

Deep Learning for Building Footprint Generation from Optical Imagery

Qingyu Li

Vollständiger Abdruck der von der TUM School of Engineering and Design der
Technischen Universität München zur Erlangung einer

Doktorin der Ingenieurwissenschaften (Dr.-Ing.)

genehmigten Dissertation.

Vorsitz: Prof. Dr.-Ing. Liqiu Meng

Prüfer*innen der Dissertation:

1. Prof. Dr.-Ing. habil. Xiao Xiang Zhu
2. Prof. Dr.-Ing. Lichao Mou
3. Prof. Dr.techn. Friedrich Fraundorfer,
Technische Universität Graz

Die Dissertation wurde am 09.06.2022 bei der Technischen Universität München
eingereicht und durch die TUM School of Engineering and Design am 10.10.2022
angenommen.

Abstract

Due to rapid urban expansion and city renewal, urban areas undergo significant changes every year. There are many new buildings being constructed in the former non-urban land, leading to adverse effects on the environment and ecology. An insight into urban development can be gained from building footprint maps. Hence, building footprint maps are essential for environmentally sustainable urbanization. Furthermore, well-established building footprint maps facilitate a wide range of applications, such as urban planning and monitoring, as well as disaster management. Remote sensing technologies with high-resolution imaging sensors provide great potential for generating building footprint maps in recent decades. The collection of high resolution (HR) and very high resolution (VHR) optical imagery allows for spatial-temporal monitoring of buildings. Therefore, this dissertation focuses on the task of building footprint generation using HR and VHR optical imagery.

During the past few years, deep learning-based methods, such as convolutional neural networks (CNNs) have contributed significantly to the task of building footprint generation. Deep learning-based methods have shown promising results for this task in terms of accuracy and efficiency, but they have two inherent limitations. **First, the extracted buildings show blurred building boundaries and blob shapes. Second, deep learning-based methods require a lot of annotated labels for network training.**

This dissertation has developed several methods to address the above issues:

- **To refine blurred building boundaries:** To preserve sharp boundaries and fine-grained segmentation, an end-to-end building footprint generation approach is developed, which integrates a convolution neural network (CNN) and graph model. A novel network that learns an attraction field map is proposed, enhancing building boundaries and suppressing the impact of background clutters.
- **To compensate for the limited supervisory information:** With the use of a large amount of labeled data in other cities, a co-segmentation learning pipeline is proposed to boost the performance of a model in the target cities. Based on consistency training, a semi-supervised network is developed to leverage a large amount of unlabeled data, improving the performance of a model.

Finally, **the developed methods are implemented in practical applications** to demonstrate that they can provide building footprint maps for urban planning and monitoring. Furthermore, sampling strategies for developed deep learning methods that aim to reduce training data size are investigated.

Zusammenfassung

Aufgrund der rasanten Stadterweiterung und Stadterneuerung unterliegen städtische Gebiete jedes Jahr erheblichen Veränderungen. Auf der ehemaligen unbewohnten Insel werden viele neue Gebäude errichtet, die zu negativen Auswirkungen auf Umwelt und Ökologie führen. Einen Einblick in die Stadtentwicklung können Gebäudegrundrisskarten gewinnen. Daher sind Gebäudegrundrisskarten für eine umweltverträgliche Urbanisierung unerlässlich. Darüber hinaus ermöglichen die etablierten Gebäudegrundrisskarten eine Vielzahl von Anwendungen, wie z. B. Stadtplanung und -überwachung sowie Katastrophenmanagement. Fernerkundungstechnologien mit hochauflösenden Bildsensoren bieten in den letzten Jahrzehnten ein großes Potenzial für die Erstellung von Gebäudegrundrisskarten. Die Erfassung von hochauflösenden (HR) und sehr hochauflösenden (VHR) optischen Bildern ermöglicht die räumlich-zeitliche Überwachung von Gebäuden. Daher konzentriert sich diese Dissertation auf die Aufgabe der Erstellung von Gebäudegrundrissen unter Verwendung optischer HR- und VHR-Bilder.

In den letzten Jahren haben auf Deep Learning basierende Methoden wie Convolutional Neural Networks (CNNs) erheblich zur Aufgabe der Gebäude-Footprint-Generierung beigetragen. Auf Deep Learning basierende Methoden haben für diese Aufgabe in Bezug auf Genauigkeit und Effizienz vielversprechende Ergebnisse gezeigt, aber sie haben zwei inhärente Einschränkungen. **Erstens zeigen die extrahierten Gebäude verschwommene Gebäudegrenzen. Zweitens erfordern auf Deep Learning basierende Methoden viele gekennzeichneten Etiketten für das Netzwerktraining.**

Diese Dissertation hat einige Methoden entwickelt, um die oben genannten Probleme anzugehen:

- **Um unscharfe Gebäudegrenzen zu verfeinern:** Um scharfe Grenzen und eine feinkörnige Segmentierung beizubehalten, wird ein End-to-End-Ansatz zur Erstellung von Gebäudegrundrissen entwickelt, der ein Convolution Neural Network (CNN) und ein Graphenmodell integriert. Ein neuartiges Netzwerk, das eine Attraktionsfeldkarte lernt, wird vorgeschlagen, um Gebäudegrenzen zu verbessern und die Auswirkungen von Hintergrundstörungen zu unterdrücken.
- **Um die begrenzten Überwachungsinformationen zu kompensieren:** Unter Verwendung einer großen Menge gekennzeichneteter Daten in anderen Städten wird eine Co-Segmentierungs-Lernpipeline vorgeschlagen, um die Leistung eines Modells in den Zielstädten zu steigern. Basierend auf einem Konsistenztraining wird ein halbüberwachtes Netzwerk entwickelt, um eine große Menge nicht gekennzeichneteter Daten zu nutzen und die Leistung eines Modells zu verbessern.

Schließlich werden **die entwickelten Methoden in praktische Anwendungen implementiert**, um zu demonstrieren, dass sie Gebäudegrundrisskarten für die Stadtplanung und -überwachung bereitstellen können. Darüber hinaus werden Sampling-Strategien für

Zusammenfassung

entwickelte Deep-Learning-Methoden untersucht, die darauf abzielen, die Größe der Trainingsdaten zu reduzieren.

Contents

Abstract	iii
Zusammenfassung	v
Acronyms	xi
1 Introduction	1
1.1 Motivations and Objectives	1
1.2 Dissertation Outline	2
2 Basics	5
2.1 Building Footprint Generation	5
2.2 Remote Sensing Data for Building Footprint Generation	6
3 Related Work	9
3.1 Early Efforts in Building Footprint Generation	9
3.1.1 Geometrical Primitive-based Methods	9
3.1.2 Index-based Methods	9
3.1.3 Oversegmentation-based Methods	10
3.1.4 Classifier-based Methods	11
3.2 Deep Learning Techniques for Building Footprint Generation	12
3.2.1 Corner-based Methods	12
3.2.2 Boundary-based Methods	12
3.2.3 Semantic Mask-based Methods	13
3.2.3.1 Multi-scale Information Aggregation	14
3.2.3.2 Building Boundary Refinement	14
3.2.3.3 Limited Supervisory Information Compensation	15
3.2.3.4 Computational Complexity Reduction	16
4 Deep Learning Methods to Refine Blurred Building Boundaries	19
4.1 Feature Pairwise Conditional Random Field	19
4.1.1 Motivation	19
4.1.2 Methodology	20
4.2 Attraction Field Representation	21
4.2.1 Motivation	21
4.2.2 Methodology	21
4.3 Summary	23
5 Deep Learning Methods to Compensate for Limited Supervisory Information	25
5.1 Cross-geolocation Co-segmentation	25
5.1.1 Motivation	25

5.1.2	Methodology	26
5.2	Semi-supervised Training	27
5.2.1	Motivation	27
5.2.2	Methodology	28
5.3	Summary	30
6	Demonstration of Developed Deep Learning Methods in Practical Applications	31
6.1	Detection of Undocumented Building Constructions	31
6.1.1	Motivation	31
6.1.2	Methodology	32
6.2	Sampling Strategies for Developed Deep Learning Methods to Reduce Training Data Size	32
6.2.1	Motivation	32
6.2.2	Methodology	34
6.2.3	Experiment	35
6.2.3.1	Dataset	35
6.2.3.2	Experimental Setup	35
6.2.3.3	Training Details	36
6.2.3.4	Evaluation Metrics	36
6.2.4	Results	37
6.2.4.1	Comparison among Different Methods	37
6.2.4.2	Results of Sampling Strategy for Labeled set	37
6.2.4.3	Results of Sampling Strategy for Unlabeled set	39
6.2.5	Discussion	39
6.2.5.1	Impact of Initial Labeled Patches	40
6.2.5.2	Impact of Newly Added Labeled Patches	40
6.2.5.3	Selection of the Optimal Number of Labeled Patches	43
6.3	Summary	45
7	Conclusion and Outlook	47
7.1	Conclusion	47
7.2	Outlook	48
7.2.1	Building Footprint Generation Using Multi-modal Data	48
7.2.2	Building Footprint Generation with Self-supervised Learning	48
7.2.3	Leverage of Building Footprint Maps	48
7.2.4	Building Height Retrieval from Optical Imagery	49
	List of Figures	51
	List of Tables	53
	Bibliography	55
	Appendices	65

A	Li, Qingyu, Lichao Mou, Yuansheng Hua, Yilei Shi, and Xiao Xiang Zhu. Building Footprint Generation Through Convolutional Neural Networks with Attraction Field Representation. <i>IEEE Transactions on Geoscience and Remote Sensing</i> , vol. 60, pp. 1-17, 2022, Art no. 5609017, doi: 10.1109/TGRS.2021.3109844.	69
B	Li, Qingyu, Lichao Mou, Yuansheng Hua, Yilei Shi, and Xiao Xiang Zhu. CrossGeoNet: A Framework for Building Footprint Generation of Label-Scarce Geographical Regions. <i>International Journal of Applied Earth Observation and Geoinformation</i> , 111 (2022): 102824.	87
C	Li, Qingyu, Yilei Shi, and Xiao Xiang Zhu. Semi-Supervised Building Footprint Generation with Feature and Output Consistency Training. <i>IEEE Transactions on Geoscience and Remote Sensing</i> , vol. 60, pp. 1-17, 2022, Art no. 5623217, doi: 10.1109/TGRS.2022.3174636	99
D	Li, Qingyu, Yilei Shi, Stefan Auer, Robert Roschlaub, Karin Möst, Michael Schmitt, Clemens Glock, and Xiaoxiang Zhu. Detection of Undocumented Building Constructions from Official Geodata Using a Convolutional Neural Network. <i>Remote Sensing</i> 12, no. 21 (2020): 3537.	119

Acronyms

2D	two-dimensional.
3D	three-dimensional.
AFMs	attraction field maps.
AP	attribute profile.
BCE	binary cross-entropy.
CAMs	class activation maps.
CNNs	convolutional neural networks.
CRF	conditional random field.
DFK	digitalcadastral map.
DSFE	deep structured feature embedding.
DSPP	dense spatial pyramid pooling.
FCNs	fully convolutional networks.
FPCRF	feature pairwise conditional random field.
FPN	feature pyramid network.
FSG	fuzzy stacked generalization.
GANs	generative adversarial networks.
GBI	geometric building index.
GCN	graph convolutional network.
GLCM	gray-level co-occurrence matrix.
GPUs	graphics processing units.
HR	high resolution.
IoU	intersection over union.
LBP	local binary pattern.
LiDAR	light detection and ranging.
LSC	least-squares classifier.
MABI	morphological attribute building index.
MAP	maximum a posteriori.
MBI	morphological building index.
MFBI	multi-scale filtering building index.

Acronyms

MMFBI	multiple channel multi-scale filtering building index.
MRF	Markov random field.
NDVI	normalized difference vegetation index.
RF	random forest.
RNN	recurrent neural network.
RPN	region proposal network.
SAR	synthetic aperture radar.
SDT	signed-distance transform.
SIFT	scale-invariant feature transform.
SSIM	structural similarity.
SVM	support vector machine.
tDSM	temporal digital surface model.
TrueDOP	orthophoto with red, green, and blue bands.
UN	united nations.
VHR	very high resolution.

1 Introduction

1.1 Motivations and Objectives

Although cities cover a small proportion of the earth’s land surface, they account for 60-80 % of energy consumption and 75 % of carbon emissions [1]. As the pace of urban expansion and city renewal continues to accelerate, significant changes occur in cities annually [2]. These changes may result in adverse impacts on the environment and ecology, such as resource depletion, greenhouse effect, and urban heat island [3] [4]. Building footprint maps characterize the planar dimension of urban structure, offering insight into urban development. For instance, the floor area is associated with energy consumption [5], greenhouse gas emission [6], and population distribution [7]. Therefore, in-depth studies of building footprint maps are the key to environmentally sustainable urbanization. Moreover, the established building footprint maps also facilitate many applications including (1) emergency responses and rescue operations, (2) undocumented building detection, (3) autonomous navigation of vehicles, (4) land use management, etc.

In the past decades, remote sensing technologies with high-resolution imaging sensors have become a fundamental approach for building footprint generation. This is because it allows collecting optical imagery on a large scale, enabling detailed analysis of buildings and spatial-temporal monitoring of newly constructed and destructed buildings. Therefore, high resolution (HR) and very high resolution (VHR) optical imagery is a reliable data source for building footprint generation.

Early studies of automatic building footprint generation from HR and VHR optical imagery rely on heuristic procedures. For instance, these methods combine different spectral, spatial, or auxiliary information to form building hypotheses. Multiple features need to be engineered, making it difficult for these methods to achieve generic, robust, and scalable building footprint maps. Recently, the emergence of deep learning methods, which are based on convolutional neural networks (CNNs), has made strong contributions to the task of building footprint generation. CNNs are artificial neural networks based on multiple processing layers. A major advantage of CNNs is their superior feature learning capability from raw data. In this regard, prior information is not required and no hand-crafted features need to be designed. As a result, deep learning methods are favorable strategies in the remote sensing community for the task of building footprint generation.

Although deep learning-based methods show promising results in terms of accuracy and efficiency, they have two inherent limitations. **One issue is that detail degradation exhibits in the extracted buildings (e.g., blurred boundaries and blob shapes)**, which can be attributed to three factors. Firstly, the downsample process of the network that compresses the input image as feature maps affect the precise boundary localization, leading to the loss of detailed spatial information. Secondly, there is an imbalance between building content and boundary pixels, resulting in irregular building boundaries. Thirdly, buildings are easily occluded by other objects such as shadows, trees, or other objects, causing fragmented and incomplete boundaries. **The other limitation is that**

training deep learning-based networks require a great amount of pixel-level annotations, which have several disadvantages. On the one hand, the acquisition of massive pixel-wise labeled data is costly and time-consuming. On the other hand, remote sensing imagery consists of complex scenes with various geographic objects, thus, annotating them with dense labels demands expertise and even field work.

Motivated by the above-mentioned facts, this dissertation aims to develop innovative deep learning algorithms for the task of building footprint generation from optical imagery. Three important objectives of this dissertation are listed as follows:

- **Development of deep learning algorithms that can refine blurred building boundaries:** this dissertation shall develop building extraction algorithms that are able to preserve low-level spatial information (e.g., sharp boundaries and fine-grained details).
- **Development of deep learning algorithms that can compensate for the limited supervisory information:** this dissertation shall develop building extraction algorithms that are able to compensate for the limited supervisory information resulting from scarce labeled samples. By doing so, the need for large amounts of dense pixel-level labeled data can be largely alleviated.
- **Demonstration of the developed deep learning algorithms in practical applications:** this dissertation shall demonstrate that building footprint maps produced by the proposed algorithms can provide useful geoinformation for urban planning and monitoring. Moreover, implementation details of developed methods in practical applications should be discussed.

1.2 Dissertation Outline

This dissertation is organized as follows. Chapter 2 introduces basic knowledge for understanding this dissertation. Chapter 3 reviews related works on the task of building footprint generation. Chapter 4 introduces the developed methodologies to refine blurred building boundaries. The proposed methods for the compensation of limited supervisory information are introduced in Chapter 5. The demonstration of developed methods in practical applications is presented in Chapter 6. Finally, chapter 7 concludes this thesis and provides outlooks on future works.

This is a cumulative dissertation based on four peer-reviewed journal papers. They are attached in the appendix and summarized as follows:

- **A. “Building Footprint Generation Through Convolutional Neural Networks With Attraction Field Representation”:** To refine blurred building boundaries, we propose a method by learning attraction field representation for building boundaries. By doing so, an enhanced representation power can be provided by the model. Our method comprises two elemental modules: an Img2AFM module and an AFM2Mask module. Img2AFM module learns an attraction field representation conditioned on an input image, which can enhance building boundaries and suppress the background. Using the learned attraction field map, the AFM2Mask module predicts segmentation masks of buildings. Experimental results show that geometric shapes and sharp boundaries of buildings are well preserved by

the proposed framework, which brings significant improvements over other competitors.

- **B. “CrossGeoNet: A Framework for Building Footprint Generation of Label-Scarce Geographical Regions”:** To compensate for limited supervisory information due to insufficient training examples in target cities, we propose to learn cross-geolocation attention maps in a co-segmentation network. By doing so, the discriminability of buildings within the target city can be improved and a more general building representation in different cities can be provided. Our method is termed as CrossGeoNet and consists of three elemental modules: a Siamese encoder, a cross-geolocation attention module, and a Siamese decoder. In the encoder, feature maps are learned using a pair of images from two different geo-locations. The cross-location attention module learns similarity based on these two feature maps and is capable of providing a global overview of common objects (e.g., buildings) in various cities. In the decoder, segmentation masks of buildings are predicted using the learned cross-location attention maps and the original convolved images. From experimental results, we find that CrossGeoNet is able to detect buildings of different sizes and alleviate false detections, which significantly outperforms other competitors.
- **C. “Semi-Supervised Building Footprint Generation with Feature and Output Consistency Training”:** A generic framework for semi-supervised semantic segmentation is proposed, which integrates the consistency of both features and outputs in the end-to-end network training of unlabeled samples. By doing so, additional constraints can be imposed to compensate for the limited supervisory information. According to the spatial resolution of input remote sensing imagery and the mean size of individual buildings in the study area, an instruction is proposed, which assigns the perturbation to the intermediate feature representations within the encoder. From experimental results, we find our approach is able to well extract more complete building structures and alleviate omission errors.
- **D. “Detection of Undocumented Building Constructions from Official Geodata Using a Convolutional Neural Network”:** A novel framework has been proposed to detect undocumented building constructions. This method utilizes a CNN model and official geodata, including VHR optical data and the normalized digital surface model (nDSM). More specifically, undocumented buildings refer to pixels predicted as “building” by CNN but do not belong to the buildings from the official cadastral map. A temporal digital surface model (tDSM) is introduced in the stage of decision fusion for the separation of the class of old or new undocumented buildings. Undocumented storey construction refers to the pixels that are “building” in both the official cadastral map and predicted results from CNN, but a height deviation is exhibited in the tDSM. Finally, a seamless map of undocumented building constructions has been produced for one-quarter of the state of Bavaria, Germany at a spatial resolution of 0.4 m. This indicated that the proposed framework is robust in large-scale practical applications.

2 Basics

2.1 Building Footprint Generation

Although urban areas cover only a small proportion of the earth’s land surface, they host 55.13 % of the world’s population according to a report from the united nations [1]. Buildings are the predominant objects that characterize the urban structure. Specifically, building footprint maps characterize the planar dimension of urban structure and form. The construction of building footprint maps in nowadays administrations of communities allows to administrate, document, and monitor urban development. Especially in less developed regions (e.g., Africa), significant changes occur in urban areas annually due to rapid urban expansion and city renewal [2]. For instance, a great number of buildings are newly constructed in the previous non-urban island, resulting in environmental and ecological problems such as resource depletion, greenhouse effect, and urban heat island [4]. Therefore, acquiring up-to-date building footprint maps is essential to urban-related analysis.

Table 2.1 summarizes common ways to acquire building footprint maps. Currently, the most reliable building footprint maps are obtained from field surveys and mapping[8], but these surveys are time-consuming due to heavy workloads. In most cities, mapping agencies already document basic building information. However, due to urban expansion and city renewal, these official cadastral maps are usually out-of-date [9]. Building footprint maps can also be acquired from open datasets provided by community-based organizations or companies, including OpenStreetMap, Google, and Microsoft. However, these datasets also suffer from two limitations. One limitation is incompleteness. Google and Microsoft only produce building footprint maps for specific countries or continents. Even though OpenStreetMap aims to provide a free building database for the entire world, stark data inequalities exist among different geographic. For instance, regions with low and medium human development are home to about half of the world’s population, but only account for 28% of the buildings on OpenStreetMap [10]. The other is incorrectness. For example, a demolished building appears in the open building footprint maps, or a newly-built building might be missing. In this regard, remote sensing techniques are more favorable, because they are able to establish update-to-date building databases in a more cost-effective way.

Table 2.1: Common ways to acquire building footprint maps

Source	Pros	Cons
Field survey and mapping	high geometrical accuracy and up-to-date	time-consuming and heavy workloads
Official cadastral maps	high geometrical accuracy	out-of-date
Community-based organizations or companies	free and large-scale	out-of-date and data inequality
Remote sensing imagery	up-to-date and cost-effective for large area coverage	require expertise

2.2 Remote Sensing Data for Building Footprint Generation

Remote sensing is the process to acquire information about an object and an area from satellites or aircraft without physical or intimate contact. Remote sensing data can be used in building footprint generation, and they usually consist of three types (cf. Table 2.2): 1) light detection and ranging (LiDAR) imagery, 2) synthetic aperture radar (SAR) imagery, and 3) optical imagery. LiDAR is a method to image objects by determining ranges with a laser, which measures the time for a reflected signal to return to the sensor. LiDAR is able to capture the precise geometry of objects but has a high acquisition cost. SAR is a form of imaging radar that transmits the successive pulses of microwaves to objects and records the echo of each pulse. The advantage of SAR imagery is that it can penetrate clouds and is insensitive to sun illumination and weather conditions. However, its side-looking geometry induces high uncertainties in recorded signals. Moreover, the visible color of objects can not be differentiated by SAR data. By contrast, optical imagery provides a cost-effective way to capture spectral information essential to understanding geo-objects at a large scale. Note that weather conditions will influence optical remote sensing, as areas under clouds can not be observed.

Table 2.2: Types of remote sensing imagery to generate building footprint maps

Source	Pros	Cons
LiDAR imagery	high geometrical accuracy	high cost
SAR imagery	insensitive to weather conditions and sun illumination	high uncertainties and no visible color information
Optical imagery	cost-effective for large area coverage	sensitive to weather conditions

Optical remote sensing makes use of the sun as the source of illumination and measures the reflected or emitted radiation from the object or scene. By doing so, optical imagery is formed by imaging equipment that detects natural energy in the wavelength range across the electromagnetic spectrum. Specifically, the wavelength ranges from 400 nm to 3000 nm, which includes visible light, near-infrared, and short-wavelength infrared. There are two essential elements in optical imagery: 1) spectral resolution, and 2) spatial resolution.

Spectral resolution is the number of wavebands, depicting the width of the wavelength range being imaged, such as red, or green. Various materials show differences in the reflection and absorption at a great variety of wavelengths, facilitating the identification of specific materials. Based on the spectral resolution, optical remote sensing imagery is categorized into the following types [11]: panchromatic, multispectral, superspectral, hyperspectral, and ultraspectral imagery.

- panchromatic imagery combines the information from full visible, and often partially the near-infrared spectrum, and only returns a single intensity value for each pixel
- multispectral imagery has a few spectral bands (less than 10)
- superspectral imagery has more spectral bands (more than 10)
- hyperspectral imagery has the bandwidth narrower than or equal to 10 nm
- multispectral imagery has the bandwidth narrower than or equal to 1 nm

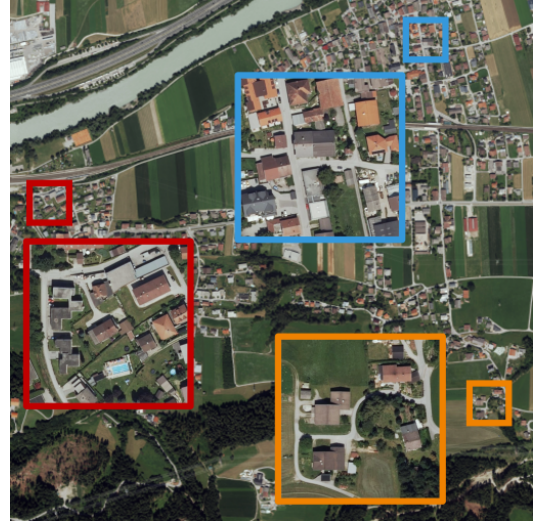
Spatial resolution is expressed as the ability to make discriminate between adjacent objects in an image. The higher the spatial resolution of a sensor, the more easily the



(a)



(b)



(c)

Figure 2.1: Buildings show differently on HR and VHR optical imagery with various spatial resolutions. (a) Planet satellite imagery (3m/pixel). (b) Aerial imagery (1 m/pixel). (c) Aerial imagery (0.3m/pixel).

objects in the image can be distinguished. According to [12], remote sensing imagery can be classified into low resolution (LR), medium resolution (MR), HR, and VHR imagery, in terms of spatial resolution.

- LR imagery has the spatial resolution lower than 100 m/pixel (e.g., the spatial resolution of MODIS data is 500 m/pixel)

- MR imagery has the spatial resolution with range from 10 m/pixel to 100 m/pixel (e.g., the spatial resolution of Landsat data is 30 m/pixel))
- HR and VHR imagery has the spatial resolution higher than 10 m/pixel (e.g., the spatial resolution of Planet data is 3 m/pixel))

For LR and MR imagery, a common issue is mixed pixels where various objects are involved in one pixel. HR and VHR imagery can alleviate this issue, as they can capture a high level of detail, allowing more precise mapping of geo-objects. Moreover, valuable spectral, texture, and geometric information can be acquired to distinguish buildings from non-building objects. Therefore, HR and VHR imagery are suitable data sources for large-scale building mapping tasks.

Crucially, there is a tradeoff between spectral resolution and spatial resolution. The higher the spatial resolution, the lower the spectral resolution. In other words, although HR and VHR imagery provide rich spatial details, they have a relatively low spectral resolution and usually belong to multispectral imagery.

Nowadays, HR and VHR satellite imagery (e.g., IKONOS, QuickBird, WorldView, Pleiades, ZiYuan, GaoFen, Planet), as well as aerial imagery provide great potential for the task of building footprint generation at a very fine scale. Figure 2.1 illustrates some examples of these images. It can be observed buildings show differently on HR and VHR optical imagery with various spatial resolutions. In this regard, building extraction methods effective on different datasets are more favored in the remote sensing community. Furthermore, the HR and VHR satellite imagery availability is unbalanced across different geographic regions. For instance, in developing or less developed regions, open source data are more favored to retrieve building information due to restrictions of financial resources. In this case, Planet satellite imagery is an ideal data source to generate building footprint maps as it is partially open access to the research community. In contrast, other HR and VHR satellite imagery or aerial imagery has a high cost.

Despite the high level of detail, HR and VHR optical imagery still bring several challenges for accurate building footprint generation. One important issue is intra-class variance. Due to differences in materials (e.g., stone, concrete, and clay), construction styles (e.g., color, height, and size), and land use functions (e.g., residential, industrial, and commercial), buildings show a variety of geometry and spectral properties. The other issue is inter-class similarity, where buildings exhibit similar spectral and spatial characteristics to other objects including paved road, bare land, rocks. Moreover, complicated background interference and the loss of relevant sensor data (i.e. shooting angle, shadows, and illumination conditions) hamper the accurate extraction of buildings from HR and VHR optical imagery.

3 Related Work

3.1 Early Efforts in Building Footprint Generation

Early efforts in building footprint generation can be categorized into four types: 1) geometrical primitive-based, 2) index-based, 3) oversegmentation-based, and 4) classifier-based methods.

3.1.1 Geometrical Primitive-based Methods

In geometrical primitive-based methods, geometric primitives (e.g., building edges and corners) are first extracted and then grouped to form closed polygons for individual buildings.

Some algorithms generate building footprints based on the building corner that refers to a point with its local neighborhoods in two varying line segment directions. Building corner is invariant to translation, rotation, and illumination [13]. With the help of some point feature operators (e.g., Harris corner detector [14] and scale-invariant feature transform (SIFT) operator [15]), building corners can be extracted. A Harris corner detector is first implemented for the detection of corner points of buildings. Afterward, these extracted corner points are connected in the order of their polar angles with respect to building central markers [16] [17], in order to construct representations of buildings. SIFT operator is employed to detect corner points, and these corner points are taken as seed points for estimating rectangular buildings with a region growing method [18].

Another commonly used geometric primitive to generate building footprints is building boundary, which is usually extracted in two steps. Firstly, the line segments that are strongly relevant to building boundaries, are detected. Secondly, closed boundaries for individual buildings are formed by grouping the extracted lines. Hough transformation [19] is a commonly used line detection algorithm, which finds straight lines in a parameter space based on a voting procedure. When compared with Hough transformation, Burns algorithm [20] has a relatively lower computation cost since it only uses gradient orientations. For instance, line segment sets are extracted with Hough transformation and Burns algorithm, respectively [21] [22]. Afterward, a structural graph is built by employing intersection nodes of the two-line segment sets. Finally, a graph search algorithm is implemented to identify building boundaries. Nevertheless, both the Hough transformation and Burns algorithm greatly rely on parameter settings, showing severe false alarms. To avoid parameter tuning, EDLines [23] is proposed with a faster computation speed and fewer false alarms. EDLines is exploited to automatically extract line segments that are later grouped by different strategies [24] [25].

3.1.2 Index-based Methods

Index-based methods aim at discriminating buildings from other objects in an index form. By doing so, buildings can be extracted with an empirical threshold. Considering that

buildings and casting shadows show high local contrast, texture-derived built-up presence index (PanTex) [26] exploits the gray-level co-occurrence matrix (GLCM) to measure anisotropic rotation-invariant characteristics of buildings. Morphological building index (MBI) [27] is a building index that describes the characteristics of buildings (e.g., contrast, brightness, size, directionality, and shape) with a set of morphological operators. However, MBI has several disadvantages. On the one hand, it has a heavy computation cost due to the multiscale and multidirectional morphological operations. On the other hand, it neglects some spectral information because it selects a maximal gray value from every spectral band. In this regard, multi-scale filtering building index (MFBI) [28] and multiple channel multi-scale filtering building index (MMFBI) [28] are developed to overcome these drawbacks, respectively. MBI and its core variants are designed for building detection in urban areas, thus, their performance is unsatisfactory in non-urban areas because morphological profiles of bare land and roads are similar to those of buildings. Hence, the morphological attribute building index (MABI) is proposed to detect buildings. MABI is based on an attribute profile (AP) that is able to fully exploit the spectral and spatial characteristics of buildings. Given that utilizing textural or spectral information is difficult to distinguish buildings in half-meter resolution data, geometric building index (GBI) [29] is proposed for automatic building detection. GBI derives the geometric saliency of buildings based on junctions and is capable of depicting meaningful structures of buildings.

3.1.3 Oversegmentation-based Methods

In oversegmentation-based methods, different segments, so-called superpixels, are obtained from the partition of an image. In this way, building regions are identified. These approaches are commonly summarized into five types: 1) region-based, 2) clustering-based, 3) graph model-based, 4) active contour model-based, and 5) watershed segmentation methods.

Region-based segmentation methods group adjacent pixels with similar attributes into unique regions, which involve the process of region splitting, region growth, and region merging. Given that the scale parameter is essential to the segmentation of buildings of various sizes, different region merging algorithms have been proposed for the estimation of the optimal scale parameter. For instance, a coarse segmentation is firstly applied, then, structural and spatial contextual information is extracted to estimate scale parameters [30]. A mathematical equation is proposed to model the relationship between the median size of buildings and the optimal scale parameter [31].

Clustering-based methods aim at grouping a set of pixels in such a way that pixels in the same class (e.g., “building”) are similar to each other. Different clustering methods can be utilized to detect buildings, e.g., mean shift [32], K-means [33] [34], ISODATA [35] [36], and ICA-based clustering methods [37] [38].

Graph model consists of vertices and edges, where vertices are the set of elements to be segmented and edges correspond to pairs of neighboring vertices. In graph model-based image segmentation methods, vertices refer to pixels and the weight of an edge is a measure to describe the similarity (e.g., color and position) between the connected two pixels. Markov random field (MRF) is a commonly used graph model to describe the spatial relations of pixels, and its energy function prefers connected pixels having the same label [39] [40] [41] [42] [43].

Contours depict the boundaries that enclose the region of interest in an image. Working towards the approximation of contours, the active contour model is able to segment images by energy forces and constraints, e.g., length and smoothness of contours. The minimization is realized in both the shape energy and the image energy. Numerous studies exploited the active contour model to approximate the building regions from remote sensing imagery [44] [45] [46] [47].

Watershed segmentation-based methods regard an image as a topographic landscape with ridges and valleys between them. The gray values or gradient magnitudes of respective pixels represent the elevation values of the landscape. Afterward, an image is decomposed into catchment basins, and watersheds delineate the boundaries between these basins. Watershed segmentation aims at assigning each pixel either to a region or a watershed. In [48] [49], watershed segmentation is proposed to detect buildings from remote sensing imagery.

3.1.4 Classifier-based Methods

Classifier-based methods mainly consist of two steps: hand-crafted feature extraction and classification. Features of each pixel are first extracted and then taken as input for classifiers that can determine its label. Compared to the other three types of methods, classifier-based methods can provide stable and generalized results, becoming the most widely used approach.

Classifiers are machine learning models that can distinguish buildings from non-building objects. Related building extraction studies rely on support vector machine (SVM) [50], random forest (RF) [51], decision tree, k-nearest neighbors (KNN) [52], Fisher’s linear discriminant analysis, least-squares classifier (LSC), and Bayes classifier. The most popular classifier for building extraction from remote sensing imagery is SVM. SVM is a supervised technique for classification. The training samples collected for each class are viewed as vectors. SVM is able to construct a hyperplane or set of hyperplanes in high-dimensional spaces that are “pushed up against” among different classes. During SVM learning, the margin between any two classes is maximized. By doing so, a good separation of various categories can be achieved when the hyperplane has the largest distance to the neighboring samples of any two classes. In [53], a variety of geometric features are exploited to characterize the geometric properties of buildings, and taken as input for SVM to discriminate between buildings and other objects. A binary SVM classification strategy is implemented to detect buildings [48], and NDVI is taken auxiliary features to improve classification results. To extract the buildings that are not detected by SVM, a histogram method is utilized, making use of the gray value distribution of building pixels that are correctly detected by SVM [54]. A gallery of feature descriptors including color histograms and local binary pattern (LBP) is implemented to distinguish buildings from non-building objects using SVM [55]. Considering that results provided by the sole classifier are biased, a hierarchical architecture, namely fuzzy stacked generalization (FSG) is proposed to combine the detected building results from multiple classifiers including SVM, KNN, and LSC [56].

3.2 Deep Learning Techniques for Building Footprint Generation

Early efforts have several limitations in extracting buildings at scale. On the one hand, they largely rely on handcrafted features that are significantly influenced by building type, scene complexity, sensor quality, and observation scale. Hence, these methods often only deal with specific data and specific building styles. On the other hand, early efforts define manually designed rules, leading to much computational complexity and cost. Moreover, with the increasing volume of remote sensing data, they are not able to generate building footprint maps efficiently and effectively.

In recent years, deep learning-based methods have significantly outperformed traditional methods on the task of building footprint generation. This is mainly contributed to significant advances in CNNs, which can automatically and adaptively learn discriminative features from raw images. The learned feature involves both low-level features and high-level semantic features. The powerful “feature learning” capability of CNNs has alleviated the heuristic feature design, promoting better generalizability. Furthermore, with the available computing resources such as graphics processing units (GPUs), deep learning methods are capable of automatically generating building footprints on large scale.

A significant number of methods have been proposed for the task of building footprint generation from remote sensing imagery. According to used visual cues, they can be categorized into three classes: corner, boundary, and semantic mask of the building.

3.2.1 Corner-based Methods

With the development of keypoint detection networks, several novel studies propose to delineate building footprints by detecting corner points using CNNs. Aiming to accurately delineate regularized building shapes, PolygonRNN [57] is a favorable architecture that consists of a CNN and a recurrent neural network (RNN), where CNN extracts corner points and RNN connects these points to realize closed polygon representations. PolyMapper [58] integrates the feature pyramid network (FPN) [59] based on PolygonRNN [57], avoiding the need for bounding box annotations. In [60], the same pipeline as PolyMapper [58] is utilized, and global context blocks and boundary refinement blocks are additionally integrated to enhance the feature extraction modules. AGPA[61] is an adaptive polygon generation algorithm that integrates local context features to yield a keypoint map indicating the locations of building corners. Afterward, the position and orientation of the building boundary are utilized to connect these keypoints to outline each building instance.

3.2.2 Boundary-based Methods

For the task of building footprint generation, some approaches are proposed to learn building boundaries in end-to-end CNNs. CLP-CNN [62] designs a concentric loop structure with bidirectional pairing loss to adjust the extracted building boundary. In [63], a GCN-based polygon prediction module is proposed for automatic building boundary extraction with the oriented bounding box. PolygonCNN [64] consists of two parts: a building segmentation network generates the initial building contour that is further improved by a modified PointNet [65]. Two studies propose to learn regularized building boundaries based on active contour models. One work is DSAC [66], which learns parameterizations

with a CNN model. The other method is DARNet [67] which learns active contour models based on polar coordinates. RCF-building [68] directly detects building boundaries using a richer convolutional features network [69]. Most studies prefer to simultaneously learn both building masks and boundaries in a multi-task learning framework. BR-Net [70] utilizes a shared backbone for both building segmentation and outline extraction. In [71], a boundary enhancement module is proposed to share mutual information about the boundary and segmentation masks. CycleNet [72] learns the structural information and the semantic information on a cyclical architecture. In [73], spatial variation fusion is introduced to build a link between building mask extraction and building boundary delineation. EANet [74] uses an edge extraction branch to leverage the features from the semantic segmentation network to learn more edge information about images. In [75], a structurally constrained module is designed to learn building boundaries from the gradient information. CGSNet [76] devises a contour-guided module to include low-level spatial information at the encoder. E-D-Net [77] consists of two sub-networks: E-Net and D-Net, where E-Net learns the edge and mask information of the images, and D-Net is followed to refine the results of E-Net. EaNet [78] proposes an edge-aware loss function, which deploys an image-level Dice loss to build an image-level association across all points. In [79], a boundary-oriented loss function is designed to focus more on pixel values near boundaries during the optimization of trainable parameters. To improve the building boundary quality, BAPANet [80] fuse three different loss functions including binary cross-entropy (BCE), intersection over union (IoU), and structural similarity (SSIM).

3.2.3 Semantic Mask-based Methods

Most methods for building footprint generation involve learning semantic masks of buildings from remote sensing imagery, and their goal is to solve a pixel-level labeling problem. In this regard, semantic segmentation networks are utilized to assign each pixel in the image with a corresponding label (i.e., “building” or “non-building”).

The commonly used semantic segmentation networks are fully convolutional networks (FCNs) [81], encoder-decoder architectures (e.g., U-Net [82], SegNet [83] and FC-DenseNet [84]), DeepLab v1 [85] /v2 [86] /v3 [87] /v3+ [88], and generative adversarial networks (GANs) [89]. FCNs are a forerunner for semantic segmentation, which replaces the fully connected layers with transposed convolutions to solve pixel labeling problems. U-Net, SegNet, and FC-DenseNet are based on an encoder-decoder structure. In the encoder, the spatial resolution of the input is downsampled to generate lower-resolution feature mappings that are learned to be highly efficient at discriminating between classes. Afterward, feature representations are upsampled into a full-resolution segmentation map in the decoder. To expand receptive fields, DeepLab models introduce the concept of dilated convolution (atrous convolution) for semantic segmentation, enabling the incorporation of more context information from neighboring pixels. GANs consist of two neural networks: a generator takes noise variables as input to generate new data instances while a discriminator decides whether each instance of data belongs to the actual training dataset or not. Discriminator and generator play a two-player minimax game to optimize both of their objective functions.

According to the addressed problems, semantic mask-based methods can also be categorized into four types: 1) multi-scale information aggregation-based, 2) building boundary refinement-based, 3) limited supervisory information compensation-based, and 4) computational complexity reduction-based approaches.

3.2.3.1 Multi-scale Information Aggregation

Buildings have large intra-class variation, e.g., size, which arises problems for the task of building footprint generation. This is because the performance of semantic segmentation networks is limited when extracting buildings of very small or large sizes. Due to the restricted receptive field, the extracted large buildings are always discontinuous and holey, while many small buildings are missed. Many methods have been proposed to address the problem of multi-scale building extraction.

Most studies focus on devising a multi-scale aggregation strategy to fuse multi-scale features. In [90], a SVM-based fusion strategy is proposed to fuse deep features produced by three different scales. SR-FCN [91] and MA-FCN [92] upsample all features from four various scales to the original one and concatenate them. When concatenating multi-scale feature maps, GMEDN [93] ignores the features of the first transposed convolutional layer that has little semantic information. ScasNet [94] proposes a novel self-cascaded architecture to aggregate global-to-local contexts. MC-FCN [95] apply three extra multi-scale constraints between three intermediate feature representations and their corresponding ground truths. SNLRUX++ [96] utilizes a cascaded multi-scale feature fusion strategy where the number of fusions depends on the scale. GRRNet [97] proposes a gated feature labeling unit to fuse multi-scale features. MSST-Net [98] performs convolutions to fuse feature information of different scales. SRI-Net [99] proposes a spatial residual inception module to aggregate multi-scale contexts. Web-Net [100] designs ultra-hierarchical sampling blocks to fuse feature maps from different levels. MHA-Net [101] proposes a multipath hybrid dilated convolution framework to aggregate multi-scale contexts. DS-Net [102] design a feature aggregation module to fuse the high-level features and the low-level features. In [103], a multi-scale fusion module with summation is implemented to ensure the local details of the buildings are preserved. The spatial attention mechanism is commonly used to take advantage of utilizing different level features [74] [104].

Some methods focus on multi-scale feature extraction. A multi-parallel dilated convolution module is often implemented to capture building features from multiple scales [105] [106] [91]. EU-Net [107] proposes a dense spatial pyramid pooling (DSPP) structure to acquire multi-scale features. MAP-Net [108] designs a parallel multipath network to extract multi-scale features with spatial localization preserved.

Some methods design specific architectures to address this issue. SiU-Net [109] introduces a Siamese network to solve the scale problem, which takes the original image tile and its down-sampled counterpart as inputs. In [110], two separate semantic-segmentation networks are trained for building with different sizes. MTAPA-Net [111] designs a multitask network for the task of both multi-label classification and building footprint generation, aiming to extract buildings with varying sizes.

3.2.3.2 Building Boundary Refinement

Buildings usually have distinct characteristics, e.g., corners and straight lines when compared to other geospatial objects. The shift and spatial characteristics of CNNs will lose detailed information for precise localization, leading to irregular and inaccurate building boundaries. Numerous approaches have been proposed to preserve sharp building boundaries and geometrical details and can be classified into three types in terms of exploited strategies, including 1) adversarial training-based, 2) graph models-based, and 3) improved output representation-based.

Adversarial training-based methods make use of GANs, which consist of a generator and a discriminator. Building-A-Net [112] implements an auto-encoder network for an adversarial discriminator, ensuring the stable learning of high-order regularities of building shapes. ASLNet [113] designs a shape discriminator to explicitly model the shape constraints of buildings. In [114] [115], GANs with the combination of three types of loss functions (adversarial loss, regularized loss, and reconstruction loss) are proposed for the automatic regularization of the building footprints obtained from an FCN.

Graph models that enable the capture of the interactions between pixels, can also be utilized to enhance building boundaries. CRF is adopted as a post-processing strategy to refine building boundaries. Some methods have proposed an end-to-end network learning strategy, where deep structured feature embedding (DSFE) is first extracted, and then graph convolutional network (GCN) is introduced to aggregate the information from neighboring pixels [116] [117] [118].

To improve network learning, some methods propose different types of output representation that can encode geometrical information about buildings. Signed-distance transform (SDT) [119] is proposed to represent the distance from a pixel to its closest point on a building boundary. This improved representation can capture information on both building boundaries and semantic masks, and has been demonstrated to refine boundaries of buildings in some studies [74] [120]. In [121], a network is designed to learn a frame field output that assigns four vectors to each building corner point.

3.2.3.3 Limited Supervisory Information Compensation

Since the building is more difficult to identify and draw, the manual annotation of buildings requires more effort than that for roads, water, bodies, and woodlands [122]. In this case, some methods have been proposed to reduce the need for a large amount of pixel-level annotations. These methods aim to compensate for the limited supervisory information, which can be categorized into four types: 1) weakly-supervised training-based, 2) pseudo-labeling-based, 3) consistency training-based, and 4) domain adaptation-based approaches.

Weakly-supervised training-based methods build models by learning with weak supervision. Apart from the limited pixel-level labels, these methods still require weaker labels including image-level labels, bounding boxes, and point labels. Image-level labels assign each patch with only one label, the patches occupying building pixels more than a certain amount of the total pixels represent the “building”, while those without building pixels correspond to “non-building”. In [123] and [124], a widely used two-stage framework is utilized, where the pixel-level pseudo labels are first produced from image-level labels and followed by a building extraction model trained by pixel-level pseudo labels. MSG-SR-Net [125] firstly learns class activation maps (CAMs) using image-level labels, and trains a building segmentation model with CAMs. In [126], bounding boxes are utilized to generate probabilistic masks for the training of weakly-supervised segmentation models. Point labels (two points inside and outside each small building, respectively) are employed in [127], which is a weakly-supervised segmentation network for small and large buildings.

Pseudo-labeling is studied in one study [122] that deals with the absence of massive annotation datasets. Specifically, a small number of labeled samples is first used to train a model that can generate pseudo-segmentation maps on the unlabeled images. Afterward, pseudo labels are selected according to some criterion and are incorporated with original training data to obtain the final fine-tuned segmentation model.

3 Related Work

Consistency training-based methods impose the consistency of the predictions when various perturbations are applied. CR [128] applies color jitter and random noise to the raw input and enforces the consistency between their outputs and original outputs. PiCoCo [129] augments the input images randomly and imposes the consistency constraint between the predictions of augmented images. Moreover, it also utilizes contrast learning on labeled images to regularize the compactness of intra- and inter-class latent representation space.

Domain adaptation aims at transferring the knowledge from the source domain to the target domain, which can address the domain shift problem. For the task of building footprint generation, the source domain dataset refers to the dataset with sufficient annotated samples, while the target domain dataset refers to the dataset with no labeled samples. In this regard, domain adaptation-based methods try to improve the performance of CNNs on the target domain by using the source domain dataset and aligning the data distribution between two domains. By doing so, the limited supervisory information in the target domain can be compensated. A two-stage method is utilized in [130] [131]. In the first stage, image-level domain adaptation is implemented, and target images are transformed into pseudo-source images. Afterward, segmentation networks trained on the source domain are utilized to obtain building semantic masks for pseudo-source images. Other studies have proposed an end-to-end framework for accomplishing tasks of both domain adaptation and semantic segmentation. BiFDANet [132] optimizes the segmentation networks in two directions including both the source-to-target direction and target-to-source direction. In [133], a segmentation network is proposed based on a domain adaptive transfer attack scheme that aims to obtain domain-adapted adversarial examples with the attack model. JPRNet [134] has two GANs, where one GAN is implemented for pixel-level domain adaptation, while the other GAN aims at representation-level domain adaptation. In [135], an adversarial entropy strategy is proposed for domain adaptation, which is capable of decreasing the entropy and the prediction uncertainty for target images. FDANet [136] is a full-level domain adaptation network for the task of building footprint generation, and it is able to effectively utilize the information from image-, feature-, and output-level.

3.2.3.4 Computational Complexity Reduction

Rapidly and accurately generating building footprint maps is vital to disaster emergency response, loss assessment, and military reconnaissance tasks. Some efficient models have been developed to reduce computation complexity and memory usage. DE-Net [137] designs a network architecture with a small number of parameters, which consists of four modules: downsampling component, encoding component, compressing component, and densely upsampling component. To improve the training speed, DR-Net [138] decreases the number of convolution kernels in networks, reducing the training parameters. ESFNet [139] proposes a separable factorized residual block, aiming at the compression of model size. ARC-Net [140] exploits the residual blocks with asymmetric convolution, which can reduce the computational complexity. RSR-Net [141] proposes a novel decoder with fewer parameters and calculations.

According to the review of related works for building footprint generation, we have defined the objective of this dissertation from two aspects. On the one hand, a lot of multi-scale aggregation-based methods can already achieve satisfactory performance. On the other hand, computational complexity reduction is at the cost of accuracy decrease, which is not what we want. Therefore, we decide to focus on other two research directions:

3.2 Deep Learning Techniques for Building Footprint Generation

1) building boundary refinement and 2) limited supervisory information compensation, which will be discussed in chapters 4 and 5, respectively.

4 Deep Learning Methods to Refine Blurred Building Boundaries

Although building footprint maps provided by existing CNNs seem to be impressive at a large scale, it is observed that such results are not that perfect when we zoom in (see results that are obtained from FC-DenseNet in Fig. 4.1). The extracted building footprints have irregular shapes which are far from their exact geometry in the cadastral maps.

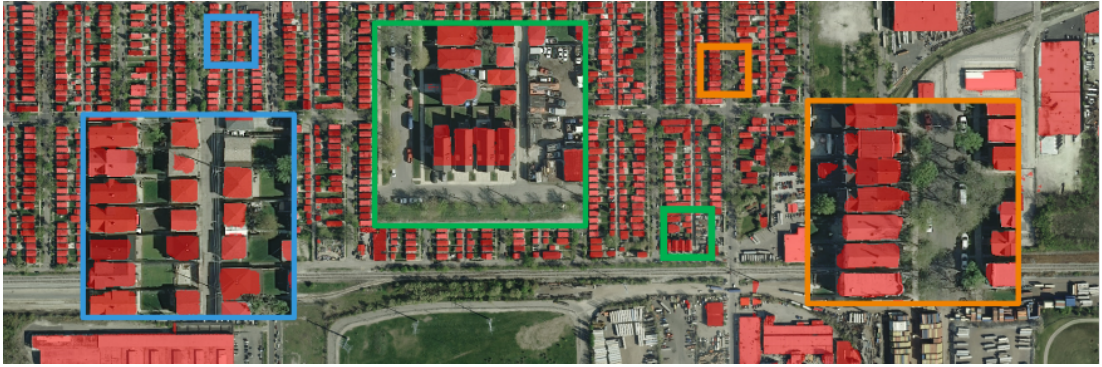


Figure 4.1: The building footprint maps generated by FC-DenseNet [84].

In this chapter, we have first proposed an end-to-end building footprint generation approach that integrates CNNs and graph models to preserve sharp boundaries and fine-grained segmentation. However, this method has not taken the geometric characteristics of buildings into account, which may lead to the generated building shapes far from the exact geometry. Therefore, we then propose a novel network that learns an attraction field map that considers the building boundary as a visual cue. This method is able to enhance building boundaries and suppress the impact of background clutters.

4.1 Feature Pairwise Conditional Random Field

4.1.1 Motivation

In order to refine the blurred building boundaries obtained by CNNs, graph model-based methods adopt a graph model such as CRF to model interactions between pixels. However, the CRF inference is usually implemented as a post-processing step and not integrated with the training of the CNNs, which fails to provide replicable and stable results. In this research, we propose an accurate and reliable building footprint generation framework, which fully integrates the graph models with CNNs in an end-to-end training framework [142]. Moreover, we propose to utilize feature pairwise conditional random field (FPCRF) as the graph model in this end-to-end framework, as it is superior to other graph models in terms of computation efficiency and completeness in feature learning.

4.1.2 Methodology

The overall architecture of the proposed method is illustrated in Figure 4.2, and has two major components including CNNs and FPCRF. The output of the CNNs is composed of two parts. One is the segmentation probability map that is obtained from the last softmax layer of CNNs, and will further be exploited as the unary potential. The other is the extracted features from CNNs, where each pixel is encoded as a fixed-length vector representation (i.e. embedding). This feature embedding is utilized for pairwise potential calculation, encouraging the assignment of similar labels to pixels with similar properties. We propose FPCRF as the graph model for the refinement of the results obtained from CNNs. FPCRF takes the feature embedding and unary potential as input, enabling the modeling of their spatial correlations. FPCRF then outputs the marginal distribution of each pixel that represents the different class labels. Our method integrates CNNs and FPCRF in an end-to-end framework, where the gradients are propagated through the entire pipeline. By doing so, CNNs and FPCRF can co-adapt, which produces the optimal output.

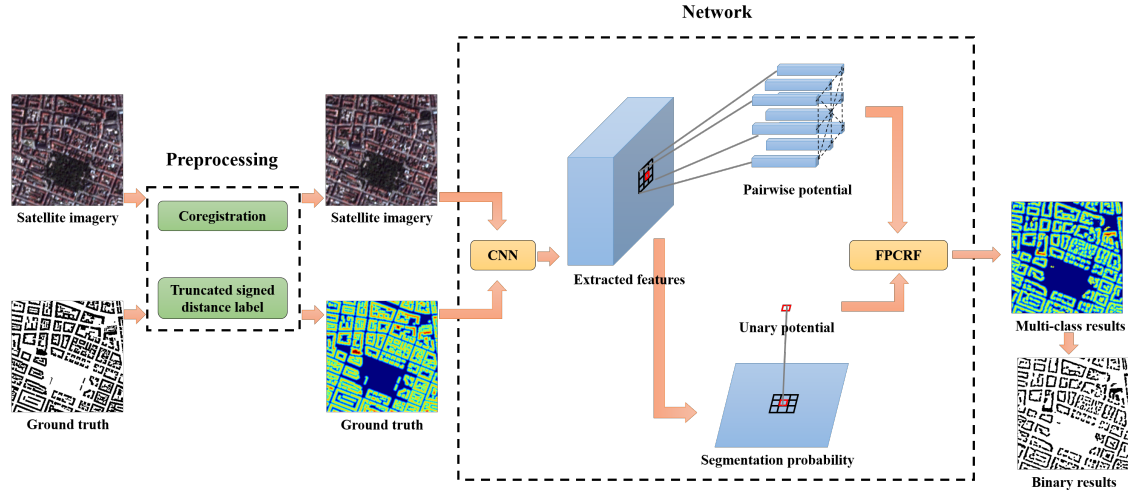


Figure 4.2: Flowchart of the proposed approach

In FPCRF, the pairwise potential $\psi_p(x_i, x_j | I)$ is defined as below,

$$\psi_p(x_i, x_j | I) = \mu(x_i, x_j) \underbrace{\sum_{m=1}^M w^{(m)} k^{(m)}(f_i, f_j)}_{k(f_i, f_j)}, \quad (4.1)$$

where $w^{(m)}$ are learnable parameters, and M is the number of kernels, which is determined by the selected kernels. The terms f_i and f_j are feature vectors for pixels i and j and may depend on the input image I . The function $\mu(x_i, x_j)$ is the compatibility transformation, capturing the compatibility between labels x_i and x_j .

However, only shallow features — the color and position of the pixel for kernels in pairwise potential terms are used in former CRF variants. In this way, the complete features extracted from CNNs have not been fully harnessed. In our research, FPCRF is proposed as a graph model in the building footprint generation framework.

In FPCRf, a pairwise potential term with localized constraints is designed, allowing complete feature learning. The kernel utilized for pairwise potential in FPCRf is a Gaussian kernel, which is defined by the feature vectors f_1, \dots, f_B , where B is the number of feature vector types. The kernel $k^{(m)}$ is defined as:

$$k^{(m)}(f_i, f_j) = \exp\left(-\sum_{b=1}^B \frac{|f_{b,i} - f_{b,j}|^2}{2\theta_b^2}\right), \quad (4.2)$$

where θ_b is a learnable parameter.

The most probable label x can be yielded by the minimization of the Gibbs energy in FPCRf. The calculation of the probability is very similar to the “many-body problem” in physics, where determining the physical behavior of systems composed of several particles is, in general, very hard. The reason is that the number of possible combinations of states increases exponentially with the number of particles. In our case, if the conditional distribution of one pixel has changed, the conditional distributions of other pixels which are linked with this changed pixel will also change, which is difficult to solve. There is a fact that the “many-body problem” is “analytically unsolvable”. This means that there is no general solution that only uses algebraic expressions and integrals. In order to solve the problem at a lower computational cost, the mean field inference will be utilized where the effect of all the other individuals on any given individual is approximated by a single averaged effect, thus reducing a many-body problem to a two-body problem. The interaction between pixels is still taken into account and mean field inference makes the problem simplified by the approximation. In this regard, many interactions of pixels are replaced by one effective interaction, so if the pixel exhibits many random interactions in the original graph, they tend to cancel each other out so the mean effective interaction and mean field theory will be more accurate. This is true in cases of high dimensionality, when the long-range pairs of pixels are included, as when the distance between pairs of pixels is larger, the interactions are weaker which could be neglected.

4.2 Attraction Field Representation

4.2.1 Motivation

Buildings usually have distinct characteristics, e.g., corners and straight lines, which inspires us to exploit geometric primitives of buildings as the most distinguishable features to extract buildings. Therefore, building boundaries are adopted as a primary visual cue to achieve our task – building boundary refinement.

Recently, attraction field representation that finds the most attractive line segment for each pixel, is used for the task of line segment detection in computer vision [143]. We have observed that when the attraction field is exploited to represent building masks, building boundaries can be greatly enhanced while background clutters (e.g., car, courtyard, and road) are suppressed. Fig. 4.3 shows an example. Motivated by this observation, we want a representation of buildings by leveraging the attraction field, which is helpful to the precise delineation of building boundaries.

4.2.2 Methodology

As shown in Figure 4.4, the proposed method consists of two modules: Img2AFM and AFM2Mask. To learn the attraction field representation, a U-Net architecture is exploited

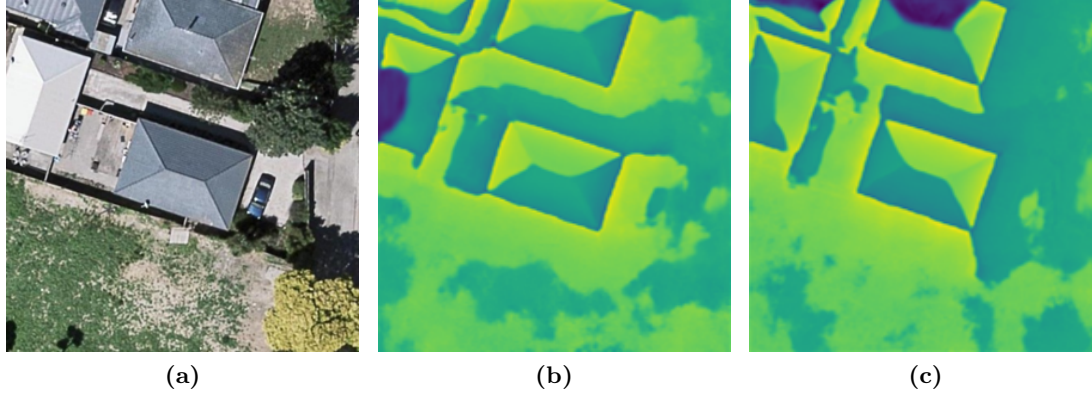


Figure 4.3: (a) The satellite imagery, and the attraction field maps in both (b) x and (c) y directions estimated by our method.

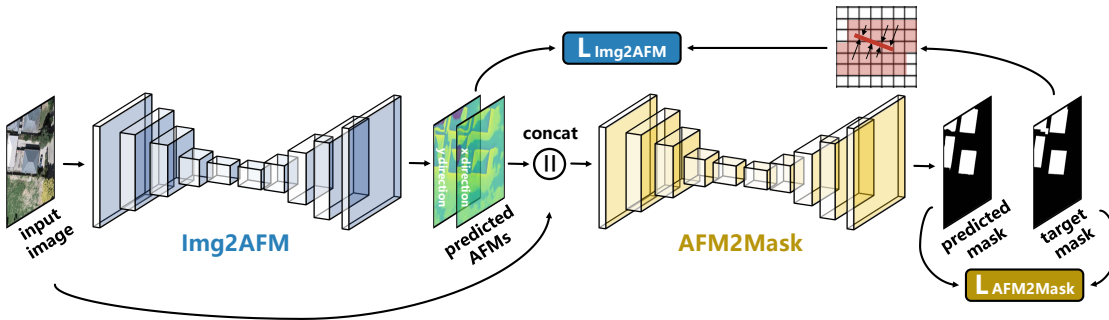


Figure 4.4: Overview of the proposed framework. The Img2AFM module takes an image as input and outputs two attraction field maps (AFMs) in x and y directions. Afterwards, the output is then fed into the AFM2Mask module along with the input image to generate a building mask. Notable that these two modules are trained in an end-to-end fashion.

in the Img2AFM module. In this way, building boundaries can be enhanced, and background clutters can be suppressed. An image is taken as the input of the Img2AFM module. Afterward, two attraction field maps (AFMs) in the x and y directions are generated and fed into the AFM2Mask module along with the input image. The AFM2Mask module aims to generate a building mask and is very flexible to utilize different semantic segmentation networks. Note that both modules are optimized jointly, thus, the co-adaption of these two modules can yield optimal output.

