UNIVERSITEIT VAN PRETORIA
UNIVERSITY OF PRETORIA
YUNIBESITHI YA PRETORIA

Automatic Building Footprint Extraction Using Remote Sensing Data within the City of Cape Town

By

Khumeleni Makungo
(22966082)

Supervisor:
Dr. Adedayo Adeleke

Submitted in fulfillment of the requirements for the degree MSc Geoinformatics in the Faculty of Natural and Agricultural Sciences,

University of Pretoria

Pretoria

November 2023

Abstract

In the City of Cape Town Metropolitan (CoCT), South Africa, GIS analysts currently delineate building footprints by digitizing aerial imagery and stereo-aerial images. This approach requires a lot of manual work. It takes a long time, is expensive, and inefficient. Recent studies have explored automatic and semi-automatic methods for extracting building footprints. Automatic extraction of building footprints from remotely sensed data is useful for urban planning, service delivery, and humanitarian efforts. However, there is currently no readily available method that can automatically extract footprints while considering the unique characteristics of the landscape, such as formal residential areas, industrial zones, and informal settlements. Therefore, the main goal of this research is to find a suitable and efficient spatial analysis method that accurately extracts building footprints of different sizes and shapes within the City of Cape Town, South Africa, using high-resolution aerial imagery and LiDAR-derived nDSM. To achieve this goal, a literature review is conducted to explore different building footprint extraction algorithms. The review identified Mask Regional Convolutional Neural Network (R-CNN) as an effective algorithm for instance segmentation and object extraction. Thus, an experiment is conducted to implement Mask R-CNN models that extract building footprints from aerial imagery and LiDAR-derived normalized Digital Surface Model (nDSM) for each of the three areas: formal residential, industrial, and informal settlements. The training focused on the Blaauwberg district, which includes formal residential areas, industrial zones, and informal settlements. Each trained model is separately tested on testing datasets for formal residential, industrial areas, and informal settlements. Evaluation metrics such as precision, recall, F1-score, and Average Precision (AP) score are calculated for each model to assess their performance in extracting building footprints from aerial imagery and LiDAR-derived nDSM in formal residential, industrial areas, and informal settlements. The Mask R-CNN algorithm proved to be very effective in extracting building footprints from high-resolution aerial imagery and LiDAR-derived nDSM in formal residential areas, achieving satisfactory precision, recall, F1-score, and AP score. In industrial areas, the Mask R-CNN algorithm is found to be highly effective in extracting footprints from LiDAR-derived nDSM. However, when extracting shacks in densely populated settlements, the Mask R-CNN algorithm performed inadequately, with an AP score of 0.28 and 0.31 from aerial imagery and LiDAR-derived nDSM, respectively. Nevertheless, the fusion of footprints extracted from LiDAR-derived nDSM and high-resolution aerial imagery improved the AP score to 0.52. Hence, this study concludes that the Mask R-CNN algorithm is highly effective in extracting building footprints in formal residential areas from both aerial imagery and LiDAR-derived nDSM, as well as industrial building footprints from LiDAR-derived nDSM. For optimal performance in informal settlements, the fusion of footprints extracted from aerial imagery and LiDAR-derived nDSM is necessary. Overall, these trained Mask R-CNN models demonstrated satisfactory performance. To enhance the existing 2D building footprint layer, these models can supplement by extracting building footprints. This updated layer will be more comprehensive and current. Various departments within the CoCT can utilize this layer for infrastructure planning, service delivery planning, land use planning, and change detection. For better performance, it is recommended to add more informal and industrial training datasets with sufficient roof variability. Fine-tuning the Mask R-CNN models

will ensure accurate extraction of shacks and industrial building footprints by allowing the models to learn effectively.

Acknowledgments

Firstly, I would like to thank my supervisor, Dr. Adedayo Adeleke, for his continuous support and encouragement. He generously spared time to provide valuable comments throughout my research for this Thesis work, accomplishing various tasks and composing the report. Dr. Adeleke's patience and immense knowledge make him an excellent mentor, and his helpfulness makes him a great person. I must also express my gratitude to my wife, Dr. Fulufhelo, who supported me unwaveringly throughout the entire journey. Finally, I am grateful to the City of Cape Town's Geospatial Unit, especially Thomas Reiner and Lara Rottcher, whose passion for geospatial work is truly inspiring. This research would not have been possible without their support. Again, I would like to express my gratitude to the City of Cape Town's Geospatial Unit for providing me with valuable resources including the latest LiDAR, High-resolution aerial imagery, and the training and validation dataset as well as the necessary hardware and software for the research and development work.

Plagiarism Declaration

I, Khumeleni Makungo declare that the dissertation, which I hereby submit for the degree MSc Geoinformatics at the University of Pretoria, is my own work and has not previously been submitted by me for a degree at this or any other tertiary institution.

Signature: Khumeleni Makungo
Date: 23 November 2023

Table of Content

1. Introduction

1.1 Background

The extraction of building footprints from remotely sensed data is increasingly important for various applications in residential and urban areas (Gilani et al., 2015). In South Africa, like in many other countries, residents are moving to cities in search of better education, job opportunities, and more. As a result, cities become overcrowded, creating a demand for formal housing among the working class. This demand, along with the needs of informal settlements, requires services from government entities such as Emergency Medical Services, the police, and water delivery, to ensure a functional environment (Esri South Africa, 2022). The private sector, focusing on utilities and real estate, also requires up-to-date information on dwelling frameworks for spatial analysis, planning, product placement, and market awareness (Haithcoat et al., 2001).

Census surveys are conducted every ten years in South Africa, but due to the dynamic nature of formal housing and informal settlements, the household count becomes outdated by the time the results are released (Esri South Africa, 2022). This poses challenges in accessing up-to-date information for planning, service delivery, and humanitarian interventions. To overcome this, up-to-date building footprint data is needed to accurately estimate the number of households in areas experiencing active migration, including the City of Cape Town. This relies on new, faster, and more reliable data acquisition and analysis techniques.

Building footprint extraction from remotely sensed data has been studied extensively worldwide. Accurate building boundaries are crucial for applications in real estate, urban planning, disaster management, 3D city modeling, cartographic mapping, and emergency responses (Sohn and Dowman, 2007; Li and Wu, 2013; K. Bittner et al., 2018). Building footprints provide visual representations of a building's location, shape, dimensions, orientation, and area. They may also include additional geospatial information such as address, latitude/longitude, place, and spatial hierarchy. With the development of smart cities, there is an increasing need for automatic or at least semi-automatic methods for digital building footprint data extraction in urban areas (Partovi et al., 2017).

Currently, in the City of Cape Town, GIS analysts manually delineate building footprints by digitizing aerial imagery and stereo-aerial images. This approach is time-consuming, expensive, and requires significant manual work (K. Bittner et al., 2018). The automatic extraction of building footprints is challenging due to variations in building shape and size, as well as the complexity of the surrounding environment (Shoko et al., 2022). High-resolution imagery, although rich in spectral information, is susceptible to noise and can be affected by contrast, illumination, occlusion, and shadow effects (Gilani et al., 2015). Extraction of building footprints becomes particularly challenging in large and densely built-up urban areas. To overcome this, researchers have focused on developing automated methods using Light Detection and Ranging (LiDAR)-derived Digital Surface Models (DSM), which provide valuable data sources for building footprint extraction and height information (Lee et al., 2003; Zhang et al., 2006; Tarantino and Figorito, 2011; Wang, 2016). The height variation captured by LiDAR data is

more suitable for detecting elevated objects and delineating building footprints than spectral and texture changes (Gilani et al., 2015). However, the under-sampling nature of LiDAR data acquisition, along with backscattering limitations, makes it difficult to extract building edges with height discontinuity, resulting in poor horizontal accuracy and geometric precision of the extracted footprints (Sohn and Dowman, 2007).

To compensate for the limitations of individual data sources, the fusion of LiDAR data and aerial imagery is used to provide complementary information and improve accuracy and robustness in building footprint extraction (K. Bittner et al., 2018). Several studies have attempted to integrate airborne LiDAR and high-resolution aerial imagery for building footprint extraction (Zhang et al., 2020; Gilani et al., 2015; Bittner et al., 2018).

In this research, a method for effectively extracting building footprints using LiDAR-derived normalized Digital Surface Models (nDSM) and high-resolution aerial images within the City of Cape Town is investigated and implemented.

1.2 The Research Problem Statement

For successful urban planning, service delivery, and humanitarian interventions, having accurate and current building footprint data is crucial. Obtaining this data relies heavily on advanced techniques for fast, reliable, and capable data acquisition and analysis. In the past, GIS analysts manually digitized building footprints from aerial and satellite imagery, which is inefficient for citywide efforts, even though it requires minimal user training (K. Bittner et al., 2018). Modern approaches prefer automated and efficient methods for extracting building footprints, aligning well with the rapidly evolving nature of big data, machine learning, deep learning, and digital extraction algorithms.

Recent studies have applied deep learning algorithms to aerial imagery or LiDAR-derived nDSM to automatically extract building footprints. However, these studies do not consider the unique characteristics of buildings in formal residential and industrial zones, and informal settlements (Zhao et al.,2018; Tiede et al., 2021; Mohamed et al.,2022). This predominantly happens in developing countries due to rapid urbanization which leads to the proliferation of informal settlements in developing countries caused by the failure to provide this rapidly urbanizing population with the necessary services and infrastructure, including planned land, for orderly development (Kironde, 2006). Thus, South Africa is a developing country and formal and informal zones co-exist in its Metropolitan. These areas consist of buildings with different characteristics, such as various roofing materials, shapes, sizes, and heights. This negatively affects the performance of deep learning algorithms. Providing a solution is necessary as this is a challenging problem. Therefore, in this research, these three areas: formal residential, industrial, and informal settlement have been separated when training the deep learning models. Two Mask R-CNN models have been trained for each of the three areas, with one using the LiDAR-derived nDSM and the other utilizing the aerial imagery, both employing the labeled 2D building footprints as training datasets.

