TUM School of Engineering and Design
Technische Universität München

Large-scale LoD1 Building Model Reconstruction from a Single SAR Image

Yao Sun

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

Doktors der Ingenieurwissenschaften (Dr.-Ing.)

genehmigten Dissertation.

Vorsitzender:
Prof. Dr. Bing Zhu

Prüfende der Dissertation:
1. Prof. Dr.-Ing. habil. Xiao Xiang Zhu
2. Prof. Dr.-Ing. habil. Richard Bamler
3. Prof. Dr. Paolo Gamba

Die Dissertation wurde am 05.07.2021 bei der Technischen Universität München eingereicht und durch die TUM School of Engineering and Design am 02.12.2021 angenommen.

Abstract

Three-dimensional (3-D) building models are widely used in public and commercial sectors for environmental researches and location-based services. For the past three decades, 3-D building reconstruction has been a hot topic in remote sensing, however, there is limited information on building models on regional and global scales. Synthetic Aperture Radar (SAR) data have been employed for modeling buildings due to their imaging capability regardless of the time or weather conditions. In addition, complete global coverages of TerraSAR-X/TanDEM-X stripmap mode data have been acquired since 2012, providing great potential as a data source for global building reconstruction. However, building interpretation from SAR data is highly challenging. Due to the side-looking geometry and one-band radar sensors, urban structures are clearly visible in SAR images but are difficult to distinguish from each other. Although extensive research has been carried out on building reconstruction using SAR data, to date, no single study investigates large-scale building reconstruction from a single SAR image.

This dissertation addresses large-scale Levels of Detail (LoD)-1 building models reconstruction from a single SAR image. Considering the characteristics of buildings in SAR images, building footprints are introduced as complementary data, and deep neural networks are employed for large-scale reconstruction. The work is developed in three stages:

- First, building footprints must be registered to SAR images for supporting SAR image interpretation. Therefore a framework is developed that automatically registers building footprints to a corresponding SAR image.

- Second, the employment of deep learning methods requires training data. Therefore an accurate Digital Surface Model (DSM)is introduced to generate individual building areas in a SAR image, and a segmentation network is proposed for predicting building areas on a large scale. The extracted building segments are then employed for LoD1 model reconstruction.

- Third, to reconstruct buildings in larger areas, more training data are needed. However, accurate DEMs are unavailable in most cases. Therefore, the LoD1 building reconstruction problem is reformulated as a bounding box regression problem so that height data from multiple sources can be employed to generate bounding boxes of buildings. A regression network is proposed and examined for four study sites using both TerraSAR-X spotlight image and stripmap mode images.

To the author's best knowledge, this is the first study investigating individual buildings in single SAR images on a large scale and the first study applying deep learning for individual building analysis using SAR images. The proposed algorithms have great potential to be applied on a regional and even global scale.

Zusammenfassung

Dreidimensionale (3-D) Gebäudemodelle werden oft im öffentlichen und kommerziellen Bereich für Umweltanalysen und standortbezogene Dienste genutzt. In den letzten dreißig Jahren war die dreidimensionale Gebäude-Rekonstruktion ein sehr aktuelles Thema für die Fernerkundung; allerdings gibt es nur sehr beschränkte Ergebnisse über Gebäudemodelle auf regionalen und globalen Maßstäben. Dabei werden SAR-Daten (Radar-Daten mit synthetischer Apertur) aufgrund ihrer Fähigkeit zu Bildgebung unabhängig von Zeit und Wetterbedingungen zur Gebäude-Modellierung eingesetzt. Seit 2012 werden vollständige globale Abdeckungen von Stripmap-Daten von TerraSAR-X und TanDEM-X erfasst, die eine wichtige potentielle Datenquelle für globale Gebäude-Rekonstruktionen darstellen, jedoch stellt eine Interpretation von Gebäuden basierend auf SAR-Daten sehr hohe Anforderungen. Aufgrund ihrer Schrägsicht-Geometrie sowie den Einzelband-Radargeräten sind urbane Strukturen in SAR-Bildern klar erkennbar, jedoch schwierig voneinander zu unterscheiden. Obwohl bereits umfangreiche Forschungsarbeiten zur Gebäude-Rekonstruktion aus SAR-Daten durchgeführt wurden, untersucht bisher keine eigene Studie die großflächige Gebäude-Rekonstruktion aus einem einzelnen SAR-Bild.

Diese Dissertation behandelt daher die großflächige Rekonstruktion von Gebäudemodellen der Detailierungsstufe 1 (LoD-1) aus einem einzelnen SAR-Bild. Basierend auf den Eigenschaften von Gebäuden in SAR-Bildern, integrieren wir die Signaturen von Gebäuden als zusätzliche Informationen und wir verwenden mehrschichtige (d. h. ,,tiefe“) neuronale Netze für großflächige Rekonstruktionen. Die Arbeiten beinhalten drei Schritte:

- Erstens müssen Gebäudesignaturen und SAR-Bilder geometrisch übereinadergelegt werden, um die Interpretation von SAR-Bildern zu unterstützen. Deshalb entwickeln wir ein Konzept, das automatisch Gebäudesignaturen und entsprechende SAR-Bilder übereinanderlegt.

- Zweitens benötigt die Verwendung von Deep-Learning-Methoden entsprechende Trainings-daten. Daher stellen wir ein präzises digitalesOberflächenmodell bereit, um die Positionen der einzelnen Gebäude in einem SAR-Bild zu generieren und wir schlagen ein Segmentierungs-netz vor, um Gebäudeflächen großräumig vorherzusagen. Die einzelnen Gebäudeteile werden dann für eine LoD-1-Modell-Rekonstruktion verwendet.

- Drittens braucht man, um Gebäude in größeren Bereichen zu rekonstruieren, zusätzliche Trainingsdaten. Allerdings sind in den meisten Fällen keine genauen digitalen Höhenmodelle verfügbar. Daher wird das LoD-1-Rekonstruktionsproblem als ein Objektlokalisierungs-Problem neu formuliert, sodass Höhendaten aus mehreren Quellen eingesetzt werden können, um Begrenzungsrahmen von Gebäuden zu generieren. Dafür schlagen wir ein Regressions-Netzwerk vor, das von uns für vier Testgebiete mit TerraSAR-X-Bildern in den Betriebsarten Spotlight und Stripmap untersucht wurde.

Nach bestem Wissen der Autorin ist dies die erste Studie, die großflächig einzelne Gebäude in einem einzelnen SAR-Bild untersucht, sowie die erste Studie, die Deep Learning zur Analyse einzelner Gebäude in SAR-Bildern einsetzt. Die vorgeschlagenen Algorithmen haben ein großes Potential, auf regionalem und sogar globalem Maßstab angewendet zu werden.

Contents

Abstract **iii**

Zusammenfassung **v**

Acronyms **xi**

1 Introduction **1**
1.1 Motivation . . . 1
1.2 Research Objectives . . . 2
1.3 Thesis Outline . . . 3

2 Background Theory **5**
2.1 SAR Basics . . . 5
2.1.1 SAR Imaging . . . 5
2.1.2 TerraSAR-X Imaging Modes . . . 5
2.1.3 Radiometric Effects and Geometrical Distortion . . . 6
2.2 Buildings in SAR Images . . . 8
2.2.1 Components of a Building and the Backscatter Contributions . . . 8
2.2.2 Buildings in SAR Images of Different Imaging Modes . . . 10
2.2.3 Challenges of Building Analyses in SAR Images . . . 11
2.3 LoD1 Building Models . . . 12
2.3.1 LoDs: Levels of Detail for Building Models . . . 12
2.3.2 The LoD1 Building Model and its Subsets . . . 12
2.4 Building Footprints as Complimentary Data . . . 13
2.4.1 Data Sources . . . 13
2.4.2 Error Sources in Radar Coding . . . 14

3 State of the Art **17**
3.1 Related Studies on Building Reconstruction from SAR Data . . . 17
3.1.1 Building Reconstruction Using Different SAR Data . . . 17
3.1.2 Research Focuses . . . 21
3.1.3 Reconstruction Approaches . . . 23
3.1.4 Problems in Applying the Existing Methods to Large Areas . . . 24
3.2 Related Registration Algorithms . . . 25
3.2.1 Registration of SAR and Optical Images . . . 25
3.2.2 Registration of SAR Images and Building Footprint Data . . . 26
3.3 Related Advances in Deep Learning . . . 27
3.3.1 Deep Learning in SAR . . . 27
3.3.2 Image Segmentation . . . 27
3.3.3 Bounding Box Regression . . . 28
3.4 Contributions of the Thesis . . . 29

4 Data Set Generation **31**
4.1 Data Set Requirements . . . 31
4.1.1 Required Data . . . 31
4.1.2 Reasons for the Required Data . . . 31
4.2 Automatic Annotating Regions of Individual Buildings in a SAR Image . . . 32
4.2.1 Scene Modelling . . . 32
4.2.2 Data Set Generation in the SAR image Coordinate System . . . 34
4.3 Automatic Labelling Bounding Boxes of Buildings in SAR Images . . . 38
4.3.1 Building Height Acquisition . . . 38
4.3.2 Data Set Generation . . . 38
4.4 Generating Ground Truth of Registered Building Footprints and a SAR Image 39

5 Automatic Registration of a Single SAR Image and Building Footprint Data on a Large Scale **41**
5.1 Problem Formulation Based on the Feature Correspondence . . . 41
5.2 Corresponding Feature Extraction . . . 42
5.2.1 Double Bounce Lines in the SAR Image . . . 42
5.2.2 Sensor-visible Edges in Building Footprint Polygons . . . 44
5.3 Progressive Feature Registration . . . 45
5.4 Experimental Results and Evaluation . . . 46
5.4.1 Test Site and Data Set . . . 47
5.4.2 Results of Feature Extraction and Registration . . . 48
5.4.3 Evaluation . . . 54
5.5 Discussion . . . 55
5.5.1 Can the Proposed Approach Work with Stripmap Images? . . . 55
5.5.2 Which Terrain Model to Use for Radar Coding? . . . 57
5.5.3 Applicable Scenario . . . 58
5.5.4 Further Applications . . . 60
5.6 Summary . . . 61

6 Conditional GIS-aware Network for Individual Building Segmentation in VHR SAR Images **63**
6.1 Conditional GIS-aware Network . . . 63
6.1.1 Multi-level Feature Extraction Module . . . 64
6.1.2 Conditional GIS-aware Normalization Module . . . 64
6.1.3 Configuration of CG-Net . . . 65
6.2 Experimental Results and Evaluation . . . 66
6.2.1 Data Set and Training Details . . . 66
6.2.2 Quantitative and Qualitative Evaluation . . . 67
6.2.3 Comparison of Complete Building Footprints and Sensor-visible Footprint Segments . . . 70
6.2.4 Can CG-Net Work with Inaccurate GIS Data? . . . 72
6.3 Discussion . . . 74
6.3.1 Further Application: Reconstruction of LoD1 Building Models from a SAR Image . . . 74
6.3.2 Can CG-Net Predict Individual Buildings from Stripmap SAR Images? 75
6.4 Summary . . . 79

7 Building Height Retrieval from Single SAR Images with Bounding Box Regression **83**
7.1 Problem Formulation Based on the Radar Viewing Geometry 84
7.2 Footprint Guided Bounding Box Regression Network 84
7.3 Experimental Results 86
7.3.1 Data Sets 86
7.3.2 Training Details 89
7.3.3 Comparative Experiments 90
7.3.4 Quantitative Evaluation 90
7.3.5 Qualitative Evaluation 92
7.4 Discussion 95
7.4.1 Can the Proposed Network Work with Inaccurate GIS Data? 95
7.4.2 Influences of the Nonlocal Filtering Procedure on SAR Data 95
7.4.3 Pros and Cons of the Segmentation Networks and Regression Networks for Building Height Retrieval 101
7.5 Summary 102

8 Conclusion and Outlook **103**
8.1 Conclusion 103
8.2 Outlook 105

List of Figures **107**

List of Tables **111**

Bibliography **113**

Acronyms

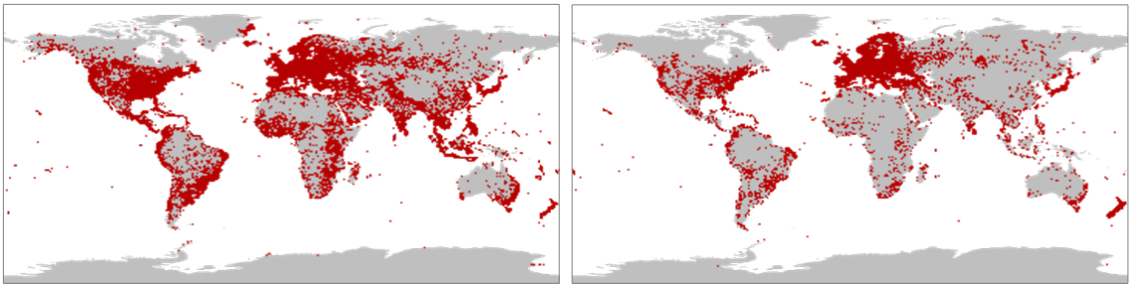
1-D	one-dimension.
2-D	two-dimension.
3-D	three-dimension.
ASTER	Advanced Spaceborne Thermal Emission and Reflection Radiometer.
CFAR	Constant False Alarm Rate.
CNN	Convolutional Neural Network.
DBSCAN	Density Based Spatial Clustering of Applications with Noise.
DEM	Digital Elevation Model.
DSM	Digital Surface Model.
FCN	Fully Convolutional Networks.
FPN	Feature Pyramid Networks.
GCP	Ground Control Point.
GDEM	Global Digital Elevation Model.
GIS	Geographic Information System.
HPR	Hidden Point Removal.
ICP	Iterative Closest Point.
InSAR	Interferometric SAR.
IWAP	Integrated Wide Area Processor.
LiDAR	Light Detection and Ranging.
LoD	Level of Detail.
NN	Nearest Neighbor.
OSM	OpenStreetMap.
PDF	Probability Density Function.
PSI	Persistent Scatterer Interferometry.
ReLU	Rectified Linear Unit.

RoI	Region of Interest.
RPN	Region Proposal Network.
SAR	Synthetic Aperture Radar.
SGD	Stochastic Gradient Descent.
SIFT	Scale Invariant Feature Transform.
SLC	Single Look Complex.
SRTM	Shuttle Radar Topography Mission.
TomoSAR	Tomographic SAR.
UTM	Universal Transverse Mercator.
VGG	Visual Geometry Group.
VHR	Very High Resolution.

1 Introduction

1.1 Motivation

Three-dimensional (3-D) building models are widely used in public and commercial sectors for environmental researches and location-based services, such as urban planning, change detection, telecommunication, solar potential analysis, driver assistance systems, virtual tourism, and many others [1]. Despite the importance, regional or even national figures of building models are hardly available or accessible. Information on the third dimension, i.e., building height, is especially limited. For instance, as of June 2021, OpenStreetMap (OSM) has recorded 458.97 million buildings, but only 2.84% of them are tagged with height information [2]. In addition, building height information in OSM is predominantly mapped in developed regions, e.g., Europe, North America, and Japan, that the spatial distribution is highly unbalanced (cf. Figure 1.1). Improving the spatial coverage of baseline geospatial data, including building heights, is not only important for emergency preparedness and prevention but also for overcoming the data gaps caused by socio-economic inequalities.



(a) Buildings (count: 458 976 276) **(b)** Building heights (count: 13 059 691)

Figure 1.1: Geographical distribution of the (a) buildings and (b) building heights recorded in OSM as of 2021/06/06 (data source: [2]).

For the past three decades, 3-D building reconstruction has been a hot topic in remote sensing [3]. Studies on building height retrieval are primarily conducted using high-resolution optical images and airborne or terrestrial LiDAR data [4]. Optical data acquisition requires the weather to be cloud-free, and airborne or terrestrial data are too expensive to collect globally. Synthetic Aperture Radar (SAR) imagery, on the other hand, is capable of providing data independently of sun illumination and insensitively to weather conditions. Such data are of great interest to applications of disaster responses [5, 6] and studies concerning regions frequently covered by clouds [7]. Since the launch of TerraSAR-X in 2007, modern SAR satellites, e.g., TerraSAR-X, TanDEM-X, and CosmoSky-Med, have been providing meter or even sub-meter resolution images, making it possible to extract and reconstruct man-made objects from spaceborne SAR data. Complete global coverages of TerraSAR-X/Tandem-X stripmap mode data have been acquired since 2012, providing great potential as a data source for global building reconstruction [8].

The study of building analysis from SAR imagery dates back to 1969, that Laprade and Leonardo derive the elevation of few buildings from shadows and layovers using simulated radar images [9]. Since then, various studies have been conducted on this topic [10–16]. However, building interpretation from SAR data is highly challenging. Due to the side-looking geometry and one-band radar sensors, urban structures are clearly visible in SAR images but are difficult to distinguish from each other. In literature, the performances of most methods are presented for a small set of test data, usually comprising one or a few buildings. Large-scale building reconstruction from SAR data is only achieved using techniques based on SAR tomography (TomoSAR) [16, 17], which often requires tens of images that are unavailable on a global scale. To date, no single study investigates large-scale building reconstruction from single SAR imagery.

In recent years, deep neural networks have become increasingly popular and have largely impacted both academia and industry. Many fields have witnessed deep learning-triggered breakthroughs, including building footprints generation towards the global scale [18–20]. Yet, the research related to building reconstruction from SAR data has not been pushed forward much, primarily due to the lack of annotation data.

Motivated by the demand for large-scale building models, the opportunity of using high-resolution SAR images, and the interest of applying the state-of-the-art techniques, this dissertation aims to reconstruct large-scale Level of Detail (LoD)-1 building models from single SAR imagery.

1.2 Research Objectives

This dissertation addresses the problem of large-scale LoD1 building model reconstruction from single SAR imagery. Considering the characteristics of buildings in SAR images, building footprints are introduced as complementary data, and deep neural networks are employed for large-scale reconstruction.

Five sub-objectives are defined towards large-scale building model reconstruction:

1. Develop a framework that automatically registers building footprints to corresponding SAR images on a large scale to enable building footprints to assist in the task of building reconstruction from SAR data.

2. Develop a workflow that automatically generates annotation data sets to enable the employment of supervised methods for building analysis using SAR data, specifically the state-of-the-art deep learning networks.

3. Develop deep neural networks suitable for our task to enable individual building analysis from single SAR imagery.

4. Investigate the performance of the developed deep neural networks in multiple areas to ensure the proposed algorithm can be generalized to more regions towards regional or even global reconstruction.

5. Investigate the impact of positioning errors in building footprint data on the proposed networks, i.e., if open-sourced building footprint data such as OSM can be exploited for individual building reconstruction in SAR images.

1.3 Thesis Outline

The remaining part of the thesis proceeds as follows: Chapter 2 provides a brief background for understanding the topic, whereas Chapter 3 reviews previous studies relevant to the thesis and summarizes the key contributions. Chapter 4 is concerned with the approaches of generating data sets used in the thesis to tackle the problem of dataset scarcity. The developed algorithms are detailed in Chapter 5 to Chapter 7, including a framework for registering building footprints to SAR images, a segmentation network for extracting areas of individual buildings, and a bounding box regression network that predicts bounding boxes of buildings in order to retrieve building heights. Chapter 8 concludes the thesis and looks into the future research and application directions.

2 Background Theory

This chapter starts with a brief introduction of SAR principles and the characteristics of buildings in SAR images, to provide a background for understanding the challenges involved in the topic, the suitable choice of SAR data for building reconstruction, and the need to employ complementary data. This chapter then presents the concept of LoD1 building models, which is the reconstruction target of the thesis. Finally, a short introduction of building footprints is given, including the data sources and the error sources when used as complementary data in SAR image interpretation.

2.1 SAR Basics

2.1.1 SAR Imaging

By illuminating the scene of interest with electromagnetic signals, a conventional SAR measures the backscattering coefficients of targets in a two-dimensional image coordinate system, with one along-track direction known as azimuth direction and one cross-track direction known as (slant) range direction.

In the range direction, the SAR system measures the time for a radar pulse transmitting to the target and returning to the radar. Along the range direction, targets are distinguished by measuring the runtime of their reflected echos, and the corresponding range resolution ρ_r is determined by the bandwidth B of the transmitted signals and the speed of light in vacuum c [21]:

$$\rho_r = \frac{c}{2B} \tag{2.1}$$

In the azimuth direction, its resolution is limited by the length of the physical antenna and the distance from the sensor to the illuminated scene. Through the forward movement of the sensor and coherent processing of the reflected echos, the synthetic aperture in the azimuth direction can be built up. By doing so, SAR system can greatly improve the azimuth resolution, which is only dependent of its physical antenna size d_A:

$$\rho_{az} = \frac{d_A}{2} \tag{2.2}$$

2.1.2 TerraSAR-X Imaging Modes

TerraSAR-X is the first German operational radar satellite mission [62]. It was launched into orbit in June 2007 and has been fully operational since January 2008. With its active radar antenna, TerraSAR-X is able to record images with different swath widths, resolutions, polarizations, and incidence angles, allowing for research perspectives in various fields, including hydrology, geology, oceanography, or ecology [63].

Since the launch, four SAR imaging modes have been operationally available, including two spotlight modes (SL and HS) with azimuth resolutions down to 1.1 m, a four-beam

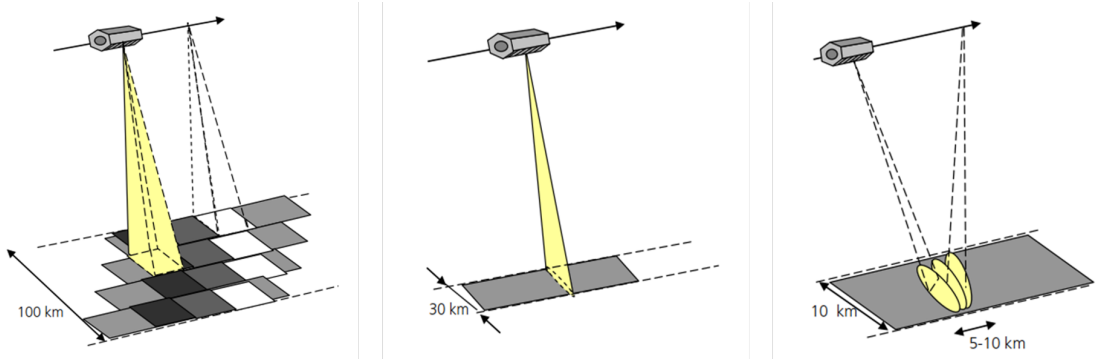


Figure 2.1: TerraSAR-X basic imaging modes [24].

Table 2.1: TerraSAR-X imaging modes' main characteristics.

Mode	**Swath width** (km)	**Azimuth scene size** (km)	**Azimuth resolution** (m)	**Full performance incidence angle** (°)
ScanSAR (*six beams*)	≥200	200	40.0	20-45
ScanSAR (*four beams*)	100	150	18.5	20-45
Stripmap (*SM*)	30	50	3.3	20-45
Spotlight (*SL/HS*)	10	10/5	1.7/1.1	20-55
Staring Spotlight (*ST*)	≈5	≈2.5	0.24	20-45

scanSAR (SC) mode with a range coverage of 100 km, and a stripmap (SM) mode with medium resolutions and medium-scale coverage. Two new modes have been added since autumn 2013 [22,23], including a wide ScanSAR mode with a range coverage of more than 200 km and a staring spotlight (ST) mode with an azimuth resolution of 0.24 m. The operational TerraSAR-X imaging modes and their important characteristics are summarized in Figure 2.1 and Table 2.1.

Figure 2.2 illustrates the scene coverage of different SAR imaging modes, with an example of the area around Munich plotted on an optical image from Google Earth. The SAR image in the ScanSAR mode is excluded as its lower resolution is unsuitable for analyzing meter-sized objects.

2.1.3 Radiometric Effects and Geometrical Distortion

The pixel intensity in a SAR image is related to the roughness, electrical conductivity, and orientation of the object relative to the sensor [25]. A particular radiometric effect of SAR imagery is speckle, which is commonly visible in areas where surface roughness is comparable to the used wavelength of the radar [26]. Speckle arises due to multi-path effects and multiple scatterers inside one resolution cell; those signal reflections from all scatterers are coherently summed to represent one scattered signal reflected from the corresponding resolution cell. The resulting amplitude of one resolution cell thus depends on the physical characteristics of scatterers and constructive and destructive phase interaction from contributing scatterers. Speckle can be mitigated by multi-looking [27] or nonlocal filtering algorithms [28,29].

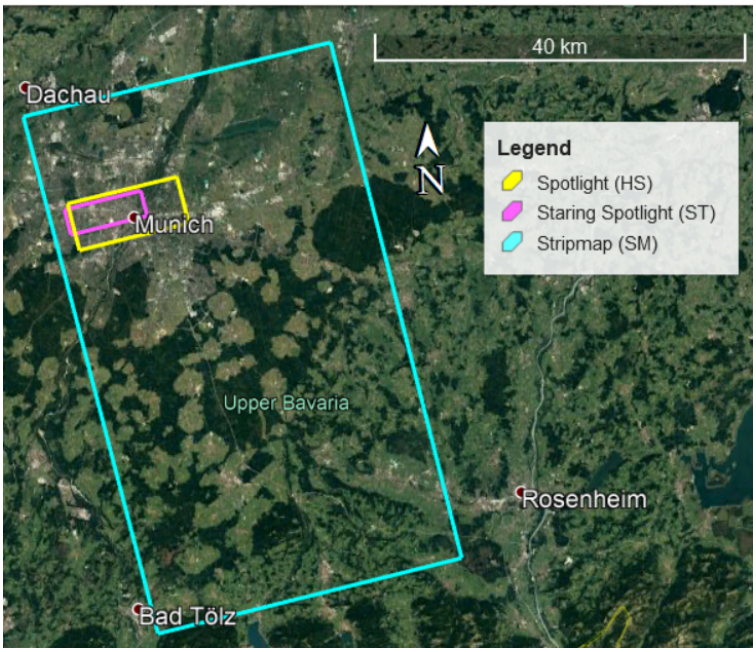


Figure 2.2: Demonstration of the scene coverage of different SAR imaging modes.

Conventional SAR imagery provides a projection of the 3-D object reflection to the 2-D azimuth-range plane. Due to the side-looking imaging geometry, this projection introduces three types of geometric distortion, i.e., layover, foreshortening, and shadowing. These geometric distortions are pervasive in urban and mountain areas, impacting the appearance of SAR images and complicating the interpretation of SAR images [25]. Figure 2.3 illustrates the foreshortening, layover, and shadowing effects on two isolated buildings. Different geometric distortions occur depending on the conditions between the incidence angle θ and the depression angle γ of the SAR sensor, the slope angle toward radar α, and the slope angle away from radar α':

a. $\theta < \alpha$: *Layover*

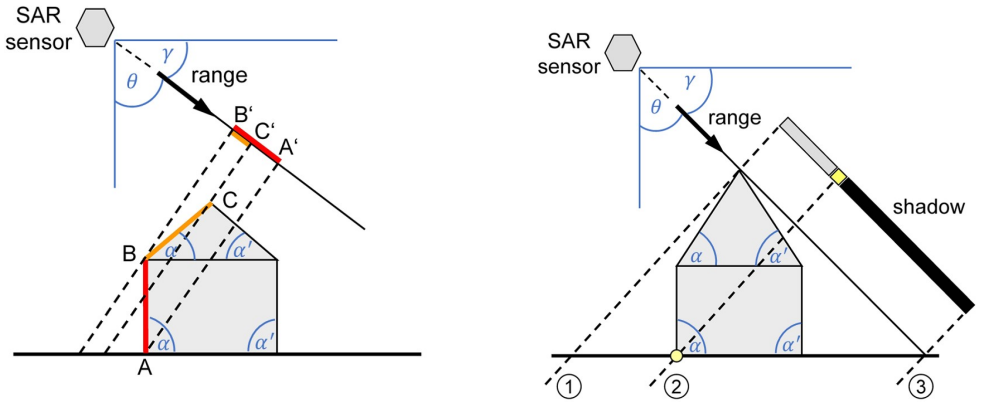
   In this case, multiple scatterers located at the same distance with respect to the sensor are mapped to the same azimuth-range pixel of the SAR image. The elevated objects are projected towards the sensor and appear bright in the SAR image.

b. $\theta > \alpha$: *Foreshortening*

   In this case, the distance between two points is shortened when projected onto the slant range direction. Foreshortening causes backscattering energy concentrated in smaller regions on SAR images, which therefore appear to be brighter.

c. $\gamma < \alpha'$: *Shadowing*

   In this case, there is no direct line-of-site from the sensor to the objects. Regions affected by shadowing appear dark in the SAR image.



(a) foreshortening (BC mapped to $B'C'$), layover(AB mapped to $A'B'$)

(b) shadowing

Figure 2.3: Three geometrical distortions in buildings in SAR data: foreshortening, layover, and shadowing. θ, γ, α, α' are the incidence angle of the SAR sensor, the depression angle of the SAR sensor, the slope angle toward radar, and the slope angle away from radar, respectively. Modified from [30].

2.2 Buildings in SAR Images

2.2.1 Components of a Building and the Backscatter Contributions

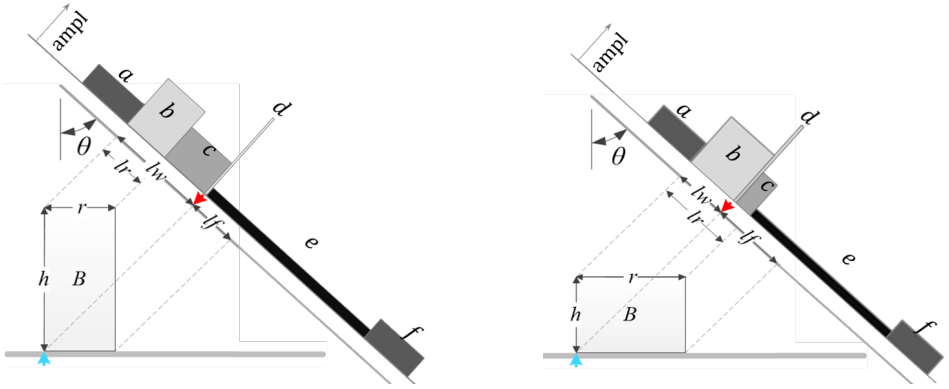
A building is composed of its roof, walls, and footprint. In orthorectified optical images where vertical walls are invisible, it is commonly agreed that reconstructing targets of buildings are the footprints. However, due to the side-looking geometry, the SAR sensor illuminates roofs, the walls facing the sensor, and the near-range side of footprints. Each part of the building shows distinct signatures in SAR images, and their appearances vary from building to building, primarily depending on the illuminate condition and the geometry of the buildings. Here, a general analysis is given under a simplified condition.

Figure 2.4 shows the projection geometry and the backscattering profiles of two flat-roof buildings in a slant-range SAR image. The buildings are both in rectangular shapes with uniformed surfaces and flat surroundings, and one is high-rise and the other is low-rise. The blue arrow marks the bottom of the sensor-facing wall, while the red arrow points at the double bounce line position on the SAR image. *lw*, *lr*, and *lf* denote the area of wall, roof, and footprint in the SAR image, respectively. The gray shades and heights of regions *a*-*f* denote expected magnitude values of intensity on the SAR image: *a* and *f* are areas that only contains backscatter from the ground; *b* is the layover area that has backscatter contributions from the ground, the front wall, and the roof; *c* results from the sum of signal returns from the ground and the roof; *d* marks the double bounce caused by the corner reflector at the intersection of the wall and the ground; and *e* is the shadow area where there is no signal return.

It can be seen that in SAR images, areas of walls, roof, and footprint of a building have different backscatter contributions:

a. *Wall*

   In Figure 2.4, the building wall *lw* covers areas *b, c, and d* for the high-rise building, and *b, d* for the low-rise building. Walls are typical layover areas that appear bright



(a) high-rise building **(b)** low-rise building

Figure 2.4: Illustration of the projection geometry and the amplitude profile (ampl) of two flat-roof buildings in a slant-range SAR image. θ is the incidence angle. h is the building height. *lw*, *lr*, and *lf* denote the area of wall, roof, and footprint in a slant-range SAR image, respectively. The gray shades and heights of regions *a-f* indicate expected magnitude values of intensity on the SAR image. The blue arrow marks the bottom of the sensor-facing wall and the red arrow points at the double bounce line on the SAR image.

in SAR images. For low-rise buildings, the wall area is covered by the roof area. Thus it might be difficult to extract.

In very high resolution (VHR) SAR images, e.g., spotlight TerraSAR-X images, parallel bright lines are often observed in wall areas (cf. Figure 2.5 (c)(d)). These bright-line signatures are referred to as corner lines [31] originated from multiple signal reflections on corner structures such as wall-ground intersections and corners on windows or balconies. Due to regularly arranged structures on building walls, i.e., windows and balconies, corner lines often form regular patterns such as parallelograms. Corresponding to walls, these parallelograms have one pair of opposite sides parallel to the slant range direction of the SAR image.

b. *Roof*

In Figure 2.4, the roof area *lr* is the area *b* for the high-rise building and *b, c, d* for the low-rise building. The magnitude of intensity value on roof area varies on SAR images, depending on the surface roughness and the number of structures on the roof. For high-rise buildings, the roof area If overlaid by wall areas thus is often hardly distinguishable.

c. *Footprint*

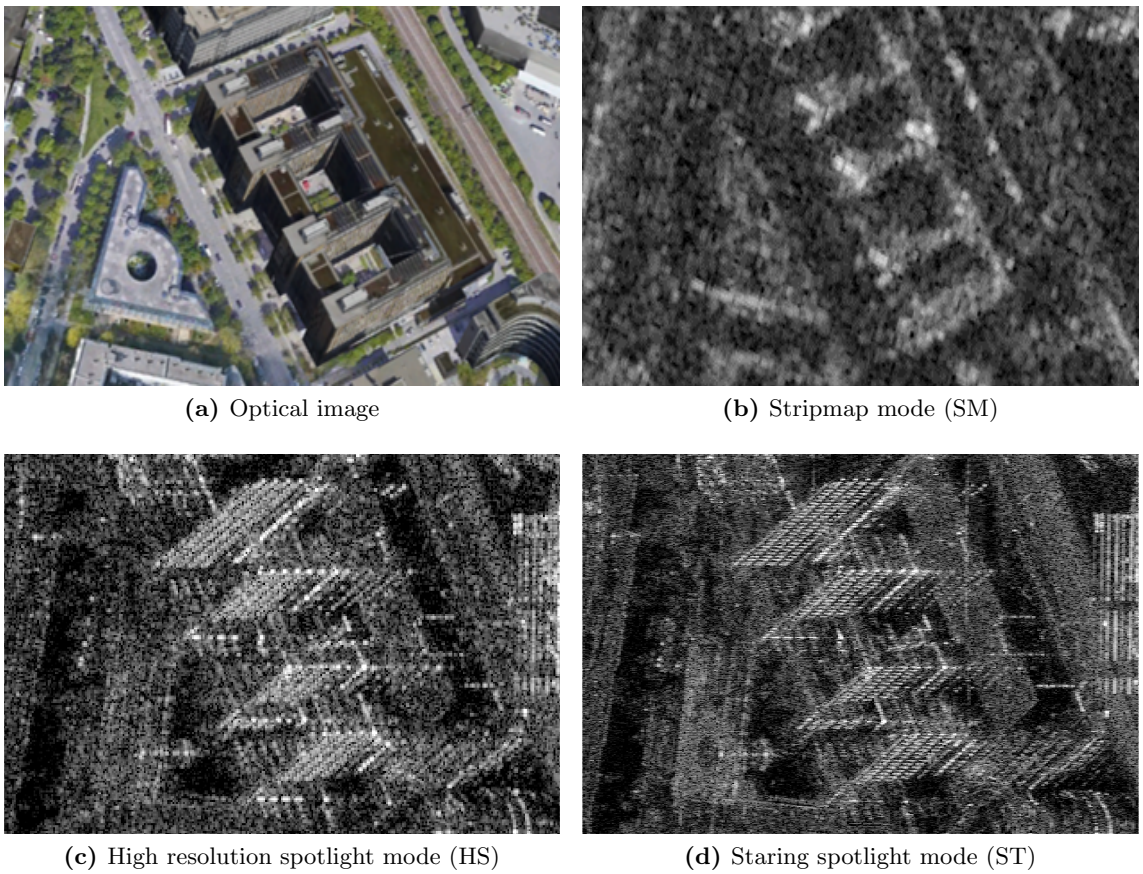
In Figure 2.4, *d* denotes the visible side of the building footprint *lf*. The bright double-bounce line is caused by strong signal responses by double-bounce scattering. The other side of the footprint connects the building shadow, which appears as a dark region *e*. For low-rise buildings, the double bounce line may not be clear if the signal from the roof is strong.

In addition, the appearances of the illuminated building components in SAR images depend on the physical properties, such as the orientation of buildings towards the sensor, roof shapes, facade structures, surface roughness, material types, etc.

2.2.2 Buildings in SAR Images of Different Imaging Modes

The appearance of a building varies in SAR images of different imaging modes, due to the changes in resolution. An example is shown in Figure 2.5. In the StripMap image (b), the building is visible without many details; in the high-resolution spotlight image (c), individual walls of the building can be recognized, indicated by the bright parallel lines; in the staring spotlight image (d), mode details are noticeable on the azimuth direction (vertical direction of the figure), e.g., the bright lines in the HS spotlight image appear to be groups of dots representing structures on the walls, such as windows.

For the analysis of individual buildings, the resolution of SAR images should be at least in meters. As for TerraSAR-X, StripMap data meet the minimum requirements.



(a) Optical image **(b)** Stripmap mode (SM)

(c) High resolution spotlight mode (HS) **(d)** Staring spotlight mode (ST)

Figure 2.5: A building shown in (b) Stripmap, (c) high resolution spotlight, and (d) staring spotlight SAR images. (a) shows the building in an optical image as a reference.

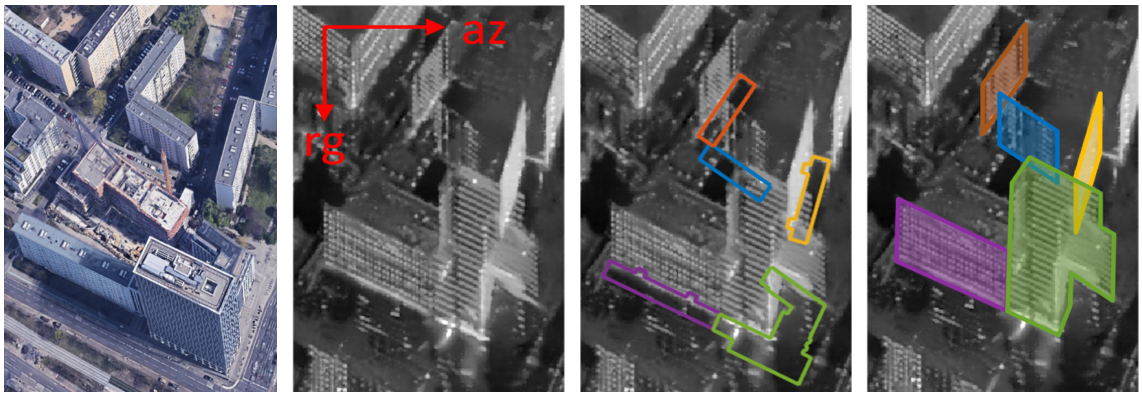
2.2.3 Challenges of Building Analyses in SAR Images

The variations of the appearances of buildings on SAR images and the occlusions between different buildings caused by the side-looking geometry pose challenges for the analysis.

For *an isolated building* in SAR images, the main interpretation challenge is to distinguish its components. As shown in Figure 2.4, the wall area l and the roof area r in SAR images are always mixed and difficult to differentiate. l covers r when the building height h is large, e.g., the case in Figure 2.4 (a), and it is covered by r when h is small, e.g., the case in Figure 2.4 (b). In addition, the roof area r overlaps the footprint area f for low-rise buildings, e.g., the case in Figure 2.4 (b). In this case, the near-range side of building footprints might be ambiguous. Moreover, the far-range side of building footprints is unknown since the footprint area connects the shadow area that also appears dark in SAR images.

For *multiple adjacent or nearby buildings*, the additional issue is to identify them. This is more crucial. In optical images, with multi-spectral band usage and often the nadir-looking geometry, the boundaries of different urban objects are clearly depicted. However, with the side-looking geometry and one-band radar sensor, the urban structures are clearly visible in SAR images but are difficult to distinguish from each other.

The intensity values in SAR images are closely related to material types and structural shapes of objects unless, in the presence of obvious material or structure changes at building boundaries, consecutive buildings in the physical world are difficult to be separated from each other in a SAR image. In addition, even if buildings in the real world are not neighboring, they may overlap with each other in the SAR image, which significantly increases the difficulty of image interpretation. Figure 2.6 shows a typical urban area in (a) an optical image and (b) a VHR SAR image. (c) and (d) show footprints and regions of buildings in the SAR image marked with different colors, respectively. In the area, the purple building and the green building are connected, and the green building has a more complex shape. From the SAR image itself, it is unlikely to tell if there are two or three buildings or only one. Besides, it is noticeable that the green building overlaps the yellow building and the blue building in the SAR image, although their footprints are not connected.



(a) Optical image **(b)** SAR image **(c)** Footprints **(d)** Building areas

Figure 2.6: A typical urban area in a SAR image (b). The corresponding optical image of the area is given in (a) as references. In (c) and (d), footprints and the corresponding building regions in the SAR image are marked in different colors for reference. rg and az denote the range direction and azimuth direction, respectively.

2.3 LoD1 Building Models

LoD1 building models are the reconstruction target of this thesis.

2.3.1 LoDs: Levels of Detail for Building Models

When speaking of building model reconstruction, the first thing to clarify is how detailed the models should be. Reconstructed building models can be presented at different levels of detail, differing from the simple prismatic model (LoD1) with single roof surface to detailed model with overhangs or balconies (LoD3), even with the interior of buildings (LoD4) (Figure 2.7), according to the official OGC standard City Geography Markup Language (CityGML) [32], an information model intending a standardized 'representation, storage, and exchange of virtual 3-D city and landscape models'.

The LoD should fit user's requirements and acceptance criteria of 3-D building models, as well as the quality of used data. In practice, usually, the larger the study area is, the coarser the building models are. And vice versa. In studies concerning local areas, most authors' definition agrees to the definition of LoD2. As for large-scale reconstruction, LoD1 building models are often the reconstruction goal.

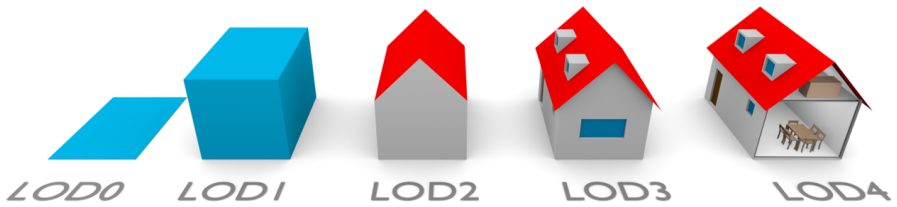


Figure 2.7: The five LoDs of building models in CityGML. From left to right, the geometric detail and the semantic complexity increase, ending with the LoD4 containing indoor features. [33]

2.3.2 The LoD1 Building Model and its Subsets

The LoD1 model is the coarsest volumetric representation of buildings, which are described as blocks models comprising prismatic buildings with flat roof structures or a set of extruded polygons that comprise a volume defined by a base height from which extrusion begins and an extrusion distance [34]. LoD1 models are usually derived with extrusion to a uniform height [35, 36] and generalization from finer LoD models [37, 38].

Providing a relatively high information content and usability compared to their geometric detail [39, 40], they are widely used in various applications, such as energy demand estimation [41], noise pollution estimation [42], floods simulation [43], shadowing simulations [41, 44, 45] and so on.

Biljecki *et al.* [33] further refine the definition of LoD1 models in four subsets, as shown in Figure 2.8:

a. *LoD 1.0:* all buildings larger than 6 m should be acquired, and neighboring buildings may be aggregated.

b. *LoD 1.1:* buildings must be individually modeled, and all large building parts shall be acquired.

c. *LoD 1.2:* smaller building parts and extensions should be acquired, and buildings are extruded to a single height.

d. *LoD 1.3:* multiple rooftop surfaces should be reconstructed if their differences are higher than a threshold, e.g., 2 meters.

In practice, the LoD1.2 is the most commonly used LoD1 model.

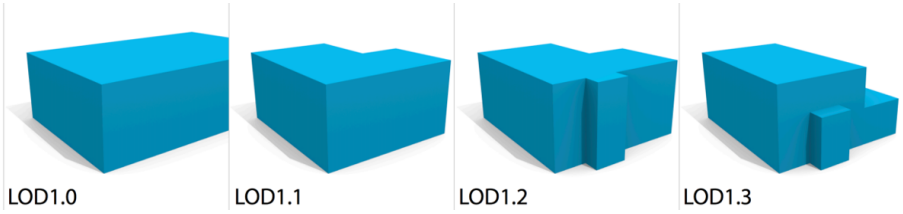


Figure 2.8: Refined LoD1s of 3-D building models. [33]

2.4 Building Footprints as Complimentary Data

Building footprints are 2-D polygonal representations of buildings, containing geometric information of buildings such as location, shape, and distribution patterns of structures in built-up areas, and often additional attributes, including building function, housing type, and building heights.

Building footprints are used by many researchers to aid SAR image interpretation. The used data are often extracted from optical images [14, 46] or obtained from GIS data [15,47,48]. In aiding individual building analysis, building footprint data provide geometric information, such as building outlines, and, more importantly, spatial relationships that are beneficial for identifying individual buildings in SAR images, which is highly challenging, as introduced in the previous sections. An example is shown in Figure 2.6, which clearly shows the advantage of using building footprint data in SAR image interpretation in urban areas.

2.4.1 Data Sources

The data sources of building footprints are mainly threefold:

a. *Official data sets from government agencies*

These data sets mostly exist in cities of developed counties, e.g., detailed building models of New York City [49] and Berlin [50]. Well documented and up to date, these data contains various attributes such as building type, usage, age or number of floors, and support activities for surveying, urban planning, and statistical offices.

b. *Crowdsourced mapping projects*

These projects are mostly conducted in less developed areas where no existing data sets are available. For instance, Humanitarian OpenStreetMap Team (HOT) [51] organizes mapping campaigns assuring high data quality of Tanzania, in which buildings are manually traced from satellite images, attributes are added by local volunteers.

c. *Building polygons extracted from optical imagery*

This approach is increasingly popular in recent years due to the advances in deep learning algorithms and the processing power of GPUs [19, 20, 52]. These methods are able to process large coverage data with consistent data quality and therefore have great potential of providing building footprint data sets on a global scale. These developments also contribute to open data sources. For example, Microsoft Building footprints are openly available in the USA, Canada, Tanzania, Uganda, and Australia to increase the coverage of building footprints available for OpenStreetMap [18]. This type of data often lacks attributes on building type, usage, and so on.