An end-to-end training pipeline is proposed for the supervised learning of our network. Specifically, the AFM2Mask module is appended after the Img2AFM module, and the two modules are jointly trained by minimizing a global loss function L that is defined as follows:

$$L = L_{Img2AFM} + \lambda \cdot L_{AFM2Mask}, \quad (4.3)$$

where $L_{Img2AFM}$ and $L_{AFM2Mask}$ are two loss functions for optimizing the Img2AFM and AFM2Mask modules, respectively. λ is a hyperparameter to introduce a weight on the second loss and can model the relative importance of two modules.

An attraction field representation is a 2D feature map that represents attraction vectors from each pixel to its projection point in x and y directions, which is feasible to be learned by a semantic segmentation network architecture. U-Net is more favorable than semantic segmentation networks for this task since multi-scale skip connections of U-Net can effectively use low-level visual cues (e.g., object edges). Moreover, we found that when other network architectures are utilized in the Img2AFM module, the learning of the attraction field map fails. Afterward, we need to remap the learned AFMs into building masks. However, we found that, in our building footprint generation task, the recovered boundary map from the heuristic algorithm [144] is not satisfactory since there is a relatively high false alarm rate. Therefore, in this work, we propose to learn this process, i.e., recovering building masks from the learned attraction field map, using a network. By doing so, the whole process can be trained in an end-to-end manner, which makes it more efficient and robust. In order to further explore how to well leverage attraction field representation, we investigate different designs [145] [120] [146] to incorporate this useful representation in network learning. From the experiment results, the recursive learning strategy has proven to be optimal, which concatenates the input image and learned attraction field representation and inputs them to a semantic segmentation network to directly generate building masks.

4.3 Summary

For the task of building footprint generation, one issue needs to be considered: detail degradation exhibits in the extracted buildings (e.g., blurred boundaries and blob shapes). The algorithms that are proposed in Appendix A and one related publication [142] for solving this issue are summarized in this chapter.

5 Deep Learning Methods to Compensate for Limited Supervisory Information

For the task of building footprint generation, CNNs can directly learn hierarchical contextual features from the raw input and surpass conventional methods in terms of accuracy of efficiency. However, there remains a challenge for generating building footprint maps on a large scale — massive data need to be collected to promote the generalization performance of CNNs. Besides, the manual annotation of reference data is a very time-consuming and costly process.

In this chapter, we have first proposed a co-segmentation learning framework to make use of a large amount of labeled data in other cities, boosting the performance of a model in target cities where annotated data is scarce. Considering that data annotation is still needed when using the co-segmentation pipeline, we have proposed a semi-supervised network based on consistency training. By doing so, a large amount of unlabeled data can be leveraged to improve the performance of a model.

5.1 Cross-geolocation Co-segmentation

5.1.1 Motivation

For target cities with scarce labeled samples, the performance of CNNs is usually restricted, because CNNs require massive strong supervisory information. Therefore, a straightforward idea to solve this issue is to transfer the knowledge from the cities with massive annotated data (hereafter called auxiliary set). However, several challenges arise due to geographic peculiarities across different geolocations. Firstly, appearances of densely or sparsely populated urban settlements are varied [120]. Secondly, buildings have large intra-class variations, e.g., shapes and colors. Thirdly, varying radiometries of remote sensing images are induced by differences in the data acquisition process (e.g., atmospheric effects and illumination conditions) [147]. Several examples are shown in Figure 5.1, and we can observe that the appearances of buildings come in a wide variety on different continents. If a network trained on the auxiliary set is directly applied to target cities, we will get unsatisfactory results from CNNs.

Co-segmentation aiming at jointly segmenting semantically similar objects in video frames or multiple images, which has been utilized for the task of object segmentation in computer vision [148] [149] [150] [151]. This is because the sequential or pair-wise relations among consecutive frames to discover common objects can be fully harnessed by co-segmentation, alleviating the dependency on strong supervisory information. Inspired by that, we want to make use of the co-segmentation framework in our cross-city building extraction task. In this paper, we propose an end-to-end trainable network—CrossGeoNet, which is able to transfer the knowledge from the auxiliary set to target cities. Since cap-

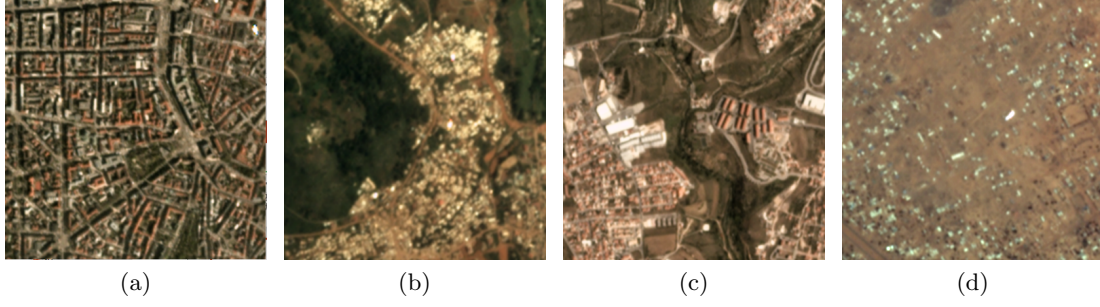


Figure 5.1: Illustration of geographic peculiarities across different geolocations. The Planet satellite images are collected from (a) Munich (Germany), (b) Yaounde (Cameroon), (c) Lisbon (Portugal), and (d) Niamey (Niger), respectively. We can see that appearances of buildings in different cities are noticeably different.

turing the relationship between the two inputs is the key element in our CrossGeoNet, we propose a cross-geolocation attention module to effectively learn the underlying similarity between different geolocations.

5.1.2 Methodology

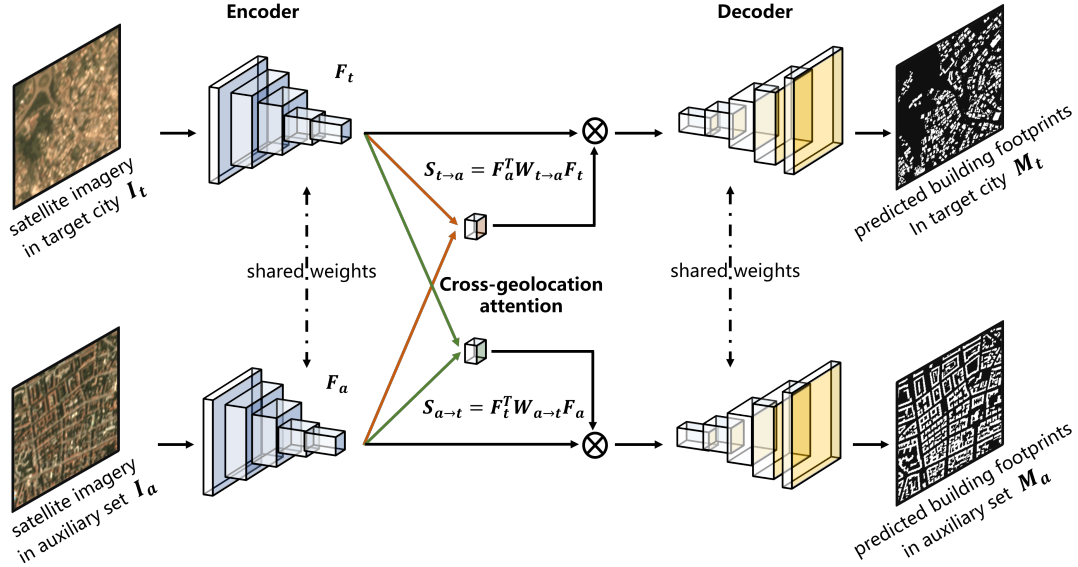


Figure 5.2: Overview of the proposed CrossGeoNet framework.

To improve model performance, co-segmentation exploits the synergistic relationship between video frames or multiple images to provide generic features of objects, which are belonging to the same class but are varies in pose, shape, or color. CrossGeoNet makes use of a co-segmentation pipeline to learn underlying similarities among various cities. By doing so, more generic representations of buildings can be extracted and the generalizability

of the model can be enhanced. As a consequence, not only building discriminability within target cities is improved, but also generic features of buildings across different cities are learned. This is helpful to compensate for the limited supervisory information in target cities.

CrossGeoNet implements a Siamese encoder-decoder architecture (see Figure 5.2), where the Siamese encoder learns high-level feature maps and the Siamese decoder predicts segmentation masks using learned feature representations. Moreover, a cross-geolocation attention module is proposed in our method, and it aims to enhance the latent features by encoding relations between the target city and cities from the auxiliary set.

The Siamese encoder consists of two identical CNNs with shared weights. It takes as input an image pair where one image \mathbf{I}_t is from a target city and the other image \mathbf{I}_a is from the auxiliary set. Afterward, their feature representations $\mathbf{F}_t \in \mathbb{R}^{C \times W \times H}$ and $\mathbf{F}_a \in \mathbb{R}^{C \times W \times H}$, are extracted by the Siamese encoder respectively. H and W are the height and width, and C is the channel dimension. Instead of that high-level features being directly decoded for inferring building masks, we propose a cross-geolocation attention module for the enhancement of the learned feature maps. Specifically, two feature maps are taken as input for this module, and two attention maps $\mathbf{S}_{t \rightarrow a}$ and $\mathbf{S}_{a \rightarrow t}$ are generated. Afterward, we fuse them with the corresponding convolved images and feed the fused feature maps into the decoder. The Siamese decoder is composed of a set of transposed convolutional layers. It upsamples the convolved images to generate two building segmentation masks \mathbf{M}_t and \mathbf{M}_a . Note that all modules are integrated into one framework and optimized in an end-to-end manner. After sufficient training, the co-adaption of these modules is expected to yield the optimal output.

We propose an end-to-end training pipeline for the supervised learning of CrossGeoNet. More specifically, the Siamese network takes a pair of the images $\{\mathbf{I}_t, \mathbf{I}_a\}$ as input, which is randomly sampled from a target city and the auxiliary set, respectively. Afterward, the corresponding segmentation masks $\{\mathbf{M}_t, \mathbf{M}_a\}$ are produced by CrossGeoNet. The whole network is trained by the following loss function:

$$L = L_t + \lambda \cdot L_a, \quad (5.1)$$

where L_t and L_a are two functions for measuring the difference between $\{\mathbf{M}_t, \mathbf{M}_a\}$ and their corresponding ground-truth masks $\{\mathbf{Q}_t, \mathbf{Q}_a\}$. λ is a hyperparameter to control the importance of the second loss. In our task, L_t and L_a are measured using the cross entropy loss function.

5.2 Semi-supervised Training

5.2.1 Motivation

Recently, several methodologies have taken advantage of semi-supervised learning to address the issue encountered by insufficient labeled training data. Among them, consistency training-based approaches not only are simple to implement but also require no additional weakly labeled examples. Consistency training-based methods exploit the teacher-student framework and encourage both the student model and teacher model to give consistent outputs for unlabeled inputs that are perturbed in various ways. By doing so, the generalization capability of the network can be improved.

However, there is still a certain gap in performance between these two models when the outputs are not completely correct during training. Inspired by [152] that more discriminative contextual information can be captured by feature maps, we propose a new consistency loss that measures the discrepancy between both feature maps and outputs of the student model and those of the teacher model, offering a strong constraint to regularize the learning of the network.

The cluster assumption where the classes must be separated by low-density regions determines the effectiveness of consistency training-based approaches. For natural images, it is observed that low-density regions are at the encoder's output [153], and the perturbation is applied at this position. Nevertheless, we have observed the presence of low-density regions separating the classes within the intermediate feature representations at a certain depth in the encoder when remote sensing imagery with low spatial resolution is utilized at the input for the task of building footprint generation. Motivated by this observation, we propose to enforce consistency over the perturbation applied to feature representations at a certain depth within the encoder that depends on the spatial resolution of remote sensing imagery and the mean size of individual buildings in the study area.

5.2.2 Methodology

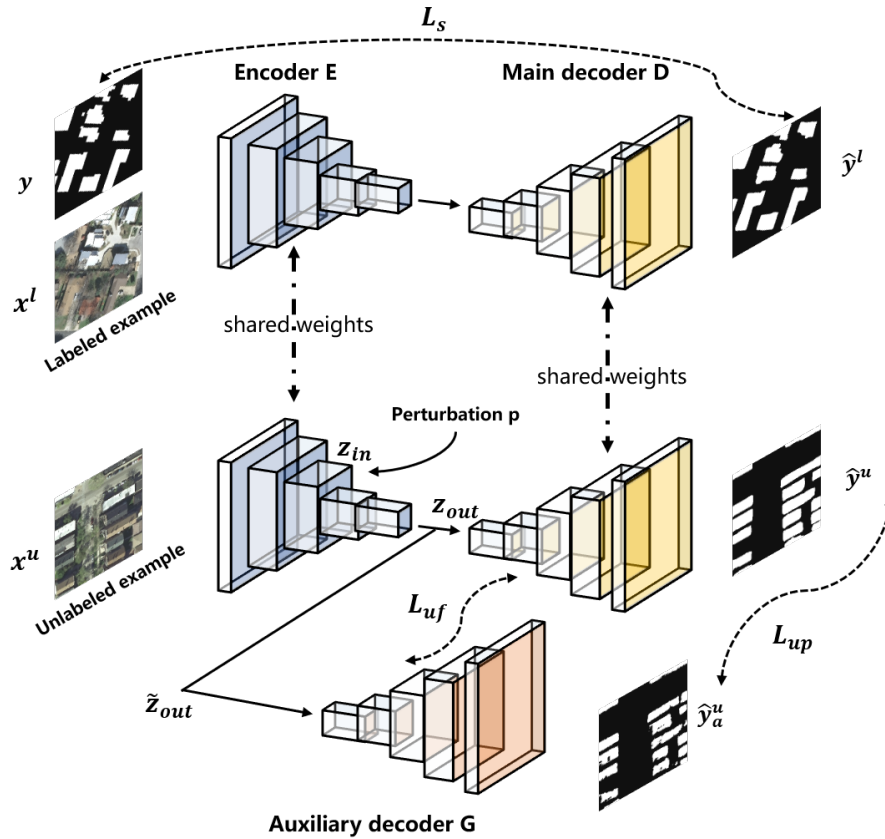


Figure 5.3: Overview of the proposed approach.

Inspired by the perceptual mechanism that leverages the extracted high-level feature maps to improve the network performance, we propose to impose a more precise consistency towards the underlying invariance of features and outputs, which can fully make use of information in deep features and output predictions. By doing so, not only the deep feature maps can be kept consistent, but also the loss of detailed information during network training can be alleviated.

As shown in Fig. 5.3, the proposed framework is composed of a shared encoder E , a main decoder D , and an auxiliary decoder G . The segmentation network F is constituted as $F = E \circ D$ and is trained on the labeled set in a fully supervised manner. The auxiliary network $A = E \circ G$ is trained on the unlabeled examples by enforcing the consistency of both features and outputs between D and G . D takes as input the encoder's output \mathbf{z}_{out} , but G is fed with its perturbed version $\tilde{\mathbf{z}}_{out}$, in which the perturbation p is applied to the output of E . By doing so, the representation learning of E can be further improved by unlabeled examples, and subsequently, that of the segmentation network F .

In each iteration during training, a labeled input image x^l and its label y as well as an unlabeled image x^u are sampled from the training dataset. x^l and x^u are passed through E and D , respectively, and then two corresponding predictions \hat{y}^l and \hat{y}^u are obtained. The supervised loss L_s is computed using y and \hat{y}^l . In order to avoid overfitting to labeled samples, η is introduced into the supervised loss. In what follows, we discuss in detail how η helps to avoid overfitting. In semi-supervised learning, labeled samples would work as anchors [154]. The labels of these samples are known with certainty, forcing the model to fit them with confidence. Therefore, the information extracted from labeled data can be reliably propagated. Moreover, mini-batches are generally evenly split between labeled and unlabeled samples. When the ratio of labeled vs unlabeled samples is very unbalanced, the few labeled samples will be seen very frequently by the optimizer. As the training process goes on, the model may become overly learned for labeled samples [155], and will thus overfit them. Therefore, labeled samples will become easy to be discriminated against, which means that the training losses for these samples will go down. At the same time, the losses for unlabeled samples keep unchanged or even go up. In other words, the model would overfit the few labeled data, ceasing to have an impact on the unlabeled samples. This will prevent effective propagation, which leads to the model under fitted to the unlabeled data. To tackle this difficulty, we only utilize a labeled example if the model's confidence in that example is lower than a predefined threshold η . Specifically, if the model's predicted probability for the correct category is higher than a threshold η , we remove that example from the loss function. This threshold η serves as a ceiling to prevent over-training on easy labeled examples.

The perturbation p is applied to \mathbf{z}_{in} that is feature representation within E for unlabeled image x^u , and $\tilde{\mathbf{z}}_{out}$ is the output from E . Afterward, G generates an auxiliary prediction \hat{y}_a^u from $\tilde{\mathbf{z}}_{out}$. The consistency loss L_{cons} is comprised of two parts L_{uf} and L_{up} . L_{uf} and L_{up} are computed between the features and outputs of G and those of D , respectively.

Based on the observation and analysis, an instruction is proposed to apply the perturbation at depth d within the encoder. d is derived according to the spatial resolution of remote sensing imagery and the mean size of individual buildings in the study area:

$$d = \lfloor \log_2 \left(\frac{l_{min} + l_{max}}{2r} \right) \rfloor, \quad (5.2)$$

where r is the spatial resolution of the remote sensing imagery, l_{min} and l_{max} are mean values of max and min length that are calculated from the ground reference of individual

buildings in the study area. $\lfloor \cdot \rfloor$ is the rounding down function, which gets the largest integer that does not exceed the original value. Afterward, we sample a noise tensor $\mathbf{N} \sim \mu(-0.3, 0.3)$ of the same size as the feature presentations \mathbf{z}_{in} , and this noise tensor is regarded as the perturbation p . We first multiply it with \mathbf{z}_{in} to adjust its amplitude, and then inject it into \mathbf{z}_{in} to derive perturbed feature maps $\tilde{\mathbf{z}}_{in}$:

$$\tilde{\mathbf{z}}_{in} = (\mathbf{z}_{in} \odot \mathbf{N}) + \mathbf{z}_{in} , \quad (5.3)$$

where \odot denotes element-wise multiplication. $\tilde{\mathbf{z}}_{in}$ will then be fed to the subsequent layers in the encoder, and the perturbed intermediate representation $\tilde{\mathbf{z}}_{out}$ of the unlabeled input sample x^u is generated.

5.3 Summary

For the task of building footprint generation, one issue needs to be considered: training deep learning-based semantic segmentation networks requires a great amount of pixel-level annotations. The algorithms that are proposed in Appendix B and C for solving this issue are summarized in this chapter.

6 Demonstration of Developed Deep Learning Methods in Practical Applications

The well-established building footprint maps can contribute to urban planning and monitoring. Therefore, we want to investigate whether the building footprint maps provided by our proposed approaches are capable of offering useful geoinformation for practical applications. Furthermore, we want to explore the implementation details of the proposed methods in a real scenario, offering insights for similar large-scale building extraction tasks.

In this chapter, we have first proposed a framework based on CNNs and decision fusion, which is able to detect undocumented building constructions. Afterward, our proposed semi-supervised training-based method is implemented in this framework. Finally, sampling strategies for this method to reduce training data size are investigated in detail.

6.1 Detection of Undocumented Building Constructions

6.1.1 Motivation

In most German cities, a basic two-dimensional (2D) building database that is known as a digital cadastral map (DFK), is provided by the official authority. The geographic coordinates of buildings documented in DFK are acquired through terrestrial surveys, which provide accurate and comprehensive information for sustainable urban planning. Nevertheless, due to urban expansion and renewal, some building constructions are not recorded via terrestrial surveying and are thus missing in the DFK. These building constructions are named undocumented building constructions. Undocumented building constructions have two types: undocumented buildings and undocumented storey construction. Undocumented buildings represent the buildings shown in airborne survey data but are missing in the cadastral map. Undocumented storey construction represents buildings that exist in both airborne survey data and a cadastral map, but show a signal of height deviation in the temporal digital surface model (tDSM) because of story buildup or demolition. Therefore, monitoring undocumented building constructions is helpful to enhance land resource management and guarantee sustainable urbanization.

Remote sensing technologies such as airborne imaging and laser scanning make it possible to identify these undocumented buildings, as they provide high-resolution data sets for detailed analysis of buildings on a large scale. In the past, identifying undocumented buildings relies on a visual comparison between aerial images with DFK. This requires a great amount of workforce and time. To alleviate the workloads, some semi-automatic strategies [156] [157] are developed for the detection of undocumented buildings. They first extract buildings based on heuristic methods and then overlay the extracted building maps on the DFK to detect undocumented buildings. However, the heuristic thresholds utilized in these strategies can not guarantee a uniform and standardized processing man-

ner, limiting their application on a large scale. Moreover, many false alarms show in the results obtained from these two methods, where vegetation is often misclassified as buildings.

Motivated by the above observation, we aim to propose an automatic and accurate framework for the detection of undocumented building constructions on a large scale.

6.1.2 Methodology

In this study, we have proposed a framework for the detection of undocumented building constructions. Figure 6.1 illustrates an overview of the proposed framework, which consists of three main tasks: (1) detection of undocumented buildings, (2) discrimination between old and new undocumented buildings, and (3) detection of undocumented storey construction.

In the proposed framework, TrueDOP (orthophotos) is utilized as the main data source in building detection, because TrueDOP is able to provide spectral information on buildings. A CNN model takes TrueDOP as input and assigns each pixel with the class label “building” or “non-building”. Note that this CNN model is a semantic segmentation network, which aims at solving pixel-level labeling problems. Afterward, we can identify the undocumented building pixels when we overlay the predicted results with DFK. The undocumented pixels are those pixels belonging to the “non-building” class in the DFK but are assigned the class label of “building” by the CNN model.

The temporal information is helpful to identify the time window of the constructions. Therefore, we introduce tDSM, which is the difference between two digital surface models (DSMs) acquired at two-time points. This information further facilitates the discrimination between different types of undocumented buildings. Here, an empiric value (1.8 m) is selected because a story or a building is usually higher than 1.8 meter. Afterward, this threshold is applied to the tDSM to identify new constructions, indicating whether a height deviation exists for this pixel within the period between two time points. New undocumented buildings are identified when there is a height deviation. This indicates that new undocumented buildings were constructed after time point 1. When there is no height deviation, these undocumented building pixels are old undocumented buildings that have been constructed before time point 1.

Undocumented story construction referring to story buildup or demolition on an existing building can result in a height deviation in two DSMs. We first overlay the predicted results from the CNN model with the DFK. If there is a height deviation in a pixel that is belonging to the class “building” in both DFK and predicted results from CNNs, this pixel is assigned the class label of undocumented storey construction.

6.2 Sampling Strategies for Developed Deep Learning Methods to Reduce Training Data Size

6.2.1 Motivation

When deep learning-based methods are implemented for large-scale building extraction tasks, the collection of training samples usually requires a large quantity of time and manual work. Therefore, we want to investigate sampling strategies that are able to reduce training data size in practical applications. In semi-supervised learning, training

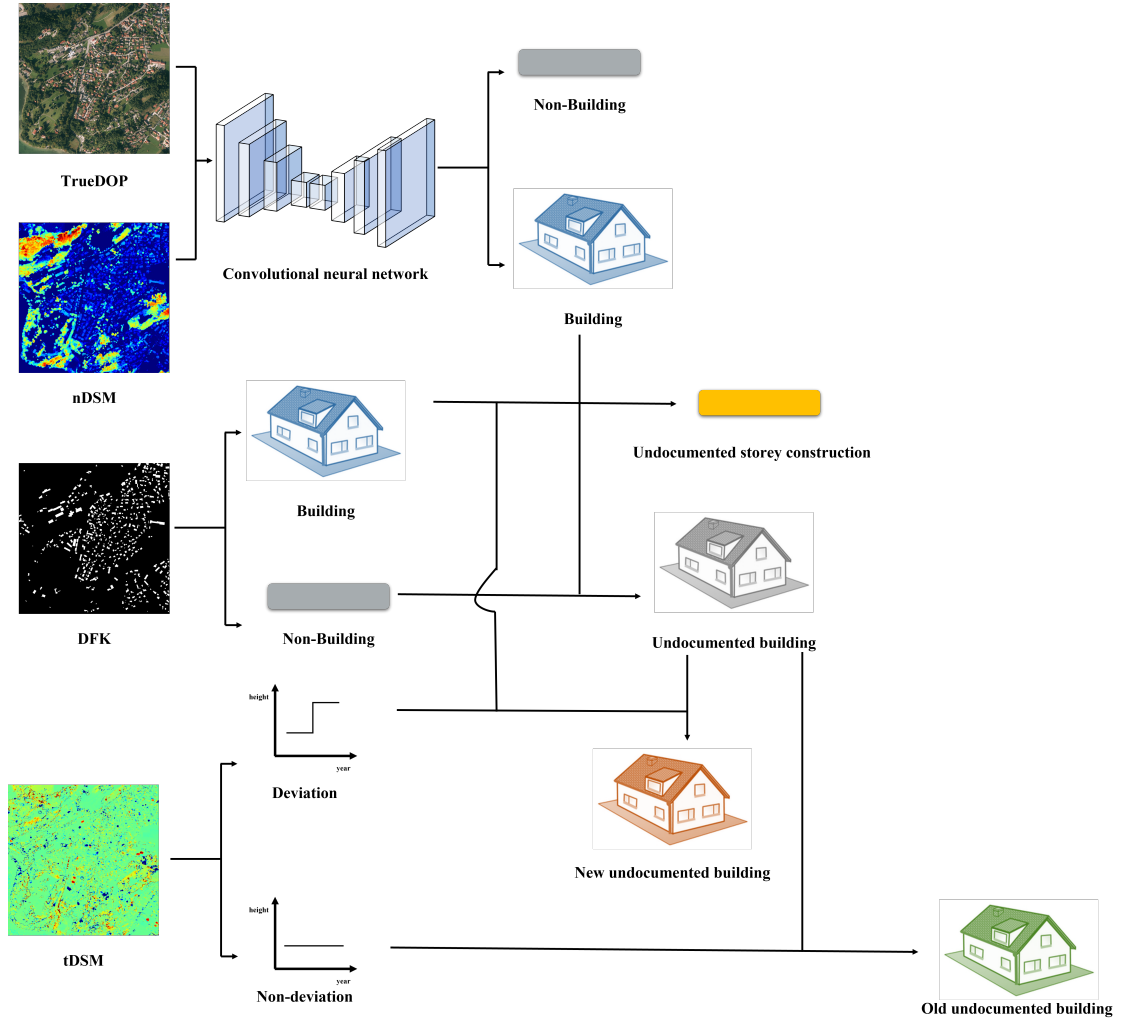


Figure 6.1: Overview of the framework of undocumented building detection.

data consists of labeled and unlabeled sets. Labeled samples would work as anchors, which are known with certainty and force the model to fit them with confidence. By doing so, the information extracted from labeled data can be reliably propagated. Moreover, additional structure about the input distribution can be learned from unlabeled data to produce an estimate of the decision boundary, better separating samples into different classes. Therefore, the investigation of sampling strategies for both labeled and unlabeled sets is vital to our proposed semi-supervised training method.

6.2.2 Methodology

Our research aims to select highly representative training samples among a large pool of patches for semi-supervised learning, which can achieve comparable performance as the supervised method using a full set of labeled patches. Inspired by active learning strategies that select informative samples to enlarge the training dataset, we propose sampling strategies to select the most informative both labeled and unlabeled samples. Specifically, we define selection criteria based on model predictions for both labeled and unlabeled patches. In this way, valuable samples that are expected to maximally boost the model performance are selected.

For the selection of the labeled set, we adopt the margin sampling strategy [158] [159], which seeks the instance that has the smallest difference between the first and second most probable label. In our case, we focus on the pixels that have the smallest difference in prediction probability between “building” and “non-building”. The margin x_M for the pixel x in the patch I is denoted as:

$$x_M = |P(\hat{y}_2|x) - P(\hat{y}_1|x)|, \quad (6.1)$$

where $P(\hat{y}_2|x)$ and $P(\hat{y}_1|x)$ represent probabilities of x that is predicted as “building” and “non-building”, respectively. In general, a smaller margin corresponds to more uncertainty.

$$x_A = \begin{cases} 1, & \text{if } x_M < \beta \\ 0, & \text{otherwise} \end{cases},$$

where β is a threshold to find the pixel that has a small margin. To measure the confidence score of the trained model on each patch, we define an image-level metric for a target labeled patch I_l , which is termed as margin ratio $R(I_l)$,

$$R(I_l) = \frac{\sum_{h=1}^H \sum_{w=1}^W x_A^{(h,w)}}{HW}, \quad (6.2)$$

where H and W are the image height and width of the patch I_l . This margin ratio suggests how well the model performs on this patch, and a higher value means a lower confidence score, indicating more pixels with low confidence are included in this patch. In other words, the higher value, we can say that the model is less confident with this labeled patch. Therefore, the inclusion of this patch into the training dataset is able to further boost the model performance.

For the selection of the unlabeled set, we follow the core idea of consistency training-based methods, which encourage the network to give consistent outputs for unlabeled inputs that are perturbed in various. In our case, we focus on the discrepancy in the outputs between the main decoder and the auxiliary decoder. To measure the confidence

score for each patch, we define an image-level metric for a target unlabeled patch I_u , which is termed as output consistency loss $S(I_u)$, which is denoted as:

$$S(I_u) = \sqrt{\frac{\sum_{h=1}^H \sum_{w=1}^W (P_d^{(h,w)} - P_g^{(h,w)})^2}{HW}}, \quad (6.3)$$

where $P_d^{(h,w)}$ and $P_g^{(h,w)}$ are the probability of the pixel from the main decoder and auxiliary decoder, respectively. This output consistency loss indicates how well the model performs on this patch. The unlabeled patch with larger output consistency loss has a lower confidence score. In other words, the model has a lower confidence level on this unlabeled patch. Hence, we can add this unlabeled patch into the training dataset, in order to improve the model performance.

Before the sample selection, we need to randomly select samples from both labeled and unlabeled sets which are taken as input for an initial training set for network learning. Afterward, we input the remaining set into the trained model for prediction and compute image-level uncertainty metrics for every target image. After the sort of these patches using image-level criterion by descending order, we pick up the top γ portion as hard samples. Finally, hard samples are combined with initial training samples for further network learning. Note that γ is a hyperparameter.

In practice, we can first fix the number of the unlabeled set where pixel-level annotations are removed. Afterward, we start to select labeled samples, where the initial sets are set as different numbers. Then we train semi-supervised models with different initial sets and follow the proposed sampling strategy to add more labeled samples. Once we find the optimal labeled set, we explore the effects of the unlabeled set on the model performance. For the initial unlabeled set selection, we first select the number of unlabeled patches as the same as that of labeled patches. Afterward, we follow the proposed sampling strategy to add more unlabeled samples.

6.2.3 Experiment

6.2.3.1 Dataset

All official geodata are preprocessed to collect training patches as input. DFK is provided as shapefiles and first converted to the raster format at 0.4 m/pixel. In this way, it has the same spatial resolution as TrueDOP. Then, all the tiles of TrueDOP and DFK as corresponding ground references are clipped into patches with a size of 256×256 pixels. Then, we collect the patches from 14 cities in the state of Bavaria, Germany, and we split the collected patches into the train and validation subset for each city. Table 6.1 shows the number of training and validation patches for the 14 selected cities. To evaluate building extraction results, we choose the city of Bad Toelz as the test area, which covers 40 square kilometers.

6.2.3.2 Experimental Setup

In order to provide an upper limit of accuracy metrics, we investigate the performance of semantic segmentation networks under the fully supervised setting where all training patches are labeled. Once we have found the optimal labeled and unlabeled sets with the proposed sampling strategy, we can compare the model performance with supervised learning methods.

Table 6.1: The numbers of training and validation patches for the 14 selected cities.

City	Number of Training Patches	Number of Validation Patches
Ansbach	67,965	18,077
Wolfratshausen	14,982	3671
Kulmbach	24,998	5679
Kronach	19,987	5112
Landau	34,964	8733
Deggendorf	38,454	9763
Landshut	60,957	15,090
Muenchen	88,364	22,213
Regensburg	47,947	11,941
Hemau	9481	2243
Rosenheim	59,141	14,789
Wasserburg	14,150	3567
Schweinfurt	54,951	13,759
Weilheim	76,959	19,202

6.2.3.3 Training Details

Our experiments are conducted within a Pytorch framework on an NVIDIA Tesla with 16 GB of memory. For all methods, the optimizer is stochastic gradient descent (SGD) with a learning rate of 0.1 and a momentum of 0.9, and the training batch size is set as 4. Detailed configurations of all methods included in our experiments are listed as follows:

(1) Efficient-UNet [160]: EfficientNet[161] is adopted as the encoder to learn feature maps. The decoder is comprised of five transposed convolutional layers that upsample the convolved image to predict segmentation masks.

(2) FC-Densenet [84]: Both the encoder and decoder in FC-DenseNet are composed of five dense blocks, and each dense block has five convolutional layers.

(3) MA-FCN [92]: This approach has proposed a feature fusion structure to aggregate multi-scale feature maps. It utilizes a Feature Pyramid Network (FPN) [59] -based structure as the backbone where the encoder is a four-layer VGG-16 [162] architecture and a corresponding decoder implements lateral connections between them.

(4) Proposed method: The hyperparameter α in the unsupervised loss weighting function λ_u is set as 0.6. The loss term weighting parameter of feature consistency ω_u is chosen as 0.2. The network architectures of F and A are the same as that of the semantic segmentation network achieving the best performance.

6.2.3.4 Evaluation Metrics

The performance of models is evaluated by two metrics: F1 score and intersection over union (IoU). They can be computed as follows.

$$\text{F1 score} = \frac{2 \times \text{precision} \times \text{recall}}{\text{precision} + \text{recall}}, \quad (6.4)$$

$$\text{IoU} = \frac{TP}{TP + FP + FN}, \quad (6.5)$$

$$\text{precision} = \frac{TP}{TP + FP}, \quad (6.6)$$

$$\text{recall} = \frac{TP}{TP + FN}, \quad (6.7)$$

where TP indicates the number of true positives, FN is the number of false negatives, and FP is the number of false positives. F1 score realizes a harmonic mean between precision and recall.

6.2.4 Results

6.2.4.1 Comparison among Different Methods

The comparisons among different semantic segmentation networks for supervised learning are presented in this section. Their respective performance is evaluated according to quantitative results in Table 6.2. The goal of this comparison is to select the best semantic segmentation network as the backbone for our semi-supervised learning method in further experiments. Among these semantic segmentation networks, Efficient-UNet [160] performs the best. The superiority of Efficient-UNet [160] can be attributed to its capability of systematically improving performance with all compound coefficients of the architecture (width, depth, and image resolution) balanced [160]. Thus, we take Efficient-UNet [160] as the backbone in both supervised learning and semi-supervised learning approaches for further experiments.

Table 6.2: Accuracies of different semantic segmentation networks for supervised learning. (%)

Method	F1 score	IoU
Efficient-UNet [160]	85.00	74.24
FC-DenseNet [84]	84.82	73.96
MA-FCN [92]	84.80	73.58

6.2.4.2 Results of Sampling Strategy for Labeled set

We first investigate whether our proposed sampling strategy is also suitable for supervised learning. Specifically, we set the labeled patches as 1400, 4200, 7000, 14000, 42000, 70000, and 126000 as initial sets. The statistical metrics are shown in Table 6.3. In a supervised setting, 70000 labeled patches (1/9 original labeled data) can show comparable results as the full set of labeled patches. This suggests there is a redundancy of this full set of labeled patches. Therefore, it is essential to select representative labeled samples for network learning.

Table 6.3: Accuracies of different numbers of labeled sets for Efficient-UNet [160]. (%)

Labeled	F1 score	IoU
1400	75.41	60.53
4200	79.06	65.37
7000	80.26	67.02
14000	81.56	68.85
42000	83.22	71.26
70000	85.26	74.25
126000	85.24	74.23

Then we follow the proposed sampling strategy to add more labeled samples for the models trained by initial sets of 1400, 4200, 7000, 14000, and 42000 labeled samples,

Table 6.4: Accuracies of different initial labeled sets for Efficient-UNet [160] using the proposed sampling strategies. (%)

Initial Labeled	Newly added labeled	F1 score	IoU	Margin ratio
1400	68600	83.69	71.95	0.005
4200	65800	85.01	73.92	0.007
7000	63000	85.41	74.48	0.008
14000	56000	85.29	74.29	0.009
42000	28000	85.09	73.99	0.013

respectively. The corresponding statistical metrics are illustrated in Table 6.4. For all initial sets of labeled samples, we can see that increasing the number of selected labeled samples to 70000 can boost network performance. However, for the initial set of 1400 labeled samples, even labeled samples are also added to 70000, but its F1 score and IoU are lower than other initial sets. This is due to the fact that the performance of its base trained model is unsatisfactory, leading to the selected labeled patches less representative. Moreover, the initial set of 1400 labeled samples has a smaller margin ratio when compared to other initial sets.

Afterward, we try to validate the proposed sampling strategy for the labeled set for semi-supervised learning. Before we investigate the impact of different labeled sets on model performance, the unlabeled set is first fixed as randomly selected 28000 patches. Specifically, we set the labeled patches as 1400, 4200, 7000, 14000, and 18200 as initial sets. The statistical metrics are shown in Table 6.5. The semi-supervised method achieves the best performance when using 14000 labeled patches. With the same amount of labeled sets, semi-supervised learning has largely improved the accuracy when compared to supervised learning, indicating the effectiveness of our semi-supervised method. Especially for the labeled set of 1400 patches, the proposed semi-supervised approach reaches improvements of above 7% in IoU. This is because the proposed method can make full use of the information provided by unlabeled data.

Table 6.5: Accuracies of different numbers of labeled sets for our semi-supervised consistency learning method. (%)

Labeled	Unlabeled	F1 score	IoU
1400	28000	80.73	67.68
4200	28000	82.68	70.48
7000	28000	83.28	71.35
14000	28000	83.92	72.30
18200	28000	83.76	72.06

Then we follow the proposed sampling strategy to add more labeled samples for the models trained by initial sets of 1400, 4200, and 7000 labeled samples. It can be seen from the results in Table 6.5, the semi-supervised method can obtain the best performance using 14000 labeled patches. Therefore, the final labeled patches for all models are added to 14000. The corresponding results are illustrated in Table 6.6. For the initial sets of 4200 and 7000 labeled samples, we can see that increasing the number of selected labeled samples can boost network performance. However, for the initial set of 1400 labeled samples, even though many labeled samples are added, no improvement is shown in accuracy metrics. The reason is that the accuracy of its base trained model is low. In this way, the selected labeled patches according to the base trained model are not informative.

It can also be observed that the margin ratio for the initial set of 1400 labeled samples is smaller than that of other initial sets. Compared to the initial 4200 labeled patches, the initial 7000 labeled patches deliver significantly better results when the final labeled patches are added to 14000. From these results, it is clear that the informativeness of patches selected by the sampling strategy is dependent on the model trained by initially labeled sets. Hence, an accurate initial model is more desirable, facilitating the selection of representative patches.

Table 6.6: Accuracies of different initial labeled sets for our semi-supervised consistency learning method using the proposed sampling strategies. (%)

Initial Labeled	Newly added labeled	Unlabeled	F1 score	IoU	Margin ratio
1400	12600	28000	80.68	67.61	0.005
4200	9800	28000	83.40	71.53	0.006
7000	7000	28000	84.82	73.64	0.01

In summary, for the sampling of labeled patches, we first need to select a relatively large number for initial model training. Afterward, the margin ratio is used to measure the informativeness of each remaining patch. The labeled patch with a large margin ratio can be incorporated into the training set for further model network learning. By doing so, the model performance can be improved.

6.2.4.3 Results of Sampling Strategy for Unlabeled set

As we have already found the optimal labeled set (14000 labeled patches), we investigate the effects of unlabeled samples on the model performance. For the initial unlabeled set selection, we first randomly select 14000 unlabeled patches and together with 14000 labeled patches to train a base model. Afterward, we follow the proposed sampling strategy to add more unlabeled samples. The quantitative results are illustrated in Table 6.7, IoU rises to a high point of 74.64% when 14000 selected unlabeled patches are additionally included. Moreover, it even obtains increments of 1% in IoU when compared to randomly selected 28000 unlabeled patches, which confirms the effectiveness of our sampling strategy for the unlabeled set. Note that after implementing our sampling strategy, the proposed semi-supervised method is able to achieve comparable performance as the supervised approach using the full set of labeled data.

Table 6.7: Accuracies of different unlabeled sets for our semi-supervised consistency learning method. (%)

Labeled	Unlabeled	F1 score	IoU	Output consistency loss
14000	14000	83.39	71.51	-
14000	21000	85.14	74.14	0.027
14000	28000	85.48	74.64	0.020
14000	35000	85.19	74.21	0.017

6.2.5 Discussion

In this section, we first investigate the impacts of initial and newly added labeled patches on the final results, respectively. Afterward, we aim to explore the strategy for selecting the optimal number of labeled patches.

6.2.5.1 Impact of Initial Labeled Patches

We first plot a percentage histogram of different labeled datasets (cf. Figure 6.2). The x-axis represents the distance between the class center of “building” and the class center of “non-building” in each patch. Specifically, the pixels belonging to the same class in each patch are taken into account, and the class center is defined as the average spectral values of these pixels. It can be observed that when the number of labeled patches is larger than 7000, the line graph will be more similar and close to that of all labeled patches.

Afterward, we plot two percentage histograms of different labeled datasets (cf. Figures 6.3 and 6.4), which can be used to examine our sampling strategies of labeled sets for supervised learning and semi-supervised consistency training, respectively. It is important to highlight the fact that our sampling strategies aim to select patches that have a lower confidence score. In other words, the selected patches are hard samples for the trained model. The smaller distance between two class centers will lead to more difficulty for class separation, which suggests the patch with a smaller distance can also be regarded as a hard sample. Therefore, we can incorporate these hard samples in the newly added labeled sets to improve network learning. It should be noted that the line graph (cf. Figure 6.4) of the set (initial 1400 labeled patches and newly added 12600 labeled patches) still shows a lower ratio of hard samples than the full labeled set. This indicates that when the number of initial labeled patches is small, the trained model can not effectively select the hard samples to improve network learning. This is consistent with results in Table 6.6, where no improvement is shown in accuracy metrics after more labeled samples are added to the initial 1400 labeled patches.

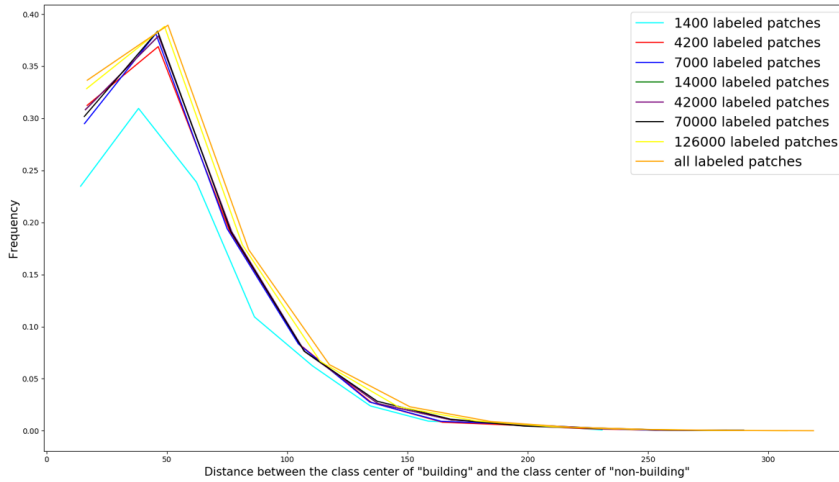


Figure 6.2: The percentage histogram of different labeled datasets.

6.2.5.2 Impact of Newly Added Labeled Patches

When the initial labeled set is fixed, we want to investigate the impacts of newly added labeled datasets (cf. Figures 6.5 and 6.6). This is helpful to explore our sampling strategies of labeled sets for supervised learning and semi-supervised learning. It is clearly seen that implementing sampling strategies can enlarge the proportion of hard samples, which are

6.2 Sampling Strategies for Developed Deep Learning Methods to Reduce Training Data Size

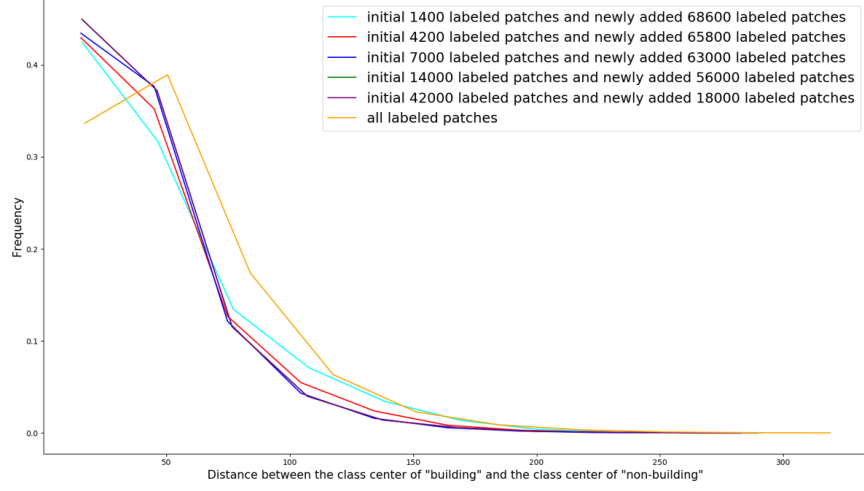


Figure 6.3: The percentage histogram of different initial labeled and newly added labeled datasets for supervised learning.

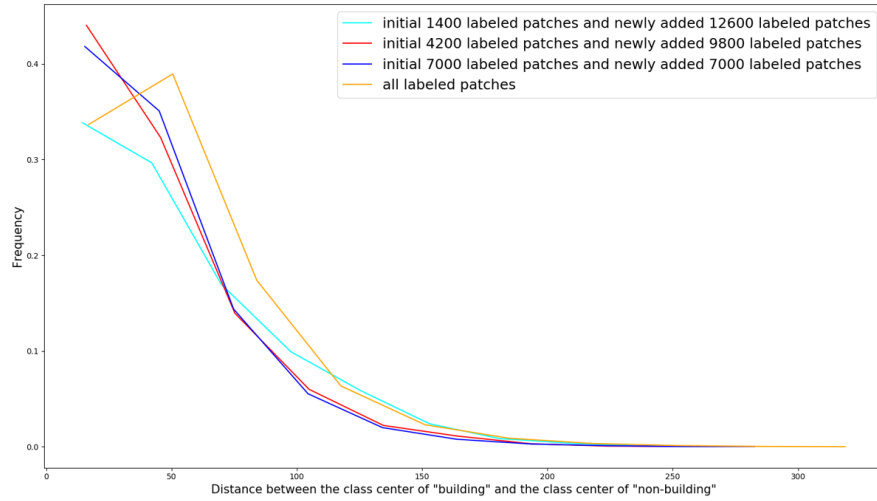


Figure 6.4: The percentage histogram of different initial labeled and newly added labeled datasets for semi-supervised learning.

Table 6.8: Accuracies of different newly added labeled sets for Efficient-UNet [160] using the proposed sampling strategies. (%)

Initial Labeled	Newly added labeled	F1 score	IoU	Margin ratio
7000	0	80.26	67.02	-
7000	35000	84.03	72.45	0.009
7000	63000	85.41	74.48	0.008

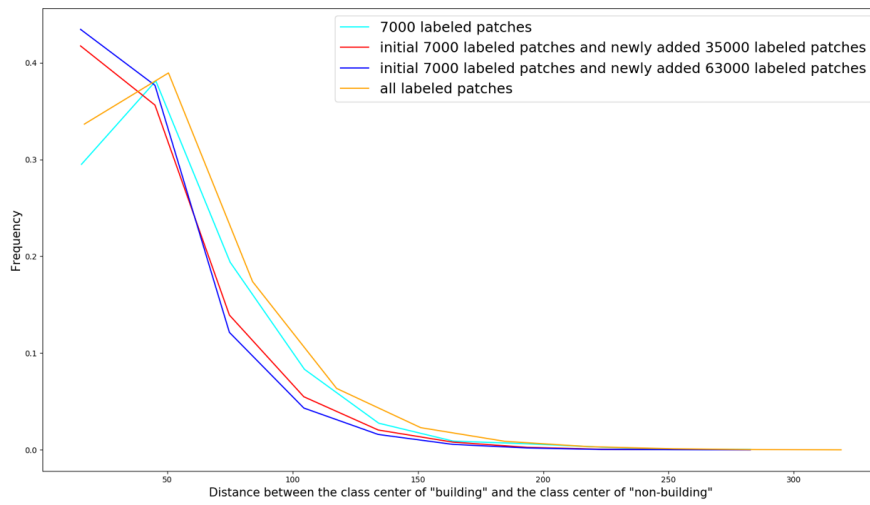


Figure 6.5: The percentage histogram of different newly added labeled datasets for supervised learning.

Table 6.9: Accuracies of different newly added labeled sets for our semi-supervised consistency training method using the proposed sampling strategies. (%)

Initial Labeled	Newly added labeled	Unlabeled	F1 score	IoU	Margin ratio
7000	0	28000	83.28	71.35	-
7000	4200	28000	84.56	72.98	0.011
7000	7000	28000	84.82	73.64	0.01

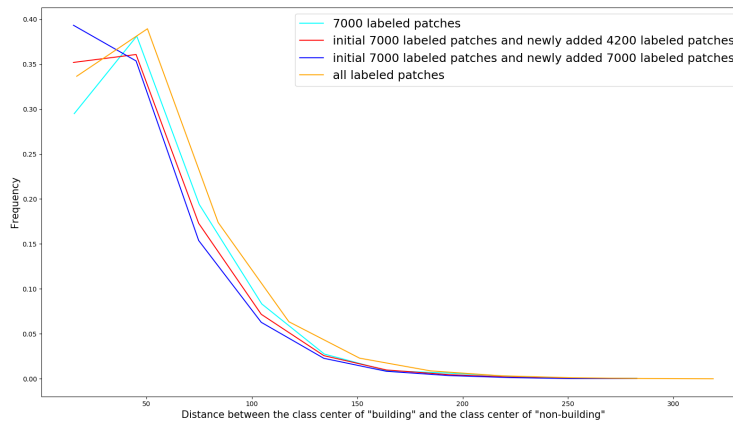


Figure 6.6: The percentage histogram of different newly added labeled datasets for semi-supervised learning.

patches with a smaller distance between two class centers. It can be observed that when more labeled patches are added, more hard samples are introduced for network learning. This is in line with the accuracy metrics presented in Tables 6.8 and 6.9, where the number of newly added labeled patches should be increased to a certain number. By doing so, the model can reach its maximum potential.

6.2.5.3 Selection of the Optimal Number of Labeled Patches

From the percentage histogram of different labeled datasets in Figure 6.2, we have observed that when the labeled set has more patches, the more similar the line graph is to the full labeled set. This inspires us that we can select the optimal number of the labeled patches by examining the similarity of line graphs between different numbers of labeled sets and the full labeled set.

To quantitatively measure the similarity, we calculate the Euclidean distance between line graphs. Specifically, we set the labeled patches as 1400, 4200, 7000, 14000, 42000, 56000, 70000, and 126000 as initial sets. We first classify all patches within the selected set into different categories and the value ranges in each category are set equal in size. In other words, the entire range of distance values (max minus min) is divided equally into the defined number of classes. Here, we set 100 categories and the data range of each category is 3. Afterward, we calculate the frequency of each category. By doing so, the Euclidean distance of frequencies in corresponding data ranges between two different labeled datasets can be calculated and summed.

Figure 6.7 shows the percentage histogram of different labeled datasets. For a fair comparison, each set is randomly 10 times selected from the full labeled set. The line represents the mean values, and the shaded area indicates a boundary at the derived standard deviation. It can be observed that when more labeled patches are included, the line graph is more smooth and similar to the full labeled set. In other words, the labeled set with more patches has a lower standard deviation. Moreover, the line representing the mean values is more similar to the line of the full labeled set.

Table 6.10 shows the similarity of line graphs between different numbers of labeled sets and the full labeled set. It can be observed that when more labeled patches are included, the distance of the line graph to the full labeled set is smaller, and the similarity of the line graph to the full labeled set is higher.

From Table 6.10, the line graph of 56000 labeled patches and 70000 labeled patches show similar distances to that of the full labeled set. This indicates that both two datasets might achieve comparable accuracy. Therefore, we further investigate the differences among various numbers of labeled sets in terms of accuracy and distance of mean value to the line graph of the full labeled set. Specifically, in 10 randomly selected sets for both 56000 labeled patches and 70000 labeled patches, we first find the datasets with the max and min distance within each dataset. Afterward, we train Efficient-UNet [160] models on the corresponding datasets. The statistical results are shown in Table 6.11. It can be observed that when the distance of the line graph to the full labeled set is smaller, accuracy metrics are higher. In this case, the distance of the line graph to the full labeled set offers an informative cue to select the optimal number of patches. For instance, 56000 labeled patches with a distance of 0.0030 show 1% improvement in IoU when compared to 70000 labeled patches with a distance of 0.0046. This again confirms that the distance of the line graph to the full labeled set can be regarded as a metric to find the optimal number of labeled patches.

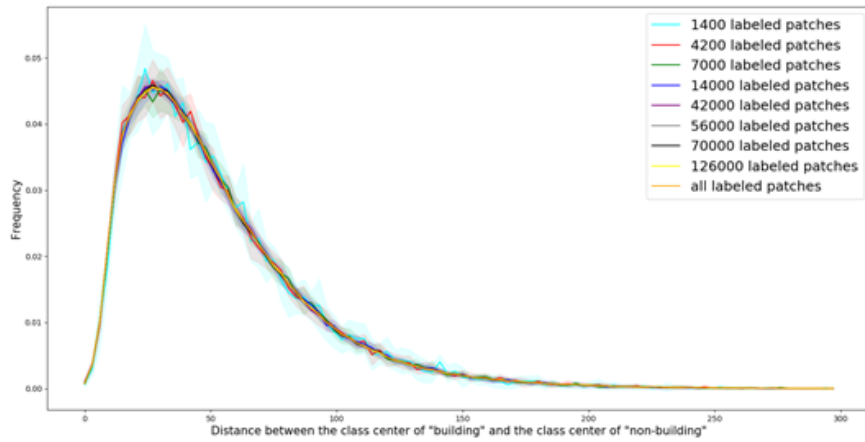


Figure 6.7: The percentage histogram of different labeled datasets. Each set is randomly 10 times selected from the full labeled set.

Table 6.10: Similarity of line graphs between different numbers of labeled sets and the full labeled set. Each set is randomly 10 times selected from the full labeled set.

Labeled	Distance of mean value to the full labeled set
1400	0.0088
4200	0.0047
7000	0.0035
14000	0.0023
42000	0.0016
56000	0.0011
70000	0.0010
126000	0.0007

Table 6.11: Comparisons of different numbers of labeled sets for Efficient-UNet [160].

Labeled	Distance of mean value to the full labeled set	F1 score	IoU
56000	0.0044	84.38 %	72.97 %
56000	0.0030	84.74 %	73.52 %
70000	0.0046	84.08 %	72.52 %
70000	0.0028	84.99 %	73.90 %

6.3 Summary

The established building footprint maps are beneficial to a wide range of practical applications, e.g., the detection of undocumented building constructions. A framework to detect undocumented building constructions is proposed in Appendix D. We have demonstrated the proposed semi-supervised training-based method is effective in this practical application. Moreover, sampling strategies of this approach are proposed.

7 Conclusion and Outlook

7.1 Conclusion

This dissertation explores and investigates modern deep learning techniques for the task of building footprint generation with the potential for practical applications. Several specific challenges induced by existing deep learning-based methods are pointed out in Chapter 1, e.g., blurred boundaries and blob shapes, and the scarcity of pixel-level annotations. Accordingly, related works are summarized in Chapter 3. Finally, in Chapters 4, 5, and 6, this dissertation provides contributions to three aspects of the task of building footprint generation: development of deep learning algorithms that can refine blurred building boundaries, development of deep learning algorithms that can compensate for limited supervisory information, and demonstration of the developed deep learning algorithms in practical applications. The following conclusions are listed as follows:

- To preserve sharp boundaries and fine-grained segmentation, graph models enabling the capture of fine local details can be integrated with CNNs in an end-to-end framework. The proposed FPCRf utilizes a pairwise potential term with localized constraints in CRF, allowing more complete feature learning and efficient message passing operation than other graph models.
- The boundary-aware attraction field can be utilized to represent building footprints, where building boundaries are enhanced and the impact of background clutters is suppressed. The superiority of attraction field representation is due to the fact that geometric properties of buildings can be encoded in 2-D (x- and y-directions), which is more reliable and accurate than other output representations.
- Semi-supervised training-based methods are capable of leveraging a large amount of unlabeled data, helping to compensate for limited supervisory information. The proposed semi-supervised network integrates the consistency training of features and outputs into a unified objective function and offers an instruction to apply the perturbation. By doing so, our method gains significant improvements when compared to other approaches.
- Co-segmentation learning is beneficial to the cross-city building extraction task, as it can jointly utilize the data from different geolocations. The proposed cross-geolocation attention module learns underlying similarities for extracting more generic representations for buildings. In this way, the limited supervisory information in target cities can be compensated.
- A framework is proposed to automatically detect undocumented building constructions, providing information for urban planning and monitoring. In this framework, our proposed semi-supervised learning network can be exploited for the task of building footprint generation, indicating that our method is robust for large-scale practical applications.

- When the proposed semi-supervised learning network is implemented in practical application, we have proposed sampling strategies for both labeled and unlabeled sets. Specifically, margin ratio and output consistency loss are utilized to select informative labeled and unlabeled samples, respectively.

7.2 Outlook

According to the studies of building footprint generation in this dissertation, a few potential topics for future deep learning-based building mapping and related applications are outlined in the following.

7.2.1 Building Footprint Generation Using Multi-modal Data

Multi-modal data refers to data from different sensors, and the joint leverage of multi-modal data enables the network to obtain more detailed information. For instance, optical sensors provide spectral and texture properties of buildings, LiDAR can acquire precise geometrical information, and SAR is insensitive to weather conditions. The main issue here is “how” and “where” to fusion multi-modal data for the task of building footprint generation. The “how” means to design a fusion strategy for the full exploitation of different data. The “where” represents the fusion level, and it includes three types: data level, feature level, and decision level. Data-level fusion means the concatenation of multi-modal data in a single data cube for processing. Feature-level fusion aims at integrating features from various data, which can learn cross-modal features. Decision-level fusion combines the predictions that are obtained from a single modality.

7.2.2 Building Footprint Generation with Self-supervised Learning

With earth observation entering an era of big data, we are able to acquire a large amount of remote sensing data. However, it is expensive and time-consuming to get pixel-level annotation for these data due to the need for expertise. Therefore, self-supervised learning methods are preferable as they are capable of actively exploiting massive unlabeled samples and harnessing the intrinsic structure of data for training. Therefore, we can utilize self-supervised learning methods to provide models pre-trained on large-scale unlabeled remote sensing data, and then transfer the trained models for the task of building footprint generation.

7.2.3 Leverage of Building Footprint Maps

The generated building footprint maps can offer important information for many practical applications on both micro and macro scale. Several examples related to the leverage of building footprint maps are given, e.g., 1) Environmental analysis, 2) hazard vulnerability analysis, and 3) High-resolution population map generation.