To effectively compare the models' performance, a new test dataset, unseen in both training and validation, is fed into the network, and evaluation metrics are calculated. For each of the

three areas, a comparison and analysis are conducted on the results obtained from the aerial imagery and LiDAR-derived nDSM to find the best-performing model for extracting building footprints throughout the CoCT.

1.3 Research Questions

The research will seek to answer the following question(s);

- Which spatial data analysis method is suitable and efficient to extract accurate and well-regularized building footprints?
- How can remote sensing data such as aerial imagery and LiDAR data be effectively used to extract accurate building footprints?
- How do building footprint extraction models perform across various urban scenes in the City of Cape Town?

1.4 Aim and Objectives

1.4.1 Aim

The main aim of this research was to automatically extract building footprint from remote sensing data in the City of Cape Town Metropolitan, South Africa.

1.4.2 Objectives

In achieving the aim stated above the following objectives were pursued;

- To conduct a literature review and identify an effective spatial data analysis method for automatic building footprint extraction.
- To use LiDAR-derived nDSM and aerial imagery to improve the accuracy and robustness of building footprint extraction.
- To discuss building footprint extraction results obtained from aerial imagery and LiDAR-derived nDSM, and compare the performance of models in formal residential, industrial area, and informal settlements.

1.5 The Research Significance

Building extraction from remote sensing data has become a crucial and challenging task in recent decades due to rapid urban growth. The identification of buildings in remote sensing data, such as high-resolution aerial imagery and LiDAR (nDSM), requires significant computational resources but holds immense importance across various industries and government institutions. Currently, the City of Cape Town manually digitizes building roofs from stereo imagery to create a building footprint layer for multiple applications. This study addresses the need for a fast and effective method to automatically extract building footprints from available remote sensing data in the City of Cape Town, municipality. Various departments within CoCT are seeking a more accurate and up-to-date building footprint dataset for a range of purposes, including:

- Electricity, Water and Sanitation, Transport departments' infrastructure planning initiatives include the development of stormwater and sewer networks, the planning of electric routes and substations, as well as road and MyCity Bus routes and stops planning

- Electricity departments use building footprints to monitor informal structures constructed beneath their power lines. This enables them to plan for relocating dwellers.
- The Catchment Stormwater and River Management branch in the Water and Sanitation department uses building footprints for flood modeling studies. Access to building footprints helps obtain more precise results for flood modeling studies.
- The Property Valuation department monitors and detects changes in buildings to proactively identify and address illegal construction activities throughout the city.
- Service delivery planning involves estimating future electricity demands, ensuring sufficient waste bins and collection services for formal residential areas with backyards, and providing bagged cleansing services for households in informal settlements. Accurate planning and service delivery depend on knowing the number of dwellings in these areas.
- The Spatial, Urban Planning, and Design departments utilize building footprints for several purposes. Firstly, uses them to estimate the floor factor in areas designated as single residential 2 for land use planning in development management schemes. Secondly, it uses them to examine changes in CoCT's densities over time. This helps determine if the densification targets set by the CoCT align with the actual situation on the ground. Lastly, the building footprint data is used to draft the Local Spatial Development Framework (LSDF), which guides decisions related to spatial development and land use management and reflects the CoCT's future development vision.

1.6 The Scope of the Research

This research is limited to the extraction of building footprint within the City of Cape Town Metropolitan in the Western Cape Province of South Africa. The City of Cape Town covers an area of about 2461 km^2, thus for this study, the Blaauwberg area is chosen for the training, validation, and testing of the deep learning models. The Blaauwberg area consists of formal residential, industrial areas, and informal settlements. As a result separate models for each area will be implemented and used to perform instance and semantic segmentation of building footprint specifically for that area (formal residential, industrial, or informal settlements) within the CoCT. The implemented deep learning models will be used to extract building footprints throughout the CoCT to supplement the existing building footprints generated through photogrammetric methods.

Considering the significant computational power and time needed to adequately train deep learning models, only the Mask R-CNN algorithm using LiDAR-derived nDSM and high-resolution aerial imagery is evaluated. The results of the models are analyzed for the three areas, namely, formal residential, industrial, and informal settlement.

1.7 Dissertation Overview

In this section, the structure of the research work is discussed.

Chapter 1: Background and motivation related to this research are presented along with the research problem statement, research questions, aim and objectives, research methodology, and scope of the research.

Chapter 2: This chapter presents the conducted literature review with a motivation behind finding a suitable and efficient spatial analysis method to extract accurate and well-regularized building footprints. Thus, this assists in achieving the research objective 1.

Chapter 3: This chapter describes the dataset and study area, important concepts about Mask R-CNN architecture to understand how it works internally are provided, and performance metrics for the experiment and model inferencing are explained. Thus, this assists in achieving the research objectives 2.

Chapter 4: This chapter presents the results of the research work and the analysis of the presented results and discussions regarding the analyzed results are mentioned. Thus, this assists in achieving the research objective 3.

Chapter 5: This chapter provides the conclusion and future work of the research work. Thus, this assists in achieving the research objective 3.

2. Literature Review

In Chapter 1, the background and objectives of this study are presented. The current chapter provides a review of research conducted on building footprint extraction using different methods, as well as the main applications of extracted building footprints from high-resolution aerial imagery with spatial resolution ranging from 0.1m to 05m, and LiDAR data. The primary purpose of this review is to analyze and identify the most suitable and efficient spatial data analysis method for accurately and consistently extracting building footprints. The review aims to explore the available algorithms and determine the appropriate spatial analysis algorithm. The remaining sections of this chapter are organized as follows:

Section 2.1 reviews the literature on studies conducted on traditional building footprint extraction methods.

Section 2.2 reviews the literature on studies conducted using deep learning-based building footprint extraction methods. This section starts by discussing the deep learning background and then reviews the literature for studies conducted using two deep learning frameworks, namely Unet and Mask R-CNN for semantic and instance object segmentation, respectively.

Section 2.3 discusses the literature on studies conducted on the key applications for building footprints extracted from high-resolution aerial imagery and LiDAR data

2.1 Traditional Building Footprint Extraction

In the past 12 years, there has been a growing trend in extracting building footprints from complex environments. This trend involves integrating high-resolution imagery and LiDAR data, which brings complementary benefits. By combining spectral and 3D surface information, a more complete description of the scene can be achieved (Gilania et al., 2015). The integration of these two data sources has been utilized to enhance the classification performance and improve the accuracy and robustness of automatic building detection (Gilania et al., 2015).

2.1.1 Rule-Based Methods

According to Siddiqui et al. (2016), rule-based building extraction methods are commonly preferred due to their simplicity and effectiveness in various environments. Here's how these methods usually work: Firstly, LiDAR data and imagery are utilized as primary cues to delineate building footprints in the pre-processing and main stages. Aerial imagery is also used to calculate features like Normalized Difference Vegetation Index (NDVI), entropy, shadow, and illumination to eliminate vegetation. Consequently, they offer better horizontal accuracy for the detected buildings (Gilania et al. 2015).

Another research reported by Li et al. (2013) uses the fusion of LiDAR data and high-resolution images with a spatial resolution of 0.1m and 0.5m, respectively. Firstly, a high-quality Digital Terrain Model (DTM) generated from the LiDAR data and highly accurate coordinates of ground control points from LiDAR intensity images are used for orthorectification by an aerial triangulation calculation. The LiDAR data is classified to filter ground points and tree points, tree points are filtered using the height difference between the first and last pulse of the point

cloud. To separate the tree points from edge and wall points, a novel criterion based on the density, connectivity, and distribution of point clusters is used. Then, coarse building footprints are extracted from the classified point cloud, edges are detected from high-resolution images, and then correct boundaries are identified within the buffer of projected coarse building footprints extracted from the LiDAR data. Subsequently, precise building footprints are generated by matching conjugate boundaries from high-resolution aerial imagery with the help of coarse building footprints from LiDAR data. The identified building footprints are further regularized using the RANdom SAmple Consensus (RANSAC) algorithm. The results obtained demonstrate the significant improvement in building footprint extraction accuracy achieved by the proposed method (Li et al., 2013).

2.1.2 Gradient-Based Methods

Siddiqui et al. (2016) developed a novel Gradient-based building extraction method that leverages both LiDAR data and aerial imagery. The LiDAR data are first divided into ground and non-ground points and then straight lines (i.e., principal orientations of buildings) are extracted from the aerial imagery using the Canny Edge detector and Gaussian function. The non-ground points are used to separate non-ground straight lines from the ground straight lines. The principal orientations of buildings are estimated using the non-ground lines. For each principal orientation, the non-ground points are employed to generate an intensity image. Then, a binary building mask is derived through a gradient analysis of the intensity image. The binary building mask enables the removal of trees through a refinement process. In the refinement process, the variance and density analysis is employed on the non-ground building points to eliminate trees, whereas, the local colour matching and shadow elimination analyses are employed on the imagery pixels to eliminate the remaining regions of trees. Remaining trees after the refinement process, the variance and density analysis are removed using the morphological filter. Finally, building footprints are extracted around each building (Siddiqui et al., 2016).

It is worth noting that these studies focused primarily on traditional methods and dealt with extracting buildings in relatively small study regions. However, their performance has not been evaluated in areas containing diverse and complex buildings.

2.2 Deep Learning Building Footprint Extraction

Building footprints are among the most prominent features of an urban setting. With the increasing availability of very high-resolution aerial imagery and LiDAR data, the research paradigm of urban feature extraction has shifted from a traditional-based approach to semantic and instance segmentation approaches using neural networks such as Convolutional Neural Networks (CNNs) (Aryal et al. 2023). This section presents CNN architectures used to extract building footprints in this research.

2.2.1 Convolutional Neural Network (CNN)

In recent years, deep learning methods have been widely used in various applications involving remote sensing images. These methods employ multiple layers in the network to represent complex characteristics (Abdollahi et al. 2020; Li et al. 2019; Liu et al. 2020). By doing so, the original input is mapped to multiple variable labels, enabling accurate classification. One common type of deep learning method is Convolutional Neural Networks (CNN), which serves

as the backbone of the networks (Alsabhan et al. 2022). CNN networks typically consist of four layers. The first layer, called the convolution layer, utilizes a filter to traverse the image. The filter acts as a sliding window, performing calculations on the pixels it covers. The output of this process is known as a feature map (Krizhevsky et al. 2012). Following the convolution layer, there is often an activation layer, such as ReLu, which sets negative values in the incoming data to zero (Anam, T. 2021). Subsequently, a pooling layer is applied. In this layer, the maximum value within a selected window is retained, reducing the data size through a process known as downsampling. Lastly, the flattening layer converts the matrix into a single vector (Anam, T. 2021).

