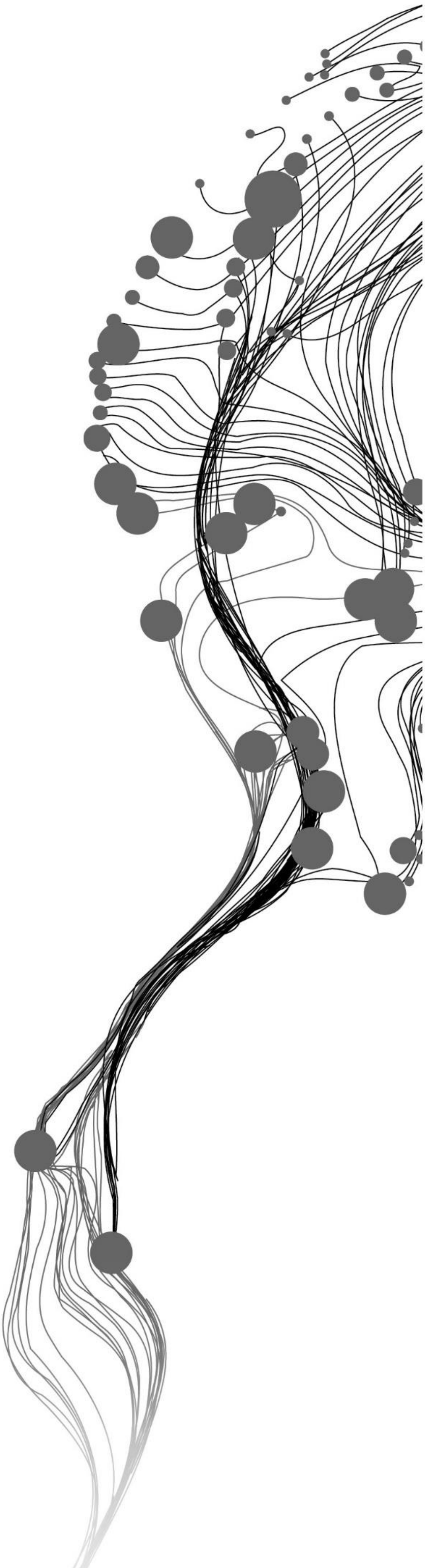
DEEP LEARNING-BASED DIGITAL SURFACE MODEL (DSM) GENERATION USING SAR IMAGE AND BUILDING FOOTPRINT DATA

NESREDIN KENZU ABDELA

August, 2023

SUPERVISORS:

dr. ing. H. Aghababaei
dr. R. Vargas Maretto



Deep Learning-based Digital Surface Model (DSM) Generation Using SAR Image and Building Footprint data.

ABDELA, NESREDIN KENZU

Enschede, The Netherlands, August 2023

Thesis submitted to the Faculty of Geo-Information Science and Earth Observation of the University of Twente in partial fulfilment of the requirements for the degree of Master of Science in Geo-information Science and Earth Observation.

Specialization: Geo-informatics

SUPERVISORS:

dr. ing. H. Aghababaei

dr. R. Vargas Maretto

THESIS ASSESSMENT BOARD:

prof.dr.ir. Francesco Nex (Chair)

dr. Thales Sehn Körting (External Examiner)

dr. L. Chang (Procedural advisor)

DISCLAIMER

This document describes work undertaken as part of a programme of study at the Faculty of Geo-Information Science and Earth Observation of the University of Twente. All views and opinions expressed therein remain the sole responsibility of the author, and do not necessarily represent those of the Faculty.

ABSTRACT

Digital Surface Models (DSMs) are crucial in urban planning, environmental monitoring, and disaster management. Although methods like stereo-photogrammetry, LiDAR, and InSAR have conventionally been used for DSM generation. Recently deep learning has been used to fill the gap between automating tasks and the need for accurate information for real-world applications. Synthetic Aperture Radar (SAR) imagery, with its unique attributes like side-looking geometry, all-weather operability, and day-and-night acquisition capabilities, present a unique opportunity for DSM reconstruction. This study has two main objectives: to assess the feasibility of deep learning for single SAR image-based DSM estimation and to explore enhancing DSM reconstruction accuracy in urban buildings by incorporating building footprint data. RADARSAT-2, TerraSAR-X SAR images, building footprint and ground truth DSM data were used. A fully convolutional neural network architecture with encoders and decoders sub-component is used for DSM estimation. The models are trained and evaluated on both single SAR images and combined SAR-image-building-footprint data using standard regression quality metrics and a structural similarity index (SISM). Results demonstrate deep learning's viability in DSM reconstruction using single SAR images, despite challenges in geometry, such as tilted building appearances and underestimation in shadow regions unseen by the sensor. Integration of building footprint data led to improved accuracy by addressing challenges faced in a single SAR model and produced well-defined building boundaries. Comparative analysis across RADARSAT-2 and TerraSAR-X datasets demonstrates competitive performance across varying spatial resolutions. Additionally, the performances of trained models were evaluated using cross-dataset and PAZ datasets promising inference was shown when using SAR and building footprint data. While accuracy from single SAR image-based predictions was limited, the trained model showed robustness when using both input data and SAR images acquired with similar frequency and looking direction. Future exploration involves adapting the model for diverse SAR datasets and data augmentation. This research demonstrates deep learning's capability in DSM reconstruction within urban settings and introduces integrating building footprint data to enhance estimation accuracy. The implications include urban planning, environmental assessment, and disaster management with a 2-3 meter RMSE accuracy level. Advanced co-registration techniques and diverse datasets are suggested for future work to enhance model performance across diverse land cover types and transferability to other SAR datasets, as a result broadening its applicability.

Keywords: Digital Surface Model (DSM), Synthetic Aperture Radar (SAR), Deep Learning, Remote Sensing, Height Estimation.

ACKNOWLEDGEMENTS

I would like to express my sincere appreciation to my first supervisor, dr. ing. H. Aghababaei, for his firm support and profound guidance in all aspects of the work. His mentorship has illuminated my academic path, and his expertise has been instrumental in shaping the outcome of this research. Equally, I am grateful to my second supervisor, dr. R. Vargas Maretto, whose insightful feedback and guidance have greatly enriched the depth and quality of this work.

I am deeply grateful to the scholarship providers, the Netherlands University Foundation for International Cooperation (NUFFIC) - Orange Knowledge Program (OKP) and ITC, for their invaluable support in making this journey possible. A special mention is reserved for the program manager, drs. J.J. Verplanke (Jeroen), whose kindness and support have given students like me a compass to navigate our academic and professional aspirations in this field.

I am profoundly thankful to my family, በበ (BABA) and ታታ (TATA), whose unconditional love and encouragement continue to motivate me. To my cherished sister, ዘቢበ (Zebiba), who has been more than a sibling - a second mother. A heartfelt thank you goes to my friends, both those studying alongside me abroad and those from back home, whether mentioned or not. Your support, encouragement, and shared experiences have been a continuous source of motivation.

Lastly, I thank Mr Christian Minet and dr. L. Chang (Ling) for providing some of the data used in this research.

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

1. INTRODUCTION

1.1. Background Information and Justification

In the current era of geoinformation, various sensor technologies with different resolutions and modalities have been introduced to capture spatial information about Earth's surface objects and terrains (Cracknell, 2019). With the emergence of large-scale datasets and improvements in computational power, different approaches are being used to extract information from remotely sensed images and have been used to assist in solving real-world problems (Cracknell, 2019). One core task in the remote sensing community is generating a digital surface/terrain model. The extraction of height information, a critical component of digital surface models (DSM) representing the Earth's surface features, including artificial structures and natural surface features, has become a research topic in the field of remote sensing (Li et al., 2014; Rizzoli et al., 2017). One can think of several significant height reconstructions from remotely sensed images to address real-world problems. For example, identifying and modelling 3D buildings and natural features plays a significant role in various applications. Among these, automated 3D building modelling (Arefi & Reinartz, 2013; Karatsiolis et al., 2021; Mahmud et al., 2013), infrastructure damage assessment and disaster preparedness (Karaer et al., 2022; Tu et al., 2016), urban planning and smart cities (Arun & Katiyar, 2013; Mahdianpari et al., 2021), modelling of flooding (Yalcin, 2018), forest height estimation and management (Lim et al., 2003; Stojanova et al., 2010) are some of the applications areas. Such applications require updated and, in some cases, real-time height information. Estimating the height information in remotely sensed images is helpful for better understanding the geometric shapes of features and their relationship within the scene (Rizzoli et al., 2017).

Detailed measurements of height information for DSM are acquired from terrestrial, aerial, and satellite sensors. Terrestrial laser scanning (TLS) and airborne Light Detection and Ranging (LiDAR) have generated the 3D dimension of Earth's surface features (Liu, 2008). TLS and airborne LiDAR uses LiDAR technologies mounted on land and aerial platform, respectively (Liu, 2008; Raber et al., 2007). Laser pulses are emitted from the sensor toward the target objects and the sensor measures the distance from the sensor to the target and provides detailed 3D cloud points of the surface objects, including buildings (Awrangjeb et al., 2010; Lu et al., 2014; Meng et al., 2009) and trees (Brandtberg et al., 2003). These methods produce high-quality, high-density 3D point clouds, representing each data point by its X-Y-Z location in a 3D coordinate system. In practice capturing three-dimensional data of features using these techniques requires high acquisition cost than taking remotely sensed images and generating third-dimension information from it (Priestnall et al., 2000). Furthermore, these datasets have limited availability and less spatial coverage (Lim et al., 2003). Compared to terrestrial and airborne platforms, images acquired from satellite platform provides vast coverage area periodically and repetitively. Authors (Ghamisi & Yokoya, 2018; Karatsiolis et al., 2021; Mahmud et al., 2013; Mou & Zhu, 2018; Zhang, 2005) have shown the feasibility of extracting 3D information on the Earth's surface from satellite images. Han et al. (2020) used information extraction and processing techniques to estimate height from optical data based on multi-view image matching and multi-stereo depth fusion method. The former approach computes correspondences of pixel information from multiple images simultaneously (Zhang and Gruen, 2006). The later technique performs fusions of point clouds detected from the binocular stereo matching of the stereo pairs of the input images (Zbontar and Lecun, 2016).

Researchers have been using datasets from optical sensors and aerial 2D images rich in color cues to reconstruct 3D features (Ghamisi & Yokoya, 2018; Karatsiolis et al., 2021; Mahmud et al., 2013; Mou & Zhu, 2018; L. Zhang, 2005). However, these datasets have limited information to understand better the scene's vertical geometric shape and feature relations (Karatsiolis et al., 2021). Furthermore, these optical images used to extract DSM are only acquired during daytime when solar illumination exists. In addition, the quality of information extracted is affected by the presence of clouds (Pohl & Van Genderen, 1998). Among remote sensing sensors, Synthetic Aperture Radar (SAR) has overcome these limitations due to the capability of radar sensors to record independently of external illumination in both day and night (Moreira et al., 2013). Together with the fact, the SAR sensors use a microwave, which is not impacted by weather conditions and can see through clouds (Fornaro & Pascazio, 2014; Moreira et al., 2013). Moreover, optical remote sensing images are overhead images with a near-nadir view. These images contain information about the building's top/roof and the ground with limited contextual information (Mou & Zhu, 2018). On the other hand, the slant range SAR images add valuable information about building facades, which can help to extract the buildings' absolute height above the ground (Fornaro & Pascazio, 2014). Owing to these advantages of SAR imagery over optical images, SAR datasets have also been used as significant datasets for height estimation of the Earth's surface, e.g., Interferometric SAR (InSAR).

InSAR technique uses the phase information in two SAR images obtained from slightly different positions and determines the relative elevation of each observed object on the Earth’s surface (Fornaro & Pascazio, 2014; Kugler et al., 2014; Sauer et al., 2009). These images can be taken by two antennas mounted on a single platform with some separation distance (single-pass interferometry) or through different passes with slightly different positions (repeat-pass interferometry). The latter case’s temporal shift affects the quality of the result due to phase delay, either a contribution from the atmospheric effect or changes in the target features (Fornaro & Pascazio, 2014).

1.2. Problem Statement

As discussed earlier, it is well understood that DSM can be reconstructed using different methods by combining two or more remotely sensed images. However, several factors motivate using a single image as input in the context of DSM reconstruction from a single SAR image. When processing multi-temporal images, errors can arise during co-registration, affecting the accuracy of the final output if high co-registration accuracy is not maintained (Han et al., 2020). Additionally, changes in target reflectance/scattering properties between images within the multi-temporal acquisition timeframe can impact the reliability of DSM generation (Fornaro & Pascazio, 2014; Mahmud et al., 2013). Temporal decoherence in InSAR images further leads to a loss of coherence and reliability of phase information and height estimation (Kugler et al., 2014). Therefore, errors related to multi-temporal processing, changes in target properties, and temporal decoherence can be minimized by focusing on a single image.

On the other hand, ordinary methods of DSM generation, like stereo photogrammetry, require several processing steps to generate DSM through complex models. The steps required are feature extractions from stereo pair images, correspondence matching, triangulation, and interpolation. These steps aim to capture non-linear relationships and complex interactions in the input remotely sensed images. These techniques still rely on human-engineered features and assumptions of linearity (Zhang & Gruen, 2006). Manually generated features may not accurately capture the underlying relationship between input images and the desired output. The recent advancement of deep learning models, especially convolutional neural networks (CNNs), has shown their potential to learn non-linear and hierarchical features from the input satellite images and map them into DSM (Schmitt & Recla, 2022). Unlike traditional approaches that rely on manual feature engineering and linear assumptions, deep learning models outperform in DSMs estimation by automatically extracting hierarchical features from satellite images and learning non-linear relationships to map the extracted features into the corresponding DSM (Li et al., 2014). Given their outperformance in learning complex features, deep learning methods are chosen for height reconstruction from remotely sensed images in this work.

While there have been numerous studies on deep learning-based height estimation from monocular single optical images, the use of deep learning for reconstructing heights from a single high-resolution intensity SAR data remains relatively unexplored. Moreover, to improve the height estimation accuracy in urban landscapes, incorporating building footprints as an additional input is proposed using a multi-modal fusion network architecture. As the building footprints contain valuable spatial information about the urban environment, such as the shape and layout of buildings. By integrating this information with SAR imagery, the model can learn prior knowledge about the physical characteristics of urban features. In addition, it allows the model to refine the estimation process, enhancing the accuracy of height reconstruction and improving the delineation of building boundaries. The combination of SAR data and building footprints facilitates capturing the connections between building geometry and radar backscattering properties, leading to a more precise height estimation. Thus, this work will assess the potential improvement can be made when building footprint information is used in the process of DSM reconstruction.

1.3. Research Questions And Objectives

1.3.1. Main Objectives

The main objectives of this study focus on DSM reconstruction from a single SAR image using deep learning methods. This study also assesses the performance of FCN methods in DSM generation in terms of absolute accuracy in edges and boundaries of the features by using a combination of SAR image and building footprint information as input.

1.3.1.1. Specific Objectives and research questions

- Implementing FCN for height reconstruction from a single SAR image.
  - What is the performance of the method for height reconstruction when a single SAR image is used?
  - How does the performance of the CNN method vary when applied to RADARSAT-2 SAR images with a 2.5m resolution compared to TerraSAR-X SAR images with a 1m resolution?
- Assessing the performance of deep learning methods (FCN) in height reconstruction from a single SAR image and a combination of SAR images and building footprint.
  - What is the performance of the deep learning method when a single SAR image is used, and to what extent the result can be improved if a combination of SAR images and building footprint information are used?
- Assessing the transferability of the models for the other SAR images acquired with different sensors and acquisition parameters.
  - Does the trained model infer accurately in different SAR images?
  - Can a deep learning model trained on C-band SAR imagery be used to predict heights from X-band SAR data and vice versa?
  - What are the limitations of the work, and what can be done to improve the results?

1.4. Scope of The Work and Contextual Framework

This experimental study explores the process of deep learning-based reconstruction of DSM from SAR images. Recently deep learning has been used to fill the gap between automating tasks and the need for accurate information for real-world applications (Persello et al., 2022). In this work, we used high-resolution SAR images, DSM and building footprint as input datasets and deep learning to automate DSM reconstruction and achieve better accuracy. The main objectives of the research include assessing the performance of CNN-based methodologies in generating DSMs from single SAR images and in combination with building footprint information. Furthermore, the study aims to assess the transferability of trained models to different datasets, particularly across SAR images with varying resolutions, such as RADARSAT-2 and TerraSAR-X, and between C-band and X-band sensors.

SAR images are used to achieve these objectives, including RADARSAT-2 and TerraSAR-X data for training and evaluating the performances of the models. The accuracy of height reconstruction will be evaluated using quantitative metrics and its performance in capturing the edges and boundaries of features. However, due to discrepancies between ground truth height data collected over a year and SAR images caused by changes in land cover, this investigation mainly focuses on built-up areas, where the most negligible changes occurred compared to vegetated areas. Furthermore, this research used relatively lower-resolution datasets to address the challenges of aligning highly detailed side-looking SAR images with ground truth DSM. RADARSAT-2 data, at 2.5x4.5 meters, and TerraSAR-X data, at 1.5x2.4 meters resolution. This is due to challenges when co-registering very high-resolution side-looking SAR images with ground truth DSM.

This research aims to contribute insights into deep learning based DSM reconstruction as a remote sensing application from SAR image. Furthermore, we proposed to use building footprint in the DSM reconstruction process and improve model prediction accuracy obtained from a single SAR image. The experimental findings of this work aim to enhance our understanding of using deep learning algorithms for height reconstruction and provide practical implications for various fields (Figure 1.1).

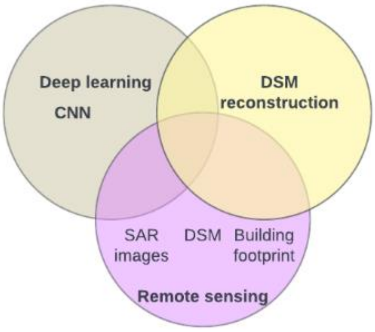


Figure 1.1 Conceptual framework of the study.

1.5. Structure of the Thesis

The first chapter provides an overview of the work, including background information, problem statement, objectives, and research questions. Chapter two, the Literature Review, highlights related works and basic concepts of SAR image and radar wave interaction with urban structures. Chapter three details the data and the methodologies used in the research. The fourth chapter presents the findings and outcomes derived from the analysis and experiment conducted throughout the study. A comprehensive analysis and interpretation, connections to previous research are made, and the implications of the findings are explored. The final chapter summarizes the key findings and answers to research questions and provides recommendations for future research and application.

2. LITERATURE REVIEW

Digital Surface Models (DSM) are essential for various applications, including urban planning, environmental assessment, and disaster preparedness and mitigation. Recent research has explored various approaches, considering diverse data sources and methodologies to reconstruct DSM. These studies have aimed to improve the accuracy and efficiency of DSM generation. By integrating data from different sources like LiDAR, photogrammetry, and remote sensing, researchers have developed advanced methods for DSM reconstruction which is representing the height of Earth's surface features. These advancements have enabled a better understanding and utilization of digital surface information for decision-making processes and spatial analysis. This chapter reviews related works using deep learning for height reconstruction and the geometric and radiometric properties of SAR images.