- **Environmental analysis:** Urbanization refers to more buildings constructed in former non-urban land. Rapid urbanization will result in adverse impacts on the environment, such as the urban heat island effect, greenhouse effect, and resource depletion. Therefore, we can first derive morphological parameters and landscape metrics of buildings using the generated building footprint maps, and then carry out

correlation analysis together with other environmental variables, e.g., carbon dioxide emission, greenhouse gas emission, and waste production.

- **Hazard vulnerability analysis:** Hazard refers to environmental phenomena that have the potential to affect humans and infrastructures negatively. There are different types of hazards leading to the damage of buildings, including extreme heat, volcanic ash, flood, earthquake, drought, tsunami, and landslide. When a hazard occurs, we can evaluate the vulnerability of buildings in a certain region. This information helps practitioners and stakeholders for a better decision-making process.
- **High-resolution population map generation:** Population data refers to population distributions and dynamics, which provides information for various applications such as estimating the population at risk, deriving health or development goals indicators, and understanding human-environmental processes. However, population data are regularly outdated and even unavailable in some areas. Considering that population is highly correlated with buildings, the generated high-resolution building footprint maps can be utilized to provide a fine-scale population map.

7.2.4 Building Height Retrieval from Optical Imagery

Building footprint maps only contain 2D information on buildings. Building height characterizes the vertical structure of urban form, providing valuable information for a comprehensive investigation of the urban process. Therefore, three-dimensional (3D) building models that link building heights and building footprints, facilitate a more wide range of applications, such as emergency responses and rescue operations, facility management, and urban monitoring. Remote sensing technologies hold great potential for building height estimation on a large scale. Generally, commonly used remote sensing data consists of two types: 1) LiDAR and 2) optical imagery. Although LiDAR allows highly accurate scene geometry measurement, it has a high cost for data acquisition. For city-scale scene analysis, optical imagery is more cost-effective to provide 3D building information with high-resolution and large spatial coverage.

List of Figures

2.1	Buildings show differently on HR and VHR optical imagery with various spatial resolutions. (a) Planet satellite imagery (3m/pixel). (b) Aerial imagery (1 m/pixel). (c) Aerial imagery (0.3m/pixel).	7
4.1	The building footprint maps generated by FC-DenseNet [84].	19
4.2	Flowchart of the proposed approach	20
4.3	(a) The satellite imagery, and the attraction field maps in both (b) x and (c) y directions estimated by our method.	22
4.4	Overview of the proposed framework. The Img2AFM module takes an image as input and outputs two attraction field maps (AFMs) in x and y directions. Afterwards, the output is then fed into the AFM2Mask module along with the input image to generate a building mask. Notable that these two modules are trained in an end-to-end fashion.	22
5.1	Illustration of geographic peculiarities across different geolocations. The Planet satellite images are collected from (a) Munich (Germany), (b) Yaounde (Cameroon), (c) Lisbon (Portugal), and (d) Niamey (Niger), respectively. We can see that appearances of buildings in different cities are noticeably different.	26
5.2	Overview of the proposed CrossGeoNet framework.	26
5.3	Overview of the proposed approach.	28
6.1	Overview of the framework of undocumented building detection.	33
6.2	The percentage histogram of different labeled datasets.	40
6.3	The percentage histogram of different initial labeled and newly added labeled datasets for supervised learning.	41
6.4	The percentage histogram of different initial labeled and newly added labeled datasets for semi-supervised learning.	41
6.5	The percentage histogram of different newly added labeled datasets for supervised learning.	42
6.6	The percentage histogram of different newly added labeled datasets for semi-supervised learning.	42
6.7	The percentage histogram of different labeled datasets. Each set is randomly 10 times selected from the full labeled set.	44

List of Tables

2.1	Common ways to acquire building footprint maps	5
2.2	Types of remote sensing imagery to generate building footprint maps	6
6.1	The numbers of training and validation patches for the 14 selected cities. .	36
6.2	Accuracies of different semantic segmentation networks for supervised learning. (%)	37
6.3	Accuracies of different numbers of labeled sets for Efficient-UNet [160]. (%)	37
6.4	Accuracies of different initial labeled sets for Efficient-UNet [160] using the proposed sampling strategies. (%)	38
6.5	Accuracies of different numbers of labeled sets for our semi-supervised consistency learning method. (%)	38
6.6	Accuracies of different initial labeled sets for our semi-supervised consistency learning method using the proposed sampling strategies. (%)	39
6.7	Accuracies of different unlabeled sets for our semi-supervised consistency learning method. (%)	39
6.8	Accuracies of different newly added labeled sets for Efficient-UNet [160] using the proposed sampling strategies. (%)	41
6.9	Accuracies of different newly added labeled sets for our semi-supervised consistency training method using the proposed sampling strategies. (%) . .	42
6.10	Similarity of line graphs between different numbers of labeled sets and the full labeled set. Each set is randomly 10 times selected from the full labeled set.	44
6.11	Comparisons of different numbers of labeled sets for Efficient-UNet [160]. .	44

Bibliography

- [1] United Nations. Sustainable development goal 11: Make cities inclusive, safe, resilient and sustainable. <https://www.un.org/sustainabledevelopment/cities/>. Accessed: 2021-12-16.
- [2] X. Huang, Y. Cao, and J. Li. An automatic change detection method for monitoring newly constructed building areas using time-series multi-view high-resolution optical satellite images. *Remote Sensing of Environment*, 244:111802, 2020.
- [3] A. Rapoport. *Human aspects of urban form: towards a man—environment approach to urban form and design*. Elsevier, 2016.
- [4] 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.
- [5] E. Resch, R. A. Bohne, T. Kvamsdal, and J. Lohne. Impact of urban density and building height on energy use in cities. *Energy Procedia*, 96:800–814, 2016.
- [6] R. Borck. Will skyscrapers save the planet? building height limits and urban greenhouse gas emissions. *Regional Science and Urban Economics*, 58:13–25, 2016.
- [7] M. Alahmadi, P. Atkinson, and D. Martin. Estimating the spatial distribution of the population of riyadh, saudi arabia using remotely sensed built land cover and height data. *Computers, Environment and Urban Systems*, 41:167–176, 2013.
- [8] B. Oshri, A. Hu, P. Adelson, X. Chen, P. Dupas, J. Weinstein, M. Burke, D. Lobell, and S. Ermon. Infrastructure quality assessment in africa using satellite imagery and deep learning. In *Proceedings of the 24th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, pages 616–625, 2018.
- [9] Q. Li, Y. Shi, S. Auer, R. Roschlaub, K. Möst, M. Schmitt, C. Glock, and X. X. Zhu. Detection of undocumented building constructions from official geodata using a convolutional neural network. *Remote Sensing*, 12(21):3537, 2020.
- [10] B. Herfort, S. Lautenbach, J. Porto de Albuquerque, J. Anderson, and A. Zipf. The evolution of humanitarian mapping within the openstreetmap community. *Scientific reports*, 11(1):1–15, 2021.
- [11] C. H. Grohnfeldt. *Multi-sensor data fusion for multi-and hyperspectral resolution enhancement based on sparse representations*. PhD thesis, Technische Universität München, 2017.
- [12] K. Tempfli, G. Huurneman, W. Bakker, L. L. Janssen, W. Feringa, A. Gieske, K. Grabmaier, C. Hecker, J. Horn, N. Kerle, et al. *Principles of remote sensing: an introductory textbook*. International Institute for Geo-Information Science and Earth Observation, 2009.
- [13] K. G. Derpanis. The harris corner detector. *York University*, pages 1–2, 2004.
- [14] C. G. Harris, M. Stephens, et al. A combined corner and edge detector. Citeseer, 1988.
- [15] D. G. Lowe. Distinctive image features from scale-invariant keypoints. *International journal of computer vision*, 60(2):91–110, 2004.
- [16] M. Cote and P. Saeedi. Automatic rooftop extraction in nadir aerial imagery of suburban regions using corners and variational level set evolution. *IEEE transactions on geoscience and remote sensing*, 51(1):313–328, 2012.

Bibliography

- [17] M. Zangrandi, E. Baccaglini, and L. Boulard. An enhanced corner-based automatic rooftop extraction algorithm leveraging drlse segmentation. In *2015 IEEE International Geoscience and Remote Sensing Symposium (IGARSS)*, pages 1024–1027. IEEE, 2015.
- [18] M. Wang, S. Yuan, and J. Pan. Building detection in high resolution satellite urban image using segmentation, corner detection combined with adaptive windowed hough transform. In *2013 IEEE International Geoscience and Remote Sensing Symposium-IGARSS*, pages 508–511. IEEE, 2013.
- [19] R. O. Duda and P. E. Hart. Use of the hough transformation to detect lines and curves in pictures. *Communications of the ACM*, 15(1):11–15, 1972.
- [20] J. B. Burns, A. R. Hanson, and E. M. Riseman. Extracting straight lines. *IEEE transactions on pattern analysis and machine intelligence*, (4):425–455, 1986.
- [21] S. Cui, Q. Yan, and P. Reinartz. Complex building description and extraction based on hough transformation and cycle detection. *Remote sensing letters*, 3(2):151–159, 2012.
- [22] M. Izadi and P. Saeedi. Three-dimensional polygonal building model estimation from single satellite images. *IEEE Transactions on Geoscience and Remote Sensing*, 50(6):2254–2272, 2011.
- [23] C. Akinlar and C. Topal. Edlines: A real-time line segment detector with a false detection control. *Pattern Recognition Letters*, 32(13):1633–1642, 2011.
- [24] J. Wang, X. Yang, X. Qin, X. Ye, and Q. Qin. An efficient approach for automatic rectangular building extraction from very high resolution optical satellite imagery. *IEEE Geoscience and Remote Sensing Letters*, 12(3):487–491, 2014.
- [25] X. Qin, S. He, X. Yang, M. Dehghan, Q. Qin, and J. Martin. Accurate outline extraction of individual building from very high-resolution optical images. *IEEE Geoscience and Remote Sensing Letters*, 15(11):1775–1779, 2018.
- [26] M. Pesaresi, A. Gerhardinger, and F. Kayitakire. A robust built-up area presence index by anisotropic rotation-invariant textural measure. *IEEE Journal of selected topics in applied earth observations and remote sensing*, 1(3):180–192, 2008.
- [27] X. Huang and L. Zhang. A multidirectional and multiscale morphological index for automatic building extraction from multispectral geoeye-1 imagery. *Photogrammetric Engineering & Remote Sensing*, 77(7):721–732, 2011.
- [28] Q. Bi, K. Qin, H. Zhang, Y. Zhang, Z. Li, and K. Xu. A multi-scale filtering building index for building extraction in very high-resolution satellite imagery. *Remote Sensing*, 11(5):482, 2019.
- [29] G.-S. Xia, J. Huang, N. Xue, Q. Lu, and X. Zhu. Geosay: A geometric saliency for extracting buildings in remote sensing images. *Computer Vision and Image Understanding*, 186:37–47, 2019.
- [30] T. Su. Scale-variable region-merging for high resolution remote sensing image segmentation. *ISPRS Journal of Photogrammetry and Remote Sensing*, 147:319–334, 2019.
- [31] S. E. Jozdani, M. Momeni, B. A. Johnson, and M. Sattari. A regression modelling approach for optimizing segmentation scale parameters to extract buildings of different sizes. *International Journal of Remote Sensing*, 39(3):684–703, 2018.
- [32] N. Jiang, J. Zhang, H. Li, and X. Lin. Semi-automatic building extraction from high resolution imagery based on segmentation. In *2008 International Workshop on Earth Observation and Remote Sensing Applications*, pages 1–5. IEEE, 2008.
- [33] B. Li and L. Yang. Clustering accuracy analysis of building area in high spatial resolution remote sensing images based on k-means algorithm. In *2017 2nd International Conference on Frontiers of Sensors Technologies (ICFST)*, pages 174–178. IEEE, 2017.

- [34] C. Ünsalan and K. L. Boyer. A system to detect houses and residential street networks in multispectral satellite images. *Computer Vision and Image Understanding*, 98(3):423–461, 2005.
- [35] H.-G. Sohn, C.-H. Park, H.-S. Kim, and J. Heo. 3-d building extraction using ikonos multispectral images. In *Proceedings. 2005 IEEE International Geoscience and Remote Sensing Symposium, 2005. IGARSS'05.*, volume 2, pages 1432–1434. IEEE, 2005.
- [36] Y. Zhang. Optimisation of building detection in satellite images by combining multispectral classification and texture filtering. *ISPRS journal of photogrammetry and remote sensing*, 54(1):50–60, 1999.
- [37] L. Agarwal and K. S. Rajan. Fast ica based algorithm for building detection from vhr imagery. In *2015 IEEE International Geoscience and Remote Sensing Symposium (IGARSS)*, pages 1889–1892. IEEE, 2015.
- [38] L. Agarwal and K. S. Rajan. Integrating mser into a fast ica approach for improving building detection accuracy. In *IGARSS 2018-2018 IEEE International Geoscience and Remote Sensing Symposium*, pages 4831–4834. IEEE, 2018.
- [39] Ö. Ö. Karadağ, C. Senaras, and F. T. Y. Vural. Segmentation fusion for building detection using domain-specific information. *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, 8(7):3305–3315, 2015.
- [40] I. Grinias, C. Panagiotakis, and G. Tziritas. Mrf-based segmentation and unsupervised classification for building and road detection in peri-urban areas of high-resolution satellite images. *ISPRS journal of photogrammetry and remote sensing*, 122:145–166, 2016.
- [41] T.-T. Ngo, C. Collet, and V. Mazet. Automatic rectangular building detection from vhr aerial imagery using shadow and image segmentation. In *2015 IEEE international conference on image processing (ICIP)*, pages 1483–1487. IEEE, 2015.
- [42] A. O. Ok, C. Senaras, and B. Yuksel. Automated detection of arbitrarily shaped buildings in complex environments from monocular vhr optical satellite imagery. *IEEE Transactions on Geoscience and Remote Sensing*, 51(3):1701–1717, 2012.
- [43] A. O. Ok. Automated detection of buildings from single vhr multispectral images using shadow information and graph cuts. *ISPRS journal of photogrammetry and remote sensing*, 86:21–40, 2013.
- [44] K. Karantza and N. Paragios. Recognition-driven two-dimensional competing priors toward automatic and accurate building detection. *IEEE Transactions on Geoscience and Remote Sensing*, 47(1):133–144, 2008.
- [45] J. Peng, D. Zhang, and Y. Liu. An improved snake model for building detection from urban aerial images. *Pattern Recognition Letters*, 26(5):587–595, 2005.
- [46] S. Ahmady, H. Ebadi, M. V. Zouj, and H. A. Moghaddam. Automatic building extraction from high resolution aerial images using active contour model. *The International Archives of the Photogrammetry, Remote Sensing and Spatial Information Sciences*, 37:453–456, 2008.
- [47] A. J. Fazan, A. Porfirio, A. J. F. Dal Poz, and D. Poz. Building roof contours extraction from aerial imagery based on snakes and dynamic programming. 2010.
- [48] M. Turker and E. Sumer. Building-based damage detection due to earthquake using the watershed segmentation of the post-event aerial images. *International Journal of Remote Sensing*, 29(11):3073–3089, 2008.
- [49] M. Pesaresi and J. A. Benediktsson. A new approach for the morphological segmentation of high-resolution satellite imagery. *IEEE transactions on Geoscience and Remote Sensing*, 39(2):309–320, 2001.

- [50] B. E. Boser, I. M. Guyon, and V. N. Vapnik. A training algorithm for optimal margin classifiers. In *Proceedings of the fifth annual workshop on Computational learning theory*, pages 144–152, 1992.
- [51] L. Breiman. Random forests. *Machine learning*, 45(1):5–32, 2001.
- [52] E. Fix and J. L. Hodges. Discriminatory analysis. nonparametric discrimination: Consistency properties. *International Statistical Review/Revue Internationale de Statistique*, 57(3):238–247, 1989.
- [53] J. Inglada. Automatic recognition of man-made objects in high resolution optical remote sensing images by svm classification of geometric image features. *ISPRS journal of photogrammetry and remote sensing*, 62(3):236–248, 2007.
- [54] H. Baluyan, B. Joshi, A. Al Hinai, and W. L. Woon. Novel approach for rooftop detection using support vector machine. *International Scholarly Research Notices*, 2013, 2013.
- [55] F. Dornaika, A. Moujahid, A. Bosaghzadeh, Y. El Merabet, and Y. Ruichek. Object classification using hybrid holistic descriptors: Application to building detection in aerial orthophotos. *Polibits*, 51:11–17, 2015.
- [56] C. Senaras, M. Ozay, and F. T. Y. Vural. Building detection with decision fusion. *IEEE journal of selected topics in applied earth observations and remote sensing*, 6(3):1295–1304, 2013.
- [57] L. Castrejon, K. Kundu, R. Urtasun, and S. Fidler. Annotating object instances with a polygon-rnn. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 5230–5238, 2017.
- [58] Z. Li, J. D. Wegner, and A. Lucchi. Topological map extraction from overhead images. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 1715–1724, 2019.
- [59] T.-Y. Lin, P. Dollár, R. Girshick, K. He, B. Hariharan, and S. Belongie. Feature pyramid networks for object detection. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 2117–2125, 2017.
- [60] W. Zhao, C. Persello, and A. Stein. Building outline delineation: From aerial images to polygons with an improved end-to-end learning framework. *ISPRS Journal of Photogrammetry and Remote Sensing*, 175:119–131, 2021.
- [61] Y. Zhu, B. Huang, J. Gao, E. Huang, and H. Chen. Adaptive polygon generation algorithm for automatic building extraction. *IEEE Transactions on Geoscience and Remote Sensing*, 2021.
- [62] S. Wei, T. Zhang, and S. Ji. A concentric loop convolutional neural network for manual delineation level building boundary segmentation from remote sensing images. *IEEE Transactions on Geoscience and Remote Sensing*, 2021.
- [63] S. Wei and S. Ji. Graph convolutional networks for the automated production of building vector maps from aerial images. *IEEE Transactions on Geoscience and Remote Sensing*, 60:1–11, 2021.
- [64] Q. Chen, L. Wang, S. L. Waslander, and X. Liu. An end-to-end shape modeling framework for vectorized building outline generation from aerial images. *ISPRS Journal of Photogrammetry and Remote Sensing*, 170:114–126, 2020.
- [65] 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.

- [66] D. Marcos, D. Tuia, B. Kellenberger, L. Zhang, M. Bai, R. Liao, and R. Urtasun. Learning deep structured active contours end-to-end. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 8877–8885, 2018.
- [67] D. Cheng, R. Liao, S. Fidler, and R. Urtasun. Darnet: Deep active ray network for building segmentation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 7431–7439, 2019.
- [68] T. Lu, D. Ming, X. Lin, Z. Hong, X. Bai, and J. Fang. Detecting building edges from high spatial resolution remote sensing imagery using richer convolution features network. *Remote Sensing*, 10(9):1496, 2018.
- [69] Y. Liu, M.-M. Cheng, X. Hu, K. Wang, and X. Bai. Richer convolutional features for edge detection. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 3000–3009, 2017.
- [70] G. Wu, Z. Guo, X. Shi, Q. Chen, Y. Xu, R. Shibasaki, and X. Shao. A boundary regulated network for accurate roof segmentation and outline extraction. *Remote Sensing*, 10(8):1195, 2018.
- [71] 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*, 2021.
- [72] Y. Zhang, W. Li, W. Gong, Z. Wang, and J. Sun. An improved boundary-aware perceptual loss for building extraction from vhr images. *Remote Sensing*, 12(7):1195, 2020.
- [73] S. He and W. Jiang. Boundary-assisted learning for building extraction from optical remote sensing imagery. *Remote Sensing*, 13(4):760, 2021.
- [74] H. L. Yang, J. Yuan, D. Lunga, M. Laverdiere, A. Rose, and B. Bhaduri. Building extraction at scale using convolutional neural network: Mapping of the united states. *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, 11(8):2600–2614, 2018.
- [75] C. Liao, H. Hu, H. Li, X. Ge, M. Chen, C. Li, and Q. Zhu. Joint learning of contour and structure for boundary-preserved building extraction. *Remote Sensing*, 13(6):1049, 2021.
- [76] S. Chen, W. Shi, M. Zhou, M. Zhang, and Z. Xuan. Cgsanet: A contour-guided and local structure-aware encoder-decoder network for accurate building extraction from very high-resolution remote sensing imagery. *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, 2021.
- [77] Y. Zhu, Z. Liang, J. Yan, G. Chen, and X. Wang. Ed-net: Automatic building extraction from high-resolution aerial images with boundary information. *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, 14:4595–4606, 2021.
- [78] X. Zheng, L. Huan, G.-S. Xia, and J. Gong. Parsing very high resolution urban scene images by learning deep convnets with edge-aware loss. *ISPRS Journal of Photogrammetry and Remote Sensing*, 170:15–28, 2020.
- [79] K. Lee, J. H. Kim, H. Lee, J. Park, J. P. Choi, and J. Y. Hwang. Boundary-oriented binary building segmentation model with two scheme learning for aerial images. *IEEE Transactions on Geoscience and Remote Sensing*, 2021.
- [80] X. Jiang, X. Zhang, Q. Xin, X. Xi, and P. Zhang. Arbitrary-shaped building boundary-aware detection with pixel aggregation network. *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, 14:2699–2710, 2020.
- [81] 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.

Bibliography

- [82] O. Ronneberger, P. Fischer, and T. Brox. U-net: Convolutional networks for biomedical image segmentation. In *International Conference on Medical image computing and computer-assisted intervention*, pages 234–241. Springer, 2015.
- [83] 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.
- [84] S. Jégou, M. Drozdal, D. Vazquez, A. Romero, and Y. Bengio. The one hundred layers tiramisu: Fully convolutional densenets for semantic segmentation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops*, pages 11–19, 2017.
- [85] L.-C. Chen, G. Papandreou, I. Kokkinos, K. Murphy, and A. L. Yuille. Semantic image segmentation with deep convolutional nets and fully connected crfs. *arXiv preprint arXiv:1412.7062*, 2014.
- [86] 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.
- [87] L.-C. Chen, G. Papandreou, F. Schroff, and H. Adam. Rethinking atrous convolution for semantic image segmentation. *arXiv preprint arXiv:1706.05587*, 2017.
- [88] L.-C. Chen, Y. Zhu, G. Papandreou, F. Schroff, and H. Adam. Encoder-decoder with atrous separable convolution for semantic image segmentation. In *Proceedings of the European conference on computer vision (ECCV)*, pages 801–818, 2018.
- [89] I. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville, and Y. Bengio. Generative adversarial nets. *Advances in neural information processing systems*, 27, 2014.
- [90] G. Sun, H. Huang, A. Zhang, F. Li, H. Zhao, and H. Fu. Fusion of multiscale convolutional neural networks for building extraction in very high-resolution images. *Remote Sensing*, 11(3):227, 2019.
- [91] S. Ji, S. Wei, and M. Lu. A scale robust convolutional neural network for automatic building extraction from aerial and satellite imagery. *International journal of remote sensing*, 40(9):3308–3322, 2019.
- [92] S. Wei, S. Ji, and M. Lu. Toward automatic building footprint delineation from aerial images using cnn and regularization. *IEEE Transactions on Geoscience and Remote Sensing*, 58(3):2178–2189, 2019.
- [93] J. Ma, L. Wu, X. Tang, F. Liu, X. Zhang, and L. Jiao. Building extraction of aerial images by a global and multi-scale encoder-decoder network. *Remote Sensing*, 12(15):2350, 2020.
- [94] Y. Liu, B. Fan, L. Wang, J. Bai, S. Xiang, and C. Pan. Semantic labeling in very high resolution images via a self-cascaded convolutional neural network. *ISPRS journal of photogrammetry and remote sensing*, 145:78–95, 2018.
- [95] G. Wu, X. Shao, Z. Guo, Q. Chen, W. Yuan, X. Shi, Y. Xu, and R. Shibasaki. Automatic building segmentation of aerial imagery using multi-constraint fully convolutional networks. *Remote Sensing*, 10(3):407, 2018.
- [96] J. Yu and S. Chan. Snlrx++ for building extraction from high-resolution remote sensing images. *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, 2021.
- [97] 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.

- [98] W. Yuan and W. Xu. Msst-net: A multi-scale adaptive network for building extraction from remote sensing images based on swin transformer. *Remote Sensing*, 13(23):4743, 2021.
- [99] P. Liu, X. Liu, M. Liu, Q. Shi, J. Yang, X. Xu, and Y. Zhang. Building footprint extraction from high-resolution images via spatial residual inception convolutional neural network. *Remote Sensing*, 11(7):830, 2019.
- [100] Y. Zhang, W. Gong, J. Sun, and W. Li. Web-net: A novel nest networks with ultra-hierarchical sampling for building extraction from aerial imageries. *Remote Sensing*, 11(16):1897, 2019.
- [101] 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*, 2021.
- [102] H. Guo, X. Su, S. Tang, B. Du, and L. Zhang. Scale-robust deep-supervision network for mapping building footprints from high-resolution remote sensing images. *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, 14:10091–10100, 2021.
- [103] C. Li, L. Fu, Q. Zhu, J. Zhu, Z. Fang, Y. Xie, Y. Guo, and Y. Gong. Attention enhanced u-net for building extraction from farmland based on google and worldview-2 remote sensing images. *Remote Sensing*, 13(21):4411, 2021.
- [104] X. Pan, F. Yang, L. Gao, Z. Chen, B. Zhang, H. Fan, and J. Ren. Building extraction from high-resolution aerial imagery using a generative adversarial network with spatial and channel attention mechanisms. *Remote Sensing*, 11(8):917, 2019.
- [105] Q. Zhu, Z. Li, Y. Zhang, and Q. Guan. Building extraction from high spatial resolution remote sensing images via multiscale-aware and segmentation-prior conditional random fields. *Remote Sensing*, 12(23):3983, 2020.
- [106] W. Deng, Q. Shi, and J. Li. Attention-gate-based encoder–decoder network for automatical building extraction. *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, 14:2611–2620, 2021.
- [107] W. Kang, Y. Xiang, F. Wang, and H. You. Eu-net: An efficient fully convolutional network for building extraction from optical remote sensing images. *Remote Sensing*, 11(23):2813, 2019.
- [108] 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*, 2020.
- [109] 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.
- [110] D. Zhou, G. Wang, G. He, R. Yin, T. Long, Z. Zhang, S. Chen, and B. Luo. A large-scale mapping scheme for urban building from gaofen-2 images using deep learning and hierarchical approach. *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, 14:11530–11545, 2021.
- [111] H. Guo, Q. Shi, B. Du, L. Zhang, D. Wang, and H. Ding. Scene-driven multitask parallel attention network for building extraction in high-resolution remote sensing images. *IEEE Transactions on Geoscience and Remote Sensing*, 59(5):4287–4306, 2020.
- [112] X. Li, X. Yao, and Y. Fang. Building-a-nets: robust building extraction from high-resolution remote sensing images with adversarial networks. *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, 11(10):3680–3687, 2018.
- [113] L. Ding, H. Tang, Y. Liu, Y. Shi, X. X. Zhu, and L. Bruzzone. Adversarial shape learning for building extraction in vhr remote sensing images. *arXiv preprint arXiv:2102.11262*, 2021.

Bibliography

- [114] S. Zorzi and F. Fraundorfer. Regularization of building boundaries in satellite images using adversarial and regularized losses. In *IGARSS 2019-2019 IEEE International Geoscience and Remote Sensing Symposium*, pages 5140–5143. IEEE, 2019.
- [115] S. Zorzi, K. Bittner, and F. Fraundorfer. Machine-learned regularization and polygonization of building segmentation masks. In *2020 25th International Conference on Pattern Recognition (ICPR)*, pages 3098–3105. IEEE, 2021.
- [116] Y. Shi, Q. Li, and X. Zhu. Building footprint extraction with graph convolutional network. In *IGARSS 2019-2019 IEEE International Geoscience and Remote Sensing Symposium*, pages 5136–5139. IEEE, 2019.
- [117] 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.
- [118] Y. Shi, Q. Li, and X. X. Zhu. Building extraction by gated graph convolutional neural network with deep structured feature embedding. In *IGARSS 2020-2020 IEEE International Geoscience and Remote Sensing Symposium*, pages 3509–3512. IEEE.
- [119] J. Yuan. Learning building extraction in aerial scenes with convolutional networks. *IEEE transactions on pattern analysis and machine intelligence*, 40(11):2793–2798, 2017.
- [120] B. Bischke, P. Helber, J. Folz, D. Borth, and A. Dengel. Multi-task learning for segmentation of building footprints with deep neural networks. In *2019 IEEE International Conference on Image Processing (ICIP)*, pages 1480–1484. IEEE, 2019.
- [121] N. Girard, D. Smirnov, J. Solomon, and Y. Tarabalka. Polygonal building extraction by frame field learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 5891–5900, 2021.
- [122] L. Xia, X. Zhang, J. Zhang, H. Yang, and T. Chen. Building extraction from very-high-resolution remote sensing images using semi-supervised semantic edge detection. *Remote Sensing*, 13(11):2187, 2021.
- [123] J. Chen, F. He, Y. Zhang, G. Sun, and M. Deng. Spmf-net: Weakly supervised building segmentation by combining superpixel pooling and multi-scale feature fusion. *Remote Sensing*, 12(6):1049, 2020.
- [124] Z. Li, X. Zhang, P. Xiao, and Z. Zheng. On the effectiveness of weakly supervised semantic segmentation for building extraction from high-resolution remote sensing imagery. *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, 14:3266–3281, 2021.
- [125] X. Yan, L. Shen, J. Wang, X. Deng, and Z. Li. Msg-sr-net: A weakly supervised network integrating multi-scale generation and super-pixel refinement for building extraction from high-resolution remotely sensed imageries. *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, 2021.
- [126] M. U. Rafique and N. Jacobs. Weakly supervised building segmentation from aerial images. In *IGARSS 2019-2019 IEEE International Geoscience and Remote Sensing Symposium*, pages 3955–3958. IEEE, 2019.
- [127] J.-H. Lee, C. Kim, and S. Sull. Weakly supervised segmentation of small buildings with point labels. In *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, pages 7406–7415, 2021.
- [128] J. Wang, C. HQ Ding, S. Chen, C. He, and B. Luo. Semi-supervised remote sensing image semantic segmentation via consistency regularization and average update of pseudo-label. *Remote Sensing*, 12(21):3603, 2020.

- [129] 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.
- [130] X. Li, M. Luo, S. Ji, L. Zhang, and M. Lu. Evaluating generative adversarial networks based image-level domain transfer for multi-source remote sensing image segmentation and object detection. *International Journal of Remote Sensing*, 41(19):7343–7367, 2020.
- [131] N. Makkar, L. Yang, and S. Prasad. Adversarial learning based discriminative domain adaptation for geospatial image analysis. *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, 15:150–162, 2021.
- [132] Y. Cai, Y. Yang, Q. Zheng, Z. Shen, Y. Shang, J. Yin, and Z. Shi. Bifdanet: Unsupervised bidirectional domain adaptation for semantic segmentation of remote sensing images. *Remote Sensing*, 14(1):190, 2022.
- [133] Y. Na, J. H. Kim, K. Lee, J. Park, J. Y. Hwang, and J. P. Choi. Domain adaptive transfer attack-based segmentation networks for building extraction from aerial images. *IEEE Transactions on Geoscience and Remote Sensing*, 59(6):5171–5182, 2020.
- [134] L. Shi, Z. Wang, B. Pan, and Z. Shi. An end-to-end network for remote sensing imagery semantic segmentation via joint pixel-and representation-level domain adaptation. *IEEE Geoscience and Remote Sensing Letters*, 2020.
- [135] X. Yao, Y. Wang, Y. Wu, and Z. Liang. Weakly-supervised domain adaptation with adversarial entropy for building segmentation in cross-domain aerial imagery. *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, 14:8407–8418, 2021.
- [136] 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*, 2021.
- [137] H. Liu, J. Luo, B. Huang, X. Hu, Y. Sun, Y. Yang, N. Xu, and N. Zhou. De-net: Deep encoding network for building extraction from high-resolution remote sensing imagery. *Remote Sensing*, 11(20):2380, 2019.
- [138] M. Chen, J. Wu, L. Liu, W. Zhao, F. Tian, Q. Shen, B. Zhao, and R. Du. Dr-net: An improved network for building extraction from high resolution remote sensing image. *Remote Sensing*, 13(2):294, 2021.
- [139] 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.
- [140] Y. Liu, J. Zhou, W. Qi, X. Li, L. Gross, Q. Shao, Z. Zhao, L. Ni, X. Fan, and Z. Li. Arc-net: An efficient network for building extraction from high-resolution aerial images. *IEEE Access*, 8:154997–155010, 2020.
- [141] H. Huang, Y. Chen, and R. Wang. A lightweight network for building extraction from remote sensing images. *IEEE Transactions on Geoscience and Remote Sensing*, 2021.
- [142] 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*, 2020.
- [143] N. Xue, S. Bai, F. Wang, G.-S. Xia, T. Wu, and L. Zhang. Learning attraction field representation for robust line segment detection. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 1595–1603, 2019.
- [144] N. Xue, S. Bai, F.-D. Wang, G.-S. Xia, T. Wu, L. Zhang, and P. H. Torr. Learning regional attraction for line segment detection. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2019.

Bibliography

- [145] S. Srivastava, M. Volpi, and D. Tuia. Joint height estimation and semantic labeling of monocular aerial images with cnns. In *2017 IEEE International Geoscience and Remote Sensing Symposium (IGARSS)*, pages 5173–5176. IEEE, 2017.
- [146] L. Mou and X. X. Zhu. Vehicle instance segmentation from aerial image and video using a multitask learning residual fully convolutional network. *IEEE Transactions on Geoscience and Remote Sensing*, 56(11):6699–6711, 2018.
- [147] O. Tasar, S. Happy, Y. Tarabalka, and P. Alliez. Colormapgan: Unsupervised domain adaptation for semantic segmentation using color mapping generative adversarial networks. *IEEE Transactions on Geoscience and Remote Sensing*, 58(10):7178–7193, 2020.
- [148] A. Mustafa and A. Hilton. Semantically coherent co-segmentation and reconstruction of dynamic scenes. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 422–431, 2017.
- [149] K. Papoutsakis, C. Panagiotakis, and A. A. Argyros. Temporal action co-segmentation in 3d motion capture data and videos. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 6827–6836, 2017.
- [150] W. Wang, X. Lu, J. Shen, D. J. Crandall, and L. Shao. Zero-shot video object segmentation via attentive graph neural networks. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 9236–9245, 2019.
- [151] W. Li, O. H. Jafari, and C. Rother. Deep object co-segmentation. In *Asian Conference on Computer Vision*, pages 638–653. Springer, 2018.
- [152] J. Johnson, A. Alahi, and L. Fei-Fei. Perceptual losses for real-time style transfer and super-resolution. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 694–711. Springer, 2016.
- [153] Y. Ouali, C. Hudelot, and M. Tami. Semi-supervised semantic segmentation with cross-consistency training. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 12674–12684, 2020.
- [154] Q. Xie, Z. Dai, E. Hovy, T. Luong, and Q. Le. Unsupervised data augmentation for consistency training. *Advances in Neural Information Processing Systems*, 33:6256–6268, 2020.
- [155] Z. Wu, C. Shen, and A. v. d. Hengel. Bridging category-level and instance-level semantic image segmentation. *arXiv preprint arXiv:1605.06885*, 2016.
- [156] R. Roschlaub, K. Möst, and T. Krey. Automated classification of building roofs for the updating of 3d building models using heuristic methods. *PFG - Journal of Photogrammetry, Remote Sensing and Geoinformation Science*, 88, 2020.
- [157] S. Geßler, T. Krey, K. Möst, and R. Roschlaub. Mit Datenfusionierung Mehrwerte schaffen – Ein Expertensystem zur Baufallerkundung. *DVW-Mitteilungen*, 2:159–187, 2019.
- [158] D. Roth and K. Small. Margin-based active learning for structured output spaces. In *European Conference on Machine Learning*, pages 413–424. Springer, 2006.
- [159] M.-F. Balcan, A. Broder, and T. Zhang. Margin based active learning. In *International Conference on Computational Learning Theory*, pages 35–50. Springer, 2007.
- [160] B. Baheti, S. Innani, S. Gajre, and S. Talbar. Eff-unet: A novel architecture for semantic segmentation in unstructured environment. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops*, pages 358–359, 2020.
- [161] M. Tan and Q. Le. Efficientnet: Rethinking model scaling for convolutional neural networks. In *International Conference on Machine Learning*, pages 6105–6114. PMLR, 2019.
- [162] K. Simonyan and A. Zisserman. Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556*, 2014.

Appendices

Appendices

A Li, Qingyu, Lichao Mou, Yuansheng Hua, Yilei Shi, and Xiao Xiang Zhu. Building Footprint Generation Through Convolutional Neural Networks with Attraction Field Representation. *IEEE Transactions on Geoscience and Remote Sensing*, vol. 60, pp. 1-17, 2022, Art no. 5609017, doi: 10.1109/TGRS.2021.3109844.

Building Footprint Generation Through Convolutional Neural Networks With Attraction Field Representation

Qingyu Li, *Student Member, IEEE*, Lichao Mou[✉], Yuansheng Hua[✉], *Graduate Student Member, IEEE*, Yilei Shi, *Member, IEEE*, and Xiao Xiang Zhu[✉], *Fellow, IEEE*

Abstract—Building footprint generation is a vital task in a wide range of applications, including, to name a few, land use management, urban planning and monitoring, and geographical database updating. Most existing approaches addressing this problem fall back on convolutional neural networks (CNNs) to learn semantic masks of buildings. However, one limitation of their results is blurred building boundaries. To address this, we propose to learn attraction field representation for building boundaries, which is capable of providing an enhanced representation power. Our method comprises two elemental modules: an Img2AFM module and an AFM2Mask module. More specifically, the former aims at learning an attraction field representation conditioned on an input image, which is capable of enhancing building boundaries and suppressing the background. The latter module predicts segmentation masks of buildings using the learned attraction field map. The proposed method is evaluated on three datasets with different spatial resolutions: the ISPRS dataset, the INRIA dataset, and the Planet dataset. From experimental results, we find that the proposed framework can well preserve geometric shapes and sharp boundaries of buildings, which brings significant improvements over other competitors. The trained model and code are available at https://github.com/lqycrystal/AFM_building.

Index Terms—Attraction field map (AFM), building footprint, convolutional neural network (CNN), semantic segmentation.

Manuscript received February 16, 2021; revised June 21, 2021; accepted August 19, 2021. Date of publication September 15, 2021; date of current version January 31, 2022. This work was supported in part by the European Research Council (ERC) through the European Union's Horizon 2020 Research and Innovation Programme under Grant Agreement ERC-2016-StG-714087 (*So2Sat*), in part by the Helmholtz Association through the Framework of Helmholtz AI under Grant ZT-I-PF-5-01 [Local Unit "Munich Unit @Aeronautics, Space and Transport (MASTr)], in part by the Helmholtz Excellent Professorship "Data Science in Earth Observation—Big Data Fusion for Urban Research" under Grant W2-W3-100, in part by the German Federal Ministry of Education and Research (BMBF) in the framework of the International Future AI Lab "AI4EO—Artificial Intelligence for Earth Observation: Reasoning, Uncertainties, Ethics and Beyond" under Grant 01DD20001, and in part by the project "Investigation of Building Cases Using AI" funded by the Bavarian State Ministry of Finance and Regional Identity (StMFH) and the Bavarian Agency for Digitization, High-Speed Internet and Surveying. (*Corresponding author: Xiao Xiang Zhu.*)

Qingyu Li, Lichao Mou, Yuansheng Hua, and Xiao Xiang Zhu are with Data Science in Earth Observation, Technische Universität München (TUM), 80333 Munich, Germany, and also with the Remote Sensing Technology Institute (IMF), German Aerospace Center (DLR), 82234 Weßling, Germany (e-mail: qingyu.li@tum.de; lichao.mou@dlr.de; yuansheng.hua@dlr.de; xiaoxiang.zhu@dlr.de).

Yilei Shi is with the Chair of Remote Sensing Technology, Technische Universität München (TUM), 80333 Munich, Germany (e-mail: yilei.shi@tum.de).

Digital Object Identifier 10.1109/TGRS.2021.3109844

I. INTRODUCTION

AUTOMATIC building footprint generation from remote sensing data has been of great interest in the community for a range of applications, such as land use management, urban planning and monitoring, and disaster management. However, accurate and reliable building footprint generation remains particularly challenging due to two reasons. On the one hand, different materials and structures lead to large variations of buildings in terms of color, shape, size, and texture. On the other hand, buildings and other man-made objects (e.g., roads and sidewalks) share similar spectral signatures, which can result in a low between-class variability.

Early efforts have been gone into seeking out hand-crafted features of being to effectively exploit spectral, structural, and context information. For example, Huang *et al.* [1] propose a framework for automatic building extraction, which utilizes spectral, geometrical, and contextual features extracted from imagery. Nonetheless, these methods still fail to satisfy accuracy requirements because they rely on a heuristic feature design procedure and usually have poor generalization capabilities.

More recently, convolutional neural networks (CNNs) have surpassed traditional methods in many remote sensing tasks [2]–[10]. CNNs can directly learn feature representations from the raw data; thus, they provide an end-to-end solution to generate building footprints from remote sensing data. Most of the studies in this field assign a label "building" or "non-building" to every pixel in the image, thus yielding semantic masks of buildings. The existing CNNs seem to be able to deliver very promising segmentation results for the purpose of building footprint generation at a large scale (cf. Fig. 1). However, when we zoom in on some segmentation masks (see results from U-Net [11] in Fig. 1), it can be clearly seen that such results are not that perfect, and the boundaries of some buildings are blurred.

We have observed that buildings usually have clear patterns (e.g., corners and straight lines). Therefore, geometric primitives of buildings can be exploited as the most distinguishable features for extraction purposes. There have been several works based on this idea [12]–[15]. In this work, we want to exploit building boundaries as a primary visual cue to achieve our task.



Fig. 1. Building footprints generated by U-Net [11] and our proposed method (U-Net with attraction field representation) at large scale and two zoomed in areas.

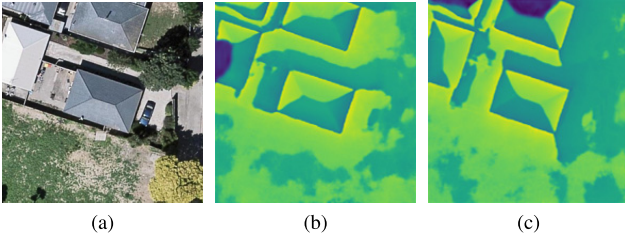


Fig. 2. (a) Satellite imagery, and the AFMs in both (b) x - and (c) y -directions estimated by our method.

Recently, attraction field representation is used for the task of line segment detection in computer vision [16], which seeks the most attractive line segment for each pixel. Our observation is that, when building boundaries in remote sensing images are represented by the attraction field, they can be greatly enhanced, while background clutters (e.g., car, courtyard, and road) are suppressed. Fig. 2 shows an example. Motivated by this observation, in this work, we want to make use of the attraction field to represent buildings and propose an end-to-end trainable network for automatic building footprint generation. This network consists of two modules: Img2AFM and AFM2Mask. The former takes as input an image and is responsible for learning a corresponding attraction field map (AFM) using a CNN. By doing so, fine-grained building boundaries can be preserved, and the impact of background clutters can be alleviated. The latter module learns another subnetwork to obtain semantic masks of buildings from augmented building edges in the learned AFM. Note that both these two modules are jointly optimized. In addition, the AFM2Mask module is flexible enough to use different semantic segmentation network architectures.

This work's contributions are threefold.

- 1) We propose to use the boundary-aware attraction field to represent building footprints in remote sensing images. This helps to enhance building boundaries while suppressing the impact of background clutters. To the best of our knowledge, it is the first work that utilizes the attraction field for the task of building footprint generation.
- 2) We propose a novel network that first learns an AFM by a subnetwork, termed Img2AFM, and then uses another subnetwork called AFM2Mask to reconstruct

segmentation masks of buildings. These two modules are trained in an end-to-end fashion.

- 3) The proposed framework obtains satisfactory performance on three datasets with different spatial resolutions, including ISPRS, INRIA, and Planet datasets. Compared with naive semantic segmentation networks and networks with other visual cues (e.g., building boundary maps), our method can significantly improve accuracies in terms of both semantic mask and boundary.

The remainder of this article is organized as follows. Related work is reviewed in Section II. Section III details the proposed framework for building footprint generation. The experiments are described in Section IV. Results and discussion are provided in Section V. Eventually, Section VI summarizes this work.

II. RELATED WORK

There are a significant number of studies working on building footprint generation from remote sensing imagery. According to used visual cues, they can be categorized into three classes: semantic mask, corner, and boundary of the building.

A. Building Footprint Generation Based on the Semantic Mask

Most methods for building footprint generation involve learning semantic masks of buildings from remote sensing imagery. Early efforts include segmentation-, classification-, and index-based methods. The segmentation-based methods extract buildings using image segmentation algorithms. For example, based on a two-level graph theory, Ok [17] proposes a segmentation approach to identify building regions. For classification-based methods, building masks are extracted by machine-learning classifiers which take spectral information and/or spatial features as input to make a prediction for each pixel. For instance, Turker and Koc-San [18] utilize a support vector machine (SVM) to identify building regions based on spectral bands and the normalized difference vegetation index (NDVI). The objective of index-based approaches is to design a feature index that can be directly applied to obtain building regions without any classification or segmentation process. Morphological building index (MBI) [19] is a widely used one, and this index integrates multiscale and multidirectional morphological operators. However, a general limitation of these early works is the use of handcrafted features and complex feature engineering, which leads to a poor generalization.

Instead of the heuristic design of features, CNNs can offer a better generalization capability. Driven by recent advances in semantic segmentation networks, results of building footprint generation have been significantly improved. These networks are usually fully convolutional network (FCN) [20] and encoder-decoder architecture, such as U-Net [11], SegNet [21], and FC-DenseNet [22]. In [23], FCN has been demonstrated to be effective in processing large amounts of remote sensing data and providing reliable building segmentation results. SegNet is used in [24] to

generate the first seamless building footprint map for the United States. In order to improve the accuracy of segmenting large buildings, a U-Net-based architecture is proposed in [25], where original images and their downsampled counterparts are taken as inputs of two branches sharing the same weights. In [26], an adversarial training strategy is proposed for building extraction from remote sensing imagery, and FC-DenseNet is exploited as a base semantic segmentation network to generate accurate building footprints.

However, many experiments show that predicted semantic masks of buildings from CNNs are still not that satisfactory, where building boundaries are blurred. In this regard, signed-distance transform (SDT) [27] is proposed to represent building footprints. The signed-distance function value is derived as the distance from a pixel to its closest point on a building boundary; positive values indicate the interior of a building and negative values otherwise. Then, the learning problem of the SDT representation can be regarded as a multiclass classification problem, which categorizes signed-distance values into a certain number of classes [24]. Compared to the widely used binary building mask, SDT can encode more fine-grained information for network learning.

B. Building Footprint Generation Based on the Corner

Some algorithms generate building footprints based on geometrical primitives, such as building corners. In these methods, geometric primitives are first detected and then grouped together to reconstruct individual building hypotheses. A building corner refers to a point with its local neighborhoods in two varying line segment directions and is invariant to translation, rotation, and illumination [28]. Early studies extract building corners with the help of some point feature operators, such as Harris corner detector [29] and scale-invariant feature transform (SIFT) operator [30]. Cote and Saeedi [12] and Zangrandi *et al.* [31] employ a Harris corner detector to extract corner points of buildings. Afterward, these detected corner points are connected in the order of their polar angles with respect to building central markers. By doing so, polygonal representations of buildings can be constructed. In [32], SIFT is exploited to extract corners that are regarded as seed points to estimate rectangle shapes of buildings with a region growing method.

With the development of keypoint detection networks, several novel studies propose to delineate building footprints by detecting corner points using CNNs. PolyMapper [33] extracts corner points with a CNN in the first stage and then connects them by a recurrent neural network (RNN) to realize closed polygon representations of individual buildings. The other research [34] utilizes the same pipeline as PolyMapper [33], and various blocks are integrated to enhance the feature extraction and object detection modules. Another method [13] also exploits a CNN to detect corners but adopts a fully geometric-based grouping strategy without any deep feature learning. Recently, Girard's method [35] proposes to learn a frame field output instead of building corners. The frame field is regarded as a geometric feature that can help to improve the segmentation of buildings.

C. Building Footprint Generation Based on the Boundary

Building boundary is another commonly used geometric primitive and can be taken as a primary visual cue to generate building footprints. Early works extract building boundaries from remote sensing data in two steps. Given that lines are strongly relevant to building boundaries, the first step is to detect line segments. Afterward, the extracted lines are grouped to form closed boundaries for individual buildings. A commonly used line detection algorithm is the Hough transformation [36] that utilizes a voting procedure to find straight lines in parameter space. Compared with the Hough transformation, the Burns algorithm [37] only uses gradient orientations and, therefore, requires a relatively lower computation cost. In [14] and [38], line segment sets are extracted with the Hough transformation and the Burns algorithm. Then, intersection nodes of the two line segment sets are employed to build a structural graph. Finally, building boundaries are identified with a graph search algorithm. However, both Hough transformation and Burns algorithm highly depend on parameter settings and have a very high false alarm rate. In this regard, EDLines [39] are proposed to avoid parameter tuning. Moreover, it has a faster computation speed and a lower false alarm rate. In [40] and [41], EDLines are, therefore, adopted for the automatic extraction of line segments, but they make use of different strategies to group these line segments.

These early works still encounter issues when dealing with more complex building shapes and large-scale applications. Considering that, nowadays, CNNs are the de facto leading approach for building footprint generation tasks, two novel works, [15] and [42], propose to learn building boundaries in their end-to-end CNNs. Marcos *et al.* [15] present a method termed deep structured active contours (DSACs), which learns active contour model (ACM) [43] parameterizations per instance using a CNN. Although DSAC improves geometric correctness, results are still not that satisfactory, e.g., there exist blob-like shapes and some self-intersections of building. Besides, the representation of boundary points in DSAC adopts Euclidean coordinates, which leads to extra computational overheads during energy minimization. On this point, another research [42] proposes to use polar coordinates, as this can not only simplify the energy function but also prevent self-intersection. However, these two methods still have two limitations. On the one hand, the initialization of them relies on external methods that are not included in an end-to-end learning process. On the other hand, their results are promising only in very high-resolution remote sensing images where strong geometric priors exist.

III. METHODOLOGY

In this work, we explicitly take building boundaries as a primary visual cue. By doing so, building footprint generation tasks can be benefited from the precise delineation of building boundaries. In this section, an overview of the proposed approach is first presented. Then, two key modules, Img2AFM and AFM2Mask, are introduced in detail, respectively. Finally, the method of integrating and jointly optimizing the two modules in an end-to-end architecture is described.

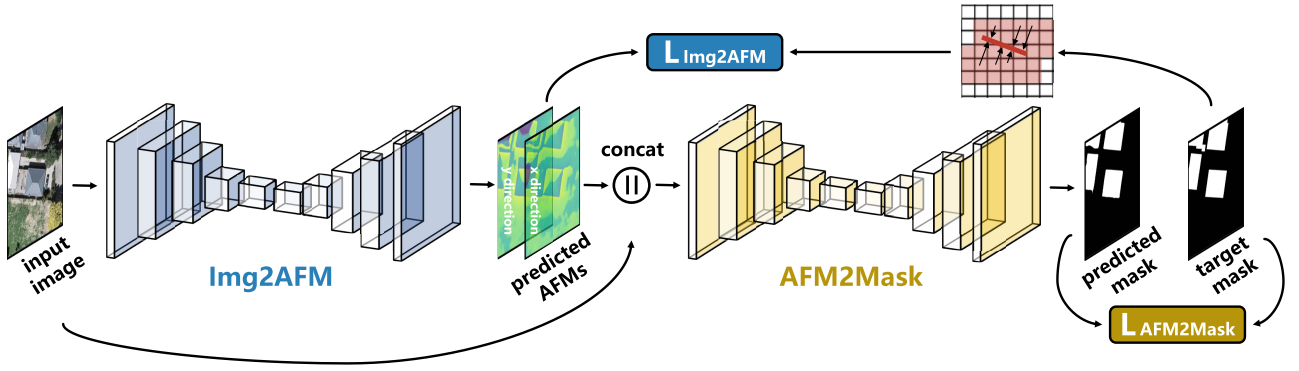


Fig. 3. Overview of the proposed framework. The Img2AFM module takes an image as input and outputs two AFMs in x - and y -directions. Afterward, the output is then fed into the AFM2Mask module along with the input image to generate a building mask. Notable that these two modules are trained in an end-to-end fashion.

A. Overview

As shown in Fig. 3, the proposed method consists of two modules. The Img2AFM module exploits a U-Net architecture to learn the attraction field representation, which can enhance building boundaries and suppress background clusters. It takes an image as input and outputs two AFMs in x - and y -directions. Afterward, the output is then fed into the AFM2Mask module along with the input image to generate a building mask. Moreover, the AFM2Mask module is very flexible to utilize different semantic segmentation networks. Note that these two modules can be integrated into an end-to-end framework and optimized jointly. In this way, the optimal output can be obtained by the coadaptation of these two modules.

B. Img2AFM Module

1) *Definition of Attraction Field Map*: An image I can be regarded a lattice. Let $E = \{\mathbf{e}_1, \mathbf{e}_2, \dots, \mathbf{e}_n\}$ be the set of building line segments in the image lattice with n being the number of building line segments. A building line segment \mathbf{e}_i is represented by two end points \mathbf{p}_i^a and \mathbf{p}_i^b . For the sake of simplicity, the set E is named boundary map in our case. The boundary map characterizing all building boundaries in the ground reference is shown in Fig. 4(c).

For each pixel, we try to find its most “attractive” building line segment that is the closest to it. Following this criterion, a region partition map R is formed by partitioning all pixels into n regions and assigning each pixel $\mathbf{x} \in I$ to its closest building line segment. R_i denotes a region for the building line segment \mathbf{e}_i in E . Specifically, in order to derive the distance between a pixel \mathbf{x} and a building line segment \mathbf{e}_i , the pixel \mathbf{x} is first projected to the straight line passing through \mathbf{e}_i . If the projection point is not on \mathbf{e}_i , the nearest endpoint is utilized as the projection point. The definition of the projection point \mathbf{p}' is

$$\mathbf{p}' = \mathbf{p}_i^a + c_x \cdot (\mathbf{p}_i^b - \mathbf{p}_i^a). \quad (1)$$

When $c_x \in (0, 1)$, \mathbf{p}' belongs to the original point-to-line projection, and if $c_x = 0$ or 1 , \mathbf{p}' is its nearest endpoint of \mathbf{e}_i .

Then, the distance $d(\mathbf{x}, \mathbf{e}_i)$ between \mathbf{x} and \mathbf{e}_i can be defined as the Euclidean distance between the pixel and the projection point. Then, R_i in the image lattice for \mathbf{e}_i can be defined as

$$R_i = \{\mathbf{x} \mid \mathbf{x} \in I; d(\mathbf{x}, \mathbf{e}_i) < d(\mathbf{x}, \mathbf{e}_j) \forall j \neq i, \mathbf{e}_j \in E\}. \quad (2)$$

It should be noted that $R_i \cap R_j = \emptyset$ and $\cup_{i=1}^n R_i = R$. Fig. 4 shows an example that, for the green building line segment, its corresponding region partition map is highlighted in green.

Afterward, the geometric property of a building line segment can be characterized by a 2-D attraction of all individual pixels in R_i . For instance, the attraction function of the pixel \mathbf{x} in R_i is defined as

$$\mathbf{a}_i(\mathbf{x}) = \mathbf{p}' - \mathbf{x}. \quad (3)$$

When $c_x \in (0, 1)$, the attraction vector is perpendicular to the line segment. Fig. 4(d) shows the attraction vectors of the green line segment.

Finally, by enumerating (3) over all pixels in I , the AFM A with respect to E can be obtained as follows:

$$A = \{\mathbf{a}(\mathbf{x}) \mid \mathbf{x} \in I\}. \quad (4)$$

The superiority of AFM lies in two aspects compared with the boundary map used in previous studies (see [15] and [42]). One is that the geometry of boundaries can be depicted more precisely by the AFM, while the boundary map is only characterized by few pixels. Thus, directly learning boundary maps can lead to a zig-zag effect that results from the extreme imbalance between the number of boundary pixels and that of nonboundary pixels. The other benefit is that the AFM associates each line segment with a support region, which avoids the blurring effect.

2) *Learning Attraction Field Map*: Each pixel in the attraction field representation has two components (x - and y -directions) that are represented by attraction vectors from it to its projection point. In this respect, an attraction field representation can be regarded as a 2-D feature map, which is feasible to be learned by a network. Hence, in this article, we view the learning of the AFM as a dense prediction problem and solve it using a semantic segmentation network architecture. Among all semantic segmentation networks, U-Net

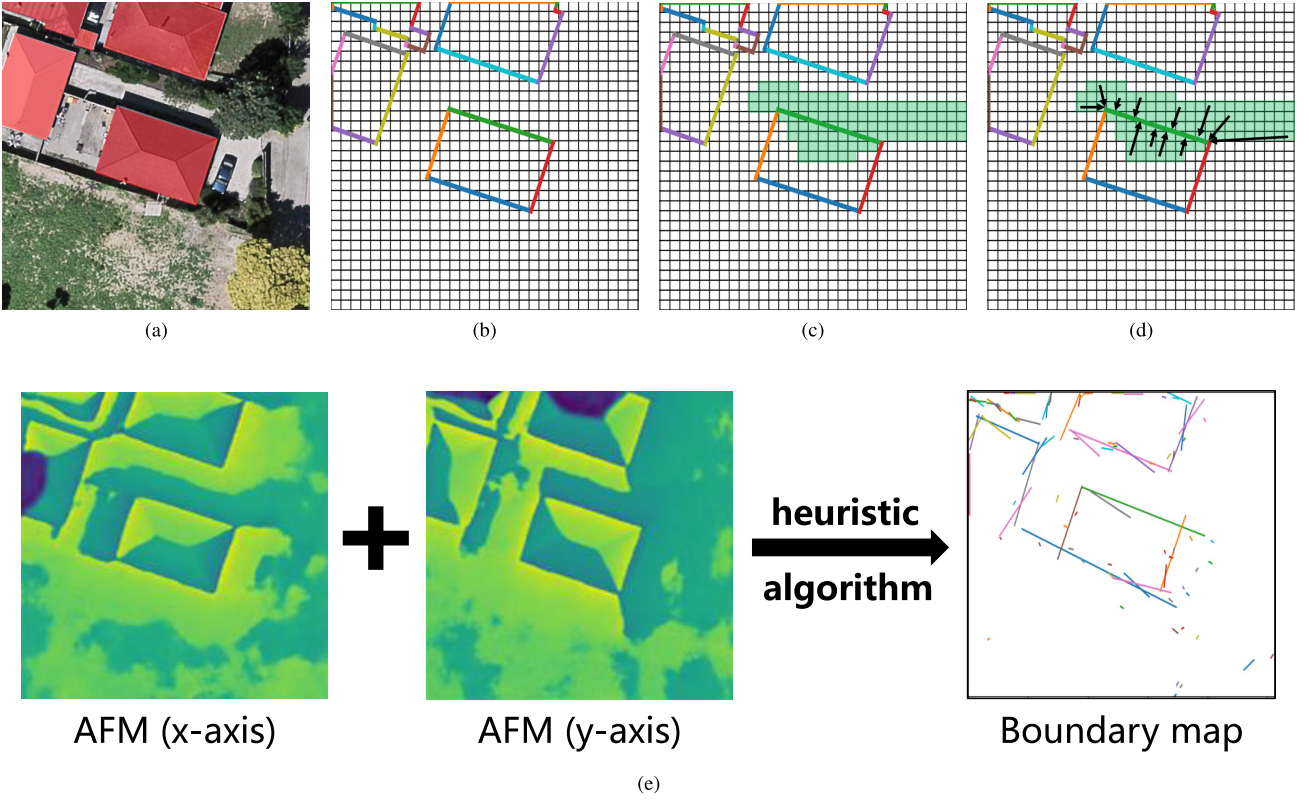


Fig. 4. (a) and (b) Semantic masks and boundaries of buildings in an image. (c) and (d) Region partition map and attraction vectors of the green building line segment according to the method in [16]. (e) Recovered boundary map obtained by the heuristic algorithm in [16].

is more favorable than others for this task. Because learning the attraction field representation relies heavily on low-level visual cues (e.g., object edges) that exist in lower layers, and multiscale skip connections of U-Net are able to effectively use such information. In fact, in our experiments, we found that taking other network architectures as the Img2AFM module fails.