CNNs are now commonly used as feature extractors in encoder-decoder networks. By learning from relevant data, CNN networks automatically extract features, such as urban building footprints, from the deep structures within the images (Alsabhan et al. 2022). Encoder-decoder networks consist of two main parts: an encoder and a decoder (Alsabhan et al. 2022). The encoder, implemented as a CNN, extracts features from the input image and creates feature maps. The decoder then transfers the low-resolution feature maps from the encoder to high-resolution feature maps that align with the input image's size for pixel-wise classification. The decoder achieves this by employing various operations, including upsampling, concatenation, and regular convolutions (Aryal et al. 2023). Upsampling, which can also be referred to as transposed convolution, up convolution, or deconvolution, is a key technique used in the process (Vincent, et al., 2018).

2.2.2 Unet

The Unet is a popular deep-learning convolutional neural network architecture for semantic segmentation and has been used in several satellite image segmentation studies (Li et al. 2019). Unet was initially developed for biomedical image segmentation by Olaf Ronneberger et al. (2015) and requires a relatively small number of training samples. The Unet architecture consists of the encoder and decoder with the addition of "skip connections". These connections enable low-level information to pass from the encoder to the decoder to concatenate the corresponding encoder layer to the output of up-convolution (Ronneberger et al, 2015). The encoder-decoder networks utilize both low-resolution and high-resolution features, conserving the spatial integrity of objects which is crucial in the semantic segmentation of features (Aryal et al. 2023).

Since its introduction in 2014, deep convolutional neural networks have been widely used for various remote sensing image analysis tasks, including road extraction, building extraction, and land cover mapping (Xu et al. 2018; Audebert et al. 2017). Deep learning has become the preferred solution for object detection and classification. For example, Li et al. (2019) applied the Unet algorithm to extract building footprints using the SpaceNet semantic labeling dataset and WorldView-3 satellite images provided in the 2018 DeepGlobe Challenge. The dataset contained 24,586 image scenes, each with a size of 200 pixels by 200 pixels. They trained the Unet-based semantic segmentation model using 302,701 fully annotated building footprint polygons.

Alsabhan et al. (2022) proposed a building extraction method using the Unet algorithm and the Massachusetts building dataset from 2013. The dataset included 151 aerial images at a resolution of 1m2/pixel covering 340 square kilometers of densely populated areas in the city of Boston. Building footprints were obtained from the OpenStreetMap project, and the dataset had two classes: Background (class 0) and Building (Class 1). Two separate Unet-based semantic segmentation models were trained for 60 epochs each, using Residual Network (ResNet50) and ResNet152 as backbones to improve the Unet model's performance. RestNet152 was considered to increase the performance of the Unet model when used as a backbone because it has deeper neural network architecture than ResNet50 architecture (Alsabhan et al., 2022).

Pan et al. (2020) trained a Unet using a Worldview-2 satellite image of 0.5m spatial resolution and a building boundary vector file from the Guangzhou Land Resources and Urban Planning Bureau. They classified four types of buildings in the Tianhe District of Guangzhou City in Southern China: old houses, old factories, iron roof buildings, and new buildings.

2.2.3 Mask R-CNN

Mask R-CNN (Regional Convolutional Neural Network) is a state-of-the-art model for instance segmentation. It builds upon Faster R-CNN by incorporating an extra branch for predicting segmentation masks on each region of interest (RoI). Mask R-CNN utilizes CNN to achieve precise object detection (Chitturi, G., 2020) and has proven to be versatile across various fields (He et al., 2017). It consists of two main phases: region proposal generation and classification (Wu et al., 2021). Mask R-CNN employs a fully convolutional network on CNN feature maps to generate a binary mask that determines if a pixel belongs to an object or not (Kaiming et al., 2020). This method is capable of detecting bounding boxes and segmenting building classes from background classes (Chitturi, G., 2020)

Esri collaborated with Nvidia and Miami-Dade Country to propose a method for 3D building reconstruction using aerial LiDAR and a Deep Neural Network. Specifically, they employed a Mask R-CNN model trained to detect and report instances of roof segments. Mask R-CNN integrates object detection, which aims to detect object classes and predict bounding boxes, with semantic segmentation, which classifies pixels within each box into predefined categories (Esri, 2015). By leveraging Mask R-CNN, objects can be detected in a raster while accurately segmenting masks for each instance (Wei et al., 2020). The training dataset was created by generating a Digital Surface Model raster from the LiDAR point cloud and manually digitizing polygons that describe the roof segments of the building. These roof segments were stored as features in a polygon feature class, and the DSM was normalized by subtracting the DTM (Dmitry et al., 2018). Using ArcGIS Pro's "Export Training Data for Deep Learning" geoprocessing tool, the human-digitized roof segments layer and nDSM with 0.21 square meter/pixel resolution were converted into deep learning training datasets, which included image chips, labels, and statistics files for instance segmentation problems (Dmitry et al., 2018). For the neural networks, Mask R-CNN architecture was employed using the TensorFlow 1.7 deep learning framework. The model was trained with a ResNet-101 backbone architecture for approximately 1,400 epochs. The trained model was then utilized to detect roofs from the raster using the "Detect Object Using Deep Learning" tool, and the raw detections were further

refined using the "Regularize Building Footprint" tool from ArcGIS Pro's 3D analyst toolbox (Dmitry et al., 2018). The results indicate that an increase in the training set can enhance prediction accuracy.

Zhao et al. (2018) also proposed Mask R-CNN ResNet-101 backbone-based building footprint extraction from high-resolution satellite images, along with a boundary regularization method to convert polygons generated by Mask R-CNN into regularized polygons due to irregular and noisy outlines.

Mohamed et al. (2022) introduced an ensemble method for efficiently extracting building footprints using LiDAR-derived DSM in densely populated rural areas of Maghagha City, Egypt. This method combined two Mask R-CNN ResNet backbones (34, 101) and employed a post-processing phase to enhance the extracted building footprints. The obtained results showed an average overall accuracy, precision, recall, and F-score of 0.95, 0.82, 0.98, and 0.88, respectively.

Tiede et al. (2021) conducted a study utilizing Mask R-CNN deep learning implemented in the Python API for Esri's ArcGIS environment to extract building footprints in Khartoum, Sudan, aiding humanitarian organizations in their response to the Covid-19 pandemic. The study achieved a recall of 0.78, precision of 0.77, and F-score of 0.78.

Furthermore, Wei et al. (2020) proposed a method for building footprint extraction from aerial images with a spatial resolution of 20cm using fully convolutional networks (FCNs). They developed a multiscale aggregation FCN (MA-FCN) with a feature pyramid network (FPN)-based structure as the backbone architecture to extract building pixels. The method also fused separately trained models to enhance segmentation accuracy, labeling each pixel as "building" or "non-building" based on consistent voting from the majority of the models.

2.3 Application of Building Footprints

In recent years, there has been significant research focused on extracting building footprints from high-resolution imagery and LiDAR data. This field of study has numerous applications, including population estimates, urban planning, disaster response, service delivery planning, and environmental monitoring. This section provides a summary of the research conducted on the main applications of extracted building footprints from high-resolution imagery and LiDAR data. The following examples highlight some key applications, but this list is not exhaustive.

2.3.1 Population Estimates

According to Esri South Africa (2022) and Shoko et al. (2022), data on building footprints can be used to accurately estimate the number of households in areas experiencing active migration. This is important because census data in South Africa is only updated every 10 years. By analyzing factors such as the number of floors, floor area, and other characteristics of the buildings, we can estimate the household count. Population estimates have various uses, including planning, service delivery, and emergency response (Biljecki et al., 2015a), as well as predicting future utility and infrastructure needs in a given area (Rajabifard et al., 2018a).

2.3.2 Urban Planning

Urban planning is one of the key applications of building footprints extracted from satellite imagery and LiDAR data. According to Zhang et al. (2022), building footprints derived from high-resolution satellite imagery can be used to analyze the spatial distribution of buildings and assess the density of urban areas. This can be useful for urban planners to identify areas of high population density and assist in transportation planning, cadastral and telecommunication network planning, and management. According to Shoko et al. (2022), building footprints derived from the LiDAR dataset can be used to separate shacks to non-shacks areas. These findings can be incorporated into urban planning frameworks, which can also be adjusted to include social and environmental factors.

2.3.3 Disaster or Emergency Response

Another important use of building footprints, extracted from satellite imagery and LiDAR data, is in disaster response. According to Zhang et al. (2022), these footprints can be utilized to assess and respond to various disasters like earthquakes, floods, and fires, particularly in informal settlements. The study discovered that satellite-derived building footprints can help identify areas where buildings have been destroyed or damaged, enabling effective planning for recovery efforts. Tiede et al. (2021) also employed a deep learning convolutional neural network to extract building footprints for supporting humanitarian response during the COVID-19 pandemic in Khartoum, Sudan.

2.3.4 Change Detection

Building footprints extracted from satellite imagery and LiDAR data are beneficial for monitoring the environment and detecting changes (Wei et al., 2020; Wei et al., 2016). They provide valuable information on alterations in land use and vegetation cover over time. Research has shown that building footprints derived from satellite imagery can effectively track changes in land cover in urban areas and assess the impact of urbanization on the environment. For example, Wei et al. (2016) utilized high-resolution aerial satellite imagery to extract building footprints and estimate the dynamics of impervious surface areas in Guangzhou, China. Their study revealed a significant 200% increase in impervious surface areas over 30 years.