2.4.2 Error Sources in Radar Coding

The joint use of building footprints with SAR images requires precise registration of these two data on a large scale. The projection process from a geographic coordinate system to a SAR image coordinate system is referred to as radar coding. Given its geographic coordinates, a building footprint polygon should be radar-coded correctly to the SAR coordinate system with imaging acquisition parameters using Doppler-Range-Ellipsoid equations [53, 54]. However, in practice, there exist three problems preventing accurate registration, namely, the geometric accuracy of the SAR image, the positioning accuracy of building footprints, and the height accuracy of the terrain used in radar coding.

a. *The geometric accuracy of SAR images*

The geometric accuracy of SAR data mainly depends on the orbit accuracy and radar timings [55]. The SAR data used in this work are TerraSAR-X / Tandem-X images with accuracy at centimeters to decimeters level [55, 55–58] so that the geometric error of the data is negligible.

b. *The positioning accuracy of building footprints*

The positioning accuracy of building footprints mainly depends on the data collecting and processing methods [59]. Official data sets provided by government agencies and commercial sources have good quality control, while Volunteered Geographic Information (VGI) data often contain positioning errors. For instance, a quality assessment study [60] shows that the average offset of building footprints in OpenStreetMap (OSM) is 4.13 m with a standard deviation of 1.71 m. In general, for individual building analysis, data with at least meter level geometric accuracy are preferred and should be chosen when possible.

c. *The height accuracy of the terrain used in radar coding*

Building footprint data usually contain 2-D geo-coordinates but not accurate heights at ground level. Due to the lack of the accurate terrain models required in coordinate projection, precise registration of 2-D building footprints and SAR images at a large scale is often difficult. As illustrated in Figure 2.9, a terrain height error (δH) leads to a radar coding error (δL). The radar coding error is proportional to the height error and the incidence angle ($\delta L = \delta H cos\theta$, where θ is the incidence angle). For TerraSAR-X, the incidence angle usually ranges from 20° to 55° [24]; Thus a height error of 10 meters results in a slant range error of 5.73 meters to 9.39 meters. The height errors are usually inconstant over the observed area by the SAR sensor; hence, so are the range errors.

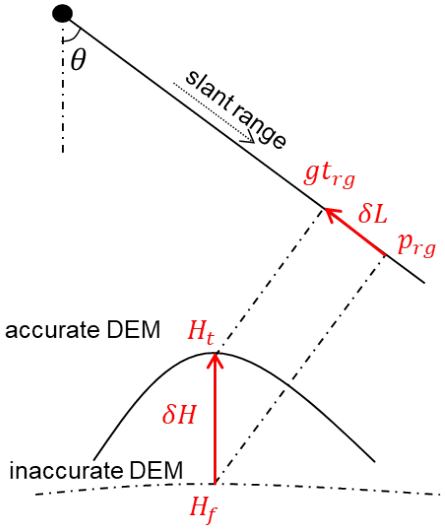


Figure 2.9: The geocoding error from inaccurate height: H_t and H_f are the accurate height and inaccurate height of a target point respectively, while gt_{rg} and p_{rg} are the corresponding accurate and inaccurate slant range. The height error $\delta H = H_t - H_f$, and θ is the incidence angle. The resulting error in slant range $\delta L = gt_{rg} - p_{rg} = \delta H cos\theta$.

3 State of the Art

This thesis aims to reconstruct LoD1 building models from a SAR image on a large scale. To achieve this objective, building footprints are introduced as complementary data, and two deep learning networks are proposed for individual building segmentation and height estimation, respectively. This chapter reviews previous studies on the topic of building reconstruction from SAR data as well as the relevant algorithms in registration and deep learning. A summary of key contributions of the thesis ends this chapter.

3.1 Related Studies on Building Reconstruction from SAR Data

The study of building analysis from SAR imagery dates back to 1969 [9]. Using simulated radar images, Laprade and Leonardo derive the elevation of few buildings from shadows and layovers based on the radar viewing geometry for the first time [9]. Since then, a considerable amount of literature has been published on the topic: for a better understanding of radar backscatter from urban areas, SAR simulators are developed to model electromagnetic scattering from buildings [61–67]; for building reconstruction, various methods are developed investigating building features in SAR data, such as layovers, shadows, double-bounce lines, and InSAR phases [68–74]; for exploiting information from different modalities, optical images and GIS data are introduced to assist in building reconstruction from SAR data [15, 71, 75, 76].

Over the past two decades, studies have been conducted using different SAR data, focusing on different research objectives, and employing different approaches. In this section, the literature is reviewed from these three aspects. In the end, the limitations of the existing methods are summarised.

3.1.1 Building Reconstruction Using Different SAR Data

In literature, building reconstruction is conducted using a single SAR image, multiple SAR images, and additional auxiliary data.

a. *Single SAR imagery*

   Building reconstruction from single SAR imagery is of great practical interest, especially in applications with stringent temporal restrictions such as emergency responses.

   With single SAR imagery, most researchers rely on salient features such as layover, shadow, and corner lines. Laprade and Leonardo [9] analyze shadows and layovers from the radar viewing geometry and estimate building heights. He *et al.* [77] employ the mean shift algorithm and conditional random field (CRF) to segment fully polarimetric SAR (PolSAR) data to shadow, layover, and other areas to extract buildings.

Zhao *et al.* [78] segment buildings based on the marker-controlled watershed algorithm. Cao *et al.* [79] employ thresholding and morphological operations to segment bright patches. The extracted shape features are employed to remove false alarms, and large patches are split using a controlled watershed algorithm to determine the number of buildings. In [80], Ferro *et al.* first extract a set of low-level primitives of bright lines and dark shadow areas, then produce a building footprint hypothesis by combining and assigning these primitives to building component classes using a fuzzy membership function. In [81], Chen *et al.* introduce a 1-D range detector to scan the range direction to detect buildings by their distinct profile. Then, along the azimuth direction, the detected building walls are processed and add to the building. In [82], Liu *et al.* develop a bottom-up/top-down hybrid algorithm. This work first extracts rectangles of different intensity levels as primitive features representing building components. Then, the bottom-up stage proposes building candidates using the extracts rectangles, and the top-down step checks the rest rectangles to predicts additional building candidates. All candidates are verified through false alarm detection to reconstruct buildings.

Some researchers reconstruct buildings in a simulating and matching manner. Jahangir *et al.* [83] simulate the shadow of a house model and continuously adjust the house parameters until the optimum delineation of the images is achieved. Brunner *et al.* [13] simulate a SAR image with pre-defined initial building heights and then match the simulated SAR image to the real SAR image. The procedure is conducted iteratively to optimize building heights. In [84], Wang *et al.* develop a similar approach. In addition to isolated buildings, this work analyses the sensor-building illumination geometry for partially occluded buildings in the simulating and matching procedure. Therefore it is able to estimate heights for both types of buildings.

Several studies perform geometrical and radiometric properties analysis on isolated buildings. Quartulli and Datcu [85] propose stochastic geometrical modeling of buildings from a high-resolution SAR image. Buildings are hierarchically modeled as a collection of radiometrically and geometrically specified object facets, and a probabilistic model is used to optimize the model parameters. Parallelepipedal and round tower objects are considered. Guida *et al.* [12, 86] estimate building shapes in SAR data via detailed modeling of their electromagnetic properties. Such techniques need extensive prior knowledge about objects, such as materials, roughness, humidity, and orientation with respect to the SAR sensor, which is generally unknown. Thus these methods serve as a tool for scene understanding rather than for building reconstruction.

b. *Multiple SAR images*

Many researchers employ multiple SAR images for building analysis, exploiting techniques including InSAR, multi-aspect SAR / InSAR, and TomoSAR.

b.1. *SAR Interferogram (InSAR)*

The opportunity to extract buildings from InSAR data has attracted the attention of many researchers.

Earlier studies make use of the InSAR DEM. Buildings are identified by horizontal roof planes in filtered InSAR DEM. Hoepfner [87] detects the far end boundary of

a building in the InSAR DEM and extracts the elevated roof area using a region growing algorithm. A rectangle is then fitted to the area representing the building. Gamba *et al.* [88] group pixels belonging to the same scan line and employ region growing for segmentation. Buildings are extracted by fitting planar surfaces to the segments.

With InSAR, researchers may exploit coherence and interferometric phase in addition to the amplitude of SAR images. Most of the work relies on building signatures in both the amplitude and phase of SAR data. In [89, 90], Bolter and Leberl estimate building footprints from shadows, then estimate the height from interferograms. Cellier *et al.* [91] first extract single/double bounces and shadows and then detect the front and backside of the building. Heights are derived from shadows and are compared to interferometry. Soergel *et al.* [92] first propose quadrangular building candidates by extracting building-shadow boundaries and shadow-ground boundaries, then calculate building heights from the InSAR DEM in the building candidate region. Thiele *et al.* [74] investigate the shadow and the phase ramps in an interferogram for detecting isolated buildings, based on the fact that rectangular buildings appear as parallelograms in SAR images whose two sides parallel the slant range direction. Dubois *et al.* [93] compute a box to bound a building wall first, then the box is iteratively sheared and scaled to build a parallelogram on InSAR phases. The best-fitted parallelogram is selected to represent the layover region, and building parameters are subsequently computed.

Other methods include stochastic geometry modeling [94, 95] and a fusion scheme for joint retrieval of heightmap and classification [96]. Quartulli and Dactu [94, 95] employ stochastic geometry and used a probabilistic model to optimize model parameters of buildings with gabled roofs, such as the slope of the roof, the length, width, and the position of the buildings. In [96], Tison *et al.* first process the interferogram, amplitude, and coherence images with different operators, e.g., classification, filtering, and structure extraction, and then employ a Markovian framework to retrieve an improved classification and a height map jointly.

b.2. *Multi-aspect SAR / InSAR data*

The side-looking geometry brings in occlusions in SAR images. To overcome this drawback and obtain complete information, researchers resort to data sets of multiple aspects. With SAR images from at least two different viewing angles of the same scene, the stereo principle is applied to SAR imagery. Depending on the flight orbits, the SAR images could be acquired from the same side or the opposite side of the scene [97, 98], from orthogonal flight paths [72, 99, 100], or even from circular trajectories [101, 102]. In published studies, multi-aspect SAR data are mostly acquired from airborne platforms as their orbits are more flexible than spaceborne satellites.

Often, building features are extracted from SAR images of each aspect, and stereo analysis is subsequently applied to identify and group corresponding information from different aspects to reconstruct building models. Simonetto *et al.* [103] investigate same-side SAR stereo. Bright L-shaped angular structures, often caused by double-bounce at buildings, are extracted and matched as features. Soergel *et al.* [99] determine the height of buildings from a pair of SAR images of orthogonal flight paths. The extracted bright lines are grouped into rectangular 2-D objects

and are subsequently matched in 3-D to reconstruct buildings. In [72, 100], Xu and Jin employ multi-aspect polarimetric SAR images from four orthogonal views and reconstruct full sides of buildings as cuboids. They mainly exploit layover areas of buildings in SAR images. Hough transform is used to identify parallel lines that are further analyzed in a probabilistic framework.

The concept of multiple aspects is incorporated with InSAR data as well. Xiao *et al.* [104] fuse four InSAR DEMs to obtain a heightmap to remove occlusion areas from one aspect. The DEM is segmented, and bounding boxes are fitted to the elevated structures representing buildings. Bolter [89, 90] uses the same data set and extends the work by analyzing shadows in the amplitude image to measure the building heights. In this work, the simulation technique is introduced to improve the understanding of the appearance of buildings in SAR and InSAR data. Soergel *et al.* [92] also employ multi-aspect InSAR data. Building candidates are first extracted from InSAR data of each aspect and then are fused to fill occluded areas. Using the building candidates, the scene is simulated and compared to the original SAR data iteratively for building reconstruction. Thiele *et al.* [105–107] analyze InSAR phase profiles and employ InSAR phase filters. Building reconstruction is supported by phase simulations of different building hypotheses and subsequent comparison of the simulated phases to the original InSAR phases. This same research team extends this approach to gable-roofed buildings in [73].

b.3. *TomoSAR*

SAR tomography (TomoSAR) aims at SAR imaging in 3-D or even higher dimensions by resolving the distinct scatterer contributions within one azimuth-range pixel of a conventional 2-D SAR image. It exploits a stack of multiple SAR images acquired from slightly different looking angles and reconstructs the 3-D position and the reflectivity of the scatterers. The first concept for 3-D imaging of volume scatterers using TomoSAR is presented in [108], and the first demonstration of spaceborne TomoSAR over a large urban environment is presented in [109].

A few studies have reconstructed buildings from TomoSAR data. In [17], using a spaceborne TomoSAR point cloud covering a large area, Shahzad *et al.* extract and reconstruct building walls. In [110], building roof shapes are extracted, and in [111, 112], multiple roof layers are reconstructed. Rambour *et al.* [113] reconstruct urban surfaces from TomoSAR point clouds using the graph-cut algorithm. Guided by building footprints, Rambour *et al.* [114] reconstruct 3-D buildings with SAR Tomography. D'Hondt *et al.* [115–117] demonstrate the building reconstruction approach from airborne TomoSAR point clouds.

A reliable TomoSAR point cloud reconstruction usually requires more than 20 SAR images. Recently, a non-local filtering method has been applied in TomoSAR [118] that reduces the required interferograms to 3–5 [119]. Combining building footprints, LoD1 building models can be reconstructed from the resulting TomoSAR point cloud.

c. *Auxiliary data*

Some researchers introduce auxiliary data to assist building reconstruction, e.g., building outlines extracted from optical images [14, 46] and footprint polygons obtained from GIS data [15, 47, 48]. Providing exact locations and geometric shapes of

buildings in the real world, footprints are highly beneficial for tasks concerning individual buildings in SAR images. An example of complex urban regions is illustrated in Figure 3.1.

In [71], Tupin estimates building heights of industrial buildings by analyzing overlay regions in a single SAR image and building outlines extracted from map data. Thiele *et al.* [47] combine building footprints with the InSAR phase to acquire building height to determine the building shapes post damage. Aided by building footprints, Liu *et al.* [120] analyses layover to estimate building heights using a single SAR image. In [48], the sensor-visible edges in building footprints are extracted to aid height estimation from a single VHR SAR image by analyzing the range profile.

Auxiliary data are also combined with InSAR data to better distinguish buildings from other elevated objects. Hepner *et al.* (1998) [121] use hyperspectral data to improve building extraction from an InSAR DEM. In [14], Sportouche *et al.* extract potential building footprints in an optical image and register them to a SAR image. Based on a log-likelihood function, building heights are retrieved through a joint optimization. Wegner *et al.* [46] extract buildings in dense urban areas using one single-aspect aerial InSAR data and one aerial image. Objects are first extracted from both the data and are then fused to obtain a building hypothesis. A threshold was set to filter only the best building hypothesis objects.

The aforementioned studies all require a precise registration of building footprint data and SAR images, which is challenging at a large scale. In fact, these studies are all conducted in a small region containing countable buildings.

3.1.2 Research Focuses

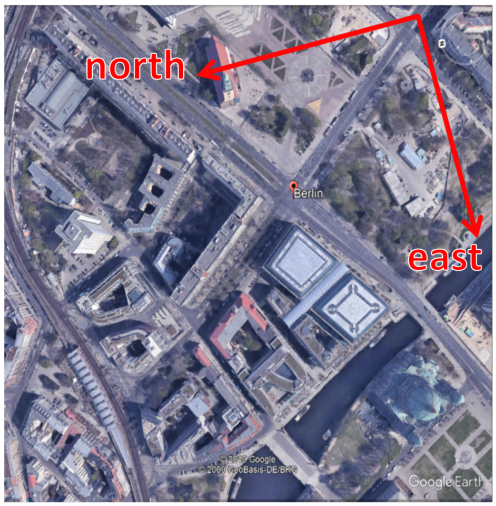
In orthorectified optical images, it is commonly agreed that buildings refer to building footprints, as vertical walls are invisible and roof outlines mark the area of building footprints. However, using side-looking SAR imagery, the research focus needs to be specified. Chapter 2 introduces building components, i.e., wall, roof, footprints, their corresponding areas in SAR images, and the backscatter contributions within each area. Studies detect/reconstruct building components depending on their signatures in SAR data.

a. *Height estimation / Wall reconstruction*[1]

   Height information is contained in the layover area, i.e., the wall area, and the shadow area. In literature, building heights are estimated by analyzing shadows [68–70] or layovers [71, 72]. Besides approaches based on these two features, model-based methods are used based on hypothesis tests for building height retrieval [13, 84]. A theoretical basis for understanding building areas in SAR images is provided in [12, 86], in which building heights are estimated based on detailed modeling of electromagnetic radar returns from isolated buildings.

   Techniques for building height retrieval from SAR imagery make use of various SAR data. Research in [9, 13, 78, 85] investigate radar viewing geometry and use the

[1] In this thesis, the term *wall* is used instead of the term *facade*. *Facade* implies an emphasis on the front wall and detailed structures on it, which is not the research focus of most related work.
Moreover, *facade reconstruction* is a different research topic that often employs terrestrial images or LiDAR data rather than SAR data.



(a) Optical image from Google Earth

(b) SAR image

(c) SAR image with building footprints

Figure 3.1: Center Berlin area (the Berliner Dom in the lower-right corner) in a SAR amplitude image with and without GIS building footprints.

layovers to measure building heights from a single SAR image. Methods in [74,92,93] make use of InSAR. Studies in [72, 97–100] employ SAR and InSAR data from multiple aspects and circular flight paths [101, 102].

Some approaches rely on building footprints extracted from optical or GIS data to support building height retrieval from SAR data [14, 15, 46–48].

b. *Building detection / Footprint reconstruction*

Caused by double-bounce scattering between building walls and the ground, bright-line features in SAR images indicate building locations. Note that not all buildings have double-bounce lines as the prominent feature. These lines are often exploited for detecting buildings and reconstructing building footprints.

Using SAR data from one aspect, only one side of the building footprint is visible. Therefore some authors extract L-shaped footprints, which are the visible side of the footprints [78,122]. Some researchers impose the rectangular assumption on footprint shapes [70, 82, 103, 123]. In [80], the authors extract and combine a set of low-level features to create building radar footprints. Employing multi-aspect SAR data, in [98, 99], bright-line segments and regular spaced point-like features are detected and subsequently grouped to building footprints. For footprint reconstruction, one crucial issue is how to assign the detected double bounce lines to different buildings and group them to building footprints. The underlying problem is distinguishing different buildings and deciding the number of buildings in the SAR data.

Recently, a multi-sensor all weather mapping (MSAW) dataset [124] is published for the purpose of building footprint extraction. Containing airborne SAR images, optical images, and building footprint annotations, this dataset promotes building footprint extraction from SAR images. More related studies are expected to be published in the future.

c. *Roof reconstruction*

Few studies extract building roofs. Chen *et al.* [125] extract low-rise buildings with gable roofs by segmenting parallelogram-like roof patches from a TerraSAR-X staring spotlight image. In SAR data, the appearance of roofs is highly dependent on physical properties, especially roof structures, building shapes, and orientations of buildings towards the sensor. Therefore, roof extraction often needs analysis on a case-by-case basis and thus is hard to generalize.

3.1.3 Reconstruction Approaches

Generally, the aforementioned approaches for building reconstruction from SAR data can be categorized into two classes: data-driven methods and model-driven methods. In addition, recent advances using deep learning networks are presented.

a. *Data-driven approaches*

Data-driven approaches extract building features and then deduce building parameters. Two solutions based on this methodology have been developed. The first one attempts detecting line- or point-like features first and extracting building regions based on these features. For example, in [11], feature lines are identified using a line detector, and layover areas are derived by extracting parallel edges; in [72], the authors exploit a constant false alarm rate (CFAR) edge detector for line feature detection and apply a Hough transform for parallelogram-like wall area extraction; in [98, 99], bright-line segments and regular spaced point-like features are detected and subsequently grouped to building footprints; and in [80], the authors extract and combine a set of low-level features to create structured primitives. The second solution directly extracts building regions using segmentation techniques, such as active contour [69], rotating mask [90], mean shift [126], and marker-controlled watershed transform [78].

b. *Model-driven approaches*

Model-based building reconstruction from high-resolution SAR images is conducted in [12, 85, 86]. With detailed modeling of isolated buildings' geometrical and radiometric properties, building shapes in SAR data are estimated. Requiring prior knowledge of the target building, such as materials, roughness, humidity, and orientation with respect to the SAR sensor, this method is difficult to be generalized for building reconstruction.

In other model-driven methods, a SAR image is simulated using geometric and radiometric hypotheses [13, 75, 84, 85, 127, 128]. The desired building parameters are progressively achieved by minimizing the difference between simulated and real data. This approach is applied to amplitude image [13, 84], InSAR data [89], and multi-aspect InSAR data [73, 92].

c. *Deep learning approaches*

In recent years, deep neural networks have become increasingly popular and shown success in remote-sensing data analysis, including a wide range of applications using SAR data, such as classification [129, 130], segmentation [131, 132], target recognition [133, 134], and change detection [135, 136]. Instead of relying on hand-crafted features, deep networks can learn effective feature representations from raw data in an end-to-end fashion.

For building footprint extraction, Shermeyer *et al.* [124] present an MSAW dataset containing airborne SAR images, optical images, and building footprint annotations, along with a deep network baseline model and benchmark.

For building area extraction, by introducing a TomoSAR point cloud, Shahzad *et al.* [137] are able to acquire accurate building areas in a SAR image and take them as ground truth annotations to train a segmentation network for building extraction. However, TomoSAR point clouds are rare, that this approach cannot be applied to other areas. To address this issue, in [138], a DEM is introduced to generate building areas in a SAR image. However, these two works cannot differentiate individual buildings.

3.1.4 Problems in Applying the Existing Methods to Large Areas

As the survey of related work shows above, most techniques have been proposed and addressed in the early years. More recent research mostly follows the existing paradigms. A few studies adopt new techniques of persistent scatterer interferometry (PSI) and TomoSAR and make innovative contributions. For instance, Gernhardt *et al.* [139] analyze PSI in urban areas, specifically on building facades; Zhu *et al.* [140] develop TomoSAR technique in an urban environment and building reconstruction from TomoSAR data [17]; Shahzad *et al.* [137] apply a deep segmentation network to extract building areas in a SAR image, and the training data are acquired with the help of TomoSAR point clouds.

Although extensive research has been carried out on building reconstruction using SAR data, to date, no single study exists which investigates large-scale building reconstruction from single SAR imagery.

Large-scaled building analysis in SAR data is restricted mainly by two reasons:

a. *Assumptions on buildings and the study scenario*

The majority of related studies are carried out on buildings with specific geometric shapes, e.g., rectangular- [82, 103, 123] or L-shaped footprints [78, 122], flat [76] or gable roofs [73,125], and different heights [125,141–143]. Only a few studies address the problem of complex-shaped buildings [98, 99].

In addition, most studies investigate simple scenarios where a minimal distance between buildings is required to ensure the scattering effects of different buildings do not interfere with each other [12–14, 81]. In complex scenarios, possible overlapping areas between two buildings are usually assigned to one building [15,144], which may cause incorrect estimations.

Moreover, the performance of the presented methods is typically presented for a small set of test data, usually comprising only one or a few buildings. The generalisability of most published research on this issue is problematic.

b. *Scarcity of annotation data*

The algorithm development is limited by the lack of annotation data. For building footprint extraction, Shermeyer *et al.* [124] present an MSAW dataset containing airborne SAR images, optical images, and building footprint annotations, along with a deep network baseline model and benchmark. By introducing a TomoSAR point cloud, Shahzad *et al.* [137] are able to acquire accurate building areas in a SAR image and take them as ground truth annotations to train a segmentation network for the purpose of building extraction. However, TomoSAR point clouds are rare, that this approach is limited to be applied to other areas. Moreover, this work cannot differentiate individual buildings.

3.2 Related Registration Algorithms

Recent developments in modern satellite missions have led to an increasing interest in the combined use of multi-sensory data. The complementary information from different sensors promotes successful remote sensing applications, such as land use and land cover classification [145, 146], urban mapping [147, 148], forest areas classification [149, 150], change detection [151, 152], and more. Data registration is the first step of all these applications, i.e., aligning data acquired at different conditions, including sensors, time, and viewpoints. In particular, the combined use of SAR data and other data is of great interest [153]. However, the registration of SAR and other data is difficult due to its completely different imaging nature of active sensing and side-looking geometry comparing with optical imagery [154]. Furthermore, the increases in spatial resolution enlarge the difference between SAR data and other data, especially in urban areas where the dense high-rise structures appear very different in SAR and optical images.

3.2.1 Registration of SAR and Optical Images

SAR and optical data are the two main spaceborne modalities and are highly complementary to each other. The literature in the field of SAR and optical image registration can be categorized into the following groups:

a. *Intensity-based methods*: In the field of SAR and optical image registration, one category of methods is intensity-based, which relies on the similarity measure between

pixel intensities in the two data, for instance, mutual information [6,155,156], implicit similarity [157], normalized cross-correlation coefficient [158], and cross-cumulative residual entropy [159]. Depending on the intensity, these methods are affected by textures, occlusions, image distortions, and illumination differences [160].

b. *Feature-based methods*: The other category of methods is feature-based that salient features, such as points or lines, are extracted from both data as control point candidates and are matched in the subsequent registration step to determine the geometric correspondence and transformation [161–163]. A widely used local feature descriptor is the scale-invariant feature transform (SIFT) [164] and its variants adapted to SAR data [165–168]. Other feature descriptors are also proposed, such as the histogram of oriented phase congruency (HOPC) descriptor [169], the Radiation-variation Insensitive Feature Transform (RIFT) [170]. However, in the presence of large geometric differences, feature correspondence can be difficult to determine.

c. *Object-based methods*: Some researchers exploit object-based features, such as road network [171–173] or building footprints shapes [11, 14, 174–176]. With such contextual information of objects, especially buildings, as the key urban feature, these methods are more robust in urban scenarios. For example, in [11], the building outlines are fused from the linear features detected in SAR image and optical image. In [14], the building footprint polygons are extracted from optical images and then projected and registered to SAR images.

d. *Deep learning approaches*: Deep learning-based registration approaches are increasingly popular in recent years. Representatives including: a two-stream network computing a similarity score for a SAR-optical patch pair for matching [177], a siamese network creating pixel-wise feature descriptors [178], translating SAR images into pseudo-optical images with Generative Adversarial Networks (GANs) first and then matching them with optical images [179].

Overall, plenty of studies focus on extracting and matching corresponding information from optical and SAR data. In theory, GIS data can usually be easily projected to orthorectified optical images, thus can be projected to SAR images once the optical image is registered to the SAR image. However, often these studies experiment in areas of flat terrain or rural environments and use data of middle resolution where the differences between SAR and optical imagery are within a certain level. These methods are limited to be used in urban regions or use VHR data.

3.2.2 Registration of SAR Images and Building Footprint Data

Two dimensional (2-D) building footprint data contain direct shape information of object boundaries that can assist SAR image interpretation, which are usually already available or alternatively can be produced from multiple sources such as maps or optical images. For example, in [180], GIS data is used to depict the borders of the area of interest for rapid damage analysis. In [181], the road network from GIS data is mapped to SAR image to assist identification of ground moving objects. In [182, 183], GIS data are combined with SAR data for flood mapping. In urban SAR interpretation, the building footprints in GIS data are particularly helpful for providing the location and footprint shape of individual buildings, for instance, for locating estimated InSAR deformation in urban area [184],

for estimating the height of buildings from SAR image [15, 48] or InSAR data [47], for identifying damages in earthquake [185].

Compared to the registration of SAR images and optical images, few studies in the area of registration of SAR images and GIS data. In [186], several building polygons are matched to SAR image by local adjustment based on the intensity value of the SAR image. However, this only works for isolated buildings with clear signatures in SAR images. In [187], GIS building footprints are registered to SAR image in a small urban area based on the building correspondence between the two data. However, it does not consider the terrain variation, thus only suitable in areas with flat terrain.

3.3 Related Advances in Deep Learning

3.3.1 Deep Learning in SAR

In recent years, deep neural networks have become increasingly popular and have shown success in many fields. In contrast to classical approaches that require expert domain knowledge and hand-crafted features, deep networks rely on a large amount of data and can learn effective feature representations from raw data in an end-to-end fashion. In applying deep networks to SAR data, most attention has focused on vision tasks such as classification [129, 130, 188–190], segmentation [131, 132], target recognition [133, 134, 191, 192], and denoising [193, 194]. Deep learning models are also adopted in the task of parameter inversion, e.g., sea ice concentration [195], rough surface parameters [196], physical scattering signatures [197]. In addition, some studies employ deep networks in matching/fusing SAR data with optical images to promote the joint use of complementary data from different modalities [177,179,198]. For a comprehensive review of deep learning methods applied to SAR data, the interested readers are referred to [199].

At present, the major problem of applying deep networks to urban SAR analysis tasks is the lack of annotation data. For instance, there is a large number of published studies applying deep learning to the task of object detection using SAR images, but the detected objects are restricted to salient objects, such as ships and vehicles, merely because those are the objects defined in the available data sets such as MSTAR [200], SAR-Ship-Dataset [201]. Other objects, however, have not received enough attention.

In the field of urban object detection, in 2020, SpaceNet 6 challenge released the MSAW data set for the task of automatic building footprint extraction using a combination of SAR and optical imagery [124]. The MSAW data set contains high-resolution SAR imagery, WorldView 2 satellite imagery, and annotations of building footprints covering the city of Rotterdam. The winning team of the challenge employed a semantic segmentation network and reported an F1 score of 0.4242 [202], indicating the difficulty involved in this task. As introduced in Section 3.1, building areas in SAR images contain more than footprints. Nevertheless, no deep learning-related investigation has been addressed on roofs or walls.

3.3.2 Image Segmentation

Image segmentation is a fundamental problem in computer vision that plays a central role in various applications. Two main tasks in image segmentation are semantic segmentation, i.e., classifying pixels with semantic labels, and instance segmentation, i.e., partitioning of individual objects. Prior to flourishing studies of deep learning, numerous segmentation algorithms have been developed, including earlier methods such as thresholding,

region growing [203], watershed methods [204], and advanced algorithms such as active contours [205], graph cut [206], conditional random fields (CRFs), and Markov random fields (MRFs) [207].

In recent years, deep learning models outperform previous algorithms exceedingly and have caused a paradigm shift in the field. Consisting of only convolutional layers, Fully Convolutional Networks (FCN) [208] is a milestone in the field. A FCN takes an image of arbitrary size and produces a segmentation map of the same size. It demonstrates that deep networks can be trained for semantic segmentation in an end-to-end manner. Following the revolutionary work of FCN, various network structures have been proposed that contribute to the field. Several networks incorporate probabilistic graphical models into deep learning architectures, such as CRFs and MRFs. A popular approach is based on the encoder-decoder architecture, including very well-known SegNet [209] and U-Net [210]. This architecture comprises two parts, that the encoder gradually reduces the spatial dimension with pooling layers, while the decoder gradually recovers the object details and spatial dimension. Dilated convolutions have been exploited in some recent works, including the well-known DeepLab family [211–213]. Dilated convolutions have gaps in the filter thus can quickly expand the effective receptive field. Another idea that has been deployed in various network architectures is multi-scale analysis. In this category, Feature Pyramid Network (FPN) [214] is one of the most prominent models. Some methods also use recurrent structures in image segmentation. With Recurrent Neural Networks (RNNs), pixels may be linked together and processed sequentially to model global contexts. For example, CNNs and RNNs are fused in ReSeg [215] and DAG-RNN [216]. Attention mechanism is another trend in image segmentation. It provides a focus on certain regions of the input when predicting a certain part of the output sequence [217]. GANs [218] have also been adopted for image segmentation. GANs consist of two different networks: one segmentation network takes the image as an input and generates per-pixel predictions, and one adversarial network discriminates segmentation maps coming either from the ground truth or from the segmentation network.

3.3.3 Bounding Box Regression

Bounding box regression refers to the task to regress the center coordinate and the size of the bounding box for each object. It is a key task in object detection and localization. Therefore, here the related work of bounding box regression is reviewed, including related object detection and localization networks and related loss functions.

Object detection is another important task in computer vision, aiming at establishing a bounding box around all objects within an image. Two of the main methods are the single-stage approach and the two-stage approach. The single-stage approach [219, 220] predicts bounding boxes directly; therefore, it is faster than the two-stage approach. In the two-stage pipeline [221, 222], the first stage generates a set of region proposals, and the second stage classifies and refines the coordinates of proposals by bounding box regression. The two-stage approach has been the leading paradigm in object detection. Recently, anchor-free methods are also proposed [223–225]. Multi-stage approaches are also adopted that iteratively refines the detection results [226, 227].

Object localization is one of the crucial modules for object detection. A widely adopted approach for object localization is to regress the center coordinate and the size of a bounding box [220–222, 228]. However, often the precision is unsatisfactory due to the large variance of the regression target. Aiming for more accurate localization, an intuitive approach

adopted by many works [226, 227, 229] is applying cascade architecture to refine bounding boxes progressively. Some methods try to reformat the object localization process. Grid R-CNN [230] adopts a grid localization mechanism to encode more clues for accurate object detection. CenterNet [231] combines the classification and regression to localize the object center. It predicts possible object centers on a keypoint heatmap and then adjusts the centers by regression. LocNet [232] predicts probabilities for object borders or locations inside the object's bounding box. SABL [233] focuses on the boundaries of the object bounding box and decomposes the localization process for each boundary with a bucketing scheme.

In addition, new losses have been proposed, including Intersection over Union (IoU) loss, Generalized IoU (GIoU) loss [234], Distance-IoU (DIoU) loss [235], and Complete IoU (CIoU) loss [236]. The insight is that the widely adopted L_n-norm loss is not tailored to the evaluation metric of IoU. Thus losses incorporating IoU are able to improve network performance. For instance, in CIoU loss, geometric factors of overlap area, normalized central point distance, and aspect ratio are formulated as invariant to regression scale. Therefore it can better distinguish difficult regression cases.

3.4 Contributions of the Thesis

As the survey of related work shows, a considerable amount of literature has been published on the topic of building reconstruction from SAR data. Nevertheless, most studies are limited to buildings with specific shapes or a small region of simple arrangements. Up to now, far too little attention has been paid to large-scale analysis.

Complementary data such as building footprints can assist building reconstruction on a large scale; however, a precise registration is needed. Deep learning algorithms have great potential in the task of urban SAR analysis, however, their application is limited by the absence of annotation data.

Aiming at LoD1 building model reconstruction on a large scale using single SAR imagery, this thesis employs building footprints as complementary data and applies deep learning algorithms. The aforementioned problems are tackled successively. In summary, the key contributions of this thesis are:

a. The first study investigating individual buildings on a large-scale SAR image and the first study employing deep networks in the problem of individual building analysis using SAR images, to the author's best knowledge.

b. An automatic registration framework that enables the joint use of building footprint data and SAR data on a large scale.

c. Two approaches to generate annotation data that enable deep learning methods to be applied for building reconstruction from SAR images. The first one produces building masks using building footprint data and an accurate DEM; the second one annotates bounding boxes of buildings using footprint and height data of buildings.

d. Two deep neural networks for individual building extraction in SAR image. The first one is a segmentation network, which is capable of improving the performance of networks by imposing constraints on the learning process. The second one is a regression network, which allows fast training by incorporating knowledge of building footprints.

e. An investigation on the robustness of the proposed networks against the positioning errors in building footprint data. These studies suggest that large open-sourced building footprint data can be exploited for individual building reconstruction in SAR images.

4 Data Set Generation

Deep learning algorithms have great potential in the task of urban SAR analysis, however, their application is limited by the absence of annotation data. To tackle the problem of dataset scarcity, this thesis proposes two approaches that automatically label the areas and bounding boxes of individual buildings in SAR images. These approaches are employed to produce the necessary data sets for training the deep networks proposed in Chapter 6 and Chapter 7. In addition, the ground truth data are required for evaluating the proposed registration algorithm in Chapter 5. This chapter also briefly presents the workflow for generating the ground truth, i.e., correctly registered building footprints and SAR images.

This chapter first clarifies the required data sets of the thesis, then presents the developed approaches for generating data sets used to train deep neural networks and evaluate the registration framework. Section 4.2 and Section 4.3 are summarised in [237] and [238], respectively.

4.1 Data Set Requirements

4.1.1 Required Data

Throughout the thesis, *SAR images* and *building footprints* in the corresponding areas are the commonly used data.

Additionally, DSMs, TomoSAR point clouds, and building height data are introduced for generating necessary training data and ground truth data. The additional data are listed in Table 4.1, corresponding to the data sets that are to be generated.

Table 4.1: Additional data employed for data set generation

Data sets to be generated	**Additional data**
Areas of individual buildings in a SAR image	An accurate DSM
Bounding boxes of individual buildings in a SAR image	A height value for each building
Correctly registered building footprints and SAR data	An accurate DSM & a TomoSAR point cloud

4.1.2 Reasons for the Required Data

The additional data are introduced for specific reasons in each work:

a. *For segmentation of individual buildings in a SAR image*

   The objective is to generate a data set for training and evaluating the segmentation networks, i.e., labeled areas of individual buildings in the SAR image.

   For this purpose, a highly accurate DSM is used to model the scene illuminated by the SAR sensor, which is subsequently projected to the SAR image coordinate system to generate individual buildings areas in the SAR image. The purpose of using a DSM is to acquire building areas without TomoSAR point clouds that are

used in [137]. DSMs are far more available and accessible than TomoSAR point clouds. Thus, with this data set generation approach, the proposed algorithm can be applied to areas where accurate DSMs are available.

b. *For localization of individual buildings in a SAR image*

The objective is to generate a data set for training and evaluating the object localization networks, i.e., the bounding box of each building.

A height value for each building is required. Since buildings heights can be acquired from multiple data sources such as city models, LiDAR data, or accurate DSMs, more training data can be generated to employ available data from different sources.

c. *For registration building footprints with a SAR image*

The objective is to generate the ground truth data to evaluate the proposed algorithm, i.e., correctly registered building footprints and a corresponding SAR image.

The additional data for data set generations are an accurate DSM and a TomoSAR point cloud. 3-D coordinates are needed in radar coding. The DSM provides the third coordinates, i.e., height values, that are added to 2-D building footprints before radar coding. The TomoSAR point cloud is in the UTM coordinate that points from other data sources are aligned to it first before projecting the corresponding SAR image. The TomoSAR point cloud is used to automate the process of the shift estimation.

4.2 Automatic Annotating Regions of Individual Buildings in a SAR Image

For training segmentation networks to extract individual buildings, annotations of building areas (as ground truth data) and building footprints (as input data) in SAR images are necessary. Therefore, a workflow is proposed to automatically label building masks and their corresponding footprints in SAR images using a highly accurate DSM and building footprints from GIS data.

The data set is generated in two stages. First, sensor-visible 3-D building models (i.e., non-occluded roofs and facades) and building footprints are prepared in the UTM coordinate system. Second, they are projected to the SAR image coordinate system in order to generate building ground truth annotations and the corresponding footprints. Figure 4.1 illustrates the workflow, and more details are presented in the following sections.

4.2.1 Scene Modelling

a. *Modeling the sensor-visible scene*

This step models a scene that can be viewed by a radar sensor in the UTM coordinate system. The procedure is conducted in three steps (cf. Figure 4.2):

a.1. *DSM is transformed to a point cloud P_{dem}.*

Specifically, each pixel in the DSM with geolocation coordinates (x, y) and a height value h is represented as a point with coordinates (x, y, h), and hence all pixels establish a nadir-looking 3-D point cloud P_{dem}.

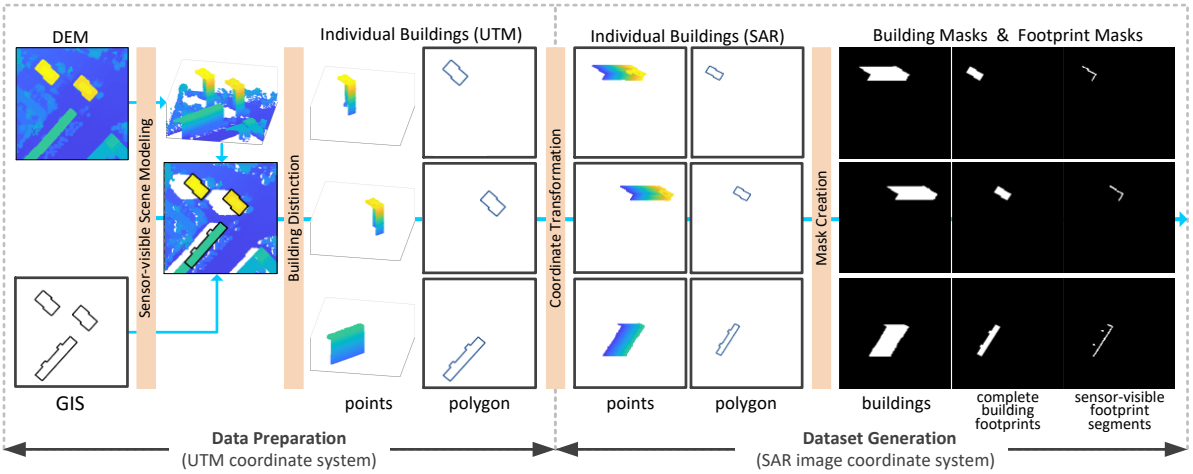


Figure 4.1: The workflow for generating pixel-wise individual buildings areas in SAR images. We first collect DSM and GIS data in the UTM coordinate system and then project them to the SAR image coordinate system in order to generate building ground truth annotations and the corresponding footprints in our study area.

a.2. *A complete 3-D point cloud P_{com} is generated by filling vertical data gaps.*

To be more specific, vertical structures such as building walls that are absent from P_{dem} are added through the following steps. We first detect building points which are located at height jumps. Afterwards, at each detected point $g(x, y, h)$, a vertical point set $G = \{g_i(x_i, y_i, h_i)|i = 1, ..., m\}$ is added, where $x_i = x, y_i = y, h_i = h_0 + i \times h_{step}, h_i < h_e$. h_0 and h_e are the minimum and maximum heights in the neighbourhood of g, h_{step} is a predefined height step, and the number of points $m = (h_e - h_0)/h_{step}$. Eventually, a complete 3-D point cloud P_{com} is built by all vertical point sets and P_{dem}. Note that the DSM is 2.5-D instead of true 3-D, i.e., each 2-D point (x, y) is assigned to a unique height value z [239], that vertical surfaces of complex objects are not represented, such as trees (cf. Figure 4.3). Therefore vertical points are only added to building areas in this step.

a.3. *A sensor-visible scene point cloud P_{svs} is obtained through a visibility test on the point cloud P_{com}.*

Since a radar sensor only sees one side of a scene, points on the other side should be removed. To this end, the hidden point removal (HPR) algorithm [240] is applied.

In our process, the viewpoint in HPR is positioned on the line of sight of the radar sensor at a large distance away from the scene in order to simulate an orthographic view in the azimuth of the radar sensor. In this way, sightlines from the viewpoint to objects in the scene are parallel to each other and orthogonal to the azimuth, enabling HPR to remove sensor-invisible points.

b. *Distinguishing buildings in the scene*

In this step, building points[1] are distinguished for individual buildings. Given one building, its building points are selected from P_{svs} using its footprint. Building

[1]Building points refer to points in a point cloud that belong to the building class.

points contain two parts: the roof points that are located within building footprints, and the wall points that are located along boundaries of building footprints.

Note that there are two possible inconsistencies between the DSM and GIS data. First, if a building is contained in P_{svs} but not in GIS data, it is not selected from P_{svs}. Second, if a building is contained in GIS data but not in P_{svs}, i.e., points in the footprint region are not elevated than surrounding ground points, we exclude this building from our dataset.

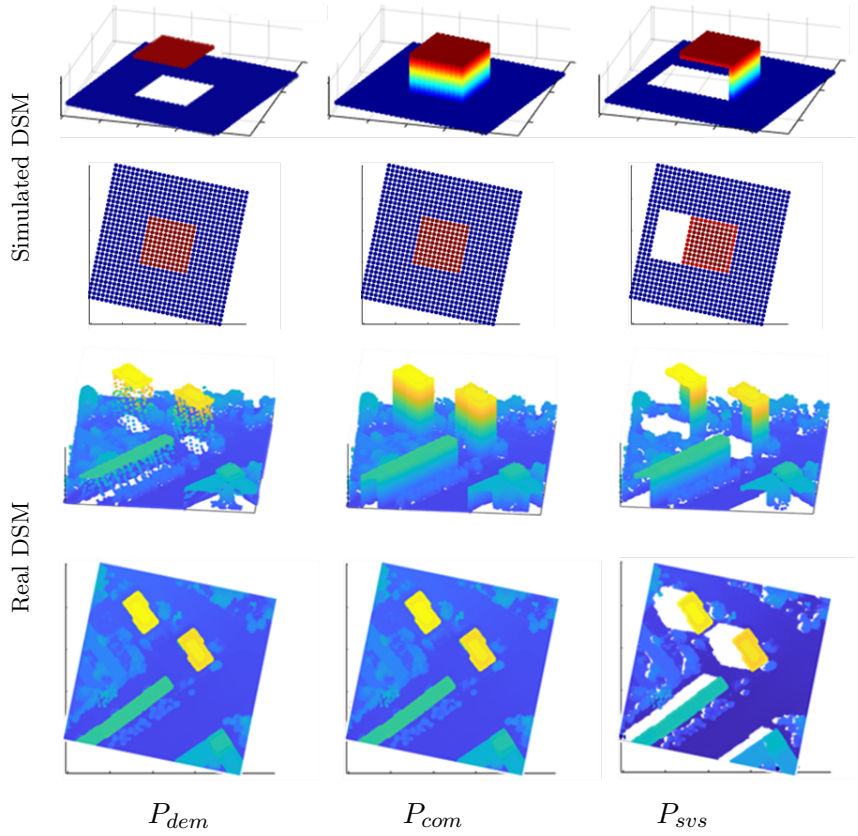


Figure 4.2: Illustration of scene modeling steps with a simulated DSM and a real DSM in 3-D (first row) and 2-D (subsecond row): (left) the DSM point cloud P_{dem}; (middle) the complete point cloud P_{com} after adding vertical points; (right) the sensor-visible point cloud P_{svs} after hidden point removal.

4.2.2 Data Set Generation in the SAR image Coordinate System

a. *Coordinate transformation*

The aforementioned procedures are carried out in the UTM coordinate system, and in our case, building points generated in the previous steps should be projected to the SAR image coordinate system; that is to say, coordinates (x, y, h) need to be transformed to $(range, azimuth)$. Moreover, building footprints are also projected to this coordinate system using ground height values obtained from the DSM. Figure 4.4 shows building footprint polygons that are superimposed on the DSM in Berlin.

