HR-Net) | 0.9412 | 0.8648 | 0.7617 |
| | SparseConvNet (co-learning with U-Net) | 0.9270 | 0.8225 | 0.6986 |

Figure 4.15 visualizes the results. The building masks predicted by source-only HR-Net in (c) and U-Net in (e) are quite noisy. Due to the disparity in color style between synthetic

and real images, the adopted image networks cannot learn generalizable features from SParis for ISPRS Potsdam images. Benefiting from the information through co-learning with point cloud modality using unlabeled target data, the qualitative performance of both HR-Net and U-Net has been significantly improved. This method has led to the successful correction of a large number of background pixels. When analyzing the results of DSMs/point clouds, the source-only network is capable of predicting a globally reasonable mask. Nonetheless, the main challenges arise in accurately capturing local details. For example, as circled in red, the baseline point cloud network misses the extraction of several building structures. This is likely because SParis cannot fully represent the diversity of all building shapes and sizes in the Potsdam dataset. Despite their shortcomings, image networks can identify these missing structures, and the information is successfully transferred to the SparseConvNet via co-learning. As a result, the co-learning-enhanced point cloud networks are able to correct those missing pixels. In addition, the example marked in blue illustrates how co-learning not only aids in filling in missing structures but also helps the point cloud network in reducing false positives.

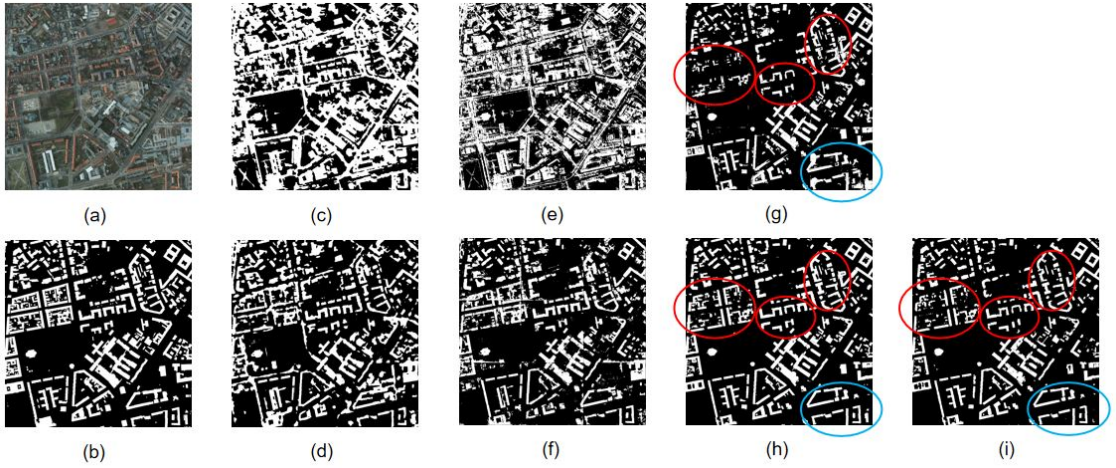


Figure 4.15: Building extraction results of ISPRS Potsdam target data. SParis of SMARS is employed as the source data. (a) RGB image. (b) Ground truth. (c) HR-Net (source only). (d) HR-Net (co-learning with SparseConvNet). (e) U-Net (source only). (f) U-Net (co-learning with SparseConvNet). (g) SparseConvNet (source only). (h) SparseConvNet (co-learning with HR-Net). (i) SparseConvNet (co-learning with U-Net).

4.3.2.4 Summary

This section presents an extended multimodal co-learning framework for building extraction with cross-domain datasets. Through two supportive experiments, the effectiveness of the proposed method is demonstrated. In this semi-supervised framework, unlabeled target data pairs are utilized to train both image and point cloud networks via a co-learning loss function that minimizes the difference between the probabilities of these two modalities. Such a procedure leads to an obvious enhancement to the networks. The experiments also reveal an interesting observation: the source-only point cloud network can exhibit satisfactory performance on unseen target DSMs, significantly outperforming the source-only image networks. This phenomenon suggests that the domain gap between

different DSMs is considerably smaller than that between different multispectral images. It also helps to explain the substantial enhancements seen with co-learning-enhanced image networks. With point clouds acting as a robust modality, more effective knowledge transfer can occur from the point cloud modality to the relatively weaker image modality. Additionally, Experiment II highlights that the SMARS dataset, particularly its DSM data, proves to be a viable training set for real-world building extraction tasks, offering a cost-effective alternative to annotating real data.

4.3.3 Case III: Multimodal Co-learning Enhances the Building Change Detection Networks with Cross-domain Data

This section gives a brief overview of the published journal paper **Appendix D**, presenting how multimodal co-learning enhances the image-based and DSM-based building change detection networks with cross-domain training data.

4.3.3.1 Background

Building change detection is an essential but challenging task that is required in a lot of real-world applications, including disaster assessment [124], urban monitoring [123], and digital mapping [125]. It aims to identify the differences in the condition of building objects within defined areas from multitemporal 2-D, 2.5-D, or 3-D data [59]. Similar to the progression in building extraction tasks, change detection has experienced rapid development due to the advancement in deep learning technologies. As introduced in section 2.2.1.3, the Siamese network is currently the mainstream deep learning architecture for change detection using 2-D images. A series of CNN-based and transformer-based Siamese networks have achieved remarkable performance on public benchmarks [7, 85, 78, 136]. However, challenges persist due to different sensors, acquisition conditions, and geographical locations, resulting in widespread domain gaps in real-world change detection datasets. This significant challenge yet has not been addressed by fully supervised methods [170, 171].

Geometric data like DSMs can provide more discriminative features to man-made objects such as buildings and their changes. Consequently, they are widely used for change detection in conventional methodologies [59, 60, 61]. The case studies presented in sections 4.3.1 and 4.3.2 have confirmed the feasibility of leveraging DSMs for building extraction using deep neural networks and revealed that DSMs exhibit superior generalization capabilities compared to multispectral imagery. This phenomenon inspires a further investigation into the application of DSMs in deep learning-driven building change detection tasks.

While DSMs are good at describing geometric features, they have inherent limitations such as unsharpened object boundaries, incomplete structures, and the presence of outliers. These issues may result in incorrect change detection [59]. Moreover, as investigated in our previous studies [153, 27], the domain gap is also a challenge to fully supervised DSM-/point cloud-based deep learning algorithms, restricting their performance on large-scale real-world datasets. Inspired by the success of knowledge transfer approaches presented in sections 4.3.1 and 4.3.2, this section proposes a multimodal co-learning framework for building change detection, utilizing the hidden mutual information from image and DSM networks to enhance the performance of each other. This section proposes three well-designed co-learning variants named *vanilla co-learning*, *fusion co-learning*, and *detached*

fusion co-learning, respectively. In addition, we propose an end-to-end transformer-based network for change detection from height difference (HDiff) maps. Two HDiff strategies including direct HDiff and robust HDiff are compared in the experiments.

4.3.3.2 Methodology

Figure 4.16 illustrates the generic image-DSM multimodal co-learning framework proposed in **Appendix D**, designed for the task of building change detection. This framework consists of two individual CNN-transformer-hybrid networks. One is a Siamese network conducted for bitemporal images. The other is a single-branch network designed for HDiff maps calculated from bitemporal DSMs. We utilize HDiff maps rather than bitemporal DSMs because HDiff demonstrates a better generalization ability, while original bitemporal DSMs cannot be satisfactorily used by the Siamese network.

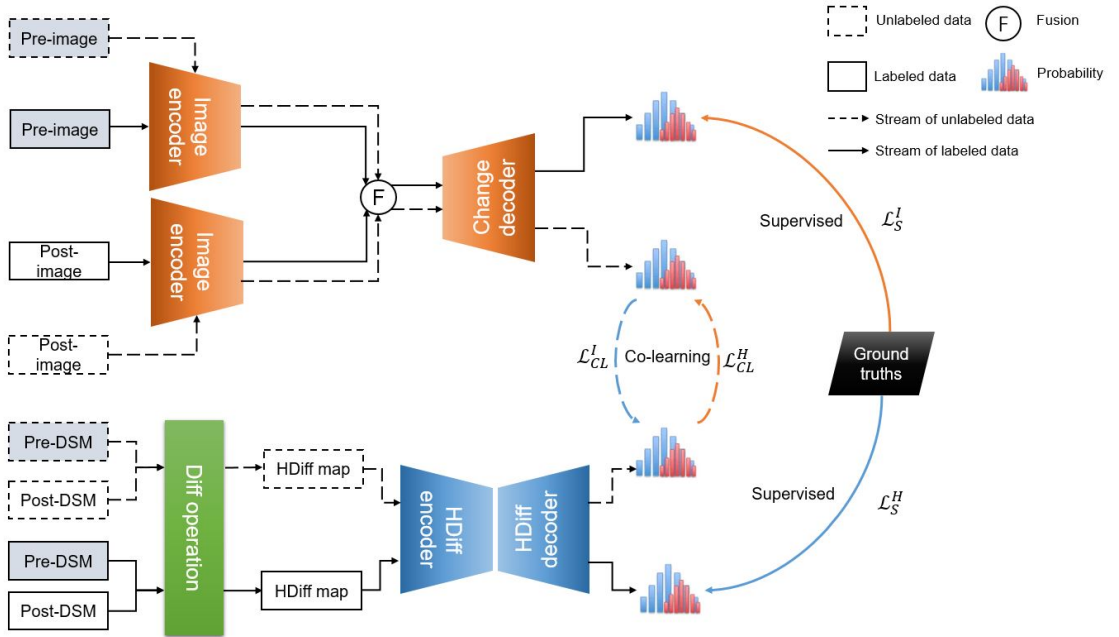


Figure 4.16: The multimodal co-learning framework for change detection from multispectral images and DSMs.

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

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

$$P_I^s = f_I(I_1^s, I_2^s), \quad (4.19)$$

4 Summary of the Work

$$P_I^t = f_I(I_1^t, I_2^t), \quad (4.20)$$

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

$$P_H^s = f_H(H_1^s, H_2^s), \quad (4.21)$$

$$P_H^t = f_H(H_1^t, H_2^t), \quad (4.22)$$

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

$$\mathcal{L}_S^I = L_S(G^s || P_I^s), \quad (4.23)$$

$$\mathcal{L}_S^H = L_S(G^s || P_H^s), \quad (4.24)$$

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

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

Vanilla Co-learning: This is the co-learning implementation following the idea presented in [6], which is based on the intuition that if both the image branch and DSM branch can produce good predictions, their building change probabilities P_I^t and P_H^t should be consistent with each other. Hence, the target co-learning problem is formulated as a generic consistency loss function L_C to minimize the distributions of P_I^t and P_H^t . The vanilla co-learning loss functions for image modality \mathcal{L}_{CL-V}^I and DSM modality \mathcal{L}_{CL-V}^H are calculated as follows:

$$\mathcal{L}_{CL-V}^I = L_C(P_{H,d}^t || P_I^t), \quad (4.25)$$

$$\mathcal{L}_{CL-V}^H = L_C(P_{I,d}^t || P_H^t), \quad (4.26)$$

where $P_{H,d}^t$ and $P_{I,d}^t$ refer to detached P_H^t and P_I^t , respectively. Detached probabilities mean they are variables removed from the gradient computational graph so they do not affect the update of the weights for the corresponding networks. They can be named shadow reference probability, utilized by the main modality network as the reference in the co-learning loss function [6].

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

$$\mathcal{L}_{CL-F}^I = L_C\left(\frac{P_I^t + P_{H,d}^t}{2} || P_I^t\right), \quad (4.27)$$

$$\mathcal{L}_{CL-F}^H = L_C\left(\frac{P_{I,d}^t + P_H^t}{2} \parallel P_H^t\right), \quad (4.28)$$

where $P_{H,d}^t$ and $P_{I,d}^t$ refer to detached P_H^t and P_I^t , respectively.

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

$$\mathcal{L}_{CL-DF}^I = L_C\left(\frac{P_{I,d}^t + P_{H,d}^t}{2} \parallel P_I^t\right), \quad (4.29)$$

$$\mathcal{L}_{CL-DF}^H = L_C\left(\frac{P_{I,d}^t + P_{H,d}^t}{2} \parallel P_H^t\right), \quad (4.30)$$

where L_C denotes a generic consistency loss function. $P_{H,d}^t$ and $P_{I,d}^t$ refer to detached P_H^t and P_I^t , respectively.

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

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

$$\mathcal{L}_{total}^I = \lambda_1 \mathcal{L}_S^I + \lambda_2 \mathcal{L}_{CL}^I, \quad (4.31)$$

$$\mathcal{L}_{total}^H = \lambda_1 \mathcal{L}_S^H + \lambda_2 \mathcal{L}_{CL}^H, \quad (4.32)$$

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

(b) *Siamese ResNet With Bitemporal Image Transformer Layer for Images:* To achieve a balance between network depth and graphics processing unit (GPU) memory, we adopt the ResNet-50 convolutional network [65] with a Siamese structure as the encoder and a bitemporal image transformer (BIT) module [7] at the bottleneck for refining the original bitemporal image features, as illustrated in Figure 4.17. This architecture unfolds in three primary steps: first, leveraging a ResNet-50 backbone to extract initial features from pre-event and post-event images; second, employing the BIT module to enhance these initial features; and third, fusing the refined features via subtraction and deploying a small change classifier to transform the fused features into change maps.

(c) *Transformer-Based UNet for HDiff Maps:* In our framework, f_H contains two steps: (1) apply an operation to calculate HDiff maps and (2) utilize an HDiff network to process HDiff maps. As HDiff rasters have 3-D information of coordinates X , Y , and ΔZ , there are two main approaches to processing them. One is to process them as point clouds

4 Summary of the Work

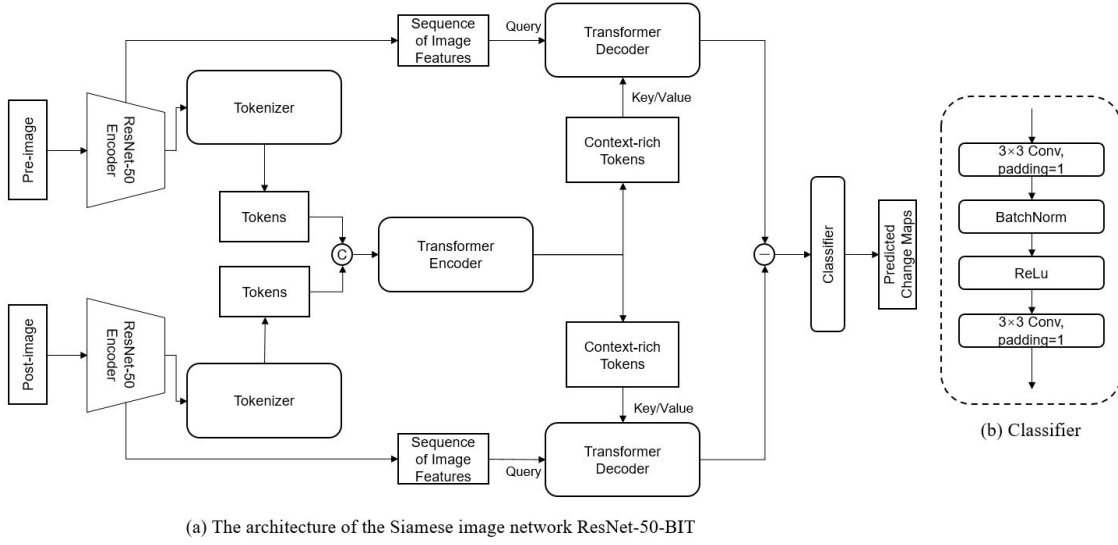


Figure 4.17: (a) The architecture of the Siamese image network employed in this work. (b) Classifier block. The modules of the tokenizer, the transformer encoder, and the transformer decoder are from the implementation of [7].

[6, 27] with 3-D neural networks, while the other is to process them as 2D rasters where the height difference values ΔZ are utilized as input channels to a 2-D neural network. Considering that the height difference values in different cities typically fall within a certain range and 2-D networks are usually more efficient than point cloud networks with the same scales [172], in this study we develop a 2D SwinTransformer-based [12] U-shape network (SwinTransUNet) for the HDiff maps. Figure 4.18 illustrates the implementation of SwinTransUNet. The encoder is conducted with Swin Transformer and patch merging blocks, generating multiscale features with a hierarchical structure. The decoder is a U-Net structure, so various scales of features can be utilized more efficiently. To control the computational cost and GPU memory usage, the dimensionality reduction blocks and upsampling blocks of the decoder are based on convolution and transposed convolution operations, respectively.

(d) *Robust Height Difference:* The quality of spaceborne images is frequently influenced by factors such as limited resolution, illumination distortion, and cloud cover. These limitations often lead to a diminished matching quality of the images, therefore affecting the generated DSMs' quality [60, 28]. The direct consequence is the presence of numerous unexpected outlier pixels in these DSMs and corresponding HDiff maps derived through direct pixelwise subtraction. Such outliers pose significant challenges to the performance of classification algorithms, notably in tasks like building extraction or change detection. To reduce the adverse effects of noise and improve the HDiff maps' quality, a robust difference method has been proposed in [60].

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

4.3 Case Studies with 2-D and 2.5-D/3-D Multimodal Learning

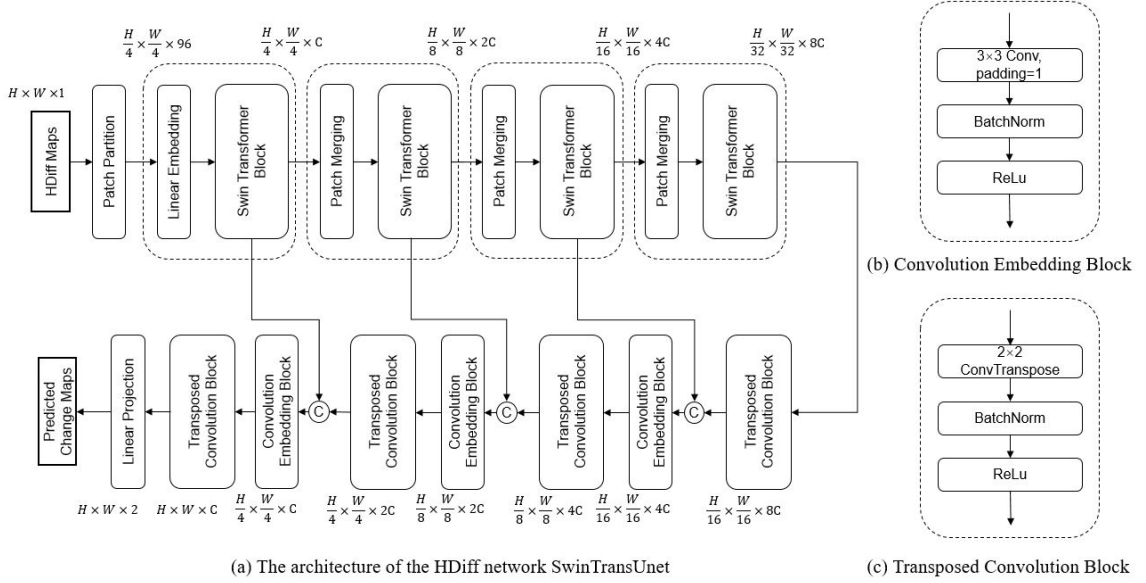


Figure 4.18: (a) The architecture of the proposed HDiff map network SwinTransUNet. (b) Convolution embedding block. (c) Transposed convolution block.

$$Diff_P^H(i, j) = \begin{cases} \min_{p, q} \{H_2(i, j) - H_1(p, q)\}, & x_2(i, j) - x_1(p, q) > 0 \\ 0, & x_2(i, j) - x_1(p, q) \leq 0 \end{cases} \quad (4.33)$$

$$Diff_N^H(i, j) = \begin{cases} 0, & x_2(i, j) - x_1(p, q) \geq 0 \\ \max_{p, q} \{H_2(i, j) - H_1(p, q)\}, & x_2(i, j) - x_1(p, q) < 0 \end{cases} \quad (4.34)$$

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

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

$$Diff_R^H(i, j) = Diff_P^H(i, j) + Diff_N^H(i, j), \quad (4.35)$$

(e) *Loss Functions:* This framework utilizes two kinds of loss functions in each training phase. First, a pixel-wise supervised loss function is used with the labeled source data for change detection. Second, an unsupervised co-learning loss function is applied to the unlabeled target data.

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

$$\begin{aligned} \mathcal{L}_S(G^s || P_I^s) &= CE(G^s || P_I^s) \\ &= \sum_{x \in \mathcal{X}} G^s(x) \log P_I^s(x), \end{aligned} \quad (4.36)$$

4 Summary of the Work

where G^s and P_I^s are defined on the same probability space \mathcal{X} . G^s is the distribution of the source domain’s ground truths. P_I^s is the predicted probability distribution of the image modality from the source domain.

Likewise, the supervised loss function for the DSM modality is

$$\begin{aligned} \mathcal{L}_S(G^s||P_H^s) &= CE(G^s||P_H^s) \\ &= \sum_{\hat{x} \in \mathcal{X}} G^s(\hat{x}) \log P_H^t(\hat{x}), \end{aligned} \quad (4.37)$$

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

2) Loss functions for unsupervised multimodal co-learning: In this work, two kinds of loss functions, KL divergence and mean square error (MSE), are adopted as the co-learning loss function. According to (a) *Problem Statement*, each of them can be seamlessly integrated into our framework and generate three co-learning combinations. It is possible that some loss functions could produce functionally equivalent combinations, meaning they have identical influences during the processes of backpropagation and parameter updating. **Appendix D** presents a method that can determine whether \mathcal{L}_{CL-V} , \mathcal{L}_{CL-F} , and \mathcal{L}_{CL-DF} are equivalent. According to its conclusion, when KL divergence is employed as L_C , \mathcal{L}_{CL-V} , \mathcal{L}_{CL-F} , and \mathcal{L}_{CL-DF} are inequivalent. So they are three different methods. When MSE is employed as L_C , \mathcal{L}_{CL-V} and \mathcal{L}_{CL-F} are equivalent. Therefore, only Vanilla co-learning and detached fusion co-learning are reported for the MSE-based experimental results in the following.

4.3.3.3 Experiments

This section reports two domain adaptation experiments involved in **Appendix D**. Experiment I is the domain adaptation experiments between two subdatasets SParis and SVenice of SMARS. Experiment II is the domain adaptation between SMARS and the real space-borne Istanbul dataset. Figure 4.19 illustrates a comparison of the examples from the two subdatasets of SMARS and the samples from the Istanbul WorldView-2 dataset. It highlights the differences in color styles and building styles among the SParis, SVenice, and Istanbul data, suggesting the domain gap challenges when attempting learning-based building change detection across different datasets.

Four metrics including F1, IoU, precision, and recall are utilized to evaluate the experimental results. Among them, OA, F1, and IoU have been introduced with Equation 4.9, 4.10, and 4.11, respectively. Precision and recall are calculated as:

$$Precision = \frac{TP}{TP + FP}, \quad (4.38)$$

$$Recall = \frac{TP}{TP + FN}, \quad (4.39)$$

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

Experiment I: Domain Adaptation with Synthetic Data

(1) *Datasets*: In this experiment, we conduct a cross-domain scenario using two sub-datasets of SMARS introduced in **Appendix C**. The 50-cm-SParis dataset is employed as the source data, while the 50-cm-SVenice dataset is utilized as the target data.

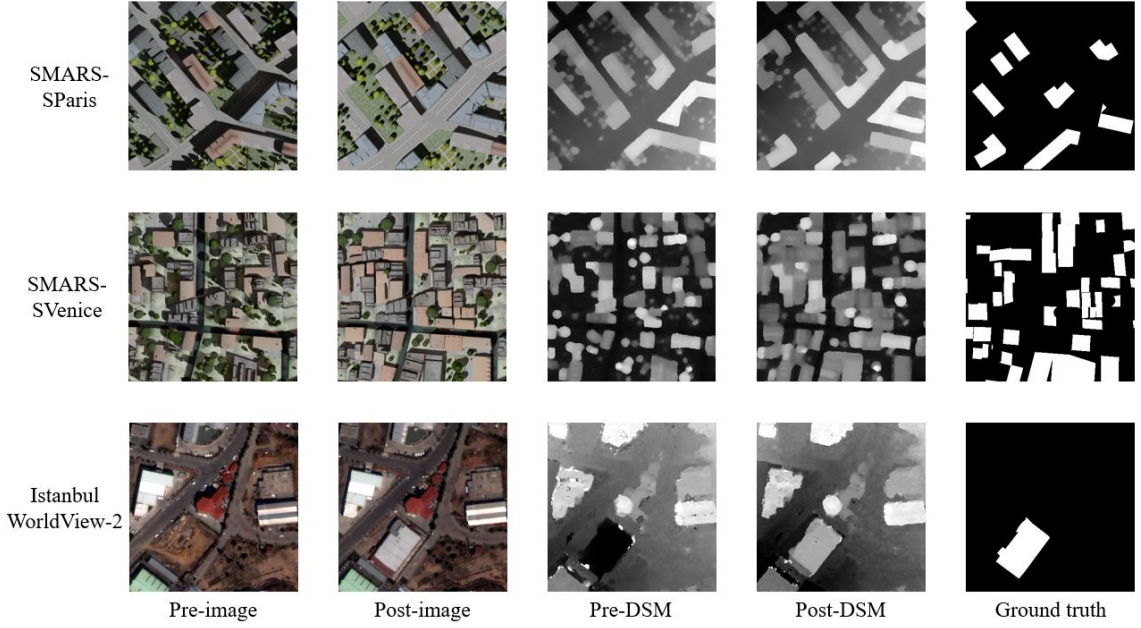


Figure 4.19: Examples of SParis and SVenice datasets in SMARS and the spaceborne Istanbul WorldView-2 dataset.

(2) *Results:* Table 4.6 lists the quantitative results of the cross-domain building change detection experiment SParis \rightarrow SVenice. Either KL-based or MSE-based co-learning combinations have achieved substantial improvements in the performance of image networks. Among the various combinations, the MSE-based detached fusion co-learning strategy stands out, yielding the most significant improvements in comparison with the baseline results — an increase of 62.19% in IoU and 63.97% in the F1 score. In those results of the DSM modality, the best performance is attained through the co-learning-enhanced network trained with MSE-based vanilla co-learning loss. This configuration leads to an F1 score of 83.52% and an IoU of 71.71%, which achieves an increase of 3.86% in F1 and 5.51% in IoU compared with the baseline results by the HDiff network.

Figure 4.20 presents the qualitative analysis of three examples in this experiment. The quantitative results in Table 4.6 and given examples show the limitations of the baseline Siamese network ResNet-50-BIT. It fails to completely identify any building changes in both bitemporal images and DSMs. In comparison, the baseline HDiff network can achieve reasonable results. However, it also results in a large number of false positives. The introduction of our proposed co-learning training strategies significantly improves the performance of the image network ResNet-50-BIT on the target domain. It also boosts the performance of the HDiff map network. When compared to the baseline single-modal learning method, the HDiff network trained in the co-learning framework exhibits a substantial reduction in false negatives.

Similar to the situation in building extraction of section 4.3.2, the DSM modality has a better generalization ability than the image modality. This is demonstrated by the phenomenon that HDiff baseline method achieves much better results than the bitemporal image baseline network. However, when applying co-learning, image networks demonstrate a higher potential for improvement. In comparison, co-learning-enhanced HDiff network has a higher tendency to produce false positive pixels, as presented in

Table 4.6: Quantitative results of different methods on the building change detection Experiment I. The best score is shown in bold.

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

example A. Since the HDiff network is designed to identify changes based on height differences in HDiff maps, some non-man-made object changes with similar geometric features to building changes are mistakenly recognized. In example A, the false positive detection of a round-shaped object at the left border is actually a tree, not a building change. Among the image-based methods, only the network trained with the KL-based fusion co-learning strategy commits the same error.

Experiment II: SMARS → Istanbul WorldView-2 Dataset

(1) *Datasets:* In this experiment, to further explore the potential of proposed co-learning methods, we conduct a challenging synthetic → real building change detection experiment. We adopt the full 50-cm-SMARS training data consisting of the training sets of both SParis and SVenice subdatasets as the source data. The real spaceborne Istanbul WorldView-2 dataset is employed as the target data.

(2) *Results:* Table 4.7 and Figure 4.21 present the quantitative and qualitative results of this experiment, respectively. To verify the effectiveness of robust height difference in improving building change detection results, two sets of comparative experiments are conducted. The first set applies the direct height difference operation to create HDiff maps for the Istanbul dataset, which is indicated with a red **D** in Table 4.7 and subsequent analysis. The second set utilizes the robust height difference method to generate optimized HDiff maps for the same Istanbul data, denoted with a blue **R** in Table Table 4.7 and in the analysis that follows.

1) Co-Learning With Direct HDiff Maps: As listed in Table 4.7 the performance of the Siamese image baseline network trained with SMARS on the novel Istanbul dataset is poor, achieving only a 4.57% F1 score and a 2.34% IoU score. This phenomenon is caused by a significant spectral domain gap between synthetic images and real WorldView-2 data. Similarly, the Siamese DSM baseline method also underperforms, indicating Siamese network is not an ideal architecture for DSM processing. In contrast, the baseline HDiff network demonstrates competent results when utilizing either robust HDiff maps or direct HDiff maps.

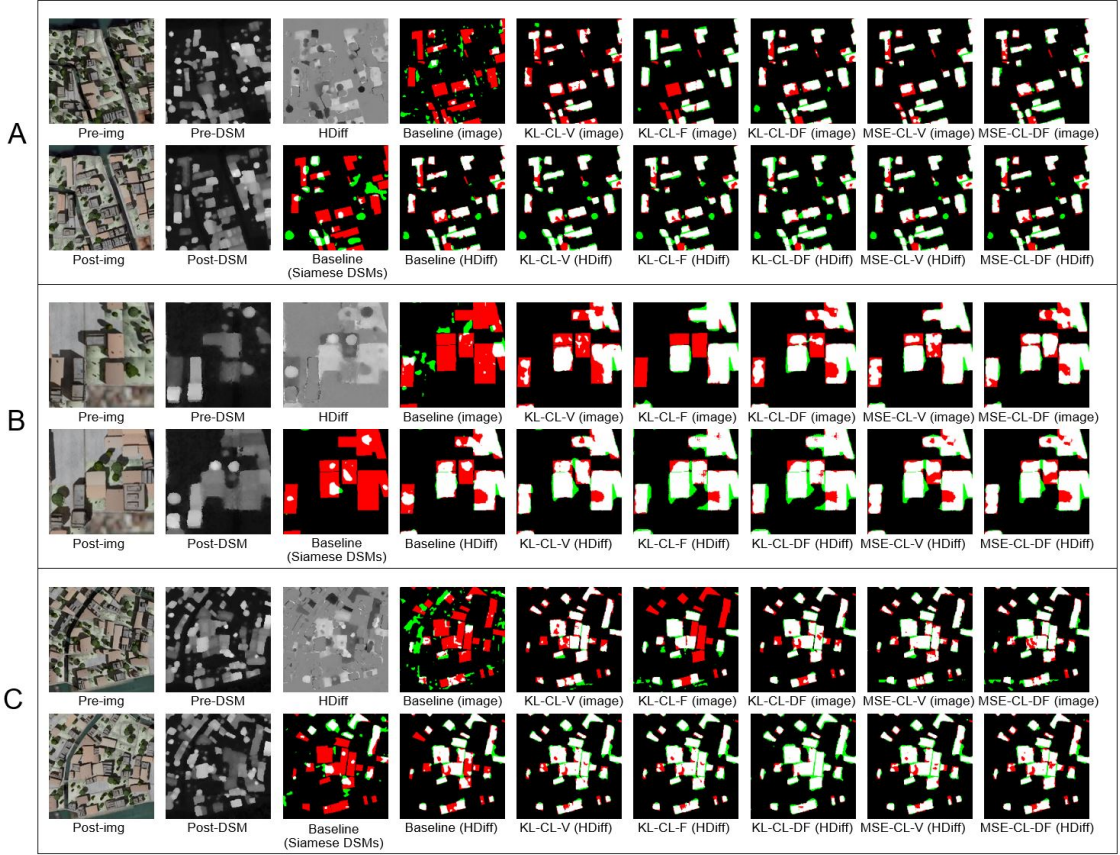


Figure 4.20: Building change detection results of SParis→SVenice. Color legend: TP TN FN FP.

The application of co-learning significantly enhances the performance of the image network. The largest improvement is obtained with the MSE-based vanilla co-learning approach, which elevates F1 to 76.97% and IoU to 62.56%. The performance of the HDiff network SwinTransUNet also benefits from co-learning. Among those methods, MSE-based detached fusion co-learning variant achieves the highest improvement, with a 12.25% increase in precision, a 5.73% rise in the F1 score, and a 7.31% improvement in IoU compared to the baseline approach.

2) Co-Learning With Robust HDiff Maps: The baseline results of **R** in Table 4.7 demonstrate the advantage of robust height difference. When compared to the baseline (**D**), which uses direct HDiff maps, baseline (**R**) that utilizes robust HDiff maps records an improvement of 1.91% in the F1 score and 2.36% in the IoU score.

Utilizing robust HDiff maps, all co-learning methods are also able to improve the performance of both the ResNet-50-BIT image network and the SwinTransUNet HDiff network. Among these methods, the MSE-CL-V variant of co-learning stands out for the image modality, achieving an F1 score of 79.29% and an IoU score of 65.68%. Meanwhile, for the DSM modality, the best performance is obtained with the MSE-CL-DF strategy, which records an F1 score of 77.62% and an IoU of 63.42%. Furthermore, the adoption of robust HDiff maps across each co-learning-enhanced HDiff network yields superior results compared with the same method using direct HDiff maps. Specifically, for the image modality,

4 Summary of the Work

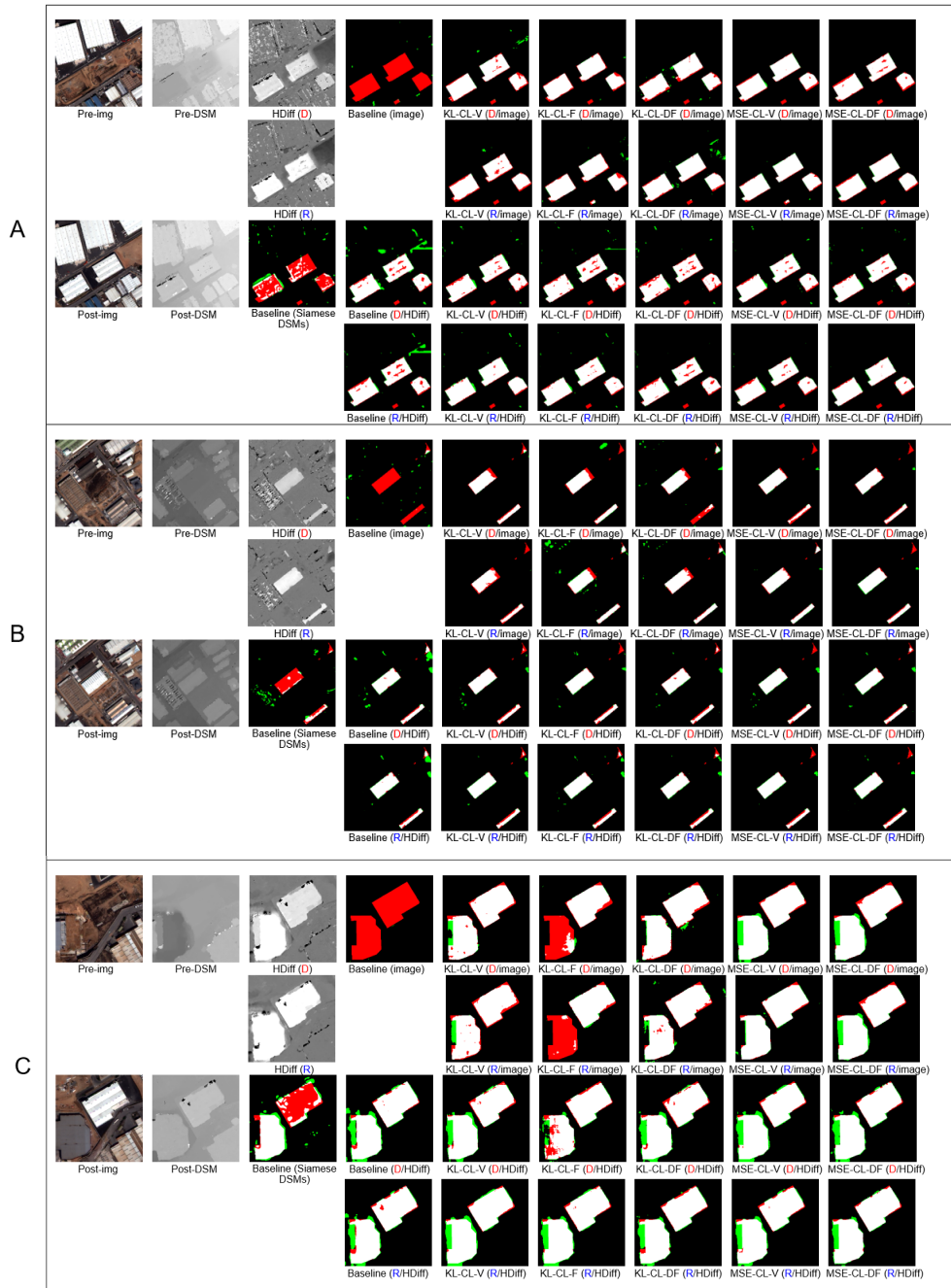


Figure 4.21: Building change detection results of SMARS→Istanbul. Color legend: TP
 TN FN FP.

Table 4.7: Quantitative results of different methods on the building change detection Experiment II.

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

the premier result from MSE-CL-V (R) surpasses MSE-CL-V (D) with a 2.32% improvement in F1 and a 3.12% increase in IoU. Similarly, for the DSM modality, the optimal performance from MSE-CL-DF (R) exhibits an enhancement of 0.13% in F1 and 0.16% in IoU over MSE-CL-DF (D).

As illustrated in Figure 4.21, the baseline (D) approach, which relies on direct HDiff maps, is more susceptible to generating false positives due to outlier values. Green clusters of examples A and B are typical examples. Conversely, the baseline (R) methodology, which employs the same model as baseline (D) but uses robust HDiff maps, demonstrates an enhanced ability to filter out such outliers, resulting in fewer false positive pixels. The application of co-learning training strategies, whether with direct HDiff maps or robust HDiff maps, significantly improves the image results achieved by ResNet-50-BIT and the DSM results achieved by SwinTransUNet. In example A of Figure 4.21, the results obtained using robust HDiff maps through various co-learning methods outperform those derived from direct HDiff maps under identical configurations. Specifically, in direct HDiff maps building change pixels are more frequently misclassified as unchanged.

In the image results, similar phenomena can be observed. Among the co-learning variations, MSE-CL-V (D/image) with direct HDiff maps achieves the highest score. However, it fails to identify a small building’s change at the bottom border of A, while the MSE-CL-V (R/image), utilizing robust HDiff maps, successfully accomplishes. Despite these

4 Summary of the Work

advantages, the use of robust HDiff maps is not without drawbacks. For example, in example C of Figure 4.21, the scenario involves a building extension where only the new section is considered a change in the ground truth. The robust HDiff map processing equates the height difference values of a narrow rectangular area with those of its adjacent extension, leading SwinTransUNet to mistakenly classify the entire area as a building change. Unfortunately, even the co-learning approach fails to rectify this mistake. In contrast, the image network trained with co-learning displays improved performance, with the MSE-CL-V(R/image) variant correctly identifying the area as unchanged.

4.3.3.4 Summary

This section presents our recent multimodal co-learning framework for building extraction from bitemporal images and DSMs, which effectively realizes domain adaptation with the help of unlabeled target data pairs. The experiments have demonstrated the effectiveness of the proposed three co-learning combinations. Similar to the performance of DSMs in the building extraction task presented in section 4.3.2, HDiff maps derived from the differences between DSMs also outperform the image modality in single-modal learning scenarios. This phenomenon further suggests that geometric features provide a more generalized and robust description of buildings compared to spectral features. For more technical details, experimental results, and comprehensive analysis, please refer to **Appendix D**.

5 Conclusion and Outlook

5.1 Conclusion

This dissertation introduces semi-supervised co-learning methodologies for remote sensing tasks of building extraction and building change detection. Proposed frameworks utilize a training mode that can exploit unlabeled data and mutual information between 2-D and 2.5-D/3-D modalities to enhance the capability of single-modal networks in challenging scenarios with limited or cross-domain labeled data. This section provides a conclusion of these works, including the following insights that deserve to be mentioned:

- **Multimodal co-learning.** Belonging to multimodal methods, our proposed co-learning frameworks and strategies have been proven to be effective in utilizing multimodal information and thereby enhance the generalization ability of image networks, point cloud networks, and HDiff map networks in different tasks. Besides, these co-learning methods circumvent the main issues of conventional data fusion techniques introduced in Chapter 1, advantageous in training single-modal networks. On the one hand, co-learning-enhanced single-modal networks are applied to single-modal testing data, suitable for testing scenarios where another modality is absent. On the other hand, single-modal networks can utilize the complete nature of each modality, avoiding irrelevant presentations that might degrade the performance.
- **Data alignment.** In this dissertation, the research data are multispectral orthophotos and photogrammetric point clouds or DSMs. Each multimodal pair originates from the same set of stereo-/multi-view images via photogrammetry techniques, ensuring perfect alignment between the different modalities. Such alignment makes this kind of pair an ideal candidate for co-learning methods, avoiding incorrect pixel-to-point mapping that might bring noise to the co-learning loss function. In reality, acquiring pairs of orthophotos and DSMs is cost-effective, leading to their widespread use in a lot of applications [59, 173, 174]. Our co-learning methods have promising potential to contribute to these multimodal applications.
- **Spectral and geometric.** Multispectral images and point clouds/DSMs provide complementary knowledge. This is the foundation of the effectiveness of co-learning methods. Through the application of deep neural networks, spectral features from images and geometric features from point clouds/DSMs can be captured and utilized to describe objects from distinct perspectives. In the cross-domain works of sections 4.3.2 and 4.3.3, the performance difference between the spectral and geometric modalities has been comprehensively compared and analyzed. In the tasks of building extraction and building change detection, geometric modality including DSMs and HDiff maps demonstrates a superior generalization ability compared to spectral modality. With the more powerful knowledge transferred from the DSM/HDiff modality by co-learning, it is natural that the performance of weaker image networks

can be improved [161]. Nevertheless, in our co-learning frameworks, the synergy is bidirectional. In most instances of sections 4.3.2 and 4.3.3, image modalities, through co-learning, can also augment the performance of point cloud and HDiff map networks. Additionally, in some co-learning-enhanced instances, the image modality may outperform the geometric modality. In both change detection experiments presented in section 4.3.3, the best quantitative results are achieved by co-learning-enhanced Siamese image methods. These phenomena further demonstrate that our proposed co-learning frameworks can effectively exchange complementary knowledge between spectral and geometric modalities, rather than unidirectional transferring.

- **Deep learning models.** In this dissertation, we select proper deep learning architectures for different tasks. For the building extraction task introduced in sections 4.3.1 and 4.3.2, widely used semantic segmentation networks are employed for processing the image modality. For the point cloud/DSM modality, efficient sparse convolutional neural networks are applied. These networks have shown to deliver reasonable performance on our experimental datasets and benefit significantly from the implementation of co-learning strategies. In the building change detection task introduced in section 4.3.3, we adopt a Siamese network for bitemporal images, while two types of architectures including the Siamese network and a single-branch semantic segmentation network are evaluated on the bitemporal DSMs and HDiff maps, respectively. Through comparative analysis, the approach of utilizing HDiff maps is more suited for the DSM modality in building change detection.
- **Synthetic data.** In **Appendix C**, we introduce a synthetic dataset named SMARS. The synthetic dataset has a similar look to real datasets. In sections 4.3.2 and 4.3.3, SMARS has been employed as the source data for exploring cross-domain building extraction and building change detection. Related experimental results demonstrate that an obvious domain gap exists between the synthetic images and real RGB images, while the difference between synthetic DSMs and real DSMs is slight. Fortunately, the performance of networks can be significantly improved by co-learning with unlabeled target data pairs. Our investigations indicate that SMARS data are feasible to be adopted to train deep learning models for realistic datasets, and proposed co-learning methods are effective means to optimize these models.

5.2 Outlook

The increasing launch of satellites equipped with diverse imaging sensors has led to unprecedented availability of spaceborne remote sensing data products, including multispectral, hyperspectral, and SAR images, as well as their derivatives like photogrammetric point clouds, DSMs, and TomoSAR point clouds. In addition, benefiting from the development of cost-effective cameras and LiDAR sensors, airborne and UAV-based image and point cloud data have become increasingly cheaper and easier to obtain. This proliferation of multisource and multimodal remote sensing data prompts an advanced discourse on how to synergistically utilize them and outperform single-modal methods. Meanwhile, the advancement of AI techniques is broadening the horizon for big data processing, potentially addressing challenging issues in complex remote sensing scenarios. As a pioneer for novel multimodal co-learning methodologies with multispectral images and photogrammetric geometric data, this dissertation investigates challenging scenarios with limited or

cross-domain labeled training data and achieves promising results in three case studies. Through studies of this dissertation and the evolving trend in remote sensing and AI technology, we outline a few challenges and opportunities for future research:

- **Multimodal co-learning in a wider context.** Except for multispectral orthophotos and photogrammetric data, many other remote sensing data modalities can provide complementary information to each other. For example, multispectral images and SAR data [175, 176], and hyperspectral images and DSMs [177] are two combinations that have been studied with other multimodal methodologies. They have the potential to be leveraged in co-learning frameworks for better performance in remote sensing tasks. Different from perfectly matched orthophotos and DSMs, original multisensor data usually do not have an accurate alignment. Under this circumstance, proper cross-modal data matching algorithms might be required as a preprocessing step to conduct qualified multimodal data pairs [178].
- **Deep learning for data regularization.** In real remote sensing applications, the diversity of domain gaps between datasets is often more than those involved in this dissertation. For complex domain gaps, extra data regularization methods are demanded to bridge the differences. For example, the resolution gap is a common issue in multisource image datasets, limiting the correct interactions between images of varying resolutions. To address this issue, additional modules like deep learning-based super resolution might be needed [104, 179]. In image-based change detection tasks, significant spectral differences or heterogeneity between pre- and post-event images can impair the effectiveness of Siamese networks. Under this circumstance, employing deep style translation [180] or feature alignment [181] strategies is crucial. Additionally, as discussed in **Appendix D**, selecting a proper height difference operation can reduce the domain gap between HDiff maps. Robust height difference [60] proves a role beneficial in this context. Recently, several deep learning-based DSM quality refinement methods have been proposed [182, 183]. They are possible to further refine the quality of HDiff maps and narrow the geometric domain gaps.
- **Further exploitation of unlabeled data.** In practical applications, a huge amount of remote sensing data are unlabeled, they are like a giant treasure to be explored. As semi-supervised learning methods, our proposed multimodal co-learning frameworks are able to utilize unlabeled data pairs in a natural way. However, a key challenge existing in semi-supervised learning methods is that not all unlabeled data can bring improved performance. Unlabeled data is useful only if it provides information that is not contained in the labeled data or cannot be easily extracted from it [184]. To conduct a more generalized framework that can effectively exploit different unlabeled data, more approaches for unlabeled data are expected. For example, self-supervised learning is a potential method to enhance our framework. By pre-training deep learning models on pretext tasks using unlabeled data, self-supervised learning potentially offers an optimized initialization for subsequent fine-tuning with downstream tasks [185].
- **Foundation models.** Recently, benefiting from the large computational power conducted in the industry, several foundation models have been developed and shown their ability to utilize multimodal data and greatly enhanced performance on downstream tasks [186, 187]. In the future, open-source pre-trained foundation models

5 *Conclusion and Outlook*

could be utilized as a constraint [188] or a teacher model [189] for developing novel methods for generic remote sensing tasks.

List of Figures

| | | |
|------|---|----|
| 1.1 | An example of the spectral confusion between buildings and vegetation in the ISPRS Potsdam dataset. | 2 |
| 2.1 | Three kinds of modalities derived from the same WorldView-2 stereo images. (a) RGB image. (b) DSM. (c) Point clouds. | 5 |
| 2.2 | A comparison between perspective projection and orthographic projection. | 6 |
| 2.3 | How 2-D convolution and 3-D convolution process the element of DSMs. | 7 |
| 3.1 | Change detection architectures. (a) Single-branch architecture. (b) Dual-branch architecture. | 15 |
| 4.1 | The difference between conventional data fusion and co-learning. (a) Early fusion. (b) Middle fusion. (c) Late fusion. (d) Multimodal co-learning. | 21 |
| 4.2 | The coverage of the Munich WorldView-2 dataset used in this dissertation. | 23 |
| 4.3 | Available data types in SMARS dataset. Scales are given as a reference for the displayed information. | 24 |
| 4.4 | The coverage of the Istanbul WorldView-2 dataset used in this dissertation. | 25 |
| 4.5 | The training phase of the proposed co-learning framework for building extraction. | 26 |
| 4.6 | 10 labeled training samples of the Munich WorldView-2 dataset. | 28 |
| 4.7 | The overview of image results obtained from 10-shot Munich WorldView-2 dataset using 10 labeled training samples and various training strategies. (a) Original image. (b) Ground truth. (c) Single-modal. (d) Standard co-learning. (e) Enhanced co-learning. | 29 |
| 4.8 | Close-up views of point cloud results obtained from 10-shot Munich WorldView-2 dataset using various training strategies. GT: Ground truth. Co-learning (S): standard co-learning. Co-learning (E): enhanced co-learning. Co-learning (E) fusion: enhanced co-learning and probability fusion. | 30 |
| 4.9 | 10 labeled training samples of the Potsdam dataset. | 31 |
| 4.10 | Close-up views of building extraction (image) results obtained from 10-shot ISPRS Potsdam dataset using various training strategies. GT: Ground truth. Co-learning (S): standard co-learning. Co-learning (E): enhanced co-learning. Co-learning (E) fusion: enhanced co-learning and probability fusion. | 32 |
| 4.11 | Close-up views of building extraction (point clouds) results obtained from 10-shot ISPRS Potsdam data set using various training strategies. (a) Ground truth. (b) Single-modality. (c) Standard co-learning. (d) Enhanced co-learning. | 32 |
| 4.12 | The workflow of the proposed multimodal and cross-domain co-learning framework for building extraction. | 35 |

List of Figures

4.13 Examples of Munich WorldView-2 (source) and ISPRS Potsdam (target) datasets. (a) A WorldView-2 image patch with RGB channels. (b) A DSM patch derived from WorldView-2 images. (c) An RGB image patch from the ISPRS Potsdam airborne dataset. (d) A DSM patch from the ISPRS Potsdam airborne dataset. 36

4.14 Building extraction results of ISPRS Potsdam target data. (a) RGB image. (b) Ground truth. Munich WorldView-2 dataset is employed as the source data. (c) HR-Net (source only). (d) HR-Net (co-learning with SparseConvNet). (e) U-Net (source only). (f) U-Net (co-learning with SparseConvNet). (g) SparseConvNet (source only). (h) SparseConvNet (co-learning with HR-Net). (i) SparseConvNet (co-learning with U-Net). 38


4.15 Building extraction results of ISPRS Potsdam target data. SParis of SMARS is employed as the source data. (a) RGB image. (b) Ground truth. (c) HR-Net (source only). (d) HR-Net (co-learning with SparseConvNet). (e) U-Net (source only). (f) U-Net (co-learning with SparseConvNet). (g) SparseConvNet (source only). (h) SparseConvNet (co-learning with HR-Net). (i) SparseConvNet (co-learning with U-Net). 39


4.16 The multimodal co-learning framework for change detection from multi-spectral images and DSMs. 41

4.17 (a) The architecture of the Siamese image network employed in this work. (b) Classifier block. The modules of the tokenizer, the transformer encoder, and the transformer decoder are from the implementation of [7]. 44

4.18 (a) The architecture of the proposed HDiff map network SwinTransUNet. (b) Convolution embedding block. (c) Transposed convolution block. 45

4.19 Examples of SParis and SVenice datasets in SMARS and the spaceborne Istanbul WorldView-2 dataset. 47

4.20 Building change detection results of SParis→SVenice. Color legend:  49

4.21 Building change detection results of SMARS→Istanbul. Color legend:  50

List of Tables

| | | |
|-----|---|----|
| 4.1 | Performance of different methods for building extraction in the experiment conducted on Munich WorldView-2 dataset utilizing only 10 labeled training sample pairs. | 30 |
| 4.2 | Performance of different methods for building extraction in the experiment conducted on ISPRS Potsdam dataset utilizing only 10 labeled training sample pairs. | 33 |
| 4.3 | Performance of single-modal learning and co-learning results in the ISPRS Potsdam data set with full labels. The results of EPUNet and ESFNet are from [167]. | 33 |
| 4.4 | Quantitative results by different methods in the building extraction experiment of spaceborne→airborne. | 37 |
| 4.5 | Quantitative results by different methods in the building extraction experiment of SParis→Potsdam. | 38 |
| 4.6 | Quantitative results of different methods on the building change detection Experiment I. The best score is shown in bold. | 48 |
| 4.7 | Quantitative results of different methods on the building change detection Experiment II. | 51 |

Bibliography

- [1] D. Mohney. Terabytes From Space: Satellite Imaging is Filling Data Centers — datacenterfrontier.com. <https://www.datacenterfrontier.com/internet-of-things/article/11429032/terabytes-from-space-satellite-imaging-is-filling-data-centers>, 2020. [Accessed 20-11-2023].
- [2] X. X. Zhu, D. Tuia, L. Mou, G.-S. Xia, L. Zhang, F. Xu, and F. Fraundorfer. Deep learning in remote sensing: A comprehensive review and list of resources. *IEEE geoscience and remote sensing magazine*, 5(4):8–36, 2017.
- [3] F. I. Diakogiannis, F. Waldner, P. Caccetta, and C. Wu. Resunet-a: A deep learning framework for semantic segmentation of remotely sensed data. *ISPRS Journal of Photogrammetry and Remote Sensing*, 162:94–114, 2020.
- [4] L. Wang, R. Li, C. Zhang, S. Fang, C. Duan, X. Meng, and P. M. Atkinson. Unetformer: A unet-like transformer for efficient semantic segmentation of remote sensing urban scene imagery. *ISPRS Journal of Photogrammetry and Remote Sensing*, 190:196–214, 2022.
- [5] Y. Shi, Q. Li, and X. X. Zhu. Building segmentation through a gated graph convolutional neural network with deep structured feature embedding. *ISPRS Journal of Photogrammetry and Remote Sensing*, 159:184–197, 2020.
- [6] Y. Xie, J. Tian, and X. X. Zhu. A co-learning method to utilize optical images and photogrammetric point clouds for building extraction. *International Journal of Applied Earth Observation and Geoinformation*, 116:103165, 2023.
- [7] H. Chen, Z. Qi, and Z. Shi. Remote sensing image change detection with transformers. *IEEE Transactions on Geoscience and Remote Sensing*, 60:1–14, 2021.
- [8] Q. Li, R. Zhong, X. Du, and Y. Du. Transunetcd: A hybrid transformer network for change detection in optical remote-sensing images. *IEEE Transactions on Geoscience and Remote Sensing*, 60:1–19, 2022.
- [9] Y. Xie, Y. Xiangtian, X. X. Zhu, and J. Tian. Multimodal co-learning for building change detection: a domain adaptation framework using vhr images and digital surface models. *IEEE Transactions on Geoscience and Remote Sensing*, 62:5402520, 2024.
- [10] L. Ma, Y. Liu, X. Zhang, Y. Ye, G. Yin, and B. A. Johnson. Deep learning in remote sensing applications: A meta-analysis and review. *ISPRS journal of photogrammetry and remote sensing*, 152:166–177, 2019.
- [11] A. Dosovitskiy, L. Beyer, A. Kolesnikov, D. Weissenborn, X. Zhai, T. Unterthiner, M. Dehghani, M. Minderer, G. Heigold, S. Gelly, et al. An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv preprint arXiv:2010.11929*, 2020.
- [12] Z. Liu, Y. Lin, Y. Cao, H. Hu, Y. Wei, Z. Zhang, S. Lin, and B. Guo. Swin transformer: Hierarchical vision transformer using shifted windows. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 10012–10022, 2021.
- [13] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, L. Kaiser, and I. Polosukhin. Attention is all you need. *Advances in neural information processing systems*, 30, 2017.
- [14] A. A. Aleissae, A. Kumar, R. M. Anwer, S. Khan, H. Cholakkal, G.-S. Xia, and F. S. Khan. Transformers in remote sensing: A survey. *Remote Sensing*, 15(7):1860, 2023.

Bibliography

- [15] Y. Xie, J. Tian, and X. X. Zhu. Linking points with labels in 3d: A review of point cloud semantic segmentation. *IEEE Geoscience and remote sensing magazine*, 8(4):38–59, 2020.
- [16] C. R. Qi, L. Yi, H. Su, and L. J. Guibas. Pointnet++: Deep hierarchical feature learning on point sets in a metric space. *Advances in neural information processing systems*, 30, 2017.
- [17] L. Landrieu and M. Simonovsky. Large-scale point cloud semantic segmentation with superpoint graphs. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 4558–4567, 2018.
- [18] H. Thomas, C. R. Qi, J.-E. Deschaud, B. Marcotegui, F. Goulette, and L. J. Guibas. Kpconv: Flexible and deformable convolution for point clouds. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 6411–6420, 2019.
- [19] C. Choy, J. Gwak, and S. Savarese. 4d spatio-temporal convnets: Minkowski convolutional neural networks. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 3075–3084, 2019.
- [20] M. Schmitt and X. X. Zhu. Data fusion and remote sensing: An ever-growing relationship. *IEEE Geoscience and Remote Sensing Magazine*, 4(4):6–23, 2016.
- [21] P. Ghamisi, B. Rasti, N. Yokoya, Q. Wang, B. Hofle, L. Bruzzone, F. Bovolo, M. Chi, K. Anders, R. Gloaguen, et al. 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, 2019.
- [22] Y. Wang, Y. Wan, Y. Zhang, B. Zhang, and Z. Gao. Imbalance knowledge-driven multi-modal network for land-cover semantic segmentation using aerial images and lidar point clouds. *ISPRS Journal of Photogrammetry and Remote Sensing*, 202:385–404, 2023.
- [23] H. Hosseinpour, F. Samadzadegan, and F. D. Javan. Cmgfnet: A deep cross-modal gated fusion network for building extraction from very high-resolution remote sensing images. *ISPRS journal of photogrammetry and remote sensing*, 184:96–115, 2022.
- [24] S. Huang, M. Usvyatsov, and K. Schindler. Indoor scene recognition in 3d. In *2020 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, pages 8041–8048. IEEE, 2020.
- [25] S. Bachhofner, A.-M. Loghin, J. Otepka, N. Pfeifer, M. Hornacek, A. Siposova, N. Schmidinger, K. Hornik, N. Schiller, O. Kähler, et al. Generalized sparse convolutional neural networks for semantic segmentation of point clouds derived from tri-stereo satellite imagery. *Remote Sensing*, 12(8):1289, 2020.
- [26] T. Peters, C. Brenner, and K. Schindler. Semantic segmentation of mobile mapping point clouds via multi-view label transfer. *ISPRS Journal of Photogrammetry and Remote Sensing*, 202:30–39, 2023.
- [27] Y. Xie and J. Tian. Multimodal co-learning: A domain adaptation method for building extraction from optical remote sensing imagery. In *2023 Joint Urban Remote Sensing Event (JURSE)*, pages 1–4. IEEE, 2023.
- [28] J. Tian, P. Reinartz, P. d’Angelo, and M. Ehlers. Region-based automatic building and forest change detection on cartosat-1 stereo imagery. *ISPRS Journal of Photogrammetry and Remote Sensing*, 79:226–239, 2013.
- [29] P. d’Angelo. Improving semi-global matching: cost aggregation and confidence measure. *The International Archives of the Photogrammetry, Remote Sensing and Spatial Information Sciences*, 41:299–304, 2016.
- [30] O. Nicolis and C. Gonzalez. Wavelet-based fractal and multifractal analysis for detecting mineral deposits using multispectral images taken by drones. In *Methods and Applications in Petroleum and Mineral Exploration and Engineering Geology*, pages 295–307. Elsevier, 2021.

- [31] J. Zhu, Y. Xu, Z. Ye, L. Hoegner, and U. Stilla. Fusion of urban 3d point clouds with thermal attributes using mls data and tir image sequences. *Infrared Physics & Technology*, 113:103622, 2021.
- [32] W. Förstner and B. P. Wrobel. *Photogrammetric computer vision*. Springer, 2016.
- [33] W. Linder. *Digital photogrammetry: theory and applications*. Springer Science & Business Media, 2013.
- [34] H. A. Al-Najjar, B. Kalantar, B. Pradhan, V. Saeidi, A. A. Halin, N. Ueda, and S. Mansor. Land cover classification from fused dsm and uav images using convolutional neural networks. *Remote Sensing*, 11(12):1461, 2019.
- [35] M. Á. Aguilar, M. del Mar Saldaña, and F. J. Aguilar. Generation and quality assessment of stereo-extracted dsm from geoeye-1 and worldview-2 imagery. *IEEE Transactions on Geoscience and Remote Sensing*, 52(2):1259–1271, 2013.
- [36] P. Chen, H. Huang, F. Ye, J. Liu, W. Li, J. Wang, Z. Wang, C. Liu, and N. Zhang. A benchmark gaofen-7 dataset for building extraction from satellite images. *Scientific Data*, 11(1):187, 2024.
- [37] Q. Li, Y. Shi, X. Huang, and X. X. Zhu. Building footprint generation by integrating convolution neural network with feature pairwise conditional random field (fpcrf). *IEEE Transactions on Geoscience and Remote Sensing*, 58(11):7502–7519, 2020.
- [38] Q. Zhu, C. Liao, H. Hu, X. Mei, and H. Li. Map-net: Multiple attending path neural network for building footprint extraction from remote sensed imagery. *IEEE Transactions on Geoscience and Remote Sensing*, 59(7):6169–6181, 2020.
- [39] J. Li, X. Huang, L. Tu, T. Zhang, and L. Wang. A review of building detection from very high resolution optical remote sensing images. *GIScience & Remote Sensing*, 59(1):1199–1225, 2022.
- [40] H. Jung, H.-S. Choi, and M. Kang. Boundary enhancement semantic segmentation for building extraction from remote sensed image. *IEEE Transactions on Geoscience and Remote Sensing*, 60:1–12, 2021.
- [41] R. Kemker, C. Salvaggio, and C. Kanan. Algorithms for semantic segmentation of multi-spectral remote sensing imagery using deep learning. *ISPRS journal of photogrammetry and remote sensing*, 145:60–77, 2018.
- [42] X. Wu, W. Li, D. Hong, J. Tian, R. Tao, and Q. Du. Vehicle detection of multi-source remote sensing data using active fine-tuning network. *ISPRS Journal of Photogrammetry and Remote Sensing*, 167:39–53, 2020.
- [43] K. Li, G. Wan, G. Cheng, L. Meng, and J. Han. Object detection in optical remote sensing images: A survey and a new benchmark. *ISPRS journal of photogrammetry and remote sensing*, 159:296–307, 2020.
- [44] T. Kattenborn, J. Leitloff, F. Schiefer, and S. Hinz. Review on convolutional neural networks (cnn) in vegetation remote sensing. *ISPRS journal of photogrammetry and remote sensing*, 173:24–49, 2021.
- [45] W. Shi, M. Zhang, R. Zhang, S. Chen, and Z. Zhan. Change detection based on artificial intelligence: State-of-the-art and challenges. *Remote Sensing*, 12(10):1688, 2020.
- [46] Y. Xu and U. Stilla. Toward building and civil infrastructure reconstruction from point clouds: A review on data and key techniques. *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, 14:2857–2885, 2021.
- [47] A. Murtiyoso, M. Veriandi, D. Suwardhi, B. Soeksmantono, and A. B. Harto. Automatic workflow for roof extraction and generation of 3d citygml models from low-cost uav image-derived point clouds. *ISPRS International Journal of Geo-Information*, 9(12):743, 2020.