2.3.5 Property Value Calculations

Building footprints extracted from satellite imagery and LiDAR data are valuable for assessing and valuing properties in the real estate industry and government. Property valuation plays a crucial role in determining real property tax, which serves as a vital source of government revenue (Isikdag et al., 2015). Consequently, as suggested by Li et al. (2013), building footprints can be utilized to evaluate the size, shape, and location of buildings, allowing for the estimation of their value based on these characteristics and location. A property's value is typically influenced by various factors, including floor area and location (Isikdag et al., 2015). In property value calculations, the floor area carries significant weight and can be approximated from a 2D building footprint.

Overall, building footprint extraction from satellite imagery and LiDAR data has a wide range of applications in various fields. With the increasing availability of high-resolution satellite imagery and LiDAR data, the use of building footprints is expected to increase in the future.

2.4 Chapter Summary

This chapter discusses relevant literature for the research questions and objectives of this study. After reviewing the literature on traditional-based and deep learning-based methods for extracting building footprints, it was found that the deep learning-based approach offers innovative ways to detect and extract building footprints using semantic and object segmentation algorithms. The Mask R-CNN algorithm, in particular, is effective for instance segmentation and object extraction. Therefore, it was chosen for this research due to its remarkable results in extracting building footprints from very high-resolution images and LiDAR, as reported by Zhao et al. (2018), Chitturi, G. (2020), Tiede et al. (2021), and Mahamed et al. (2022). Furthermore, Mask R-CNN provides GIS-ready building footprints and outperforms other approaches in this regard (Tiede et al. 2021).

Moreover, the reviewed literature in this study does not consider the unique characteristics of different landscapes, such as formal residential, industrial areas, and informal settlements. Thus, in this research, the uniqueness of these landscapes is taken into account, and formal residential, industrial, and informal settlements are separated accordingly. This consideration informed the study design and methodologies adopted in the subsequent chapters.

3. Methodology

Having reviewed relevant literature covering major aspects of building footprint extractions, this chapter presents the techniques adopted in achieving the aim of this research by answering the research objectives.
Section 3.1 provides an insight into the study area's location and extent. The study area's data acquisition and preparation is presented, detailing steps followed to prepare aerial imagery and generation of LIDAR-derived nDSM. Furthermore, the creation of training and validation sets is presented.

Section 3.2 provides an insight into building footprint extraction from high-resolution aerial imagery and LIDAR-derived nDSM using instance object segmentation. Object instance segmentation is discussed and mask R-CNN architecture is presented. Furthermore, the steps adopted for training the Mask R-CNN model are presented.

Section 3.3 outlines in detail the steps adopted to extract the building footprint using the models trained in sections 3.2 and 3.3, from aerial imagery and LiDAR-derived nDSM. Furthermore, the adopted steps to regularise building footprints are presented.

3.1 Data Acquisition and Preparation

3.1.1 Study Area and Dataset Descriptions

This research focuses on extracting building footprints within the City of Cape Town Metropolitan in the Western Cape Province of South Africa. South Africa is a developing country and its metropolitan consists mainly of formal and informal zones. This is due to rapid urbanization which leads to the proliferation of informal settlements in developing countries caused by the failure to provide this rapidly urbanizing population with the necessary services and infrastructure, including planned land, for orderly development (Kironde, 2006). Thus, formal and informal zones co-exist in the City of Cape Town Metropolitan which is located at a latitude of -33.55°S and a longitude of 18.25°E along the south-western coast of South Africa. It covers approximately 2700 km2, with a built-up area of about 1400 km2.

The primary data inputs for this research include high-resolution aerial imagery and LiDAR covering the entire metropolitan. The aerial imagery is captured annually, while the LiDAR data is captured in two batches over two years. The LiDAR data, captured between 2020 and 2021 using a Reigl 1560 LiDAR sensor, has an average point density of 10 points per square meter. The aerial imagery is ortho-rectified and has three spectral bands: visible red, visible green, and visible blue, with a spatial resolution of 8cm. The aerial imagery is captured using Vexcel UltraCam large-format metric digital camera. The Vexcel cameras frame dimensions are: UCE 20,010 by 13,080 pixels. Vexcel cameras use 'forward motion compensation' which adjusts for the forward speed of the aircraft so that "image movement" on the sensor is reduced to 0 micrometres. Both the LiDAR data and aerial imagery are subdivided into manageable 5km2 tiles to facilitate computational processing.

Given the extensive study area and the large data size (approximately 68.4GB for compressed aerial imagery and 1.2TB for the LiDAR data of the build-up area), substantial computational power and time are required to properly train deep learning models. Consequently, the Blaauwberg district within the City of Cape Town was selected for training and validation of the deep learning models used to detect building footprints. The Blaauwberg district covers an area of approximately 550.27 km2 and is located along the northwestern coastal boundary of the CoCT. It encompasses formal residential, industrial zones, and informal settlements with buildings of various roofing materials, sizes, and shapes. Figure 1 depicts the boundaries of the City of Cape Town metropolitan area and the Blaauwberg district.

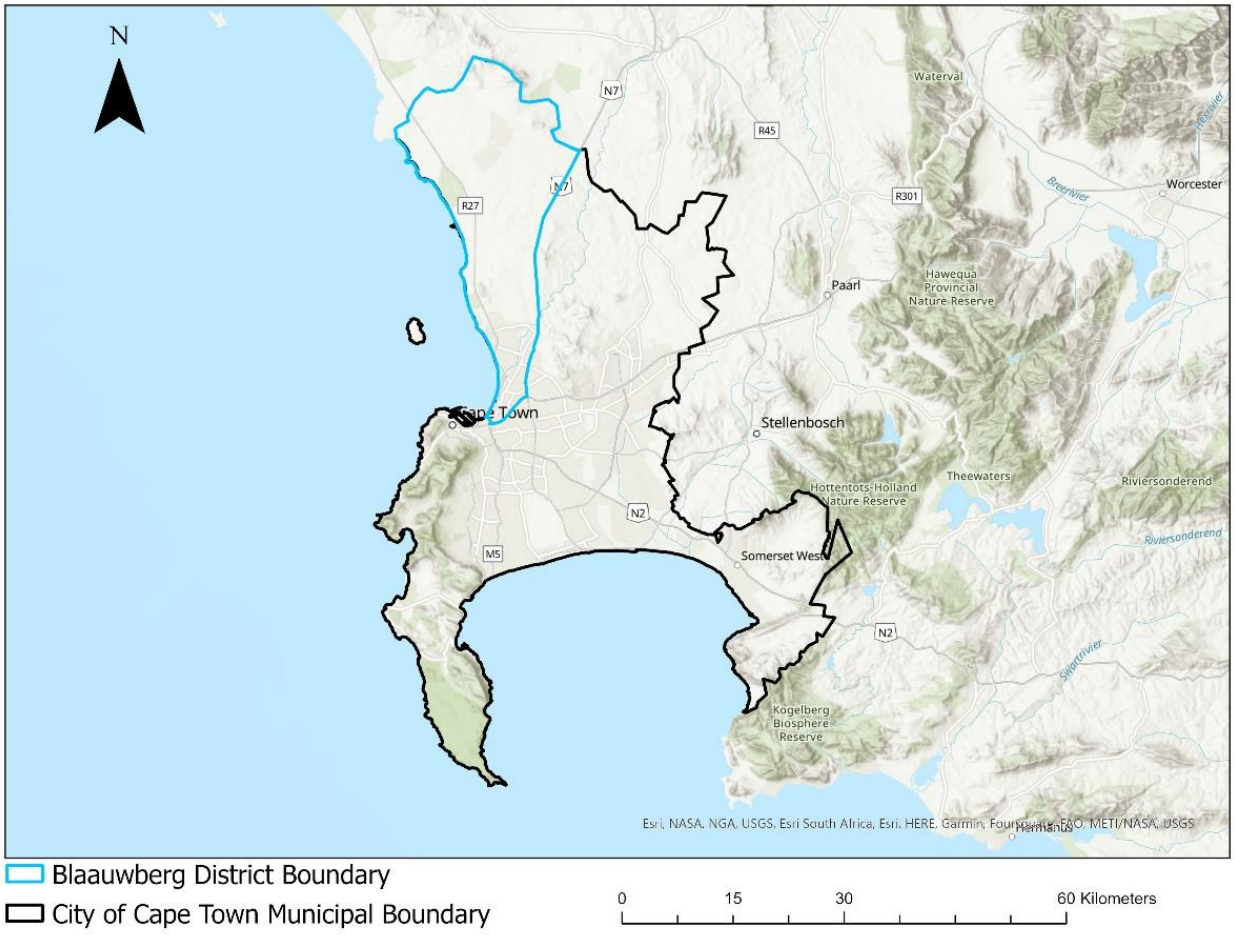


Figure 1: Location of the study area in the City of Cape Town Metropolitan, Western Cape Province of South Africa: Blue represents the Blaauwberg district boundary, and black represents the City of Cape Town boundary.

The aerial imagery and LiDAR data for formal residential, industrial, and informal settlements are extracted from the City's dataset. The Blaauwberg district is divided into these areas to train separate models for each, as they contain buildings with distinct characteristics

3.1.2 Data Labelling

The labelled building footprints used to train and validate the models in these areas are derived from the CoCT's 3D building models of 2022. To create 2D building footprints, the roof details of the 3D models are converted using the "Multipatch Footprint" geoprocessing tool in ArcGIS.

These roof details are obtained from stereo-aerial images captured in 2021 using a digital photogrammetric workstation. The roof details are captured using SocetGXP software. Figure 2 illustrates the workflow of the processes involved.