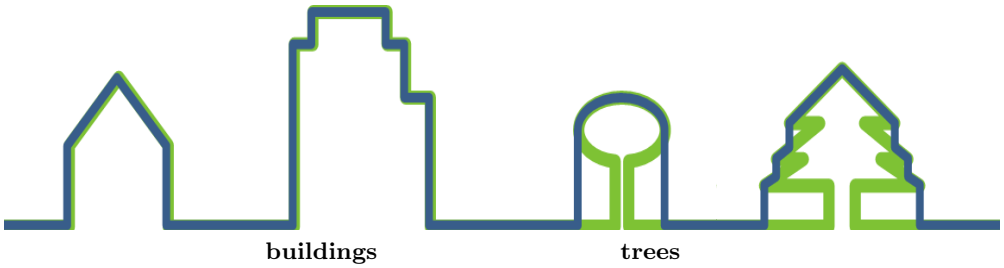


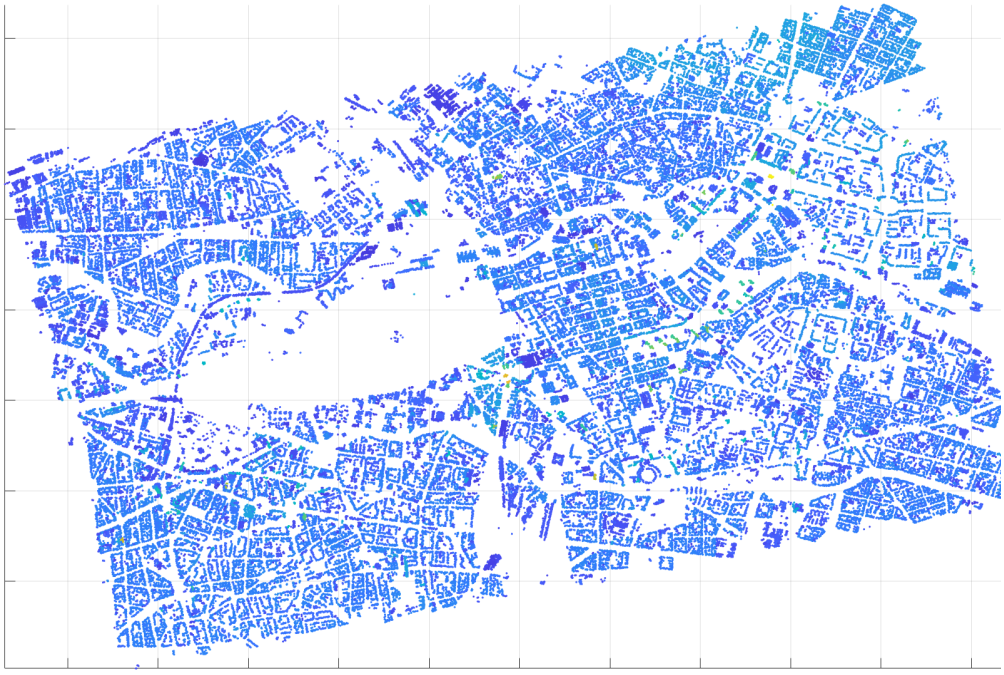
Figure 4.3: Illustration of 2.5-D (dark blue) and 3-D (green) surface models. In 2.5-D representation, each 2-D point (x, y) is assigned to a unique height value z. Therefore 2.5-D DSM can represent vertical walls of buildings, but not vertical surfaces of complex objects, such as trees.

Figure 4.4: Building footprint polygons superimposed on a DSM in Berlin that used in the thesis.

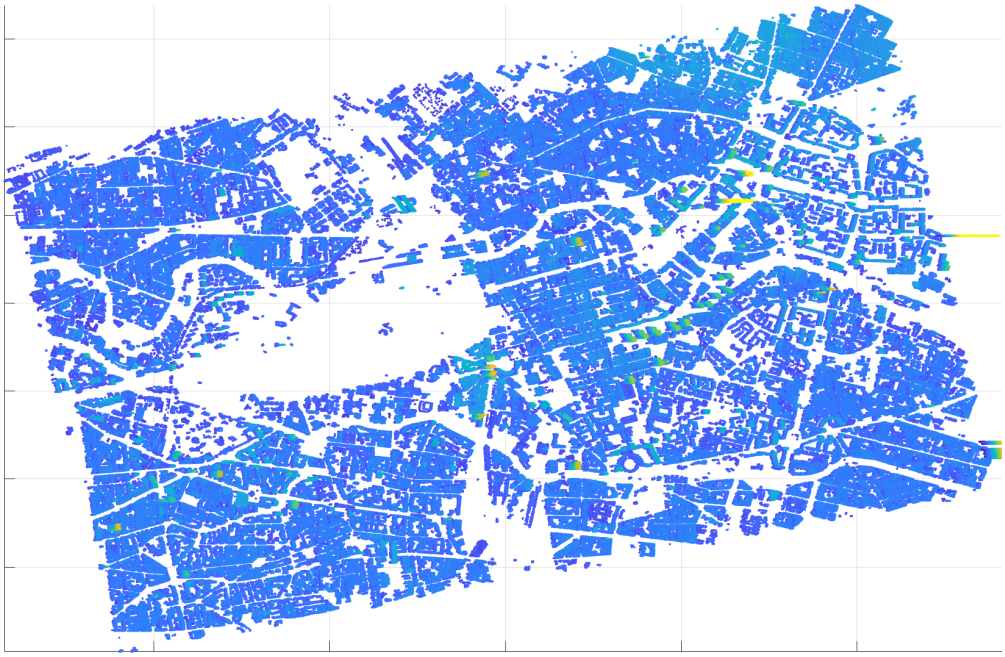
Combining these two data, building points can be extracted. In Figure 4.5, the extracted building points of all buildings from the DSM in Figure 4.4 are shown in the SAR image coordinate system, in which the roof points (a) and wall points (b) are plotted separately. Added from the DSM, the wall points are now visible and can be used for building analysis together with roof points.

Generally, the coordinate transformation of the point cloud from the UTM coordinate system to the SAR imaging coordinate system includes iterative solving Doppler-Range-Ellipsoid equations that can be implemented with different approaches [53, 241–243]. In this work, radar coding was performed using DLR's Integrated Wide Area Processor (IWAP) [244].

b. *Mask creation*



(a) Roof points



(b) Wall points

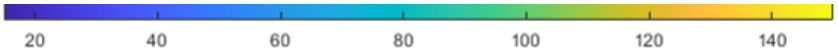


Figure 4.5: Building points (roof points and wall points) are extracted using building footprints from DSM points and are radar-coded to the SAR image coordinate system. Height is color-coded (meter).

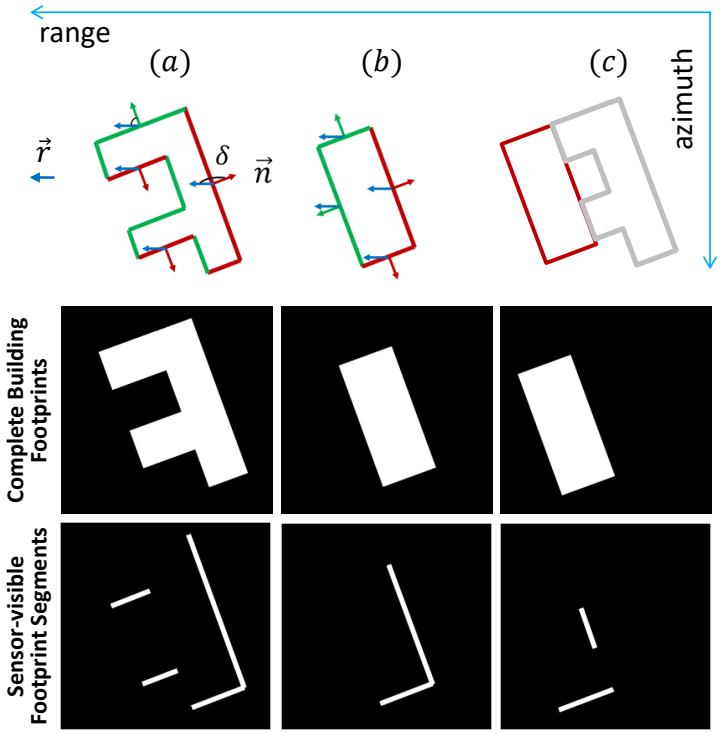


Figure 4.6: Examples of (top) the visibility test of building footprints and (middle and bottom) the two footprint representations. (a) and (b) show footprints of isolated buildings: red edges are sensor-visible, as the angle δ between the outward normal vector of an edge $\overrightarrow{n}$ and the range direction vector $\overrightarrow{r}$ is in the range of $(90°, 180°]$, while green ones are invisible. (c) shows a case that a footprint is touching another one; hence common edges are sensor-invisible.

Finally, according to *range* − *azimuth* coordinates of building points, we generate building ground truth masks, in which buildings are indicated by 1 and backgrounds are marked as 0. In addition, building footprint masks in the SAR image coordinate system are also created. Notably, to find an effective way of using building footprints, we create two representations, namely complete building footprints and sensor-visible footprint segments. The latter is generated via a visibility test (see Figure 4.6). Formally, let $\overrightarrow{n}$ be the outward normal vector of a polygon edge, $\overrightarrow{r}$ be the range direction vector, and $\delta \in [0°, 180°]$ be the angle between $\overrightarrow{n}$ and $\overrightarrow{r}$. A polygon edge is sensor-visible if $\delta \in (90°, 180°]$, and if a footprint is touching other footprints, common edges are invisible because they do not exist in the real world (e.g., Figure 4.6(c)).

c. *Post-processing*

Since the used SAR image and DSM are collected at different times, there might be inconsistencies resulted from urban changes, such as building construction and deconstruction. This leads to inaccurate ground truth data. We cope with the problem using the intensity values of the given SAR image. In the SAR image, the intensity values are generally larger in building areas than in ground areas.

Therefore, a threshold is set to be the mode of the intensity values of the SAR image to exclude buildings of which mean intensity values are smaller than the threshold.

4.3 Automatic Labelling Bounding Boxes of Buildings in SAR Images

For training bounding box regression networks, building bounding boxes (as ground truth data) and building footprints (as input data) in SAR images are necessary. For this reason, a workflow is proposed that employs building footprint and height data to automatically label building bounding boxes and their corresponding footprints in SAR images. The proposed workflow is illustrated in Figure 4.7 and comprises two simple steps: building heights acquisition, and data set generation. In the following sections, we explain the details.

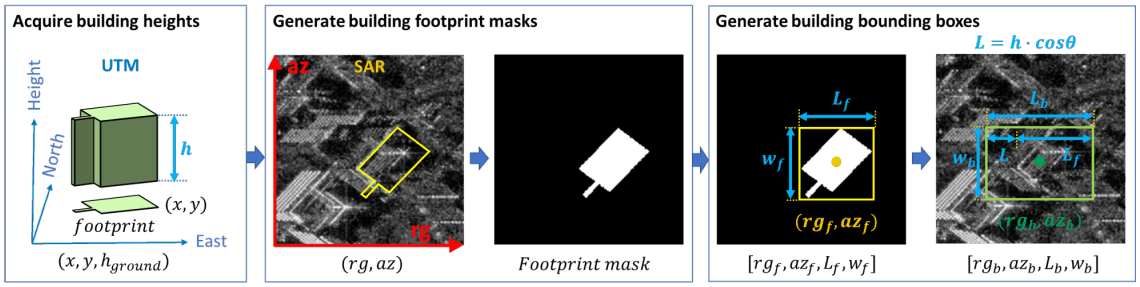


Figure 4.7: The workflow for generating bounding boxes and building footprint masks for individual buildings. Building footprints and height data are first collected in the UTM coordinate system and then projected to the SAR image coordinate system to generate building footprint masks and building bounding boxes.

4.3.1 Building Height Acquisition

The first step is to collect building data. For each building, we collect its footprint coordinate (x, y), the ground height h_{ground}, and the building height h.

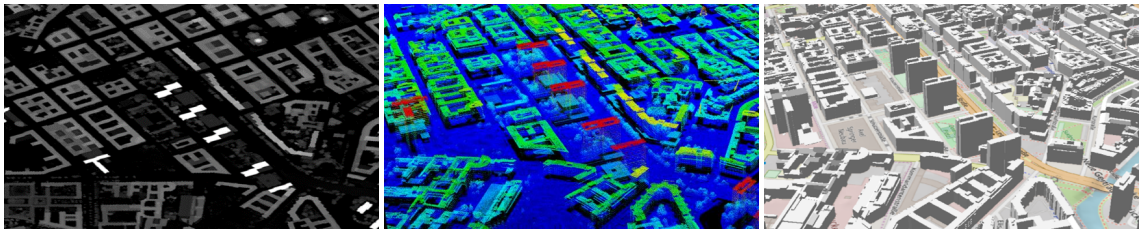
The proposed workflow requires only one height value for one building, which can be acquired from multiple data sources, such as LiDAR data, accurate DSMs, and city models. Figure 4.8 shows an example of some possible data sources of building heights in the same region. This design maximizes the scale of ground truth data that can be generated. In this way, the generation of our data sets can be supported by all these data sources, including publicly available data sets, e.g., Berlin city models [50], NYC open data [49], 3-D Buildings and Addresses of the Netherlands [245].

In LoD1 city models, usually, there is one height value for each building. As for DSMs and LiDAR data, in this thesis, the average roof height is regarded as the building height[2].

4.3.2 Data Set Generation

In the previous step, building footprints and building heights are acquired in the UTM coordinate system. For our task, building footprints need to be projected to the SAR image

[2] http://en.wiki.quality.sig3d.org/index.php/Modeling_Guide_for_3D_Objects_Part_2:_Modeling_of_Buildings_(LoD1,_LoD2,_LoD3)



(a) DSM **(b)** LiDAR point clouds [246] **(c)** Building models [50]

Figure 4.8: Examples of building height sources of the same area.

coordinate system; that is to say, for each building footprint, its coordinates (x, y, h_{ground}) need to be transformed to (rg, az), rg and az denote range and azimuth coordinate, respectively. Radar coding was performed using DLR's Integrated Wide Area Processor (IWAP) [244].

Then, building footprint masks are generated according to *range-azimuth* coordinates of building footprint polygons, in which building footprints are indicated by 1 and backgrounds are marked as 0.

For each building, we generate its bounding box B_b in two steps (cf. Figure 4.7):

The first step is to compute the footprint bounding box B_f. B_f is defined by four values in pixels $[rg_f, az_f, L_f, w_f]$, in which (rg_f, az_f) are coordinates of the center point of the bounding box, and L_f and w_f are the width and height of the bounding box.

Second, the building bounding box B_b is computed from B_f. The difference between B_b and B_f results from the added width L, which is the layover length corresponding to the building height h: $L = h \cdot cos\theta$, $L_b = L + L_f$. Therefore, the bounding box $B_b = [rg_b, az_b, L_b, w_b]$ is generated, where:

$$\begin{cases} rg_b = rg_f - \frac{1}{2}L \\ az_b = az_f \\ L_b = L + L_f \\ w_b = w_f \end{cases} \tag{4.1}$$

Finally, same as in Section 4.2.2, possible inaccurate bounding boxes need to be removed. This is also coped with the problem using the intensity values of the given SAR image. A threshold is set to be the mode of the intensity values of the SAR image to exclude bounding boxes which mean intensity values in them are smaller than the threshold.

4.4 Generating Ground Truth of Registered Building Footprints and a SAR Image

Considering the problems preventing correct registration explained in Section 2.4.2, two auxiliary data sets are used: an accurate DSM of the study area and a TomoSAR point cloud produced from the SAR images. Derived from SAR images, each point in the TomoSAR point cloud directly corresponds with the pixels in the SAR image. Therefore, if the GIS polygons are correctly aligned with the TomoSAR point cloud in the Universal Transverse Mercator (UTM) coordinate system, GIS polygons will have direct correspondences to SAR image coordinates. The DSM provides accurate heights for radar coding

GIS polygons, also provides the transformation parameters for aligning the GIS data to the TomoSAR point cloud by 3-D matching itself with TomoSAR point cloud.

The original GIS data in the UTM coordinate system is transformed three times to obtain the ground truth in the SAR image coordinate system, as shown in Figure 4.9:

a. 2-D GIS data with coordinates of *(East, North)* is aligned with the accurate DSM, by adding the ground height H from the DSM;

b. 3 -D GIS data with coordinates of *(East, North, Height)* is projected to TomoSAR point cloud, using transformation parameters derived from 3-D matching of the DSM and the TomoSAR point cloud [247];

c. GIS data *(East', North', Height')* is projected to SAR image coordinate system with the coordinate of *(Range, Azimuth)*, using the geocoding parameters between the TomoSAR point cloud and the SAR image.

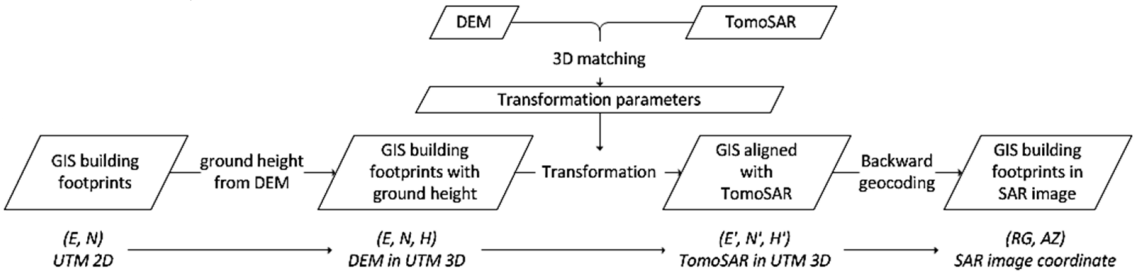


Figure 4.9: Procedures of ground truth generation for building footprints and the SAR image registration.

5 Automatic Registration of a Single SAR Image and Building Footprint Data on a Large Scale

In order to jointly use building footprints and SAR images, these two data must be registered first. Due to the lack of accurate terrain models, these two data are often misaligned when projecting one to the other. As explained in Section 2.4.2, for TerraSAR-X, a height error of 10 m results in a slant range error of 5.73 m to 9.39 m. And the errors are inconstant over the SAR image due to terrain variations in the observed area by the SAR sensor.

This chapter presents a framework for automatically registering 2-D building footprints to a corresponding SAR image. This work lays a foundation for the algorithms presented in Chapter 6 and Chapter 7, which require building footprints to be well registered to SAR images in pre-processing steps.

The main contents of this chapter are summarised in [248], and additional experiments with a TerraSAR-X stripmap image are provided in Section 5.5.1.

5.1 Problem Formulation Based on the Feature Correspondence

As aforementioned in Section 2.4.2, in radar coding, there exist three problems preventing accurate registration: the geometric accuracy of the SAR image, the positioning accuracy of building footprints, and the height accuracy of the terrain used in radar coding. The first two are negligible for TerraSAR-X/Tandem-X images and official GIS data sets[1], and the registration error to solve in this chapter is caused by the errors in the terrain.

Practically, registering the two data sets is challenging. The first challenge is to find the correspondence between them. Due to the geometry difference of the two data sets, objects depicted in one may not be presented in the other. The next challenge is to extract correct features. GIS data consist of building boundaries only; however, explicit boundary extraction of objects in SAR images is difficult since the high-intensity values are more related to structures and materials than object boundaries. The ambiguity of object boundaries further increases due to the existing speckle noise. In addition, the registration problem is non-rigid because of the aforementioned inconstant terrain errors. The registration process needs to discover local deformation between the two data sets.

This work is based on the building correspondences between the two data sets, as illustrated in Figure 5.1. A building in Universal Transverse Mercator (UTM) coordinate system (*North-East-Height*) and its signature in a SAR image coordinate system (*range-azimuth*) are plotted: the GIS footprint in the SAR coordinate system is depicted in yellow,

[1] As for the unavoidable positioning errors in building footprint data, Chapter 6 and Chapter 7 will discuss the impact of them on building analysis.

while the building signatures in the SAR image of the sensor-facing walls are outlined in green. The orange lines connect the near-range side boundary of the GIS polygon and the double bounce lines in the SAR image: they both represent the bottom of the sensor-facing walls and therefore correspond to each other. Based on the correspondence, the features from both data sets are extracted and registered.

Since the objective is to register all the GIS polygons to their correct location in the corresponding SAR image, the algorithm is performed in the SAR image coordinate system. Before the main workflow, the GIS data are radar-coded to the SAR image coordinate system with heights from a coarse terrain model.

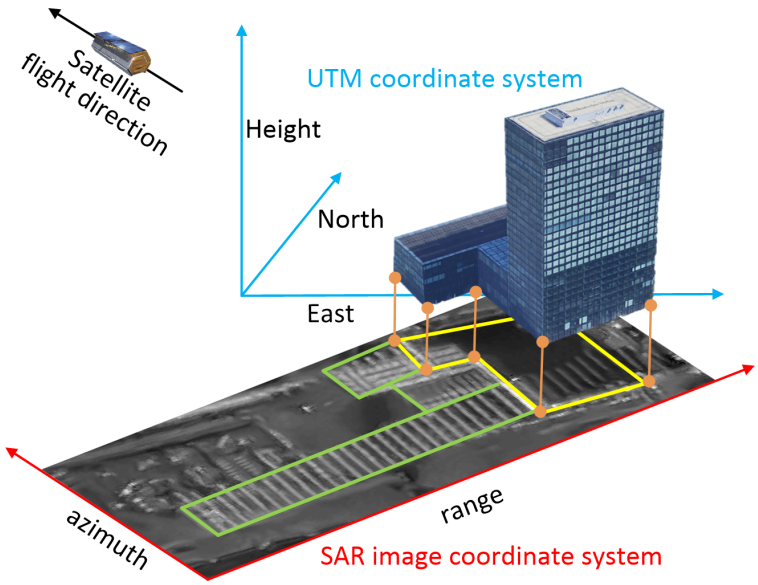


Figure 5.1: Illustration of the building correspondence between SAR and GIS data: the near-range side of the GIS footprint (yellow polygon) corresponds to the double bounce line in the SAR image, which is approximately the far-range side of the wall signatures (green polygons).

5.2 Corresponding Feature Extraction

5.2.1 Double Bounce Lines in the SAR Image

In SAR images, the corresponding features are the double bounce lines, which are the bright linear features from signal double bounces at the wall-ground intersections. They cannot be extracted by intensity values only, as other geometries may also appear as bright lines, e.g., regular windows or balconies on building walls, namely, corner lines [31]. However, for one visible wall, its double bounce line is usually located at the far-range side of the parallel corner line group. Therefore, our approach is based on the geometric relationship of the double bounce lines and other corner lines: first, the wall areas are segmented; then the far-range side boundaries of the wall segments are extracted as the preliminary estimates of the double bounce lines. Finally, the preliminary double bounce

lines are refined by exploiting the intensity of the SAR image. The detailed algorithms are explained below.

a. *Wall segmentation*

The SAR image is first segmented using the Potts model. The Potts model [249–251] formulates segmentation as an optimization problem:

$$u^* = \arg\min_u \gamma\|\nabla u\|_0 + \|u - F\|_2^2, \tag{5.1}$$

where F is the measured image and the data fidelity is measured by the L^2 norm. The empirical model parameter $\gamma > 0$ controls the balance between data fidelity and the regularizing term. The term u is a piecewise constant function whose discontinuity set encodes the boundaries of the corresponding segmentation. The term $\|\nabla u\|_0$ denotes the total length of the segment boundaries induced by u. The Potts problem is NP-hard. In this study, the minimization strategy of [252, 253] is adopted, where readers can find the details of implementation.

The Potts model is unsupervised. Hence the wall segments are selected by the following criteria: (1) the area of segments: the largest segments are excluded as background, and tiny segments are excluded to reduce outliers in the subsequent registration procedure; (2) the SAR image intensity in the area of the segments: the average intensity value in the building segments should be greater than the intensity value in ground area, which is approximately the mean intensity of the image; (3) the shape of segments: the near-range and the far-range sides of building segments should be roughly parallel, tested using the correlation coefficient.

b. *Far-range boundary extraction*

After segment selection, the contours of the segments are extracted. To extract the far-range feature lines, a visibility test is performed on the contour of the building segments using the algorithm described in Section 5.2.2 from the far-range side.

c. *Intensity based refinement*

Due to the smear effect [31], the double bounce lines do not perfectly overlay the far-range side contours of building segments, introducing a bias requiring compensation. Figure 5.2 (right) illustrates such a bias, where the red line is the desired double bounce line, and the blue line is the estimated one. Such bias is often systematically shifted towards the far-range direction. Since the double bounce line is often the brightest of its neighborhood, the intensity profile of the SAR image is utilized to estimate the bias.

On SAR images, the corner line repeats itself on every one floor of the building. Thus the distance between two corner lines on SAR images is $hcos\theta/ps_{rg}$, where h is the floor height, θ is the incidence angle, and ps_{rg} is the pixel spacing of the SAR image in the range direction. For the far-range line $P = \bigcup_{i=1}^{n} p_i, i = 1, ..., n$ consists of boundary points p_i, the bias s is estimated:

$$s^* = \arg\max_s \sum_{i=1}^{n} I(r_i, c_i - s), \quad 0 < s < hcos\theta/ps_{rg}, s \in \mathbb{N}, \tag{5.2}$$

where for $p_i(r_i, c_i)$ at the r_i-th row and c_i-th column of the SAR image, s^* is the range bias to be estimated, and $I(r_i, c_i)$ is the intensity value at $p_i(r_i, c_i)$. The refined double bounce lines are obtained by adding each bias to the corresponding far-range boundary lines.

A SAR point-set is sampled from the extracted double bounce lines.

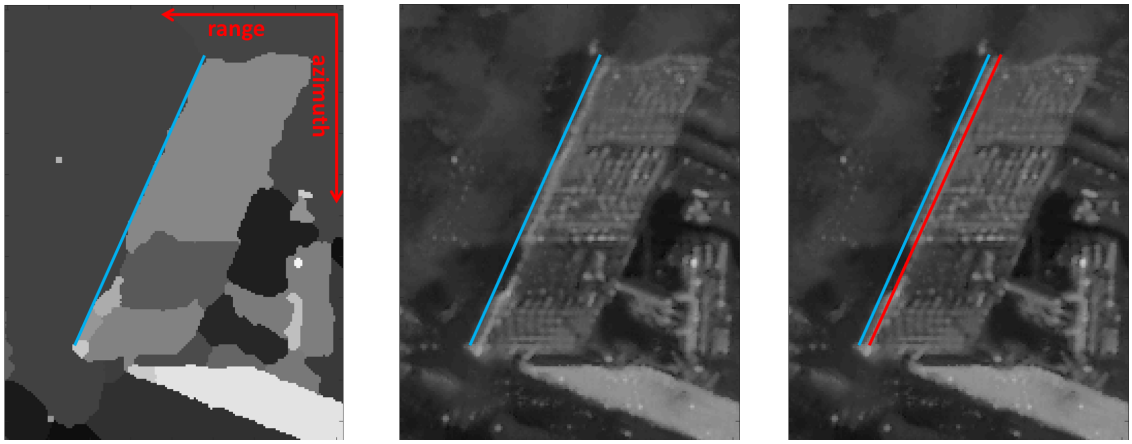


Figure 5.2: Illustration of the intensity-based refinement of the extracted SAR features. (left) SAR image segmentation result with the extracted far-range boundary of segments (blue) overlaid on it; (middle) SAR amplitude image with the extracted far-range boundary of segments (blue) overlaid on it; (right) SAR amplitude image with the extracted far-range boundary of segments (blue) and the double bounce line (red) overlaid on it. Apparently, there is a bias between the far-range segment boundary and the double bounce line that needs to be estimated.

5.2.2 Sensor-visible Edges in Building Footprint Polygons

The corresponding features in the GIS building footprint that are also visible in the SAR image are the 'sensor-visible' edges in GIS polygons, which represent the bottom of the illuminated wall in 3-D.

The visibility of polygon edges can be tested via the angle between the sensor line of sight and the edge's normal direction. Let $\overrightarrow{n}$ be the outward normal vector of an edge in a GIS polygon, $\overrightarrow{r}$ be the vector in range direction, and δ be the angle between them clockwise from $\overrightarrow{r}$, $\delta \in [0°, 360°)$. According to δ and polygon geometry, there are the following three visibility cases, as illustrated in Figure 5.3 (a)(b): (1) Visible: $\delta \in (90°, 270°)$, and the edge locates at the near-range side of the exterior boundary; (2) Invisible: $\delta \in [0°, 90°]$, or $\delta \in [270°, 360°)$; (3) Partially visible: $\delta \in (90°, 270°)$, and the edge locates at the far-range side of the exterior boundary or the interior boundary. The partially visible edges are often not visible in the SAR image due to layover and shadow. Therefore, they are not extracted as features.

The above visibility test is designed for isolated polygons, e.g., the polygons in Figure 5.3(a)(b). Sometimes, two polygons are connected, e.g., in Figure 5.3(c). In such cases, the connected polygons are merged into a single polygon before the visibility test is performed, as shown in Figure 5.3(c) and Figure 5.3(e).

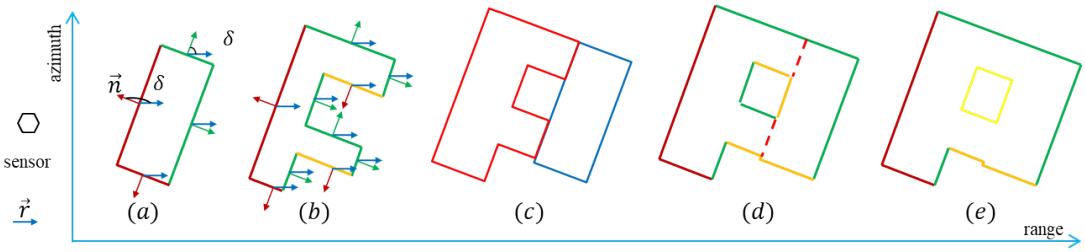


Figure 5.3: Visibility test: building polygons in a range-azimuth coordinate system (nadir-view). (a) and (b) show two isolated building polygons: $\overrightarrow{n}$ is the outward normal vector of an edge in building polygons, $\overrightarrow{r}$ is the vector in range direction, and δ is the angle between $\overrightarrow{r}$ and $\overrightarrow{n}$. $\delta \in (90°, 270°)$ holds for the red and orange edges: the red edges are visible, as they are at the near-range side, while the orange edges are partially visible, as they are at the far-range side. The green edges are sensor invisible, as $\delta \in (90°, 270°)$ does not hold. (c) shows two connected building polygons (red and blue) that should be merged to eliminate the connecting edges (the red dash lines in (d)). (e) shows the merged polygon from (c). The orange edges located at the interior boundary in (d) are partially visible.

Finally, the GIS point-set is sampled from the extracted GIS features, which are to be used for registration.

5.3 Progressive Feature Registration

As introduced in Section 2.4.2, height error δH causes a range error of $\delta L = \delta H cos\theta$. Due to terrain variation and the different accuracy of each GIS building footprint, the shift is not constant over a whole city. However, the shift is considered to be constant for individual building instances. Consequently, the registration problem is rigid at the polygon level but non-rigid at the global level.

However, many building polygons cannot be registered at the polygon level since there may not be sufficient SAR points to perform registration. To this end, a three-step registration strategy is proposed: first, a global registration is performed to recover a global transformation that brings the two point-sets as close as possible; second, a subarea registration is performed to recover local transformation for polygons in subareas with similar height error; third, individual polygon registration is performed to recover large local transformations that are not estimated previously. Therefore, the transformation is rigid at each step but altogether is non-rigid since they target different subsets of the whole point-sets. Each registration step is explained in the following sections.

A. *Global Registration*

To recover the global transformation, the rigid registration is solved with the Iterative Closest Point (ICP) algorithm [254, 255]. ICP iterates over two steps: (1) find correspondence set $\mathcal{K} = \{(\mathbf{p}, \mathbf{q})\}$ from target point-set $\mathbf{P}$ and source point-set $\mathbf{Q}$ transformed with current translation $\mathbf{t}$ and rotation $\mathbf{R}$; (2) update translation $\mathbf{t}$ and rotation $\mathbf{R}$ by minimizing an objective function $E(\mathbf{R}, \mathbf{t})$ defined over the correspondence set $\mathcal{K}$. The point-to-point ICP is used [255], with an objective:

$$E(\mathbf{R}, \mathbf{t}) = \sum_{(\mathbf{p},\mathbf{q})\in\mathcal{K}} \|\mathbf{p} - \mathbf{R}\mathbf{q} - \mathbf{t}\|^2. \tag{5.3}$$

The ICP algorithm is performed to register the GIS point-set to the SAR point-set in SAR image coordinates, thus the SAR point-set is fixed. After calculating the transformation, the GIS polygons are updated by applying the transformation to them.

B. *Subarea Registration*

In subarea registration, a set of grids is evenly distributed over the whole region. The δH in each grid is assumed to be constant. Therefore, the grid size should be large enough to contain a sufficient amount of points in each grid for registration, meanwhile as small as possible to promote the constant δH assumption. In practice, the grid size is chosen to be larger than the largest polygon (after merging).

Let dp be the distance between one GIS point and its closest SAR point. For all the GIS points and the corresponding SAR points in one grid, the distance set is $D = \bigcup_{i=1}^{n} dp_i$, where n is the number of GIS and SAR point-pairs. If the assumption of constant δH holds, the distribution of D will be unimodal with one clear peak center at C. In one grid, the point-sets are already registered, and no further processing is needed if $C = 0$; while the point-sets need further registration, if $C \neq 0$. To avoid discontinuity of the translation parameters between grids, the connected grids with a similar mode of their distribution D are clustered into subareas, using DBSCAN [256], before performing registration using ICP.

In the same way as in global registration, after subarea registration, the transformation is applied to the corresponding GIS polygons. For the GIS polygons that cross two or multiple subareas, several transformations may be suitable. After each transformation, the polygons are calculated, and the one that permits the smallest point-pair distance is chosen.

C. *Polygon Registration*

When the distribution of D does not show a clear center, the constant δH assumption does not hold. In this case, the registration proceeds to the polygon level, i.e., finding a rigid transformation for each polygon.

For each polygon, there can be two possible situations: (a) There are sufficient SAR points to perform further registration. Then the polygon is further registered using ICP. In practice, the following two criteria are employed to check the feasibility of ICP registration: the ratio of the number of near SAR points and the GIS points should be large, in our experiments larger than 0.7; and the corresponding SAR point shape and the GIS point shape should be similar. In practice, a correlation coefficient is larger than 0.8. (b) There are not enough SAR points around this polygon, then its nearest neighbor (NN) polygon is queried, and the transformation parameters of the NN polygon are adopted for this polygon. If the polygon has more than one NN polygon, the one that permits the smallest point-pair distance is chosen.

5.4 Experimental Results and Evaluation

In the following, the applicability of the proposed automatic registration framework is demonstrated.

5.4.1 Test Site and Data Set

a. *Study area and the used data*

Berlin is selected as the study area, and it is also the primary study area throughout the thesis. Containing typical urban forms, such as compact middle-rise area, open high-rise area, open middle-rise area, according to the definition of Local Climate Zones [257, 258], Frequently monitored and studied using TerraSAR-X data, Berlin is a representative of large-scale urban regions [58, 109, 247, 259]. Our study area is shown in the intersection area of the two rectangles in Figure 5.4: the yellow rectangle shows the area of the SAR image, while the red rectangle shows the area of DEM used for ground truth generation.

Several forms of data covering Berlin are collected for algorithm development and validation. Table 5.1 lists the used data for the Berlin study area. In our dataset, a descending TerraSAR-X image was acquired in the high-resolution spotlight mode with the pixel spacing[2] of 0.871 m in the azimuth direction and 0.455 m in the slant range direction. The incidence angle of this SAR image is 36°, and the heading angle is 194.34°. To reduce the speckle effect, the SAR image was filtered using a nonlocal InSAR algorithm [29].

The GIS building footprints data in the study area were obtained from the Berlin 3D-Download Portal [50]. The building footprints polygons are merged and result in 2414 polygons for the visibility test described in Section 5.2.2.

For radar coding, a constant height value of 77.5 m is added to all the GIS polygons. Figure 5.5 shows the radar-coded GIS polygons (red) superimposed on the SAR image, and three small areas in the green rectangles are selected to inspect the change of the two point-sets at different steps in the registration procedure (cf. Figure 5.7). In the following text, the figures are all shown in the SAR image coordinate system, with the range and azimuth direction the same as in Figure 5.5, unless otherwise specified.

b. *Ground truth data set*

For evaluation, the ground truth data set is generated using the workflow described in Section 4.4.

The DEM has been created from aerial UltraCam-D[3] images using the Semi-Global Matching stereo algorithm [260–262]. The ground resolution of the images and the resolution of the DEM are 7 cm/pixel. The TomoSAR point cloud was generated from 109 images using Tomo-GENESIS software developed at DLR [263].

The DEM provides precise relative terrain heights for radar coding the GIS polygons. The TomoSAR point cloud is used to calibrate the height of the DEM with respect to our InSAR processor, as each point in the TomoSAR point cloud has a direct correspondence with the pixels coordinates in the SAR image.

[2] In SAR images, *pixel spacing* represents the length one pixel corresponds to in the real world, while *resolution* indicates the minimum distance at which the radar can distinguish two close scatters.

[3] `http://download.microsoft.com/download/2/E/7/2E7FCE24-E085-4A64-B568-25BA956FCB60/UltraCamSpecifications_UCD_UCX_UCXp_UCXpWA_UCL_UCLp.pdf`

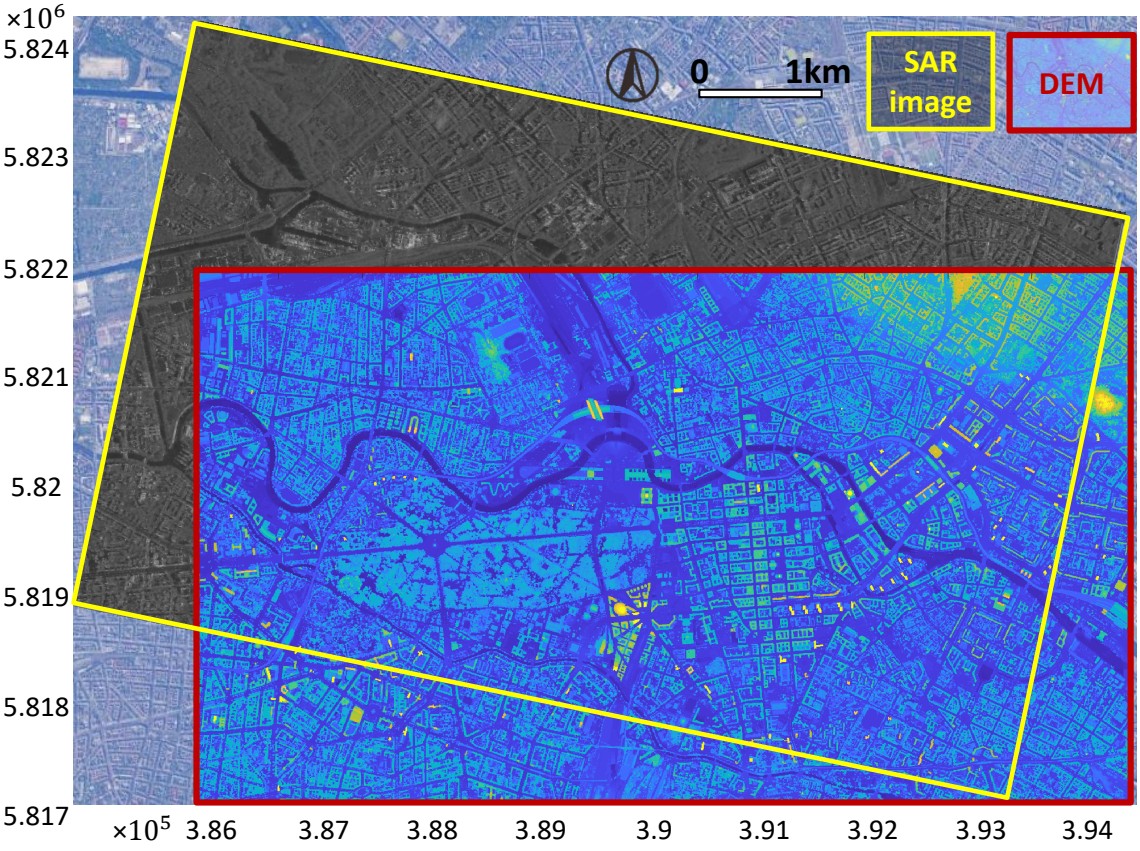


Figure 5.4: The study area in UTM coordinate system. The yellow rectangle indicates the area of the SAR image, while the red rectangle is the area of DEM used as ground truth height. Thus the intersection area of the two rectangles is our test area.

Table 5.1: Data used in the study area Berlin.

Data to be registered	**Description**
SAR data	One TerraSAR-X image in HS mode
Building footprints	Downloaded from Berlin 3D-Download Portal [50]
Data for ground truth generation	**Description**
a TomoSAR point cloud	Generated from 109 TerraSAR-X images using Tomo-GENESIS software developed at DLR [263]
an accurate DEM	Resolution: 7cm/pixel Obtained via the stereo processing of aerial images [261]

5.4.2 Results of Feature Extraction and Registration

a. *Feature extraction results*

a.1 *Features extracted from the SAR image*

Figure 5.6 shows the steps of feature extraction steps from a SAR amplitude image (Figure 5.6(a)). First, the SAR image was segmented using the Potts model: the segmentation results are shown in Figure 5.6(d). Second, as shown in Figure 5.6(b), the wall segments are selected, and the coarse double bounce lines, i.e., the far-range boundaries of the wall segments, are extracted (plotted in blue). Finally, the

refined double bounce lines are obtained and are superimposed on the SAR image in Figure 5.6(e), with the coarse double bounce lines plotted in blue for comparison. Figures 5.6(c) and (f) show an example of the intensity-based refinement of the double bounce line: the coarse double bounce lines in Figure 5.6(c) are shifted towards the near-range direction in the given neighborhood; based on the intensity, the red line with estimated bias is chosen, as shown in Figure 5.6(f). In Equation 5.2, assuming the height of one floor h is around 3 m, hence its neighborhood on the SAR image $hcos\theta/ps_{rg} = 3 \times cos36°/0.455 = 5.33(pixels)$. Therefore, the bias is defined: $s \in [1, 2, 3, 4, 5, 6]$. Based on the intensity value, the bias s for each far-range line is estimated, and most of them are estimated to be 2 or 3.

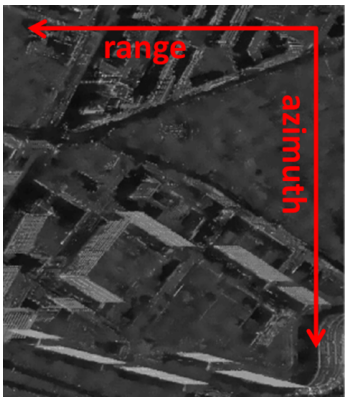
a.2. *Features extracted from GIS data*

The near-range side segments from all the GIS polygons are extracted. In Figure 5.7, the first column shows that the GIS building polygons from the selected areas are the areas 1-3 marked by the green rectangles in Figure 5.5. Their extracted GIS feature lines are plotted in red.

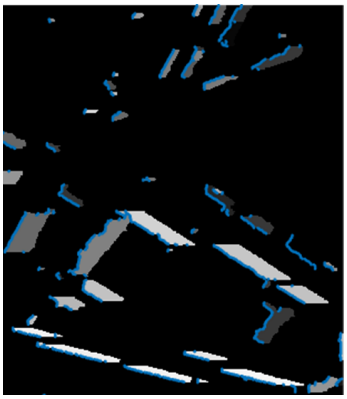
Figure 5.5: The study area in the SAR image coordinate system. GIS building footprints (red) are radar coded in the SAR image coordinate system with a constant height of 77.5 m. For detailed inspection of the two point-sets before and after registration, three small areas in the green rectangles are selected: Area 1, 2, and 3 represent areas where registration procedures are needed at global, subarea, and polygon levels, respectively. The point-sets in each area are shown in Figure 5.7.

b. *Feature registration results*

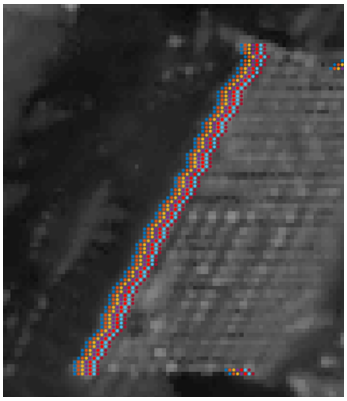
b.1 *Registration results at the global level*



(a) SAR image.

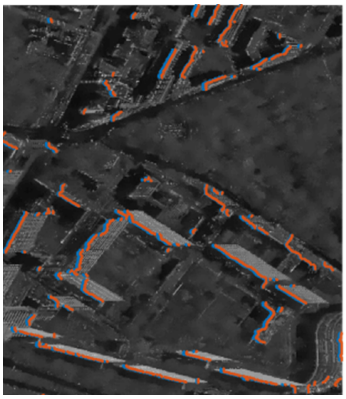


(b) Selected wall segments with far-range boundary lines (blue).

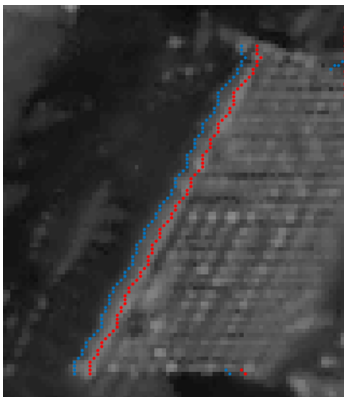


(c) Bias estimation example: the far-range boundary is shifted toward the near-range direction. One shift is shown in one color.

(d) Segmentation result.



(e) Estimated double bounce line (red) and extracted far-range boundary (blue) superimposed on SAR image.



(f) Bias estimation result example: the estimated double bounce line (red) and extracted far-range boundary (blue) for one building segment.

Figure 5.6: SAR feature line extraction steps. (a) shows an area in a SAR image; (d) shows the Potts segmentation result, and (b) shows the selected wall segments, with the coarse double bounce lines superimposed on the segments; (e) shows the refined double bounce lines in the SAR image; (d) and (f) show examples of the bias estimation process and result.

ICP was performed on the whole GIS point-set and the whole SAR point-set to determine the global transformation. Since a TerraSAR-X image and official GIS building footprints are used, whose geometric errors are negligible, the registration error mainly comes from the inaccurate height used in the radar coding. Consequently, only errors in the range direction are introduced. Therefore, in our experiments, the rotation matrix $\mathbf{R}$ and the translation in the azimuth direction $\mathbf{t_{az}}$ in equation (5.3) were not considered, and the objective was reduced to solve the translation in the range direction $\mathbf{t_{rg}}$.

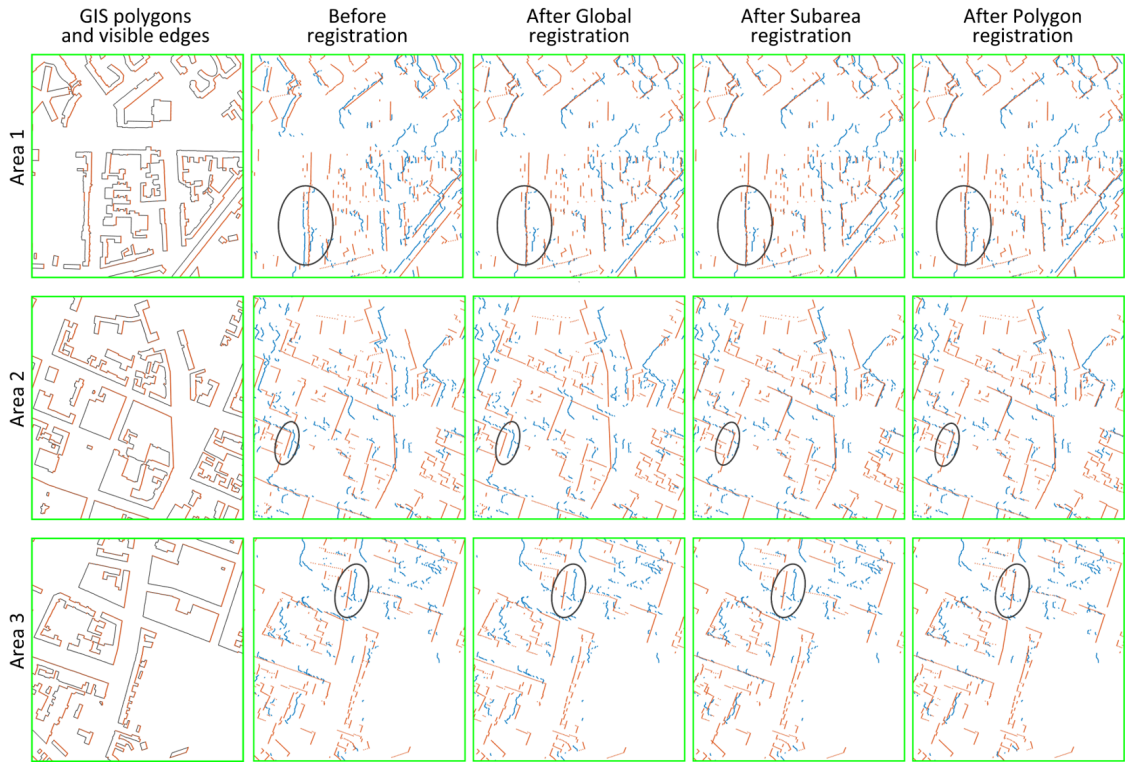


Figure 5.7: GIS features and registration results of each step in Areas 1-3 of Figure 5.5. The first column shows the GIS polygons and the extracted GIS features (red). The second to the last column shows the GIS (red) and SAR (blue) point-sets before registration, after global registration, after subarea registration, and after polygon registration. After global registration, the distance between the two point-sets in Area 1 decreased but increased in Area 2 and Area 2; after subarea registration, the distance decreased in Area 2; after polygon registration, the distance decreased in Area 3.