Bibliography

- [48] T.-A. Teo, Y.-J. Fu, K.-W. Li, M.-C. Weng, and C.-M. Yang. Comparison between image-and surface-derived displacement fields for landslide monitoring using an unmanned aerial vehicle. *International Journal of Applied Earth Observation and Geoinformation*, 116:103164, 2023.
- [49] Q. Hu, B. Yang, S. Khalid, W. Xiao, N. Trigoni, and A. Markham. Sensaturban: Learning semantics from urban-scale photogrammetric point clouds. *International Journal of Computer Vision*, 130(2):316–343, 2022.
- [50] R. Huang, Y. Xu, L. Hoegner, and U. Stilla. Semantics-aided 3d change detection on construction sites using uav-based photogrammetric point clouds. *Automation in Construction*, 134:104057, 2022.
- [51] X. X. Zhu and M. Shahzad. Facade reconstruction using multiview spaceborne tomosar point clouds. *IEEE Transactions on Geoscience and Remote Sensing*, 52(6):3541–3552, 2013.
- [52] M. Shahzad and X. X. Zhu. Robust reconstruction of building facades for large areas using spaceborne tomosar point clouds. *IEEE Transactions on Geoscience and Remote Sensing*, 53(2):752–769, 2014.
- [53] X. X. Zhu, S. Montazeri, C. Gisinger, R. F. Hanssen, and R. Bamler. Geodetic sar tomography. *IEEE Transactions on Geoscience and Remote Sensing*, 54(1):18–35, 2015.
- [54] Y. Shi, R. Bamler, Y. Wang, and X. X. Zhu. Sar tomography at the limit: Building height reconstruction using only 3-5 tandem-x bistatic interferograms. *IEEE Transactions on Geoscience and Remote Sensing*, 58(11):8026–8037, 2020.
- [55] C. Wang, C. Wen, Y. Dai, S. Yu, and M. Liu. Urban 3d modeling with mobile laser scanning: a review. *Virtual Reality & Intelligent Hardware*, 2(3):175–212, 2020.
- [56] A. Nguyen and B. Le. 3d point cloud segmentation: A survey. In *2013 6th IEEE conference on robotics, automation and mechatronics (RAM)*, pages 225–230. IEEE, 2013.
- [57] E. Arnold, O. Y. Al-Jarrah, M. Dianati, S. Fallah, D. Oxtoby, and A. Mouzakitis. A survey on 3d object detection methods for autonomous driving applications. *IEEE Transactions on Intelligent Transportation Systems*, 20(10):3782–3795, 2019.
- [58] F. Remondino, M. G. Spera, E. Nocerino, F. Menna, F. Nex, and S. Gonizzi-Barsanti. Dense image matching: Comparisons and analyses. In *2013 Digital Heritage International Congress (DigitalHeritage)*, volume 1, pages 47–54. IEEE, 2013.
- [59] R. Qin, J. Tian, and P. Reinartz. 3d change detection—approaches and applications. *ISPRS Journal of Photogrammetry and Remote Sensing*, 122:41–56, 2016.
- [60] J. Tian, S. Cui, and P. Reinartz. Building change detection based on satellite stereo imagery and digital surface models. *IEEE Transactions on Geoscience and Remote Sensing*, 52(1):406–417, 2013.
- [61] J. Tian and J. Dezert. Fusion of multispectral imagery and dsms for building change detection using belief functions and reliabilities. *International Journal of Image and Data Fusion*, 10(1):1–27, 2019.
- [62] Z. Li, F. Liu, W. Yang, S. Peng, and J. Zhou. A survey of convolutional neural networks: analysis, applications, and prospects. *IEEE transactions on neural networks and learning systems*, 2021.
- [63] K. Simonyan and A. Zisserman. Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556*, 2014.
- [64] G. Huang, Z. Liu, L. Van Der Maaten, and K. Q. Weinberger. Densely connected convolutional networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 4700–4708, 2017.

- [65] K. He, X. Zhang, S. Ren, and J. Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778, 2016.
- [66] M. Sandler, A. Howard, M. Zhu, A. Zhmoginov, and L.-C. Chen. Mobilenetv2: Inverted residuals and linear bottlenecks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 4510–4520, 2018.
- [67] J. Long, E. Shelhamer, and T. Darrell. Fully convolutional networks for semantic segmentation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 3431–3440, 2015.
- [68] O. Ronneberger, P. Fischer, and T. Brox. U-net: Convolutional networks for biomedical image segmentation. In *Medical Image Computing and Computer-Assisted Intervention—MICCAI 2015: 18th International Conference, Munich, Germany, October 5–9, 2015, Proceedings, Part III 18*, pages 234–241. Springer, 2015.
- [69] H. Zhao, J. Shi, X. Qi, X. Wang, and J. Jia. Pyramid scene parsing network. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2881–2890, 2017.
- [70] L.-C. Chen, G. Papandreou, I. Kokkinos, K. Murphy, and A. L. Yuille. Deeplab: Semantic image segmentation with deep convolutional nets, atrous convolution, and fully connected crfs. *IEEE transactions on pattern analysis and machine intelligence*, 40(4):834–848, 2017.
- [71] L.-C. Chen, G. Papandreou, F. Schroff, and H. Adam. Rethinking atrous convolution for semantic image segmentation. *arXiv preprint arXiv:1706.05587*, 2017.
- [72] T. Xiao, Y. Liu, B. Zhou, Y. Jiang, and J. Sun. Unified perceptual parsing for scene understanding. In *Proceedings of the European conference on computer vision (ECCV)*, pages 418–434, 2018.
- [73] Z. Zhang, Q. Liu, and Y. Wang. Road extraction by deep residual u-net. *IEEE Geoscience and Remote Sensing Letters*, 15(5):749–753, 2018.
- [74] D. Jha, P. H. Smedsrud, M. A. Riegler, D. Johansen, T. De Lange, P. Halvorsen, and H. D. Johansen. Resunet++: An advanced architecture for medical image segmentation. In *2019 IEEE international symposium on multimedia (ISM)*, pages 225–2255. IEEE, 2019.
- [75] W. Wang, E. Xie, X. Li, D.-P. Fan, K. Song, D. Liang, T. Lu, P. Luo, and L. Shao. Pyramid vision transformer: A versatile backbone for dense prediction without convolutions. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 568–578, 2021.
- [76] E. Xie, W. Wang, Z. Yu, A. Anandkumar, J. M. Alvarez, and P. Luo. Segformer: Simple and efficient design for semantic segmentation with transformers. *Advances in Neural Information Processing Systems*, 34:12077–12090, 2021.
- [77] L. Gao, H. Liu, M. Yang, L. Chen, Y. Wan, Z. Xiao, and Y. Qian. Stransfuse: Fusing swin transformer and convolutional neural network for remote sensing image semantic segmentation. *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, 14:10990–11003, 2021.
- [78] C. Zhang, L. Wang, S. Cheng, and Y. Li. Swinsunet: Pure transformer network for remote sensing image change detection. *IEEE Transactions on Geoscience and Remote Sensing*, 60:1–13, 2022.
- [79] X. He, Y. Zhou, J. Zhao, D. Zhang, R. Yao, and Y. Xue. Swin transformer embedding unet for remote sensing image semantic segmentation. *IEEE Transactions on Geoscience and Remote Sensing*, 60:1–15, 2022.
- [80] M. Contributors. MMSegmentation: Openmmlab semantic segmentation toolbox and benchmark. <https://github.com/open-mmlab/mms Segmentation>, 2020.

Bibliography

- [81] K. Han, Y. Wang, H. Chen, X. Chen, J. Guo, Z. Liu, Y. Tang, A. Xiao, C. Xu, Y. Xu, et al. A survey on vision transformer. *IEEE transactions on pattern analysis and machine intelligence*, 45(1):87–110, 2022.
- [82] R. C. Daudt, B. Le Saux, and A. Boulch. Fully convolutional siamese networks for change detection. In *2018 25th IEEE International Conference on Image Processing (ICIP)*, pages 4063–4067. IEEE, 2018.
- [83] R. Zhang, H. Zhang, X. Ning, X. Huang, J. Wang, and W. Cui. Global-aware siamese network for change detection on remote sensing images. *ISPRS Journal of Photogrammetry and Remote Sensing*, 199:61–72, 2023.
- [84] M. Zhang and W. Shi. A feature difference convolutional neural network-based change detection method. *IEEE Transactions on Geoscience and Remote Sensing*, 58(10):7232–7246, 2020.
- [85] W. G. C. Bandara and V. M. Patel. A transformer-based siamese network for change detection. In *IGARSS 2022-2022 IEEE International Geoscience and Remote Sensing Symposium*, pages 207–210. IEEE, 2022.
- [86] A. Boulch, J. Guerry, B. Le Saux, and N. Audebert. Snapnet: 3d point cloud semantic labeling with 2d deep segmentation networks. *Computers & Graphics*, 71:189–198, 2018.
- [87] H. Zhao, L. Jiang, J. Jia, P. H. Torr, and V. Koltun. Point transformer. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 16259–16268, 2021.
- [88] M.-H. Guo, J.-X. Cai, Z.-N. Liu, T.-J. Mu, R. R. Martin, and S.-M. Hu. Pct: Point cloud transformer. *Computational Visual Media*, 7:187–199, 2021.
- [89] L. Tchapmi, C. Choy, I. Armeni, J. Gwak, and S. Savarese. Segcloud: Semantic segmentation of 3d point clouds. In *2017 international conference on 3D vision (3DV)*, pages 537–547. IEEE, 2017.
- [90] G. Riegler, A. Osman Ulusoy, and A. Geiger. Octnet: Learning deep 3d representations at high resolutions. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 3577–3586, 2017.
- [91] P.-S. Wang, Y. Liu, Y.-X. Guo, C.-Y. Sun, and X. Tong. O-cnn: Octree-based convolutional neural networks for 3d shape analysis. *ACM Transactions On Graphics (TOG)*, 36(4):1–11, 2017.
- [92] B. Graham, M. Engelcke, and L. Van Der Maaten. 3d semantic segmentation with sub-manifold sparse convolutional networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 9224–9232, 2018.
- [93] C. R. Qi, H. Su, K. Mo, and L. J. Guibas. Pointnet: Deep learning on point sets for 3d classification and segmentation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 652–660, 2017.
- [94] O. Vinyals, S. Bengio, and M. Kudlur. Order matters: Sequence to sequence for sets. *arXiv preprint arXiv:1511.06391*, 2015.
- [95] F. Engelmann, T. Kontogianni, A. Hermans, and B. Leibe. Exploring spatial context for 3d semantic segmentation of point clouds. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 716–724, 2017.
- [96] H. Zhao, L. Jiang, C.-W. Fu, and J. Jia. Pointweb: Enhancing local neighborhood features for point cloud processing. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 5565–5573, 2019.
- [97] Y. Li, R. Bu, M. Sun, W. Wu, X. Di, and B. Chen. Pointcnn: Convolution on x-transformed points. *Advances in neural information processing systems*, 31, 2018.

- [98] W. Wu, Z. Qi, and L. Fuxin. Pointconv: Deep convolutional networks on 3d point clouds. In *Proceedings of the IEEE/CVF Conference on computer vision and pattern recognition*, pages 9621–9630, 2019.
- [99] M. Reba and K. C. Seto. A systematic review and assessment of algorithms to detect, characterize, and monitor urban land change. *Remote sensing of environment*, 242:111739, 2020.
- [100] M. Ghanea, P. Moallem, and M. Momeni. Building extraction from high-resolution satellite images in urban areas: Recent methods and strategies against significant challenges. *International journal of remote sensing*, 37(21):5234–5248, 2016.
- [101] C. Wang, M. Ferrando, F. Causone, X. Jin, X. Zhou, and X. Shi. Data acquisition for urban building energy modeling: A review. *Building and Environment*, 217:109056, 2022.
- [102] S. Crommelinck, R. Bennett, M. Gerke, F. Nex, M. Y. Yang, and G. Vosselman. Review of automatic feature extraction from high-resolution optical sensor data for uav-based cadastral mapping. *Remote Sensing*, 8(8):689, 2016.
- [103] V. Badrinarayanan, A. Kendall, and R. Cipolla. Segnet: A deep convolutional encoder-decoder architecture for image segmentation. *IEEE transactions on pattern analysis and machine intelligence*, 39(12):2481–2495, 2017.
- [104] J. Wang, K. Sun, T. Cheng, B. Jiang, C. Deng, Y. Zhao, D. Liu, Y. Mu, M. Tan, X. Wang, et al. Deep high-resolution representation learning for visual recognition. *IEEE transactions on pattern analysis and machine intelligence*, 43(10):3349–3364, 2020.
- [105] S. Ji, S. Wei, and M. Lu. Fully convolutional networks for multisource building extraction from an open aerial and satellite imagery data set. *IEEE Transactions on geoscience and remote sensing*, 57(1):574–586, 2018.
- [106] J. Cai and Y. Chen. Mha-net: Multipath hybrid attention network for building footprint extraction from high-resolution remote sensing imagery. *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, 14:5807–5817, 2021.
- [107] Y. Zhou, Z. Chen, B. Wang, S. Li, H. Liu, D. Xu, and C. Ma. Bomsc-net: Boundary optimization and multi-scale context awareness based building extraction from high-resolution remote sensing imagery. *IEEE Transactions on Geoscience and Remote Sensing*, 60:1–17, 2022.
- [108] F. Chen, N. Wang, B. Yu, and L. Wang. Res2-unet, a new deep architecture for building detection from high spatial resolution images. *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, 15:1494–1501, 2022.
- [109] Y. Quan, A. Yu, X. Cao, C. Qiu, X. Zhang, B. Liu, and P. He. Building extraction from remote sensing images with dog as prior constraint. *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, 15:6559–6570, 2022.
- [110] Q. Li, L. Mou, Y. Hua, Y. Shi, and X. X. Zhu. Building footprint generation through convolutional neural networks with attraction field representation. *IEEE Transactions on Geoscience and Remote Sensing*, 60:1–17, 2021.
- [111] C. Qiu, H. Li, W. Guo, X. Chen, A. Yu, X. Tong, and M. Schmitt. Transferring transformer-based models for cross-area building extraction from remote sensing images. *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, 15:4104–4116, 2022.
- [112] H. Huang, J. Liu, and R. Wang. Easy-net: A lightweight building extraction network based on building features. *IEEE Transactions on Geoscience and Remote Sensing*, 2023.
- [113] L. Wang, S. Fang, X. Meng, and R. Li. Building extraction with vision transformer. *IEEE Transactions on Geoscience and Remote Sensing*, 60:1–11, 2022.

Bibliography

- [114] Z. Chen, Y. Luo, J. Wang, J. Li, C. Wang, and D. Li. Dpenet: Dual-path extraction network based on cnn and transformer for accurate building and road extraction. *International Journal of Applied Earth Observation and Geoinformation*, 124:103510, 2023.
- [115] L. Xu, Y. Li, J. Xu, Y. Zhang, and L. Guo. Bctnet: Bi-branch cross-fusion transformer for building footprint extraction. *IEEE Transactions on Geoscience and Remote Sensing*, 61:1–14, 2023.
- [116] J. Kang, Z. Wang, R. Zhu, X. Sun, R. Fernandez-Beltran, and A. Plaza. Picoco: Pixelwise contrast and consistency learning for semisupervised building footprint segmentation. *IEEE journal of selected topics in applied earth observations and remote sensing*, 14:10548–10559, 2021.
- [117] Q. Li, Y. Shi, and X. X. Zhu. Semi-supervised building footprint generation with feature and output consistency training. *IEEE Transactions on Geoscience and Remote Sensing*, 60:1–17, 2022.
- [118] X. Yan, L. Shen, J. Pan, J. Wang, C. Chen, and Z. Li. Alnet: Auxiliary learning based network for weakly supervised building extraction from high-resolution remote sensing images. *IEEE Transactions on Geoscience and Remote Sensing*, 2023.
- [119] H. Chen, L. Cheng, Q. Zhuang, K. Zhang, N. Li, L. Liu, and Z. Duan. Structure-aware weakly supervised network for building extraction from remote sensing images. *IEEE Transactions on Geoscience and Remote Sensing*, 60:1–12, 2022.
- [120] Performance evaluation of fusion techniques for cross-domain building rooftop segmentation. *The International Archives of the Photogrammetry, Remote Sensing and Spatial Information Sciences*, 43:501–508, 2022.
- [121] D. Peng, H. Guan, Y. Zang, and L. Bruzzone. Full-level domain adaptation for building extraction in very-high-resolution optical remote-sensing images. *IEEE Transactions on Geoscience and Remote Sensing*, 60:1–17, 2021.
- [122] Q. Li, L. Mou, Y. Hua, Y. Shi, and X. X. Zhu. Crossgeonet: A framework for building footprint generation of label-scarce geographical regions. *International Journal of Applied Earth Observation and Geoinformation*, 111:102824, 2022.
- [123] I. R. Hegazy and M. R. Kaloop. Monitoring urban growth and land use change detection with gis and remote sensing techniques in daqahlia governorate egypt. *International Journal of Sustainable Built Environment*, 4(1):117–124, 2015.
- [124] Z. Zheng, Y. Zhong, J. Wang, A. Ma, and L. Zhang. Building damage assessment for rapid disaster response with a deep object-based semantic change detection framework: From natural disasters to man-made disasters. *Remote Sensing of Environment*, 265:112636, 2021.
- [125] Z. Ali, A. Tuladhar, and J. Zevenbergen. An integrated approach for updating cadastral maps in pakistan using satellite remote sensing data. *International Journal of Applied Earth Observation and Geoinformation*, 18:386–398, 2012.
- [126] K. Yang, G.-S. Xia, Z. Liu, B. Du, W. Yang, M. Pelillo, and L. Zhang. Asymmetric siamese networks for semantic change detection in aerial images. *IEEE Transactions on Geoscience and Remote Sensing*, 60:1–18, 2021.
- [127] Z. Zheng, Y. Zhong, S. Tian, A. Ma, and L. Zhang. Changemask: Deep multi-task encoder-transformer-decoder architecture for semantic change detection. *ISPRS Journal of Photogrammetry and Remote Sensing*, 183:228–239, 2022.
- [128] P. F. Alcantarilla, S. Stent, G. Ros, R. Arroyo, and R. Gherardi. Street-view change detection with deconvolutional networks. *Autonomous Robots*, 42:1301–1322, 2018.
- [129] D. Peng, Y. Zhang, and H. Guan. End-to-end change detection for high resolution satellite images using improved unet++. *Remote Sensing*, 11(11):1382, 2019.

- [130] H. Jiang, M. Peng, Y. Zhong, H. Xie, Z. Hao, J. Lin, X. Ma, and X. Hu. A survey on deep learning-based change detection from high-resolution remote sensing images. *Remote Sensing*, 14(7):1552, 2022.
- [131] C. Zhang, P. Yue, D. Tapete, L. Jiang, B. Shangguan, L. Huang, and G. Liu. A deeply supervised image fusion network for change detection in high resolution bi-temporal remote sensing images. *ISPRS Journal of Photogrammetry and Remote Sensing*, 166:183–200, 2020.
- [132] P. Chen, B. Zhang, D. Hong, Z. Chen, X. Yang, and B. Li. Fccdnet: Feature constraint network for vhr image change detection. *ISPRS Journal of Photogrammetry and Remote Sensing*, 187:101–119, 2022.
- [133] W. Zhao, L. Mou, J. Chen, Y. Bo, and W. J. Emery. Incorporating metric learning and adversarial network for seasonal invariant change detection. *IEEE Transactions on Geoscience and Remote Sensing*, 58(4):2720–2731, 2019.
- [134] B. Hou, Q. Liu, H. Wang, and Y. Wang. From w-net to edgan: Bitemporal change detection via deep learning techniques. *IEEE Transactions on Geoscience and Remote Sensing*, 58(3):1790–1802, 2019.
- [135] M. Zhang, G. Xu, K. Chen, M. Yan, and X. Sun. Triplet-based semantic relation learning for aerial remote sensing image change detection. *IEEE Geoscience and Remote Sensing Letters*, 16(2):266–270, 2018.
- [136] X. Peng, R. Zhong, Z. Li, and Q. Li. Optical remote sensing image change detection based on attention mechanism and image difference. *IEEE Transactions on Geoscience and Remote Sensing*, 59(9):7296–7307, 2020.
- [137] H. Chen, C. Wu, B. Du, L. Zhang, and L. Wang. Change detection in multisource vhr images via deep siamese convolutional multiple-layers recurrent neural network. *IEEE Transactions on Geoscience and Remote Sensing*, 58(4):2848–2864, 2019.
- [138] T. Yan, Z. Wan, P. Zhang, G. Cheng, and H. Lu. Transy-net: Learning fully transformer networks for change detection of remote sensing images. *IEEE Transactions on Geoscience and Remote Sensing*, 2023.
- [139] X. Zhang, S. Cheng, L. Wang, and H. Li. Asymmetric cross-attention hierarchical network based on cnn and transformer for bitemporal remote sensing images change detection. *IEEE Transactions on Geoscience and Remote Sensing*, 61:1–15, 2023.
- [140] Z. Yang, W. Jiang, B. Xu, Q. Zhu, S. Jiang, and W. Huang. A convolutional neural network-based 3d semantic labeling method for als point clouds. *Remote Sensing*, 9(9):936, 2017.
- [141] Z. Yang, B. Tan, H. Pei, and W. Jiang. Segmentation and multi-scale convolutional neural network-based classification of airborne laser scanner data. *Sensors*, 18(10):3347, 2018.
- [142] M. Yousefhussein, D. J. Kelbe, E. J. Ientilucci, and C. Salvaggio. A multi-scale fully convolutional network for semantic labeling of 3d point clouds. *ISPRS journal of photogrammetry and remote sensing*, 143:191–204, 2018.
- [143] R. Huang, Y. Xu, D. Hong, W. Yao, P. Ghamisi, and U. Stilla. Deep point embedding for urban classification using als point clouds: A new perspective from local to global. *ISPRS Journal of Photogrammetry and Remote Sensing*, 163:62–81, 2020.
- [144] R. Huang, Y. Xu, and U. Stilla. Granet: Global relation-aware attentional network for semantic segmentation of als point clouds. *ISPRS Journal of Photogrammetry and Remote Sensing*, 177:1–20, 2021.
- [145] T. Zeng, F. Luo, T. Guo, X. Gong, J. Xue, and H. Li. Recurrent residual dual attention network for airborne laser scanning point cloud semantic segmentation. *IEEE Transactions on Geoscience and Remote Sensing*, 2023.

Bibliography

- [146] Y. Lin, G. Vosselman, Y. Cao, and M. Y. Yang. Local and global encoder network for semantic segmentation of airborne laser scanning point clouds. *ISPRS journal of photogrammetry and remote sensing*, 176:151–168, 2021.
- [147] S. Oehmcke, L. Li, K. Trepekli, J. C. Revenga, T. Nord-Larsen, F. Gieseke, and C. Igel. Deep point cloud regression for above-ground forest biomass estimation from airborne lidar. *Remote Sensing of Environment*, 302:113968, 2024.
- [148] J. Hu, L. Shen, and G. Sun. Squeeze-and-excitation networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 7132–7141, 2018.
- [149] U. Stilla and Y. Xu. Change detection of urban objects using 3d point clouds: A review. *ISPRS Journal of Photogrammetry and Remote Sensing*, 197:228–255, 2023.
- [150] I. de Gélis, S. Lefèvre, and T. Corpetti. Siamese kpconv: 3d multiple change detection from raw point clouds using deep learning. *ISPRS Journal of Photogrammetry and Remote Sensing*, 197:274–291, 2023.
- [151] I. de Gélis, S. Lefèvre, and T. Corpetti. 3d urban change detection with point cloud siamese networks. *International Archives of the Photogrammetry, Remote Sensing and Spatial Information Sciences*, 43:879–886, 2021.
- [152] L. Fang, J. Liu, Y. Pan, Z. Ye, and X. Tong. Semantic supported urban change detection using als point clouds. *International Journal of Applied Earth Observation and Geoinformation*, 118:103271, 2023.
- [153] Y. Xie, K. Schindler, J. Tian, and X. X. Zhu. Exploring cross-city semantic segmentation of als point clouds. *International Archives of the Photogrammetry, Remote Sensing and Spatial Information Sciences*, 43:247–254, 2021.
- [154] M. Howe, B. Repasky, and T. Payne. Effective utilisation of multiple open-source datasets to improve generalisation performance of point cloud segmentation models. In *2022 International Conference on Digital Image Computing: Techniques and Applications (DICTA)*, pages 1–8. IEEE, 2022.
- [155] M. F. Reyes, Y. Xie, X. Yuan, P. d’Angelo, F. Kurz, D. Cerra, and J. Tian. A 2d/3d multimodal data simulation approach with applications on urban semantic segmentation, building extraction and change detection. *ISPRS Journal of Photogrammetry and Remote Sensing*, 205:74–97, 2023.
- [156] Y. Sun, Z. Fu, C. Sun, Y. Hu, and S. Zhang. Deep multimodal fusion network for semantic segmentation using remote sensing image and lidar data. *IEEE Transactions on Geoscience and Remote Sensing*, 60:1–18, 2021.
- [157] P. Zhang, P. Du, C. Lin, X. Wang, E. Li, Z. Xue, and X. Bai. A hybrid attention-aware fusion network (hafnet) for building extraction from high-resolution imagery and lidar data. *Remote Sensing*, 12(22):3764, 2020.
- [158] X. Yuan, J. Tian, and P. Reinartz. Building change detection based on deep learning and belief function. In *2019 Joint Urban Remote Sensing Event (JURSE)*, pages 1–4. IEEE, 2019.
- [159] K. Bittner, F. Adam, S. Cui, M. Körner, and P. Reinartz. Building footprint extraction from vhr remote sensing images combined with normalized dsms using fused fully convolutional networks. *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, 11(8):2615–2629, 2018.
- [160] Y.-C. Li, H.-C. Li, W.-S. Hu, and H.-L. Yu. Dspcanet: Dual-channel scale-aware segmentation network with position and channel attentions for high-resolution aerial images. *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, 14:8552–8565, 2021.

- [161] T. Baltrušaitis, C. Ahuja, and L.-P. Morency. Multimodal machine learning: A survey and taxonomy. *IEEE transactions on pattern analysis and machine intelligence*, 41(2):423–443, 2018.
- [162] D. Hong, L. Gao, N. Yokoya, J. Yao, J. Chanussot, Q. Du, and B. Zhang. More diverse means better: Multimodal deep learning meets remote-sensing imagery classification. *IEEE Transactions on Geoscience and Remote Sensing*, 59(5):4340–4354, 2020.
- [163] J. Huang, X. Zhang, Q. Xin, Y. Sun, and P. Zhang. Automatic building extraction from high-resolution aerial images and lidar data using gated residual refinement network. *ISPRS journal of photogrammetry and remote sensing*, 151:91–105, 2019.
- [164] S. Zhou, Y. Feng, S. Li, D. Zheng, F. Fang, Y. Liu, and B. Wan. Dsm-assisted unsupervised domain adaptive network for semantic segmentation of remote sensing imagery. *IEEE Transactions on Geoscience and Remote Sensing*, 2023.
- [165] A. Rahate, R. Walambe, S. Ramanna, and K. Kotecha. Multimodal co-learning: Challenges, applications with datasets, recent advances and future directions. *Information Fusion*, 81:203–239, 2022.
- [166] M. Jaritz, T.-H. Vu, R. d. Charette, E. Wirbel, and P. Pérez. xmuda: Cross-modal unsupervised domain adaptation for 3d semantic segmentation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 12605–12614, 2020.
- [167] Q. Li, S. Zorzi, Y. Shi, F. Fraundorfer, and X. X. Zhu. Reggan: An end-to-end network for building footprint generation with boundary regularization. *Remote Sensing*, 14(8):1835, 2022.
- [168] H. Guo, Q. Shi, A. Marinoni, B. Du, and L. Zhang. Deep building footprint update network: A semi-supervised method for updating existing building footprint from bi-temporal remote sensing images. *Remote Sensing of Environment*, 264:112589, 2021.
- [169] J. Lin, W. Jing, H. Song, and G. Chen. Esfnet: Efficient network for building extraction from high-resolution aerial images. *IEEE Access*, 7:54285–54294, 2019.
- [170] H. Chen, C. Wu, B. Du, and L. Zhang. Dsdanet: Deep siamese domain adaptation convolutional neural network for cross-domain change detection. *arXiv preprint arXiv:2006.09225*, 2020.
- [171] P. J. S. Vega, G. A. O. P. da Costa, R. Q. Feitosa, M. X. O. Adarme, C. A. de Almeida, C. Heipke, and F. Rottensteiner. An unsupervised domain adaptation approach for change detection and its application to deforestation mapping in tropical biomes. *ISPRS Journal of Photogrammetry and Remote Sensing*, 181:113–128, 2021.
- [172] Y. Cui, R. Chen, W. Chu, L. Chen, D. Tian, Y. Li, and D. Cao. Deep learning for image and point cloud fusion in autonomous driving: A review. *IEEE Transactions on Intelligent Transportation Systems*, 23(2):722–739, 2021.
- [173] J. Höhle. Automated mapping of buildings through classification of dsm-based ortho-images and cartographic enhancement. *International Journal of Applied Earth Observation and Geoinformation*, 95:102237, 2021.
- [174] C. Lissak, A. Bartsch, M. De Michele, C. Gomez, O. Maquaire, D. Raucoules, and T. Roul-land. Remote sensing for assessing landslides and associated hazards. *Surveys in Geophysics*, 41:1391–1435, 2020.
- [175] P. Ebel, A. Meraner, M. Schmitt, and X. X. Zhu. Multisensor data fusion for cloud removal in global and all-season sentinel-2 imagery. *IEEE Transactions on Geoscience and Remote Sensing*, 59(7):5866–5878, 2020.

Bibliography

- [176] Y. Wang, X. X. Zhu, B. Zeisl, and M. Pollefeys. Fusing meter-resolution 4-d insar point clouds and optical images for semantic urban infrastructure monitoring. *IEEE Transactions on Geoscience and Remote Sensing*, 55(1):14–26, 2016.
- [177] D. Hong, J. Hu, J. Yao, J. Chanussot, and X. X. Zhu. Multimodal remote sensing benchmark datasets for land cover classification with a shared and specific feature learning model. *ISPRS Journal of Photogrammetry and Remote Sensing*, 178:68–80, 2021.
- [178] L. H. Hughes, D. Marcos, S. Lobry, D. Tuia, and M. Schmitt. A deep learning framework for matching of sar and optical imagery. *ISPRS Journal of Photogrammetry and Remote Sensing*, 169:166–179, 2020.
- [179] X. Dong, X. Sun, X. Jia, Z. Xi, L. Gao, and B. Zhang. Remote sensing image super-resolution using novel dense-sampling networks. *IEEE Transactions on Geoscience and Remote Sensing*, 59(2):1618–1633, 2020.
- [180] X. Li, Z. Du, Y. Huang, and Z. Tan. A deep translation (gan) based change detection network for optical and sar remote sensing images. *ISPRS Journal of Photogrammetry and Remote Sensing*, 179:14–34, 2021.
- [181] C. Zhang, Y. Feng, L. Hu, D. Tapete, L. Pan, Z. Liang, F. Cigna, and P. Yue. A domain adaptation neural network for change detection with heterogeneous optical and sar remote sensing images. *International Journal of Applied Earth Observation and Geoinformation*, 109:102769, 2022.
- [182] K. Bittner, P. d’Angelo, M. Körner, and P. Reinartz. Dsm-to-lod2: Spaceborne stereo digital surface model refinement. *Remote Sensing*, 10(12):1926, 2018.
- [183] C. Stucker and K. Schindler. Resdepth: A deep residual prior for 3d reconstruction from high-resolution satellite images. *ISPRS Journal of Photogrammetry and Remote Sensing*, 183:560–580, 2022.
- [184] J. E. Van Engelen and H. H. Hoos. A survey on semi-supervised learning. *Machine learning*, 109(2):373–440, 2020.
- [185] Y. Wang, C. M. Albrecht, N. A. A. Braham, L. Mou, and X. X. Zhu. Self-supervised learning in remote sensing: A review. *IEEE Geoscience and Remote Sensing Magazine*, 10(4):213–247, 2022.
- [186] X. Guo, J. Lao, B. Dang, Y. Zhang, L. Yu, L. Ru, L. Zhong, Z. Huang, K. Wu, D. Hu, et al. Skysense: A multi-modal remote sensing foundation model towards universal interpretation for earth observation imagery. *arXiv preprint arXiv:2312.10115*, 2023.
- [187] A. Kirillov, E. Mintun, N. Ravi, H. Mao, C. Rolland, L. Gustafson, T. Xiao, S. Whitehead, A. C. Berg, W.-Y. Lo, et al. Segment anything. *arXiv preprint arXiv:2304.02643*, 2023.
- [188] X. Ma, Q. Wu, X. Zhao, X. Zhang, M.-O. Pun, and B. Huang. Sam-assisted remote sensing imagery semantic segmentation with object and boundary constraints. *arXiv preprint arXiv:2312.02464*, 2023.
- [189] X. He, Y. Chen, L. Huang, D. Hong, and Q. Du. Foundation model-based multimodal remote sensing data classification. *IEEE Transactions on Geoscience and Remote Sensing*, 62, 2024.

Appendices

A Yuxing Xie, Jiaojiao Tian, and Xiao Xiang Zhu. “Linking points with labels in 3D: A review of point cloud semantic segmentation.” *IEEE Geoscience and remote sensing magazine* 8.4 (2020): 38-59.

<https://doi.org/10.1109/MGRS.2019.2937630>

Linking Points With Labels in 3D: A Review of Point Cloud Semantic Segmentation

Yuxing Xie, Jiaojiao Tian, *Member, IEEE* and Xiao Xiang Zhu, *Senior Member, IEEE*

Abstract—3D Point Cloud Semantic Segmentation (PCSS) is attracting increasing interest, due to its applicability in remote sensing, computer vision and robotics, and due to the new possibilities offered by deep learning techniques. In order to provide a needed up-to-date review of recent developments in PCSS, this article summarizes existing studies on this topic. Firstly, we outline the acquisition and evolution of the 3D point cloud from the perspective of remote sensing and computer vision, as well as the published benchmarks for PCSS studies. Then, traditional and advanced techniques used for Point Cloud Segmentation (PCS) and PCSS are reviewed and compared. Finally, important issues and open questions in PCSS studies are discussed.

Index Terms—review, point cloud, segmentation, semantic segmentation, deep learning.

I. MOTIVATION

Semantic segmentation, in which pixels are associated with semantic labels, is a fundamental research challenge in image processing. Point Cloud Semantic Segmentation (PCSS) is the 3D form of semantic segmentation, in which regular or irregular distributed points in 3D space are used instead of regular distributed pixels in a 2D image. The point cloud can be acquired directly from sensors with distance measurability, or generated from stereo- or multi-view imagery. Due to recently developed stereovision algorithms and the deployment of all kinds of 3D sensors, point clouds, basic 3D data, have become easily accessible. High-quality point clouds provide a way to connect the virtual world to the real one. Specifically, they generate 2.5D/3D geometric structures, with which modeling is possible.

A. Segmentation, classification, and semantic segmentation

Research on PCSS has a long tradition involving different fields and defining distinct concepts for similar tasks. A brief clarification of some concepts is therefore necessary to avoid misunderstandings. The term PCSS is widely used in computer vision, especially in recent deep learning applications [1]–[3]. However, in photogrammetry and remote sensing, PCSS is usually called “point cloud classification” [4]–[6]. Or in some cases, this task is also called “point labeling” [7]–[9]. In this article, to avoid confusion and to make this literature review keep up with latest deep learning techniques, we refer to point cloud semantic segmentation/classification/labeling, i.e., the task of associating each point of a point cloud with a semantic label, as PCSS.

Before effective supervised learning methods were widely applied in semantic segmentation, unsupervised Point Cloud Segmentation (PCS) was a significant task for 2.5D/3D data.

PCS aims at grouping points with similar geometric/spectral characteristics without considering semantic information. In the PCSS workflow, PCS can be utilized as a presegmentation step, influencing the final results. Hence, PCS approaches are also included in this paper.

Single objects or the same classes of structures cannot be acquired from a raw point cloud directly. However, instance-level or class-level objects are required for object recognition. For example, urban planning and Building Information Modeling (BIM) need buildings and other man-made ground objects for reference [10], [11]. Forest remote sensing monitoring needs individual tree information based on their geometric structures [12], [13]. Robotics applications, like Simultaneous Localization And Mapping (SLAM), need detailed indoor objects for mapping [7], [14]. In some applications related to computer vision, such as autonomous driving, object detection, segmentation, and classification are necessary with the construction of a High Definition (HD) Map [15]. For the mentioned cases, PCSS and PCS are basic and critical tasks for 3D applications.

B. New challenges and possibilities

Papers [16] and [17] provide two of the best available reviews for PCS and PCSS, but lack detailed information, especially for PCSS. Furthermore, in the past two years, deep learning has largely driven studies in PCSS. To meet the demand of deep learning, 3D datasets have improved, both in quality and diversity. Therefore, an updated study on current PCSS techniques is necessary. This paper starts with the introduction of existing techniques to acquire point clouds and the existing benchmarks for point cloud study (section II). In section III and IV, the major categories of algorithms are reviewed, for both PCS and PCSS. In section V, some issues related to data and techniques are discussed. Section VI concludes this paper with a technical outlook.

II. AN INTRODUCTION TO POINT CLOUD

A. Point cloud data acquisition

In computer vision and remote sensing, point clouds can be acquired with four main techniques: 1) Image-derived methods; 2) Light Detection And Ranging (LiDAR) systems; 3) Red Green Blue -Depth (RGB-D) cameras; and 4) Synthetic Aperture Radar (SAR) systems. Due to the differences in survey principles and platforms, their data features and application ranges are very diverse. A brief introduction to these techniques is provided below.

1) *Image-derived point cloud*: Image-derived methods generate a point cloud indirectly from spectral imagery. First, they acquire stereo- or multi-view images through electro-optical systems, e.g., cameras. Then they calculate 3D isolated point information according to principles in photogrammetry or computer vision theory, either automatically or semi-automatically [18], [19]. Based on distinct platforms, stereo- and multi-view image-derived systems can be divided into airborne, spaceborne, UAV-based, and close-range categories.

Early aerial traditional photogrammetry produced 3D points with semi-automatic human-computer interaction in digital photogrammetric systems, characterized by strict geometric constraints and high survey accuracy [20]. To produce this type of point data was time expensive due to many manual works. Therefore it was not feasible to generate dense points for large areas in this way. In the surveying and remote sensing industry, those early-form “point clouds” were used in mapping and producing Digital Surface Models (DSMs) and Digital Elevation Models (DEMs). Due to the limitation of image resolution and the ability of processing multi-view images, traditional photogrammetry could only acquire close to nadir views with few building façades from aerial/satellite platforms, which only generated a 2.5D point cloud rather than full 3D. At this stage, photogrammetry principles could also be applied as close-range photogrammetry in order to obtain points from certain objects or small-area scenes, but manual editing would also be necessary in the point cloud generating procedure.

Dense matching [21]–[23], Multiple View Stereovision (MVS) [24], [25], and Structure from Motion (SfM) [19], [26], [27], changed the image-derived point cloud, and opened the era of multiple view stereovision. SfM can estimate camera positions and orientations automatically, making it capable of processing multiview images simultaneously, while dense matching and MVS algorithms provide the ability to generate large volume of point clouds. In recent years, city-scale full 3D dense point clouds can be acquired easily through an oblique photography technique based on SfM and MVS. However, the quality of point clouds from SfM and MVS is not as good as those generated by traditional photogrammetry or LiDAR techniques, and it is especially unreliable for large regions [28].

Compared to airborne photogrammetry, satellite stereo system is disadvantaged in terms of spatial resolution and availability of multi-view imagery. However, satellite cameras are able to map large regions in a short period of time with relatively lower cost. Also due to new dense matching techniques and their improved spatial resolution, satellite imagery is becoming an important data source for image-derived point clouds.

2) *LiDAR point cloud*: Light Detection And Ranging (LiDAR) is a surveying and remote sensing technique. As its name suggests, LiDAR utilizes laser energy to measure the distance between the sensor and the object to be surveyed [29]. Most LiDAR systems are pulse-based. The basic principle of pulse-based measuring is to emit a pulse of laser energy and then measure the time it takes for that energy to travel to a target. Depending on sensors and platforms, the point

density or resolution varies greatly, from less than 10 points per m^2 (pts/m^2) to thousands of points per m^2 [30]. Based on platforms, LiDAR systems are divided into airborne LiDAR scanning (ALS), terrestrial LiDAR scanning (TLS), mobile LiDAR scanning (MLS) and unmanned LiDAR scanning (ULS) systems.

ALS operates from airborne platforms. Early ALS LiDAR data are 2.5D point clouds, which are similar to traditional photogrammetric point clouds. The density of ALS points is normally low, as the distance from an airborne platform to the ground is large. In comparison to traditional photogrammetry, ALS point clouds are more expensive to acquire and normally contain no spectral information. Vaihingen point cloud semantic labeling dataset [31] is a typical ALS benchmark dataset. Multispectral airborne LiDAR is a special form of an ALS system that obtains data using different wavelengths. Multispectral LiDAR performs well for the extraction of water, vegetation and shadows, but the data are not easily available [32], [33].

TLS, also called static LiDAR scanning, scans with a tripod-mounted stationary sensor. Since it is used in a middle- or close-range environment, the point cloud density is very high. Its advantage is its ability to provide real, high quality 3D models. Until now TLS has been commonly used for modeling small urban or forest sites, and heritage or artwork documentation. Semantic3D.net [34] is a typical TLS benchmark dataset.

MLS operates from a moving vehicle on the ground, with the most common platforms being cars. Currently, research and development on autonomous driving is a hot topic, for which HD maps are essential. The generation of HD maps is therefore the most significant application for MLS. Several mainstream point cloud benchmark datasets belong to MLS [35], [36].

ULS systems are usually deployed on drones or other unmanned vehicles. Since they are relatively cheap and very flexible, this recent addition to the LiDAR family is currently becoming more and more popular. Compared to ALS, where the platform is working above the objects, ULS can provide a shorter-distance LiDAR survey application, collecting denser point clouds with higher accuracy. Thanks to the small size and light weight of its platform, ULS offers high operational flexibility. Therefore, in addition to traditional LiDAR tasks (e.g., acquiring DSMs), ULS has advantages in agriculture and forestry surveying, disaster monitoring and mining surveying [37]–[39].

For LiDAR scanning, since the system is always moving with the platform, it is necessary to combine points’ positions with Global Navigation Satellite System (GNSS) and Inertial Measurement Unit (IMU) data to ensure a high-quality matching point cloud. Until now, LiDAR has been the most important data source for point cloud research and has been used to provide ground truth to evaluate the quality of other point clouds.

3) *RGB-D point cloud*: An RGB-D camera is a type of sensor that can acquire both RGB and depth information. There are three kinds of RGB-D sensors, based on different principles: (a) structured light [40], (b) stereo [41], and (c) time of flight [42]. Similar to LiDAR, the RGB-D camera can

measure the distance between the camera to the objects, but pixel-wise. However, an RGB-D sensor is much cheaper than a LiDAR system. Microsoft's Kinect is a well-known and widely used RGB-D sensor [40], [42]. In an RGB-D camera, relative orientation elements between or among different sensors are calibrated and known, so co-registered synchronized RGB images and depth maps can be easily acquired. Obviously, the point cloud is not the direct product of RGB-D scanning. But since the position of the camera's center point is known, the 3D space position of each pixel in a depth map can be easily obtained, and then directly used to generate the point cloud. RGB-D cameras have three main applications: object tracking, human pose or signature recognition, and SLAM-based environment reconstruction. Since mainstream RGB-D sensors are close-range, even much closer than TLS, they are usually employed in indoor environments. Several mainstream indoor point cloud segmentation benchmarks are RGB-D data [43], [44].

4) *SAR point cloud*: Interferometric Synthetic Aperture Radar (InSAR), a radar technique crucial to remote sensing, generates maps of surface deformation or digital elevation based on the comparison of multiple SAR image pairs. A rising star, InSAR-based point cloud has showed its value over the past few years and is creating new possibilities for point cloud applications [45]–[49]. Synthetic Aperture Radar tomography (TomoSAR) and Persistent Scatterer Interferometry (PSI) are two major techniques that generate point clouds with InSAR, extending the principle of SAR into 3D [50], [51]. Compared with PSI, TomoSAR's advantage is its detailed reconstruction and monitoring of urban areas, especially man-made infrastructure [51]. The TomoSAR point cloud has a point density that is comparable to ALS LiDAR [52], [53]. These point clouds can be employed for applications in building reconstruction in urban areas, as they have the following features [46]:

(a) TomoSAR point clouds reconstructed from spaceborne data have a moderate 3D positioning accuracy on the order of 1 m [54], even able to reach a decimeter level by geocoding error correction techniques [55], while ALS LiDAR provides accuracy typically on the order of 0.1 m [56].

(b) Due to their coherent imaging nature and side-looking geometry, TomoSAR point clouds emphasize different objects with respect to LiDAR systems: a) The side-looking SAR geometry enables TomoSAR point clouds to possess rich façade information: results using pixel-wise TomoSAR for the high-resolution reconstruction of a building complex with a very high level of detail from spaceborne SAR data are presented in [57]; b) temporarily incoherent objects, e.g., trees, cannot be reconstructed from multipass spaceborne SAR image stacks; and c) to obtain the full structure of individual buildings from space, façade reconstruction using TomoSAR point clouds from multiple viewing angles is required [45], [58].

(c) Complementary to LiDAR and optical sensors, SAR is so far the only sensor capable of providing fourth dimension information from space, i.e., temporal deformation of the building complex [59], and microwave scattering properties of the façade reflect geometrical and material features.

InSAR point clouds have two main shortcomings that affect their accuracy: (1) Due to limited orbit spread and the small number of images, the location error of TomoSAR points is highly anisotropic, with an elevation error typically one or two orders of magnitude higher than in range and azimuth; (2) Due to multiple scattering, ghost scatterers may be generated, appearing as outliers far away from a realistic 3D position [60].

Compared with the aforementioned image-derived, LiDAR-based, and RGB-D-based point cloud, the data from SAR have not yet been widely used for studies and applications. However, mature SAR satellites, such as TerraSAR-X, have collected rich global SAR data, which are available for InSAR-based reconstruction at global scale [61]. Hence, the SAR point cloud can be expected to play a conspicuous role in the future.

B. Point cloud characters

From the perspective of sensor development and various applications, we have cataloged point clouds into: (a) sparse (less than 20 pts/m^2), (b) dense (hundreds of pts/m^2), and (c) multi-source.

(a) In their early stage, which was limited by matching techniques and computation ability, photogrammetric point clouds were sparse and small in volume. At that time, laser scanning systems had limited types and were not widely used. ALS point clouds, mainstream laser data, were also sparse. Limited by the point density, point clouds at this stage were not able to represent land covers in object level. Therefore there was no specific demand for precise PCS or PCSS. Researchers mainly focused on 3D mapping (DEM generation), and simple object extraction (e.g., rooftops).

(b) Computer vision algorithms, such as dense matching, and high-efficiency point cloud generators, such as various LiDAR systems and RGB-D sensors, opened the big data era of the dense point cloud. Dense and large-volume point clouds created more possibilities in 3D applications but also had a stronger desire for practicable algorithms. PCS and PCSS were newly proposed and became increasingly necessary, since only a class-level or instance-level point cloud further connect virtual word to the real one. Both computer vision and remote sensing need PCS and PCSS solutions to develop class-level interactive applications.

(c) From the perspective of general computer vision, research on the point cloud and its related algorithms remain at stage (b). However, as a benefit to the development of spaceborne platforms and multi-sensors, remote sensing researchers developed a new understanding of the point cloud. New-generation point clouds, such as satellite photogrammetric point clouds and TomoSAR point clouds, stimulated demand for relevant algorithms. Multi-source data fusion has become a trend in remote sensing [62]–[64], but current algorithms in computer vision are insufficient for such remote sensing datasets. To fully exploit multi-source point cloud data, more research is needed.

As we have reviewed, different point clouds have different features and application environments. Table I provides an

overview of basic information about various point clouds, including point density, advantages, disadvantages, and applications.

C. Point cloud application

In the studies on PCS and PCSS, data and algorithm selections are driven by the requirements of specific applications. In this section, we outline most of the studies focusing on PCS and PCSS reviewed in this article (see Table II). These works are classified according to their point cloud data types and working environments. The latter include urban, forest, industry, and indoor settings. In Table II, texts in brackets, after each reference, contain the corresponding publishing year and main methods. Algorithm types are represented as abbreviations.

Several issues can be summarized from Table II: (a) LiDAR point clouds are the most commonly used data in PCS. They have been widely used for buildings (urban environments) and trees (forests). Buildings are also the most popular research objects in traditional PCS. As buildings are usually constructed with regular planes, plane segmentation is a fundamental topic in building segmentation.

(b) Image-derived point clouds have been frequently used in real-world scenarios. However, mainly due to the limitation of available annotated benchmarks, there are not many PCS and PCSS studies on image-based data. Currently, there is only one public influential dataset based on image-derived points, whose range is only a very small area around one single building [132]. More efforts are therefore needed in this area.

(c) RGB-D sensors are limited by their close range, so they are usually applied in an indoor environment. In PCS studies, plane segmentation is the main task for RGB-D data. In PCSS studies, since there are several benchmark datasets from RGB-D sensors, many deep learning-based methods are tested on them.

(d) As for InSAR point clouds, although there are not many PCS or PCSS studies, these have shown potential in urban monitoring, especially building structure segmentation.

D. Benchmark datasets

Public standard benchmark datasets achieve significant effectiveness for algorithm development, evaluation and comparison. It should be noted that most of them are labeled for PCSS rather than PCS. Since 2009, several benchmark datasets have been available for PCSS. However, early datasets have plenty of shortcomings. For example, the Oakland outdoor MLS dataset [96], the Sydney Urban Objects MLS dataset [133], the Paris-rue-Madame MLS dataset [134], the IQmulus & TerraMobilita Contest MLS dataset [35] and ETHZ CVL RueMonge 2014 multiview stereo dataset [132] can not sufficiently provide both different object representations and labeled points. KITTI [135] and NYUv2 [136] have more objects and points than the aforementioned datasets, but they do not provide a labeled point cloud directly. These must be generated from 3D bounding boxes in KITTI or depth images in NYUv2.

To overcome the drawbacks of early datasets, new benchmark data have been made available in recent years. Currently, mainstream PCSS benchmark datasets are from either LiDAR or RGB-D sensors. A nonexhaustive list of these datasets follows.

1) *Semantic3D.net*: The semantic3D.net [34] is a representative large-scale outdoor TLS PCSS dataset. It is a collection of urban scenes with over four billion labeled 3D points in total for PCSS purposes only. Those scenes contain a range of diverse urban objects, divided into eight classes, including man-made terrain, natural terrain, high vegetation, low vegetation, buildings, hardscape, scanning artefacts, and cars. In consideration of the efficiency of different algorithms, two types of sub-datasets were designed, semantic-8 and reduced-8. Semantic-8 is the full dataset, while reduced-8 uses training data in the same way as semantic-8, but only includes four small-sized subsets as test data. This dataset can be downloaded at <http://www.semantic3d.net/>. To learn the performance of different algorithms on this dataset, readers are recommended to refer to [2], [67], [112].

2) *Stanford Large-scale 3D Indoor Spaces Dataset (S3DIS)*: Unlike semantic3D.net, S3DIS [44] is a large-scale indoor RGB-D dataset, which is also a part of the 2D-3D-S dataset [137]. It is a collection of over 215 million points, covering an area of over 6,000 m^2 in six indoor regions originating from three buildings. The main covered areas are for educational and office use. Annotations in S3DIS have been prepared at an instance level. Objects are divided into structural and movable elements, which are further classified into 13 classes (structural elements: ceiling, floor, wall, beam, column, window, door; movable elements: table, chair, sofa, bookcase, board, clutter for all other elements). The dataset can be requested from <http://buildingparser.stanford.edu/dataset.html>. To learn the performance of different algorithms on this dataset, readers are recommended to refer to [2], [70], [100], [119].

3) *Vaihingen point cloud semantic labeling dataset*: This dataset [31] is the most well-known published benchmark dataset in the remote sensing field in recent years. It is a collection of ALS point cloud, consisting of 10 strips captured by a Leica ALS50 system with a 45° field of view and 500 m mean flying height over Vaihingen, Germany. The average overlap between two neighboring strips is around 30% and the median point density is 6.7 $points/m^2$ [31]. This dataset had no label at a point level at first. Niemeyer et al. [87] first used it for a PCSS test and labeled points in three areas. Now the labeled point cloud is divided into nine classes as an algorithm evaluation standard. Although this dataset has significantly fewer points compared with semantic3D.net and S3DIS, it is an influential ALS dataset for remote sensing. The dataset can be requested from <http://www2.isprs.org/commissions/comm3/wg4/3d-semantic-labeling.html>.

4) *Paris-Lille-3D*: The Paris-Lille-3D [36] is a brand new benchmark for PCSS, as it was published in 2018. It is an MLS point cloud dataset with more than 140 million labelled points, including 50 different urban object classes along 2 km of streets in two French cities, Paris and Lille. As an MLS dataset, it also could be used for autonomous vehicles. As this

TABLE I
AN OVERVIEW OF VARIOUS POINT CLOUDS

| | | Point density | Advantages | Disadvantages | Applications |
|--------------|----------------------|--|--|---|---|
| | Image-derived | From sparse ($<10pts/m^2$) to very high ($>400pts/m^2$), depending on the spatial resolution of the stereo- or multi-view images | With color (RGB, multi-spectral) information; suitable for large area (airborne, spaceborne) | Influenced by light; accuracy depends on available precise camera models, image matching algorithms, stereo angles, image resolution and image quality; not suitable for areas or objects without texture, such as water or snow-covered regions; influenced by shadows in images | Urban monitoring; vegetation monitoring; 3D object reconstruction; etc. |
| LiDAR | ALS | Sparse ($<20pts/m^2$); when the survey distance is shorter, the density is higher | High accuracy ($<15cm$); suitable for large area; not affected by weather | Expensive; affected by mirror reflection; long scanning time | Urban monitoring; vegetation monitoring; power line detection; etc. |
| | MLS | Dense ($>100pts/m^2$); when the survey distance is shorter, the density is higher | High accuracy (cm-level) | | HD map; urban monitoring |
| | TLS | Dense ($>100pts/m^2$); when the survey distance is shorter, the density is higher | High accuracy (mm-level) | | Small-area 3D reconstruction |
| | ULS | Dense ($>100pts/m^2$); when the survey distance is shorter, the density is higher | High accuracy (cm-level) | | Forestry survey; mining survey; disaster monitoring; etc. |
| | RGB-D | Middle-density | Cheap; flexible | Close-range; limited accuracy | Indoor reconstruction; object tracking; human pose recognition; etc. |
| | InSAR | Sparse ($<20pts/m^2$) | Global data is available; compared to ALS, complete building façade information is available; 4D information; middle-accuracy; not affected by weather | Expensive data; ghost scatterers; preprocessing techniques are needed | Urban monitoring; forest monitoring; etc. |

TABLE II
AN OVERVIEW OF PCS AND PCSS APPLICATIONS SORTED ACCORDING TO DATA ACQUISITIONS

RG is short for Region Growing. HT is short for Hough Transform. R is short for RANSAC. C is short for Clustering-based. O is short for Oversegmentation. ML is short for Machine Learning. DL is short for Deep Learning.

| | Urban | Forest | Industry | Indoor |
|---------------------------|---|--|--|---|
| Image-derived | Building façades: [65] (2018/RG), [66] (2005/RG); PCSS: [67] (2018/DL), [68] (2018/DL), [69] (2017/DL), [70] (2019/DL) | | | Plane PCS: [71] (2015/HT) |
| ALS | Building plane PCS: [72] (2015/R), [73] (2014/R), [74] (2007/R, HT), [75] (2002/HT), [76] (2006/C), [77] (2010/C), [78] (2012/C), [79] (2014/C); Urban scene: [80] (2007/C), [81] (2009/C); PCSS: [82] (2007/ML), [83] (2009/ML), [84] (2009/ML), [85] (2010/ML), [86] (2012/ML), [87] (2014/ML), [88] (2017/HT, R, ML), [89] (2011/ML), [90] (2014/ML), [4] (2013/HT, ML) | Tree structure PCS: [91](2004/C); Forest structure: [92] (2010/C) | | |
| MLS | Buildings: [93] (2015/RG); Urban objects: [94] (2012/RG); PCSS: [89] (2011/ML), [95] (2015/ML), [5] (2015/ML), [8] (2012/ML), [90] (2014/ML), [96] (2009/ML), [97] (2017/ML), [98] (2017/DL), [99] (2018/DL), [100] (2019/O, DL) | | | Plane PCS: [101] (2013/R), [102] (2017/R) |
| TLS | Building/building structure PCS: [103] (2007/R), [93] (2015/RG), [104] (2018/RG, C), [105] (2008/C); Buildings and trees: [106] (2009/RG); Urban scene: [107] (2016/O, C), [108] (2017/O, C), [109] (2018/O, C); PCSS: [6] (2015/ML), [110] (2009/O, ML), [111] (2016/ML), [67] (2018/DL), [98] (2017/DL), [2] (2018/O, DL), [112] (2019/DL) [70] (2019/DL) | Tree PCSS: [113] (2005/ML) | | Plane PCS: [114] (2011/HT) |
| RGB-D | | | | Plane PCS: [115] (2014/HT), [104] (2018/RG, C); PCSS: [116] (2012/ML), [117] (2013/ML), [118] (2018/DL), [119] (2018/DL), [98] (2017/DL), [1] (2017/DL), [120] (2017/DL), [3] (2018/DL), [2] (2018/DL), [99] (2018/DL), [121] (2018/DL), [70] (2019/DL), [112] (2019/DL), [122] (2019/DL), [123] (2019/DL), [124] (2019/DL), [125] (2019/DL), [126] (2019/DL), [100] (2019/O, DL); Instance segmentation: [127] (2018/DL), [128] (2019/DL), [123] (2019/DL), [124] (2019/DL) |
| InSAR | Building/building structure: [47] (2015/C), [45] (2012/C), [46] (2014/C) | Tree PCS: [48] (2015/C) | | |
| Not mentioned data | | | [129](2005/HT), [130] (2015/R), [131] (2018/R) | |

is a recent dataset, only a few validated results are shown on the related website. This dataset is available at <http://npm3d.fr/paris-lille-3d>.

5) *ScanNet*: ScanNet [43] is an instance-level indoor RGB-D dataset that includes both 2D and 3D data. In contrast to the benchmarks mentioned above, ScanNet is a collection of labeled voxels rather than points or objects. Up to now, ScanNet v2, the newest version of ScanNet, has collected 1513 annotated scans with an approximate 90% surface coverage. In the semantic segmentation task, this dataset is marked in 20 classes of annotated 3D voxelized objects. Each class corresponds to one category of furniture. This dataset can be requested from <http://www.scan-net.org/index#code-and-data>. To learn the performance of different algorithms on this dataset, readers are recommended to refer to [70], [120], [123], [124].

III. POINT CLOUD SEGMENTATION TECHNIQUES

PCS algorithms are mainly based on strict hand-crafted features from geometric constraints and statistical rules. The main process of PCS aims at grouping raw 3D points into non overlapping regions. Those regions correspond to specific structures or objects in one scene. Since no supervised prior knowledge is required in such a segmentation procedure, the delivered results have no strong semantic information. Those approaches could be categorized into four major groups: edge-based, region growing, model fitting, and clustering-based.

A. Edge-based

Edge-based PCS approaches were directly transferred from 2D images to 3D point clouds, which were mainly used in the very early stage of PCS. As the shapes of objects are described by edges, PCS can be solved by finding the points that are close to the edge regions. The principle of edge-based methods is to locate the points that have a rapid change in intensity [16], which is similar to some 2D image segmentation approaches.

According to the definition from [138], an edge-based segmentation algorithm is formed by two main stages: (1) edge detection, where the boundaries of different regions are extracted, and (2) grouping points, where the final segments are generated by grouping points inside the boundaries from (1). For example, in [139], the authors designed a gradient-based algorithm for edge detection, fitting 3D lines to a set of points and detecting changes in the direction of unit normal vectors on the surface. In [140], the authors proposed a fast segmentation approach based on high-level segmentation primitives (curve segments), in which the amount of data could be significantly reduced. Compared to the method presented in [139], this algorithm is both accurate and efficient, but it is only suitable for range images, and may not work for uneven-density point clouds. Moreover, paper [141] extracted close contours from a binary edge map for fast segmentation. Paper [142] introduced a parallel edge-based segmentation algorithm extracting three types of edges. An algorithm optimization mechanism, named reconfigurable multiRing network, was applied in this algorithm to reduce its runtime.

The edge-based algorithms enable a fast PCS due to its simplicity, but their good performance can only be maintained when simple scenes with ideal points are provided (e.g., low noise, even density). Some of them are only suitable for range images rather than 3D points. Thus this approach is rarely applied for dense and/or large-area point cloud datasets nowadays. Besides, in 3D space, such methods often deliver disconnected edges, which cannot be used to identify closed segments directly, without a filling or interpretation procedure [17], [143].

B. Region growing

Region growing is a classical PCS method, which is still widely used today. It uses criteria, combining features between two points or two region units in order to measure the similarities among pixels (2D), points (3D), or voxels (3D), and merge them together if they are spatially close and have similar surface properties. Besl and Jain [144] introduced a two-step initial algorithm: (1) coarse segmentation, in which seed pixels are selected based on the mean and Gaussian curvature of each point and its sign; and (2) region growing, in which interactive region growing is used to refine the result of step (1) based on a variable order bivariate surface fitting. Initially, this method was primarily used in 2D segmentation. As in the early stage of PCS research most point clouds were actually 2.5D airborne LiDAR data, in which only one layer has a view in the z direction, the general preprocessing step was to transform points from 3D space into a 2D raster domain [145]. With the more easily available real 3D point clouds, region growing was soon adopted directly in 3D space. This 3D region growing technique has been widely applied in the segmentation of building plane structures [75], [93], [94], [101], [104].

Similar to the 2D case, 3D region growing comprises two steps: (1) select seed points or seed units; and (2) region growing, driven by certain principles. To design a region growing algorithm, three crucial factors should be taken into consideration: criteria (similarity measures), growth unit, and seed point selection. For the criteria factor, geometric features, e.g., Euclidean distance or normal vectors, are commonly used. For example, Ning et al. [106] employed the normal vector as criterion, so that the coplanar may share the same normal orientation. Tovari et al. [146] applied normal vectors, the distance of the neighboring points to the adjusting plane, and the distance between the current point and candidate points as the criteria for merging a point to a seed region that was randomly picked from the dataset after manually filtering areas near edges. Dong et al. [104] chose normal vectors and the distance between two units.

For growth unit factor, there are usually three strategies: (1) single points, (2) region units, e.g., voxel grids and octree structures, and (3) hybrid units. Selecting single points as region units was the main approach in the early stages [106], [138]. However, for massive point clouds, point-wise calculation is time-consuming. To reduce the data volume of the raw point cloud and improve calculation efficiency, e.g., neighborhood search with a k -d tree in raw data [147], the

region unit is an alternative idea of direct points in 3D region growing. In a point cloud scene, the number of voxelized units is smaller than the number of points. In this way, the region growing process can be accelerated significantly. Guided by this strategy, Deschaud et al. [147] presented a voxel-based region growing algorithm to improve efficiency by replacing points with voxels during the region growing procedure. Vo et al. [93] proposed an adaptive octree-based region growing algorithm for fast surface patch segmentation by incrementally grouping adjacent voxels with a similar saliency feature. As a balance of accuracy and efficiency, hybrid units were also proposed and tested by several studies. For example, Xiao et al. [101] combined single points with subwindows as growth units to detect planes. Dong et al. [104] utilized a hybrid region growing algorithm, based on units of both single points and supervoxels, to realize coarse segmentation before global energy optimization.

For Seed point selection, since many region growing algorithms aim at plane segmentation, a usual practice is designing a fitting plane for a certain point and its neighbor points first, and then choosing the point with minimum residual to the fitting plane as a seed point [106], [138]. The residual is usually estimated by the distance between one point and its fitting plane [106], [138] or the curvature of the point [94], [104].

Nonuniversality is a nontrivial problem for region growing [93]. The accuracy of these algorithms relies on the growth criteria and locations of the seeds, which should be predefined and adjusted for different datasets. In addition, these algorithms are computationally intensive and may require a reduction in data volume for a trade-off between accuracy and efficiency.

C. Model fitting

The core idea of model fitting is matching the point clouds to different primitive geometric shapes, thus it has been normally regarded as a shape detection or extraction method. However, when dealing with scenes with parameter geometric shapes/models, e.g., planes, spheres, and cylinders, model fitting can also be regarded as a segmentation approach. Most widely used model-fitting methods are built on two classical algorithms, Hough Transform (HT) and RANdom SAmple Consensus (RANSAC).

1) *HT*: HT is a classical feature detection technique in digital image processing. It was initially presented in [148] for line detection in 2D images. There are three main steps in HT [149]: (1) mapping every sample (e.g., pixels in 2D images and points in point clouds) of the original space into a discretized parameter space; (2) laying an accumulator with a cell array on the parameter space and then, for each input sample, casting a vote for the basic geometric element of which they are inliers in the parameter space; and (3) selecting the cell with the local maximal score, of which parameter coordinates are used to represent a geometric segment in original space. The most basic version of HT is Generalized Hough Transform (GHT), also called the Standard Hough Transform (SHT), which is introduced in [150]. GHT uses

an angle-radius parameterization instead of the original slope-intercept form, in order to avoid the infinite slope problem and simplify the computation. The GHT is based on:

$$\rho = x \cos(\theta) + y \sin(\theta) \quad (1)$$

where x and y are the image coordinates of a corresponding sample pixel, ρ is the distance between the origin and the line through the corresponding pixel, and θ is the angle between the normal of the above-mentioned line and the x -axis. Angle-radius parameterization can also be extended into 3D space, and thus can be used in 3D feature detection and regular geometric structure segmentation. Compared with the 2D form, in 3D space, there is one more angle parameter, ϕ :

$$\rho = x \cos(\theta) \sin(\phi) + y \sin(\theta) \sin(\phi) + z \cos(\phi) \quad (2)$$

where x , y , and z are corresponding coordinates of a 3D sample (e.g., one specific point from the whole point cloud), and θ and ϕ are polar coordinates of the normal vector of the plane, which includes the 3D sample.

One of the major disadvantages of GHT is the lack of boundaries in the parameter space, which leads to high memory consumption and long calculation time [151]. Therefore, some studies have been conducted to improve the performance of HT by reducing the cost of the voting process [71]. Such algorithms include Probabilistic Hough transform (PHT) [152], Adaptive probabilistic Hough transform (APHT) [153], Progressive Probabilistic Hough Transform (PPHT) [154], Randomized Hough Transform (RHT) [149], and Kernel-based Hough Transform (KHT) [155]. In addition to computational costs, choosing a proper accumulator representation is also a way to optimize HT performance [114].

Several review articles involving 3D HT are available [71], [114], [151]. As with region growing in the 3D field, planes are the most frequent research objects in HT-based segmentation [71], [74], [115], [156]. In addition to planes, other basic geometric primitives can also be segmented by HT. For example, Rabbani et al. [129] used a Hough-based method to detect cylinders in point clouds, similar to plane detection. In addition, a comprehensive introduction to sphere recognition based on HT methods is presented in [157].