C. AFM2Mask Module

By learning the AFM, a representation encoding building boundaries can be obtained. Then, we need to remap the learned AFM into building masks. In [16], a heuristic algorithm has been proposed to recover line segments from the AFM. In this heuristic algorithm, attraction vectors are rearranged mathematically to generate a proposal map of line segments, and final line segments are then extracted with a greedy grouping strategy. However, we found that, in our building footprint generation task, the recovered boundary map from this algorithm is not satisfactory [cf. Fig. 4(e)] since there is a relatively high false alarm rate [see short line segments in Fig. 4(e)]. The reason is that predicted attraction vectors from CNNs are not mathematically precise enough. In this case, some potential outliers have been included in the following heuristic method, which leads to inaccurate line segment detections. Another reason is that this heuristic algorithm is not robust to imprecise estimates of the AFM. Furthermore, it requires a set of heuristics and makes the

whole process inefficient. Therefore, in this work, we propose to learn this process, i.e., recovering building masks from the learned AFM, using a network. By doing so, the whole process can be trained in an end-to-end manner, which makes it more efficient and robust.

In the AFM2Mask module, the input image and learned attraction field representation from the previous module are concatenated as the input to this module. Afterward, the network can directly generate building masks without using math heuristics (that do not work well in our case). It is noteworthy that different semantic segmentation network architectures are quite flexible to be utilized in this module, which makes full use of the power of state-of-the-art networks to generate building footprint maps.

D. End-to-End Network Learning

We propose an end-to-end training pipeline for the supervised learning of our network. More specifically, the Img2AFM module is appended before the AFM2Mask module, and the two modules are jointly trained by minimizing a global loss function. The global loss function L is defined as follows:

$$L = L_{\text{Img2AFM}} + \lambda \cdot L_{\text{AFM2Mask}} \quad (5)$$

where L_{Img2AFM} and L_{AFM2Mask} are two loss functions for optimizing the Img2AFM and AFM2Mask modules, respectively. λ is a hyperparameter to introduce a weight on the second loss and can model the relative importance of two modules.

For the first term, an L_1 loss function is utilized, and it is defined as follows:

$$L_{\text{Img2AFM}} = \sum_{x \in I} \|\hat{\mathbf{a}}(\mathbf{x}) - \mathbf{a}(\mathbf{x})\|_1 \quad (6)$$

where $\hat{\mathbf{a}}(\mathbf{x})$ is the predicted AFM and $\mathbf{a}(\mathbf{x})$ is ground reference AFM for the input image.

For the AFM2Mask module, we make use of a cross entropy loss function to guide the learning. L_{AFM2Mask} is defined as

$$L_{\text{AFM2Mask}} = \sum_{x \in I} \begin{cases} -\log(f(x)) & \text{if } y = 1 \\ -\log(1 - f(x)) & \text{if } y = 0 \end{cases} \quad (7)$$

where y is the ground truth of pixel x , $y = 1$ denotes building, and $y = 0$ represents non-building. $f(x) \in [0, 1]$ is the output probability value of x .

In the backward propagation, L_{AFM2Mask} is first backpropagated through the AFM2Mask module and then together with $\lambda \cdot L_{\text{Img2AFM}}$ propagated backward through the Img2AFM module.

IV. EXPERIMENTS

A. Dataset

We validate the proposed method on three datasets with different spatial resolutions, i.e., the ISPRS dataset, the INRIA dataset, and the Planet dataset.

1) *ISPRS Dataset*: The ISPRS dataset [44] is a benchmark dataset consisting of 38 tiles of aerial imagery over the city of Potsdam [cf. Fig. 5(a)]. Each aerial imagery includes 6000×6000 pixels at a spatial resolution of 5 cm/pixel. The provided ground reference has six land cover classes. In this work, we only use RGB bands of aerial images, and for the ground reference, the class of building is a positive class, while the other five categories are viewed as the class of non-building. Following the training/validation/test split in [45], 20 tiles (tile 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 used for training, four tiles (tile id: 7-8, 4-10, 2-11, and 5-11) are for validation, and the remaining 14 tiles are used to test models.

2) *INRIA Dataset*: The INRIA dataset [46] is composed of 360 large-scale aerial images that are collected over ten different cities. The size of each imagery is 5000×5000 , and each image consists of three bands (RGB) at a spatial resolution of 30 cm/pixel. A sample aerial image is showed in Fig. 5(b). The ground reference data of this dataset provide building masks but are only publicly available for five cities (Austin, Chicago, Kitsap County, Western Tyrol, and Vienna). In this article, data are split into training and test sets according to the setup in [46] and [47]. For each city, images with ids 1–5 are used for validation, and the remaining 31 images are for training. The statistics are derived from the validation set.

3) *Planet Dataset*: In addition to the aforementioned two public datasets, we create a Planet dataset by collecting PlanetScope satellite images and their corresponding building footprints from OpenStreetMap. The PlanetScope satellite images are gathered from eight European cities (Munich, Berlin, Amsterdam, Paris, Cologne, Milan, Rome, and Zurich)



(a)



(b)



(c)

Fig. 5. (a) Aerial imagery in the ISPRS dataset (spatial resolution: 5 cm/pixel). (b) Aerial image in the INRIA dataset (spatial resolution: 30 cm/pixel). (c) Satellite imagery in the Planet dataset (spatial resolution: 3 m/pixel).

with three bands (RGB) at 3-m spatial resolution. Compared to the former two datasets, the Planet dataset is more challenging due to its coarser spatial resolution. Fig. 5(c) shows an example of Munich. In our experiment, the image of Munich is used as the test set to evaluate the performance of models. The remaining seven cities are utilized as training and validation

sets. Specifically, for each city, 80% of samples are used for training, while 20% of samples are for validation purposes.

B. Experiment Setup

Our proposed model consists of two modules in an end-to-end framework, where the Img2AFM module utilizes a U-Net to learn the attraction field representation of an image with respect to building edges, and the AFM2Mask module can learn building masks from the representation using different semantic segmentation networks. To explore the flexibility of the AFM2Mask module, we select four state-of-the-art semantic segmentation networks: FCN-8s [20], SegNet [21], U-Net [11], and FC-DenseNet [22]. The attraction field representation encodes the geometric relation between pixels and building boundaries in an image, and it can be considered as a variant of distance transform, such as SDT [27] that measures the distance from the pixel to the boundary. Hence, we compare our model with existing works [24], [27] learning SDT representations of buildings. On the other hand, it is clearly seen that the learned AFMs from the Img2AFM module can well enhance building boundaries. In this aspect, the function of the attraction field representation seems to be similar to other visual cues, such as building boundaries and SDT masks. Thus, we also compare our network with two methods, SDT-recursive and boundary-recursive, where, basically, we incorporate SDT/edge learning into the proposed framework (cf. Fig. 3). Comparing the proposed approach and the two models can verify whether the attraction field representation is effective. Besides, the sensitivity of the hyper-parameter λ , being the coefficient of loss of the AFM2Mask module, is investigated.

C. Training Details

Our experiments are conducted within a Pytorch framework on an NVIDIA Tesla P100 GPU with 16 GB of memory. For the model training, remote sensing images and their corresponding ground reference building masks are cropped into small patches with a size of 256×256 pixels. Afterward, the boundaries, SDT, and AFMs are generated from the ground-truth building masks for further experiments as a ground reference in the training set. All models are trained for 100 epochs, and the optimizer is stochastic gradient descent (SGD) with a learning rate of 0.00001. The training batch size of all models is set as 4. The cross-entropy function is used as the loss function for other competitors.

The configurations of competitors included in experiments are listed as follows.

- 1) FCN-8s adopts a VGG16 architecture [48] as the backbone.
- 2) The encoder in SegNet is based on VGG16, and the decoder utilizes a reversed VGG16 architecture.
- 3) U-Net is composed of five blocks in both the encoder and the decoder. Each block in the encoder has two convolution layers, and in the decoder, it has one transposed convolution layer.

- 4) Both the encoder and the decoder in FC-DenseNet consist of five dense blocks, and each dense block has five convolutional layers.
- 5) For the SDT-based network that directly learns the SDT representations of buildings, they utilize the aforementioned four semantic segmentation networks and, finally, convert the learned SDT representations of buildings to semantic masks by definition [24], [27].
- 6) The SDT-recursive model or boundary-recursive model first utilizes a U-Net to learn the SDT representation or boundaries of buildings. Afterward, they also utilize the aforementioned four semantic segmentation networks to reconstruct semantic masks of building. It should be noted that the whole method is trained in an end-to-end fashion.

D. Evaluation Metrics

The performance of models is evaluated from two aspects. Mask metrics are focused on building masks, while boundary metrics are exploited to measure the quality of boundaries of the predicted building masks.

1) *Mask Metrics*: In our experiments, F1 score and intersection over union (IoU) are selected as two mask metrics. They can be computed as follows:

$$F1 \text{ score} = \frac{2 \times \text{precision} \times \text{recall}}{\text{precision} + \text{recall}} \quad (8)$$

$$IoU = \frac{TP}{TP + FP + FN} \quad (9)$$

$$\text{precision} = \frac{TP}{TP + FP} \quad (10)$$

$$\text{recall} = \frac{TP}{TP + FN} \quad (11)$$

where TP indicates the number of true positives, FN is the number of false negatives, and FP is the number of false positives. Notable that these metrics are calculated based on building pixels rather than building objects. F1 score realizes a harmonic between precision and recall.

2) *Boundary Metrics*: In order to assess building boundaries, structural similarity index (SSIM) [49] and F-measure [50] are exploited as two evaluation criteria. SSIM is a measure to calculate the similarity between two images, which can be used for the quality assessment of boundaries [51]. Before the calculation of F-measure, building boundaries are extracted first from predicted semantic masks by the Sobel edge operator [52]. F-measure is used to score the boundary and is defined as the geometric mean of the precision and recall

$$\text{precision}' = \frac{TP'}{TP' + FP'} \quad (12)$$

$$\text{recall}' = \frac{TP'}{TP' + FN'} \quad (13)$$

$$F\text{-measure} = \frac{2 \times \text{precision}' \times \text{recall}'}{\text{precision}' + \text{recall}'} \quad (14)$$

where TP' is the number of correctly identified boundary pixels, FN' is the number of boundary pixels in the ground

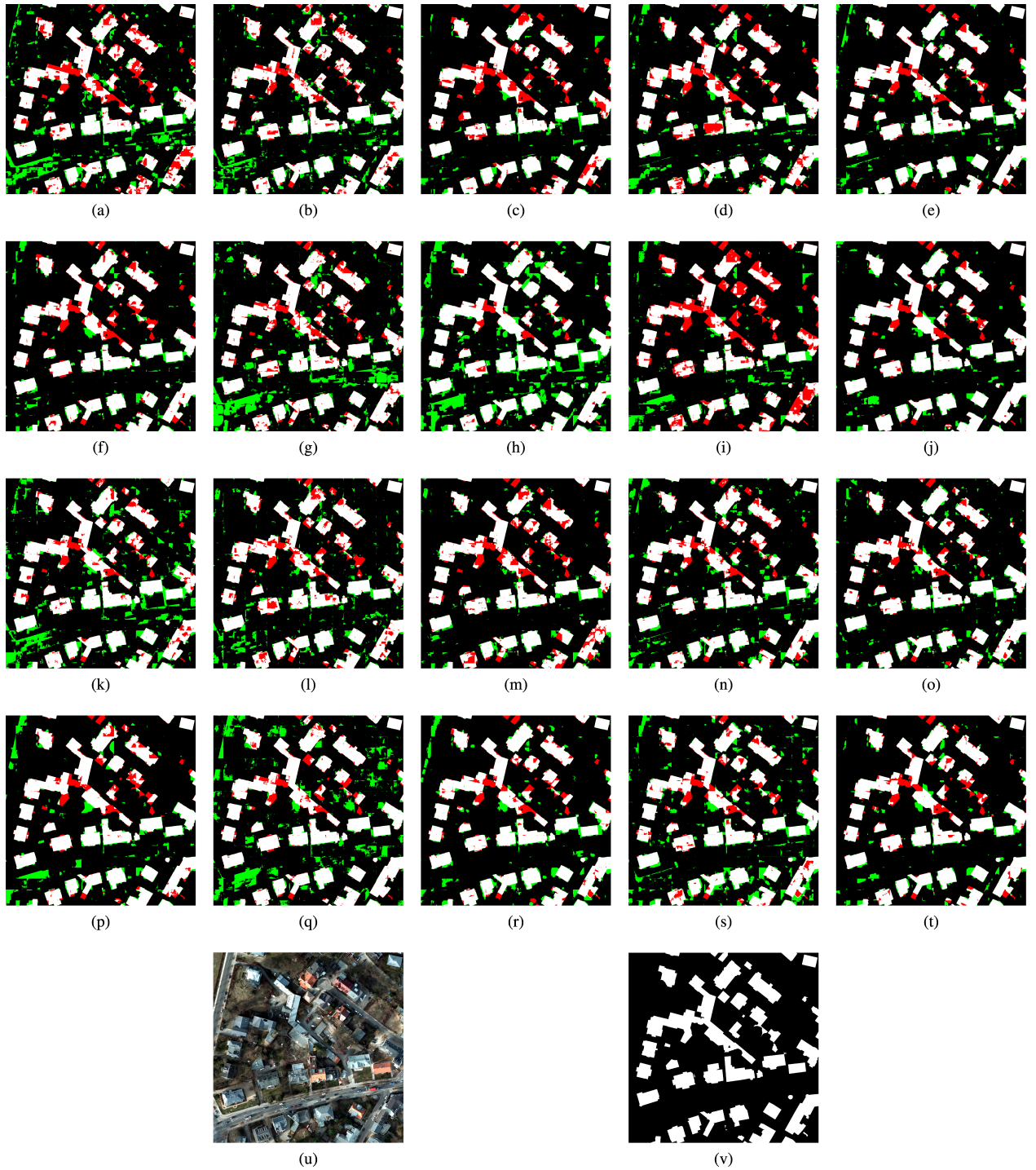


Fig. 6. Predicted results obtained from (a) FCN-8s, (b) FCN-8s-SDT, (c) FCN-8s-SDT-recursive, (d) FCN-8s-boundary-recursive, (e) proposed FCN-8s-AFM, (f) SegNet, (g) SegNet-SDT, (h) SegNet-SDT-recursive, (i) SegNet-boundary-recursive, (j) proposed SegNet-AFM, (k) U-Net, (l) U-Net-SDT, (m) U-Net-SDT-recursive, (n) U-Net-boundary-recursive, (o) proposed U-Net-AFM, (p) FC-DenseNet, (q) FC-DenseNet-SDT, (r) FC-DenseNet-SDT-recursive, (s) FC-DenseNet-boundary-recursive, and (t) proposed FC-DenseNet-AFM. Pixel-based true positives, false positives, and false negatives are marked in white, green, and red, respectively. (u) and (v) Aerial imagery and ground reference from the ISPRS dataset (spatial resolution: 5 cm/pixel).

reference but being failed to be detected, and FP' is the number of nonboundary pixels mislabeled as “boundary.”

V. RESULTS AND DISCUSSION

A. Comparison With Other Competitors

The comparisons among the proposed method, naive semantic segmentation networks, SDT-based networks, SDT-learning

methods, and boundary-learning methods are presented in this section. Their respective performance is evaluated according to both quantitative (cf. Tables I–III) and qualitative results (see Figs. 6–8) on three datasets.

Naive semantic segmentation networks that are regarded as baseline methods are first compared with the proposed framework. Specifically, we implement four baseline models,

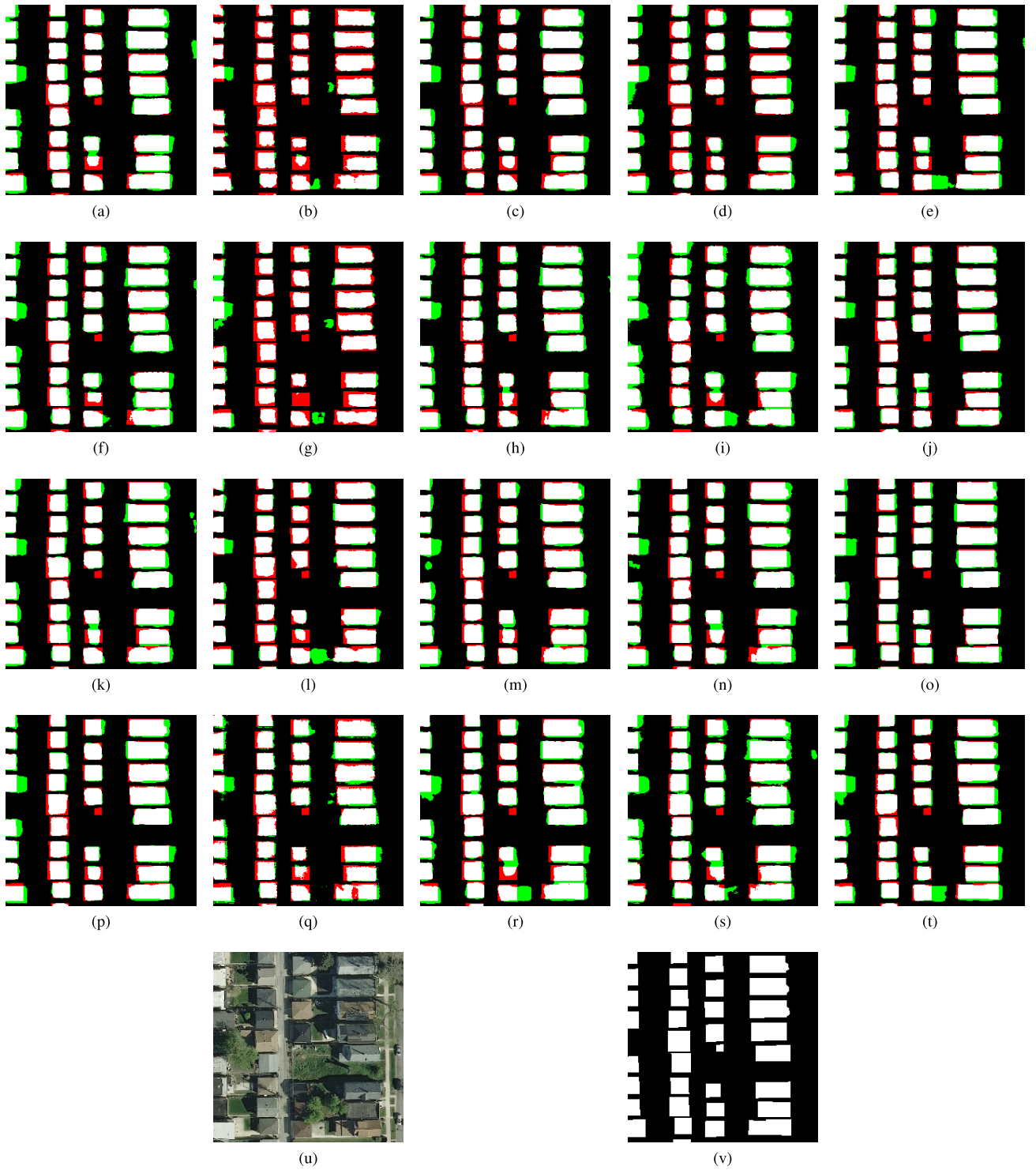


Fig. 7. Predicted results obtained from (a) FCN-8s, (b) FCN-8s-SDT, (c) FCN-8s-SDT-recursive, (d) FCN-8s-boundary-recursive, (e) proposed FCN-8s-AFM, (f) SegNet, (g) SegNet-SDT, (h) SegNet-SDT-recursive, (i) SegNet-boundary-recursive, (j) proposed SegNet-AFM, (k) U-Net, (l) U-Net-SDT, (m) U-Net-SDT-recursive, (n) U-Net-boundary-recursive, (o) proposed U-Net-AFM, (p) FC-DenseNet, (q) FC-DenseNet-SDT, (r) FC-DenseNet-SDT-recursive, (s) FC-DenseNet-boundary-recursive, and (t) proposed FC-DenseNet-AFM. Pixel-based true positives, false positives, and false negatives are marked in white, green, and red, respectively. (u) and (v) Aerial imagery and ground reference from the INRIA dataset (spatial resolution: 30 cm/pixel).

i.e., FCN-8s, SegNet, U-Net, and FC-DenseNet. For a fair comparison, the AFM2Mask module is instantiated with these four networks separately. It can be seen from the statistics of three datasets that the proposed approach significantly

boosts performance in both mask and boundary metrics compared with baseline networks. This indicates that the integration of learning attraction field representation is effective, and our framework can offer more robust results for the

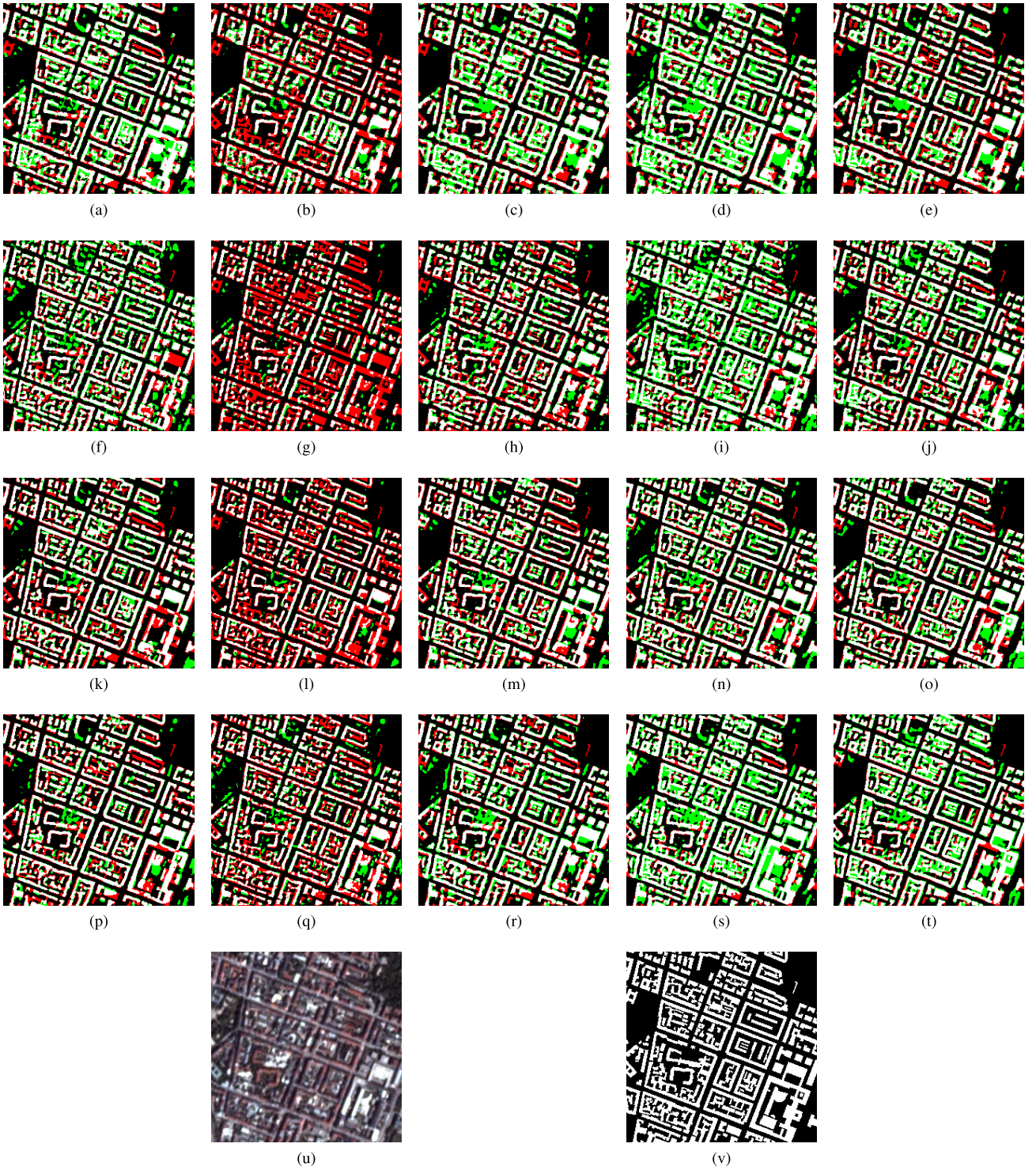


Fig. 8. Predicted results obtained from (a) FCN-8s, (b) FCN-8s-SDT, (c) FCN-8s-SDT-recursive, (d) FCN-8s-boundary-recursive, (e) proposed FCN-8s-AFM, (f) SegNet, (g) SegNet-SDT, (h) SegNet-SDT-recursive, (i) SegNet-boundary-recursive, (j) proposed SegNet-AFM, (k) U-Net, (l) U-Net-SDT, (m) U-Net-SDT-recursive, (n) U-Net-boundary-recursive, (o) proposed U-Net-AFM, (p) FC-DenseNet, (q) FC-DenseNet-SDT, (r) FC-DenseNet-SDT-recursive, (s) FC-DenseNet-boundary-recursive, and (t) proposed FC-DenseNet-AFM. Pixel-based true positives, false positives, and false negatives are marked in white, green, and red, respectively. (u) and (v) Satellite imagery and ground reference from the Planet dataset (spatial resolution: 3 m/pixel).

task of building footprint generation. For the ISPRS dataset (cf. Table I), our proposed FCN-8s-AFM obtains increments of 6.65% and 10.1% in F1 score and IoU, respectively. Moreover, the proposed U-Net-AFM reaches improvements

of 4.65% and 4.18% in SSIM and F-measure, respectively. The increases in boundary metrics suggest that our method can better preserve geometric details. The spatial resolution and image quality of the Planet dataset are much lower

TABLE I

ACCURACIES (%) OF DIFFERENT NETWORKS FOR BUILDING FOOTPRINT GENERATION IN THE ISPRS DATASET (SPATIAL RESOLUTION: 5 cm/pixel)

Method	Mask		Boundary	
	F1 score	IoU	SSIM	F-measure
FCN-8s	81.82	69.23	79.00	18.71
FCN-8s-SDT	84.38	72.97	79.52	22.94
FCN-8s-SDT-recursive	86.53	76.26	84.67	20.28
FCN-8s-boundary-recursive	84.83	73.66	82.67	15.97
proposed FCN-8s-AFM	88.47	79.33	86.15	19.39
SegNet	87.81	78.28	85.92	17.11
SegNet-SDT	83.88	72.24	80.86	20.61
SegNet-SDT-recursive	87.09	77.14	84.12	15.01
SegNet-boundary-recursive	81.79	69.20	80.36	14.38
proposed SegNet-AFM	90.56	82.75	88.76	20.34
U-Net	85.37	74.48	82.59	19.32
U-Net-SDT	84.77	73.57	82.65	19.98
U-Net-SDT-recursive	86.65	76.44	84.65	18.11
UNet-boundary-recursive	86.33	75.94	83.63	19.18
proposed U-Net-AFM	89.30	80.67	87.24	23.50
FC-DenseNet	88.34	79.11	86.24	20.76
FC-DenseNet-SDT	88.03	78.61	85.16	23.95
FC-DenseNet-SDT-recursive	87.98	78.53	85.62	18.96
FC-DenseNet-boundary-recursive	85.63	74.88	82.75	19.20
proposed FC-DenseNet-AFM	89.38	80.81	87.70	21.17

TABLE II

ACCURACIES (%) OF DIFFERENT NETWORKS FOR BUILDING FOOTPRINT GENERATION IN THE INRIA DATASET (SPATIAL RESOLUTION: 30 cm/pixel)

Method	Mask		Boundary	
	F1 score	IoU	SSIM	F-measure
FCN-8s	84.79	73.60	85.65	27.01
FCN-8s-SDT	78.07	64.03	82.12	20.58
FCN-8s-SDT-recursive	85.11	74.08	86.06	28.16
FCN-8s-boundary-recursive	84.84	73.68	85.73	26.86
proposed FCN-8s-AFM	85.92	75.31	86.47	29.92
SegNet	84.43	73.05	85.47	28.16
SegNet-SDT	76.72	62.23	82.07	20.34
SegNet-SDT-recursive	84.78	73.58	85.79	27.39
SegNet-boundary-recursive	83.08	71.06	84.63	23.40
proposed SegNet-AFM	86.18	75.72	86.88	31.38
U-Net	84.83	73.66	86.76	28.98
U-Net-SDT	83.27	71.33	85.12	28.28
U-Net-SDT-recursive	85.41	74.54	86.48	28.06
U-Net-boundary-recursive	85.06	74.01	86.18	28.79
proposed U-Net-AFM	86.68	76.49	87.07	33.77
FC-DenseNet	84.66	73.41	85.92	28.96
FC-DenseNet-SDT	77.90	63.80	81.38	27.68
FC-DenseNet-SDT-recursive	84.86	73.70	85.81	27.67
FC-DenseNet-boundary-recursive	83.69	71.95	84.61	27.52
proposed FC-DenseNet-AFM	85.41	74.53	86.20	29.72

than the other two datasets, which may lead to a negative effect on accurately extracting buildings. In this case, although improvements in both mask and boundary metrics on the Planet dataset are less significant than those on the other two datasets, the nearly 1% gain is still not trivial.

From qualitative results, we can observe that building boundaries obtained from naive semantic segmentation networks are blurred, which is also pointed out in [53]–[55]. The visual comparisons (cf. Figs. 6–8) demonstrate the effectiveness of the proposed method. As illustrated in Fig. 7, semantic masks provided by naive networks have blob-like shapes. Even with skip connections that help compensate spatial details in networks, U-Net and FC-DenseNet fail to

achieve accurate building boundaries. Moreover, this scene is a residential area, and some consecutive buildings are identified as a large building by most of the baseline models. Note that building boundaries produced by our algorithm are more rectilinear and precise. Even for buildings with complex structures (cf. Fig. 6 and 8), building footprints generated from our framework are more adherent to the ground reference. These observations suggest that our model really benefits from learning attraction field representation, enabling us to gain more geometric details of buildings.

The attraction field representation can be considered as a type of distance transform, which represents the relationship between the pixel and the boundary. Therefore, we also

TABLE III
ACCURACIES (%) OF DIFFERENT NETWORKS FOR BUILDING FOOTPRINT GENERATION IN THE PLANET DATASET (SPATIAL RESOLUTION: 3 m/pixel)

Method	Mask		Boundary	
	F1 score	IoU	SSIM	F-measure
FCN-8s	60.45	43.32	66.86	48.84
FCN-8s-SDT	47.87	31.47	64.43	33.68
FCN-8s-SDT-recursive	60.96	43.85	67.43	46.37
FCN-8s-boundary-recursive	59.80	42.66	66.47	45.26
proposed FCN-8s-AFM	61.32	44.22	68.42	49.67
SegNet	56.30	39.18	63.11	52.48
SegNet-SDT	46.11	29.96	61.47	44.46
SegNet-SDT-recursive	57.01	39.87	63.66	49.52
SegNet-boundary-recursive	56.52	39.40	61.49	50.33
proposed SegNet-AFM	60.38	43.25	66.21	53.31
U-Net	63.74	46.78	70.63	54.03
U-Net-SDT	50.99	34.22	66.32	39.39
U-Net-SDT-recursive	63.21	46.21	69.40	53.84
U-Net-boundary-recursive	64.14	47.21	68.74	54.83
proposed U-Net-AFM	65.03	48.18	69.71	56.62
FC-DenseNet	64.54	47.65	70.80	54.35
FC-DenseNet-SDT	55.62	38.52	64.63	49.90
FC-DenseNet-SDT-recursive	61.60	44.51	67.68	52.52
FC-DenseNet-boundary-recursive	64.68	47.80	68.82	55.10
proposed FC-DenseNet-AFM	65.68	48.90	70.56	56.67

take another type of distance transform: SDT as competitors. One competitor is an SDT-based network that utilizes variant backbones to learn the SDT representation of buildings and then convert this representation to semantic masks by definition [24], [27]. Compared to baseline networks, the SDT-based network can contribute to the F-measure only on the ISPRS dataset. However, there are even decreases in mask metrics. This suggests that directly learning SDT labels as final output have the potential for the improvement of geometric details only in remote sensing data with very high resolution (e.g., 5 cm/pixel). The other competitor is the SDT-recursive model, which first learns the SDT representation of buildings with a U-Net and then reconstructs semantic masks by different backbones. Notable that the whole method is trained in an end-to-end fashion. The SDT-recursive model that feeds the learned SDT representations into semantic segmentation networks is much superior to the SDT-based network, as we can see gains in both mask and boundary metrics. This may be because the SDT representation learned from the remote sensing imagery carries useful information to capture the global semantic context in semantic segmentation networks, which indicates the potential of SDT in a recursive learning way for building footprint generation. It is worthy to note that the performances of both SDT-based network and SDT-recursive model are more sensitive to the backbone semantic segmentation networks. For the ISPRS dataset (see Table III), when the backbone is FCN-8s, both SDT-based network and SDT-recursive model can boost the performance. However, the performance of SegNet-SDT and SegNet-SDT-recursive is worse than that of SegNet.

The geometric property of building boundaries can be significantly enhanced by AFMs (see Fig. 2). From Tables I–III, it can be observed that our framework can improve results in terms of both mask and boundary metrics, which confirms that

explicitly encoding geometric information is essential to building footprint generation tasks. In this regard, we investigate another competitor, the boundary-recursive model, to further validate the effectiveness of the attraction field representation. This method first learns building boundaries from remote sensing images with a U-Net and then uses them as auxiliary information to extract building masks by variant semantic segmentation networks. Notable that these two subnetworks are jointly optimized. Experimental results show that this model does not bring this task any benefits in terms of building boundary quality, and we can see decreases in boundary metrics and more blurred boundaries compared to the naive semantic segmentation network. This may be because building boundaries are characterized with very few pixels, and this class imbalance leads to ambiguity in the network learning.

By contrast, our method can always provide significant gains, regardless of which semantic segmentation network architecture is chosen as the AFM2Mask module, and the proposed approach outperforms other competitors in most of the statistical metrics for three datasets. This is due to two facts. One is that the attraction field representation can encode geometric properties in 2-D (x - and y -directions), but SDT only relies on the Euclidean distance and, thus, characterizes the information in 1-D. This indicates that the use of the information in different dimensions is more reliable and accurate. Fig. 9(a) and (b) presents the AFM learned by the proposed U-Net-AFM, and Fig. 9(c) shows the SDT representation learned by U-Net-SDT-recursive. It can be observed that attraction field representation can better delineate sharp building boundaries. The other factor is that the attraction field representation takes the nonboundary pixels into account, which have addressed the challenges of class imbalance in boundary-learning methods.

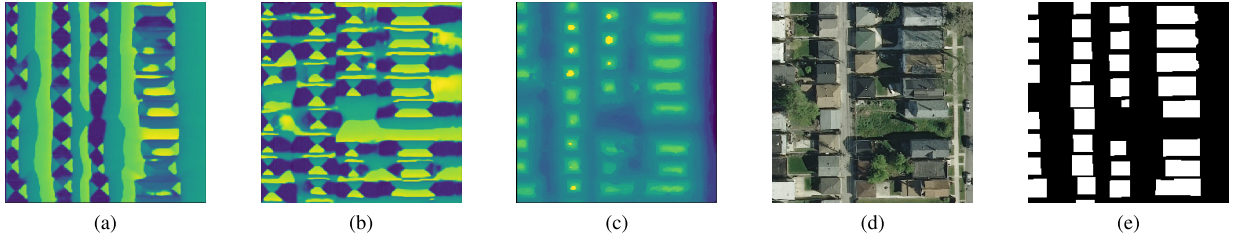


Fig. 9. (a) AFM (x-axis) and (b) AFM (y-axis) are learned by the proposed method (U-Net-AFM). (c) SDT representation learned by the U-Net-SDT-recursive. (d) and (e) Aerial imagery and ground reference from the INRIA dataset (spatial resolution: 0.3 m/pixel).

TABLE IV
ACCURACIES (%) OF DIFFERENT COEFFICIENTS OF AFM2MASK LOSS
(λ) FOR BUILDING FOOTPRINT GENERATION IN THE INRIA
DATASET (SPATIAL RESOLUTION: 30 cm/pixel)

λ	Mask		Boundary	
	F1 score	IoU	SSIM	F-measure
10	86.04	75.50	86.77	31.80
1	86.37	76.01	87.06	33.22
0.1	86.68	76.49	87.07	33.77

B. Analysis of Hyperparameter Tuning

As shown in the results on three datasets, taking U-Net as the AFM2Mask module can deliver relatively satisfactory results on all three datasets. Therefore, in this section, we use U-Net-AFM for further studies. Moreover, the INRIA dataset is selected as an example dataset to carry out the following experiments.

In the proposed framework, the global loss function is utilized to guide the end-to-end learning of building masks from remote sensing data. This function is a sum of losses from two separate modules, where the hyperparameter λ is the coefficient of the AFM2Mask module. Here, λ is set as three different numbers, i.e., 0.1, 1, and 10, to investigate its impact on final results.

The statistical results with different values of λ are shown in Table IV. We can see that our model is insensitive to this parameter, and networks with all different λ values outperform the naive U-Net. Furthermore, increasing the value of λ will lead to a slight reduction in both mask and boundary metrics. The best result is obtained when $\lambda = 0.1$. A small value of λ indicates more significance of the Img2AFM module than the AFM2Mask module, which places an emphasis on the attraction field representation learning in the whole framework. It can be clearly seen from the Fig. 10 that gradually lowering λ can reduce false detections. This is mainly because the attraction field representation can alleviate the impact of background clutters.

C. Different Methods to Incorporate Attraction Field Representation

It is worth noting that building boundaries leaned by the proposed method are significantly improved due to the exploitation of attraction field representation. In order to further explore how to well leverage attraction field representation,

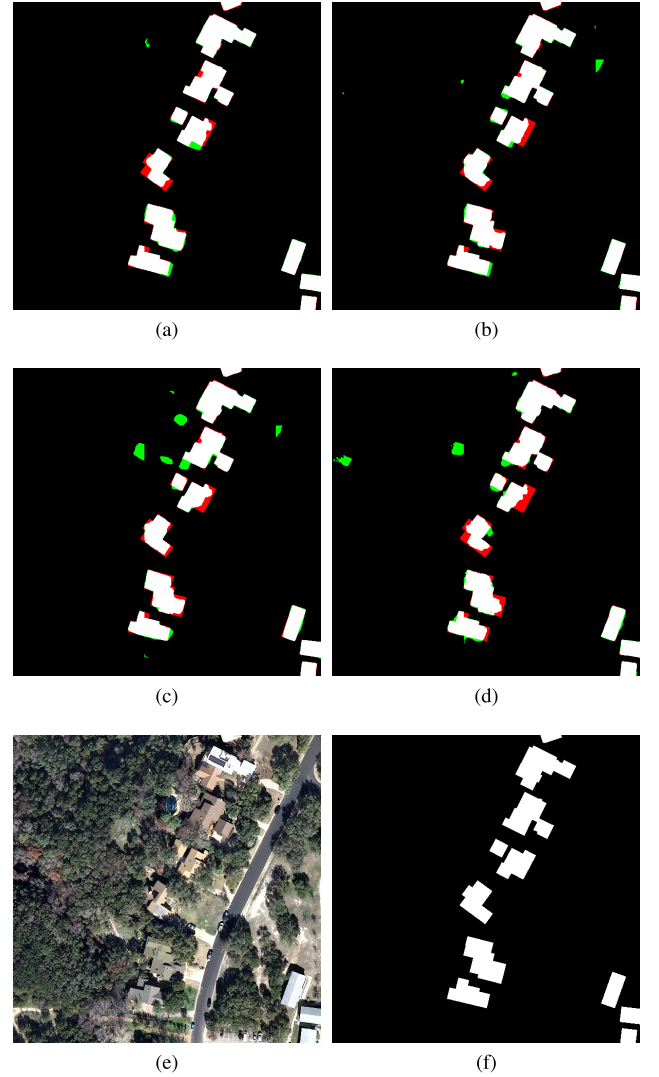


Fig. 10. Results obtained by the proposed method (U-Net-AFM) with coefficient $\lambda =$ (a) 0.1, (b) 1, and (c) 10. (d) Result obtained by the naive U-Net. Pixel-based true positives, false positives, and false negatives are marked in white, green, and red, respectively. (e) and (f) Corresponding aerial imagery and ground reference from the INRIA dataset (spatial resolution: 30 cm/pixel).

we investigate another three designs to incorporate this useful representation in network learning.

- 1) **Srivastava *et al.* [56]:** It uses a U-Net architecture followed by two separate fully connected layers to

TABLE V
ACCURACIES (%) OF DIFFERENT DESIGNS FOR THE INCORPORATION OF
ATTRACTION FIELD REPRESENTATION IN THE INRIA DATASET
(SPATIAL RESOLUTION: 30 cm/pixel)

Method	Mask		Boundary	
	F1 score	IoU	SSIM	F-measure
U-Net	84.83	73.66	86.76	28.98
proposed U-Net-AFM	86.68	76.49	87.07	33.77
Srivastava <i>et al.</i> [56]	85.97	75.39	86.54	29.83
Bischke <i>et al.</i> [47]	86.10	75.59	86.67	29.80
Mou & Zhu [57]	85.48	74.63	86.29	30.06

TABLE VI
ACCURACIES (%) OF DIFFERENT METHODS FOR BUILDING
FOOTPRINT GENERATION IN THE ISPRS DATASET
(SPATIAL RESOLUTION: 5 cm/pixel)

Method	Mask		Boundary	
	F1 score	IoU	SSIM	F-measure
proposed SegNet-AFM	90.56	82.75	88.76	20.34
Griffiths & Boehm [58]	85.00	-	-	-
Lin <i>et al.</i> [59]	88.47	79.94	85.73	18.41
Wei <i>et al.</i> [60]	88.65	80.23	86.40	19.67

learn semantic masks and attraction field representation, respectively.

- 2) **Bischke *et al.* [47]:** It takes a U-Net as the backbone and first adds one convolutional layer after the decoder to learn the attraction field representation. Afterward, this learned attraction field representation and feature maps produced by the decoder are concatenated and fed into another convolutional layer to learn final segmentation masks.
- 3) **Mou and Zhu [57]:** It utilizes an encoder and two separate decoders to jointly optimize two complementary tasks, namely, building semantic segmentation and attraction field representation learning. Note that the architecture of encoder and decoders in this design is the same as those in U-Net.

The statistical and visual results are reported in Table V and Fig. 11, respectively. From both mask and boundary metrics in Table V, all methods have shown superior results than naive U-Net, which again confirms the significance of attraction field representation in our task. Among all design options, the proposed framework has achieved the best performance. In particular, the F-measure achieved by our approach is increased by more than 3% when compared to the other methods. Besides, it can be seen that the building boundaries and corners learned by the proposed framework are more accurate than its competitors. This suggests that our approach is able to effectively leverage information of attraction field representation, which is attributed to our recursive learning strategy.

D. Comparison With State-of-the-Art Methods

To verify the superiority of our approach on datasets with different spatial resolutions, we make a comparison with other

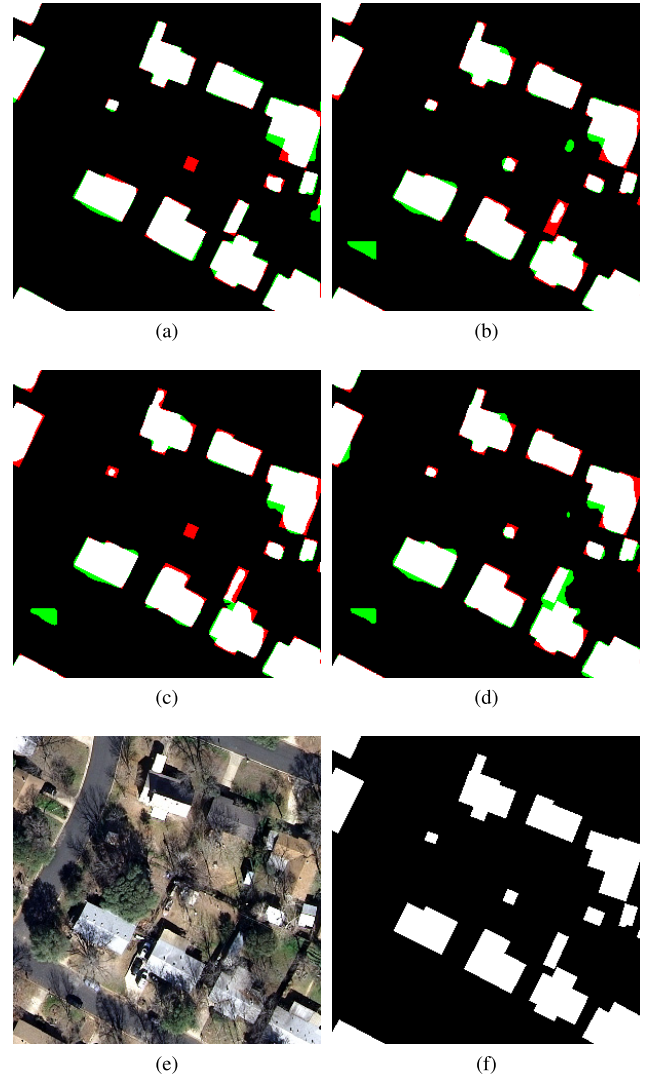


Fig. 11. Results obtained by (a) proposed U-Net-AFM, (b) Srivastava *et al.* [56], (c) Bischke *et al.* [47], and (d) Mou and Zhu [57]. Pixel-based true positives, false positives, and false negatives are marked in white, green, and red, respectively. (e) and (f) Corresponding aerial imagery and ground reference from the INRIA dataset (spatial resolution: 30 cm/pixel).

state-of-the-art methods on the ISPRS, INRIA, and Planet datasets. The statistical results of different algorithms on three datasets are shown in Tables VI–VIII, respectively. On both ISPRS and Planet datasets, the proposed method surpasses all other models in both mask and boundary metrics. For the INRIA dataset, our approach achieves the highest scores in boundary metrics and comparative performance in mask prediction. Compared to our methods, Girard's method [35] gains a marginal improvement in mask metrics at the cost of additional ground-truth annotations (i.e., vector format of building footprints). For an intuitive comparison, the visual results of our method and Girard's method [35] are illustrated in Fig. 12. As we can see, Girard's method [35] fails to recover detailed structures of complicated buildings. On the contrary, our approach can accurately capture more geometric details,

TABLE VII
ACCURACIES (%) OF DIFFERENT METHODS FOR BUILDING FOOTPRINT
GENERATION IN THE INRIA DATASET (SPATIAL
RESOLUTION: 30 cm/pixel)

Method	Mask		Boundary	
	F1 score	IoU	SSIM	F-measure
proposed U-Net-AFM	86.68	76.49	87.07	33.77
Ji <i>et al.</i> [25]	-	71.40	-	-
Liu <i>et al.</i> [61]	-	71.76	-	-
Bischke <i>et al.</i> [47]	-	73.00	-	-
Audebert <i>et al.</i> [62]	-	74.17	-	-
Girard <i>et al.</i> [35]	86.82	76.71	86.49	32.00

TABLE VIII
ACCURACIES (%) OF DIFFERENT METHODS FOR BUILDING FOOTPRINT
GENERATION IN THE PLANET DATASET (SPATIAL
RESOLUTION: 3 m/pixel)

Method	Mask		Boundary	
	F1 score	IoU	SSIM	F-measure
proposed FC-DenseNet-AFM	65.68	48.90	70.56	56.67
Lin <i>et al.</i> [59]	59.54	42.39	64.53	53.03
Wei <i>et al.</i> [60]	64.85	47.98	69.06	54.85

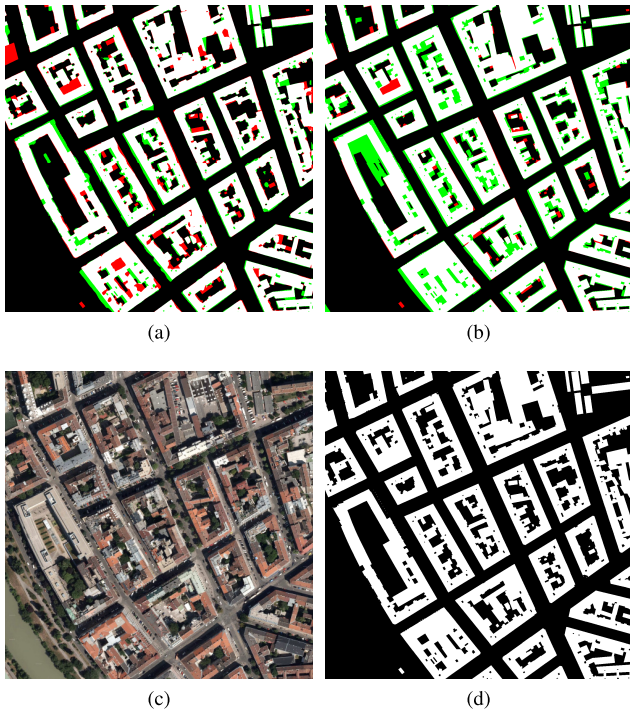


Fig. 12. Results obtained by (a) proposed U-Net-AFM and (b) Girard *et al.* [35]. Pixel-based true positives, false positives, and false negatives are marked in white, green, and red, respectively. (c) and (d) Corresponding aerial imagery and ground reference from the INRIA dataset (spatial resolution: 30 cm/pixel).

which again demonstrates the strength of the AFM for the task of building footprint generation.

VI. CONCLUSION

Considering that building boundaries are easily blurred when using semantic segmentation networks to directly

learn building footprints, a new end-to-end building footprint generation method through learning the attraction field representation is proposed in this article. The proposed model comprises two modules: an Img2AFM module and an AFM2Mask module. More specifically, the former is designed to learn the attraction field representation, which enables not only the enhancement of building boundaries but also the suppression of background clutters. Afterward, the latter exploits the input remote sensing image and learned AFM to reconstruct building masks. The performance of the proposed end-to-end network is assessed on three datasets with different spatial resolutions: the ISPRS dataset (5 cm/pixel), the INRIA dataset (30 cm/pixel), and the Planet dataset (3 m/pixel). Experimental results suggest that the incorporation of the attraction field representation in our framework can offer more satisfactory building footprint maps. On the one hand, sharp boundaries and geometric details of buildings can be better preserved. On the other hand, non-building objects that are wrongly detected as buildings can be avoided to a large extent. Thus, we believe that our method has the potential to be a robust solution for building footprint generation at a large scale. Looking into the future, we intend to investigate the potential of the attraction field representation in other tasks, e.g., road extraction and vehicle detection.

REFERENCES

- [1] X. Huang, W. Yuan, J. Li, and L. Zhang, "A new building extraction postprocessing framework for high-spatial-resolution remote-sensing imagery," *IEEE J. Sel. Topics Appl. Earth Observ. Remote Sens.*, vol. 10, no. 2, pp. 654–668, Feb. 2017.
- [2] X. X. Zhu *et al.*, "Deep learning in remote sensing: A comprehensive review and list of resources," *IEEE Geosci. Remote Sens. Mag.*, vol. 5, no. 4, pp. 8–36, Dec. 2017.
- [3] Y. Hua, L. Mou, and X. X. Zhu, "Recurrently exploring class-wise attention in a hybrid convolutional and bidirectional LSTM network for multi-label aerial image classification," *ISPRS J. Photogramm. Remote Sens.*, vol. 149, pp. 188–199, Mar. 2019.
- [4] C. Qiu, L. Mou, M. Schmitt, and X. X. Zhu, "Local climate zone-based urban land cover classification from multi-seasonal sentinel-2 images with a recurrent residual network," *ISPRS J. Photogramm. Remote Sens.*, vol. 154, pp. 151–162, Aug. 2019. [Online]. Available: <https://www.sciencedirect.com/science/article/pii/S0924271619301261>
- [5] C. Qiu, M. Schmitt, C. Geiß, T.-H.-K. Chen, and X. X. Zhu, "A framework for large-scale mapping of human settlement extent from sentinel-2 images via fully convolutional neural networks," *ISPRS J. Photogramm. Remote Sens.*, vol. 163, pp. 152–170, May 2020.
- [6] Y. Hua, L. Mou, and X. X. Zhu, "Relation network for multilabel aerial image classification," *IEEE Trans. Geosci. Remote Sens.*, vol. 58, no. 7, pp. 4558–4572, Jul. 2020.
- [7] J. Kang, R. Fernandez-Beltran, Z. Ye, X. Tong, P. Ghamisi, and A. Plaza, "Deep metric learning based on scalable neighborhood components for remote sensing scene characterization," *IEEE Trans. Geosci. Remote Sens.*, vol. 58, no. 12, pp. 8905–8918, Dec. 2020.
- [8] J. Kang, R. Fernandez-Beltran, P. Duan, X. Kang, and A. J. Plaza, "Robust normalized softmax loss for deep metric learning-based characterization of remote sensing images with label noise," *IEEE Trans. Geosci. Remote Sens.*, early access, Dec. 16, 2020, doi: [10.1109/TGRS.2020.3042607](https://doi.org/10.1109/TGRS.2020.3042607).
- [9] P. Wang, X. Sun, W. Diao, and K. Fu, "FMSSD: Feature-merged single-shot detection for multiscale objects in large-scale remote sensing imagery," *IEEE Trans. Geosci. Remote Sens.*, vol. 58, no. 5, pp. 3377–3390, Dec. 2020.
- [10] X. Sun, Y. Liu, Z. Yan, P. Wang, W. Diao, and K. Fu, "SRAF-net: Shape robust anchor-free network for garbage dumps in remote sensing imagery," *IEEE Trans. Geosci. Remote Sens.*, vol. 59, no. 7, pp. 6154–6168, Jul. 2021.
- [11] O. Ronneberger, P. Fischer, and T. Brox, "U-Net: Convolutional networks for biomedical image segmentation," in *Proc. Int. Conf. Med. Image Comput. Comput.-Assist. Intervent.* Cham, Switzerland: Springer, 2015, pp. 234–241.

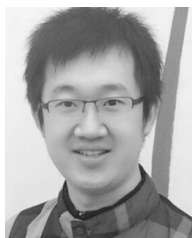
- [12] M. Cote and P. Saeedi, "Automatic rooftop extraction in nadir aerial imagery of suburban regions using corners and variational level set evolution," *IEEE Trans. Geosci. Remote Sens.*, vol. 51, no. 1, pp. 313–328, Jan. 2013.
- [13] Q. Li *et al.*, "Instance segmentation of buildings using keypoints," in *Proc. IEEE Int. Geosci. Remote Sens. Symp.*, Sep./Oct. 2020, pp. 1452–1455.
- [14] S. Cui, Q. Yan, and P. Reinartz, "Complex building description and extraction based on Hough transformation and cycle detection," *Remote Sens. Lett.*, vol. 3, no. 2, pp. 151–159, Mar. 2012.
- [15] L. Zhang *et al.*, "Learning deep structured active contours end-to-end," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2018, pp. 8877–8885.
- [16] N. Xue, S. Bai, F. Wang, G.-S. Xia, T. Wu, and L. Zhang, "Learning attraction field representation for robust line segment detection," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2019, pp. 1595–1603.
- [17] A. O. Ok, "Automated detection of buildings from single VHR multispectral images using shadow information and graph cuts," *ISPRS J. Photogramm. Remote Sens.*, vol. 86, pp. 21–40, Dec. 2013.
- [18] M. Turker and D. Koc-San, "Building extraction from high-resolution optical spaceborne images using the integration of support vector machine (SVM) classification, Hough transformation and perceptual grouping," *Int. J. Appl. Earth Observ. Geoinf.*, vol. 34, pp. 58–69, Feb. 2015.
- [19] X. Huang and L. Zhang, "A multidirectional and multiscale morphological index for automatic building extraction from multispectral geospatial imagery," *Photogramm. Eng. Remote Sens.*, vol. 77, no. 7, pp. 721–732, 2011.
- [20] J. Long, E. Shelhamer, and T. Darrell, "Fully convolutional networks for semantic segmentation," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2015, pp. 3431–3440.
- [21] V. Badrinarayanan, A. Kendall, and R. Cipolla, "SegNet: A deep convolutional encoder-decoder architecture for image segmentation," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 39, no. 12, pp. 2481–2495, Dec. 2017.
- [22] S. Jegou, M. Drozdal, D. Vazquez, A. Romero, and Y. Bengio, "The one hundred layers tiramisu: Fully convolutional DenseNets for semantic segmentation," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. Workshops (CVPRW)*, Jul. 2017, pp. 11–19.
- [23] E. Maggiori, Y. Tarabalka, G. Charpiat, and P. Alliez, "Convolutional neural networks for large-scale remote-sensing image classification," *IEEE Trans. Geosci. Remote Sens.*, vol. 55, no. 2, pp. 645–657, Feb. 2017.
- [24] H. L. Yang, J. Yuan, D. Lunga, M. Laverdiere, A. Rose, and B. Bhaduri, "Building extraction at scale using convolutional neural network: Mapping of the United States," *IEEE J. Sel. Topics Appl. Earth Observ. Remote Sens.*, vol. 11, no. 8, pp. 2600–2614, Aug. 2018.
- [25] S. Ji, S. Wei, and M. Lu, "Fully convolutional networks for multisource building extraction from an open aerial and satellite imagery data set," *IEEE Trans. Geosci. Remote Sens.*, vol. 57, no. 1, pp. 574–586, Jan. 2019.
- [26] X. Li, X. Yao, and Y. Fang, "Building-a-nets: Robust building extraction from high-resolution remote sensing images with adversarial networks," *IEEE J. Sel. Topics Appl. Earth Observ. Remote Sens.*, vol. 11, no. 10, pp. 3680–3687, Oct. 2018.
- [27] J. Yuan, "Learning building extraction in aerial scenes with convolutional networks," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 40, no. 11, pp. 2793–2798, Nov. 2018.
- [28] K. G. Derpanis, *The Harris Corner Detector*. Toronto, ON, Canada: York Univ., 2004, pp. 1–2.
- [29] C. Harris and S. Mike, "A combined corner and edge detector," in *Proc. Alvey Vis. Conf.*, Manchester, U.K., vol. 15, no. 50, pp. 10–5244, Sep. 1988.
- [30] D. G. Lowe, "Distinctive image features from scale-invariant keypoints," *Int. J. Comput. Vis.*, vol. 60, no. 2, pp. 91–110, 2004.
- [31] M. Zangrandi, E. Baccaglini, and L. Boulard, "An enhanced corner-based automatic rooftop extraction algorithm leveraging drlse segmentation," in *Proc. IEEE Int. Geosci. Remote Sens. Symp. (IGARSS)*, Jul. 2015, pp. 1024–1027.
- [32] M. Wang, S. Yuan, and J. Pan, "Building detection in high resolution satellite urban image using segmentation, corner detection combined with adaptive windowed Hough transform," in *Proc. IEEE Int. Geosci. Remote Sens. Symp. - IGARSS*, Jul. 2013, pp. 508–511.
- [33] Z. Li, J. D. Wegner, and A. Lucchi, "Topological map extraction from overhead images," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Oct. 2019, pp. 1715–1724.
- [34] W. Zhao, C. Persello, and A. Stein, "Building outline delineation: From aerial images to polygons with an improved end-to-end learning framework," *ISPRS J. Photogramm. Remote Sens.*, vol. 175, pp. 119–131, May 2021.
- [35] N. Girard, D. Smirnov, J. Solomon, and Y. Tarabalka, "Polygonal building segmentation by frame field learning," 2020, *arXiv:2004.14875*. [Online]. Available: <http://arxiv.org/abs/2004.14875>
- [36] R. O. Duda and R. E. Hart, "Use of the Hough transformation to detect lines and curves in pictures," *Commun. ACM*, vol. 15, no. 1, pp. 11–15, Jan. 1972.
- [37] J. B. Burns, A. R. Hanson, and E. M. Riseman, "Extracting straight lines," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. PAMI-8, no. 4, pp. 425–455, Jul. 1986.
- [38] M. Izadi and P. Saeedi, "Three-dimensional polygonal building model estimation from single satellite images," *IEEE Trans. Geosci. Remote Sens.*, vol. 50, no. 6, pp. 2254–2272, Jun. 2012.
- [39] C. Akinlar and C. Topal, "EDLines: A real-time line segment detector with a false detection control," *Pattern Recognit. Lett.*, vol. 32, no. 13, pp. 1633–1642, Oct. 2011.
- [40] J. Wang, X. Yang, X. Qin, X. Ye, and Q. Qin, "An efficient approach for automatic rectangular building extraction from very high resolution optical satellite imagery," *IEEE Geosci. Remote Sens. Lett.*, vol. 12, no. 3, pp. 487–491, Mar. 2015.
- [41] X. Qin, S. He, X. Yang, M. Dehghan, Q. Qin, and J. Martin, "Accurate outline extraction of individual building from very high-resolution optical images," *IEEE Geosci. Remote Sens. Lett.*, vol. 15, no. 11, pp. 1775–1779, Nov. 2018.
- [42] D. Cheng, R. Liao, S. Fidler, and R. Urtasun, "DARNet: Deep active ray network for building segmentation," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2019, pp. 7431–7439.
- [43] M. Kass, A. Witkin, and D. Terzopoulos, "Snakes: Active contour models," *Int. J. Comput. Vis.*, vol. 1, no. 4, pp. 321–331, 1988.
- [44] *ISPRS*. Accessed: Dec. 15, 2018. [Online]. Available: <http://www2.isprs.org/commissions/comm3/wg4/semantic-labeling.html>
- [45] D. Marmanis, K. Schindler, J. D. Wegner, S. Galliani, M. Datcu, and U. Stilla, "Classification with an edge: Improving semantic image segmentation with boundary detection," *ISPRS J. Photogramm. Remote Sens.*, vol. 135, pp. 158–172, Jan. 2018.
- [46] E. Maggiori, Y. Tarabalka, G. Charpiat, and P. Alliez, "Can semantic labeling methods generalize to any city? The inria aerial image labeling benchmark," in *Proc. IEEE Int. Geosci. Remote Sens. Symp. (IGARSS)*, Jul. 2017, pp. 3226–3229.
- [47] B. Bischke, P. Helber, J. Folz, D. Borth, and A. Dengel, "Multi-task learning for segmentation of building footprints with deep neural networks," in *Proc. IEEE Int. Conf. Image Process. (ICIP)*, Sep. 2019, pp. 1480–1484.
- [48] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," 2014, *arXiv:1409.1556*. [Online]. Available: <http://arxiv.org/abs/1409.1556>
- [49] Z. Wang, A. C. Bovik, H. R. Sheikh, and E. P. Simoncelli, "Image quality assessment: From error visibility to structural similarity," *IEEE Trans. Image Process.*, vol. 13, no. 4, pp. 600–612, Apr. 2004.
- [50] I. Kokkinos, "Boundary detection using F-measure-, filter- and feature-(F3) boost," in *Proc. Eur. Conf. Comput. Vis.* Berlin, Germany: Springer, 2010, pp. 650–663.
- [51] D. Sadykova and A. P. James, "Quality assessment metrics for edge detection and edge-aware filtering: A tutorial review," in *Proc. Int. Conf. Adv. Comput., Commun. Informat. (ICACCI)*, Sep. 2017, pp. 2366–2369.
- [52] I. Sobel, *An Isotropic 3×3 Gradient Operator, Machine Vision for Three-Dimensional Scenes*. New York, NY, USA: Academic, 1990, pp. 376–379.
- [53] Y. Shi, Q. Li, and X. X. Zhu, "Building segmentation through a gated graph convolutional neural network with deep structured feature embedding," *ISPRS J. Photogramm. Remote Sens.*, vol. 159, pp. 184–197, Jan. 2020.
- [54] 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 Trans. Geosci. Remote Sens.*, vol. 58, no. 11, pp. 7502–7519, Nov. 2020.
- [55] 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 J. Sel. Topics Appl. Earth Observ. Remote Sens.*, vol. 11, no. 8, pp. 2615–2629, Aug. 2018.