2.1. Data Sources for DSM Reconstruction

The primary data sources used for DSM generation are LiDAR point clouds and remote sensing imagery (Arefi & Reinartz, 2013). Due to their precise elevation measurements, LiDAR point clouds serve as a direct source for DSM generation. However, LiDAR data availability and affordability pose challenges, along with potential issues such as outdatedness and difficulty in differentiating building boundaries from neighbouring objects (Karatsiolis et al., 2021). Whereas remote sensing imagery, particularly very high-resolution (VHR) aerial and satellite images offer valuable textural and spatial information that can be used for DSM reconstruction. These images provide an indirect source for DSM generation by using techniques such as photogrammetry or stereo-matching to derive elevation information from image pairs (Arefi & Reinartz, 2013).

Additionally, fusing different datasets has been proposed to enhance the accuracy of DSM reconstruction (Zhang & Lin, 2017). This fusion approach involves combining LiDAR data with RS imagery or other ancillary data sources and using each dataset's strengths and compensating for their limitations. However, data fusion introduces challenges related to the registration process, differing spatial resolutions, and simultaneous availability of the datasets (Liu et al., 2020).

This research uses a single SAR imagery for DSM reconstruction using deep learning techniques. The primary objective is to use SAR data as the primary source for generating accurate and detailed DSMs. Then, a deep learning based fusion approach that combines SAR data with building footprints as a complementary data source will be explored. By incorporating building footprints, we will assess the precision and reliability of the generated DSMs.

2.2. State-of-the-art in DSM Reconstruction Using Deep Learning

Along with the rise of the computational performance of computers and the availability of vast datasets, tasks are being automated. With this advancement, there is a parallel desire to economize the task of AI-based navigation and automated information extraction by reducing the number of input data required. Thus, efforts have been made to reduce the minimum input to a single image (Schmitt & Recla, 2022). Deep learning has found applications in various areas of computer vision such as motion detection, pose estimation, and object and face recognition. More recently, deep learning techniques have also been utilized in autonomous driving to estimate the depth of features from side-looking images (Eigen et al., 2014; Ma et al., 2019). This technique has attracted researchers in the remote sensing field and is used for top-viewing remotely sensed images for object detection, land use mapping, change detection, classification tasks, and height estimation (Li et al., 2014). Some of the works that used deep learning for height reconstruction are Mahmud et al. (2020), Karatsiolis et al (2021), Liu et al. (2020), Ghamisi and Yokoya (2018), Li et al (2022), Mou and Zhu, (2018), Recla and Schmitt (2022) and Schmitt and Recla (2022).

Mou and Zhu (2018) proposed the IM2HEIGHT neural network architecture to predict building height from top-viewing high-resolution optical single images. The proposed IM2HEIGHT architecture has two parts namely an encoder and decoder as sub-network architecture. The encoder performs feature extraction and abstraction, encoding the input data into a compact representation. Whereas, the decoder, a symmetrical counterpart of the encoder, generates a height map from the compact representation of the input high-resolution image. The authors experimented with different arrangements in the convolutional layers, including plain convolutional layers, batch normalization, and activation functions in the encoder and decoder components. They also explored the use of ResNet, introduced by He et al (2016) as a backbone of the network architecture. Additionally, they have experimented by using a skip-connection to pass low-level information from the first part of the encoder block to the last decoder block of the model architecture with the ResNet block. Based on their experiments, the network architecture containing ResNet as the convolutional block and the skip-connection between the encoder and the decoder block demonstrated to have better performance in reconstructing DSMs from a single image.

The significance of using both the residual blocks and skip layer in enhancing model performance has been emphasized by researchers. This observation aligns with the findings of Amirkolaee and Arefi (2019) and the work of Liu et al. (2020), who conducted height reconstruction tasks using similar CNN with encoder-decoder architecture. In addition, Liu et al. further improved their model by incorporating data pre-processing and registration techniques. However, these works are done using optical dataset and there is a need to explore the potential of the other modalities of satellite image, which is SAR images.

Ghamisi and Yokoya (2018) introduced the IMG2DSM method, which aims to estimate heights from high-resolution optical images. The IM2DSM approach uses a generative adversarial network architecture consisting of generative and discriminative components. The generative component adopts a U-Net architecture, while the discriminator contains a series of convolutional layers that assess the probability of the generated data likeness with the ground truth DSM. These components operate in an adversarial manner, where the generator generates synthetic height data and the discriminator tries to distinguish between real and the generated synthetic data. Through this process, the generative component learns to produce height data that closely resembles real-world values.

Mahmud et al. (2020) proposed a boundary-aware method for 3D building modelling from overhead optical images. While their main objective was to generate a 3D building model, they also obtained height maps as a byproduct of their work. Their approach involved developing multi-task, multi-feature deep learning frameworks using shared feature representations. These frameworks enabled tasks such as building boundary prediction, ground and building surface recovery, and accurate identification of buildings in aerial images.

The works discussed above focused on using a deep learning method for height reconstruction from VHR single optical images. However, there is a need to explore the potential of SAR data for height estimation. SAR imaging offers unique capabilities, such as cloud penetration and side-looking geometry, which can give an enhanced height estimation.

The first attempt to map a single SAR image to DSM has been recently proposed by Recla and Schmitt (2022). The authors used deep learning for height construction from VHR SAR intensity data based on the modified IM2HEIGHT network architecture, initially developed by Mou and Zhu (2018), to work with the SAR dataset and generate building height from a single SAR intensity image. The authors trained the model with ground truth DSM by projecting it into a side-looking geometry of the SAR image. Their work demonstrated the potential of deep learning for height reconstruction and its applicability to SAR acquisition modes. In addition, a comparison of height reconstruction from a single SAR and a single optical image was done by Schmitt and Recla (2022) using the same model architecture and reported that better performance is shown on a model trained using a single SAR image. Schmitt and Recla (2022) argued that the SAR image's side-looking nature provides more information about the building facades than the roof area. Its complementary shadowing phenomena provide a better height cue for the model to learn and map the building's height.

However, there are only a few studies conducted on single SAR images. There is room to enhance the accuracy by incorporating additional input data to enhance the performance of DSM generation, particularly in complex regions such as building edges and boundaries. Thus, this work will explore a multi-modal fusion deep learning model, which can be trained using a combination of a SAR image and building footprint data to estimate DSMs.

Unlike the work of Recla and Schmitt (2022), this work does not consider projecting the height data into SAR side-looking geometry. Instead, the side-looking SAR images are co-registered and used for model training. Even though, as discussed earlier, the side-looking nature of SAR image brings an advantage for height estimation. It has different geometry from the ground truth DSM and needs to consider their difference during pre-processing (co-registration) properly. Therefore, before going into the details of the methodology, it is essential to highlight the effects arising from the ranging geometry of radar sensors, including layover, shadowing, and double bounce. Subsequently, the next section will provide an overview of SAR data's geometric and radiometric properties, focusing on the interaction between radar waves and urban structures.

2.3. Geometric and Radiometric Properties of SAR Data

The collection of information about the Earth's surface involves various measurement techniques depending on the sensor and its configurations. The SAR, optical and LiDAR sensors view objects with distinct characteristics in terms of geometry and radiometry. Integrating data from multiple modalities can present challenges. However, these sensor configurations also offer valuable opportunities, providing complementary perspectives of the objects under study (Moreira et al., 2013). Understanding their configuration and their drawback is necessary to use data collected by those sensors. When working with SAR imagery for urban scenes, it is important to take into account certain effects due to the specific geometry of radar sensors. Some of the effects, namely layover, shadowing, and double bounce, significantly impact the interpretation and analysis of SAR images. Therefore, understanding these phenomena is essential for accurately interpreting and analyzing SAR images. In this section, a summary of these phenomena is addressed. For more detail, readers are referred to the following works (Flores-Anderson et al., 2019; Fornaro & Pascazio, 2014; Moreira et al., 2013)

A common phenomenon observed in SAR images is a layover, which occurs when tall structures, such as buildings or towers, extend beyond their actual footprint in the image (Flores-Anderson et al., 2019; Moreira et al., 2013). This phenomenon happens because the radar signal can bounce off these structures and reach the sensor at a different angle, resulting in distorted representations of the structure. Adjusting the incidence angle of the sensor to a higher value makes it possible to lower the layover effect. However, an increase in incidence angle increases the shadow effect (Flores-Anderson et al., 2019). This relationship is shown in Eq.(1)and Eq.(2). Another phenomenon in SAR images is the double-bounce effect, which occurs when the radar signal reflects off a structure and then reflects to the ground before reaching the sensor (Flores-Anderson et al., 2019). Double bounce is common in urban scenes, particularly when the radar signal encounters buildings or structures with multiple reflective surfaces. Thus, this lead to misinterpretation in interpreting SAR images with duplicate signals caused by the double-bounce effect (Fornaro & Pascazio, 2014; Moreira et al., 2013). Shadowing is another effect observed in SAR images of urban scenes. It refers to areas in the image where tall structures block or attenuate radar waves, resulting in darker regions (Flores-Anderson et al., 2019). Shadowing occurs when the radar signal is obstructed by buildings, causing less radar energy to reach the sensor in these areas. These unseen regions have an effect, particularly when assessing the brightness or intensity of objects or features (Flores-Anderson et al., 2019).

Figure 2.1 Diagram adapted from Brunner et al. (2010) provides a detailed visual representation of how the radar wave interacts with various parts of urban structures, including the ground, corners, walls, roof, and shadowed areas. The transmitted radar waves interact with simple flat roofs with fixed widths *w* and varying heights *h* of the building above the ground. The building has a flat rooftop with fixed width (*w*), with varying heights (*h*) and is sensed at an incidence angle of *θ*. The line *a* in the diagram shows the reflected radar wave back from the surface. Line *b* shows a double-bounce effect when the transmitted radar wave bounces off at the corner of the building. This creates a double bounce effect when the transmitted radar wave bounces both at the ground surface and the façade of the building. Line *c* demonstrates the transmitted wave reflecting back to the sensor from the wall of the structure facing the sensor. Line *d* shows the radar wave returning from the roof of the building. The shadow effect is illustrated using the area below line *e*, the shadow area where neither the building nor the ground reflects the radar wave, resulting in no return signal. The cases are when $h < w \cdot \tan(\theta)$ (Figure 2.1 a), $h = w \cdot \tan(\theta)$ (Figure 2.1b), and $h > w \cdot \tan(\theta)$ (Figure 2.1 c).

Areas affected by layover (*l*) and shadow (*s*), which are given by Eq.(1) and Eq.(2) are shown in the diagram.

$$l = h * \cot(\theta) \tag{1}$$

$$s = h * \tan(\theta) \tag{2}$$

Based on Eq.(1) and Eq.(2), areas influenced by layover and shadow in SAR imagery significantly relate to the structures' incidence angle and height. When the incidence angle is larger, meaning the radar wave arrives at a steeper angle, it results in a greater potential for layover effects. This is because the radar wave interacts with the vertical surfaces of structures, causing the radar signal to extend beyond the actual ground location. Consequently, structures with taller heights are more susceptible to layover, as the radar wave is more likely to strike their sides, causing the radar signal to extend beyond the actual ground location. On the other hand, low-incidence angles tend to create fewer shadow areas. This is because the radar wave approaches the structures more vertically, causing less shadowing. In contrast, higher-incidence angles (θ) result in larger shadow regions. Thus layover and shadow effects have an inverse relationship, and the incidence angle and height of the feature determine the amount of effect they pose in SAR images.

In addition to the geometric distortions, SAR images are characterized by speckles that arise due to the coherent nature of the radar signal. They manifest as granular noise-like patterns, introducing unwanted variations in the intensity of the image (Flores-Anderson et al., 2019). The speckle effect can be filtered and minimized their effect in the quality of the SAR image.

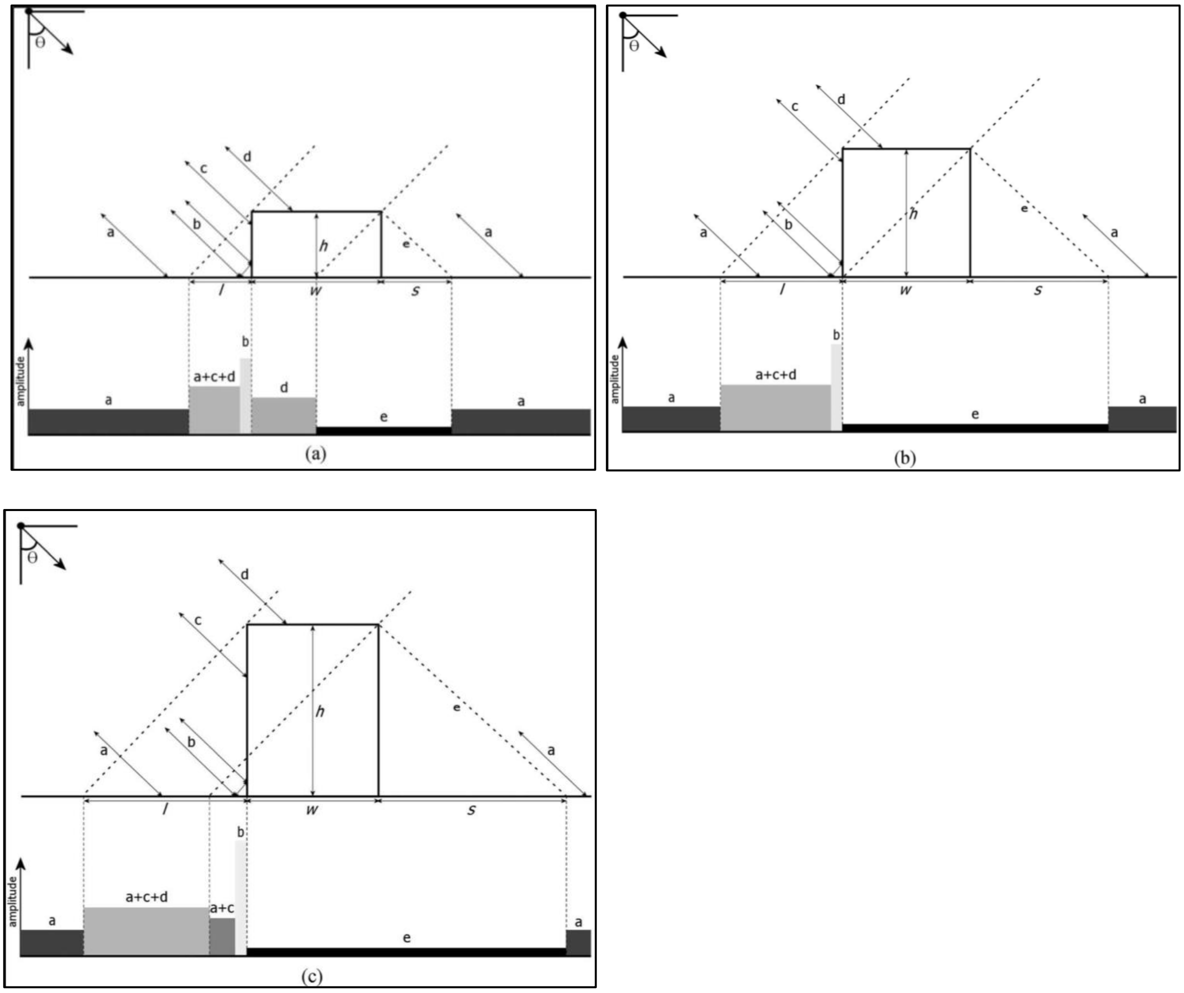


Figure 2.1 Diagram of radar signal interaction with a flat-roof building model. a) when the building height *h* is less than the product of the width *w* and the tangent of the incidence angle *θ*, b) when the building height *h* is equal to the product of the width *w* and the tangent of the incidence angle *θ*, and c) when the building height *h* is greater than the product of the width *w* and the tangent of the incidence angle (Source: Brunner et al. (2010)).

3. MATERIALS AND METHOD

3.1. Description of Study Area

The research is conducted within selected cities in the Netherlands as the study area. The study area is selected based on the availability of a freely accessible ground truth DSM and other geospatial layers for ground truth data and free Radarsat-2 SAR images in the Netherlands. The reference DSM used for model training and evaluation was obtained from airborne LiDAR point clouds. This high-resolution DSM provides accurate information about the terrain and building heights in the study area, serving as a reliable source of ground truth data for model training and validation and evaluation of the model during the reconstruction of DSM. Figure 3.1 shows the location map of the study area.

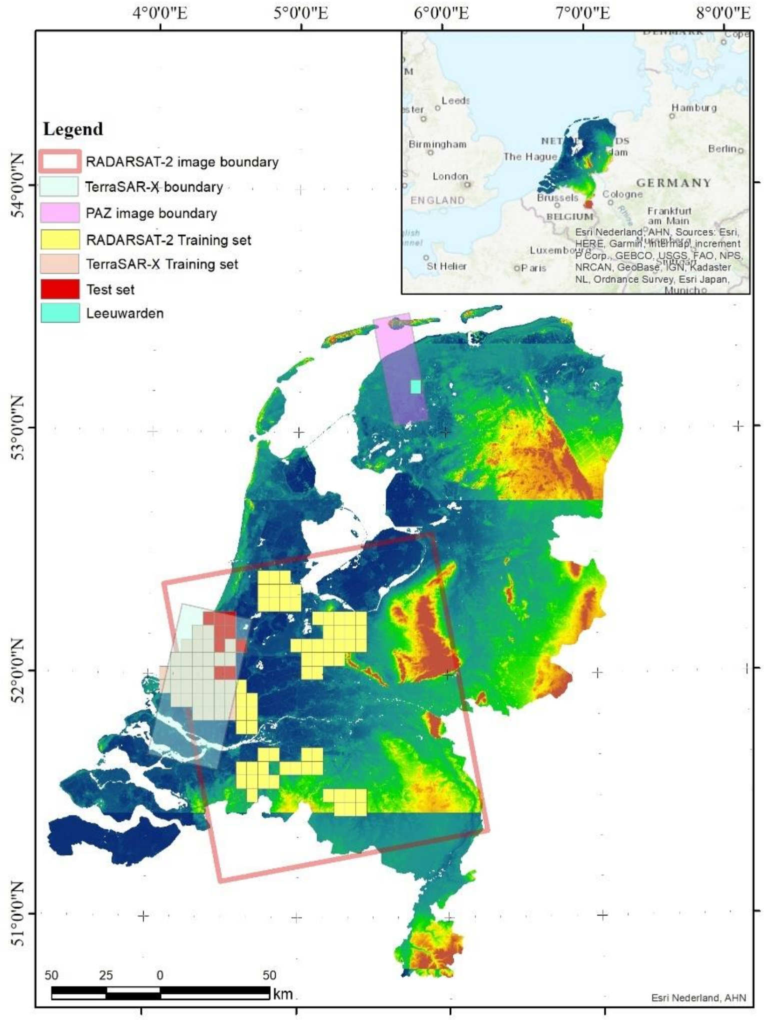


Figure 3.1 Location map of the study area and the footprint of RADARSAT-2, TerraSAR-X, PAZ SAR images and the location of training and testing set.

3.2. Data

This study used RADARSAT-2, TerraSAR-X SAR images and building footprint shapefiles to reconstruct DSMs. Furthermore, additional SAR datasets, such as Sentinel-1 and PAZ, were used to evaluate the transferability of the trained models to data collected from different sensors. A reference height dataset derived from airborne LiDAR point clouds was used for model training and assessment. Further details about the datasets, including their characteristics and sources, are presented in this section.

3.2.1. SAR images

The SAR image data used in this research include RADARSAT-2, TerraSAR-X, Sentinel-1, and PAZ SAR images. Each dataset has unique characteristics and imaging parameters, allowing for a comprehensive analysis and assessment of the deep learning-based DSM reconstruction process. This work used the RADARSAT-2 and TerraSAR-X images for model training, validation, and testing. Whereas the Sentinel-1 and PAZ SAR images are used only to evaluate the transferability of the trained model to other SAR data collected at different sensor and acquisition parameters.