To evaluate different HT algorithms on point clouds, Borrmann et al. [114] compared improved HT algorithms and concluded that RHT was the best one for PCS at that time, due to its high efficiency. Limberger et al. [71] extended KHT [155] to 3D space, and proved that 3D KHT performed better than previous HT techniques, including RHT, for plane detection. The 3D KHT approach is also robust to noise and even to irregularly distributed samples [71].

2) *RANSAC*: The RANSAC technique is the other popular model fitting method [158]. Several reviews about general RANSAC-based methods have been published. Learning more about the RANSAC family and their performance is highly recommended, particularly in [159]–[161]. The RANSAC-based algorithm has two main phases: (1) generate a hypothesis from random samples (hypothesis generation), and (2) verify it to the data (hypothesis evaluation/model verification) [159], [160]. Before step (1), as in the case of HT-based

methods, models have to be manually defined or selected. Depending on the structure of 3D scenes, in PCS, these are usually planes, spheres, or other geometric primitives that can be represented by algebraic formulas.

In hypothesis generation, RANSAC randomly chooses N sample points and estimates a set of model parameters using those sample points. For example, in PCS, if the given model is a plane, then $N = 3$ since 3 non-collinear points determine a plane. The plane model can be represented by:

$$aX + bY + cZ + d = 0 \quad (3)$$

where $[a, b, c, d]^T$ is the parameter set to be estimated.

In hypothesis evaluation, RANSAC chooses the most probable hypothesis from all estimated parameter sets. RANSAC uses Eq. 4 to solve the selection problem, which is regarded as an optimization problem [159]:

$$\hat{M} = \arg \min_M \left\{ \sum_{d \in \mathcal{D}} \text{Loss}(\text{Err}(d; M)) \right\} \quad (4)$$

where \mathcal{D} is data, Loss represents a loss function, and Err is an error function such as geometric distance.

As an advantage of random sampling, RANSAC-based algorithms do not require complex optimization or high memory resources. Compared to HT methods, efficiency and the percentage of successful detected objects are two main advantages for RANSAC in 3D PCS [74]. Moreover, RANSAC algorithms have the ability to process data with a high amount of noise, even outliers [162]. For PCS, as with HT and region growing, RANSAC is widely used in plane segmentation, such as building façades [65], [66], [103], building roofs [73], and indoor scenes [102]. In some fields there is demand for the segmentation of more complex structures than planes. Schnabel et al. [162] proposed an automatic RANSAC-based algorithm framework to detect basic geometric shapes in unorganized point clouds. Those shapes include not only planes, but also spheres, cylinders, cones, and tori. RANSAC-based PCS segmentation algorithms were also utilized for cylinder objects in [130] and [131].

RANSAC is a nondeterministic algorithm, and thus its main shortcoming is its spurious surface: the probability exists that models detected by RANSAC-based algorithm do not exist in reality (Fig. 1). To overcome the adverse effect of RANSAC in PCS, a soft-threshold voting function was presented to improve the segmentation quality in [72], in which both the point-plane distance and the consistency between the normal vectors were taken into consideration. Li et al. [102] proposed an improved RANSAC method based on NDT cells [163], also in order to avoid spurious surface problem in 3D PCS.

As with HT, many improved algorithms based on RANSAC have emerged over the past decades to further improve its efficiency, accuracy and robustness. These approaches have been categorized by their research objectives and are shown in Fig. 2. The figure has been originally described in [159], in which seven subclasses according to seven strategies are used. Venn diagrams are utilized here to describe connections between methods and strategies, since a method may use two strategies. For detail description and explanation on those

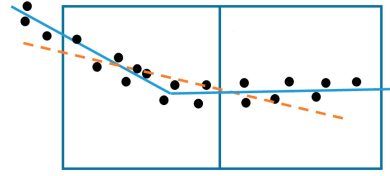


Fig. 1. An example of a spurious plane [102]. Two well-estimated hypothesis planes are shown in blue. A spurious plane (in orange) is generated using the same threshold.

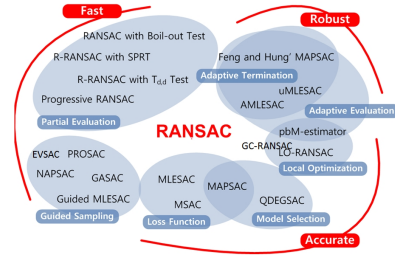


Fig. 2. RANSAC family with algorithms categorized according to their performance and basic strategies [159], [164], [165].

strategies, please refer to [159]. Considering that [159] is obsolete, we add two recently published methods, EVSAC [164] and GC-RANSAC [165] on original figure to make it keep up with the times.

D. Unsupervised clustering-based

Clustering-based methods are widely used for unsupervised PCS task. Strictly speaking, clustering-based methods are not based on a specific mathematical theory. This methodology family is a mixture of different methods that share a similar aim, which is grouping points with similar geometric features, spectral features or spatial distribution into the same homogeneous pattern. Unlike region growing and model fitting, these patterns usually are not defined in advance [166], and thus clustering-based algorithms can be employed for irregular object segmentation, e.g., vegetation. Moreover, seed points are not required by clustering-based approaches, in contrast to region growing methods [109]. In the early stage, K -means [45], [46], [76], [77], [91], mean shift [47], [48], [80], [92], and fuzzy clustering [77], [105] were the main algorithms in the clustering-based point cloud segmentation family. For each clustering approach, several similarity measures with different features can be selected, including Euclidean distance, density, and normal vector [109]. From the perspective of mathematics and statistics, the clustering problem can be regarded as a graph-based optimization problem, so several graph-based methods have been experimented in PCS [78], [79], [167].

1) K -means: K -means is a basic and widely used unsupervised cluster analysis algorithm. It separates the point cloud dataset into K unlabeled classes. The clustering centers of K -means are different than the seed points of region growing. In K -means, every point should be compared to every cluster center in each iteration step, and the cluster centers will change when absorbing a new point. The process of K -means

is “clustering” rather than “growing”. It has been adopted for single tree crown segmentation on ALS data [91] and planar structure extraction from roofs [76]. Shahzad et al. [45] and Zhu et al. [46] utilized K -means for building façade segmentation on TomoSAR point clouds.

One advantage of K -means is that it can be easily adapted to all kinds of feature attributes, and can even be used in a multidimensional feature space. The main drawback of K -means is that it is sometimes difficult to predefine the value of K properly.

2) *Fuzzy clustering*: Fuzzy clustering algorithms are improved versions of K -means. K -means is a hard clustering method, which means the weight of a sample point to a cluster center is either 1 or 0. In contrast, fuzzy methods use soft clustering, meaning a sample point can belong to several clusters with certain nonzero weights.

In PCS, a no-initialization framework was proposed in [105], by combining two fuzzy algorithms, Fuzzy C -Means (FCM) algorithm and Possibilistic C -Means (PCM). This framework was tested on three point clouds, including a one-scan TLS outdoor dataset with building structures. Those experiments showed that fuzzy clustering segmentation worked robustly on planer surfaces. Sampath et al. [77] employed fuzzy K -means for segmentation and reconstruction of building roofs from an ALS point cloud.

3) *Mean-shift*: In contrast to K -means, mean-shift is a classic nonparametric clustering algorithm and hence avoids the predefined K problem in K -means [168]–[170]. It has been applied effectively on ALS data in urban and forest terrain [80], [92]. Mean-shift have also been adopted on TomoSAR point clouds, enabling building façades and single trees to be extracted [47], [48].

As both the cluster number and the shape of each cluster are unknown, mean-shift delivers with high-probability oversegmented result [81]. Hence, it is usually used as a presegmentation step before partitioning or refinement.

4) *Graph-based*: In 2D computer vision, introducing graphs to represent data units such as pixels or superpixels has proven to be an effective strategy for the segmentation task. In this case, the segmentation problem can be transformed into a graph construction and partitioning problem. Inspired by graph-based methods from 2D, some studies have applied similar strategies in PCS and achieved results in different datasets.

For instance, Golovinskiy and Funkhouser [167] proposed a PCS algorithm based on min-cut [171], by constructing a graph using k -nearest neighbors. The min-cut was then successfully applied for outdoor urban object detection [167]. Ural et al. [78] also used min-cut to solve the energy minimization problem for ALS PCS. Each point is considered to be a node in the graph, and each node is connected to its 3D voronoi neighbors with an edge. For the roof segmentation task, Yan et al. [79] used an extended α -expansion algorithm [172] to minimize the energy function from the PCS problem. Moreover, Yao et al. [81] applied a modified normalized cut (N-cut) in their hybrid PCS method.

Markov Random Field (MRF) and Conditional Random Field (CRF) are machine learning approaches to solve graph-

based segmentation problems. They are usually used as supervised methods or postprocessing stages for PCSS. Major studies using CRF and supervised MRFs belong to PCSS rather than PCS. For more information about supervised approaches, please refer to section IV-A.

E. Oversegmentation, supervoxels, and presegmentation

To reduce the calculation cost and negative effects from noise, a frequently used strategy is to oversegment a raw point cloud into small regions before applying computationally expensive algorithms. Voxels can be regarded as the simplest oversegmentation structures. Similar to superpixels in 2D images, supervoxels are small regions of perceptually similar voxels. Since supervoxels can largely reduce the data volume of a raw point cloud with low information loss and minimal overlapping, they are usually utilized in presegmentation before executing other computationally expensive algorithms. Once oversegments like supervoxels are generated, these are fed to postprocessing PCS algorithms rather than initial points.

The most classical point cloud oversegmentation algorithm is Voxel Cloud Connectivity Segmentation (VCCS) [173]. In this method, a point cloud is first voxelized by the octree. Then a K -means clustering algorithm is employed to realize supervoxel segmentation. However, since VCCS adopts fixed resolution and relies on initialization of seed points, the quality of segmentation boundaries in a non-uniform density cannot be guaranteed. To overcome this problem, Song et al. [174] proposed a two-stage supervoxel oversegmentation approach, named Boundary-Enhanced Supervoxel Segmentation (BESS). BESS preserves the shape of the object, but it also has an obvious limitation for the assumption that points are sequentially ordered in one direction. Recently, Lin et al. [175] summarized the limitations of previous studies, and formalized oversegmentation as a subset selection problem. This method adopts an adaptive resolution to preserve boundaries, a new practice in supervoxel generation. Landrieu and Boussaha [100] presented the first supervised framework for 3D point cloud oversegmentation, achieving significant improvements compared to [173], [175]. For PCS tasks, several studies have been based on supervoxel-based presegmentation [107]–[109], [176], [177].

As mentioned in section III-D, in addition to supervoxels, other methods can also be employed as presegmentation. For example, Yao et al. [81] utilized mean-shift to oversegment ALS data in urban areas.

IV. POINT CLOUD SEMANTIC SEGMENTATION TECHNIQUES

The procedure of PCSS is similar to clustering-based PCS. But in contrast to non-semantic PCS methods, PCSS techniques generate semantic information for every point, and are not limited to clustering. Therefore, PCSS is usually realized by supervised learning methods, including “regular” supervised machine learning and state-of-the-art deep learning.

A. Regular supervised machine learning

In this section, regular supervised machine learning refers to non-deep supervised learning algorithms. Comprehensive

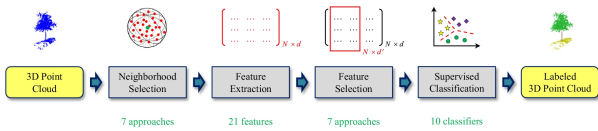


Fig. 3. The PCSS framework by [95]. The term “semantic segmentation” in our review is defined as “supervised classification” in [95].

and comparative analysis on different PCSS methods based on regular supervised machine learning has been provided by previous researchers [87], [88], [95], [97].

Paper [5] pointed out that supervised machine learning applied to PCSS could be divided into two groups. One group, individual PCSS, classifies each point or each point cluster based only on its individual features, such as Maximum Likelihood classifiers based on Gaussian Mixture Models [113], Support Vector Machines [4], [111], AdaBoost [6], [82], a cascade of binary classifiers [83], Random Forests [84], and Bayesian Discriminant Classifiers [116]. The other group is statistical contextual models, such as Associative and Non-Associative Markov Networks [85], [90], [96], Conditional Random Fields [86]–[88], [110], [178], Simplified Markov Random Fields [8], multistage inference procedures focusing on point cloud statistics and relational information over different scales [89], and spatial inference machines modeling mid- and long-range dependencies inherent in the data [117].

The general procedure of the individual classification for PCSS has been well described in [95]. As Fig. 3 shows, the procedure entails four stages: neighborhood selection, feature extraction, feature selection, and semantic segmentation. For each stage, paper [95] summarized several crucial methods and tested different methods on two datasets to compare their performance. According to the authors’ experiment, in individual PCSS, the Random Forest classifier had a good trade-off between accuracy and efficiency on two datasets. It should be noted that [95] used a so-called “deep learning” classifier in their experiments, but that is an old neural network appearing in the time of regular machine learning, not the recent deep learning methods described in section IV-B.

Since individual PCSS does not take contextual features of points into consideration, individual classifiers work efficiently but generate unavoidable noise that cause unsmooth PCSS results. Statistical context models can mitigate this problem. Conditional Random Fields (CRF) is the most widely used context model in PCSS. Niemeyer et al. [87] provided a very clear introduction about how CRF has been used on PCSS, and tested several CRF-based approaches on the Vaihingen dataset. Based on the individual PCSS framework [95], Landrieu et al. [97] proposed a new PCSS framework that combines individual classification and context classification. As shown in Fig. 4, in this framework a graph-based contextual strategy was introduced to overcome the noise problem of initial labeling, from which the process was named structured regularization or “smoothing”.

For the regularization process, Li et al. [111] utilized a multilabel graph-cut algorithm to optimize the initial segmentation result from Support Vector Machine (SVM). Landrieu et al.

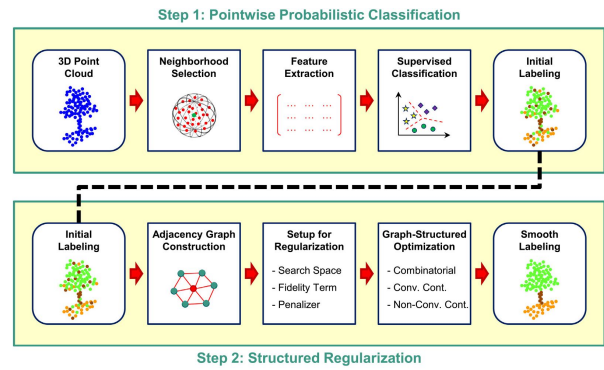


Fig. 4. The PCSS framework by [97]. The term “semantic segmentation” in our review is defined as “supervised classification” in [97].

[97] compared various postprocess methods in their studies, which proved that regularization indeed improved the accuracy of PCSS.

B. Deep learning

Deep learning is the most influential and fastest-growing current technique in pattern recognition, computer vision, and data analysis [179]. As its name indicates, deep learning uses more than two hidden layers to obtain high-dimension features from training data, while traditional handcrafted features are designed with domain-specific knowledge. Before being applied in 3D data, deep learning appeared as an effective power in a variety of tasks in 2D computer vision and image processing, such as image recognition [180], [181], object detection [182], [183], and semantic segmentation [184], [185]. It has been attracting more interest in 3D analysis since 2015, driven by the multiview-based idea proposed by [186], and voxel-based 3D Convolutional Neural Network (CNN) by [187].

Standard convolutions originally designed for raster images cannot easily be directly applied to PCSS, as the point cloud is unordered and unstructured/irregular/non-raster. Thus, in order to solve this problem, a transformation of the raw point cloud becomes essential. Depending on the format of the data ingested into neural networks, deep learning-based PCSS approaches can be divided into three categories: multiview-based, voxel-based, and point-based, respectively.

1) *Multiview-based*: One of the early solutions to applying deep learning in 3D is dimensionality reduction. In short, the 3D data is represented by multi-view 2D images, which can be processed based on 2D CNNs. Subsequently, the classification results can be restored into 3D. The most influential multi-view deep learning in 3D analysis is MVCNN [186]. Although the original MVCNN algorithm did not experiment on PCSS, it is a good example for learning about the multiview concept.

The multiview-based methods have solved the structuring problems of point cloud data well, but there are two serious shortcomings in these methods. Firstly, they cause numerous limitations and a loss in geometric structures, as 2D multiview images are just an approximation of 3D scenes. As a result, complex tasks such as PCSS could yield limited and unsatisfactory performances. Secondly, multiview projected images

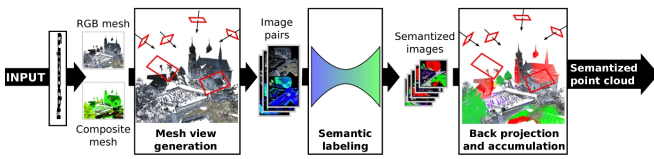


Fig. 5. The Workflow of SnapNet [67].

must cover all spaces containing points. For large, complex scenes, it is difficult to choose enough proper viewpoints for multiview projection. Thus, few studies used multiview-based deep learning architecture for PCSS. One of exceptions is SnapNet [9], [67], which uses full dataset semantic-8 of semantic3D.net as the test dataset. Fig. 5 shows the workflow of SnapNet. In SnapNet, the preprocessing step aims at decimating the point cloud, computing point features and generating a mesh. Snap generation is to generate RGB images and depth composite images of the mesh, based on various virtual cameras. Semantic labeling is to realize image semantic segmentation from the two types of input images, by image-based deep learning. The last step is to project 2D semantic segmentation results back to 3D space, thereby 3D semantics can be acquired.

2) *Voxel-based*: Combining voxels with 3D CNNs is the other early approach in deep learning-based PCSS. Voxelization solves both unordered and unstructured problems of the raw point cloud. Voxelized data can be further processed by 3D convolutions, as in the case of pixels in 2D neural networks.

Voxel-based architectures still have serious shortcomings. In comparison to the point cloud, the voxel structure is a low-resolution form. Obviously, there is a loss in data representation. In addition, voxel structures not only store occupied spaces, but also store free or unknown spaces, which can result in high computational and memory requirements.

The most well-known voxel-based 3D CNN is VoxNet [187], but this was only tested for object detection. On the PCSS task, some papers, like [69], [98], [188] and [189], proposed representative frameworks. SegCloud [98] is an end-to-end PCSS framework that combines 3D-FCNN, trilinear interpolation (TI), and fully connected Conditional Random Fields (FC-CRF) to accomplish the PCSS task. Fig. 6 shows the framework of SegCloud, which also provides a basic pipeline of voxel-based semantic segmentation. In SegCloud, the preprocessing step is to voxelize raw point clouds. Then a 3D fully convolutional neural network is applied to generate downsampled voxel labels. After that, a trilinear interpolation layer is employed to transfer voxel labels back to 3D point labels. Finally, a 3D fully connected CRF method is utilized to regularize previous 3D PCSS results, and acquire final results. SegCloud used to be the state-of-art approach in both S3DIS and semantic3D.net, but it did not take any steps to optimize high computational and memory problem from fixed-sized voxels. With more advanced methods springing up, SegCloud has fallen from favor in recent years.

To reduce unnecessary computation and memory consumption, the flexible octree structure is an effective replacement for fixed-size voxels in 3D CNNs. OctNet [69] and O-CNN

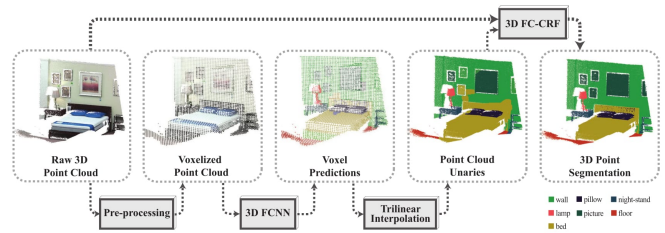


Fig. 6. The Workflow of SegCloud [98].

[188] are two representative approaches. Recently, VV-NET [189] extended the use of voxels. VV-Net utilized a radial basis function-based Variational Auto-Encoder (VAE) network, which provided a more information-rich representation for point cloud compared with binary voxels. What is more, Choy et al. [70] proposed 4-dimensional convolutional neural networks (MinkowskiNets) to process 3D-videos, which are a series of CNNs for high-dimensional spaces including the 4D spatio-temporal data. MinkowskiNets can also be applied on 3D PCSS tasks. They have achieved good performance on a series of PCSS benchmark datasets, especially a significant accuracy improvement on ScanNet [43].

3) *Directly process point cloud data*: As there are serious limitations in both multiview- and voxel-based methods (e.g., loss in structure resolution), exploring PCSS methods directly on point is a natural choice. Up to now, many approaches have emerged and are still emerging [1]–[3], [119], [120]. Unlike employing separated pretransformation operation in multiview-based and voxel-based cases, in these approaches the canonicalization is binding with the neural network architecture.

PointNet [1] is a pioneering deep learning framework which has been performed directly on point. Different with recently published point cloud networks, there is no convolution operator in PointNet. The basic principle of PointNet is:

$$f(\{x_1, \dots, x_n\}) \approx g(h(x_1), \dots, h(x_n)) \quad (5)$$

where $f : 2^{\mathbb{R}^N} \rightarrow \mathbb{R}$ and $h : \mathbb{R}^N \rightarrow \mathbb{R}^K$. $g : \underbrace{\mathbb{R}^K \times \dots \times \mathbb{R}^K}_n \rightarrow \mathbb{R}$ is a symmetric function, used to solve

the ordering problem of point clouds. As Fig. 7 shows, PointNet uses MultiLayer Perceptrons (MLPs) to approximate h , which represents the per-point local features corresponding to each point. The global features of point sets g are aggregated by all per-point local features in a set, through a symmetric function, max pooling. For the classification task, output scores for k classes can be produced by a MLP operation on global features. For the PCSS task, in addition to global features, per-point local features are demanded. PointNet concatenates aggregated global features and per-point local features into combined point features. Subsequently, new per-point features are extracted from the combined point features by MLPs. On their basis, semantic labels are predicted.

Although more and more newly published networks outperform PointNet on various benchmark datasets, PointNet is still a baseline for PCSS research. The original PointNet

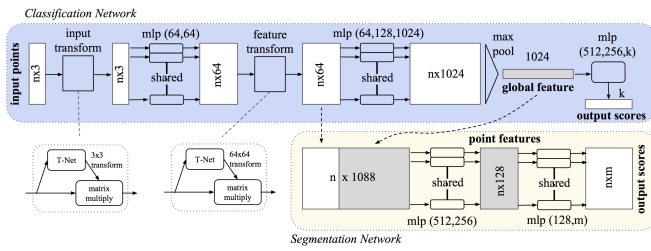


Fig. 7. The Workflow of PointNet [1]. In this figure, “Classification Network” is used for object classification. “Segmentation Network” is applied for the PCSS mission.

uses no local structure information within neighboring points. In a further study, Qi et al. [120] used a hierarchical neural network to capture local geometric features to improve the basic PointNet model and proposed PointNet++. Drawing inspiration from PointNet/PointNet++, studies on 3D deep learning focus on feature augmentation, especially to local features/relationships among points, utilizing knowledge from other fields to improve the performance of the basic PointNet/PointNet++ algorithms. For example, Engelmann et al. [190] employed two extensions on the PointNet to incorporate larger-scale spatial context. Wang et al. [3] considered that missing local features was still a problem in PointNet++, since it neglected the geometric relationships between a single point and its neighbors. To overcome this problem, Wang et al. [3] proposed Dynamic Graph CNN (DGCNN). In this network, the authors designed a procedure called EdgeConv to extract edge features while maintaining permutation invariance. Inspired by the idea of the attention mechanism, Wang et al. [112] designed a Graph Attention Convolution (GAC), of which kernels could be dynamically adapted to the structure of an object. GAC can capture the structural features of point clouds while avoiding feature contamination between objects. To exploit richer edge features, Landrieu and Simonovsky [2] introduced the SuperPoint Graph (SPG), offering both compact and rich representation of contextual relationships among object parts rather than points. The partition of the superpoint can be regarded as a nonsemantic presegmentation step. After SPG construction, each superpoint is embedded in a basic PointNet network and then refined in Gated Recurrent Units (GRUs) for PCSS. Benefiting from information-rich downsampling, SPG is highly efficient for large-volume datasets.

Also in order to overcome the drawback of no local features represented by neighboring points in PointNet, 3P-RNN [99] adopted a Pointwise Pyramid Pooling (3P) module to capture the local feature of each point. In addition, it employed a two-direction Recurrent Neural Network (RNN) model to integrate long-range context in PCSS tasks. The 3P-RNN technique has increased overall accuracy at a negligible extra overhead. Komarichev et al. [125] introduced an annular convolution, which could capture the local neighborhood by specifying the ring-shaped structures and directions in the computation, and adapt to the geometric variability and scalability at the signal processing level. Due to the fact that the K -nearest neighbor search in PointNet++ may lead to the K neighbors falling in one orientation, Jiang et al. [121] designed PointSIFT to

capture local features from eight orientations. In the whole architecture, the PointSIFT module achieves multiscale representation by stacking several Orientation-Encoding (OE) units. The PointSIFT module can be integrated into all kinds of PointNet-based 3D deep learning architectures to improve the representational ability for 3D shapes. Built upon PointNet++, PointWeb [126] utilized the Adaptive Feature Adjustment (AFA) module to find the interaction between points. The aim of AFA is also to capture and aggregate local features of points.

Besides, based on PointNet/PointNet++, instance segmentation can also be realized, even accompanied by PCSS. For instance, Wang et al. [127] presented the Similarity Group Proposal Network (SGPN). SGPN is the first published point cloud instance segmentation framework. Yi et al. [128] presented a Region-based PointNet (R-PointNet). The core module of R-PointNet is named as Generative Shape Proposal Network (GSPN), of which the base is PointNet. Pham et al. [124] applied a Multi-task Pointwise Network (MT-PNet) and a Multi-Value Conditional Random Field (MV-CRF) to address PCSS and instance segmentation simultaneously. MV-CRF jointly realized the optimization of semantics and instances. Wang et al. [123] proposed an Associatively Segmenting Instances and Semantics (ASIS) module, making PCSS and instance segmentation take advantage of each other, leading to a win-win situation. In [123], the backbone that networks employed are also PointNet and PointNet++.

An increasing number of researchers have chosen an alternative to PointNet, employing the convolution as a fundamental and significant component. Some of them, like [3], [112], [125], have been introduced above. In addition, PointCNN used a \mathcal{X} -transformation instead of symmetric functions to canonicalize the order [119], which is a generalization of CNNs to feature learning from unordered and unstructured point clouds. Su et al. [68] provided a PCSS framework that could fuse 2D images with 3D point clouds, named SParse LATtice Networks (SPLATNet), preserving spatial information even in sparse regions. Recurrent Slice Networks (RSN) [118] exploited a sequence of multiple 1×1 convolution layers for feature learning, and a slice pooling layer to solve the unordered problem of raw point clouds. A RNN model was then applied on ordered sequences for the local dependency modeling. Te et al. [191] proposed Regularized Graph CNN (RGCNN) and tested it on a part segmentation dataset, ShapeNet [192]. Experiments show that RGCNN can reduce computational complexity and is robust to low density and noise. Regarding convolution kernels as nonlinear functions of the local coordinates of 3D points comprised of weight and density functions, Wu et al. [122] presented PointConv. PointConv is an extension to the Monte Carlo approximation of the 3D continuous convolution operator. PCSS is realized by a deconvolution version of PointConv.

As SPG [2], DGCNN [3], RGCNN [191] and GAC [112] employed graph structures in neural networks, they can also be regarded as Graph Neural Networks (GNNs) in 3D [193], [194].

The research on PCSS based on deep learning is still ongoing. New ideas and approaches on the topic of 3D deep

learning-based frameworks are keeping popping up. Current achievements have proved that it is a great boost for the accuracy of 3D PCSS.

C. Hybrid methods

In PCSS, hybrid segment-wise methods have been attracting researchers' attention in recent years. A hybrid approach is usually made up of at least two stages: (1) utilize an oversegmentation or PCS algorithm (introduced in section III) as the presegmentation, and (2) apply PCSS on segments from (1) rather than points. In general, as with presegmentation in PCS, presegmentation in PCSS also has two main functions: to reduce the data volume and to conduct local features. Oversegmentation for supervoxels is a kind of presegmentation algorithm in PCSS [110], since it is an effective way to reduce the data volume with light accuracy loss. In addition, because nonsemantic PCS methods can provide rich natural local features, some PCSS studies also use them as presegmentation. For example, Zhang et al. [4] employed region growing before SVM. Vosselman et al. [88] applied HT to generate planar patches in their PCSS algorithm framework as the presegmentation. In deep learning, Landrieu and Simonovsky [2] exploited a superpoint structure as the presegmentation step, and provided a contextual PCSS network combining superpoint graphs with PointNet and contextual segmentation. Landrieu and Boussaha [100] used a supervised algorithm to realize the presegmentation, which is the first supervised framework for 3D point cloud oversegmentation.

V. DISCUSSION

A. Open issues in segmentation techniques

1) *Features*: One of the core questions in pattern recognition is how to obtain effective features. Essentially, the biggest differences among the various methods in PCSS or PCS are the differences of feature design, selection, and application. Feature selection is a trade-off between algorithm accuracy and efficiency. Focusing on PCSS, Weinmann et al. [5] analyzed features from three aspects: neighborhood selection (fixed or individual); feature extraction (single-scale or multi-scale); and classifier selection (individual classifier or contextual classifier). Deep learning-based algorithms face similar problems. The local feature is a significant aspect to be improved after the birth of PointNet [1].

Even in a PCS task, different methods also show different understandings of features. Model fitting is actually searching for a group of points connected with certain geometric primitives, which also can be defined as features. For this reason, deep learning has been introduced into model fitting recently [195]. The criteria or the similarity measure in region growing or clustering is the feature of a point essentially. The improvement of an algorithm reflects its ability to more strongly capture features.

2) *Hybrid*: As mentioned in section IV-C, hybrid is a strategy for PCSS. Presegmentation can provide local features in a natural way. Once the development of neural network architectures stabilizes, nonsemantic presegmentation might become a progressive course for PCSS.

3) *Contextual information*: In PCSS tasks, contextual models are crucial tools for regular supervised machine learning, widely exploited as a smoothing postprocessing step. In deep learning, several methods, like [98], [2], [124] and [70], have employed contextual segmentation, but there is still room for further improvements.

4) *PCSS with GNNs*: GNN is becoming increasingly popular in 2D image processing [193], [194]. For PCSS tasks, its excellent performance has been shown in [2], [3], [191] and [112]. Similar to contextual models, the GNN might also have some surprises for PCSS. But more research is required in order to evaluate its performance.

5) *Regular machine learning vs. deep learning*: Before deep learning emerged, regular machine learning was the choice of supervised PCSS. Deep learning has changed the way a point cloud is handled. Compared with regular machine learning, deep learning has notable advantages: (1) it is more efficient at handling large-volume datasets; (2) there is no need to handcraft feature design and selection, a difficult task in regular machine learning; and (3) it yields high ranks (high-accuracy results) on public benchmark datasets. Nevertheless, deep learning is not a universal solution. Firstly, its principal shortcoming is poor interpretability. Currently, it is well known how each type of layers (e.g., convolution, pooling) works in a neural network. In pioneering PCSS works, such knowledge has been used to develop a series of functional networks [1], [119], [122]. However, a detailed internal decision-making process for deep learning is not yet understood, and therefore cannot be fully described. As a result, certain fields demanding high-level safety or stability cannot trust deep learning completely. A typical example that is relevant to PCSS is autonomous driving. Secondly, data limit the application of deep learning-based PCSS. Compared with annotating 2D images, acquiring and annotating a point cloud is much more complicated. Finally, although current public datasets provide several indoor and outdoor scenes, they cannot meet the demand in real applications sufficiently.

B. Remote sensing meets computer vision

Remote sensing and general computer vision might be two of the most active groups interested in point clouds, having published many pioneering studies. The main difference between these two groups is that computer vision focuses on new algorithms to further improve the accuracy of the results. Remote sensing researchers, on the other hand, are trying to apply these techniques on different types of datasets. However, in many cases the algorithms proposed by computer vision studies cannot be adopted in remote sensing directly.

1) *Evaluation system*: In generic computer vision, in order to evaluate the accuracy, the overall accuracy is a significant index. However, some remote sensing applications care more about the accuracy of certain objects. For instance, for urban monitoring the accuracy of buildings is crucial, while the segmentation or the semantic segmentation of other objects is less important. Thus, compared to computer vision, remote sensing needs a different evaluation system for selecting proper algorithms.

2) *Multi-source Data*: As discussed in section II, point clouds in remote sensing and computer vision appear differently. For example, airborne/spaceborne 2.5D and/or sparse point clouds are also crucial components of remote sensing data, while computer vision focuses on denser full 3D.

3) *Remote sensing algorithms*: Published computer vision algorithms are usually tested on a small-area dataset with limited categories of objects. However, for remote sensing applications, large-area data with more complex and specific ground object categories are demanded. For example, in agricultural remote sensing, vegetation is expected to be separated into certain specific species, which is difficult for current computer vision algorithms to solve.

4) *Noise and outliers*: Current computer vision algorithms do not pay much attention to noise, while in remote sensing, sensor noise is unavoidable. Currently, noise adaptive algorithms are unavailable.

C. Limitation of public benchmark datasets

In section II-D, several popular benchmark datasets are listed. Obviously, in comparison to the situation several years ago, the number of large-scale datasets with dense point clouds and rich information available to researchers has increased considerably. Some datasets, such as semantic3D.net and S3DIS, have hundreds of millions of points. However, those benchmark datasets are still insufficient for PCSS tasks.

1) *Limited data types*: Despite the fact that several large datasets for PCSS are available, there is still demand for more varied data. In the real world, there are much more object categories than the ones considered in current benchmark datasets. For example, semantic3D.net provides a large-scale urban point cloud benchmark. However, it only covers one kind of cities. If researchers chose a different city for a PCSS task, in which building styles, vegetation species, and even ground object types would differ, algorithm results might in turn be different.

2) *Limited data sources*: Most mainstream point cloud benchmark datasets are acquired from either LiDAR or RGB-D. But in practical applications, image-derived point clouds cannot be ignored. As previously mentioned, in remote sensing the airborne 2.5D point cloud is an important category, but for PCSS tasks only the Vaihingen dataset [31], [87] is published as a benchmark dataset. New data types, such as satellite photogrammetric point clouds, InSAR point clouds, and even multi-source fusion data, are also necessary to establish corresponding baselines and standards.

VI. CONCLUSION

This paper provided a review of current PCSS and PCS techniques. This review not only summarizes the main categories of relevant algorithms, but also briefly introduces the acquisition methodology and evolution of point clouds. In addition, the advanced deep learning methods that have been proposed in recent years are compared and discussed. Due to the complexity of the point cloud, PCSS is more challenging than 2D semantic segmentation. Although many approaches are available, they have each been tested on very

limited and dissimilar datasets, so it is difficult to select the optimal approach for practical applications. Deep learning-based methods have ranked high for most of the benchmark-based evaluations, yet there is no standard neural network publicly available. Improved neural networks for the solution of PCSS problems can be expected to be designed in coming years.

Most current methods have only considered point features, but in practical applications such as remote sensing the noise and outliers are still problems that cannot be avoided. Improving the robustness of current approaches, and combining initial point-based algorithms with different sensor theories to denoise the data are two potential future fields of research for semantic segmentation.

ACKNOWLEDGMENT

The authors would like to thank Dr. D. Cerra and P. Schwind for proof-reading this paper, and the anonymous reviewers and the associate editor for commenting and improving this paper.

The work of Yuxing Xie is supported by the DLR-DAAD research fellowship (No. 57424731), which is funded by the German Academic Exchange Service (DAAD) and the German Aerospace Center (DLR).

The work of Xiao Xiang Zhu is jointly supported by the European Research Council (ERC) under the European Union's Horizon 2020 research and innovation programme (grant agreement No. [ERC-2016-StG-714087], Acronym: *So2Sat*), Helmholtz Association under the framework of the Young Investigators Group "SiPEO" (VH-NG-1018, www.sipeo.bgu.tum.de), and the Bavarian Academy of Sciences and Humanities in the framework of Junges Kolleg.

REFERENCES

- [1] C. R. Qi, H. Su, K. Mo, and L. J. Guibas, "Pointnet: Deep learning on point sets for 3d classification and segmentation," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 652–660, 2017.
- [2] L. Landrieu and M. Simonovsky, "Large-scale point cloud semantic segmentation with superpoint graphs," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 4558–4567, 2018.
- [3] Y. Wang, Y. Sun, Z. Liu, S. E. Sarma, M. M. Bronstein, and J. M. Solomon, "Dynamic graph cnn for learning on point clouds," *arXiv preprint arXiv:1801.07829*, 2018.
- [4] J. Zhang, X. Lin, and X. Ning, "Svm-based classification of segmented airborne lidar point clouds in urban areas," *Remote Sensing*, vol. 5, no. 8, pp. 3749–3775, 2013.
- [5] M. Weinmann, A. Schmidt, C. Mallet, S. Hinz, F. Rottensteiner, and B. Jutzi, "Contextual classification of point cloud data by exploiting individual 3d neighbourhoods," *ISPRS Annals of the Photogrammetry, Remote Sensing and Spatial Information Sciences II-3 (2015)*, Nr. W4, vol. 2, no. W4, pp. 271–278, 2015.
- [6] Z. Wang, L. Zhang, T. Fang, P. T. Mathiopoulos, X. Tong, H. Qu, Z. Xiao, F. Li, and D. Chen, "A multiscale and hierarchical feature extraction method for terrestrial laser scanning point cloud classification," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 53, no. 5, pp. 2409–2425, 2015.
- [7] H. S. Koppula, A. Anand, T. Joachims, and A. Saxena, "Semantic labeling of 3d point clouds for indoor scenes," in *Advances in neural information processing systems*, pp. 244–252, 2011.
- [8] Y. Lu and C. Rasmussen, "Simplified markov random fields for efficient semantic labeling of 3d point clouds," in *2012 IEEE/RSJ International Conference on Intelligent Robots and Systems*, pp. 2690–2697, IEEE, 2012.

- [9] A. Boulch, B. Le Saux, and N. Audebert, "Unstructured point cloud semantic labeling using deep segmentation networks.," in *3DOR*, 2017.
- [10] P. Tang, D. Huber, B. Akinci, R. Lipman, and A. Lytle, "Automatic reconstruction of as-built building information models from laser-scanned point clouds: A review of related techniques," *Automation in construction*, vol. 19, no. 7, pp. 829–843, 2010.
- [11] R. Volk, J. Stengel, and F. Schultmann, "Building information modeling (bim) for existing buildings literature review and future needs," *Automation in construction*, vol. 38, pp. 109–127, 2014.
- [12] K. Lim, P. Treitz, M. Wulder, B. St-Onge, and M. Flood, "Lidar remote sensing of forest structure," *Progress in physical geography*, vol. 27, no. 1, pp. 88–106, 2003.
- [13] L. Wallace, A. Lucieer, C. Watson, and D. Turner, "Development of a uav-lidar system with application to forest inventory," *Remote Sensing*, vol. 4, no. 6, pp. 1519–1543, 2012.
- [14] R. B. Rusu, Z. C. Marton, N. Blodow, M. Dolha, and M. Beetz, "Towards 3d point cloud based object maps for household environments," *Robotics and Autonomous Systems*, vol. 56, no. 11, pp. 927–941, 2008.
- [15] X. Chen, H. Ma, J. Wan, B. Li, and T. Xia, "Multi-view 3d object detection network for autonomous driving," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 1907–1915, 2017.
- [16] A. Nguyen and B. Le, "3d point cloud segmentation: A survey," in *2013 6th IEEE conference on robotics, automation and mechatronics (RAM)*, pp. 225–230, IEEE, 2013.
- [17] E. Grilli, F. Menna, and F. Remondino, "A review of point clouds segmentation and classification algorithms," in *The International Archives of Photogrammetry, Remote Sensing and Spatial Information Sciences*, vol. 42, p. 339, 2017.
- [18] E. P. Baltsavias, "A comparison between photogrammetry and laser scanning," *ISPRS Journal of photogrammetry and Remote Sensing*, vol. 54, no. 2-3, pp. 83–94, 1999.
- [19] M. J. Westoby, J. Brasington, N. F. Glasser, M. J. Hambrey, and J. Reynolds, "structure-from-motion photogrammetry: A low-cost, effective tool for geoscience applications," *Geomorphology*, vol. 179, pp. 300–314, 2012.
- [20] E. M. Mikhail, J. S. Bethel, and J. C. McGlone, "Introduction to modern photogrammetry," *New York*, 2001.
- [21] H. Hirschmüller, "Accurate and efficient stereo processing by semiglobal matching and mutual information," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 807–814, 2005.
- [22] H. Hirschmüller, "Stereo processing by semiglobal matching and mutual information," *IEEE Transactions on pattern analysis and machine intelligence*, vol. 30, no. 2, pp. 328–341, 2008.
- [23] H. Hirschmüller and D. Scharstein, "Evaluation of cost functions for stereo matching," in *2007 IEEE Conference on Computer Vision and Pattern Recognition*, pp. 1–8, IEEE, 2007.
- [24] Y. Furukawa and J. Ponce, "Accurate, dense, and robust multiview stereopsis," *IEEE transactions on pattern analysis and machine intelligence*, vol. 32, no. 8, pp. 1362–1376, 2010.
- [25] F. Nex and F. Remondino, "Uav for 3d mapping applications: a review," *Applied geomatics*, vol. 6, no. 1, pp. 1–15, 2014.
- [26] N. Snavely, S. M. Seitz, and R. Szeliski, "Photo tourism: exploring photo collections in 3d," in *ACM transactions on graphics (TOG)*, vol. 25, pp. 835–846, ACM, 2006.
- [27] N. Snavely, S. M. Seitz, and R. Szeliski, "Modeling the world from internet photo collections," *International journal of computer vision*, vol. 80, no. 2, pp. 189–210, 2008.
- [28] J. Xiao, A. Owens, and A. Torralba, "Sun3d: A database of big spaces reconstructed using sfm and object labels," in *Proceedings of the IEEE International Conference on Computer Vision*, pp. 1625–1632, 2013.
- [29] J. Shan and C. K. Toth, *Topographic laser ranging and scanning: principles and processing*. CRC press, 2018.
- [30] R. Qin, J. Tian, and P. Reinartz, "3d change detection—approaches and applications," *ISPRS Journal of Photogrammetry and Remote Sensing*, vol. 122, pp. 41–56, 2016.
- [31] F. Rottensteiner, G. Sohn, M. Gerke, and J. D. Wegner, "Isprs test project on urban classification and 3d building reconstruction," *Commission III-Photogrammetric Computer Vision and Image Analysis, Working Group III/4-3D Scene Analysis*, pp. 1–17, 2013.
- [32] F. Morsdorf, C. Nichol, T. Malthus, and I. H. Woodhouse, "Assessing forest structural and physiological information content of multi-spectral lidar waveforms by radiative transfer modelling," *Remote Sensing of Environment*, vol. 113, no. 10, pp. 2152–2163, 2009.
- [33] A. Wallace, C. Nichol, and I. Woodhouse, "Recovery of forest canopy parameters by inversion of multispectral lidar data," *Remote Sensing*, vol. 4, no. 2, pp. 509–531, 2012.
- [34] T. Hackel, N. Savinov, L. Ladicky, J. Wegner, K. Schindler, and M. Pollefeys, "Semantic3d. net: a new large-scale point cloud classification benchmark," *ISPRS Annals of Photogrammetry, Remote Sensing and Spatial Information Sciences*, pp. 91–98, 2017.
- [35] M. Brédif, B. Vallet, A. Serna, B. Marcotegui, and N. Paparoditis, "Terramobilita/iqmulus urban point cloud classification benchmark," in *Workshop on Processing Large Geospatial Data*, 2014.
- [36] X. Roynard, J.-E. Deschaud, and F. Goulette, "Paris-lille-3d: A large and high-quality ground-truth urban point cloud dataset for automatic segmentation and classification," *The International Journal of Robotics Research*, vol. 37, no. 6, pp. 545–557, 2018.
- [37] T. Sankey, J. Donager, J. McVay, and J. B. Sankey, "Uav lidar and hyperspectral fusion for forest monitoring in the southwestern usa," *Remote Sensing of Environment*, vol. 195, pp. 30–43, 2017.
- [38] X. Zhang, R. Gao, Q. Sun, and J. Cheng, "An automated rectification method for unmanned aerial vehicle lidar point cloud data based on laser intensity," *Remote Sensing*, vol. 11, no. 7, p. 811, 2019.
- [39] J. Li, B. Yang, Y. Cong, L. Cao, X. Fu, and Z. Dong, "3d forest mapping using a low-cost uav laser scanning system: Investigation and comparison," *Remote Sensing*, vol. 11, no. 6, p. 717, 2019.
- [40] J. Han, L. Shao, D. Xu, and J. Shotton, "Enhanced computer vision with microsoft kinect sensor: A review," *IEEE transactions on cybernetics*, vol. 43, no. 5, pp. 1318–1334, 2013.
- [41] S. Mattoccia and M. Poggi, "A passive rgbd sensor for accurate and real-time depth sensing self-contained into an fpga," in *Proceedings of the 9th International Conference on Distributed Smart Cameras*, pp. 146–151, ACM, 2015.
- [42] E. Lachat, H. Macher, M. Mittet, T. Landes, and P. Grussenmeyer, "First experiences with kinect v2 sensor for close range 3d modelling," in *The International Archives of Photogrammetry, Remote Sensing and Spatial Information Sciences*, vol. 40, p. 93, 2015.
- [43] A. Dai, A. X. Chang, M. Savva, M. Halber, T. Funkhouser, and M. Nießner, "ScanNet: Richly-annotated 3d reconstructions of indoor scenes," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 5828–5839, 2017.
- [44] I. Armeni, O. Sener, A. R. Zamir, H. Jiang, I. Brilakis, M. Fischer, and S. Savarese, "3d semantic parsing of large-scale indoor spaces," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 1534–1543, 2016.
- [45] M. Shahzad, X. X. Zhu, and R. Bamler, "Facade structure reconstruction using spaceborne tomosar point clouds," in *2012 IEEE International Geoscience and Remote Sensing Symposium*, pp. 467–470, IEEE, 2012.
- [46] X. X. Zhu and M. Shahzad, "Facade reconstruction using multiview spaceborne tomosar point clouds," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 52, no. 6, pp. 3541–3552, 2014.
- [47] M. Shahzad and X. X. Zhu, "Robust reconstruction of building facades for large areas using spaceborne tomosar point clouds," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 53, no. 2, pp. 752–769, 2015.
- [48] M. Shahzad, M. Schmitt, and X. X. Zhu, "Segmentation and crown parameter extraction of individual trees in an airborne tomosar point cloud," in *International Archives of Photogrammetry, Remote Sensing and Spatial Information Sciences*, vol. 40, pp. 205–209, 2015.
- [49] M. Schmitt, M. Shahzad, and X. X. Zhu, "Reconstruction of individual trees from multi-aspect tomosar data," *Remote Sensing of Environment*, vol. 165, pp. 175–185, 2015.
- [50] R. Bamler, M. Eineder, N. Adam, X. X. Zhu, and S. Gernhardt, "Interferometric potential of high resolution spaceborne sar," *Photogrammetrie-Fernerkundung-Geoinformation*, vol. 2009, no. 5, pp. 407–419, 2009.
- [51] X. X. Zhu and R. Bamler, "Very high resolution spaceborne sar tomography in urban environment," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 48, no. 12, pp. 4296–4308, 2010.
- [52] S. Gernhardt, N. Adam, M. Eineder, and R. Bamler, "Potential of very high resolution sar for persistent scatterer interferometry in urban areas," *Annals of GIS*, vol. 16, no. 2, pp. 103–111, 2010.
- [53] S. Gernhardt, X. Cong, M. Eineder, S. Hinz, and R. Bamler, "Geometrical fusion of multitrack ps point clouds," *IEEE Geoscience and Remote Sensing Letters*, vol. 9, no. 1, pp. 38–42, 2012.
- [54] X. X. Zhu and R. Bamler, "Super-resolution power and robustness of compressive sensing for spectral estimation with application to spaceborne tomographic sar," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 50, no. 1, pp. 247–258, 2012.

- [55] S. Montazeri, F. Rodríguez González, and X. X. Zhu, "Geocoding error correction for insar point clouds," *Remote Sensing*, vol. 10, no. 10, p. 1523, 2018.
- [56] F. Rottensteiner and C. Briese, "A new method for building extraction in urban areas from high-resolution lidar data," in *International Archives of Photogrammetry Remote Sensing and Spatial Information Sciences*, vol. 34, pp. 295–301, 2002.
- [57] X. X. Zhu and R. Bamler, "Demonstration of super-resolution for tomographic sar imaging in urban environment," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 50, no. 8, pp. 3150–3157, 2012.
- [58] X. X. Zhu, M. Shahzad, and R. Bamler, "From tomosar point clouds to objects: Facade reconstruction," in *2012 Tyrrhenian Workshop on Advances in Radar and Remote Sensing (TyWRRS)*, pp. 106–113, IEEE, 2012.
- [59] X. X. Zhu and R. Bamler, "Let's do the time warp: Multicomponent nonlinear motion estimation in differential sar tomography," *IEEE Geoscience and Remote Sensing Letters*, vol. 8, no. 4, pp. 735–739, 2011.
- [60] S. Auer, S. Gernhardt, and R. Bamler, "Ghost persistent scatterers related to multiple signal reflections," *IEEE Geoscience and Remote Sensing Letters*, vol. 8, no. 5, pp. 919–923, 2011.
- [61] Y. Shi, X. X. Zhu, and R. Bamler, "Nonlocal compressive sensing-based sar tomography," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 57, no. 5, pp. 3015–3024, 2019.
- [62] Y. Wang and X. X. Zhu, "Automatic feature-based geometric fusion of multiview tomosar point clouds in urban area," *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, vol. 8, no. 3, pp. 953–965, 2014.
- [63] M. Schmitt and X. X. Zhu, "Data fusion and remote sensing: An ever-growing relationship," *IEEE Geoscience and Remote Sensing Magazine*, vol. 4, no. 4, pp. 6–23, 2016.
- [64] Y. Wang, X. X. Zhu, B. Zeisl, and M. Pollefeys, "Fusing meter-resolution 4-d insar point clouds and optical images for semantic urban infrastructure monitoring," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 55, no. 1, pp. 14–26, 2017.
- [65] A. Adam, E. Chatzilari, S. Nikolopoulos, and I. Kompatsiaris, "H-ransac: A hybrid point cloud segmentation combining 2d and 3d data," *ISPRS Annals of Photogrammetry, Remote Sensing & Spatial Information Sciences*, vol. 4, no. 2, 2018.
- [66] J. Bauer, K. Karner, K. Schindler, A. Klaus, and C. Zach, "Segmentation of building from dense 3d point-clouds," in *Proceedings of the ISPRS. Workshop Laser scanning Enschede*, pp. 12–14, 2005.
- [67] A. Boulch, J. Guerry, B. Le Saux, and N. Audebert, "Snapnet: 3d point cloud semantic labeling with 2d deep segmentation networks," *Computers & Graphics*, vol. 71, pp. 189–198, 2018.
- [68] H. Su, V. Jampani, D. Sun, S. Maji, E. Kalogerakis, M.-H. Yang, and J. Kautz, "Splatnet: Sparse lattice networks for point cloud processing," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 2530–2539, 2018.
- [69] G. Riegler, A. Osman Ulusoy, and A. Geiger, "Octnet: Learning deep 3d representations at high resolutions," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 3577–3586, 2017.
- [70] C. Choy, J. Gwak, and S. Savarese, "4d spatio-temporal convnets: Minkowski convolutional neural networks," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 3075–3084, 2019.
- [71] F. A. Limberger and M. M. Oliveira, "Real-time detection of planar regions in unorganized point clouds," *Pattern Recognition*, vol. 48, no. 6, pp. 2043–2053, 2015.
- [72] B. Xu, W. Jiang, J. Shan, J. Zhang, and L. Li, "Investigation on the weighted ransac approaches for building roof plane segmentation from lidar point clouds," *Remote Sensing*, vol. 8, no. 1, p. 5, 2015.
- [73] D. Chen, L. Zhang, P. T. Mathiopoulos, and X. Huang, "A methodology for automated segmentation and reconstruction of urban 3-d buildings from als point clouds," *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, vol. 7, no. 10, pp. 4199–4217, 2014.
- [74] F. Tarsha-Kurdi, T. Landes, and P. Grussenmeyer, "Hough-transform and extended ransac algorithms for automatic detection of 3d building roof planes from lidar data," in *ISPRS Workshop on Laser Scanning 2007 and SilviLaser 2007*, vol. 36, pp. 407–412, 2007.
- [75] B. Gorte, "Segmentation of tin-structured surface models," in *International Archives of Photogrammetry Remote Sensing and Spatial Information Sciences*, vol. 34, pp. 465–469, 2002.
- [76] A. Sampath and J. Shan, "Clustering based planar roof extraction from lidar data," in *American Society for Photogrammetry and Remote Sensing Annual Conference, Reno, Nevada, May*, pp. 1–6, 2006.
- [77] A. Sampath and J. Shan, "Segmentation and reconstruction of polyhedral building roofs from aerial lidar point clouds," *IEEE Transactions on geoscience and remote sensing*, vol. 48, no. 3, pp. 1554–1567, 2010.
- [78] S. Ural and J. Shan, "Min-cut based segmentation of airborne lidar point clouds," in *International Archives of the Photogrammetry, Remote Sensing and Spatial Information Sciences*, pp. 167–172, 2012.
- [79] J. Yan, J. Shan, and W. Jiang, "A global optimization approach to roof segmentation from airborne lidar point clouds," *ISPRS journal of photogrammetry and remote sensing*, vol. 94, pp. 183–193, 2014.
- [80] T. Melzer, "Non-parametric segmentation of als point clouds using mean shift," *Journal of Applied Geodesy Jag*, vol. 1, no. 3, pp. 159–170, 2007.
- [81] W. Yao, S. Hinz, and U. Stilla, "Object extraction based on 3d-segmentation of lidar data by combining mean shift with normalized cuts: Two examples from urban areas," in *2009 Joint Urban Remote Sensing Event*, pp. 1–6, IEEE, 2009.
- [82] S. K. Lodha, D. M. Fitzpatrick, and D. P. Helmbold, "Aerial lidar data classification using adaboost," in *Sixth International Conference on 3-D Digital Imaging and Modeling (3DIM 2007)*, pp. 435–442, IEEE, 2007.
- [83] M. Carlberg, P. Gao, G. Chen, and A. Zakhor, "Classifying urban landscape in aerial lidar using 3d shape analysis," in *2009 16th IEEE International Conference on Image Processing (ICIP)*, pp. 1701–1704, IEEE, 2009.
- [84] N. Chehata, L. Guo, and C. Mallet, "Airborne lidar feature selection for urban classification using random forests," in *International Archives of Photogrammetry, Remote Sensing and Spatial Information Sciences*, vol. 38, pp. 207–212, 2009.
- [85] R. Shapovalov, E. Velizhev, and O. Barinova, "Nonassociative markov networks for 3d point cloud classification," in *International Archives of the Photogrammetry, Remote Sensing and Spatial Information Sciences*, vol. 38, pp. 103–108, 2010.
- [86] J. Niemeyer, F. Rottensteiner, and U. Soergel, "Conditional random fields for lidar point cloud classification in complex urban areas," in *ISPRS annals of the photogrammetry, remote sensing and spatial information sciences*, vol. 3, pp. 263–268, 2012.
- [87] J. Niemeyer, F. Rottensteiner, and U. Soergel, "Contextual classification of lidar data and building object detection in urban areas," *ISPRS journal of photogrammetry and remote sensing*, vol. 87, pp. 152–165, 2014.
- [88] G. Vosselman, M. Coenen, and F. Rottensteiner, "Contextual segment-based classification of airborne laser scanner data," *ISPRS journal of photogrammetry and remote sensing*, vol. 128, pp. 354–371, 2017.
- [89] X. Xiong, D. Munoz, J. A. Bagnell, and M. Hebert, "3-d scene analysis via sequenced predictions over points and regions," in *2011 IEEE International Conference on Robotics and Automation*, pp. 2609–2616, IEEE, 2011.
- [90] M. Najafi, S. T. Namin, M. Salzmann, and L. Petersson, "Non-associative higher-order markov networks for point cloud classification," in *European Conference on Computer Vision*, pp. 500–515, Springer, 2014.
- [91] F. Morsdorf, E. Meier, B. Kötz, K. I. Itten, M. Dobbertin, and B. Allgöwer, "Lidar-based geometric reconstruction of boreal type forest stands at single tree level for forest and wildland fire management," *Remote Sensing of Environment*, vol. 92, no. 3, pp. 353–362, 2004.
- [92] A. Ferraz, F. Bretar, S. Jacquemoud, G. Gonçalves, and L. Pereira, "3d segmentation of forest structure using a mean-shift based algorithm," in *2010 IEEE International Conference on Image Processing*, pp. 1413–1416, IEEE, 2010.
- [93] A.-V. Vo, L. Truong-Hong, D. F. Lafer, and M. Bertolotto, "Octree-based region growing for point cloud segmentation," *ISPRS Journal of Photogrammetry and Remote Sensing*, vol. 104, pp. 88–100, 2015.
- [94] A. Nurunnabi, D. Belton, and G. West, "Robust segmentation in laser scanning 3d point cloud data," in *2012 International Conference on Digital Image Computing Techniques and Applications (DICTA)*, pp. 1–8, IEEE, 2012.
- [95] M. Weinmann, B. Jutzi, S. Hinz, and C. Mallet, "Semantic point cloud interpretation based on optimal neighborhoods, relevant features and efficient classifiers," *ISPRS Journal of Photogrammetry and Remote Sensing*, vol. 105, pp. 286–304, 2015.
- [96] D. Munoz, J. A. Bagnell, N. Vandapel, and M. Hebert, "Contextual classification with functional max-margin markov networks," in *2009 IEEE Conference on Computer Vision and Pattern Recognition*, pp. 975–982, IEEE, 2009.