- [56] S. Srivastava, M. Volpi, and D. Tuia, "Joint height estimation and semantic labeling of monocular aerial images with CNNs," in *Proc. IEEE Int. Geosci. Remote Sens. Symp. (IGARSS)*, Jul. 2017, pp. 5173–5176.
- [57] L. Mou and X. X. Zhu, "Vehicle instance segmentation from aerial image and video using a multitask learning residual fully convolutional network," *IEEE Trans. Geosci. Remote Sens.*, vol. 56, no. 11, pp. 6699–6711, Nov. 2018.
- [58] D. Griffiths and J. Boehm, "Improving public data for building segmentation from convolutional neural networks (CNNs) for fused airborne lidar and image data using active contours," *ISPRS J. Photogramm. Remote Sens.*, vol. 154, pp. 70–83, Aug. 2019.
- [59] J. Lin, W. Jing, H. Song, and G. Chen, "ESFNet: Efficient network for building extraction from high-resolution aerial images," *IEEE Access*, vol. 7, pp. 54285–54294, 2019.
- [60] S. Wei, S. Ji, and M. Lu, "Toward automatic building footprint delineation from aerial images using CNN and regularization," *IEEE Trans. Geosci. Remote Sens.*, vol. 58, no. 3, pp. 2178–2189, Mar. 2020.
- [61] P. Liu, X. Liu, M. Liu, Q. Shi, J. Yang, X. Xu, and Y. Zhang, "Building footprint extraction from high-resolution images via spatial residual inception convolutional neural network," *Remote Sens.*, vol. 11, no. 7, p. 830, 2019.
- [62] N. Audebert, A. Boulch, B. Le Saux, and S. Lefèvre, "Distance transform regression for spatially-aware deep semantic segmentation," *Comput. Vis. Image Understand.*, vol. 189, Dec. 2019, Art. no. 102809.



Qingyu Li (Student Member, IEEE) received the bachelor's degree in remote sensing science and technology from Wuhan University, Wuhan, China, in 2015, and the master's degree in earth oriented space science and technology (ESPACE) from the Technische Universität München (TUM), Munich, Germany, in 2018. She is currently pursuing the Ph.D. degree with TUM and the German Aerospace Center (DLR), Weßling, Germany.

Her research interests include deep learning, remote sensing mapping, and remote sensing applications.



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

In 2015, he spent six months at the Computer Vision Group, University of Freiburg, Freiburg im Breisgau, Germany. In 2019, he was a Visiting

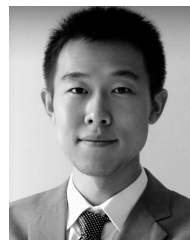
Researcher with the Cambridge Image Analysis Group (CIA), University of Cambridge, Cambridge, U.K. He is currently a Guest Professor with the Munich AI Future Lab AI4EO, TUM, and the Head of the Visual Learning and Reasoning Team, Department "EO Data Science," Remote Sensing Technology Institute (IMF), German Aerospace Center (DLR), Weßling, Germany. Since 2019, he has been a Research Scientist with DLR-IMF and an AI Consultant for the Helmholtz Artificial Intelligence Cooperation Unit (HAICU).

Dr. Mou was a recipient of the First Place in the 2016 IEEE GRSS Data Fusion Contest and the finalist for the Best Student Paper Award at the 2017 Joint Urban Remote Sensing Event and the 2019 Joint Urban Remote Sensing Event.



Yuansheng Hua (Graduate Student Member, IEEE) received the bachelor's degree in remote sensing science and technology from Wuhan University, Wuhan, China, in 2014, and the double master's degrees in earth oriented space science and technology (ESPACE) and photogrammetry and remote sensing from the Technical University of Munich (TUM), Munich, Germany, and Wuhan University in 2018 and 2019, respectively. He is currently pursuing the Ph.D. degree with the German Aerospace Center (DLR), Weßling, Germany, and TUM.

In 2019, he was a Visiting Researcher with Wageningen University & Research, Wageningen, The Netherlands. His research interests include remote sensing, computer vision, and deep learning, especially their applications in remote sensing.



Yilei Shi (Member, IEEE) received the Dipl.Ing degree in mechanical engineering and the Dr.Ing degree in signal processing from the Technische Universität München (TUM), Munich, Germany, in 2010 and 2019, respectively.

In April and May 2019, he was a Guest Scientist with the Department of Applied Mathematics and Theoretical Physics, University of Cambridge, Cambridge, U.K. He is currently a Senior Scientist with the Chair of Remote Sensing Technology, TUM. His research interests include fast solver and

parallel computing for large-scale problems, high-performance computing and computational intelligence, advanced methods on synthetic-aperture radar (SAR) and interferometric SAR (InSAR) processing, machine learning and deep learning for a variety of data sources, such as SAR, optical images, and medical images, and partial differential equation (PDE)-related numerical modeling and computing.



Xiao Xiang Zhu (Fellow, IEEE) received the M.Sc., Dr.Ing., and Habilitation degrees in signal processing from the Technical University of Munich (TUM), Munich, Germany, in 2008, 2011, and 2013, respectively.

She was a Guest Scientist or a Visiting Professor with the Italian National Research Council (CNR-IREA), Naples, Italy, Fudan University, Shanghai, China, The University of Tokyo, Tokyo, Japan, and the University of California at Los Angeles, Los Angeles, CA, USA, in 2009, 2014,

2015, and 2016, respectively. Since 2019, she has been a Co-Coordinator of the Munich Data Science Research School, TUM. Since 2019, she has been the Head of the Helmholtz Artificial Intelligence—Research Field "Aeronautics, Space and Transport." Since May 2020, she has been the Director of the International Future AI Lab "AI4EO—Artificial Intelligence for Earth Observation: Reasoning, Uncertainties, Ethics and Beyond," Munich. Since October 2020, she has been the Co-Director of the Munich Data Science Institute (MDSI), TUM. She is currently a Professor of data science in earth observation (former: signal processing in earth observation) with TUM and the Head of the Department "EO Data Science," Remote Sensing Technology Institute, German Aerospace Center (DLR), Weßling, Germany. She is currently also an AI Professor with ESA's Phi-Lab, Frascati, Italy. Her main research interests are remote sensing and earth observation, signal processing, machine learning, and data science, with a special application focus on global urban mapping.

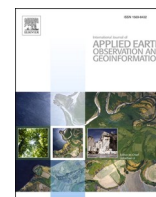
Dr. Zhu is also a member of the young academy (Junge Akademie/Junges Kolleg) at the Berlin-Brandenburg Academy of Sciences and Humanities and the German National Academy of Sciences Leopoldina and the Bavarian Academy of Sciences and Humanities. She is also an Associate Editor of IEEE TRANSACTIONS ON GEOSCIENCE AND REMOTE SENSING and serves as an Area Editor responsible for special issues of *IEEE Signal Processing Magazine*.

B Li, Qingyu, Lichao Mou, Yuansheng Hua, Yilei Shi, and Xiao Xiang Zhu. CrossGeoNet: A Framework for Building Footprint Generation of Label-Scarce Geographical Regions. International Journal of Applied Earth Observation and Geoinformation, 111 (2022): 102824.



Contents lists available at ScienceDirect

International Journal of Applied Earth Observations and Geoinformation

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

CrossGeoNet: A Framework for Building Footprint Generation of Label-Scarce Geographical Regions

Qingyu Li^{a,b}, Lichao Mou^{a,b}, Yuansheng Hua^{a,b}, Yilei Shi^c, Xiao Xiang Zhu^{a,b,*}^a Data Science in Earth Observation, Technical University of Munich (TUM), Arcisstr. 21, 80333 Munich, Germany^b Remote Sensing Technology Institute (IMF), German Aerospace Center (DLR), Oberpfaffenhofen, 82234 Wessling, Germany^c Remote Sensing Technology (LMF), Technical University of Munich (TUM), Arcisstr. 21, 80333 Munich, Germany

ARTICLE INFO

Keywords:

Building footprint
Semantic segmentation
Convolutional neural network
Co-segmentation
Planet satellite

ABSTRACT

Building footprints are essential for understanding urban dynamics. Planet satellite imagery with daily repetition frequency and high resolution has opened new opportunities for building mapping at large scales. However, suitable building mapping methods are scarce for less developed regions, as these regions lack massive annotated samples to provide strong supervisory information. To address this problem, we propose to learn cross-geolocation attention maps in a co-segmentation network, which is able to improve the discriminability of buildings within the target city and provide a more general building representation in different cities. In this way, the limited supervisory information resulting from insufficient training examples in target cities can be compensated. Our method is termed as CrossGeoNet, and consists of three elemental modules: a Siamese encoder, a cross-geolocation attention module, and a Siamese decoder. More specifically, the encoder learns feature maps from a pair of images from two different geo-locations. The cross-geolocation attention module aims at learning similarity based on these two feature maps and can provide a global overview of common objects (e.g., buildings) in different cities. The decoder predicts segmentation masks of buildings using the learned cross-geolocation attention maps and the original convolved images. The proposed method is evaluated on two datasets with different spatial resolutions, i.e., Planet dataset (3 m/pixel) and Inria dataset (0.3 m/pixel), which are collected from various locations around the world. Experimental results show that CrossGeoNet can well extract buildings of different sizes and alleviate false detections, which significantly outperforms other competitors.

1. Introduction

Building footprint maps offer insights for the comprehensive understanding of urban development. In less developed regions (e.g., Africa), significant changes occur in urban areas annually due to rapid urban expansion and city renewal (Huang et al., 2020), resulting in environmental and ecological problems (Guo et al., 2021a). Therefore, acquiring up-to-date building footprint maps for these regions is essential to the urban-related analysis.

In recent decades, high spatial resolution satellite images are capable of deriving spatial details of individual buildings. However, there are some weaknesses in high-resolution commercial satellites, e.g., high cost and low revisit frequency. This limits the regional or global building footprint generation. Planet is a new micro-satellite constellation, which consists of more than 120 satellites in orbit and is able to collect meter-

level spatial resolution imagery on a daily basis at low-cost (Houborg and McCabe, 2016). Its high revisit capability also helps to acquire low cloud cover observations for the regions with above-average cloud cover (Asner et al., 2017). To date, most high-resolution building footprint generation studies are limited to aerial imagery (Bischke et al., 2019, Bischke et al., 2019; Maggiori et al., 2017; Maggiori et al., 2017; Li et al., 2020; Li et al., 2020) or WorldView satellite imagery (Pan et al., 2020b; Pan et al., 2020b; Tonbul and Kavzoglu, 2020; Tonbul and Kavzoglu, 2020), and the investigation on Planet satellite imagery is lacking.

Although some approaches (Ivanovsky et al., 2019; Ivanovsky et al., 2019; Li et al., 2020; Li et al., 2020; Li et al., 2021; Li et al., 2021; Shi et al., 2020; Shi et al., 2020) are capable of delivering very promising results on Planet satellite imagery, they are mostly developed for Europe. To the best of our knowledge, few are dedicated to the cities in less developed regions represented by Africa, South America, and Asia,

* Corresponding author.

E-mail addresses: qingyu.li@tum.de (Q. Li), lichao.mou@dlr.de (L. Mou), yuansheng.hua@dlr.de (Y. Hua), yilei.shi@tum.de (Y. Shi), xiaoxiang.zhu@dlr.de (X.X. Zhu).<https://doi.org/10.1016/j.jag.2022.102824>

Received 18 February 2022; Received in revised form 11 May 2022; Accepted 13 May 2022

Available online 11 June 2022

1569-8432/© 2022 The Authors. Published by Elsevier B.V. This is an open access article under the CC BY license (<http://creativecommons.org/licenses/by/4.0/>).

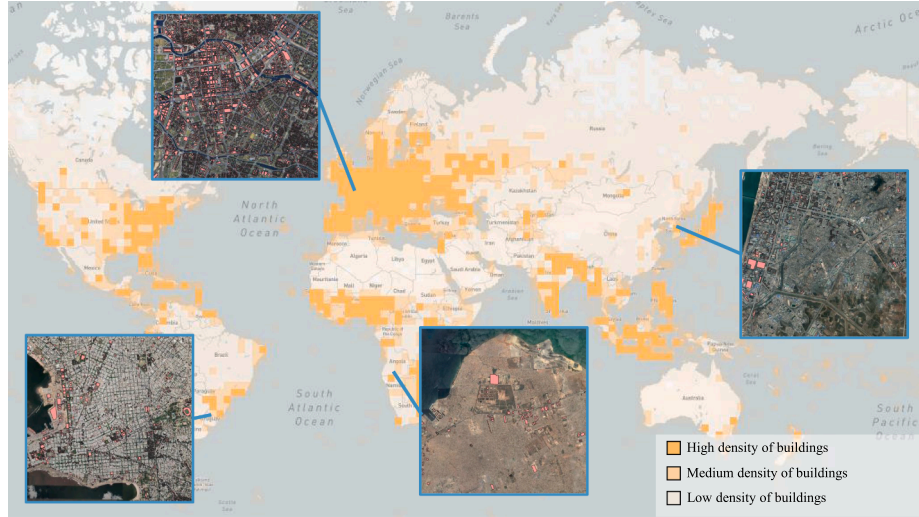


Fig. 1. The annotated building footprints in OpenStreetMap (counted by continents), and four examples of cities in Europe, Africa, South America, and Asia. The base map about building densities on OpenStreetMap is obtained from OpenStreetMap Analytics ([osm, 2021-08-24](https://osm-analytics.com/)).

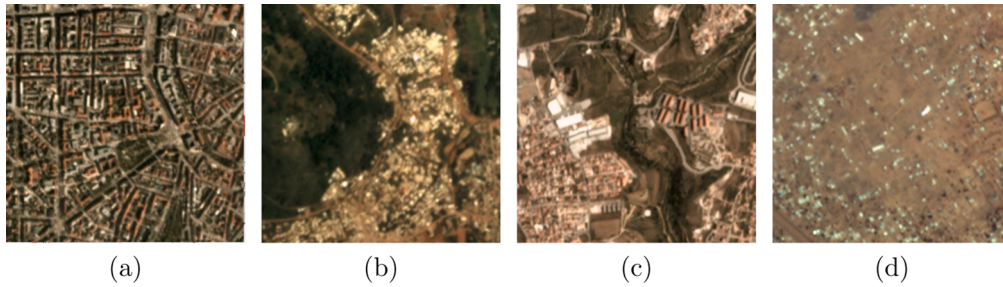


Fig. 2. Illustration of geographic peculiarities across different geolocations. The Planet satellite images are collected from (a) Munich (Germany), (b) Yaounde (Cameroon), (c) Lisbon (Portugal), and (d) Niamey (Niger), respectively. We can see that appearances of buildings in different cities are noticeably different.

where buildings differ substantially in size and type from those in Europe.

To generate building footprint maps from Planet satellite imagery, existing studies use convolutional neural networks (CNNs) that can effectively learn high-level features from raw data without heuristic feature design. Nevertheless, there remains a challenge for extracting building footprints on target cities — massive data need to be collected to promote the performance of CNNs. Considering that the manual annotation of reference data is a very time-consuming and costly process, OpenStreetMap (OSM) could be considered as a good source for acquiring manually annotated building footprints for training networks (Kaiser et al., 2017). By analyzing available building annotation data in OSM, we observe that they have an extremely uneven distribution across cities in different continents (see Fig. 1). For example, there are abundant labeled samples in European cities, while for cities in Africa, South America, and Asia, annotated data are quite limited. The lack of annotated data usually restricts the performance of existing methods in these regions, as they require a lot of strong supervisory information for network learning.

In this paper, we aim to generate building footprint maps using Planet satellite imagery for target cities that suffer from data deficit of labeled samples. In order to improve the performance of a network trained on the target city with scarce labeled data, a straightforward idea is to take advantage of the cities with massive annotated data (hereafter called auxiliary set). Nonetheless, geographic peculiarities across different geolocations raise several challenges. As shown in Fig. 2, appearances of buildings in different continents are noticeably different. This induces CNNs to produce unsatisfactory results when we directly

apply a network trained on the auxiliary set to target cities. In this regard, some works (Maggiori et al., 2016) utilize transfer learning that fine-tunes a pre-trained model with a few labeled instances in target cities. Domain adaptation methods (Vu et al., 2019) aim to transfer the knowledge learned from a domain to improve performance on target cities. Other works (He et al., 2020) utilize a new learning strategy, where the model is first pre-trained with a large number of unlabeled images in a self-supervised way and then transferred to the task of semantic segmentation with very few labeled samples.

Recently, co-segmentation is proposed for the object segmentation in computer vision, aiming at jointly segmenting semantically similar objects in video frames (Papoutsakis et al., 2017; Papoutsakis et al., 2017; Wang et al., 2019; Wang et al., 2019) or multiple images (Li et al., 2018). The success of these works suggests that co-segmentation can fully harness the sequential or pair-wise relations among consecutive frames to discover common objects, which helps to alleviate the dependency of strong supervisory information. This gives us an incentive that the co-segmentation framework may benefit our cross-city building extraction task. Therefore, we propose an end-to-end trainable network—CrossGeoNet, which consists of three modules: a Siamese encoder, a cross-geolocation attention module, and a Siamese decoder. The encoder takes as input a pair of images from two different geolocations and is responsible for learning feature representations for both images. The cross-geolocation attention module learns to explicitly encode correlations between the features of the two images, enabling the network to attend more to common objects (i.e., building in our case). The decoder combines convolved images with the cross-geolocation attention maps to generate segmentation masks through a series of

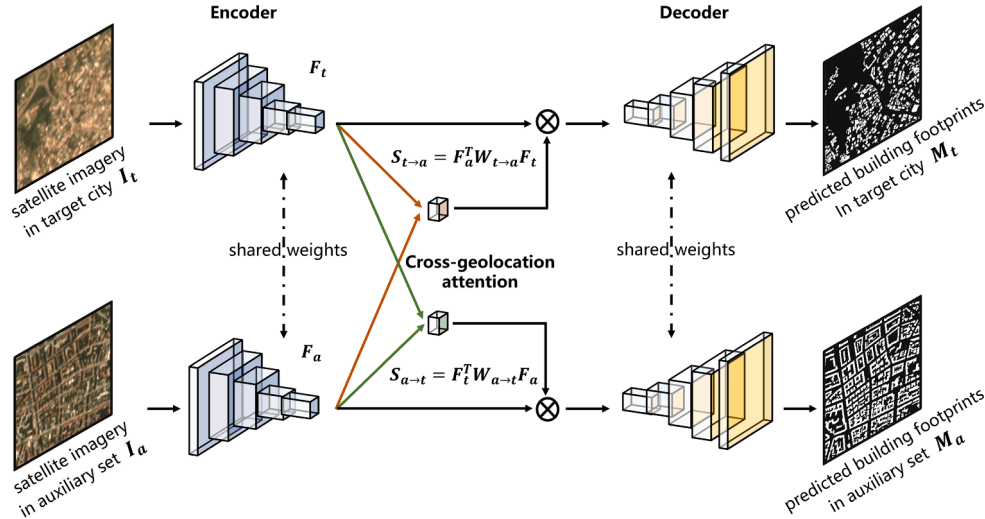


Fig. 3. Overview of the proposed CrossGeoNet framework.

deconvolutional layers. Note that the three components are jointly optimized in our method. This work's contributions are threefold.

- (1) The proposed CrossGeoNet examines the potential of Planet satellite imagery for building mapping in less developed regions (e.g., cities in Africa, South America, and Asia).
- (2) To tackle the problem of insufficient labeled samples in target cities, we propose to use a co-segmentation learning framework that can leverage a large amount of labeled data in other cities to improve the performance of a model in the target cities. To the best of our knowledge, our work is the first one that exploits co-segmentation learning to generate building footprint maps.
- (3) Since capturing the relationship between the two inputs is the key element in our CrossGeoNet, we propose a cross-geolocation attention module to effectively learn the underlying similarity between different geolocations, which is superior to other existing methods (e.g. mutual correlation (Li et al., 2018) and Fourier domain correlation (Danelljan et al., 2014)). Compared with other competitors, our approach gains significantly better results. The codes of CrossGeoNet will be made publicly available in https://github.com/lqycrystal/coseg_building.

This article is organized as follows. Section 2 presents the framework of CrossGeoNet for building footprint generation. The experiments are described in Section 3. Results are provided in Section 4. The performance of CrossGeoNet on another data source is investigated in Section 5. Eventually, Section 6 summarizes this work.

2. Methodology

In this section, the co-segmentation pipeline of CrossGeoNet is first presented. Afterward, we present the proposed cross-geolocation attention module in detail. Finally, the end-to-end network learning procedure is described.

2.1. Co-segmentation Pipeline

When objects of the same class vary in pose, shape, or color, the idea of co-learning can exploit the synergistic relationship between video frames or multiple images to provide generic features, improving model performance. In this work, our motivation is that by jointly viewing common objects (i.e., building in our case) in different geolocations, networks can learn underlying similarities for extracting more generic representations for buildings. In this regard, we propose to integrate co-

segmentation learning into the framework of building footprint generation, which is capable of fully harnessing information from various locations and further enhancing the generalizability of the model. Specifically, we propose a cross-geolocation attention module in the co-segmentation pipeline that learns to enhance latent features by encoding relations between the target city and cities from the auxiliary set. As a consequence, our co-segmentation network is able to not only improve building discriminability within target cities but also learn generic features of buildings across different cities. By doing so, the limited supervisory information in target cities can be compensated.

As shown in Fig. 3, a Siamese encoder-decoder architecture is adopted in CrossGeoNet. The Siamese encoder is composed of two identical CNNs with shared weights for the purpose of feature extraction. The input of the encoder is an image pair, where one image I_t is from a target city and the other image I_a is from the auxiliary set, and their feature representations are denoted as $F_t \in \mathbb{R}^{C \times W \times H}$ and $F_a \in \mathbb{R}^{C \times W \times H}$, respectively. H and W represent the height and width, and C denotes the channel dimension. Unlike conventional semantic segmentation networks, where high-level features are directly decoded for inferring building masks, here we enhance the learned feature maps through the proposed cross-geolocation attention module. Specifically, this module takes two feature maps as input and outputs two attention maps S_{t-a} and S_{a-t} . Afterward, they are fused with the corresponding convolved images and fed into the decoder. The Siamese decoder is comprised of a set of transposed convolutional layers that upsample the convolved images to generate two building segmentation masks M_t and M_a . Note that all modules are integrated into one framework and optimized in an end-to-end manner.

2.2. Cross-geolocation Attention

The feature maps learned from the Siamese encoder contain abstract semantic information, and when the input images contain the common object (e.g., building), their features should also include similar information. The key component of co-segmentation learning is to find similarities in feature vectors among various images. In the literature, there have been several commonly used similarity measures, e.g., mutual correlation (Li et al., 2018) and Fourier domain correlation (Danelljan et al., 2014).

Inspired by the success of self-attention (Hu et al., 2018) in capturing long-range interactions among input signals, we propose a cross-geolocation attention module that can adaptively learn the similarity between target cities and the auxiliary set. By doing so, semantic information of the common object (e.g., building) can be enhanced. More

Table 1
Statistics of the datasets utilized in this research.

	Continent	Name	The number of patches		
			train	validation	test
Target city	Africa	Yaounde	100	100	300
	South America	Porto Alegre	100	100	300
	Asia	Kyoto	100	100	300
Auxiliary set	Europe	Madrid	2971	743	0
		London	2256	565	0
		Rome	2303	576	0
		Lisbon	2043	511	0
		Munich	2271	568	0
		Zurich	1849	463	0

specifically, we calculate the cross-geolocation attention map $S_{t \rightarrow a} \in \mathbb{R}^{(WH) \times (WH)}$ between F_t and F_a as:

$$S_{t \rightarrow a} = F_a^T W_{t \rightarrow a} F_t, \quad (1)$$

where $W_{t \rightarrow a} \in \mathbb{R}^{C \times C}$ is a weight matrix. Here F_t and F_a are flattened into vectors with the size of $C \times WH$ and can be represented as:

$$F_t = [f_t^1, f_t^2, \dots, f_t^p, \dots, f_t^{WH}], \quad (2)$$

$$F_a = [f_a^1, f_a^2, \dots, f_a^q, \dots, f_a^{WH}], \quad (3)$$

where f_t^p is a C -dimensional feature vector at position $p \in \{1, 2, \dots, WH\}$ in F_t , and f_a^q is a C -dimensional feature vector at position $q \in \{1, 2, \dots, WH\}$ in F_a . Thus, the entry (q, p) of $S_{t \rightarrow a}$ reflects the similarity between f_a^q and f_t^p , and can be learned automatically. $S_{t \rightarrow a}$ is capable of capturing the dependencies between any two positions of feature maps without regard for their distance in the spatial dimension. Therefore, our cross-geolocation module can model rich contextual dependencies, which is superior to other similarity measures that only consider local features.

Since the weight matrix $W_{t \rightarrow a}$ is a square matrix, the diagonalization of $W_{t \rightarrow a}$ can be represented as follows:

$$W_{t \rightarrow a} = P_{t \rightarrow a}^{-1} D_{t \rightarrow a} P_{t \rightarrow a}, \quad (4)$$

where $P_{t \rightarrow a}$ is an invertible matrix and $D_{t \rightarrow a}$ is a diagonal matrix. Then, Eq. (1) can be rewritten as:

$$S_{t \rightarrow a} = F_a^T P_{t \rightarrow a}^{-1} D_{t \rightarrow a} P_{t \rightarrow a} F_t. \quad (5)$$

According to Eq. (5), a learnable linear transformation is first applied to the feature representation of each image, and then the similarity between these two feature representations is dynamically captured by the dot product. Similarly, the cross-geolocation attention map $S_{a \rightarrow t}$ between F_a and F_t is computed as:

$$S_{a \rightarrow t} = F_t^T P_{a \rightarrow t}^{-1} D_{a \rightarrow t} P_{a \rightarrow t} F_a, \quad (6)$$

where $P_{a \rightarrow t}$ is an invertible matrix, and $D_{a \rightarrow t}$ is a diagonal matrix.

Note that $S_{t \rightarrow a}^q$ is the q -th row of $S_{t \rightarrow a}$, which is a vector with length WH and represents the similarity between each feature vector in F_t and f_a^q . If the p -th element in $S_{t \rightarrow a}^q$ has a larger value than others, f_t^p is more similar to f_a^q than other feature vectors in F_t , which indicates a very high probability of having the common object in f_t^p and f_a^q .

Afterward, we obtain the cross-geolocation attention-enhanced features Z_t by allocating the learned cross-geolocation attention map to F_t , which is computed with the following equations:

$$Z_t = S_{t \rightarrow a} F_t^T. \quad (7)$$

And Z_a is calculated in the same manner:

$$Z_a = S_{a \rightarrow t} F_a^T. \quad (8)$$

Finally, Z_t and Z_a are reshaped into the size of $C \times H \times W$ and fed into

the Siamese decoder to produce final segmentation masks M_t and M_a , respectively.

In what follows, we discuss in detail why the proposed approach can improve the performance of a model in target cities. It is well known that contextual information is able to offer important cues for semantic segmentation tasks. In conventional CNNs, convolutions are used to extract such information. However, the performance might be limited due to their local receptive fields. Also, inadequate samples affect the learning of CNNs. On the contrary, the proposed cross-geolocation module explores global contextual information by learning cross-geolocation attention maps. Specifically, for a pixel in a sample from the target city, the cross-geolocation attention map can effectively capture relations between it and not only all other pixels in the same sample but also all pixels in a sample from the auxiliary set. Afterward, CrossGeoNet selectively aggregates global contextual information to provide a global view of common objects (i.e., building), alleviating the influence of background. In other words, we leverage the auxiliary set to provide additional supervisory information to enhance the discriminability of building, which improves building extraction results on the target city.

2.3. Network Learning

We propose an end-to-end training pipeline for the supervised learning of CrossGeoNet. The whole network is trained by the following loss function:

$$L = L_t + \lambda \cdot L_a, \quad (9)$$

where L_t and L_a are two cross-entropy loss functions for measuring the difference between segmentation masks and their corresponding ground-truth masks. λ is a hyperparameter to control the importance of the second loss.

3. Experiments

3.1. Dataset

In this work, we collect Planet satellite images and their corresponding OSM building footprints from different cities all over the globe. Planet satellite images have 3 bands (i.e., red, green, blue), and their spatial resolution is 3 m/pixel. In the pre-processing step, all images and ground-truth masks are cropped into small patches with the size of 256×256 pixels. To thoroughly investigate the performance of CrossGeoNet, we select three target cities from different continents: Yaounde (Cameroon), Porto Alegre (Brazil), and Kyoto (Japan). As to the auxiliary set, 6 European cities, Madrid (Spain), London (UK), Rome (Italy), Lisbon (Portugal), Munich (Germany), and Zurich (Switzerland), are selected due to their massive building footprint annotations. The numbers of patches collected from each city for network training, validation, and test are reported in Table 1.

3.2. Experimental Setup

To verify the effectiveness of CrossGeoNet for building footprint generation, we compare it with several commonly-used network learning methods, i.e., Baseline-t, Baseline-a, Baseline-a+t, fine-tuning, ADVENT (Vu et al., 2019) IntraDA (Pan et al., 2020a), MetaCorrection (Guo et al., 2021b), MoCo (He et al., 2020), DenseCL (Wang et al., 2021), U-Net-AFM (Li et al., 2021), CBRNet (Guo et al., 2022), EPU-Net (Guo et al., 2021a), and CSGANet (Chen et al., 2021). Note that experiments are independently conducted in three target cities. That is to say, for the experiment in one target city, training samples consist of only patches from that target city and the auxiliary set. For the evaluation of our cross-geolocation attention module, we conduct comparisons with the aforementioned two similarity measures, i.e., mutual correlation (Li

Table 2

Accuracies (%) of different learning methods for building footprint generation on tagert cities.

Method	Yaounde		Porto Alegre		Kyoto	
	F1 score	IoU	F1 score	IoU	F1 score	IoU
Baseline-t	63.85	46.90	58.57	41.41	59.80	42.65
Baseline-a	1.90	0.96	27.41	15.88	36.35	22.21
Baseline-a+t	64.95	48.10	60.44	43.31	62.76	45.72
Fine-tuning	63.35	46.36	60.12	42.98	59.31	42.16
ADVENT(Vu et al., 2019)	55.26	38.18	31.13	18.43	46.89	30.63
IntraDA (Pan et al., 2020a)	56.59	39.46	40.86	25.67	53.05	36.10
MetaCorrection (Guo et al., 2021b)	55.44	38.35	51.68	34.84	49.27	32.69
MoCo(He et al., 2020)	60.98	43.87	57.59	40.44	58.22	41.06
DenseCL(Wang et al., 2021)	60.99	43.88	59.12	39.00	58.10	40.94
U-Net-AFM (Li et al., 2021)	61.32	44.19	53.64	36.72	52.86	36.01
CBRNet (Guo et al., 2022)	63.52	46.54	59.98	42.84	61.78	44.70
EPU-Net (Guo et al., 2021a)	52.45	35.55	45.72	29.64	50.04	33.37
CSGANet (Chen et al., 2021)	61.51	44.42	56.69	39.55	58.07	40.92
CrossGeoNet	67.77	51.26	62.12	45.05	65.28	48.46

et al., 2018) and Fourier domain correlation (Danelljan et al., 2014).

3.3. Training Details

CrossGeoNet is implemented on PyTorch framework and trained on an NVIDIA Quadro P4000 GPU with 8 GB memory. The training epochs of all models are set as 100 epochs, and stochastic gradient descent (SGD) with a learning rate of 0.001 is set as the optimizer. The size of the training batch for all models is 4. Detailed configurations of all methods in our experiments are presented as follows:

- (1) CrossGeoNet: Since our model is trained for each target city independently, we select I_t and I_a from one target city and the auxiliary set, respectively, in the training phase. To enlarge the number of training pairs, for each patch in the target city, we create 100 duplicates and pair them with 100 samples randomly selected from one city in the auxiliary set. In the inference stage, I_t and I_a are both selected from test patches of the target city. The loss term weighting parameter λ in Eq. (9) is set as 0.00001 empirically.
- (2) Baseline-t: An Efficient-UNet is trained and tested with training and test sets of the target city.
- (3) Baseline-a: An Efficient-UNet is trained with samples collected from the auxiliary set and tested on test instances in the target city.
- (4) Baseline-a+t: An Efficient-UNet is trained using samples from training sets of the target city and the auxiliary set, and tested on test samples from the target city.
- (5) Fine-tuning: It consists of two steps. Firstly, all samples from the auxiliary set are used to pre-train an Efficient-UNet. Secondly, the pre-trained network is fine-tuned with the training set of the target city.
- (6) ADVENT (Vu et al., 2019), IntraDA (Pan et al., 2020a), and MetaCorrection (Guo et al., 2021b): They aim at addressing the task of domain adaptation in semantic segmentation. The auxiliary set is regarded as the source domain, and the target city is the target domain.
- (7) MoCo (He et al., 2020) and DenseCL (Wang et al., 2021): They first learn knowledge from a large number of unlabeled images in a self-supervised way. Afterward, the weights are transferred to

the task of semantic segmentation. In our research, MoCo (He et al., 2020) learns from the auxiliary set, while for DenseCL (Wang et al., 2021), we use its pre-trained weights (Deng et al., 2009).

- (8) U-Net-AFM (Li et al., 2021), CBRNet (Guo et al., 2022), EPU-Net (Guo et al., 2021a), and CSGANet (Chen et al., 2021): They are semantic segmentation networks for the task of building footprint generation.

Note that for MoCo (He et al., 2020), DenseCL (Wang et al., 2021), U-Net-AFM (Li et al., 2021), CBRNet (Guo et al., 2022), EPU-Net (Guo et al., 2021a), and CSGANet (Chen et al., 2021), we have separately organized the training set according to three experiment procedures (i. e., Baseline-t, Baseline-a+t, and Fine-tuning), and the best result among three cases is reported.

We evaluate the performance of all models using two metrics: F1 score and intersection over union (IoU).

4. Results

4.1. Comparison of Different Learning Methods

This section presents the comparisons among CrossGeoNet, Baseline-t, Baseline-a, Baseline-a+t, fine-tuning, ADVENT (Vu et al., 2019), IntraDA (Pan et al., 2020a), MetaCorrection (Guo et al., 2021b), Moco (He et al., 2020), DenseCL (Wang et al., 2021), U-Net-AFM (Li et al., 2021), CBRNet (Guo et al., 2022), EPU-Net (Guo et al., 2021a), and CSGANet (Chen et al., 2021). Their performance is evaluated from quantitative (cf. Tables 2) and and qualitative (see Figs. 4–6) perspectives in three target cities.

Compared with Baseline-t, the proposed method has largely improved the accuracy. It can be seen from numerical results in three target cities that CrossGeoNet reaches improvements of above 3% in both F1 score and IoU. Especially for the target city of Kyoto, our method obtains increments of 5.48% in F1 score and 5.81% in IoU, respectively. As shown in Fig. 4, Baseline-t fails to recover complete masks of large buildings. This is due to the fact that limited training samples can not represent the true class distribution comprehensively (Hou et al., 2019). Although Baseline-a exploits massive annotated samples of the auxiliary set, it still performs worse than CrossGeoNet. For instance, in the target city of Yaounde (see Table 2), Baseline-a only achieves 1.90% in F1 score and 0.96% in IoU. Moreover, these results are worse than those of Baseline-t. This is caused by significant differences between the target cities and the auxiliary set, e.g., variant morphological appearance of human settlements and material available for building construction (Li et al., 2020).

Afterward, we select another seven competitors (Baseline-a+t, fine-tuning, ADVENT (Vu et al., 2019), IntraDA (Pan et al., 2020a), MetaCorrection (Guo et al., 2021b), MoCo (He et al., 2020), and DenseCL (Wang et al., 2021)) to make a further comparison, as these methods also jointly utilize training samples of both the target city and the auxiliary set. Fine-tuning is a commonly used method to handle the issue of scarce training data in target datasets (Maggiori et al., 2016). Nevertheless, compared with Baseline-t, fine-tuning even leads to decreases in accuracy metrics for Yaounde and Kyoto. A possible explanation is that the gap between target cities and auxiliary set is quite large, making it difficult to transfer the knowledge learned from the auxiliary set to target cities. Domain adaptation methods are also capable of transferring the knowledge from the auxiliary set to the target city. From the results in Table 2, it can be seen that ADVENT (Vu et al., 2019), IntraDA (Pan et al., 2020a), and MetaCorrection (Guo et al., 2021b) perform worse than fine-tuning in knowledge transfer. One important reason is that the labels in the target domain are not utilized by domain adaptation methods. It can be observed from statistical results that MoCo (He et al., 2020) and DenseCL (Wang et al., 2021) are even inferior to Baseline-t on all three cities. This might be attributed to two factors. On

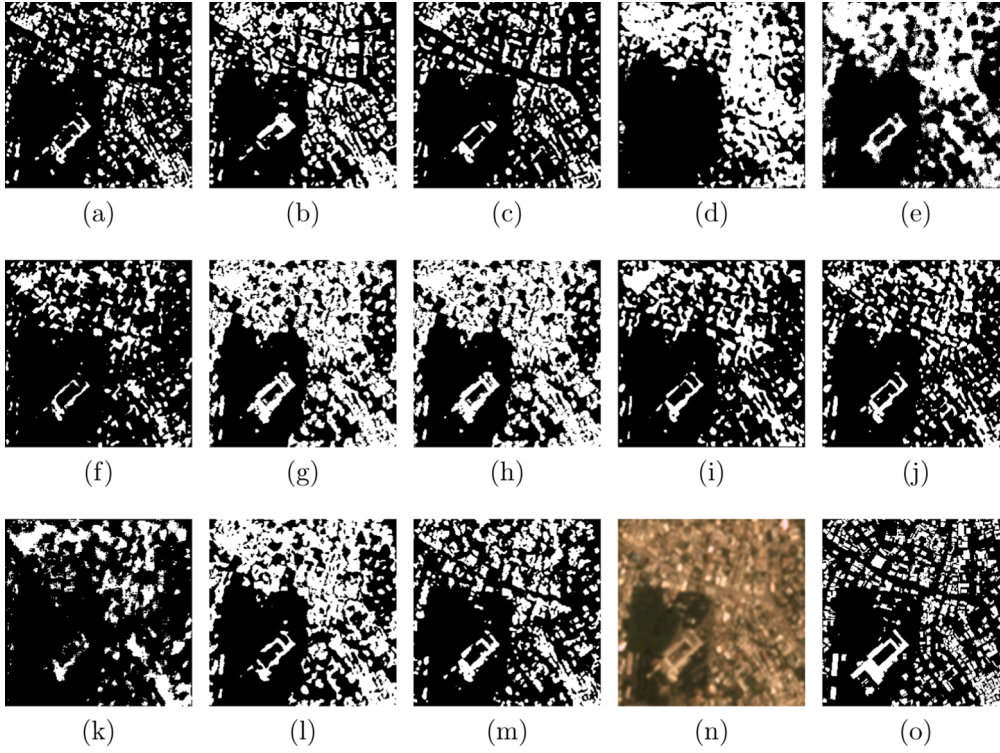


Fig. 4. Examples of building extraction results obtained by different learning methods. (a) Baseline-t. (b) Baseline-a+t. (c) Fine-tuning. (d) ADVENT (Vu et al., 2019). (e) IntraDA (Pan et al., 2020a). (f) MetaCorrection (Guo et al., 2021b). (g) MoCo (He et al., 2020). (h) DenseCL (Wang et al., 2021). (i) U-Net-AFM (Li et al., 2021). (j) CBRNet (Guo et al., 2022). (k) EPU-Net (Guo et al., 2021a). (l) CSGANet (Chen et al., 2021). (m) CrossGeoNet. (n) and (o) are Planet satellite imagery and ground reference from Yaounde.

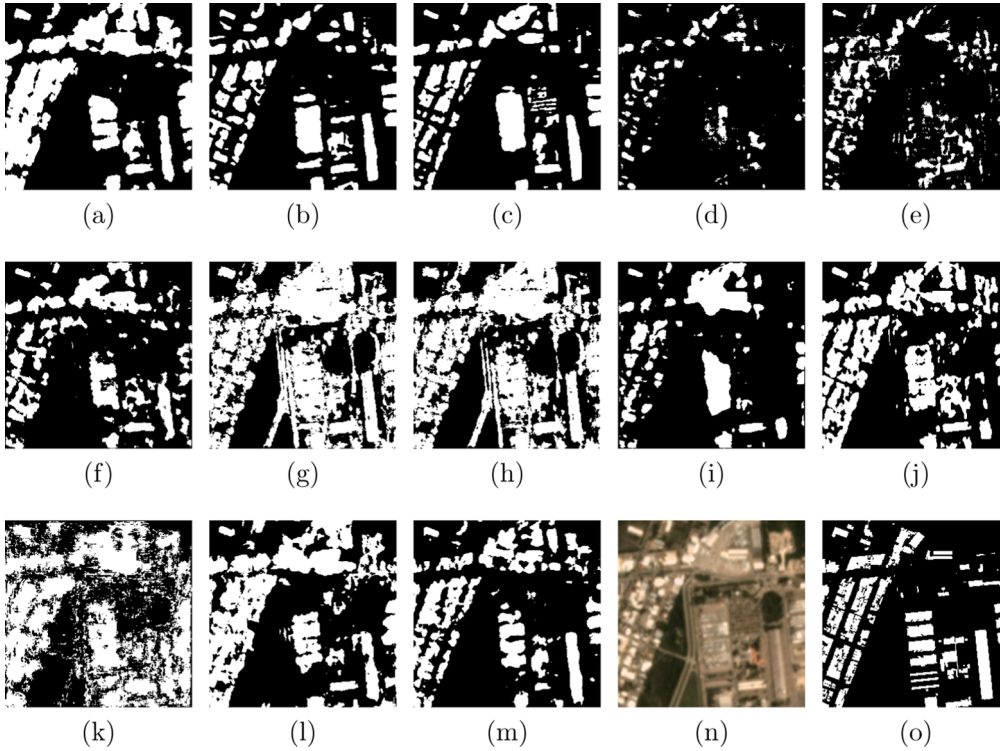


Fig. 5. Examples of building extraction results obtained by different learning methods. (a) Baseline-t. (b) Baseline-a+t. (c) Fine-tuning. (d) ADVENT (Vu et al., 2019). (e) IntraDA (Pan et al., 2020a). (f) MetaCorrection (Guo et al., 2021b). (g) MoCo (He et al., 2020). (h) DenseCL (Wang et al., 2021). (i) U-Net-AFM (Li et al., 2021). (j) CBRNet (Guo et al., 2022). (k) EPU-Net (Guo et al., 2021a). (l) CSGANet (Chen et al., 2021). (m) CrossGeoNet. (n) and (o) are Planet satellite imagery and ground reference from Porto Alegre.

the one hand, the annotated information of the auxiliary set has not been leveraged in self-supervised learning. On the other hand, large differences existing between the auxiliary set and target cities might impair the model performance when migrated to target cities.

CrossGeoNet has achieved the highest accuracies among all methods, and it shows nearly 2% improvements of F1 score and IoU on all target cities compared to Baseline-a+t. From qualitative results, we can

observe that Baseline-a+t fails to detect some small buildings (cf. Fig. 6). This can be explained by the imbalanced number of training samples collected from target cities and the auxiliary set. When training samples of the auxiliary set dominate the learning procedure, the network fails to guarantee accurate segmentation in target cities. On the contrary, our method is able to avoid these omission errors and reconstruct complete building structures to a large extent. These observations suggest that

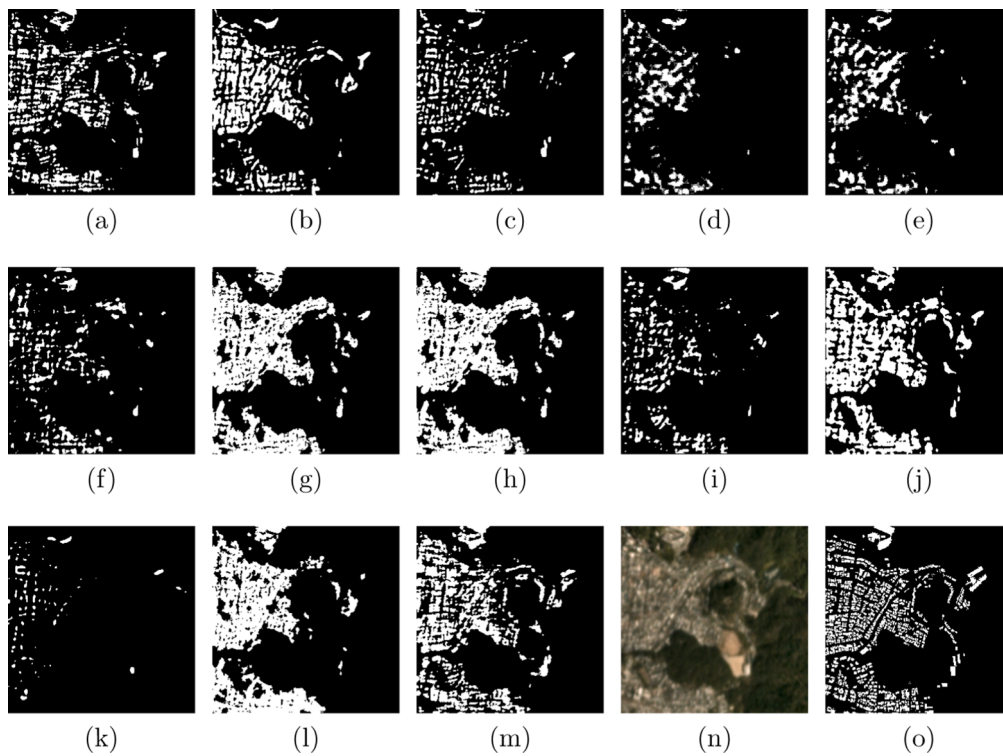


Fig. 6. Examples of building extraction results obtained by different learning methods. (a) Baseline-t. (b) Baseline-a+t. (c) Fine-tuning. (d) ADVENT (Vu et al., 2019). (e) IntraDA (Pan et al., 2020a). (f) MetaCorrection (Guo et al., 2021b). (g) MoCo (He et al., 2020). (h) DenseCL (Wang et al., 2021). (i) U-Net-AFM (Li et al., 2021). (j) CBRNet (Guo et al., 2022), (k) EPU-Net (Guo et al., 2021a). (l) CSGANet (Chen et al., 2021). (m) CrossGeoNet. (n) and (o) are Planet satellite imagery and ground reference from from Kyoto.

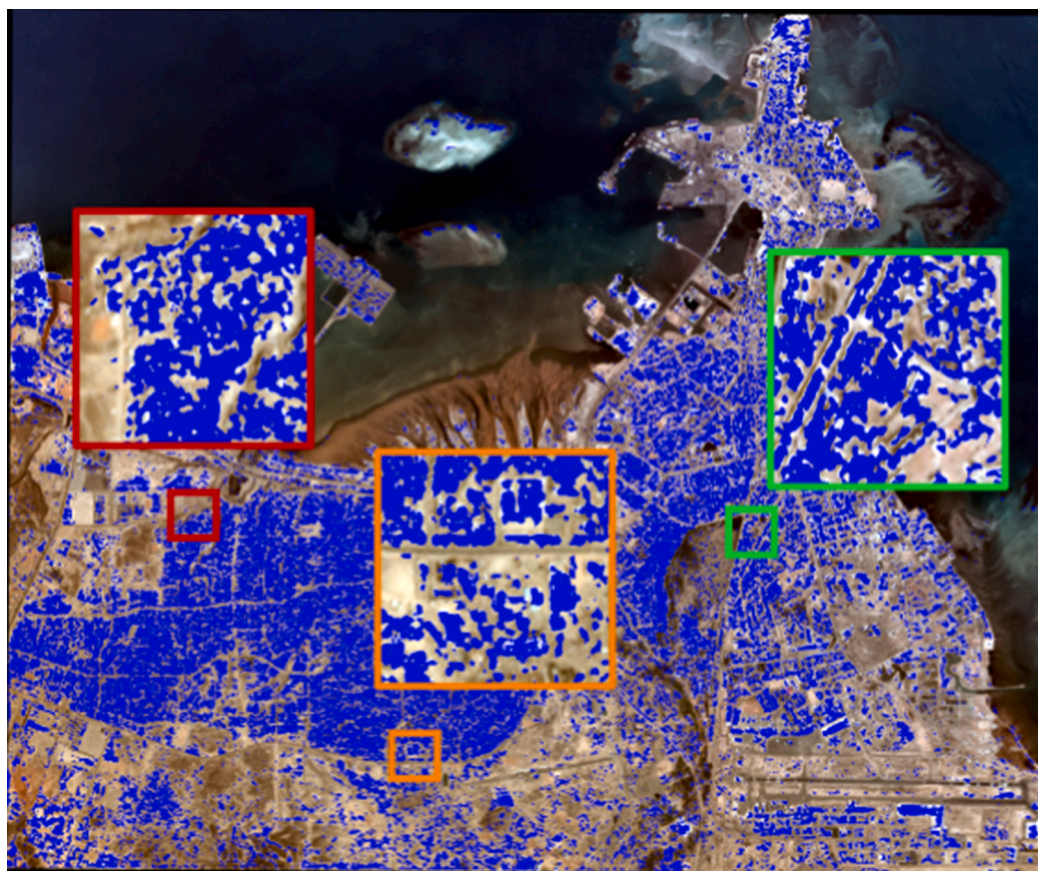


Fig. 7. Building extraction results (in blue) obtained by CrossGeoNet from Djibouti and three zoomed in areas.

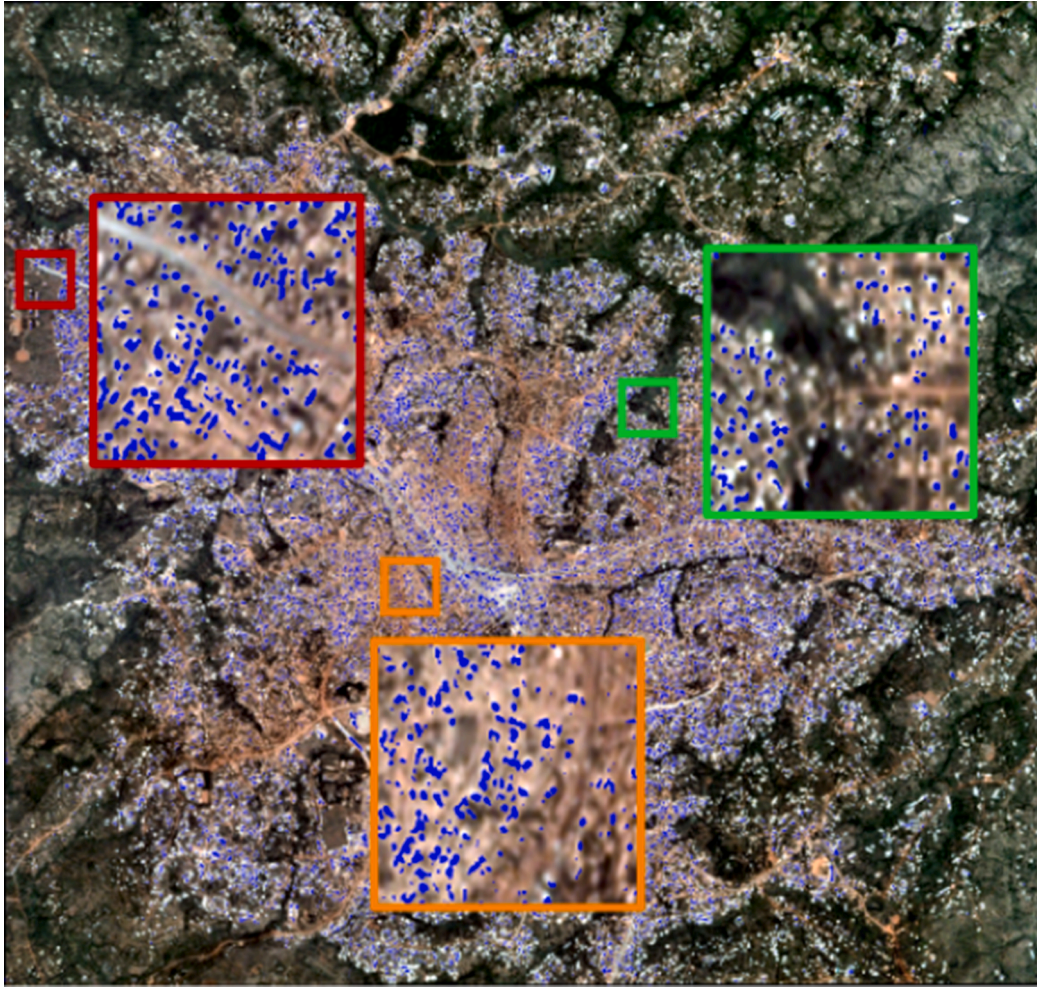


Fig. 8. Building extraction results (in blue) obtained by CrossGeoNet from Bafoussam and three zoomed in areas.

Table 3

Accuracies (%) of different similarity measures on Yaounde.

Method	F1 score	IoU
Mutual correlation (Li et al., 2018)	66.78	50.13
Fourier domain correlation (Danelljan et al., 2014)	65.76	48.99
Proposed cross-geolocation attention module	67.77	51.26

CrossGeoNet benefits from the learning of the cross-geolocation attention module, enabling the leverage of rich relationships between target cities and the auxiliary set.

We then compare CrossGeoNet with U-Net-AFM (Li et al., 2021), CBRNet (Guo et al., 2022), EPU-Net (Guo et al., 2021a), and CSGANet (Chen et al., 2021), which are four state-of-the-art methods for the task of building footprint generation. It can be observed from the statistical

and visual results on three cities that our method surpasses all other building extraction methods.

We further explore the generalizability of model trained by CrossGeoNet and test it on unseen cities (which are neither from the target city nor from the auxiliary set). Note that we directly apply the trained model to the unseen cities. Specifically, we select two African cities, Djibouti (Republic of Djibouti) and Bafoussam (Cameroon). In the training phase, we select Yaounde as the target city due to its high similarity with Djibouti and Bafoussam. Figs. 7 and 8 illustrate visual results on these two cities. CrossGeoNet is promising to provide building footprint maps in other unseen geographic regions.

4.2. Comparison With Different Similarity Measures

Explicitly capturing similarities among various cities is essential for

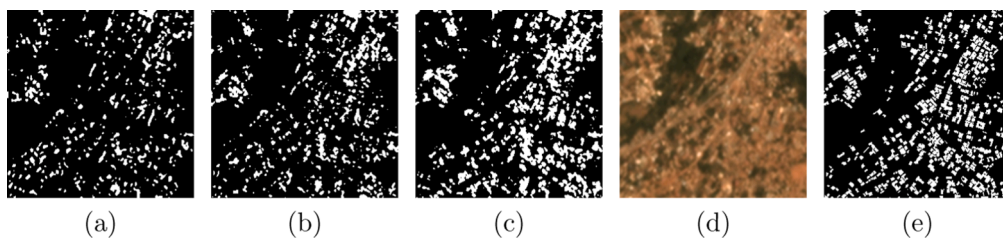


Fig. 9. Examples of building extraction results obtained by different similarity measures. (a) Mutual correlation (Li et al., 2018). (b) Fourier domain correlation (Danelljan et al., 2014). (c) Proposed cross-geolocation attention module. (d) and (e) are Planet satellite imagery and ground reference from Yaounde.

Table 4

Accuracies (%) of different learning methods for building footprint generation on Vienna.

Method	F1 score	IoU
Baseline-t	82.32	69.96
Baseline-a	78.75	64.95
Baseline-a+t	85.02	73.95
Fine-tuning	85.38	74.49
ADVENT(Vu et al., 2019)	81.07	68.17
IntraDA (Pan et al., 2020a)	82.44	70.12
MetaCorrection (Guo et al., 2021b)	83.93	72.31
MoCo (He et al., 2020)	85.66	74.91
DenseCL (Wang et al., 2021)	86.52	76.25
U-Net-AFM (Li et al., 2021)	86.64	76.42
CBRNet (Guo et al., 2022)	86.46	76.09
EPU-Net (Guo et al., 2021a)	86.04	75.50
CSGANet (Chen et al., 2021)	86.59	76.35
CrossGeoNet	87.51	77.79

co-segmentation methods. Therefore, we further investigate the aforementioned two similarity measures, i.e., mutual correlation (Li et al., 2018) and Fourier domain correlation (Danelljan et al., 2014), to make a comparison with our cross-geolocation attention module.