In Figure 5.7, the second column shows the GIS (red) and the SAR (blue) point-sets before global registration, while the third column shows the point-sets after global registration, in area 1 to area 3 in Figure 5.5. Details can be seen in areas marked by the black ellipses. As can be seen, after global registration, the distance between the two data sets in Area 1 decreased, while the distance increased in Area 2 and Area 3. This is because the global registration aligns the two data sets to minimize overall distance instead of local distance.

b.2 *Registration results at the subarea level*

A set of 16×16 regular grids is defined on the whole region. To contain sufficient points from both the SAR and GIS point-sets, the size of one grid is defined to be larger than the largest GIS polygon (after polygon merging). In each grid area, the distribution of its point-pair distance $D = \bigcup_{i=1}^{n} dp_i$ is calculated.

Based on the distribution of D, the grid cells are classified into three types, shown in Figure 5.8 (left). For magenta grids, the mode of the distribution curve is approximately at 0. Thus no further processing is required. For yellow grids: the mode of the distribution is non-zero. Then adjacent grids with similar distributions are clustered into subareas for subsequent registration. For cyan grids, the distribution

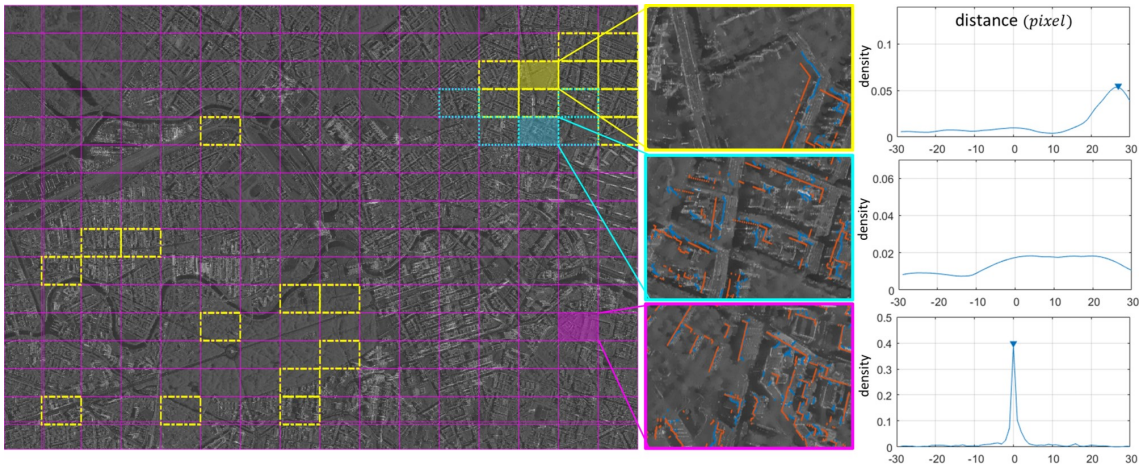


Figure 5.8: The three types of grids in the study area based on the distance distribution. Magenta grids: the peak value of the distance distribution is at 0; yellow grids: the peak value of the distance distribution is at C (non-zero constant); cyan grids: the distance distribution has no clear peak value. An example of each type is given: (middle) the GIS points (red) and the SAR points (blue) are shown in the grids; (right) the distance distribution between the two point-sets that the peak positions are shown in the grids.

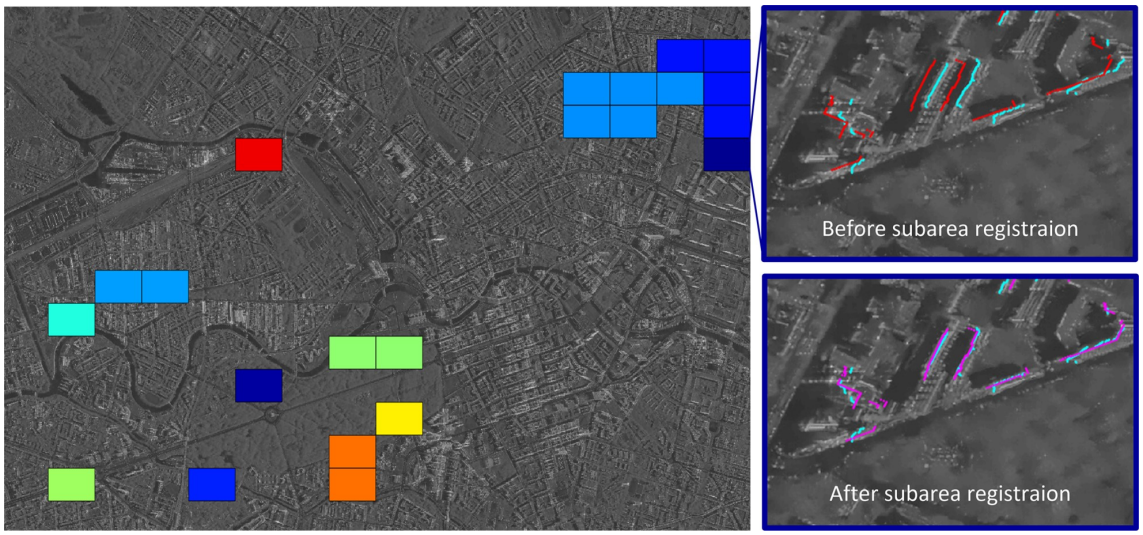


Figure 5.9: Subareas clustering and registration. The yellow grids in Figure 5.8 (left) are clustered into subareas, and each subarea is represented by one color. The GIS and SAR point-sets inside each subarea are then registered. An example before (up) and after (down) subarea registration is shown on the right. The GIS points before (red) are registered to the SAR points (cyan). The transformed GIS points are plotted in magenta.

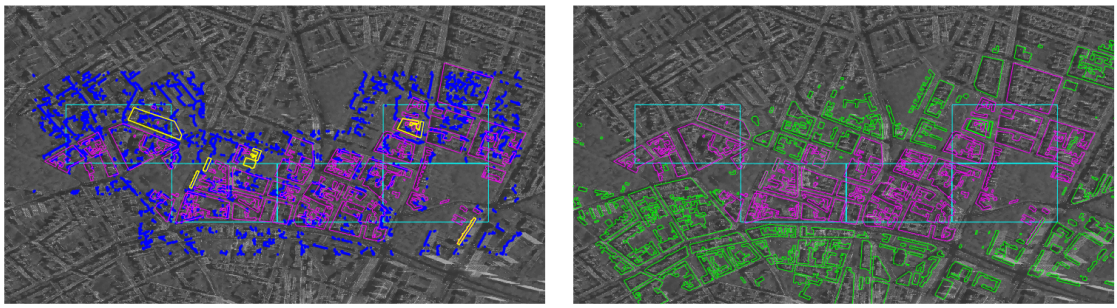


Figure 5.10: Polygon registration in two cases. (left) Among the magenta polygons in the grids, only the yellow polygons have enough corresponding SAR points (blue) to perform polygon registration. (right) For the remaining polygons (magenta), the transformation parameters are searched from their nearest neighbor polygon (candidates polygons shown in green).

has no clear peak. In this case, the constant δH assumption does not hold, meaning the polygons need to be examined further.

Figure 5.8 provides examples of the three abovementioned types of grids. The distributions of their corresponding point-pair distance are plotted on the right side of Figure 5.8: in the magenta grid, D has a clear peak at 0; in the yellow grid, D has a clear peak at around 27 pixels; however, the cyan grid has no clear peak value.

Subareas are clustered from the yellow grids, using DBSCAN, as shown in Figure 5.9 (left). After subarea clustering, ICP was performed in each subarea. Figure 5.9 (right) shows examples of subarea registration: the GIS and SAR point-sets are well-matched after subarea registration.

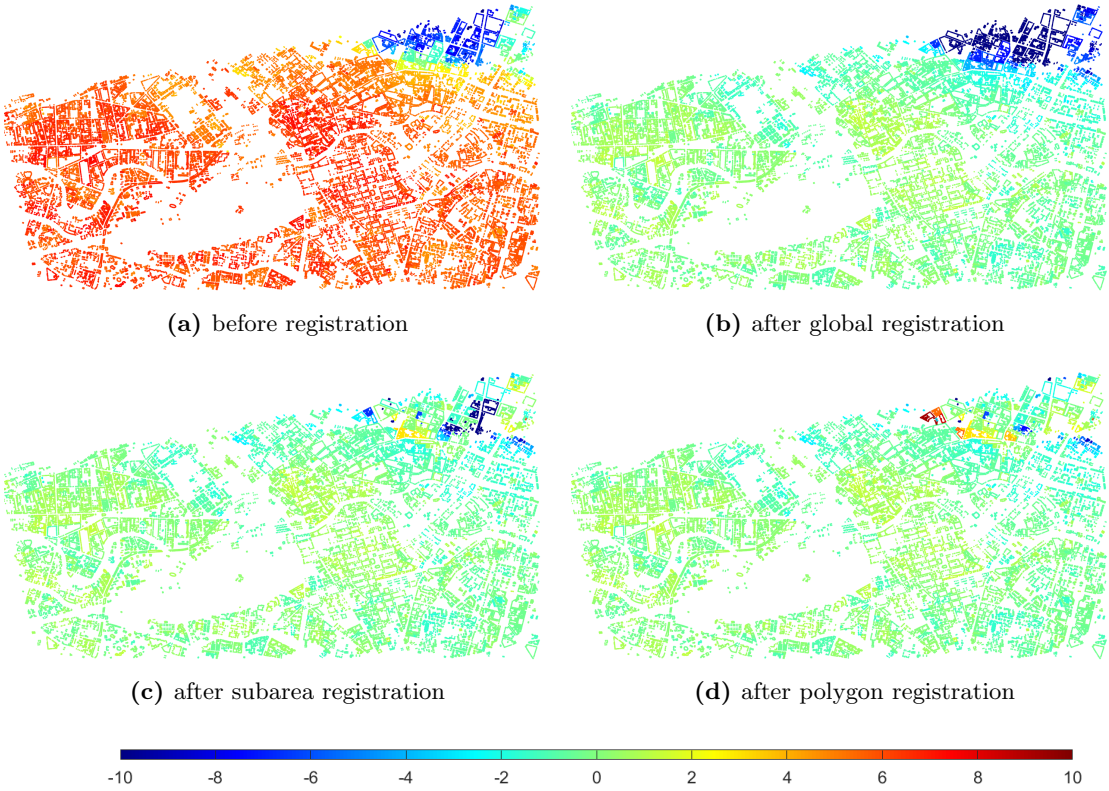
In Figure 5.7, the fourth column shows the GIS (red) and the SAR (blue) point-sets after subarea registration of the three selected areas in Figure 5.5. As can be seen, in Area 2, the distance between the two data sets increased after global registration but decreased after subarea registration, while in Area 3, the distance was still large, so that further registration is needed.

b.3 *Registration results at the polygon level*

When the distribution of D does not show a clear peak, the constant δH assumption does not hold. The registration proceeds to the polygon level, i.e., we seek to find a rigid transformation for each polygon.

Figure 5.10 shows the polygon level registration process in the cyan grids. As shown in Figure 5.10 (a), for all the polygons (magenta) in the grids, the ones with enough corresponding SAR points (blue) are further registered, and the polygons after registration are plotted in yellow. In Figure 5.10 (b), for the rest of the polygons (magenta), their nearest neighbor polygon is searched from all the candidate polygons (shown in green), and the transformation parameters are adopted.

In Figure 5.7, the last column shows the final registration result of the selected areas in Figure 5.5. Compared to the results in previous steps, the distance between the two data sets decreased in Area 3.



(a) before registration **(b)** after global registration

(c) after subarea registration **(d)** after polygon registration

Figure 5.11: Range error maps of vertices in building polygons between registration results and ground truth: (a) before registration, (b) after global registration, (c) after subarea registration, and (d) after polygon registration. Errors are color-coded (meters), in the interval [-10 10].

5.4.3 Evaluation

The performance of the proposed algorithm is evaluated using the registration range error δrg.

The δrg of each vertex in GIS polygons is shown in Figure 5.11, where subfigures (a) through (d) are the error maps before registration, after global registration, after subarea registration, and after polygon registration, respectively. As can be seen, the range difference is not centered at 0 before registration: the majority of δrg is around positive 6 meters, whereas δrg is negative in the upper-right hand corner of the study area. After global registration, this bias is such that the majority of δrg is shifted to 0. Subarea registration and polygon registration further decreases local δrg variations.

The final result shows that the average range difference is reduced from 5.91 m to -0.08 m, and the standard deviation of the range difference is reduced from 2.77 m to 1.12 m. The bias and the standard deviation of the errors are listed in Table 5.2, and the histograms of the range errors before and after registration are plotted in Figure 5.12.

Table 5.2: The bias and standard deviation of the registration errors, comparing the registration results in each step to the ground truth.

Error (m)	Bias	Standard deviation
before registration	5.91	2.77
global registration	-0.18	2.77
subarea registration	-0.11	1.43
polygon registration	-0.08	1.12

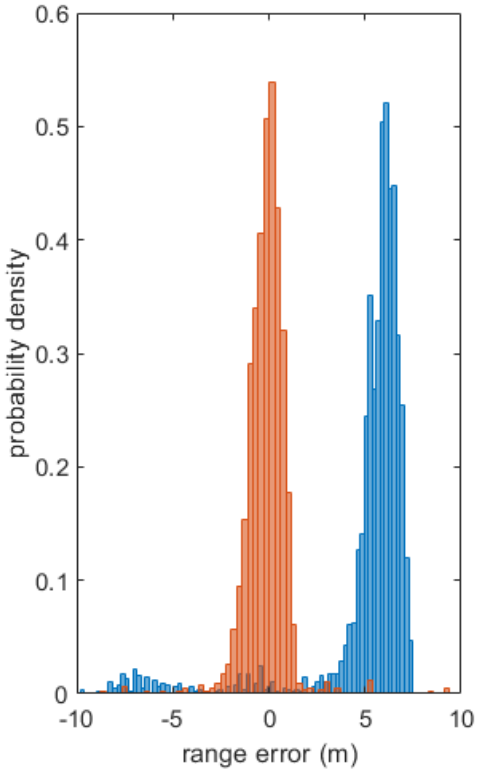


Figure 5.12: The probability density of vertex distance before (blue) and after registration (orange), compared with ground truth.

5.5 Discussion

5.5.1 Can the Proposed Approach Work with Stripmap Images?

So far, the proposed algorithm is tested on a high-resolution spotlight TerraSAR-X image. Since stripmap images are globally acquired, it is very interesting to know if the proposed registration approach can be applied to stripmap data. Therefore, a test is performed using a TanDEM-X stripmap image in the central Munich area. Figure 5.13 shows the used SAR image and building footprints. The building footprints are extracted from Planet images [19] and are subsequently radar-coded to the SAR image with a constant height value of 520 m. Features are extracted from the SAR image and the GIS data and are subsequently registered.

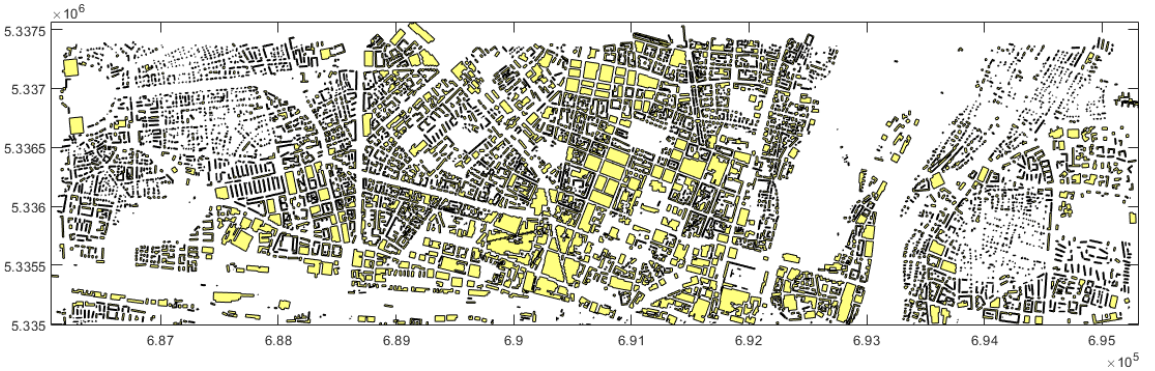
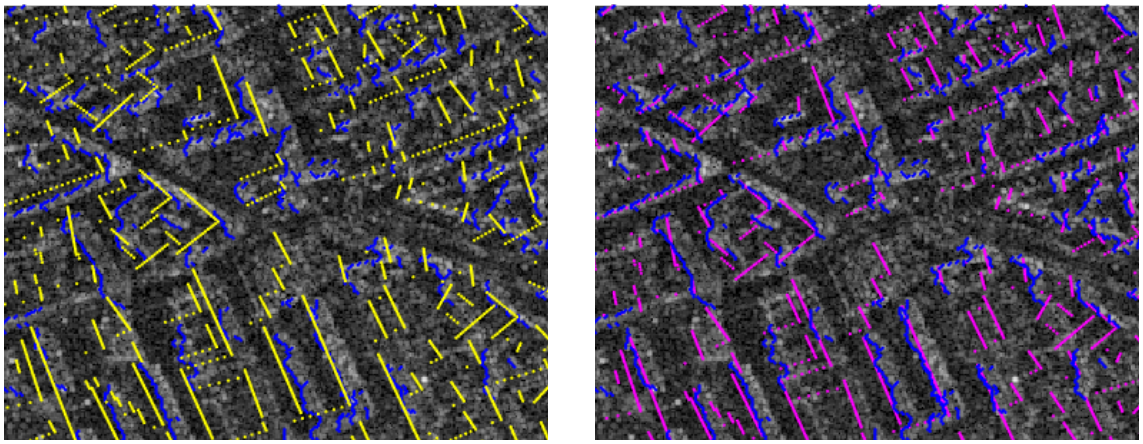


Figure 5.13: Data to be registered: (up) Tandem-X stripmap image and (down) building footprint polygons in central Munich area.



(a) before registration **(b)** after registration

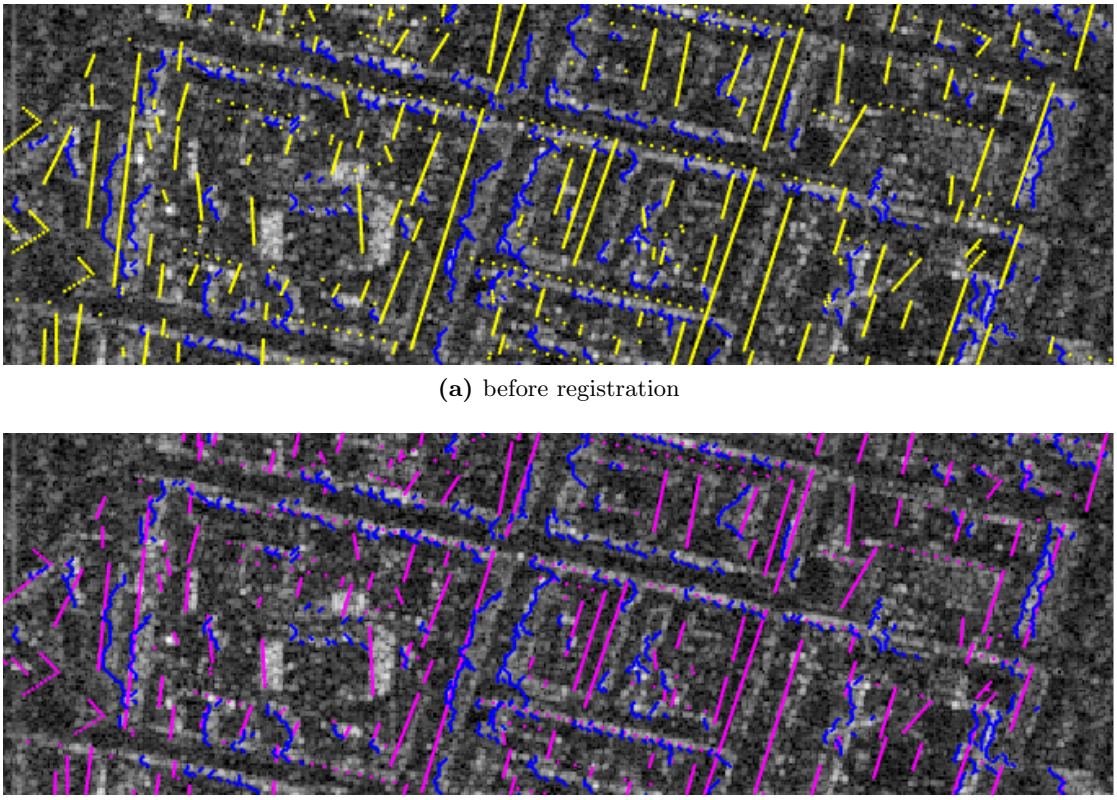
Figure 5.14: Comparison of the SAR point-set (blue) and GIS point-set (yellow) before and (magenta) after registration in Area 1.

Due to the lack of TomoSAR point clouds generated from stripmap images in this area, the ground truth cannot be generated for evaluation. Instead, the results are visually examined by comparing the building footprint and SAR point-sets before and after registration.

The registration results in two subareas of the study region are plotted in Figure 5.14 and Figure 5.15. The SAR point-set is fixed (plotted in blue). Comparing the building

footprint point-set (yellow) before and (magenta) after registration, it can be observed that the registration approach has aligned the two point-sets well in both the areas.

This test shows that the proposed registration method is capable of dealing with SAR images in both spotlight and stripmap modes. Relying on the registration methods, both spotlight and stripmap SAR images can be used with building footprint data for analysis on individual buildings.



(a) before registration

(b) after registration

Figure 5.15: Comparison of the SAR point-set (blue) and GIS point-set (yellow) before and (magenta) after registration in Area 2.

5.5.2 Which Terrain Model to Use for Radar Coding?

In general, an accurate terrain model for radar coding is preferred to minimize the range error caused by the height error. However, such terrain models are not available for most areas. For globally available digital elevation models (DEMs), such as the Shuttle Radar Topography Mission (SRTM), the Advanced Spaceborne Thermal Emission and Reflection Radiometer (ASTER) Global Digital Elevation Model (GDEM), and TanDEM-X DEM, the vertical accuracy is often limited [264]. In addition, the DEMs are surface models instead of terrain models. To acquire terrain heights, a filtering algorithm must be applied. Consequently, the accuracy of the terrain heights is further influenced by the choice of filtering algorithm.

To show the effect of radar coding using different terrain models, three terrain models for radar coding are demonstrated in the study area: a filtered SRTM DEM, a filtered TanDEM-X DEM, and a flat terrain model. Figure 5.16 (left) shows the range errors over

the whole region using filtered SRTM DEM, Figure 5.16 (middle) shows the range errors using filtered TanDEM-X DEM, and Figure 5.16 (right) shows the range errors using a flat terrain. Obviously, in the case of radar coding using a flat terrain model, the range errors are similar in most areas that can be corrected with global registration. However, in the case of the filtered DEMs, the range error varies more significantly, which increases the difficulty of registration.

Our experience is that a flat terrain model is the first choice for radar coding in urban areas. Since most cities are located in relatively flat areas, the height error is constant for most areas, and the resulting constant range error can be corrected in the global registration step. The residual range errors in non-flat areas are inconstant and can then be captured and further corrected in the local registration steps. In the presence of highly varied terrain, a constant height is not enough. A heavily filtered DEM is therefore preferred.

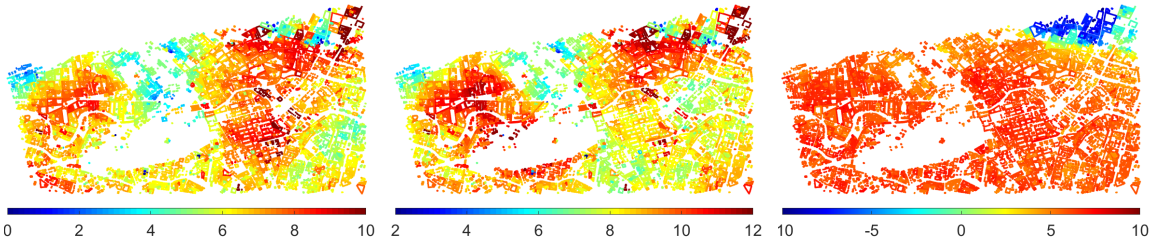


Figure 5.16: Range error maps compared to the ground truth, resulting from radar coding using different heights: (left) height from SRTM DEM (median filtered with window size 50×50), (middle) height from TanDEM-X DEM (median filtered with window size 50×50), and (right) constant height of 77.5 m. Errors are color-coded (meters). Note that the ranges in the three colormaps are different.

Figure 5.11(d) shows that after registration, the majority of range differences are around 0. In comparison to Figure 5.11(c), in the polygon registration step, the range difference further decreased for most polygons, except for several polygons. The reason is that these polygons do not have enough nearby SAR points to perform ICP and thus adopted the transformation parameters of their neighboring polygons. However, these transformation parameters are incorrect if the ground heights of the neighboring polygon differ from the original polygons' heights. Therefore, these errors are inevitable in such cases.

5.5.3 Applicable Scenario

The proposed method relies on the correspondence between the SAR image and the GIS building footprints. This requires sufficient corresponding features in the two data sets. In SAR images, the double bounce lines are mainly affected by the layover, shadowing areas, which are explained in detail as follows. Figure 5.17 illustrates SAR imaging geometries in different situations, where θ is the incidence angle; h and r are the height and width of building B, respectively; d is the distance between two buildings; and lr, lw, and lf are the layover areas of the roof, wall, and footprint, respectively. The blue arrow marks the bottom of the sensor-facing wall, while the red arrow points to the double bounce line position. We can, in general, consider the following two cases:

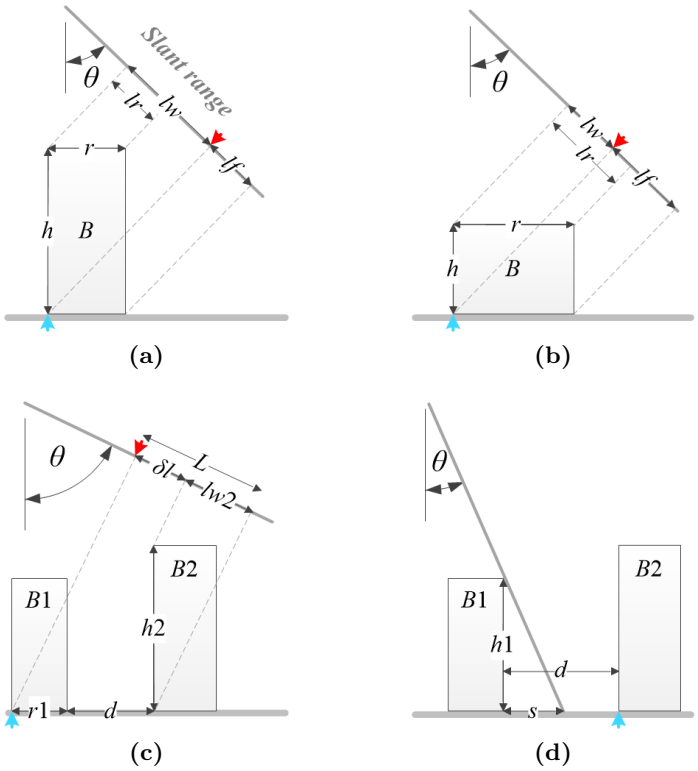


Figure 5.17: SAR imaging geometry of buildings under different settings, where θ is the incidence angle; h is the building height, r is the building width, lr, lw, and lf are the roof area, the wall area, and the footprint area in the SAR image, respectively. The blue arrow marks the bottom of the sensor-facing wall, while the red arrow points at the double bounce line position. (a) and (b) show a single building with different heights and widths. In (a), the double bounce line is detectable, while it is not in (b). (c) and (d) show two buildings. In (c), the double bounce line of building $B1$ is detectable because $\delta l > 0$. In (d), the double bounce line of building $B2$ is detectable because the distance between the two buildings exceeds the shadow area of building $B1$.

a. Single building: the double bounce line should not be overlaid with the building roof.

   Figures 5.17(a)(b) show a building under the same θ, but with different h and r. Usually, the SAR image intensity is low at lf and high at lr and lw, depending on the number of structures on the roof and the wall. As can be seen, in Figure 5.17(a), lr lies in lw, and so the double bounce line between lw and lf is detectable. In contrast, in Figure 5.17(b), lw lies in lr, and thus the double bounce line does not show a clear signature when roof area lr shows high intensity.

   To ensure that the double bounce line is detectable, lw should not be covered by lr, i.e., $lw \geqslant lr$. Since $lw = h \cdot cos\theta$, $lr = r \cdot sin\theta$, the requirement is therefore: $\frac{h}{r} \geqslant \tan\theta$.

b. Multiple buildings:

This case is simplified to two buildings: $B1$ as the front building and $B2$ as the rear building, with respect to the sensor. There are two considerations.

First, the double bounce line of the front building shall not layover with the rear building. As shown in Figure 5.17(c), to ensure that the double bounce line of $B1$ is detectable, it should not be mixed with the layover area of $B2$, i.e., there must be an area $\delta l > 0$ with lower intensity between the double bounce line and $lw2$. Since $\delta l = L - lw2, L = (r1 + d) \cdot \sin\theta, lw2 = h2 \cdot \cos\theta$, the requirement is therefore: $\tan\theta > \frac{h2}{r1+d}$.

Second, the bottom of the sensor-facing wall of the rear building should not be occluded by the front buildings. As shown in Figure 5.17(d), to ensure that the double bounce line of $B2$ is detectable, $B2$ should not be occluded by the front building $B1$, i.e., the shadow s of $B1$ should be no bigger than the distance d between $B1$ and $B2$. Since $s = h1 \cdot \tan\theta$, the requirement is therefore: $d \geqslant h1 \cdot \tan\theta$.

Hence, the multiple building case requires: $\frac{h2}{r1+d} < \tan\theta \leqslant \frac{d}{h1}$.

The SAR geometries are reflected by the urban morphology of the study area, i.e., the compactness and the building heights, as well as the incidence angle of the SAR image used. Buildings in open areas can be regarded as the single building case, i.e., the constraint is $\frac{h}{r} \geqslant \tan\theta$. When θ is fixed, study areas with taller buildings are preferred. If most buildings are low-rise or have a larger width, SAR images with smaller θ should be chosen. Buildings in a compact area can be regarded as the multiple buildings case, i.e., in addition to $\frac{h}{r} \geqslant \tan\theta$, the constraints are $\frac{h2}{r1+d} < \tan\theta \leqslant \frac{d}{h1}$. When θ is fixed, the average building height in the study area should be small. If most buildings in the study area are high-rise, the SAR image should be chosen so that θ meets the above inequity constraints. However, when $\frac{h2}{r1+d} = \frac{d}{h1}$, there is no solution for θ, meaning that if the study area is very compact and has many high-rise buildings, the severe occlusion together with layover effects prevent the extraction of sufficient double bounce lines. Our method cannot handle this situation.

In practice, 3-D information of study areas is often unknown, so that it is impossible to analyze layover and occlusion for individual buildings. Therefore, pre-knowledge of the urban morphology in the study area is of great benefits, such as a local climate zone classification map [258].

5.5.4 Further Applications

The result of automatic registration of one SAR image and GIS building footprints can be used in data fusion for different applications. Next two chapters detail the LoD1 building reconstruction using building footprints and SAR images. Besides, two potential applications are particularly interesting:

First, after being registered to a SAR image, the GIS building footprint polygons can be used as iso-height lines in the range direction for object-level reconstruction. The iso-height lines can provide shape prior to building height estimation from a SAR image [15,48]. The iso-height lines can also be used to group pixels for tomographic inversion using the joint sparsity [186]. Second, the registered GIS and SAR data offer the potential of generating large training datasets for building classification. The attributes contained in GIS data can be directly used as ground truth and be learned from SAR image classification. For example, the building function can be learned from a large dataset containing SAR images

and GIS ground truth labels. Indeed, SAR image classification at the building level is difficult in comparison to using optical images. However, the huge data quantity has potential, especially in areas where cloud coverage limits the use of optical images.

5.6 Summary

This chapter presents a framework for automatically registering 2-D building footprints to a corresponding SAR image. The proposed framework relies on the corresponding features in building footprints and SAR images and registers the two data progressively in three levels, allowing the algorithm to cope with variations in the local terrain. The experiments in Berlin using a TerraSAR-X high-resolution spotlight image shows that the proposed algorithm reduced the average distance error to -0.08 m and the standard deviation to 1.12 m. Such accuracy, better than half of the typical urban floor height (3m), is significant for building height reconstruction on a large scale. Further experiments using a stripmap image also show promising results.

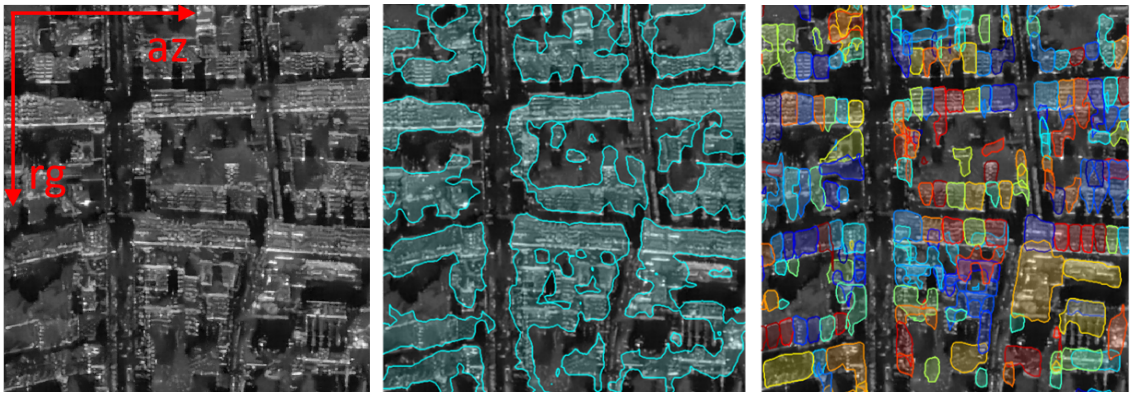
The proposed registration framework lays the groundwork for building footprints to assist SAR image interpretation on a large scale. The next two chapters utilize the registered building footprints and SAR images for the purpose of LoD1 building reconstruction.

6 Conditional GIS-aware Network for Individual Building Segmentation in VHR SAR Images

Once the building footprints are registered to the SAR image on a large scale, as presented in Chapter 5, they are ready to be employed to assist building analysis in SAR images. With pixel-wise ground truth annotations generated using the method in Section 4.2, now the conditions are met for applying deep neural networks.

This chapter proposes a novel conditional GIS-aware network (CG-Net) that takes a SAR image and a building footprint as inputs and predicts the building area associated with the footprint in the SAR image. This work targets at individual buildings, as shown in Figure 6.1.

The main contents of this chapter are summarised in [237]. The approach of using segmentation results for LoD1 building reconstruction is presented and discussed in Section 6.3.1, and additional experiments are provided in Section 6.3.2.



(a) SAR image **(b)** Semantic segmentation results in [138] **(c)** Individual building segmentation results in this chapter

Figure 6.1: Illustration of the difference between building semantic segmentation and individual building segmentation. In (b), all buildings are assigned the same label, while in (c), each individual building is identified as one class.

6.1 Conditional GIS-aware Network

This work aims to train a network that takes a SAR image and building footprint as inputs and predicts the building area associated with the footprint in the SAR image. Aiming to find out a more suitable way of using building footprints for this task, two

representations of building footprints are compared, namely complete building footprints and sensor-visible footprint segments.

Since footprints and visible segments generated from GIS data can provide precise geometry and location information, we resort to exploiting such cues in our task and devise a network module that performs a conditional GIS-aware normalization. By utilizing the CG module, our network, termed as CG-Net, can learn feature representations from not only SAR but also GIS data. Specifically, VGG-16 [265] is employed as the backbone of CG-Net to learn multi-level features from SAR images. Afterward, outputs of the last three convolutional blocks are upsampled and fed into the CG module separately. Meanwhile, footprints or visible segments are imported into the CG module as complementary inputs in order to yield final predictions. In what follows, Section 6.1.1 illustrates the procedure of multi-level feature extraction. Section 6.1.2 introduces details of our CG module, and Section 6.1.3 details the configuration of our CG-Net.

6.1.1 Multi-level Feature Extraction Module

VGG-16 [265] is used as the backbone of our network to extract features from multiple layers, as these multi-level features help in recognizing buildings with variant scales. The backbone consists of five convolutional blocks, and each of them contains two or three convolutional layers. The size of their filters is 3×3. Outputs of all convolutional layers are activated by ReLU [266], and 2×2 max-pooling layers with a pooling stride of 2 are interleaved among these blocks. Features learned from deep layers are considered to include high-level semantics, while those from shallow layers are low-level. Therefore, in this task, we utilize features learned from the last three blocks, i.e., *Block3*, *Block4*, and *Block5* (see Figure 6.2). Afterward, the extracted features are fed into the CG module separately.

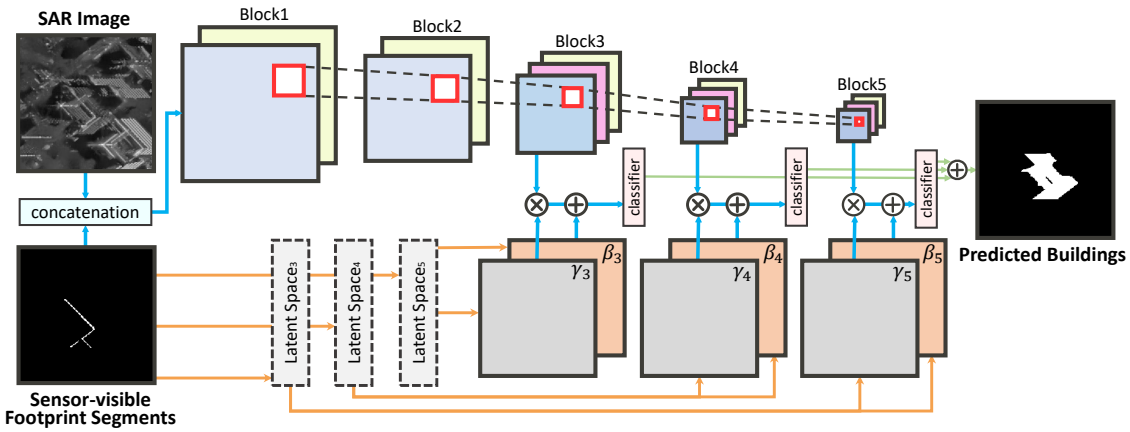


Figure 6.2: Overview of the CG-Net architecture.

6.1.2 Conditional GIS-aware Normalization Module

An intuitive way to make use of GIS data is to simply concatenate them with SAR images and then feed them to a vanilla semantic segmentation network, such as fully convolutional networks (FCN). However, such a method might suffer from the inefficient use of GIS data and leads to unstructured predictions (see the third column in Figure 6.7). To address this

issue, we propose a conditional GIS-aware normalization module to distill the geometry information of individual buildings from GIS data and normalize final predictions with such information. Formally, let $\boldsymbol{m}_{gis}$ be the mask of the complete building footprint or sensor-visible footprint segments with a spatial size of $W \times H$, and $\boldsymbol{x}_b$ denotes feature maps extracted from the b-th convolutional block. The width and height of $\boldsymbol{x}_b$ are represented as W' and H', respectively. The number of channels is denoted as C'. We consider a naive conditional normalization procedure as follows:

$$\hat{\boldsymbol{x}}_b = \gamma_b \boldsymbol{x}_b + \beta_b, \tag{6.1}$$

where, γ_b and β_b represent a scale factor and a bias, respectively, and they indicate to what extent $\boldsymbol{x}_b$ should be scaled and shifted. The normalized $\boldsymbol{x}_b$ is denoted as $\hat{\boldsymbol{x}}_b$. A commonly-used measure of γ and β is to calculate the standard deviation and mean of $\boldsymbol{x}_b$. Since $\boldsymbol{x}_b$ consists of more than one channel, γ and β are often computed in a channel-wise manner, and thus, Eq. (6.1) can be rewritten as

$$\hat{\boldsymbol{x}}_{b,c} = \gamma_{b,c}(\boldsymbol{x}_{b,c}) \cdot \boldsymbol{x}_{b,c} + \beta_{b,c}(\boldsymbol{x}_{b,c}), \tag{6.2}$$

where c denotes the c-th channel of $\boldsymbol{x}_b$ and ranges from 1 to C'. This equation can be easily extended to the batch normalization [267] by computing the standard deviation and mean of each $\boldsymbol{x}_{b,c}$ in a batch.

In our case, we want to normalize feature representations learned from SAR images, conditioned on GIS data. Our insight is that the GIS data imply coarse localization cues, and their use can guide the network to segment individual buildings accurately. Therefore, we reformulate Eq. (6.2) as follows:

$$\hat{\boldsymbol{x}}_{b,c,p,q} = \gamma_{b,c,p,q}(\boldsymbol{m}_{gis}) \cdot \boldsymbol{x}_{b,c,p,q} + \beta_{b,c,p,q}(\boldsymbol{m}_{gis}), \tag{6.3}$$

where $\gamma_{b,c,p,q}$ and $\beta_{b,c,p,q}$ indicate the scale factor and bias *learned* specifically for the pixel located at (p, q) in the c-th channel of $\boldsymbol{x}_b$. As a consequence, normalization parameters $\boldsymbol{\gamma}_b$ and $\boldsymbol{\beta}_b$ are formatted as matrices with a size of $W' \times H' \times C'$. Such a design enjoys an advantage that normalization parameters are learned in a data-driven manner, and thus these parameters are expected to be more adapted to $\boldsymbol{x}_b$. As to the implementation of Eq. (6.3), we first project $\boldsymbol{m}_{gis}$ onto a latent space through 3×3 convolutions and then employ two convolutional layers to learn γ_b and β_b from the encoded $\boldsymbol{m}_{gis}$. Subsequently, the element-wise multiplication of $\gamma_b(\boldsymbol{m}_{gis})$ and $\boldsymbol{x}_b$ is performed, and the output is added to $\beta_b(\boldsymbol{m}_{gis})$ pixel by pixel. Figure 6.3 illustrates the architecture of our CG module.

6.1.3 Configuration of CG-Net

In order to fully exploit GIS data at multiple scales, we append three CG modules to the last three convolutional blocks of the backbone (see Figure 6.2). However, a question is that spatial and channel dimensions of the extracted multi-level features are inconsistent with those of complete building footprints/sensor-visible footprint segments. To address this issue, we upsample these multi-level feature maps to match the spatial resolution of $\boldsymbol{m}_{gis}$ via bilinear interpolation. Note that doing so would significantly increase the computation overhead of subsequent operations. Hence we reduce the number of feature channels through 1×1 convolutions and modify the CG module (see Figure 6.4) accordingly. Outputs of the CG module are squashed into the number of classes 2, and added

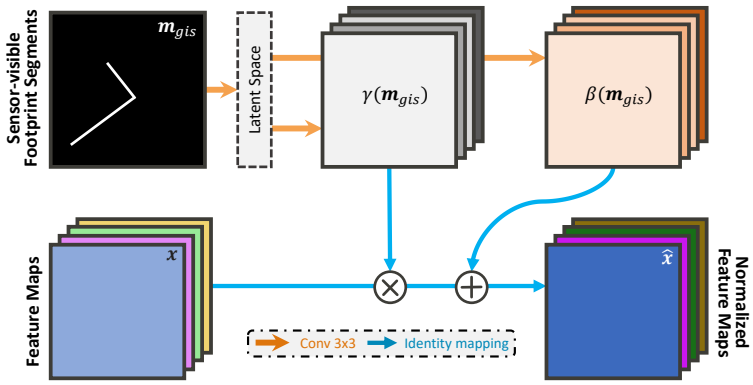


Figure 6.3: Architecture of the proposed CG module. Here, we take the sensor-visible footprint segments as an example. γ and β are normalization parameters learned from the sensor-visible footprint segments and used to normalize input feature maps with Eq. (6.3).

via an element-wise addition operation to produce final segmentation results. Figure 6.2 illustrates the architecture of the proposed CG-Net. Furthermore, we note that the proposed CG module is in a plug-and-play fashion and is flexible enough to enhance other semantic segmentation network architectures, e.g., DeepLabv3. For DeepLabv3, since it already fuses features from different layers in its architecture, we simply add our module right before the last layer.

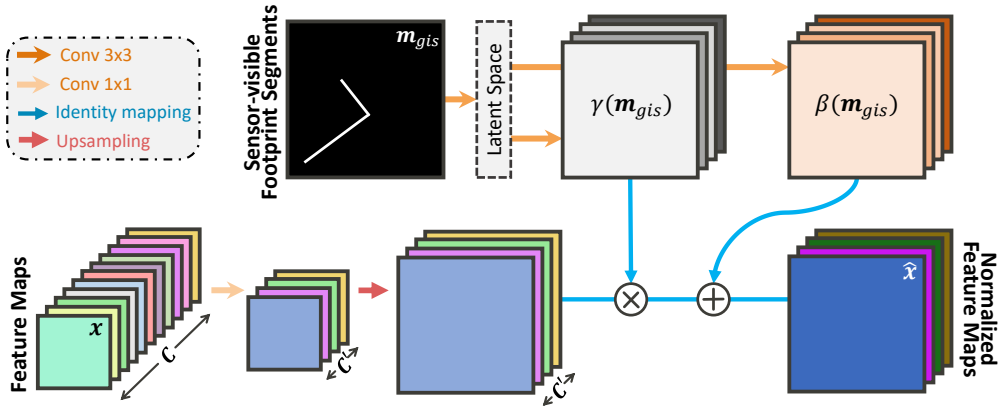


Figure 6.4: Architecture of the final CG module. In advance of performing normalization, the channel of input feature maps is first reduced, and the spatial size is enlarged according to that of sensor-visible footprint segments.

6.2 Experimental Results and Evaluation

6.2.1 Data Set and Training Details

A. *Data Set*

In our data set, one TerraSAR-X image and building footprints are acquired over Berlin, which are the same data used in Chapter 5. Note that unlike in Chapter 5, the connected building footprint polygons are not merged but kept as individual

building polygons. In order to yield ground truth annotations, a highly accurate DEM is employed, which is the same DEM used for ground truth generation in Chapter 5. More details of the used data are described in Section 5.4.1. The study area is the same area tested for the registration framework in Chapter 5. Figure 5.4 shows the study area of Berlin in the intersection area of the two rectangles: the yellow rectangle shows the area of the SAR image, while the red rectangle shows the area of DEM used for ground truth generation. Notably, only data covering the study region are used for generating our dataset.