- [97] L. Landrieu, H. Raguét, B. Vallet, C. Mallet, and M. Weinmann, "A structured regularization framework for spatially smoothing semantic labelings of 3d point clouds," *ISPRS Journal of Photogrammetry and Remote Sensing*, vol. 132, pp. 102–118, 2017.
- [98] L. Tchapmi, C. Choy, I. Armeni, J. Gwak, and S. Savarese, "Segcloud: Semantic segmentation of 3d point clouds," in *2017 International Conference on 3D Vision (3DV)*, pp. 537–547, IEEE, 2017.
- [99] X. Ye, J. Li, H. Huang, L. Du, and X. Zhang, "3d recurrent neural networks with context fusion for point cloud semantic segmentation," in *Proceedings of the European Conference on Computer Vision (ECCV)*, pp. 403–417, 2018.
- [100] L. Landrieu and M. Boussaha, "Point cloud oversegmentation with graph-structured deep metric learning," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 7440–7449, 2019.
- [101] J. Xiao, J. Zhang, B. Adler, H. Zhang, and J. Zhang, "Three-dimensional point cloud plane segmentation in both structured and unstructured environments," *Robotics and Autonomous Systems*, vol. 61, no. 12, pp. 1641–1652, 2013.
- [102] L. Li, F. Yang, H. Zhu, D. Li, Y. Li, and L. Tang, "An improved ransac for 3d point cloud plane segmentation based on normal distribution transformation cells," *Remote Sensing*, vol. 9, no. 5, p. 433, 2017.
- [103] H. Boulaassal, T. Landes, P. Grussenmeyer, and F. Tarsha-Kurdi, "Automatic segmentation of building facades using terrestrial laser data," in *ISPRS Workshop on Laser Scanning 2007 and SilviLaser 2007*, pp. 65–70, 2007.
- [104] Z. Dong, B. Yang, P. Hu, and S. Scherer, "An efficient global energy optimization approach for robust 3d plane segmentation of point clouds," *ISPRS Journal of Photogrammetry and Remote Sensing*, vol. 137, pp. 112–133, 2018.
- [105] J. M. Biosca and J. L. Lerma, "Unsupervised robust planar segmentation of terrestrial laser scanner point clouds based on fuzzy clustering methods," *ISPRS Journal of Photogrammetry and Remote Sensing*, vol. 63, no. 1, pp. 84–98, 2008.
- [106] X. Ning, X. Zhang, Y. Wang, and M. Jaeger, "Segmentation of architecture shape information from 3d point cloud," in *Proceedings of the 8th International Conference on Virtual Reality Continuum and its Applications in Industry*, pp. 127–132, ACM, 2009.
- [107] Y. Xu, S. Tattas, and U. Stilla, "Segmentation of 3d outdoor scenes using hierarchical clustering structure and perceptual grouping laws," in *2016 9th IAPR Workshop on Pattern Recognition in Remote Sensing (PRRS)*, pp. 1–6, IEEE, 2016.
- [108] Y. Xu, L. Hoegner, S. Tattas, and U. Stilla, "Voxel-and graph-based point cloud segmentation of 3d scenes using perceptual grouping laws," in *ISPRS Annals of Photogrammetry, Remote Sensing & Spatial Information Sciences*, vol. 4, 2017.
- [109] Y. Xu, W. Yao, S. Tattas, L. Hoegner, and U. Stilla, "Unsupervised segmentation of point clouds from buildings using hierarchical clustering based on gestalt principles," *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, no. 99, pp. 1–17, 2018.
- [110] E. H. Lim and D. Suter, "3d terrestrial lidar classifications with super-voxels and multi-scale conditional random fields," *Computer-Aided Design*, vol. 41, no. 10, pp. 701–710, 2009.
- [111] Z. Li, L. Zhang, X. Tong, B. Du, Y. Wang, L. Zhang, Z. Zhang, H. Liu, J. Mei, X. Xing, et al., "A three-step approach for tfs point cloud classification," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 54, no. 9, pp. 5412–5424, 2016.
- [112] L. Wang, Y. Huang, Y. Hou, S. Zhang, and J. Shan, "Graph attention convolution for point cloud semantic segmentation," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 10296–10305, 2019.
- [113] J.-F. Lalonde, R. Unnikrishnan, N. Vandapel, and M. Hebert, "Scale selection for classification of point-sampled 3d surfaces," in *Fifth International Conference on 3-D Digital Imaging and Modeling (3DIM'05)*, pp. 285–292, IEEE, 2005.
- [114] D. Borrmann, J. Elseberg, K. Lingemann, and A. Nüchter, "The 3d hough transform for plane detection in point clouds: A review and a new accumulator design," *3D Research*, vol. 2, no. 2, p. 3, 2011.
- [115] R. Hulik, M. Spanel, P. Smrz, and Z. Materna, "Continuous plane detection in point-cloud data based on 3d hough transform," *Journal of visual communication and image representation*, vol. 25, no. 1, pp. 86–97, 2014.
- [116] K. Khoshelham and S. O. Elberink, "Accuracy and resolution of kinect depth data for indoor mapping applications," *Sensors*, vol. 12, no. 2, pp. 1437–1454, 2012.
- [117] R. Shapovalov, D. Vetrov, and P. Kohli, "Spatial inference machines," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 2985–2992, 2013.
- [118] Q. Huang, W. Wang, and U. Neumann, "Recurrent slice networks for 3d segmentation of point clouds," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 2626–2635, 2018.
- [119] Y. Li, R. Bu, M. Sun, W. Wu, X. Di, and B. Chen, "Pointcnn: Convolution on x-transformed points," in *Advances in Neural Information Processing Systems*, pp. 828–838, 2018.
- [120] C. R. Qi, L. Yi, H. Su, and L. J. Guibas, "Pointnet++: Deep hierarchical feature learning on point sets in a metric space," in *Advances in Neural Information Processing Systems*, pp. 5099–5108, 2017.
- [121] M. Jiang, Y. Wu, and C. Lu, "Pointsift: A sift-like network module for 3d point cloud semantic segmentation," *arXiv preprint arXiv:1807.00652*, 2018.
- [122] W. Wu, Z. Qi, and L. Fuxin, "Pointconv: Deep convolutional networks on 3d point clouds," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 9621–9630, 2019.
- [123] X. Wang, S. Liu, X. Shen, C. Shen, and J. Jia, "Associatively segmenting instances and semantics in point clouds," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 4096–4105, 2019.
- [124] Q.-H. Pham, T. Nguyen, B.-S. Hua, G. Roig, and S.-K. Yeung, "Jsis3d: Joint semantic-instance segmentation of 3d point clouds with multi-task pointwise networks and multi-value conditional random fields," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 8827–8836, 2019.
- [125] A. Komarichev, Z. Zhong, and J. Hua, "A-cnn: Annularly convolutional neural networks on point clouds," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 7421–7430, 2019.
- [126] H. Zhao, L. Jiang, C.-W. Fu, and J. Jia, "Pointweb: Enhancing local neighborhood features for point cloud processing," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 5565–5573, 2019.
- [127] W. Wang, R. Yu, Q. Huang, and U. Neumann, "Sgpn: Similarity group proposal network for 3d point cloud instance segmentation," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 2569–2578, 2018.
- [128] L. Yi, W. Zhao, H. Wang, M. Sung, and L. J. Guibas, "Gspn: Generative shape proposal network for 3d instance segmentation in point cloud," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 3947–3956, 2019.
- [129] T. Rabbani and F. Van Den Heuvel, "Efficient hough transform for automatic detection of cylinders in point clouds," in *International Archives of the Photogrammetry, Remote Sensing and Spatial Information Sciences*, vol. 3, pp. 60–65, 2005.
- [130] T.-T. Tran, V.-T. Cao, and D. Laurendeau, "Extraction of cylinders and estimation of their parameters from point clouds," *Computers & Graphics*, vol. 46, pp. 345–357, 2015.
- [131] V.-H. Le, H. Vu, T. T. Nguyen, T.-L. Le, and T.-H. Tran, "Acquiring qualified samples for ransac using geometrical constraints," *Pattern Recognition Letters*, vol. 102, pp. 58–66, 2018.
- [132] H. Riemenschneider, A. Bódis-Szomorú, J. Weissenberg, and L. Van Gool, "Learning where to classify in multi-view semantic segmentation," in *European Conference on Computer Vision*, pp. 516–532, Springer, 2014.
- [133] M. De Deuge, A. Quadros, C. Hung, and B. Douillard, "Unsupervised feature learning for classification of outdoor 3d scans," in *Australasian Conference on Robotics and Automation*, vol. 2, 2013.
- [134] A. Serna, B. Marcotegui, F. Goulette, and J.-E. Deschaud, "Paris-rue-madame database: a 3d mobile laser scanner dataset for benchmarking urban detection, segmentation and classification methods," in *4th International Conference on Pattern Recognition, Applications and Methods ICPRAM 2014*, 2014.
- [135] A. Geiger, P. Lenz, C. Stiller, and R. Urtasun, "Vision meets robotics: The kitti dataset," *The International Journal of Robotics Research*, vol. 32, no. 11, pp. 1231–1237, 2013.
- [136] N. Silberman, D. Hoiem, P. Kohli, and R. Fergus, "Indoor segmentation and support inference from rgb-d images," in *European Conference on Computer Vision*, pp. 746–760, Springer, 2012.
- [137] I. Armeni, S. Sax, A. R. Zamir, and S. Savarese, "Joint 2d-3d-semantic data for indoor scene understanding," *arXiv preprint arXiv:1702.01105*, 2017.
- [138] T. Rabbani, F. Van Den Heuvel, and G. Vosselmann, "Segmentation of point clouds using smoothness constraint," in *International archives*

- of photogrammetry, remote sensing and spatial information sciences, vol. 36, pp. 248–253, 2006.
- [139] B. Bhanu, S. Lee, C.-C. Ho, and T. Henderson, “Range data processing: Representation of surfaces by edges,” in *Proceedings of the Eighth International Conference on Pattern Recognition*, pp. 236–238, IEEE Computer Society Press, 1986.
- [140] X. Y. Jiang, U. Meier, and H. Bunke, “Fast range image segmentation using high-level segmentation primitives,” in *Proceedings Third IEEE Workshop on Applications of Computer Vision*, pp. 83–88, IEEE, 1996.
- [141] A. D. Sappa and M. Devy, “Fast range image segmentation by an edge detection strategy,” in *Proceedings Third International Conference on 3-D Digital Imaging and Modeling*, pp. 292–299, IEEE, 2001.
- [142] M. A. Wani and H. R. Arabnia, “Parallel edge-region-based segmentation algorithm targeted at reconfigurable multiring network,” *The Journal of Supercomputing*, vol. 25, no. 1, pp. 43–62, 2003.
- [143] E. Castillo, J. Liang, and H. Zhao, “Point cloud segmentation and denoising via constrained nonlinear least squares normal estimates,” in *Innovations for Shape Analysis*, pp. 283–299, Springer, 2013.
- [144] P. J. Besl and R. C. Jain, “Segmentation through variable-order surface fitting,” *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 10, no. 2, pp. 167–192, 1988.
- [145] R. Geibel and U. Stilla, “Segmentation of laser altimeter data for building reconstruction: different procedures and comparison,” in *International Archives of Photogrammetry and Remote Sensing*, vol. 33, pp. 326–334, 2000.
- [146] D. Tóvári and N. Pfeifer, “Segmentation based robust interpolation—a new approach to laser data filtering,” in *International Archives of Photogrammetry, Remote Sensing and Spatial Information Sciences*, vol. 36, pp. 79–84, 2005.
- [147] J.-E. Deschaud and F. Goulette, “A fast and accurate plane detection algorithm for large noisy point clouds using filtered normals and voxel growing,” in *3DPVT*, 2010.
- [148] P. V. Hough, “Method and means for recognizing complex patterns,” 1962. US Patent 3,069,654.
- [149] L. Xu, E. Oja, and P. Kultanen, “A new curve detection method: randomized hough transform (rht),” *Pattern recognition letters*, vol. 11, no. 5, pp. 331–338, 1990.
- [150] R. O. Duda and P. E. Hart, “Use of the hough transformation to detect lines and curves in pictures,” *Communications of the ACM*, vol. 15, no. 1, pp. 11–15, 1972.
- [151] A. Kaiser, J. A. Ybanez Zepeda, and T. Boubekeur, “A survey of simple geometric primitives detection methods for captured 3d data,” in *Computer Graphics Forum*, vol. 38, pp. 167–196, Wiley Online Library, 2019.
- [152] N. Kiryati, Y. Eldar, and A. M. Bruckstein, “A probabilistic hough transform,” *Pattern recognition*, vol. 24, no. 4, pp. 303–316, 1991.
- [153] A. Yla-Jaaski and N. Kiryati, “Adaptive termination of voting in the probabilistic circular hough transform,” *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 16, no. 9, pp. 911–915, 1994.
- [154] C. Galamhos, J. Matas, and J. Kittler, “Progressive probabilistic hough transform for line detection,” in *Proceedings. 1999 IEEE computer society conference on computer vision and pattern recognition*, vol. 1, pp. 554–560, IEEE, 1999.
- [155] L. A. Fernandes and M. M. Oliveira, “Real-time line detection through an improved hough transform voting scheme,” *Pattern recognition*, vol. 41, no. 1, pp. 299–314, 2008.
- [156] G. Vosselman, B. G. Gorte, G. Sithole, and T. Rabbani, “Recognising structure in laser scanner point clouds,” in *International archives of photogrammetry, remote sensing and spatial information sciences*, vol. 46, pp. 33–38, 2004.
- [157] M. Camurri, R. Vezzani, and R. Cucchiara, “3d hough transform for sphere recognition on point clouds,” *Machine vision and applications*, vol. 25, no. 7, pp. 1877–1891, 2014.
- [158] M. A. Fischler and R. C. Bolles, “Random sample consensus: a paradigm for model fitting with applications to image analysis and automated cartography,” *Communications of the ACM*, vol. 24, no. 6, pp. 381–395, 1981.
- [159] S. Choi, T. Kim, and W. Yu, “Performance evaluation of ransac family,” in *Proceedings of the British Machine Vision Conference*, 2009.
- [160] R. Raguram, J.-M. Frahm, and M. Pollefeys, “A comparative analysis of ransac techniques leading to adaptive real-time random sample consensus,” in *European Conference on Computer Vision*, pp. 500–513, Springer, 2008.
- [161] R. Raguram, O. Chum, M. Pollefeys, J. Matas, and J.-M. Frahm, “Usac: a universal framework for random sample consensus,” *IEEE transactions on pattern analysis and machine intelligence*, vol. 35, no. 8, pp. 2022–2038, 2013.
- [162] R. Schnabel, R. Wahl, and R. Klein, “Efficient ransac for point-cloud shape detection,” in *Computer graphics forum*, vol. 26, pp. 214–226, Wiley Online Library, 2007.
- [163] P. Biber and W. Straßer, “The normal distributions transform: A new approach to laser scan matching,” in *Proceedings 2003 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS 2003)(Cat. No. 03CH37453)*, vol. 3, pp. 2743–2748, IEEE, 2003.
- [164] V. Fragoso, P. Sen, S. Rodriguez, and M. Turk, “Evsac: accelerating hypotheses generation by modeling matching scores with extreme value theory,” in *Proceedings of the IEEE International Conference on Computer Vision*, pp. 2472–2479, 2013.
- [165] D. Barath and J. Matas, “Graph-cut ransac,” in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 6733–6741, 2018.
- [166] S. Filin, “Surface clustering from airborne laser scanning data,” in *International Archives of Photogrammetry Remote Sensing and Spatial Information Sciences*, vol. 34, pp. 119–124, 2002.
- [167] A. Golovinskiy and T. Funkhouser, “Min-cut based segmentation of point clouds,” in *IEEE 12th International Conference on Computer Vision Workshops, ICCV Workshops*, pp. 39–46, IEEE, 2009.
- [168] D. Comaniciu and P. Meer, “Mean shift analysis and applications,” in *Proceedings of the Seventh IEEE International Conference on Computer Vision*, vol. 2, pp. 1197–1203, IEEE, 1999.
- [169] D. Comaniciu and P. Meer, “Mean shift: A robust approach toward feature space analysis,” *IEEE Transactions on Pattern Analysis & Machine Intelligence*, no. 5, pp. 603–619, 2002.
- [170] Y. Cheng, “Mean shift, mode seeking, and clustering,” *IEEE transactions on pattern analysis and machine intelligence*, vol. 17, no. 8, pp. 790–799, 1995.
- [171] Y. Boykov and G. Funka-Lea, “Graph cuts and efficient nd image segmentation,” *International journal of computer vision*, vol. 70, no. 2, pp. 109–131, 2006.
- [172] A. Delong, A. Osokin, H. N. Isack, and Y. Boykov, “Fast approximate energy minimization with label costs,” *International journal of computer vision*, vol. 96, no. 1, pp. 1–27, 2012.
- [173] J. Papon, A. Abramov, M. Schoeler, and F. Worgotter, “Voxel cloud connectivity segmentation-supervoxels for point clouds,” in *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 2027–2034, 2013.
- [174] S. Song, H. Lee, and S. Jo, “Boundary-enhanced supervoxel segmentation for sparse outdoor lidar data,” *Electronics Letters*, vol. 50, no. 25, pp. 1917–1919, 2014.
- [175] Y. Lin, C. Wang, D. Zhai, W. Li, and J. Li, “Toward better boundary preserved supervoxel segmentation for 3d point clouds,” *ISPRS journal of photogrammetry and remote sensing*, vol. 143, pp. 39–47, 2018.
- [176] S. Christoph Stein, M. Schoeler, J. Papon, and F. Worgotter, “Object partitioning using local convexity,” in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 304–311, 2014.
- [177] B. Yang, Z. Dong, G. Zhao, and W. Dai, “Hierarchical extraction of urban objects from mobile laser scanning data,” *ISPRS Journal of Photogrammetry and Remote Sensing*, vol. 99, pp. 45–57, 2015.
- [178] A. Schmidt, F. Rottensteiner, and U. Sörgel, “Classification of airborne laser scanning data in wadden sea areas using conditional random fields,” in *International Archives of the Photogrammetry, Remote Sensing and Spatial Information Sciences*, vol. 39, pp. 161–166, 2012.
- [179] X. X. Zhu, D. Tuia, L. Mou, G.-S. Xia, L. Zhang, F. Xu, and F. Fraundorfer, “Deep learning in remote sensing: A comprehensive review and list of resources,” *IEEE Geoscience and Remote Sensing Magazine*, vol. 5, no. 4, pp. 8–36, 2017.
- [180] K. Simonyan and A. Zisserman, “Very deep convolutional networks for large-scale image recognition,” *arXiv preprint arXiv:1409.1556*, 2014.
- [181] K. He, X. Zhang, S. Ren, and J. Sun, “Deep residual learning for image recognition,” in *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 770–778, 2016.
- [182] R. Girshick, “Fast r-cnn,” in *Proceedings of the IEEE international conference on computer vision*, pp. 1440–1448, 2015.
- [183] S. Ren, K. He, R. Girshick, and J. Sun, “Faster r-cnn: Towards real-time object detection with region proposal networks,” in *Advances in neural information processing systems*, pp. 91–99, 2015.
- [184] J. Long, E. Shelhamer, and T. Darrell, “Fully convolutional networks for semantic segmentation,” in *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 3431–3440, 2015.
- [185] L.-C. Chen, G. Papandreou, I. Kokkinos, K. Murphy, and A. L. Yuille, “DeepLab: Semantic image segmentation with deep convolutional nets, atrous convolution, and fully connected crfs,” *IEEE transactions on*

- pattern analysis and machine intelligence*, vol. 40, no. 4, pp. 834–848, 2018.
- [186] H. Su, S. Maji, E. Kalogerakis, and E. Learned-Miller, “Multi-view convolutional neural networks for 3d shape recognition,” in *Proceedings of the IEEE international conference on computer vision*, pp. 945–953, 2015.
- [187] D. Maturana and S. Scherer, “Voxnet: A 3d convolutional neural network for real-time object recognition,” in *IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, pp. 922–928, 2015.
- [188] P.-S. Wang, Y. Liu, Y.-X. Guo, C.-Y. Sun, and X. Tong, “O-cnn: Octree-based convolutional neural networks for 3d shape analysis,” *ACM Transactions on Graphics (TOG)*, vol. 36, no. 4, p. 72, 2017.
- [189] H.-Y. Meng, L. Gao, Y. Lai, and D. Manocha, “Vv-net: Voxel vae net with group convolutions for point cloud segmentation,” *arXiv preprint arXiv:1811.04337*, 2018.
- [190] F. Engelmann, T. Kontogianni, A. Hermans, and B. Leibe, “Exploring spatial context for 3d semantic segmentation of point clouds,” in *Proceedings of the IEEE International Conference on Computer Vision*, pp. 716–724, 2017.
- [191] G. Te, W. Hu, A. Zheng, and Z. Guo, “Rgcnn: Regularized graph cnn for point cloud segmentation,” in *ACM Multimedia Conference on Multimedia Conference*, pp. 746–754, ACM, 2018.
- [192] A. X. Chang, T. Funkhouser, L. Guibas, P. Hanrahan, Q. Huang, Z. Li, S. Savarese, M. Savva, S. Song, H. Su, *et al.*, “Shapenet: An information-rich 3d model repository,” *arXiv preprint arXiv:1512.03012*, 2015.
- [193] J. Zhou, G. Cui, Z. Zhang, C. Yang, Z. Liu, and M. Sun, “Graph neural networks: A review of methods and applications,” *arXiv preprint arXiv:1812.08434*, 2018.
- [194] Z. Wu, S. Pan, F. Chen, G. Long, C. Zhang, and P. S. Yu, “A comprehensive survey on graph neural networks,” *arXiv preprint arXiv:1901.00596*, 2019.
- [195] L. Li, M. Sung, A. Dubrovina, L. Yi, and L. J. Guibas, “Supervised fitting of geometric primitives to 3d point clouds,” in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 2652–2660, 2019.

B Yuxing Xie, Jiaojiao Tian, and Xiao Xiang Zhu. “A co-learning method to utilize optical images and photogrammetric point clouds for building extraction.” *International Journal of Applied Earth Observation and Geoinformation* 116 (2023): 103165.

<https://doi.org/10.1016/j.jag.2022.103165>



Contents lists available at ScienceDirect

International Journal of Applied Earth Observation and Geoinformation

journal homepage: www.elsevier.com/locate/jag

A co-learning method to utilize optical images and photogrammetric point clouds for building extraction

Yuxing Xie^{a,b,*}, Jiaojiao Tian^b, Xiao Xiang Zhu^a^a Data Science in Earth Observation, Technical University of Munich, Munich, 80333, Germany^b The Remote Sensing Technology Institute, German Aerospace Center (DLR), Muenchener Strasse 20, Wessling, 82234, Germany

ARTICLE INFO

Keywords:

Building extraction
Co-learning
Multimodality learning
Multispectral images
Point clouds
Remote sensing

ABSTRACT

Although deep learning techniques have brought unprecedented accuracy to automatic building extraction, several main issues still constitute an obstacle to effective and practical applications. The industry is eager for higher accuracy and more flexible data usage. In this paper, we present a co-learning framework applicable to building extraction from optical images and photogrammetric point clouds, which can take the advantage of 2D/3D multimodality data. Instead of direct information fusion, our co-learning framework adaptively exploits knowledge from another modality during the training phase with a soft connection, via a predefined loss function. Compared to conventional data fusion, this method is more flexible, as it is not mandatory to provide multimodality data in the test phase. We propose two types of co-learning: a standard version and an enhanced version, depending on whether unlabeled training data are employed. Experimental results from two data sets show that the methods we present can enhance the performance of both image and point cloud networks in few-shot tasks, as well as image networks when applying fully labeled training data sets.

1. Introduction

Automatic building extraction from remotely sensed data is an important task in the photogrammetry and remote sensing field. It plays a vital role in many practical applications, such as building information modeling, urban monitoring and planning, and digital twins. Recently, advanced deep learning algorithms with high-quality data sets have achieved unprecedented performance in building extraction. However, there are still numerous problems restricting the generalization. When the deep neural network is trained with insufficient training samples, overfitting will occur and the network cannot perform accurately against unseen data. To meet the requirements of industry applications, better accuracy and less dependency on annotated training data sets are among the most urgent needs. Annotating a large amount of training data is labor intensive. Hence, studies on automatic building extraction are still ongoing, but researchers' attention has shifted from simply stacking different networks to developing targeted algorithms in order to better regularize the results, as well as designing flexible architectures to efficiently utilize multimodality data in networks, resulting in less dependent on the quantity of annotated data.

Based on the applications and corresponding data types employed, building extraction tasks are usually divided into three categories: 2D image based, 3D geometric data (point clouds/DSMs) based, and multimodality data based. Image-based automatic building extraction is

the most widely studied case, as the acquisition cost of optical images is relatively low. In recent years, deep learning-based methods, especially convolutional neural networks (CNNs), have taken the place of the traditional algorithms and became the most widely utilized, as their performance is superior on various data sets (Zhu et al., 2020; Shi et al., 2020; Li et al., 2021).

Although 2D remotely sensed images are widely used in practical applications, they have several obvious limitations. Remotely sensed images captured by airborne or spaceborne sensors usually cover much larger area, which may cause scale variation of buildings, thereby influencing the performance of algorithms. Furthermore, unavoidable reflection of light, shadows, and obstructions can also have negative effects on building extraction results (Tian et al., 2014; Sun et al., 2021b). Due to the development of LiDAR sensors and dense image matching algorithms, 3D geometric data such as point clouds and DSMs have brought new possibilities to the building extraction field and compensate for deficiencies in the images, as they can provide geometric features that are not affected by spectral distortion. Also Driven by the success of deep learning techniques, researchers have recently been keen to apply all kinds of point cloud neural networks to urban point cloud processing (Xie et al., 2020), such as PointNet (Qi et al., 2017; Huang et al., 2020a; Yousefhussein et al., 2018), KPConv (Lin et al.,

* Corresponding author at: The Remote Sensing Technology Institute, German Aerospace Center (DLR), Muenchener Strasse 20, Wessling, 82234, Germany.
E-mail address: yuxing.xie@dlr.de (Y. Xie).

<https://doi.org/10.1016/j.jag.2022.103165>

Received 25 August 2022; Received in revised form 6 December 2022; Accepted 16 December 2022

Available online 31 December 2022

1569-8432/© 2023 The Authors. Published by Elsevier B.V. This is an open access article under the CC BY-NC-ND license (<http://creativecommons.org/licenses/by-nc-nd/4.0/>).

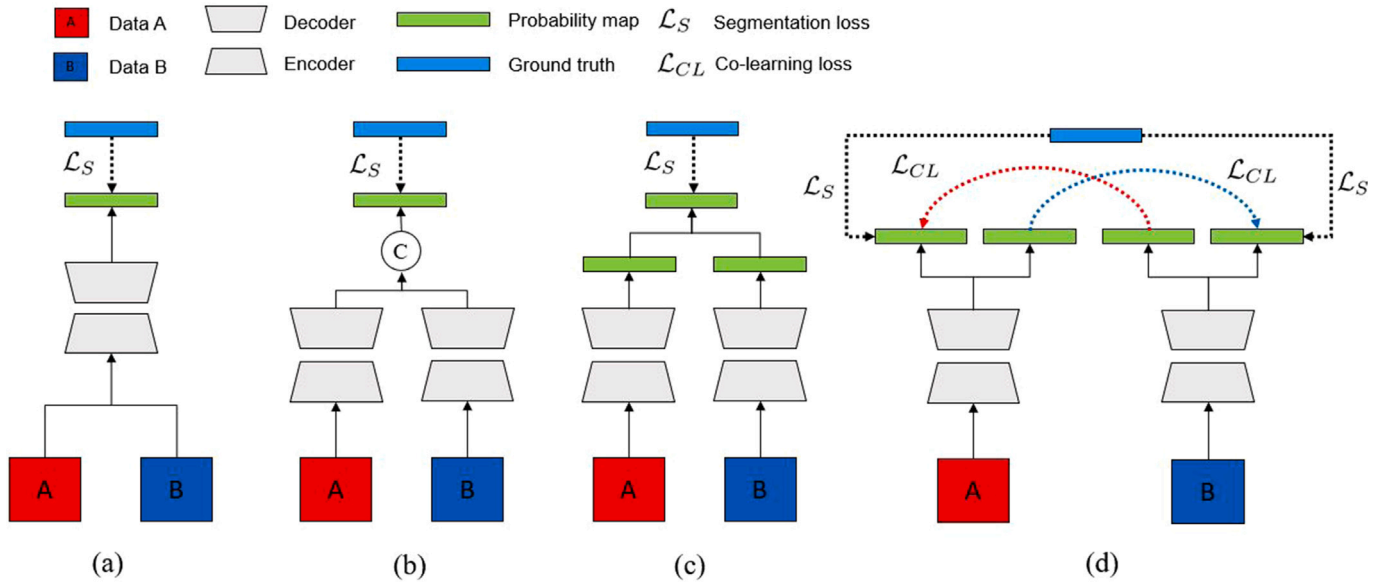


Fig. 1. The difference between conventional data fusion and co-learning. (a) Early fusion. (b) Middle fusion. (c) Late fusion. (d) Multimodality co-learning in our work.

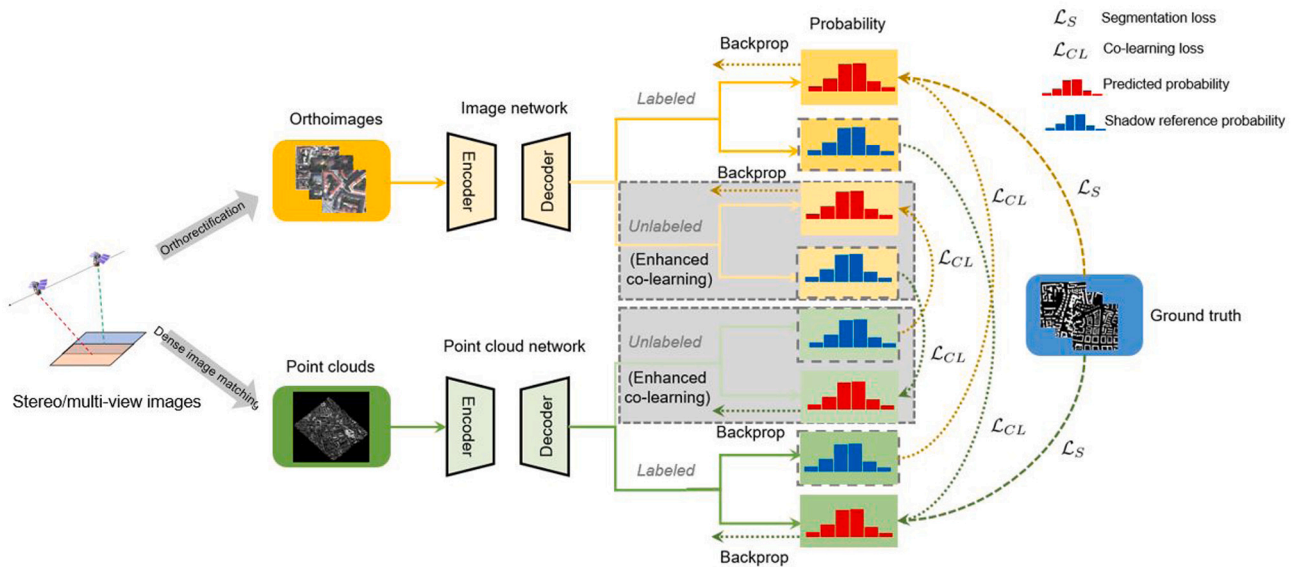


Fig. 2. The training phase of the proposed co-learning framework. In our work, images used for building extraction are orthoimages. The forward propagation, loss functions, and backward propagation of the image network and the point cloud network are indicated by yellow and green arrows, respectively. Point clouds are generated from raw stereo- or multi-view images. In the procedure of optional enhanced co-learning (framed by gray boxes), unlabeled data do not participate in the optimization of the supervised semantic segmentation loss function. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

2021; Thomas et al., 2019), and sparse CNN (Graham et al., 2018; Bachhofner et al., 2020).

Unfortunately, 3D data also have limitations in applications to building extraction. Point clouds are discrete, which leads to the problems of missing building structure, as well as boundaries that are not sufficiently sharp. Since 2D images and 3D geometric data can provide information complementary to each other, which could benefit the accuracy of building extraction, methods of multimodality learning have attracted the attention of researchers (Bittner et al., 2018; Sun et al., 2021b). Most available multimodality learning works in the remote sensing field concentrate on data fusion, including early fusion (input fusion or observation-level fusion), middle fusion (feature fusion or feature-level fusion), and late fusion (probability fusion or decision-level fusion) (Schmitt and Zhu, 2016).

As shown in Fig. 1(a), early fusion is usually carried out at the data input stage. In one popular case in the remote sensing field,

multispectral images are fused with DSMs for semantic segmentation (Paisitkriangkrai et al., 2015). In this approach, spectral channels of optical images and geometric information such as the height values of DSMs are concatenated as combined input features to a single-modality network. In Fig. 1(b), middle fusion is operated at the stage of feature embedding, concatenating deep features learned by different network streams to a composite stream (Zhou et al., 2021). Following operations are based on the concatenated feature vectors. Late fusion is employed at the decision stage, which operates on the probability maps output from multiple algorithms, as shown in Fig. 1(c).

Data fusion takes the benefit of multiple information sources and improves the performance of semantic segmentation algorithms, including building extraction algorithms. But these techniques have strict requirements for both data amount and data quality, and assume that all modalities are present, aligned, and noiseless during the training and the test phase (Rahate et al., 2022). However, 2D images and

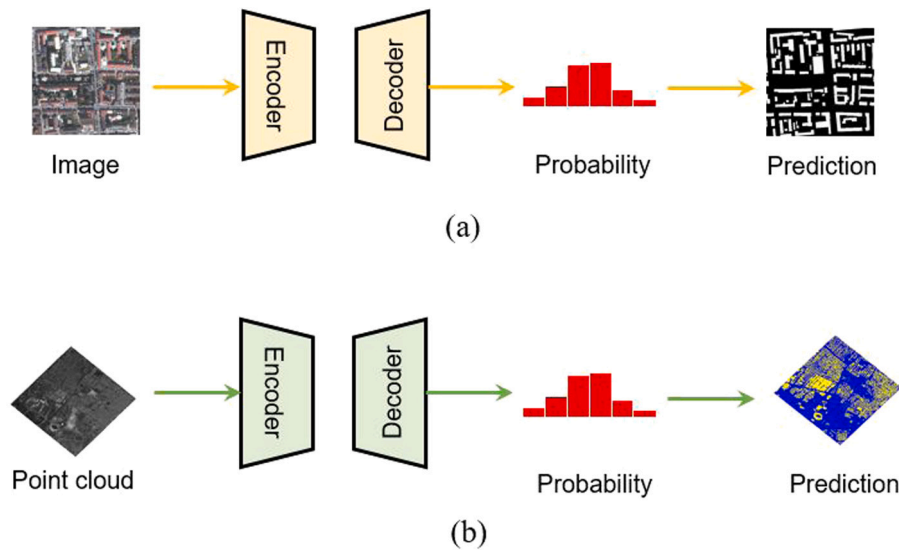


Fig. 3. In the test phase, networks are used individually as normal single-modality networks.

3D data are not always simultaneously available in diverse data sets. In addition, LiDAR-based data are expensive and time consuming to acquire, so are not suitable for projects involving large-scale areas. Imagery-derived 3D data require a certain amount of overlapped high-resolution optical images for the dense image matching algorithm. This is also a challenge for applications involving historic orthophotos in which raw stereo/multi-view images are missing and matched 3D data cannot be obtained. Well-performing single-modality networks are essential for practical applications. On the other hand, the architectures of networks that process fused data are usually complex and bloated, resulting in low efficiency and requiring high computational ability. By contrast, methods with simple and efficient architectures that consume few annotated data would be welcome in practical applications that demand real-time data processing capability.

Recently, co-learning methods are proposed in the generic artificial intelligence field, aiming to aid the modeling of one modality by exploiting knowledge from another and offering a tradeoff between the advantage of multimodalities and strict input data requirements. Co-learning explores how knowledge learning from one modality can help a deep learning model trained on other different modalities, especially when one modality has limited resources, such as missing modality, noisy modality, and lacking annotated data (Rahate et al., 2022; Zheng et al., 2021). As reviewed in (Baltrušaitis et al., 2019) and (Rahate et al., 2022), co-learning-based methods have been employed in several cross-modality applications (e.g., audio-visual Zadeh et al., 2020, visual-text Ma et al., 2021). Fig. 1(d) presents a type of co-learning architectures based on loss functions, which is applicable to multimodality semantic segmentation. The step of knowledge transfer bridging multimodality networks in this architecture is realized by the co-learning loss function rather than direct addition or concatenation. Each single-modality network is trained individually, where corresponding parameters can be better optimized with the help of another modality via co-learning loss. Unlike traditional data fusion approaches, a semantic segmentation model trained with a multimodality data set through this way can be also performed on single-modality test data, thus effectively solving the problem of insufficient availability of multimodality test data.

In computer vision, cross-modality unsupervised domain adaptation (xMUDA) is the first work to adaptively transfer information among multimodality data sets to improve the segmentation results of mobile LiDAR point clouds (Jaritz et al., 2020). As its name suggests, xMUDA aims to address the problem of domain adaptation for point cloud semantic segmentation. In our article, we combine the theoretical

background of generic co-learning and xMUDA, and propose an elegant framework applicable to automatic building extraction from spectral images and corresponding photogrammetric point clouds. Fig. 2 shows the architecture of our proposed co-learning model. The architecture of the training model contains a 2D network to process images and a 3D network to work on point clouds. As shown in Fig. 3(a) and (b), these two networks can be used individually in the test phase. As another difference from xMUDA, there is no self-training step involved in our method, so it would be more friendly to software development. In addition, our architecture can utilize unlabeled training data, thereby reducing the dependence on the amount of annotated data. Hence, it is especially suitable for the case with fewer annotated training data. The main contributions of our work are as follows:

- We present a co-learning framework to handle the case in which one modality is missing during the testing time. In particular we exemplify the framework with photogrammetric point clouds and corresponding optical images, because in practice these two modalities are one of the most widely-used pairs.
- We apply the proposed co-learning framework in few-shot tasks to solve the problem of scarcity of labeled training data. We investigate the effects of unlabeled data in our framework.
- We evaluate our co-learning framework on two data sets: the ISPRS Potsdam public airborne data set, and a data set of Munich collected by the WorldView-2 satellite. Experimental results demonstrate the effectiveness of our method for the task of automatic building extraction.

2. Methodology

2.1. Overview

The detailed flowchart of proposed co-learning based network architecture is shown in Fig. 2. As it shows, the co-learning method we applied transfers knowledge from one modality to another based on the probability maps. The intuition behind this approach is that the better results the networks achieve, the smaller the prediction gap between two modalities. To meet this requirement, a co-learning loss function is proposed to learn the similarity between the predictions of the 2D and 3D networks. In the training phase, the target is to minimize two loss functions, a supervised loss function for semantic segmentation purpose, and the unsupervised co-learning loss function to measure the distance between two predictions. In the implementation, each network

outputs two types of probability maps. One, *predicted probability*, is used in the loss functions of the same network, and influences the backward propagation. In order to distinguish from real reference (ground truth), the other is named *shadow reference probability*, and is actually utilized by the other modality network as the reference in the co-learning loss function. The training data involved in two loss functions could be asymmetric, which means only part of the data needs to be annotated. Unlabeled data pairs are also beneficial to the minimization of co-learning loss.

2.2. 2D and 3D feature learning

As building extraction can be regarded as a branch of semantic segmentation, convolutional encoder–decoder neural networks are mainstream architectures applied for feature learning from raw images and/or point clouds. In our work, we employ a 2D U-Net (Ronneberger et al., 2015) with residual blocks of ResNet34 (He et al., 2016) as the backbone to learn 2D features from multispectral images. A U-Net-like sparse convolutional neural network (Graham et al., 2018; Choy et al., 2019) is employed as the backbone to learn 3D features from point clouds.

CNNs are a category of deep learning models that have been successfully utilized in image and point cloud processing, and consist of multiple convolutional layers. In each layer, the input feature maps are convolved by a kernel with learned weights. In image cases, the convolutional kernel is usually naturally dense, and is defined as (Choy et al., 2019)

$$\mathbf{x}_{\mathbf{u}}^{\text{out}} = \sum_{\mathbf{i} \in \mathcal{V}^D} W_{\mathbf{i}} \mathbf{x}_{\mathbf{u}+\mathbf{i}}^{\text{in}}, \quad (1)$$

where $\mathbf{x}_{\mathbf{u}}^{\text{in}} \in \mathbb{R}^{N^{\text{in}}}$ is the input feature vector of coordinate $\mathbf{u} \in \mathbb{Z}^D$ in a D -dimensional space. \mathcal{V}^D is the list of offset elements in the hypercube centered at the origin, which is covered by the convolution kernel. $W_{\mathbf{i}}$ is the kernel weight corresponding to the offset element $\mathbf{u} + \mathbf{i}$. For 2D images, $D = 2$.

In the real world, most 3D spaces are not occupied by any objects. As a result, corresponding point clouds and converted voxels contain large empty areas (Xu et al., 2021). If we adopt conventional dense convolutions to process such sparse data, the calculation would be time-consuming and memory-intensive. Sparse convolution presented by Bachhofner et al. (2020) and Choy et al. (2019) is a solution to this problem. Arbitrary kernel shapes instead of conventional dense shapes are utilized in sparse convolutions, which only take those non-empty grids into the convolving calculation. By defining the existing offset grids covered by the convolution as $\mathcal{N}^D(\mathbf{u}, C^{\text{in}})$, the output feature vector $\mathbf{x}_{\mathbf{u}}^{\text{out}}$ is presented as (Choy et al., 2019)

$$\mathbf{x}_{\mathbf{u}}^{\text{out}} = \sum_{\mathbf{i} \in \mathcal{N}^D(\mathbf{u}, C^{\text{in}})} W_{\mathbf{i}} \mathbf{x}_{\mathbf{u}+\mathbf{i}}^{\text{in}}, \quad (2)$$

where $\mathcal{N}^D(\mathbf{u}, C^{\text{in}}) = \{\mathbf{i} | \mathbf{u} + \mathbf{i} \in C^{\text{in}}, \mathbf{i} \in \mathcal{N}^D\}$. C^{in} is the predefined sparse tensors to be convolved. In the case of a point cloud, $D = 3$.

2.3. Co-learning

The co-learning method in our work is a flexible framework that makes use of different categories of training data. As mentioned above, both labeled and unlabeled training data can be employed in this framework. The labeled data and unlabeled data can be asymmetric. According to the availability of ground truth and multimodality pairs, the training data in our co-learning framework can be classified into three categories: *labeled pairs*, *unlabeled pairs*, and *labeled singles*, separately named in our work.

- *labeled pairs* refer to the data with the ground truth that are co-registered with another modality; these pairs are involved in both supervised loss function and the co-learning loss function.

- *unlabeled pairs* refer to the samples that are without ground truth but have co-registered multimodalities, which means they can benefit the co-learning loss function.
- *labeled singles* are the single-modality training data with ground truth, that are involved only in the supervised loss function not the co-learning loss function.

In our work, we mainly explore the influence of labeled pairs and unlabeled pairs. The effect of labeled singles is obvious, as they have been widely investigated in works on conventional single-modality learning, which can be regarded as architectures only with labeled singles. We name the setting with only labeled pairs as *standard co-learning*, and the situation trained partly with additional unlabeled pairs as *enhanced co-learning*. Fig. 2 contains the training procedures of both standard and enhanced cases.

2.3.1. Standard co-learning

The intuition behind our co-learning method is that unsupervised mutual information from the other modality would be a positive factor to the target networks. Apart from the difference between the prediction and the ground truth (i.e., supervised segmentation loss function), the similarity between multimodality data could also be potentially valid information benefiting the training phase and helping find more proper deep model parameters. This is realized by a co-learning loss function. As shown in Fig. 2, standard co-learning adopts the labeled training samples in the learning procedure. For each backpropagation step, the gradients of the combination of the supervised segmentation loss function and co-learning function are computed. Algorithm 1 shows how the standard co-learning is implemented. For each iteration, first the predicted probability of images p_{2D} and the predicted probability of point clouds p_{3D} are calculated by the forward propagations of two networks, respectively. Then supervised segmentation loss functions and co-learning functions are computed. In the calculation of co-learning loss for images, p_{3D} is used as the shadow reference probability. In the computation of co-learning loss for point clouds, p_{2D} is employed as the shadow reference probability. Finally, backpropagation operations are carried out and the parameters of the image network W_{2D} as well as the parameters of the point cloud network W_{3D} are updated.

Algorithm 1 Standard co-learning

Input: $(D_{2D}, L_{2D}), (D_{3D}, L_{3D})$

Output: W_{2D}, W_{3D}

- 1: **Initialize** W_{2D}, W_{3D}
 - 2: **while** $i < I$ **do** ▷ I is the number of iterations
 - 3: Randomly sample labeled training pairs d_{2D} and d_{3D} from D_{2D} and D_{3D}
 - 4: $p_{2D} \leftarrow \text{net}_{2D}(d_{2D})$ ▷ forward pass of the image network
 - 5: $p_{3D} \leftarrow \text{net}_{3D}(d_{3D})$ ▷ forward pass of the point cloud network
 - 6: **Calculate** $\mathcal{L}_S^{2D}(l_{2D} || p_{2D})$ ▷ image segmentation loss
 - 7: **Calculate** $\mathcal{L}_{CL}^{2D}(p_{3D} || p_{2D})$ ▷ image co-learning loss
 - 8: **Calculate** $\mathcal{L}_S^{3D}(l_{3D} || p_{3D})$ ▷ point cloud segmentation loss
 - 9: **Calculate** $\mathcal{L}_{CL}^{3D}(p_{2D} || p_{3D})$ ▷ point cloud co-learning loss
 - 10: 2D backward pass
 - 11: **Update** W_{2D}
 - 12: 3D backward pass
 - 13: **Update** W_{3D}
 - 14: **end while**
 - 15: **Return** W_{2D}, W_{3D}
-

2.3.2. Enhanced co-learning

Annotating a large amount of training data is always a challenge in deep learning-based tasks, and is both expensive and time-consuming. Thus few-shot learning, which serves as a low-cost solution, is attracting more attention in deep learning related research (Sun et al., 2021a).

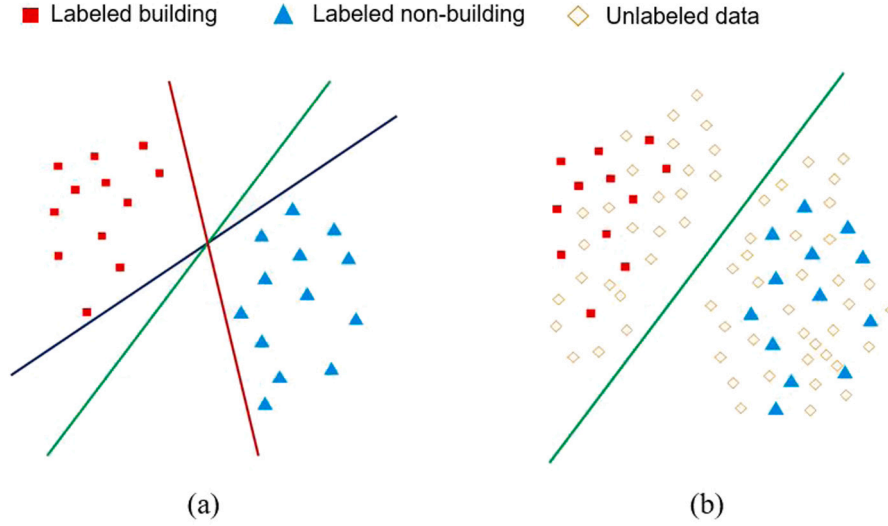


Fig. 4. (a) Learning with few data. (b) Enhanced co-learning. Lines with different colors represent different classifiers/models. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

A main drawback of conventional few-shot learning is the restricted beforehand knowledge. As shown in Fig. 4(a), we simulate this problem based on a building extraction task in a simple two-dimensional feature space. If there is no interference on the learning case with fewer training data, multiple models with different parameters can yield a reasonable classification. However, most of those models are prone to overfitting. They may have reasonable prediction results on the training samples, but they are likely to fail to predict unseen test data.

In reality, there is a huge amount of unlabeled data exist, but they are difficult to use directly in supervised learning. One advantage of the co-learning function is that it can employ unlabeled pairs. If unlabeled pairs are able to assist the clustering procedure, more accurate and less ambiguous models with better generalization ability could be obtained, as Fig. 4(b) shows. This is the intuition behind enhanced co-learning. Enhanced co-learning utilizes data in a more efficient way than conventional semi-supervised self-training that employs unlabeled training samples. Self-training is a procedure with several individual steps: training an initial model with a few labeled training samples, predicting on several unlabeled data, and re-training a model with unlabeled data and predicted pseudo labels (Zoph et al., 2020; Zhang et al., 2021). In order to obtain more accurate and stable models, sometimes users have to design extra algorithms to select proper samples with pseudo labels, and the training procedure has to be repeated several times (Zhang et al., 2021; Tong et al., 2020). In contrast, our enhanced co-learning is a one-step operation requiring no extra algorithm, which is much more user-friendly in practice.

For these reasons, our work incorporates an enhanced co-learning structure into our design by adopting both labeled and unlabeled training samples. Algorithm 2 demonstrates the implementation of the enhanced co-learning. For each iteration, enhanced co-learning carries out two forward propagations for each modality. One is with labeled training data. The other is with unlabeled training data.

2.4. Loss functions

Our method employs two categories of loss functions: the supervised loss function for the purpose of building extraction and the unsupervised loss function to realize co-learning. As mentioned above, we mainly consider two categories of training data: labeled pairs and unlabeled pairs. Hence, we describe our proposed loss functions accordingly.

Algorithm 2 Enhanced co-learning

Input: $(D_{2D}, L_{2D}), (D_{3D}, L_{3D}), U_{2D}, U_{3D}$

Output: W_{2D}, W_{3D}

```

1: Initialize  $W_{2D}, W_{3D}$ 
2: while  $i < I$  do
3:   Randomly sample labeled training pairs  $d_{2D}$  and  $d_{3D}$  from  $D_{2D}$ 
   and  $D_{3D}$ 
4:    $p_{2D}^{labeled} \leftarrow net_{2D}(d_{2D})$   $\triangleright$  forward pass of the image network
5:    $p_{3D}^{labeled} \leftarrow net_{3D}(d_{3D})$   $\triangleright$  forward pass of the point cloud network
6:   Calculate  $\mathcal{L}_S^{2D}(I_{2D} || p_{2D}^{labeled})$   $\triangleright$  segmentation loss for labeled
   images
7:   Calculate  $\mathcal{L}_{CL}^{2D-labeled}(p_{3D}^{labeled} || p_{2D}^{labeled})$   $\triangleright$  co-learning loss for
   labeled images
8:   Calculate  $\mathcal{L}_S^{3D}(I_{3D} || p_{3D}^{labeled})$   $\triangleright$  segmentation loss for labeled
   point clouds
9:   Calculate  $\mathcal{L}_{CL}^{3D-labeled}(p_{2D}^{labeled} || p_{3D}^{labeled})$   $\triangleright$  co-learning loss for
   labeled point clouds
10:  2D backward pass
11:  3D backward pass
12:  Randomly sample unlabeled training pairs  $u_{2D}$  and  $u_{3D}$  from  $U_{2D}$ 
   and  $U_{3D}$ 
13:   $p_{2D}^{unlabeled} \leftarrow net_{2D}(u_{2D})$   $\triangleright$  forward pass of the image network
14:   $p_{3D}^{unlabeled} \leftarrow net_{3D}(u_{3D})$   $\triangleright$  forward pass of the point cloud network
15:  Calculate  $\mathcal{L}_{CL}^{2D-unlabeled}(p_{3D}^{unlabeled} || p_{2D}^{unlabeled})$   $\triangleright$  co-learning loss for
   unlabeled images
16:  Calculate  $\mathcal{L}_{CL}^{3D-unlabeled}(p_{2D}^{unlabeled} || p_{3D}^{unlabeled})$   $\triangleright$  co-learning loss for
   unlabeled point clouds
17:  2D backward pass
18:  3D backward pass
19:  Update  $W_{2D}$ 
20:  Update  $W_{3D}$ 
21: end while
22: Return  $W_{2D}, W_{3D}$ 

```

Building extraction is a branch of supervised semantic segmentation. In our work, a cross-entropy loss function is used for this purpose:

$$\mathcal{L}_S(P || Q) = H(P || Q) \quad (3)$$

$$= - \sum_{x \in \mathcal{X}} P(x) \log(Q(x)), \quad (4)$$

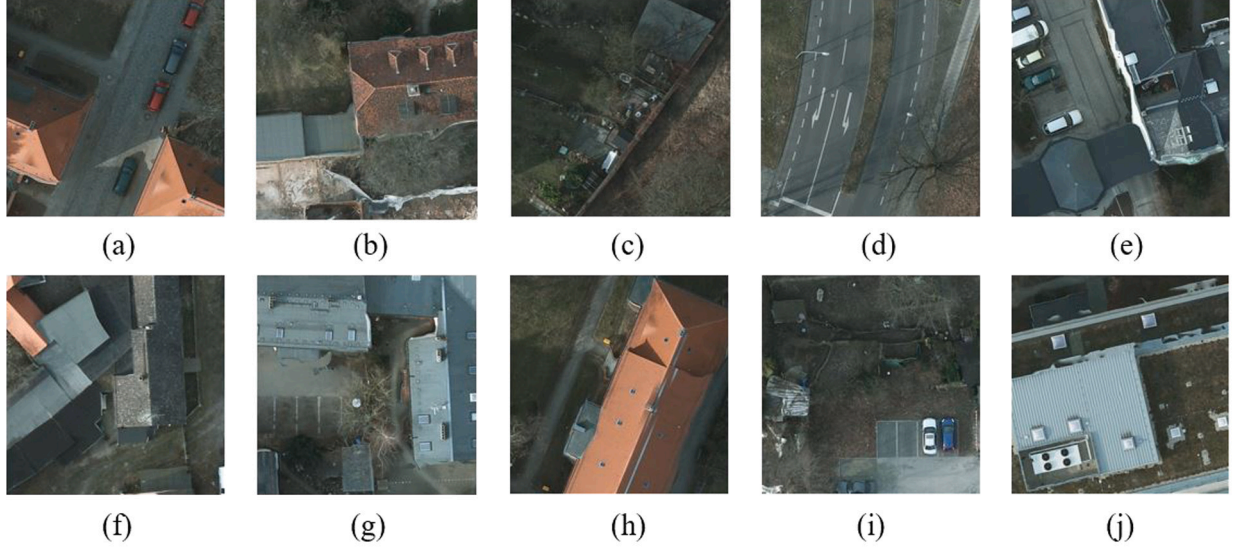


Fig. 5. 10-shot training samples of the ISPRS Potsdam data set.

where P and Q are defined on the same probability space \mathcal{X} . The term P denotes the distribution of the ground truth, while Q is the probability distribution of the predicted output.

The co-learning function is designed to transfer mutual information from one modality to another. When both networks are optimized, the difference in the building extraction results between the 2D and 3D modalities should be minimized. In other words, the probability distribution of one modality should be consistent with the distribution of the other. This step can be realized by a similarity loss function. Referring to Jaritz et al. (2020), we adopted KL divergence to realize this optimization.

$$\mathcal{L}_{CL}(P \parallel Q) = D_{KL}(P \parallel Q) \quad (5)$$

$$= \sum_{x \in \mathcal{X}} P(x) \log\left(\frac{P(x)}{Q(x)}\right), \quad (6)$$

where P and Q are defined on the same probability space \mathcal{X} . The item P denotes the probability distribution of the target data, while Q is the probability distribution of the predicted output. In our co-learning framework, P and Q are from two different modalities. P is the shadow reference probability, while Q is the predicted probability.

Combining a co-learning loss function \mathcal{L}_{CL} with semantic segmentation loss function \mathcal{L}_S , the total standard co-learning loss function \mathcal{L}_{total} for each single-modality network is derived. For a 2D image network, the total loss function is:

$$\mathcal{L}_{total} = \mathcal{L}_S + \lambda_1 \mathcal{L}_{CL}(P_{3D} \parallel P_{2D}), \quad (7)$$

where λ_1 is the hyperparameter to weight the co-learning loss function. Here the probability map of point clouds P_{3D} is set as the shadow reference, which is regarded as constant coefficients in the co-learning loss function for the image network.

For a 3D point cloud network, the total loss function is

$$\mathcal{L}_{total} = \mathcal{L}_S + \lambda_1 \mathcal{L}_{CL}(P_{2D} \parallel P_{3D}), \quad (8)$$

where λ_1 is the hyperparameter to weight co-learning loss function. Here the probability map of images P_{2D} is set as the shadow reference, which is regarded as constant coefficients in the co-learning loss function for the point cloud network.

For the case of enhanced co-learning, unlabeled pairs are also taken into consideration by the co-learning loss. The total image network loss function combining enhanced co-learning loss function $\mathcal{L}_{CL}^{unlabeled}$ is:

$$\mathcal{L}_{total} = \mathcal{L}_S + \lambda_1 \mathcal{L}_{CL}^{labeled}(P_{3D} \parallel P_{2D}) + \lambda_2 \mathcal{L}_{CL}^{unlabeled}(P_{3D} \parallel P_{2D}), \quad (9)$$

The total loss function combining enhanced co-learning employed in a 3D point cloud network is

$$\mathcal{L}_{total} = \mathcal{L}_S + \lambda_1 \mathcal{L}_{CL}^{labeled}(P_{2D} \parallel P_{3D}) + \lambda_2 \mathcal{L}_{CL}^{unlabeled}(P_{2D} \parallel P_{3D}), \quad (10)$$

3. Experiments

In this section, we introduce the data sets utilized for the evaluation of the proposed co-learning methodology, as well as our experimental setup. Two remotely sensed data sets are utilized for the evaluation.

3.1. Data description

ISPRS Potsdam is a public benchmark for 2D/3D semantic labeling (ISPRS, 2022). It is also widely used as a building detection benchmark (Li et al., 2021, 2022). This data set provides airborne orthoimages and corresponding DSMs generated via dense image matching. The ground sampling distance of images and DSMs is 5 cm. In our experiment, we convert these DSMs to 3D point clouds. Thus we can evaluate our methodology on a public benchmark, as there is no well-known public data set providing both annotated airborne images and well-matched original point clouds. Furthermore, we crop images from this data set into patches with a size of 512×512 pixels. The overlap between two up-and-down or left-and-right neighboring patches is 256 pixels. In our main experiments, a 10-shot learning case is investigated, which means only 10 randomly selected labeled patches of images and point clouds are used as the training samples. The training samples used in our 10-shot learning experiments are shown in Fig. 5.

Munich WorldView-2 is a collection of WorldView-2 satellite imagery captured over the city center of Munich, Germany. It contains two parts: orthoimages with only RGB channels, and unregistered colorless 3D point clouds. The 3D point clouds are generated from the stereo WorldView-2 panchromatic images using the improved semi-global matching approach (Tian et al., 2013; d'Angelo, 2016). Rasterized DSMs from point clouds are adopted to orthorectify the multispectral and panchromatic images. After pansharpening, we select the red (5th), green (3th), and blue (2nd) channels from multispectral images to generate the orthoimages. The ground sampling distance of the orthoimages is 0.5 m. As Fig. 6 shows, the test region marked as A4 has a size of 6000×6000 pixels. The images denoted as A1, A2, and A3, each with a size of 6000×6000 pixels, comprise the full training data. The images marked as A5 and A6 are used as the validation sets, each of which has a size of 6000×3200 , respectively. The building masks

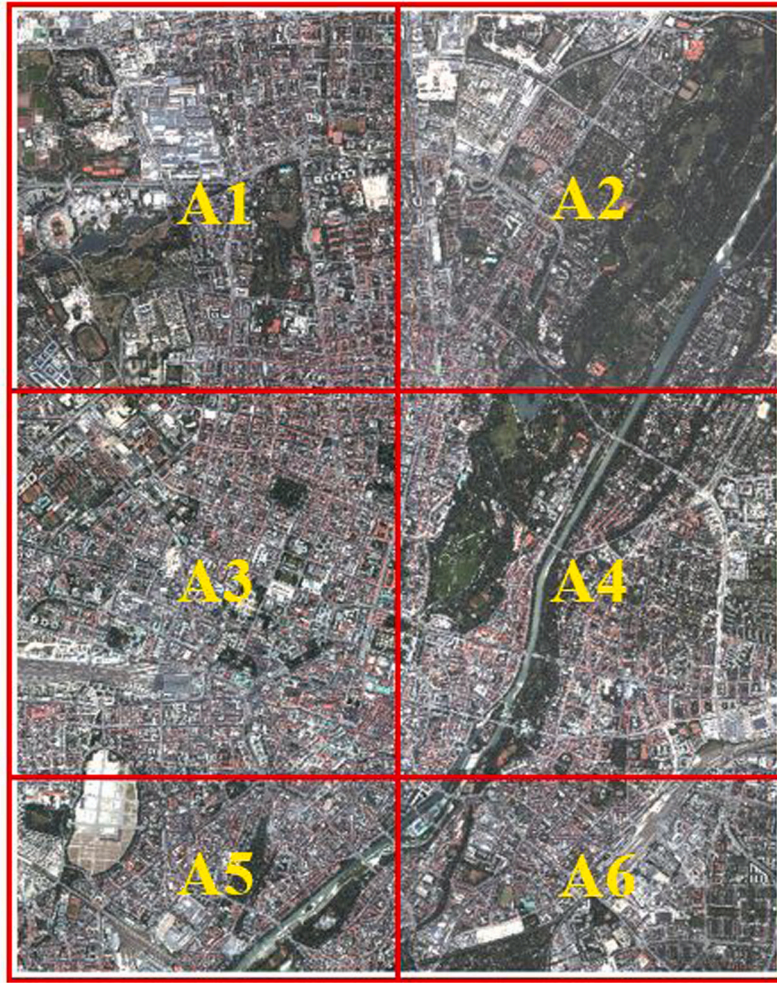


Fig. 6. The coverage of the Munich data set used in our experiment. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

(ground truths) of original images are manually annotated by using the open street map as a basis. The ground truths of point clouds are obtained through an affine transformation from the building masks. To satisfy the limitation of GPU memory, the full training data set has been cropped into patches with a size of 512×512 pixels and an overlap of 256 pixels in rows and columns. Fig. 7 shows those 10 training samples utilized in our 10-shot learning experiments on the Munich WorldView-2 data set.

3.2. Experiment setup

Our experiments are conducted within the PyTorch deep learning framework. We adopt the SparseConvNet library presented by Graham et al. (2018) to implement the sparse convolutional neural network. Training and testing are performed on a Geforce RTX 2080 Ti GPU with 11 GB RAM. All models are trained with the Adam optimizer until convergence is achieved. The scaling factor controlling the input resolution of voxels is an important parameter for sparse convolutional neural networks. Referring to the resolution of the original images, we set the input voxel size of Munich WorldView-2 to 0.5 m, and the input size of ISPRS Potsdam to 0.05 m. The learning rate is set to 0.001. The batch size of the training models is set as 4. In our experiments, the input features to the image network are red, green, and blue channels. Because in real applications the expected point cloud test data sometimes have no spectral information, only coordinate values (X, Y, and

Z) are employed as input features to the point cloud neural network, ignoring potential color information provided by multispectral images.

We test both of the standard and enhanced co-learning approaches in our experiments. In order to explore the learning ability of the co-learning architecture, we do not carry out any pre-training or data augmentation operations. In the experiments of 10-shot labeled training pairs, baseline methods and standard co-learning only utilizes 10 labeled patches in the training phase. Enhanced co-learning employs 10 labeled patches as well as all remaining patches of original training data as unlabeled pairs. In short, for the ISPRS Potsdam 10-shot experiment, we used 10 labeled and 10,570 unlabeled training pairs. While for the Munich WorldView-2 experiment we employ 10 labeled and 1577 unlabeled training pairs.

Following Li et al. (2021), the F1-score and intersection over union (IoU) of the building class are selected as the evaluation metrics. In order to better evaluate the confusion between the background and buildings, overall accuracy (OA), false negative rate (FNR), and false positive rate (FPR) are reported in our work. These metrics are calculated as follows:

$$OA = \sum_{i=1}^n \left(\frac{TP_i + TN_i}{TP_i + TN_i + FP_i + FN_i} \right), \quad (11)$$

$$F1 = \frac{2TP}{2TP + FP + FN}, \quad (12)$$

$$IoU = \frac{TP}{TP + FP + FN}, \quad (13)$$

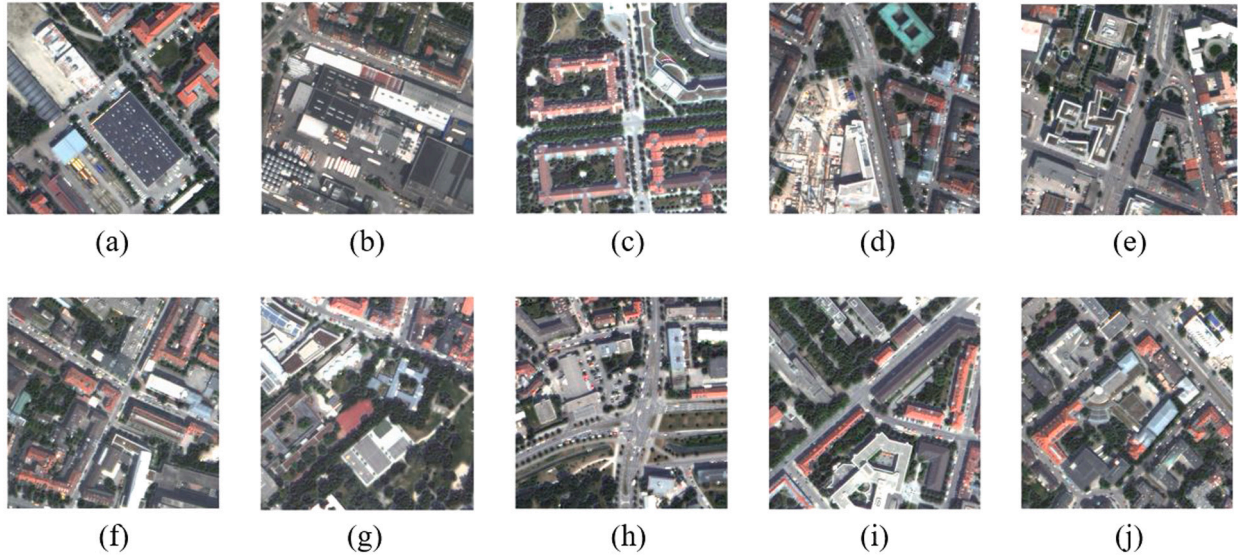


Fig. 7. 10-shot training samples of the Munich WorldView-2 data set.

Table 1

Performance of different methods for building extraction in the 10-shot ISPRS Potsdam data set.

| | Methods | OA | IoU | F1 | FNR | FPR |
|--------------|---|---------------|---------------|---------------|---------------|---------------|
| Image | Single-modality U-Net (baseline) | 0.8795 | 0.5633 | 0.7202 | 0.3502 | 0.0483 |
| | Early fusion U-Net (RGB + elevation) | 0.9004 | 0.6471 | 0.7857 | 0.2364 | 0.0566 |
| | Co-learning U-Net (standard) | 0.8850 | 0.6018 | 0.7514 | 0.2734 | 0.0652 |
| | Co-learning U-Net (enhanced) | 0.9370 | 0.7439 | 0.8532 | 0.2349 | 0.0089 |
| Point clouds | Single-modality SparseConvNet (baseline) | 0.9409 | 0.7773 | 0.8747 | 0.1379 | 0.0343 |
| | Early fusion SparseConvNet (colorized point clouds) | 0.9167 | 0.6958 | 0.8206 | 0.2034 | 0.0455 |
| | Co-learning SparseConvNet (standard) | 0.9450 | 0.7906 | 0.8831 | 0.1321 | 0.0307 |
| | Co-learning SparseConvNet (enhanced) | 0.9504 | 0.8059 | 0.8925 | 0.1390 | 0.0215 |

Table 2

Performance of probability enhanced image results in the 10-shot ISPRS Potsdam data set.

| Methods | OA | IoU | F1 | FNR | FPR |
|-------------------------------|---------------|---------------|---------------|---------------|---------------|
| Enhanced co-learning | 0.9370 | 0.7439 | 0.85326 | 0.2349 | 0.0089 |
| Enhanced co-learning (fusion) | 0.9581 | 0.8291 | 0.9066 | 0.1509 | 0.0076 |

$$FNR = \frac{FN}{TP + FN}, \quad (14)$$

$$FPR = \frac{FP}{TN + FP}, \quad (15)$$

where i is the class index and n is the total number of classes; in our case $n = 2$. TP refers to the number of true positives, FP the false positives, TN the true negatives, and FN the false negatives.

4. Results and discussion

In this section, the results of experiments for single-modality learning (as the baseline), and proposed co-learning methods are presented on the two data sets. In the experiments using the ISPRS Potsdam data set, the point cloud network is superior to the image network. In the Munich WorldView-2 experiment, the image network has a better performance. Therefore, we also explore a late fusion operation by averaging probabilities to improve the initial result of the weaker modality. Furthermore, we investigate how co-learning works on the full data set.

4.1. Comparison on the 10-shot ISPRS potsdam data set

We perform four approaches on the 10-shot ISPRS Potsdam data set and compare their results. The first is the baseline approach trained with the single-modality network. The second is with the standard co-learning. The third is with the enhanced co-learning strategy utilizing 10 labeled training pairs and all unlabeled pairs. These three approaches are conducted separately on the 2D images and 3D point

clouds. The fourth approach, probability fusion, is performed only on the image modality, which has a inferior performance compared to 3D point cloud modality.

4.1.1. Quantitative evaluation

The performance metrics are shown in Table 1. The best results are achieved with enhanced co-learning. Compared to the results obtained by single-modality learning, the best results of images achieved by enhanced co-learning gain increments of 5.75%, 18.06%, and 13.30% in OA, IoU, and F1, respectively. Enhanced co-learning also demonstrates an improvement over the results achieved by standard co-learning. In addition, the best FNR and FPR scores are also obtained by enhanced co-learning. When testing on point clouds, the differences among the three models are rather limited. Compared to the baseline result, the best performance achieved by enhanced co-learning strategy has an improvement of 0.95%, 2.86%, and 1.78% in OA, IoU, and F1, respectively. Enhanced co-learning also achieves the best FPR among all the methods and an FNR score very close to the best.

It should be noted that our experiments on the ISPRS Potsdam data set proved that 3D point clouds outperform images. To further explore whether the results of the weaker data type could be improved by probability fusion, we average the 2D building probability and 3D building

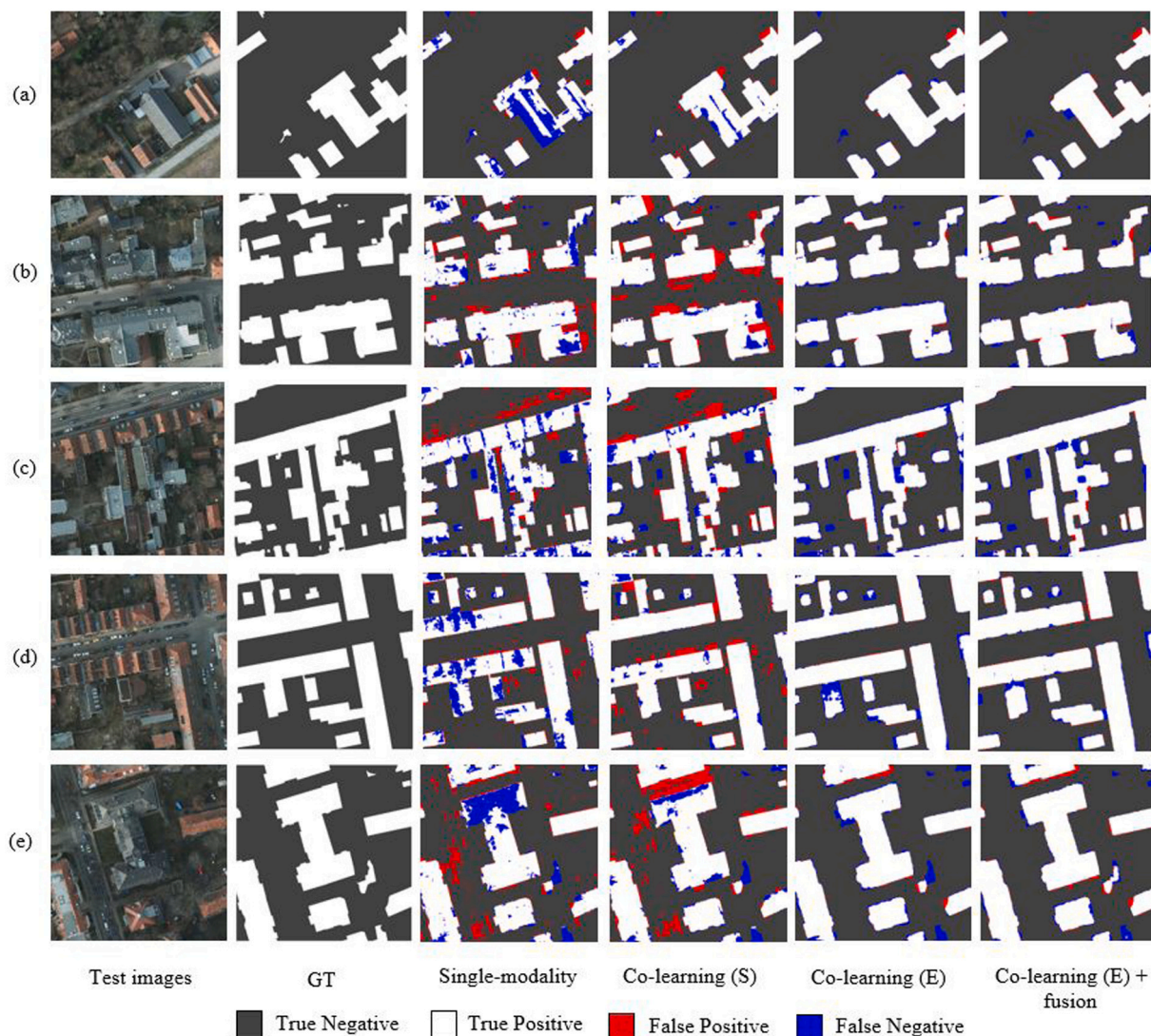


Fig. 8. 2D building extraction results obtained from the ISPRS Potsdam data set using 10 labeled training samples and various training strategies. GT: Ground truth. Co-learning (S): standard co-learning. Co-learning (E): enhanced co-learning. Co-learning (E) + fusion: enhanced co-learning and probability fusion. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

probability maps as a new probability map for the images. An image network and point cloud network trained by enhanced co-learning are used. Table 2 compares the best 2D building extraction result achieved by the image network (enhanced co-learning) and the result obtained from the fused probability map with image and DSM-derived point clouds. The probability fusion operation has a further enhancement on building extraction result, which gains an improvement of 2.11% on OA, 8.52% on IoU, and 5.34% on F1, as well as a decrease of 8.4% on FNR and 0.13% on FPR, compared with the results without fusion.