The statistical results on Yaounde are reported in Table 3. The proposed module outperforms the other two methods by over 1% in statistical metrics. In Fig. 9, the building masks obtained by CrossGeoNet are much closer to ground-truth masks. However, the results provided by Fourier domain correlation show many omitted detection. One reason is that mutual correlation (Li et al., 2018) and Fourier domain correlation (Danelljan et al., 2014) operate on a local neighborhood, leading to the loss of global information. In contrast, our cross-geolocation attention module can capture long-range dependencies, enabling the leverage of useful information from more remote regions in the target image and those from the auxiliary set. This is beneficial to the reduction of semantic noise and the enhancement of semantic information of buildings. Another reason is that these two methods simply concatenate correlation maps with original convolved images to generate new features, while our module updates features by selectively

aggregating contexts according to the learned attention maps. By doing so, mutual gains can be achieved through similar features, providing more representative features for building footprint generation.

5. Performance Investigation on Another Data Source

In this section, we further investigate the performance of CrossGeoNet on another dataset, INRIA Aerial Image Labeling data (Maggiori et al., 2017), comprising images captured by airborne sensors. The INRIA dataset is a benchmark dataset, which consists of 360 tiles of aerial imagery. Each aerial image has 5000×5000 pixels at a spatial resolution of 30 cm/pixel. In this dataset, only ground reference data for five cities (Austin, Chicago, Kitsap County, Western Tyrol, and Vienna) are made publicly available, and hence we only conduct experiments on these cities. According to the setup in (Bischke et al., 2019), data are split into training and validation sets in our research. We observe that buildings in Vienna have very different structures and sizes in comparison with the other four cities. Therefore, we select Vienna as the target city and the other four cities as the auxiliary set. To verify the effectiveness of CrossGeoNet on INRIA dataset, we make a comparison of different learning methods, i.e., Baseline-t, Baseline-a, Baseline-a+t, fine-tuning, ADVENT (Vu et al., 2019) IntraDA (Pan et al., 2020a), MetaCorrection (Guo et al., 2021b), MoCo (He et al., 2020), DenseCL (Wang et al., 2021), U-Net-AFM (Li et al., 2021), CBRNet (Guo et al., 2022), EPU-Net (Guo et al., 2021a), CSGANet (Chen et al., 2021), and CrossGeoNet. Note the statistics are computed from the validation set of

Table 5

Accuracies (%) of different learning methods on Vienna. Auxiliary and target sets are chosen from Vienna for ensuring similar data distribution.

Method	F1 score	IoU
Baseline-t	78.93	65.19
Baseline-a	81.27	68.45
Baseline-a+t	82.32	69.96
CrossGeoNet	86.38	76.03

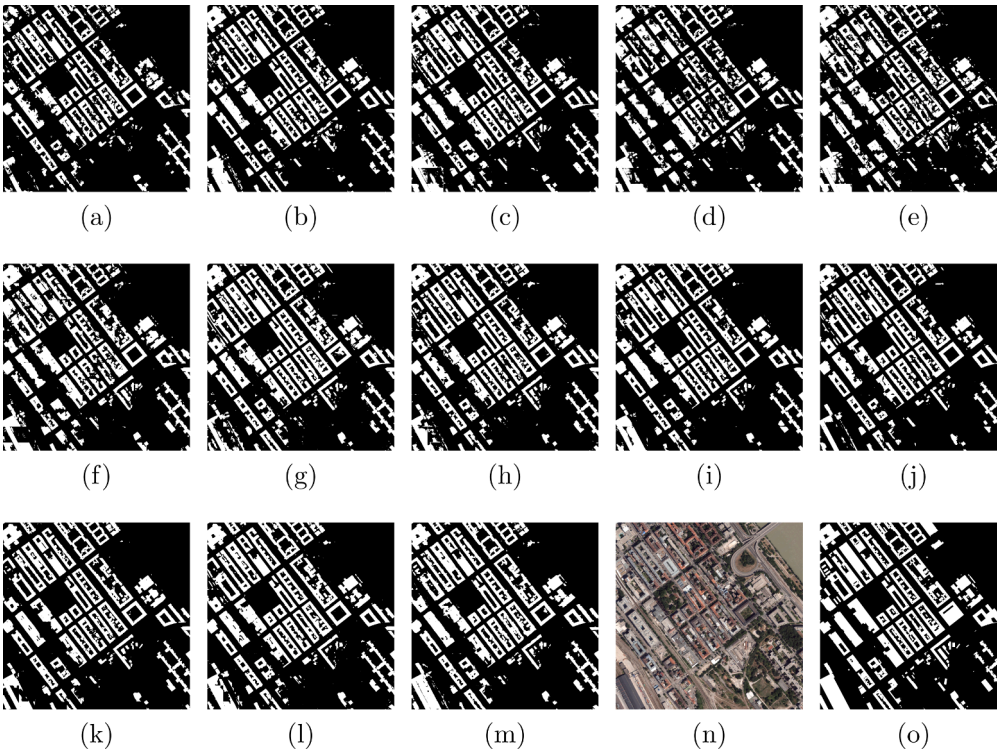


Fig. 10. Examples of building extraction results obtained by different learning methods. (a) Baseline-t. (b) Baseline-a+t. (c) Fine-tuning. (d) ADVENT (Vu et al., 2019). (e) IntraDA (Pan et al., 2020a). (f) MetaCorrection (Guo et al., 2021b). (g) MoCo (He et al., 2020). (h) DenseCL (Wang et al., 2021). (i) U-Net-AFM (Li et al., 2021). (j) CBRNet (Guo et al., 2022). (k) EPU-Net (Guo et al., 2021a). (l) CSGANet (Chen et al., 2021). (m) CrossGeoNet. (n) and (o) are INRIA aerial imagery and ground reference from Vienna.



Fig. 11. Examples of building extraction results obtained by different learning methods. (a) Baseline-t. (b) Baseline-a. (c) Baseline-a+t. (d) CrossGeoNet. (e) and (f) are INRIA aerial imagery and ground reference from Vienna. Auxiliary and target sets are chosen from Vienna for ensuring similar data distribution.

Vienna.

We first compare the proposed method against the Baseline-a. It is observed from the statistical results in Table 4, our network obtains increments of 12.84% in IoU. Moreover, CrossGeoNet surpasses Baseline-t by 7.83% in IoU. This indicates that the proposed approach is able to boost the network performance by the joint use of training samples from both the target city and the auxiliary set. From accuracy metrics in Table 4, the proposed method has achieved better performance than other learning methods that aim at transferring the knowledge learned from the auxiliary set to the target city. This demonstrates the effectiveness and robustness of the proposed method for this task, as cross-geolocation co-segmentation learning is able to improve the results on different data sources. When compared with state-of-the-art building extraction methods, CrossGeoNet shows above 1.3% improvement in IoU.

Fig. 10 presents a visual comparison among different learning methods on Vienna. The building footprints generated by CrossGeoNet are more accurate and reliable, as they coincide better with the ground reference when compared with the other methods. For instance, most methods detect only a part of the large building in the bottom left area. In contrast, the proposed approach is capable of accurately capturing a more complete roof outline. Furthermore, for buildings in complex shapes, buildings masks obtained by our network contain more detailed structures, which suggests that CrossGeoNet is still promising in such challenging situations.

In order to investigate the performance of CrossGeoNet when target and auxiliary sets are similar, we have split the original training data of Vienna into two parts, i.e., auxiliary set and target set. Furthermore, we explore the performance of models trained by different learning methods. Specifically, we compare CrossGeoNet with three competitors (i.e., Baseline-t, Baseline-a, and Baseline-a+t) quantitatively and qualitatively. The quantitative results are shown in Table 5. Baseline-t performs poorly than Baseline-a. This is because the number of training patches in the target set is smaller than that in the auxiliary set, which makes it difficult for Baseline-t to achieve good results. Baseline-a+t provides better results than both Baseline-a and Baseline-t, as all training patches are jointly utilized during network learning. It should be noted that CrossGeoNet significantly outperforms Baseline-a+t, with the IoU improved by 6.07%. This demonstrates that our cross-geolocation co-segmentation learning helps to improve model performance. Moreover, this improvement is more significant than that in the case where target and auxiliary sets are less similar. This is because the similarity between target and auxiliary contributes to extracting more generic representations for buildings. Fig. 11 illustrates visual comparisons of different learning methods. Baseline-t and Baseline-a fail to detect some building footprints on the top area. On the contrary, CrossGeoNet is able to alleviate omission errors.

6. Conclusion

Planet satellite imagery holds potentials for generating high-resolution building footprint maps at a large scale. However, generating building footprint maps from Planet satellite imagery is difficult for less developed regions because of the lack of massive annotated

samples. Given these issues, we have proposed a novel end-to-end building mapping method, namely CrossGeoNet, aiming at exploring the use of Planet satellite images in detecting buildings on the target city with scarce labeled samples. CrossGeoNet comprises three modules: a Siamese encoder, a cross-geolocation attention module, and a Siamese decoder. More specifically, the encoder is designed to learn features from a pair of images from different geolocations. Afterward, the cross-geolocation attention module learns to encode similarities between them, enabling the capture of a more discriminative and generic representation of the common object (i.e., building in our case). Finally, the decoder exploits the original feature maps and the learned cross-geolocation attention maps to predict building masks. We investigate the proposed approach on two datasets with different spatial resolutions, i.e., Planet dataset (3 m/pixel) and Inria dataset (0.3 m/pixel), which are collected from diverse cities across the globe. Experimental results suggest that the incorporation of the proposed cross-geolocation attention module in co-segmentation learning can offer more satisfactory building footprints than other competitors. Thus, we believe that CrossGeoNet is a robust solution for the task of building footprint generation when dealing with scarce training samples within target cities.

Acknowledgement

The work 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*), by the Helmholtz Association through the Framework of Helmholtz AI (grant number: ZT-I-PF-5-01) - Local Unit "Munich Unit @Aeronautics, Space and Transport (MASTr)" and Helmholtz Excellent Professorship "Data Science in Earth Observation - Big Data Fusion for Urban Research" (grant number: W2-W3-100), by the German Federal Ministry of Education and Research (BMBF) in the framework of the international future AI lab "AI4EO – Artificial Intelligence for Earth Observation: Reasoning, Uncertainties, Ethics and Beyond" (grant number: 01DD20001) and by German Federal Ministry of Economics and Technology in the framework of the "national center of excellence ML4Earth" (grant number: 50EE2201C).

References

- OpenStreetMap Analytics analysis map. <http://osm-analytics.org/#/>. Accessed: 2021-08-24.
- Asner, G.P., Martin, R.E., Mascaro, J., 2017. Coral reef atoll assessment in the south china sea using planet dove satellites. *Remote Sensing in Ecology and Conservation* 3, 57–65.
- Bischke, B., Helber, P., Folz, J., Borth, D., Dengel, A., 2019. Multi-task learning for segmentation of building footprints with deep neural networks, in: 2019 IEEE International Conference on Image Processing (ICIP), IEEE. pp. 1480–1484.
- Chen, S., Shi, W., Zhou, M., Zhang, M., Xuan, Z., 2021. Cgsanet: A contour-guided and local structure-aware encoder-decoder network for accurate building extraction from very high-resolution remote sensing imagery. *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing* 15, 1526–1542.
- Danelljan, M., Häger, G., Khan, F., Felsberg, M., 2014. Accurate scale estimation for robust visual tracking, in: British Machine Vision Conference, Nottingham, September 1–5, 2014, BMVA Press.
- Deng, J., Dong, W., Socher, R., Li, L.J., Li, K., Fei-Fei, L., 2009. Imagenet: A large-scale hierarchical image database, in: 2009 IEEE conference on computer vision and pattern recognition, Ieee. pp. 248–255.

- Guo, H., Du, B., Zhang, L., Su, X., 2022. A coarse-to-fine boundary refinement network for building footprint extraction from remote sensing imagery. *ISPRS Journal of Photogrammetry and Remote Sensing* 183, 240–252.
- Guo, H., Shi, Q., Marinoni, A., Du, B., Zhang, L., 2021a. 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.
- Guo, X., Yang, C., Li, B., Yuan, Y., 2021b. Metacorection: Domain-aware meta loss correction for unsupervised domain adaptation in semantic segmentation.
- He, K., Fan, H., Wu, Y., Xie, S., Girshick, R., 2020. Momentum contrast for unsupervised visual representation learning. In: *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 9729–9738.
- Hou, R., Chang, H., Ma, B., Shan, S., Chen, X., 2019. Cross attention network for few-shot classification. *arXiv preprint arXiv:1910.07677*.
- Houborg, R., McCabe, M.F., 2016. High-resolution ndvi from planet's constellation of earth observing nano-satellites: A new data source for precision agriculture. *Remote Sensing* 8, 768.
- Hu, J., Shen, L., Sun, G., 2018. Squeeze-and-excitation networks, in: *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 7132–7141.
- Huang, X., Cao, Y., Li, J., 2020. An automatic change detection method for monitoring newly constructed building areas using time-series multi-view high-resolution optical satellite images. *Remote Sensing of Environment* 244, 111802.
- Ivanovsky, L., Khryashchev, V., Pavlov, V., Ostrovskaya, A., 2019. Building detection on aerial images using u-net neural networks, in: *2019 24th Conference of Open Innovations Association (FRUCT)*, IEEE. pp. 116–122.
- Kaiser, P., Wegner, J.D., Lucchi, A., Jaggi, M., Hofmann, T., Schindler, K., 2017. Learning aerial image segmentation from online maps. *IEEE Transactions on Geoscience and Remote Sensing* 55, 6054–6068.
- Li, Q., Mou, L., Hua, Y., Shi, Y., Zhu, X.X., 2021. Building footprint generation through convolutional neural networks with attraction field representation. *IEEE Transactions on Geoscience and Remote Sensing*.
- Li, Q., Shi, Y., Auer, S., Roschlaub, R., Möst, K., Schmitt, M., Glock, C., Zhu, X.X., 2020. Detection of undocumented building constructions from official geodata using a convolutional neural network. *Remote Sensing* 12, 3537.
- Li, Q., Shi, Y., Huang, X., Zhu, X.X., 2020. Building footprint generation by integrating convolution neural network with feature pairwise conditional random field (fpcrf). *IEEE Transactions on Geoscience and Remote Sensing*.
- Li, W., Jafari, O.H., Rother, C., 2018. Deep object co-segmentation, in: *Asian Conference on Computer Vision*, Springer. pp. 638–653.
- Maggiori, E., Tarabalka, Y., Charpiat, G., Alliez, P., 2016. Convolutional neural networks for large-scale remote-sensing image classification. *IEEE Transactions on Geoscience and Remote Sensing* 55, 645–657.
- Maggiori, E., Tarabalka, Y., Charpiat, G., Alliez, P., 2017. Can semantic labeling methods generalize benchmark to any city? the inria aerial image labeling, in: *IEEE International Geoscience and Remote Sensing Symposium (IGARSS)*, IEEE.
- Pan, F., Shin, I., Rameau, F., Lee, S., Kweon, I.S., 2020a. Unsupervised intra-domain adaptation for semantic segmentation through self-supervision, in: *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*.
- Pan, Z., Xu, J., Guo, Y., Hu, Y., Wang, G., 2020b. Deep learning segmentation and classification for urban village using a worldview satellite image based on u-net. *Remote Sensing* 12, 1574.
- Papoutsakis, K., Panagiotakis, C., Argyros, A.A., 2017. Temporal action co-segmentation in 3d motion capture data and videos, in: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 6827–6836.
- Shi, Y., Li, Q., Zhu, X.X., 2020. Building segmentation through a gated graph convolutional neural network with deep structured feature embedding. *ISPRS Journal of Photogrammetry and Remote Sensing* 159, 184–197.
- Tonbul, H., Kavzoglu, T., 2020. Semi-automatic building extraction from worldview-2 imagery using taguchi optimization. *Photogrammetric Engineering & Remote Sensing* 86, 547–555.
- Vu, T.H., Jain, H., Bucher, M., Cord, M., Pérez, P., 2019. Advent: Adversarial entropy minimization for domain adaptation in semantic segmentation, in: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 2517–2526.
- Wang, W., Lu, X., Shen, J., Crandall, D.J., Shao, L., 2019. Zero-shot video object segmentation via attentive graph neural networks, in: *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 9236–9245.
- Wang, X., Zhang, R., Shen, C., Kong, T., Li, L., 2021. Dense contrastive learning for self-supervised visual pre-training, in: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 3024–3033.

C Li, Qingyu, Yilei Shi, and Xiao Xiang Zhu. Semi-Supervised Building Footprint Generation with Feature and Output Consistency Training. IEEE Transactions on Geoscience and Remote Sensing, vol. 60, pp. 1-17, 2022, Art no. 5623217, doi: 10.1109/TGRS.2022.3174636

Semi-Supervised Building Footprint Generation with Feature and Output Consistency Training

Qingyu Li, *Student Member, IEEE*, Yilei Shi, *Member, IEEE*, and Xiao Xiang Zhu, *Fellow, IEEE*

Abstract—Accurate and reliable building footprint maps are vital to urban planning and monitoring, and most existing approaches fall back on convolutional neural networks (CNNs) for building footprint generation. However, one limitation of these methods is that they require strong supervisory information from massive annotated samples for network learning. State-of-the-art semi-supervised semantic segmentation networks with consistency training can help to deal with this issue by leveraging a large amount of unlabeled data, which encourages the consistency of model output on data perturbation. Considering that rich information is also encoded in feature maps, we propose to integrate the consistency of both features and outputs in the end-to-end network training of unlabeled samples, enabling to impose additional constraints. Prior semi-supervised semantic segmentation networks have established the cluster assumption, in which the decision boundary should lie in the vicinity of low sample density. In this work, we observe that for building footprint generation, the low-density regions are more apparent at the intermediate feature representations within the encoder than the encoder's input or output. Therefore, we propose an instruction to assign the perturbation to the intermediate feature representations within the encoder, which considers the spatial resolution of input remote sensing imagery and the mean size of individual buildings in the study area. The proposed method is evaluated on three datasets with different resolutions: Planet dataset (3 m/pixel), Massachusetts dataset (1 m/pixel), and Inria dataset (0.3 m/pixel). Experimental results show that the proposed approach can well extract more complete building structures and alleviate omission errors.

Index Terms—building footprint, semantic segmentation, semi-supervised, consistency training

This work is 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*), by the Helmholtz Association through the Framework of Helmholtz AI (grant number: ZT-I-PF-5-01) - Local Unit "Munich Unit @Aeronautics, Space and Transport (MASTr)" and Helmholtz Excellent Professorship "Data Science in Earth Observation - Big Data Fusion for Urban Research" (grant number: W2-W3-100), by the German Federal Ministry of Education and Research (BMBF) in the framework of the international future AI lab "AI4EO - Artificial Intelligence for Earth Observation: Reasoning, Uncertainties, Ethics and Beyond" (grant number: 01DD20001) and by German Federal Ministry of Economics and Technology in the framework of the "national center of excellence ML4Earth" (grant number: 50EE2201C). This work is also part of the project "Investigation of building cases using AI" funded by Bavarian State Ministry of Finance and Regional Identity (StMFH) and the Bavarian Agency for Digitization, High-Speed Internet and Surveying.

Corresponding author: Xiao Xiang Zhu.

Q. Li, and X.X. Zhu are with the Chair of Data Science in Earth Observation, Technische Universität München (TUM), 80333 Munich, Germany and the Remote Sensing Technology Institute (IMF), German Aerospace Center (DLR), 82234 Weßling, Germany (e-mails: qingyu.li@tum.de; xiaoxiang.zhu@dlr.de)

Y. Shi is with the Chair of Remote Sensing Technology, Technische Universität München (TUM), 80333 Munich, Germany (e-mail: yilei.shi@tum.de)

I. INTRODUCTION

Building footprint generation is a hot topic in the community of remote sensing, which involves numerous applications such as identifying undocumented buildings and assessing building damage after natural disasters. Remote sensing imagery that offers potential for meaningful geospatial target extraction on a large scale, becomes a fundamental data source for building footprint generation. However, obtaining accurate and reliable building footprint maps from remote sensing imagery is still challenging due to several reasons. On the one hand, the complex and heterogeneous appearance of buildings leads to internal variability. On the other hand, the mixed backgrounds and other objects with similar spectral signatures further limit the class separability.

Nowadays, convolutional neural networks (CNNs) have been widely used for remote sensing tasks [1] [2] [3], as they surpass conventional methods in terms of accuracy of efficiency. CNNs are capable of directly learning hierarchical contextual features from the original input, which have greater generalization capabilities for the building footprint generation from remote sensing imagery. Although the existing CNNs are able to deliver very promising results [2] [4] [5] [6], there remains a challenge for extracting building footprints on a large scale. This challenge arises from that CNNs require massive annotated data to obtain strong supervisory information. However, manual annotation of reference data is a time-consuming and costly process.

To address this issue, a straightforward idea is to utilize semi-supervised learning, which can leverage a large amount of unlabeled data and alleviate the need for labeled examples. In general, semi-supervised semantic segmentation methods are summarized into three types: weakly-supervised training-based, adversarial training-based, and consistency training-based. Nevertheless, weakly-supervised training-based methods need additional annotations, e.g. image-level labels or region-level labels. Adversarial training-based methods are able to make use of the unlabeled data but are difficult to train. Consistency training-based approaches, while not only are simple to implement, but also require no additional weakly labeled examples. The core idea of consistency training-based methods is to encourage the network to give consistent outputs for unlabeled inputs that are perturbed in various ways, thus, improving the generalization of the network [7].

The state-of-the-art consistency training-based methods exploit the teacher-student framework [8]. Specifically, a student model is applied to the unlabeled sample, while a teacher model is applied to a perturbed version of the same sample.

Afterward, the consistency is imposed between the outputs of two models to improve the performance of the student model [8]. However, there is still a certain gap in performance between these two models when the outputs are not completely correct during training. Inspired by [9] that feature maps can capture more discriminative contextual information, we further improve the performance of consistency training by proposing a new consistency loss that measures the discrepancy between both feature maps and outputs of student model and those of teacher model. By doing so, it can offer a strong constraint to regularize the learning of the network.

The effectiveness of consistency training-based approaches depends heavily on the behavior of the data distribution, i.e., the cluster assumption, where the classes must be separated by low-density regions. However, the low-density regions separating the classes are not within the inputs, which offers an explanation for why semi-supervised is a challenging problem for semantic segmentation [10]. [11] observes that for natural images low-density regions separating the classes are present at the encoder's output, thus, proposing to assign the perturbation at this position. However, for remote sensing imagery with low spatial resolution, we observe the presence of low-density regions separating the classes is within the intermediate feature representations in the encoder rather than the encoder's input or output. Motivated by this observation, in this work, we propose to enforce the consistency over the perturbation applied to feature representations at a certain depth within the encoder, where this depth should be in line with the spatial resolution of remote sensing imagery and the mean size of individual buildings in the study area.

Specifically, we consider a shared encoder and a main decoder that are trained together using the labeled examples. To leverage unlabeled data, we then consider an auxiliary decoder whose inputs are perturbed versions of the shared encoder's output. The consistency is imposed between outputs and feature maps of the main decoder and those of the auxiliary decoder. By doing so, the shared encoder's representation is enhanced by using the additional training signal extracted from the unlabeled data.

This work's contributions are threefold.

(1) We propose a semi-supervised network for building footprint generation, which has not been adequately addressed in the current literature. When the annotated samples are insufficient, the proposed method can leverage a large amount of unlabeled data to improve the performance of a model.

(2) Our proposed method integrates the consistency training of features and outputs into a unified objective function, which formulates an efficient end-to-end training framework. Compared with other competitors, our approach gains significant improvements.

(3) Observing that the low-density regions separating the classes are within the intermediate feature representations in the encoder, we propose an instruction, in which the perturbation is applied on the feature representations at a certain depth within the encoder according to the spatial resolution of input remote sensing imagery and the mean size of individual buildings in the study area.

The remainder of the paper is organized as follows. Related

work is reviewed in Section II. Section III details the proposed network for building footprint generation. The experiments are described in Section IV. Results and Discussions are provided in Section V and VI, respectively. Eventually, Section VII summarizes this work.

II. RELATED WORK

A. Building Footprint Generation

A tremendous amount of remote sensing imagery can be collected with recent technological advances, providing huge potential for mapping buildings. A variety of methods have been proposed to generate building footprints from remote sensing imagery.

Early studies can be categorized into four types: geometrical primitive-based, index-based, segmentation-based, and classification-based methods. The geometrical primitive-based methods [12] first extract geometric primitives (e.g., building edges and corners) and then group them to form building hypotheses. In the index-based methods [13], an index is designed to discriminate buildings from other objects. Afterward, buildings are extracted by selecting an empirical threshold. By utilizing over-segmentation algorithms, the segmentation-based methods [14] aims at partitioning an image into different segments, so-called superpixels, and identify those belonging to buildings. In the classification-based methods [15], spectral and/or spatial features of each pixel are taken as input of classifiers to differentiate building from other classes. Nonetheless, a general limitation of these methods is that they rely heavily on manually defined rules and handcrafted features, resulting in a decrease in accuracy and efficiency.

In the past few years, deep learning-based methods have shown remarkable performance on this task, as discriminative features from raw images can be automatically and adaptively learned. Early methods [16] [17] employ a patch-wise classification framework, and assign the label to each pixel according to the class of its enclosing patch. However, the large overlap among patches leads to redundant operation and low efficiency. Therefore, semantic segmentation networks that can efficiently perform pixel-wise segmentation, becomes more popular in the task of building footprint generation [18] [19] [20] [5] [6] [21] [22] [23] [24] [25]. The commonly used network architectures involve fully convolutional networks (FCNs) [26] and encoder-decoder based architectures (e.g., DeepLabv3+ [27] Efficient-UNet [28], FC-DenseNet [29]). In order to take the characteristics of buildings in remote sensing imagery into account, some methods (e.g., ESFNet [30], MA-FCN [31], HA U-Net [32], and Multi-task [33]) have made some specific adaptations to these network architectures, e.g., attention block and multi-scale feature aggregation. More recently, instance segmentation networks are exploited to delineate individual building instances in several novel studies [34] [35]. Instance segmentation networks can not only assign a semantic label to each pixel with the class of its enclosing object but also distinguish different instances. The commonly used instance segmentation architecture for this task is Mask R-CNN [36].

B. Semi-Supervised Semantic Segmentation

Deep learning methods require strong supervisory information for network training, however, the collection of large volumes of annotated data is time-consuming and costly. Especially for the task of semantic segmentation, the acquisition of pixel-level labels is more expensive and laborious. Therefore, semi-supervised learning is favored in this task, and it can leverage a large amount of unlabeled data to compensate for limited supervisory information. In general, semi-supervised semantic segmentation methods are summarized into three types: weakly-supervised training-based, adversarial training-based, and consistency training-based.

Weakly-supervised training-based methods [37] [38] [39] [40] integrate weakly-supervised learning in their approaches. Apart from the limited pixel-level labels, they still require weaker labels that can be regarded as supervisory information for network training. For the application of building footprint generation, weaker labels include image-level labels, bounding boxes, and point labels. The image-level label has two classes, where “building” refers to the images occupying building pixels more than a certain amount of the total pixels, and “non-building” corresponds to images without building pixels [41] [42]. In [43], bounding box annotations are utilized to generate probabilistic masks using bivariate Gaussian distribution for every image. Point labels (two points inside and outside each small building, respectively) are employed in [44], which is helpful to detect small buildings. Nevertheless, weakly-supervised training-based methods fail to take advantage of massive unlabeled data. Adversarial training-based methods [45] [46] are able to exploit unlabeled samples, which adapt generative adversarial networks (GANs) [47] for semi-supervised semantic segmentation. Both the generator and the discriminator are first trained by labeled samples. Afterward, the generator outputs the segmentation masks of unlabeled images, while the discriminator distinguishes trustworthy regions in their predicted results to provide additional supervisory signals. Considering that the adversarial training strategy may be insufficient to guide network training, pseudo labels are generated by selecting high-confident segmentation predictions for unlabeled images [48]. Afterward, pseudo-building masks are incorporated to expand the training data and the generator is retrained. However, adversarial training-based methods are very hard to train due to the instability of GANs [49]. By contrast, consistency training-based methods not only can leverage unlabeled images to improve the performance of the segmentation network but also are simple and efficient to implement. The goal of consistency training is to enforce the consistency of the model’s predictions for unlabeled inputs that are applied by small perturbations. By doing so, the robustness of the learned model will be enhanced.

Recently, several consistency training-based methods are proposed for the task of semi-supervised semantic segmentation, e.g., CutMix [10] and CCT [11]. CutMix [10] applies the perturbations to the raw input and uses MixUp [50] to enforce the consistency between the mixed outputs and the outputs from the mixed inputs. CCT [11] imposes an invariance of the model’s outputs over small perturbations applied to

the encoder’s output. In the remote sensing community, two consistency training-based methods have been proposed for the application of building footprint generation, i.e., CR [51] and PiCoCo [52]. Color jitter and random noise are chosen as the perturbation for CR [51], and are applied to the raw input. Then, the consistency of their outputs is enforced. PiCoCo [52] is also an input perturbation method, which augment the input images randomly and impose the consistency constraint between the predictions of augmented images. In addition, it implements contrast learning on labeled images, which can regularize the compactness of intra- and interclass latent representation space [52].

However, these consistency training-based methods still have two limitations. On the one hand, these methods ignore the rich information encoded in feature maps and generally impose consistency only over the outputs of the models. On the other hand, they add perturbations over the raw input or encoder’s output for all types of data, failing to take the characteristics of target objects into consideration when selecting the optimal position to apply perturbations.

III. METHODOLOGY

In this section, consistency training-based methods are first introduced. Afterward, the proposed framework in the end-to-end network learning procedure is described. Finally, we propose an instruction to assign perturbation for the task of building footprint generation, which is based on our observation and analysis of cluster assumption.

A. Consistency Training-based Methods

Given a small set of n input-target pairs $S_l = \{(x_1^l, y_1), \dots, (x_n^l, y_n)\}$ sampled from an unknown joint distribution $\beta(x, y)$, the goal of supervised learning is to derive a prediction function $f_\theta(x)$ parametrized by θ , and this prediction function is able to assign the correct target y to an unseen sample from $\beta(x)$. In semi-supervised learning, a larger set of m unlabeled examples $S_u = \{x_1^u, \dots, x_m^u\}$ is additionally provided. Semi-Supervised learning aims to derive a more accurate prediction function than what is obtained by only using S_l . For instance, additional structure about the input distribution $\beta(x)$ can be learned from S_u to produce a estimate of the decision boundary, which makes a better separation of samples into different classes [53].

Consistency training-based methods follow an intuitive goal to perform semi-supervised learning: when a perturbation is assigned to the data points $x \in S_u$ as \hat{x} , the output of $f_\theta(x)$ should not be significantly changed. Therefore, the objective of consistency training-based methods is to minimize the following loss function:

$$L = L_s + \lambda_u \cdot L_{cons}, \quad (1)$$

where L_s is a supervised loss on labeled data. λ_u is a weighting function to control the importance of a consistency loss term L_{cons} which is formalized as:

$$L_{cons} = \mathbf{T}(f_\theta(x), f_\theta(\hat{x})), \quad (2)$$

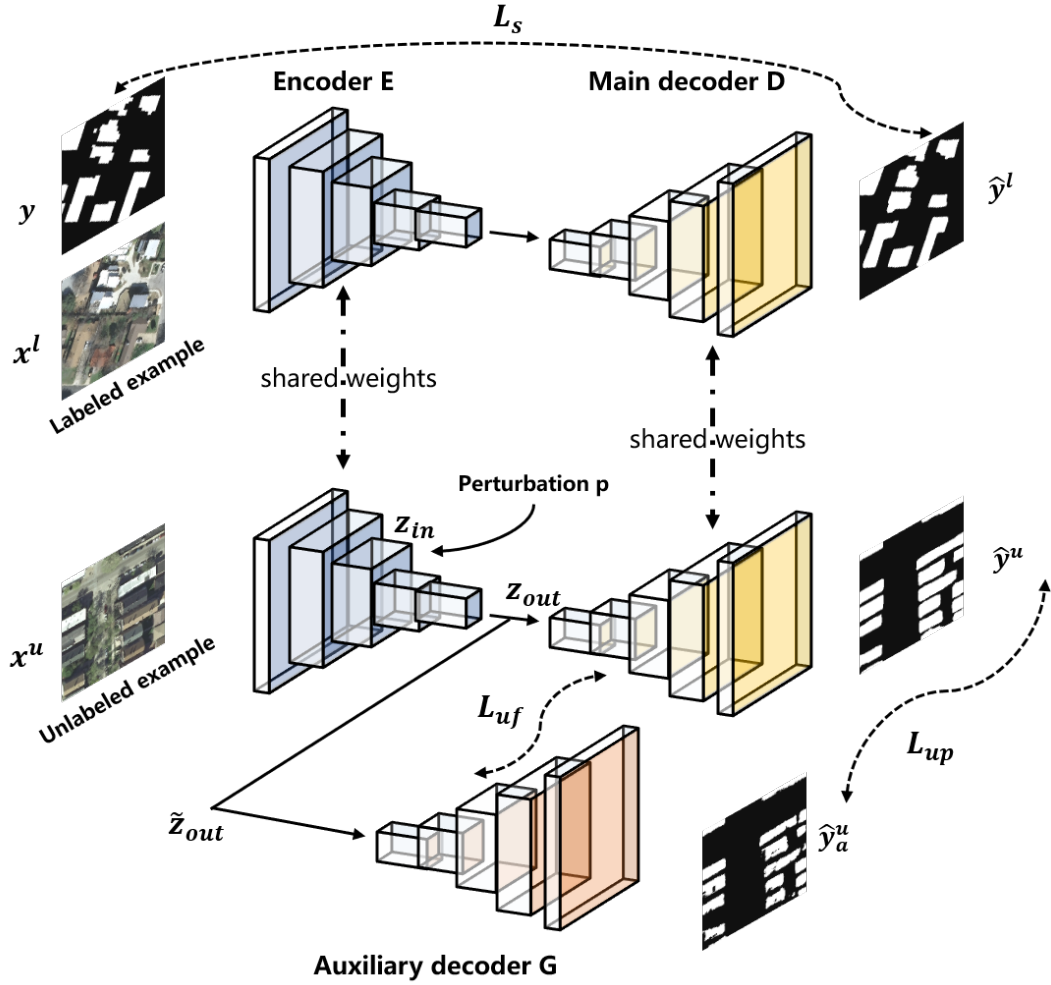


Fig. 1. Overview of the proposed semi-supervised building footprint generation network.

where $T(.,.)$ measures a discrepancy between the outputs of the prediction functions. In this regard, the unlabeled data can be leveraged to find a smooth manifold where the dataset lies [54].

Different settings in assigning perturbation or minimizing the L_{cons} lead to a wide variety of approaches for semi-supervised classification, e.g., Virtual Adversarial Training (VAT) [55] and Interpolation Consistency Training (ICT) [56], and those from semi-supervised semantic segmentation, e.g., CutMix[10], CCT [11], CR [51], and PiCoCo [52]. These methods are conducted in teacher-student frameworks, where a teacher model is first constructed from data perturbation, and then the output of the teacher model on unlabeled data is utilized to supervise a student model [8]. However, they have not fully leveraged the information of the teacher model. This is because they fail to use intermediate feature maps of the teacher model that can also be regarded as knowledge to guide the learning of the student model. Therefore, a more precise consistency towards the underlying invariance of features and outputs between the student model and the teacher model is preferable in our research.

B. Proposed Framework in End-to-End Network Learning

Recently, the perceptual mechanism has achieved promising results for image reconstruction [9], and they make use of the extracted high-level feature maps to improve the network performance. Inspired by it, we propose to impose consistency on both features and predictions for the training of unlabeled data, which is capable of fully harnessing information in deep features and output predictions. As a consequence, our network can guarantee that the deep feature maps are consistent, alleviating the loss of detailed information during network training.

As shown in Fig. 1, the proposed framework is composed of a shared encoder E , a main decoder D , and an auxiliary decoder G . The segmentation network F is constituted as $F = E \circ D$ and is trained on the labeled set in a fully supervised manner. The auxiliary network $A = E \circ G$ is trained on the unlabeled examples by enforcing the consistency of both features and outputs between D and G . D takes as input the encoder's output z_{out} , but G is fed with its perturbed version \tilde{z}_{out} , in which the perturbation p is applied to feature representations z_{in} at a certain depth within E .

By doing so, the representation learning of E can be further improved by unlabeled examples, and subsequently, that of the segmentation network F .

For each iteration of training, a labeled input image x^l and its label y together are sampled together with an unlabeled image x^u . Both x^l and x^u are passed through E and D , obtaining two main predictions \hat{y}^l and \hat{y}^u , respectively. The supervised loss L_s is computed with y and \hat{y}^l . For x^u , the perturbation p is applied to \mathbf{z}_{in} with \mathbf{z}_{in} being its feature representation within E and its output from E is $\tilde{\mathbf{z}}_{out}$. Afterward, an auxiliary prediction \hat{y}_a^u is generated from G using the $\tilde{\mathbf{z}}_{out}$. The consistency loss L_{cons} consists of two parts L_{uf} and L_{up} , where L_{uf} is computed between the features of G and those of D , and L_{up} is computed between the outputs of G and that of D .

In the proposed approach, S_l and S_u are jointly trained by minimizing a global loss function L as Eq. 1. Following [57], λ_u is set to ramp up starting from zero along a Gaussian curve up to a fixed weight α , which can avoid the use of the initial noisy output from the main encoder. The total loss L is derived and back-propagated to train the segmentation network F and the auxiliary network A . Note that L_{cons} is not backpropagated through D , and D is trained only by labeled examples. By doing so, D is only trained on original input data. This is helpful from two aspects. On the one hand, it can avoid collapsing solutions. If L_{cons} is backpropagated through both main decoder D and auxiliary decoder G , main decoder D will collapse since L_{cons} will be minimized if predictions of both D and G are zeros. On the other hand, the method can be better adapted to the test stage since no perturbation is applied to test images.

For the labeled set, a supervised loss L_s is exploited to train the segmentation network F . In order to avoid overfitting, an annealed version of the bootstrapped Cross-Entropy loss [11] is chosen to compute the supervised loss L_s , and it is denoted as:

$$L_s = \frac{1}{|S_l|} \sum_{x_i^l, y_i \in S_l} \{F(x_i^l) < \eta\} \mathbf{H}(y_i, F(x_i^l)), \quad (3)$$

where $F(x_i)$ is the output probability from F for a labeled example x_i , y_i is its ground reference label, and $\mathbf{H}(\cdot, \cdot)$ is the cross entropy-based loss. In semi-supervised learning, the model is often overfitted to the limited amount of labeled data while being under-fitted to the unlabeled data. To address this issue, a labeled example is utilized only if the model's confidence in it is lower than a predefined threshold η . In other words, L_s is computed only over the pixels with a probability less than the threshold η that serves as a ceiling to prevent over-training on easy labeled data [58]. Following [11], we gradually increase η from 0.5 to 0.9 during the beginning of training.

For an unlabeled example x_i^u , \mathbf{z}_{out} is derived as the output from the shared encoder E . One contribution in our approach is to apply the perturbation to the feature representation \mathbf{z}_{in} for x^u within the encoder E according to our proposed instruction. Afterward, the perturbed feature representations $\tilde{\mathbf{z}}_{in}$ will be fed to the subsequent layers in the encoder to generate the

perturbed encoder's output $\tilde{\mathbf{z}}_{out}$. Finally, \mathbf{z}_{out} and $\tilde{\mathbf{z}}_{out}$ are taken as input for D and G , respectively.

The training objective of the unlabeled set is to minimize a consistency loss L_{cons} , which is defined as:

$$L_{cons} = L_{up} + \omega_u \cdot L_{uf}, \quad (4)$$

where L_{uf} and L_{up} measure the discrepancy between the features and outputs of D and those of G , respectively. ω_u is a hyperparameter to introduce a weight to model the relative importance of two losses. More specifically, L_{up} is defined as:

$$L_{up} = \frac{1}{|S_u|} \sum_{x_i^u \in S_u} \mathbf{T}(D(\mathbf{z}_{out}), G(\tilde{\mathbf{z}}_{out})), \quad (5)$$

with $\mathbf{T}(\cdot, \cdot)$ as mean squared error-based loss.

Note that a contribution of our approach is that a loss term L_{uf} is introduced into the proposed network by imposing the consistency on features between the main decoder and auxiliary decoder, which is able to harness the detailed information in the feature maps. Let $\phi_j(q)$ be the activations of the j th layer of the network ϕ when processing the input q . For D and G , $D_j(\mathbf{z}_{out})$ and $G_j(\tilde{\mathbf{z}}_{out})$ will be the corresponding feature maps at j th depth in the decoder. Here, j represents the position where upsampling operations are applied in the decoder. Then, L_{uf} is denoted as:

$$L_{uf} = \frac{1}{|S_u|} \sum_{x_i^u \in S_u} \sum_{j=1}^J \mathbf{T}(D_j(\mathbf{z}_{out}), G_j(\tilde{\mathbf{z}}_{out})), \quad (6)$$

where J is the total number of depth in the decoder. In other words, J represents how many upsampling operations are applied in the decoder.

The proposed semi-supervised method can be summarized by the following Algorithm 1:

C. An Instruction to Assign Perturbation for the Task of Building Footprint Generation

The effectiveness of consistency training-based methods relies on the cluster assumption, i.e., two samples belonging to the same cluster in the input distribution are likely to have the same label [59]. In this case, the decision boundary should lie in the low-density regions [60]. In other words, if a decision boundary crosses a high-density region, it will divide a cluster into two different classes, which violates the cluster assumption. From the formal analysis, the expected value of L_{cons} is proportional to the squared magnitude of the Jacobian of the network's outputs with respect to its inputs [7]. Therefore, minimizing L_{cons} indicates that the decision function in the regions of unsupervised samples will be flattened, and the decision boundary will be moved into the vicinity of low sample density [10].

The cluster assumption has inspired many recent consistency training-based methods for semi-supervised semantic segmentation [10] [11] which propose to assign the perturbation to the raw input or encoder's output. However, they are not suitable for the task of building footprint generation, as the characteristics of both building objects and remote sensing imagery haven't been taken into account. Therefore,

Algorithm 1 Algorithm for Feature and Output Consistency Training

Input: Labeled image x^l and pixel-level label y , as well as unlabeled image x^u

Require: Shared encoder E , main decoder D with the total depth number J , and auxiliary decoder G

```

1: Forward  $x^l$  through  $E$  and  $D$ :  $\hat{y}^l = D(E(x^l))$ 
2: Forward  $x^u$  through  $E$ :  $\mathbf{z}_{out} = E(x^u)$ 
3: Generate the main decoder's feature maps for  $\mathbf{z}_{out}$ :
4: for  $j = 1$  to  $J$  do
    Derive  $D_j(\mathbf{z}_{out})$ 
5: end for
6: Generate the main decoder's output for  $\mathbf{z}_{out}$ :
    Derive  $D(\mathbf{z}_{out})$ 
7: Forward  $x^u$  through  $E$  and apply a noise perturbation  $\mathbf{N}$ 
   to feature representations  $\mathbf{z}_{in}$ :  $\tilde{\mathbf{z}}_{in} = (\mathbf{z}_{in} \odot \mathbf{N}) + \mathbf{z}_{in}$ 
8: Forward  $\tilde{\mathbf{z}}_{in}$  through the subsequent layers in  $E$  to generate
   the perturbed encoder's output  $\tilde{\mathbf{z}}_{out}$ 
9: Generate the auxiliary decoder's feature maps for  $\tilde{\mathbf{z}}_{out}$ :
10: for  $j = 1$  to  $J$  do
    Derive  $G_j(\tilde{\mathbf{z}}_{out})$ 
11: end for
12: Generate the auxiliary decoder's output for  $\tilde{\mathbf{z}}_{out}$ :
    Derive  $G(\tilde{\mathbf{z}}_{out})$ 
13: Training the network.
     $L_s = \{\hat{y}^l < \eta\} \mathbf{H}(y, \hat{y}^l)$ 
     $L_{up} = \mathbf{T}(D(\mathbf{z}_{out}), G(\tilde{\mathbf{z}}_{out}))$ 
     $L_{uf} = \sum_{j=1}^J \mathbf{T}(D_j(\mathbf{z}_{out}), G_j(\tilde{\mathbf{z}}_{out}))$ 
    Update network by  $L = L_s + \lambda_u \cdot (L_{up} + \omega_u \cdot L_{uf})$ 

```

we propose an instruction to assign perturbation for this task, which is inspired by the observation and analysis of the cluster assumption in building footprint generation from remote sensing imagery. In order to examine the cluster assumption, the local variations at an encoder depth d are measured between the value of each pixel and its local neighbors, and local variations with high values depict the presence of low-density regions [10]. Here, d represents the position where how many downsampling operations are applied in the encoder. For instance, when $d = 1$, the spatial size (i.e., height and width) of feature representation is half of that of the raw input. Similarly, when $d = 2$, the spatial size (i.e., height and width) of feature representation is 1/4 of that of the raw input. Following [11], the average Euclidean distance at each spatial location and its 8 intermediate neighbors is computed for the encoder's input ($d = 0$), and the feature representations of both intermediate layer ($d = 2$) and encoder's output ($d = 5$). Both feature representations are first resampled to the input size, and then the average distance between the neighboring activations is calculated. Fig. 2 illustrates the example results for Planet satellite imagery (3 m/pixel). The feature representations from intermediate layer and encoder's output are 24-dimensional and 1280-dimensional feature vectors learned from Efficient-UNet [28], respectively. It can be observed that the low-density regions are not aligned with the class boundaries at the encoder's input or encoder's output, where

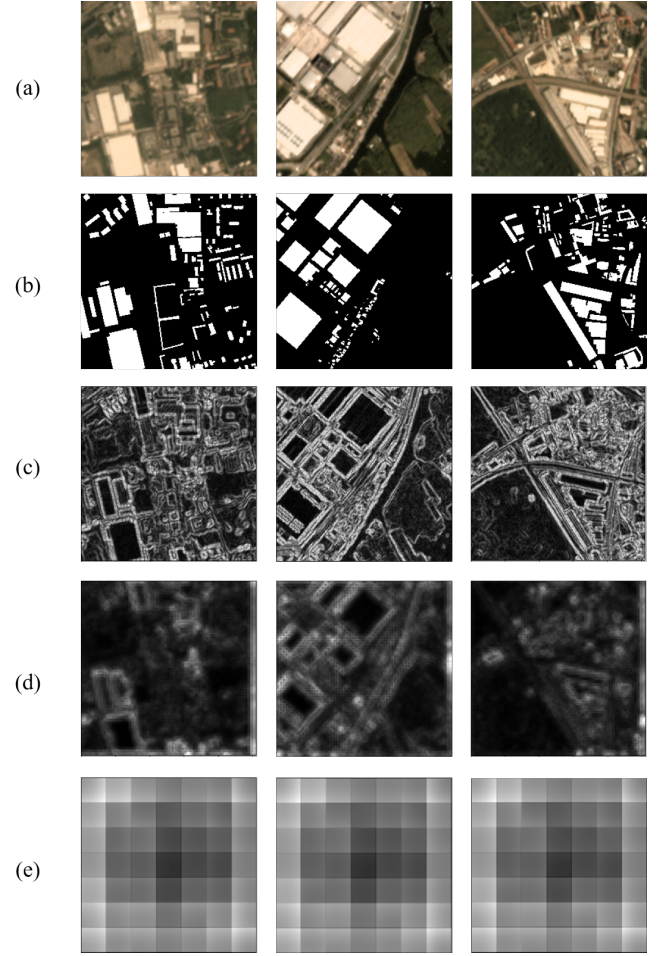


Fig. 2. The cluster assumption in consistency training-based methods for building footprint generation. Examples from (a) Planet satellite imagery (3m/pixel), (b) pixel-level labels, as well as local variations at (c) encoder's input, (d) intermediate layer in the encoder, and (e) encoder's output. Bright regions indicate large variation.

the cluster assumption is violated. By contrast, the cluster assumption is maintained at the intermediate layer, given that the class boundaries with high average distance coincide with low-density regions. This observation may be related to the receptive field of the network. The receptive field will be enlarged when the depth increases within the encoder, but when the receptive field exceeds a certain value that is much beyond the size of target objects, it might introduce more noise for network learning [61]. Furthermore, for remote sensing imagery with varying resolutions, the receptive fields of the network are various at the same depth within the encoder, when the unit is meter.

Based on the above observation and analysis, we propose an instruction to assign the perturbation. The perturbation should be added to the feature presentations at depth d within the encoder according to the spatial resolution of remote sensing imagery and the mean size of individual buildings in the study area. More specifically, d is computed as:



Fig. 3. The satellite imagery of Lisbon in the Planet dataset (spatial resolution: 3m/pixel) and three zoomed in areas.

$$d = \lfloor \log_2 \left(\frac{l_{min} + l_{max}}{2r} \right) \rfloor, \quad (7)$$

where r is the spatial resolution of the remote sensing imagery, l_{min} and l_{max} are mean values of max and min length that are derived from the ground reference of individual buildings in the study area. $\lfloor \cdot \rfloor$ is the rounding down function, which aims to get the largest integer that does not exceed the original value.

A noise tensor $\mathbf{N} \sim \mu(-0.3, 0.3)$ of the same size as the feature presentations \mathbf{z}_{in} is uniformly sampled as the perturbation p . It is first multiplied with \mathbf{z}_{in} to adjust its amplitude, and then injected into \mathbf{z}_{in} to get perturbed feature maps $\tilde{\mathbf{z}}_{in}$:

$$\tilde{\mathbf{z}}_{in} = (\mathbf{z}_{in} \odot \mathbf{N}) + \mathbf{z}_{in}, \quad (8)$$

where \odot denotes element-wise multiplication. Afterward, it will be fed to the subsequent layers in the encoder to generate the perturbed intermediate representation $\tilde{\mathbf{z}}_{out}$ of the unlabeled input sample x^u .

IV. EXPERIMENT

A. Dataset

The effectiveness of the proposed method is validated on three datasets with different spatial resolutions, i.e., Planet dataset [62], Massachusetts dataset [16], and Inria dataset [18].

1) Planet dataset: In this research, PlanetScope satellite imagery is collected from 8 European cities (Amsterdam, Berlin, Lisbon, Madrid, London, Paris, Milan, and Zurich) to create a Planet dataset. The PlanetScope satellite images have three bands (i.e., red, green, blue) at a spatial resolution of 3 m/pixel. The corresponding building footprints that are stored as vector files are acquired from OpenStreetMap. Fig. 3 presents example imagery of Lisbon.

2) Massachusetts dataset: The Massachusetts dataset is composed of 151 tiles of aerial imagery over the city of Boston. Each aerial imagery has three bands (i.e., red, green, blue) at a spatial resolution of 1 m/pixel, and its size is 1500×1500 pixels. A sample aerial image is illustrated in



Fig. 4. An aerial image in the Massachusetts dataset (spatial resolution: 1 m/pixel) and three zoomed in areas.

TABLE I
THE STATISTICS OF THE SELECTED DATASETS UTILIZED IN THIS RESEARCH.

Dataset	City	The number of patches		
		train	validation	test
Planet dataset	Amsterdam	4800	1600	2400
	Berlin			
	Lisbon			
	Madrid			
	London			
	Paris			
	Milan			
	Zurich			
Massachusetts dataset	Boston	3424	100	250
Inria dataset	Austin	39852	6044	6044
	Chicago			
	Kitsap County			
	Western Tyrol			
	Vienna			

Fig. 4. The corresponding ground reference building masks are also included in this benchmark dataset.

3) Inria dataset: The Inria dataset is a benchmark dataset consisting of 360 large-scale aerial images, in which each image is of the size of 5000×5000 and has three bands (i.e., red, green, blue) at a spatial resolution of 0.3 m/pixel. A sample aerial image is showed in Fig. 5. The ground reference building masks of this dataset are only publicly released for five cities (Austin, Chicago, Kitsap County, Western Tyrol, and Vienna).

For all three datasets, all remote sensing images and ground-truth building masks are cut into small patches with the size of 256×256 pixels. For the Planet dataset, we have

TABLE II
THE SETTINGS OF ALL METHODS UTILIZED IN THIS RESEARCH. λ_u AND ω_u REPRESENT THE WEIGHTS OF CONSISTENCY LOSS TERM AND FEATURE CONSISTENCY LOSS TERM, RESPECTIVELY.

Method	λ_u	ω_u	The position of the assigned perturbation
Supervised Learning (SL)	= 0	= 0	-
Supervised Learning + Data Augmentation (SL+DA)	= 0	= 0	-
ICT [56]	> 0	= 0	encoder's input
VAT [55]	> 0	= 0	encoder's input
CutMix[10]	> 0	= 0	encoder's input
CCT [11]	> 0	= 0	encoder's output
CR [51]	> 0	= 0	encoder's input
PiCoCo [52]	> 0	= 0	encoder's input
Proposed method	> 0	> 0	encoder's intermediate feature representations

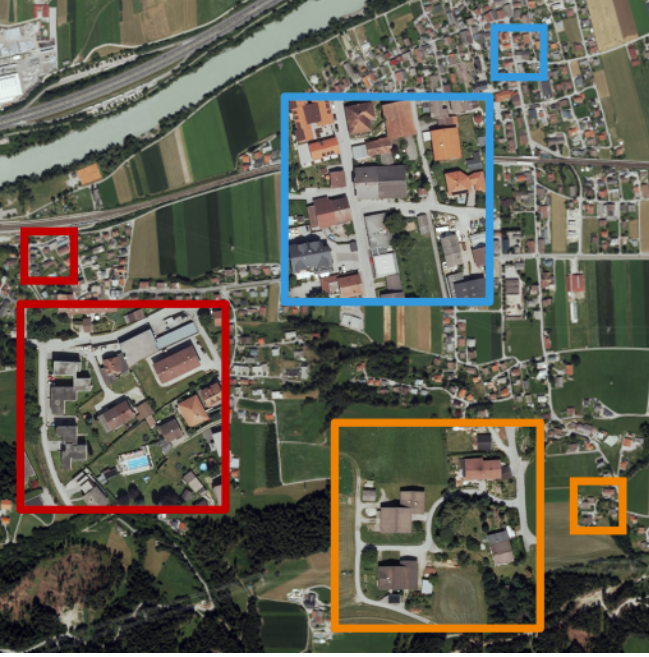


Fig. 5. An aerial image in the Inria dataset (spatial resolution: 0.3 m/pixel) and three zoomed in areas.

manually selected 1100 pairs of proper patches for each of eight European cities. The selected pairs are then separated into three parts, and the ratio of train, validation, and test set is 6:2:3. Data split in the Inria dataset is according to the setup in [18] [33]. More specifically, for each city, images with ids 1-5 are used for validation, and the remaining 31 images are for training. The statistics are derived from the validation set. The training/validation/test split of the Massachusetts dataset follows [16], where 137 tiles are used for training, 4 tiles are for validation, and the remaining 10 tiles are used to test models. The numbers of patches collected from each dataset for network training, validation, and test are reported in Table I.

B. Experiment Setup

Since the semantic segmentation network is an essential part of our approach, we first investigate which CNN model (i.e.,

Efficient-UNet [28], FC-DenseNet [29], DeepLabv3+ [27], ESFNet [30], MA-FCN [31], HA U-Net [32], and Multi-task [33]) has better performance for the task of building footprint generation. The CNN model achieving the best results under the fully supervised setting is selected as the backbone. Afterward, for each dataset, we randomly split the training data into two parts, which are labeled set and unlabeled set, and the pixel-level annotations are excluded in the unlabeled set. Under the semi-supervised setting, the ratios of labeled data to unlabeled data are set as three different ratios (e.g., 1:2, 1:5, 1:10). To validate the superiority of the proposed method, we make a comparison with other competitors, including Supervised Learning (SL), Supervised Learning + Data Augmentation (SL+DA), ICT [56], VAT [55], CutMix [10], CCT [11], CR [51] and PiCoCo [52]. The settings of λ_u , ω_u being the weights of consistency loss term and feature consistency loss term, and the position of the assigned perturbation in different methods are shown in Table II for a better understanding of their differences. Furthermore, the effectiveness of our proposed feature and output consistency, being imposed between the main decoder and the auxiliary decoder, is analyzed. The position within the encoder to apply perturbation is also carefully investigated for different datasets. Finally, we explore whether the auxiliary decoder is able to improve the performance of the proposed method.

C. Training Details

Our experiments are conducted within a Pytorch framework on an NVIDIA Tesla with 16 GB of memory. For all methods, the optimizer is stochastic gradient descent (SGD) with a learning rate of 0.1 and a momentum of 0.9, and the training batch size is set as 4. Detailed configurations of all methods included in our experiments are listed as follows:

(1) Efficient-UNet [28]: EfficientNet[63] is adopted as the encoder to learn feature maps. The decoder is comprised of five transposed convolutional layers that upsample the convolved image to predict segmentation masks.

(2) DeepLabv3+ [27]: The feature extractor in DeepLabv3+ is the Xception model [64].

(3) FC-Densenet [29]: Both the encoder and decoder in FC-DenseNet are composed of five dense blocks, and each dense block has five convolutional layers.

(4) ESFNet [30]: This method employs Separable Factorized Residual Block (SFRB) as the core module. The encoder is composed of 16 blocks, where 3 blocks are downsampling blocks and 13 blocks are SFRB. The decoder consists of 7 blocks for transposed convolutions and SFRB.

(5) MA-FCN [31]: This approach has proposed a feature fusion structure to aggregate multi-scale feature maps. It utilizes a Feature Pyramid Network (FPN) [65] -based structure as the backbone where the encoder is a four-layer VGG-16 [66] architecture and a corresponding decoder implements lateral connections between them.

(6) HA U-Net [32]: The encoder of this network adopts ResNet34 [67]. The decoder is comprised of four modules that include up-sampling module, attention module, overall nesting module, and auxiliary loss module.

(7) Multi-task [33]: This method is based on SegNet [68]. It first adds one convolutional layer after the decoder to learn the distance to the border of buildings. Afterward, this learned distance mask and feature maps produced by the decoder are concatenated and fed into another convolutional layer to learn the final building masks.

(8) Proposed method: The hyperparameter α in the unsupervised loss weighting function λ_u is set as 0.6. The loss term weighting parameter of feature consistency ω_u is chosen as 0.2. The network architectures of F and A are the same as that of the backbone.

(9) SL: The backbone is learned from labeled samples. Note that unlabeled samples are not considered during training.

(10) SL+DA: Following [69], data augmentation is first performed by randomly horizontally or vertically flipping, or rotating the image patches before training. Afterward, the backbone is trained on labeled samples.

(11) ICT [56] and VAT [55]: Following [10], we adapt these two semi-supervised classification methods for the task of semantic segmentation. The CNN model is the same as the backbone in our proposed method.

(12) CutMix [10], CCT [11], CR [51] and PiCoCo [52]: For a fair comparison, we replace the CNN model with the same backbone in our proposed method.

D. Evaluation Metrics

The performance of models is evaluated by two metrics: F1 score and intersection over union (IoU). They can be computed as follows.

$$\text{F1 score} = \frac{2 \times \text{precision} \times \text{recall}}{\text{precision} + \text{recall}}, \quad (9)$$

$$\text{IoU} = \frac{TP}{TP + FP + FN}, \quad (10)$$

$$\text{precision} = \frac{TP}{TP + FP}, \quad (11)$$

$$\text{recall} = \frac{TP}{TP + FN}, \quad (12)$$

where TP indicates the number of true positives, FN is the number of false negatives, and FP is the number of false positives. F1 score realizes a harmonic mean between precision and recall.

V. RESULTS

A. Results of Different Semantic Segmentation Networks for Supervised Learning

The comparisons among different semantic segmentation networks for supervised learning are presented in this section. Their respective performance is evaluated according to both quantitative (cf. Table III) and qualitative results (cf. Fig. 6, 7, and 8) on three datasets, respectively. The goal of this comparison is to select the best semantic segmentation network as the backbone for different learning methods in further experiments. In this case, we can avoid potential impacts due to convolutional layers and architectural differences.

Among these semantic segmentation networks, Efficient-UNet [28] performs better than DeepLabv3+ [27], FC-DenseNet [29], ESFNet [30], HA U-Net [32], and Multi-task [33] on all three datasets. Especially for the Planet dataset that has a relatively low spatial resolution, Efficient-UNet [28] obtains increments of 13.04% and 12.01% in F1 score and IoU when compared with DeepLabv3+ [27]. Although MA-FCN [31] is superior to Efficient-UNet [28] on the Massachusetts dataset, Efficient-UNet surpasses it by about 0.5% in IoU on both Planet and Inria datasets. Fig. 8 presents a visual comparison among different methods on three datasets. For the Inria dataset with relatively high spatial resolution, some non-building objects are wrongly identified as buildings by other methods. On the contrary, Efficient-UNet [28] is able to avoid such false alarms. The superiority of Efficient-UNet [28] on different resolution data can be attributed to its capability of systematically improving performance with all compound coefficients of the architecture (width, depth, and image resolution) balanced [28]. Thus, we take Efficient-UNet [28] as the backbone in both supervised learning and semi-supervised learning approaches for further comparisons.