3.2.1.1. TerraSAR-X-Image Products

The other satellite image used for this work is TerraSAR-X images. The TerraSAR-X satellite is designed for Earth observation and operated by the German Aerospace Center (DLR) and Airbus Defence and Space. The TerraSAR-X sensor uses the X-band to gather backscattering of Earth's features. Together with its twin satellite, TanDEM-X, TerraSAR-X contributes to the creation of the WorldDEM, a global Digital Elevation Model (DEM) with uniform resolution. The sensor images the Earth with different imaging modes.

The first imaging mode is spotlight mode, which provides a very high image resolution of up to 1 meter. Even though this imaging mode covers relatively smaller areas, it allows to acquire of very-fine resolution information for detailed image analysis and information extraction. The second imaging mode is the StripMap Mode which offers high-resolution images with a spatial resolution of up to 3 meters. This imaging mode is suitable for detailed imaging of specific areas. The other mode is a ScanSAR that acquire a wider coverage with a swath width of up to 100 kilometers but at the expense of a lower spatial resolution of up to 18 meters (Rizzoli et al., 2017). For this investigation we have made use of a StripMap TerraSAR-X image with ground range spacing of 1[1]X1.7[2] meter. The image has HH polarization and is collected with an incidence angle of ~24^0 and a right-look direction (Figure 3.1, Table 3.1). The TerraSAR-X data was provided by German Aerospace Center (DLR) through the TerraSAR-X Science Proposal RES3721.

[1] range

[2] azimuth

3.2.1.2. RADARSAT-2 image product

RADARSAT-2 is a Canadian Earth observation satellite designed for high-resolution imaging using SAR technology. Similar to its predecessor, RADARSAT-1, it operates at a C-band frequency of 5.405 GHz. The satellite's SAR sensor is capable of transmitting radar waves in both vertical and horizontal orientations, allowing for the generation of products with different polarization modes such as co-polarization (HH or VV), cross-polarization (HV, VH), and even double/quadruple polarization. By using these imaging capabilities, RADARSAT-2 can capture detailed information about various terrains and features on Earth, providing valuable insights (Morena et al., 2004).

The SAR images acquired by RADARSAT-2 are delivered in a complex data format representing high-resolution backscattered intensity measurements. These measurements contain fine spatial details and much information about the Earth's surface. For this research, we used a single Look Complex (SLC) RADARSAT-2 image scene with VV polarization acquired in July and August 2020. The incidence angle of these scenes ranged from 31.530 to 31.570, and the spatial resolution was 2.5x4.9 meters (Table 3.1, Figure 3.1). The acquisition parameters of RADARSAT-2 and Terra-SAR-X are different. We will analyze the effect of these acquisition parameters in model training and test it on other datasets. The acquisition of these images was made possible through the Netherlands Space Office[3].

3.2.1.3. Sentinel – 1 image product

With the launch of the Sentinel-1A and Sentinel-1B satellites as part of the Copernicus program by the European Space Agency (ESA) in 2014 and 2016, respectively, a new era of openly available radar data began. These C-band radar missions provide high-resolution images with a fine spatial-temporal resolution, making them ideal for various applications. The SAR technology used by Sentinel-1 allows for acquiring single and dual polarization images, including VV, VH, HH, and HV, regardless of weather conditions. Furthermore, Sentinel-1 offers a high revisit time, ensuring frequent data acquisition.

This study used Sentinel-1 SAR imagery for evaluating the performance of a deep learning model trained for height prediction using a RADARSAT-2 images. The SAR images acquired from Sentinel-1 satellites, with their high spatial-temporal resolutions and open-source availability, serve as a valuable resource for evaluating the performance of our trained models. By utilizing the Sentinel-1 SAR imagery, we aim to assess the transferability of the trained model to other SAR sensor images and acquisition parameters in the xity of Leeuwarden (Figure 3.1). Table 3.1 gives the characteristics of the used sentinel image.

[3] https://www.spaceoffice.nl

3.2.1.4. PAZ image product

The other SAR image used for model evaluation in this investigation is the PAZ SAR image. As part of the PNOTS (Spanish Earth Observation National Program), the PAZ satellite is an earth observation satellite equipped with a SAR sensor. Its launch in February 2018 marked a significant advancement in radar imaging capabilities within the Copernicus program. With its active phased array antenna technology operating at the X-band frequency range, PAZ can operate in multiple imaging modes, including StripMap (SM), ScanSAR (SC), Spotlight (SL), and High-Resolution Spotlight (HS) modes. Except for ScanSAR, which supports only single polarization, the other modes support different levels of resolution and can be used in single or dual polarization. The PAZ data has a spatial resolution of 2.5X1meter and acquired at an incidence angle of 36^0 with right look-direction (Table 3.1). The data acquired in September 2019 around Leeuwarden City, northern Netherlands (Figure 3.1)

This study uses the PAZ SAR data to evaluate the performance of a deep learning model trained on VV-polarized RADARSAT-2 data and HH TerraSAR-X images. By comparing the model's predictions with the PAZ SAR data acquired in both VV and HH polarizations, we aim to assess its generalization capability and ability to handle different acquisition parameters. This evaluation contributes to a better understanding of the model's performance under different SAR acquisition parameters and suitability for height prediction applications using SAR imagery. The PAZ SAR image is obtained through project number AO-001-030 (with Instituto Nacional De Técnica Aeroespacial (INTA)).

3.2.2. Height Data and Building Information

Two height datasets were used in this work: the first elevation data used in the early pre-processing stage of SAR images and the other is the primary height data used for model training and evaluation. The first height dataset used was the Shuttle Radar Topography Mission (SRTM) data, which has a spatial resolution of 30 meters (1 arc second). This data was used explicitly during the SAR image pre-processing stage, particularly for geometric correction. The SRTM data were seamlessly integrated into the pre-processing workflow by automatically downloading it using the Sentinels Application Platforms (SNAP) tool.

The Actueel Hoogtebestand Nederland-4 (AHN4) DSM, obtained through airborne LiDAR technology, serves as the ground truth dataset for model training, validation, and evaluation. This dataset offers high-resolution height information about surface features and serves as the ground truth in this investigation. The AHN4, collected between 2020 and 2022, is the most recent height dataset available in the study area, with a point density ranging from 10-14 points per square meter and even higher densities of 20-24 points per square meter in certain locations. The Dutch government provides access to this height dataset as a raster image and classified cloud points through the Platform for High-Quality Geodata (PDOK). The AHN4 DSM is highly accurate, with a standard deviation of no more than five centimetres (Table 3.1).

Researchers and practitioners can access these datasets through PDOK to perform a wide range of geospatial analyses. The AHN datasets, known for their high accuracy, provide valuable insights for terrain modelling, infrastructure planning, environmental monitoring, and other scientific research and decision-making process. In this work, we used AHN4 as a basis for model training and evaluation throughout the investigation.

The other dataset is the building footprint which will be used for model training to assess the model performance in the DSM reconstruction process when additional information is added. A file with Geospatial Markup Language (GML) extension representing the building footprints provided by PDOK. This information layer represents the spatial extent and boundaries of buildings in the study area. Furthermore, it provides information on buildings' location, size, and shape, which will be helpful for model training as it gives prior knowledge about urban structures. Thus, we considered integrating building footprint data with SAR images so that the model can benefit from contextual information and improve DSM accuracy.

3.3. Method

The methodology section outlines the step-by-step approach taken to address the research problem. Firstly, the satellite and height data were pre-processed to ensure their quality and suitability for subsequent analysis. Subsequently, the pre-processed data was prepared for input into the deep learning model. This includes creating patches and normalizing the data for deep learning algorithm input. The next aspect discussed in the methodology section is the model architecture, referring to the specific design and structure of the deep learning model used for height reconstruction. Additionally, the evaluation process of the model is described by outlining how its performance and accuracy were assessed. Finally, the experimental setup is detailed, explaining the specific quality metrics and datasets (Figure 3.2).

3.3.1. Data Pre-processing

Despite their numerous advantages, SAR images have intrinsic properties that can affect their data quality and accuracy. These limitations include radiometric distortions, speckles, and geometric distortions (Flores-Anderson et al., 2019). Applying pre-processing techniques to address these limitations and enhance the reliability of SAR-based height reconstruction is essential.

The SAR images underwent radiometric calibration to transform the pixel values into meaningful measurements of backscatter intensity. Gamma-naught (γ0) representation provides the radar signal's amplitude or intensity information, which helps reduce dependence to look angle. Therefore, it is used as the representation after calibration. After calibration of the raw SAR images, we can work with and compare the radar signal strength across different images with consistent interpretation of the backscatter data. Thus, the radiometric calibration was applied to all SAR images using the SNAP tool and the γ0 band was created and used for subsequent pre-processing steps.

Table 3.1 Characteristics of the dataset used in this study.

Data	Acquisition date	Technical characteristics	usage	Source
RADARSAT-2	July 26, 2020 August 02, 2020 August 19, 2020	product type: SLC Polarization: VV Resolution: 2.5x4.9m Incidence angle: 31.5^{0} Looking direction: right PassingDirection: Ascending	Training and Evaluation	[4]
TerraSAR-X	August 04, 2020	product type: SSC StripMap Polarization: HH Ground range spacing: 1x1.7m Incidence angle: ~24^{0} Looking direction: right PassingDirection: Descending	Training and Evaluation	[5]
Sentinel-1	September 03, 2020	Product: IW – GRD Polarization: VH, VV Resolution: 10x10m Incidence angle: 38^{o} Looking direction: right PassingDirection: Descending	Evaluation	[6]
PAZ	September 28, 2019	Product type: SSC - Stripmap Polarization: VV and HH Resolution: 2.5x1m Incidence angle: 36^{o}	Evaluation	[7]
DSMs (AHN4)	2020 - 2022	format: Geotiff Point density: 10-14/m^2	Ground truth	PDOK[8]
BAG Building footprint	December 18, 2015	Format: GML	Training	

[4] https://www.satellietdataportaal.nl

[5] TerraSAR-X data was provided by German Aerospace Center (DLR) through the TerraSAR-X Science Proposal RES3721.

[6] Copernicus Sentinel data 2020. Retrieved from Open Access Hub 2023, processed by ESA.

[7] PAZ SAR image is obtained through project number AO-001-030 (with Instituto Nacional De Técnica Aeroespacial (INTA))

[8] PDOK: https://www.pdok.nl

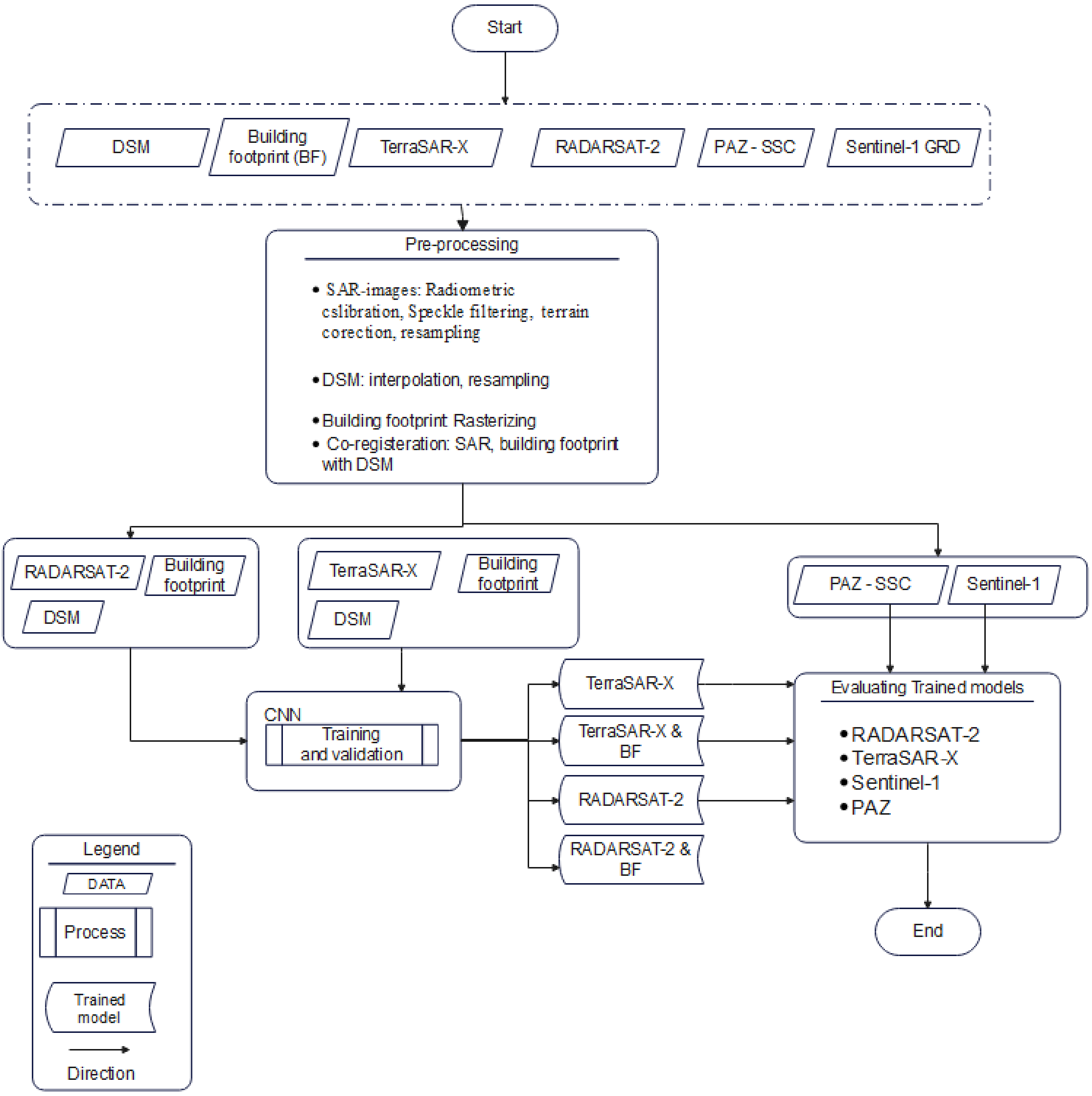


Figure 3.2 Methodological flowchart showing steps taken to achieve the task.

SAR images are commonly affected by speckles, which appear as salt and pepper patterns and degrade the visual quality, making accurate interpretation and analysis challenging (section 2.3). To mitigate the impact of speckle noise, we applied speckle reduction filters as a pre-processing step before extracting information from the SAR images. The objective of speckle filtering is twofold. Firstly, it aims to minimize the intensity variance caused by speckles. Secondly, it tries to preserve the underlying image features and structures while reducing speckle noise. For this purpose, we selected the improved Lee sigma filter, a local statistic-based speckle filtering method. By using window sizes of 9x9 and a target window size of 3x3 to calculate the local statistics within these window sizes, the Lee sigma filter effectively reduces speckle noise, improving the visual quality of the SAR images while preserving essential image details (Lee et al., 2009). The selection of the enhanced Lee Sigma filter was motivated by its demonstrated capability in speckle (Figure 3.3).

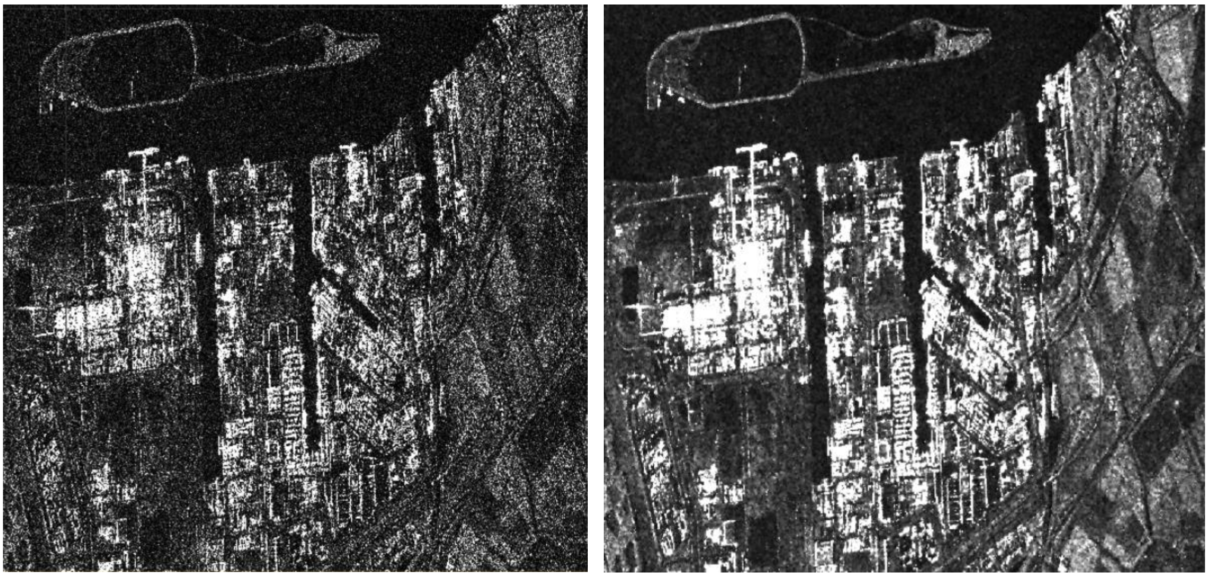


Figure 3.3 Comparison of a cropped section of RADARSAT-2 image before (left) and after (right) applying speckle filtering using Lee sigma. The speckle filtering effectively minimizes the speckle effect, resulting in a visually improved image with reduced speckle artifacts.

The SAR images are initially geocoded to the satellite's position, which assigns geographic coordinates to each pixel based on the satellite's location. However, to align the SAR images with the DSM dataset and enable accurate co-registration, the images need to be reprojected to a real-world geographic coordinate system. In this study, the SAR images were reprojected to the RD New Amersfoort projection system (EPSG:28992), which is a local projection system specific to the Netherlands. This reprojection is key for effective co-registration and alignment between the satellite images and the reference DSM. For this purpose, Copernicus 30 meter resolution global Digital Elevation Model (DEM) was auto-downloaded and used in the SNAP tool. Additionally, during the geometric correction, water bodies are masked out. It is important to note that this study does not consider reconstructing height information for regions covered by water bodies. This pre-processing is applied to all SAR images used in this study.

Pre-processing is applied to the ground truth DSM like the SAR images. The study area has water bodies such as rivers and canals, and there is no height information in areas covered by the water bodies; instead, they were represented as NaN values in the provided DSM. These NaN values are the masked-out equivalent in the SAR images. However, the masked-out NaN values are assigned a zero value to make these values compatible with model training. Furthermore, there were NaN values in some parts of the DSM, like at the edge of buildings, near trees and filled using bilinear interpolation from neighbouring pixels. Subsequently, the DSM is converted into normalized DSM (nDSM) by subtracting DEM from the DSM. In this case, the height values represent the height of the features above ground instead of above mean sea level. In addition to the DSM, The building footprint was provided in a vector format and converted to a raster image to make it compatible with other input data formats.

3.3.2. Co-registering of Input Data

Co-registration is carried out to align SAR images with the DSM and buildings' footprint, ensuring precision in subsequent analysis. In the geometric correction section discussed above, the SAR images have a similar coordinate system with the reference DSM which is necessary for subsequent co-registration. The co-registration task is implemented using the SNAP tool (Figure 3.4). During the co-registration process, the SAR images serve as the master image and the DSM and building footprint as the slave image, ensuring the DSM and building footprint images are aligned with the SAR image. The co-registration process consists of three main steps: creating a stack of the inputs, followed by coarse and cross-correlation, and warping. During stack creation, the SAR and DSM inputs are collocated into a single reference geometry, where the DSM is aligned to the SAR image based on product geolocation. Cross-correlation between the input images is computed using coarse and fine registration by computing image coherence at different window sizes.