4.1.2. Qualitative evaluation

Single-modality learning is more sensitive to the quantity and quality of the training samples: thus the performance of deep learning models is restricted by the limited amount of training samples. In all five examples presented in Fig. 8, almost every building has defects, due to the poor features learned from only 10 annotated samples. Standard co-learning shows some improvement on buildings. However, many background pixels are wrongly classified as buildings. In contrast, the enhanced co-learning strategy with a large quantity of unlabeled training data achieves excellent results. In those examples, only building boundaries, small buildings, and auxiliary structures have

apparent flaws. The probability fusion approach with enhanced co-learning is superior to all three of the abovementioned cases, especially at recognizing small-sized buildings, as presented in (d) and (e), which are ignored by the enhanced co-learning without fusion operation.

For building extraction from DSM-derived point clouds, a main drawback shared by all three methods is that some points of high objects are easily misclassified as buildings, since there is no spectral textural information as a constraint. Fortunately, with the mutual knowledge transferred from the image neural network, such errors are eliminated. Fig. 9 is one typical example. As shown in the circled area, both results by two types of co-learning strategies have fewer false positive points than what single-modality learning achieves. Enhanced co-learning performs the best among the training strategies.

4.2. Comparison on 10-shot Munich WorldView-2 data set

The proposed approach was also applied and evaluated on Munich WorldView-2 data set with the same experimental setting.

4.2.1. Quantitative evaluation

Table 3 shows the performance of co-learning strategies in 10-shot settings, as well as the performance of the baseline. As with the first

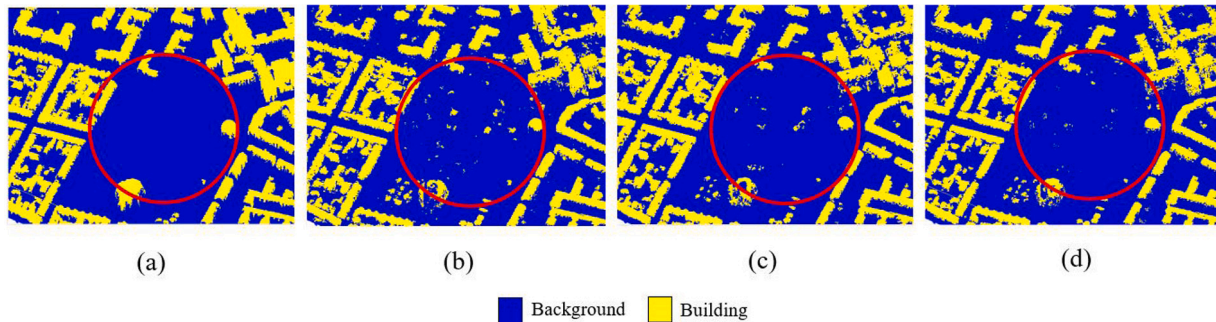


Fig. 9. Point cloud segmentation results obtained from the ISPRS Potsdam data set using 10 labeled training samples and various training strategies. (a) Ground truth. (b) Single-modality. (c) Standard co-learning. (d) Enhanced co-learning. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

Table 3
Performance of different methods for building extraction in the 10-shot Munich WorldView-2 data set.

| | Methods | OA | IoU | F1 | FNR | FPR |
|--------------|---|---------------|---------------|---------------|---------------|---------------|
| Image | Single-modality U-Net (baseline) | 0.8903 | 0.5979 | 0.7484 | 0.1940 | 0.0883 |
| | Co-learning U-Net (standard) | 0.9245 | 0.6847 | 0.8129 | 0.1899 | 0.0465 |
| | Co-learning U-Net (enhanced) | 0.9224 | 0.6682 | 0.8011 | 0.2282 | 0.0393 |
| Point clouds | Single-modality SparseConvNet (baseline) | 0.8465 | 0.4753 | 0.6443 | 0.3958 | 0.0811 |
| | Early fusion SparseConvNet (colorized point clouds) | 0.7938 | 0.4756 | 0.6446 | 0.1874 | 0.2118 |
| | Co-learning SparseConvNet (standard) | 0.8492 | 0.5024 | 0.6688 | 0.3388 | 0.0946 |
| | Co-learning SparseConvNet (enhanced) | 0.8790 | 0.5746 | 0.7298 | 0.2902 | 0.0703 |

Table 4
Performance of probability enhanced point cloud results in the 10-shot Munich data set.

| Methods | OA | IoU | F1 | FNR | FPR |
|-------------------------------|---------------|---------------|---------------|---------------|---------------|
| Enhanced co-learning | 0.8790 | 0.5746 | 0.7298 | 0.2902 | 0.0703 |
| Enhanced co-learning (fusion) | 0.9371 | 0.7456 | 0.8543 | 0.1984 | 0.0224 |

experiment, the trained models are separately tested on images and 3D point clouds. According to the comparison results, both standard and enhanced co-learning strategies can largely improve building extraction results. For the image-based results, in comparison to the baseline method, standard co-learning achieves a 3.42% higher OA, an 8.68% higher IoU, and a 6.45% higher F1, while FNR and FPR are reduced by 0.41% and 4.18%, respectively. However, the enhanced co-learning model trained by involving unlabeled training pairs as well as labeled pairs is slightly inferior to the standard version in overall performance.

For point clouds, the improvement achieved by standard co-learning includes 0.27% in OA, 2.71% in IoU, and 2.45% in F1 score, respectively. The best performance is achieved by the enhanced co-learning strategy, where IoU and F1 are increased by 9.93% and 8.55%, and FNR and FPR are decreased by 10.56% and 1.08% in comparison with the results by the single-modality method.

Unlike the ISPRS Potsdam data set, image results are better than point cloud results in the Munich WorldView-2 data set. At this point in the experiment, we fused the probability map of point clouds and corresponding image pixels to improve the building extraction results of the point clouds. In the probability fusion experiment of the Munich WorldView-2 data set, an image network and a point cloud network trained by enhanced co-learning are utilized. As reported in Table 4, the probability fusion operation improves the point cloud results by 5.81%, 17.1%, 12.45%, 9.18%, and 4.79% on OA, IoU, F1, FNR, and FPR, respectively.

4.2.2. Qualitative evaluation

As shown in Fig. 10, many non-building areas are distinguished as buildings by the single-modality baseline method. Some of those errors are continuous areas, while others are presented as dispersed spots, so

the corresponding predicted building mask looks quite noisy. For example, inside the red oval marked area, low vegetation and partial water with a light color and regular boundary can easily be distinguished as buildings by the baseline method. The explanation is that using only 10 labeled images cannot provide sufficient spectral and textural information to the deep learning models. With the help of the co-learning strategy's transferred geometric knowledge from corresponding point clouds, such false positives can be largely eliminated.

Close-up views of several image segmentation examples are presented in Fig. 11. The prediction results of the single-modality network contain a significant amount of false positive pixels. Although enhanced co-learning does not achieve the best scores in evaluation metrics, it shows better performance on complex buildings. As can be observed in (a), (b), and (c), there are more missing building structures predicted by single-modality and standard co-learning methods. However, enhanced co-learning is prone to ignore small individual houses in our experiment. As shown in (d), a large number of small-sized buildings are classified as the background by the enhanced co-learning strategy. This phenomenon commonly happens in the full test image. That is why the standard co-learning strategy has a slightly better performance than the enhanced version in the quantitative evaluation.

For point clouds, three examples are presented in Fig. 12. The second, third, and fourth columns compare results obtained by the single-modality baseline method, standard co-learning, and enhanced co-learning, respectively. As shown in (a), (b), and red and green circled areas of (c), many building structures are ignored by single-modality learning. The standard co-learning-based network can recognize more building points correctly. Enhanced co-learning achieves better accuracy in identifying complete building structures than standard co-learning. However, it sometimes results in more false positive points, as highlighted in (a) by the red and (c) by the yellow. The fourth and fifth columns of Fig. 12 qualitatively analyze the probability fusion approach and the corresponding original enhanced co-learning method on point clouds. With the help of probability fusion, many of the abovementioned errors can be eliminated, such as those circled in (a) and (c). In addition, the probability fusion approach can benefit several inconspicuous buildings, such as those highlighted by the green oval in (b).

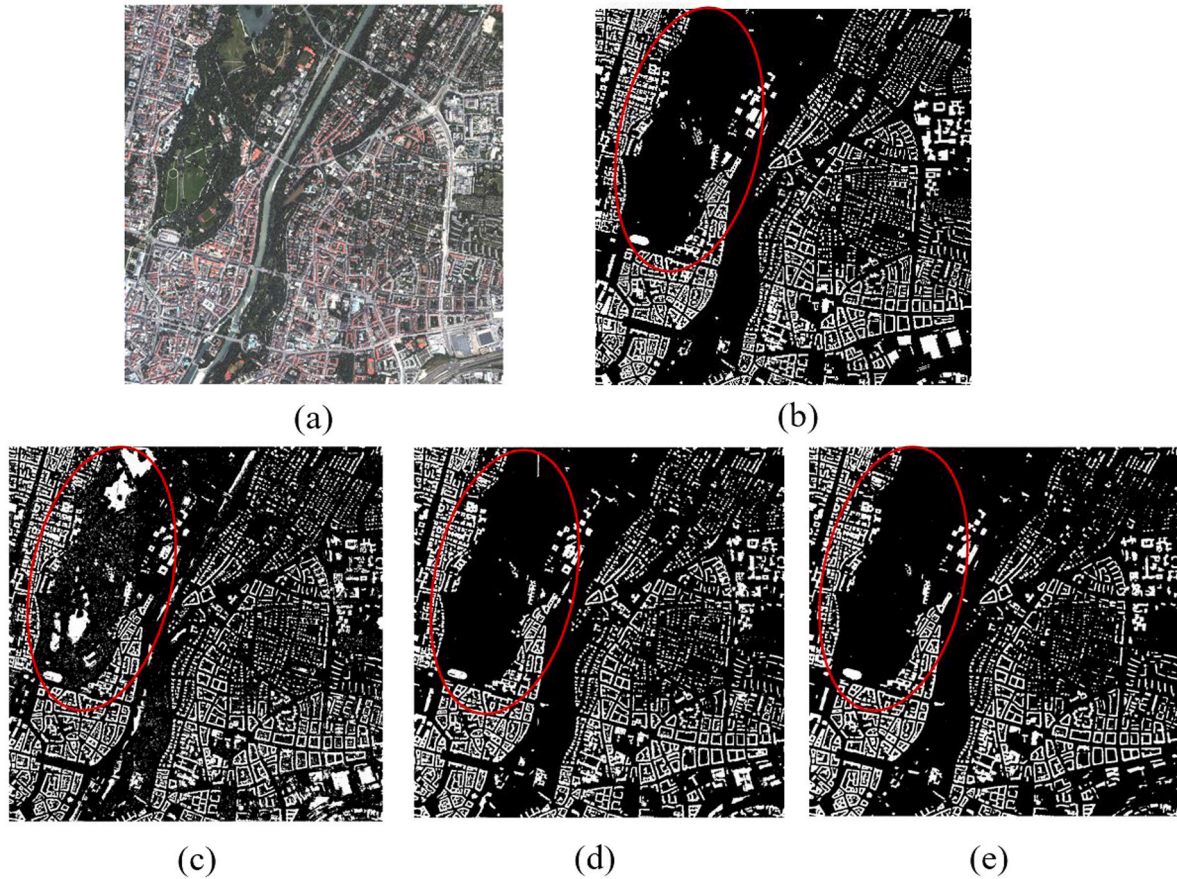


Fig. 10. The overview of image results obtained from the Munich WorldView-2 data set using 10 labeled training samples and various training strategies. (a) Original image. (b) Ground truth. (c) Single-modality. (d) Standard co-learning. (e) Enhanced co-learning. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

Table 5

Performance of single-modality learning and co-learning results in the ISPRS Potsdam data set with full labels. The results of EPUNet and ESFNet are from Li et al. (2022).

| Methods | OA | IoU | F1 | FNR | FPR |
|--|---------------|---------------|---------------|---------------|---------------|
| EPUNet (Guo et al., 2021) | – | 0.7941 | 0.8852 | – | – |
| ESFNet (Lin et al., 2019) | – | 0.8023 | 0.8865 | – | – |
| RegGAN (Li et al., 2022) | – | 0.8248 | 0.9040 | – | – |
| SegNet-8s-AFM (Li et al., 2021) | – | 0.8275 | 0.9056 | – | – |
| Single-modality U-Net | 0.9486 | 0.7928 | 0.8844 | 0.1770 | 0.0120 |
| Early fusion U-Net (RGB + elevation) | 0.9678 | 0.8686 | 0.9297 | 0.1092 | 0.0080 |
| Co-learning U-Net (standard) | 0.9623 | 0.8484 | 0.9180 | 0.1183 | 0.0123 |
| Co-learning U-Net (enhanced + test) | 0.9673 | 0.8676 | 0.9291 | 0.1048 | 0.0100 |
| Co-learning U-Net (enhanced + test + fusion) | 0.9759 | 0.9025 | 0.9488 | 0.0683 | 0.0102 |

4.3. Comparison on data sets with fully labeled training data

To further investigate the potentials of the co-learning framework and compare it with the state-of-the-art single-modality networks, we conduct the experiment for building extractions based on 2D images, using fully labeled training data from the ISPRS Potsdam and Munich WorldView-2 data sets.

4.3.1. 2D building extraction from fully labeled ISPRS potsdam data set

We follow the data splitting settings of Li et al. (2021, 2022). No pre-training operation or data augmentation is carried out. Table 5 describes our results and state-of-the-art results reported by Li et al. (2021, 2022). Compared with the result achieved by our single-modality learning, the OA of the standard co-learning method is 1.37% higher, and the IoU and F1 of the building class is increased by 5.56% and 1.63%, respectively. Our 2D U-Net trained with the standard co-learning strategy achieves higher scores than the state-of-the-art results by single-modality networks reported by Li et al. (2021, 2022).

In addition, we investigate the enhanced co-learning and probability fusion operation employing the test data with images and point clouds as the unlabeled pairs. Among them, the enhanced co-learning slightly outperforms the standard co-learning approach, while the probability fusion operation achieves the best scores of OA, IoU, and F1 among all co-learning strategies. The main problem in single-modality learning with fully annotated training data is that a few building structures are classified incorrectly as the background. It has the highest FNR among all the methods. Fig. 13 gives four examples. In (b) and (d), co-learning-based methods are capable of successfully recognizing more building structures. Fig. 13(a) is an area with several industrial buildings. Due to the lack of valid training samples, it is quite challenging for the single-modality 2D U-Net model to detect these buildings correctly. It should be noted that co-learning strategies have a better performance, especially on small-sized objects. Fig. 13(c) is an extreme example: the color of two buildings is close to the color of vegetation, so they are completely wrongly classified as “background” by the model trained

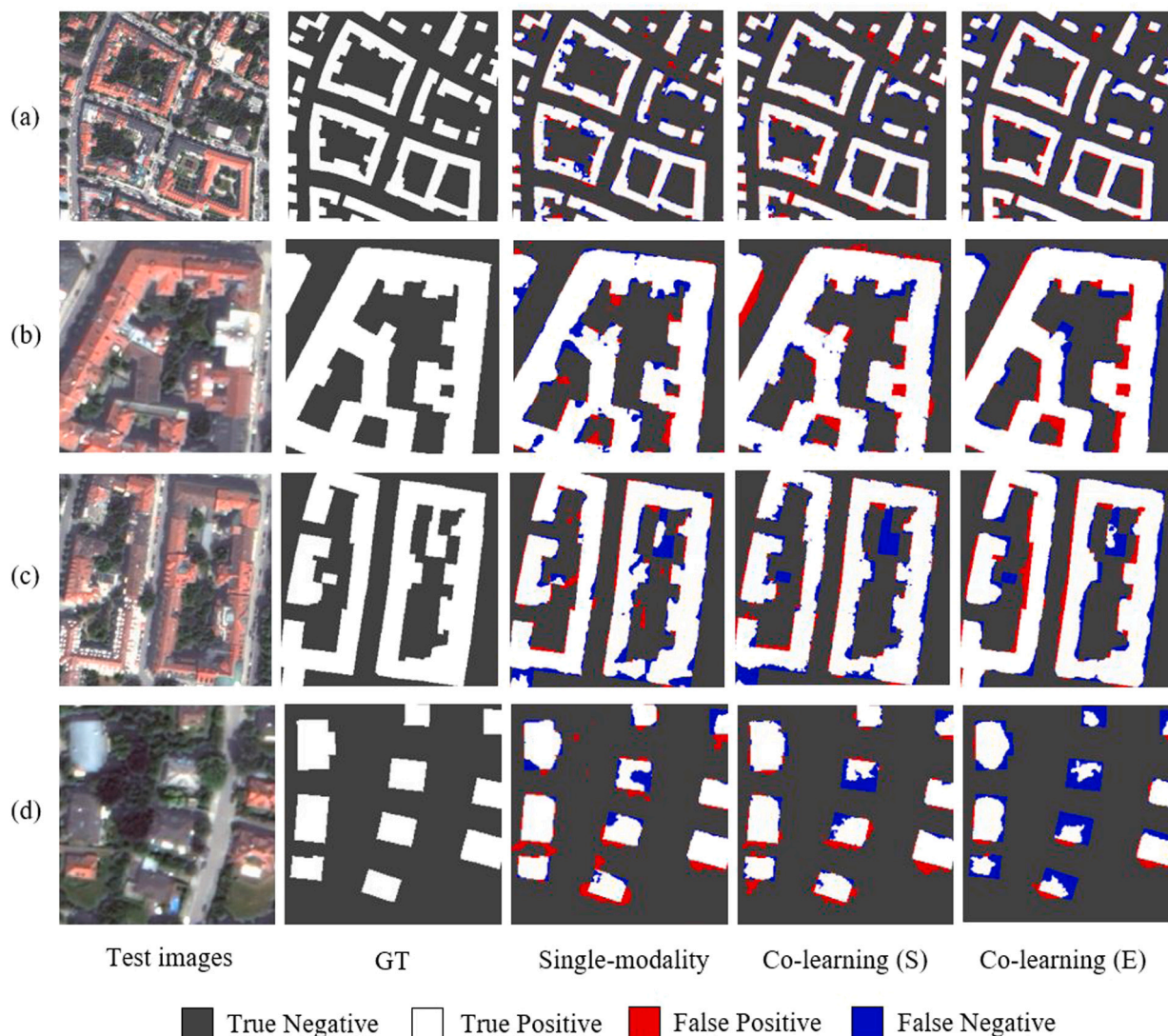


Fig. 11. Close-up views of image results obtained from the 10-shot Munich WorldView-2 data set using various training strategies. GT: Ground truth. Co-learning (S): standard co-learning. Co-learning (E): enhanced co-learning. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

by single-modality learning. With the transferred geometric knowledge from standard or enhanced co-learning, the results are slightly improved. Benefiting from the point cloud network, the probability fusion operation successfully eliminates most false negative pixels and outperforms other methods.

4.3.2. 2D building extraction from fully labeled Munich WorldView-2 data set

As shown in Table 6, the image network trained by co-learning is superior to the one trained by single-modality learning. OA, IoU, and F1 scores of the building class are increased by 1.08%, 5.64%, and 3.73%, respectively. In addition, co-learning method contributes an 8.45% lower FNR, which means it can correct many building pixels classified as non-buildings by the baseline U-Net. In Fig. 14, four visualization examples of predicted results are given. In (a) and (b), the co-learning-based network achieves greater completeness on buildings, especially at boundaries. Example (c) is an example of small-sized buildings, where the co-learning method is able to detect building structures that are more complete, although it also presents a few false positives. Example (d) is a rare case that includes round buildings and a multi-tiered square building, where standard co-learning approaches also have better performances and predict building structures that are more complete than single-modality learning.

Table 6 Performance of single-modality learning and co-learning results in the full 2D Munich data set.

| Methods | OA | IoU | F1 | FNR | FPR |
|------------------------------|---------------|---------------|---------------|---------------|---------------|
| Single-modality U-Net | 0.9370 | 0.7099 | 0.8304 | 0.2384 | 0.0185 |
| Co-learning U-Net (standard) | 0.9478 | 0.7663 | 0.8677 | 0.1539 | 0.0264 |

4.4. Discussion

Our experiments have clearly demonstrated the advantages of the proposed co-learning framework. At first, it reduces the dependence on the quantity of annotated data. Another advantage of the proposed co-learning framework is its flexibility. First, the training data and the test data can be asymmetric. Co-learning utilizes multimodality data to train the neural networks, while the test data can be single-modality. Second, both labeled and unlabeled training pairs can be fed to the neural network, and they can be asymmetric. There is no specific requirement for the ratio of labeled to unlabeled training samples. Third, the framework can also accept conventional single-modality labeled data. As this is a generally accepted strategy to improve the generalization ability of networks, it is not tested in our paper.

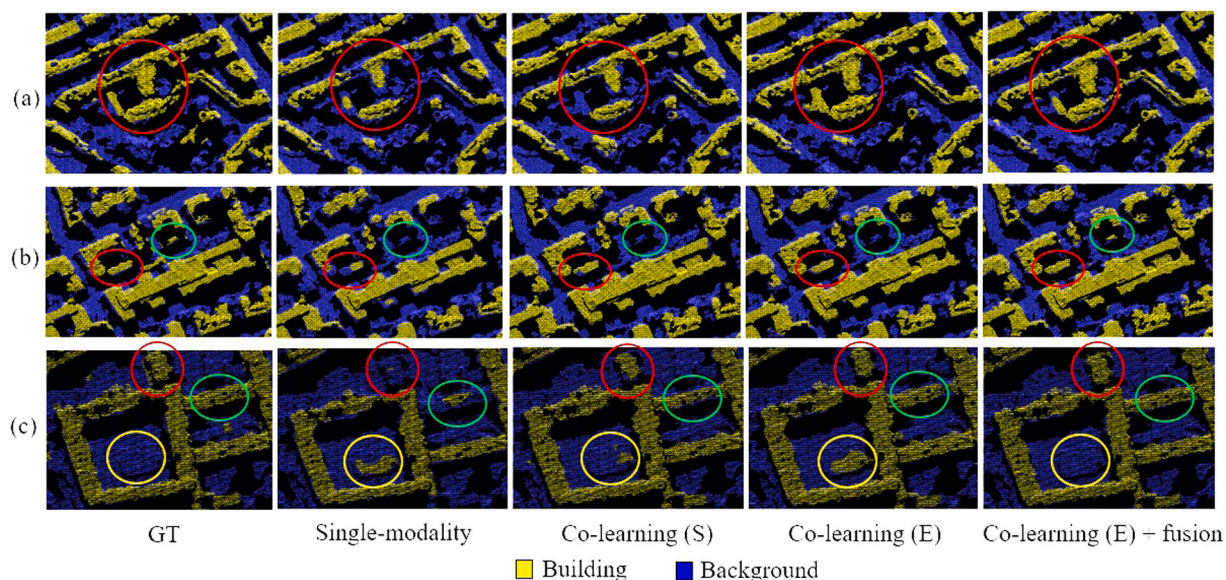


Fig. 12. Close-up views of point cloud results obtained from 10-shot Munich WorldView-2 data set using various training strategies. GT: Ground truth. Co-learning (S): standard co-learning. Co-learning (E): enhanced co-learning. Co-learning (E) + fusion: enhanced co-learning and probability fusion. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

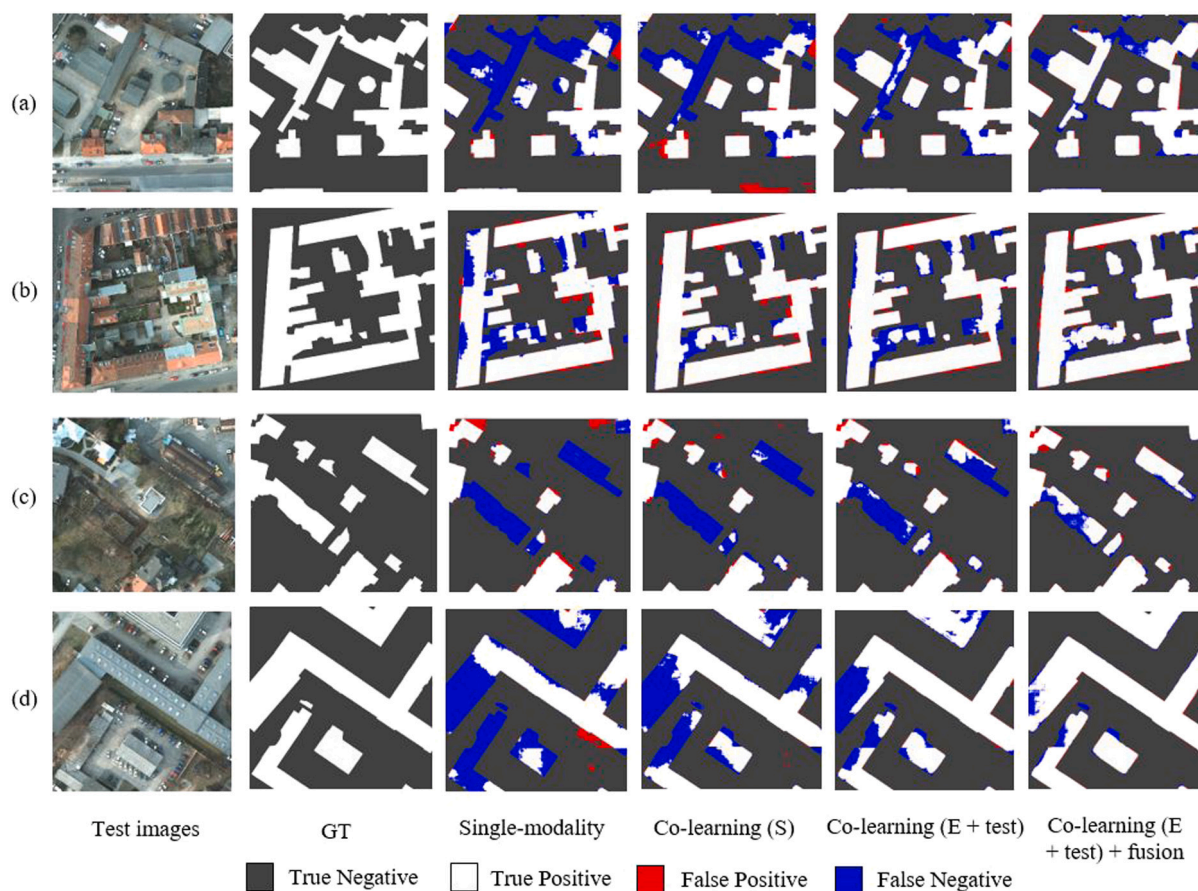


Fig. 13. 2D building extraction results obtained from the ISPRS Potsdam data set using fully labeled training data and various training strategies. Co-learning (S): standard co-learning. Co-learning (E + test): enhanced co-learning with test data. Co-learning (E+test) + fusion: enhanced co-learning with test data, and probability fusion. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

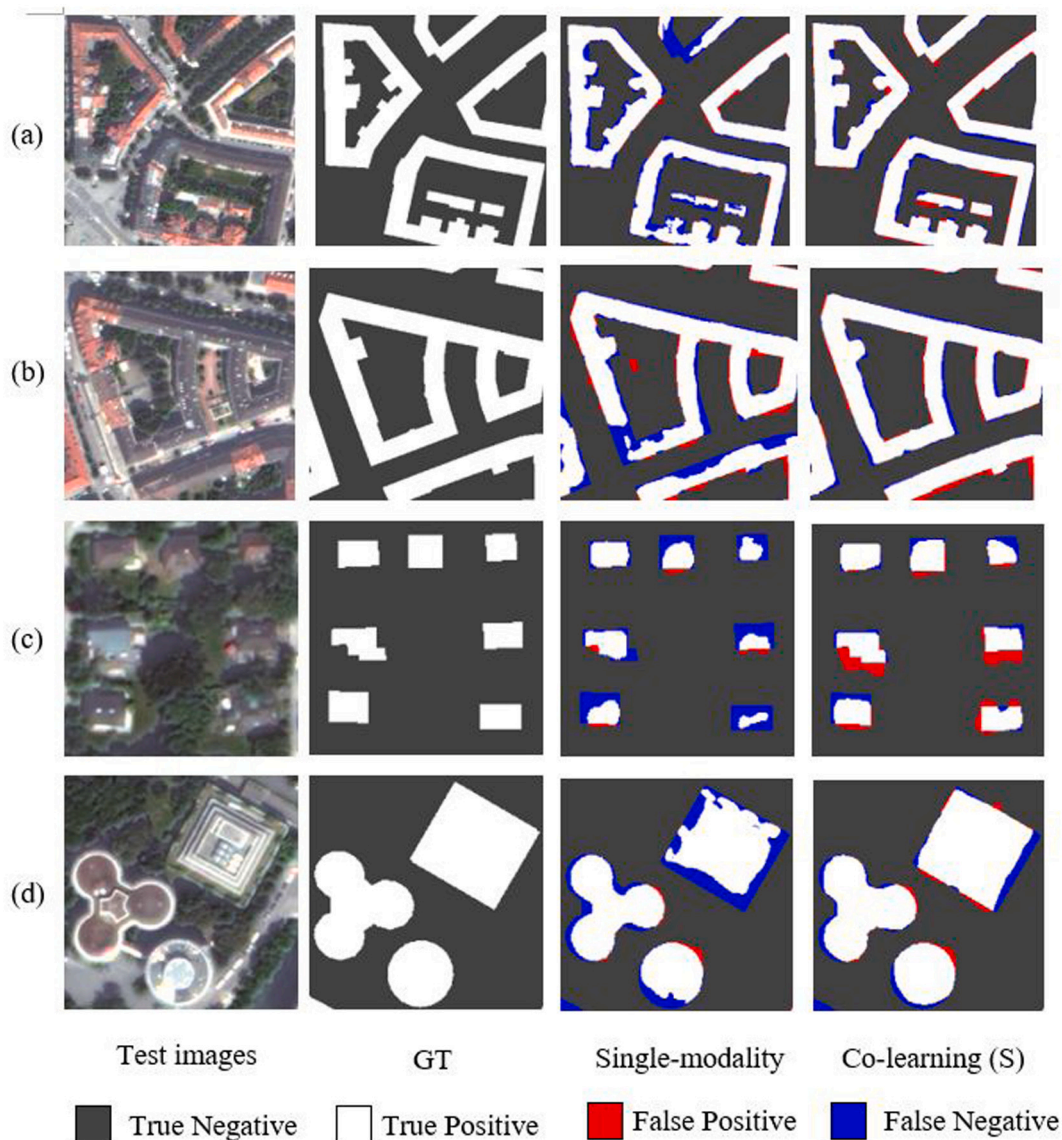


Fig. 14. 2D building extraction results obtained from the WorldView-2 Munich data set using fully labeled training data and various training strategies. GT: Ground truth. Co-learning (S): Standard co-learning. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

Depending on the available data sets, co-learning can be performed in various ways. Our experiments demonstrate that co-learning is well suited for few-shot tasks. In 4.1 and 4.2, we conducted four groups of experiments with 10 labeled 2D and 3D training samples. Both standard and enhanced co-learning methods achieve superior performance compared to single-modality learning. In three of them, enhanced co-learning is superior to standard co-learning. Only in one study case is the result of enhanced co-learning slightly worse than standard co-learning. These results demonstrate that mutual information by unsupervised learning can benefit building extraction to a large extent. According to the example in 4.3, standard co-learning is also able to improve the capacity of the image network trained with full training samples. Benefiting from transferred geometric knowledge from DSM-derived point clouds, even an essential U-Net has better performance than state-of-the-art networks on ISPRS Potsdam benchmark. As presented in Section 4.3.1, test data can be used as unlabeled training data by the enhanced co-learning framework, further improving the performance of image models.

Tables 1, 3, and 5 have also reported the quantitative results by conventional early fusion. In the experiments of images, early fusion is to concatenate the elevation values from DSMs with RGB channels of

corresponding images as a 4-channel input for the 2D U-Net. In the experiments of point clouds, early fusion is to utilize RGB channels projected from images as extra initial features of point clouds for the 3D SparseConvNet. When applying the early fusion strategy, both two modalities including images and point clouds/DSMs are also required in the testing phase. It has a more stringent data requirement than our co-learning methods without probability fusion. In Tables 1 and 3, early fusion is inferior to the same image backbone and point cloud backbone trained with standard and/or enhanced co-learning. For the results of point clouds in Table 1, early fusion even causes an obviously negative effect, inferior to the single-modality baseline. Color information does not lead to an increase in general performance for deep learning-based point cloud semantic segmentation. Sometimes it even reduces the performance of point cloud neural networks (Huang et al., 2020b; Bachhofner et al., 2020). In Table 5, the results obtained by early fusion are close to corresponding scores achieved by standard co-learning and enhanced co-learning. However, if multimodality test data pairs are involved, enhanced co-learning with probability fusion has a better performance than early fusion. The above phenomenons indicate co-learning methods are comparable to conventional data

fusion strategies. They can even replace conventional data fusion in some cases, with lower requirements for the test data.

Apart from the data fusion, previous deep learning-related works for building extraction mostly introduce extra modules to enhance the recognition ability of backbones (Lin et al., 2019; Guo et al., 2021; Li et al., 2021). Such methods usually target specific issues such as blur building boundaries (Guo et al., 2021; Li et al., 2021) and can achieve considerable enhancement compared with backbones. The cost of doing so is introducing more parameters for models and causing bigger model sizes as well as lower efficiency. Co-learning does not influence the structure of backbones. It exploits hidden knowledge via the communication between different modalities to optimize backbones, but it does not create redundant parameters. The usage method of models trained by co-learning is the same as the usage method of single-modality backbones. The main drawback of co-learning is more GPU memory usage and more training time, as there are two neural networks for different modalities that are trained in one GPU in parallel.

Our experiments also suggest the novel idea of utilizing photogrammetric point clouds or DSMs, which are inexpensive and cheap to obtain when stereo- or multi-view high-resolution imagery is available. By comparing the results between the ISPRS Potsdam airborne data set and the Munich WorldView-2 spaceborne data set, we find that a point cloud network trained by the former yields better performance than the WorldView-2 data set. The image resolution directly influences the stereo matching results (Tian et al., 2017). With 5 cm resolution, the Potsdam point cloud data present not only sharper building boundaries, but also rich geometric features. Therefore, the buildings and trees can be well separated without the assistance of spectral information, which is the reason that the 3D point cloud in the Potsdam data set contributes to co-learning better than the Munich WorldView-2 satellite data. The absorption of more reliable transferred point cloud information by enhanced co-learning has a greater improvement on the image results of the ISPRS Potsdam data set.

5. Conclusion

In this paper, we proposed a co-learning framework for automatic building extraction from remotely sensed images and corresponding stereo/multi-view point clouds. The experiments indicate that co-learning is able to enhance the ability of a single-modality neural network by transferring mutual information from another modality with spaceborne or airborne data, and therefore is especially suitable for situations with insufficient labels. Enhanced co-learning, which is superior to standard co-learning in most experiments, shows great potential in learning with unlabeled data pairs. Fusing the prediction results from the multimodality data sets can further improve the building extraction results. Using a fully labeled data set, our method is able to further enhance the capability of the image network with the help of knowledge from corresponding photogrammetric point clouds. The experiments also show that both dense-image-matching and DSM-derived point clouds can benefit a 2D image network via co-learning. In the future, we will explore more architectures of co-learning, and introduce our framework to more diverse remote sensing tasks, such as multi-class semantic segmentation and change detection. In addition, more advanced fusion strategies will be investigated to combine the prediction results from multimodality data.

CRedit authorship contribution statement

Yuxing Xie: Conceptualization, Methodology, Software, Investigation, Writing – original draft, Writing – review & editing, Visualization. **Jiaojiao Tian:** Supervision, Resources, Writing – original draft, Writing – review & editing, Project administration, Methodology. **Xiao Xiang Zhu:** Writing – review & editing, Supervision.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Data availability

Data will be made available on request.

Acknowledgments

This work was supported by a DLR-DAAD research fellowship (57424731) funded by German Aerospace Center (DLR) and German Academic Exchange Service (DAAD). The authors would like to thank Prof. Dr. Peter Reinartz for the provision of necessary data and hardware. The authors would like to thank the German Society for Photogrammetry and Remote Sensing for providing the Potsdam data set. The authors thank Dr. Pablo d'Angelo for generating point clouds from WorldView-2 images, and Xiangtian Yuan for proofreading the manuscript.

References

- Bachhofner, S., Loghin, A.-M., Otepka, J., Pfeifer, N., Hornacek, M., Siposova, A., Schmidinger, N., Hornik, K., Schiller, N., Kähler, O., et al., 2020. Generalized sparse convolutional neural networks for semantic segmentation of point clouds derived from tri-stereo satellite imagery. *Remote Sens.* 12 (8), 1289.
- Baltrušaitis, T., Ahuja, C., Morency, L.-P., 2019. Multimodal machine learning: A survey and taxonomy. *IEEE Trans. Pattern Anal. Mach. Intell.* 41 (2), 423–443.
- Bittner, K., Adam, F., Cui, S., Körner, M., Reinartz, P., 2018. Building footprint extraction from VHR remote sensing images combined with normalized DSMs using fused fully convolutional networks. *IEEE J. Sel. Top. Appl. Earth Obs. Remote Sens.* 11 (8), 2615–2629.
- Choy, C., Gwak, J., Savarese, S., 2019. 4D spatio-temporal convnets: Minkowski convolutional neural networks. In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. pp. 3075–3084.
- d'Angelo, P., 2016. Improving semi-global matching: cost aggregation and confidence measure. In: *XXIII ISPRS Congress, Technical Commission I, Vol. 41. The International Archives of the Photogrammetry, Remote Sensing and Spatial Information Sciences*, pp. 299–304.
- Graham, B., Engelcke, M., Van Der Maaten, L., 2018. 3D semantic segmentation with submanifold sparse convolutional networks. In: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. pp. 9224–9232.
- Guo, H., Shi, Q., Marinoni, A., Du, B., Zhang, L., 2021. Deep building footprint update network: A semi-supervised method for updating existing building footprint from bi-temporal remote sensing images. *Remote Sens. Environ.* 264, 112589.
- He, K., Zhang, X., Ren, S., Sun, J., 2016. Deep residual learning for image recognition. In: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. pp. 770–778.
- Huang, S., Usvyatsov, M., Schindler, K., 2020b. Indoor scene recognition in 3D. In: *2020 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*. IEEE, pp. 8041–8048.
- Huang, R., Xu, Y., Hong, D., Yao, W., Ghamisi, P., Stilla, U., 2020a. Deep point embedding for urban classification using ALS point clouds: A new perspective from local to global. *ISPRS J. Photogramm. Remote Sens.* 163, 62–81.
- ISPRS, 2022. 2D semantic labeling contest - potsdam. URL <https://www.isprs.org/education/benchmarks/UrbanSemLab/2d-sem-label-potsdam.aspx>.
- Jaritz, M., Vu, T.-H., Charette, R.d., Wirbel, E., Pérez, P., 2020. xmuda: Cross-modal unsupervised domain adaptation for 3d semantic segmentation. In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. pp. 12605–12614.
- Li, Q., Mou, L., Hua, Y., Shi, Y., Zhu, X.X., 2021. Building footprint generation through convolutional neural networks with attraction field representation. *IEEE Trans. Geosci. Remote Sens.*
- Li, Q., Zorzi, S., Shi, Y., Fraundorfer, F., Zhu, X.X., 2022. RegGAN: An end-to-end network for building footprint generation with boundary regularization. *Remote Sens.* 14 (8), 1835.
- Lin, J., Jing, W., Song, H., Chen, G., 2019. ESFNet: Efficient network for building extraction from high-resolution aerial images. *IEEE Access* 7, 54285–54294.
- Lin, Y., Vosselman, G., Cao, Y., Yang, M.Y., 2021. Local and global encoder network for semantic segmentation of Airborne laser scanning point clouds. *ISPRS J. Photogramm. Remote Sens.* 176, 151–168.
- Ma, M., Ren, J., Zhao, L., Tulyakov, S., Wu, C., Peng, X., 2021. SMIL: Multimodal learning with severely missing modality. In: *Proceedings of the AAAI Conference on Artificial Intelligence, Vol. 35*. pp. 2302–2310.

- 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. pp. 36–43.
- Qi, C.R., Su, H., Mo, K., Guibas, L.J., 2017. Pointnet: Deep learning on point sets for 3d classification and segmentation. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 652–660.
- Rahate, A., Walambe, R., Ramanna, S., Kotecha, K., 2022. Multimodal Co-learning: Challenges, applications with datasets, recent advances and future directions. *Inf. Fusion* 81, 203–239.
- Ronneberger, O., Fischer, P., Brox, T., 2015. U-net: Convolutional networks for biomedical image segmentation. In: International Conference on Medical Image Computing and Computer-Assisted Intervention. Springer, pp. 234–241.
- Schmitt, M., Zhu, X.X., 2016. Data fusion and remote sensing: An ever-growing relationship. *IEEE Geosci. Remote Sens. Mag.* 4 (4), 6–23.
- Shi, Y., Li, Q., Zhu, X.X., 2020. Building segmentation through a gated graph convolutional neural network with deep structured feature embedding. *ISPRS J. Photogramm. Remote Sens.* 159, 184–197.
- Sun, Y., Fu, Z., Sun, C., Hu, Y., Zhang, S., 2021b. Deep multimodal fusion network for semantic segmentation using remote sensing image and LiDAR data. *IEEE Trans. Geosci. Remote Sens.* 60, 1–18.
- Sun, X., Wang, B., Wang, Z., Li, H., Li, H., Fu, K., 2021a. Research progress on few-shot learning for remote sensing image interpretation. *IEEE J. Sel. Top. Appl. Earth Obs. Remote Sens.* 14, 2387–2402.
- Thomas, H., Qi, C.R., Deschaud, J.-E., Marcotegui, B., Goulette, F., Guibas, L.J., 2019. Kpconv: Flexible and deformable convolution for point clouds. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 6411–6420.
- Tian, J., Cui, S., Reinartz, P., 2014. Building change detection based on satellite stereo imagery and digital surface models. *IEEE Trans. Geosci. Remote Sens.* 52 (1), 406–417.
- Tian, J., Reinartz, P., d'Angelo, P., Ehlers, M., 2013. Region-based automatic building and forest change detection on Cartosat-1 stereo imagery. *ISPRS J. Photogramm. Remote Sens.* 79, 226–239.
- Tian, J., Schneider, T., Straub, C., Kugler, F., Reinartz, P., 2017. Exploring digital surface models from nine different sensors for forest monitoring and change detection. *Remote Sens.* 9 (3), 287.
- Tong, X.-Y., Xia, G.-S., Lu, Q., Shen, H., Li, S., You, S., Zhang, L., 2020. Land-cover classification with high-resolution remote sensing images using transferable deep models. *Remote Sens. Environ.* 237, 111322.
- Xie, Y., Tian, J., Zhu, X.X., 2020. Linking points with labels in 3D: A review of point cloud semantic segmentation. *IEEE Geosci. Remote Sens. Mag.* 8 (4), 38–59.
- Xu, Y., Tong, X., Stilla, U., 2021. Voxel-based representation of 3D point clouds: Methods, applications, and its potential use in the construction industry. *Autom. Constr.* 126, 103675.
- Yousefhussein, M., Kelbe, D.J., Ientilucci, E.J., Salvaggio, C., 2018. A multi-scale fully convolutional network for semantic labeling of 3D point clouds. *ISPRS J. Photogramm. Remote Sens.* 143, 191–204.
- Zadeh, A., Liang, P.P., Morency, L.-P., 2020. Foundations of multimodal co-learning. *Inf. Fusion* 64, 188–193.
- Zhang, L., Lan, M., Zhang, J., Tao, D., 2021. Stagewise unsupervised domain adaptation with adversarial self-training for road segmentation of remote-sensing images. *IEEE Trans. Geosci. Remote Sens.* 60, 1–13.
- Zheng, Z., Ma, A., Zhang, L., Zhong, Y., 2021. Deep multisensor learning for missing-modality all-weather mapping. *ISPRS J. Photogramm. Remote Sens.* 174, 254–264.
- Zhou, W., Jin, J., Lei, J., Hwang, J.-N., 2021. CEGFNet: Common extraction and gate fusion network for scene parsing of remote sensing images. *IEEE Trans. Geosci. Remote Sens.* 60, 1–10.
- Zhu, Q., Liao, C., Hu, H., Mei, X., Li, H., 2020. MAP-Net: Multiple attending path neural network for building footprint extraction from remote sensed imagery. *IEEE Trans. Geosci. Remote Sens.* 59 (7), 6169–6181.
- Zoph, B., Ghiasi, G., Lin, T.-Y., Cui, Y., Liu, H., Cubuk, E.D., Le, Q., 2020. Rethinking pre-training and self-training. *Adv. Neural Inf. Process. Syst.* 33, 3833–3845.

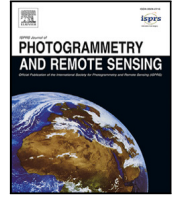
C Mario Fuentes Reyes*, Yuxing Xie*,
Xiangtian Yuan*, Pablo d'Angelo,
Franz Kurz, Daniele Cerra, and
Jiaojiao Tian. “A 2D/3D multimodal
data simulation approach with
applications on urban semantic
segmentation, building extraction and
change detection.” *ISPRS Journal of
Photogrammetry and Remote Sensing*
205 (2023): 74-97. (* equal
contribution)

<https://doi.org/10.1016/j.isprsjprs.2023.09.013>



Contents lists available at ScienceDirect

ISPRS Journal of Photogrammetry and Remote Sensing

journal homepage: www.elsevier.com/locate/isprsjprs

A 2D/3D multimodal data simulation approach with applications on urban semantic segmentation, building extraction and change detection

Mario Fuentes Reyes ^{a,1}, Yuxing Xie ^{b,a,1}, Xiangtian Yuan ^{a,1}, Pablo d'Angelo ^a, Franz Kurz ^a, Daniele Cerra ^a, Jiaojiao Tian ^{a,*}

^a Remote Sensing Technology Institute, German Aerospace Center (DLR), Muenchener Strasse 20, Weßling, 82234, Germany

^b Chair of Data Science in Earth Observation, Technical University of Munich, Munich, 80333, Germany

ARTICLE INFO

Keywords:

3D change detection
Building extraction
Urban semantic segmentation
Synthetic datasets

ABSTRACT

Advances in remote sensing image processing techniques have further increased the demand for annotated datasets. However, preparing annotated multi-temporal 2D/3D multimodal data is especially challenging, both for the increased costs of the annotation step and the lack of multimodal acquisitions available on the same area. We introduce the Simulated Multimodal Aerial Remote Sensing (SMARS) dataset, a synthetic dataset aimed at the tasks of urban semantic segmentation, change detection, and building extraction, along with a description of the pipeline to generate them and the parameters required to set our rendering. Samples in the form of orthorectified photos, digital surface models and ground truth for all the tasks are provided. Unlike existing datasets, orthorectified images and digital surface models are derived from synthetic images using photogrammetry, yielding more realistic simulations of the data. The increased size of SMARS, compared to available datasets of this kind, facilitates both traditional and deep learning algorithms. Reported experiments from state-of-the-art algorithms on SMARS scenes yield satisfactory results, in line with our expectations. Both benefits of the SMARS datasets and constraints imposed by its use are discussed. Specifically, building detection on the SMARS-real Potsdam cross-domain test demonstrates the quality and the advantages of proposed synthetic data generation workflow. SMARS is published as an ISPRS benchmark dataset and can be downloaded from https://www2.isprs.org/commissions/comm1/wg8/benchmark_smars/.

1. Introduction

Recent years have seen dramatic progress in the development of image processing algorithms. Deep neural networks have outperformed traditional image processing approaches on most of the classical image understanding and interpretation problems (Minaee et al., 2021; Xie et al., 2020).

At the early stages of computer vision, high quality manually labeled data series were published as benchmark datasets for computer vision tasks including classification and recognition, such as PASCAL Visual Object Classes (VOC) 150 (Everingham et al., 2010), KITTI (Geiger et al., 2013), Microsoft Common Objects in Context (MS COCO) (Lin et al., 2014), and Cityscapes (Cordts et al., 2016). These large-scale benchmark datasets have been then used to develop and validate deep learning algorithms. The performance of these networks highly depends on the amount and the quality of the available training data, which are expensive and sometimes difficult to acquire. The vast majority

of newly published papers are dealing with the “Training” phase, as the collection of training data represents often the bottleneck for these applications (Zhou, 2018; Pourpanah et al., 2022). The performance of artificial intelligence (AI) algorithms is severely limited whenever insufficient data with low number of samples, unbalanced classes, or inaccurate annotations are available (Li et al., 2020; Xie et al., 2023).

Today, advanced neural network architectures are adopted in many other fields such as medical image analysis and remote sensing. Many excellent approaches originally proposed in the computer vision community have been successfully applied and further developed for earth observation tasks, including building/road extraction, semantic classification, and change detection (Zhu et al., 2017; Xie et al., 2020). Additional hindrances are added to the ones listed above regarding the availability of training datasets for specific problems in remote sensing.

The nature of remote sensing data is often multimodal, as the different sensors usually provide complementary information on a target

* Corresponding author.

E-mail addresses: Mario.FuentesReyes@dlr.de (M. Fuentes Reyes), Yuxing.Xie@dlr.de (Y. Xie), Xiangtian.Yuan@dlr.de (X. Yuan), Pablo.Angelo@dlr.de (P. d'Angelo), franz.kurz@dlr.de (F. Kurz), Daniele.Cerra@dlr.de (D. Cerra), Jiaojiao.Tian@dlr.de (J. Tian).

¹ These (the first three) authors contributed equally to this work.

<https://doi.org/10.1016/j.isprsjprs.2023.09.013>

Received 24 April 2023; Received in revised form 15 September 2023; Accepted 16 September 2023

Available online 6 October 2023

0924-2716/© 2023 The Author(s). Published by Elsevier B.V. on behalf of International Society for Photogrammetry and Remote Sensing, Inc. (ISPRS). This is an open access article under the CC BY license (<http://creativecommons.org/licenses/by/4.0/>).

on the ground, by measuring the backscattered radiation in different frequency ranges (including visible, infrared, thermal emissions, and microwaves), and estimating ground and canopy height parameters yielding 3D information. Additionally, this information is seldom acquired in a single acquisition from the same observation platform, therefore introducing variations in viewing angle, sensing geometry, acquisition time, atmospheric conditions, and position of the source of radiation. The availability of different sources of information on the same area is often beneficial for remote sensing applications: for example, the fusion of 2D/3D data is advantageous for image classification (Ghamisi et al., 2016), building extraction (Hosseinpour et al., 2022), and change detection (Tian et al., 2013; Qin et al., 2016). In addition to the limited availability of the corresponding 2D/3D training data, sufficient variability in the data must be present in order to train a valid deep learning (DL) neural network. Furthermore, annotating changes in a large scale remote sensing images is time-consuming and error-prone. To our knowledge, there is no 2D/3D multimodal building change detection benchmark dataset available until now, which in part limits the implementation of effective AI techniques for 3D change detection.

To this end, synthetic data have been proposed in order to fill this gap in a less expensive way. Currently, several available studies highlight the advantages of using synthetic data for solving real-world problems, especially in the fields of medicine and healthcare, for which real physical experiments are often linked to expensive retrieval costs (Chen et al., 2021). Besides avoiding data acquisition problems, using synthetic data has an increased flexibility when coping with data balancing, in particular for the studying of rare diseases (Chen et al., 2021). Several other studies experience similar problems, such as the ones conducted in the field of physics research, where the process of observing real data may be particularly long and expensive (Stoecklein et al., 2017; Li et al., 2021). Existing literature in remote sensing use synthetic data in order to evaluate their algorithms or fuse them with real data, yielding an efficient training for augmentation tasks. However, in addition to evaluating AI models, the synthetic data should be suitable for integration with real data in order to solve application oriented problems (Nikolenko, 2021). Thus, the domain gap between synthetic and real data should be limited.

Directly rendering of digital surface models (DSM) from a 3D environment retrieves highly accurate products as presented in Fig. 1(a), exhibiting sharp boundaries around the buildings without any occlusions or gaps. Such precise DSM can be hardly achieved using real data with the currently available optical acquisition and stereo matching techniques, as results obtained from photogrammetry pipeline are characterized by blurred boundaries and contain outliers (Fig. 1(b)). In order to reduce the gaps between rendered and real data, we aim at defining a novel approach generating synthetic DSMs with the same limitations of real ones, as for the DSM reported in Fig. 1(c), which more closely resembles the level of detail in Fig. 1(b) with respect to the generation using directly rendered samples.

Considering all the points above, we propose a novel synthetic photogrammetric data generation procedure with special focus on the application of 2D/3D multimodal classification (or segmentation), building detection and 3D change detection. We use this dataset as source and real remote sensing imagery as target for domain adaptation experiments. The main contributions of our paper are the following:

- A workflow to produce synthetic data with higher level of realism.
- A 2D/3D multimodal remote sensing dataset, which we name the Simulated Multimodal Aerial Remote Sensing (SMARS).
- A systematic evaluation of the performance of SMARS on building extraction, multi-class semantic classification and change detection.

This paper is organized as follows. Section 2 presents an overview of the state of the art of synthetic data used in remote sensing and the related studies in virtual city synthetic data generation. In Section 3, the

proposed synthetic data generation, which include the multi-temporal stereo imagery simulation as well as the data process procedure is introduced in detail. Section 4 illustrates the proposed method in details. In Section 4, we further describe the details of the generated SMARS dataset and the tasks to be addressed, including building extraction (Section 5), multi-class semantic segmentation (Section 6) and building change detection (Section 7). Moreover, extensive discussions are presented in Section 8. Section 9 provides the conclusions.

2. State of the art

2.1. Existing real 2D/3D multimodal benchmark datasets

Due to the aforementioned reasons, the number of available 2D/3D multimodal benchmark datasets is limited. The ISPRS Potsdam dataset² is at the moment of writing the most popular public benchmark for 2D/3D semantic labeling, and it is also widely used to test and validate building extraction methods (Xie et al., 2023). This dataset provides airborne orthoimages and corresponding DSMs generated via dense image matching. The ground sampling distance of both images and DSMs is 5 cm. The original training data have 24 pairs of tiles, each having a size of 6000 × 6000 pixels (300 m × 300 m). The ISPRS Vaihingen³ is another airborne benchmark dataset containing both 2D images and DSMs. However, its limitation of having only near-infrared, red, and green bands restricts its applicability in mainstream applications requiring RGB images, as the blue band is not available. DroneDeploy⁴ is a 2D/3D multimodal dataset containing aerial scenes captured from drones. Its main limitation is that it provides only original irregular mosaics, furthermore, it lacks a clear separation between training, validation, and test sets. Hence, it is not widely used in the community.

On the subject of change detection, there are a number of several single modal benchmark datasets available (Caye Daudt et al., 2018; Gupta et al., 2019; Caye Daudt et al., 2019; Shao et al., 2021). To the best of our knowledge, 3DCD is currently the only benchmark that provides 2D/3D multimodal data suitable for evaluating deep learning algorithms in remote sensing change detection (Coletta et al., 2022; Marsocci et al., 2023). Nonetheless, in this dataset, DSMs are obtained by LiDAR sensors, whose acquisition dates as well as the Ground Sample Distance (GSD) differ from the corresponding optical images. This may potentially affect their paired use in multimodal algorithms. Apart from the voids in the DSM, the changes are not exclusively defined for buildings but also for other land cover changes. In addition, the dataset only covers the urban center of the city of Valladolid in Spain, and therefore is not suitable for domain adaptation experiments.

2.2. Synthetic data in remote sensing

Curating real 2D/3D multimodal datasets requires valid data acquisition and processing, which is then compounded by the time consuming and costly step of manual annotation. Therefore, the generation of synthetic data for remote sensing applications is preferred whenever real-world data are not available or difficult to collect. Authors in Börner et al. (2001) propose SENSOR (Software Environment for the Simulation of Optical Remote Sensing Systems) to simulate hyperspectral images. Artificial orbit and attitude data are used in Schwind et al. (2012) to analyze the co-registration errors between visible and near-infrared (VNIR) and short-wavelength infrared (SWIR) imagery for the design of the EnMAP (Environmental Mapping and Analysis Program) satellite. Simulated SAR images are generated in Tao et al.

² <https://www.isprs.org/education/benchmarks/UrbanSemLab/2d-sem-label-potsdam.aspx>.

³ <https://www.isprs.org/education/benchmarks/UrbanSemLab/2d-sem-label-vaihingen.aspx>.

⁴ <https://github.com/dronedeploy/dd-ml-segmentation-benchmark>.

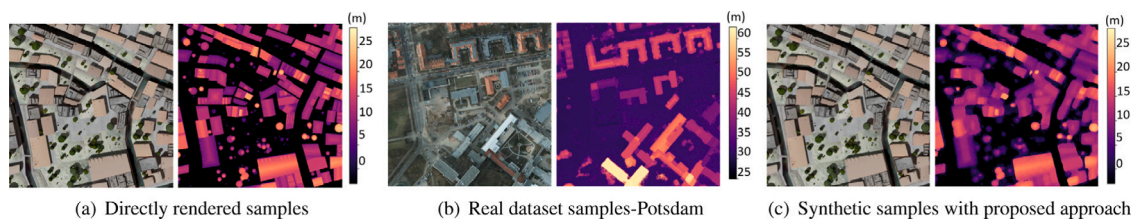


Fig. 1. Quality differences between synthetic and real data. Elevation scale for the DSM is in meters.

(2013) for change detection. Synthetic data have been explored in vegetation studies. Li and Strahler (1985) proposed a geometric-optical forest canopy model to explain the variance of a pixel in low resolution images of forest stands. The model represents conifers with Lambertian surfaces shaped as cones, which cast shadows on the ground.

Moreover, multi-temporal datasets are more costly to prepare and the annotation is more challenging with respect to single images: the number of public change detection benchmarks is therefore rather limited, furthermore, most of them are single modal data and some are characterized either by small size or a low ground sampling distance (Shi et al., 2020). The described difficulties in curating the described multi-temporal datasets can be mitigated by relying on synthetic data. For instance, Townshend et al. (1992) simulate a set of different misregistrations degrees to find out their impact on vegetation change detection. Simulated change detection datasets have been used in Almutairi and Warner (2010) to compare state-of-the-art algorithms. The simulated data therein are rather simple with few shape patterns and additional artificial noise. A real LiDAR point cloud is used in de Gélis et al. (2021) to generate one Level of Detail 2 (LoD2) model as a pre-emptive dataset. By manually adding or removing buildings in the model, the construction or demolition of buildings can thus be simulated. This results in a time-consuming process, and with buildings as the only objects present in the 3D model, the results have a large domain gap with real urban 3D models.

In order to close the domain gap between simulated and real data, Hoeser and Kuenzer (2022) propose an artificial data generation procedure by including expert knowledge in a highly structured manner to control the automatic image and label generation, by employing an ontology in the process. However, with more complex background information, urban change detection is more difficult to simulate and control.

Radiative transfer models have been explored to simulate remote sensing data. Recently, in order to analyze vegetation behaviors, several synthetic data generation tools have been introduced based on the radiative transfer model (Qi et al., 2019; Disney et al., 2006). As one of the most representative software for radiative transfer modeling, the Discrete Anisotropic Radiative Transfer (DART) can accurately simulate 3D radiative budget and chlorophyll fluorescence of vegetation (Gastellu-Etchegorry et al., 2015), as well as passive remote sensing and LiDAR signals of natural and urban scenarios. It is capable of precisely simulating the vegetation reflectance in several wavelengths and also works for dense forests with complex canopy structures (Janoutová et al., 2019). However, rendering more realistic urban scenes using DART is quite challenging for inexperienced users due to its complex parameters requiring expert knowledge in the relevant fields. In contrast, 3D rendering engines such as Blender or Unity are considerably easier to use, offer more sophisticated rendering features, and support several formats of 3D models and materials (Richter et al., 2016; Shah et al., 2018; Fabbri et al., 2021). Moreover, in order to construct a large urban scene, many detailed and realistic 3D models for vegetation and buildings are needed. 3D rendering engines not only have more large-scale 3D city models but can also edit those models while DART does not support editing, which poses difficulty in simulating urban changes. In comparison, the 3D rendering engine is more advantageous in creating multi-temporal urban scenes of large regions.

2.3. Virtual city synthetic data

Generating data from a virtual model is currently becoming more popular in computer vision due to the capabilities of modeling software. However, the application of synthetic data is rather limited if the domain gap with the real data is too large. A virtual model can contain anything from a small object to a city. For example, building models can be used to create indoor based point clouds (Ma et al., 2020) or depth and semantics, as in Hypersim (Roberts et al., 2021). Studies related to autonomous driving have also benefited from the developments of synthetic data creation. A widely known example is the SYNTHIA dataset (Ros et al., 2016) that provides synthetic images of urban scenes labeled for semantic segmentation. Such scenes are rendered from a virtual New York City 3D model with the Unity game engine. The dataset includes segmentation annotations for 13 classes including pedestrians, cyclists, buildings, and roads. Another approach is used by CARLA (Dosovitskiy et al., 2017), an open source simulator that supports the training, prototyping, and validating of autonomous driving models. CARLA facilitates the data acquisition from street view for the generation of segmentation and depth maps. Similarly, the ParallelEye dataset (Li et al., 2019) generates images from the CityEngine software with depth and optical flow as part of the ground truth.

A similar setting can be considered for the simulation of aerial or satellite imagery. The Synthinel-1 dataset (Kong et al., 2020), also based on CityEngine, targets the building/no-building classification from an airplane perspective. The article also addresses the advantages of synthetic imagery by ablation studies. The VALID dataset (Chen et al., 2020), on the other hand, focuses on panoptic segmentation and depth estimation over urban infrastructure. Furthermore, the SyntCities dataset (Fuentes Reyes et al., 2022) provides semantics and disparity maps, making the data suitable for stereo reconstruction. The STPLS3D dataset (Chen et al., 2022) provides point clouds, and semantic and instance maps based on open geospatial data sources. Authors in Xiao et al. (2022) simulate LiDAR acquisition for an urban environment and deliver the dataset as point clouds.

However, further applications of synthetic data are limited by the large differences in characteristics between real and synthetic data. A remarkable example is the much higher quality of the DSM obtained from the virtual 3D models in comparison with the one generated from photogrammetric matching. Edges are usually sharper in the simulated data, and the occlusions are absent in the generated ground truth. In addition, images from real scenarios show imperfect textures, light reflection, seasonal changes, the presence of temporary objects (cars, pedestrians, street advertisements, etc.), atmospheric effects, and other elements that cannot be easily modeled in software. Hence, the simulation is mostly restricted to the geometry of the scene, textures, and camera properties. Still, the rendered images can visually resemble real cases and help to compensate for the limits of real sensors (such as sparsity) and reduce the costs to generate ground truth.

3. Methodology on synthetic data generation

To close the gap between synthetic data collection and remote sensing applications we combine two techniques, airborne data collection from virtual city and photogrammetric stereo data preparation. In this

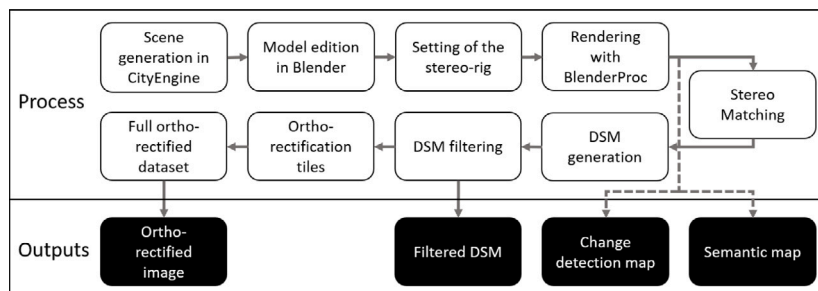


Fig. 2. Basic description of the pipeline used to generate the dataset.

section, we propose a novel workflow to generate a 2D–3D multimodal dataset. A diagram to summarize it is shown in Fig. 2. It consists of three parts: 3D virtual city design, imagery simulation, and data processing.