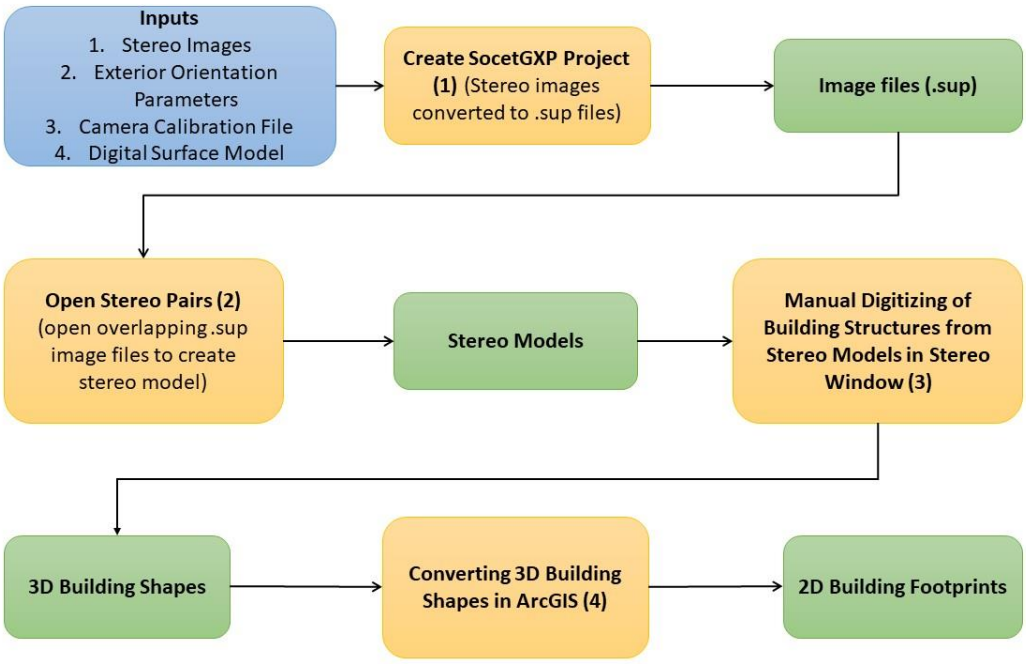


Figure 2: Workflow process for generating training dataset in SocetGXP and ArcGIS. Yellow boxes represent tools/processes and green boxes represent generated outputs from those tools/processes.

The stereo images are imported into the socetGXP software along with the adjusted exterior orientation parameters from aerial triangulation and the camera calibration file. A socetGXP project is created, and all imported stereo images are converted to .sup format. A GeoTIFF DSM of the area, with a spatial resolution of 2m, is imported into the socetGXP project to determine depth, which elevates the stereo images to DSM height. Stereo models are created by opening overlapping image files (.sup), also known as stereo pairs. The generated stereo models are accessed, and 3D building shapes are manually digitized by identifying and capturing building structures and roof details from these models. Special tools are used to ensure the roofs maintain parallel shapes and consistent height for geometric accuracy. Once captured, the shapes are exported to a GIS workstation, where they are transformed into 2D building footprints using ArcGIS's "Multipatch Footprint" geoprocessing tool. These footprints are then checked against the land parcel boundaries. The resulting 2D building footprints are stored as features in a polygon feature class within a local file geodatabase.

3.1.3 Aerial Imagery, DSM, DTM, and nDSM Generation

This section discusses the methods used to process aerial imagery and LiDAR data for extracting building footprints. ArcGIS Pro 3.0 is utilized for processing LiDAR data, while Global Mapper Pro is used for aerial imagery. These software packages are widely adopted in the geospatial industry.

The ortho-rectified aerial imagery of the Blaauwberg district consists of three spectral bands. It is extracted from the compressed aerial imagery (.ecw format) of the entire CoCT. The aerial imagery has a spatial resolution of 8cm, which is then downsampled to 20cm using Global Mapper Pro. The resulting aerial imagery is a three-band (8-bit unsigned) GeoTIFF. This

downsampling helps reduce the computational power and time required for training deep learning models (as discussed in sections 3.2.3 and 3.3.3). Initially a formal residential model was trained from 8cm aerial imagery and the training time was over a month for only 30 epochs. Thus, the model was then tested and compared against a model trained on 20cm aerial imagery. The downsampling did not have significant impact on accuracy, but reduced the computational power and time significantly. The choice of 20cm spatial resolution is based on the observation that similar studies reviewed in chapter 2 used aerial imagery and nDSM with spatial resolutions of 20cm or 30cm (Mohamed et al., 2022; Wei et al., 2020)

Furthermore, the workflow processes shown in Figure 3 are followed to generate DSM, DTM, and nDSM.

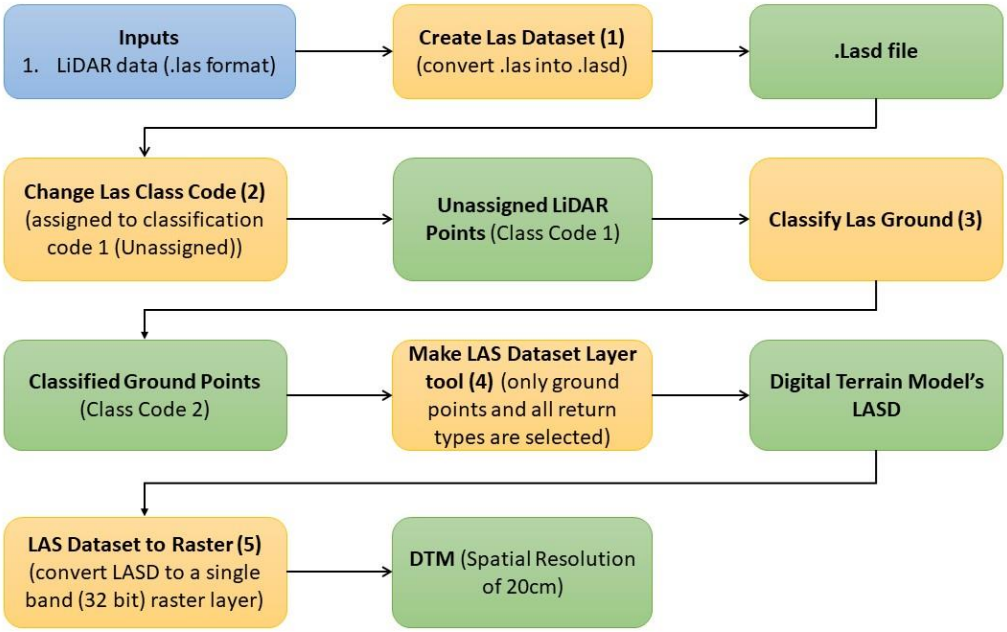


Figure 3: Workflow processes for generating DTM from LiDAR data. Yellow boxes represent ArcGIS geoprocessing tools and green boxes represent generated outputs.

The LiDAR data is in raw format (.las 1.2) and has been preprocessed by the contractor/data provider. The preprocessing conducted by the data provider includes the classification and removal of noise points as well as the classification of overlaps. This dataset is then used to create DSM, DTM, and nDSM. To process the LiDAR data in ArcGIS Pro, the "Create LAS dataset" tool is utilized to convert the format from .las to .lasd. Initially, all LiDAR points are reclassified and assigned to classification code 1 (Unassigned) using the "Change LAS Class Code" tool. Ground points are then classified as such and assigned classification code 2 using the "Classify LAS Ground" tool. The 'Standard Classification' method is used to detect ground points. This method has a tolerance for slope variation that allows it to capture gradual undulations in the ground's topography (Esri, 2023)

From the above-created LAS dataset, two other LAS datasets are generated: DTM and DSM LAS datasets. For the DTM LAS dataset, only ground points and all return types are selected to create an LASD consisting of ground points only. For the DSM LAS dataset, all class codes, along with 'First of Many Return' and '1st Return,' are chosen to create an LASD depicting the first object

the LiDAR encounters (e.g., tree tops, building tops, and ground). The "Make LAS Dataset Layer" tool is employed for this purpose.

Next, the DTM and DSM LAS datasets are converted into a single band (32-bit) raster layer with a spatial resolution of 20cm using the "LAS Dataset to Raster" geoprocessing tool. This conversion results in separate DTM and DSM raster layers. DSM is a three-dimensional (3D) representation of the Earth's surface, including various features like trees and buildings (Zhang et al., 2020). In contrast, DTM represents the bare Earth and excludes above-ground features. When creating DSM and DTM, the interpolation type is binning, the cell (pixel) assignment method is average values, and the void fill method is linear. Furthermore, the DSM is normalized by subtracting the DTM from it. The resulting nDSM raster represents above-ground features in a 3D format.

3.1.4 Creating Training and Validation Sets

The labelled 2D building footprints polygon features, the 20cm aerial imagery, and LiDAR-derived nDSM discussed in chapters 3.1.1 and 3.1.2 are used to create training and validation sets. Training and validation sets are created for each formal residential, industrial area, and, informal settlement. This is to ensure that models for these areas are trained separately since they constitute buildings of different characteristics. Figure 4 shows the three training and validation areas of interest (AOI) chosen to train the Mask R-CNN models.