By using the workflow described in Section 4.2, building annotations and footprints are generated. Since one objective is to explore how GIS data can be effectively used for individual building segmentation, these two versions of footprint masks are produced, namely complete building footprints and sensor-visible footprint segments. The dataset therefore contains a 5736×10312 SAR image, two versions of footprint masks, and ground truths of individual buildings.

B. *Training Details*

In order to train an effective and robust segmentation network, we crop the SAR image into patches of 256×256 pixels with a stride of 150 pixels. Note that patches, including incomplete footprints or ground truth annotations are discarded. Consequently, 30056 buildings are remaining, and each of them has three patches: a SAR image patch, a footprint patch, and a ground truth mask. Among all buildings, 19434 of them are utilized for training networks, and the others are test samples. Note that training and test regions do not overlap. The network takes one SAR patch and the corresponding GIS patch for one building as inputs. After predicting the masks of all buildings, overlapping areas are obtained by overlaying all masks.

During the training phase, components of the proposed CG-Net are initialized with different strategies. Specifically, the multi-level feature extraction module is initialized with weights pre-trained on ImageNet [268], and all convolutional layers in the CG modules are initialized with a Glorot uniform initializer. The network is implemented on TensorFlow and trained on one NVIDIA Tesla P100 16GB GPU for 155k iterations. During the training procedure, all weights are updated through back-propagation, and we select Netrov Adam [269] as the optimizer. Parameters of this optimizer are set as recommended: $\epsilon = 1\text{e}{-}08$, $\beta_1 = 0.9$, and $\beta_2 = 0.999$. The loss is defined as binary cross-entropy, as only two classes are considered in our dataset, i.e., building segments and background. We initialize the learning rate as $2e-3$ and reduce it by a factor of $\sqrt{10}$ once the loss stops to decrease for two epochs. Moreover, we utilize a small batch size of 5 in our experiments.

6.2.2 Quantitative and Qualitative Evaluation

A. *Quantitative Evaluation*

To evaluate the performance of networks, we calculate the F1 score as follows:

$$F1 = 2 \cdot \frac{P \cdot R}{P + R}, P = \frac{tp}{tp + fp}, R = \frac{tp}{tp + fn}, \tag{6.4}$$

where P and R denote the precision and recall, respectively. In addition, the intersection over union (IoU) and overall accuracy (OA) are also calculated for a

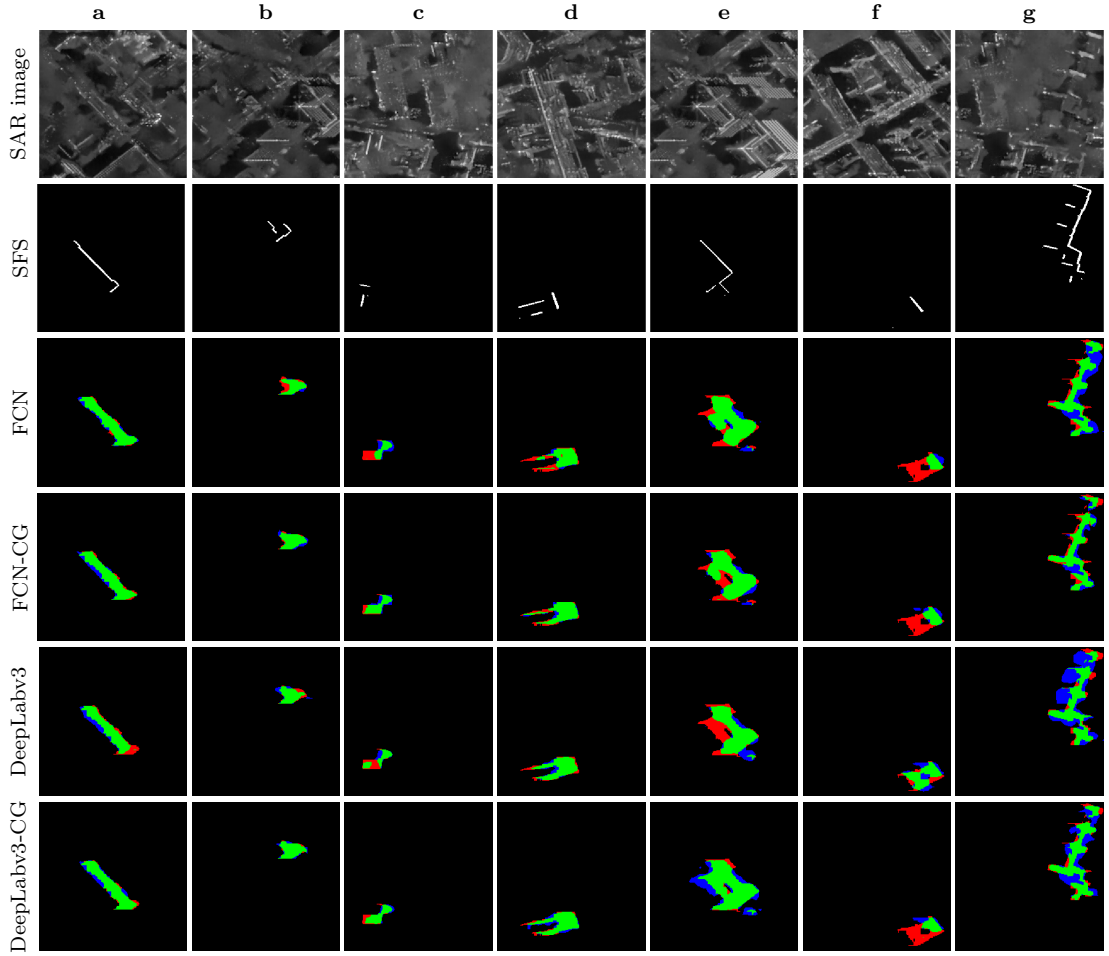


Figure 6.5: Examples of segmentation results using sensor-visible footprint segments (abbreviated as SFS). Pixel-based true positives, false positives, and false negatives are marked in green, red, and blue, respectively.

comprehensive comparison:

$$IoU = \frac{tp}{tp + fp + fn}, OA = \frac{tp + tn}{tp + tn + fp + fn}. \tag{6.5}$$

tp, fp, tn, fn represent pixel-based true positives, false positives, true negatives, and false negatives for buildings, respectively.

In our experiments, we compare four models: FCN, FCN-CG, DeepLabv3, and DeepLabv3-CG. It is worth mentioning that FCN and DeepLabv3 are regarded as baselines, and their inputs are concatenations of SAR patches and their corresponding footprint patches. Both FCN-CG and DeepLabv3-CG are our proposed networks with different backbones.

Table 6.1 reports numerical results of different models on our dataset, where sensor-visible footprint segments are used. Comparison of these results corroborates that the proposed CG module can improve the performance of individual building segmen-

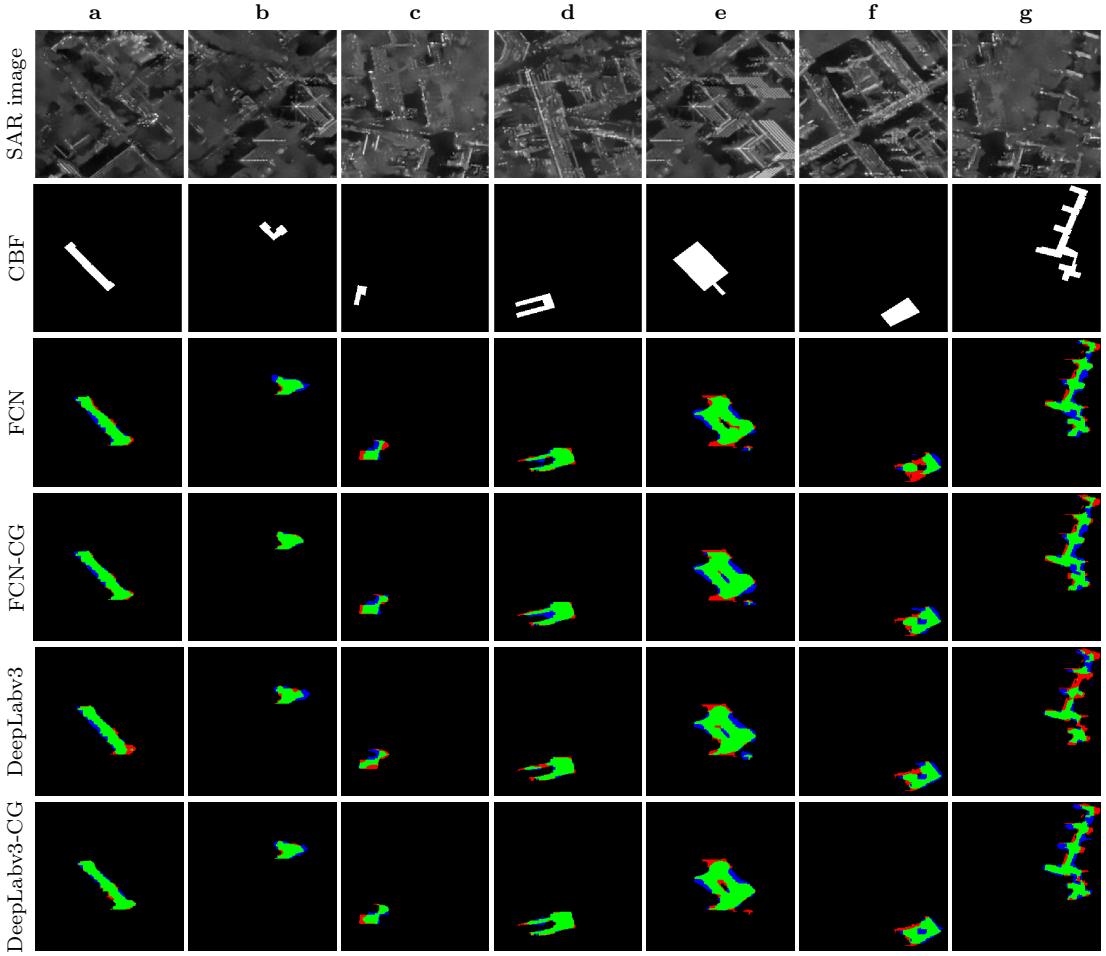


Figure 6.6: Examples of segmentation results using complete building footprints (abbreviated as CBF). Pixel-based true positives, false positives, and false negatives are marked in green, red, and blue, respectively.

tation. Specifically, compared to FCN and DeepLabv3, FCN-CG and DeepLabv3-CG achieve improvements of 0.75% and 2.17% in the precision, respectively. Besides, increments of 1.23% and 1.65% in the mean F1 score and IoU can be observed by comparing FCN-CG and FCN, while improvements of 0.97% and 1.14% in the same metrics are achieved by introducing the CG module to DeepLabv3.

Table 6.2 presents results of variant models using complete building footprints. We can see that the results are consistent with those using sensor-visible footprint segments. For example, with the CG module, the precision improves 1.95% and 3.94% with the backbone, FCN and DeepLabv3, and the IoU increases 1.50% and 2.16%. To summarize, improvements achieved by FCN-CG and DeepLabv3-CG demonstrate the effectiveness of the proposed CG module, and DeepLabv3-CG can achieve the best performance in all four metrics on our dataset. Moreover, we note that all models achieve relatively high OAs, and even the worst model can achieve an OA of

Table 6.1: Numerical results using sensor-visible footprint segments. The highest values of different metrics are highlighted in **bold**.

Model	P	F1 score	IoU	OA
FCN	0.6478	0.6808	0.5138	0.8340
FCN-CG	0.6553	0.6931	0.5303	0.9926
DeepLabv3	0.6635	0.6971	0.5351	0.9927
DeepLabv3-CG	**0.6852**	**0.7068**	**0.5465**	**0.9928**

Table 6.2: Numerical results using complete building footprints. The highest values of different metrics are highlighted in **bold**.

Model	P	F1 score	IoU	OA
FCN	0.7045	0.7242	0.5676	0.9932
FCN-CG	0.7240	0.7362	0.5826	0.9935
DeepLabv3	0.7129	0.7337	0.5794	0.9935
DeepLabv3-CG	**0.7523**	**0.7508**	**0.6010**	**0.9937**

83.40%. This is because OA is computed by considering all pixels, while non-building pixels, which are easily recognized, account for a large proportion.

B. *Qualitative Evaluation*

In addition to the quantitative evaluation, we visualize several segmentation results in Figure 6.5 and 6.6. Pixel-based true positives, false positives, and false negatives are presented in green, red, and blue, respectively.

Figure 6.5 shows results of models using sensor-visible footprint segments. We can observe a general improvement in quality from FCN/DeepLabv3 to FCN-CG/DeepLabv3-CG, especially for buildings in columns b, c, and g. For buildings with simple structures (e.g., the building in column a), all models are able to offer satisfactory segmentation results, while for those with complicated shapes (see column e), large under-segmentation areas (cf. red pixels) can be seen in predicted building masks. Besides, the utilization of the proposed CG module can effectively reduce over-segmentation in final predictions.

Figure 6.6 presents results of models using complete footprints. They indicate that our CG module can ease both over-segmentation (cf. blue pixels in column b) and under-segmentation (cf. red pixels in column e) problems to a considerable extent. Moreover, examples in the third row, column f and the fifth row, column f show that the connectivity of segmentation results are disrupted (cf. green pixels), while the integration of the CG module can alleviate such a problem. A similar phenomenon can also be seen in columns d and g that exploiting the CG module can enhance the connectivity of predictions. In summary, the proposed CG module effectively improves segmentation results.

6.2.3 Comparison of Complete Building Footprints and Sensor-visible Footprint Segments

From Table 6.1 and 6.2, it can be seen that models trained with complete building footprints surpass those trained with sensor-visible footprint segments. For instance,

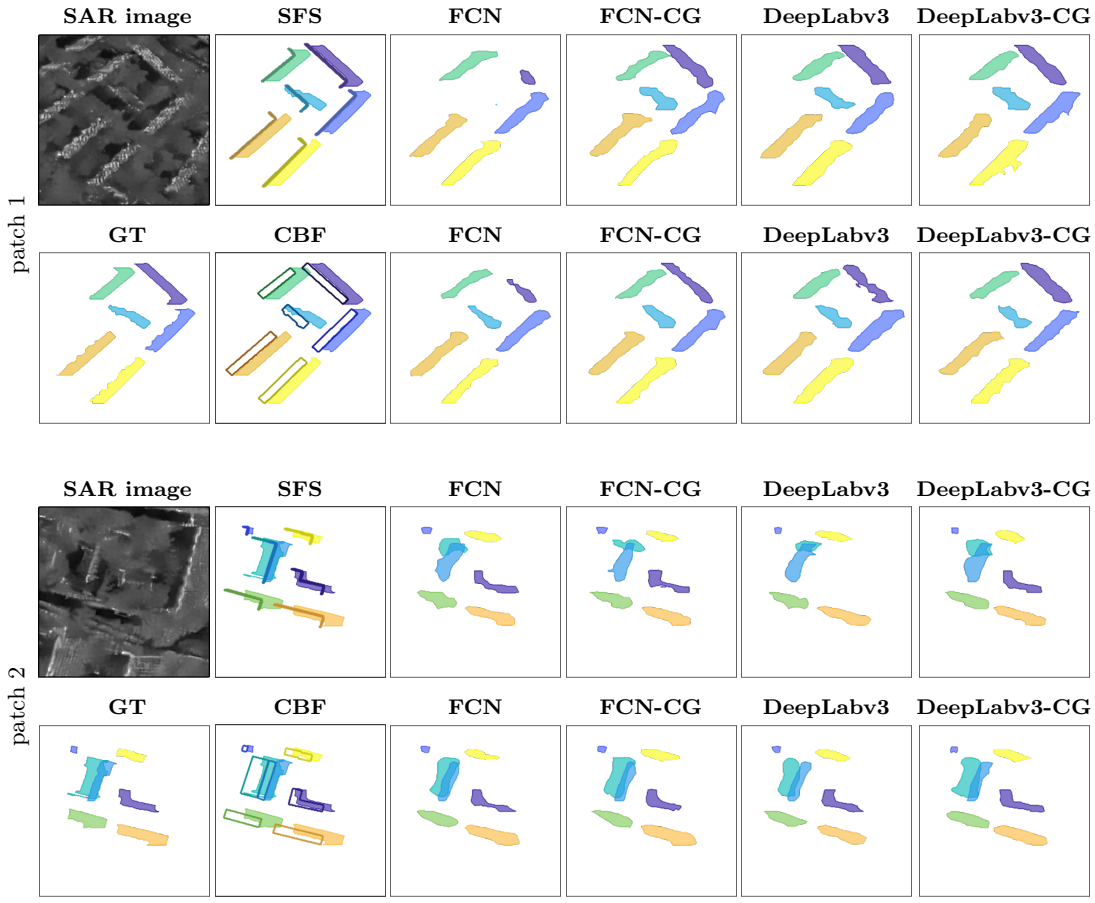


Figure 6.7: Examples of segmentation results from different models on two patches, using complete building footprints (abbreviated as CBF) and sensor-visible footprint segments (abbreviated as SFS). In the second column, CBF and SFS are overlaid on the ground truth (GT) to visualize the difference between building footprints and buildings. Different buildings are plotted in different colors (50% transparency).

DeepLabv3-CG trained on complete footprints improves the F1 score and IoU by 4.40% and 5.45%, respectively, compared to that learned with sensor-visible segments.

Figure 6.7 provides segmentation results of two patches using two versions of footprint masks, and different buildings are marked in different colors (50% transparency). Note that individual building masks are predicted separately, and then masks of buildings in the same patch are plotted together to visualize the overlapping areas. Here, patch 1 presents a simple scenario in which buildings are isolated and show clear signatures in the SAR image. In this case, all models can obtain good segmentation results. Patch 2 shows a fairly complicated scene, where two consecutive buildings exist in the center (cf. buildings in cyan and blue), and SAR signatures are unclear. Although all networks can still successfully segment isolated buildings, the two overlapped buildings are not correctly segmented by models trained with sensor-visible footprint segments (see the third row of Figure 6.7). This is because the mask of sensor-visible footprint segments for the

building on the left contains only one edge, which does not provide adequate information. Moreover, we notice that the overlapping region between these two buildings can only be well identified by models trained with complete building footprints.

Overall, these results suggest that complete building footprints are more befitting for the segmentation of individual buildings than sensor-visible footprint segments. This may be because the former delivers more information, especially for low-rise buildings.

6.2.4 Can CG-Net Work with Inaccurate GIS Data?

So far, building footprints used in our experiments are highly accurate as they are acquired from official GIS data. However, most openly available GIS data, such as OSM, often contain positioning errors. To test the performance of CG-Net in such cases, we conduct supplementary experiments on training our CG-Net with inaccurate building footprints, and discuss the impact of positioning errors in GIS data.

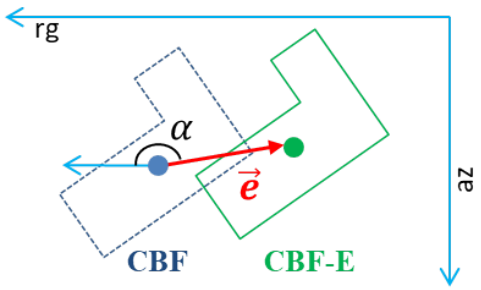


Figure 6.8: Illustration of the process generating building footprints with positioning errors. Positioning error $\vec{e}$ is added to building footprint CBF, resulting in CBF-E. rg and az denote the range direction and azimuth direction, respectively. α is the angle between $\vec{e}$ and rg.

First, we generate inaccurate CBF, termed CBF-E, by injecting positioning errors. As illustrated in Figure 6.8, $\vec{e}$ is the added positioning error, and α is the angle between $\vec{e}$ and the range direction. According to the quality assessment study of OSM in [60], the average offset of building footprints is 4.13 m with a standard deviation of 1.71 m. Therefore we consider the positioning error as a variable whose magnitude is Gaussian distributed, i.e., $|\vec{e}| \sim \mathcal{N}(\mu = 4.13, \sigma^2 = 1.71^2)$. Since the offset may point to different directions, we assume the direction of $\vec{e}$ is uniformly distributed, i.e., α is uniformaly distributed in the range of $[0°, 360°)$. For simplicity, let α be discrete: $\alpha \sim \text{DiscreteUniform}(0°, 359°)$. Note that this is the most difficult case that all footprints contain positioning errors.

Then, we train DeepLabv3-CG using CBF-E and SAR patches and test the trained network with a clean test set. DeepLabv3-CG is chosen because it performs best among all the networks. The parameter settings of the network remain the same as previous experiments, as described in Section 6.2.1.

The results are listed in Table 6.3. As can be seen, compared to results using CBF, the precision of the network trained on CBF-E is decreased by 3.02%, the F1 score is reduced by 3.62%, and the IoU is decreased by 4.5%. However, it still gives competent segmentation results. For visual comparison, Figure 6.9 shows results of DeepLabv3-CG trained with CBF-E and CBF. For the building in column c, DeepLabv3-CG trained with CBF performs much better than that with CBF-E. However, the predictions for buildings

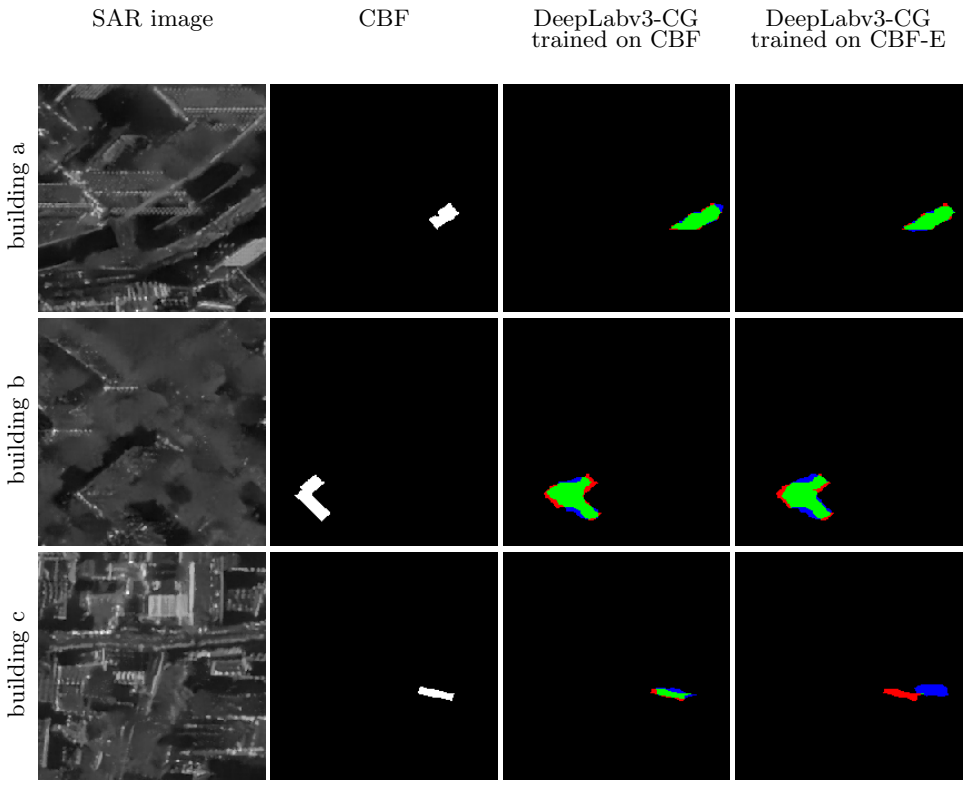


Figure 6.9: Examples of segmentation results of networks trained using complete building footprints (abbreviated as CBF) and networks trained using building footprints with positioning errors (abbreviated as CBF-E). Pixel-based true positives, false positives, and false negatives are marked in green, red, and blue, respectively.

Table 6.3: Numerical results of DeepLabv3-CG trained using CBF and CBF-E.

GIS data used for training	P	F1 score	IoU	OA
CBF	0.7523	0.7508	0.6010	0.9937
CBF-E	0.7221	0.7146	0.5560	0.9927

in columns a and b are visually very similar. Moreover, we observed that predictions from DeepLabv3-CG trained on CBF-E are satisfactory for most buildings.

The experiments show that although weakened by positioning errors in GIS data, the proposed CG-Net is robust even in the most difficult case. This finding suggests that a large amount of existing open-sourced GIS data, such as OSM, can be exploited for segmenting individual buildings in SAR images.

6.3 Discussion

6.3.1 Further Application: Reconstruction of LoD1 Building Models from a SAR Image

Building models can be created at different LoDs. This section demonstrates the process of reconstructing LoD1 models using our predicted individual building masks. Here, the building height is regarded as the average roof height[1].

Figure 6.10 illustrates the projection geometry of two flat-roof buildings in a constant azimuth profile of a SAR image. θ is the incidence angle. l, r, and f denote the length of layover, roof, and footprint areas in the slant-range SAR image, respectively. Notably, the building region in the SAR image contains both the layover and the roof areas. The layover area coincides with the building region when the building height h is large, e.g., the case in Figure 6.10 (left), and it is covered by the building region when h is small, e.g., the case in Figure 6.10 (right). In both cases, the layover area can be calculated by subtracting the footprint from the building region. Therefore, l is estimated to be the length of the layover area in the slant-range direction, and h can be computed with the following equation:

$$h = l/cos\theta. \tag{6.6}$$

From the predicted individual building masks (cf. Figure 6.12(a)), building heights are calculated with Eq. (6.6), and the results are shown in Figure 6.12(b). We further evaluate the estimated height against the mean height from the accurate DEM for each building. The mean absolute height error we achieve in the study site is 2.39 m. The histogram of height errors is shown in Figure 6.11.

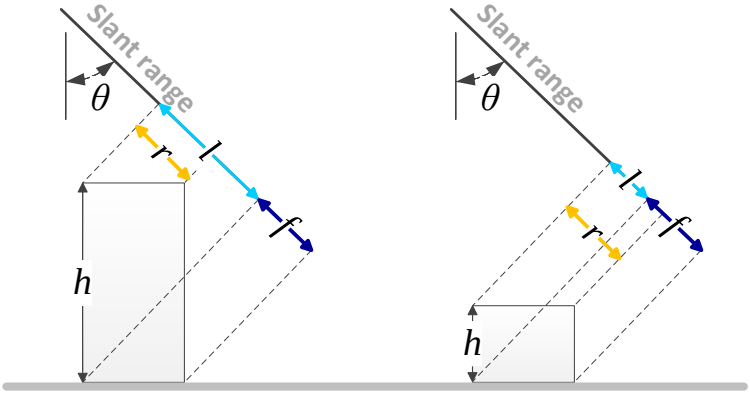


Figure 6.10: The projection geometry of two flat-roof buildings in a slant-range SAR image. θ is the incidence angle. h is the building height. l, r, and f denote the length of layover, roof, and footprint areas in a slant-range SAR image, respectively.

Afterwards, LoD1 building models are created by extruding building footprints with the obtained heights. Figure 6.13 presents example LoD1 models superimposed on the SAR image in the yellow rectangle area in Figure 6.12(b). It can be observed that buildings with large l (pointed by yellow arrows) are predicted as high-rise, while those with small l (pointed by red arrows) are reconstructed as low-rise buildings. This is in line with reality.

[1] http://en.wiki.quality.sig3d.org/index.php/Modeling_Guide_for_3D_Objects_Part_2:_Modeling_of_Buildings_(LoD1,_LoD2,_LoD3)

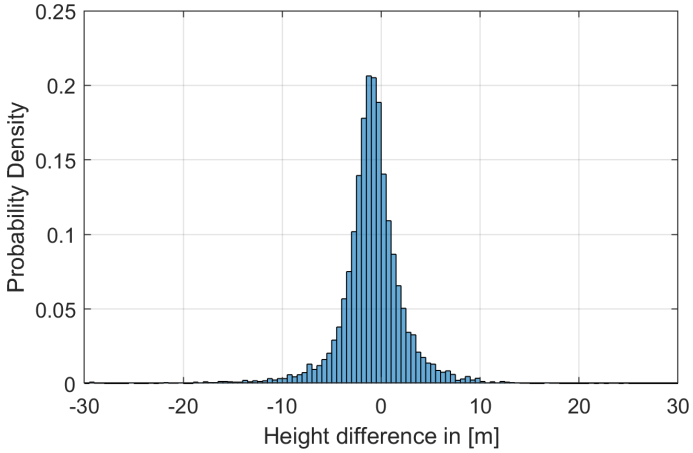


Figure 6.11: Histogram of building height errors in the study area.

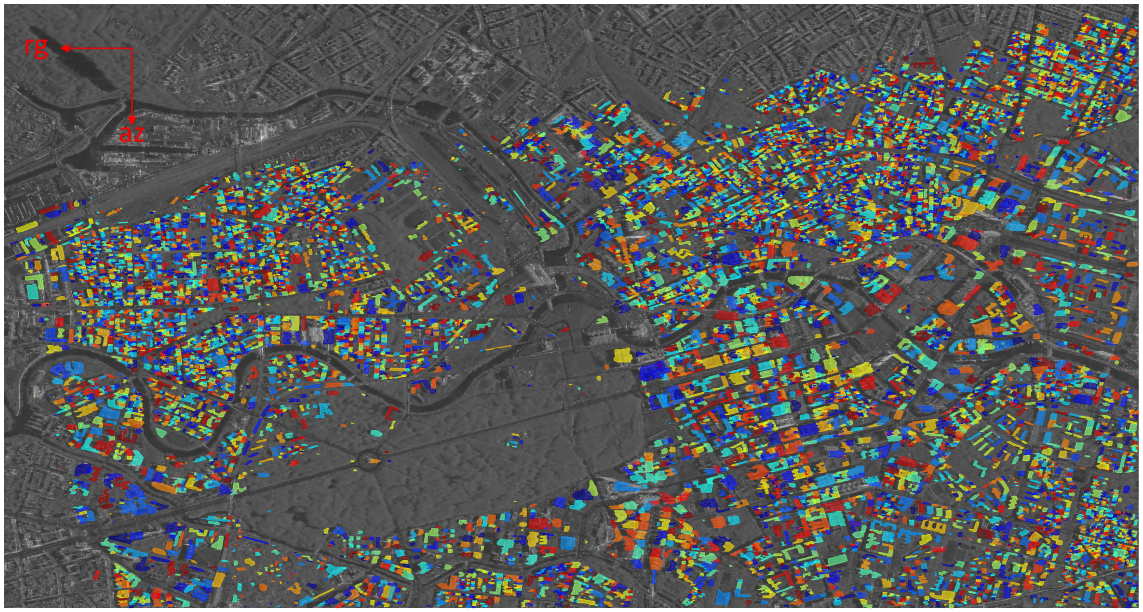
It is worth mentioning that the conversion from the layover areas to building heights is based on the simplified case of flat-roof buildings that ensures the geometric relationship between the estimated height hl and the mean height hm (cf. Figure 6.14 (a)). For buildings with different roof types or with tall surrounding objects, however, this geometric relationship may not hold. Although used for height estimation, this conversion is an approximation. Figure 6.14 (b) - (f) shows several examples. Two buildings with multiple flat roof surfaces in (b) and (e). As can be seen, $hl > hm$ in (b) and $hl < hm$ in (e). The same observation found for gable-roof buildings are shown in (c) and (f): $hl > hm$ in (c), and $hl < hm$ in (f). Figure 6.14 (d) shows a situation that a low-rise building that is partially occluded by an object in front of it. In this case, $hl < hm$. Figure 6.15 shows three examples of real buildings in 2-D and 3-D. Therefore, when applying this conversion from building segmentation results, one should consider the building shapes in the study region. In theory, the height estimation results are better in areas where most buildings have only one flat roof and are not severely occluded.

6.3.2 Can CG-Net Predict Individual Buildings from Stripmap SAR Images?

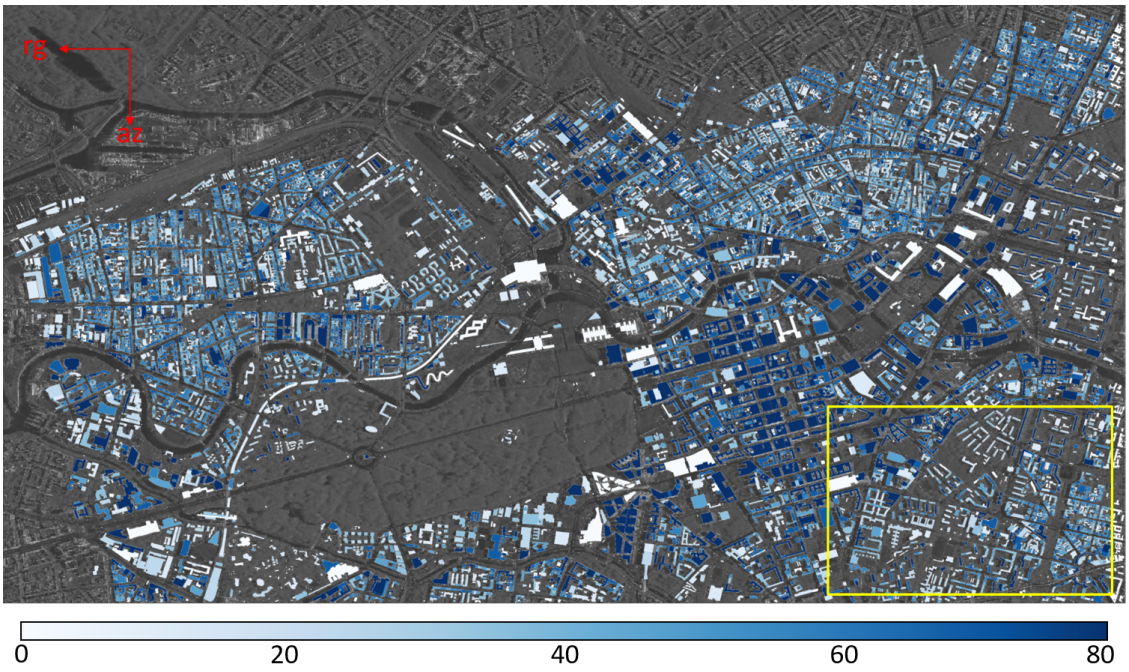
In the previous section, LoD1 building models are reconstructed on a large scale, using the segmentation results of CG-Net trained on a spotlight SAR image. Aiming at a larger scale, one has to ask the question: can CG-Net produce LoD1 models using stripmap images?

Due to the lack of TomoSAR point clouds generated from stripmap images, we are not able to generate new data sets for training the networks and evaluating the segmentation results. We can, however, deploy the trained CG-Net for inference on stripmap images. The segmentation results can then be visually examined, and the height results can be evaluated using the mean building height.

Two inference experiments are conducted, and DeepLabv3-CG is chosen because it performs best among all the networks. The CG-Net is trained using a TerraSAR-X spotlight image in Berlin. Thus, the first inference experiment is designed using a TerraSAR-X stripmap image in Berlin, i.e., the data is acquired in the same region but with a different imaging mode comparing to the training data. Building footprints are projected



(a) Segmentation results in the study area obtained by DeepLabv3-CG. The building segments are plotted with different colors translucently for visualising the layover areas between buildings.



(b) Estimated building heights in the study area obtained by DeepLabv3-CG. Height is color-coded.

Figure 6.12: (a) Segmentation results and (b) estimated building heights obtained by DeepLabv3-CG. The LoD1 models in the yellow rectangle are shown in Figure 6.13. rg and az denote the range direction and azimuth direction, respectively.

to the SAR image to generate footprint masks, and the TerraSAR-X stripmap image is cropped into patches of 128 × 128 pixels with a stride of 70 pixels. 4120 buildings are remaining, each of them has two patches: a SAR image patch and a footprint mask patch.

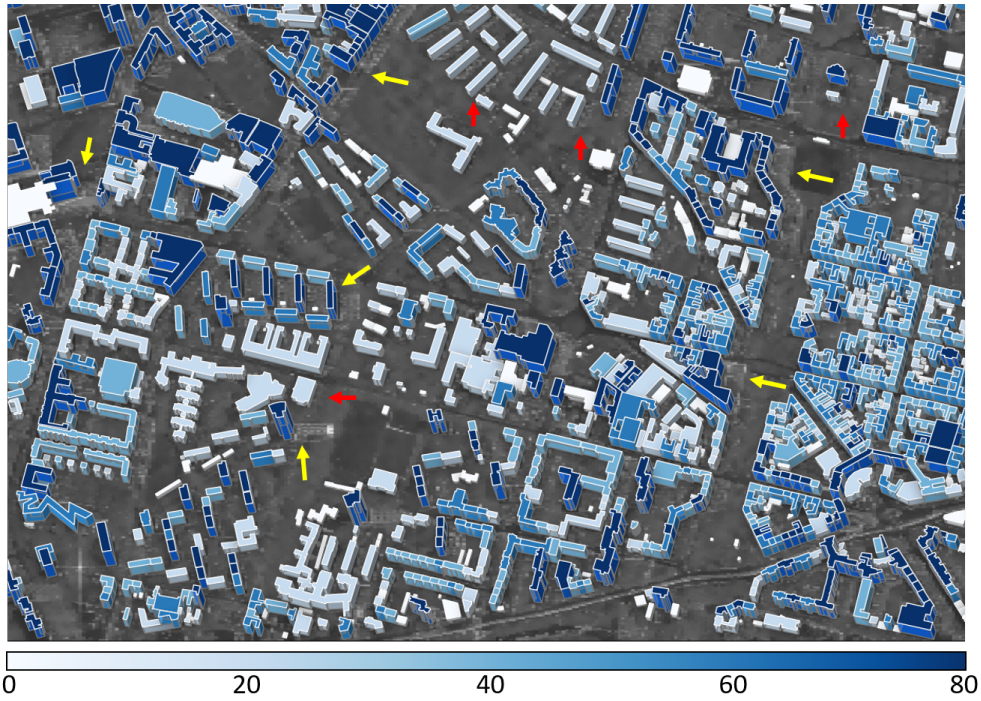


Figure 6.13: Reconstructed LoD1 building models in the yellow rectangle area in Figure 6.12(b) superimposed on the SAR image. Layover areas of some buildings are visible, as pointed by the yellow and red arrows. Building heights are color-coded.

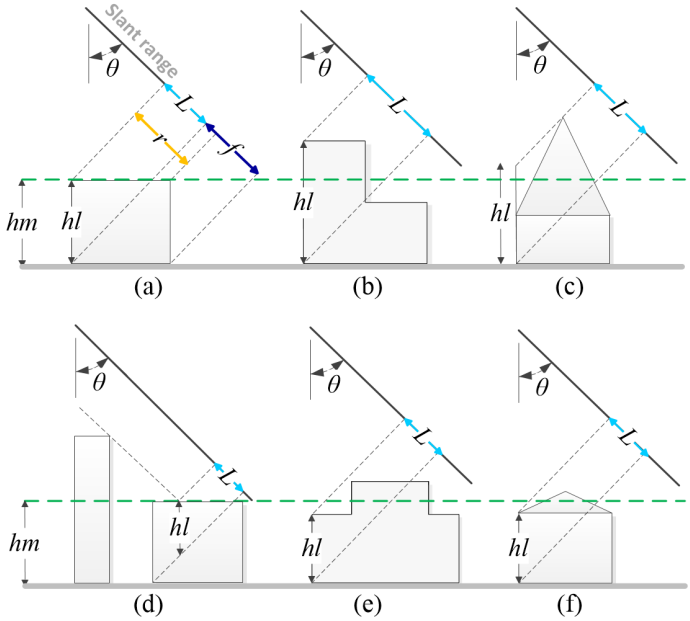


Figure 6.14: The SAR imaging geometry of buildings under different settings. hm and hl are the mean height and height transformed from the layover area. In (a), $hl = hm$; in (b) and (c), $hl > hm$; in (d)-(f), $hl < hm$. θ is the incidence angle, L, r, and f are the layover, roof, and footprint areas in the SAR image, respectively.

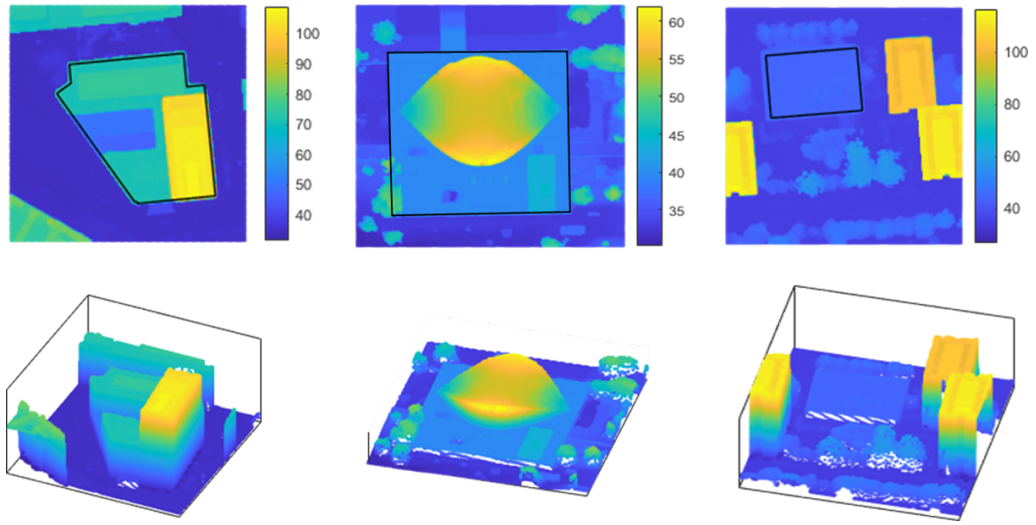


Figure 6.15: Examples of buildings with different roof types. The first row shows three buildings in a DEM in 2-D, in which the black polygons mark the building footprints. The second row shows the corresponding buildings in a completed 3-D point cloud. Height is color-coded.

The network takes one SAR patch and the corresponding GIS patch for one building as inputs for inference. Figure 6.16 shows the segmentation results of four buildings. The segmentation results of the same buildings from the spotlight image are also plotted for comparison. The inference result, building footprint, and their intersection are marked in dark gray, light gray, and white, respectively. As can been seen, the segmentation results from the stripmap image and the spotlight image seem quite similar. From the segmentation results, building heights are computed and evaluated against the mean height from the accurate DEM for each building. The mean absolute height error we achieve in the study site is 3.89 m. The histogram of height errors is shown in Figure 6.17.

The second inference experiment is performed in New York City, using a TerraSAR-X stripmap image, i.e., the region and the imaging mode are both different from the training data. Building footprints and heights are acquired from NYC open data [49]. Same as the previous experiment, building footprint masks are generated, and the SAR image is cropped to produce 3482 building samples for inference. Figure 6.18 shows the segmentation results of five buildings. In the prediction patches, the segmentation result, building footprints, and their intersection are marked in dark gray, light gray, and white, respectively. For comparison, in the SAR patches, building footprint polygons are plotted in blue, and the corresponding location of roofs are plotted in green, i.e., the building segments should be the areas between the roof polygon and the footprint polygon. As can be observed, the segmentation results on the first three buildings are far from correct, and those of the last two buildings are reasonable. Building heights are then computed and evaluated with the mean absolute height error of 14.24 m. The histogram of height errors is shown in Figure 6.19.

The two inference experiments on stripmap SAR images show quite different results in Berlin and New York City. Trained using a spotlight SAR image in Berlin, the CG-Net inferences well on the stripmap image of the same area, but the results are far from satisfactory when the area is changed, e.g., New York containing different urban forms

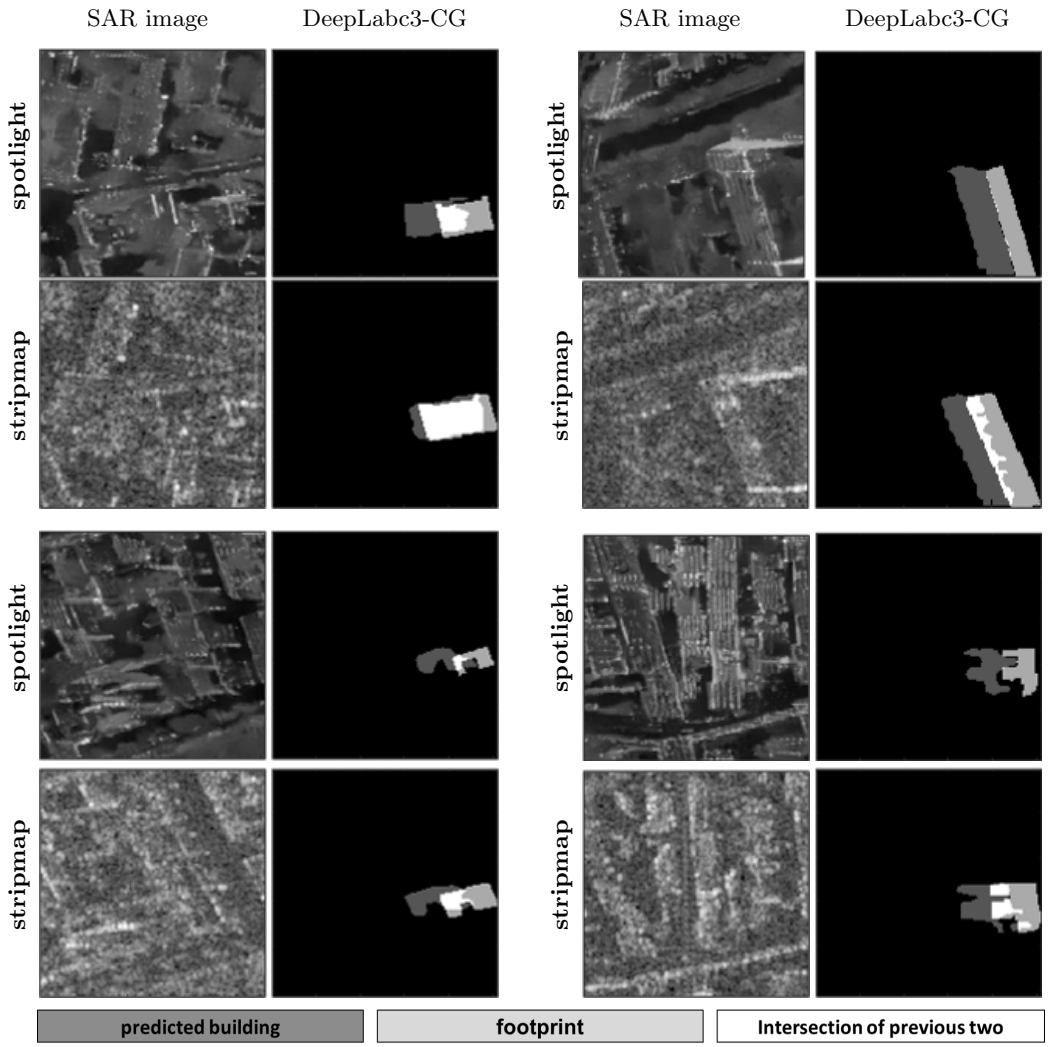


Figure 6.16: Examples of segmentation results of the same buildings from networks trained using spotlight and stripmap SAR images in Berlin. In the prediction patches, the segmentation result, building footprints, and their intersection are marked in dark gray, light gray, and white, respectively.

than Berlin. The stripmap image and the spotlight image in Berlin are both collected from the descending orbits with similar incidence angles. It can thus be suggested that our CG-Net is not sensitive to the image resolution.

In general, the inference performance of deep neural networks can be improved by increasing the scale and diversity of the training samples. Therefore, it needs more annotation data to improve the transferability of CG-Net.

6.4 Summary

This chapter presents a conditional GIS-aware network (CG-Net) to segment individual buildings from a large-scale VHR SAR image.