3.1. 3D virtual city design

In order to produce a realistic change scenario we used a 3D virtual city as a starting point to simulate the scene growth process, instead of directly generating artificial images. We built the 3D scenes based on the CityEngine software,⁵ a suite facilitating the modeling of urban environments based on the computer-generated Architecture (CGA) shape grammar language. The software was also used to develop the above-mentioned ParallelEye and Synthinel-1 datasets (Li et al., 2019). CityEngine supports building a city model from land cover maps, such as Open Street Map, or a manually designed base map. However, designing a virtual world with carefully customized features would require relevant expert knowledge and would be time-consuming. Therefore, we selected two predefined city models from ESRI and further refined them accordingly.

In this paper, we chose two typical European cities: Paris and Venice. Henceforth we refer to them as SParis and SVenice, respectively. The selected city models have a variety of textures and architectures resembling the original cities, as well as a large surface that allows the inclusion of a large number of buildings in the subsequent rendered images. The buildings are defined in terms of roof type, roof angle (if any), height, number of floors, floor height, and size of the parcel. In order to have a lifelike view, we further edited the 3D model of the cities by modifying the streets in order to have a more realistic topography, as the original version has streets with the shape of letters. The trees were replaced with textured ellipsoids instead of the original ones represented with a uniform color. Additionally, some areas were manually corrected in order to ensure that any parcel in the area included urban content.

A large pool of textures has been used in the provided models, namely 219 for buildings (rooftops and facades) and 87 for vegetation. For the latter ones, we edited the default textures of the ellipsoids by creating a dense representation of leaves in order to resemble canopies. While still limited with respect to the full variety of the real world, these refinements helped produce a scene with sufficient variability.

As the dataset is mainly intended for change detection applications in urban areas, each city model was generated with two versions simulating the city's growth:

- A case where approximately 50% of the parcels are covered by buildings. This is considered the model before changes happen and we refer to it as pre-model in the remainder of the paper.
- A case with approximately 70% of the parcels covered by buildings. Some areas defined previously as gardens are replaced by constructions, while some buildings have been removed and substituted with green areas. This model contains the changes to be detected, and is therefore named post-model.

In Fig. 3, we show samples for both the pre- and post-model, respectively Fig. 3(a) and Fig. 3(b). The central image exhibits a higher number of buildings and less vegetation cover. Also, some of the original buildings have been replaced with lawns or vegetation.

According to the requirements described above, we adapted a total of four city models (two cities, two epochs) and exported all cases in the Wavefront (with extension .obj) format for further editing. The manipulation of the scenes in CityEngine demands about 17 GB of RAM memory.

Subsequently, we loaded the Wavefront files in Blender, an open source tool for modeling, simulation, and rendering. We applied the BlenderProc pipeline (Denninger et al., 2020) to render the images. Our approach for the rendering is based on the one described in SyntCities (Fuentes Reyes et al., 2022) and we generated for this case the colored images (we refer henceforth to them as “optical”) and the semantic maps.

Within Blender we split the geometry of the scenes according to their textures, separating all the surfaces into the required semantic labels. The available categories include: vegetation, streets, rooftops (mansard, gambrel, gable, hip and flat styles), facades, grass, landmarks, cars, and background. We combined them into five typical land cover classes used for urban mapping, including buildings (all rooftops, facades and landmarks), streets, high vegetation (trees), grass (lawns) and others (cars, water, bare soil or background).

We simulate different illumination conditions by setting an artificial Sun in two specific positions for the pre-/after-event models, reproducing two different times for data acquisition. The selected angles were 70° for elevation, and 217° (pre-model) and 160° (post-model) for azimuth. The same conditions were applied to both cities. Finally, we added a homogeneous plane under the ground level of each scene to avoid undefined regions (no value pixels) in the rendering process, which is assigned to the “other” category. Without it, distance would be considered to be infinite if there is an empty region in the objects. This plane guarantees a color and depth value for each rendered pixel.

3.2. Airborne stereo imagery simulation

SMARS is designed to resemble aerial imagery and the simulated camera is constrained by a stereo rig, which helps to later generate a digital surface model (DSM). In this part, we provide more details on the simulated data acquisition and camera parameters.

Firstly, the simulated camera is located 2 km above the origin of the scenes. Since we used synthetic models that are not georeferenced, the origin of the coordinate system assigned by City Engine is used by default. An arbitrary point located at the center of the model and on the terrain level is taken as a reference for the rendering process.

In Fig. 4(a), we show the configuration of the stereo rig. In order to simulate the stereo imagery acquisition procedure, two cameras are located at the same distance from the rig center with a baseline of

⁵ <https://www.esri.com/en-us/arcgis/products/esri-cityengine/overview>.

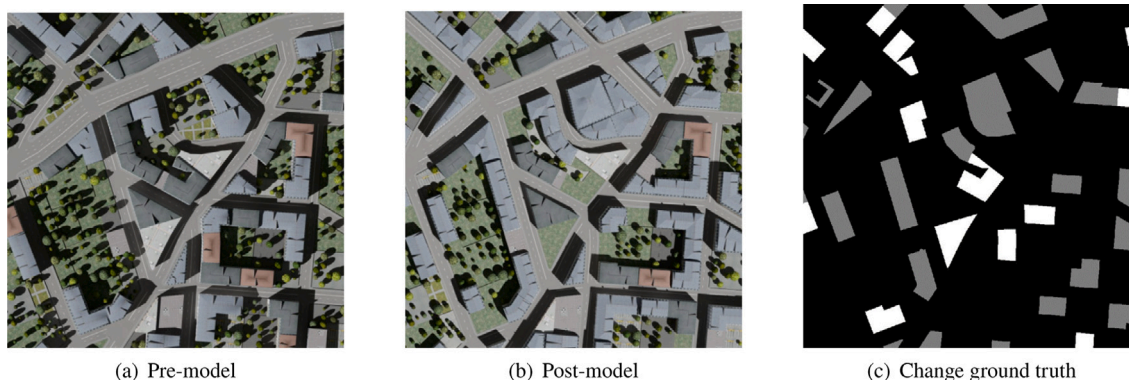


Fig. 3. Samples from the pre- and post-models after rendering with associated ground truth for change detection. The pre-model has a lower building density and different illumination conditions. Black regions in the ground truth exhibit no change, while gray indicates new buildings and white replaced ones, respectively.

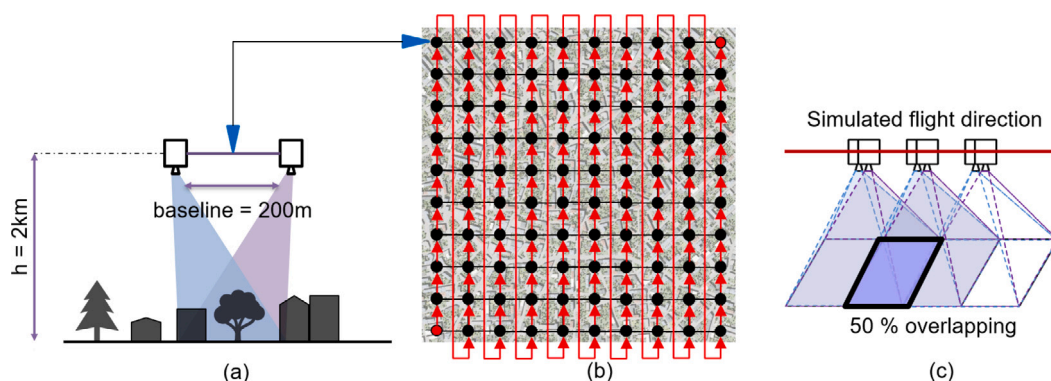


Fig. 4. Simulated stereo configuration. (a) Stereo rig, where the converge distance and baseline of the cameras have been adapted to cover the same area on the ground. (b) The trajectory of the simulated camera above the scene. (c) Overlapping between adjacent samples is 50% for both horizontal and vertical directions. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

200 m in all cases. Both cameras follow the pinhole model and have the same focal length. As image resolution plays an essential role in transfer learning, we aim to provide this image dataset in two GSDs, namely 30 cm and 50 cm. Following Eq. (1), we set the focal length of the cameras to 234.37 mm and 140.62 mm, respectively.

$$f = \frac{height \cdot sensor_width}{covered_area} \tag{1}$$

where f is the focal length, $height = 2000$ m as described above, $sensor_width = 36$ mm for the simulated camera and $covered_area = 1024 * GSD$, being 1024 the size in pixels of the output image. The converge distance is set to 2 km (same as the simulated height) with an off-axis camera, which allows us to cover the same area on the ground from two different points of view. This configuration is also modeled with the offset of the principal point in the intrinsic matrix of the camera.

In Fig. 4(b) we illustrate the trajectory of the simulated camera above the scene. We rendered images at 100 positions within a regular square grid, with strides set as 153.6 m and 256 m for 30 cm and 50 cm GSD, respectively. The center of the grid is set to be close to one of the scenes, so most of the content is included. In order to simulate a real-world airborne data acquisition campaign, the pair of stereo-cameras are moved from the lower-left to the upper-right corner with a constant stride. The points belonging to the grid represent the location of the center of the stereo rig (see the arrow with blue extremes). This means that the cameras are located symmetrically to the left and right side of each point.

Overlapping between adjacent samples is set to 50% in both the horizontal and vertical directions of the grid. A visual representation of the overlapping is given in Fig. 4(c), where the camera pairs along the

simulated flight direction are also included. The images are rendered with a size of 1024×1024 pixels.

After rendering, a semantic segmentation map to be used as ground truth (GT) is delivered with the categories described previously (buildings, streets, vegetation, lawns and others). For the building extraction GT map, we combine all categories except building to no-building, enabling binary semantic segmentation. With the pre-/post-event building extraction GT maps, we calculate the building change detection map by taking only the building class for comparison. Three change classes are included:

- No change: buildings or no-buildings have the same semantic label pre/post-event images.
- Construction: pixels labeled as building in the post event images are no-building in the pre-event images.
- Demolition: pixels labeled as building in the pre-event images are replaced by the no-building label.

The change detection ground truth is directly rendered from the 3D model with an orthographic view. Labels for the semantic categories are also directly rendered from Blender, as BlenderProc generates a category for each object in the scene. We assigned all geometric elements to the desired categories. The building masks are a simplified version of the category maps considering a binary building/non-building case. For the change detection mask, building masks are compared and labeled according to their difference. In this case, all generated ground truth is generated in the rendering step, and therefore perfectly matches the original images. Due to the orthorectification process described in Section 3.4, the alignment will not be perfect as this simulates the quality obtained from a photogrammetric pipeline. Results show that the alignment differences do not have a significant impact on the three evaluated tasks.

3.3. Stereo matching and DSM generation

Although very precise 3D point clouds and DSMs can be directly delivered with the rendering software, the quality of these data for all cases will be higher than the real-world 3D point clouds generated by stereo matching techniques, where many mismatching errors and occlusions occur. Thus, in this work we only take the synthetic stereo image pairs and generate the orthophotos and 3D point clouds with a traditional approach. First, we assign a fake UTM projection to all synthetic airborne stereo images, in order to enable the photogrammetric processing. Concretely, we assign the tiles to the UTM zone 31N coordinate system (EPSG:32631), even though the simulated model does not match any region on a real map, this area would match the city of Paris. Additionally, for the photogrammetric pipeline we enter the camera extrinsic and intrinsic matrices, including focal length, principal points, and camera rotation and translation parameters. The extrinsic and intrinsic parameters of the synthetic data are precise and there was no artificial noise added. We assume that the deviation of the positional accuracy is negligible, as the relative accuracy of real-world aerial images used for stereo matching is better than 0.2 pixels.

A DSM is generated from all tiles by using the CATENA pipeline (Krauß, 2014), which is used for multiple tasks related to the processing of satellite imagery. The disparity estimation, which is the first step, is computed via Semi-Global Matching (SGM) (Hirschmüller, 2008), an algorithm widely used for stereo matching due to its good balance between accuracy and computational costs. SGM takes a rectified stereo image pair as input and estimates a disparity map. We apply the implementation of SGM described in d’Angelo and Reinartz (2011), which takes satellite data as input, and set the penalty parameters $P_1 = 400$, $P_2 = 800$ and the window size for the Census transform (Zabih and Woodfill, 1994) to 7×9 .

After the matching and the use of the camera parameters to determine the 3D location of each pixel, we retrieve a georeferenced DSM for each stereo pair. We subsequently merge all the stereo pair DSMs by using the median of all values belonging to the same location, resulting in one final DSM for each virtual city.

As a real DSM generation procedure, gaps are present due to matching failures or occlusions. We apply an inverse distance weighted interpolation in order to fill the remaining holes (Bartier and Keller, 1996).

3.4. Orthophoto and reference data

The orthorectification process for the rendered optical tiles is implemented in a GPU as described in Kurtz et al. (2012), considering as input the generated DSM, and the intrinsic and extrinsic parameters of the optical images. The outputs are take into account occlusions by buildings and vegetation. Bilinear interpolation is used to resample the orthorectified images to a given ground sampling distance.

We merge all the tiles into a single large image with the warp utility from the GDAL library (GDAL/OGR contributors, 2022), having as a result a complete orthorectified optical image, corresponding to the DSMs at pixel level.

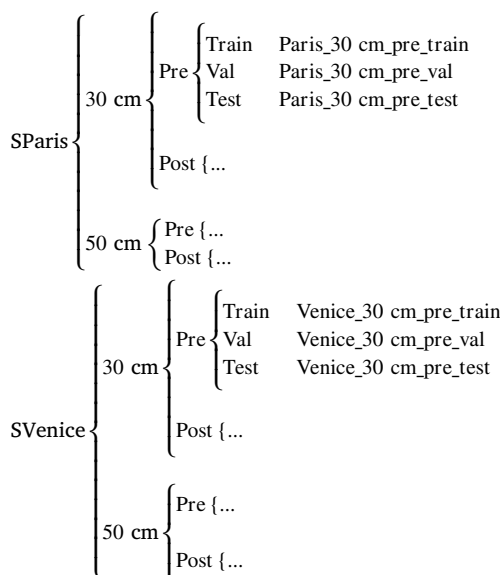
4. Experimental design

In this section we describe some additional details of the generated SMARS dataset and the delimitation of the regions used for training and testing in the deep learning algorithms for both cities. Additionally, we explain the tasks to be addressed with our generated data to show the advantages and constraints of SMARS.

4.1. SParis and SVenice multimodal data structure

The pre- and post-event DSMs and orthophotos are generated using the workflow described in Section 3. All the datasets are projected to the UTM zone 31N coordinate system and cropped in order to cover the same regions. Fig. 5 reports examples of the generated DSMs. Buildings appear well delimited and easy to identify in most cases, while other elements such as streets or vegetation appear incomplete or blurred. There is a clear difference between the models obtained using 30 cm and 50 cm GSD respectively, as the former exhibits sharper edges with individual trees easy to identify, while the latter exhibits some blobs merging different objects. Despite some artifacts or the presence of outliers, the DSMs still have a high quality in all cases due to their synthetic nature.

The final dataset splittings are summarized in the diagram below. We list all possible subsets but report the names for only three of them for each city in order to simplify the diagram, with the remaining cases following the same nomenclature. For each subset, we have available optical images, DSMs, semantic maps, and building masks for both pre- and post-event scenarios. Additionally, we have building change detection masks for the difference between pre- and post-images. All these cases are shown in Fig. 6.



Figs. 7 and 8 illustrate the pre-event training, validation and test areas for SParis and SVenice, respectively. For post-event data, the splitting in training, validation and test data follows the same process. The size of both SParis and SVenice rasters with 30 cm GSD is 5600×5600 pixels. For SParis (Fig. 7) 30% of the coverage is used for training (marked in yellow as P1), 30% for validation (P2), and 40% for testing (P3). Training, validation and testing data are marked in yellow as P1, P2, and P3, respectively. For SVenice (Fig. 8), 50% of the coverage is set as training as it contains a large area of water, belonging to the class “others” (V1, marked in blue), while 15% is used for validation (V2) and 35% (V3) for testing. The footprints of the images with a GSD of 50 cm are larger with respect to the ones of 30 cm, namely 4500×3560 pixels (SParis) and 5600×5600 pixels (SVenice). The splitting boundaries of the 50 cm datasets are the same as the ones in the 30 cm datasets. In Fig. 7 and Fig. 8, P4/V4, P5/V5 and P6/V6 represent respectively training, validation and testing areas for the 50 cm datasets.

The released version of SMARS includes the above-mentioned rasters all in GeoTIFF format. Optical images are stored in three Band (RGB) uint8 format, DSMs with float precision and ground truth maps/masks with discrete values. The released version includes 9.0 GB of GeoTIFF data, covering the original rasters and split training, validation, and test tiles. According to our splitting approach, each

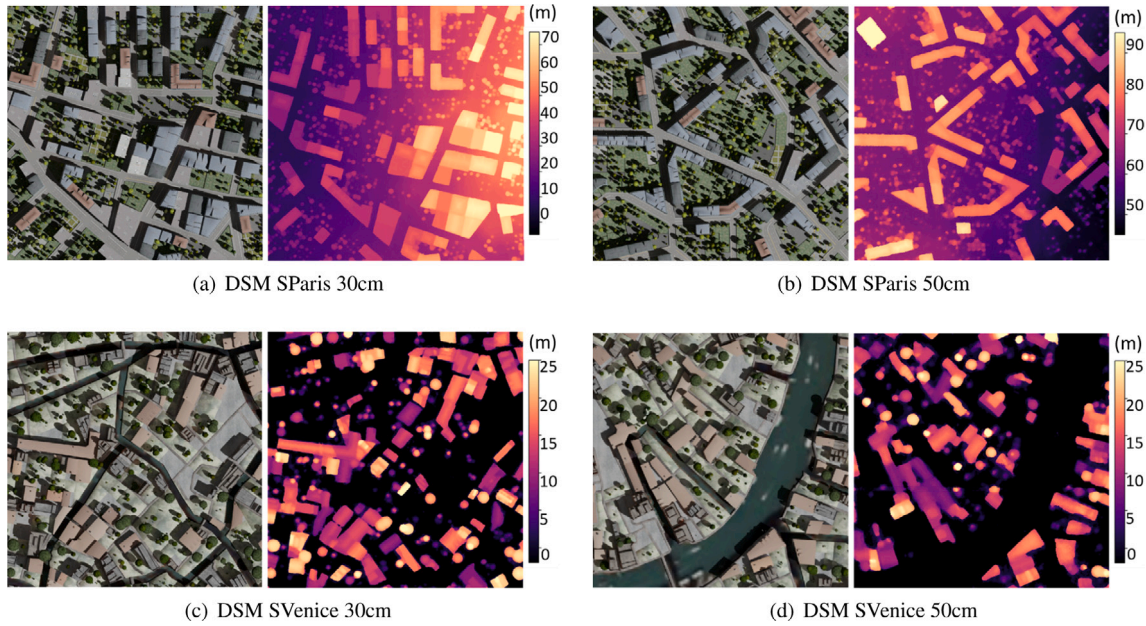


Fig. 5. Example regions of the DSMs generated for SMARS besides the paired orthorectified images. All samples are taken from the pre-event models. Elevation scale for the DSM is in meters.

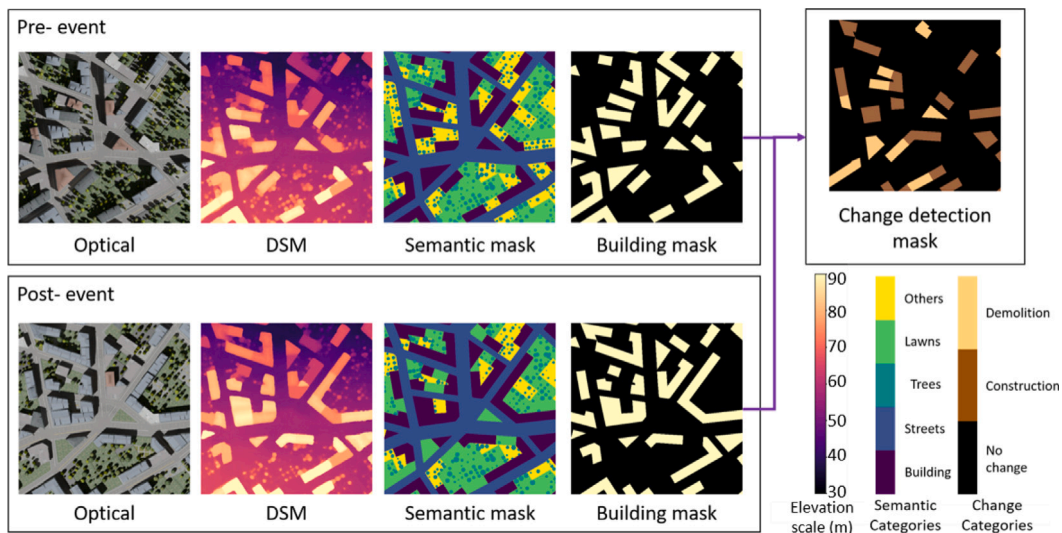


Fig. 6. Available information for each tile in pre and post-events scenarios. For each case, an optical image, a DSM and semantic and building masks are included. For the change detection, the difference between the two events is used for the ground truth mask. Scales are given as a reference for displayed information. The elevation scale for the DSM is in meters.

city_GSD data group consists of 27 tiles (6 pre-/post-event optical image tiles, 6 pre-/post-event DSMs, 6 pre-/post-event building masks, 6 pre-/post-event semantic maps, and 3 pre-/post-event building change detection masks). With four city_GSD combinations, there are 108 tiles in total. To make it easier for users to start with this data, a Python tool for patch cropping is included in the release version. The default training patches in our work have a size of 512×512 pixels with 50% overlapping, but users can customize training and validation patches as required. In addition, the DSM rasters can be converted to point cloud formats with another released Python tool, so users can use SMARS data with point cloud building extraction/semantic segmentation networks directly.

We employ the pre-event version with a GSD of 30 cm in the building extraction and 5-class semantic segmentation test design. In order to better visualize the testing results in this paper, the test region of each dataset is split into two regions, I and II (Fig. 9).

4.2. Data quality evaluation design

The proposed SMARS dataset focuses on building extraction, semantic segmentation, and 3D change detection. The building types and distributions of SParis and Venice are distinct and resemble those of the corresponding real cities. In addition, the building blocks of Venice are often separated by water channels instead of roads. The distinct features between the SParis and SVenice data result in large domain gaps for learning tasks, making SMARS a feasible data source for domain adaptation tests.

We experiment with state-of-the-art deep learning neural networks on the SMARS dataset for three tasks: (1) building extraction, (2) multi-class semantic segmentation, and (3) building change detection.

As buildings are dense in the scenes and resemble the architecture of real cities, the SMARS samples are an adequate input for building extraction and multi-class semantic segmentation tasks. Several effective

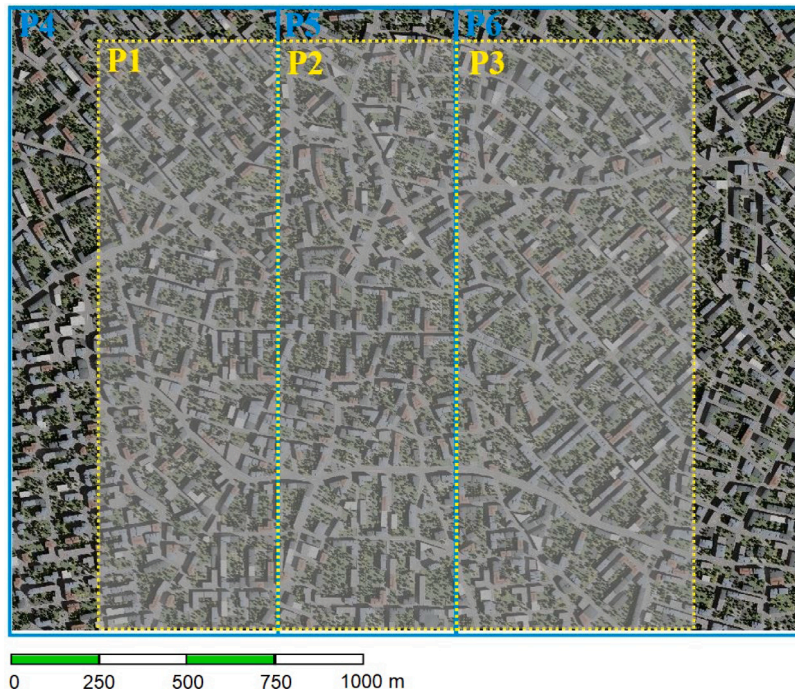


Fig. 7. Layout of the SParis images. Yellow dotted lines represent the splitting of the 30 cm resolution dataset (1.68 km by 1.68 km). Blue solid lines represent the splitting of the 50 cm resolution images (2.25 km by 1.78 km). (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

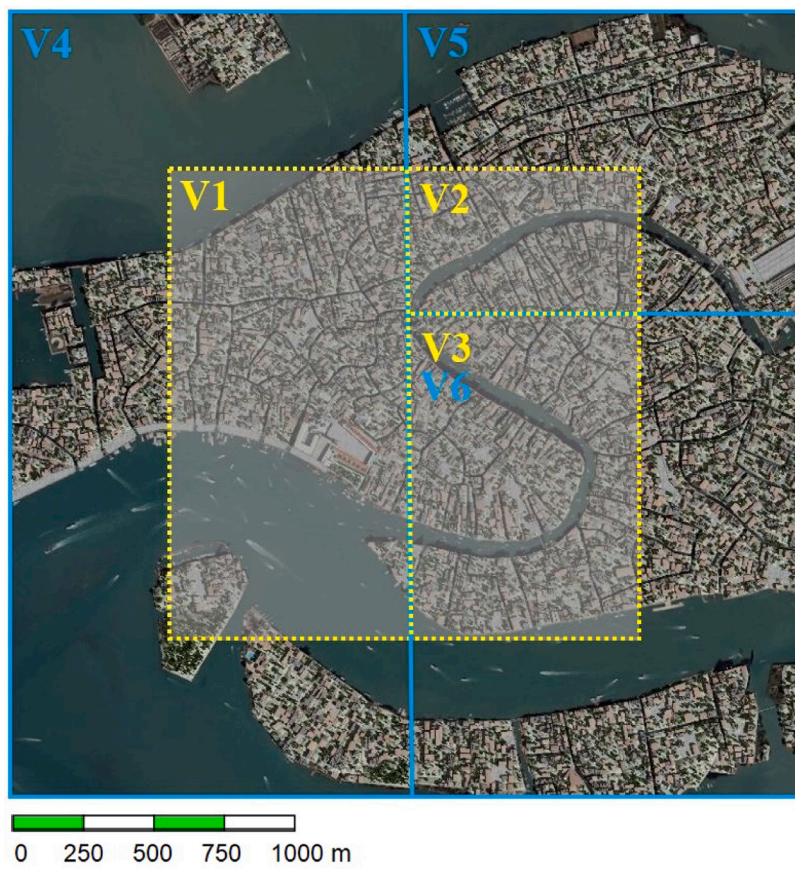


Fig. 8. Layout of the SVenice images. Yellow dotted lines represent the splitting of the 30 cm resolution dataset (1.68 km by 1.68 km). Blue solid lines represent the splitting of the 50 cm resolution images (2.8 km by 2.8 km). (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

deep learning approaches are available for these tasks. For the first two tasks, we work on two situations. The first is the single domain test with the provided train/Val/Test data from each synthetic city separately. In addition, we perform synthetic data cross-domain experiments by using SParis and SVenice separately for training, and test on the other model. Finally, we evaluate the predictions of samples from real sensors in the building segmentation task, which represents the most interesting experiment. In this case, we take samples from the Potsdam dataset for testing. We use as input either the images or the point clouds, which are addressed by 2D and 3D approaches respectively. Aspects to be studied include the correct detection and completeness of the buildings, as well as the transferability to previously unseen data.

Considering that the data are rendered with different semantic classes (*buildings, streets, trees, lawns and others*), we assess the performance of different neural networks using both 2D and 3D data, relying respectively on the images and their associated DSMs. Samples from both models have been generated with the same classes, enabling both single and cross-domain strategies to be tested. As the scenes are based on two different architectures, we expect some difficulties in the cross-domain case. Unfortunately, these experiments cannot be applied to real data due to the incompatibility of the available classes. Apart from the usual metrics such as Jaccard score (intersection over union (IoU)) and accuracy, we investigate the effect of the different nature of the data in relation to the obtained results.

The third task, change detection, is a key aspect to evaluate as the virtual scenes are constructed in order to simulate changes caused by city growth. The objective of this task is to localize the regions where the landscape has a significant change, whether because of new constructions or demolitions of buildings. The quality of the processed DSM plays a relevant role in the performance of this task, therefore we expect a difference in performance for the two cities, where the heights and space between buildings are significantly different. A comprehensive analysis based on the results highlights which approaches performed better on SMARS, and the approaches yielding a superior performance. As there are no unanimously accepted deep-learning based 3D change detection approaches available, we apply machine learning based approaches and are not able in this case to evaluate the transferability as in the previous tasks.

The following sections describe how the applied algorithms have been adapted for our experiments, the metrics to assess the performance on the different tasks, and a discussion of the capabilities and constraints of our dataset.

5. Building extraction

To examine the similarities between the SParis and SVenice datasets, and the domain gaps between the subsets of SMARS data and the real multimodal data in a deep learning context, we conduct building extraction experiments using different combinations of training and testing data, as detailed below:

- SMARS-to-SMARS single domain test
 - SParis→SParis: SParis used for training, SParis for testing
 - SVenice→SVenice: SVenice used for training, SVenice for testing
- SMARS-to-SMARS cross domain test
 - SParis→SVenice: SParis used for training, SVenice for testing
 - SVenice→SParis: SVenice used for training, SParis for testing
- SMARS-to-real cross domain test
 - SParis→Potsdam: SParis used for training, ISPRS Potsdam for testing
 - SVenice→Potsdam: SVenice used for training, ISPRS Potsdam for testing

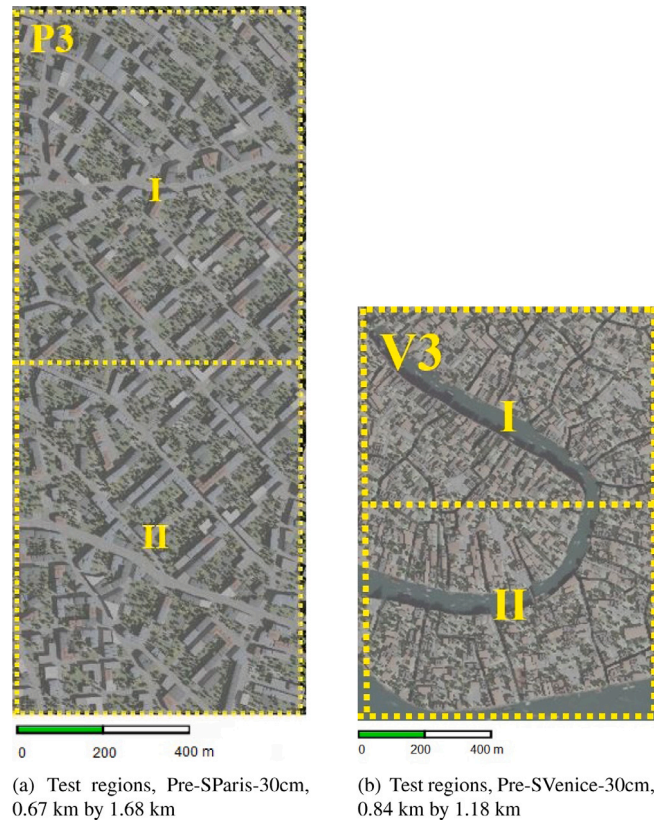


Fig. 9. The test regions of the 30 cm datasets.

- Potsdam→Potsdam: ISPRS Potsdam used for training, ISPRS Potsdam for testing (reference)

In order to assess the building extraction task from optical images, we report results obtained by applying the state-of-the-art Swin Transformer (Liu et al., 2021). We also employ the widely-used point cloud network SparseConvNet (Graham et al., 2018) to investigate the domain gaps between DSMs. Point cloud networks are proven to have a reasonable performance in urban scenes (Xie et al., 2020), even in the semantic segmentation task of photogrammetric point clouds (Bachhofner et al., 2020) or DSMs (Xie et al., 2023). We downsample the resolution of both optical imagery and DSM-derived point clouds from the Potsdam dataset from 30 cm from 5 cm in order to reduce the impact of differences in spatial resolution on the results.

5.1. Single domain test: 2D data

In order to verify whether deep learning methods can be applied to the SMARS data for remote sensing tasks such as building extraction from earth observation data, we train the Swin Transformer with the optical images of SParis and SVenice separately, using the data split described in chapter Section 4.2, and test the models with the corresponding test sets. Results are listed in Table 1, rows 1 and 3. The segmentation results are reported in Figs. 10 and 11(a) and (b). Within the same dataset, the building extraction IoUs of SParis and SVenice are above 95% and 92% respectively, indicating very satisfactory results. We can conclude that the synthetic data can be used for remote sensing tasks with deep learning approaches, yielding results similar to the ones obtained using real data.

5.2. Cross domain test: SMARS-SMARS 2D data

In order to investigate domain shifts between the two synthetic datasets, and how these affect in turn the downstream task of building

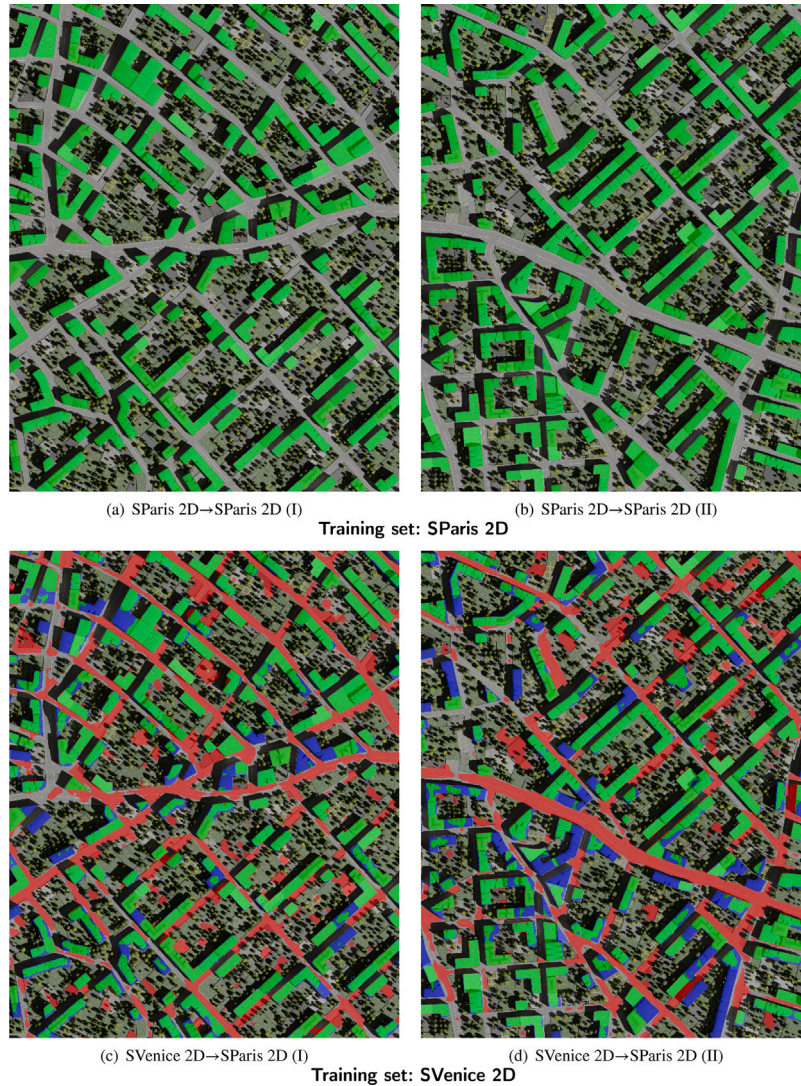


Fig. 10. The image building extraction results of SVenice: (a) and (b) Swin Transformer trained on SParis; (c) and (d) Swin Transformer trained on SVenice. Legend: True Positive False Positive False Negative. True Negative is not displayed.

Table 1
SMARS optical imagery building extraction results.

| Train | Test | Precision [%] | Recall [%] | F1 score [%] | IoU [%] |
|---------|---------|---------------|------------|--------------|---------|
| SParis | SParis | 97.27 | 98.38 | 97.82 | 95.73 |
| SParis | SVenice | 65.62 | 81.89 | 72.86 | 57.30 |
| SVenice | SVenice | 95.92 | 95.84 | 95.88 | 92.09 |
| SVenice | SParis | 47.28 | 88.69 | 61.68 | 44.59 |

extraction, we test the Swin Transformer trained with one of the two sets on the other one, according to the data split described in chapter Section 4.2. Results are presented in Table 1, rows 2 and 4. The segmentation results are reported in Figs. 10 and 11(c) and (d). With respect to the results presented in Section 5.1, the building IoU scores are significantly degraded, from 95% and 92% to 57.37% and 44.59%, respectively. The decrease in performance can be attributed to large domain shifts, as evidenced by the distinct architectural styles, street appearance and ground features of the two scenes. The decrease in performance when training and testing data have distinct distributions can also be observed in real remote sensing data.

5.3. Cross domain test: SMARS-real 2D data

In order verify the suitability of employing a synthetic dataset to assess algorithms to be applied to real data, we test our network trained

with SMARS with the ISPRS Potsdam data (for brevity named Potsdam thereafter). In addition, we apply the CIELAB color space transformation (He et al., 2021) to the SMARS 2D data in order to reduce the domain gaps between the synthetic and real datasets. Adopting similar workflow and settings as in our previous work (Li et al., 2022), we select 10 images from the Potsdam data to be used as reference, and transform the SParis and SVenice data to the Potsdam data in the LAB color space (LAB), and then convert SParis and SVenice back to RGB colorspace. Quantitative results are listed in Table 2. The result of Potsdam to Potsdam is listed in the last row for reference. Surprisingly, the test results on Potsdam data yield better performance than the SParis/SVenice cross domain experiments. This can be explained by the fact that the buildings in Potsdam are more similar to SParis than SVenice in terms of their structure and appearance. The CIELAB transformation does not lead to consistent performance changes. For the SParis trained model, the IoU score of building extraction in Potsdam dataset increases 4%, while for SVenice trained model decreases over 3%. Another performance discrepancy is observed in the relationship between precision and recall. For model trained on SParis, precision is significantly lower than recall, while the opposite is observed for model trained on SVenice. Results of building extraction in Potsdam is shown in Fig. 12. In spite of being far from perfect for the Potsdam dataset, the majority of buildings is correctly extracted, suggesting that

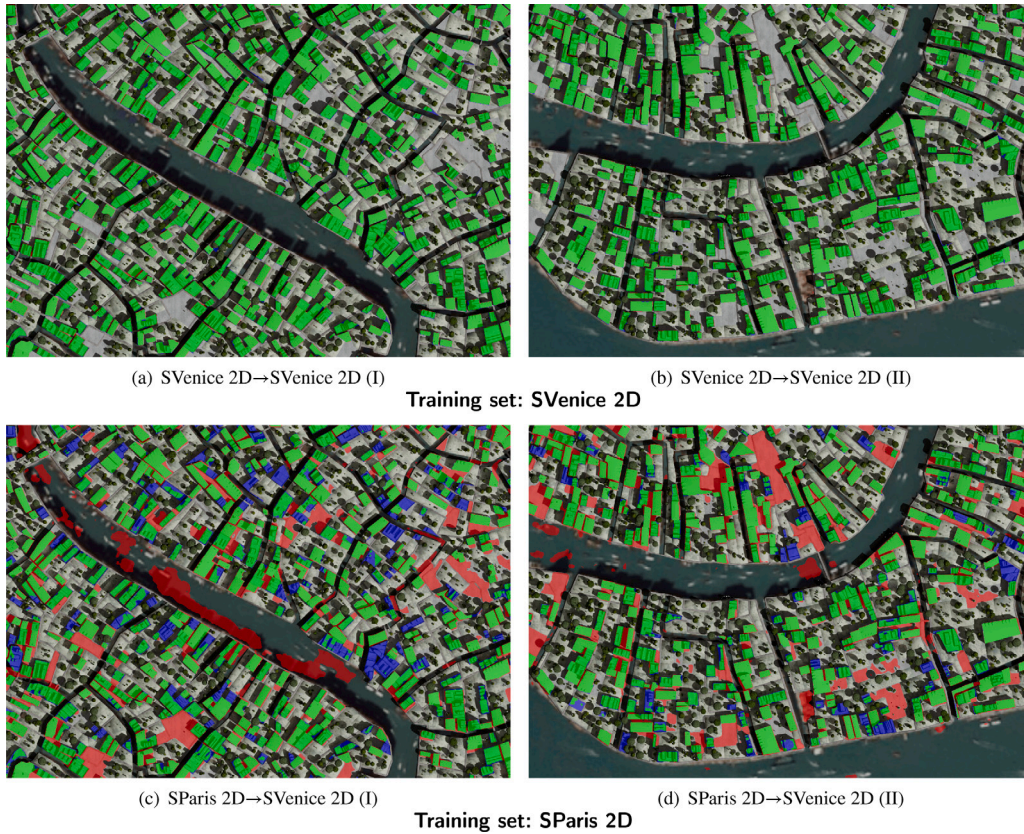


Fig. 11. The image building extraction results of SVenice: (a) and (b) Swin Transformer trained on SVenice; (c) and (d) Swin Transformer trained on SParis. Legend: ■ True Positive ■ False Positive ■ False Negative. True Negative is not displayed.

simulated optical data can be suited to train a neural network for building extraction and other tasks employing real earth observation data. To further validate the suitability of the SMARS as training data for building extraction, we include the cross-domain test results of Potsdam reported by Peng et al. (2022), which address the difficulties in cross-domain building extraction. The results are shown in Table 2. Two real datasets, namely WHU (Ji et al., 2019) and MASS (Mnih, 2013) are used as source domain data for training. Without any domain adaptation strategy (denoted w/o DA), the models trained with SMARS significantly outperform those trained with WHU and MASS datasets notwithstanding that they are real data. In the same work by Peng et al. (2022), a unsupervised domain adaptation method named FDANet was proposed, which consists of Wallis filter, adversarial learning and consistency regularization to tackle domain shift. Nevertheless, our model trained with CIELAB-transformed-SParis data outperforms FDANet trained with MASS data, further demonstrating the potential of SMARS as training data. In addition, the result of intra-domain experiment that used Potsdam as training data is listed in the last row of Table 2.

5.4. Single domain test: 3D data

As mentioned above, we also employ the point cloud network SparseConvNet (Graham et al., 2018) as a reference in order to examine the quality of the DSMs. The quantitative results of single-domain building extraction from DSM-based point clouds are listed in Table 3, rows 1 and 3. The classification results are presented in Fig. 13(a) and (b), and Fig. 14(a) and (b). The IoU scores of SParis→SParis and SVenice→SVenice are 95.16% and 91.03%, respectively, which are slightly inferior to the results obtained by the Swin Transformer with the simulated optical imagery but still satisfactory. Based on the evaluation metrics and visual quality, the synthetic data can be

Table 2

2D cross-domain study, row 1–4: SMARS and ISPRS Potsdam as training and testing sets, respectively; row 5–8: WHU and MASS as training data, and Potsdam as testing data from Peng et al. (2022), where ‘w/o DA’ denotes without domain adaptation and FDANet is described in Section 5.3; the last row: Potsdam as training and testing.

| Train | Test | Precision [%] | Recall [%] | F1 score [%] | IoU [%] |
|------------------|---------|---------------|------------|--------------|---------|
| SParis | Potsdam | 69.47 | 84.57 | 76.28 | 61.65 |
| SParis (CIELAB) | Potsdam | 73.18 | 86.70 | 79.37 | 65.79 |
| SVenice | Potsdam | 81.68 | 73.37 | 77.30 | 63.00 |
| SVenice (CIELAB) | Potsdam | 78.28 | 71.44 | 74.68 | 59.59 |
| WHU (w/o DA) | Potsdam | – | – | 68.83 | 52.47 |
| WHU (FDANet) | Potsdam | – | – | 88.87 | 79.96 |
| MASS (w/o DA) | Potsdam | – | – | 39.05 | 24.26 |
| MASS (FDANet) | Potsdam | – | – | 78.63 | 64.78 |
| Potsdam | Potsdam | 94.45 | 95.30 | 94.88 | 90.25 |

considered a valid substitute or integration for the training of deep networks for the considered tasks, whenever sufficient annotated real earth observation data are not available.

5.5. Cross domain test: SMARS-SMARS 3D data

Using a similar workflow as described in Section 5.2, we carry out experiments SParis→SVenice and SVenice→SParis by integrating the DSM-based point clouds, in order to investigate domain shifts between the two synthetic DSMs. Rows 2 and 4 of Table 3 list the quantitative results. Compared to the single-domain case, the score of each metric decreases. In the experiment of SVenice→SParis, precision, F1, and IoU decrease 2.55%, 1.1%, and 2.07% compared with the results of SParis→SParis, respectively. Such decreases in performance appear to be acceptable. According to the qualitative results shown in Fig. 13(c) and (d), the SparseConvNet model trained on the SVenice data correctly



Fig. 12. SMARS 2D→Potsdam results, trained respectively with SParis(a), SParis-CIELAB(b), SVenice(d), and SVenice-CIELAB(e). Legend: True Positive False Positive False Negative. True Negative is not displayed. A detailed view of the area misclassified as buildings by all models is shown in (f). The area is highlighted in (a)-(e) with a white rectangle.

Table 3
SMARS building extraction using DSM-derived point clouds as the input.

| Train | Test | Precision [%] | Recall [%] | F1 score [%] | IoU [%] |
|---------|---------|---------------|------------|--------------|---------|
| SParis | SParis | 97.74 | 97.29 | 97.52 | 95.16 |
| SParis | SVenice | 86.13 | 85.01 | 85.57 | 74.78 |
| SVenice | SVenice | 95.91 | 94.70 | 95.30 | 91.03 |
| SVenice | SParis | 95.19 | 97.68 | 96.42 | 93.09 |

covers all building objects. Compared with the building masks generated by the model of SParis→SParis, it contains more false negative pixels on several building instances. For the SParis→SVenice case, precision, recall, F1, and IoU decrease 11.61%, 12.28%, 11.95%, and 20.38% compared with the results of SVenice→SVenice, respectively. The predicted building masks exhibit non-negligible noise (Fig. 14(c) and (d)). Several pixels belonging to other classes which are adjacent to buildings are here misclassified as buildings, with the same happening for some pixels belonging to the water semantic class. This phenomenon can be explained by two factors. Firstly, the majority of buildings in the SVenice dataset are smaller with respect to the ones contained in SParis. Consequently, the SparseConvNet model trained on SParis fails at recognizing them. Secondly, the water class is not present in SParis. As a result, several flat water areas in SVenice’s DSMs are more easily misidentified as rooftops by the point cloud building extraction network trained on SParis data.

5.6. Cross domain test: SMARS-real 3D data

As illustrated by the qualitative results in Fig. 15, models trained with synthetic data achieve reasonable performance on the ISPRS Potsdam dataset when inspected visually. Nevertheless, partial building structures, which are seldom found in synthetic data, are often not

Table 4
3D cross-domain study, with SMARS and ISPRS Potsdam datasets as training and testing set, respectively.

| Train | Test | Precision [%] | Recall [%] | F1 score [%] | IoU [%] |
|---------|---------|---------------|------------|--------------|---------|
| SParis | Potsdam | 67.60 | 89.50 | 77.02 | 62.63 |
| SVenice | Potsdam | 80.58 | 88.00 | 84.13 | 72.61 |
| Potsdam | Potsdam | 93.75 | 92.54 | 93.14 | 87.17 |

detected, such as four quadrilateral building clusters in Fig. 15(b). In Fig. 15(c), such errors appear considerably reduced. Table 4 shows that the IoU and F1 scores for SVenice→Potsdam are 9.98% and 7.11% higher than those for SParis→Potsdam, respectively. This indicates that the SparseConvNet trained on SVenice has better generalization capabilities on real data with respect to the model trained on SParis. However, when compared to the reference results of Potsdam→Potsdam, both models trained on synthetic data still yield a decreased performance. Results further show a good capability for transfer learning, especially for SVenice. This could also be used as a step for pre-training neural networks, and supplementing it with a few additional samples for fine-tuning might alleviate the domain gap. Additionally, including a larger variety of building models in the training data might help to correctly identify some missing shapes.

6. Multi-class semantic segmentation

In order to assess the performance in semantic classes different from buildings, we carry out multi-class semantic segmentation on the 2D optical and point cloud data, with both single- and cross-domain tests.



Fig. 13. Building extraction results of SParis test data using DSM-derived point clouds as the input: (a) and (b) SparseConvNet trained on SParis; (c) and (d) SparseConvNet trained on SVenice. Legend: True Positive False Positive False Negative. True Negative is not displayed.

6.1. 2D multi-class semantic segmentation

The SwinTransformer is here trained on SParis and SVenice using all 5 semantic classes. Quantitative results are listed in Table 5, while segmentation maps are reported in Fig. 16. For the model trained on SParis, all classes except *trees* achieve IoU over 90% on the SParis test set; however, when tested with the SVenice test set, the performance significantly decreases, with the exception of the *trees* class. This indicates a large domain gap between the two datasets, especially regarding *buildings*, *streets*, *lawns*, and *others*. On the contrary, *trees* in both datasets have relatively uniform appearance, thing which can explain the comparatively smaller performance degradation in the cross-domain setting. In the SVenice→SVenice results, the lowest accuracy and IoU scores are associated to the class *streets*: probably, this can be due to the small number of instances of the classtreets in SVenice, as well as to their different structure. Interestingly, the *lawns* class appears to be the least affected in the SVenice→SParis experiment, while the exact opposite happens for the SVenice→SParis results. Fig. 17(e) and (f) show the river (belonging to the *others* class) as being mostly misclassified as *street*, due to the absence of water in the SParis dataset; meanwhile, the majority of *streets* and *lawns* are misclassified as *others*, as their structure is different in the SParis dataset.

Table 5

Transferability study of SMARS optical images, 5 classes.

| Train | Test | | Building | Street | Tree | Lawns | Other | Mean |
|---------|---------|---------|----------|--------|-------|-------|-------|-------|
| SParis | SParis | IoU [%] | 96.07 | 96.80 | 85.49 | 92.97 | 92.70 | 92.81 |
| | | Acc [%] | 97.90 | 98.54 | 93.06 | 96.11 | 95.88 | 96.30 |
| SParis | SVenice | IoU [%] | 45.37 | 0.37 | 83.45 | 0.01 | 13.58 | 28.56 |
| | | Acc [%] | 72.80 | 0.66 | 92.62 | 0.02 | 29.84 | 39.19 |
| SVenice | SVenice | IoU [%] | 86.55 | 56.68 | 78.95 | 87.19 | 87.85 | 79.45 |
| | | Acc [%] | 94.87 | 64.92 | 86.97 | 94.84 | 93.43 | 87.01 |
| SVenice | SParis | IoU [%] | 15.67 | 10.16 | 54.06 | 66.73 | 17.41 | 32.81 |
| | | Acc [%] | 20.02 | 10.23 | 66.22 | 87.92 | 54.51 | 47.78 |

6.2. 3D multi-class semantic segmentation

In these experiments we train the model including SParis and SVenice DSMs with the described 5 semantic classes using SparseConvNet. Table 6 reports a quantitative assessment of the results for SParis→SParis, SParis→SVenice, SVenice→SVenice, and SVenice→SParis models. In the two experiments having the same source for training and test data, namely SParis→SParis and SVenice→SVenice, SparseConvNet achieves a satisfactory performance for the classes *buildings* and *trees*. SParis→SParis exhibits clearly superior results with



Fig. 14. Building extraction results of SVenice test data using DSM-derived point clouds as the input: (a) and (b) SparseConvNet trained with SVenice data. (c) and (d) SparseConvNet trained with SParis data. Legend: True Positive False Positive False Negative. True Negative is not displayed.

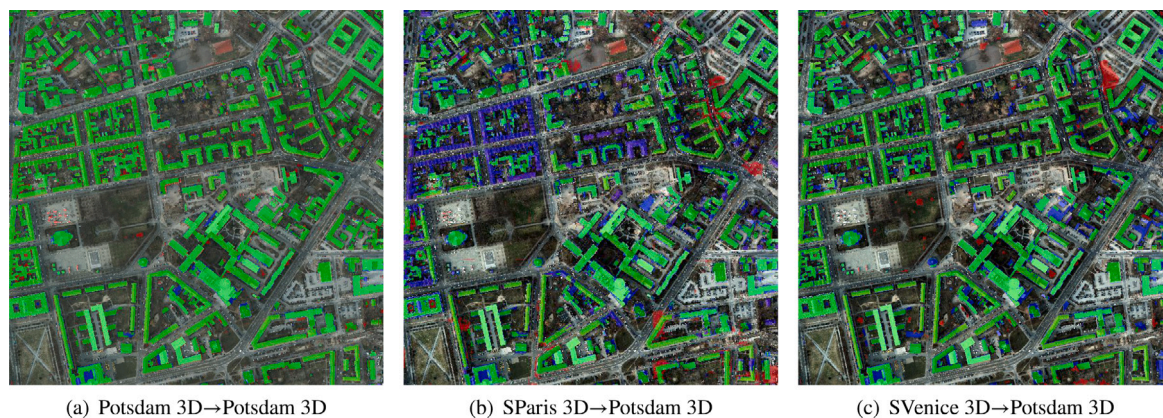


Fig. 15. Building extraction results of ISPRS Potsdam data using DSM-derived point clouds as the input. (a) SparseConvNet trained on ISPRS Potsdam. (b) SparseConvNet trained on SParis. (c) SparseConvNet trained on SVenice. Legend: True Positive False Positive False Negative. True Negative is not displayed.

Table 6
Transferability study of SMARS DSM-derived point clouds, 5 class.

| Train | Test | | Building | Street | Tree | Lawns | Other | Mean |
|---------|---------|---------|----------|--------|-------|-------|-------|-------|
| SParis | SParis | IoU [%] | 94.85 | 90.00 | 78.66 | 46.39 | 47.48 | 71.48 |
| | | Acc [%] | 96.99 | 96.98 | 83.26 | 58.59 | 69.66 | 81.10 |
| SParis | SVenice | IoU [%] | 72.81 | 9.26 | 37.55 | 33.78 | 2.33 | 31.15 |
| | | Acc [%] | 83.85 | 48.23 | 54.55 | 44.23 | 2.77 | 46.73 |
| SVenice | SVenice | IoU [%] | 90.04 | 25.71 | 80.20 | 62.89 | 69.09 | 65.59 |
| | | Acc [%] | 97.54 | 38.64 | 87.45 | 83.26 | 76.15 | 76.61 |
| SVenice | SParis | IoU [%] | 93.19 | 4.54 | 75.30 | 39.45 | 4.07 | 43.31 |
| | | Acc [%] | 96.48 | 4.75 | 86.06 | 84.54 | 8.43 | 56.05 |

respect to SVenice→SVenice for the class *streets*. As mentioned, this is due to the limited number of samples for this class available for training in SVenice. In the cross-domain experiment SParis→SVenice, the performance of each class decreases severely. Among the results of SVenice→SParis, the IoU and accuracy scores for the class *buildings* are excellent, and comparable to the scores achieved in SParis→SParis. This is in line with the results presented for SVenice→SParis building extraction in Section 5.5, as SVenice features a wide variety of building sizes covering most of their variability for the respective class in SParis data. The generalization capability of recognizing buildings is preserved in the 5-class semantic segmentation point cloud model. The visual assessment of the results presented in Fig. 18 suggests that the model trained with SVenice is not optimal to recognize streets in the DSMs of SParis. In Fig. 19, most of the areas covered by water are classified

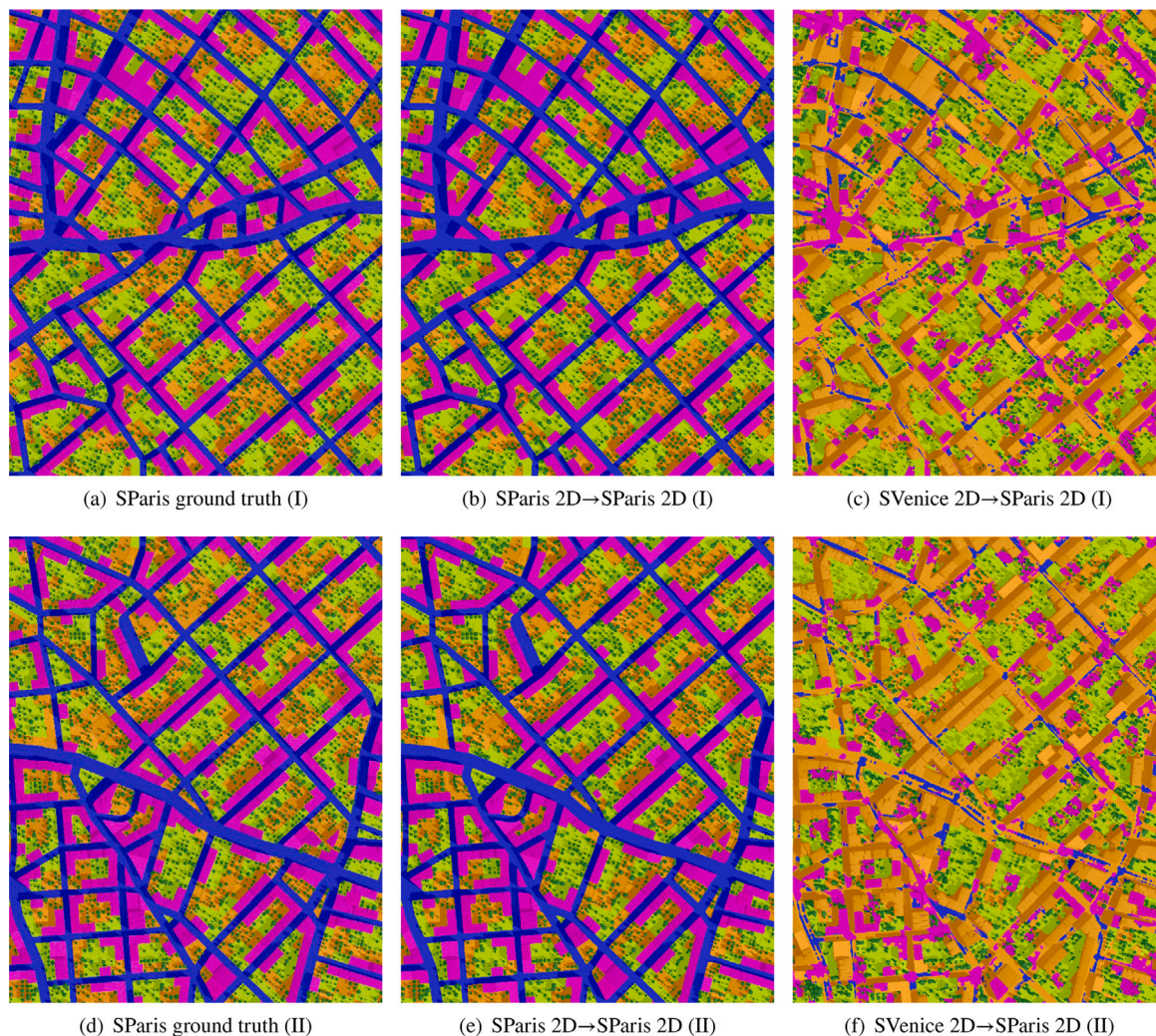


Fig. 16. Results of image semantic segmentation for SParis (5 classes). Legend: ■ Buildings ■ Street ■ Trees ■ Lawns ■ Other. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

as streets in SParis→SVenice. This is because SParis lacks training data for this class, as discussed in Section 5.5.

7. Building change detection

In order to assess the feasibility of SMARS for 3D change detection applications, in this section change indicators from both 2D and 3D data are extracted and evaluated. In addition, we present several state of the art change detection approaches for comparison (Tian and Dezert, 2019).

7.1. Robust height differences

As detailed in our previous work (Tian et al., 2013), the quality of the pre- and post-event DSMs can exhibit relevant differences according to GSD, sensors characteristics, illumination conditions, stereo viewing angles and other parameters of the multi-view images from which the DSMs are generated. Hence, methods based on pixel-based subtraction do not in all cases deliver ideal results (Tian et al., 2013; Qin et al., 2016). Thus, robust distance measurements yielding a refined height change indicator have been proposed. The main motivation of the experiments reported in this section is assessing the differences between DSMs generated from synthetic and real data, along with their impact on practical applications. We compare the robust height differences proposed in Tian et al. (2013) (window size set to $w = 5$) to the use

of direct height difference (considering only positive height changes). In addition, the pre- and post-event images are “acquired” with similar settings by the virtual camera, such as GSD and different illumination conditions, lowering the impact of the sources of errors when using methods based on direct subtraction of the DSMs. Nevertheless, in Fig. 20, results obtained by applying robust height differences appear superior, as they exhibit reduced noise in the building boundary regions.

7.2. Building change mask generation

In order to further assess the quality of the proposed data for 2D and 3D change detection applications, extended experiments with different change detection approaches are summarized in this section. In this paper, we test direct height differences with threshold values manually and automatically selected to generate positive building change masks for the test regions. In addition, 2D change detection results are extracted and evaluated using the state of art Interactively Reweighted Multivariate Alteration Detection (IR-MAD) (Nielsen, 2007). For the case of fusion-based change detection approaches, we follow the method proposed in Tian and Dezert (2019), which employs the decision fusion model to combine the 2D and 3D change indicators. Three decision criteria are considered, including Maximum of Belief (MaxBel), Maximum of Plausibility (MaxPl) and Maximum of Betting Probability (MaxBetP). In order to calculate the Basis Belief

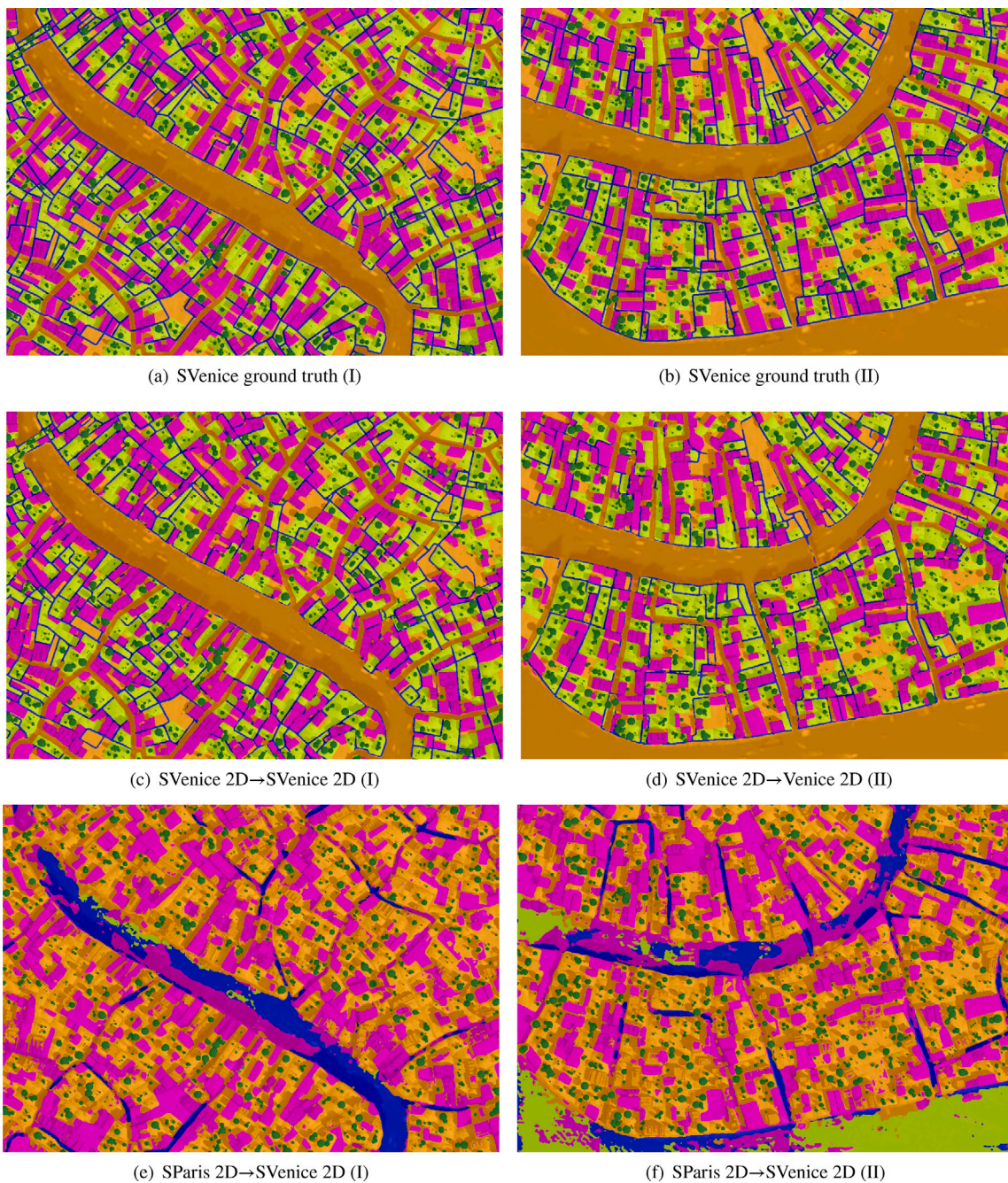


Fig. 17. Results of image semantic segmentation for SVenice (5 classes). Legend: Buildings Street Trees Lawns Other.

Assignments (BBAs) of the concordance and discordance indices, two thresholds are required. In our previous work (Tian and Dezert, 2019), we adopt an extension of Otsu thresholding to project the change indicators to a sigmoid distribution. As here the training data are provided by SMARS, we use them to automatically calculate the two thresholds for each change indicator, namely the mean value of the change indicators for each class (change (T_0), no-change(T_1)). We refer to this approach as automatic threshold values selection (AUTO), and set $T = \text{mean}(T_0, T_1)$ for the height differences and IR-MAD, separately.