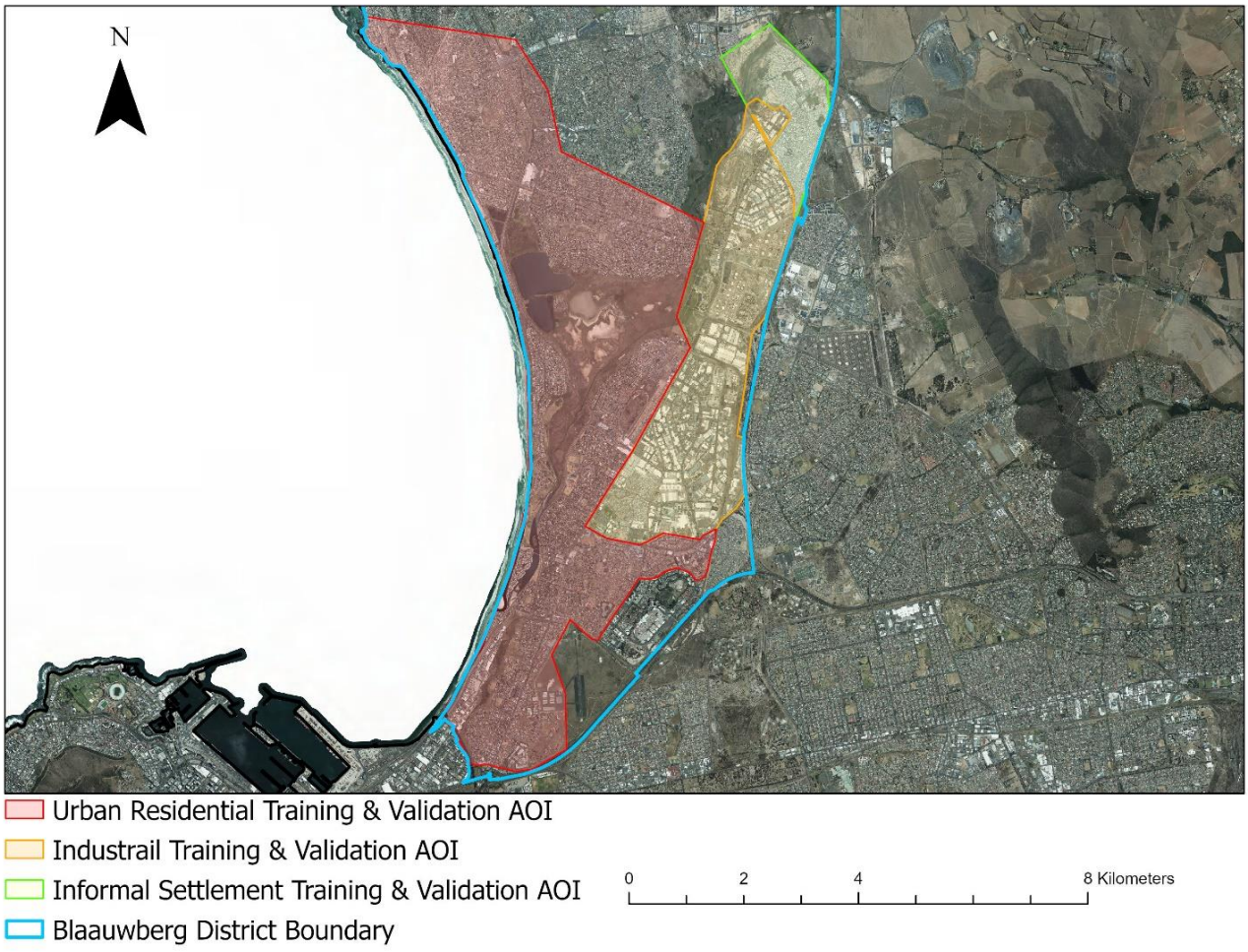


Figure 4: Training and validation areas.

These training areas constitute a total of 21714, 1501, and 14217 labelled building footprints for formal residential, industrial areas, and informal settlements respectively.

A workflow process involved in creating training and validation sets for object instance segmentation is shown in Figure 5.

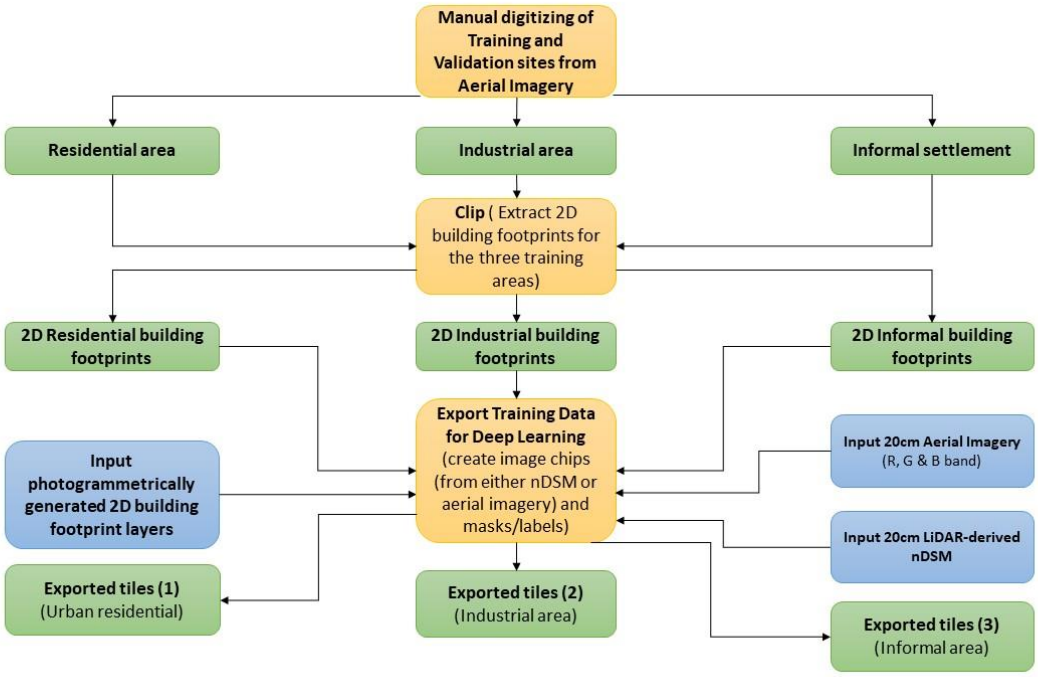


Figure 5: Workflow processes for creating Mask-RCNN's training and validation sets

The labelled 2D building footprint vector files for formal residential, industrial areas, and informal settlements are used in semantic segmentation. These building footprint vector files are used together with either aerial imagery or LiDAR-derived nDSM associated with the label data (vector files) to create deep-learning training datasets. The '" Export Training Data for Deep Learning" geoprocessing tool is used to generate Masks R-CNN's deep learning training datasets using the photogrammetrically generated building footprint vector files and the 20cm aerial imagery associated with the label data. "RCNN Masks" is chosen as a metadata format when running the tool. This step is repeated using the LiDAR-derived nDSM instead of the 20cm aerial imagery. The results are aerial imagery and LiDAR-derived nDSM's deep learning training and validation sets for each formal residential, industrial area, and informal settlement. This resulted in each of the three areas with two sets of training and validation sets from aerial imagery and LiDAR-derived nDSM. As mentioned above, the output from the "Export Training Data for Deep Learning" tool is a folder of tiles, masks, and statistics files. The tiles and masks are cropped into small patches with a size of 256 x 256 pixels containing image chips of rasterized LiDAR or aerial imagery, and masks representing the building footprint in each image chip.

It is worth mentioning that the LiDAR, high-resolution aerial imagery and 2D labeled building footprints used in this research were acquired from the Geospatial Unit of the Information and Knowledge Management department in the City of Cape Town.

3.2 Building Footprint Detection

3.2.1 Architecture of Mask R-CNN:

Object instance segmentation integrates object detection tasks where the goal is to detect objects along with bounding box prediction in an image and semantic segmentation task, which classifies each pixel into pre-defined categories (Esri, 2023). As a result, it enables the detection of objects in a raster while precisely segmenting a mask for each object instance.

In this research, a Mask R-CNN model trained to detect building footprints is used. ResNet-101 is used as the backbone of the model. The model training and inferencing are done through the integration of ArcGIS Pro 3.0 and ArcGIS API for Python in Jupyter Notebook. Mask R-CNN is a state-of-the-art model for instance segmentation, developed on top of Faster R-CNN with an additional branch for predicting segmentation masks on each Region of Interest (RoI). Mask R-CNN uses a fully convolutional network on CNN feature maps to generate a binary mask that identifies if a pixel belongs to an object or not (Kaiming et al. 2020).

In Faster R-CNN, the Rol pool is replaced by RoIAlign to ensure spatial information is preserved which gets misaligned in the case of the Rol pool. RolAlign uses binary interpolation to create fixed-size feature maps. The RolAlign layer's output is fed to the Mask head which consists of two convolution layers. These layers generate masks for each Rol and thus produce a pixel-level segmentation.

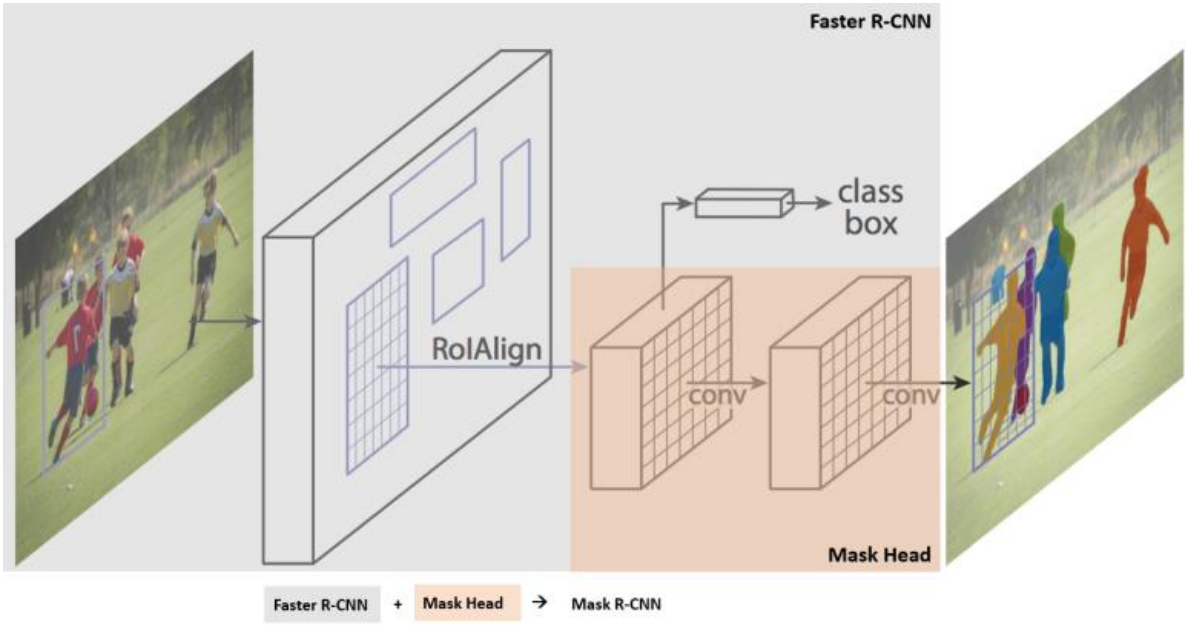


Figure 6: Mask R-CNN (Kaiming et al. 2020)

3.2.2 Mask R-CNN Model Training

The image chips and building masks created in Chapter 3.1.4 are used to perform object instance segmentation using Mask R-CNN. The models for formal residential, industrial areas, and informal settlements are trained using the ArcGIS API for Python in Jupyter Notebook. Pre-trained ResNet101 is used as the backbone of the models. To preserve as many training samples as possible, the original training sets are split into non-overlapping training and validation subsets, by default the validation subset is 0.2 or 20% of the full training data and the remaining

80% goes into the training subset. Traditionally, validation samples are used to verify the loss convergence at the end of each training epoch. Four (4) batch size is used when training all six (6) models. The optimum learning rate is calculated using the lr_find() method. The learning rate is a very important parameter, while training a deep learning model it sees the training data several times and adjusts itself (the weights of the network) (Hafidz, Z., 2018). Too high a learning rate leads to the convergence of the model to a suboptimal solution and too low a learning rate slows down the convergence of the model (Hafidz, Z., 2018). Table 1 shows the optimum learning rates used to train the six (6) individual Mask R-CNN models.