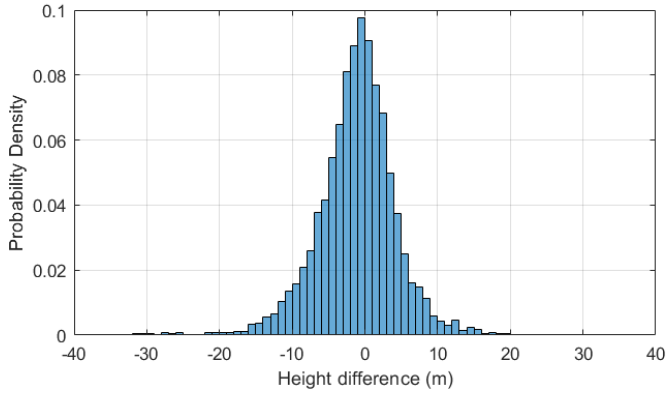


Figure 6.17: Histogram of height errors of 4120 buildings in Berlin from inference experiments on a TerraSAR-X stripmap image.

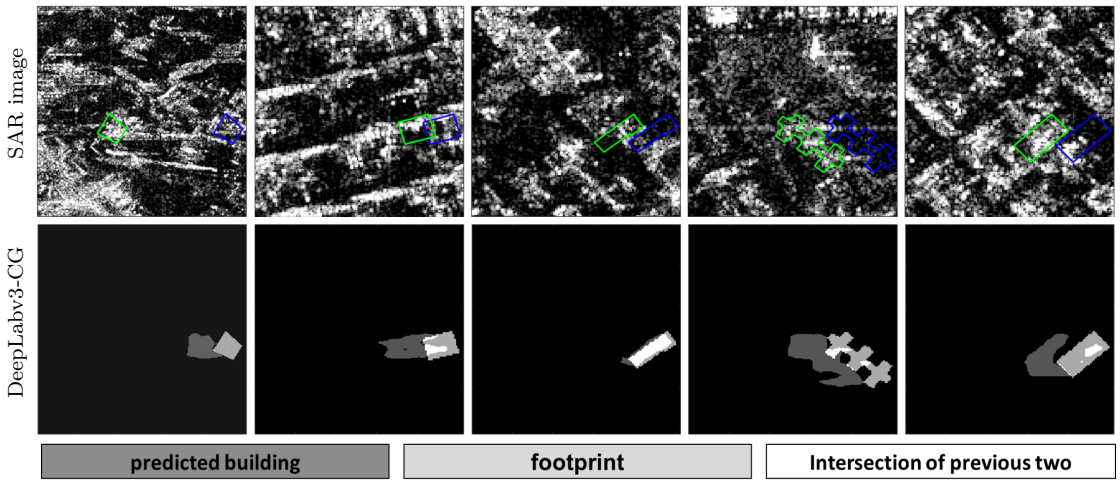


Figure 6.18: Examples of segmentation results in networks trained using stripmap SAR imagery in New York. In the SAR patches, building footprint polygons are plotted in blue, and the corresponding location of roofs are plotted in green. In the prediction patches, the segmentation result, building footprints, and their intersection are marked in dark gray, light gray, and white, respectively.

The proposed method is evaluated in the Berlin area, using a high-resolution spotlight TerraSAR-X image and building footprints obtained from GIS data. Both qualitative and quantitative results demonstrate the effectiveness of the proposed CG module. Compared to competitors, DeepLabv3-CG achieves the best F1 score of 75.08%. In addition, we compare two building footprint representations, namely complete building footprints and sensor-visible footprint segments. Experimental results suggest that the use of complete building footprints leads to better results. Further experiments of training the networks using inaccurate GIS data suggest that CG-Net is robust in the presence of positioning errors in GIS data.

As application, the predicted building segments are used to compute building heights for the LoD1 building model are reconstruction. The mean absolute height error we achieve in the study site is 2.39 m.

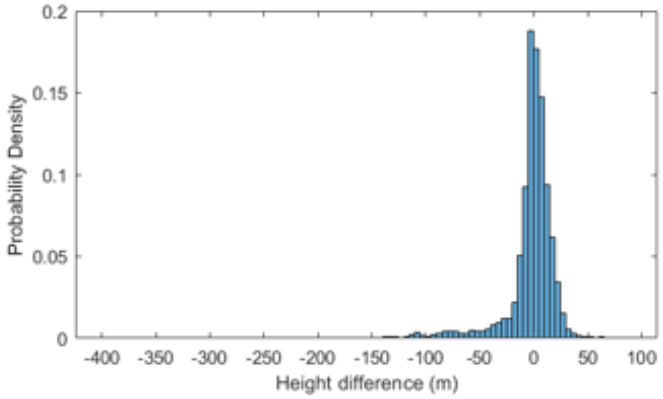


Figure 6.19: Histogram of height errors of 3482 buildings in New York City from an inference experiment on a TerraSAR-X stripmap image.

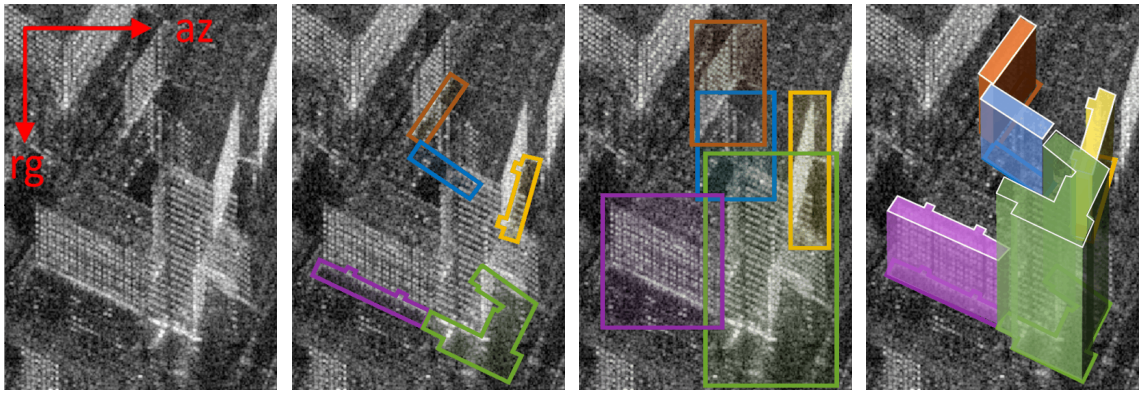
Two inference experiments are conducted using stripmap TerraSAR-X images in Berlin and New York. The evaluation of building heights show that the CG-Net inferences well on the stripmap image of the same area, but the results are far from satisfactory when the area is changed. For improving the transferability of CG-Net, more annotation data are needed to train the networks.

7 Building Height Retrieval from Single SAR Images with Bounding Box Regression

In Chapter 6, individual buildings are segmented on a large scale, and building heights are estimated subsequently. However, pixel-wise labels are still expensive. The unavailability of accurate DEMs in most areas limits the algorithm to be generalized to more regions. On the other hand, in addition to DEMs, the height value of buildings can be acquired from other data types, such as city models, or LiDAR data, and different data sources, including publicly available data sets for some cities.

To overcome the problem and improving transferability, this chapter proposes to employ building heights from multiple data sources and develops a network to learn building heights. The problem of building height retrieval is formulated as a bounding box regression, i.e., a task to regress the center coordinate and the size of the bounding box for each building. As demonstrated in Figure 7.1, this network takes SAR images and building footprints as input and retrieves building heights by predicting bounding boxes of buildings, and LoD1 building models are subsequently reconstructed.

The contents of this chapter are summarised in [238].



(a) SAR image **(b)** Footprints **(c)** Bounding boxes **(d)** LoD1 models

Figure 7.1: Illustration of the input and output of our method in a typical urban area[1]. (a) a SAR image and (b) building footprints in the SAR image are the input data. (c) shows the detected bounding boxes of these buildings. Building heights are then computed from the bounding boxes and building footprints, and build LoD1 models shown in (d). rg and az denote the slant range direction and azimuth direction, respectively.

[1]Interested readers are referred to Figure 2.6 for the optical image and building regions in the same ares.

7.1 Problem Formulation Based on the Radar Viewing Geometry

Consider a vertical line in the UTM coordinate system. Under the radar viewing geometry, all points on this vertical line have the same azimuth distance with respect to the SAR sensor. Consequently, projected to a SAR image, these points share the same azimuth coordinate, i.e., this vertical line parallels the slant-range direction of the SAR image. For buildings, this means that each vertical wall is projected to a SAR image into a parallelogram with one pair of opposite sides paralleling the slant-range direction, which be observed in the SAR image in Figure 7.1. The extent of a building in a SAR image, therefore, is bounded by two of the vertical lines from its walls in the azimuth direction and the region of the layover and footprint in the range direction.

Figure 7.2 illustrates this geometric relationship by two buildings in the UTM and the SAR image coordinate systems. On the left of the figure, b1 and b2 are two buildings in the UTM coordinate system, and their projections on a SAR image plane. As can be seen, the sensor-visible walls (yellow and blue) are projected into the SAR image as parallelogram shapes, and the vertical sides of the walls parallel the slant-range direction. The building height h is directly related to the layover length L:

$$h = L/cos\theta, \tag{7.1}$$

where θ is the incidence angle.

On the right of Figure 7.2, b1 and b2 and their bounding boxes (green) are shown in the SAR image coordinate system. As can be seen, for both b1 and b2, the layover length L is the width difference between the building bounding box and the footprint bounding box:

$$L = L_{building} - L_{footprint}. \tag{7.2}$$

Therefore, for a building in SAR images, its height can be obtained once its footprint is known and its bounding box is detected. Based on the geometry relationships, the problem of building height retrieval from SAR images can be formulated as a bounding box regression problem, i.e., given a SAR image and a building footprint, find the bounding box of the building, and then derive the building height from it.

7.2 Footprint Guided Bounding Box Regression Network

This chapter proposes a bounding box regression network for building height retrieval that utilizes the location information contained in building footprints. Figure 7.3 provides an overview of the proposed approach. Specifically, the network takes concatenated SAR images and building footprint masks as the inputs. ResNet-101 [270] is employed as the backbone. First, *conv1* to *conv4* in ResNet are utilized to extract feature maps. The footprint bounding boxes are extracted from the building footprint masks and enlarged and mapped on the feature maps and are regarded as the Region of interest (RoI) of each building, i.e., the initial bounding boxes to be corrected. For each RoI, local features are pooled by RoI-Align [222]. Then, *conv5* of ResNet takes the pooled features, and a global average pooling layer and a fully connected layer proceed to predict corrections for the RoI with respect to the ground truth bounding box. The corrections are then added to the RoI of each building to produce its bounding box. Finally, building heights are derived

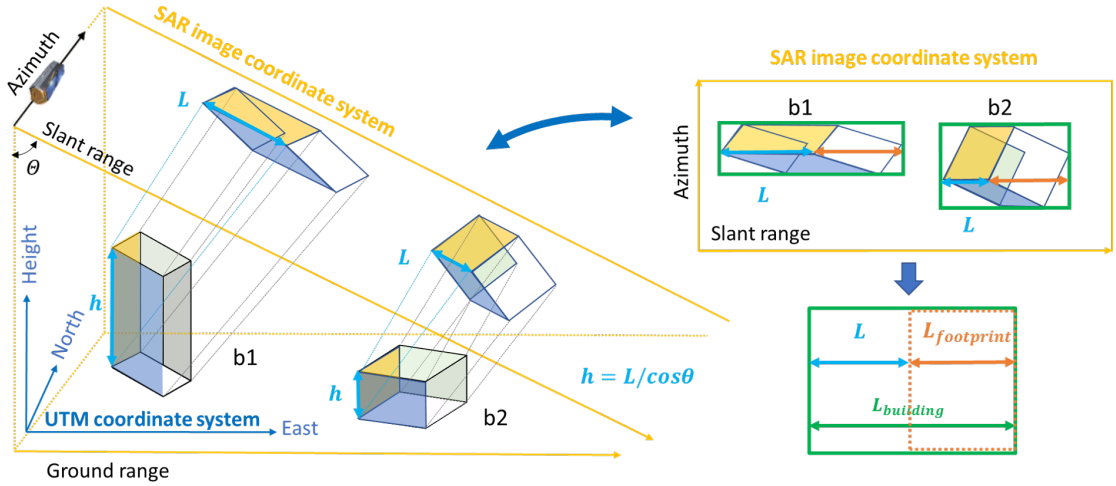


Figure 7.2: Illustration of bounding boxes of two buildings in a slant-range SAR image. On the left, two buildings, b1 and b2, in the UTM coordinate system are imaged in a SAR image plane. On the right, bounding boxes of b1 and b2 are shown in the SAR image coordinate system.

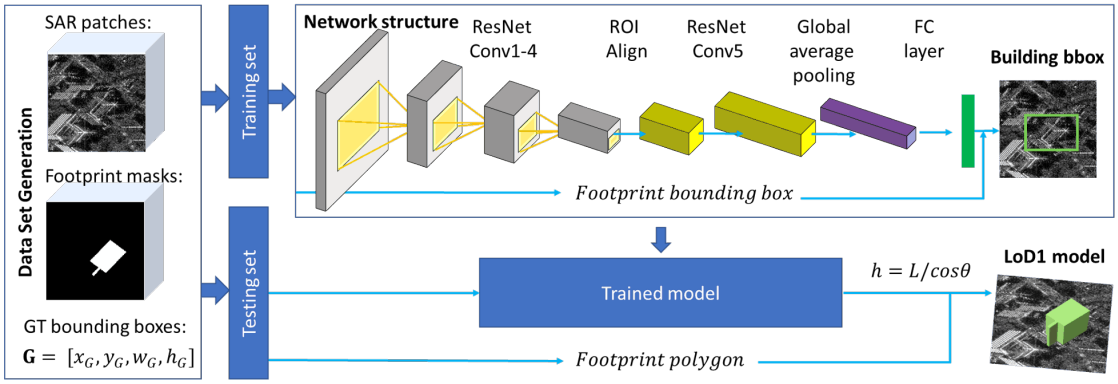


Figure 7.3: Overview of the proposed method. The input of our network is the concatenated SAR image patches and building footprint masks. The network employs ResNet-101 [270] as the backbone and predicts corrections for footprint bounding boxes with respect to the building bounding boxes. The corrections are then added to the RoI of each building to produce its bounding box and subsequently reconstruct LoD1 building models.

from the predicted bounding boxes and are used to extrude LoD1 building models from the building footprint polygons.

For the parameterizations of bounding boxes, the (x, y, w, h) coordinates used by R-CNN [271] are adopted. Let $\mathbf{B} = [x_B, y_B, w_B, h_B] \in \mathcal{R}^4$ be the bounding box representation as a 4-dimensional vector, where x, y, w, and h denote the box's center coordinates and its width and height in an image patch. The task of bounding box regression is to regress a candidate bounding box $\mathbf{B}$ into a target bounding box $\mathbf{G} = [x_G, y_G, w_G, h_G]$. In our case, $\mathbf{B}$ is the footprint bounding box, and $\mathbf{G}$ is the building bounding box. The

network predicts the distance vector $\mathbf{\Delta} = [\delta_x, \delta_y, \delta_w, \delta_h]$:

$$\begin{cases} \delta_x = (x_G - x_B)/w_B, \\ \delta_y = (y_G - y_B)/h_B, \\ \delta_w = log(w_G/w_B), \\ \delta_h = log(h_G/h_B). \end{cases} \tag{7.3}$$

The Complete Intersection over Union (CIoU) loss [235] is employed, which considers three geometric factors of the bounding boxes: the overlap area, the central point distance, and the aspect ratio. CIoU is defined as:

$$\mathcal{L}_{CIoU} = 1 - IoU - \frac{\rho^2(\mathbf{b}, \mathbf{g})}{c^2} + \alpha v, \tag{7.4}$$

where $\mathbf{b}$ and $\mathbf{g}$ denote the central points of $\mathbf{B}$ and $\mathbf{G}$, ρ is the Euclidean distance, c is the diagonal length of the smallest enclosing box covering the two boxes, α is a positive trade-off parameter, and v measures the consistency of the aspect ratio. They are defined as follows:

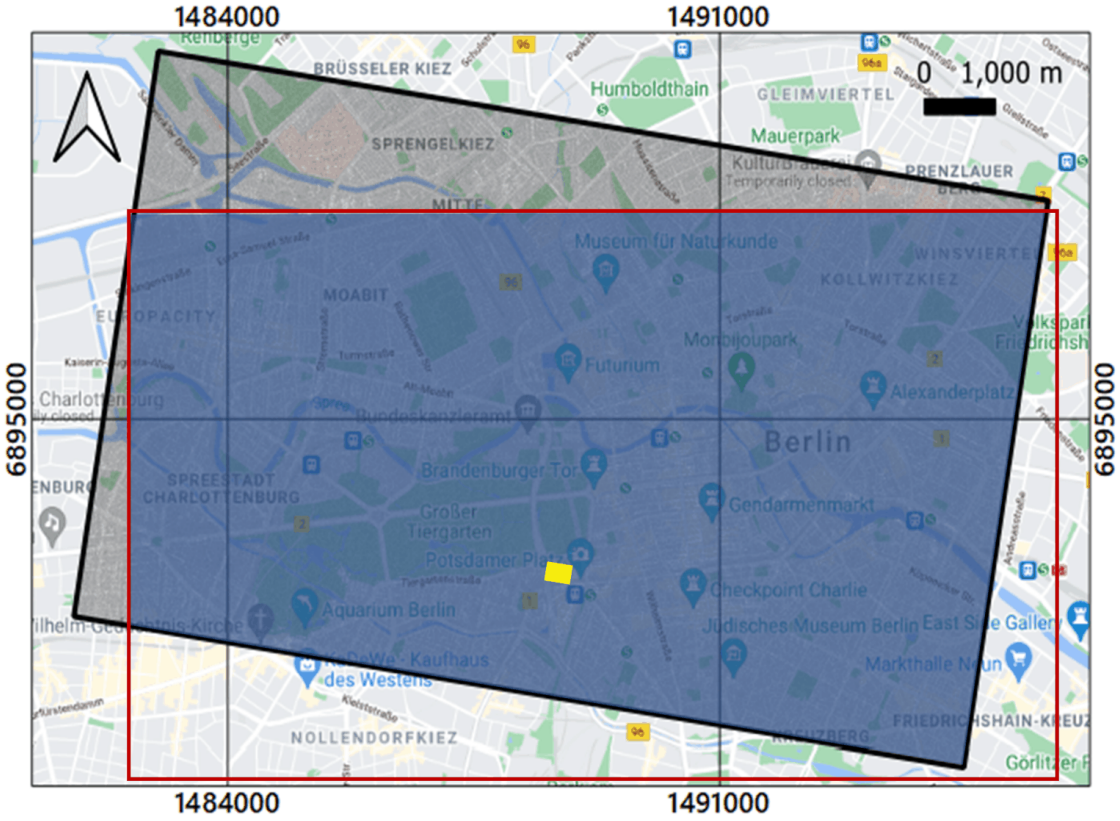
$$\begin{aligned} IoU &= \frac{|\mathbf{B} \cap \mathbf{G}|}{|\mathbf{B} \cup \mathbf{G}|}, \\ \alpha &= \frac{v}{(1 - IoU) + v}, \\ v &= \frac{4}{\pi^2}(arctan\frac{w^g}{h^g} - arctan\frac{w^b}{h^b})^2. \end{aligned} \tag{7.5}$$

7.3 Experimental Results

7.3.1 Data Sets

The performance of the proposed method is evaluated on four data sets, including one TerraSAR-X high-resolution spotlight (HS) image acquired over Berlin and three TerraSAR-X stripmap (SM) images acquired over Berlin, Rotterdam, and New York. The four data sets are termed Berlin HS, Berlin SM, Rotterdam, and New York. The study region in Berlin is the intersection of the SAR image area (black rectangle) and the DEM area (red rectangle), and the SAR images in Berlin HS and Berlin SM data sets are cropped to cover the same region, as illustrated in Figure 7.4 (a). Figure 7.4 (b) and (c) show the stripmap image and the spotlight image in the yellow rectangle in (a), respectively. The study regions in Rotterdam and New York are shown in Figure 7.5 and Figure 7.6, respectively.

Table 7.1 and Table 7.2 list the data sources of building footprints and heights, as well as the main characteristics of the used SAR images in each data set. In this work, height data are acquired from LoD1 building models and DEMs. LoD1 models represent buildings as blocks with flat roof structures and contain one height for each building [32]. As for DEMs, same as in Chapter 6, the average roof height is regarded as the building height. By using the workflow described in Section 4.3, building bounding boxes and footprint masks were generated. For each building, the data set contains a SAR image patch, a footprint mask, and a bounding box of the building.



(a) Study area (blue region) in Berlin: the intersection of the spotlight SAR image (black rectangle)) and the DEM area (red rectangle).

(b) Spotlight SAR image in the yellow rectangle in (a).

(c) Stripmap SAR image in the yellow rectangle in (a).

Figure 7.4: The study area of both Berlin HS and Berlin SM data sets. (a) shows the study area (blue) in the UTM coordinate system (UTM zone 32N). (b) and (c) show a comparison of the TerraSAR-X spotlight image and the stripmap image in the yellow rectangle in (a), respectively.

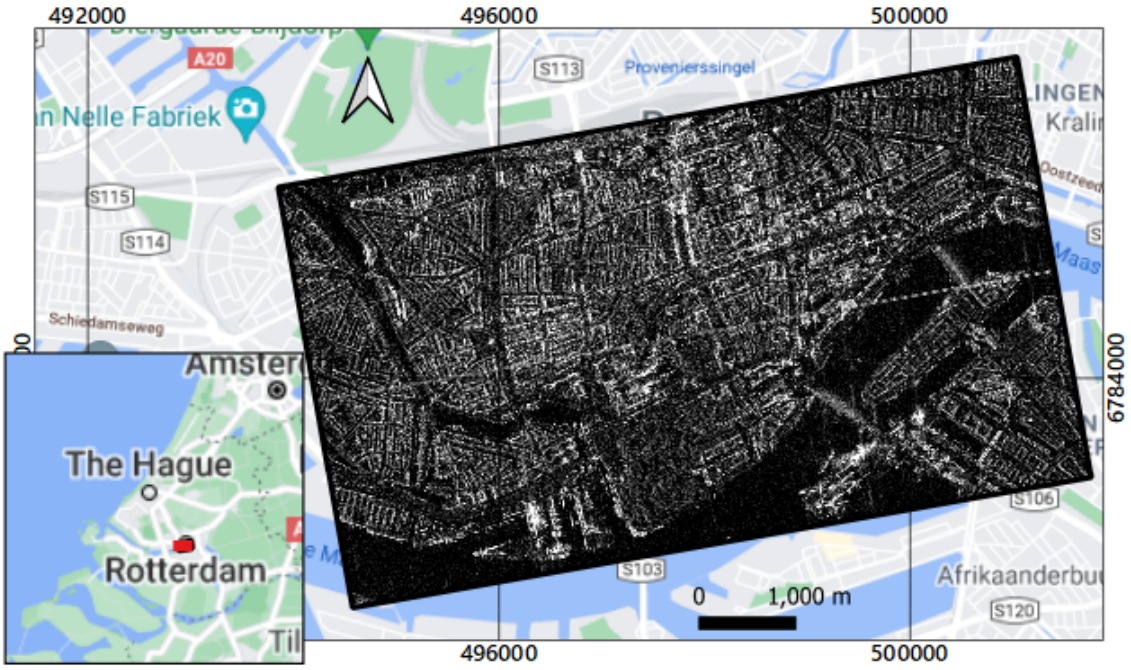


Figure 7.5: The Rotterdam study area in the UTM coordinate system (zone 31N).

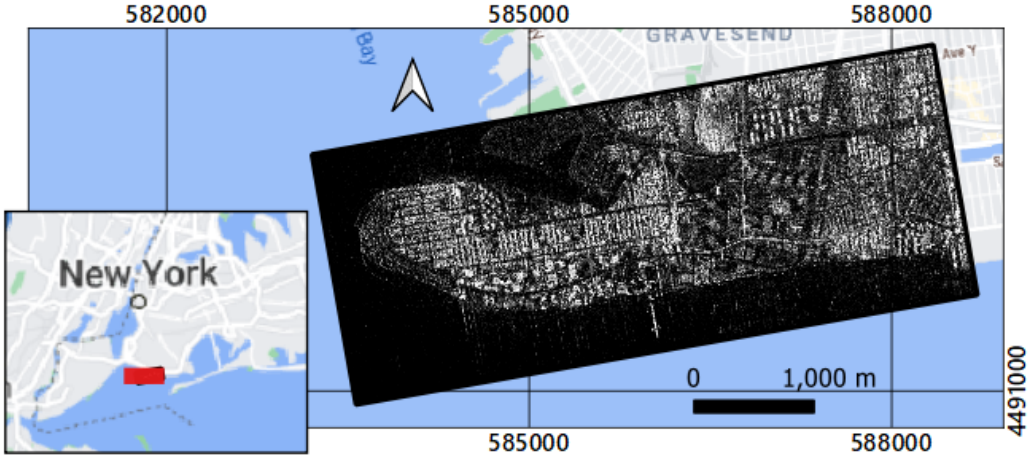


Figure 7.6: The New York study area in the UTM coordinate system (zone 18N).

Table 7.1: Main characteristics of the used SAR images in each data set.

	TerraSAR-X imaging mode	Pixel spacing: rg direction (m)	Pixel spacing: az direction (m)	Incidence angle (°)
Berlin HS	spotlight	0.45	0.871	36.08
Berlin SM	stripmap	0.909	1.836	46.68
Rotterdam	stripmap	1.364	1.852	39.28
New York	stripmap	1.364	2.203	42.65

Table 7.2: Data sources of building footprints and heights in each data set.

	Building footprints	Building heights
Berlin HS	Berlin 3D-Download Portal [50]	DEM (7cm/pixel)
Berlin SM	Berlin 3D-Download Portal [50]	DEM (7cm/pixel)
Rotterdam	3D BAG [245]	3D BAG [245]
New York	NYC open data [49]	NYC open data [49]

7.3.2 Training Details

To train an effective and robust network, first, the SAR images are cropped into patches. Patches containing incomplete footprints or bounding boxes are discarded. In the four datasets, the high-resolution spotlight SAR images are cropped into patches of 256×256 pixels with a stride of 150 pixels, and the stripmap SAR images are cropped into patches of 128×128 pixels with a stride of 70 pixels. Consequently, building data in the study areas are prepared, and each building has a ground truth bounding box and two patches: a SAR image patch and a footprint mask patch. All the building samples are then divided to build the training set and the testing set. The training and test regions do not overlap. Table 7.3 lists the patch size and sample numbers of training/testing sets of each data set. Before feeding data into models, all of the data sets used in the experiments were normalized into a range of [0, 1]. For the data sets generated using stripmap SAR images, the image patches are re-scaled to 256×256 pixels.

The network is implemented on Pytorch and trained on one NVIDIA Tesla P100 16GB GPU for 10 epochs. During the training procedure, the backbone module is initialized with weights pre-trained on ImageNet [268], and all new layers are randomly initialized by drawing weights from a zero-mean Gaussian distribution with a standard deviation of 0.01. All weights are updated through back-propagation, and stochastic gradient descent (SGD) [272] is selected as the optimizer. The learning rate is initialized as 0.001 and reduced by a factor of 0.1 once the loss stops to decrease for three epochs. A momentum of 0.9 and a weight decay of 0.0005 are used. In the experiments, a small batch size of 4 is utilized.

Table 7.3: Patch size and sample numbers in each data set.

	Patch size (pixel)	Cropping stride	Total samples	Training samples	Testing samples
Berlin HS	256×256	150	29842	19251	10591
Berlin SM	128×128	70	17183	15863	1321
Rotterdam	128×128	70	15054	13368	1686
New York	128×128	70	7922	7318	604

7.3.3 Comparative Experiments

For our problem, the major focus is to correctly locate the bounding boxes of buildings. As bounding box regression is also an important task for object detection, object detection networks can be employed for our problem by deriving building heights from the predicted bounding boxes.

In the experiments, five object detection models are utilized to estimate building heights and compare the results with the proposed network. The object detection networks include three one-stage networks, i.e., SSD [220], YOLOv3 [273], RetinaNet [274], and a two-stage network, i.e., Faster R-CNN [222]. Additionally, feature pyramid networks (FPN) [214] are combined with Faster R-CNN in its backbone, termed as Faster R-CNN w. FPN, for better detecting objects at different scales. The procedures of object detection and height estimation are denoted as SSD_h, YOLOv3_h, RetinaNet_h, Faster R-CNN_h, Faster R-CNN w. FPN_h.

For implementation, MMdetection [275] is employed for SSD_h, YOLOv3_h, RetinaNet_h, and Faster R-CNN w. FPN_h, and the implementation in [276] is utilized for Faster R-CNN_h. ResNet-101 is used as the backbone for RetinaNet_h, Faster R-CNN_h, and Faster R-CNN w. FPN_h. For all the networks, the input image patches are the concatenated SAR images and building footprint masks, and the input image patches are all re-scaled to 256×256 pixels. Other default parameters in each network are kept.

7.3.4 Quantitative Evaluation

The performance of networks is evaluated based on two criteria: height accuracy and training time. We record the training time that each model takes for training on each data set and calculate building heights from the predicted bounding boxes, as stated in Section 7.1. The accuracy of retrieved building heights is measured by the mean absolute (he_{mae}) and the standard deviation (he_{std}) of height errors of all buildings H_e:

$$\begin{aligned} he_{mae} &= mean(|H_e|), \\ he_{std} &= std(H_e). \end{aligned} \tag{7.6}$$

$H_e = \{h^i_{true} - h^i_{predict} | i = 1, ..., n\}$, where h^i_{true} and $h^i_{predict}$ are the ground truth height and the predicted height for each buildings, respectively, and n is the number of test samples.

Table 7.4 reports numerical results of different models on the four data sets, and Figure 6.11 shows the histograms of height errors predicted by our network. It can be observed that Faster R-CNN_h performs the best in all four data sets among all the networks, however only 0.1-0.2 m better than the results achieved by our network, which are trivial compared to the mean absolute height errors (in the range of 4.3 m to 5.6 m). The results show that our networks, RetinaNet_h, and Faster R-CNN_h, outperform SSD_h and YOLOv3_h in height accuracy. Interestingly, FPN did not bring improvement to Faster R-CNN_h. One reason could be that the difference in the scale of building footprints is not particularly large.

On Berlin HS data set, all networks achieve the best performance in terms of height accuracy comparing to other data sets, owing to the higher spatial resolution of the spotlight image than the stripmap images. However, we notice that the differences are not significant. For instance, using stripmap images, the mean absolute height error achieved

by Faster R-CNN$_h$ ranges from 4.7 m to 5.6 m, and the standard deviation from 7.1 m to 7.6 m, depending on the data set. While using spotlight data (Berlin HS), the mean absolute height error achieved is 4.3 m, and the standard deviation is 6.2 m.

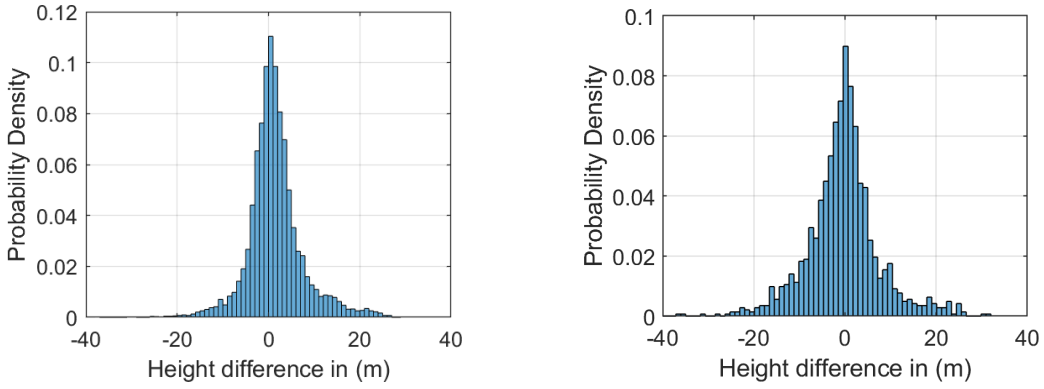
In terms of the speed, our methods significantly outperform not only the two-stage networks such as Faster R-CNN$_h$ but also the fast networks like SSD$_h$ and YOLOv3. Comparing to Faster R-CNN$_h$, the training time of our network reduces about 80%. The computation of the networks is reduced owing to the utilization of footprint bounding boxes.

Our network outperforms the detection-based networks mainly due to the tailored use of building footprints, i.e., the module is designed to extract the footprint bounding box as the initial bounding box specifically for our task. The detection networks, on the other hand, lack the module specified for extracting building footprint information. They rely on a large number of region proposals to obtain possible initial bounding boxes. In addition, our network provides one initial proposal, i.e., footprint bounding box, for each bounding box. However, the detection networks must provide multiple proposals in the earlier stage and rely on the classification scores to select the final bounding box in the later stage. Therefore, the computational cost of our network is much smaller.

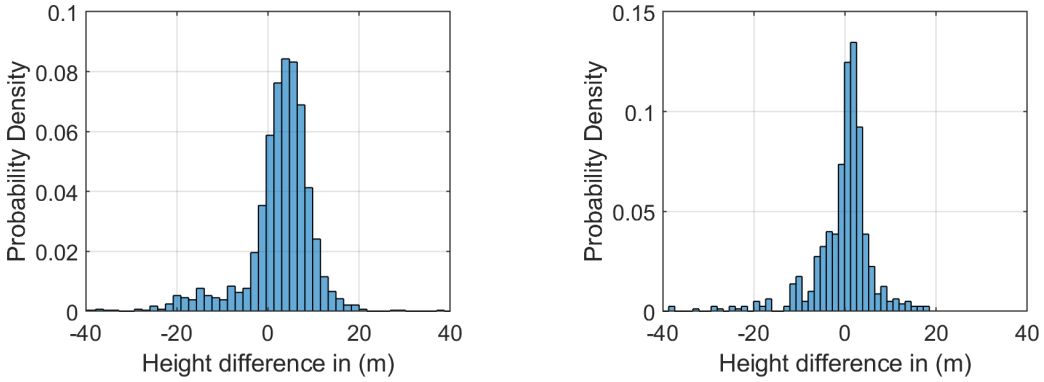
To sum up, our network achieves accuracy comparable with Faster R-CNN$_h$ and much superior performance on speed by involving multi-modal information involved in GIS data. Comparison of these results corroborates that the proposed network can significantly reduce the computation cost while keeping the height accuracy.

Table 7.4: Numerical results on four data sets. The highest values of different metrics are highlighted in **bold**.

Data set	Model name	he_{mae} (m)	he_{std} (m)	Training time
Berlin HS	SSD$_h$	6.6	9.4	3h26mins
	YOLOv3$_h$	6.0	8.1	4h16mins
	RetinaNet$_h$	4.7	6.5	5h22mins
	Faster R-CNN w.FPN$_h$	5.0	7.3	5h10mins
	Faster R-CNN$_h$	**4.3**	**6.2**	5h26mins
	Ours	**4.3**	6.3	**1h01mins**
Berlin SM	SSD$_h$	7.9	10.3	1h59mins
	YOLOv3$_h$	6.5	9.8	2h22mins
	RetinaNet$_h$	5.9	9.0	3h32mins
	Faster R-CNN w.FPN$_h$	6.1	8.7	3h25mins
	Faster R-CNN$_h$	**5.6**	**7.1**	3h28mins
	Our	5.7	7.2	**52mins**
Rotterdam	SSD$_h$	6.4	9.5	1h47mins
	YOLOv3$_h$	5.9	8.3	2h13mins
	RetinaNet$_h$	5.4	7.6	3h23mins
	Faster R-CNN w.FPN$_h$	5.8	7.8	3h14mins
	Faster R-CNN$_h$	**5.4**	**7.6**	3h40mins
	Our	5.5	**7.6**	**44mins**
New York	SSD$_h$	6.2	12.2	57mins
	YOLOv3$_h$	6.2	13.2	1h15mins
	RetinaNet$_h$	4.8	7.3	1h55mins
	Faster R-CNN w.FPN$_h$	5.0	7.8	1h30mins
	Faster R-CNN$_h$	**4.7**	**7.3**	1h59mins
	Our	4.9	7.6	**26mins**



(a) Berlin HS **(b)** Berlin SM



(c) Rotterdam **(d)** New York

Figure 7.7: Histogram of building height errors predicted with our network in the study areas.

7.3.5 Qualitative Evaluation

In addition to the quantitative evaluation, several segmentation results are visualized in Figure 7.8 and 7.9. In both the two figures, the first two rows show the building footprint masks and the SAR image patches, and Row 3 to 8 present the predicted bounding boxes from each model, in which the building footprint mask and the SAR image are both plotted. The ground truth boxes and the predicted boxes are presented in green and red, respectively.

Figure 7.8 presents results of models in Berlin HS and Berlin SM data sets. We can observe a general improvement in quality from one-stage detectors to two-stage detectors, especially for buildings in columns $b2$ and $b6$. For buildings with larger footprints and clear signatures in the SAR image (e.g., the building in column $b4$), all models can offer satisfactory results. In contrast, for those with a small footprint (see column $b1$) or ambiguous signatures (see column $b6$), one-stage models did not recognize full buildings. Besides, despite the resolution differences between the spotlight image and the stripmap image, the performance of all networks seems consistent.

Figure 7.9 visualizes results of models in Rotterdam and New York data sets. Similar results can be seen in columns $b7$ and $b12$ that all networks perform well when the building signatures clearly distinguish with the surroundings. On the contrary, the predictions for the building in column $b8$ are not satisfactory. The same can be observed on column $b11$

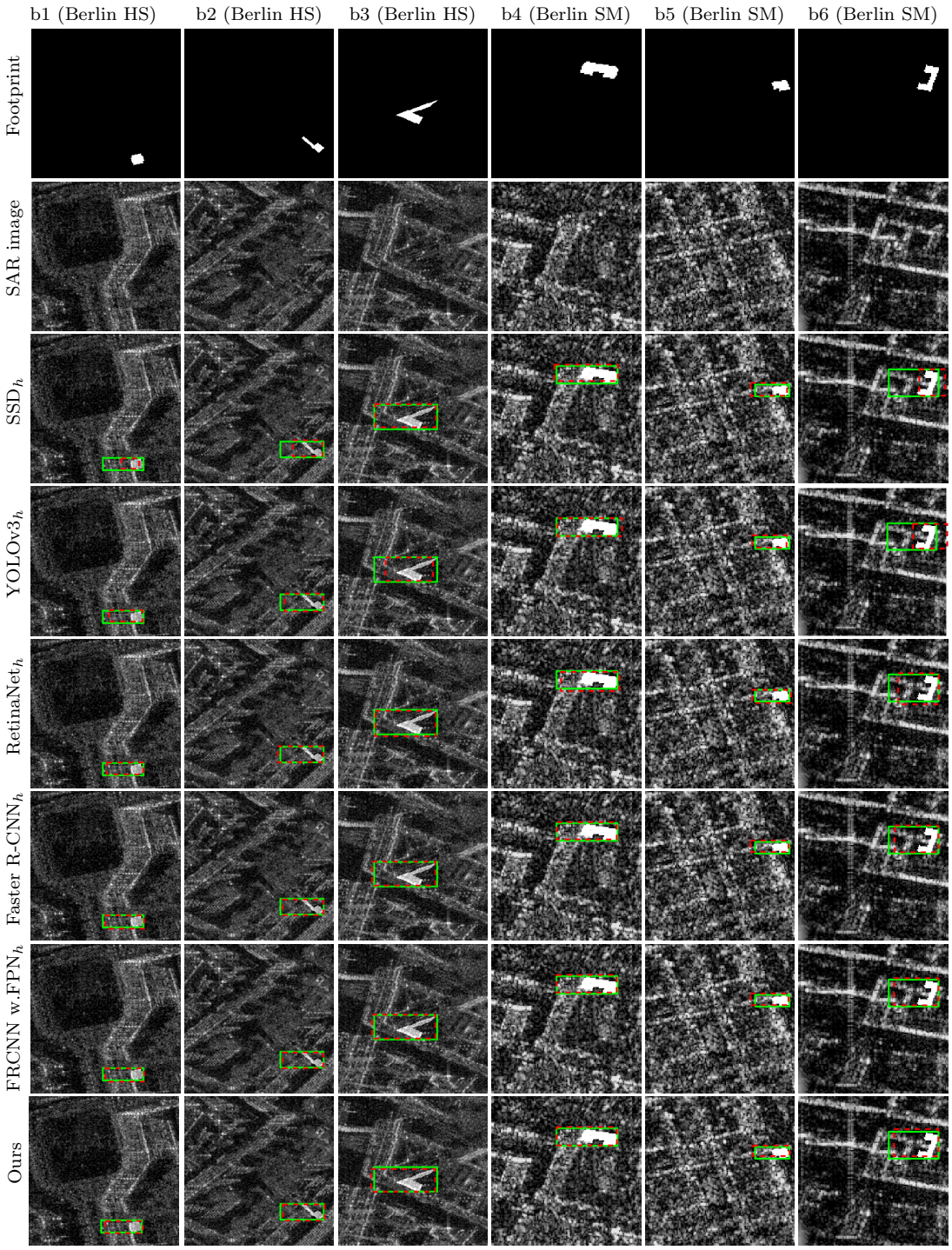


Figure 7.8: Examples of predicted bounding boxes using different networks in Berlin HS and Berlin SM datasets. The predicted and ground truth bounding boxes and are marked in red and green, respectively.

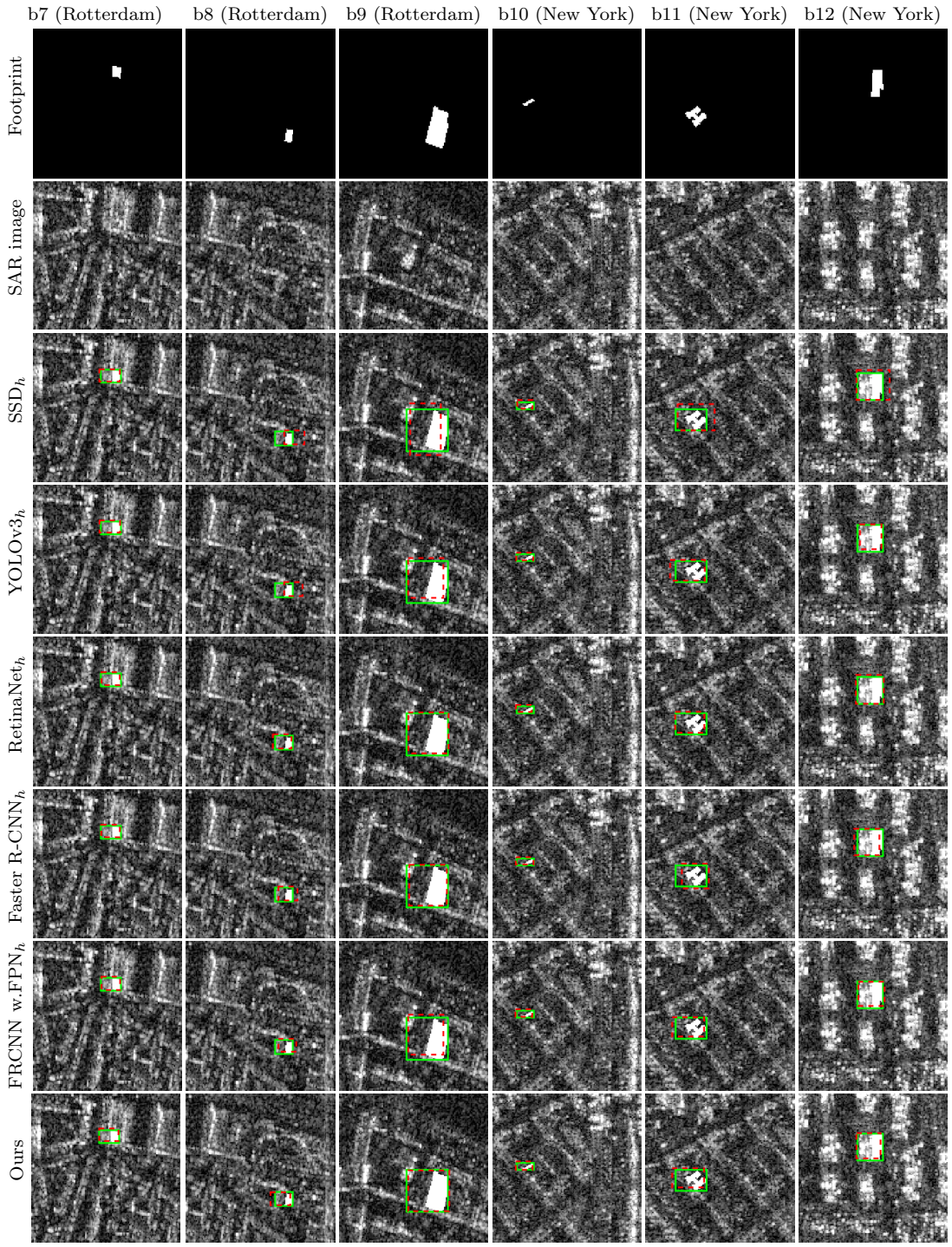


Figure 7.9: Examples of predicted bounding boxes using different networks in Rotterdam and New York datasets. The predicted and ground truth bounding boxes and are marked in red and green, respectively.

Table 7.5: Numerical results on trained on Berlin HS and Berlin HS-E data sets.

Data set	he_{mae} (m)	he_{std} (m)	Training time
Berlin HS	4.3	6.3	1h01mins
Berlin HS-E	4.6	6.8	1h03mins

in a building with a complex shape. Moreover, examples in columns $b9$ and $b10$ show two buildings are both well detected by all the networks despite the distinct differences in their footprints' sizes, which also indicates that the FPN does not enhance the precision of the bounding boxes. In summary, the proposed network has a similar performance with Faster R-CNN$_h$.

From the predicted bounding boxes of the individual building, building heights are retrieved and LoD1 building models are reconstructed in the four data sets, as shown in Figure 7.10, Figure 7.11, Figure 7.12, and Figure 7.13.

7.4 Discussion

7.4.1 Can the Proposed Network Work with Inaccurate GIS Data?

To examine the robustness of the proposed network against positioning errors in building footprints, similar experiments are conducted as Section 6.2.4. The proposed network is trained with inaccurate building footprints, and we discuss the impact of positioning errors in building footprints.

The Berlin HS data set with positioning errors in building footprints (termed as Berlin HS-E) are generated with the same process under the same magnitude and direction distribution of the positioning errors as explained in Section 6.2.4 and as illustrated in Figure 6.8. Note that this is the most difficult case that all footprints contain positioning errors. Then, we train our network on Berlin HS-E and test the trained network with a clean test set. The parameter settings of the network remain the same as previous experiments, as described in Section 7.3.2.

The results are listed in Table 7.5. As can be seen, comparing to results trained on Berlin HS, the mean absolute height error is increased by 0.3 m, and the standard deviation of the height error is increased by 0.5 m. However, it still gives competent height estimation results. For visual comparison, Figure 7.14 shows the network results trained with Berlin HS and Berlin HS-E. As can be seen, for building b and building c, the network trained with Berlin HS performs better, and the network trained with Berlin HS-E seems to predict better for building d. The predictions for buildings a and e are visually very similar. We observed that predictions from the network trained on Berlin HS-E are visually satisfactory for most buildings.

The experiments show that the proposed network is robust against the positioning errors in building footprint data. This finding suggests that a large amount of existing open-sourced GIS data, such as OSM, can be exploited for localizing bounding boxes of individual buildings in SAR images.