The performance of the difference change detection approaches is evaluated based on overall accuracy (OA), kappa accuracy (KA) and IoU (Table 7). Each synthetic image has two test regions, which are marked as AOI (I) and AOI (II) in Table 7, respectively. SParis appears to be an easier test region, featuring mainly high-rise and well-separated buildings. In addition, the buildings are considerably higher

than most of the trees, introducing a relevant increase in height in the transitions from trees to buildings. Therefore, the direct height differences with automatic thresholding approach (Hdiff (AUTO)) achieve the best accuracy according to the figures of merit listed in Table 7. However, a visual assessment of Fig. 21 reveals that the decision fusion results present a reduced amount of false positives, especially around building boundary regions. Further details are reported in Fig. 22. The best results are achieved by directly comparing the two building masks derived from Section Section 5.5: we refer to this case as “Post-classification” in Table 7.

SVenice is a challenging test region for 3D change detection compared to SParis, as it features small-sized buildings and narrow streets. In addition, the trees are sometimes taller than nearby residential buildings, resulting in negative height changes for newly constructed buildings. This rarely occurs in the SParis dataset (see in Fig. 22).

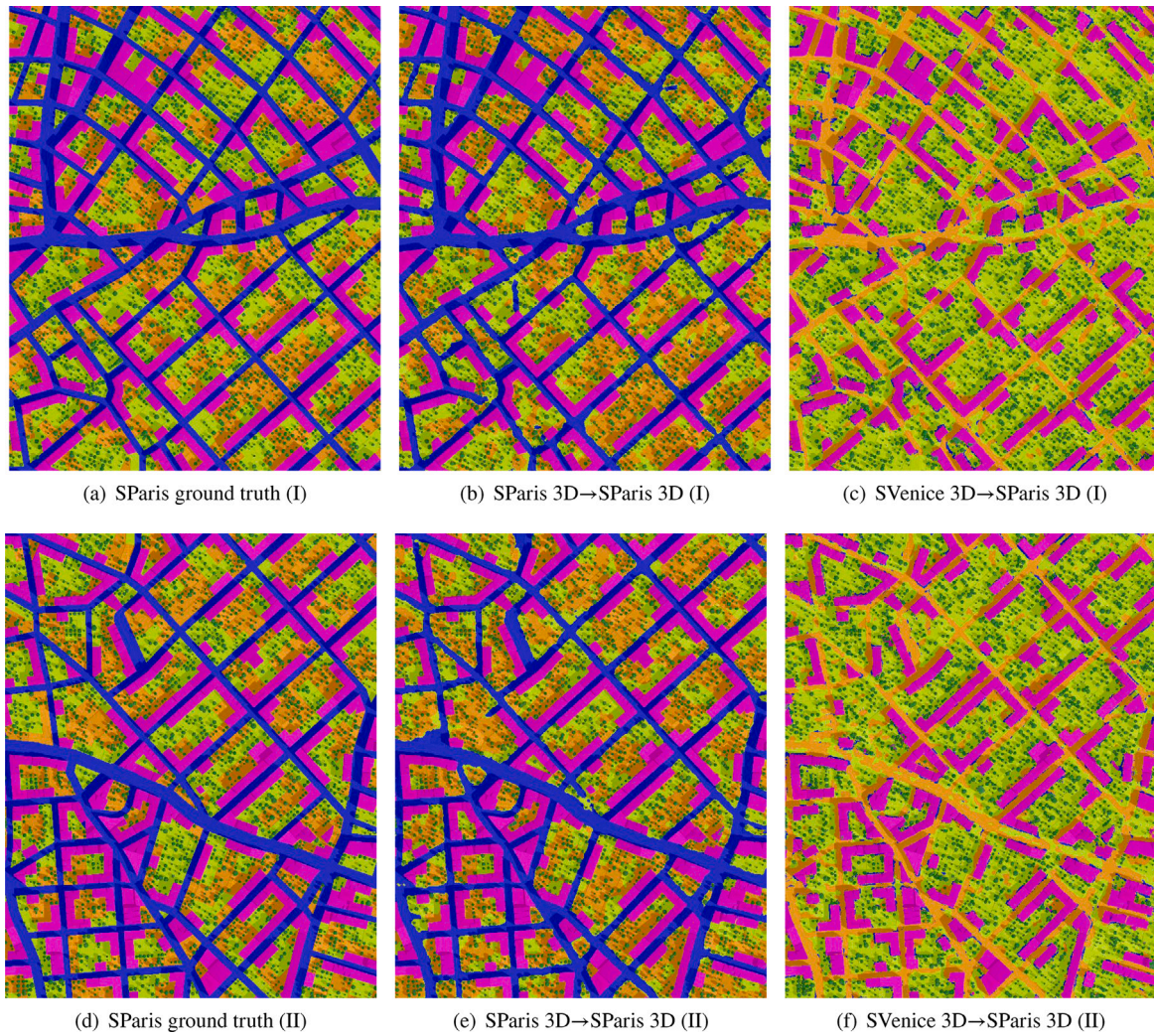


Fig. 18. 5-class semantic segmentation results of SParis test data using DSM-derived point clouds as the input. (a) and (d) Ground truth. (b) and (e) SparseConvNet trained with SParis data. (c) and (f) SparseConvNet trained with SVenice data. Legend: Buildings Street Trees Lawns Other.

Table 7
Results of different change detection approaches on SParis and SVenice.

| Test regions | Methods | AOI(I) | | | AOI(II) | | |
|--------------|-------------------------|--------|--------|---------|---------|--------|---------|
| | | OA [%] | KA [%] | IoU [%] | OA [%] | KA [%] | IoU [%] |
| SParis | HDiff > 5 m | 97.83 | 90.85 | 85.36 | 97.91 | 90.12 | 84.00 |
| | Robust HDiff (AUTO) | 98.41 | 92.84 | 88.24 | 98.49 | 92.34 | 87.25 |
| | Hdiff (AUTO) | 98.45 | 93.22 | 88.87 | 98.49 | 92.55 | 87.63 |
| | IR-MAD (AUTO) | 85.87 | 43.84 | 35.10 | 86.42 | 40.05 | 31.29 |
| | Decision Fusion-MaxBel | 98.08 | 91.07 | 85.47 | 98.13 | 90.07 | 83.67 |
| | Decision Fusion-MaxPl | 98.04 | 90.89 | 85.18 | 98.10 | 89.94 | 83.48 |
| | Decision Fusion-MaxBetP | 98.09 | 91.10 | 85.51 | 98.14 | 90.16 | 83.82 |
| | Region- DS-MaxBel | 91.46 | 67.15 | 56.29 | 91.54 | 62.50 | 50.66 |
| | Post-classification | 98.86 | 94.99 | 91.64 | 98.78 | 93.95 | 89.83 |
| SVenice | HDiff > 5 m | 93.30 | 77.26 | 68.54 | 94.30 | 76.47 | 66.37 |
| | Robust HDiff (AUTO) | 93.54 | 77.07 | 68.02 | 94.40 | 75.68 | 65.15 |
| | Hdiff (AUTO) | 93.19 | 77.02 | 68.13 | 94.24 | 76.39 | 66.31 |
| | IR-MAD (AUTO) | 85.90 | 36.26 | 27.27 | 87.36 | 32.74 | 24.19 |
| | Decision Fusion-MaxBel | 93.38 | 75.84 | 66.36 | 94.39 | 75.01 | 64.21 |
| | Decision Fusion-MaxPl | 93.39 | 75.84 | 66.36 | 94.36 | 74.81 | 63.98 |
| | Decision Fusion-MaxBetP | 93.37 | 75.80 | 66.32 | 94.37 | 74.89 | 64.07 |
| | Region- DS-MaxBel | 90.70 | 69.01 | 59.60 | 91.45 | 66.12 | 55.17 |
| | Post-classification | 96.94 | 89.53 | 84.14 | 97.57 | 90.03 | 84.24 |

Moreover, the *water* class occupying around 5% of each test region is not defined for this dataset. Differences between *water* and other semantic classes are particularly evident in the synthesized optical images, which were simulated relying on low-resolution satellite data.

The 2D change detection results have an associated IoU of 27.27% and 24.17% in the two test regions, respectively, confirming the impact of differences in illumination conditions between the pre- and post-event images on the final results. When applied on the SVenice data,

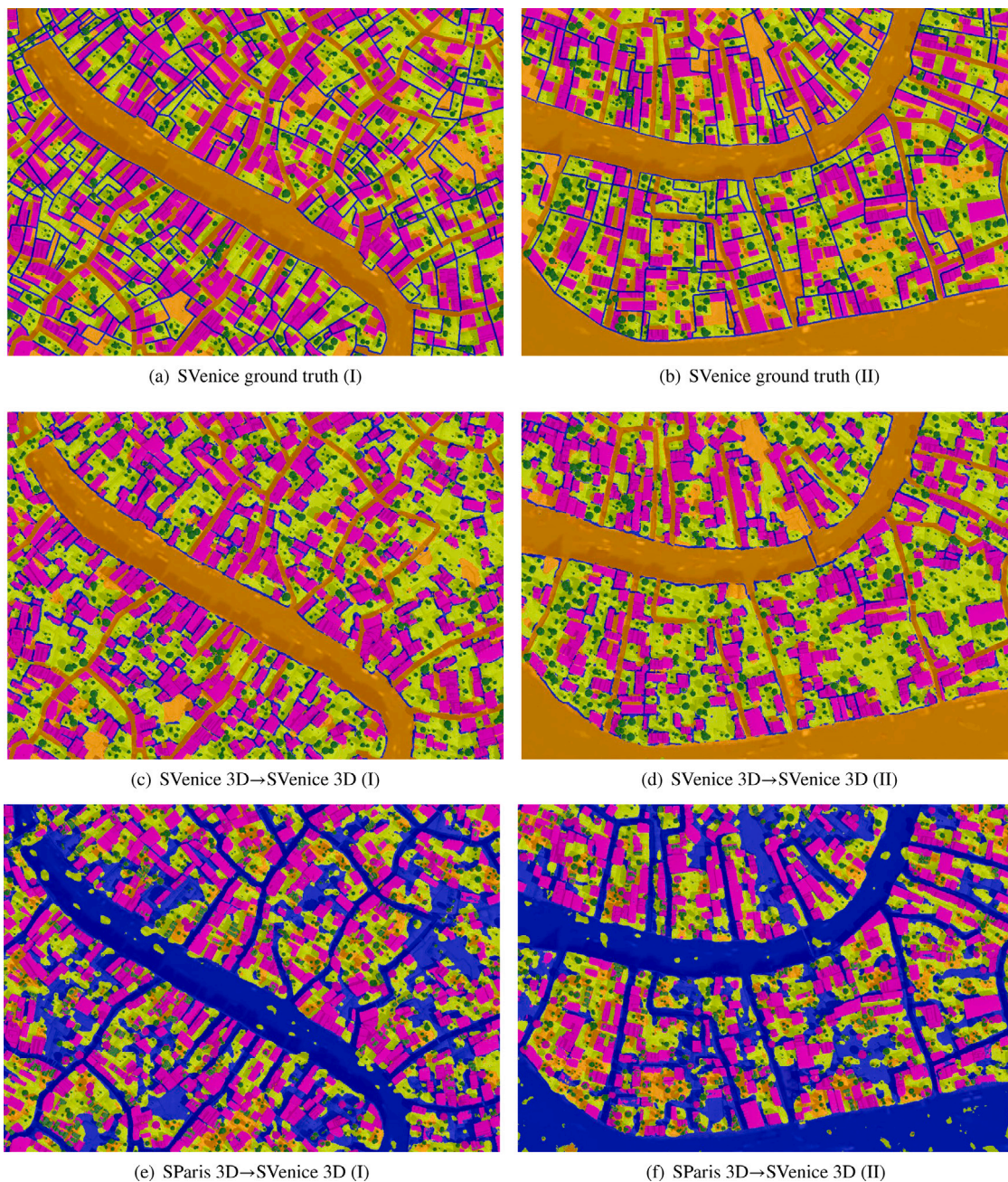


Fig. 19. 5-class semantic segmentation results of SVenice test data using DSM-derived point clouds as the input. (a) and (b) Ground truth. (c) and (d) SparseConvNet trained with SVenice data. (e) and (f) SparseConvNet trained with SParis data. Legend: Buildings Street Trees Lawns Other.

robust height differences achieve slightly higher accuracy with respect to fusion-based approaches. Nevertheless, Fig. 23 shows that both height differences and DS fusion detect regions as false negatives, if a tall tree is replaced by a building in the post-image. Additionally, a relevant number of new trees is detected as newly constructed buildings (highlighted in red in Fig. 23), as these match both conditions of having an increased height and exhibiting changes in the spectrum of the optical data. In a similar way to the experiments carried out on SParis, the differences in performance between the three decision approaches are not obvious. Relying on the accurate 2D/3D multimodal building detection result of section Section 5.5, post-classification clearly outperforms other approaches, achieving an IoU equal to 84.14% and 84.24% in the two test regions, respectively. The second test region of SVenice is presented in Fig. 24(b), in which most of the newly constructed buildings are correctly identified.

In order to reduce false negatives for newly constructed buildings, we test region-based 3D change detection by fusing the post-event building mask with the fusion-based change detection results. As all three DS fusion methods yield similar results, we only report results obtained with Decision Fusion-MaxBel for the following region-based change detection experiment. Buildings belonging to the post-event building mask are considered as newly constructed if more than 30% of their pixels belongs to the “building change” category in the pixel-based change detection results. The performance of the region-based change detection approach is rather poor for both SParis and SVenice, as shown in Table 7. This can be explained by examining Fig. 24, where a relevant number of newly constructed buildings are connected to the unchanged buildings in the virtually simulated environment. Therefore, a relevant number of both false positives and negatives are introduced when averaging the change decisions in these regions.

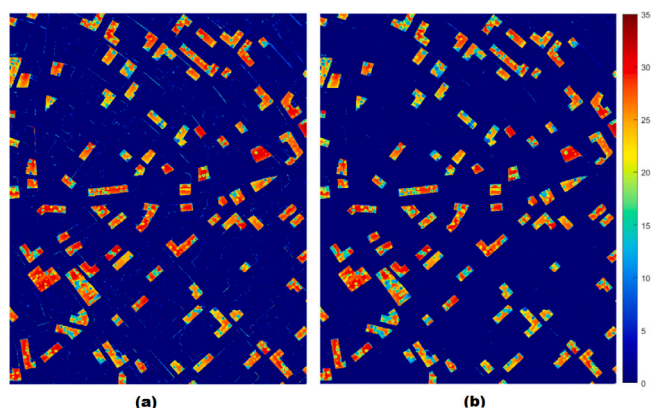


Fig. 20. Positive height differences: (a) direct subtraction; (b) Robust height differences with ($w = 5$).

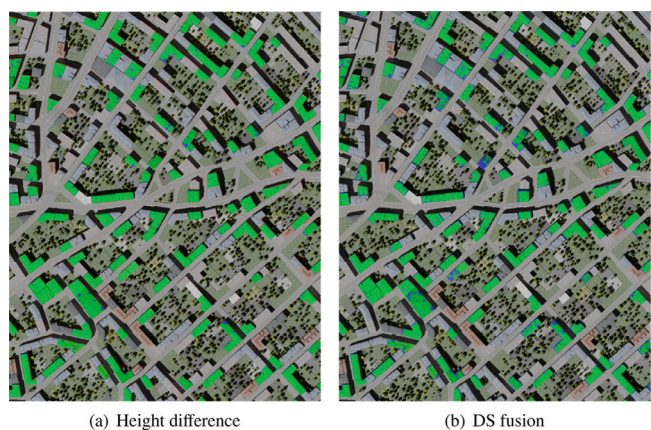


Fig. 21. 3D change detection results of SPariS (I) generated by direct height difference (a) and decision fusion (b). Legend: ■ True Positive ■ False Positive ■ False Negative. True Negative is not displayed.

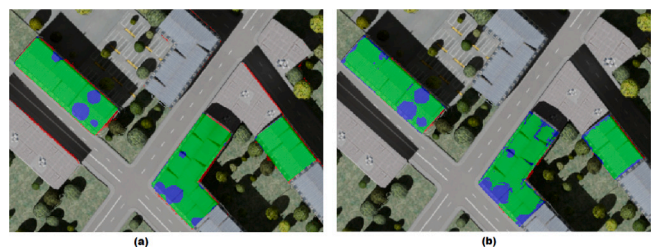


Fig. 22. Comparison of results obtained for single buildings: (a) direct height differences (b) DS fusion. Legend: ■ True Positive ■ False Positive ■ False Negative. True Negative is not displayed.

8. Discussion

This paper proposes a novel workflow for synthetic data generation filling the gaps in the available 2D/3D multimodal data for building extraction, multi-class semantic segmentation and 3D change detection. Our data analysis goes in two directions: (1) the feasibility of using SMARS to evaluate the efficiency of existing approaches for building extraction, multi-class semantic segmentation and building change detection and (2) the effects of the domain gap when the models trained on our synthetic data are tested on real data.

8.1. Quality of the synthetic dataset

This subsection discusses the main advantages and disadvantages of the rendered images described in Section 3. The proposed SMARS dataset meets our expectations in most of the reported experiments. Nevertheless, it also presents some limitations. Both will be discussed below for each of the available semantic categories in SMARS.

8.1.1. Buildings

The buildings generated by CityEngine exhibit good quality in terms of geometry, architectural appearance, and textures. They can be favorably compared to models with LoD2 and LoD3, as some rooftops have additional features such as chimneys. Moreover, the buildings resemble the expected distribution of a city in terms of size and arrangement and contribute to creating realistic scenarios. Taking into account the options to manipulate the building properties, it is easy to simulate the city growth as required for the change detection task. Furthermore, as *buildings* achieve a very good reconstruction in the DSMs, they can be easily detected by the algorithms considered in this article.

Nonetheless, the pool of textures to generate the *buildings* is limited and might lead to overfitting in the learning process. Besides, no construction sites are part of the dataset, as would be the case for real images; these regions represent a challenge for change detection depending on the progress of the constructions. Another constraint is given by the generation of mostly residential buildings, as facilities such as commercial buildings, parks, sports centers, or transport stations are not included in our dataset.

In the experiments, we notice that the discrepancy in height between the two city models leads to errors for prediction in the learning models, as the DSMs values have different ranges. With traditional approaches, the similarity in height between *trees* and *buildings* can also increase the challenges of classification, especially when they are close to each other. In the SMARS dataset, the building roofs are generally well visible and do not suffer from occlusion problems as in real data, making the task of building extraction easier.

8.1.2. Street

A major difference between the two models is the street category. In SPariS the streets match the common design with sidewalks, concrete material, and broken and solid lines. Besides, streets in this model are wide and have a height profile different from all other elements, with the exception of lawns.

SVenice is more difficult in this category. In the same way as the real city, the streets are designed for pedestrians, and are therefore narrow, causing sidewalks to be absent and are not marked either by broken or solid lines. Additionally, the width of the streets is comparable to the one of the multiple canals crossing the city. This problem is aggravated by the similarities in terms of height between the “others” (where canals and sea are included) and street categories. Because of that, we can notice in the semantic segmentation task that cross-domain experiments drop significantly in performance for this category. For learning models trained with SPariS, the canals of SVenice are considered streets and the lawns are predicted as “others”. Likewise, for learning models trained with SVenice, the streets of SPariS are many times wrongly labeled as “others” and only a few streets are actually detected.

As width and height are within the expected ranges for streets, a suitable solution would be to enhance the available categories in order to incorporate canals, squares, roundabouts, alleys, and other elements that could be confused with roads.

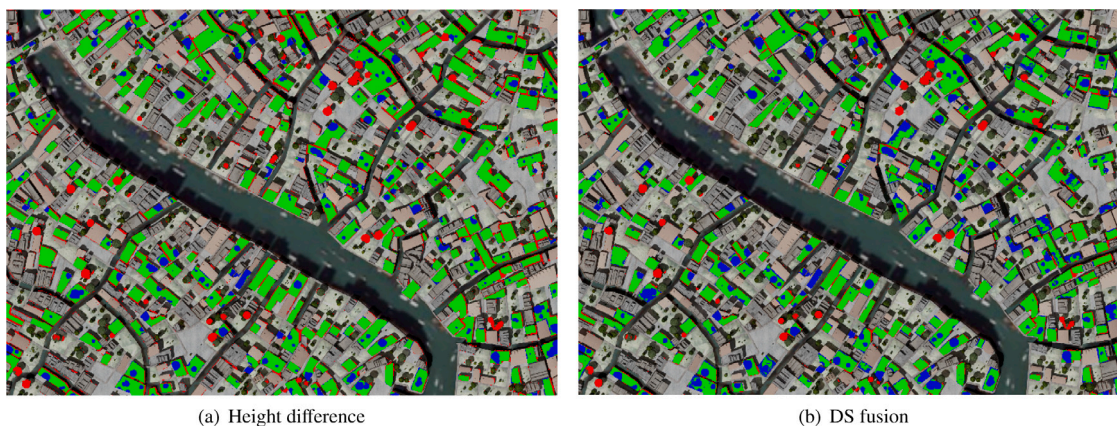


Fig. 23. 3D change detection result for SVenice (I) generated by direct height difference (a) and decision fusion. Legend: True Positive (green), False Positive (red), False Negative (blue). True Negative is not displayed.

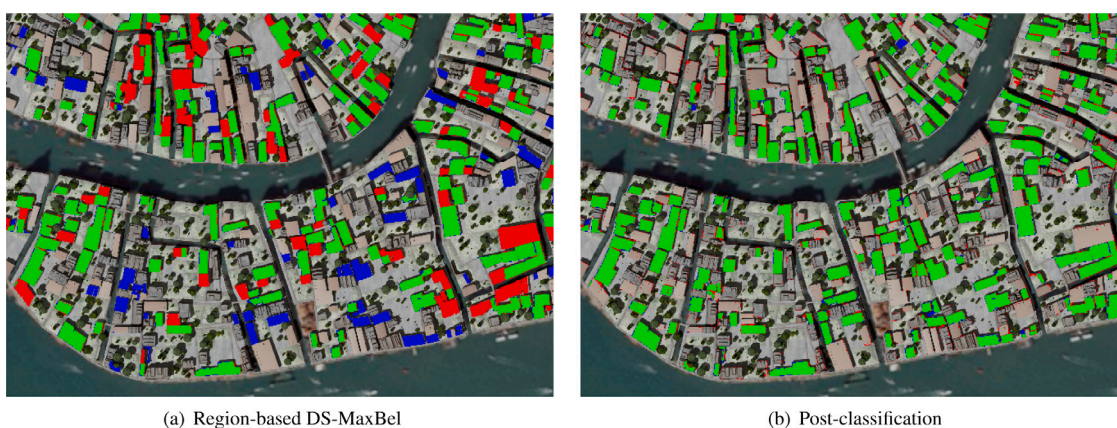


Fig. 24. 3D change detection result of SVenice-AOI2 relying on region-based approaches (a) and post-classification results (b). Legend: True Positive (green), False Positive (red), False Negative (blue). True Negative is not displayed.

8.1.3. Vegetation and lawns

Representation of shapes and structures of trees and bushes in 3D is a critical issue. A detailed representation requires a complicated geometric definition leading to high computational costs. A common simplified case with only two intersected vertical planes greatly reduces the memory requirements but exhibits poor visual quality in the models. Due to the trade-off between memory and appearance, we used textured ellipsoids. This allows the inclusion of a large number of trees and bushes in the virtual scenes. We include many textures, but these are limited to a specific number of plant species.

Yet, the *vegetation* regions largely suffer from the domain gaps between synthetic and real data. Real scenes have no simplified geometry (with the exception of man-trimmed trees) and cannot be easily modeled. Using only ellipsoids makes the learning biased towards this shape, and cannot adequately lead to correct predictions of other types of vegetation. Also, seasonal effects (such as leaf colors, snow covering, or fallen leaves) are not considered.

On top of that, the *lawns* category has been simplified too. Actual grass has a non-negligible height (even if this is relatively small in comparison to the other objects), no uniform texture, and can include small vegetation such as low bushes. For the simulated cities, the *lawns* are simplified by a flat area with grass-like texture, which appears realistic enough in the orthophotos. Without the texture, the *lawns* would be similar to the *roads* or *bare soil* category, as the height information of *lawns* is set close to 0.

In DART, *trees* are defined by tree species, various attributes of trunk and crown, and are simulated using turbid voxels or isosceles triangles (Gastellu-Etchegorry et al., 2015). Tree crown shapes can be

chosen from ellipsoidal, ellipsoid-composed, truncated cone, trapezoid, and cylinder with truncated cone. In addition, branches and twigs can be added. However, the tree modeling requires many manual input and is still not realistic as desired. Nevertheless, there is still potential to improve the quality of the *trees* class by using existing detailed 3D tree models. For example, the RAdiation transfer Model Intercomparison (RAMI) experiments derived detailed and realistic 3D models of various tree species by in situ measurements. The 3D models have been exported to DART, and can be edited in Blender as well. But those tree models do not include enough typical urban tree species to represent the urban tree scenario. For the reasons described above, we did not adopt these accurate 3D tree models.

8.1.4. Water

Water is not an annotated category in our SMARS dataset. However, it is an important land cover type in the SVenice scene. In the provided Venice city model of CityEngine, the water bodies are actually covered by a real low-resolution satellite image, exhibiting shadows that might not correspond to the simulated sun conditions. In addition, elements present in the water (such as boats and bridges) do not have an above ground height, so the captured multi-view images do not present a meaningful disparity in the epipolar image pairs. Therefore, in the generated DSMs the surface of water bodies is rather flat and smooth. In reality, the elements present in the water would have a height value larger than zero.

On the other hand, the SParis model has no *water*, so these are absent in the ground truth for either city, an aspect which can lead to errors in the semantic segmentation task, especially for cross domain

experiments. It is particularly complex for the algorithms to separate water from streets in the SVenice model, where the canals have similar contextual features as the *streets* in SParis. The collection of a larger number of samples with labeled *water* coverage might help solve this issue.

Finally, since we use an aerial photo as the source for the water areas, these do not change between the pre- and post-models and remain also constant within the simulated flight campaigns. In reality, the waves and tides produce an irregular surface, causing the matching algorithms to yield poor results. Usually, the DSM pipeline would fail to reconstruct such regions, while our DSM has a constant value. As discussed above, a physical simulation of water would lead to enhanced realism in the scenarios. Since our work focuses mainly on buildings, this is currently left out of our studies.

8.2. Single domain test

In single-domain building extraction experiments, both optical image and point cloud methods produce satisfactory results. As training and testing data share common features, very precise results are therefore derived for the task of building segmentation. The optical images have slightly better performance concerning 3D point clouds, as the images exhibit denser features in comparison while the point clouds have sparse representations.

Buildings in SParis and SVenice exhibit large variability in roof texture and their details, or the size and shape of the buildings. As a result, the evaluation metrics show that building extraction is less complex for SParis with respect to SVenice. The situation is more complex for the multi-class segmentation experiments. In SParis→SParis, the use of optical images achieves a mean IoU above 90%, while information from the 3D point cloud underperforms, with a mean IoU of 71%. The most problematic classes appear to be *lawns* and *background* classes. This suggests that point cloud features alone are not sufficient to represent some of the classes. For the case of SVenice→SVenice, both 2D and 3D methods exhibit relatively poor performance for the class *street*, as these are predominantly narrow pedestrian walks, which can be easily confused visually with the stone-paved square, belonging to the class *background* (Fig. 9(b)). Similarly, point cloud features of *streets* are not discriminative enough to allow separating this semantic class from the others. Different results are obtained for the class *buildings*: here, the 3D method can achieve satisfactory results not only for building extraction but also for multi-class semantic segmentation, indicating a good ability of point clouds to characterize features relating to man-made regular objects. However, it is worth noting that optical image analysis still outperforms the 3D method, achieving slightly higher IoU scores in all single-domain test scenarios, except for SVenice→SVenice multi-class semantic segmentation. These differences are observed in the building extraction experiments of SParis→SParis and SVenice→SVenice, as well as the multi-class semantic segmentation experiment of SParis→SParis, with differences of 0.57%, 1.06%, and 1.22%, respectively. This is due to the reason that building objects in DSMs are easily confused with other classes having similar heights by geometric features. In Fig. 13(b), an evident false positives area is noticeable at the right border of the image, where several trees are incorrectly recognized as buildings. In Fig. 10(b), no such error is present. In addition, due to limitations of the matching algorithms, some building boundaries in DSMs are incomplete and missing a few pixels (Tian et al., 2013; d'Angelo and Reinartz, 2011), leading to more false negatives in a 3D single-domain test when compared with optical image analysis.

The difference in performance between the binary and the multi-class segmentation lies partly in the optimization. It is intrinsically more difficult to optimize a multi-class problem with respect to a binary one, which results in a longer convergence time and less definite decision boundary. In addition, as the optimizers take into account the loss values of all classes, the gradient for weight update is different from the binary building extraction experiment. In conclusion, from

the single domain experiments performed we do not observe particular differences from the use of real multimodal data. Therefore, we can conclude that the SMARS dataset could be suitable as a training dataset for multimodal remote sensing tasks. Compared with SParis data, SVenice dataset is more challenging.

8.3. Cross domain test

In the remote sensing field, domain gap or shift is a common challenge for deep learning models. Preparing labeled datasets is normally costly and time-consuming, therefore many weakly and semi-supervised learning approaches are proposed by utilizing existing benchmark datasets (Li et al., 2022). However, target and source domain datasets may be different in terms of city styles, ground object types, seasonal changes, or characteristics of the acquiring sensors, leading to widespread attention of domain adaptation in recent years (Tuia et al., 2016). The lack of benchmark datasets hinders in-depth research in this field, especially for domain adaptation of the joint use of 2D/3D multimodal datasets. The experiments show that the two synthetic data generated using the proposed approach, namely SParis and SVenice, have clear domain gaps, and the results of 5-class semantic segmentation still have significant room for improvements in both 2D and 3D experiments. For example, for the SParis→SVenice and SVenice→SParis scenarios, it is common for streets to be confused with other classes.

For building extraction tasks, the 2D version is suitable for testing domain adaptation methods, while the 3D version of the SParis→SVenice case can be further refined based on baseline methods. The synthetic→real workflow is a challenge presenting wide opportunities for its exploration. Training with synthetic data and testing on real data can significantly reduce the cost of annotating training samples. Likewise, training with real data and testing on synthetic data for evaluating models can greatly reduce the cost of annotating testing samples, which typically require higher accuracy. Furthermore, the reference data associated to the generated synthetic data is ensured to be free from annotation errors. Therefore, this benchmark provides a starting point for the remote sensing community to investigate such topics.

When using different baseline methods, 3D data are more robust to domain shifts for buildings with respect to optical 2D images, while the opposite happens for single-domain tests. Point cloud networks, which are based on geometric features, have better generalization abilities in building extraction tasks for unseen domains, as they are not influenced by possible confusion between spectral features. For instance, in Fig. 16, the image network wrongly recognizes several roads as buildings, as their colors and 2D geometry are similar, while such errors do not occur in the results derived from the point cloud network. The point cloud network SparseConvNet outperforms the image network Swin Transformer for the building class in the synthetic→real building extraction and 5-class semantic segmentation cases. As illustrated in Fig. 12(f), a non-building object is misclassified as a building due to the lack of geometric information, while the prediction from the point cloud network is correct.

Cross-domain results are similar to what would be expected to achieve using real data, demonstrating the feasibility of the SMARS datasets to be integrated into practical applications employing real images. In this paper, no new domain adaptation approach is proposed: we encourage other researchers to test their approaches on this dataset, or prepare their own synthetic data with the proposed approach for their test regions of interest.

8.4. 3D building change detection

Recent years witnessed an increase in demand for accessible and high quality 3D dataset (Tian et al., 2013; Tian and Dezert, 2019; Xie et al., 2020). Their multi-temporal availability represents a desired

feature enabling applications to 3D change detection, where the accuracy of the results is increased by the provided information on targets height, complementary to the spectral information conveyed by optical earth observation data (Qin et al., 2016). Nevertheless, the lack of available benchmark datasets of this kind makes the development of 3D change detection approaches difficult, especially the ones relying on deep learning, as demonstrated by their scarcity in literature. The production of data for 3D change detection presents several problems. On the one hand, large cities in developed countries have limited changes, not sufficient to train a deep network (Tian et al., 2013). On the other hand, in developing countries building changes are often confused with different categories of changes, such as construction of highways and train stations, hindering their correct annotation. In addition, 2D/3D multimodal multi-temporal datasets are generally expensive to acquire, and several research institutes collect new data in the frame of specific projects: therefore, they cannot easily disclose them as publicly available benchmarks.

This paper presents a novel workflow to generate synthetic data suitable for training classification algorithms for 3D change detection. The illumination conditions of the simulated optical images present relevant differences, making this task non-trivial for algorithms relying solely on spectral changes. Pre- and post-event data are almost perfectly co-registered, allowing the user to remove this source of error propagation in their change detection workflow, which must be dealt with when using real data.

The introduced SMARS dataset presents aspects which may be improved in the future. Regarding the intrinsic quality and rendering of the data, results show that DSMs exhibit sharp boundaries and a reduced number of occluded areas with respect to typical real digital elevation models. Regarding the content of the scenes, in SPARIS most of the building blocks have been extended or partially removed in the transition from the simulated pre- to the post-images, and the changes are evenly distributed throughout the entire virtual city. This usually does not correspond to the pace and distribution of urban pattern changes in the real world.

The reported experiments suggest that traditional machine learning approaches are not optimal at detecting building changes relying on optical images only, as no elevation data are available. The use of high quality DSMs increases the accuracy of the results: however, when using only the generated synthetic DSMs, changes in buildings are often confused with changes in trees, keeping this task highly challenging. In this paper, the best change detection results are obtained by employing both simulated optical data and their associated DSM, by directly comparing the pre- and post-event building masks generated by multimodal co-learning approaches.

9. Conclusion

In this paper we introduce SMARS, a synthetic large and accurately annotated 2D/3D multi-temporal earth observation dataset, as an effort to meet the demand for multimodal benchmark data suitable for change detection applications in urban areas. In addition to 3D change detection, we provide orthorectified images, DSMs and ground truth for semantic segmentation, along with a pipeline to generate similar synthetic images resembling the characteristics of real aerial acquisitions, including their limitations. By modifying the scenes within the pipeline, it is easy to set and adjust the changes between two simulated acquisition times, which is a difficult task when using real data. As a result, the pipeline has the potential to create larger samples with high variability. As the main goal of this paper is the generation of synthetic 2D/3D multimodal data as similar as possible to real data, deep-learning based 3D change detection approaches are not discussed here.

The ground truth associated to the dataset is free from wrongly annotated labels or confusion between classes, being generated during the rendering process. This aspect propagates its advantages to the

change detection applications, where a large number of modifications can be handled and are ensured to be correct in the change mask to be used as reference. The quality of the presented synthetic data has been investigated in several experiments, which yielded results similar to what would be expected using real data. The quality of SMARS data is high in terms of coregistration, orthorectification and ground truth quality.

In addition to testing segmentation and change detection approaches, the presented synthetic data can be adapted to train a valid building extraction or semantic segmentation model that can be applied to real datasets. For instance, building extraction shows a good performance on the ISPRS Potsdam dataset, even without a fine-tuning step. Considering the 3D case, most of the buildings are properly classified with sharp boundaries. However, land cover classes not present in the synthetic data were not properly handled by the networks and lead to wrong classification. In terms of multi-class semantic segmentation, we observed a good performance within the same domain, but this decreased when using cross-domain datasets. Besides, it is not a trivial task to evaluate the transferability since the semantic classes present are different in the considered datasets. Further reducing the domain gaps between real and synthetic data, as well as increasing the available number of classes could help to overcome these difficulties. On the other hand, for the building semantic segmentation experiments, we observe good results as most of the classes have been properly predicted, with the exception of building edges and vegetation for some cases. In general, the synthetic data represent a feasible option for training neural networks for building detection, semantic segmentation, and change detection tasks, in spite of the described constraints due to domain gaps.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgments

Mario Fuentes Reyes is currently funded by a DLR-DAAD Research Fellowship (No. 57478193) to pursue his PhD studies. Yuxing Xie was supported by a DLR-DAAD Research Fellowship (No. 57424731).

References

- Almuntairi, A., Warner, T.A., 2010. Change detection accuracy and image properties: a study using simulated data. *Remote Sens.* 2 (6), 1508–1529.
- Bachhofner, S., Loghin, A.-M., Otepka, J., Pfeifer, N., Hornacek, M., Siposova, A., Schmidinger, N., Hornik, K., Schiller, N., Kähler, O., et al., 2020. Generalized sparse convolutional neural networks for semantic segmentation of point clouds derived from tri-stereo satellite imagery. *Remote Sens.* 12 (8), 1289.
- Bartier, P.M., Keller, C.P., 1996. Multivariate interpolation to incorporate thematic surface data using inverse distance weighting (IDW). *Comput. Geosci.* 22 (7), 795–799.
- Börner, A., Wiest, L., Keller, P., Reulke, R., Richter, R., Schaepman, M., Schläpfer, D., 2001. SENSOR: a tool for the simulation of hyperspectral remote sensing systems. *ISPRS J. Photogramm. Remote Sens.* 55 (5–6), 299–312.
- Caye Daudt, R., Le Saux, B., Boulch, A., Gousseau, Y., 2018. Urban change detection for multispectral earth observation using convolutional neural networks. In: *IEEE International Geoscience and Remote Sensing Symposium (IGARSS)*. pp. 2115–2118.
- Caye Daudt, R., Le Saux, B., Boulch, A., Gousseau, Y., 2019. Multitask learning for large-scale semantic change detection. *Comput. Vis. Image Underst.* 187, 102783.
- Chen, M., Hu, Q., Yu, Z., Thomas, H., Feng, A., Hou, Y., McCullough, K., Ren, F., Soibelman, L., 2022. STPLS3D: A large-scale synthetic and real aerial photogrammetry 3D point cloud dataset. In: *33rd British Machine Vision Conference 2022, BMVC 2022*, London, UK, November 21–24, 2022. BMVA Press.
- Chen, L., Liu, F., Zhao, Y., Wang, W., Yuan, X., Zhu, J., 2020. VALID: A comprehensive virtual aerial image dataset. In: *2020 IEEE International Conference on Robotics and Automation (ICRA)*. pp. 2009–2016.
- Chen, R.J., Lu, M.Y., Chen, T.Y., Williamson, D.F., Mahmood, F., 2021. Synthetic data in machine learning for medicine and healthcare. *Nat. Biomed. Eng.* 5 (6), 493–497.

- Coletta, V., Marsocci, V., Ravanelli, R., 2022. 3DCD: a new dataset for 2D and 3D change detection using deep learning techniques. *Int. Arch. Photogramm. Remote Sens. Spat. Inf. Sci. XLIII-B3-2022*, 1349–1354.
- Cordts, M., Omran, M., Ramos, S., Rehfeld, T., Enzweiler, M., Benenson, R., Franke, U., Roth, S., Schiele, B., 2016. The cityscapes dataset for semantic urban scene understanding. In: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 3213–3223.
- d'Angelo, P., Reinartz, P., 2011. Semiglobal matching results on the ISPRS stereo matching benchmark. *Int. Arch. Photogramm. Remote Sens. Spat. Inf. Sci. XXXVIII-4/W19*, 79–84.
- Denninger, M., Sundermeyer, M., Winkelbauer, D., Olefir, D., Hodan, T., Zidan, Y., Elbadrawy, M., Knauer, M., Katam, H., Lodhi, A., 2020. Blenderproc: Reducing the reality gap with photorealistic rendering. In: *International Conference on Robotics: Science and Systems, RSS 2020*.
- Disney, M., Lewis, P., Saich, P., 2006. 3D modelling of forest canopy structure for remote sensing simulations in the optical and microwave domains. *Remote Sens. Environ.* 100 (1), 114–132.
- Dosovitskiy, A., Ros, G., Codevilla, F., Lopez, A., Koltun, V., 2017. CARLA: An open urban driving simulator. In: *Proceedings of the 1st Annual Conference on Robot Learning*, PMLR, pp. 1–16.
- Everingham, M., Van Gool, L., Williams, C.K., Winn, J., Zisserman, A., 2010. The pascal visual object classes (VOC) challenge. *Int. J. Comput. Vis.* 88 (2), 303–338.
- Fabbri, M., Brasó, G., Mauter, G., Cetintas, O., Gasparini, R., Ošep, A., Calderara, S., Leal-Taixé, L., Cucchiara, R., 2021. Motsynth: How can synthetic data help pedestrian detection and tracking? In: *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 10849–10859.
- Fuentes Reyes, M., D'Angelo, P., Fraundorfer, F., 2022. SyntCities: A large synthetic remote sensing dataset for disparity estimation. *IEEE J. Sel. Top. Appl. Earth Obs. Remote Sens.* 15, 10087–10098.
- Gastellu-Etchegorry, J.-P., Yin, T., Lauret, N., Cajfinger, T., Gregoire, T., Grau, E., Feret, J.-B., Lopes, M., Guilleux, J., Dedieu, G., et al., 2015. Discrete anisotropic radiative transfer (DART 5) for modeling airborne and satellite spectroradiometer and LIDAR acquisitions of natural and urban landscapes. *Remote Sens.* 7 (2), 1667–1701.
- GDAL/OGR contributors, 2022. GDAL/OGR Geospatial Data Abstraction Software Library. Open Source Geospatial Foundation, <http://dx.doi.org/10.5281/zenodo.5884351>, URL: <https://gdal.org>.
- Geiger, A., Lenz, P., Stillér, C., Urtasun, R., 2013. Vision meets robotics: The KITTI dataset. *Int. J. Robot. Res.* 32 (11), 1231–1237.
- de Gélis, L., Lefèvre, S., Corpetti, T., 2021. Change detection in urban point clouds: An experimental comparison with simulated 3D datasets. *Remote Sens.* 13 (13), 2629.
- Ghamisi, P., Höfle, B., Zhu, X.X., 2016. Hyperspectral and LiDAR data fusion using extinction profiles and deep convolutional neural network. *IEEE J. Sel. Top. Appl. Earth Obs. Remote Sens.* 10 (6), 3011–3024.
- Graham, B., Engelcke, M., Van Der Maaten, L., 2018. 3D semantic segmentation with submanifold sparse convolutional networks. In: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 9224–9232.
- Gupta, R., Goodman, B., Patel, N., Hosfelt, R., Sajejev, S., Heim, E., Doshi, J., Lucas, K., Choset, H., Gaston, M., 2019. Creating xBD: A dataset for assessing building damage from satellite imagery. In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops*, pp. 10–17.
- He, J., Jia, X., Chen, S., Liu, J., 2021. Multi-source domain adaptation with collaborative learning for semantic segmentation. In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 11008–11017.
- Hirschmuller, H., 2008. Stereo processing by semiglobal matching and mutual information. *IEEE Trans. Pattern Anal. Mach. Intell.* 30 (2), 328–341.
- Hoeser, T., Kuenzer, C., 2022. SyntEO: Synthetic dataset generation for earth observation and deep learning—Demonstrated for offshore wind farm detection. *ISPRS J. Photogramm. Remote Sens.* 189, 163–184.
- Hosseinpour, H., Samadzadegan, F., Javan, F.D., 2022. CMGFNet: A deep cross-modal gated fusion network for building extraction from very high-resolution remote sensing images. *ISPRS J. Photogramm. Remote Sens.* 184, 96–115.
- Janoutová, R., Homolová, L., Malenkovský, Z., Hanuš, J., Lauret, N., Gastellu-Etchegorry, J.-P., 2019. Influence of 3D spruce tree representation on accuracy of airborne and satellite forest reflectance simulated in DART. *Forests* 10 (3), 292.
- Ji, S., Wei, S., Lu, M., 2019. Fully convolutional networks for multisource building extraction from an open aerial and satellite imagery data set. *IEEE Trans. Geosci. Remote Sens.* 57 (1), 574–586.
- Kong, F., Huang, B., Bradbury, K., Malof, J., 2020. The Synthinel-1 dataset: a collection of high resolution synthetic overhead imagery for building segmentation. In: *2020 Winter Conference on Applications of Computer Vision*, pp. 1814–1823.
- Krauß, T., 2014. Six years operational processing of satellite data using CATENA at DLR: Experiences and recommendations. *KN-J. Cartogr. Geogr. Inf.* 64 (2), 74–80.
- Kurz, F., Türmer, S., Meynberg, O., Rosenbaum, D., Runge, H., Reinartz, P., Leitloff, J., 2012. Low-cost optical camera systems for real-time mapping applications. *Photogramm.-Fernerkund.-Geoinf.* 159–176.
- Li, R., Jiao, Q., Cao, W., Wong, H.-S., Wu, S., 2020. Model adaptation: Unsupervised domain adaptation without source data. In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 9641–9650.
- Li, X., Strahler, A.H., 1985. Geometric-optical modeling of a conifer forest canopy. *IEEE Trans. Geosci. Remote Sens.* GE-23 (5), 705–721.
- Li, H., Tian, J., Xie, Y., Li, C., Reinartz, P., 2022. Performance evaluation of fusion techniques for cross-domain building rooftop segmentation. *Int. Arch. Photogramm. Remote Sens. Spat. Inf. Sci.* 501–508.
- Li, H., Wang, Z., Hong, T., 2021. A synthetic building operation dataset. *Sci. Data* 8 (1), 1–13.
- Li, X., Wang, K., Tian, Y., Yan, L., Deng, F., Wang, F.-Y., 2019. The ParallelEye dataset: A large collection of virtual images for traffic vision research. *IEEE Trans. Intell. Transp. Syst.* 20 (6), 2072–2084.
- Lin, T.-Y., Maire, M., Belongie, S., Hays, J., Perona, P., Ramanan, D., Dollár, P., Zitnick, C.L., 2014. Microsoft COCO: Common objects in context. In: *European Conference on Computer Vision*, Springer, pp. 740–755.
- Liu, Z., Lin, Y., Cao, Y., Hu, H., Wei, Y., Zhang, Z., Lin, S., Guo, B., 2021. Swin transformer: Hierarchical vision transformer using shifted windows. In: *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 10012–10022.
- Ma, J.W., Czerniawski, T., Leite, F., 2020. Semantic segmentation of point clouds of building interiors with deep learning: Augmenting training datasets with synthetic BIM-based point clouds. *Autom. Constr.* 113, 103144.
- Marsocci, V., Coletta, V., Ravanelli, R., Scardapane, S., Crespi, M., 2023. Inferring 3D change detection from bitemporal optical images. *ISPRS J. Photogramm. Remote Sens.* 196, 325–339.
- Minaee, S., Boykov, Y., Porikli, F., Plaza, A., Kehtarnavaz, N., Terzopoulos, D., 2021. Image segmentation using deep learning: A survey. *IEEE Trans. Pattern Anal. Mach. Intell.* 44 (7), 3523–3542.
- Mnih, V., 2013. *Machine Learning for Aerial Image Labeling*. University of Toronto, Canada.
- Nielsen, A.A., 2007. The regularized iteratively reweighted MAD method for change detection in multi-and hyperspectral data. *IEEE Trans. Image Process.* 16 (2), 463–478.
- Nikolenko, S.I., 2021. *Synthetic Data for Deep Learning*, Vol. 174. Springer.
- Peng, D., Guan, H., Zang, Y., Bruzzone, L., 2022. Full-level domain adaptation for building extraction in very-high-resolution optical remote-sensing images. *IEEE Trans. Geosci. Remote Sens.* 60, 1–17.
- Pourpanah, F., Abdar, M., Luo, Y., Zhou, X., Wang, R., Lim, C.P., Wang, X.-Z., Wu, Q.J., 2022. A review of generalized zero-shot learning methods. *IEEE Trans. Pattern Anal. Mach. Intell.*
- Qi, J., Xie, D., Yin, T., Yan, G., Gastellu-Etchegorry, J.-P., Li, L., Zhang, W., Mu, X., Norford, L.K., 2019. LESS: Large-Scale remote sensing data and image simulation framework over heterogeneous 3D scenes. *Remote Sens. Environ.* 221, 695–706.
- Qin, R., Tian, J., Reinartz, P., 2016. 3D change detection—approaches and applications. *ISPRS J. Photogramm. Remote Sens.* 122, 41–56.
- Richter, S.R., Vineet, V., Roth, S., Koltun, V., 2016. Playing for data: Ground truth from computer games. In: *Computer Vision—ECCV 2016: 14th European Conference, Amsterdam, The Netherlands, October 11–14, 2016, Proceedings, Part II 14*. Springer, pp. 102–118.
- Roberts, M., Ramapuram, J., Ranjan, A., Kumar, A., Bautista, M.A., Paczan, N., Webb, R., Susskind, J.M., 2021. Hypersim: A photorealistic synthetic dataset for holistic indoor scene understanding. In: *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 10912–10922.
- Ros, G., Sellart, L., Materzynska, J., Vazquez, D., Lopez, A.M., 2016. The SYNTHIA dataset: A large collection of synthetic images for semantic segmentation of urban scenes. In: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 3234–3243.
- Schwind, P., Müller, R., Palubinskas, G., Storch, T., 2012. An in-depth simulation of EnMAP acquisition geometry. *ISPRS J. Photogramm. Remote Sens.* 70, 99–106.
- Shah, S., Dey, D., Lovett, C., Kapoor, A., 2018. Airsim: High-fidelity visual and physical simulation for autonomous vehicles. In: *Field and Service Robotics: Results of the 11th International Conference*, Springer, pp. 621–635.
- Shao, R., Du, C., Chen, H., Li, J., 2021. SUNet: Change detection for heterogeneous remote sensing images from satellite and UAV using a dual-channel fully convolution network. *Remote Sens.* 13 (18), 3750.
- Shi, W., Zhang, M., Zhang, R., Chen, S., Zhan, Z., 2020. Change detection based on artificial intelligence: State-of-the-art and challenges. *Remote Sens.* 12 (10), 1688.
- Stoecklein, D., Lore, K.G., Davies, M., Sarkar, S., Ganapathysubramanian, B., 2017. Deep learning for flow sculpting: Insights into efficient learning using scientific simulation data. *Sci. Rep.* 7 (1), 1–11.
- Tao, J., Auer, S., Palubinskas, G., Reinartz, P., Bamler, R., 2013. Automatic SAR simulation technique for object identification in complex urban scenarios. *IEEE J. Sel. Top. Appl. Earth Obs. Remote Sens.* 7 (3), 994–1003.
- Tian, J., Cui, S., Reinartz, P., 2013. Building change detection based on satellite stereo imagery and digital surface models. *IEEE Trans. Geosci. Remote Sens.* 52 (1), 406–417.
- Tian, J., Dezert, J., 2019. Fusion of multispectral imagery and DSMS for building change detection using belief functions and reliabilities. *Int. J. Image Data Fusion* 10 (1), 1–27.
- Townshend, J.R., Justice, C.O., Gurney, C., McManus, J., 1992. The impact of misregistration on change detection. *IEEE Trans. Geosci. Remote Sens.* 30 (5), 1054–1060.

- Tuia, D., Persello, C., Bruzzone, L., 2016. Domain adaptation for the classification of remote sensing data: An overview of recent advances. *IEEE Geosci. Remote Sens. Mag.* 4 (2), 41–57.
- Xiao, A., Huang, J., Guan, D., Zhan, F., Lu, S., 2022. Transfer learning from synthetic to real LiDAR point cloud for semantic segmentation. In: *Proceedings of the AAAI Conference on Artificial Intelligence*. pp. 2795–2803.
- Xie, Y., Tian, J., Zhu, X.X., 2020. Linking points with labels in 3D: A review of point cloud semantic segmentation. *IEEE Geosci. Remote Sens. Mag.* 8 (4), 38–59.
- Xie, Y., Tian, J., Zhu, X.X., 2023. A co-learning method to utilize optical images and photogrammetric point clouds for building extraction. *Int. J. Appl. Earth Obs. Geoinf.* 116, 103165.
- Zabih, R., Woodfill, J., 1994. Non-parametric local transforms for computing visual correspondence. In: *European Conference on Computer Vision*. Springer, pp. 151–158.
- Zhou, Z.-H., 2018. A brief introduction to weakly supervised learning. *Natl. Sci. Rev.* 5 (1), 44–53.
- Zhu, X.X., Tuia, D., Mou, L., Xia, G.-S., Zhang, L., Xu, F., Fraundorfer, F., 2017. Deep learning in remote sensing: A comprehensive review and list of resources. *IEEE Geosci. Remote Sens. Mag.* 5 (4), 8–36.

D Yuxing Xie, Xiangtian Yuan, Xiao Xiang Zhu, and Jiaojiao Tian.
“Multimodal Co-learning for Building Change Detection: A Domain Adaptation Framework Using VHR Images and Digital Surface Models.”
IEEE Transactions on Geoscience and Remote Sensing (2024) 62 (2024): 5402520.

<https://doi.org/10.1109/TGRS.2024.3362680>

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

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

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

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

I. INTRODUCTION

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

commonly consists of two steps: feature extraction and change detection [5].

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

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

The rapid advancement of deep neural networks in recent years has set new standards in supervised 2D change detection [5], [20]–[23]. Specifically, the success of convolutional neural networks (CNNs) in other remote sensing and computer vision tasks [24]–[27] has established CNNs as the backbone for change detection in numerous studies. Few of them are based on single-stream architectures [21], [28], [29], which take as input image differencing, hand-crafted change features, or concatenation of bitemporal images. Due to the large variability between the pre- and post-event images, the single-stream methods often suffer from noise and loss of information from the input, inhibiting a wider application in remote sensing change detection. Consequently, the mainstream methods are based on the Siamese architecture [21], [30], which extracts features from bitemporal images via two parallel encoders with the same network structure. The Siamese approaches not only

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

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

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

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

maintain the information lost in single-stream approaches but also exhibit more robust and distinctive representations for the object of interest [31]–[40]. More recently, vision transformer (ViT) [41], [42] has achieved further success in deep learning-based image processing topics and attracted attention from the remote sensing community. Transformer-based methods have also been introduced in several recent change detection studies [30], [43]–[45] and achieved stunning results. With the development of foundation models, recent studies like [46]–[48] have successfully incorporated them into the change detection task, paving a way for further improvement in this area.

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

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

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

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

- 1) We propose a co-learning framework for bitemporal images and DSMs modalities, focusing on the building change detection task. Three well-designed co-learning combinations (vanilla co-learning, fusion co-learning, and detached fusion co-learning) are proposed, defined, and investigated in this work. Furthermore, we present a way to determine whether these co-learning combinations are equivalent for different loss functions.
- 2) This work highlights the advantages of photogrammet-

ric DSMs in the task of building change detection. Compared with 2D optical imagery, existing studies on photogrammetric DSMs are limited. We propose an end-to-end transformer-based network for change detection from HDiff maps and investigate the difference from 2D change detection in cross-domain scenarios.

- 3) This work also involves synthetic-to-real domain adaptation, a novel topic in remote sensing. To the best of our knowledge, this is the first study to address this topic specifically for the change detection task. We utilize co-learning as a domain adaptation method and explore the potential of using the recently published synthetic benchmark dataset SMARS [72] to train change detection deep neural networks for a real dataset.

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

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

II. RELATED WORKS

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

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

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

Early fusion is carried out at the data input stage. In remote sensing tasks, 2D multispectral images are concatenated with the height values of DSMs or normalized DSMs (nDSMs) as the input channels to a single-modal network. For example, [74] proposes the gated residual refinement network (GRRNet) using multispectral images and LiDAR-derived nDSMs as the input. A gated feature labeling (GFL) unit is designed in the decoder to refine the semantic segmentation results. In a few early fusion studies, spectral information from images is added directly to 3D point clouds as per-point values, and colored point clouds are processed in a three-dimensional domain with point cloud deep neural networks. However, till

now no consensus has been reached on whether coloring the 3D point clouds brings advantages [75]. Some earlier studies found such fusion operations can even lead to a decline in the performance of point cloud networks [51], [76], [77].

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

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

Knowledge transfer does not directly operate on the data or extracted features. There are two principles of knowledge transfer methods: (1) employing different network branches for different data modalities. (2) Bridging the relationships between different modalities by soft connections (usually loss functions). Each network only influences others in the training phase and can be utilized alone for testing single-modal data.

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

B. Change Detection with Multimodal Data

Compared with the single-modal image or DSM data, multimodal data provide more stable and accurate change features. Therefore, several studies have introduced multimodal strategies for change detection. For example, in our previous works the decision fusion method belief functions have been proven to be an efficient fusion module for multimodal change detection [60], [83], [84], which can effectively improve the building change detection results compared with single-modal change indicators. The paper [83] proposes a change detection pipeline based on the robust height differences between DSMs and the similarity measurement between corresponding optical image pairs. A fusion module based on the Dempster-Shafer theory is adopted to fuse these two change indicators, which significantly improves the change accuracy compared with the results of either single modality. Additionally, vegetation and shadow classification results are introduced as extra information to refine the initial change detection results, and a building extraction method based on shape features is performed to get more accurate building change maps. [84] proposes another multimodal change detection framework. First, it uses a refined basic belief assignments (BBAs) model to calculate the BBAs of the change indicators from optical images and DSMs.

Then a building change detection decision fusion approach is applied to fuse these BBAs. Finally, four decision-making criteria are employed to convert the fused global BBAs to building change maps. [60] extends the framework in [84] and employs initial building probabilities extracted by the deep neural network Deeplabv3+ for the change decision, which shows better generalization ability than the previous version. Also based on the Dempster-Shafer theory, [85] introduces a complementary evidence fusion framework. In this framework, the image change indicator is calculated with the subtraction of the normalized difference vegetation index (NDVI) of bitemporal images. A complementary evidence combination rule is employed for the decision fusion to alleviate the conflicts between the change evidence from optical images and DSMs. Recently, [86] utilizes the morphological building index (MBI) as the image change feature and robust height difference proposed in [83] as the height change feature and proposes a co-segmentation framework for building change detection. The changed areas and unchanged areas are distinguished by a graph-cut-based energy minimization method.

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

III. METHODOLOGY

A. Overview

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

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

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

Assume that there are two datasets in a cross-domain scenario, the source dataset \mathbb{D}_s and the target dataset \mathbb{D}_t . Each dataset includes bitemporal data. In the following text, we use subscripts 1 and 2 to denote pre- and post-event data,

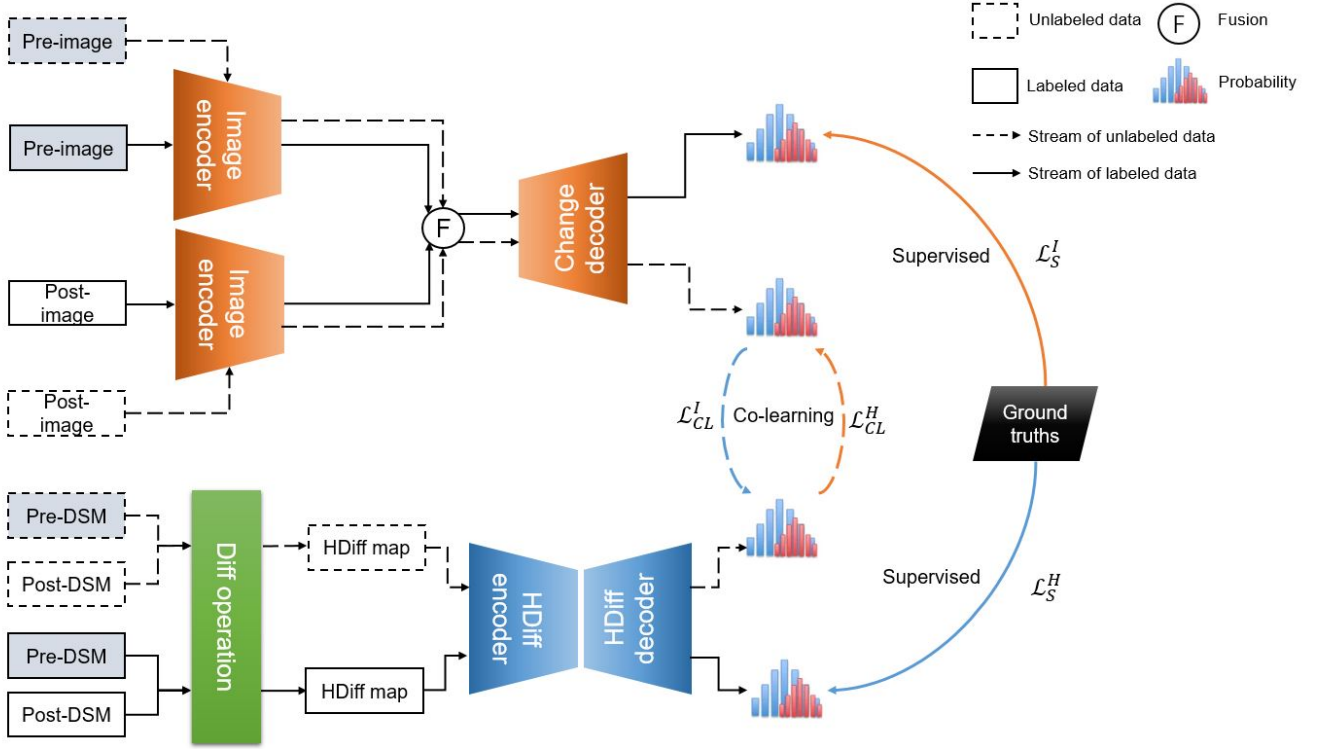


Fig. 1. Our proposed co-learning change detection framework.

respectively. \mathbb{D}_s consists of labeled source samples $\{\{I_1^s, I_2^s\}, \{H_1^s, H_2^s\}, G^s\}$, including pre-images I_1^s , post-images I_2^s , pre-DSMs H_1^s , post-DSMs H_2^s , and the change detection ground truths G^s . \mathbb{D}_t consists of unlabeled target samples $\{\{I_1^t, I_2^t\}, \{H_1^t, H_2^t\}\}$, including pre-images I_1^t , post-images I_2^t , pre-DSMs H_1^t , and post-DSMs H_2^t .

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

$$P_I^s = f_I(I_1^s, I_2^s), \quad (1)$$

$$P_I^t = f_I(I_1^t, I_2^t), \quad (2)$$

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

$$P_H^s = f_H(H_1^s, H_2^s), \quad (3)$$

$$P_H^t = f_H(H_1^t, H_2^t), \quad (4)$$

1) *Supervised Change Detection with Labeled Source Data:* To supervise the pixel-wise change detection, a generic loss function L_S measuring the difference between the source

building change probability P_I^s/P_H^s and ground truth G_s is needed:

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

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

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

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

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

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

$$\mathcal{L}_{CL-V}^H = L_C(P_{I,d}^t || P_H^t), \quad (8)$$

where $P_{H,d}^t$ and $P_{I,d}^t$ refer to detached P_H^t and P_I^t , respectively. Detached probabilities mean they are variables removed from the gradient computational graph so they do not affect the update of the weights for the corresponding networks. They

can be named shadow reference probability, utilized by the main modality network as the reference in the co-learning loss function [51].

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

$$\mathcal{L}_{CL-F}^I = L_C\left(\frac{P_I^t + P_{H,d}^t}{2} || P_I^t\right), \quad (9)$$

$$\mathcal{L}_{CL-F}^H = L_C\left(\frac{P_{I,d}^t + P_H^t}{2} || P_H^t\right), \quad (10)$$

where $P_{H,d}^t$ and $P_{I,d}^t$ refer to detached P_H^t and P_I^t , respectively.

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

$$\mathcal{L}_{CL-DF}^I = L_C\left(\frac{P_{I,d}^t + P_{H,d}^t}{2} || P_I^t\right), \quad (11)$$

$$\mathcal{L}_{CL-DF}^H = L_C\left(\frac{P_{I,d}^t + P_{H,d}^t}{2} || P_H^t\right), \quad (12)$$

where L_C denotes a generic consistency loss function. $P_{H,d}^t$ and $P_{I,d}^t$ refer to detached P_H^t and P_I^t , respectively.

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

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

$$\mathcal{L}_{total}^I = \lambda_1 \mathcal{L}_S^I + \lambda_2 \mathcal{L}_{CL}^I, \quad (13)$$

$$\mathcal{L}_{total}^H = \lambda_1 \mathcal{L}_S^H + \lambda_2 \mathcal{L}_{CL}^H, \quad (14)$$

where $\mathcal{L}_{CL}^I \in \{\mathcal{L}_{CL-V}^I, \mathcal{L}_{CL-F}^I, \mathcal{L}_{CL-DF}^I\}$ and $\mathcal{L}_{CL}^H \in \{\mathcal{L}_{CL-V}^H, \mathcal{L}_{CL-F}^H, \mathcal{L}_{CL-DF}^H\}$. \mathcal{L}_{total}^I , \mathcal{L}_S^I , and \mathcal{L}_{CL}^I are the total loss function, the supervised loss function, and the co-learning loss function for the image modality, respectively. \mathcal{L}_{total}^H , \mathcal{L}_S^H , and \mathcal{L}_{CL}^H are the total loss function, the supervised loss function, and the co-learning loss function for the DSM modality, respectively. λ_1 and λ_2 are the hyperparameters to

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

Input: $\mathbb{D}_s, \mathbb{D}_t$

Output: W_I, W_H

- 1: Initialize W_I, W_H
 - 2: **while** $n < N$ **do**
 - 3: Part 1: Learning with labeled source samples
 - 4: (1) Randomly sample B labeled source data samples $\{\{I_1^s, I_2^s\}, \{H_1^s, H_2^s\}, G^s\}$ from the source dataset \mathbb{D}_s .
 - 5: (2) Forward pass:
 - 6: $P_I^s \leftarrow f_I(I_1^s, I_2^s)$
 - 7: $P_H^s \leftarrow f_H(H_1^s, H_2^s)$
 - 8: (3) Calculate supervised loss:
 - 9: $\mathcal{L}_S^I \leftarrow L_S(G^s || P_I^s)$
 - 10: $\mathcal{L}_S^H \leftarrow L_S(G^s || P_H^s)$
 - 11:
 - 12: Part 2: Learning with unlabeled target samples
 - 13: (1) Randomly sample B unlabeled target data samples $\{\{I_1^t, I_2^t\}, \{H_1^t, H_2^t\}\}$ from the target dataset \mathbb{D}_t .
 - 14: (2) Forward pass:
 - 15: $P_I^t \leftarrow f_I(I_1^t, I_2^t)$
 - 16: $P_H^t \leftarrow f_H(H_1^t, H_2^t)$
 - 17: (3) Calculate co-learning loss:
 - 18: $\mathcal{L}_{CL}^I \leftarrow L_C(P_I^t, P_{H,d}^t)$
 - 19: $\mathcal{L}_{CL}^H \leftarrow L_C(P_H^t, P_{I,d}^t)$
 - 20:
 - 21: Part 3: Backward propagation and updating network parameters
 - 22: (1) Calculate total loss:
 - 23: $\mathcal{L}_{total}^I \leftarrow \lambda_1 \mathcal{L}_S^I + \lambda_2 \mathcal{L}_{CL}^I$
 - 24: $\mathcal{L}_{total}^H \leftarrow \lambda_1 \mathcal{L}_S^H + \lambda_2 \mathcal{L}_{CL}^H$
 - 25: (2) Backward pass:
 - 26: Calculate the backward pass for the image change detection network.
 - 27: Calculate the backward pass for the DSM change detection network.
 - 28: (3) Update: W_I, W_H
 - 29: **end while**
 - 30: **Return** W_I, W_H
-

weigh the supervised loss function and the co-learning loss function.