Table 1: Optimum learning rate used to train Mask R-CNN models fast enough

(a) High-Resolution Aerial Imagery

Area Type	Formal Residential	Industrial	Informal Settlements
Learning Rate	3.6308e-05	2.0893e-05	2.5119e-05

(b) LiDAR-derived nDSM

Area Type	Formal Residential	Industrial	Informal Settlements
Learning Rate	3.6308e-05	2.0893e-05	2.5119e-05

The ArcGIS API (arcgis.learn) provides the Mask R-CNN model for instance segmentation tasks. The MaskRCNN() function is used to define a Mask R-CNN model in Jupyter Notebook. All six (6) models are trained using the "model.fit()" function on a single NVidia Quadro T2000 GPU with CUDA 11.7 and 8GB of memory (RAM).

The training and validation loss is calculated from the loss function for each batch processed inside an epoch for all six (6) models. The training loss helps to optimize model parameters during training while validation loss helps to be aware of how well the model generalizes on data that it has never seen and prevents overfitting of the model. The workflow processes followed are detailed in Figure 7 below:

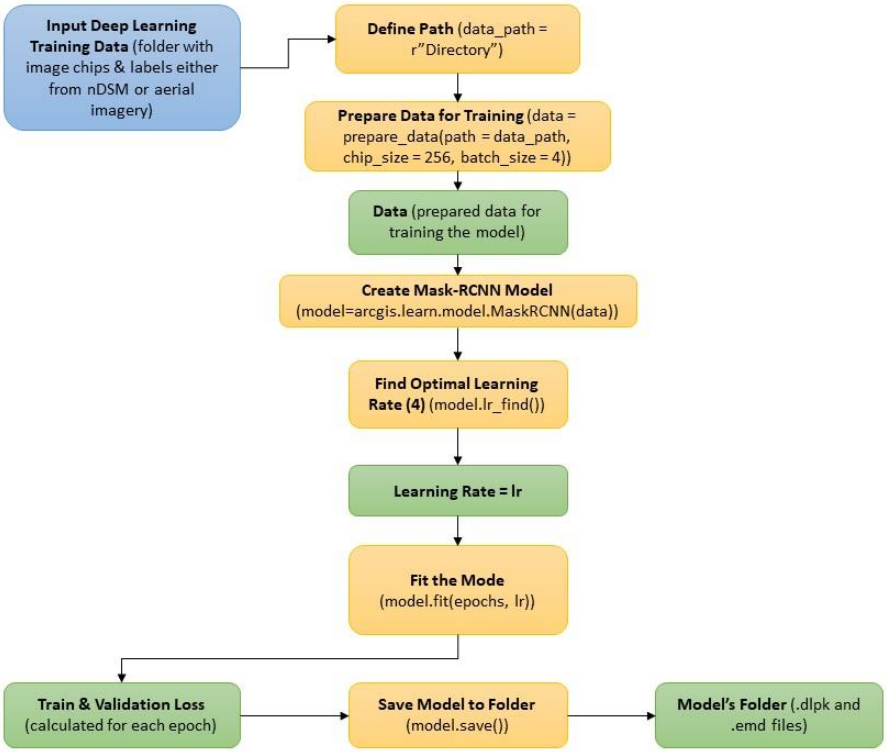


Figure 7: Workflow processes for training the Mask R-CNN model

The "save()" method is used to save Mask R-CNN models and by default, the models are saved to a folder 'models' inside the training data folder. Models are saved in a deep learning model package (.dlpk). The package contains an Esri model definition file (.emd) and a trained model file. The .emd file is a JSON file that describes the trained deep learning model. The file contains the required model definition parameters to run the inference tools. These parameters include deep learning framework, model configuration, model type, inference function, model description, extract bands, labelled training raster classes, projections, labelled training raster, and aerial imagery cell sizes. In addition, the folder contains model metrics used to quantitatively analyze the accuracy of the model segmentation.

3.3 Model Evaluation

In this section, the training and validation loss calculated for each batch of images processed during model training and validation is discussed as well as the model metrics used to quantitatively analyze the performance and accuracy of the models.

3.3.1 Training and Validation Loss

The training and validation loss helps with model optimization and being aware of how the model generalizes on data that it has never seen and prevents overfitting of the model. The loss quantifies the error the model produces. A high loss value means the model's predicted output of a given input is erroneous, while a low loss indicates that there are fewer erroneous predicted outputs by the model. The cross-entropy loss function is used to calculate the training and validation loss. Typically, it is defined as the average of the difference between the model predictions and ground truth for a set of training examples. The cross-entropy loss function is shown in the following equation:

$$\boldsymbol{L(g,p) = -\frac{1}{n}\sum_{i=1}^{m} g^i \ln p^i + (1-g^i)\ln(1-p^i)} \quad (1)$$

Where p^i denotes the predicted probability distribution for category i, g^i denotes the probability distribution of the corresponding ground truth, and m is the total number of training images.

i. ***Training Loss***

Training loss is a metric used to assess how a deep learning model fits training data. In other words, it assesses the error of the model prediction in the training set. Training loss is the value of the loss function which is the sum of the errors for each example in the training set (Kingma et al., 2015). It is used to optimize the model parameters during the training process, the model adjusts its parameters to minimize the value of the loss function. It is calculated from the training set which normally is 80% of the training data. The training set is passed through the neural network in small batches and training loss is measured after each batch has been processed. This is usually visualized by plotting a curve of the training loss.

ii. ***Validation Loss***

Validation loss is a metric used to measure how well a trained model can generalize on new, unseen data that is not used during training. Validation loss is important to ensure that the model is not overfitting to the training data. Hence, a separate set of data called the validation set, which is normally 20% of the training data is used to calculate the validation loss. Validation loss is the value of the loss function on the validation set which is the average loss over all the validation examples (Kingma et al., 2015).

3.3.2 Mask R-CNN Evaluation Metrics

In Mask-RCNN, the Average Precision (AP) Score is taken as a key evaluation indicator to quantitatively analyze the performance and accuracy of the model segmentation (Anwar, 2022). AP is calculated with the help of several other metrics such as Intersection over Union (IoU), confusion matrix (true positive (TP), false positive (FP), false negative (FN)), precision, and recall. Figure 8 shows the road map to calculate the Average Precision score.

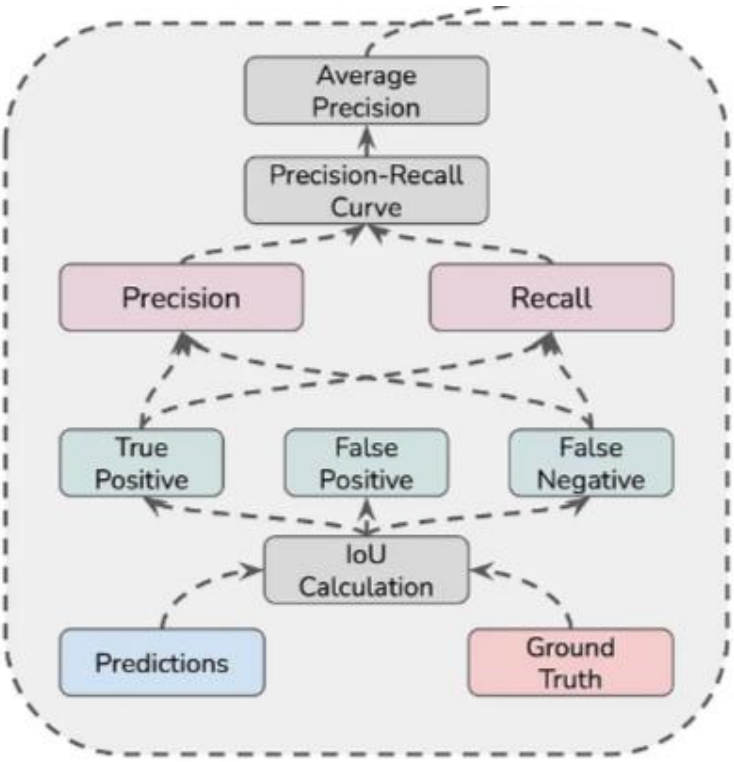


Figure 8: A road map to calculate the Average Precision score of the classification (Anwar, 2022)

i. *Intersection over Union (IoU)*

IoU quantifies the closeness of the prediction to the ground truth. IoU is also referred to as the Jaccard index. IoU metric is the area of overlap between the ground truth and the prediction to the area of union between the ground truth and the prediction. The formula below shows how it is calculated (Anam, 2021).

$$\boldsymbol{IoU} = \frac{\boldsymbol{A} \cap \boldsymbol{B}}{\boldsymbol{A} \cup \boldsymbol{B}} \tag{2}$$

Where A denotes the ground truth value, and B denotes the prediction.

ii. *Confusion Matrix (TP, FP, FN)*

The confusion matrix measures the performance of the model after the classification in a matric form. It shows how many predictions are correct and incorrect per class. It helps in understanding the classes that are being confused by the model as other classes (Tiwari, 2022). Figure 9 shows a sample confusion matrix for the binary classification problem.