7.4.2 Influences of the Nonlocal Filtering Procedure on SAR Data

In this work, we have employed original SAR amplitude images in our experiments. However, previous studies in [137, 237] perform nonlocal filtering [29] on SAR images prior

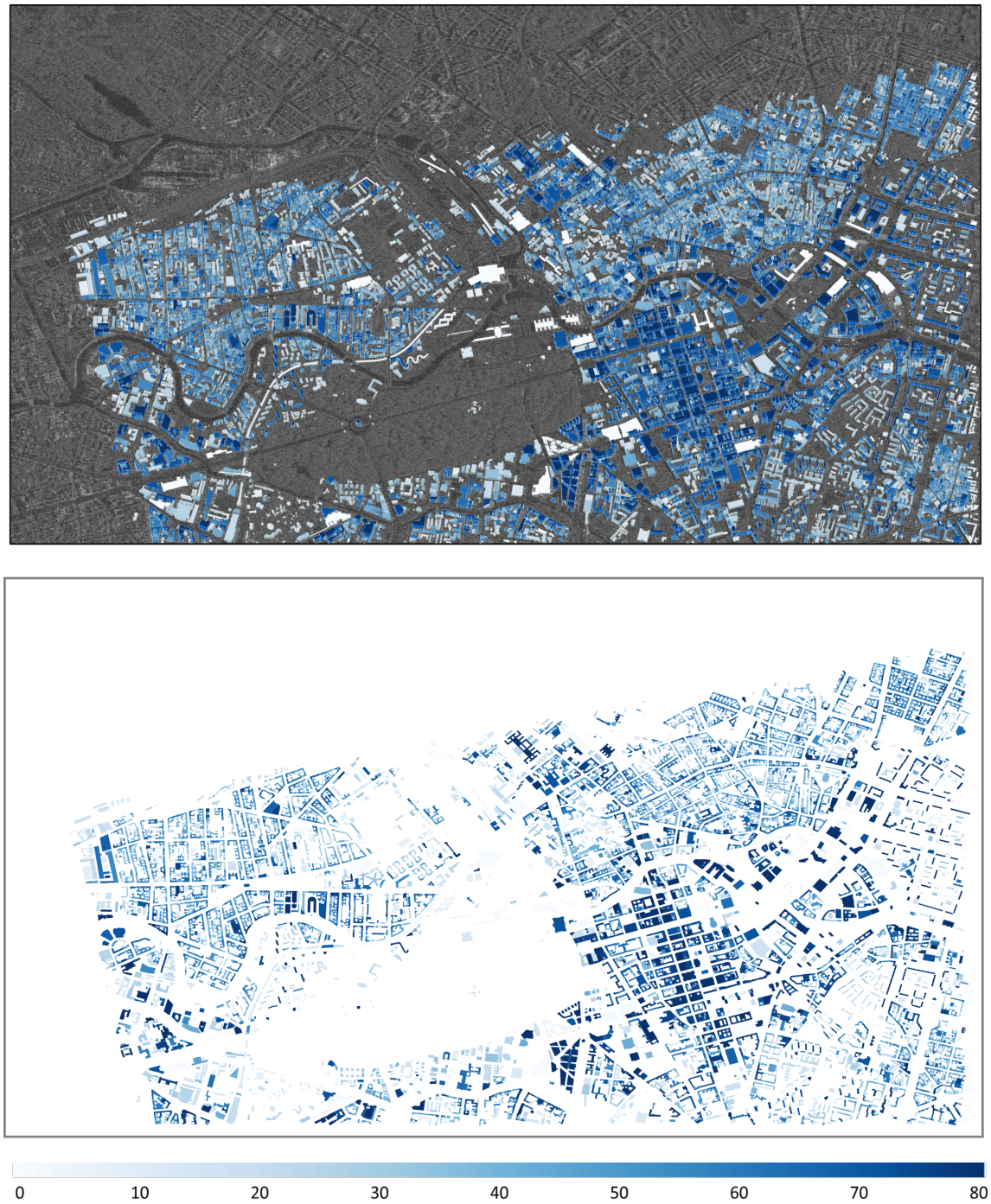


Figure 7.10: Height prediction map in Berlin HS dataset. (up) reconstructed LoD1 building overlaid on the SAR image. (down) Height prediction map in the SAR image coordinate system. Height is color-coded.

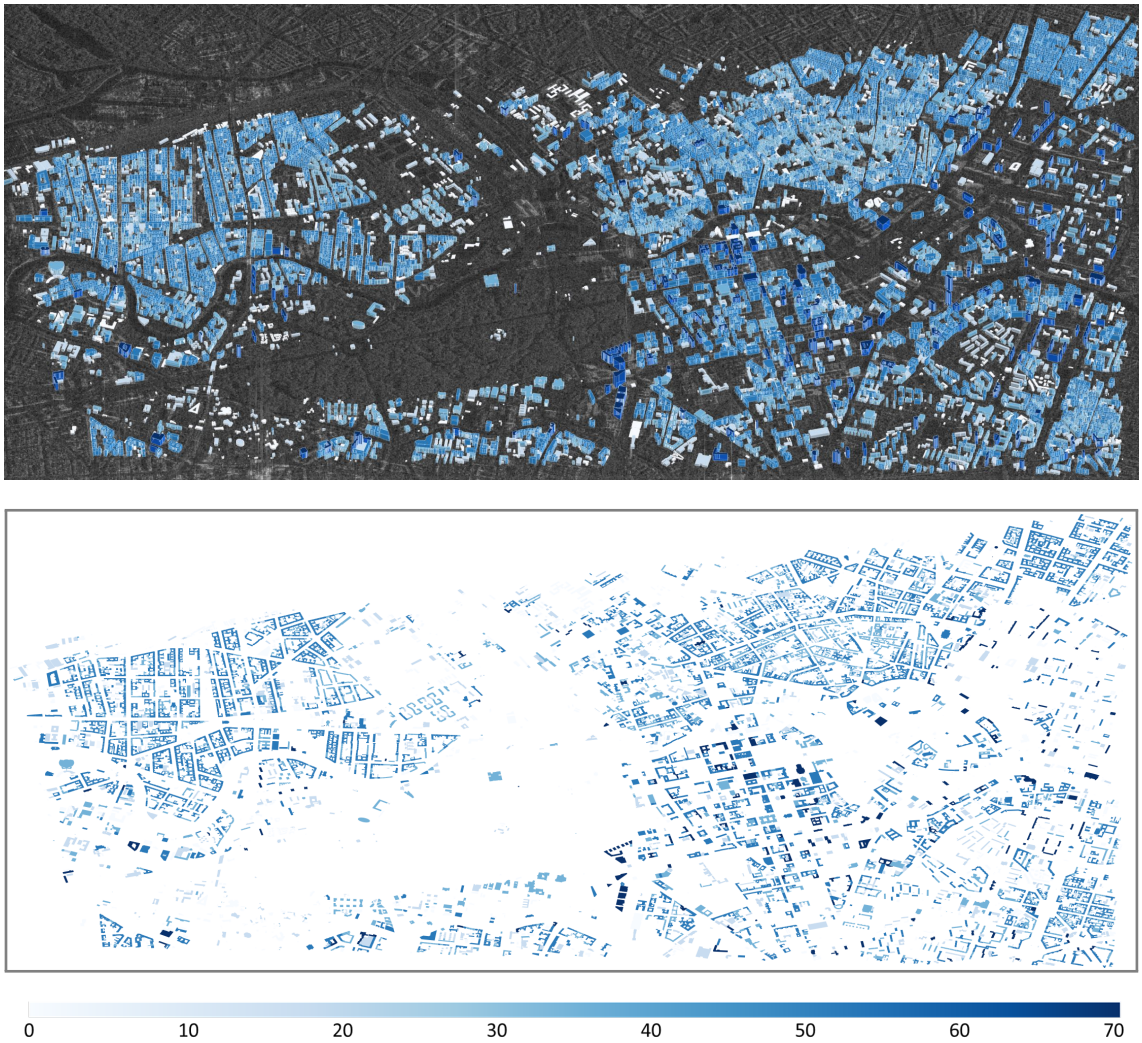


Figure 7.11: Height prediction map in Berlin SM dataset. (up) reconstructed LoD1 building overlaid on the SAR image. (down) Height prediction map in the SAR image coordinate system. Height is color-coded.

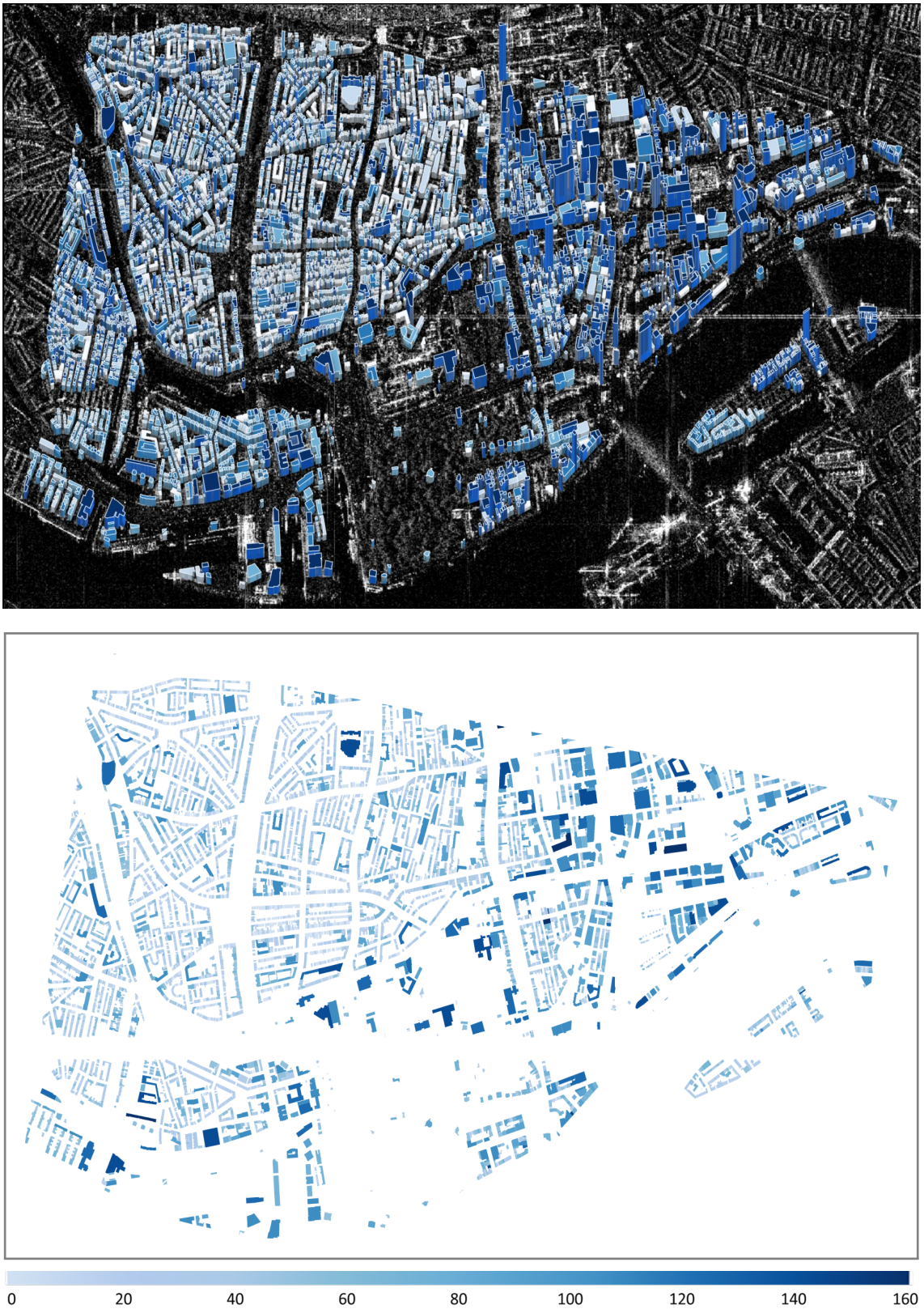


Figure 7.12: Height prediction map in Rotterdam dataset. (up) reconstructed LoD1 building overlaid on the SAR image. (down) Height prediction map in the SAR image coordinate system. Height is color-coded.

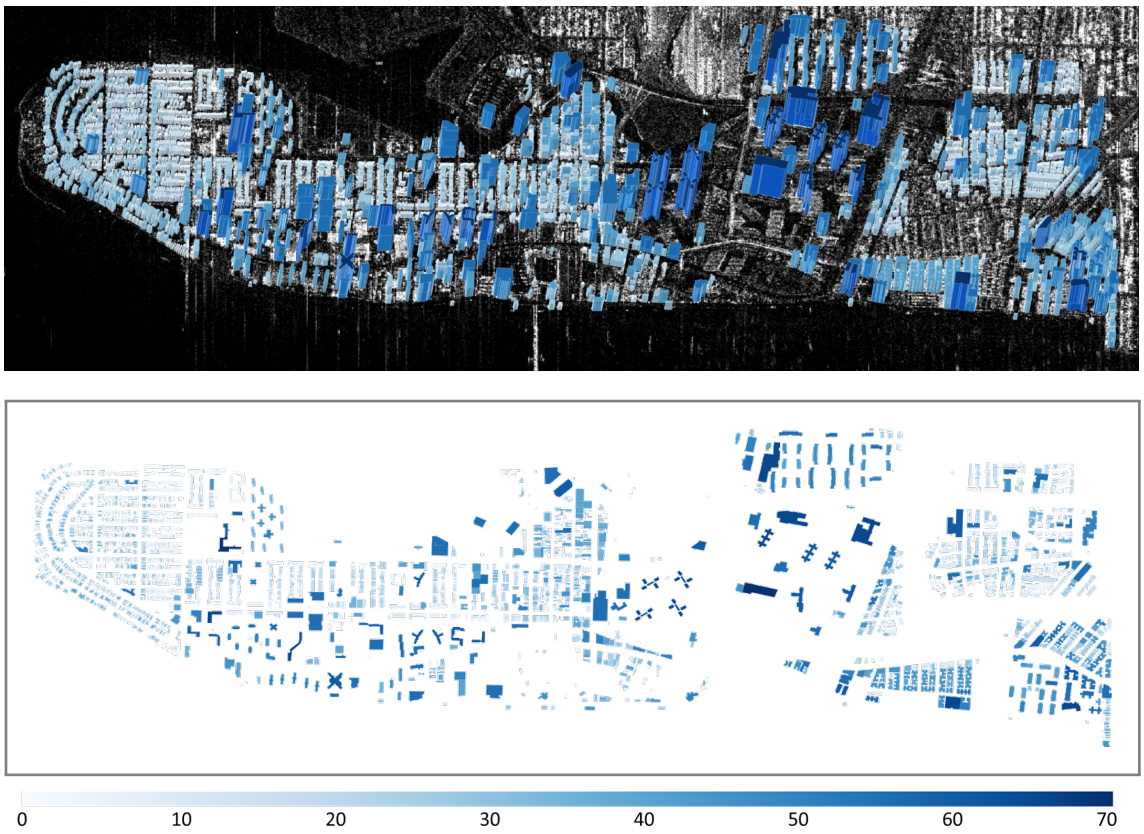


Figure 7.13: Height prediction map in New York dataset. (up) reconstructed LoD1 building overlaid on the SAR image. (down) Height prediction map in the SAR image coordinate system. Height is color-coded.

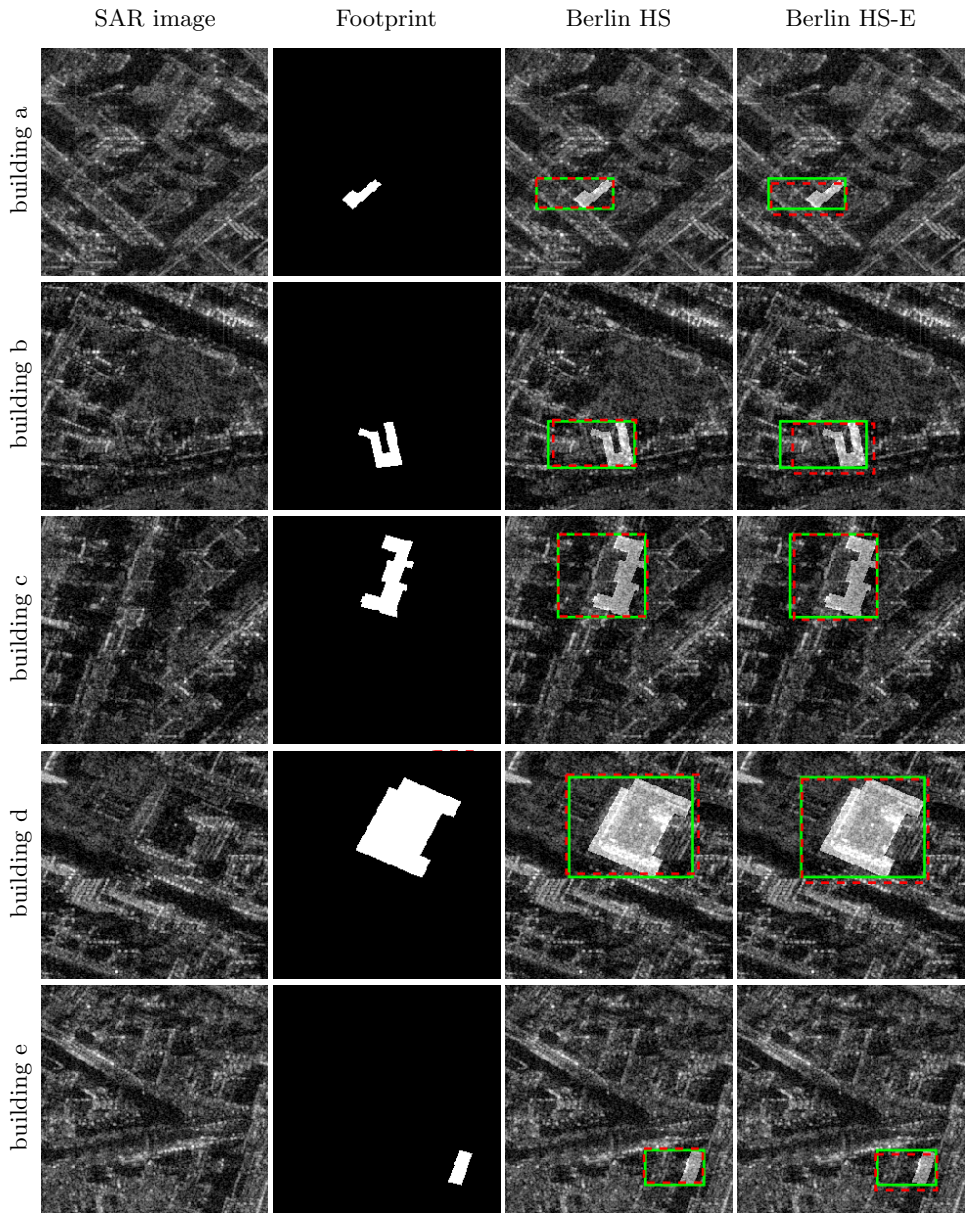


Figure 7.14: Examples of results of the proposed network trained using Berlin HS and Berlin HS-E (building footprints with positioning errors). The predicted and ground truth bounding boxes and are marked in red and green, respectively.

to training to reduce the speckle effect. To test the influence of the nonlocal filtering procedure for our networks, we conduct supplementary experiments to train the proposed network with nonlocal filtered SAR images.

We perform denoising on SAR images using a nonlocal InSAR algorithm [29]. Berlin HS data set is chosen for this experiment, and the nonlocal filtered data set is termed Berlin HS-NL. Then, we train and test our network and all the comparative networks on Berlin HS-NL dataset. The parameter settings of the networks remain the same as previous experiments, as described in Section 7.3.2 and Section 7.3.3.

Table 7.6: Numerical results on Berlin HS and Berlin HS-NL data sets. The highest values of different metrics are highlighted in **bold**.

Data set	Model name	he_{mae} (m)	he_{std} (m)	Training time
Berlin HS	SSD_h	6.6	9.4	3h26mins
	$YOLOv3_h$	6.0	8.1	4h16mins
	$RetinaNet_h$	4.7	6.5	5h22mins
	Faster R-CNN w.FPN_h	5.0	7.3	5h10mins
	Faster R-CNN_h	**4.3**	**6.2**	5h26mins
	Ours	**4.3**	6.3	**1h01mins**
Berlin HS-NL	SSD_h	6.7	9.4	3h28mins
	$YOLOv3_h$	5.9	8.1	4h15mins
	$RetinaNet_h$	4.7	6.6	5h29mins
	Faster R-CNN w.FPN_h	5.1	7.3	5h15mins
	Faster R-CNN_h	**4.3**	**6.4**	5h28mins
	Ours	**4.3**	6.5	**1h04mins**

Table 7.6 lists the results. As can be seen, results from Berlin HS and Berlin HS-NL data sets are very similar on all networks. The experiments show that the filtering procedure does not improve the results. We think the reason might be that the large amount of filters in CNNs, in fact, have filtering effects on speckle noises.

This finding suggests that the filtering step is not needed for our task. Therefore, the computational cost for pre-processing can be largely reduced, which benefits especially for larger-scale processing.

7.4.3 Pros and Cons of the Segmentation Networks and Regression Networks for Building Height Retrieval

So far, both the CG-Net presented in Chapter 6 and the regression network is able to reconstruct LoD1 building models. On Berlin HS-NL data set, the mean absolute height error achieved by the regression network is 4.3 m. In CG-Net, the building height achieved using a segmentation network from the same SAR data is 2.39 m. The advantage of CG-Net in terms of height accuracy is obvious. However, as aforementioned, pixel-wise labels are expensive and it is not possible to generate training data for areas without accurate DEMs. Thus the applicability of CG-Net is restricted.

The regression network has two advantages. First, since the building height retrieval problem is formulated as a bounding box regression problem, the proposed method is capable of employing building height data from multiple sources. This enables the generation of annotation data on a larger scale and improves the transferability of the proposed networks. Second, the data set generation approach of bounding boxes is much simpler than building areas, as explained in Section 4.2 and Section 4.3. This is crucial when processing large data sets, e.g., on a regional or larger scale. The disadvantage is that the building heights it predicted have lower accuracy, compared to the results of CG-Net.

In summary, these comparisons suggest that the proposed bounding box regression network has great potential for applications aiming at large scales, e.g., to reconstruct baseline models on a regional or even global scale. When accurate DEMs are available, segmentation networks such as CG-Net are preferred for the higher accuracy on the reconstructed building heights.

7.5 Summary

This chapter presents a method for retrieving building height on a large scale. The problem of building height retrieval is formulated as a bounding box regression problem, i.e., a task to regress the center coordinate and the size of the bounding box for each building. Requiring building footprints and only one height value for generating reference data, the proposed method is able to integrate building height data from multiple sources, such as open building models, LiDAR, and DEMs, for the generation of annotation data on a larger scale.

Four study sites are used to test the proposed networks, including one high-resolution spotlight TerraSAR-X image in Berlin and three stripmap TerraSAR-X images in Berlin, Rotterdam, and south Brooklyn in New York City. The mean absolute height error achieved in the four sites ranges from 4.3 m to 5.7 m. Achieved from a single stripmap TerraSAR-X image, the results are significant. Compared to methods utilizing object detection networks for building height retrieval, the proposed network can significantly reduce computation cost while keeping the height accuracy of individual buildings compared to Faster R-CNN$_h$.

Further experiments of training the networks using inaccurate building footprint data suggest that the proposed network is robust in the presence of positioning errors in building footprints. The proposed algorithm has the potential to be applied on a regional and even global scale.

The comparison of building height retrieved from the proposed network and the CG-Net presented in Chapter 6 suggests that the bounding box regression network has great potential for larger-scale processing; however, CG-Net is preferred if an accurate DEM is available.

8 Conclusion and Outlook

8.1 Conclusion

This Ph.D. work aims to reconstruct LoD1 building models from single SAR imagery on a large scale. Considering the characteristics of buildings in SAR images, building footprints are introduced as complementary data, and deep neural networks are employed for large-scale reconstruction. Towards the goal, five sub-objectives are defined, as summarised in Table 8.1:

Table 8.1: Summary of the objectives defined in Section 1.2.

Objective	
1.	Registering building footprints to SAR images on a large scale;
2.	Generating annotation data sets for applying deep neural networks;
3.	Developing deep learning algorithms for individual building analysis;
4.	Investigating building height reconstruction in multiple regions;
5.	Investigating the impact of positioning errors in building footprint data on the proposed algorithms, if unavoidable.

During the algorithm development in this thesis, the objectives are fulfilled progressively in three stages and resulted in three papers:

1. *Automatic registration of a single SAR image and building footprints in a large urban area*

   - This work addresses Objective 1 and is summarised in [248].

   A framework is proposed to automatically register building footprints to a SAR image by exploiting the features representing the intersection of ground and visible walls in both data, i.e., the near-range boundaries of building footprint polygons and the double bounce lines in the SAR image. Based on those features, the two data sets are progressively registered in three stages, allowing the algorithm to cope with variations in the local terrain.

   The proposed framework is tested in Berlin using a high-resolution spotlight TerraSAR-X image and GIS building footprints. Comparing to the ground truth, the proposed algorithm reduced the average distance error to -0.08 m and the standard deviation to 1.12 m. Such accuracy, better than half of the typical urban floor height (3m), is significant for building height reconstruction on a large scale.

   Further experiments using a TerraSAR-X stripmap image in Munich also show promising registration results.

2. *Large-scale individual building segmentation from a very high-resolution SAR image*

   - This work addresses Objectives 2, 3 and 5, and is summarised in [237].

First, a dataset generation approach is proposed that utilizes an accurate DEM to annotate individual buildings in a SAR image automatically.

For individual building segmentation, a novel conditional GIS-aware segmentation network (CG-Net) is proposed that learns multi-level visual features and employs building footprints to normalize the features for predicting building areas in the SAR image. Then, the segmentation results are applied to retrieve building heights.

Comparative experiments are designed to train the networks on accurate and inaccurate GIS data for investigating the impact of positioning errors in building footprints on the proposed networks.

The proposed network is validated using a high-resolution spotlight TerraSAR-X image collected over Berlin and building footprints obtained from GIS data. Several networks are tested, and the best model achieves the F1 score of 75.08% for segmentation. The segmentation results are employed for LoD1 building model reconstruction and achieved the mean absolute height error of 2.39 m in the study site. This work also investigates two building footprint representations, namely complete building footprints and sensor-visible footprint segments. Experimental results suggest that the use of complete building footprints leads to better results. Further experiments of training the networks using inaccurate GIS data suggest that CG-Net is robust in the presence of positioning errors in GIS data.

Two inference experiments are conducted using TerraSAR-X stripmap images in Berlin and New York. The evaluation of building heights shows that the CG-Net inferences well on the stripmap image of the same area, but the results are far from satisfactory when the area is changed. For improving the transferability of CG-Net, more annotation data are needed to train the networks.

3. *Building height retrieval from a very high-resolution SAR image based on bounding box regression networks*

   - This work addresses Objective 4 and enriches Objective 2, 3, 5. This work is summarised in [238].

   The previous work needs accurate DEMs for generating pixel-wise labels. The unavailability of accurate DEMs in most areas limits the algorithm from being generalized to more regions. To overcome the problem and improve transferability, the third work is proposed that employs multiple height sources and develops a network to learn building heights. Specifically, the problem of building height estimation is formulated as a bounding box regression problem, i.e., a task to regress the center coordinate and the size of the bounding box for each building.

   This work proposes to generate data incorporating height data from multiple data types, such as city models, or LiDAR data, and different data sources, including publicly available data sets for some cities, which allows the network to be applied to larger areas. Correspondingly, a bounding box regression network is proposed that employs the generated bounding building boxes using building footprints and building heights.

   Same as in the previous work, inaccurate building footprints are generated for testing the robustness of the proposed networks against positioning errors in building footprints.

Four study sites are used to test the proposed networks, including one high-resolution spotlight TerraSAR-X image in Berlin and three stripmap TerraSAR-X images in Berlin, Rotterdam, and south Brooklyn in New York City. The mean absolute height error achieved in the four sites ranges from 4.3 m to 5.7 m. Using a single stripmap TerraSAR-X image, the results are significant. Further experiments of training the networks using inaccurate building footprint data suggest that the proposed network is robust in the presence of positioning errors in building footprint.

The proposed algorithm has the potential to be applied on a regional and even global scale. The comparison of building height retrieved from the CG-Net and the regression network suggest that the regression network has great potential for larger-scale processing; however, CG-Net is preferred if an accurate DEM is available.

To the author's best knowledge, this is the first study investigating individual buildings in single SAR images on a large scale and the first study applying deep learning for individual building analysis using SAR images. The algorithms developed in the thesis can be used to improve the spatial coverage of baseline geospatial data, including building heights.

8.2 Outlook

The proposed algorithms are only the first steps towards global-scale building reconstruction from SAR data. Future investigations need to be done in three directions:

1. First and foremost, to improve the inference performance of the proposed networks, it is necessary to generate larger data sets covering different urban forms. In this regard, both the building area data sets and the building bounding boxes data sets need to be expanded.

2. Second, future research should be undertaken to expand the application scenario of segmentation networks, as our experiments show that the segmentation network is able to reconstruct building heights with higher accuracy.

   For instance, *domain adaptation* uses labeled data in one or more source domains to solve new tasks in a target domain, mitigating the data differences between different areas. Another promising direction is *weekly supervised segmentation*, which may take coarse annotations such as bounding boxes of buildings for training and assign pixel-wise labels to building areas.

3. Third, a fruitful area for further work would be to explore and combine SAR data of different imaging modes for reconstructing building models on larger scales and with more details.

   In this thesis, high-resolution spotlight images and stripmap images are investigated for LoD1 building reconstruction on a large scale. An important piece of information that has not been investigated is the large number of details on building facades, as shown in Figure 2.1. Visible in high-resolution spotlight and staring spotlight SAR images, these facade patterns have been studied in one building in [277]. It would be very appealing to reconstruct baseline LoD1 models using stripmap data or high-resolution spotlight data, and add detailed facade models from high-resolution or staring spotlight data to achieve building models with details close to LoD3.

Last but not least, the SAR industry is changing, and the time ahead is exciting. To date, SAR satellites have been primarily built and operated by space agencies. This situation is changing now. Not only traditional aerospace companies, such are Airbus, Maxar, and Lockheed Martin, are interested in acquiring and distributing SAR data, startups are entering the game. In January 2018, Iceye launched the first commercial SAR satellite. In October 2020, Capella Space began releasing imagery from its X-band satellites, and two months later, the first X-band Japanese SAR satellite by Synspective and the first C-band Chinese commercial SAR satellite by Spacety were both launched in December 2020. Several other companies, such as Umbra Lab, Trident Space, and PredaSAR Corp, have announced their first launch in 2021. The list of SAR companies is growing and dozens of small SAR satellites are planned for launch within several years.

On the ground segment side, big tech companies, such as AWS and Microsoft, are becoming key infrastructures for hosting and delivering data. Soon, near-real-time SAR images in meter or sub-meter resolution mapping the globe will be available.

These changes in the SAR field, will accelerate relevant research and development, expand SAR applications, and boost the demands of SAR data and products.

With these great opportunities ahead, now is the time for the SAR researchers to take advantage of deep learning techniques and big data to solve problems in researches and society with their domain knowledge.

List of Figures

1.1 Geographical distribution of (a) buildings and (b) building heights recorded in OSM as of 2021/06/06. . . . 1

2.1 TerraSAR-X basic imaging modes. . . . 6
2.2 Demonstration of the scene coverage of different SAR imaging modes. . . . 7
2.3 The geometric distortions of foreshortening, layover, and shadowing on buildings. . . . 8
2.4 Illustration of the projection geometry and the amplitude profile of two flat-roof buildings in a slant-range SAR image. . . . 9
2.5 Example of a building in SAR images of different imaging modes. . . . 10
2.6 Example of the building footprints and building regions in a typical urban area in a SAR image. . . . 11
2.7 Illustration of building models of different LoDs in CityGML. . . . 12
2.8 Refined LoD1s of 3-D building models. . . . 13
2.9 Illustration of the geocoding error from inaccurate height. . . . 15

3.1 Center Berlin area in a SAR image with and without GIS building footprints. . . . 22

4.1 The proposed workflow for generating pixel-wise ground truth of individual buildings in SAR images. . . . 33
4.2 Illustration of scene modeling steps with a simulated DSM and a real DSM in 3-D and 2-D. . . . 34
4.3 Illustration of 2.5-D and 3-D surface models. . . . 35
4.4 Example of building footprint polygons superimposed on a DSM. . . . 35
4.5 Example of extracted roof points and wall points of buildings in the SAR image coordinate system. . . . 36
4.6 Illustration of the visibility test of building footprints and the two footprint representations. . . . 37
4.7 The proposed workflow for generating the bounding box for individual buildings. . . . 38
4.8 Examples of building height sources of the same area. . . . 39
4.9 Procedures of ground truth generation for building footprints and the SAR image registration. . . . 40

5.1 Illustration of the building correspondence between SAR and GIS data. . . . 42
5.2 Illustration of the intensity-based refinement of the extracted SAR features. 44
5.3 Illustration of the visibility test of building footprint polygons. . . . 45
5.4 The Berlin study area in the UTM coordinate system for testing the registration framework. . . . 48
5.5 The test site in the SAR image coordinate system for registering building footprints to a corresponding SAR image. . . . 49

5.6 Example of the SAR feature line extraction steps. 50
5.7 Examples of the GIS features and registration results in each step. 51
5.8 The three types of grid cells in the study area based on the distance distribution. 52
5.9 Subareas clustering and registration. 52
5.10 Two cases of the polygon registration. 53
5.11 Range error maps of vertices in building polygons between registration results and ground truth. 54
5.12 The probability density of vertex distance before and after registration, compared with ground truth. 55
5.13 The Tandem-X stripmap image and building footprint polygons in central Munich area for testing the registration results on stripmap images. 56
5.14 Comparison of the SAR point-set (stripmap) and GIS point-set before and after registration. (Area 1) 56
5.15 Comparison of the SAR point-set (stripmap) and GIS point-set before and after registration. (Area 2) 57
5.16 Range error maps compared to the ground truth, resulting from radar coding using different heights. 58
5.17 SAR imaging geometry of buildings under different settings. 59

6.1 Illustration of the difference between building semantic segmentation and individual building segmentation. 63
6.2 Overview of the CG-Net architecture. 64
6.3 Architecture of the proposed CG module. 66
6.4 Architecture of the final CG module. 66
6.5 Examples of segmentation results using sensor-visible footprint segments. . 68
6.6 Examples of segmentation results using complete building footprints. . . . 69
6.7 Examples of segmentation results from different models using complete building footprints and sensor-visible footprint segments. 71
6.8 Illustration of the process generating building footprints with positioning errors. 72
6.9 Comparison of segmentation results of networks trained using building footprints with/without positioning errors. 73
6.10 The projection geometry of two flat-roof buildings in a slant-range SAR image. 74
6.11 Histogram of building height errors in the Berlin area using a spotlight image. 75
6.12 Segmentation results and the estimated building heights in the study area obtained by DeepLabv3-CG. 76
6.13 Examples of the reconstructed LoD1 building models. 77
6.14 Examples of buildings on which the geometric relationship between the estimated height and the mean height does not hold. 77
6.15 Examples of buildings with different roof types. 78
6.16 Examples of segmentation results in networks trained using spotlight SAR images and stripmap SAR images in Berlin. 79
6.17 Histogram of height errors of 4120 buildings in Berlin from inference experiments on a TerraSAR-X stripmap image. 80

6.18 Examples of segmentation results in networks trained using stripmap SAR imagery in New York. 80
6.19 Histogram of height errors of 3482 buildings in New York City from inference experiments on a TerraSAR-X stripmap image. 81

7.1 Illustration of the input and output of the object localization network in a typical urban area. 83
7.2 Illustration of bounding boxes of two buildings in a slant-range SAR image. 85
7.3 Overview of the proposed method. 85
7.4 The Berlin study area in the UTM coordinate system. 87
7.5 The Rotterdam study area in the UTM coordinate system. 88
7.6 The New York study area in the UTM coordinate system. 88
7.7 Histogram of building height errors predicted with our network. 92
7.8 Examples of predicted bounding boxes of different networks in Berlin HS and Berlin SM datasets. 93
7.9 Examples of predicted bounding boxes of different networks in Rotterdam and New York datasets. 94
7.10 Height prediction map in Berlin HS dataset. 96
7.11 Height prediction map in Berlin SM dataset. 97
7.12 Height prediction map in Rotterdam dataset. 98
7.13 Height prediction map in New York dataset. 99
7.14 Comparison of results of the proposed network trained using Berlin HS and Berlin HS-E. 100

List of Tables

2.1 TerraSAR-X imaging modes' main characteristics. 6

4.1 Additional data employed for data set generation. 31

5.1 Data used in the study area Berlin. 48
5.2 The bias and standard deviation of the registration errors. 55

6.1 Numerical results using sensor-visible footprint segments. 70
6.2 Numerical results using complete building footprints. 70
6.3 Numerical results of DeepLabv3-CG trained using CBF and CBF-E. 73

7.1 Main characteristics of the used SAR images in each data set. 89
7.2 Data sources of building footprints and heights in each data set. 89
7.3 Patch size and sample numbers in each data set. 89
7.4 Numerical results on four data sets. 91
7.5 Numerical results on trained on Berlin HS and Berlin HS-E data sets. . . . 95
7.6 Numerical results on Berlin HS and Berlin HS-NL data sets. 101

8.1 Summary of the objectives. 103

Bibliography

[1] V. Verma, R. Kumar, and S. Hsu. 3D building detection and modeling from aerial LiDAR data. In *2006 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'06)*, volume 2. IEEE, 2006.

[2] OpenStreetMap Wiki. Taginfo - OpenStreetMap Wiki. `https://taginfo.openstreetmap.org/`, 2021. Accessed: 2021-06-06.

[3] F. Rottensteiner, G. Sohn, J. Jung, M. Gerke, C. Baillard, S. Benitez, and U. Breitkopf. The ISPRS benchmark on urban object classification and 3D building reconstruction. *ISPRS Annals of the Photogrammetry, Remote Sensing and Spatial Information Sciences I-3 (2012), Nr. 1*, 1(1):293–298, 2012.

[4] C. Brenner. Building reconstruction from images and laser scanning. *International Journal of Applied Earth Observation and Geoinformation*, 6(3):187–198, 2005.

[5] D. Brunner, G. Lemoine, and L. Bruzzone. Earthquake damage assessment of buildings using VHR optical and SAR imagery. *IEEE Transactions on Geoscience and Remote Sensing*, 48(5):2403–2420, 2010.

[6] T.-L. Wang and Y.-Q. Jin. Postearthquake building damage assessment using multi-mutual information from pre-event optical image and postevent SAR image. *IEEE Geoscience and Remote Sensing Letters*, 9(3):452–456, 2012.

[7] B. Huang, Y. Li, X. Han, Y. Cui, W. Li, and R. Li. Cloud removal from optical satellite imagery with SAR imagery using sparse representation. *IEEE Geoscience and Remote Sensing Letters*, 12(5):1046–1050, 2015.

[8] X. X. Zhu, Y. Sun, Y. Shi, Y. Wang, and N. Ge. Towards global 3d/4d urban modeling using tandem-x data. In *12th European Conference on Synthetic Aperture Radar (EUSAR)*, 2018.

[9] G. L. Laprade and E. S. Leonardo. Elevations from radar imagery. *Photogrammetric Engineering*, 35(4):336–371, 1969.

[10] G. Franceschetti, A. Iodice, and D. Riccio. A canonical problem in electromagnetic backscattering from buildings. *IEEE Transactions on Geoscience and Remote Sensing*, 40(8):1787–1801, 2002.

[11] F. Tupin and M. Roux. Detection of building outlines based on the fusion of SAR and optical features. *ISPRS Journal of Photogrammetry and Remote Sensing*, 58(1-2):71–82, 2003.

[12] R. Guida, A. Iodice, and D. Riccio. Height retrieval of isolated buildings from single high-resolution sar images. *IEEE Transactions on Geoscience and Remote Sensing*, 48(7), 2010.

[13] D. Brunner, G. Lemoine, L. Bruzzone, and H. Greidanus. Building height retrieval from VHR SAR imagery based on an iterative simulation and matching technique. *IEEE Transactions on Geoscience and Remote Sensing*, 48(3):1487–1504, 2010.

[14] H. Sportouche, F. Tupin, and L. Denise. Extraction and three-dimensional reconstruction of isolated buildings in urban scenes from high-resolution optical and SAR spaceborne images. *IEEE Transactions on Geoscience and Remote Sensing*, 49(10):3932–3946, 2011.

[15] L. Wen and F. Yamazaki. Building height detection from high-resolution TerraSAR-X imagery and GIS data. In *Joint Urban Remote Sensing Event (JURSE)*, 2013.

[16] X. X. Zhu and M. Shahzad. Facade reconstruction using multiview spaceborne TomoSAR point clouds. *IEEE Transactions on Geoscience and Remote Sensing*, 52:3541–3552, 2014.

[17] M. Shahzad and Xiao Xiang Zhu. Robust Reconstruction of Building Facades for Large Areas Using Spaceborne TomoSAR point clouds. *IEEE Transactions on Geoscience and Remote Sensing*, 53(2):752–769, 2015.

[18] OpenStreetMap Wiki. Microsoft Building Footprint Data - OpenStreetMap Wiki. `https://wiki.openstreetmap.org/w/index.php?title=Microsoft_Building_Footprint_Data&oldid=2135179`, 2021. Accessed: 30-April-2021.

[19] 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.

[20] Q. Li, Y. Shi, X. Huang, and X. X. Zhu. Building footprint generation by integrating convolution neural network with feature pairwise conditional random field (fpcrf). *IEEE Transactions on Geoscience and Remote Sensing*, 58(11):7502–7519, 2020.

[21] A. Moreira, P. Prats-Iraola, M. Younis, G. Krieger, I. Hajnsek, and K. P. Papathanassiou. A tutorial on synthetic aperture radar. *IEEE Geoscience and remote sensing magazine*, 1(1):6–43, 2013.

[22] H. Breit, M. Fischer, U. Balss, and T. Fritz. TerraSAR-X Staring Spotlight Processing and Products. In *10th European Conference on Synthetic Aperture Radar (EUSAR)*, 2014.

[23] U. Steinbrecher, T. Kraus, G. C. Alfonzo, C. Grigorov, D. Schulze, and B. Bräutigam. Terrasar-x: Design of the new operational widescansar mode. In *10th European Conference on Synthetic Aperture Radar (EUSAR)*, 2014.

[24] M. Eineder, T. Fritz, J. Mittermayer, A. Roth, E. Boerner, and H. Breit. TerraSAR-X ground segment, basic product specification document. Technical report, Cluster Applied Remote Sensing (CAF) Oberpfaffenhofen (Germany), 2008.

[25] J. C. Curlander. Utilization of spaceborne sar data for mapping. *IEEE Transactions on Geoscience and Remote Sensing*, 22(2):106–112, 1984.

[26] J. W. Goodman. Statistical properties of laser speckle patterns. In *Laser speckle and related phenomena*, pages 9–75. Springer, 1975.

[27] G. Franceschetti and R. Lanari. *Synthetic aperture radar processing*. CRC press, 2018.

[28] C.-A. Deledalle, L. Denis, and F. Tupin. Iterative weighted maximum likelihood denoising with probabilistic patch-based weights. *IEEE Transactions on Image Processing*, 18(12):2661–2672, 2009.

[29] G. Baier, X. X. Zhu, M. Lachaise, H. Breit, and R. Bamler. Nonlocal InSAR filtering for DEM generation and addressing the staircasing effect. In *11th European Conference on Synthetic Aperture Radar (EUSAR)*, 2016.

[30] S. J. Auer. *3D synthetic aperture radar simulation for interpreting complex urban reflection scenarios*. PhD thesis, Technische Universität München, 2011.

[31] S. Auer and S. Gernhardt. Linear signatures in urban SAR images - partly misinterpreted? *IEEE Geoscience and Remote Sensing Letters*, 11(10):1762–1766, 2014.

[32] T. H. Kolbe, G. Gröger, and L. Plümer. CityGML: Interoperable access to 3D city models. In *International Symposium on Geo-Information for Disaster Management (Gi4DM)*, 2005.

[33] F. Biljecki, H. Ledoux, and J. Stoter. An improved LOD specification for 3D building models. *Computers, Environment and Urban Systems*, 59:25–37, 2016.

[34] I. Guskov and B. Brewington. Automatic generation of 2.5D extruded polygons from full 3D models, 2015. US Patent App. 13/024,855.

[35] K. A. Ohori, H. Ledoux, F. Biljecki, and J. Stoter. Modeling a 3D city model and its levels of detail as a true 4d model. *ISPRS International Journal of Geo-Information*, 4(3):1055–1075, 2015.

[36] M. Over, A. Schilling, S. Neubauer, and A. Zipf. Generating web-based 3D City Models from OpenStreetMap: The current situation in Germany. *Computers, Environment and Urban Systems*, 34(6):496–507, 2010.

[37] S. U. Baig and A. Abdul-Rahman. Generalization of buildings within the framework of CityGML. *Geo-spatial Information Science*, 16(4):247–255, 2013.

[38] A. Forberg. Generalization of 3D building data based on a scale-space approach. *ISPRS Journal of Photogrammetry and Remote Sensing*, 62(2):104–111, 2007.

[39] A. Henn, C. Römer, G. Gröger, and L. Plümer. Automatic classification of building types in 3D city models. *GeoInformatica*, 16(2):281–306, 2012.

[40] J. Hofierka and M. Zlocha. A new 3-D solar radiation model for 3-D city models. *Transactions in GIS*, 16(5):681–690, 2012.

[41] A. Strzalka, N. Alam, E. Duminil, V. Coors, and U. Eicker. Large scale integration of photovoltaics in cities. *Applied Energy*, 93:413–421, 2012.

[42] J. Stoter, H. De Kluijver, and V. Kurakula. 3D noise mapping in urban areas. *International Journal of Geographical Information Science*, 22(8):907–924, 2008.

[43] V. Varduhn, R.-P. Mundani, and E. Rank. Multi-resolution models: Recent progress in coupling 3D geometry to environmental numerical simulation. *3D Geoinformation Science*, pages 55–69, 2015.

[44] N. Alam, V. Coors, and S. Zlatanova. Detecting shadow for direct radiation using citygml models for photovoltaic potentiality analysis. In E. Claire, Z. Sisi, R. Massimo, and L. Robert, editors, *Urban and Regional Data Management: UDMS Annual 2013 (1st ed.)*, pages 191–196. CRC Press London, UK, 2013.

[45] Z. Li, Z. Zhang, and K. Davey. Estimating geographical PV potential using liDAR data for buildings in downtown San Francisco. *Transactions in GIS*, 19(6):930–963, 2015.

[46] J. D. Wegner, U. Soergel, and A. Thiele. Building extraction in urban scenes from high-resolution InSAR data and optical imagery. In *Joint Urban Remote Sensing Event (JURSE)*, 2009.

[47] A. Thiele, S. Hinz, and E. Cadario. Combining GIS and InSAR data for 3D building reconstruction. In *IEEE International Geoscience and Remote Sensing Symposium (IGARSS)*, 2010.

[48] Y. Sun, M. Shahzad, and X. X. Zhu. Building height estimation in single SAR image using OSM building footprints. In *Joint Urban Remote Sensing Event (JURSE)*, 2017.