During the co-registration of SAR, building footprint and DSM, a first-degree polynomial is chosen for warping considering the flatness of the study area. This polynomial ensures an accurate transformation between the SAR images, DSM, and building footprint. To perform the warping, cubic convolution interpolation is used. Furthermore, the distribution of Ground Control Points (GCPs) used in the warping step is assessed before applying the warping process. Only GCPs with a root mean square error (RMSE) of less than 0.5 are considered for warping, ensuring a precise co-registration process at the sub-pixel level (Figure 3.4). However, there are challenges when co-registering the SAR and DSM/building footprint. The challenge arises due to the geometry of the inputs being co-registered, where the SAR image is a side-looking image containing more information about the wall of the building and there is a layover in tall buildings where the building is extended beyond its footprint. The other issue is shadow areas in the SAR image where there is no information in these areas. In contrast, the DSM and building footprint have overhead geometry providing information about the height and extent of the building in top viewing geometry, respectively. This difference in geometry is challenging to co-register these layers and this issue is addressed by reducing the coregistering inputs into small tiles and applying a co-registration (Figure 6). Tiles with high misregistered are removed as it has an impact during the training of the deep learning model.

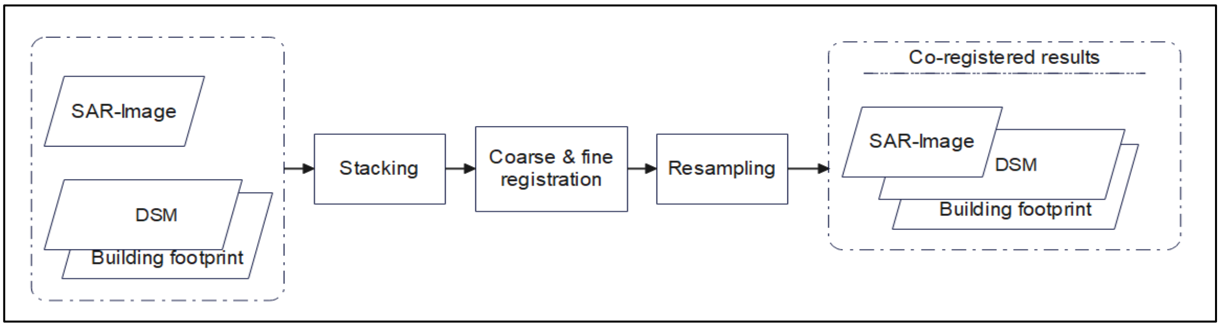


Figure 3.4 A workflow to co-register SAR images with DSM and building footprint.

3.3.3. Data preparation for model input

After co-registering the SAR images with building footprint and reference DSM, the next step is preparing the dataset for the deep learning model, including normalizing the data, creating patches and splitting the inputs into training, validation and testing sets.

A logarithmic transformation was applied to prepare the SAR images for model training by converting the pixel values to the logarithmic scale (dB). This conversion was performed to achieve a distribution that approximates a Gaussian normal distribution, facilitating the subsequent analysis and modelling tasks. The threshold-based approach, where pixel values below -30dB were clipped to -30dB and values exceeding +10dB were clipped to +10dB, was implemented to make the pixel values within a specific range suitable for the subsequent normalization task. Let β_min and β_max represent the 2nd and the 98th percentile of the radar brightness values in the SAR image. If pixel value < β_{min}, then pixel value = β_{min}, If pixel value > β_{max}, then pixel value = β_{max}.

It is worth mentioning that this step does not alter the overall information content of the image. Essentially, the main information contained in most pixels remains unchanged. This process aids in aligning the data with the training requirements of the model and improves the suitability of input for subsequent processing steps, which is input data normalization.

Subsequently, normalization is applied to bring image pixel values to the scale of 0-1, where 0 is the lowest intensity value and 1 is assigned to the maximum intensity value of the SAR image. This step helps for fast model convergence and stability during model training. This step ensures that all features in the image contribute equally to the learning process, preventing dominance by any particular feature or pixel value range. Furthermore, it allows the model to generalize better to new, unseen data, as it becomes less sensitive to variations in absolute pixel values and can adapt to different intensity ranges. To avoid variation of maximum and minimum value from patch to patch, normalization is done at the scene level by taking global maximum and minimum values, and the task is represented mathematically using the following equation:

$$\beta_{\text{normalized}} = \frac{(Pixel_{value} - \beta_{\text{min}})}{(\beta_{\text{max}} - \beta_{\text{min}})} \tag{3}$$

When training the model, we used a patch-based approach to optimize memory usage and effectively capture the local spatial characteristics of height features by dividing the input data into smaller patches with 256 x 256 pixels dimensions. This helps to facilitate efficient data processing and allows the deep learning model to capture complex patterns within the features at a fine-grained level. Patches for TerraSAR-X images are generated with a 50% overlap, ensuring the availability of sufficient training data. In contrast, there is no overlap between patches of the RADARSAT-2 SAR images. Additionally, the patches are further partitioned into separate sets for training and validation. Whereas the test dataset is taken from spatially separated areas, this partitioning ensures proper evaluation and performance assessment of the deep learning model. The training dataset, which consists of a larger portion of patches and the validation set, is used for optimizing and fine-tuning the model's parameters during training.

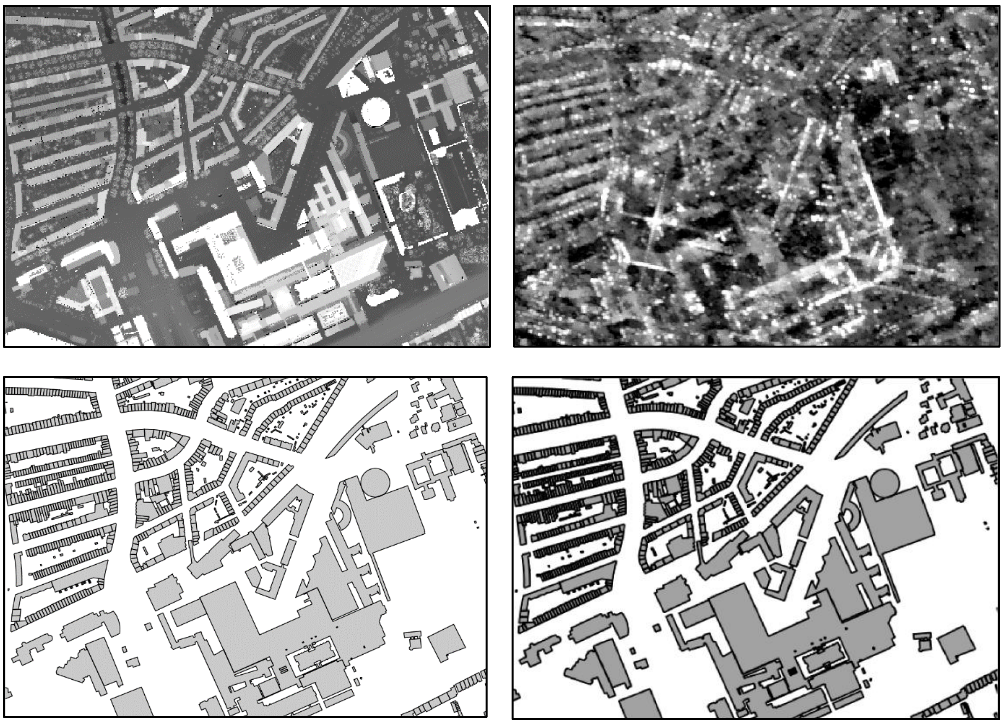


Figure 3.5 Example of co-registered DSM (top right?), SAR (top left) and Building footprint (bottom). The left bottom and right bottom represent a vector and rasterize the building footprints, respectively.

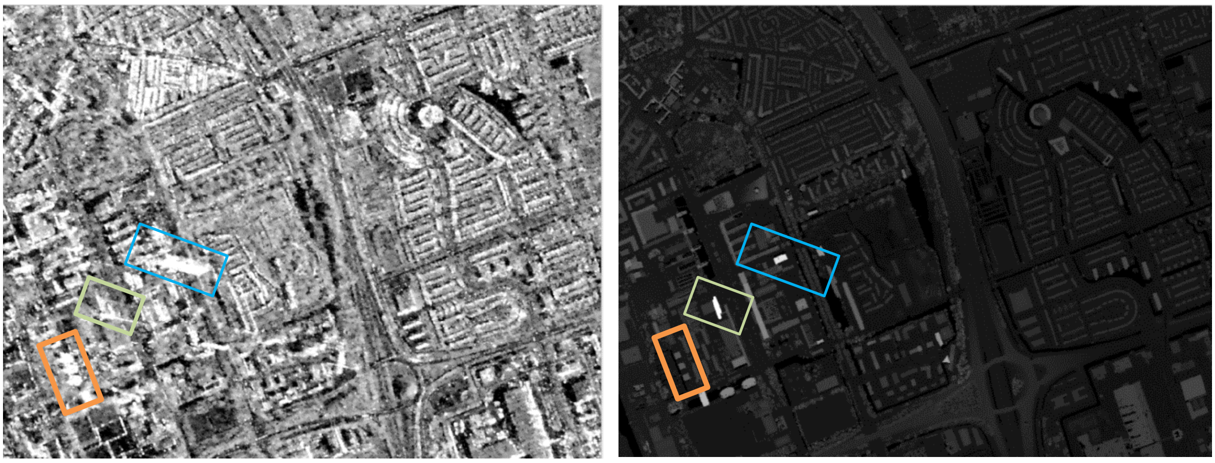


Figure 3.6 Co-registered RADARSAT-2 SAR image (left) and corresponding DSM (right) images. The images also highlight the observed layover effect. The colour boxes indicate areas where the layover effect is visible, particularly tall buildings. In the SAR image, the layover effect is visible as a bright intensity extending beyond the building footprint. In contrast, the corresponding buildings are represented as high elevations in the DSM. For more details, refer to section 2.3 see Figure 2.1.

Table 3.2 Image patches used for training, validation and evaluating the model.

Type	RADARSAT-2	TerraSAR-X	PAZ
Training and Validation	10290	11120	----
Testing	870	1095	356

3.3.4. The Architecture of Deep Learning Method

The architecture used in this study utilizes the power of deep learning to reconstruct heights from remotely sensed images through a fully convolutional neural network (CNN). This network structure is composed of a convolutional layer, max-pooling, and un-pooling/deconvolutional layers, omitting fully connected layers. Such a design allows the network to handle inputs of varying sizes and generate outputs with similar dimensions. The network architecture consists of two main components: the encoder and decoder blocks (Figure 3.7a)).

The encoder block is responsible for encoding the input data. The model takes single-channel image patches for height reconstruction from a single SAR image. The input data is convolved using multiple convolutional layers, each with a specific number of filters. The number of filters starts at 64 in the first convolutional block and doubles in subsequent encoder blocks up to 512, resulting in a richer representation of the input data at the centre with tensor depth of 512. The encoder part of the architecture uses a custom ResNet block, which includes skip-connections to address the issue of vanishing gradients. The skip connection allows the input data to bypass certain layers and be summed with the output, facilitating gradient flow and enabling the creation of deeper networks (He et al., 2016).

The implementation of this custom ResNet architecture for height reconstruction was inspired by the work of He et al. (2016), who introduced the concept of residual learning. This concept has been shown to be effective in training deep networks, as demonstrated by Mou and Zhu (2018) and Recla and Schmitt (2022). Their works have shown that incorporating a ResNet block in the architecture yields superior results compared to using plain convolutional blocks without such a block. In this work, the custom ResNet consists of a sequence of convolutional layers (Conv2D), followed by batch normalization (BN) and activation functions (AF) used. The skip connection in the ResNet used takes the input, convolves it with a 1x1 kernel size convolutional layer, and sums it to the output of the ResNet block (Figure 3.8).

After each encoder block, max-pooling operations are applied to reduce the dimensionality of the input and retain essential features (Boureau et al., 2010). The pooling operation uses a stride size of 2, resulting in a halving of the input dimensions. Initially, the input patch size was 256 by 256 pixels, and after each max-pooling operation, the dimension was reduced by half. At the centre of the model architecture, the tensor size reaches a dimension of 32x32 with 512 depths.

The decoder block, acting as the counterpart to the encoder block, decodes the compressed information and reconstructs it into a DSM with dimensions and resolutions similar to the input image. It begins with the unpooling operation, where the output from the last encoder block is passed through the deconvolutional layer (Conv2DTranspose) with a stride of 2, doubling the size of the tensor. This unpooling operation allows for the recovery of spatial information lost during the max-pooling process (Dumoulin & Visin, 2016).

Following the unpooling operation, the decoder block incorporates a skip connection between symmetrical blocks. This involves concatenating the encoded features from the encoder to the decoder block immediately after the un-pooling layer. This skip connection facilitates the transfer of information across the model bottleneck. By connecting the corresponding feature maps in the encoder and decoder, the skip connection ensures the transfer of high and low-level information (such as edge information) from the convolved inputs SAR image across the model bottleneck. This helps to keep the resolution and enhance the accuracy of the reconstructed height information (Ronneberger et al., 2021).

The decoder block then applies a series of layers, including convolution, batch normalization, and activation functions. These layers further refine the reconstructed features and by combining these operations, the decoder block decodes the compressed information and generates a reconstructed DSM with dimensions and resolution similar to the input image. Unlike the encoder block, the number of filters progressively reduced from 512 down to 64.

In the unpooling operation, the choice of using the deconvolution (Conv2DTranspose) operation for unpooling is motivated by its ability to recover spatial information and reconstruct fine-grained details (Dumoulin and Visin, 2016). Unlike simple unpooling operations that replicate values, Conv2DTranspose incorporates learnable parameters and performs a learned deconvolution. This adaptive nature allows the network to upsample the feature maps effectively, improving the overall reconstruction quality. Additionally, by using a stride of 2x2, Conv2DTranspose doubles the size of the tensor to match the dimensions of the corresponding symmetrical encoder blocks. This uniform expansion preserves spatial resolution and prevents any loss of critical spatial information during the unpooling process (Dumoulin & Visin, 2016).

Finally, the single-channel DSM emerges as the final output, achieved via a convolutional layer featuring a 1x1 kernel size and a linear activation function. This output layer produces a height map output that can be directly compared with ground truth DSMs.

Throughout the convolution operation in both the encoder and decoder, a convolutional filter with a kernel size of 3x3 was applied to the input data. The choice of a 3x3 kernel size is a common practice in convolutional neural networks (CNNs) for image analysis tasks. This kernel size creates a balance between capturing local patterns and fine-grained details while also considering larger global structures. By choosing a smaller kernel size, the model can capture spatial dependencies and avoid excessive smoothing or blurring of the edges of features (Dumoulin & Visin, 2016). The network architecture employed in this work shares a similar structure with the works of Mou and Zhu (2018) and Recla and Schmitt (2022), both of which also used a skip connections between the encoder and decoder blocks to reconstruct the height of buildings from a single channel optical and SAR images, respectively. However, in contrast to their approach, this work implements skip connections across each symmetrical encoder and decoder block. This helps to enhance the information flow and improves the preservation of fine-grained details throughout the height reconstruction process.

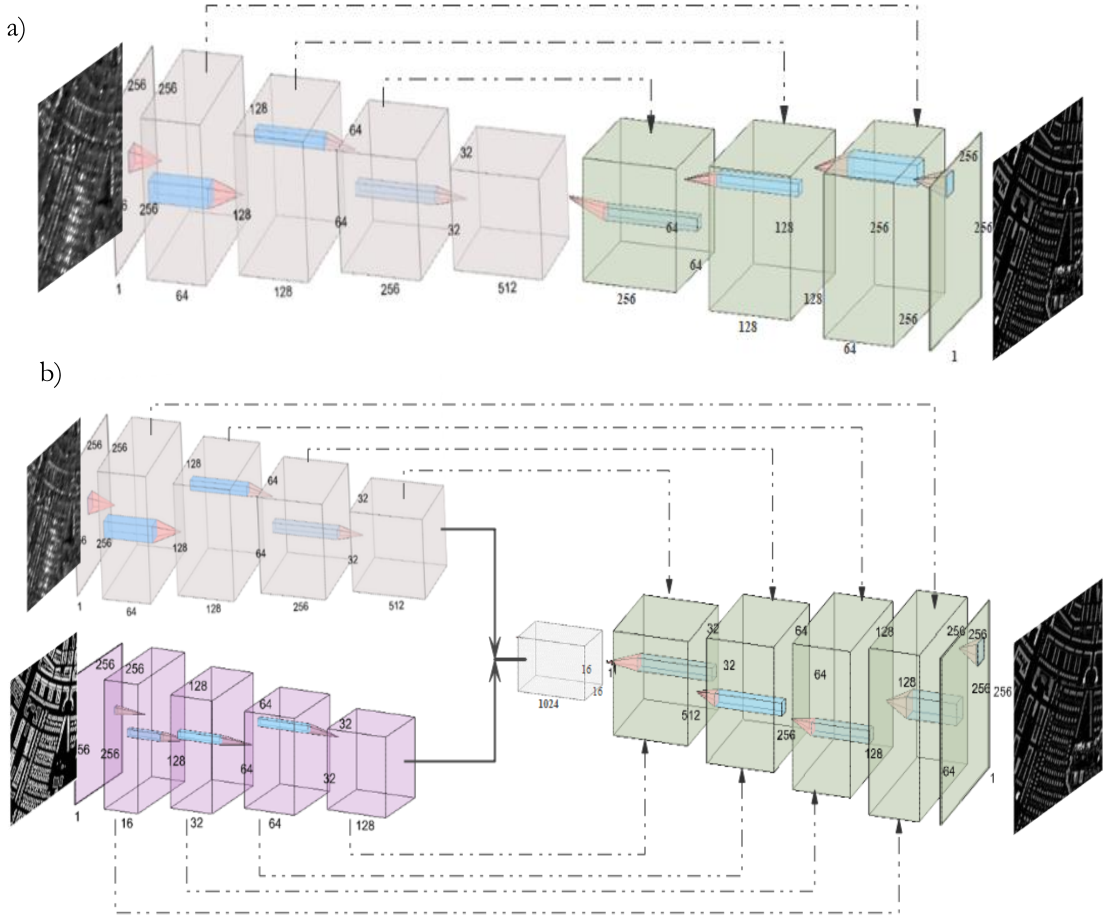


Figure 3.7 Fully Convolutional Network architecture used in this study, where a) for single-channel SAR image and b) for SAR and building footprint, dashed-line represent the skip-connections between the encoder and decoder blocks.

The model architecture for height reconstruction from SAR and building footprint inputs uses a multi-modal fusion strategy, wherein fusion occurs at an intermediate point within the architecture (Li et al., 2022). The SAR and building footprint inputs undergo separate and parallel encoding paths using dedicated encoder blocks (Figure 3.7 b)), as previously discussed. In this framework, greater emphasis is placed on extracting information from SAR images using a higher filter count of 512 at the core of the encoder sequence. Concurrently, the building footprint encoder reaches a filter count of 128 before fusion. This configuration ensures that the model derives a larger portion of its knowledge from SAR imagery. Although it is feasible to provide the model with a dual-channel input of both SAR and building footprint data, the chosen methodology prioritizes extracting information from SAR imagery. We are interested in exploring the potential of SAR images for height reconstruction and using building footprints as supplementary data to enhance the model performance. Thus, we used this fusion method.