B. Comparison with Other Competitors

Furthermore, we make comparisons among the proposed method, SL, SL+DA, ICT [56], VAT [55], CutMix [10], CCT [11], CR [51] and PiCoCo [52]. Here, the ratios of labeled data to unlabeled data are designed as 1:2, 1:5, and 1:10, respectively. SL is regarded as the baseline method that is only trained with labeled data, while SL+DA is trained on the labeled data that are already augmented. Labeled and unlabeled data are jointly trained for the proposed method, ICT [56], VAT [55], CutMix [10], CCT [11], CR [51] and PiCoCo [52]. Their performance is evaluated from quantitative (cf. Tables IV, V, and VI) perspectives. As an example, experiments are carried out for five runs on the Massachusetts dataset where the ratio of labeled data to unlabeled data is 1:2. This provides a fair comparison, and the corresponding F1 score and IoU are shown as mean and variance. Fig. 9, 10, and 11 illustrate visual results obtained by different methods for the ratio 1:10.

It can be seen from the statistics of three datasets that the proposed approach significantly boosts performance in F1 score and IoU when compared with other methods. The challenge induced by the ratio of 1:10 is the limited data representation for buildings, however, we notice that the

TABLE III
ACCURACIES OF DIFFERENT SEMANTIC SEGMENTATION NETWORKS FOR SUPERVISED LEARNING ON THREE DATASETS. (%)

Method	Planet dataset (3 m/pixel) 4800 labeled		Massachusetts dataset (1 m/pixel) 3424 labeled		INRIA dataset (0.3 m/pixel) 39852 labeled	
	F1 score	IoU	F1 score	IoU	F1 score	IoU
Efficient-UNet [28]	59.03	41.87	68.70	52.32	85.34	74.42
DeepLabv3+ [27]	45.99	29.86	65.96	49.21	80.67	67.61
FC-DenseNet [29]	55.78	38.68	68.17	51.87	84.66	73.41
ESFNet [30]	56.55	39.42	67.37	50.80	83.65	71.90
MA-FCN [31]	58.40	41.35	68.95	52.62	85.03	74.27
HA U-Net [32]	53.70	36.70	64.87	48.00	84.28	72.82
Multi-task [33]	48.05	31.62	65.43	48.63	84.56	73.26

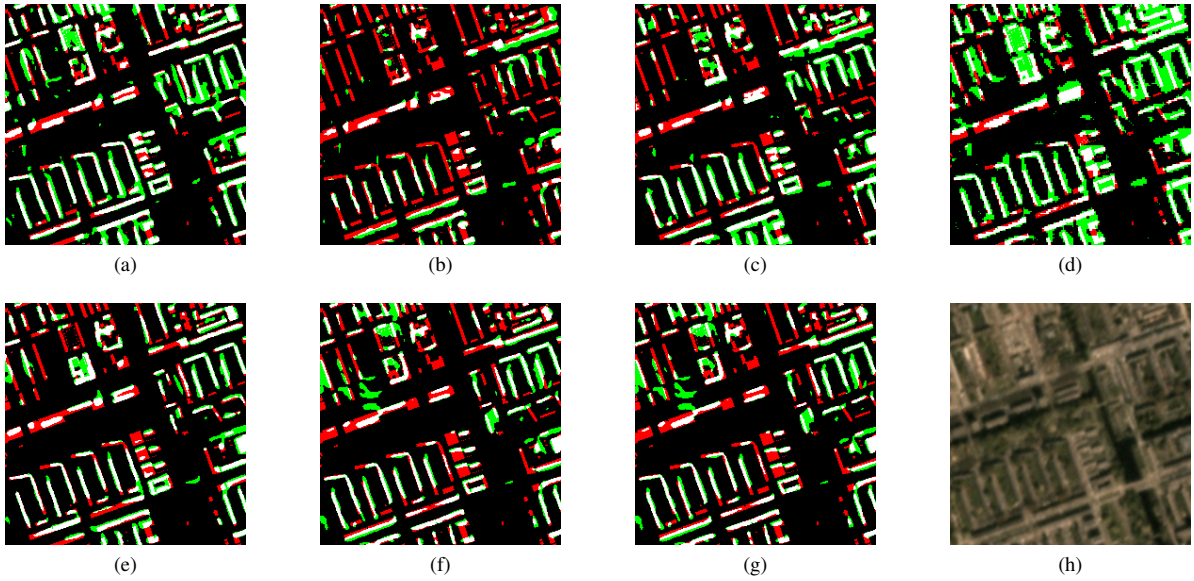


Fig. 6. Results obtained from (a) Efficient-UNet [28], (b) DeepLabv3+ [27], (c) FC-DenseNet [29], (d) ESFNet [30], (e) MA-FCN [31], (f) HA U-Net [32], and (g) Multi-task [33]. (h) is satellite imagery from the Planet dataset (spatial resolution: 3 m/pixel). Pixel-based true positives, false positives, and false negatives are marked in white, green, and red, respectively.

TABLE IV
ACCURACIES OF DIFFERENT METHODS ON PLANET DATASET (3 M/PIXEL). (%)

Method	labeled:unlabeled $\approx 1 : 2$ (1600 labeled, 3200 unlabeled)		labeled:unlabeled $\approx 1 : 5$ (800 labeled, 4000 unlabeled)		labeled:unlabeled $\approx 1 : 10$ (400 labeled, 4400 unlabeled)	
	F1 score	IoU	F1 score	IoU	F1 score	IoU
Proposed method	59.35	42.20	56.19	39.07	53.78	36.78
SL	53.15	36.19	51.80	34.95	48.03	31.60
SL + DA	53.84	36.83	52.87	35.93	48.53	32.04
ICT [56]	54.23	37.20	52.20	35.32	49.87	33.22
VAT [55]	36.25	22.13	34.25	20.67	33.77	20.32
CutMix [10]	54.10	37.08	52.43	35.53	49.86	33.21
CCT [11]	56.09	38.97	53.10	36.15	50.81	34.06
CR [51]	47.17	30.86	44.60	28.77	41.38	26.08
PiCoCo [52]	54.12	37.10	52.43	35.54	46.94	30.67

proposed method still manages to perform better on three datasets when compared to its competitors. Our method gains improvements of 5.18%, 10.40%, 7.91% in IoU than SL for the Planet, Massachusetts, and Inria datasets, respectively. In particular, on the Massachusetts dataset, the IoU of the proposed approach is improved by more than 7% when compared to other methods. When the ratio of labeled data to unlabeled data is 1:2, the number of labeled samples is already sufficient for SL, but our method still provides advantages over it. Note

that the proposed approach performs even better than the other semantic segmentation networks (cf. Table III) that are trained on the full labeled sets. This proves that the effectiveness and robustness of the proposed approach for the task of building footprint generation.

The accuracy metric of IoU obtained by our method for the ratio of 1:2 is higher than that for the ratio of 1:10. This suggests that using more labeled samples increases the overall performances (42.20% vs. 36.78% in the Planet dataset, 54.15

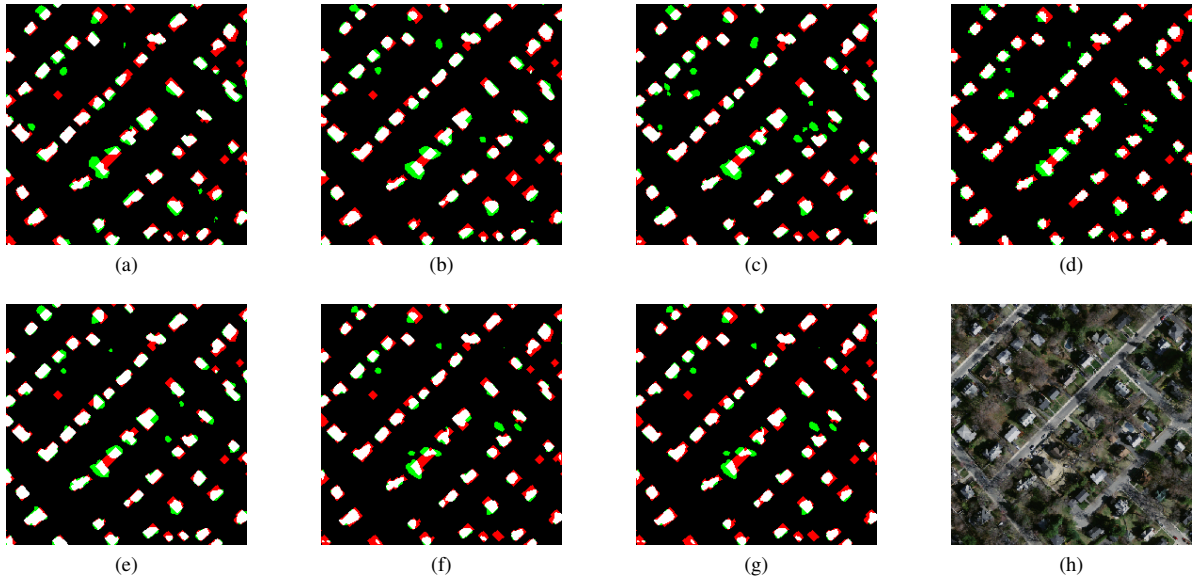


Fig. 7. Results obtained from (a) Efficient-UNet [28], (b) DeepLabv3+ [27], (c) FC-DenseNet [29], (d) ESFNet [30], (e) MA-FCN [31], (f) HA U-Net [32], and (g) Multi-task [33]. (h) is satellite imagery from the Massachusetts dataset (spatial resolution: 1 m/pixel). Pixel-based true positives, false positives, and false negatives are marked in white, green, and red, respectively.

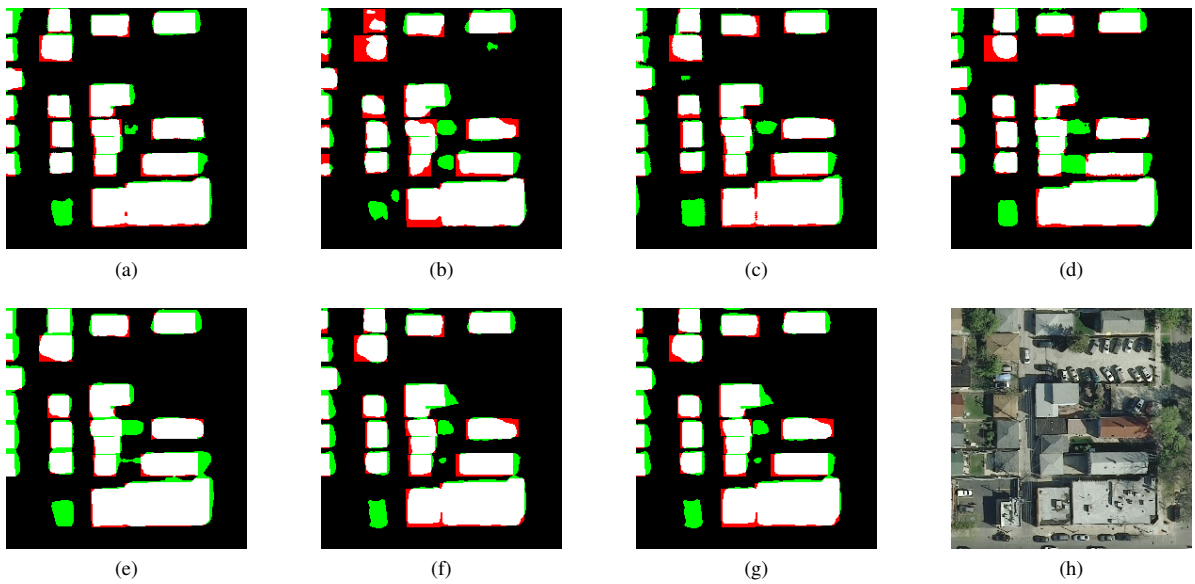


Fig. 8. Results obtained from (a) Efficient-UNet [28], (b) DeepLabv3+ [27], (c) FC-DenseNet [29], (d) ESFNet [30], (e) MA-FCN [31], (f) HA U-Net [32], and (g) Multi-task [33]. (h) is satellite imagery from the Inria dataset (spatial resolution: 0.3 m/pixel). Pixel-based true positives, false positives, and false negatives are marked in white, green, and red, respectively.

$\pm 0.68\%$ vs. 51.16% in the Massachusetts dataset, 75.22% vs. 72.03% in the Inria dataset). It should be mentioned that the proposed approach is capable of reducing the gap between the different ratios. For instance, Table V shows that the IoU produced by our method, which is trained on the data of ratio of 1:10, only drops 1% than that of ratio of 1:5. This demonstrates that our method can obtain reliable segmentation results even when there is only a small number of annotated samples.

The visual results on the Planet dataset are illustrated in Fig.

9. There is a lot of missed detection in results provided by SL, VAT [55], CCT [11], CR [51] and PiCoCo [52], as the number of labeled samples is insufficient. On the contrary, our method can extract more building structures. Fig. 11 presents results on the Inria dataset. It can be clearly seen that our method is able to avoid more false alarms than its competitors. This suggests that the proposed method has a better capability of utilizing unlabeled data to improve network performance.

TABLE V
ACCURACIES OF DIFFERENT METHODS ON MASSACHUSETTS DATASET (1 M/PIXEL). (%)

Method	labeled:unlabeled $\approx 1 : 2$ 1100 labeled, 2324 unlabeled		labeled:unlabeled $\approx 1 : 5$ 560 labeled, 2864 unlabeled		labeled:unlabeled $\approx 1 : 10$ 300 labeled, 3124 unlabeled	
	F1 score	IoU	F1 score	IoU	F1 score	IoU
Proposed method	70.26 \pm 0.57	54.15 \pm 0.68	68.59	52.20	67.69	51.16
SL	66.31 \pm 0.40	49.65 \pm 0.47	62.75	45.72	57.91	40.76
SL + DA	66.56 \pm 0.44	49.85 \pm 0.52	63.26	46.26	58.76	41.60
ICT [56]	67.03 \pm 0.19	50.42 \pm 0.23	63.33	46.34	60.19	43.05
VAT [55]	66.10 \pm 0.70	49.40 \pm 0.84	64.45	47.55	60.77	43.65
CutMix [10]	66.84 \pm 0.32	50.22 \pm 0.38	63.13	46.13	59.41	42.26
CCT [11]	67.79 \pm 0.33	51.30 \pm 0.39	64.54	47.64	60.70	43.58
CR [51]	65.83 \pm 0.55	51.08 \pm 0.66	63.91	46.96	60.68	43.56
PiCoCo [52]	68.76 \pm 0.52	52.39 \pm 0.62	65.73	48.96	64.76	47.88

TABLE VI
ACCURACIES OF DIFFERENT METHODS ON INRIA DATASET (0.3 M/PIXEL). (%)

Method	labeled:unlabeled $\approx 1 : 2$ 13000 labeled, 26852 unlabeled		labeled:unlabeled $\approx 1 : 5$ 6000 labeled, 33852 unlabeled		labeled:unlabeled $\approx 1 : 10$ 3600 labeled, 36252 unlabeled	
	F1 score	IoU	F1 score	IoU	F1 score	IoU
Proposed method	85.86	75.22	84.65	73.39	83.74	72.03
SL	81.94	69.41	80.38	67.61	77.87	64.12
SL + DA	82.40	70.07	81.01	68.08	78.56	64.69
ICT [56]	82.53	70.26	81.10	68.21	78.60	64.75
VAT [55]	82.79	70.63	81.42	68.66	78.48	64.58
CutMix [10]	82.89	70.77	81.34	68.55	78.59	64.74
CCT [11]	85.21	74.23	83.74	72.02	83.00	70.93
CR [51]	82.38	70.04	81.00	68.05	78.27	64.30
PiCoCo [52]	84.59	73.29	83.65	71.90	80.91	67.94

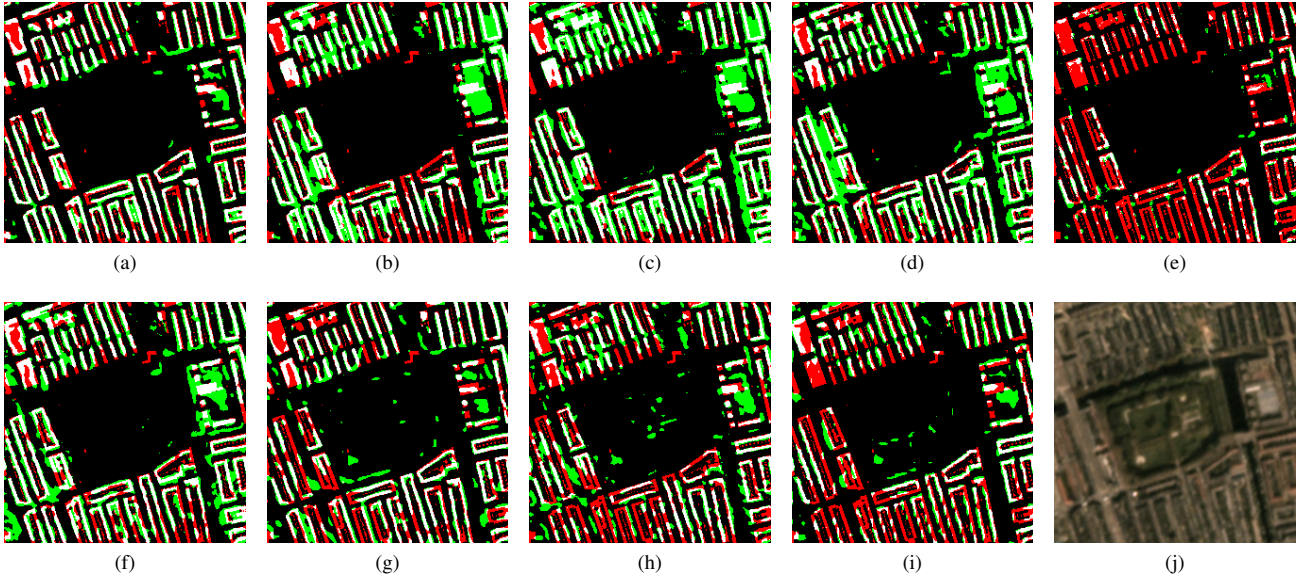


Fig. 9. Results obtained from (a) proposed method, (b) SL, (c) SL+DA, (d) ICT [56], (e) VAT [55], (f) CutMix [10], (g) CCT [11], (h) CR [51], and (i) PiCoCo [52]. In this experiment, the ratio of labeled data to unlabeled data is 1:10 (400 labeled, 4400 unlabeled). (j) is satellite imagery from the Planet dataset (spatial resolution: 3 m/pixel). Pixel-based true positives, false positives, and false negatives are marked in white, green, and red, respectively.

TABLE VII
ABLATION STUDY OF THE IMPOSED CONSISTENCY ON THREE DATASETS. (%)

The type of the imposed consistency	Planet dataset (3 m/pixel) 1600 labeled, 3200 unlabeled		Massachusetts dataset (1 m/pixel) 1100 labeled, 2324 unlabeled		Inria dataset (0.3 m/pixel) 13000 labeled, 26852 unlabeled	
	F1 score	IoU	F1 score	IoU	F1 score	IoU
Feature and output consistency	59.35	42.20	70.26	54.16	85.86	75.22
Output consistency	57.95	40.80	69.15	52.84	85.21	74.23

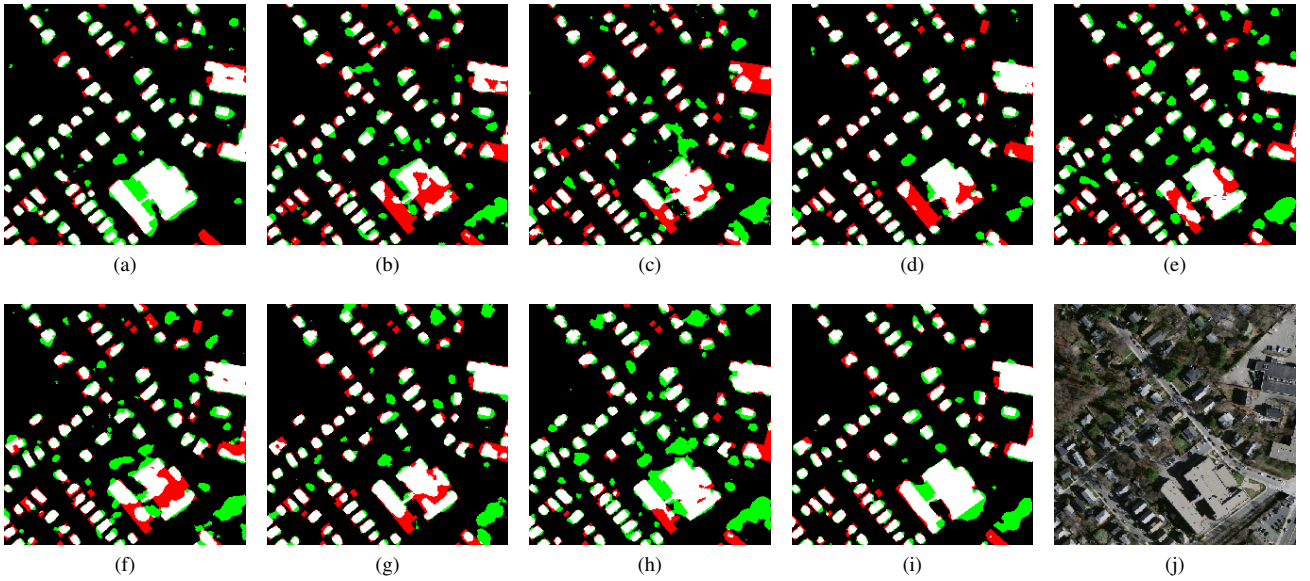


Fig. 10. Results obtained from (a) proposed method, (b) SL, (c) SL+DA, (d) ICT [56], (e) VAT [55], (f) CutMix [10], (g) CCT [11], (h) CR [51], and (i) PiCoCo [52]. In this experiment, the ratio of labeled data to unlabeled data is 1:10 (300 labeled, 3124 unlabeled). (j) is aerial imagery from the Massachusetts dataset (spatial resolution: 1 m/pixel). Pixel-based true positives, false positives, and false negatives are marked in white, green, and red, respectively.

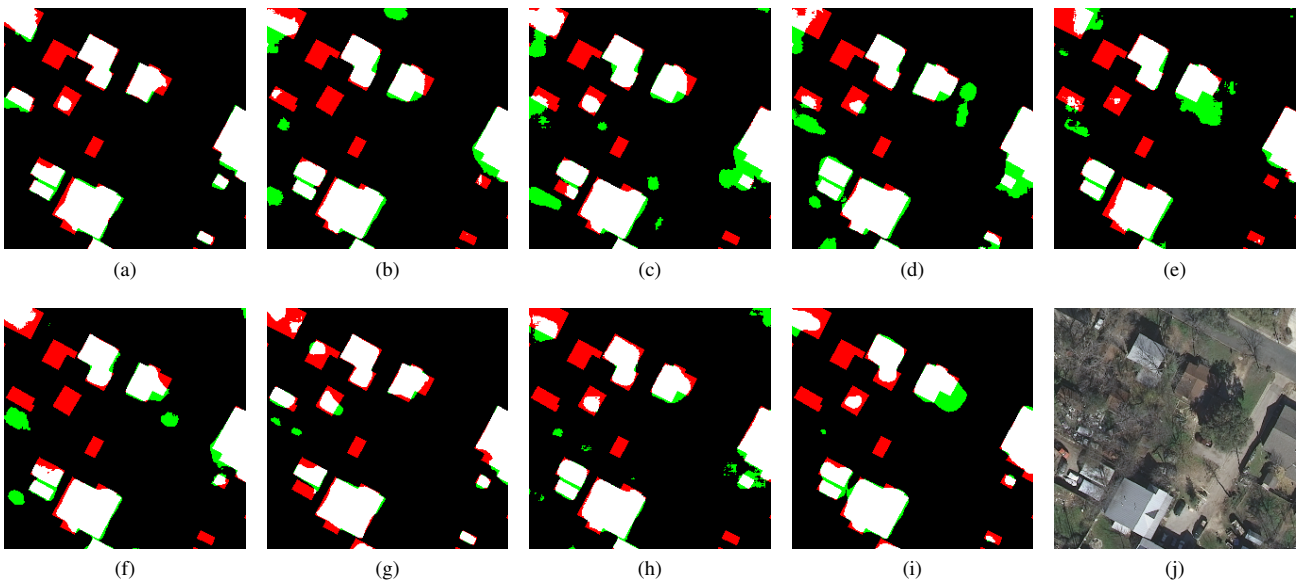


Fig. 11. Results obtained from (a) proposed method, (b) SL, (c) SL+DA, (d) ICT [56], (e) VAT [55], (f) CutMix [10], (g) CCT [11], (h) CR [51], and (i) PiCoCo [52]. In this experiment, the ratio of labeled data to unlabeled data is 1:10 (3600 labeled, 36252 unlabeled). (j) is aerial imagery from the Inria dataset (spatial resolution: 0.3 m/pixel). Pixel-based true positives, false positives, and false negatives are marked in white, green, and red, respectively.

VI. DISCUSSION

As shown in the results on three datasets for a semi-supervised setting, our proposed method with the ratio of 2:1 can deliver the best results. Therefore, in this section, we carry out ablation studies of the proposed method under this data split.

A. Ablation Study of the Imposed Consistency

One contribution of our approach worthy of being highlighted is that we introduce a novel objective function by

imposing consistency on both features and outputs between the main decoder and the auxiliary decoder.

The statistical results of different types of the imposed consistency are reported in Table VII. Experimental results show that implementing feature and output consistency for this task is helpful to improve the network performance, and we can see nearly 1% gains in IoU on all datasets when compared to solely output consistency. This may be because that more abstract and invariant information are included in the feature representations [70], and the network is able to learn more

TABLE VIII
ABLATION STUDY OF THE ASSIGNED PERTURBATION ON THREE DATASETS. (%)

The position of the assigned perturbation	Planet dataset (3 m/pixel) 1600 labeled, 3200 unlabeled		Massachusetts dataset (1 m/pixel) 1100 labeled, 2324 unlabeled		Inria dataset (0.3 m/pixel) 13000 labeled, 26852 unlabeled	
	F1 score	IoU	F1 score	IoU	F1 score	IoU
d=1	57.72	40.47	67.67	51.46	84.65	73.59
d=2	59.35	42.20	68.66	52.27	84.44	73.28
d=3	57.32	40.18	67.02	51.08	84.74	73.73
d=4	56.58	39.45	70.26	54.16	84.67	73.63
d=5	59.63	39.50	67.65	51.11	85.86	75.22

TABLE IX
ABLATION STUDY OF THE AUXILIARY DECODER ON THREE DATASETS. (%)

Method	Planet dataset (3 m/pixel) 1600 labeled, 3200 unlabeled		Massachusetts dataset (1 m/pixel) 1100 labeled, 2324 unlabeled		Inria dataset (0.3 m/pixel) 13000 labeled, 26852 unlabeled	
	F1 score	IoU	F1 score	IoU	F1 score	IoU
With auxiliary decoder	59.35	42.20	70.26	54.16	85.86	75.22
Without auxiliary decoder	58.37	41.20	68.73	52.34	84.84	73.67

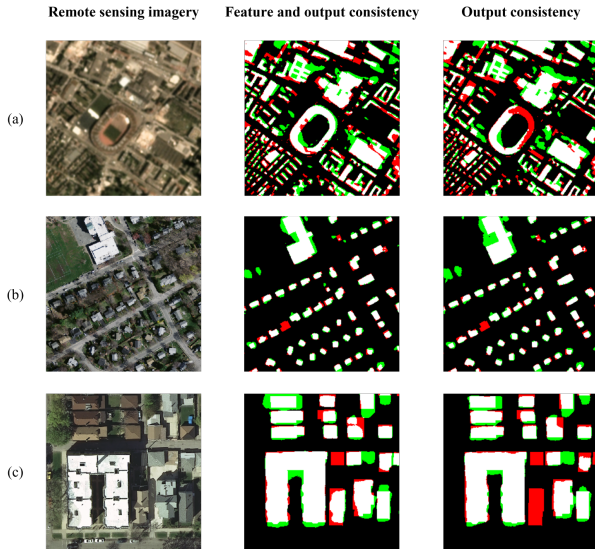


Fig. 12. Results obtained from different methods on (a) Planet dataset (spatial resolution: 3 m/pixel), (b) Massachusetts dataset (spatial resolution: 1 m/pixel), and (c) Inria dataset (spatial resolution: 0.3 m/pixel). Pixel-based true positives, false positives, and false negatives are marked in white, green, and red, respectively.

knowledge when feature consistency is additionally imposed.

Fig. 12 illustrates a visual comparison between different types of the imposed consistency. Some buildings are omitted in the results provided by sole output consistency in the example areas of the INRIA dataset. The reason is that the sole output consistency ignores the rich information in feature representations. On the contrary, building masks obtained by the feature and output consistency are much closer to real building shapes. This suggests that our method can capture information in both feature representations and outputs, enabling the enhancement of semantic information of buildings.

B. Ablation Study of the Assigned Perturbation

For the perturbation being assigned to the feature representations within the encoder, we propose an instruction to select the optimal position: the encoder depth d . To verify this instruction, we apply the perturbation to five different positions within the encoder, respectively. Specifically, d is first set as five numbers i.e., 1, 2, 3, 4, and 5, to investigate its impact on final results. The spatial size of their corresponding feature maps is 128×128 , 64×64 , 32×32 , 16×16 , 8×8 .

The statistical results of the perturbation applied to different depths within the encoder are shown in Table VIII. We can see that the best position to assign the perturbation is varied across different datasets. Moreover, increasing the value of the depth will promote the improvement of results on the higher resolution dataset (Inria dataset). However, we note that a large value of d will lead to a reduction in accuracy metrics on the relatively low-resolution dataset (Planet dataset). The best results are obtained when $d = 2$ for the Planet dataset, $d = 4$ for the Massachusetts dataset, and $d = 5$ for the Inria dataset. This coincides with our proposed instruction to apply the perturbation.

Taking the spatial resolution of remote sensing imagery into consideration, the respective field of these positions are corresponding to $3 \times 2^2 = 12m$ (Planet dataset), $1 \times 2^4 = 16m$ (Massachusetts dataset), $0.3 \times 2^5 = 9.6m$ (Inria dataset), which are close to the size of a building that usually has a length within the range from 10 m to 20 m. Afterward, we calculate the statistics of individual buildings of all three datasets, i.e., max length and min length (cf. Fig. 13). We found that the mean values of the max length of individual buildings are 19 m for the Planet dataset, 17 m for the Massachusetts dataset, 16 m for the Inria dataset. Mean values of the min length of individual buildings are 17 m for the Planet dataset, 14 m for the Massachusetts dataset, 12 m for the Inria dataset. That is to say, mean values of max length and min length of individual buildings also range from 10 m to 20 m among all datasets. This indicates the geometrical characteristics of the building are related to the effective receptive field of the network, which may place an emphasis on how to select the optimal position to

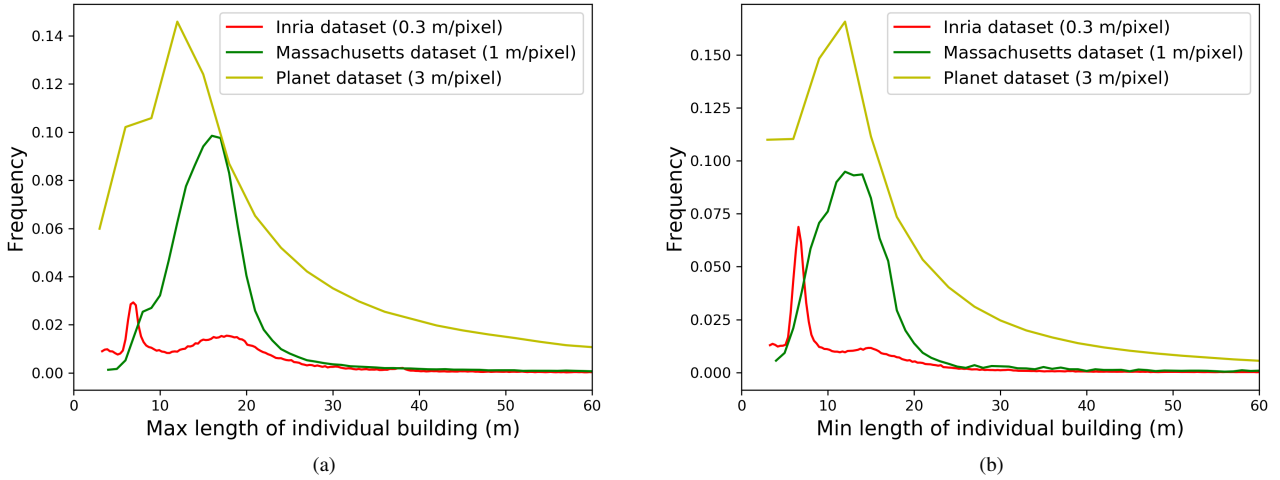


Fig. 13. Summarized statistics of (a) max length and (b) min length of individual buildings on three datasets.

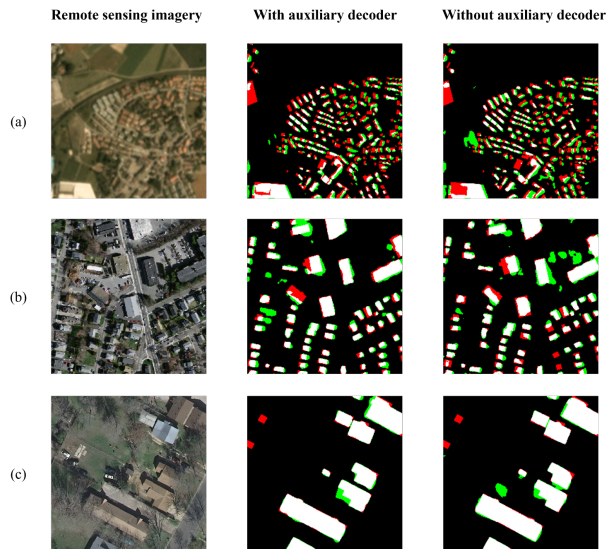


Fig. 14. Results obtained from different methods on (a) Planet dataset (spatial resolution: 3 m/pixel), (b) Massachusetts dataset (spatial resolution: 1 m/pixel), and (c) Inria dataset (spatial resolution: 0.3 m/pixel). Pixel-based true positives, false positives, and false negatives are marked in white, green, and red, respectively.

assign the perturbation in the whole framework. Therefore, we infer that the perturbation should be assigned to the different positions within the encoder according to the spatial resolution of remote sensing imagery and the mean size of the individual buildings in the study area.

C. Ablation Study of the Auxiliary Decoder

In our approach, an auxiliary decoder is employed to train the unlabeled set, and additional training signals can be extracted by enforcing the consistency of features and predictions

between the main decoder and the auxiliary decoders. In order to validate the effectiveness of the auxiliary decoder, we perform an ablation study with another competitor, i.e., the proposed method without auxiliary decoder. That is to say, the auxiliary decoder is removed, and the main decoder takes as input both an uncorrupted and perturbed version of the encoder's output to impose consistency on their features and outputs.

The ablation study is carried out on Planet, Massachusetts, and Inria datasets. Numerical results are shown in Tables IX. As can be seen in statistical results on all three datasets, an auxiliary decoder brings a nearly 1% improvement in IoU, leading to a positive influence on the performance of our network. Fig. 14 shows a visual comparison of segmentation results, which demonstrates that the performance of our approach can be boosted up by the leverage of an auxiliary decoder. In Fig. 14 (e) and (h), the method without auxiliary decoder wrongly identifies cars as buildings on both Massachusetts and Inria datasets. This is because, the colors of cars are similar to those of buildings, which leads to a misjudgment. The use of an auxiliary decoder is able to avoid such false alarms. The main reason is that supervision from the same decoder might guide the network to better approximate the features and outputs of the perturbed inputs, making the network converges in the wrong direction. In contrast, supervision by the features and predictions from the other decoder is able to avoid over-fitting the wrong direction.

VII. CONCLUSION

Considering that the performance of semantic segmentation networks is limited when the annotated training samples are insufficient, a novel semi-supervised building footprint generation method with feature and output consistency training is proposed in this paper. The proposed model comprises three modules: a shared encoder, a main decoder, and an auxiliary decoder. More specifically, the shared encoder and

the main decoder are designed to learn from labeled data in a fully supervised manner. Afterward, we assign the perturbation at the intermediate feature representations within the encoder and aims to encourage the auxiliary decoder to give consistent predictions for unlabeled inputs as the main decoder. The consistency is imposed between outputs and features of the main decoder and those of the auxiliary decoder. The performance of the proposed end-to-end network is assessed on three datasets with different resolutions: Planet dataset (3 m/pixel), Massachusetts dataset (1 m/pixel), and Inria dataset (0.3 m/pixel). Experimental results suggest that the incorporation of both feature and output consistency in our method can offer more satisfactory building footprints, where omission errors can be alleviated to a large extent. Therefore, We believe that our method is a robust solution for building footprint generation when dealing with scarce training samples. Furthermore, the best position to assign the perturbation has been investigated that the perturbation should be applied to the different depths within the encoder according to the spatial resolution of input remote sensing imagery and the mean size of the individual buildings in the study area. This practical strategy is beneficial to other semi-supervised building footprint generation works that use remote sensing imagery. A subsequent study will intend to investigate the potential of the feature and output consistency training in the instance segmentation of buildings.

REFERENCES

- [1] 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.
- [2] Y. Shi, Q. Li, and X. X. Zhu, "Building footprint generation using improved generative adversarial networks," *IEEE Geoscience and Remote Sensing Letters*, vol. 16, no. 4, pp. 603–607, 2018.
- [3] Y. Hua, L. Mou, and X. X. Zhu, "Relation network for multilabel aerial image classification," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 58, no. 7, pp. 4558–4572, 2020.
- [4] 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*, vol. 159, pp. 184–197, 2020.
- [5] 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*, 2020.
- [6] 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*, 2021.
- [7] B. Athiwaratkun, M. Finzi, P. Izmailov, and A. G. Wilson, "There are many consistent explanations of unlabeled data: Why you should average," *International Conference on Learning Representations*, 2018.
- [8] G.-J. Qi and J. Luo, "Small data challenges in big data era: A survey of recent progress on unsupervised and semi-supervised methods," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2020.
- [9] J. Johnson, A. Alahi, and L. Fei-Fei, "Perceptual losses for real-time style transfer and super-resolution," in *Proceedings of the European Conference on Computer Vision (ECCV)*. Springer, 2016, pp. 694–711.
- [10] G. French, S. Laine, T. Aila, M. Mackiewicz, and G. Finlayson, "Semi-supervised semantic segmentation needs strong, varied perturbations," in *British Machine Vision Conference*, no. 31, 2020.
- [11] Y. Ouali, C. Hudelot, and M. Tami, "Semi-supervised semantic segmentation with cross-consistency training," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages=12674–12684, year=2020.
- [12] X. Qin, S. He, X. Yang, M. Dehghan, Q. Qin, and J. Martin, "Accurate outline extraction of individual building from very high-resolution optical images," *IEEE Geoscience and Remote Sensing Letters*, vol. 15, no. 11, pp. 1775–1779, 2018.
- [13] X. Huang and L. Zhang, "A multidirectional and multiscale morphological index for automatic building extraction from multispectral geocye-1 imagery," *Photogrammetric Engineering & Remote Sensing*, vol. 77, no. 7, pp. 721–732, 2011.
- [14] A. O. Ok, "Automated detection of buildings from single vhr multispectral images using shadow information and graph cuts," *ISPRS Journal of Photogrammetry and Remote Sensing*, vol. 86, pp. 21–40, 2013.
- [15] M. Turker and D. Koc-San, "Building extraction from high-resolution optical spaceborne images using the integration of support vector machine (svm) classification, hough transformation and perceptual grouping," *International Journal of Applied Earth Observation and Geoinformation*, vol. 34, pp. 58–69, 2015.
- [16] V. Mnih, *Machine learning for aerial image labeling*. Citeseer, 2013.
- [17] R. Alshehhi, P. R. Marpu, W. L. Woon, and M. Dalla Mura, "Simultaneous extraction of roads and buildings in remote sensing imagery with convolutional neural networks," *ISPRS Journal of Photogrammetry and Remote Sensing*, vol. 130, pp. 139–149, 2017.
- [18] E. Maggiori, Y. Tarabalka, G. Charpiat, and P. Alliez, "Can semantic labeling methods generalize benchmark to any city? the inria aerial image labeling," in *IGARSS 2017-2017 IEEE International Geoscience and Remote Sensing Symposium*. IEEE, 2017.
- [19] X. Liu, Y. Chen, M. Wei, C. Wang, W. N. Gonçalves, J. Marcato, and J. Li, "Building instance extraction method based on improved hybrid task cascade," *IEEE Geoscience and Remote Sensing Letters*, vol. 19, pp. 1–5, 2021.
- [20] H. Huang, Y. Chen, and R. Wang, "A lightweight network for building extraction from remote sensing images," *IEEE Transactions on Geoscience and Remote Sensing*, 2021.
- [21] H. Guo, Q. Shi, B. Du, L. Zhang, D. Wang, and H. Ding, "Scene-driven multitask parallel attention network for building extraction in high-resolution remote sensing images," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 59, no. 5, pp. 4287–4306, 2020.
- [22] 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*, vol. 59, no. 7, pp. 6169–6181, 2020.
- [23] 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*, 2021.
- [24] K. Lee, J. H. Kim, H. Lee, J. Park, J. P. Choi, and J. Y. Hwang, "Boundary-oriented binary building segmentation model with two scheme learning for aerial images," *IEEE Transactions on Geoscience and Remote Sensing*, 2021.
- [25] 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*, vol. 60, pp. 1–17, 2021.
- [26] 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 (CVPR)*, 2015, pp. 3431–3440.
- [27] L.-C. Chen, Y. Zhu, G. Papandreou, F. Schroff, and H. Adam, "Encoder-decoder with atrous separable convolution for semantic image segmentation," in *Proceedings of the European Conference on Computer Vision (ECCV)*, 2018, pp. 801–818.
- [28] B. Baheti, S. Innani, S. Gajre, and S. Talbar, "Eff-unet: A novel architecture for semantic segmentation in unstructured environment," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR) Workshops*, 2020, pp. 358–359.
- [29] S. Jégou, M. Drozdal, D. Vazquez, A. Romero, and Y. Bengio, "The one hundred layers tiramisu: Fully convolutional densenets for semantic segmentation," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR) Workshops*, 2017, pp. 11–19.
- [30] J. Lin, W. Jing, H. Song, and G. Chen, "Esfnet: Efficient network for building extraction from high-resolution aerial images," *IEEE Access*, vol. 7, pp. 54 285–54 294, 2019.
- [31] S. Wei, S. Ji, and M. Lu, "Toward automatic building footprint delineation from aerial images using cnn and regularization," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 58, no. 3, pp. 2178–2189, 2019.
- [32] L. Xu, Y. Liu, P. Yang, H. Chen, H. Zhang, D. Wang, and X. Zhang, "Ha u-net: Improved model for building extraction from high resolution

- remote sensing imagery," *IEEE Access*, vol. 9, pp. 101 972–101 984, 2021.
- [33] B. Bischke, P. Helber, J. Folz, D. Borth, and A. Dengel, "Multi-task learning for segmentation of building footprints with deep neural networks," in *2019 IEEE International Conference on Image Processing (ICIP)*. IEEE, 2019, pp. 1480–1484.
- [34] Y. Liu, D. Chen, A. Ma, Y. Zhong, F. Fang, and K. Xu, "Multiscale u-shaped cnn building instance extraction framework with edge constraint for high-spatial-resolution remote sensing imagery," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 59, no. 7, pp. 6106–6120, 2020.
- [35] Q. Li, L. Mou, Y. Hua, Y. Sun, P. Jin, Y. Shi, and X. X. Zhu, "Instance segmentation of buildings using keypoints," in *IGARSS 2020-2020 IEEE International Geoscience and Remote Sensing Symposium*. IEEE, 2020, pp. 1452–1455.
- [36] K. He, G. Gkioxari, P. Dollár, and R. Girshick, "Mask r-cnn," in *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, 2017, pp. 2961–2969.
- [37] Y. Wei, J. Feng, X. Liang, M.-M. Cheng, Y. Zhao, and S. Yan, "Object region mining with adversarial erasing: A simple classification to semantic segmentation approach," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2017, pp. 1568–1576.
- [38] Y. Wei, H. Xiao, H. Shi, Z. Jie, J. Feng, and T. S. Huang, "Revisiting dilated convolution: A simple approach for weakly-and semi-supervised semantic segmentation," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2018, pp. 7268–7277.
- [39] J. Lee, E. Kim, S. Lee, J. Lee, and S. Yoon, "Ficklenet: Weakly and semi-supervised semantic image segmentation using stochastic inference," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2019, pp. 5267–5276.
- [40] Z. Huang, X. Wang, J. Wang, W. Liu, and J. Wang, "Weakly-supervised semantic segmentation network with deep seeded region growing," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2018, pp. 7014–7023.
- [41] J. Chen, F. He, Y. Zhang, G. Sun, and M. Deng, "Spmf-net: Weakly supervised building segmentation by combining superpixel pooling and multi-scale feature fusion," *Remote Sensing*, vol. 12, no. 6, p. 1049, 2020.
- [42] Z. Li, X. Zhang, P. Xiao, and Z. Zheng, "On the effectiveness of weakly supervised semantic segmentation for building extraction from high-resolution remote sensing imagery," *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, vol. 14, pp. 3266–3281, 2021.
- [43] M. U. Rafique and N. Jacobs, "Weakly supervised building segmentation from aerial images," in *IGARSS 2019-2019 IEEE International Geoscience and Remote Sensing Symposium*. IEEE, 2019, pp. 3955–3958.
- [44] J.-H. Lee, C. Kim, and S. Sull, "Weakly supervised segmentation of small buildings with point labels," in *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, 2021, pp. 7406–7415.
- [45] W.-C. Hung, Y.-H. Tsai, Y.-T. Liou, Y.-Y. Lin, and M.-H. Yang, "Adversarial learning for semi-supervised semantic segmentation," 2018.
- [46] N. Souly, C. Spampinato, and M. Shah, "Semi supervised semantic segmentation using generative adversarial network," in *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, 2017, pp. 5688–5696.
- [47] I. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville, and Y. Bengio, "Generative adversarial nets," *Advances in neural information processing systems*, vol. 27, 2014.
- [48] X. Yao, Y. Wang, Y. Wu, and Z. Liang, "Weakly-supervised domain adaptation with adversarial entropy for building segmentation in cross-domain aerial imagery," *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, vol. 14, pp. 8407–8418, 2021.
- [49] M. Arjovsky, S. Chintala, and L. Bottou, "Wasserstein generative adversarial networks," in *International Conference on Machine Learning (ICML)*. PMLR, 2017, pp. 214–223.
- [50] H. Zhang, M. Cisse, Y. N. Dauphin, and D. Lopez-Paz, "mixup: Beyond empirical risk minimization," in *International Conference on Learning Representations (ICLR)*, 2018.
- [51] J. Wang, C. HQ Ding, S. Chen, C. He, and B. Luo, "Semi-supervised remote sensing image semantic segmentation via consistency regularization and average update of pseudo-label," *Remote Sensing*, vol. 12, no. 21, p. 3603, 2020.
- [52] 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*, vol. 14, pp. 10 548–10 559, 2021.
- [53] A. Oliver, A. Odena, C. A. Raffel, E. D. Cubuk, and I. Goodfellow, "Realistic evaluation of deep semi-supervised learning algorithms," *Advances in neural information processing systems*, vol. 31, 2018.
- [54] M. Belkin, P. Niyogi, and V. Sindhwani, "Manifold regularization: A geometric framework for learning from labeled and unlabeled examples," *Journal of Machine Learning Research*, vol. 7, no. 11, 2006.
- [55] T. Miyato, S.-i. Maeda, M. Koyama, and S. Ishii, "Virtual adversarial training: a regularization method for supervised and semi-supervised learning," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 41, no. 8, pp. 1979–1993, 2018.
- [56] V. Verma, K. Kawaguchi, A. Lamb, J. Kannala, Y. Bengio, and D. Lopez-Paz, "Interpolation consistency training for semi-supervised learning," in *Proceedings of the International Joint Conference on Artificial Intelligence (IJCAI)*, 2019.
- [57] S. Laine and T. Aila, "Temporal ensembling for semi-supervised learning," *arXiv preprint arXiv:1610.02242*, 2016.
- [58] Q. Xie, Z. Dai, E. Hovy, T. Luong, and Q. Le, "Unsupervised data augmentation for consistency training," *Advances in Neural Information Processing Systems*, vol. 33, pp. 6256–6268, 2020.
- [59] O. Chapelle and A. Zien, "Semi-supervised classification by low density separation," in *AISTATS*, vol. 2005. Citeseer, 2005, pp. 57–64.
- [60] Y. Luo, J. Zhu, M. Li, Y. Ren, and B. Zhang, "Smooth neighbors on teacher graphs for semi-supervised learning," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2018, pp. 8896–8905.
- [61] W. Luo, Y. Li, R. Urtasun, and R. Zemel, "Understanding the effective receptive field in deep convolutional neural networks," *Advances in neural information processing systems*, vol. 29, 2016.
- [62] Planetscope. [Online]. Available: <https://www.planet.com>
- [63] M. Tan and Q. Le, "Efficientnet: Rethinking model scaling for convolutional neural networks," in *International Conference on Machine Learning*. PMLR, 2019, pp. 6105–6114.
- [64] F. Chollet, "Xception: Deep learning with depthwise separable convolutions," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2017, pp. 1251–1258.
- [65] T.-Y. Lin, P. Dollár, R. Girshick, K. He, B. Hariharan, and S. Belongie, "Feature pyramid networks for object detection," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2017, pp. 2117–2125.
- [66] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," in *International Conference on Learning Representations (ICLR)*, 2015.
- [67] 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 (CVPR)*, 2016, pp. 770–778.
- [68] 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*, vol. 39, no. 12, pp. 2481–2495, 2017.
- [69] Z. Huang, G. Cheng, H. Wang, H. Li, L. Shi, and C. Pan, "Building extraction from multi-source remote sensing images via deep deconvolution neural networks," in *IGARSS 2016-2016 IEEE International Geoscience and Remote Sensing Symposium*. IEEE, 2016, pp. 1835–1838.
- [70] Y. Bengio, A. Courville, and P. Vincent, "Representation learning: A review and new perspectives," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 35, no. 8, pp. 1798–1828, 2013.



Qingyu Li (S'21) received the bachelor's degree in remote sensing science and technology from the Wuhan University, Wuhan, China, in 2015, and the master's degree in Earth Oriented Space Science and Technology (ESPAC) from the Technische Universität München (TUM), Munich, Germany, in 2018.

She is currently pursuing the Ph.D. degree with the TUM, Munich, Germany and the German Aerospace Center (DLR), Weßling, Germany. Her research interests include deep learning, remote sensing mapping, and remote sensing applications.



Yilei Shi (M'18) received the Dipl.-Ing degree in mechanical engineering and Dr.-Ing degree in signal processing from the Technische Universität München (TUM), Munich, Germany, in 2010 and 2019, respectively. In April and May 2019, he was a guest scientist with the department of applied mathematics and theoretical physics, University of Cambridge, United Kingdom. He is currently a senior scientist with the Chair of Remote Sensing Technology, TUM.

His research interests include fast solver and parallel computing for large-scale problems, high performance computing and computational intelligence, advanced methods on SAR and InSAR processing, machine learning and deep learning for variety of data sources, such as SAR, optical images, and medical images, and PDE-related numerical modeling and computing.



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

She is currently the Professor for Data Science in Earth Observation (former: Signal Processing in Earth Observation) at Technical University of Munich (TUM) and the Head of the Department “EO Data Science” at the Remote Sensing Tech-




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

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

D Li, Qingyu, Yilei Shi, Stefan Auer, Robert Roschlaub, Karin Möst, Michael Schmitt, Clemens Glock, and Xiaoxiang Zhu. Detection of Undocumented Building Constructions from Official Geodata Using a Convolutional Neural Network. *Remote Sensing* 12, no. 21 (2020): 3537.

Article

Detection of Undocumented Building Constructions from Official Geodata Using a Convolutional Neural Network

Qingyu Li ^{1,2}, Yilei Shi ³, Stefan Auer ² , Robert Roschlaub ⁴, Karin Möst ⁴,
Michael Schmitt ^{1,5} , Clemens Glock ⁴ and Xiaoxiang Zhu ^{1,2,*} 

¹ Signal Processing in Earth Observation (Sipeo), Technical University of Munich (TUM), 80333 Munich, Germany; qingyu.li@tum.de (Q.L.); michael.schmitt@hm.edu (M.S.)

² Remote Sensing Technology Institute (IMF), German Aerospace Center (DLR), 82234 Wessling, Germany; stefan.auer@dlr.de

³ Remote Sensing Technology (LMF), Technical University of Munich (TUM), 80333 Munich, Germany; yilei.shi@tum.de

⁴ Bavarian Agency for Digitization, High-Speed Internet and Surveying (LDBV), 80538 Munich, Germany; robert.roschlaub@ldbv.bayern.de (R.R.); karin.moest@ldbv.bayern.de (K.M.); clemens.glock@ldbv.bayern.de (C.G.)

⁵ Department of Geoinformatics, Munich University of Applied Sciences, 80333 Munich, Germany

* Correspondence: xiaoxiang.zhu@dlr.de; Tel.: +49-(0)8153-28-3531

Received: 27 September 2020; Accepted: 23 October 2020; Published: 28 October 2020



Abstract: Undocumented building constructions are buildings or stories that were built years ago, but are missing in the official digital cadastral maps (DFK). The detection of undocumented building constructions is essential to urban planning and monitoring. The state of Bavaria, Germany, uses two semi-automatic detection methods for this task that suffer from a high false alarm rate. To solve this problem, we propose a novel framework to detect undocumented building constructions using a Convolutional Neural Network (CNN) and official geodata, including high resolution optical data and the Normalized Digital Surface Model (nDSM). More specifically, an undocumented building pixel is labeled as “building” by the CNN but does not overlap with a building polygon of the DFK. The class of old or new undocumented building can be further separated when a Temporal Digital Surface Model (tDSM) is introduced in the stage of decision fusion. In a further step, undocumented story construction is detected as the pixels that are “building” in both DFK and predicted results from CNN, but shows a height deviation from the tDSM. By doing so, we have produced a seamless map of undocumented building constructions for one-quarter of the state of Bavaria, Germany at a spatial resolution of 0.4 m, which has proved that our framework is robust to detect undocumented building constructions at large-scale. Considering that the official geodata exploited in this research is advantageous because of its high quality and large coverage, a transferability analysis experiment is also designed in our research to investigate the sampling strategies for building detection at large-scale. Our results indicate that building detection results in unseen areas at large-scale can be improved when training samples are collected from different districts. In an area where training samples are available, local training samples collection and training can save much time and effort.

Keywords: building detection; Convolutional Neural Network; deep learning; semantic segmentation; decision fusion

1. Introduction

The creation and maintenance of databases of buildings have numerous applications, which involve urban planning and monitoring as well as three-dimensional (3D) city modeling.

In particular, the complete documentation of buildings in official cadastral maps is essential to the transparent management of land properties, which can guarantee the legal and secure acquisition of properties. In Germany, the boundary of a building is acquired through a terrestrial survey by the official authority and then a two-dimensional (2D) ground plan of buildings is documented in the official cadastral map, which is known as the digital cadastral map (DFK).

However, due to the lack of information from owners about some building construction projects, some building constructions are never recorded via terrestrial surveying and are thus missing in the DFK. These building constructions are called undocumented building constructions, and include both undocumented buildings and undocumented story construction. Undocumented buildings have two types, old undocumented buildings and new undocumented buildings. Old undocumented buildings are buildings that were constructed many years ago but never recorded in the cadastral maps. New undocumented buildings are buildings that have only recently been erected. In this regard, the building ground plans of both old and new undocumented buildings are missing in the DFK. Both old and new undocumented buildings should be terrestrially surveyed by the official authority, but they may only charge the terrestrial survey fee for new undocumented buildings, due to Germany's regulations. In undocumented story construction, there are some changes on site, such as a newly built story or story demolition, that were not documented in the records of the official authority. Undocumented story construction will not lead to changes in the DFK, but this information is crucial to updating 3D building models. Therefore, collecting this undocumented building constructions is necessary to continue and complete these databases.

The technologies of airborne imaging and laser scanning show great potential in the task of building detection for nationwide 3D building model derivation [1,2]. The high resolution airborne data sets make detailed analysis of the geospatial targets more convenient and efficient. In the past, identifying undocumented buildings entailed a visual comparison of aerial images from different flying periods with DFK, enabling a comprehensive and timely interactive survey of changes in buildings. However, the visual interpretation of the aerial photos required a great amount of workforce and time.

In order to reduce the amount of work, two semi-automatic strategies are currently used by the state of Bavaria, Germany for the detection of undocumented buildings: the filter-based method [3] and the comparison-based method [4]. Both of these methods first detect buildings in remote sensing data. In the filter-based method, various filters, including a height filter, color filter, noise filter, and geometry filter, are applied to the data to detect the buildings. The comparison-based method detects all buildings with the aid of heuristically defined threshold values for the colors of buildings in the representative RGB color space and for the height in the Normalized Digital Surface Model (nDSM). Then both methods overlay the building detection results on the DFK to identify undocumented buildings. With the help of a Temporal Digital Surface Model (tDSM) derived from two Digital Surface Models (DSMs) in different epochs, new undocumented buildings can be discriminated from old undocumented buildings. Both methods are based on heuristic methods [3]. However, the heuristic definition of threshold values is not standardized, and have to be determined individually for different flight campaigns. Therefore, the data covering a large area cannot be processed in a uniform and standardized manner. Moreover, there are many false alarms in the results obtained from these two methods, where vegetation is frequently misclassified as buildings. For instance, the results of undocumented buildings obtained from the filter-based method also involve isolated vegetation (see Figure 1). In addition, these two methods do not provide any evidence of undocumented story construction.



■ undocumented buildings ■ buildings from DFK

Figure 1. Building detection results obtained from the filter-based method overlaid on the DFK (gray) to identify undocumented buildings (blue).

Recently, deep learning methods such as the Convolutional Neural Network (CNN) have been favored by the remote sensing community [5,6] in applications such as land cover classification [7,8], change detection [9,10], multi-label classification [11,12], and human settlement extraction [13,14]. CNN comprises multiple processing layers, which can learn hierarchical feature representations from the input without any prior knowledge. For the task of building detection from remote sensing data, CNN has also proven to achieve remarkable performances that far exceed those of traditional methods [15–17]. This is due to their superiority in generalization and accuracy without hand-crafted features. A key ingredient of CNN is training data. The amount of training data can be reduced if the pretrained transferable model is applicable in another unseen area [18], a property that is called transferability [19,20]. However, due to the limited size and quality of existing publicly available data sets, transferability cannot be well investigated in the task of building detection.

In this paper, our unique contributions are three-fold:

- (1) A new framework for the automatic detection of undocumented building constructions is proposed, which has integrated the state-of-the-art CNNs and fully harnessed official geodata. The proposed framework can identify old undocumented buildings, new undocumented buildings, and undocumented story construction according to their year and type of construction. Specifically, a CNN model is firstly exploited for the semantic segmentation of stacked nDSM and orthophoto with RGB bands (TrueDOP) data. Then, this derived binary map of “building” and “non-building” pixels is utilized to identify different types of undocumented building constructions through automatic comparison with the DFK and tDSM.
- (2) Our building detection results are compared with those obtained from two conventional solutions utilized in the state of Bavaria, Germany. With a large collection of reference data, this comparison has statistical sense. Our method can significantly reduce the false alarm rate, which has demonstrated the use of CNN for the robust detection of buildings at large-scale.
- (3) In order to offer insights for similar large-scale building detection tasks, we have investigated the transferability issue and sampling strategies further by using reference data of selected districts in the state of Bavaria, Germany and employing CNNs. It should be noted that this work is in an

advanced position to study the practical strategies for the task of large-scale building detection, as we implement such high quality and resolution official geodata at large-scale.