[49] City of New York. Building Footprints. NYC Open Data. `https://data.cityofnewyork.us/Housing-Development/Building-Footprints/nqwf-w8eh.`, 2021. Accessed: 29-04-2021.

[50] Berlin Partner für Wirtschaft und Technologie GmbH. Berlin3D - Downloadportal. `http://www.businesslocationcenter.de/berlin3d-downloadportal/`, 2014. Accessed: 20-02-2017.

[51] OpenStreetMap Wiki. METEOR Project in Tanzania - OpenStreetMap Wiki. `https://wiki.openstreetmap.org/w/index.php?title=METEOR_Project_in_Tanzania&oldid=2024407`, 2020. Accessed: 30-April-2021.

[52] Q. Li, L. Mou, Y. Hua, Y. Sun, P. Jin, Y. Shi, and X. X. Zhu. Instance segmentation of buildings using keypoints. In *IEEE International Geoscience and Remote Sensing Symposium IGARSS*, 2020.

[53] M. Schwabisch. A fast and efficient technique for SAR interferogram geocoding. In *International Geoscience and Remote Sensing Symposium Proceedings (IGARSS)*, 1998.

[54] M. Eineder. Efficient simulation of SAR interferograms of large areas and of rugged terrain. *IEEE Transactions on Geoscience and Remote Sensing*, 41(6):1415–1427, 2003.

[55] H. Breit, T. Fritz, U. Balss, M. Lachaise, A. Niedermeier, and M. Vonavka. TerraSAR-X SAR processing and products. *IEEE Transactions on Geoscience and Remote Sensing*, 48(2):727–740, 2009.

[56] M. Jehle, D. Perler, D. Small, A. Schubert, and E. Meier. Estimation of atmospheric path delays in TerraSAR-X data using models vs. measurements. *Sensors*, 8(12):8479–8491, 2008.

[57] U. Balss, C. Gisinger, and M. Eineder. Measurements on the absolute 2-D and 3-D localization accuracy of TerraSAR-X. *Remote Sensing*, 10(4):656, 2018.

[58] S. Montazeri, F. Rodríguez González, and X. X. Zhu. Geocoding error correction for InSAR point clouds. *Remote Sensing*, 10(10):1523, 2018.

[59] P. A. Zandbergen. Positional accuracy of spatial data: Non-normal distributions and a critique of the national standard for spatial data accuracy. *Transactions in GIS*, 12(1):103–130, 2008.

[60] H. Fan, A. Zipf, Q. Fu, and P. Neis. Quality assessment for building footprints data on openstreetmap. *International Journal of Geographical Information Science*, 28(4):700–719, 2014.

[61] J.-M. Nasr and D. Vidal-Madjar. Image simulation of geometric targets for spaceborne synthetic aperture radar. *IEEE transactions on geoscience and remote sensing*, 29(6):986–996, 1991.

[62] N. Taket, S. Howarth, and R. E. Burge. A model for the imaging of urban areas by synthetic aperture radar. *IEEE Transactions on Geoscience and Remote Sensing*, 29(3):432–443, 1991.

[63] Y. Dong, B. Forster, and C. Ticehurst. Radar backscatter analysis for urban environments. *International journal of remote sensing*, 18(6):1351–1364, 1997.

[64] G. Franceschetti, A. Iodice, D. Riccio, and G. Ruello. SAR raw signal simulation for urban structures. *IEEE Transactions on Geoscience and Remote Sensing*, 41(9):1986–1995, 2003.

[65] F. Xu and Y.-Q. Jin. Imaging simulation of polarimetric SAR for a comprehensive terrain scene using the mapping and projection algorithm. *IEEE Transactions on Geoscience and Remote Sensing*, 44(11):3219–3234, 2006.

[66] T. Balz and U. Stilla. Hybrid GPU-based single-and double-bounce SAR simulation. *IEEE Transactions on Geoscience and Remote Sensing*, 47(10):3519–3529, 2009.

[67] S. Auer, S. Hinz, and R. Bamler. Ray-Tracing Simulation Techniques for Understanding High-Resolution SAR Images. *IEEE Transactions on Geoscience and Remote Sensing*, 48:1445–1456, 2010.

[68] A. Bennett and D. Blacknell. The extraction of building dimensions from high resolution SAR imagery. In *2003 Proceedings of the International Conference on Radar (IEEE Cat. No.03EX695)*, 2003.

[69] R. Hill, C. Moate, and D. Blacknell. Estimating building dimensions from synthetic aperture radar image sequences. *IET Radar, Sonar & Navigation*, 2(3):189–199, 2008.

[70] C. Tison, F. Tupin, and H. Maitre. Retrieval of building shapes from shadows in high resolution SAR interferometric images. In *International Geoscience and Remote Sensing Symposium (IGARSS)*, volume 3. IEEE, 2004.

[71] F. Tupin. Extraction of 3D information using overlay detection on SAR images. In *2nd GRSS/ISPRS Joint Workshop on Remote Sensing and Data Fusion over Urban Areas*, 2003.

[72] F. Xu and Y.-Q. Jin. Automatic reconstruction of building objects from multiaspect meter-resolution SAR images. *IEEE Transactions on Geoscience and Remote Sensing*, 45(7):2336–2353, 2007.

[73] A. Thiele, E. Cadario, K. Schulz, and U. Soergel. Analysis of gable-roofed building signature in multiaspect InSAR data. *IEEE Geoscience and Remote Sensing Letters*, 7(1):83–87, 2010.

[74] A. Thiele, M. M. Wurth, M. Even, and S. Hinz. Extraction of building shape from TanDEM-X data. *ISPRS - International Archives of the Photogrammetry, Remote Sensing and Spatial Information Sciences*, XL-1/W1:345–350, 2013.

[75] H. Sportouche, F. Tupin, and L. Denise. Building extraction and 3D reconstruction in urban areas from high-resolution optical and SAR imagery. In *Joint Urban Remote Sensing Event (JURSE)*, 2009.

[76] J. D. Wegner, J. R. Ziehn, and U. Soergel. Combining high-resolution optical and InSAR features for height estimation of buildings with flat roofs. *IEEE Transactions on Geoscience and Remote Sensing*, 52(9):5840–5854, 2014.

[77] W. He, M. Jäger, A. Reigber, and O. Hellwich. Building extraction from polarimetric SAR data using mean shift and conditional random fields. In *7th European Conference on Synthetic Aperture Radar (EUSAR)*, 2008.

[78] L. Zhao, X. Zhou, and G. Kuang. Building detection from urban SAR image using building characteristics and contextual information. *EURASIP Journal on Advances in Signal Processing*, 2013(1):56, 2013.

[79] Y. Cao, C. Su, and G. Yang. Detecting the number of buildings in a single high-resolution SAR image. *European Journal of Remote Sensing*, 47:513–535, 2014.

[80] A. Ferro, D. Brunner, and L. Bruzzone. Automatic detection and reconstruction of building radar footprints from single VHR SAR images. *IEEE Transactions on Geoscience and Remote Sensing*, 51(2):935–952, 2013.

[81] S. Chen, H. Wang, F. Xu, and Y.-Q. Jin. Automatic recognition of isolated buildings on single-aspect SAR image using range detector. *IEEE Geoscience and Remote Sensing Letters*, 12(2):219–223, 2015.

[82] B. Liu, K. Tang, and J. Liang. A bottom-up/top-down hybrid algorithm for model-based building detection in single very high resolution SAR image. *IEEE Geoscience and Remote Sensing Letters*, 14(6):926–930, 2017.

[83] M. Jahangir, D. Blacknell, C. Moate, and R. Hill. Extracting information from shadows in SAR imagery. In *2007 International Conference on Machine Vision*. IEEE, 2007.

[84] Z. Wang, L. Jiang, L. Lin, and W. Yu. Building height estimation from high resolution SAR imagery via model-based geometrical structure prediction. *Progress In Electromagnetics Research*, 41:11–24, 2015.

[85] M. Quartulli and M. Datcu. Stochastic geometrical modeling for built-up area understanding from a single SAR intensity image with meter resolution. *IEEE Transactions on Geoscience and Remote Sensing*, 42(9):1996–2003, 2004.

[86] R. Guida, A. Iodice, D. Riccio, and U. Stilla. Model-based interpretation of high-resolution SAR images of buildings. *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, 1(2):107–119, 2008.

[87] K. Hoepfner. Recovery of building structure from IFSAR-derived elevation maps. *Computer Science*, 1999.

[88] P. Gamba, B. Houshmand, and M. Saccani. Detection and extraction of buildings from interferometric SAR data. *IEEE Transactions on Geoscience and Remote Sensing*, 38(1):611–617, 2000.

[89] R. Bolter and F. Leberl. Detection and reconstruction of human scale features from high resolution interferometric SAR data. In *Proceedings 15th International Conference on Pattern Recognition. ICPR-2000*, volume 4. IEEE, 2000.

[90] R. Bolter and F. Leberl. Shape-from-shadow building reconstruction from multiple view SAR images. In *24th workshop of the Austrian Association for Pattern Recognition (ÖAGM/AAPR)*, 2000.

[91] F. Cellier, H. Oriot, and J.-M. Nicolas. Hypothesis management for building reconstruction from high resolution InSAR imagery. In *2006 IEEE International Symposium on Geoscience and Remote Sensing*, 2006.

[92] U. Soergel, U. Thoennessen, and U. Stilla. Iterative building reconstruction from multi-aspect InSAR data. In *3-D reconstruction from airborne laserscanner and InSAR data*, 2003.

[93] C. Dubois, A. Thiele, and S. Hinz. Building detection and building parameter retrieval in InSAR phase images. *ISPRS Journal of Photogrammetry and Remote Sensing*, 114:228–241, 2016.

[94] M. Quartulli and M. Datcu. Bayesian model based city reconstruction from high resolution ISAR data. In *IEEE/ISPRS Joint Workshop on Remote Sensing and Data Fusion over Urban Areas (Cat. No. 01EX482)*. IEEE, 2001.

[95] M. Quartulli and M. Datcu. Information fusion for scene understanding from interferometric SAR data in urban environments. *IEEE Transactions on geoscience and remote sensing*, 41(9):1976–1985, 2003.

[96] C. Tison, F. Tupin, and H. Maître. A fusion scheme for joint retrieval of urban height map and classification from high-resolution interferometric SAR images. *IEEE Transactions on Geoscience and remote Sensing*, 45(2):496–505, 2007.

[97] F. Leberl. *Radargrammetric image processing*. Artech House, 1989.

[98] E. Michaelsen, U. Soergel, and U. Thoennessen. Perceptual grouping for automatic detection of man-made structures in high-resolution SAR data. *Pattern Recognition Letters*, 27(4):218–225, 2006.

[99] U. Soergel, E. Michaelsen, A. Thiele, E. Cadario, and U. Thoennessen. Stereo analysis of high-resolution SAR images for building height estimation in cases of orthogonal aspect directions. *ISPRS Journal of Photogrammetry and Remote Sensing*, 64(5):490–500, 2009.

[100] Y.-Q. Jin and F. Xu. *Polarimetric scattering and SAR information retrieval*. John Wiley & Sons, 2013.

[101] H. Oriot and H. Cantalloube. Circular SAR imagery for urban remote sensing. In *7th European conference on synthetic aperture radar (EUSAR)*, 2008.

[102] S. Palm, H. M. Oriot, and H. M. Cantalloube. Radargrammetric DEM Extraction Over Urban Area Using Circular SAR Imagery. *IEEE Transactions on Geoscience and Remote Sensing*, 50(11):4720–4725, 2012.

[103] E. Simonetto, H. Oriot, and R. Garello. Rectangular building extraction from stereoscopic airborne radar images. *IEEE Transactions on Geoscience and Remote Sensing*, 43(10):2386–2395, 2005.

[104] R. Xiao, C. Lesher, and B. Wilson. Building detection and localization using a fusion of interferometric synthetic aperture radar and multispectral image. In *ARPA Image Understanding Workshop*, 1998.

[105] A. Thiele, E. Cadario, K. Schulz, U. Thoennessen, and U. Soergel. Feature extraction of gable-roofed buildings from multi-aspect high-resolution InSAR data. In *2007 IEEE International Geoscience and Remote Sensing Symposium*, Barcelona, Spain, 2007. IEEE.

[106] A. Thiele, E. Cadario, K. Schulz, U. Thoennessen, and U. Soergel. Reconstruction of residential buildings by detail analysis of multi-aspect InSAR data. In U. Michel, D. L. Civco, M. Ehlers, and H. J. Kaufmann, editors, *Remote Sensing for Environmental Monitoring, GIS Applications, and Geology VIII*, Cardiff, Wales, United Kingdom, 2008.

[107] A. Thiele, J. D. Wegner, and U. Soergel. Building reconstruction from multi-aspect InSAR data. In *Radar Remote Sensing of Urban Areas*, pages 187–214. Springer, 2010.

[108] A. Reigber and A. Moreira. First demonstration of airborne SAR tomography using multibaseline l-band data. *IEEE Transactions on Geoscience and Remote Sensing*, 38(5):2142–2152, 2000.

[109] X. X. Zhu and R. Bamler. Very high resolution spaceborne SAR tomography in urban environment. *IEEE Transactions on Geoscience and Remote Sensing*, 48(12):4296–4308, 2010.

[110] M. Shahzad and X. X. Zhu. Automatic detection and reconstruction of 2-D/3-D building shapes from spaceborne TomoSAR point clouds. *IEEE Transactions on Geoscience and Remote Sensing*, 54(3):1292–1310, 2015.

[111] Y. Sun, M. Shahzad, and X. X. Zhu. First prismatic building model reconstruction from tomosar points clouds. *The International Archives of the Photogrammetry, Remote Sensing and Spatial Information Sciences*, 41:381–386, 2016.

[112] Y. Sun. 3D building reconstruction from spaceborne TomoSAR point cloud. Master's thesis, Technical University of Munich, 2016.

[113] C. Rambour, L. Denis, F. Tupin, H. Oriot, Y. Huang, and L. Ferro-Famil. Urban surface reconstruction in SAR tomography by graph-cuts. *Computer vision and image understanding*, 188:102791, 2019.

[114] C. Rambour, L. Denis, and F. Tupin. 3D Buildings Reconstruction with SAR Tomography Guided by Partial Footprints Information. In *13th European Conference on Synthetic Aperture Radar (EUSAR)*, 2021.

[115] O. D'Hondt, S. Guillaso, and O. Hellwich. Automatic extraction of geometric structures for 3D reconstruction from tomographic SAR data. In *2012 IEEE International Geoscience and Remote Sensing Symposium*, 2012.

[116] O. D'Hondt, S. Guillaso, and O. Hellwich. Geometric primitive extraction for 3D reconstruction of urban areas from tomographic SAR data. In *Joint Urban Remote Sensing Event 2013*, 2013.

[117] A. Ley, O. D'Hondt, and O. Hellwich. Regularization and completion of tomosar point clouds in a projected height map domain. *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, 11(6):2104–2114, 2018.

[118] O. D'Hondt, C. López-Martínez, S. Guillaso, and O. Hellwich. Nonlocal filtering applied to 3-d reconstruction of tomographic SAR data. *IEEE Transactions on Geoscience and Remote Sensing*, 56(1):272–285, 2018.

[119] Y. Shi, R. Bamler, Y. Wang, and X. X. Zhu. SAR tomography at the limit: Building height reconstruction using only 3–5 tandem-x bistatic interferograms. *IEEE Transactions on Geoscience and Remote Sensing*, 58(11):8026–8037, 2020.

[120] W. Liu, F. Yamazaki, B. Adriano, E. Mas, and S. Koshimura. Development of Building Height Data in Peru from High-Resolution SAR Imagery. *Journal of Disaster Research*, 9:1042–1049, 2014.

[121] G. F. Hepner, B. Houshmand, I. Kulikov, and N. Bryant. Investigation of the potential for the integration of AVIRIS and IFSAR for urban analysis. *Photogrammetric engineering and remote sensing*, 64(8):813–820, 1998.

[122] F. Zhang, Y. Shao, X. Zhang, and T. Balz. Building L-shape footprint extraction from high resolution SAR image. In *Joint Urban Remote Sensing Event (JURSE)*, 2011.

[123] Y. Wang, F. Tupin, C. Han, and J.-M. Nicolas. Building detection from high resolution PolSAR data by combining region and edge information. In *IEEE International Geoscience and Remote Sensing Symposium (IGARSS)*, 2008.

[124] J. Shermeyer, D. Hogan, J. Brown, A. Van Etten, N. Weir, F. Pacifici, R. Haensch, A. Bastidas, S. Soenen, T. Bacastow, et al. SpaceNet 6: Multi-sensor all weather mapping dataset. *arxiv:2004.06500*, 2020.

[125] J. Chen, C. Wang, H. Zhang, F. Wu, B. Zhang, and W. Lei. Automatic detection of low-rise gable-roof building from single submeter SAR images based on local multilevel segmentation. *Remote Sensing*, 9(3):263, 2017.

[126] F. Cellier, H. Oriot, and J.-M. Nicolas. Introduction of the mean shift algorithm in SAR imagery: Application to shadow extraction for building reconstruction. In *EARSeL Workshop 3D-Remote Sensing*, 2005.

[127] A. Thiele, C. Dubois, E. Cadario, and S. Hinz. GIS-supported iterative filtering approach for building height estimation from InSAR data. In *9th European Conference on Synthetic Aperture Radar (EUSAR)*, 2012.

[128] Y. Zhang, X. Sun, A. Thiele, and S. Hinz. Stochastic geometrical model and monte carlo optimization methods for building reconstruction from InSAR data. *ISPRS Journal of Photogrammetry and Remote Sensing*, 108:49–61, 2015.

[129] Q. Lv, Y. Dou, X. Niu, J. Xu, J. Xu, and F. Xia. Urban land use and land cover classification using remotely sensed SAR data through deep belief networks. *Journal of Sensors*, 2015:1–10, 2015.

[130] Z. Zhang, H. Wang, F. Xu, and Y.-Q. Jin. Complex-valued convolutional neural network and its application in polarimetric SAR image classification. *IEEE Transactions on Geoscience and Remote Sensing*, 55(12):7177–7188, 2017.

[131] Y. Duan, F. Liu, L. Jiao, P. Zhao, and L. Zhang. SAR image segmentation based on convolutional-wavelet neural network and markov random field. *Pattern Recognition*, 64:255–267, 2017.

[132] F. Mohammadimanesh, B. Salehi, M. Mahdianpari, E. Gill, and M. Molinier. A new fully convolutional neural network for semantic segmentation of polarimetric SAR imagery in complex land cover ecosystem. *ISPRS Journal of Photogrammetry and Remote Sensing*, 151:223–236, 2019.

[133] S. Chen and H. Wang. SAR target recognition based on deep learning. In *International Conference on Data Science and Advanced Analytics (DSAA)*, 2014.

[134] F. Gao, T. Huang, J. Sun, J. Wang, A. Hussain, and E. Yang. A new algorithm for SAR image target recognition based on an improved deep convolutional neural network. *Cognitive Computation*, 11(6):809–824, 2019.

[135] M. Gong, J. Zhao, J. Liu, Q. Miao, and L. Jiao. Change detection in synthetic aperture radar images based on deep neural networks. *IEEE Transactions on Neural Networks and Learning Systems*, 27(1):125–138, 2016.

[136] J. Geng, X. Ma, X. Zhou, and H. Wang. Saliency-guided deep neural networks for SAR image change detection. *IEEE Transactions on Geoscience and Remote Sensing*, 57(10):7365–7377, 2019.

[137] M. Shahzad, M. Maurer, F. Fraundorfer, Y. Wang, and X. X. Zhu. Buildings detection in VHR SAR images using fully convolution neural networks. *IEEE Transactions on Geoscience and Remote Sensing*, 57(2):1100–1116, 2019.

[138] Y. Sun, Y. Hua, L. Mou, and X. X. Zhu. Large-scale building height estimation from single VHR SAR image using fully convolutional network and GIS building footprints. In *Joint Urban Remote Sensing Event (JURSE)*, 2019.

[139] S. Gernhardt, N. Adam, M. Eineder, and R. Bamler. Potential of very high resolution SAR for persistent scatterer interferometry in urban areas. *Annals of GIS*, 16(2):103–111, 2010.

[140] X. X. Zhu, Y. Wang, S. Gernhardt, and R. Bamler. Tomo-GENESIS: DLR's tomographic SAR processing system. In *Joint Urban Remote Sensing Event (JURSE)*, 2013.

[141] R. Guo and X. X. Zhu. High-rise building feature extraction using high resolution spotlight TanDEM-X data. In *10th European Conference on Synthetic Aperture Radar (EUSAR)*, 2014.

[142] W. Liu, K. Suzuki, and F. Yamazaki. Height estimation for high-rise buildings based on InSAR analysis. In *Joint Urban Remote Sensing Event (JURSE)*, 2015.

[143] K. Tang, B. Liu, and B. Zou. High-rise building detection in dense urban area based on high resolution SAR images. In *IEEE International Geoscience and Remote Sensing Symposium (IGARSS)*, 2016.

[144] P. Lu, K. Du, W. Yu, and H. Feng. New building signature extraction method from single very high-resolution synthetic aperture radar images based on symmetric analysis. *Journal of Applied Remote Sensing*, 9(1):095072, 2015.

[145] H. Zhang, H. Lin, and Y. Li. Impacts of feature normalization on optical and SAR data fusion for land use/land cover classification. *IEEE Geoscience and Remote Sensing Letters*, 12(5):1061–1065, 2015.

[146] D. Tuia, G. Moser, B. Le Saux, B. Bechtel, and L. See. 2017 IEEE GRSS data fusion contest: Open data for global multimodal land use classification [technical committees]. *IEEE Geoscience and Remote Sensing Magazine*, 5(1):70–73, 2017.

[147] F. Pacifici, F. Del Frate, W. J. Emery, P. Gamba, and J. Chanussot. Urban mapping using coarse SAR and optical data: Outcome of the 2007 GRSS data fusion contest. *IEEE Geoscience and Remote Sensing Letters*, 5(3):331–335, 2008.

[148] X. Huang, T. Hu, J. Li, Q. Wang, and J. A. Benediktsson. Mapping urban areas in china using multisource data with a novel ensemble svm method. *IEEE Transactions on Geoscience and Remote Sensing*, 56(8):4258–4273, 2018.

[149] M. Dalponte, L. Bruzzone, and D. Gianelle. Fusion of hyperspectral and lidar remote sensing data for classification of complex forest areas. *IEEE Transactions on Geoscience and Remote Sensing*, 46(5):1416–1427, 2008.

[150] P. Kempeneers, F. Sedano, L. Seebach, P. Strobl, and J. San-Miguel-Ayanz. Data fusion of different spatial resolution remote sensing images applied to forest-type mapping. *IEEE Transactions on Geoscience and Remote Sensing*, 49(12):4977–4986, 2011.

[151] D. Li. Remotely sensed images and GIS data fusion for automatic change detection. *International Journal of Image and Data Fusion*, 1(1):99–108, 2010.

[152] G. Camps-Valls, L. Gómez-Chova, J. Muñoz-Marí, J. L. Rojo-Álvarez, and M. Martínez-Ramón. Kernel-based framework for multitemporal and multisource remote sensing data classification and change detection. *IEEE Transactions on Geoscience and Remote Sensing*, 46(6):1822–1835, 2008.

[153] M. Schmitt, F. Tupin, and X. X. Zhu. Fusion of SAR and optical remote sensing data—Challenges and recent trends. In *Geoscience and Remote Sensing Symposium (IGARSS), 2017 IEEE International*. IEEE, 2017.

[154] G. Palubinskas, P. Reinartz, and R. Bamler. Image acquisition geometry analysis for the fusion of optical and radar remote sensing data. *International Journal of Image and Data Fusion*, 1(3):271–282, 2010.

[155] M. Chini, N. Pierdicca, and W. J. Emery. Exploiting SAR and VHR optical images to quantify damage caused by the 2003 bam earthquake. *IEEE Transactions on Geoscience and Remote Sensing*, 47(1):145–152, 2008.

[156] S. Suri and P. Reinartz. Mutual-information-based registration of TerraSAR-X and ikonos imagery in urban areas. *IEEE Transactions on Geoscience and Remote Sensing*, 48(2):939–949, 2009.

[157] Y. Keller and A. Averbuch. Multisensor image registration via implicit similarity. *IEEE transactions on pattern analysis and machine intelligence*, 28(5):794–801, 2006.

[158] W. Shi, F. Su, R. Wang, and J. Fan. A visual circle based image registration algorithm for optical and SAR imagery. In *International Geoscience and Remote Sensing Symposium (IGARSS)*, 2012.

[159] M. Hasan, M. R. Pickering, and X. Jia. Robust automatic registration of multimodal satellite images using CCRE with partial volume interpolation. *IEEE Transactions on Geoscience and Remote Sensing*, 50(10):4050–4061, 2012.

[160] A. Sedaghat, M. Mokhtarzade, and H. Ebadi. Uniform robust scale-invariant feature matching for optical remote sensing images. *IEEE Transactions on Geoscience and Remote Sensing*, 49(11):4516–4527, 2011.

[161] P. Schwind, S. Suri, P. Reinartz, and A. Siebert. Applicability of the SIFT operator to geometric SAR image registration. *International Journal of Remote Sensing*, 31(8):1959–1980, 2010.

[162] J. Fan, Y. Wu, F. Wang, Q. Zhang, G. Liao, and M. Li. SAR image registration using phase congruency and nonlinear diffusion-based SIFT. *IEEE Geoscience and Remote Sensing Letters*, 12(3):562–566, 2014.

[163] H. Sui, C. Xu, J. Liu, and F. Hua. Automatic optical-to-SAR image registration by iterative line extraction and voronoi integrated spectral point matching. *IEEE Transactions on geoscience and remote sensing*, 53(11):6058–6072, 2015.

[164] D. G. Lowe. Distinctive image features from scale-invariant keypoints. *International journal of computer vision*, 60(2):91–110, 2004.

[165] F. Dellinger, J. Delon, Y. Gousseau, J. Michel, and F. Tupin. SAR-SIFT: a SIFT-like algorithm for SAR images. *IEEE Transactions on Geoscience and Remote Sensing*, 53(1):453–466, 2014.

[166] S. Wang, H. You, and K. Fu. BFSIFT: A novel method to find feature matches for SAR image registration. *IEEE Geoscience and Remote Sensing Letters*, 9(4):649–653, 2011.

[167] B. Fan, C. Huo, C. Pan, and Q. Kong. Registration of optical and SAR satellite images by exploring the spatial relationship of the improved SIFT. *IEEE Geoscience and Remote Sensing Letters*, 10(4):657–661, 2012.

[168] Y. Xiang, F. Wang, and H. You. OS-SIFT: A robust SIFT-like algorithm for high-resolution optical-to-SAR image registration in suburban areas. *IEEE Transactions on Geoscience and Remote Sensing*, 56(6):3078–3090, 2018.

[169] Y. Ye and L. Shen. HOPC: A novel similarity metric based on geometric structural properties for multi-modal remote sensing image matching. *ISPRS Annals of the Photogrammetry, Remote Sensing and Spatial Information Sciences*, 3:9, 2016.

[170] J. Li, Q. Hu, and M. Ai. RIFT: Multi-modal image matching based on radiation-variation insensitive feature transform. *IEEE Transactions on Image Processing*, 29:3296–3310, 2019.

[171] P. Gamba, F. Dell'Acqua, and G. Lisini. Improving urban road extraction in high-resolution images exploiting directional filtering, perceptual grouping, and simple topological concepts. *IEEE Geoscience and Remote Sensing Letters*, 3(3):387–391, 2006.

[172] V. Poulain, J. Inglada, M. Spigai, J.-Y. Tourneret, and P. Marthon. High resolution optical and SAR image fusion for road database updating. In *International Geoscience and Remote Sensing Symposium (IGARSS)*, 2010.

[173] T. Perciano, F. Tupin, R. Hirata Jr, and R. M. Cesar Jr. A two-level markov random field for road network extraction and its application with optical, SAR, and multitemporal data. *International Journal of Remote Sensing*, 37(16):3584–3610, 2016.

[174] J. D. Wegner, R. Hansch, A. Thiele, and U. Soergel. Building detection from one orthophoto and high-resolution InSAR data using conditional random fields. *IEEE Journal of selected topics in applied Earth Observations and Remote Sensing*, 4(1):83–91, 2010.

[175] V. Poulain, J. Inglada, M. Spigai, J.-Y. Tourneret, and P. Marthon. High-resolution optical and SAR image fusion for building database updating. *IEEE Transactions on Geoscience and Remote Sensing*, 49(8):2900–2910, 2011.

[176] M. Teimouri, M. Mokhtarzade, and M. J. Valadan Zoej. Optimal fusion of optical and SAR high-resolution images for semiautomatic building detection. *GIScience & Remote Sensing*, 53(1):45–62, 2016.

[177] L. Mou, M. Schmitt, Y. Wang, and X. X. Zhu. A CNN for the identification of corresponding patches in SAR and optical imagery of urban scenes. In *2017 Joint Urban Remote Sensing Event (JURSE)*. IEEE, 2017.

[178] N. Merkle, W. Luo, S. Auer, R. Müller, and R. Urtasun. Exploiting deep matching and SAR data for the geo-localization accuracy improvement of optical satellite images. *Remote Sensing*, 9(6):586, 2017.

[179] N. Merkle, S. Auer, R. Müller, and P. Reinartz. Exploring the potential of conditional adversarial networks for optical and SAR image matching. *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, 11(6):1811–1820, 2018.

[180] P. Gamba, F. Dell'Acqua, and G. Trianni. Rapid damage detection in the bam area using multitemporal SAR and exploiting ancillary data. *IEEE Transactions on Geoscience and Remote Sensing*, 45(6):1582–1589, 2007.

[181] S. Suchandt, M. Eineder, H. Breit, and H. Runge. Analysis of ground moving objects using srtm/x-SAR data. *ISPRS Journal of Photogrammetry and Remote Sensing*, 61(3-4):209–224, 2006.

[182] S. B. Borah, T. Sivasankar, M. Ramya, and P. Raju. Flood inundation mapping and monitoring in kaziranga national park, assam using sentinel-1 SAR data. *Environmental monitoring and assessment*, 190(9):520, 2018.

[183] M. R. Rahman and P. K. Thakur. Detecting, mapping and analysing of flood water propagation using synthetic aperture radar (SAR) satellite data and GIS: A case study from the kendrapara district of orissa state of india. *The Egyptian Journal of Remote Sensing and Space Science*, 21:S37–S41, 2018.

[184] R. Lanari, G. Zeni, M. Manunta, S. Guarino, P. Berardino, and E. Sansosti. An integrated SAR/GIS approach for investigating urban deformation phenomena: a case study of the city of naples, italy. *International Journal of Remote Sensing*, 25(14):2855–2867, 2004.

[185] Y. Dong, Q. Li, A. Dou, and X. Wang. Extracting damages caused by the 2008 Ms 8.0 Wenchuan earthquake from SAR remote sensing data. *Journal of Asian Earth Sciences*, 40(4):907–914, 2011.

[186] X. X. Zhu, N. Ge, and M. Shahzad. Joint sparsity in SAR tomography for urban mapping. *IEEE Journal of Selected Topics in Signal Processing*, 9(8):1498–1509, 2015.

[187] Y. Sun, Y. Wang, and X. Zhu. Automatic registration of SAR image and GIS building footprints data in dense urban area. In *International Geoscience and Remote Sensing Symposium (IGARSS)*, 2019.

[188] Y. Zhou, H. Wang, F. Xu, and Y.-Q. Jin. Polarimetric SAR image classification using deep convolutional neural networks. *IEEE Geoscience and Remote Sensing Letters*, 13(12):1935–1939, 2016.

[189] Z. Zhao, L. Jiao, J. Zhao, J. Gu, and J. Zhao. Discriminant deep belief network for high-resolution SAR image classification. *Pattern Recognition*, 61:686–701, 2017.

[190] J. Geng, H. Wang, J. Fan, and X. Ma. Deep supervised and contractive neural network for SAR image classification. *IEEE Transactions on Geoscience and Remote Sensing*, 55(4):2442–2459, 2017.

[191] J. Ding, B. Chen, H. Liu, and M. Huang. Convolutional neural network with data augmentation for SAR target recognition. *IEEE Geoscience and Remote Sensing Letters*, pages 1–5, 2016.

[192] M. Kang, K. Ji, X. Leng, and Z. Lin. Contextual region-based convolutional neural network with multilayer fusion for SAR ship detection. *Remote Sensing*, 9(8):860, 2017.

[193] Q. Zhang, Q. Yuan, J. Li, Z. Yang, and X. Ma. Learning a dilated residual network for SAR image despeckling. *Remote Sensing*, 10(2):196, 2018.

[194] X. Ma, C. Wang, Z. Yin, and P. Wu. SAR image despeckling by noisy reference-based deep learning method. *IEEE Transactions on Geoscience and Remote Sensing*, 58(12):8807–8818, 2020.

[195] L. Wang, K. A. Scott, L. Xu, and D. A. Clausi. Sea ice concentration estimation during melt from dual-pol SAR scenes using deep convolutional neural networks: A case study. *IEEE Transactions on Geoscience and Remote Sensing*, 54(8):4524–4533, 2016.

[196] T. Song, L. Kuang, L. Han, Y. Wang, and Q. H. Liu. Inversion of rough surface parameters from SAR images using simulation-trained convolutional neural networks. *IEEE Geoscience and Remote Sensing Letters*, 15(7):1130–1134, 2018.

[197] J. Zhao, M. Datcu, Z. Zhang, H. Xiong, and W. Yu. Contrastive-regulated CNN in the complex domain: a method to learn physical scattering signatures from flexible PolSAR images. *IEEE Transactions on Geoscience and Remote Sensing*, 57(12):10116–10135, 2019.

[198] C. Grohnfeldt, M. Schmitt, and X. Zhu. A conditional generative adversarial network to fuse SAR and multispectral optical data for cloud removal from sentinel-2 images. In *IGARSS 2018-2018 IEEE International Geoscience and Remote Sensing Symposium*, 2018.

[199] X. X. Zhu, S. Montazeri, M. Ali, Y. Hua, Y. Wang, L. Mou, Y. Shi, F. Xu, and R. Bamler. Deep learning meets SAR. *IEEE Geoscience and Remote Sensing Magazine*, pp(pp):1–26, 2021.

[200] Y. Wang, C. Wang, H. Zhang, Y. Dong, and S. Wei. A SAR dataset of ship detection for deep learning under complex backgrounds. *remote sensing*, 11(7):765, 2019.

[201] T. D. Ross, S. W. Worrell, V. J. Velten, J. C. Mossing, and M. L. Bryant. Standard SAR ATR evaluation experiments using the MSTAR public release data set. In *Algorithms for Synthetic Aperture Radar Imagery V*, volume 3370. International Society for Optics and Photonics, 1998.

[202] J. Shermeyer. SpaceNet 6: Announcing the Winners. `https://medium.com/the-downlinq/spacenet-6-announcing-the-winners-df817712b515/`. Accessed: 2021-05-05.

[203] R. Adams and L. Bischof. Seeded region growing. *IEEE Transactions on pattern analysis and machine intelligence*, 16(6):641–647, 1994.

[204] L. Najman and M. Schmitt. Watershed of a continuous function. *Signal Processing*, 38(1):99–112, 1994.

[205] M. Kass, A. Witkin, and D. Terzopoulos. Snakes: Active contour models. *International journal of computer vision*, 1(4):321–331, 1988.

[206] Y. Boykov, O. Veksler, and R. Zabih. Fast approximate energy minimization via graph cuts. *IEEE Transactions on pattern analysis and machine intelligence*, 23(11):1222–1239, 2001.

[207] N. Plath, M. Toussaint, and S. Nakajima. Multi-class image segmentation using conditional random fields and global classification. In *Proceedings of the 26th Annual International Conference on Machine Learning*, 2009.

[208] 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*, 2015.

[209] 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.

[210] 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*, 2015.

[211] 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.

[212] 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.

[213] L.-C. Chen, G. Papandreou, F. Schroff, and H. Adam. Rethinking atrous convolution for semantic image segmentation. *arXiv preprint arXiv:1706.05587*, 2017.

[214] 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*, 2017.

[215] F. Visin, M. Ciccone, A. Romero, K. Kastner, K. Cho, Y. Bengio, M. Matteucci, and A. Courville. Reseg: A recurrent neural network-based model for semantic segmentation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops*, 2016.

[216] B. Shuai, Z. Zuo, B. Wang, and G. Wang. Dag-recurrent neural networks for scene labeling. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2016.

[217] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, L. Kaiser, and I. Polosukhin. Attention is all you need. *arXiv preprint arXiv:1706.03762*, 2017.

[218] I. J. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville, and Y. Bengio. Generative adversarial networks. *arXiv preprint arXiv:1406.2661*, 2014.

[219] J. Redmon, S. Divvala, R. Girshick, and A. Farhadi. You only look once: Unified, real-time object detection. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2016.

[220] W. Liu, D. Anguelov, D. Erhan, C. Szegedy, S. Reed, C.-Y. Fu, and A. C. Berg. SSD: Single shot multibox detector. In *European conference on computer vision*. Springer, 2016.

[221] R. Girshick. Fast R-CNN. In *Proceedings of the IEEE international conference on computer vision*, 2015.

[222] S. Ren, K. He, R. Girshick, and J. Sun. Faster R-CNN: Towards real-time object detection with region proposal networks. *arXiv preprint arXiv:1506.01497*, 2015.

[223] T. Kong, F. Sun, H. Liu, Y. Jiang, L. Li, and J. Shi. Foveabox: Beyound anchor-based object detection. *IEEE Transactions on Image Processing*, 29:7389–7398, 2020.

[224] H. Law and J. Deng. Cornernet: Detecting objects as paired keypoints. In *Proceedings of the European conference on computer vision (ECCV)*, 2018.

[225] Z. Tian, C. Shen, H. Chen, and T. He. Fcos: Fully convolutional one-stage object detection. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2019.

[226] M. Najibi, M. Rastegari, and L. S. Davis. G-CNN: an iterative grid based object detector. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2016.

[227] Z. Cai and N. Vasconcelos. Cascade R-CNN: Delving into high quality object detection. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2018.

[228] J. Dai, Y. Li, K. He, and J. Sun. R-FCN: Object detection via region-based fully convolutional networks. *arXiv preprint arXiv:1605.06409*, 2016.

[229] K. Chen, J. Pang, J. Wang, Y. Xiong, X. Li, S. Sun, W. Feng, Z. Liu, J. Shi, W. Ouyang, et al. Hybrid task cascade for instance segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2019.

[230] X. Lu, B. Li, Y. Yue, Q. Li, and J. Yan. Grid R-CNN. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2019.

[231] X. Zhou, D. Wang, and P. Krähenbühl. Objects as points. *arXiv preprint arXiv:1904.07850*, 2019.

[232] S. Gidaris and N. Komodakis. Locnet: Improving localization accuracy for object detection. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2016.

[233] J. Wang, W. Zhang, Y. Cao, K. Chen, J. Pang, T. Gong, J. Shi, C. C. Loy, and D. Lin. Side-aware boundary localization for more precise object detection. In *European Conference on Computer Vision*. Springer, 2020.

[234] H. Rezatofighi, N. Tsoi, J. Gwak, A. Sadeghian, I. Reid, and S. Savarese. Generalized intersection over union: A metric and a loss for bounding box regression. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2019.

[235] Z. Zheng, P. Wang, W. Liu, J. Li, R. Ye, and D. Ren. Distance-iou loss: Faster and better learning for bounding box regression. arxiv 2019. *arXiv preprint arXiv:1911.08287*, 2020.

[236] Z. Zhaohui, W. Ping, R. Dongwei, L. Wei, Y. Rongguang, H. Qinghua, and Z. Wangmeng. Enhancing geometric factors in model learning and inference for object detection and instance segmentation. *arXiv:2005.03572*, 2020.

[237] Y. Sun, Y. Hua, L. Mou, and X. X. Zhu. CG-Net: Conditional GIS-aware network for individual building segmentation in VHR SAR images. *IEEE Transactions on Geoscience and Remote Sensing*, 60:1–15, 2022.

[238] Y. Sun, L. Mou, Y. Wang, S. Montazeri, and X. X. Zhu. Large-scale building height retrieval from single SAR imagery based on bounding box regression networks. *ISPRS Journal of Photogrammetry and Remote Sensing*, 184:79–95, 2022.

[239] R. Weibel. Digital terrain modelling. *Geographical Information Systems: Principles and Applications*, pages 269–297, 1991.

[240] S. Katz, A. Tal, and R. Basri. Direct visibility of point sets. *ACM Transactions on Graphics*, 26(3):24, 2007.

[241] J. C. Curlander. Location of spaceborne SAR imagery. *IEEE Transactions on Geoscience and Remote Sensing*, GE-20(3):359–364, 1982.

[242] T. Toutin. Geometric processing of remote sensing images: models, algorithms and methods. *International journal of remote sensing*, 25(10):1893–1924, 2004.

[243] A. Roth, M. Huber, and D. Kosmann. Geocoding of TerraSAR-X data. In *International Congress of the ISPRS*, 2004.

[244] F. R. Gonzalez, N. Adam, A. Parizzi, and R. Brcic. The Integrated Wide Area Processor (IWAP): A processor for wide area persistent scatterer interferometry. In *ESA Living Planet Symposium*, 2013.

[245] 3D BAG by 3D geoinformation research group, TU Delft. `https://3dbag.nl`, note = Accessed: 20-05-2021.

[246] Y. Wang and X. X. Zhu. InSAR forensics: Tracing InSAR scatterers in high resolution optical image. In *Proceedings of FRINGE 2015 workshop*, 2015.

[247] Y. Wang, X. X. Zhu, B. Zeisl, and M. Pollefeys. Fusing meter-resolution 4-D InSAR point clouds and optical images for semantic urban infrastructure monitoring. *IEEE Transactions on Geoscience and Remote Sensing*, 55(1):14–26, 2017.

[248] Y. Sun, S. Montazeri, Y. Wang, and X. X. Zhu. Automatic registration of a single SAR image and GIS building footprints in a large-scale urban area. *ISPRS Journal of Photogrammetry and Remote Sensing*, 170:1–14, 2020.

[249] R. B. Potts. Some generalized order-disorder transformations. In *Mathematical proceedings of the cambridge philosophical society*, 1952.

[250] S. Geman and D. Geman. Stochastic relaxation, gibbs distributions, and the bayesian restoration of images. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, PAMI-6(6):721–741, 1984.

[251] D. Mumford and J. Shah. Optimal approximations by piecewise smooth functions and associated variational problems. *Communications on pure and applied mathematics*, 42(5):577–685, 1989.

[252] M. Storath, A. Weinmann, and M. Unser. Unsupervised texture segmentation using monogenic curvelets and the Potts model. In *International Conference on Image Processing (ICIP)*, 2014.

[253] M. Storath, A. Weinmann, J. Frikel, and M. Unser. Joint image reconstruction and segmentation using the Potts model. *Inverse Problems*, 31(2):025003, 2015.

[254] Y. Chen and G. Medioni. Object modelling by registration of multiple range images. *Image and vision computing*, 10(3):145–155, 1992.

[255] P. J. Besl and N. D. McKay. Method for registration of 3-D shapes. In *Sensor fusion IV: control paradigms and data structures*, 1992.

[256] D. Birant and A. Kut. St-dbscan: An algorithm for clustering spatial–temporal data. *Data & Knowledge Engineering*, 60(1):208–221, 2007.

[257] I. D. Stewart and T. R. Oke. Local climate zones for urban temperature studies. *Bulletin of the American Meteorological Society*, 93(12):1879–1900, 2012.

[258] X. X. Zhu, J. Hu, C. Qiu, Y. Shi, J. Kang, L. Mou, H. Bagheri, M. Haberle, Y. Hua, R. Huang, L. Hughes, H. Li, Y. Sun, G. Zhang, S. Han, M. Schmitt, and Y. Wang. So2sat lcz42: A benchmark data set for the classification of global local climate zones [software and data sets]. *IEEE Geoscience and Remote Sensing Magazine*, 2020.

[259] S. Montazeri, X. X. Zhu, M. Eineder, and R. Bamler. Three-dimensional deformation monitoring of urban infrastructure by tomographic SAR using multitrack TerraSAR-X data stacks. *IEEE Transactions on Geoscience and Remote Sensing*, 54(12):6868–6878, 2016.

[260] F.-M. Adolf and H. Hirschmüller. Meshing and simplification of high resolution urban surface data for UAV path planning. *Journal of Intelligent & Robotic Systems*, 61(1-4):169–180, 2011.

[261] H. Hirschmuller. Stereo processing by semiglobal matching and mutual information. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 30(2):328–341, 2008.

[262] H. Hirschmuller and D. Scharstein. Evaluation of stereo matching costs on images with radiometric differences. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 31(9):1582–1599, 2009.

[263] X. X. Zhu, Y. Wang, S. Gernhardt, and R. Bamler. Tomo-GENESIS: DLR's tomographic SAR processing system. In *Joint Urban Remote Sensing Event (JURSE)*. IEEE, 2013.

[264] G. J.-P. Schumann and P. D. Bates. The need for a high-accuracy, open-access global DEM. *Frontiers in Earth Science*, 6:225, 2018.

[265] K. Simonyan and A. Zisserman. Very deep convolutional networks for large-scale image recognition. In *IEEE International Conference on Learning Representation (ICLR)*, 2015.

[266] A. Krizhevsky, I. Sutskever, and G. E. Hinton. ImageNet classification with deep convolutional neural networks. In *Advances in Neural Information Processing Systems (NIPS)*, 2012.

[267] S. Ioffe and C. Szegedy. Batch normalization: Accelerating deep network training by reducing internal covariate shift. In *International Conference on Machine Learning (ICML)*, 2015.

[268] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei. ImageNet: A large-scale hierarchical image database. In *IEEE International Conference on Computer Vision and Pattern Recognition (CVPR)*, 2009.

[269] T. Dozat. Incorporating Nesterov momentum into Adam. `http://cs229.stanford.edu/proj2015/054_report.pdf`. Online.

[270] K. He, X. Zhang, S. Ren, and J. Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2016.

[271] R. Girshick, J. Donahue, T. Darrell, and J. Malik. Rich feature hierarchies for accurate object detection and semantic segmentation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2014.

[272] S. Zhang, A. Choromanska, and Y. LeCun. Deep learning with elastic averaging sgd. *arXiv preprint arXiv:1412.6651*, 2014.

[273] J. Redmon and A. Farhadi. Yolov3: An incremental improvement. *arXiv preprint arXiv:1804.02767*, 2018.

[274] T.-Y. Lin, P. Goyal, R. Girshick, K. He, and P. Dollár. Focal loss for dense object detection. In *Proceedings of the IEEE international conference on computer vision*, 2017.

[275] K. Chen, J. Wang, J. Pang, Y. Cao, Y. Xiong, X. Li, S. Sun, W. Feng, Z. Liu, J. Xu, Z. Zhang, D. Cheng, C. Zhu, T. Cheng, Q. Zhao, B. Li, X. Lu, R. Zhu, Y. Wu, J. Dai, J. Wang, J. Shi, W. Ouyang, C. C. Loy, and D. Lin. MMDetection: Open mmlab detection toolbox and benchmark. *arXiv preprint arXiv:1906.07155*, 2019.

[276] J. Yang, J. Lu, D. Batra, and D. Parikh. A Faster Pytorch Implementation of Faster R-CNN. *https://github.com/jwyang/faster-rcnn.pytorch*, 2017.

[277] S. Auer, C. Gisinger, and R. Bamler. Characterization of SAR image patterns pertinent to individual façades. In *2012 IEEE International Geoscience and Remote Sensing Symposium*, 2012.