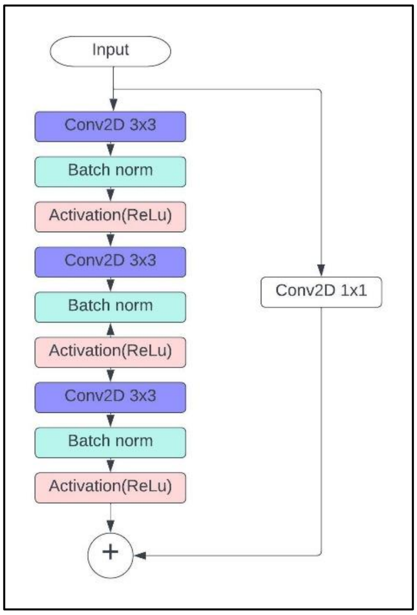


Figure 3.8 Custom ResNet block used in the encoder sub-component of the model.

The subsequent phase involves the fusion of outputs from the SAR and building footprint encoders by concatenating them just before initiating the deconvolution block. This merged encoder output undergoes convolution with 1024 filters. This fusion of information from both SAR and building footprint data helps to enhance the model's accuracy in reconstructing height information by adding additional physical boundaries of the buildings. Following the fusion at the core block, a decoder block, used in the architecture for a single SAR image, as previously discussed, decodes the encoded information. This multi-modal fusion also uses a similar final output layer with a kernel size of 1x1 and linear activation. Thus, we will have a single-channel DSM with dimensions and resolution similar to the input image.

3.3.5. Implementation of Deep Learning Task

3.3.5.1. Training And Model Optimization

The network architectures at hand are ready for training, and the next step is to compile the model with the necessary parameters. The Adam optimizer is chosen to optimize the model, which is justified based on its effectiveness in similar tasks by Mou and Zhu (2018), Ghamisi and Yokoya (2018), and Recla and Schmitt (2022).

Setting the learning rate during model compilation is a key step for model optimization as it determines the magnitude of steps taken towards the minimum loss during training. In this case, a learning rate 0.001 is used for the first few epochs to allow a faster convergence. Subsequently, the learning rate is exponentially decreased throughout training. This approach helps to balance the convergence speed and avoids overshooting the minimum loss. By applying an exponential learning rate decay, the model benefits from a higher initial learning rate and gradually decreases throughout training. Specifically, the learning rate starts at 0.001 for the first few epochs and decreases exponentially to the minimum value. Since ReLu activation is used as an activation function to introduce non-linearity, all weights are initialized based on he_normal kernel initialization.

Two commonly used loss functions for regression tasks are l1 and l2 loss namely Mean Absolute Error (MAE) and Mean Squared Error (MSE), respectively. MSE is the average squared difference between the predicted and true values (eq3), while MAE measures the average absolute difference (eq4).

$$MSE = \frac{1}{N} \sum_{i=0}^{n} (y - \hat{y})^2 \tag{4}$$

$$MAE = \frac{1}{N} \sum_{i=0}^{n} (y - \hat{y}) \tag{5}$$

where N is the number of samples, y represents the true values, and $\hat{y}$ represents the predicted values.

In an experiment using the loss functions in this work, it was observed that the MAE loss function showed fast convergence than MSE. MAE is advantageous when outliers or extreme values are present in the data, as MAE is less sensitive to outliers. Thus, MAE was chosen as the preferred loss function for this task. To moderate model overfitting, a dropout rate of 30% and a weight decay with a coefficient of 0.001 were applied during model training.

3.3.5.2. Model Evaluation and Experimental Setup

This section describes the experimental setup and metrics used to assess the performance of deep learning-based DSM reconstruction model. For this purpose, spatially separated test datasets were used to ensure an unbiased evaluation, avoid spatial auto-correlation, and provide a reliable assessment of the model's performance.

The experimental setup for this work includes several key experiments. The first experiment assesses and compares the model's performance using a single SAR image and a combination of SAR image and building footprint. This comparison allows us to analyse the benefits of incorporating additional information in the reconstruction process. Same comparison will be made by training the model with two different SAR images, namely RDAARSAT-2 and TerraSAR-x SAR images. The other experiment aims to assess the transferability and generalization capability of the trained model using cross-datasets to predict DSMs. Further, the model trained on the RADARSAT-2 and TerraSAR-X dataset and prediction will be made using the Sentinel-1 and PAZ datasets. This allows us to evaluate the model's performance across different SAR sensors and assess its ability to perform well on unseen datasets. Furthermore, a qualitative assessment is conducted to evaluate the model's performance across different land cover types. This assessment provides insights into the model's performance and consistency across different SAR images acquired with different sensors and frequencies and considering various land cover categories (urban areas and vegetation).

For quantitative assessment, several quality metrics commonly used in similar tasks of height reconstruction from remotely sensed images are employed to evaluate the performance of the deep learning models (Amirkolaee and Arefi, 2019; Eigen et al., 2014; Karatsiolis et al., 2021; Mou and Zhu, 2018; Recla and

Schmitt, 2022). These metrics provide measures that quantify the reconstructed height information's accuracy, allowing comparisons with different models and other works. The metrics used for evaluation include RMSE, logarithmic Root Mean Squared Error (logRMSE), relative error, logarithmic relative error (relLog), and Structural Similarity Index (SSIM) (Amirkolaee & Arefi, 2019; Eigen et al., 2014; Karatsiolis et al., 2021; Mou & Zhu, 2018; Recla & Schmitt, 2022).

RMSE is a commonly used metric that measures the average Euclidean distance between the predicted height values and the corresponding ground truth values. It measures the mean squared difference between the predicted and ground truth height information. Let's assume that y, ŷ, n are the ground truth height values, predicted height values and number of pixels, respectively. The RMSE can be calculated using the following equation:

$$\text{RMSE} = \sqrt{\frac{1}{n}\sum_{k=1}^{n}(y_k - \hat{y}_k)^2} \tag{6}$$

where n is the total number of height values, $\hat{y}_k$ y_k represents the predicted and the ground truth height values.

LogRMSE is a variant of RMSE that operates on the logarithm of the height values. It is particularly useful when the range of height values is large, as it accounts for the relative differences rather than the absolute differences. The logRMSE can be calculated as follows:

$$\text{logRMSE} = \sqrt{\frac{1}{n}\sum_{k=1}^{n}|log_{10}(y_k + 1) - log_{10}(\hat{y}_k + 1)|^2} \tag{7}$$

Another metric is the Relative error, which measures the percentage difference between the predicted and ground truth height values. It indicates the magnitude of the errors relative to the true height values. The relative error can be computed using the equation:

$$\text{Relative error} = \sqrt{\frac{1}{n}\sum_{k=1}^{n}\frac{|y_k - \hat{y}_k|}{|y_k| + 1}} \tag{8}$$

Rel_{log}, the logarithmic variant of the relative error, operates on the logarithm of the height values. The Rel_{log} can be calculated as follows:

$$\text{Rel}_{log} = \frac{1}{n}\sum_{k=1}^{n}|log_{10}(y_k + 1) - log_{10}(\hat{y}_k + 1)| \tag{9}$$

To avoid division by zero when calculating the logRMSE, Relative error and Rel_{log} a value of '1' is added and has no effect in the interpretation of the metrics.

SSIM is a similarity metric that assesses the structural similarity between the predicted and ground truth height maps. It considers luminance, contrast, and structural similarity to measure the similarity between two images. The SSIM ranges from -1 to 1, where 1 with perfect similarity, 0 indicates no similarity and -1 indicates perfect dissimilarity. The calculation of SSIM involves comparing local neighbourhoods in the predicted and ground truth images.

$$SSIM = \frac{\left(2\,\mu_{\mathrm{y}}\mu_{\hat{\mathrm{y}}}\ + c_1\right)\left(2\sigma_{\mathrm{y}\hat{\mathrm{y}}},+ c_2\right)}{(\ \mu_y^2 + \mu_{\hat{\mathrm{y}}}^2 + \mathrm{c1})\ (\sigma_y^2 + \sigma_{y\hat{\mathrm{y}}} + c_2)} \tag{10}$$

Where: y and ŷ represent the predicted and ground truth height maps, respectively.

μ_{y} and $\mu_{\hat{\mathrm{y}}}$ denote the mean values of y and ŷ, respectively.

σy and σŷ represent the standard deviations of y and ŷ, respectively.

σyŷ denotes the covariance of y and ŷ

c1 and c2 are small constants used for stability and to avoid division by zero.

3.3.6. Software, Hardware, and Libraries

The implementation of the task was carried out using the Python programming language. Specifically, TensorFlow version 2.9.2 with Keras version 2.9.0 was employed as the primary framework for developing and training the model. TensorFlow and Keras provided a high-level interface for efficiently building and training deep learning models. These libraries offered various pre-defined layers, optimizers, and loss functions, simplifying the implementation process. The model training and evaluation were conducted on the university's computing platform, CRIB (www.crib.utwente.nl). The platform featured an Nvidia GPU, accelerating the computational tasks required to train deep learning models.

4. RESULTS AND DISCUSSION

This chapter presents the results obtained from the conducted experiments and their discussion. The results are structured following the experimental setup discussed in section (3.3.5.2), with each section corresponding to a specific research objective of this study.

In Case-1, the model's performance was evaluated using a single SAR image trained on RADARSAT-2 and tested on the RADARSAT-2 image test set, and the model was trained on a combination of RADARSAT-2 and building footprint data. Case-2 presents the result and discussion of the trained deep learning model using TerraSAR-x and a combination of TerraSAR-x and building footprint data. Then, comparisons of the model's performance and potential limitations are discussed. Finally, the model's transferability to other SAR dataset models trained using the two SAR datasets (i.e. RADARSAT-2 and TerraSAR-x) is evaluated, and their transferability to other SAR datasets with different acquisition parameters is presented. For this purpose, we have used Sentinel-1 and an X-band dual-polarization PAZ dataset. The model's performance on these datasets was measured using standard quality metrics and SSIM. The model performance in all cases is assessed using the quality metrics mentioned in section (3.3.5.2): RMSE, logRMSE, relative error, relLog. Quantitative differences between the predicted and ground truth data were analysed based on these metrics. Additionally, the structural similarity between the predicted and ground truth data was assessed using SSIM, which measures the perceptual similarity of the predicted and ground truth DSM.

The model used for prediction was fine-tuned to optimize its performance for height reconstruction. When testing the trained model on the test dataset, we follow multiple training ad predictions which is five times for each case experiment. The average values of the quality metrics were computed based on the predictions from these models. When analysing the performance, the influence of random fluctuations caused by randomness in model initialization, random weight assignments, and stochastic optimization during model training can be mitigated by training the model multiple times. Thus, the quality metric discussed hereafter is not on a single training and prediction outcome but the average of five training and prediction.

4.1. Performance Evaluation on RADARSAT-2 Dataset

This section presents the results of the model's performance evaluation on the RADARSAT-2 dataset. The primary objective is to assess the model's accuracy in height reconstruction using a single SAR image, followed by improvements made when a combination of SAR and building footprint is used during model training.

4.1.1. RADARSAT-2 Single SAR Image DSM Estimation

This section presents the results obtained from Case-1, which focuses on height reconstruction using a single SAR image trained on the RADARSAT-2 dataset. The primary objective of this experiment is to evaluate the model's performance when using a single SAR image for height estimation. The trained model is assessed using the RADARSAT-2 hold-out test dataset.

The RADARSAT-2 test set consists of 870 SAR image patches covering urban and vegetated regions. The patch size is 256x256 pixels with a resolution of 2.5 meters, similar to training patches. The model's performance on the RADARSAT-2 test set was quantitatively evaluated using the quality metrics shown in Table 4.1 and a visual results are presented in Figure 4.1.

Table 4.1 Quality metrics of a model trained on single RADARSAT-2 and tested on RADARSAT-2 SAR image.

DATA	RMSE	logRMSE (m)	Rel Error	logRel	SSIM (-1:1)
Single SAR	3.19 ± 1.6	0.55 ± 0.2	0.56 ± 0.32	0.14 ± 0.07	0.37
SAR + BF	2.24 ± 1.2	0.53 ± 0.2	0.27 ± 0.18	0.08 ± 0.05	0.63

Table 4.1 shows the model's performance in reconstructing height information from a single SAR image trained using a RADARSAT-2 image and tested on a RADARSAT-2 test set. The predicted height values are compared against the corresponding reference height values. The average RMSE score for the test set was 3.19 ± 1.16 meters, indicating the average squared Euclidean distance between the predicted and ground truth height values. The logRMSE of 0.55 ± 0.2 considers the relative differences in a logarithmic scale, while the relative error of 0.56 ± 0.32 represents the percentage difference between the predicted and ground truth height values. The relLog, a logarithmic variant of the relative error, achieved a value of 0.14 ± 0.07, highlighting its lower value is better. To assess the structural similarity of the predicted height, SSIM was computed, resulting in a moderate similarity score of 0.37 out of a perfect similar value of 1.

Figure 4.1 first row visually illustrates the model's predictions using a single SAR image. This visualization provides insights into the model's ability to capture height information and its performance in different land cover types. The first, second, and third column shows the single SAR image patches used for model prediction, the corresponding ground truth DSM, and the predicted DSM generated by a model trained on a single RADARSAT-2 image, respectively. The visual observation of the results shows the model's ability to capture the height information and produce visually acceptable results. The rightmost column shows the distribution of error and model performance where it is over-estimation in blue and under-estimation in red. The model performs relatively well on buildings than the vegetated areas, as seen in the ground truth height map.

4.1.2. RADARSAT-2 SAR Image and Building Footprint DSM Estimation

The multi-modal fusion model was trained using RADARSAT-2 SAR images and building footprint information as input, aiming to use the additional spatial context provided by the building boundaries. The model was trained using the RADARSAT-2 dataset and corresponding building footprint layers. The model's performance was subsequently evaluated on the same RADARSAT-2 test set, and the results are presented here.

Table 4.1 displays the quantitative evaluation metrics for the model trained on RADARSAT-2 SAR and building boundaries. The same metrics are used for evaluation as discussed in Section (3.3.5.2). The evaluation indicates very promising results, where the model achieved significant improvements over the previous case, where only a single SAR image was used for training. The average RMSE score for the test set was 2.24 ± 1.2 meters, indicating the model's enhanced accuracy in height reconstruction. Additionally, logRMSE of 0.56 ± 0.2 and relative error of 0.27 ± 0.18 demonstrate the model's ability to capture better relative height differences same smaller value of LogRel of 0.08 ± 0.05 were achieved. Remarkable improvement can be observed in the structural similarity (SSIM) score, which scored 0.63, indicating a higher perceptual similarity between the predicted height map and the ground truth.

Figure 4.1 second row presents visual examples of the model's predictions when trained on RADARSAT-2 SAR and building boundary data. The leftmost column shows the SAR image patches used for prediction, followed by the corresponding ground truth DSM and the predicted DSM generated by the model. The figure shows that the model produces more accurate and sharper height estimations. The rightmost column shows the distribution of errors, where over-prediction is highlighted in blue and under-prediction is marked in red. Where in this case less discrepancies is observed in the urban landscape. The visual observations show that the model performs exceptionally well in capturing the height information, especially in the buildings.

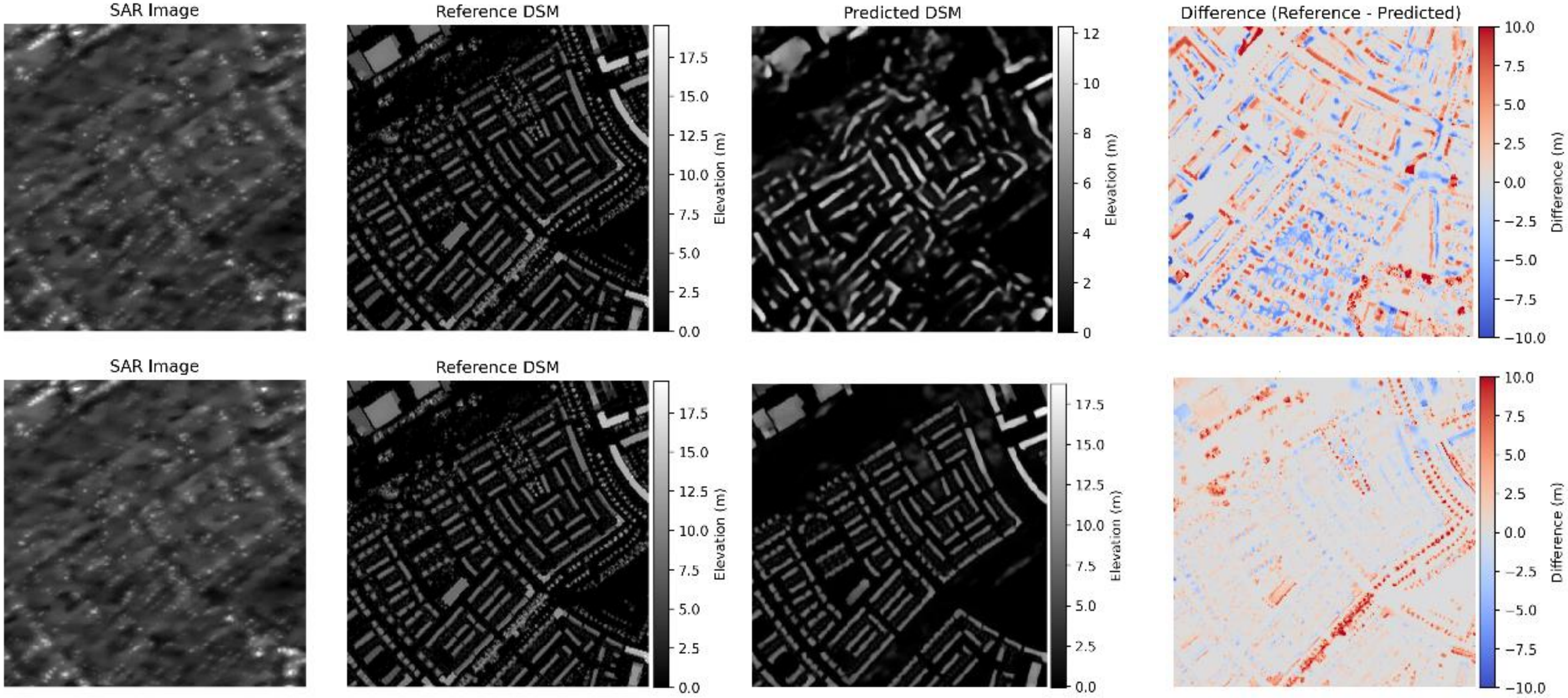


Figure 4.1 Visual example of a result from a model trained using RADARSAT-2 and tested on the RADARSAT-2 test set. The top row represent results from an single RADARSAT image, while the bottom row displays results from a model trained on RADARSAT-2 data and building footprints. Images shown here have 256x256 pixels with a resolution of 2.5 meters. Columns represents SAR image, reference nDSM, predicted nDSM, and the difference between reference and prediction.

4.2. Performance Evaluation on TerraSAR-X Dataset

4.2.1. TerraSAR-X Single SAR image DSM Estimation

This section presents the results obtained using a single SAR image trained on the TerraSAR-X dataset. The TerraSAR-X test set consists of 1095 SAR image patches with a patch size of 256x256 pixels and a resolution of 1 meter, similar to training patches. The model's performance on the TerraSAR-X test set was quantitatively evaluated using the quality metrics shown in Table 4.2 and a visual results are presented in Figure 4.2.

Table 4.2 Quality metrics of the model trained using TerraSAR-X SAR image and building footprint data.