The remainder of the paper is organized as follows: Related work is reviewed in Section 2. The study area and official geodata utilized in this work are described in Section 3. Section 4 details the proposed framework for the detection of undocumented building constructions. The experiments are described in Section 5. The results and discussion are provided in Sections 6 and 7, respectively. Eventually, Section 8 summarizes this work.

2. Related Work

2.1. Two Conventional Strategies for the Detection of Undocumented Buildings

In the state of Bavaria, Germany, there are two conventional strategies utilized to detect undocumented buildings, the filter-based method [3] and the comparison-based method [4]. For both methods, the detection of undocumented buildings is carried out by first detecting all buildings in the remote sensing data and then identifying undocumented buildings within the DFK by overlaying the results with the DFK. Finally, the detected undocumented buildings are separated into two classes by introducing a tDSM, i.e., they are classified as old undocumented buildings and new undocumented buildings.

The filter-based method detects buildings from remote sensing data based on multiple filters, which include height, color, and geometric filters. Considering that buildings are elevated objects, a “height filter” is first applied in an nDSM, in order to remove all points with height less than an empirically determined threshold. Then, the second filter “color filter” takes the color values of the individual points into account. It is assumed that all pixels belonging to the class “building” are normally distributed in an individual color channels. Thus, the values of the individual color channel from the TrueDOP for each building are calculated to derive a confidence range for the buildings. If the color values of the examined pixel are beyond this confidence range, it will be removed. The Normalized Difference Vegetation Index (NDVI) is then calculated to remove vegetation. The third filter, the “noise filter”, is implemented by comparing its height with neighboring points in a defined area. This is a further separation of those vegetation points. The last filter, the “geometry filter”, recognizes buildings according to their area, the number of breakpoints, the ratio of area to circumference, and elongation (angularity).

In the comparison-based method, all buildings at present are delineated by setting heuristic threshold values based on color and height information. The building footprints from the DFK are first intersected with the TrueDOP to derive the training areas of buildings. Then, the RGB color values from the training areas are collected from the TrueDOP as a reference [4], where the frequency and distribution of the individual RGB combination are utilized in order to separate buildings from vegetation with an empirically chosen threshold. Finally, with the help of the nDSM, incorrect classifications between buildings and other objects such as streets are avoided by an empirically determined height threshold.

In order to minimize the incorrect detection of non-building cases that can be caused by the height noise of the nDSM or by vegetation, the filter-based method utilizes “color filters” and the comparison-based method exploits a RGB cube. However, aerial imaging is carried out with different airplanes and opposite trajectory directions at different times and with different lighting conditions, where the color channels for the same objects can also have varied values. The color values for each individual building are also largely dependent on the amount of current sunlight. Therefore, the confidence range or thresholds are not sufficient to identify buildings. For these two methods, buildings can only be identified through different heuristic thresholds for different districts, which is still not a fully automatic strategy. Furthermore, these two methods do not provide a more detailed type of undocumented building construction case—undocumented story construction.

2.2. Shallow Learning Methods for Building Detection

Building detection is a favored topic in the remote sensing community. Over the past decades, a large number of shallow learning methods have been proposed, which can be summarized into four general types [15]: (1) edge-based, (2) region-based, (3) index-based, and (4) classification-based methods.

The edge-based methods recognize the buildings based on geometric details of buildings. In [21], the edges of buildings are first detected using the edge operator, and then are grouped based on perceptual groupings to construct the boundary of the buildings. In the region-based methods, the region of buildings is identified based on image segmentation methods, using a two-level graph theory framework enhanced by shadow information [22]. The index-based methods indicate the presence of buildings by a number of proposed indices to depict the building features. The morphological building index (MBI) [23] is a building index that extracts buildings automatically, and describes the characteristics of buildings by using multiscale and multidirectional morphological operators. In the classification-based methods, buildings are extracted by feeding the spectral information and spatial features into a classifier to make a prediction. In [24], automatic recognition of buildings is achieved through a Support Vector Machine (SVM) classification of a great quantity of geometric image features.

The shallow learning methods have shown some good results in the task of building detection by combining different spectral, spatial, or auxiliary information or assuming building hypotheses. However, the prior information and hand-crafted features of shallow learning methods make it difficult to achieve generic, robust, and scalable building detection results at large-scale. Moreover, the optimization of parameters in the shallow learning-based methods also leads to inefficiency in processing.

2.3. Deep Learning Methods for Building Detection

Recently, the emergence of deep learning methods, which are based on artificial neural networks, have made strong contributions to the task of building detection. The use of multiple layers in the network allows the automatic learning of representations from raw data. Prior information is not required in deep learning methods for hand-crafted feature design, which indicates that deep learning methods can generalize well over large areas. CNNs are deep learning architectures, that are commonly used and have been exploited as a preferred framework for the task of building detection, as they have demonstrated more powerful generalization capability and better performance than traditional methods [25]. The task of building detection using CNNs is related to the task of semantic segmentation in computer vision, which aims at performing pixel-wise labelling in an image [26]. This indicates that a CNN can assign a class label to every pixel in the image. Different CNN architectures, such as fully convolutional networks (FCN) [27] and encoder-decoder based architectures (e.g., U-Net [28], SegNet [29] and others), are commonly used for the task of semantic segmentation, which outperform shallow learning approaches marginally [30].

FCN is a pioneer work for semantic segmentation that effectively converts popular classification CNN models to generate pixel-level prediction maps with the transposed convolutions. In [31], the spectral and height information from different data sets are combined as the input for FCN to generate building footprints. In addition to FCN, the encoder-decoder based architectures are another popular variant. Spatial resolution has been gradually reduced for highly efficient feature mapping in the encoder, while feature representations are recovered into a full-resolution segmentation map in the decoder. In U-Net, the skip connections, which links the encoder and the decoder, is beneficial to the preservation of the spatial details. Considering that the results of FCN-based methods are sensitive to the size of buildings, the U-Net structure implemented in [32] increases scale invariance of algorithms for the task of building detection. SegNet is another encoder-decoder based architecture, where the max-pooling indices from the encoders are transferred to the corresponding decoders. By reusing max-pooling indices, SegNet requires less memory than U-Net. In [25], SegNet is exploited

to produce the first seamless building footprint map of America at the spatial resolution of 1 m. Currently, FC-DenseNet [33] is a favoured method among different CNN architectures for the semantic segmentation of geospatial scenes, and is superior to many other networks in accuracy [17,34] due to its better feature extraction capability [16].

3. Study Area and Official GeoData

In our research, the study area covers one-quarter of the state of Bavaria, Germany (see Figure 2), which includes 16 districts: Ansbach, Bad Toelz, Deggendorf, Hemau, Kulmbach, Kronach, Landau, Landshut, Muenchen, Nuernburg, Regensburg, Rosenheim, Wasserburg, Schweinfurt, Weilheim, and Wolfratshausen. Bavaria is a federal state of Germany located in the southeast of the country. It is the state with the largest land area and the second most populous state in Germany. The 16 selected districts include both urban and rural areas, where different types of buildings are covered.

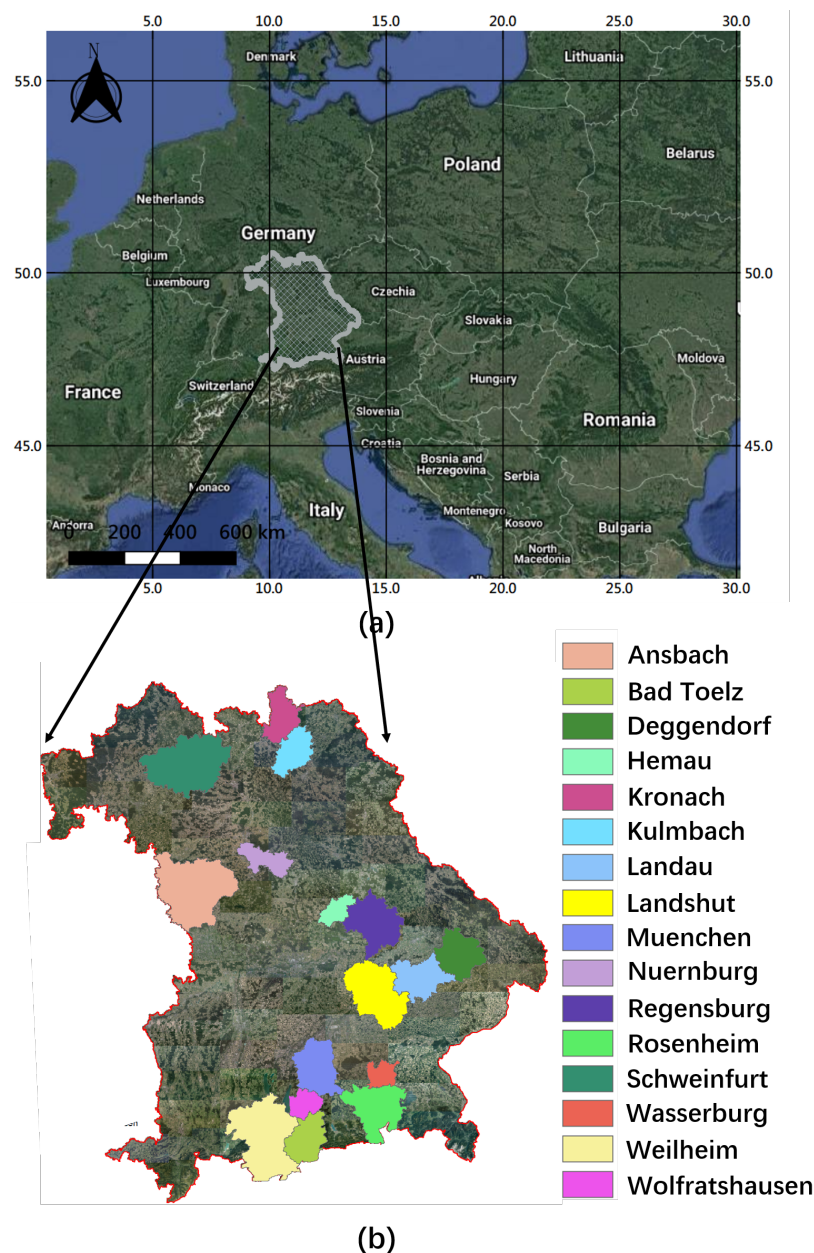


Figure 2. (a) The location of the state of Bavaria, Germany, (b) The study sites in this research, which cover 16 districts in the state of Bavaria, Germany.

Four types of official geodata are used in this study: nDSM, tDSM, TrueDOP, and DFK. The sample data sets are illustrated in Figure 3 and their related details are shown in Table 1. In the state of Bavaria, Germany, aerial flight campaigns are acquired through both aerial photographs and Airborne Laser Scanning (ALS). A regular point grid from ALS can be derived as the Digital Terrain Model (DTM). The DSM is obtained from a point cloud generated from optical data with the dense matching method [35]. The nDSM utilized in this research is a difference model between a current DSM at time point 2 (year 2017) and the DTM of the scene, which highlights elevated objects above the ground, such as buildings and trees. In this research, the tDSM is the difference model of two DSMs captured at two time points, i.e., time points 1 (year 2014) and 2 (year 2017). The TrueDOP is an orthophoto with RGB bands acquired in time point 2 (year 2017); ortho projection and geo-localization has been achieved corresponding to the DSM. Thus, all buildings and elevated objects in TrueDOP lie in position without geometric distortion. Each district is covered by a large number of tiles of TrueDOP, nDSM, and tDSM, where each tile has a size of 2500×2500 pixel at 0.4 m. The DFK is the cadastral 2D ground plan where the footprint of buildings is delineated. It is acquired via a terrestrial surveying in the field with accuracy in the range of cm. One of the limitations of a publicly available data set is the lack of high quality ground truth data [36], where inaccurate locations of building annotations lead to the misalignment between the building footprint and the data used for analysis [37]. It should be noted that, the DFK exploited as ground reference in our research is accurate: the buildings shown in a TrueDOP coincide the corresponding building footprint in the DFK.

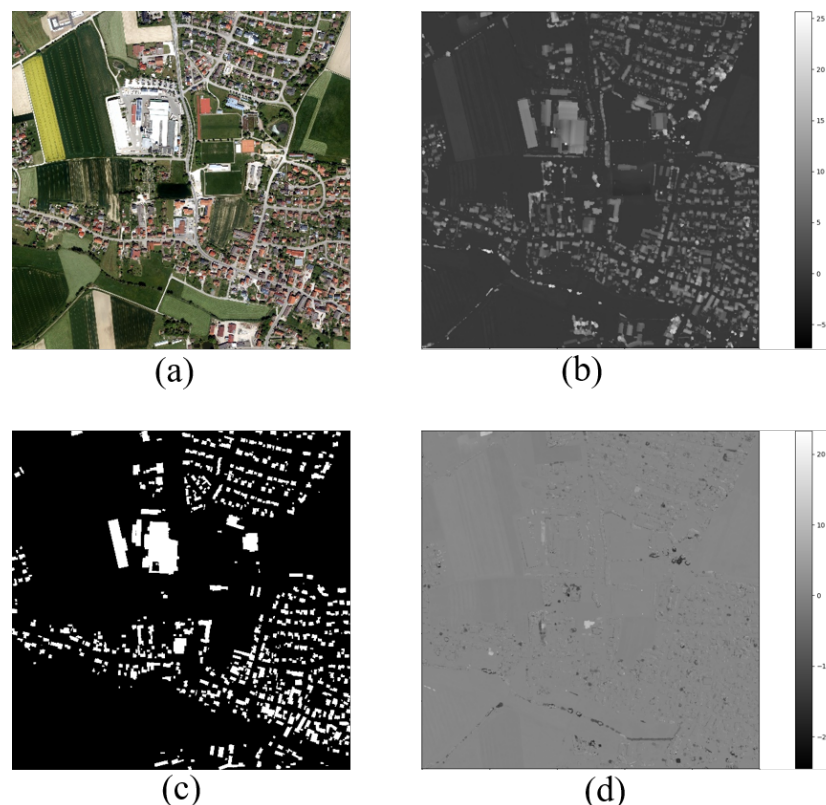


Figure 3. Sample data from (a) TrueDOP, (b) nDSM, (c) rasterized DFK, and (d) tDSM.

Table 1. Detailed information of data sets utilized in this research.

Data Set	Temporal Information	Spatial Resolution	Size	Channels
Normalized Digital Surface Model (nDSM)	year 2017	0.4 m	2500×2500	1
Temporal Digital Surface Model (tDSM)	from year 2014 to year 2017	0.4 m	2500×2500	1
Orthophoto with RGB bands (TrueDOP)	year 2017	0.4 m	2500×2500	3
Digital Cadastral Map (DFK)	year 2017	0.4 m	2500×2500	1

4. Methodology

4.1. The Proposed Framework for the Detection of Undocumented Building Constructions

Undocumented building constructions comprise two cases: undocumented buildings and undocumented story construction. Undocumented buildings are the buildings that exist in airborne survey data (nDSM and TrueDOP), but are not recorded in the cadastral 2D ground plan (DFK). Undocumented story construction represents buildings that exist in both airborne survey data (nDSM and TrueDOP) and the cadastral 2D ground plan (DFK), but show a signal of height deviation in the tDSM due to story buildup or demolition. We propose a framework to detect undocumented building constructions that is able to identify both undocumented buildings and undocumented story construction. This proposed framework is carried out based on CNN and decision fusion, and can be implemented as a routine strategy in large-scale object detection works.

An overview of the proposed framework is illustrated in Figure 4. The framework proposed in this study consists of three main tasks: (1) detection of undocumented buildings, (2) discrimination between old and new undocumented buildings, and (3) detection of undocumented story construction.

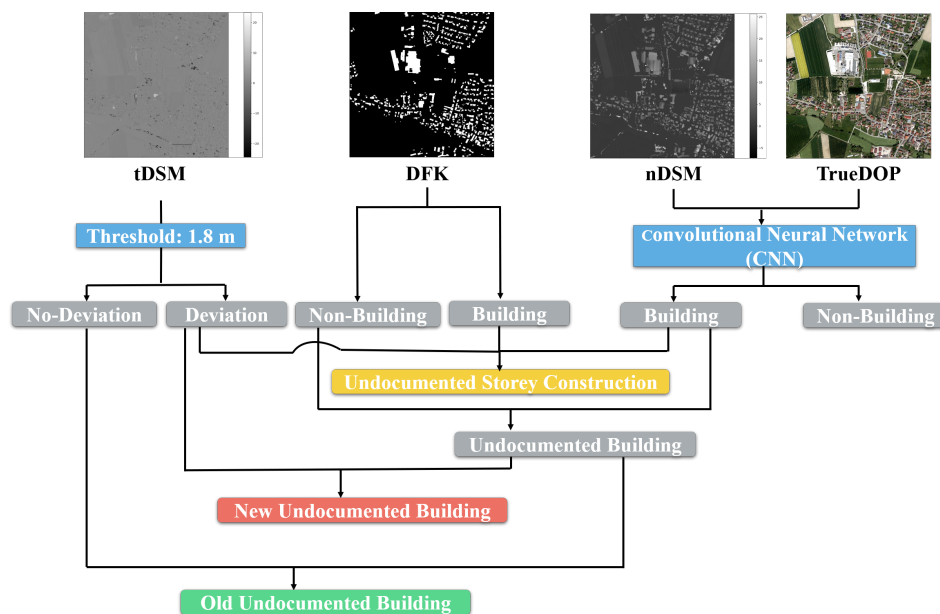


Figure 4. Flowchart of the proposed approach for the detection of undocumented building constructions.

In the proposed framework, TrueDOP stacked with the nDSM are utilized as the two main data sources in the first stage, building detection. These were chosen because that individual data sources may lead to biased building detection results. In the TrueDOP, the buildings share very similar spectral and texture characteristics with other areas, such as sidewalks. Moreover, varied light intensities due to atmospheric and seasonal effects, as well as shadow, can result in the variation in the appearance of buildings [38], which is largely dependent on the time of data acquisition. The nDSM data derived from the DSM and ALS data can directly inference the scene geometry, avoiding the influence of environmental variables. However, some issues emerge when relying solely on the nDSM, including marked occluded surfaces and planar surfaces that are split up [36]. In this case, buildings and other elevated objects above the ground can not be discriminated well by purely nDSM methods. Therefore, in order to make full use of both data sets, we stack the TrueDOP and the nDSM as input of the CNN model, which assigns the class label “building” or “non-building” to each pixel. The undocumented building pixels can then be identified when we overlay the predicted results with DFK, highlighting those pixels that are assigned the class label of “building” from the CNN model but belongs to the “non-building” class in the DFK.

In order to further distinguish between different types of undocumented buildings, the temporal information is essential to identifying the time window of the constructions. In this regard, the tDSM, which is the difference between two DSMs acquired at two time points, is introduced as an additional source of information. New constructions can be identified with an empiric value (1.8 m) applied to the tDSM, which indicates that there is a height deviation for this pixel within the period between two time points (from year 2014 to year 2017 in this research). This is due to the fact that a story or a building is usually higher than 1.8 m. If there is a height deviation within this period, the obtained undocumented building pixels from the previous stage will be assigned to the class as new undocumented building. It indicates that this undocumented building was constructed after time point 1 (year 2014). Otherwise it will be assigned the class of old undocumented building, which indicates that there was an undocumented building constructed before time point 1 (in this case, the year 2014).

Another case of building construction that can lead to a height deviation in two DSMs, is the undocumented story construction, which refers to story buildup or demolition on an existing building. The predicted results from the CNN model are first overlaid with the DFK. When the pixel in both data sources corresponds to the class “building” and if there is a height deviation identified in the tDSM, this pixel is placed in the class of undocumented story construction.

4.2. A CNN Model for Building Detection

Considering that the spatial resolution of airborne data is relatively high, massive quantities of data can be collected within the area of one-quarter of the state of Bavaria, Germany. CNNs, the most favored methods for many large-scale tasks [39], are therefore implemented as the most essential part of our proposed framework. FC-DenseNet is exploited as the base semantic segmentation network for building detection in the proposed framework, the goal of which is to assign the class label of “building” or “non-building” to each pixel.

Network Architecture

FC-DenseNet is also an encoder-decoder architecture, where the key ingredient is the DenseNet block. DenseNet [40] is a network that has proven to achieve superior performance for scene classification tasks [41]. In this regard, FC-DenseNet (see Figure 5) is proposed in [33], where the DenseNet is extended to a fully convolutional network for semantic segmentation tasks. The DenseNet block has introduced a new connective pattern between layers, where the input of each layer is all preceding features, and the output features from this layer are then transferred to all subsequent layers. Instead of ResNet [42], which combines features by summation, DenseNet combines features using iterative concatenation. This provides a more efficient flow of information through the network. The feature concatenation in the DenseNet block reuses all features, which makes the connections within layers shorter. In this regard, the intermediate layers will be enforced to learn distinguished feature maps for easier training. Another important design element of FC-DenseNet is the skip connections [43] between the encoder and the decoder, where higher resolution information can be passed. The spatial details can be well recovered in the decoder from the encoder with the help of the skip connection.

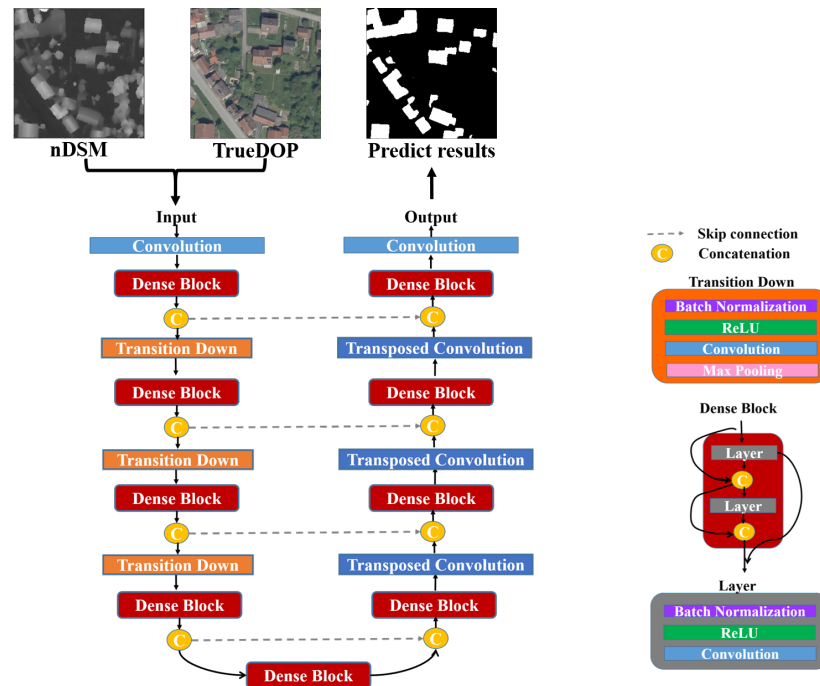


Figure 5. The implemented CNN architectures: FC-DenseNet.

5. Experiment

5.1. Data Preprocessing

The crucial element of our proposed framework is the CNN method that can predict buildings at current state. Training data is essential for CNN learning, and thus all the official geodata are preprocessed to collect training patches as input. DFK is provided as shape files, and first converted to the raster format at 0.4 m, which is the same spatial resolution as TrueDOP, nDSM, and tDSM. Then, all the tiles of TrueDOP, nDSM, and the DFK as corresponding ground reference are clipped into patches with a size of 256×256 pixels, where each patch has an overlap of 124 pixels with its neighboring patches.

Then, we collect the patches from 14 districts in the state of Bavaria, Germany, except the districts of Bad Toelz and Nuernburg. And for each district among the 14 selected districts, we split the collected patches into the train and validation subset. Table 2 shows the number of training and validation patches for the 14 selected districts.

Table 2. The numbers of training and validation patches for the 14 selected districts.

District	Number of Training Patches	Number of Validation Patches
Ansbach	67,965	18,077
Wolfratshausen	14,982	3671
Kulmbach	24,998	5679
Kronach	19,987	5112
Landau	34,964	8733
Deggendorf	38,454	9763
Landshut	60,957	15,090
Muenchen	88,364	22,213
Regensburg	47,947	11,941
Hemau	9481	2243
Rosenheim	59,141	14,789
Wasserburg	14,150	3567
Schweinfurt	54,951	13,759
Weilheim	76,959	19,202

5.2. Experiment Setup

Using the training and validation data collected from the 14 selected districts, we have firstly trained a FC-DenseNet model to get building detection results. Then, with the aid of tDSM, we have generated a seamless map of undocumented detection for one-quarter of the state of Bavaria, Germany.

To validate our building detection results, we choose the district “Bad Toelz” as the test area. Firstly, we compare our results in the district of Bad Toelz with those obtained from two conventional solutions (filter-based method and comparison-based method) utilized in the state of Bavaria, Germany. Furthermore, we also make a comparison among different CNNs. Thus, we implement another two commonly used networks (FCN-8s [27] and U-Net [28]) in the remote sensing community for building detection.

As one contribution of our work, the transferability issues with training data from selected districts around the state of Bavaria, Germany are explored. In this regard, transferability is examined by training another FC-DenseNet model with the training and validation data only from the district of Ansbach. Then we evaluate the two FC-DenseNet models on the districts of Bad Toelz and Nuernburg, respectively. Note that the districts of Bad Toelz and Nuernburg are not included from the 14 selected districts, which is helpful to investigate the transferability of these two trained models.

In order to investigate the sampling strategy in a local area where training samples are available, we also test the two trained FC-DenseNet models on the district of Ansbach, since the district of Ansbach is included in training and validation data of both trained models.

5.3. Training Details

In this study, all networks are applied under a Pytorch framework and trained for 100 epochs. All models are trained from scratch by a stochastic gradient descent (SGD) optimizer with a learning rate of 0.000001. The cross entropy loss is utilized as the loss function, and the batch size is 5. A Tesla P100 GPU with 16 G memory is used to train our models.

The configurations of CNNs included in experiments are listed as follows;

- (1) FC-DenseNet is composed of four DenseNet blocks in both encoder and decoder, and one bottleneck block connecting them, which is also a DenseNet block. In each DenseNet block, we utilize 5 convolutional layers.
- (2) FCN-8s adopts a VGG16 architecture [44] as the backbone.
- (3) U-Net is composed of five blocks in both the encoder and decoder. Each block in the encoder has two convolution layers, and in the decoder it has one transposed convolution layer.

5.4. Evaluation Metrics

For building detection, the model performance is evaluated by calculating the accuracy metrics, which include overall accuracy, precision, recall, F1 score, and intersection over union (IoU), which are defined as:

$$\text{Overall accuracy} = \frac{TP + TN}{TP + FP + FN + TN} \quad (1)$$

$$\text{precision} = \frac{TP}{TP + FP} \quad (2)$$

$$\text{recall} = \frac{TP}{TP + FN} \quad (3)$$

$$\text{F1 score} = \frac{2 * \text{precision} * \text{recall}}{\text{precision} + \text{recall}} \quad (4)$$

$$\text{IoU} = \frac{TP}{TP + FP + FN} \quad (5)$$

where TP (true positive) is the number of pixels correctly identified with the class label “building”, FN (false negative) denotes the number of omitted pixels with the class label of “building”. FP (false

positive) represents the number of “non-building” pixels in the ground reference, but are mislabeled as “building” by the model. TN (true negative) is the number of the correctly detected pixels with the class label of “non-building”. Precision denotes the fraction of identified “building” pixels that are correct with ground reference, and recall represents how many “building” pixels in the ground reference are correctly predicted. The F1 score denotes a harmonic mean between precision and recall.

6. Results

6.1. Results of Undocumented Building Constructions from Proposed Framework

In our research, we have generated a seamless map of undocumented building constructions for one-quarter of the state of Bavaria, Germany. Due to the limited space, the zoom-in visual examples of the large-scale undocumented building constructions can only be presented at block level here (see Figure 6).



Figure 6. Zoomed-in results of undocumented building constructions for one-quarter of the state of Bavaria, Germany at block level.

To evaluate the undocumented detections in a more targeted manner, we collected all the undocumented buildings in the district of Bad Toelz. Each undocumented building was reevaluated by manual photo interpretation to determine the correctness. Among the 1545 undocumented buildings from our results in the district of Bad Toelz, 1271 undocumented buildings were correctly detected.

A detailed visual analysis of undocumented building constructions in the district of Bad Toelz is given as an example in Figure 7, including (a) old undocumented building, (b) new undocumented building, (c) undocumented story construction. Note that the training data set excludes the data for the district of Bad Toelz, but it can still provide satisfying results in this district. Case (a) represents old undocumented buildings (green), which are clearly distinguishable in the TrueDOP and are shown as elevated objects in the nDSM. However, they are not contained in the DFK. Considering that no height deviation is present in the tDSM, these undocumented buildings belong to the class of old undocumented building, which indicates that they were built before time point 1 (year 2014). In case (b), a new undocumented building (red) is depicted well in our detection results. From the TrueDOP and nDSM, it can be clearly seen that this is a building, however, it is not present in the DFK. Since there is an obvious signal of height deviation from tDSM, this new undocumented building was built in the period covered by the tDSM (from year 2014 to year 2017). For the undocumented story construction illustrated in case (c), a strong signal of height deviation is present in the tDSM. This site corresponds to a building that has been recorded in the DFK; thus, we can conclude that this height deviation results from story buildup.

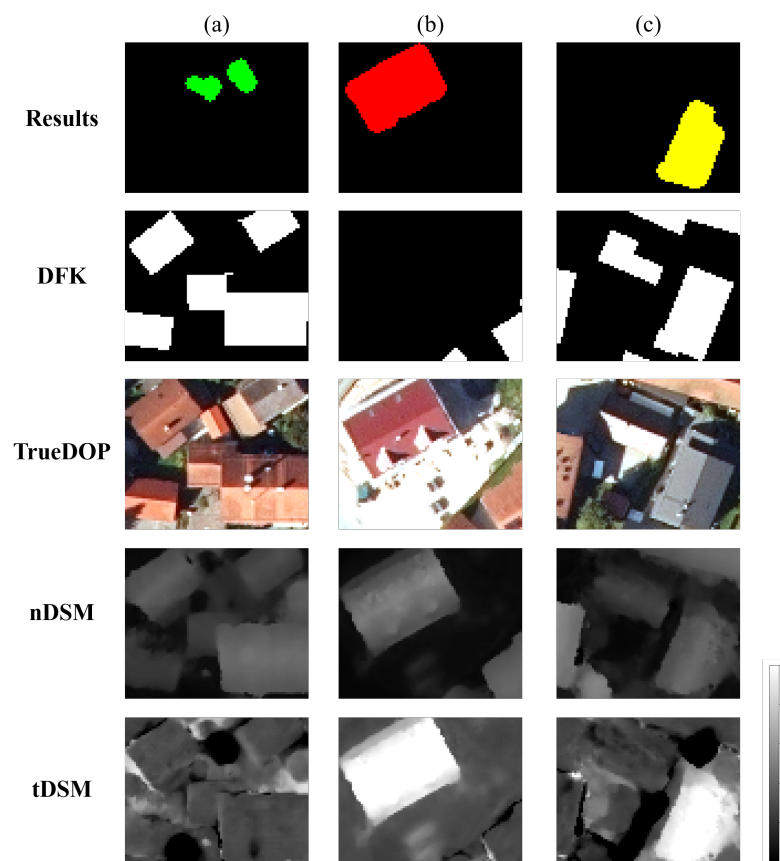


Figure 7. Example of detection results of undocumented building reconstructions for (a) old undocumented building, (b) new undocumented building, and (c) undocumented story construction.

6.2. Results of Building Detections from Proposed Framework

In our proposed framework, the module of CNN plays a vital role, and its performance has an impact on the final undocumented building detections results. In order to evaluate the CNN performance of the proposed framework, we compare our building detection results in the district of Bad Toelz with those acquired from two conventional solutions (filter-based method and comparison-based method) utilized in the state of Bavaria, Germany. A comparison among different CNNs (FC-DenseNet, FCN-8s, and U-Net) is also presented in this section.

6.2.1. Comparison with Two Conventional Solutions

The visual building detection results from the proposed framework and two other conventional solutions (the filter-based method and the comparison-based method) are shown in Figure 8. For further verification, a statistical analysis of the results from these three methods on the district of Bad Toelz is carried out (see Table 3). As a comparative measure, the F1 score is clearly more objective here, since it takes both false alarms and omitted detections into consideration.

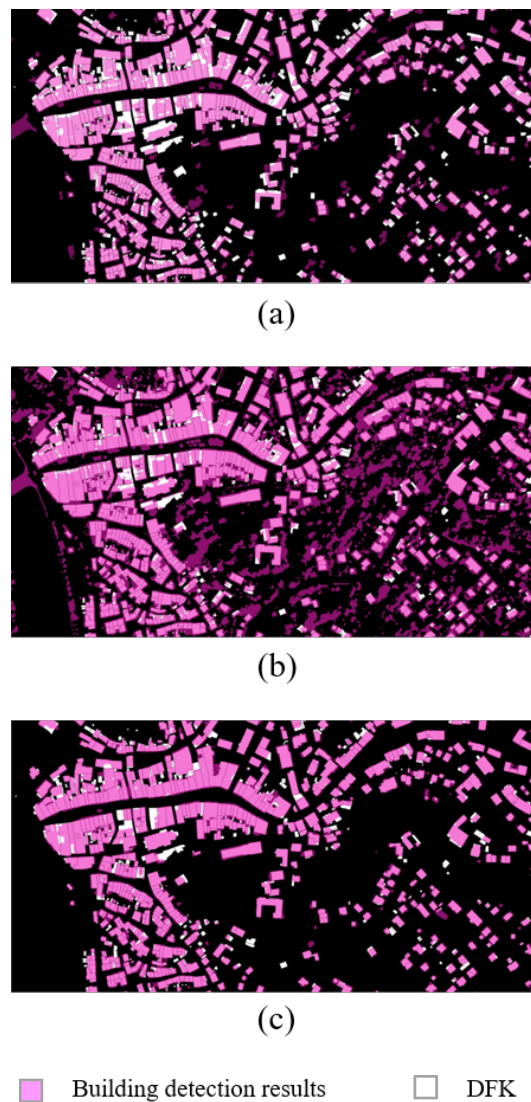


Figure 8. Building detection results from (a) filter-based method, (b) comparison-based method, and (c) CNN model.

Table 3. Statistical accuracy of building detection results among different methods.

Method	Overall Accuracy	Precision	Recall	F1 Score	IoU
Filter-based method	97.6%	59.7%	82.3%	69.3%	53.0%
Comparison-based method	90.4%	24.1%	89.0%	37.9%	23.4%
CNN method	99.0%	84.6%	85.5%	85.1%	74.0%

For the filter-based method, the low precision rate results from some false detection. One reason is that the nDSM naturally delivers all elevated objects, such as vegetation and trucks, in addition to buildings. The other reason is that the color filter is mostly affected by aerial imaging conditions, which means that vegetation can be also misclassified as buildings under some uncertainties. Some omission errors in the results also reduce the recall value, which may be due to the confidence intervals of the color filter. This interval may be insufficient to identify buildings, since the RGB values for an individual building are significantly dependent on the amount of sunlight. In this case, there are some buildings whose colors are in the peripheral areas, e.g., very bright white roofs or very dark roofs, which can not be identified as buildings.

In the results obtained from the comparison-based method, the precision value is much lower than the other two methods, which indicates that many non-building pixels are mislabeled as buildings. After a further detailed visual check, we have found that there is a lot of confusion between trees and buildings. Since some trees grow above the roofs, the RGB color cube in TrueDOP collected from reference buildings also involve RGB color values of vegetation. In this regard, the reference for buildings in the RGB color cube will be distorted by these vegetation components, and thus vegetation can be wrongly classified as buildings. Moreover, the color values of vegetation and dark roofs are also similar in shadow areas, which produces misclassifications between vegetation and buildings.

The CNN method yields the highest precision values, which indicates that it can suppress false alarms well. The CNN model clearly outperforms the other two methods with respect to accuracy (F1 score). This proves that, in a comparison of the building detectors examined, reliable building detection and a good separation from vegetation are only possible with the CNN model. This is due to the powerful generalization capability of CNNs, which are independent from prior knowledge and hand-crafted features.

6.2.2. Comparison with Other CNNs

In order to compare with other CNNs, two networks including FCN-8s, and U-Net are also trained with the training and validation samples collected from 14 districts. Their respective performance is then tested on the district of Bad Toelz.

Statistical results of three networks are shown in Table 4. It is demonstrated that FC-DenseNet outperforms other two methods in terms of both F1 score and IoU. Specifically, comparisons with FCN-8s and U-Net, where FC-DenseNet obtain increments of 3.9% and 3.2% in F1 score, respectively, validates its superiority in the task of building detection. Compared to U-Net, FC-DenseNet reaches improvements of 3.2% and 4.6% in F1 score and IoU, which indicates that the DenseNet block is more effective than the normal block.

Figure 9 shows a few examples of building detection results of three networks. In all these three scenes, FC-DenseNet is able to capture more buildings, whereas U-Net and FCN-8s suffer from more omission errors. This is mainly because, in FC-DenseNet, the DenseNet block reuses features, which leads to a better judgment of buildings. Thanks to the architecture of skip connection, FC-DenseNet is capable of preserving sharper building boundaries than FCN-8s.

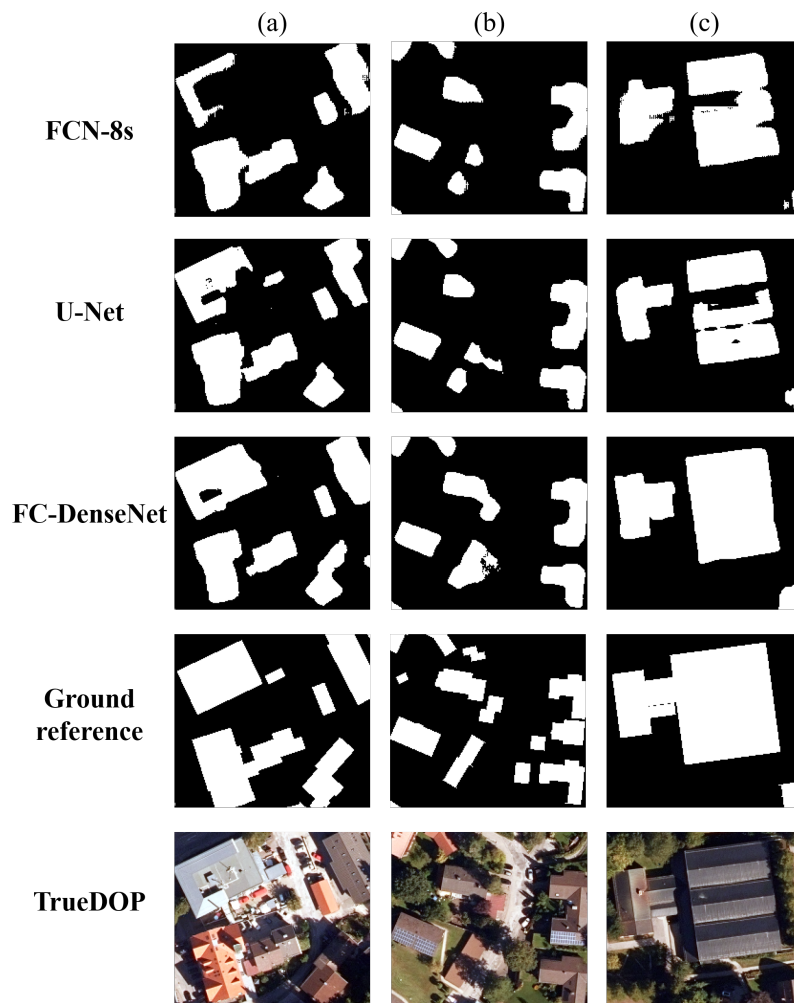


Figure 9. Three examples (a–c) represent the building detection results from three CNNs: FCN-8s, U-Net, and FC-DenseNet.

Table 4. Statistical accuracy of building detection results among different CNNs.

Method	Overall Accuracy	Precision	Recall	F1 Score	IoU
FCN-8s	98.8%	82.5%	80.1%	81.2%	68.4%
U-Net	98.8%	81.5%	82.3%	81.9%	69.4%
FC-DenseNet	99.0%	84.6%	85.5%	85.1%	74.0%

7. Discussion

The collection of training samples for large-scale building detection takes a large quantity of time and manual work. Therefore, the investigation of transferability issues and sampling strategies for building detection at large-scale is vital in practical use. In this regard, we have trained two FC-DenseNet models with different training and validation sets, and named them as the trained model 1 and 2, respectively. In the trained model 1, the training samples are only collected from the district of Ansbach. In the trained model 2, the training samples are collected not only from the district of Ansbach, but also another 13 districts.

7.1. Transferability Investigation

The transferability of trained models is examined by evaluating the performances of the two trained models in the districts of Bad Toelz and Nuernburg, respectively. For both trained models, neither training data nor validation data include the data from these two districts, which is considered

as a more realistic test for the task of large-scale building detection, since training data can only be collected from limited areas. Table 5 proves that the trained model 2 has superior transferability. In the district Bad Toelz, F1 score and IoU of the trained model 2 shows a large improvement of 12.8% and 17.4% in comparison to the trained model 1, respectively. In the district of Nuernburg, the trained model 2 surpasses the trained model 1 by 3.9% and 5.8% in the F1 score and IoU score, respectively.

Table 5. Accuracy of two different trained models evaluated in the districts of Bad Toelz and Nuernburg.

Trained Model	Train and Validation District	Test District	Overall Accuracy	Precision	Recall	F1 Score	IoU
1	Ansbach	Bad Toelz	98.2%	75.3%	69.4%	72.3%	56.6%
2	14 districts	Bad Toelz	99.0%	84.6%	85.5%	85.1%	74.0%
1	Ansbach	Nuernburg	92.4%	86.9%	78.0%	82.2%	69.8%
2	14 districts	Nuernburg	94.6%	87.6%	84.7%	86.1%	75.6%

Some visual examples of these two trained models in the districts of Bad Toelz and Nuernburg are illustrated in Figure 10 for comparison. The visual results are consistent with the statistical results of Table 5, where the trained model 2 shows higher increments of precision and recall than the trained model 1. This indicates that when the evaluation data is unseen by both the training and the validation set, the optimal sampling strategy is to collect training data from different districts rather than from only one. This improvement is due to the fact that the trained model 2 collects the training samples from 14 different districts in the state of Bavaria, Germany, where the variety in the types of buildings facilitates the learning of CNN. This again confirms that a diverse training set is beneficial to the generalization capability of CNN. Since CNN is focused on learning location-specific building patterns, a diverse training set can mitigate this effect and enable the CNN to learn more generic patterns, where the semantic segmentation in an unseen area can be improved [45].

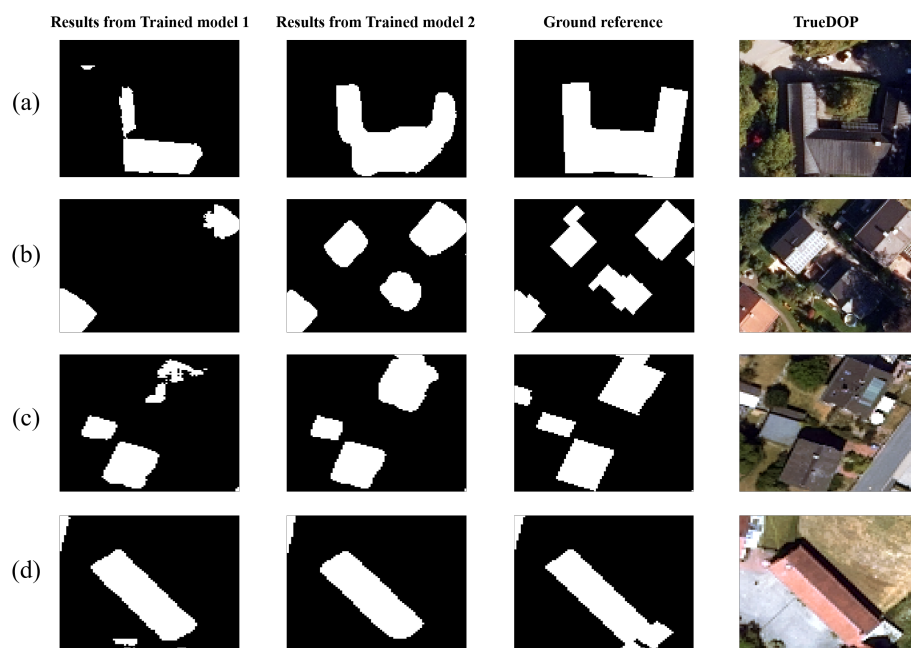


Figure 10. Two examples (a,b) represent the buildings detection results in the district of Bad Toelz obtained from trained model 1 and trained model 2, respectively. Two examples (c,d) represent the buildings detection results in the district of Nuernburg obtained from trained model 1 and trained model 2, respectively.

7.2. Sampling Strategy Investigation

In order to investigate the sampling strategy in a local area where training samples are available, we test the two trained models on the district of Ansbach. This is due to the fact that the district of Ansbach is included in both two trained models. The evaluation data in the district of Ansbach is the same as the validation data in the trained model 1 (18,077 patches). Table 6 presents a comparison of the statistical accuracy of two trained models. An interesting finding is that, statistical metrics of the two trained models only show slight differences, which indicates that local training sample collection and training can achieve comparative performance as collecting extensive training samples from different districts. This is because training data in the trained model 1 share a similar data distribution with evaluation data in the district of Ansbach, which can also lead to a good fit of the model. This provides a sampling strategy in a local area where the training samples are available, so that we can just use only local training samples to obtain the building detection results in this area rather than collecting extensive training samples from multiple districts. This sampling strategy can save much more effort and time in a local area with available training samples.

Table 6. Accuracy of two different trained models evaluated in the district of Ansbach.

Trained Model	Train and Validation District	Test District	Overall Accuracy	Precision	Recall	F1 Score	IoU
1	Ansbach	Ansbach	98.9%	90.9%	90.3%	90.5%	82.7%
2	14 districts	Ansbach	98.8%	91.3%	89.3%	90.3%	82.3%

8. Conclusions

In order to ensure the transparent management of land properties, buildings as vital terrestrial objects, need an official terrestrial survey to be documented in the cadastral maps. For this purpose, we have proposed a framework for the detection of undocumented building constructions from official geodata, which includes nDSM, TrueDOP, and DFK. Moreover, the proposed framework categorizes detected undocumented building constructions into three types: old undocumented building, new undocumented building, and undocumented story construction with the aid of tDSM. This can contribute to the management of different construction cases.

Our framework is based on a CNN and decision fusion, and has shown greater potential for updating the building model in geographic information system than two strategies used so far in the state of Bavaria, Germany.

We investigated the transferability issue and sampling strategies for building detection at large-scale. In an unseen area, the model that collects diverse training samples from multiple districts has better transferability than the model that collects training data from only one district. However, in a local area where training samples are already available, the local samples collection and training can achieve comparative performance as the model that collects extensive training samples from different districts. These practical strategies are beneficial to other large-scale object detection works that use remote sensing data.

Furthermore, the seamless map of undocumented building constructions generated in our research covers one-quarter of the state of Bavaria, Germany at a spatial resolution of 0.4 m, and is beneficial to efficient land resource management and sustainable urban development.

Author Contributions: Conceptualization, X.Z., R.R., S.A. and M.S.; methodology, Q.L. and Y.S.; software, Q.L. and K.M.; validation, Q.L. and K.M.; formal analysis, Q.L.; investigation, Q.L.; resources, X.Z. and R.R.; data curation, Q.L. and K.M.; writing—original draft preparation, Q.L.; writing—review and editing, Q.L., Y.S., S.A., R.R., K.M., M.S., C.G. and X.Z.; visualization, Q.L. and K.M.; supervision, X.Z.; project administration, X.Z.; funding acquisition, X.Z. All authors have read and agreed to the published version of the manuscript.

Funding: This work is supported by the European Research Council (ERC) under the European Union’s Horizon 2020 research and innovation program (grant agreement no. ERC-2016-StG-714087, acronym: So2Sat, www.so2sat.eu), the Helmholtz Association under the framework of the Young Investigators Group “SiPEO” (VH-NG-1018, www.sipeo.bgu.tum.de) and Helmholtz Excellent Professorship “Data Science in Earth Observation-Big Data Fusion for Urban Research”. This work is also part of the project “Investigation of building cases using AI” funded by Bavarian State Ministry of Finance and Regional Identity (StMFH) and the Bavarian Agency for Digitization, High-Speed Internet and Surveying.

Conflicts of Interest: The authors declare no conflict of interest. The founding sponsors had no role in the design of the study; in the collection, analyses, or interpretation of data; in the writing of the manuscript, and in the decision to publish the results.

References

1. Arlinger, K.; Roschlaub, R. Calculation and update of a 3D building model of Bavaria using LiDAR, image matching and cadastre information. In Proceedings of the 8th International 3D GeoInfo Conference, Istanbul, Turkey, 27–29 November 2013; pp. 28–29.
2. Aringer, K.; Roschlaub, R. Bavarian 3D building model and update concept based on LiDAR, image matching and cadastre information. In *Innovations in 3D Geo-Information Sciences*; Springer: Berlin/Heidelberg, Germany, 2014; pp. 143–157.
3. Geßler, S.; Krey, T.; Möst, K.; Roschlaub, R. Mit Datenfusionierung Mehrwerte schaffen—Ein Expertensystem zur Baufallerkundung. *DVW Mitt.* **2019**, *2*, 159–187.
4. Roschlaub, R.; Möst, K.; Krey, T. Automated Classification of Building Roofs for the Updating of 3D Building Models Using Heuristic Methods. *PFG J. Photogramm. Remote Sens. Geoinf. Sci.* **2020**, *88*, 85–97.
5. Zhu, X.X.; Tuia, D.; Mou, L.; Xia, G.S.; Zhang, L.; Xu, F.; Fraundorfer, F. Deep learning in remote sensing: A comprehensive review and list of resources. *IEEE Geosci. Remote Sens. Mag.* **2017**, *5*, 8–36.
6. Li, J.; Huang, X.; Gong, J. Deep neural network for remote-sensing image interpretation: Status and perspectives. *Natl. Sci. Rev.* **2019**, *6*, 1082–1086.
7. Mou, L.; Zhu, X.X. Learning to Pay Attention on Spectral Domain: A Spectral Attention Module-Based Convolutional Network for Hyperspectral Image Classification. *IEEE Trans. Geosci. Remote Sens.* **2019**, *58*, 110–122.
8. Qiu, C.; Mou, L.; Schmitt, M.; Zhu, X.X. Local climate zone-based urban land cover classification from multi-seasonal Sentinel-2 images with a recurrent residual network. *ISPRS J. Photogramm. Remote Sens.* **2019**, *154*, 151–162. [[CrossRef](#)]
9. Mou, L.; Bruzzone, L.; Zhu, X.X. Learning spectral-spatial-temporal features via a recurrent convolutional neural network for change detection in multispectral imagery. *IEEE Trans. Geosci. Remote Sens.* **2018**, *57*, 924–935. [[CrossRef](#)]
10. Caye Daudt, R.; Le Saux, B.; Boulch, A.; Gousseau, Y. Guided anisotropic diffusion and iterative learning for weakly supervised change detection. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops, Long Beach, CA, USA, 16–20 June 2019.
11. Hua, Y.; Mou, L.; Zhu, X.X. Relation Network for Multilabel Aerial Image Classification. *IEEE Trans. Geosci. Remote Sens.* **2020**. [[CrossRef](#)]
12. Hua, Y.; Mou, L.; Zhu, X.X. Recurrently exploring class-wise attention in a hybrid convolutional and bidirectional LSTM network for multi-label aerial image classification. *ISPRS J. Photogramm. Remote Sens.* **2019**, *149*, 188–199.
13. Qiu, C.; Schmitt, M.; Geiß, C.; Chen, T.H.K.; Zhu, X.X. A framework for large-scale mapping of human settlement extent from Sentinel-2 images via fully convolutional neural networks. *ISPRS J. Photogramm. Remote Sens.* **2020**, *163*, 152–170. [[CrossRef](#)]
14. He, C.; Liu, Z.; Gou, S.; Zhang, Q.; Zhang, J.; Xu, L. Detecting global urban expansion over the last three decades using a fully convolutional network. *Environ. Res. Lett.* **2019**, *14*, 034008. [[CrossRef](#)]
15. Shi, Y.; Li, Q.; Zhu, X.X. Building Footprint Generation Using Improved Generative Adversarial Networks. *IEEE Geosci. Remote Sens. Lett.* **2018**, *16*, 603–607. [[CrossRef](#)]
16. Shi, Y.; Li, Q.; Zhu, X.X. Building segmentation through a gated graph convolutional neural network with deep structured feature embedding. *ISPRS J. Photogramm. Remote Sens.* **2020**, *159*, 184–197. [[CrossRef](#)]

17. Li, Q.; Shi, Y.; Huang, X.; Zhu, X.X. Building Footprint Generation by Integrating Convolution Neural Network With Feature Pairwise Conditional Random Field (FPCRF). *IEEE Trans. Geosci. Remote Sens.* **2020**. [\[CrossRef\]](#)
18. Wurm, M.; Stark, T.; Zhu, X.X.; Weigand, M.; Taubenböck, H. Semantic segmentation of slums in satellite images using transfer learning on fully convolutional neural networks. *ISPRS J. Photogramm. Remote Sens.* **2019**, *150*, 59–69. [\[CrossRef\]](#)
19. Demuzere, M.; Bechtel, B.; Mills, G. Global transferability of local climate zone models. *Urban Clim.* **2019**, *27*, 46–63. [\[CrossRef\]](#)
20. Li, Q.; Qiu, C.; Ma, L.; Schmitt, M. Mapping the Land Cover of Africa at 10 m Resolution from Multi-Source Remote Sensing Data with Google Earth Engine. *Remote Sens.* **2020**, *12*, 602. [\[CrossRef\]](#)
21. San, D.K.; Turker, M. Building extraction from high resolution satellite images using Hough transform. *Int. Arch. Photogramm. Remote Sens. Spat. Inf. Sci.* **2010**, *38*, 1063–1068.
22. Ok, A.O. Automated detection of buildings from single VHR multispectral images using shadow information and graph cuts. *ISPRS J. Photogramm. Remote Sens.* **2013**, *86*, 21–40. [\[CrossRef\]](#)
23. Huang, X.; Zhang, L. A multidirectional and multiscale morphological index for automatic building extraction from multispectral GeoEye-1 imagery. *Photogramm. Eng. Remote Sens.* **2011**, *77*, 721–732. [\[CrossRef\]](#)
24. Inglada, J. Automatic recognition of man-made objects in high resolution optical remote sensing images by SVM classification of geometric image features. *ISPRS J. Photogramm. Remote Sens.* **2007**, *62*, 236–248. [\[CrossRef\]](#)
25. Yang, H.L.; Yuan, J.; Lunga, D.; Laverdiere, M.; Rose, A.; Bhaduri, B. Building extraction at scale using convolutional neural network: Mapping of the united states. *IEEE J. Sel. Top. Appl. Earth Obs. Remote Sens.* **2018**, *11*, 2600–2614. [\[CrossRef\]](#)
26. Garcia-Garcia, A.; Orts-Escolano, S.; Oprea, S.; Villena-Martinez, V.; Garcia-Rodriguez, J. A review on deep learning techniques applied to semantic segmentation. *arXiv* **2017**, arXiv:1704.06857.
27. Long, J.; Shelhamer, E.; Darrell, T. Fully convolutional networks for semantic segmentation. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Boston, MA, USA, 7–12 June 2015; pp. 3431–3440.
28. Ronneberger, O.; Fischer, P.; Brox, T. U-net: Convolutional networks for biomedical image segmentation. In Proceedings of the International Conference on Medical Image Computing and Computer-Assisted Intervention, Munich, Germany, 5–9 October 2015; Springer: Berlin/Heidelberg, Germany, 2015; pp. 234–241.
29. Badrinarayanan, V.; Kendall, A.; Cipolla, R. Segnet: A deep convolutional encoder-decoder architecture for image segmentation. *IEEE Trans. Pattern Anal. Mach. Intell.* **2017**, *39*, 2481–2495. [\[CrossRef\]](#)
30. Bischke, B.; Helber, P.; Folz, J.; Borth, D.; Dengel, A. Multi-task learning for segmentation of building footprints with deep neural networks. In Proceedings of the 2019 IEEE International Conference on Image Processing (ICIP), Taipei, Taiwan, 22–25 September 2019; pp. 1480–1484.
31. Bittner, K.; Adam, F.; Cui, S.; Körner, M.; Reinartz, P. 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.* **2018**, *11*, 2615–2629. [\[CrossRef\]](#)
32. Ji, S.; Wei, S.; Lu, M. Fully convolutional networks for multisource building extraction from an open aerial and satellite imagery data set. *IEEE Trans. Geosci. Remote Sens.* **2018**, *57*, 574–586. [\[CrossRef\]](#)
33. Jégou, S.; Drozdal, M.; Vazquez, D.; Romero, A.; Bengio, Y. The one hundred layers tiramisu: Fully convolutional densenets for semantic segmentation. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops, Honolulu, HI, USA, 21–26 July 2017; pp. 11–19.
34. Li, X.; Yao, X.; Fang, Y. Building-A-Nets: Robust Building Extraction from High-Resolution Remote Sensing Images with Adversarial Networks. *IEEE J. Sel. Top. Appl. Earth Obs. Remote Sens.* **2018**, *11*, 3680–3687. [\[CrossRef\]](#)
35. Ressel, C.; Brockmann, H.; Mandlbürger, G.; Pfeifer, N. Dense Image Matching vs. Airborne Laser Scanning—Comparison of two methods for deriving terrain models. *Photogramm. Fernerkund. Geoinf.* **2016**, *2016*, 57–73. [\[CrossRef\]](#)
36. Griffiths, D.; Boehm, J. Improving public data for building segmentation from Convolutional Neural Networks (CNNs) for fused airborne lidar and image data using active contours. *ISPRS J. Photogramm. Remote Sens.* **2019**, *154*, 70–83. [\[CrossRef\]](#)

37. Vargas-Muñoz, J.E.; Lobry, S.; Falcão, A.X.; Tuia, D. Correcting rural building annotations in OpenStreetMap using convolutional neural networks. *ISPRS J. Photogramm. Remote Sens.* **2019**, *147*, 283–293. [[CrossRef](#)]
38. Sirmacek, B.; Unsalan, C. Building detection from aerial images using invariant color features and shadow information. In Proceedings of the 2008 23rd International Symposium on Computer and Information Sciences, Istanbul, Turkey, 27–29 October 2008; pp. 1–5.
39. Maggiori, E.; Tarabalka, Y.; Charpiat, G.; Alliez, P. Convolutional neural networks for large-scale remote-sensing image classification. *IEEE Trans. Geosci. Remote Sens.* **2016**, *55*, 645–657.
40. Huang, G.; Liu, Z.; Van Der Maaten, L.; Weinberger, K.Q. Densely connected convolutional networks. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Honolulu, HI, USA, 21–26 July 2017; pp. 4700–4708.
41. Gu, J.; Wang, Z.; Kuen, J.; Ma, L.; Shahroudy, A.; Shuai, B.; Liu, T.; Wang, X.; Wang, G.; Cai, J.; et al. Recent advances in convolutional neural networks. *Pattern Recognit.* **2018**, *77*, 354–377.
42. He, K.; Zhang, X.; Ren, S.; Sun, J. Deep residual learning for image recognition. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Las Vegas, NV, USA, 27–30 June 2016; pp. 770–778.
43. Drozdal, M.; Vorontsov, E.; Chartrand, G.; Kadoury, S.; Pal, C. The importance of skip connections in biomedical image segmentation. In *Deep Learning and Data Labeling for Medical Applications*; Springer: Berlin/Heidelberg, Germany, 2016; pp. 179–187.
44. Simonyan, K.; Zisserman, A. Very deep convolutional networks for large-scale image recognition. *arXiv* **2014**, arXiv:1409.1556.
45. Kaiser, P.; Wegner, J.D.; Lucchi, A.; Jaggi, M.; Hofmann, T.; Schindler, K. Learning aerial image segmentation from online maps. *IEEE Trans. Geosci. Remote Sens.* **2017**, *55*, 6054–6068. [[CrossRef](#)]

Publisher’s Note: MDPI stays neutral with regard to jurisdictional claims in published maps and institutional affiliations.



© 2020 by the authors. Licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license (<http://creativecommons.org/licenses/by/4.0/>).