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

C. Siamese ResNet with Bitemporal Image Transformer Layer

Considering the balance between the network depth and GPU memory, we employ the ResNet-50 convolutional net-

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

D. Transformer-based UNet for HDiff Maps

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

E. Robust Height Difference

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

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

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

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

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

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

$$Diff_R^H(i, j) = Diff_P^H(i, j) + Diff_N^H(i, j), \quad (17)$$

F. Loss Functions

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

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

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

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

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

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

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

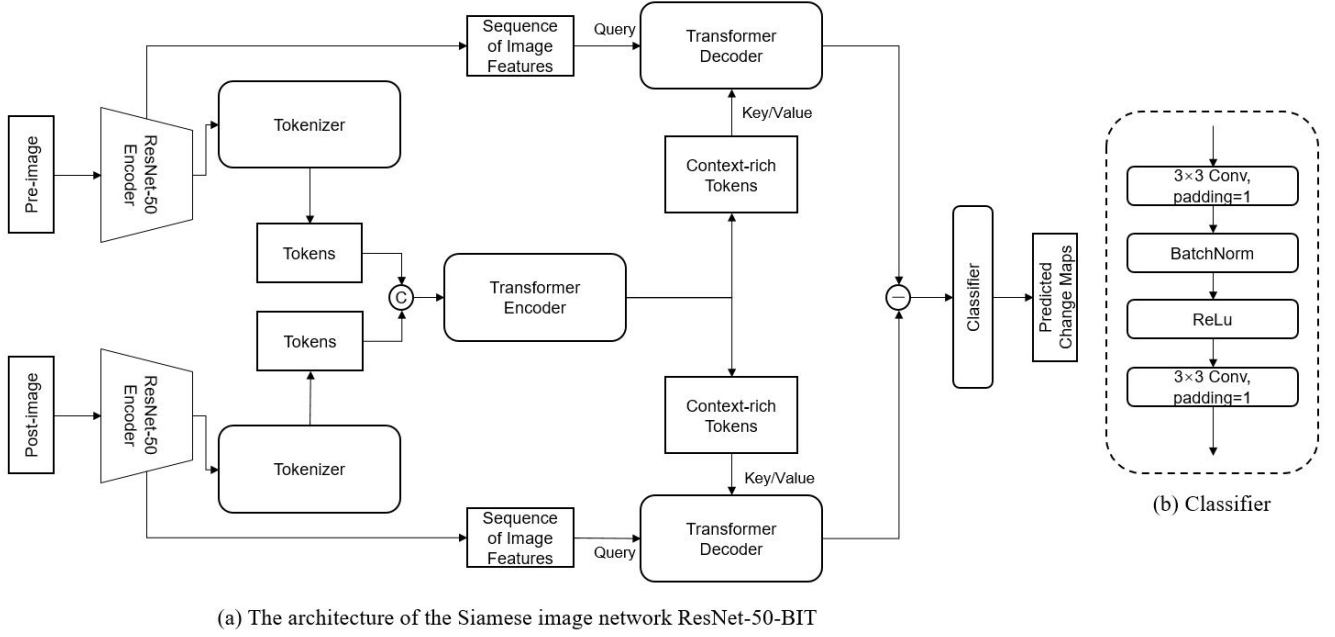


Fig. 2. (a) The architecture of the Siamese image network ResNet-50-BIT. (b) The classifier block. The modules of the tokenizer, transformer encoder, and transformer decoder are forked from the official implementation of [30] https://github.com/justchenhao/BIT_CD.

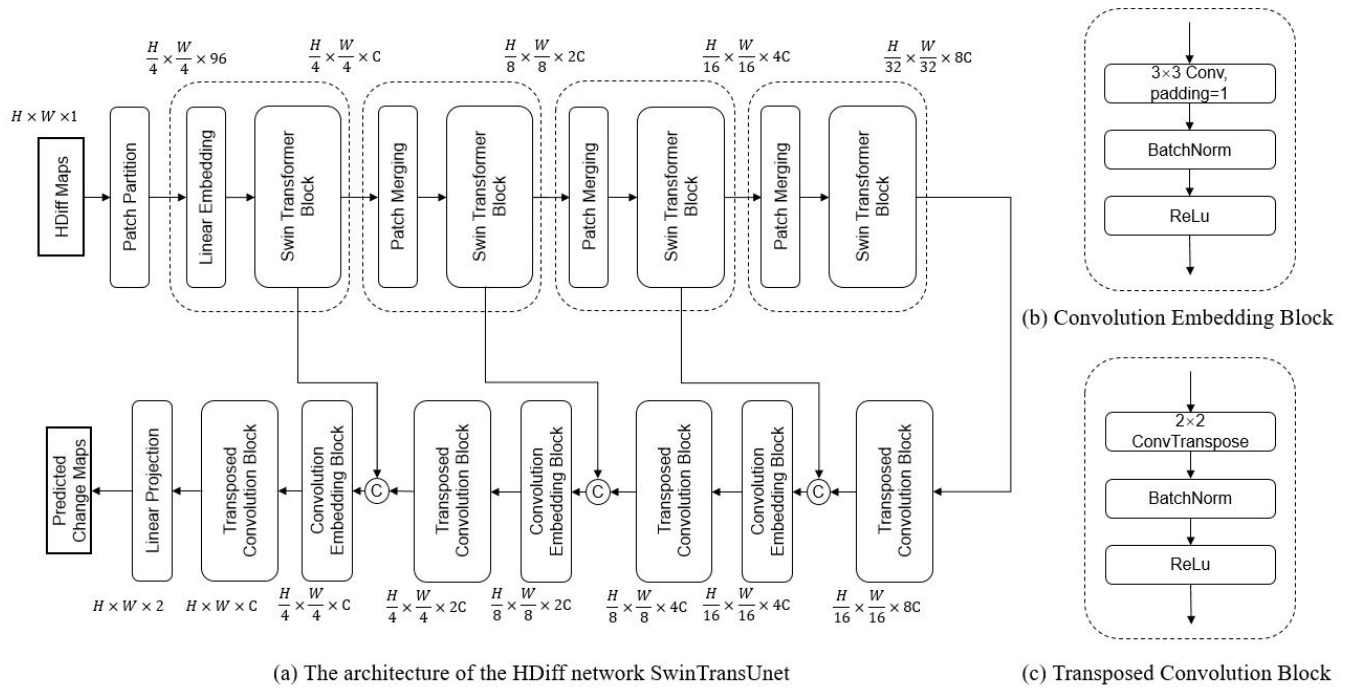


Fig. 3. (a) The architecture of the proposed HDiff network SwinTransUNet. (b) Convolution Embedding Block. (c) Transposed Convolution Block. The swin transformer encoder modules are forked from the official implementation of [42] <https://github.com/microsoft/Swin-Transformer>.

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

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

IV. EXPERIMENTS

A. Datasets

1) *Simulated Multimodal Aerial Remote Sensing (SMARS) dataset:* SMARS¹ is a recently published synthetic aerial remote sensing dataset by the German Aerospace Center (DLR) and the International Society for Photogrammetry and Remote Sensing (ISPRS) [72]. This dataset is designed for multimodal urban semantic segmentation, building extraction, and building change detection tasks. Its feasibility of being employed as a benchmark for algorithm training and evaluation has been proven [72]. It consists of two sub-datasets with distinct urban styles. One simulated city is named Synthetic Paris (SParis). The other is named Synthetic Venice (SVenice). Each sub-dataset includes bitemporal orthophotos, bitemporal photogrammetric DSMs, corresponding semantic maps, and corresponding building change maps. SMARS provides two versions, with resolutions of 30cm and 50cm, respectively. In this work, we employ the version of 50cm to evaluate the co-learning-based cross-domain change detection experiments. The training, validation, and testing raster sizes of the 50cm-SParis dataset are 1518×3560 pixels, 1008×3560 pixels, and 1974×3560 pixels, respectively. The training, validation, and testing raster sizes of the 50cm-SVenice dataset are 2800×5600, 2800×2128, and 2800×3472, respectively. Based on SParis and SVenice data, two groups of cross-domain experiments are conducted in this work: (1) SParis→SVenice: SParis used for training, SVenice for testing, and (2) SVenice→SParis: SVenice used for training, SParis for testing.

2) *Istanbul WorldView-2 dataset:* The Istanbul WorldView-2 dataset is a building change detection dataset covering two areas of Istanbul, Türkiye with a GSD of 50 cm. This dataset consists of 100 pairs of bitemporal orthophotos with RGB channels and photogrammetric DSMs from 2011 and 2012 and the corresponding building change ground truth annotated by hand. The orthophotos and photogrammetric DSMs are

generated from stereo WorldView-2 satellite images using the improved semi-global matching approach [90], [91]. Each patch has a pixel size of 400×400. In this work, the Istanbul WorldView-2 dataset is used as the testing data in a series of synthetic→real experiments, of which the training set is the SMARS dataset.

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

B. Experiment Setup

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

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

$$F1 = \frac{2TP}{2TP + FP + FN}, \quad (20)$$

$$IoU = \frac{TP}{TP + FP + FN}, \quad (21)$$

$$Precision = \frac{TP}{TP + FP}, \quad (22)$$

$$Recall = \frac{TP}{TP + FN}, \quad (23)$$

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

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

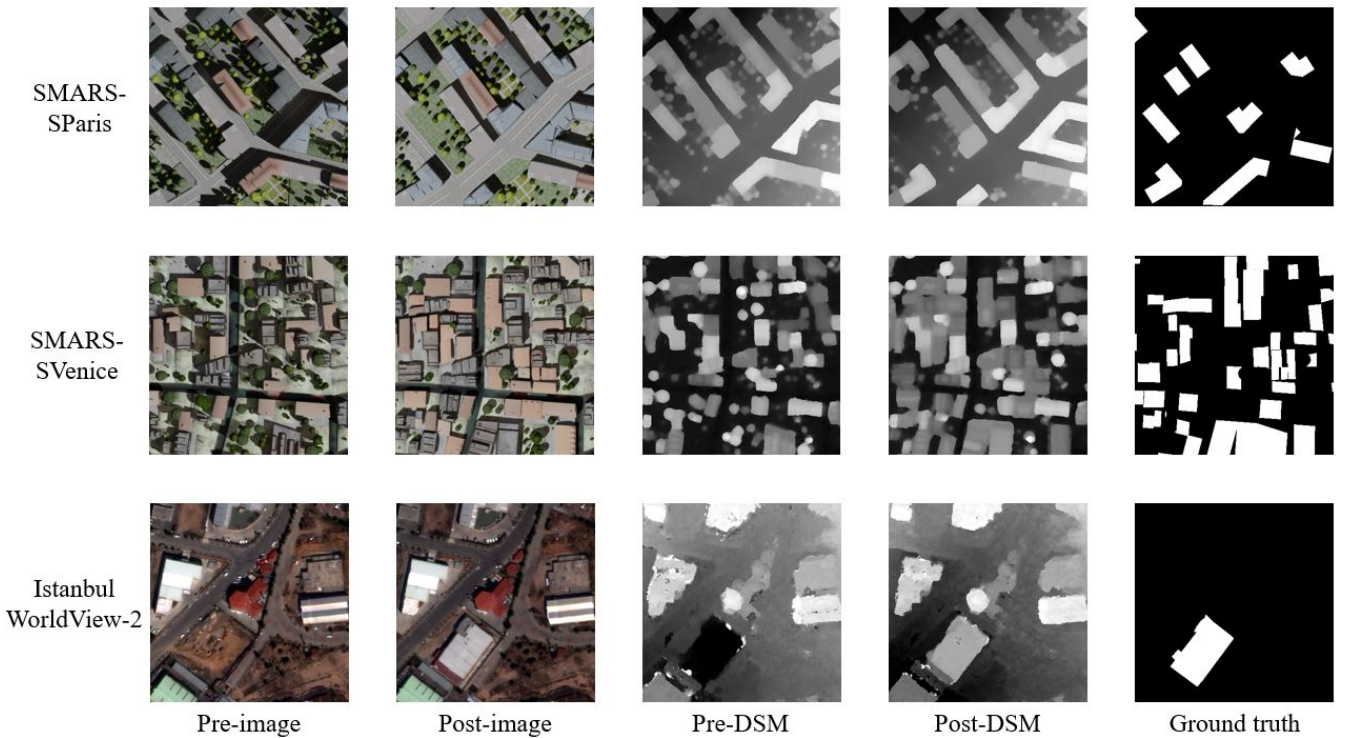


Fig. 4. Samples of SParis dataset and SVenice dataset of SMARS, as well as the Istanbul WorldView-2 dataset.

C. Experiment I: Domain Adaptation with Synthetic Data

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

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

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

Fig. 5 shows the qualitative results of SParis→SVenice. From the given examples, the baseline bitemporal method

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

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

Single-modal HDiff baseline method achieves much better results than the single-modal bitemporal image baseline method. Yet, image networks possess greater improvement potential when co-learning is applied. HDiff network is more prone to generate more false positive pixels, as shown in example A in Fig. 5 and example B in Fig. 6. Since the HDiff network is designed to detect the shapes with certain height differences in the HDiff map, some non-man-made

object changes have similar geometric features with changes in buildings and are therefore wrongly recognized. In Fig. 5 A, a noticeable false positive object of round shape at the left border of all results by the HDiff network is the change of a tree rather than a building. In the results of image-based methods, however, only the network trained with the KL-CL-F strategy makes the same mistake.

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

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

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

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

D. Experiment II: SMARS→istanbul WorldView-2

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

1) *Co-learning with direct HDiff maps*: As presented in Table III, the Siamese image baseline network ResNet-50-BIT trained with SMARS has abysmal performance on the unseen Istanbul dataset, in which only 4.57% of the F1 score and 2.34% of the IoU score are obtained. This performance can be attributed to the significant spectral domain gap between the synthetic images and real WorldView-2 images. The

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

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

Siamese DSM baseline method also produces poor results, again demonstrating that the Siamese DSM approach has a poor generalization ability. By contrast, the baseline HDiff network can achieve reasonable results with either **R** or **D**.

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

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

With robust HDiff maps, all co-learning methods can also improve the performance of both the ResNet-50-BIT image network and the SwinTransUNet HDiff network. The MSE-CL-V co-learning variety achieves the best image modality result, with an F1 score of 79.29% and an IoU score of 65.68%. For the DSM modality, the best result is achieved by MSE-CL-DF, with the F1 score of 77.62% and the IoU of 63.42%. In addition, each co-learning-enhanced HDiff network with robust HDiff maps yields better results compared with the same method utilizing direct HDiff maps. For the image modality, the best result achieved by MSE-CL-V (**R**) has a

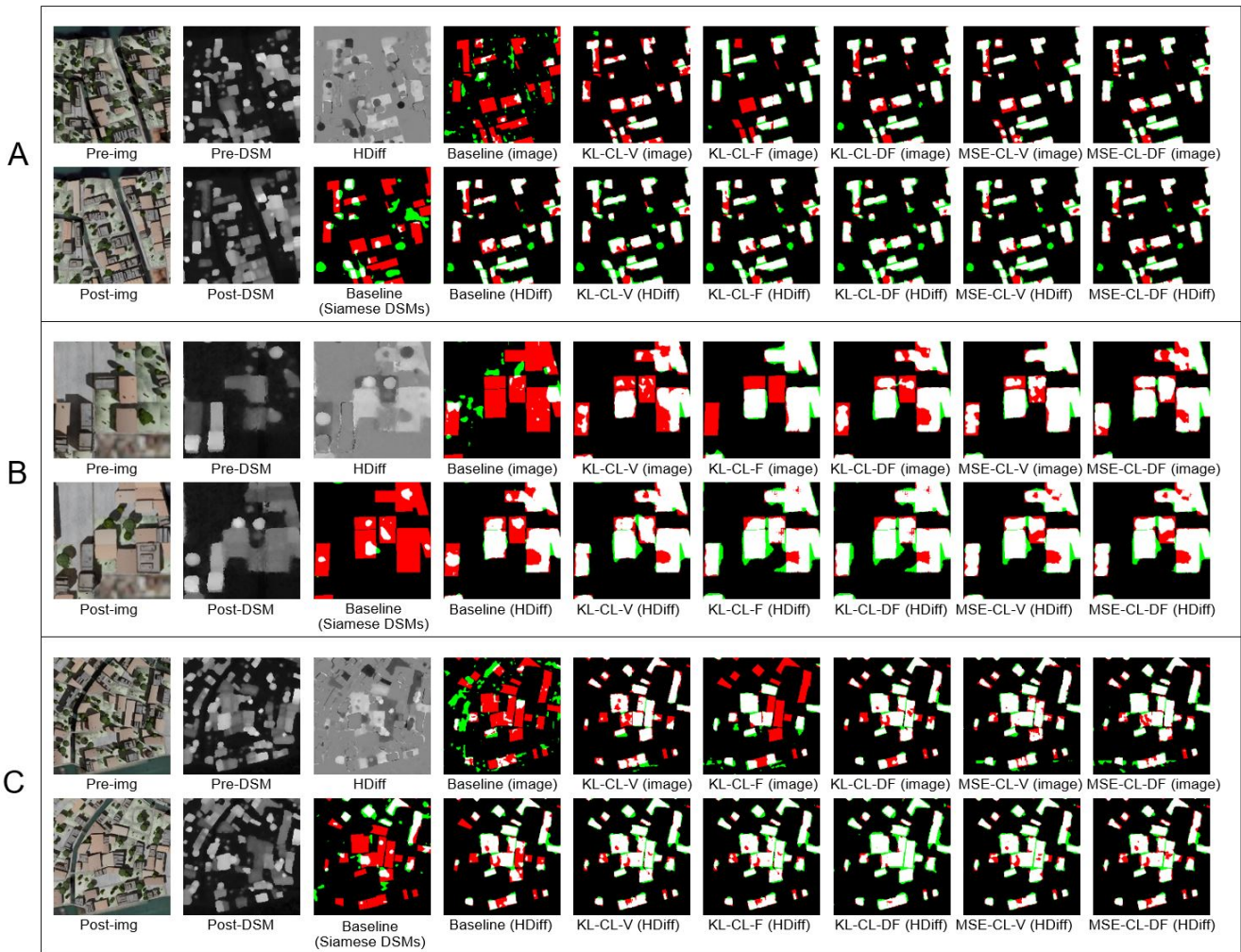


Fig. 5. Building change detection results of SParis→SVenice. Color legend: TP TN FN FP.

2.32% higher F1 and a 3.12% higher IoU compared with MSE-CL-V (D). For the DSM modality the best result achieved by MSE-CL-DF (R) has a 0.13% higher F1 and a 0.16% higher IoU compared with MSE-CL-DF (D).

According to the visualization results presented in Fig. 7, Baseline (D) is more prone to generate obvious false positives due to the outlier values in direct HDiff maps, especially as exemplified by the green clusters in A and B. As the robust height difference approach can filter out a portion of such outliers, Baseline (R) (using the same model with Baseline (D)) results contain fewer false positive pixels. Whether using Direct HDiff maps or Robust HDiff maps, the co-learning training approaches lead to significant improvements in image results by ResNet-50-BIT and HDiff results by SwinTransUNet. In Fig. 7 A, the results of robust HDiff maps with co-learning varieties are superior to those of direct HDiff maps with the same approach. In the results of direct HDiff maps, more building change pixels are wrongly recognized as unchanged pixels. In the image results, similar phenomena can be observed. MSE-CL-V (D/image), which achieves the

highest score among all co-learning varieties with direct HDiff maps, cannot recognize the change of a small building at the bottom border of A, while MSE-CL-V (R/image) is capable.

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

V. DISCUSSION

A. Domain Gaps in Different Modalities

Due to the differences in imaging sensors, capturing conditions, and preprocessing operations for the raw data, the

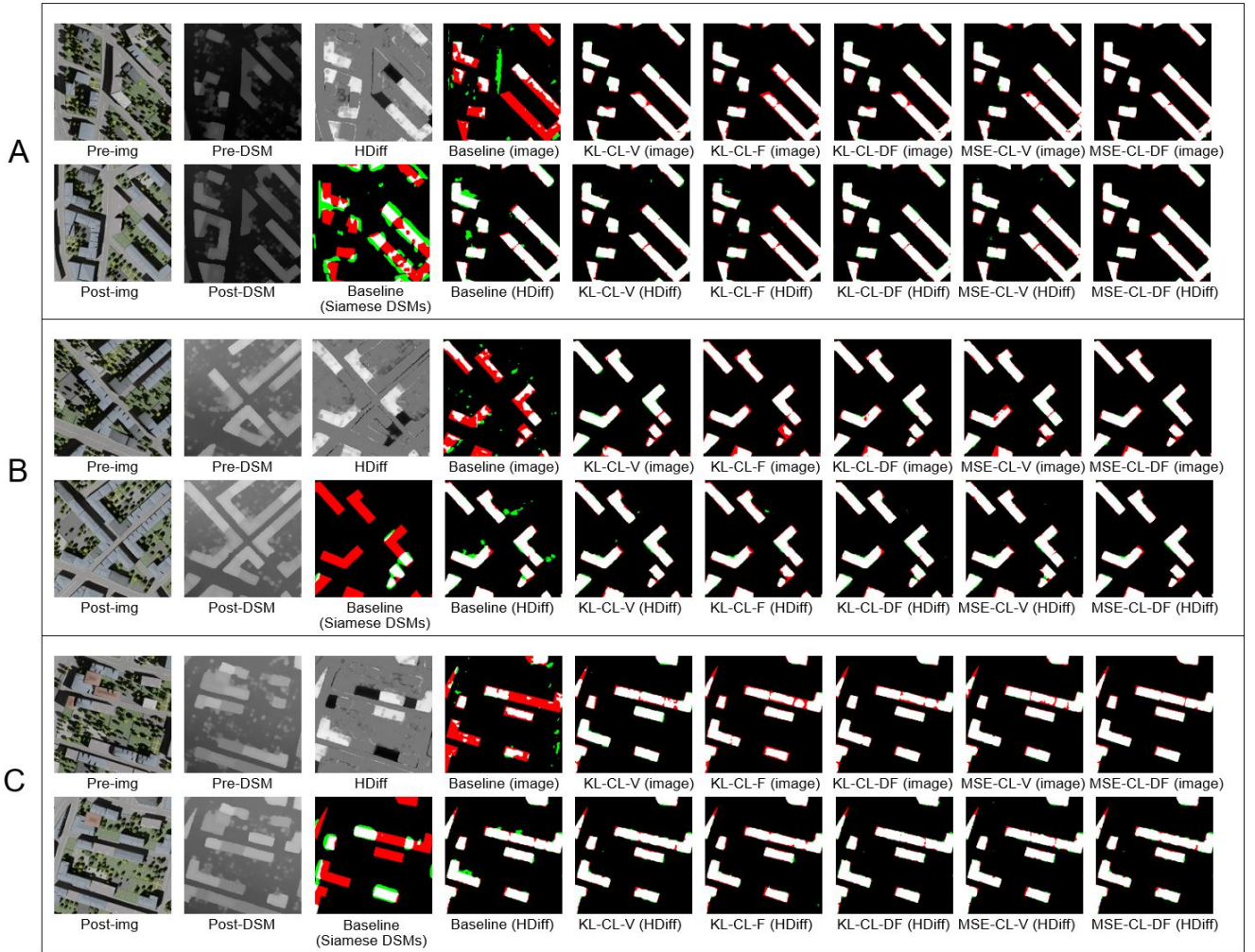


Fig. 6. Building change detection results of SVenice→SParis. Color legend: TP TN FN FP.

domain gaps of spectral distribution widely exist between different source and target datasets in remote sensing tasks [93]. Therefore, domain adaptation is becoming an essential topic.

This study presented the building change detection results of three baseline networks across three variants of two modalities: Siamese optical images, Siamese DSMs, and HDiff maps. Among the three paradigms, the HDiff maps demonstrate the most remarkable generalization ability in cross-domain scenarios. On the contrary, Siamese images and Siamese DSMs fail to produce reasonable results in our experiments. This phenomenon underscores the domain gap issues in these Siamese modalities, including synthetic→synthetic and synthetic→real cases, which is less pronounced in the HDiff maps for building change detection tasks. The superior cross-domain generalizability of HDiff maps can be attributed to its explicit geometric features, which excel in representing building changes. As a result, SwinTransUNet can learn robust knowledge and yield reasonable results in HDiff map single-modal learning mode. Nevertheless, the domain gaps of HDiff

maps between different synthetic data and those between synthetic and real data are different. Since the two sub-datasets of SMARS focus on urban scenes and have similar building geometry, the domain gaps in HDiff maps between them are not significant. The baseline method for HDiff maps can yield commendable results, with the F1 score of 79.66% in SParis→SVenice and 93.17% in SVenice→SParis. As mentioned in section IV-C SVenice has a higher building diversity than SParis, which causes the main difference in building changes between these two sub-datasets. Larger domain gaps exist between SMARS and Istanbul datasets. First, Istanbul data are derived from space borne WorldView-2 data that are under the influence of real-world capturing conditions, which could also lead to variation in the quality of DSMs. Second, the Istanbul dataset encompasses not only urban scenes but also suburban industrial areas, where the building and the building change characteristics differ from those in the urban scenes of SMARS. The aforementioned points present the challenge for cross-domain experiments as exemplified by the case C in Fig. 7.

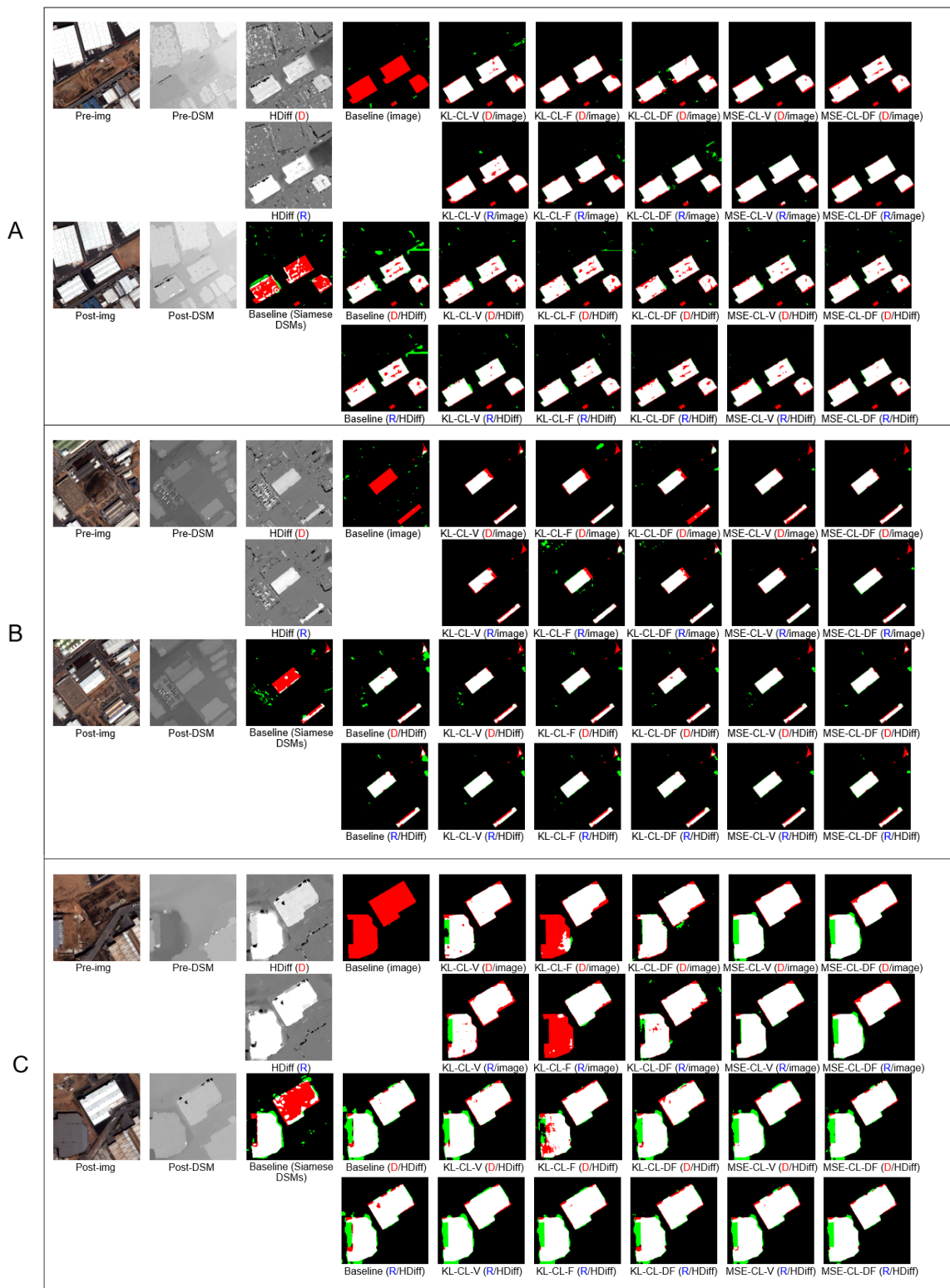


Fig. 7. Building change detection results of SMARS→Istanbul. Color legend: TP TN FN FP.

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

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

C. Multimodal Co-learning

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

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

The co-learning framework is versatile and easily extendable, allowing for integration with other multimodal learning

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

D. Efficiency and Computational Complexity

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

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

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

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

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

Utilizing the co-learning framework with bitemporal image and HDiff map modalities, four distinct models can be acquired: a single-modal Siamese image network, a single-modal HDiff map network, a co-learning-enhanced Siamese image network, and a co-learning-enhanced HDiff map network. This is especially useful when the training data and test data do not have the same modalities, which poses great constraints for the Siamese methods. In addition, the co-learning change

detection framework is flexible to extend. As depicted in Fig. 1, besides the change detection backbones for images and HDiff maps (comprising encoders and decoders), modules like the fusion operation in the Siamese network, the height difference operation for DSMs, and co-learning loss functions can be tailored for specific scenarios.

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

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

VI. CONCLUSION

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

In the future, we would like to investigate more multimodal learning methods for remote sensing tasks. Specifically

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

APPENDIX A

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

$$P_I^t = W_I^T X_I^t + b_I, \quad (24)$$

$$P_H^t = W_H^T X_H^t + b_H, \quad (25)$$

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

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

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

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

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

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

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

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

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

$$\frac{\partial \mathcal{L}_{CL-V}^I}{\partial X_I^t} = \frac{\partial \mathcal{L}_{CL-V}^I}{\partial P_I^t} \frac{\partial P_I^t}{\partial X_I^t}, \quad (29)$$

$$\frac{\partial \mathcal{L}_{CL-F}^I}{\partial X_I^t} = \frac{\partial \mathcal{L}_{CL-F}^I}{\partial P_I^t} \frac{\partial P_I^t}{\partial X_I^t}, \quad (30)$$

$$\frac{\partial \mathcal{L}_{CL-DF}^I}{\partial X_I^t} = \frac{\partial \mathcal{L}_{CL-DF}^I}{\partial P_I^t} \frac{\partial P_I^t}{\partial X_I^t}, \quad (31)$$

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

APPENDIX B

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

A. KL-divergence

When KL-divergence is employed as the co-learning loss function, \mathcal{L}_C^I for image modality is as follows:

$$\mathcal{L}_C^I = P_S^I \ln \frac{P_S^I}{P_I^t}, \quad (32)$$

where P_S^I is the shadow reference probability of the image modality. $P_S^I \in \{P_{H,d}^t, \frac{P_I^t + P_{H,d}^t}{2}, \frac{P_{I,d}^t + P_{H,d}^t}{2}\}$, depending on which co-learning combination is employed.

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

(1) Vanilla co-learning:

$$\mathcal{L}_{CL-V}^I = P_{H,d}^t \ln \frac{P_{H,d}^t}{P_I^t}, \quad (33)$$

so,

$$\begin{aligned} \frac{\partial \mathcal{L}_{CL-V}^I}{\partial X_I^t} &= \frac{\partial \mathcal{L}_{CL-V}^I}{\partial P_I^t} \frac{\partial P_I^t}{\partial X_I^t} \\ &= \frac{\partial P_{H,d}^t \ln \frac{P_{H,d}^t}{P_I^t}}{\partial P_I^t} \frac{\partial P_I^t}{\partial X_I^t} \\ &= -\frac{P_{H,d}^t}{P_I^t} \frac{\partial P_I^t}{\partial X_I^t}, \end{aligned} \quad (34)$$

(2) Fusion co-learning

$$\mathcal{L}_{CL-F}^I = \frac{P_I^t + P_{H,d}^t}{2} \ln \frac{P_I^t + P_{H,d}^t}{2P_I^t}, \quad (35)$$

so,

$$\begin{aligned} \frac{\partial \mathcal{L}_{CL-F}^I}{\partial X_I^t} &= \frac{\partial \mathcal{L}_{CL-F}^I}{\partial P_I^t} \frac{\partial P_I^t}{\partial X_I^t} \\ &= \frac{\partial \frac{P_I^t + P_{H,d}^t}{2} \ln \frac{P_I^t + P_{H,d}^t}{2P_I^t}}{\partial P_I^t} \frac{\partial P_I^t}{\partial X_I^t} \\ &= \frac{1}{2} [\ln (P_{H,d}^t + P_I^t) - \ln P_{H,d}^t \\ &\quad - \frac{P_{H,d}^t}{P_I^t} - \ln 2] \frac{\partial P_I^t}{\partial X_I^t}, \end{aligned} \quad (36)$$

(3) Detached fusion co-learning

$$\mathcal{L}_{CL-DF}^I = \frac{P_{I,d}^t + P_{H,d}^t}{2} \ln \frac{P_{I,d}^t + P_{H,d}^t}{2P_I^t}, \quad (37)$$

so,

$$\begin{aligned} \frac{\partial \mathcal{L}_{CL-DF}^I}{\partial X_I^t} &= \frac{\partial \mathcal{L}_{CL-DF}^I}{\partial P_I^t} \frac{\partial P_I^t}{\partial X_I^t} \\ &= \frac{\partial \frac{P_{I,d}^t + P_{H,d}^t}{2} \ln \frac{P_{I,d}^t + P_{H,d}^t}{2P_I^t}}{\partial P_I^t} \frac{\partial P_I^t}{\partial X_I^t} \\ &= -\frac{P_{I,d}^t + P_{H,d}^t}{2P_I^t} \frac{\partial P_I^t}{\partial X_I^t}, \end{aligned} \quad (38)$$

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

B. MSE

When MSE is employed as the co-learning loss function, \mathcal{L}_C^I for image modality is as follows:

$$\mathcal{L}_C^I = |P_I^t - P_S^I|^2, \quad (39)$$

where P_S^I is the shadow reference probability of the image modality. $P_S^I \in \{P_{H,d}^t, \frac{P_I^t + P_{H,d}^t}{2}, \frac{P_{I,d}^t + P_{H,d}^t}{2}\}$, depending on which co-learning combination is employed.

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

(1) Vanilla co-learning:

$$\mathcal{L}_{CL-V}^I = |P_I^t - P_{H,d}^t|^2, \quad (40)$$

so,

$$\begin{aligned} \frac{\partial \mathcal{L}_{CL-V}^I}{\partial X_I^t} &= \frac{\partial \mathcal{L}_{CL-V}^I}{\partial P_I^t} \frac{\partial P_I^t}{\partial X_I^t} \\ &= \frac{\partial |P_I^t - P_{H,d}^t|^2}{\partial P_I^t} \frac{\partial P_I^t}{\partial X_I^t} \\ &= 2(P_I^t - P_{H,d}^t) \frac{\partial P_I^t}{\partial X_I^t}, \end{aligned} \quad (41)$$

(2) Fusion co-learning

$$\begin{aligned} \mathcal{L}_{CL-F}^I &= |P_I^t - \frac{P_I^t + P_{H,d}^t}{2}|^2 \\ &= \frac{|P_I^t - P_{H,d}^t|^2}{4}, \end{aligned} \quad (42)$$

so,

$$\begin{aligned} \frac{\partial \mathcal{L}_{CL-F}^I}{\partial X_I^t} &= \frac{\partial \mathcal{L}_{CL-F}^I}{\partial P_I^t} \frac{\partial P_I^t}{\partial X_I^t} \\ &= \frac{\partial \frac{|P_I^t - P_{H,d}^t|^2}{4}}{\partial P_I^t} \frac{\partial P_I^t}{\partial X_I^t} \\ &= \frac{P_I^t - P_{H,d}^t}{4} \frac{\partial P_I^t}{\partial X_I^t}, \end{aligned} \quad (43)$$

(3) Detached fusion co-learning

$$\mathcal{L}_{CL-DF}^I = |P_I^t - \frac{P_{I,d}^t + P_{H,d}^t}{2}|^2, \quad (44)$$

so,

$$\begin{aligned} \frac{\partial \mathcal{L}_{CL-DF}^I}{\partial X_I} &= \frac{\partial \mathcal{L}_{CL-DF}^I}{\partial P_I} \frac{\partial P_I}{\partial X_I} \\ &= \frac{\partial |P_I^t - \frac{P_{I,d}^t + P_{H,d}^t}{2}|^2}{\partial P_I^t} \frac{\partial P_I^t}{\partial X_I^t} \quad (45) \\ &= (2P_I^t - P_{I,d}^t - P_{H,d}^t) \frac{\partial P_I^t}{\partial X_I^t}, \end{aligned}$$

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

ACKNOWLEDGMENT

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

REFERENCES

- [1] R. Qin, J. Tian, and P. Reinartz, "3d change detection—approaches and applications," *ISPRS Journal of Photogrammetry and Remote Sensing*, vol. 122, pp. 41–56, 2016.
- [2] I. R. Hegazy and M. R. Kaloop, "Monitoring urban growth and land use change detection with gis and remote sensing techniques in daqahlia governorate egypt," *International Journal of Sustainable Built Environment*, vol. 4, no. 1, pp. 117–124, 2015.
- [3] Z. Zheng, Y. Zhong, J. Wang, A. Ma, and L. Zhang, "Building damage assessment for rapid disaster response with a deep object-based semantic change detection framework: From natural disasters to man-made disasters," *Remote Sensing of Environment*, vol. 265, p. 112636, 2021.
- [4] Z. Ali, A. Tuladhar, and J. Zevenbergen, "An integrated approach for updating cadastral maps in pakistan using satellite remote sensing data," *International Journal of Applied Earth Observation and Geoinformation*, vol. 18, pp. 386–398, 2012.
- [5] D. Wen, X. Huang, F. Bovolo, J. Li, X. Ke, A. Zhang, and J. A. Benediktsson, "Change detection from very-high-spatial-resolution optical remote sensing images: Methods, applications, and future directions," *IEEE Geoscience and Remote Sensing Magazine*, vol. 9, no. 4, pp. 68–101, 2021.
- [6] A. A. Nielsen, K. Conradsen, and J. J. Simpson, "Multivariate alteration detection (mad) and maf postprocessing in multispectral, bitemporal image data: New approaches to change detection studies," *Remote Sensing of Environment*, vol. 64, no. 1, pp. 1–19, 1998.
- [7] J. Deng, K. Wang, Y. Deng, and G. Qi, "Pca-based land-use change detection and analysis using multitemporal and multisensor satellite data," *International Journal of Remote Sensing*, vol. 29, no. 16, pp. 4823–4838, 2008.
- [8] L. Bruzzone and D. F. Prieto, "An adaptive semiparametric and context-based approach to unsupervised change detection in multitemporal remote-sensing images," *IEEE Transactions on image processing*, vol. 11, no. 4, pp. 452–466, 2002.
- [9] T. Lei, J. Wang, H. Ning, X. Wang, D. Xue, Q. Wang, and A. K. Nandi, "Difference enhancement and spatial-spectral nonlocal network for change detection in vhr remote sensing images," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 60, pp. 1–13, 2021.
- [10] L. Bruzzone and F. Bovolo, "A novel framework for the design of change-detection systems for very-high-resolution remote sensing images," *Proceedings of the IEEE*, vol. 101, no. 3, pp. 609–630, 2012.
- [11] Z. Lei, T. Fang, H. Huo, and D. Li, "Bi-temporal texton forest for land cover transition detection on remotely sensed imagery," *IEEE Transactions on Geoscience and remote sensing*, vol. 52, no. 2, pp. 1227–1237, 2013.
- [12] F. Bovolo, L. Bruzzone, and M. Marconcini, "A novel approach to unsupervised change detection based on a semisupervised svm and a similarity measure," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 46, no. 7, pp. 2070–2082, 2008.
- [13] C. Wu, L. Zhang, and L. Zhang, "A scene change detection framework for multi-temporal very high resolution remote sensing images," *Signal Processing*, vol. 124, pp. 184–197, 2016.
- [14] K. J. Wessels, F. Van den Bergh, D. P. Roy, B. P. Salmon, K. C. Steenkamp, B. MacAlister, D. Swanepoel, and D. Jewitt, "Rapid land cover map updates using change detection and robust random forest classifiers," *Remote sensing*, vol. 8, no. 11, p. 888, 2016.
- [15] H. Nemmour and Y. Chibani, "Multiple support vector machines for land cover change detection: An application for mapping urban extensions," *ISPRS Journal of Photogrammetry and Remote Sensing*, vol. 61, no. 2, pp. 125–133, 2006.
- [16] L. Zhou, G. Cao, Y. Li, and Y. Shang, "Change detection based on conditional random field with region connection constraints in high-resolution remote sensing images," *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, vol. 9, no. 8, pp. 3478–3488, 2016.
- [17] T. Kasetkasem and P. K. Varshney, "An image change detection algorithm based on markov random field models," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 40, no. 8, pp. 1815–1823, 2002.
- [18] W. Gu, Z. Lv, and M. Hao, "Change detection method for remote sensing images based on an improved markov random field," *Multimedia Tools and Applications*, vol. 76, pp. 17719–17734, 2017.
- [19] G. Cao, L. Zhou, and Y. Li, "A new change-detection method in high-resolution remote sensing images based on a conditional random field model," *International Journal of Remote Sensing*, vol. 37, no. 5, pp. 1173–1189, 2016.
- [20] X. X. Zhu, D. Tuia, L. Mou, G.-S. Xia, L. Zhang, F. Xu, and F. Fraundorfer, "Deep learning in remote sensing: A comprehensive review and list of resources," *IEEE geoscience and remote sensing magazine*, vol. 5, no. 4, pp. 8–36, 2017.
- [21] W. Shi, M. Zhang, R. Zhang, S. Chen, and Z. Zhan, "Change detection based on artificial intelligence: State-of-the-art and challenges," *Remote Sensing*, vol. 12, no. 10, p. 1688, 2020.
- [22] A. Shafique, G. Cao, Z. Khan, M. Asad, and M. Aslam, "Deep learning-based change detection in remote sensing images: A review," *Remote Sensing*, vol. 14, no. 4, p. 871, 2022.
- [23] H. Jiang, M. Peng, Y. Zhong, H. Xie, Z. Hao, J. Lin, X. Ma, and X. Hu, "A survey on deep learning-based change detection from high-resolution remote sensing images," *Remote Sensing*, vol. 14, no. 7, p. 1552, 2022.
- [24] X. Wu, D. Hong, and J. Chanussot, "Uiu-net: U-net in u-net for infrared small object detection," *IEEE Transactions on Image Processing*, vol. 32, pp. 364–376, 2022.
- [25] C. Li, B. Zhang, D. Hong, J. Yao, and J. Chanussot, "Lrr-net: An interpretable deep unfolding network for hyperspectral anomaly detection," *IEEE Transactions on Geoscience and Remote Sensing*, 2023.
- [26] L. Ma, Y. Liu, X. Zhang, Y. Ye, G. Yin, and B. A. Johnson, "Deep learning in remote sensing applications: A meta-analysis and review," *ISPRS journal of photogrammetry and remote sensing*, vol. 152, pp. 166–177, 2019.
- [27] L. Alzubaidi, J. Zhang, A. J. Humaidi, A. Al-Dujaili, Y. Duan, O. Al-Shamma, J. Santamaría, M. A. Fadhel, M. Al-Amidie, and L. Farhan, "Review of deep learning: Concepts, cnn architectures, challenges, applications, future directions," *Journal of big Data*, vol. 8, pp. 1–74, 2021.
- [28] J. Zhao, M. Gong, J. Liu, and L. Jiao, "Deep learning to classify difference image for image change detection," in *2014 International Joint Conference on Neural Networks (IJCNN)*. IEEE, 2014, pp. 411–417.
- [29] N. Lv, C. Chen, T. Qiu, and A. K. Sangaiah, "Deep learning and superpixel feature extraction based on contractive autoencoder for change detection in sar images," *IEEE transactions on industrial informatics*, vol. 14, no. 12, pp. 5530–5538, 2018.
- [30] H. Chen, Z. Qi, and Z. Shi, "Remote sensing image change detection with transformers," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 60, pp. 1–14, 2021.
- [31] Y. Zhan, K. Fu, M. Yan, X. Sun, H. Wang, and X. Qiu, "Change detection based on deep siamese convolutional network for optical aerial images," *IEEE Geoscience and Remote Sensing Letters*, vol. 14, no. 10, pp. 1845–1849, 2017.
- [32] R. C. Daudt, B. Le Saux, and A. Boulch, "Fully convolutional siamese networks for change detection," in *2018 25th IEEE International Conference on Image Processing (ICIP)*. IEEE, 2018, pp. 4063–4067.
- [33] C. Zhang, P. Yue, D. Tapete, L. Jiang, B. Shanguan, L. Huang, and G. Liu, "A deeply supervised image fusion network for change

- detection in high resolution bi-temporal remote sensing images,” *ISPRS Journal of Photogrammetry and Remote Sensing*, vol. 166, pp. 183–200, 2020.
- [34] H. Chen, C. Wu, B. Du, and L. Zhang, “Dsdanet: Deep siamese domain adaptation convolutional neural network for cross-domain change detection,” *arXiv preprint arXiv:2006.09225*, 2020.
- [35] Q. Shi, M. Liu, S. Li, X. Liu, F. Wang, and L. Zhang, “A deeply supervised attention metric-based network and an open aerial image dataset for remote sensing change detection,” *IEEE transactions on geoscience and remote sensing*, vol. 60, pp. 1–16, 2021.
- [36] S. Fang, K. Li, J. Shao, and Z. Li, “Snunet-cd: A densely connected siamese network for change detection of vhr images,” *IEEE Geoscience and Remote Sensing Letters*, vol. 19, pp. 1–5, 2021.
- [37] G. Cheng, G. Wang, and J. Han, “Isnet: Towards improving separability for remote sensing image change detection,” *IEEE Transactions on Geoscience and Remote Sensing*, vol. 60, pp. 1–11, 2022.
- [38] Z. Chen, Y. Zhou, B. Wang, X. Xu, N. He, S. Jin, and S. Jin, “Egde-net: A building change detection method for high-resolution remote sensing imagery based on edge guidance and differential enhancement,” *ISPRS Journal of Photogrammetry and Remote Sensing*, vol. 191, pp. 203–222, 2022.
- [39] Z. Li, C. Yan, Y. Sun, and Q. Xin, “A densely attentive refinement network for change detection based on very-high-resolution bitemporal remote sensing images,” *IEEE Transactions on Geoscience and Remote Sensing*, vol. 60, pp. 1–18, 2022.
- [40] R. Zhang, H. Zhang, X. Ning, X. Huang, J. Wang, and W. Cui, “Global-aware siamese network for change detection on remote sensing images,” *ISPRS Journal of Photogrammetry and Remote Sensing*, vol. 199, pp. 61–72, 2023.
- [41] A. Dosovitskiy, L. Beyer, A. Kolesnikov, D. Weissenborn, X. Zhai, T. Unterthiner, M. Dehghani, M. Minderer, G. Heigold, S. Gelly *et al.*, “An image is worth 16x16 words: Transformers for image recognition at scale,” *arXiv preprint arXiv:2010.11929*, 2020.
- [42] Z. Liu, Y. Lin, Y. Cao, H. Hu, Y. Wei, Z. Zhang, S. Lin, and B. Guo, “Swin transformer: Hierarchical vision transformer using shifted windows,” in *Proceedings of the IEEE/CVF international conference on computer vision*, 2021, pp. 10 012–10 022.
- [43] W. G. C. Bandara and V. M. Patel, “A transformer-based siamese network for change detection,” in *IGARSS 2022-2022 IEEE International Geoscience and Remote Sensing Symposium*. IEEE, 2022, pp. 207–210.
- [44] C. Zhang, L. Wang, S. Cheng, and Y. Li, “Swinsunet: Pure transformer network for remote sensing image change detection,” *IEEE Transactions on Geoscience and Remote Sensing*, vol. 60, pp. 1–13, 2022.
- [45] Q. Li, R. Zhong, X. Du, and Y. Du, “Transunetcd: A hybrid transformer network for change detection in optical remote-sensing images,” *IEEE Transactions on Geoscience and Remote Sensing*, vol. 60, pp. 1–19, 2022.
- [46] D. Hong, B. Zhang, X. Li, Y. Li, C. Li, J. Yao, N. Yokoya, H. Li, X. Jia, A. Plaza *et al.*, “Spectralgpt: Spectral foundation model,” *arXiv preprint arXiv:2311.07113*, 2023.
- [47] K. Li, X. Cao, and D. Meng, “A new learning paradigm for foundation model-based remote sensing change detection,” *arXiv preprint arXiv:2312.01163*, 2023.
- [48] X. Guo, J. Lao, B. Dang, Y. Zhang, L. Yu, L. Ru, L. Zhong, Z. Huang, X. Wu, D. Hu *et al.*, “Skysense: A multi-modal remote sensing foundation model towards universal interpretation for earth observation imagery,” *arXiv preprint arXiv:2312.10115*, 2023.
- [49] D. Hong, B. Zhang, H. Li, Y. Li, J. Yao, C. Li, M. Werner, J. Chanussot, A. Zipf, and X. X. Zhu, “Cross-city matters: A multimodal remote sensing benchmark dataset for cross-city semantic segmentation using high-resolution domain adaptation networks,” *Remote Sensing of Environment*, vol. 299, p. 113856, 2023.
- [50] P. J. S. Vega, G. A. O. P. da Costa, R. Q. Feitosa, M. X. O. Adarme, C. A. de Almeida, C. Heipke, and F. Rottensteiner, “An unsupervised domain adaptation approach for change detection and its application to deforestation mapping in tropical biomes,” *ISPRS Journal of Photogrammetry and Remote Sensing*, vol. 181, pp. 113–128, 2021.
- [51] Y. Xie, J. Tian, and X. X. Zhu, “A co-learning method to utilize optical images and photogrammetric point clouds for building extraction,” *International Journal of Applied Earth Observation and Geoinformation*, vol. 116, p. 103165, 2023.
- [52] M. Turker and B. Cetinkaya, “Automatic detection of earthquake-damaged buildings using dems created from pre-and post-earthquake stereo aerial photographs,” *International Journal of Remote Sensing*, vol. 26, no. 4, pp. 823–832, 2005.
- [53] L. Zhu, H. Shimamura, K. Tachibana, Y. Li, and P. Gong, “Building change detection based on object extraction in dense urban areas,” *International Archives of the Photogrammetry, Remote Sensing and Spatial Information Sciences*, 2008.
- [54] F. Jung, “Detecting building changes from multitemporal aerial stereopairs,” *ISPRS Journal of Photogrammetry and Remote Sensing*, vol. 58, no. 3-4, pp. 187–201, 2004.
- [55] A. Sasagawa, E. Baltsavias, S. Kocaman-Aksakal, and J. D. Wegner, “Investigation on automatic change detection using pixel-changes and dsm-changes with alos-prism triplet images,” *International archives of the photogrammetry, remote sensing and spatial information sciences*, vol. 40, no. 7/W2, pp. 213–217, 2013.
- [56] J. Tian, H. Chaabouni-Chouayakh, and P. Reinartz, “3d building change detection from high resolution spaceborne stereo imagery,” in *2011 International Workshop on Multi-Platform/Multi-Sensor Remote Sensing and Mapping*. IEEE, 2011, pp. 1–7.
- [57] G. Dini, K. Jacobsen, F. Rottensteiner, M. Al Rajhi, and C. Heipke, “3d building change detection using high resolution stereo images and a gis database,” *The International Archives of the Photogrammetry, Remote Sensing and Spatial Information Sciences; XXXIX-B7*, vol. 39, pp. 299–304, 2012.
- [58] Y. Xie and J. Tian, “Multimodal co-learning: A domain adaptation method for building extraction from optical remote sensing imagery,” in *2023 Joint Urban Remote Sensing Event (JURSE)*. IEEE, 2023, pp. 1–4.
- [59] Y. Xie, K. Schindler, J. Tian, and X. X. Zhu, “Exploring cross-city semantic segmentation of als point clouds,” *International Archives of the Photogrammetry, Remote Sensing and Spatial Information Sciences*, vol. 43, pp. 247–254, 2021.
- [60] X. Yuan, J. Tian, and P. Reinartz, “Building change detection based on deep learning and belief function,” in *2019 Joint Urban Remote Sensing Event (JURSE)*. IEEE, 2019, pp. 1–4.
- [61] A. Rahate, R. Walambe, S. Ramanna, and K. Kotecha, “Multimodal co-learning: Challenges, applications with datasets, recent advances and future directions,” *Information Fusion*, vol. 81, pp. 203–239, 2022.
- [62] T. Baltrušaitis, C. Ahuja, and L.-P. Morency, “Multimodal machine learning: A survey and taxonomy,” *IEEE transactions on pattern analysis and machine intelligence*, vol. 41, no. 2, pp. 423–443, 2018.
- [63] Y. Wang, Y. Wan, Y. Zhang, B. Zhang, and Z. Gao, “Imbalance knowledge-driven multi-modal network for land-cover semantic segmentation using aerial images and lidar point clouds,” *ISPRS Journal of Photogrammetry and Remote Sensing*, vol. 202, pp. 385–404, 2023.
- [64] M. J. Westoby, J. Brasington, N. F. Glasser, M. J. Hambrey, and J. M. Reynolds, “‘structure-from-motion’ photogrammetry: A low-cost, effective tool for geoscience applications,” *Geomorphology*, vol. 179, pp. 300–314, 2012.
- [65] S. Gehrke, K. Morin, M. Downey, N. Boehrer, and T. Fuchs, “Semi-global matching: An alternative to lidar for dsm generation,” in *Proceedings of the 2010 Canadian Geomatics Conference and Symposium of Commission I*, vol. 2, no. 6, 2010.
- [66] R. Perko, H. Raggam, and P. M. Roth, “Mapping with pléiades—end-to-end workflow,” *Remote Sensing*, vol. 11, no. 17, p. 2052, 2019.
- [67] H. A. Al-Najjar, B. Kalantar, B. Pradhan, V. Saeidi, A. A. Halin, N. Ueda, and S. Mansor, “Land cover classification from fused dsm and uav images using convolutional neural networks,” *Remote Sensing*, vol. 11, no. 12, p. 1461, 2019.
- [68] M. Á. Aguilar, M. del Mar Saldaña, and F. J. Aguilar, “Generation and quality assessment of stereo-extracted dsm from geoeye-1 and worldview-2 imagery,” *IEEE Transactions on Geoscience and Remote Sensing*, vol. 52, no. 2, pp. 1259–1271, 2013.
- [69] P. d’Angelo and J. Tian, “Geometric evaluation of gaofen-7 stereo data,” *ISPRS Annals of the Photogrammetry, Remote Sensing and Spatial Information Sciences*, vol. 10, pp. 805–811, 2023.
- [70] N. Audebert, B. Le Saux, and S. Lefèvre, “Beyond rgb: Very high resolution urban remote sensing with multimodal deep networks,” *ISPRS journal of photogrammetry and remote sensing*, vol. 140, pp. 20–32, 2018.
- [71] H. Hosseinpour, F. Samadzadegan, and F. D. Javan, “Cmgfnet: A deep cross-modal gated fusion network for building extraction from very high-resolution remote sensing images,” *ISPRS journal of photogrammetry and remote sensing*, vol. 184, pp. 96–115, 2022.
- [72] M. Fuentes Reyes, Y. Xie, X. Yuan, P. d’Angelo, F. Kurz, D. Cerra, and J. Tian, “A 2d/3d multimodal data simulation approach with applications on urban semantic segmentation, building extraction and change detection,” *ISPRS Journal of Photogrammetry and Remote Sensing*, vol. 205, pp. 74–97, 2023.

- [73] M. Schmitt and X. X. Zhu, "Data fusion and remote sensing: An ever-growing relationship," *IEEE Geoscience and Remote Sensing Magazine*, vol. 4, no. 4, pp. 6–23, 2016.
- [74] J. Huang, X. Zhang, Q. Xin, Y. Sun, and P. Zhang, "Automatic building extraction from high-resolution aerial images and lidar data using gated residual refinement network," *ISPRS journal of photogrammetry and remote sensing*, vol. 151, pp. 91–105, 2019.
- [75] T. Peters, C. Brenner, and K. Schindler, "Semantic segmentation of mobile mapping point clouds via multi-view label transfer," *ISPRS Journal of Photogrammetry and Remote Sensing*, vol. 202, pp. 30–39, 2023.
- [76] S. Huang, M. Usvyatsov, and K. Schindler, "Indoor scene recognition in 3d," in *2020 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*. IEEE, 2020, pp. 8041–8048.
- [77] S. Bachhofner, A.-M. Loghin, J. Otepka, N. Pfeifer, M. Hornacek, A. Siposova, N. Schmidinger, K. Hornik, N. Schiller, O. Kähler *et al.*, "Generalized sparse convolutional neural networks for semantic segmentation of point clouds derived from tri-stereo satellite imagery," *Remote Sensing*, vol. 12, no. 8, p. 1289, 2020.
- [78] C. Hazirbas, L. Ma, C. Domokos, and D. Cremers, "Fusenet: Incorporating depth into semantic segmentation via fusion-based cnn architecture," in *Computer Vision—ACCV 2016: 13th Asian Conference on Computer Vision, Taipei, Taiwan, November 20–24, 2016, Revised Selected Papers, Part I 13*. Springer, 2017, pp. 213–228.
- [79] K. Bittner, F. Adam, S. Cui, M. Körner, and P. Reinartz, "Building footprint extraction from vhr remote sensing images combined with normalized dsms using fused fully convolutional networks," *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, vol. 11, no. 8, pp. 2615–2629, 2018.
- [80] Y.-C. Li, H.-C. Li, W.-S. Hu, and H.-L. Yu, "Dspcnet: Dual-channel scale-aware segmentation network with position and channel attentions for high-resolution aerial images," *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, vol. 14, pp. 8552–8565, 2021.
- [81] P. Zhang, P. Du, C. Lin, X. Wang, E. Li, Z. Xue, and X. Bai, "A hybrid attention-aware fusion network (hafnet) for building extraction from high-resolution imagery and lidar data," *Remote Sensing*, vol. 12, no. 22, p. 3764, 2020.
- [82] S. Zhou, Y. Feng, S. Li, D. Zheng, F. Fang, Y. Liu, and B. Wan, "Dsm-assisted unsupervised domain adaptive network for semantic segmentation of remote sensing imagery," *IEEE Transactions on Geoscience and Remote Sensing*, 2023.
- [83] J. Tian, S. Cui, and P. Reinartz, "Building change detection based on satellite stereo imagery and digital surface models," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 52, no. 1, pp. 406–417, 2013.
- [84] J. Tian and J. Dezert, "Fusion of multispectral imagery and dsms for building change detection using belief functions and reliabilities," *International Journal of Image and Data Fusion*, vol. 10, no. 1, pp. 1–27, 2019.
- [85] S. Tian, Y. Zhong, A. Ma, and L. Zhang, "Three-dimensional change detection in urban areas based on complementary evidence fusion," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 60, pp. 1–13, 2021.
- [86] H. Wang, X. Lv, K. Zhang, and B. Guo, "Building change detection based on 3d co-segmentation using satellite stereo imagery," *Remote Sensing*, vol. 14, no. 3, p. 628, 2022.
- [87] Q. Li, Y. Shi, S. Auer, R. Roschlaub, K. Möst, M. Schmitt, C. Glock, and X. Zhu, "Detection of undocumented building constructions from official geodata using a convolutional neural network," *Remote Sensing*, vol. 12, no. 21, p. 3537, 2020.
- [88] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2016, pp. 770–778.
- [89] Y. Cui, R. Chen, W. Chu, L. Chen, D. Tian, Y. Li, and D. Cao, "Deep learning for image and point cloud fusion in autonomous driving: A review," *IEEE Transactions on Intelligent Transportation Systems*, vol. 23, no. 2, pp. 722–739, 2021.
- [90] J. Tian, P. Reinartz, P. d'Angelo, and M. Ehlers, "Region-based automatic building and forest change detection on cartosat-1 stereo imagery," *ISPRS Journal of Photogrammetry and Remote Sensing*, vol. 79, pp. 226–239, 2013.
- [91] P. d'Angelo, "Improving semi-global matching: cost aggregation and confidence measure," *The International Archives of the Photogrammetry, Remote Sensing and Spatial Information Sciences*, vol. 41, pp. 299–304, 2016.
- [92] A. Paszke, S. Gross, S. Chintala, G. Chanan, E. Yang, Z. DeVito, Z. Lin, A. Desmaison, L. Antiga, and A. Lerer, "Automatic differentiation in pytorch," in *NIPS-W*, 2017.
- [93] D. Hong, N. Yokoya, J. Chanussot, and X. X. Zhu, "An augmented linear mixing model to address spectral variability for hyperspectral unmixing," *IEEE Transactions on Image Processing*, vol. 28, no. 4, pp. 1923–1938, 2018.
- [94] L. Zhang, M. Lan, J. Zhang, and D. Tao, "Stagewise unsupervised domain adaptation with adversarial self-training for road segmentation of remote-sensing images," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 60, pp. 1–13, 2021.
- [95] X.-Y. Tong, G.-S. Xia, Q. Lu, H. Shen, S. Li, S. You, and L. Zhang, "Land-cover classification with high-resolution remote sensing images using transferable deep models," *Remote Sensing of Environment*, vol. 237, p. 111322, 2020.
- [96] Z. Huang, Q. Liu, H. Zhou, G. Gao, T. Xu, Q. Wen, and Y. Wang, "Building detection from panchromatic and multispectral images with dual-stream asymmetric fusion networks," *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, vol. 16, pp. 3364–3377, 2023.
- [97] E. J. Hoffmann, K. Abdulahhad, and X. X. Zhu, "Using social media images for building function classification," *Cities*, vol. 133, p. 104107, 2023.
- [98] G. Kyriakaki, A. Doulamis, N. Doulamis, M. Ioannides, K. Makantasis, E. Protopapadakis, A. Hadjiprocopis, K. Wenzel, D. Fritsch, M. Klein *et al.*, "4d reconstruction of tangible cultural heritage objects from web-retrieved images," *International Journal of Heritage in the Digital Era*, vol. 3, no. 2, pp. 431–451, 2014.
- [99] J. E. Van Engelen and H. H. Hoos, "A survey on semi-supervised learning," *Machine learning*, vol. 109, no. 2, pp. 373–440, 2020.
- [100] Y. Wang, C. M. Albrecht, N. A. A. Braham, L. Mou, and X. X. Zhu, "Self-supervised learning in remote sensing: A review," *IEEE Geoscience and Remote Sensing Magazine*, vol. 10, no. 4, pp. 213–247, 2022.
- [101] V. Stojnic and V. Risojevic, "Self-supervised learning of remote sensing scene representations using contrastive multiview coding," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2021, pp. 1182–1191.
- [102] E. Protopapadakis, A. Doulamis, N. Doulamis, and E. Maltezos, "Stacked autoencoders driven by semi-supervised learning for building extraction from near infrared remote sensing imagery," *Remote Sensing*, vol. 13, no. 3, p. 371, 2021.
- [103] D. Hong, L. Gao, R. Hang, B. Zhang, and J. Chanussot, "Deep encoder–decoder networks for classification of hyperspectral and lidar data," *IEEE Geoscience and Remote Sensing Letters*, vol. 19, p. 5500205, 2020.
- [104] Z. Qiu, H. Shen, L. Yue, and G. Zheng, "Cross-sensor remote sensing imagery super-resolution via an edge-guided attention-based network," *ISPRS Journal of Photogrammetry and Remote Sensing*, vol. 199, pp. 226–241, 2023.