DATA	RMSE	logRMSE (m)	Rel Error	logRel	SSIM (-1:1)
Single SAR	3.1 ± 1.42	0.5 ± 0.2	0.56 ± 0.32	0.19 ± 0.06	0.35
SAR + BF	2.8 ± 1.36	0.47 ± 0.17	0.37 ± 0.32	0.11 ± 0.06	0.6

4.2.2. TerraSAR-X SAR image and Building Footprint Data

A similar TerraSAR-X dataset along with the building footprint dataset were used to train the multi-modal fusion model. The performance of deep learning based height reconstruction from TerraSAR-X image assisted with complementary information from the building footprint dataset assessed qualitatively and presented in Table 4.2. Visual examples are also presented using Figure 4.2.

4.3. Comparative Analysis: Single SAR Image vs. SAR + Building Footprint Data

4.3.1. Quantitative Analysis

This section presents a comparative analysis of the height reconstruction models trained using RADARSAT-2 and TerraSAR-X datasets. The objective is to assess the impact of using building footprint (BF) data with SAR imagery on the accuracy of Digital Surface Model (DSM) reconstruction and further assess the quality metrics across SAR datasets.

Table 4.3 provides an overview of the quality metrics for both datasets under the Single SAR and SAR + BF models scenarios. The results show valuable insights into the models' performance and improvements achieved by incorporating BF data.

For the RADARSAT-2 dataset, using the SAR + BF model showed improvement across all quality metrics. The RMSE was significantly reduced from 3.19 meters in the Single SAR model to 2.24 meters in the SAR + BF model, substantially improving height estimation accuracy. The logarithmic variants, logRMSE and logRel, also showed reductions, indicating a better ability to capture relative height differences. Particularly in urban areas, where building boundaries play a significant role in enhancing height estimation accuracy.

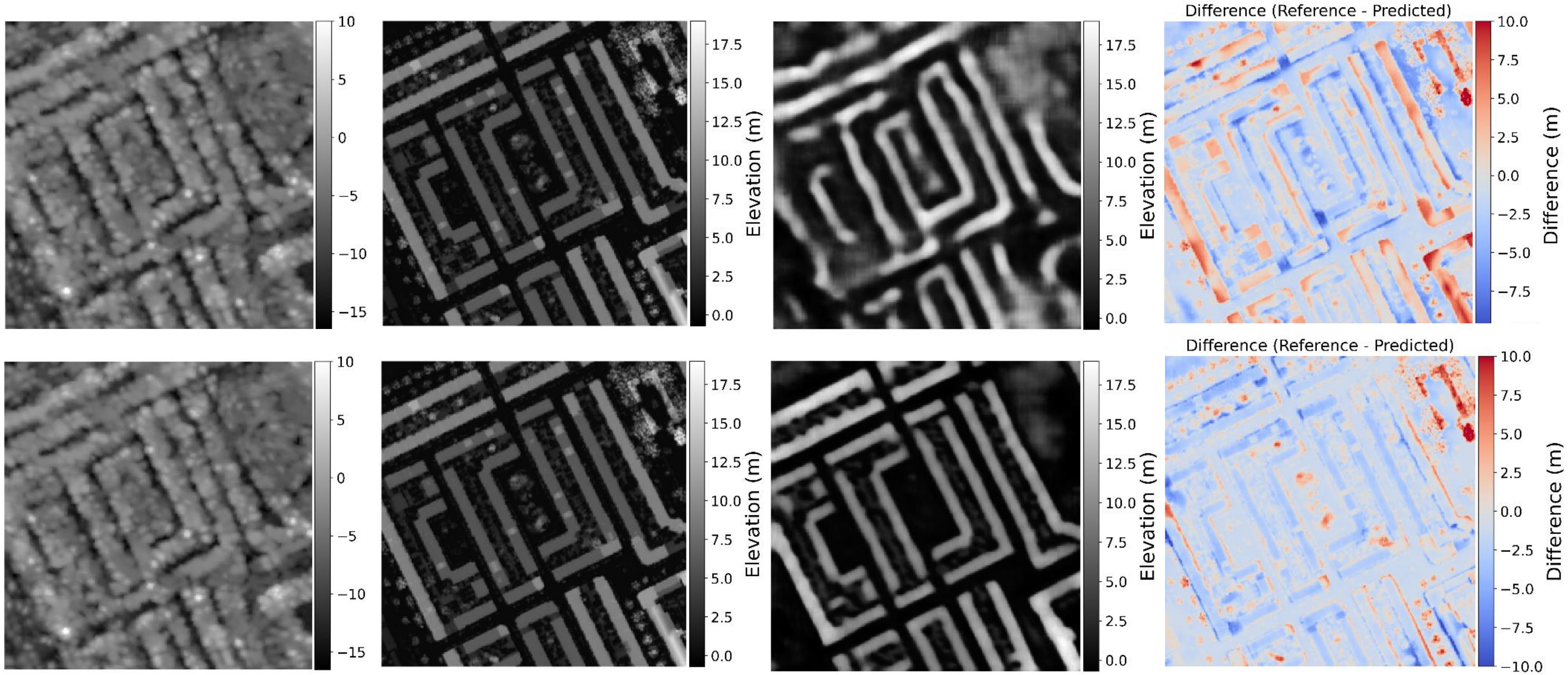


Figure 4.2 Visual example of a model trained on TerraSAR-X and tested on the identical TerraSAR-X dataset. The top row represents results from a single TerraSAR-X image, while the bottom row displays results from a model trained on TerraSAR-X data and building footprints. Images shown here have 256x256 pixels with a resolution of 1 meters. Columns represents SAR image, reference nDSM, predicted nDSM, and the difference between reference and prediction.

For the case of the TerraSAR-X dataset, a similar trend was observed. The SAR + BF model outperformed the Single SAR model in the RMSE, logRMSE, and relative error metrics. The RMSE score decreased from 3.1 meters to 2.21 meters, highlighting the effectiveness of incorporating BF data. Moreover, the logRel metric decreased, showing improvements in capturing relative height differences. The structural similarity index (SSIM) increased, indicating enhanced perceptual similarity between the predicted height map and the ground truth.

Regarding perceptual similarity, both datasets' trained using SAR + BF models showed considerable improvements over their respective Single SAR models. For the RADARSAT-2 dataset, the SSIM increased from 0.37 to 0.63, while for the TerraSAR-X dataset, it increased from 0.35 to 0.6. These SSIM values indicate that the models trained with SAR + BF data better preserve the visual likeness to the ground truth, resulting in more accurate and precise height representations.

Table 4.3 Quality metrics comparison between the two models.

DATA		RMSE	logRMSE	Rel Error	logRel	SSIM(-1:1)
RADARSAT-2	SAR	3.19 ± 1.6	0.55 ± 0.2	0.56 ± 0.32	0.14 ± 0.07	0.37
	SAR + BF	2.24 ± 1.2	0.53 ± 0.2	0.27 ± 0.18	0.08 ± 0.05	0.63
TerraSAR-X	SAR	3.1 ± 1.42	0.5 ± 0.2	0.56 ± 0.32	0.19 ± 0.06	0.35
	SAR + BF	2.21 ± 1.36	0.47 ± 0.17	0.37 ± 0.32	0.11 ± 0.06	0.6

To visualize the impact further, Figure 4.4 provides histograms illustrating the prediction errors for both scenarios. Comparing these histograms shows the error reduction achieved during training when BF information is used. The histograms for the SAR + BF models shows narrower and zero centred distributions, confirming the reduction in prediction errors.

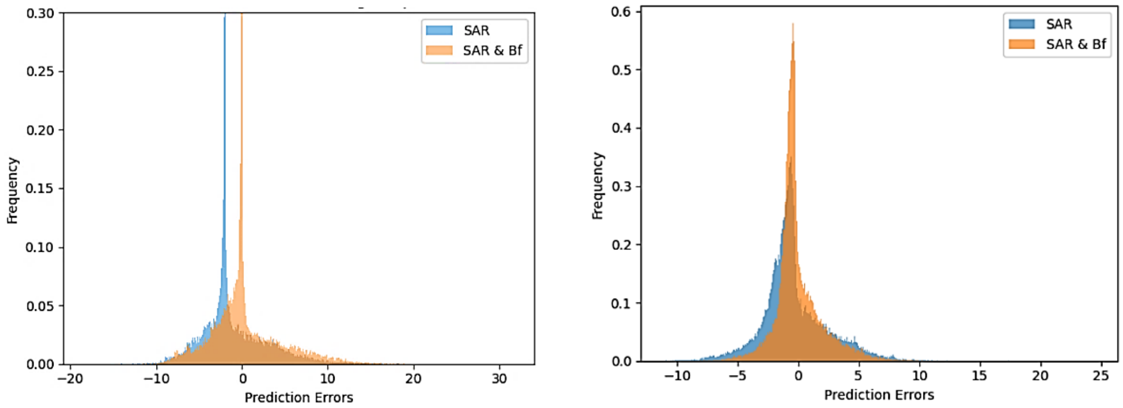


Figure 4.4: Histogram of prediction errors between the Single SAR Image and SAR + Building Boundary models trained using RADARSAT-2 images (left) and TerraSAR-X image (right).

Comparing RADARSAT-2 and TerraSAR-X datasets, Table 4.3 highlights differences in model performance. For the TerraSAR-X dataset, both Single SAR and SAR + BF models achieved relatively better results than RADARSAT-2 models. The RADARSAT-2 SAR + BF model had a slightly higher SSIM value of 0.63 than the TerraSAR-X SAR + BF model at 0.6. However, the differences were subtle, indicating that BF information is effective across both datasets and that there are parallel improvements in quality metrics when BF is used during the training process.

In similar works done by Recla and Schmitt (2022) to reconstruct building height from a single SAR image, authors be able to estimate the height of building in slant-range geometry similar to SAR image and they have achieved comparable accuracy. They have used different resolution TerraSAR-X images and good results were achieved with very high resolution images. Even though the test site is different, with similar resolution images used our work has comparatively better accuracy, especially when a combination of SAR and building footprint data is used. Unlike their work, the result obtained in this work is overhead geometry nDSM. A final remark here is that there is high prediction erros in both case when estimating tall buildings.

4.3.2. Model's Prediction in Different Land Cover Types

The analysis of the model's performance across land cover types showed its strengths and limitations. Experiments conducted using single SAR images from RADARSAT and TerraSAR-X datasets indicate that models trained on a single SAR data achieve better height estimations for vegetated areas than models integrating SAR with building footprint data. Whereas the model trained on SAR and building footprint models performs better in predicting heights in built-up areas. However, this advantage comes at the expense of underestimating heights in vegetated regions. This discrepancy is shown in difference maps where vegetation in the street has high error (Figure 4.1, 4.2). High errors in urban vegetation areas arise from building footprint data treating vegetated regions as zeros, suppressing SAR data's vegetation-related information.

On the other hand, urban buildings benefited from refined predictions due to building footprint assistance in boundary delineation. The outcome highlights a trade-off between urban and vegetated predictions when using building footprint data. Even though it enhances prediction accuracy in buildings, its application to vegetation leads to underestimations due to input data characteristics. Thus, considering specific objectives and land cover types is crucial when applying the model. This underperformance in vegetation areas worsens when the building footprint is used as a second channel during model training, where information is extracted equally from both inputs. Thus, using Multi-fusion reduced the effect of poor performance in vegetated areas. This is because the model that incorporated the building footprint as the second channel to the SAR image tended to extract features equally from both inputs instead of emphasising the vegetation related information extracted from SAR images. Therefore, using multi-modal fusion with different input weights helped to balance results. It is good to note that no observation is made concerning the height due to the topography since the experiment is conducted in a flat landscape.

4.3.3. Observed Effects Due to SAR Image and DSM Geometry in DSM Reconstruction

The SAR image's side-looking perspective presents both advantages and challenges in building height reconstruction. While high-intensity signals in the SAR image effectively capture building facades, the side-looking geometry can cause buildings to appear tilted and extend beyond their actual footprints, affecting accuracy (Figure 3.6). Moreover, the sensor's limitations in capturing shadow areas lead to underpredicted regions, resulting in height discrepancies. This is observed in the difference map of single SAR models (for both RADARSAT-2 and TerraSAR-X), where there is under prediction marked as blue right next to the buildings, initially shadow areas in the SAR image. Overestimation is represented in the blue where there is a layover in the SAR image.

Integration of building footprint data acts as valuable complementary information, bridging the gap between SAR images and DSM geometry. This alleviates the side-looking geometry issues by establishing building footprints as height-transition boundaries. As a result, though over/under-predictions remain, the impact of side-looking geometry in single SAR image predictions is minimized. Visual assessment and quality metrics affirm the enhanced accuracy and reduced discrepancies in height estimations (Figure 4.1, Table 4.1, Figure 4.2, Table 4.2). This approach addresses the unique challenges of SAR imagery's geometry, enhancing prediction accuracy through a simplified representation of building boundaries.

Such effects can also be exaggerated when there is poor co-registration. As discussed earlier in the data pre-processing section (section 3.3.2, Figure 3.6). The experiment is conducted in relatively coarser SAR images. The effect of the difference in geometry is higher and hard to achieve high co-registration accuracy. However, in this work, we tried to use relatively coarser input data to minimize the fine detail observed in the facades of the buildings. The TerraSAR-X image is relatively high resolution than RADARSAT-2 data. However, the TerraSAR-X images has low incidence angle ~24 (near overhead) than RADARSAT-2 ~31. Even though high incidence angle is better and provides rich information about the height of the building. The issue of achieving good co-registration accuracy leads to poor model prediction. As a result, for the high-resolution images, we have used images acquired with low-incidence angle.

In the context of the co-registration accuracy, the effects mentioned can be magnified as previously discussed in the data pre-processing section (section 3.3.2). The experiment was conducted using relatively coarser SAR images, where the disparities between the SAR and DSM geometries are less than in finer-resolution SAR images. However, the challenge of achieving high co-registration accuracy persists. To counteract these challenges, a straightforward approach was taken by using relatively lower-resolution input data. This aims to minimize the observed fine details of building facades which may not be consistently reflected in the DSM. A comparison between TerraSAR-X and RADARSAT-2 shows this fact. Despite TerraSAR-X's higher image resolution, we have used a lower incidence angle (~24 degrees compared to RADARSAT-2's ~31 degrees). While higher incidence angles can yield richer height information, they also present difficulties in achieving accurate co-registration with the co-registration method used in this work.

Thus, there is a need to explore co-registration methods to achieve high co-registration in order to use very-high resolution and high incidence angle SAR images for better height prediction.

4.3.4. Effect of Looking Direction and Building Orientation on DSM Prediction

The evaluation of estimating the heights of buildings through a model trained on a single RADARSAT-2 image demonstrated favourable results. However, there is poor prediction performance mainly due to the influence of both viewing direction and building orientation on the accuracy of height predictions. Figure 4.3 showed that buildings aligned in a north-south direction showed relatively good prediction accuracy than that oriented east-west, with fuzzy boundaries and poor prediction.

The model trained on the right-looking RADARSAT-2 and TerraSAR-X images captures the structural characteristics of urban landscapes, particularly buildings with a north-south alignment. Due to the geometry involved in SAR imaging, buildings in this orientation produce distinct and well-defined backscatter patterns in the SAR image. Buildings aligned in an east-west direction show high brightness levels in the SAR image and form continuous building appearance for a separate building, making it challenging to define building boundaries during the prediction process accurately. Consequently, this led to incorrect estimations of height (Figure 4.3, Figure 4.4). However, training the model using both SAR and building footprint data addressed these challenges related to building orientation. Including building footprint information provided valuable supplementary information and improved the model's understanding of the geometric arrangement of buildings within the scene. Thus, prediction accuracy is improved for buildings with an east-west orientation, which previously had poor prediction accuracy in a model trained only using SAR images (Figure 4.3, Figure 4.4). The proposed approach, which includes the additional context of the buildings, improved the overall result of DSM reconstruction.

As outlined in the data and methodology section, both training and predictions are performed on a patch-based approach. The model processes patches with 256x256 pixels and generates image patches of the same dimensions. However, a post-processing step is required when the objective involves making predictions at a scene level or beyond the initially used patch size. This post-processing phase involves merging each predicted patch and applying smoothing techniques to avoid any potential artefacts that might arise at the boundaries between adjacent patches.

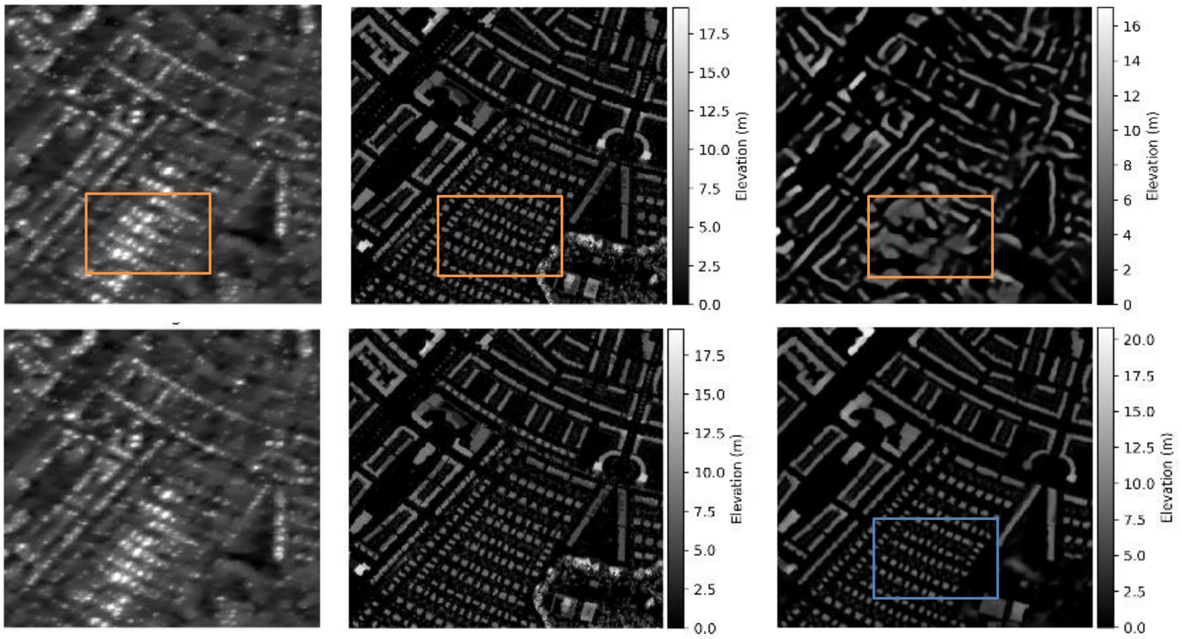


Figure 4.3 Effects of building orientation and looking direction example in the DSM Predictions between a single RADARSAT-2 and SAR + Building Footprint. The top row displays 256X256 pixels with 2.5 meter resolution patches of SAR, ground truth DSM, and predicted DSM using RADARSAT-2 only (the orange box highlights regions of suboptimal performance). The bottom row shows improved results after incorporating building footprints along with SAR image during model training (blue box indicates enhanced predictions).

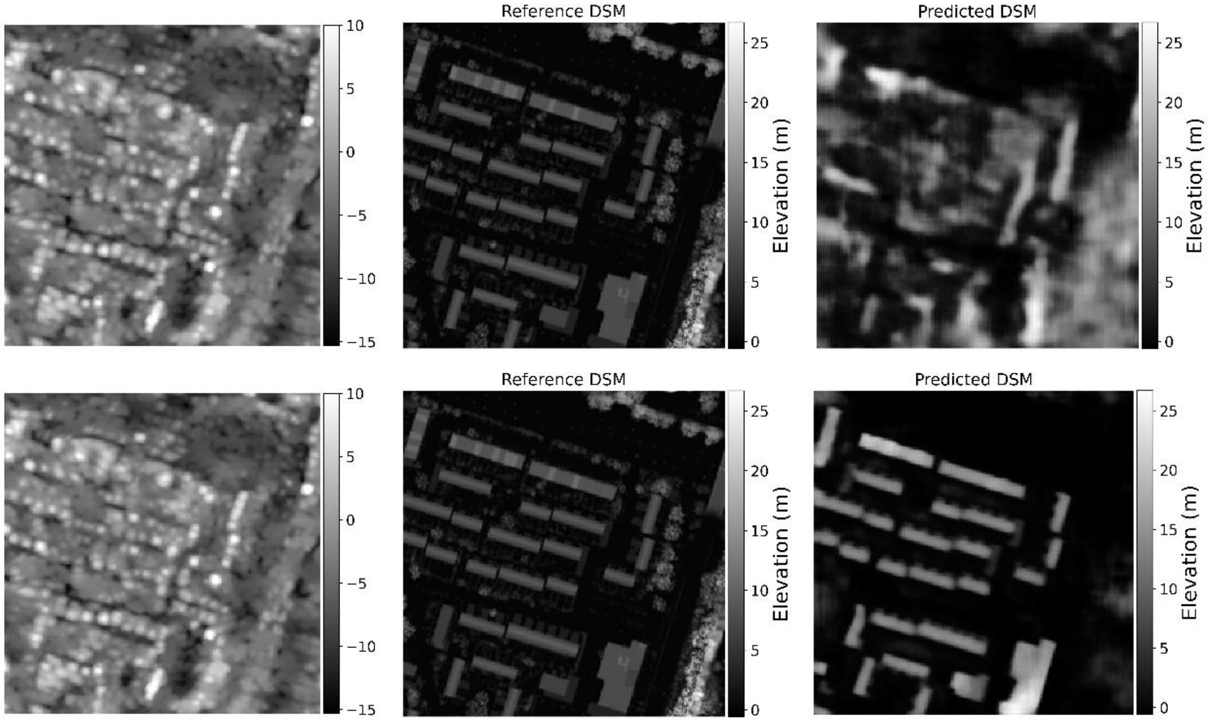


Figure 4.4 Effects of building orientation and looking direction example in the DSM Predictions between a single TerraSAR-X and SAR + Building Footprint. The columns are SAR, ground truth nDSM, and predicted nDSM. The first row is a result of a single TerraSAR-X model and second row shows improved results after incorporating building footprints along with SAR image during model training.

4.4. Evaluation on cross-dataset and other SAR dataset

To address the research question on assessing the trained model's transferability on other SAR datasets acquired using different frequencies and acquisition parameters. We evaluated each model on corss-dataset (i.e. model trained using RADARSAT-2 to TerraSAR-X data and vice-versa) and PAZ and Sentinel-1 dataset. Table 4.4 shows the quality metrics obtained when assessing the models' performances using other SAR images. When predicting DSM using a single SAR image, the models had poor performance and only results obtained using SAR and building footprint are presented here. The quality metric showed relatively less accuracy than when the model is used to predict on the same dataset as the training set. Comparatively, the TerraSAR-X model performs better in predicting other SAR datasets than the model trained on RADARSAT-2.

Table 4.4 Quality metrics of model transferability to other SAR datasets.

Training	Testing	RMSE	logRMSE	Rel Error	logRel	SSIM(-1:1)
RADARSAT-2	TerraSAR-X	4.75 ± 1.85	0.76 ± 0.20	0.72 ± 0.2	0.26 ± 0.07	**0.29**
TerraSAR-X	RADARSAT-2	**3.42 ± 1.69**	0.56 ± 0.22	0.47 ± 0.3	**0.13 ± 0.09**	0.27
	PAZ (VV)	5.03 ± 1.25	0.67 ± 0.08	1.13 ± 0.23	0.42 ± 0.06	0.21
	PAZ (HH)	4.74 ± 1.25	**0.43 ± 0.07**	**0.38 ± 0.07**	0.15 ± 0.02	0.21

Figure 4.5 shows the result obtained when TerraSAR-X is used to evaluate the model trained on RADARSAT-2 and building footprint. The figure shows the result obtained with two different scenarios: (1) the first and third row of Figure 4.5 shows a result of a model trained on the same TerraSAR-X dataset (TerraSAR-X + BF), and (2) the second and fourth row shows a result obtained when a model trained on a different dataset (RADARSAT-2 + BF).

In the first case, the error distribution remains relatively consistent despite some underestimation and overestimation. However, the second and fourth rows show the result when the same TerraSAR-X image is used for prediction using a model trained on RADARSAT-2 and BF dataset. The error map shows unique error distribution and a shift in building estimations. This difference is attributed to the dissimilar acquisition parameters and side-looking geometry between the two datasets. The model's training data, derived from RADARSAT-2 with right-looking and ascending pass direction, contrasts with the descending orientation of the TerraSAR-X image used for prediction. Consequently, building orientations diverge, leading to visible differences in prediction accuracy. This observation highlights the model's sensitivity to varying acquisition parameters. Similar observations are made in Figure 4.6 when the model trained using TerraSAR-X is used to predict the RADARSAT-2 dataset with changes in error direction with respect to the looking direction of the image used to test. This variation aligns with the orientation of the image used for testing, reinforcing the model's reliance on contextual information from shadows for prediction.

During the evaluation of the models using the PAZ dataset, it was observed that only the model trained on TerraSAR-X data, along with building footprint information, could successfully predict the Digital Surface Model (DSM). Furthermore, HH polarization has relatively better accuracy when examining the metric values than VV polarization (Table 4.4). This result can be due to the fact that the PAZ dataset shares similarities in terms of the X-band frequency with the TerraSAR-X data and the acquired polarization of the datasets. However, it is essential to note that the accuracy of the predicted DSM using the PAZ VV polarization was relatively lower, it did not match the accuracy achieved when predicting using TerraSAR-X data (Figure 4.7).

In a study conducted by Recla and Schmitt (2022), the authors explored the transferability of trained models across different imaging modes of the same sensor. Their findings indicated that the trained model showed the ability to estimate DSM with various imaging modes from the same sensor. However, it was observed that the accuracy of predictions achieved using the same dataset remained higher compared to using a model trained on one imaging mode to predict images acquired with a different mode. While this research did not evaluate the generalization capability of trained models using similar SAR sensor images with different imaging modes, similar trends were observed when the trained models made inferences with lower accuracy to images characterized by similar looking direction, pass direction and frequency. Nevertheless, it is noteworthy that the model consistently achieved better quality predictions across all cases when tested on similar datasets.

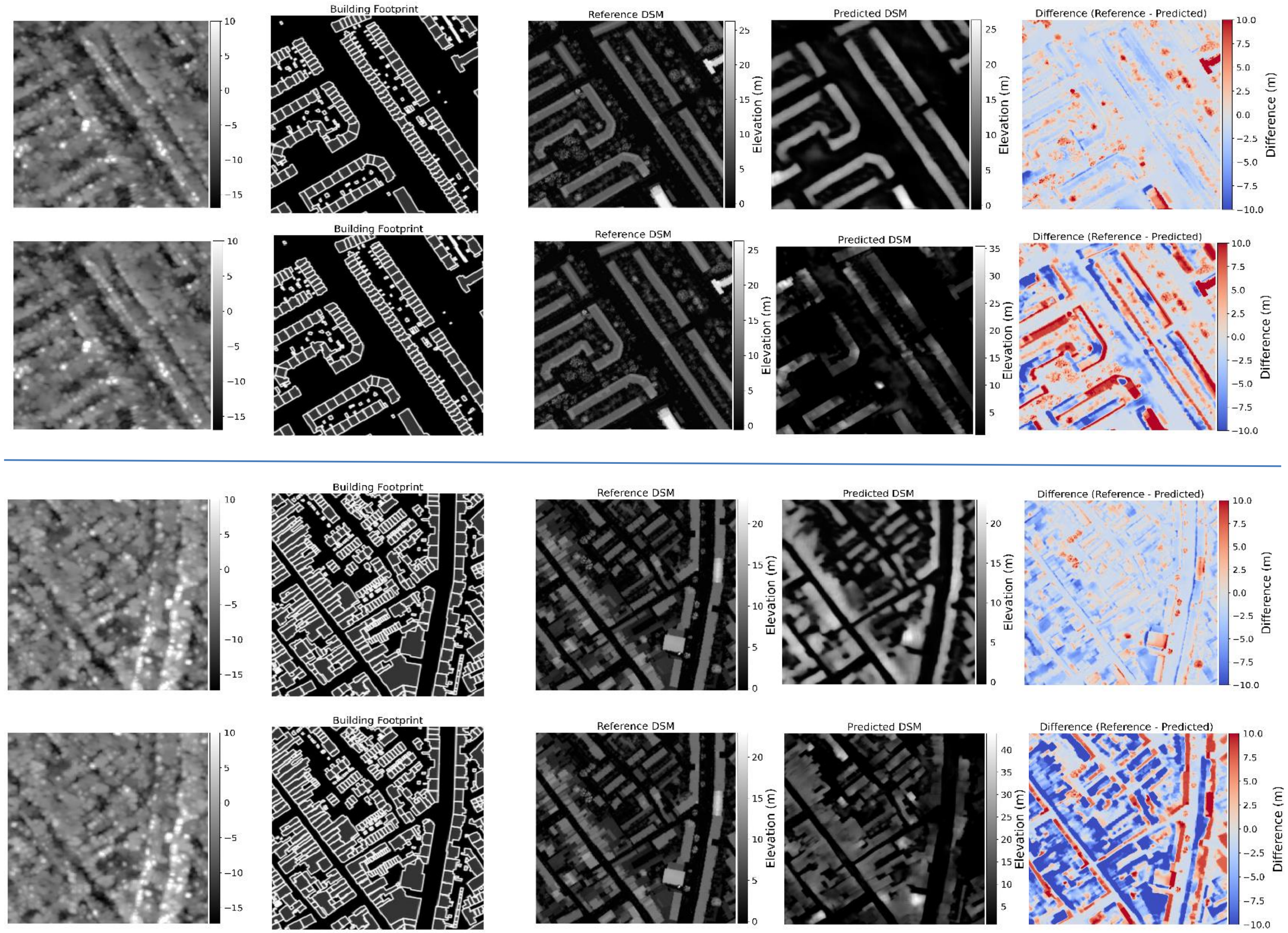


Figure 4.5 Two Example of Prediction results for terraSAR-x images using model trained on RADARSAT-2. For comparison, the first and third row represent result from model trained on terraSAR-X + BF model and second and fourth row shows when RADARSAT-2 + BF model is used to predict using TerraSAR-X.

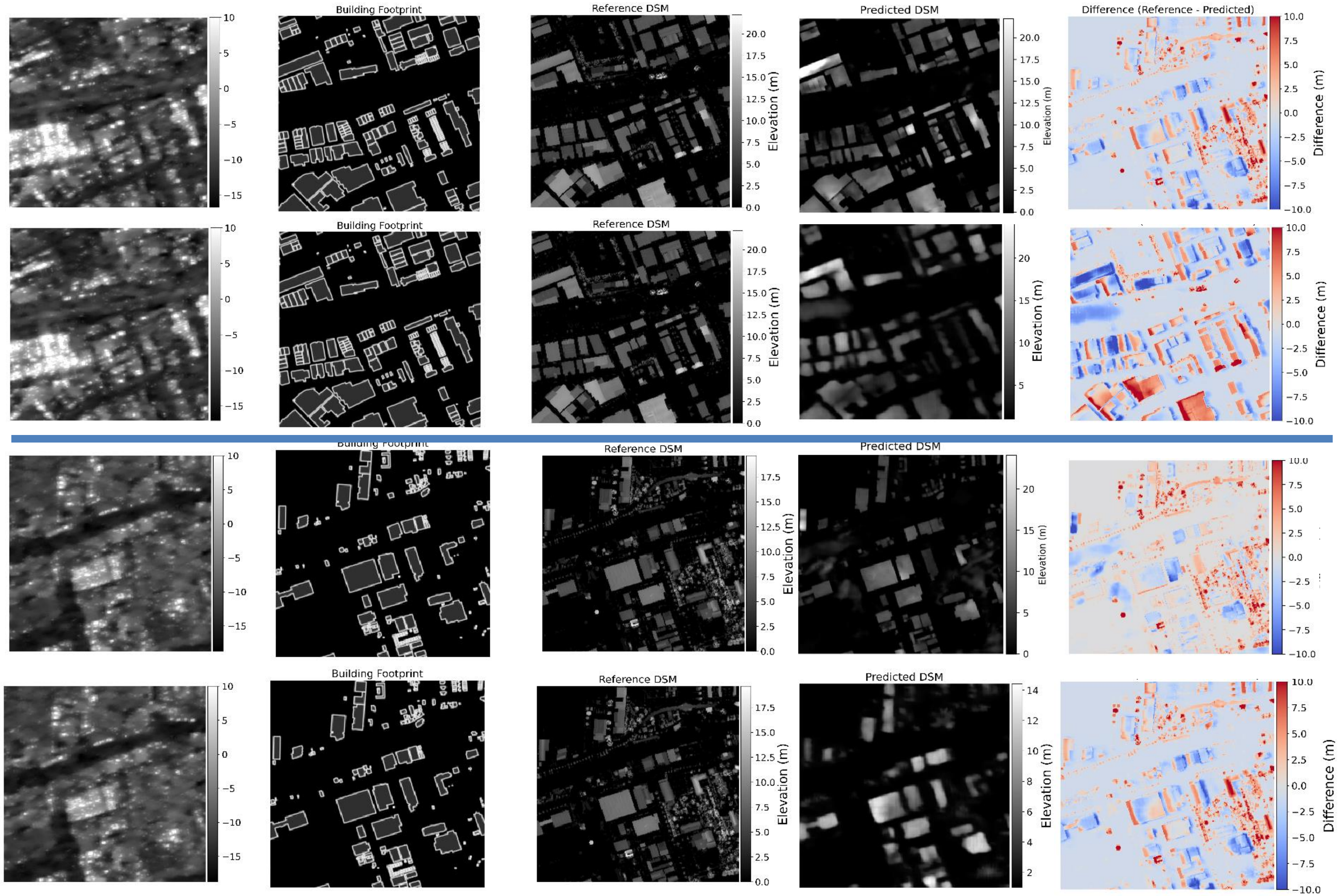


Figure 4.6 Two Example of prediction results for RADARSAT-2 images using model trained on TerraSAR-X +BF. First and third row represent result from model trained on RADARSAT-2 + BF model and second and fourth row shows when TerraSAR-X + BF model is used to predict using RADARSAT-2 image.

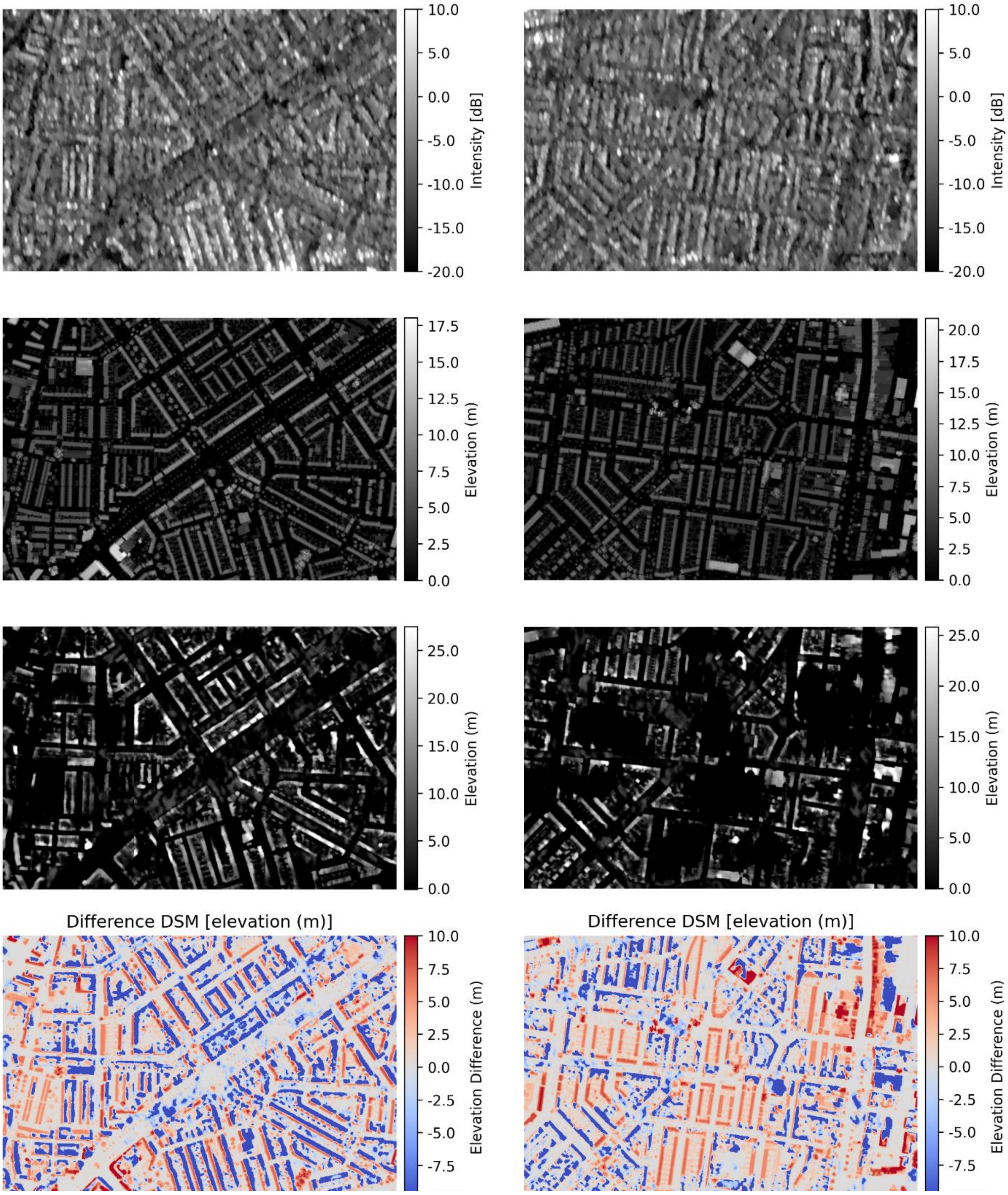


Figure 4.7 Example of a model trained using TerraSAR-X and BF evaluated using PAZ-VV (left) and HH (right) images.

5. CONCLUSION AND RECOMMENDATIONS

The key findings obtained from the experiments and discussions presented in the previous sections will be summarised in this concluding section. Additionally, it provides recommendations based on the observations made and outlines further research directions.

5.1. Conclusion

The main goal of this experimental study was to use a deep learning Method for Digital Surface Model (DSM) reconstruction using a single SAR image and to evaluate the model's performance when incorporating building footprint data during the training process. RADARSAT-2, TerraSAR-X SAR images, and building footprint and ground truth DSM data were used to achieve this objective. The datasets were pre-processed, co-registered, and made ready input for the deep learning model.

For the task of DSM reconstruction from a single SAR image, a fully convolutional neural network having encoder and decoder subcomponents with skip connections across each equivalent encoder and decoder blocks was used. When considering the model taking both SAR and building footprint data, a similar network architecture was adopted, comprising two encoder blocks corresponding to each input, with varying filter numbers to extract high-level information. As an experimental work, this study focussed on assessing the potential of SAR images in reconstructing DSM using a deep learning method, more features were extracted from SAR images and the building footprint data were used as complementary input to enhance the model performance in the urban landscape. For this reason, a multi modal-fusion with two encoders was used and the outputs of the encoders were concatenated before the deconvolutional subcomponent. The trained the performance of the trained model assessed using a standard regression error metrics. The results of this study showed the applicability of deep learning techniques in DSM estimation, highlighting the significant enhancement achieved by incorporating additional information such as building footprints. Thus, this research contributed advancement in the field of deep learning based DSM reconstruction in urban landscapes using SAR images and building footprint information.

5.2. Summary on Research Questions

The first research question aimed to assess the applicability of deep learning techniques in the DSM reconstruction from a single SAR images. This involved training CNNs using single SAR images and evaluating their performance using standard quality metrics. The models trained on single SAR images, particularly from RADARSAT-2 and TerraSAR-X datasets, showed promising results in height estimation. However, challenges emerged due to SAR image geometry and co-registration method used. Buildings often appeared tilted to the sensor, leading to overestimation errors, and underestimation occurred in shadow areas which is not well captured by the models. To assess variation in performance between RADARSAT-

2 and TerraSAR-X dataset a comparative analysis between RADARSAT-2 and TerraSAR-X datasets was carried out to observe potential performance variations from spatial resolutions (2.5 meters and 1 meter, respectively). The results showed competitive performance by CNN models across both datasets. In both SAR datasets poor performance observed in vegetated areas while relatively good results were achieved in urban buildings.

The second research question was assessing the potential of integrating building footprint data in enhancing DSM reconstruction accuracy from SAR images. This required training models with both SAR images and building footprint data and comparing their performance against the previous models. Incorporating building footprint data improved the DSM reconstruction accuracy. The model's ability to define boundaries between building tops and ground surfaces reduced overestimation errors, which were poorly delineated in the models trained using single SAR images. Moreover, adding building boundary information mitigated the challenges posed by SAR image geometry, such as tilted building appearances and underestimation in shadow areas. Furthermore, it gives a crisp boundary and edges in buildings. However, adding this data during model training affects the model's performance in estimating vegetation height, where the single SAR performs relatively better in vegetated areas.

The third research question focuses on evaluating the transferability of the trained deep-learning models across different datasets with different acquisition parameters. When using the model trained using a single SAR image, no observation is achieved, including the coarse resolution sentinel-1 image. However results were achieved when using models trained on Sar and building footprint data. The models show distinct error distribution highlighting the model's reliance on the training SAR acquisition parameters, including looking direction and the shadow regions.

5.3. Recommendations

Throughout this study, we have explored the potential of deep learning and SAR data in DSM reconstruction. It is worth noting that limitations are observed and possible recommendations are given here. The first point is the co-registration of the SAR and ground truth DSM. The quality of co-registration results has been a challenging task. While efforts were made to mitigate mis-registration issues by using relatively coarser images and images acquired using low incidence angle, the use of high incidence angle and very high-resolution data might yield good results, as the performance of with TerraSAR-X data evidences it. Thus, we can potentially minimize mis-registration errors and achieve better alignment by investigating advanced techniques, such as deep learning-based methods, for more accurate height predictions.

We explored and showed the significance of additional spatial information during the training of deep learning. When integrating complementary data for better performance in DSM estimation, it is essential to align the study's objectives and the effects of the data used to achieve the objective. This study shows that incorporating building footprint enhances DSM estimation accuracy for urban buildings while undermining the model performance on vegetated areas. Thus, aligning the objectives and strength of the added data in achieving the objective without affecting model performance in other land cover types is necessary. Another point is that there is a trade-off between smoothing and loss of information when applying preprocessing steps such as speckle filtering. Using speckle filters can reduce the speckle effect but might also result in the smoothing of fine details. Therefore, it is important to consider the option of assessing the method without speckle reduction techniques. This would be interesting to evaluate how the model performs with raw SAR data.

As discussed earlier in assessing the model transferability to other SAR datasets, the models must be trained on diverse SAR images acquired with different acquisition parameters to enhance its robustness and generalization capabilities. Other possibilities, like diverse training data through data augmentation, can be used to enhance the generalization capabilities of the model.

The contribution of this work is notable, considering the limited number of studies that have explored deep learning for DSM reconstruction, especially in the context of integrating building footprint data along with a single SAR image. As far as our knowledge, this study is the first to show the use of building footprint data to enhance DSM accuracy in urban areas from a single SAR image. Even though the achieved accuracy is good and comparable with the works of Recla and Schmitt (2022) and Amirkolaee and Arefi (2019), it is important to acknowledge the limitations of the current approach. The accuracy achieved in this study may not be suitable for applications that require sub-meter or even up to 2-3 meter level accuracy. Therefore, the practical implications of this research extend to where the achieved accuracy aligns with the requirements of specific applications, particularly those that do not require very high precision.

Declaration: The PAZ SAR image is obtained through project number AO-001-030 (with Instituto Nacional De Técnica Aeroespacial (INTA)). Moreover, the TerraSAR-X data was provided by German Aerospace Center (DLR) through the TerraSAR-X Science Proposal RES3721. These data are used only for this work with restricted distribution and will not be published publicly.

LIST OF REFERENCES

Amirkolaee, H. A., & Arefi, H. (2019). Height estimation from single aerial images using a deep convolutional encoder-decoder network. *ISPRS Journal of Photogrammetry and Remote Sensing*, *149*, 50–66. https://doi.org/10.1016/J.ISPRSJPRS.2019.01.013

Arefi, H., & Reinartz, P. (2013). Building Reconstruction Using DSM and Orthorectified Images. *Remote Sensing*, *5*(4), 1681–1703. https://doi.org/10.3390/rs5041681

Arun, P. V., & Katiyar, S. K. (2013). An intelligent approach towards automatic shape modelling and object extraction from satellite images using cellular automata-based algorithms. *GIScience & Remote Sensing*, *50*(3), 337–348. https://doi.org/10.1080/15481603.2013.802870

Awrangjeb, M., Ravanbakhsh, M., & Fraser, C. S. (2010). Automatic detection of residential buildings using LIDAR data and multispectral imagery. *ISPRS Journal of Photogrammetry and Remote Sensing*, *65*(5), 457–467. https://doi.org/10.1016/J.ISPRSJPRS.2010.06.001

Boureau, Y. L., Bach, F., LeCun, Y., & Ponce, J. (2010). Learning mid-level features for recognition. *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, 2559–2566. https://doi.org/10.1109/CVPR.2010.5539963

Brandtberg, T., Warner, T. A., Landenberger, R. E., & McGraw, J. B. (2003). Detection and analysis of individual leaf-off tree crowns in small footprint, high sampling density lidar data from the eastern deciduous forest in North America. *Remote Sensing of Environment*, *85*(3), 290–303. https://doi.org/10.1016/S0034-4257(03)00008-7

Brunner, D., Lemoine, G., Bruzzone, L., & Greidanus, H. (2010). Building height retrieval from VHR SAR imagery based on an iterative simulation and matching technique. *IEEE Transactions on Geoscience and Remote Sensing*, *48*(3 PART2), 1487–1504. https://doi.org/10.1109/TGRS.2009.2031910

Cracknell, A. P. (2019). The development of remote sensing in the last 40 years. *International Journal of Remote Sensing*, *39*(23), 8387–8427. https://doi.org/10.1080/01431161.2018.1550919

Dumoulin, V., & Visin, F. (2016). *A guide to convolution arithmetic for deep learning*. 1–28. http://arxiv.org/abs/1603.07285

Eigen, D., Puhrsch, C., & Fergus, R. (2014). Depth map prediction from a single image using a multi-scale deep network. *Advances in Neural Information Processing Systems*, *3*(January), 2366–2374.

Flores-Anderson, A. I., Herndon, K. E., Thapa, R. B., & Cherrington, E. (2019). SAR Handbook: Comprehensive Methodologies for Forest Monitoring and Biomass Estimation. *THE SAR HANDBOOK Comprehensive Methodologies for Forest Monitoring and Biomass Estimation*, 1–307. https://doi.org/10.25966/nr2c-s697

Fornaro, G., & Pascazio, V. (2014). SAR Interferometry and Tomography: Theory and Applications. *Academic Press Library in Signal Processing*, *2*, 1043–1117. https://doi.org/10.1016/B978-0-12-396500-4.00020-X

Ghamisi, P., & Yokoya, N. (2018). IMG2DSM: Height Simulation from Single Imagery Using Conditional Generative Adversarial Net. *IEEE Geoscience and Remote Sensing Letters*, *15*(5), 794–798. https://doi.org/10.1109/LGRS.2018.2806945

Han, Y., Wang, S., Gong, D., Wang, Y., Wang, Y., & Ma, X. (2020a). State of the art in digital surface modelling from multi-view high-resolution satellite images. *ISPRS Annals of the Photogrammetry, Remote Sensing and Spatial Information Sciences*, *2*, 351–356. https://doi.org/https://doi.org/10.5194/isprs-annals-V-2-2020-351-2020

Han, Y., Wang, S., Gong, D., Wang, Y., Wang, Y., & Ma, X. (2020b). State of the Art in Digital Surface Modelling from Multi-View High-Resolution Satellite Images. *ISPRS Annals of the Photogrammetry, Remote Sensing and Spatial Information Sciences*, *5*(2), 351–356. https://doi.org/10.5194/ISPRS-ANNALS-V-2-2020-351-2020

He, K., Zhang, X., Ren, S., & Sun, J. (2016). Deep residual learning for image recognition. *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, *2016-Decem*, 770–778.

https://doi.org/10.1109/CVPR.2016.90

Karaer, A., Chen, M., Gazzea, M., Ghorbanzadeh, M., Abichou, T., Arghandeh, R., & Ozguven, E. E. (2022). Remote sensing-based comparative damage assessment of historical storms and hurricanes in Northwestern Florida. *International Journal of Disaster Risk Reduction*, *72*, 102857. https://doi.org/10.1016/J.IJDRR.2022.102857

Karatsiolis, S., Kamilaris, A., & Cole, I. (2021). IMG2nDSM: Height Estimation from Single Airborne RGB Images with Deep Learning. *Remote Sensing*, *13*, 2417–2438. https://doi.org/10.3390/RS13122417

Kugler, F., Schulze, D., Hajnsek, I., Pretzsch, H., & Papathanassiou, K. P. (2014). TanDEM-X Pol-InSAR performance for forest height estimation. *IEEE Transactions on Geoscience and Remote Sensing*, *52*(10), 6404–6422. https://doi.org/10.1109/TGRS.2013.2296533

Lee, J. Sen, Wen, J. H., Ainsworth, T. L., Chen, K. S., & Chen, A. J. (2009). Improved sigma filter for speckle filtering of SAR imagery. *IEEE Transactions on Geoscience and Remote Sensing*, *47*(1), 202–213. https://doi.org/10.1109/TGRS.2008.2002881

Li, J., Hong, D., Gao, L., Yao, J., Zheng, K., Zhang, B., & Chanussot, J. (2022). Deep learning in multimodal remote sensing data fusion: A comprehensive review. *International Journal of Applied Earth Observation and Geoinformation*, *112*(April), 102926. https://doi.org/10.1016/j.jag.2022.102926

Li, X., Wang, M., & Fang, Y. (2014). Height estimation from single aerial images using a deep ordinal regression network. *IEEE Geoscience and Remote Sensing Letters*, *13*(9), 1–5.

Li, X., Wang, M., & Fang, Y. (2022). Height Estimation from Single Aerial Images Using a Deep Ordinal Regression Network. *IEEE Geoscience and Remote Sensing Letters*, *19*(9), 1–5. https://doi.org/10.1109/LGRS.2020.3019252

Lim, K., Treitz, P., Wulder, M., St-Onge, B., & Flood, M. (2003). LiDAR remote sensing of forest structure. *Progress in Physical Geography*, *27*(1), 88–106. https://doi.org/10.1191/0309133303pp360ra

Liu, C. J., Krylov, V. A., Kane, P., Kavanagh, G., & Dahyot, R. (2020). IM2ELEVATION: Building height estimation from single-view aerial imagery. *Remote Sensing*, *12*(17), 1–22. https://doi.org/10.3390/RS12172719

Liu, X. (2008). Airborne LiDAR for DEM generation: some critical issues: *Progress in Physical Geography: Earth and Environment*, *32*(1), 31–49. https://doi.org/10.1177/0309133308089496

Lu, Z., Im, J., Rhee, J., & Hodgson, M. (2014). Building type classification using spatial and landscape attributes derived from LiDAR remote sensing data. *Landscape and Urban Planning*, *130*(1), 134–148. https://doi.org/10.1016/J.LANDURBPLAN.2014.07.005

Ma, X., Wang, Z., Li, H., Zhang, P., Ouyang, W., & Fan, X. (2019). Accurate monocular 3D object detection via color-embedded 3D reconstruction for autonomous driving. *Proceedings of the IEEE International Conference on Computer Vision*, *2019-Octob*, 6850–6859. https://doi.org/10.1109/ICCV.2019.00695

Mahdianpari, M., Granger, J. E., Mohammadimanesh, F., Warren, S., Puestow, T., Salehi, B., & Brisco, B. (2021). Smart solutions for smart cities: Urban wetland mapping using very-high resolution satellite imagery and airborne LiDAR data in the City of St. John's, NL, Canada. *Journal of Environmental Management*, *280*, 111676. https://doi.org/10.1016/J.JENVMAN.2020.111676

Mahmud, J., Price, T., Bapat, A., & Frahm, J.-M. (2013). Boundary-aware 3D Building Reconstruction from a Single Overhead Image Modified Signed Distance Function Pixel-wise Height Prediction Pixel-wise Semantic Segmentation Building Refinement 3D Building Modeling. *GIScience & Remote Sensing*, *50*(3), 337–348.

Mahmud, J., Price, T., Bapat, A., & Frahm, J. M. (2020). Boundary-Aware 3D Building Reconstruction from a Single Overhead Image. *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, 438–448. https://doi.org/10.1109/CVPR42600.2020.00052

Meng, X., Wang, L., & Currit, N. (2009). Morphology-based building detection from airborne lidar data. *Photogrammetric Engineering and Remote Sensing*, *75*(4), 437–442. https://doi.org/10.14358/PERS.75.4.437

Moreira, A., Prats-Iraola, P., Younis, M., Krieger, G., Hajnsek, I., & Papathanassiou, K. P. (2013). A tutorial on synthetic aperture radar. *IEEE Geoscience and Remote Sensing Magazine*, *1*(1), 6–43. https://doi.org/10.1109/MGRS.2013.2248301

Morena, L. C., James, K. V., & Beck, J. (2004). An introduction to the RADARSAT-2 mission. *Canadian Journal of Remote Sensing*, *30*(3), 221–234. https://doi.org/10.5589/m04-004

Mou, L., & Zhu, X. X. (2018). IM2HEIGHT: Height Estimation from Single Monocular Imagery via Fully Residual Convolutional-Deconvolutional Network. *IEEE Transactions on Geoscience and Remote Sensing*, *52*(10). http://arxiv.org/abs/1802.10249

Persello, C., Wegner, J. D., Hansch, R., Tuia, D., Ghamisi, P., Koeva, M., & Camps-Valls, G. (2022). Deep Learning and Earth Observation to Support the Sustainable Development Goals: Current approaches, open challenges, and future opportunities. *IEEE Geoscience and Remote Sensing Magazine*, *10*(2), 172–200. https://doi.org/10.1109/MGRS.2021.3136100

Pohl, C., & Van Genderen, J. L. (1998). Review article Multisensor image fusion in remote sensing: Concepts, methods and applications. In *International Journal of Remote Sensing* (Vol. 19, Issue 5). https://doi.org/10.1080/014311698215748

Priestnall, G., Jaafar, J., & Duncan, A. (2000). Extracting urban features from LiDAR digital surface models. *Computers, Environment and Urban Systems*, *24*(2), 65–78. https://doi.org/10.1016/S0198-9715(99)00047-2

Raber, G. T., Jensen, J. R., Hodgson, M. E., Tullis, J. A., Davis, B. A., & Berglund, J. (2007). Impact of Lidar Nominal Post-spacing on DEM Accuracy and Flood Zone Delineation. *Photogrammetric Engineering and Remote Sensing*, *73*(7), 793–804.

Recla, M., & Schmitt, M. (2022). Deep-learning-based single-image height reconstruction from very-high-resolution SAR intensity data. *ISPRS Journal of Photogrammetry and Remote Sensing*, *183*, 496–509. https://doi.org/10.1016/j.isprsjprs.2021.11.012

Rizzoli, P., Martone, M., Gonzalez, C., Wecklich, C., Borla Tridon, D., Bräutigam, B., Bachmann, M., Schulze, D., Fritz, T., Huber, M., Wessel, B., Krieger, G., Zink, M., & Moreira, A. (2017). Generation and performance assessment of the global TanDEM-X digital elevation model. *ISPRS Journal of Photogrammetry and Remote Sensing*, *132*, 119–139. https://doi.org/10.1016/J.ISPRSJPRS.2017.08.008

Ronneberger, O., Fischer, P. F., & Brox, T. (2021). U-Net: Convolutional Networks for Biomedical Image Segmentation. *IEEE Access*, *9*, 16591–16603. https://doi.org/10.1109/ACCESS.2021.3053408

Sauer, S., Ferro-Famil, L., Reigber, A., & Pottier, E. (2009). Polarimetric dual-baseline INSAR building height estimation at l-band. *IEEE Geoscience and Remote Sensing Letters*, *6*(3), 408–412. https://doi.org/10.1109/LGRS.2009.2014571

Schmitt, M., & Recla, M. (2022). Comparison of Single-Image Urban Height Reconstruction From Optical and Sar Data. *International Archives of the Photogrammetry, Remote Sensing and Spatial Information Sciences - ISPRS Archives*, *43*(B2-2022), 1139–1144. https://doi.org/10.5194/isprs-archives-XLIII-B2-2022-1139-2022

Stojanova, D., Panov, P., Gjorgjioski, V., Kobler, A., & Džeroski, S. (2010). Estimating vegetation height and canopy cover from remotely sensed data with machine learning. *Ecological Informatics*, *5*(4), 256–266. https://doi.org/10.1016/J.ECOINF.2010.03.004

Tu, J., Sui, H., Feng, W., & Song, Z. (2016). Automatic building damage detection method using high-resolution remote sensing images and 3D GIS model. *ISPRS Annals of the Photogrammetry, Remote Sensing and Spatial Information Sciences*, *3*, 43–50. https://doi.org/10.5194/ISPRS-ANNALS-III-8-43-2016

Zbontar, J., & Lecun, Y. (2016). Stereo Matching by Training a Convolutional Neural Network to Compare Image Patches. *Journal of Machine Learning Research*, *17*(1), 2287–2318. https://doi.org/10.5555/2946645

Zhang, J., & Lin, X. (2017). Advances in fusion of optical imagery and LiDAR point cloud applied to photogrammetry and remote sensing. *International Journal of Image and Data Fusion*, *8*(1), 1–31. https://doi.org/10.1080/19479832.2016.1160960

Zhang, L. (2005). *Automatic Digital Surface Model (DSM) generation from linear array images* [Swiss Federal Institute of Technology (ETH) CH-8093]. https://doi.org/https://doi.org/10.3929/ETHZ-A-005055636

Zhang, Li, & Gruen, A. (2006). Multi-image matching for DSM generation from IKONOS imagery. *ISPRS Journal of Photogrammetry & Remote Sensing*, *60*(2006), 195–211. https://doi.org/10.1016/j.isprsjprs.2006.01.001