		Actual class	
		P	**N**
Predicted class	**P**	**TP**	FP
	N	FN	**TN**

Figure 9: Confusion Matric example (Tiwari, 2022)

True positive is the number of correctly classified and extracted building pixels, false positive is the number of erroneously classified and extracted building pixels, false negative is the number of missed building pixels i.e. building existed but not classified, and true negative is the number of correctly extracted non-building pixels. To calculate true positives, false positives, and false negatives, the IoU threshold value is considered. The prediction-ground truth mask pair is considered to be true positive if it has an IoU score greater than the threshold. Conversely, if it has an IoU score of less than the threshold value, it is considered a false positive. A false negative is when the ground truth mask has no respective predicted mask (Anam, 2021).

iii. ***Precision and Recall***

Precision, Recall, and F1-score are calculated based on TP, FP, and FN. Precision refers to the ratio between correctly classified pixels (true positives) and the total number of pixels that are predicted to be true (equivalently the sum of true positives and false positives). It is a measure that tells what proportion of the buildings that are detected as buildings, were buildings. Recall refers to the ratio between the number of correctly classified pixels (true positives) and the total number of actually correct pixels (equivalently the sum of true positives and false negatives) (Zhang et al., 2022). It is a measure that tells the total proportion of the buildings detected by the model as buildings. In addition, the F1-score is used to balance precision and recall parameters. The equations below show how these metrics are calculated.

$$\boldsymbol{Precision} = \frac{\boldsymbol{TP}}{\boldsymbol{TP} + \boldsymbol{FP}} \tag{3}$$

$$\boldsymbol{Recall} = \frac{\boldsymbol{TP}}{\boldsymbol{TP} + \boldsymbol{FN}} \tag{4}$$

$$\boldsymbol{F_1 - Score} = \mathbf{2} \times \frac{\boldsymbol{Precision} \times \boldsymbol{Recall}}{\boldsymbol{Precision} + \boldsymbol{Recall}} \tag{5}$$

iv. ***Average Precision Score***

The AP is calculated from the Precision-Recall (PR) curve, it is the area under the Precision–Recall curve. For each precision-recall pair, the area under the PR curve can be found by approximating the curve using rectangles. The higher the precision and recall, the higher the AP (Anwar, 2022). The formula below shows how it is calculated.

$$\boldsymbol{Average\ Precision}\ (\boldsymbol{AP}) = \int_{\boldsymbol{r=0}}^{\mathbf{1}} \boldsymbol{p}(\boldsymbol{r})\boldsymbol{dr} \tag{6}$$

Where *p* denotes precision*, r* denotes recall and *d(r)* shows that the equation is being integrated with respect to variable *r*, which is recall.

3.4 Inferencing and Post-processing

The high-resolution aerial imagery and LiDAR-derived nDSM with 20cm spatial resolution of the formal residential, industrial area, and informal settlement testing areas are passed through a neural network to extract footprints in those areas. The testing areas also contain human-digitized building footprints and they are used as ground truth masks together with the prediction results to evaluate and analyze the performance of each Mask R-CNN model.

3.4.1 Inferencing

To extract building footprints from high-resolution aerial imagery and LiDAR-derived nDSM rasters while segmenting a mask for each building footprint instance precisely 'Detect Objects Using Deep Learning" geoprocessing tool in ArcGIS Pro is used. The result is a polygon feature class in a file geodatabase of raw building footprints detected from the input raster using trained Mask R-CNN models.

3.4.2 Boundary Regularization

The raw detected polygons show irregular and noisy outlines due to the locality of pixel-wise labeling conducted by Mask R-CNN. In addition, when a neural network is used for pixel-level semantic segmentation, the output-building boundaries are irregular (Xie et al. 2020). To convert the initial polygons into regularized ones, an advanced ArcGIS's "Regularize Building Footprints" geoprocessing tool is used. Building footprints were regularized using the right-angles and diagonals method, and the densification (sampling interval) and tolerance values of 0.5m were used. Figure 10, shows the workflow process of detecting building footprints using the 'Detect Object Using Deep Learning' and detected building footprints regularization using 'Regularize Building Footprint'. These processes are combined and executed using an ArcGIS model builder. ArcGIS models are workflows that string together sequences of geoprocessing tools. They can be thought of as a visual programming language for building workflows.

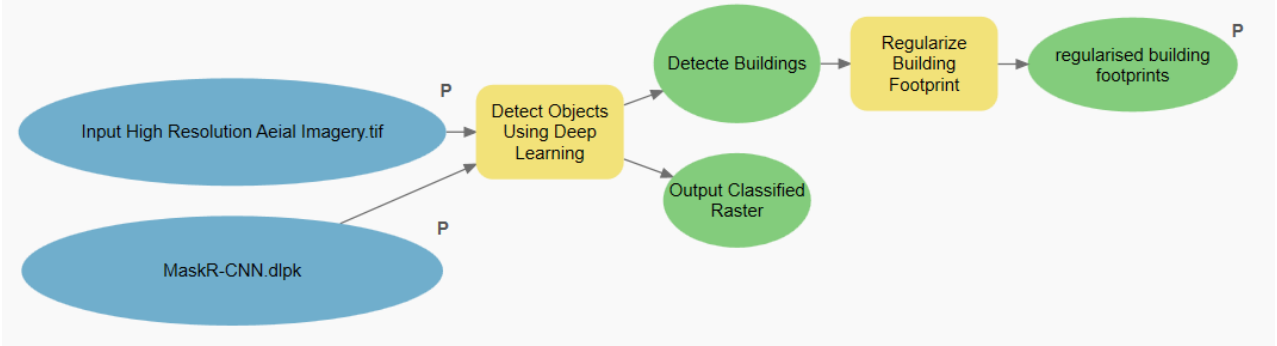


Figure 10: ArcGIS Model for Detecting Building Footprint using Mask R-CNN, Regularize and Refine detected building footprint. Blue colour represents inputs. Yellow colour represents ArcGIS geoprocessing tools. The green colour represents generated outputs.

The results from this ArcGIS model are two 2D building footprints polygon feature classes for each formal residential, industrial, and informal settlement from LiDAR-derived nDSM and high-resolution aerial imagery. These results are thoroughly analyzed and discussed in Chapter 4.

4. Results and Discussion

In this chapter, research question 2 is addressed: How can remote sensing data be effectively used to accurately extract building footprints? In addition, research objective 3 is addressed. This is done by discussing and evaluating the performance of Mask R-CNN in extracting building footprints from aerial imagery and LiDAR-derived nDSM for different cases, including formal residential areas, industrial areas, and informal settlements. The results are analyzed based on the performance and evaluation metrics discussed in Section 3.3. The rest of this chapter is structured as follows:
Section 4.1 outlines the steps followed to preprocess aerial imagery and generation of LiDAR-derived nDSM and presents the preprocessing results. It further presents the samples of the 2D labelled training datasets.

Section 4.2 presents the preprocessed training and validation datasets used in the training and validation of the Mask R-CNN models.

Section 4.3 presents and discusses model training results, it starts by presenting and discussing the training and validation loss results of each trained Mask R-CNN model. Then, it presents and discusses the model evaluation metrics on the training and testing dataset.

Section 4.4 provides an analysis of the experiment. The effectiveness of the Mask R-CNN in extracting building footprints in formal residential, industrial areas, and informal settlements from high-resolution aerial imagery and LiDAR-derived nDSM is discussed and analyzed.

4.1 Data Preprocessing

4.1.1 Aerial Imagery resampling and nDSM generation

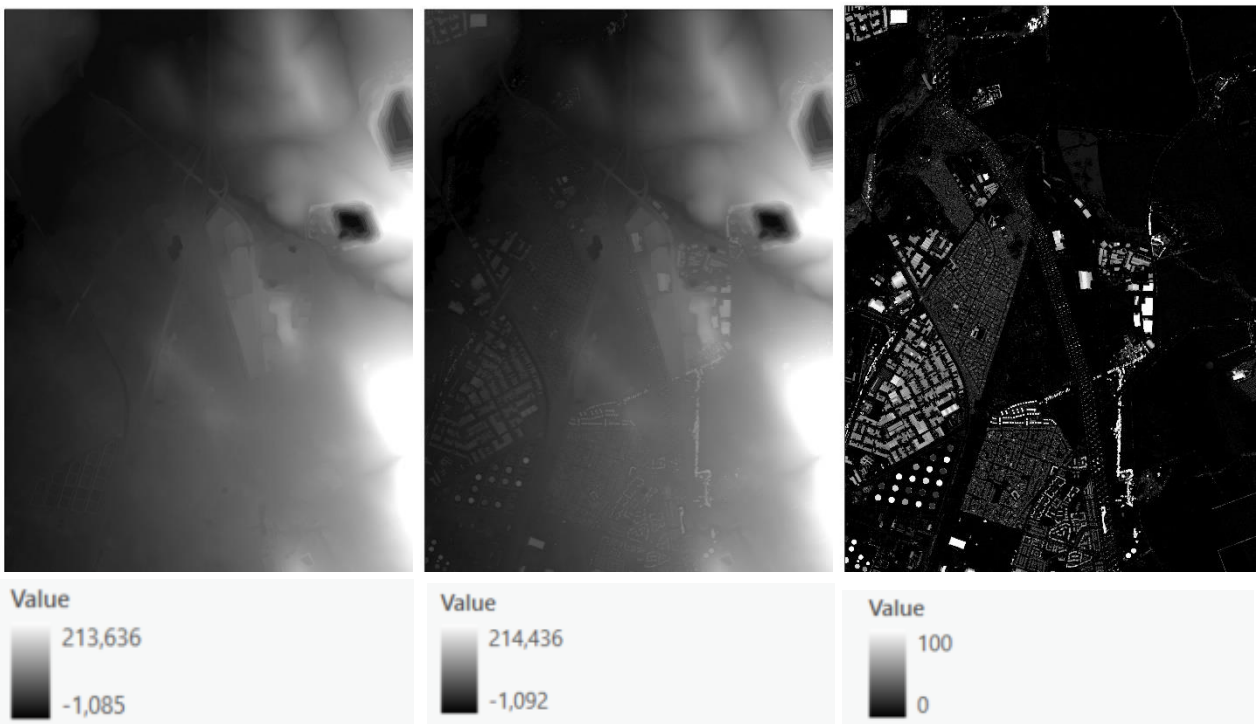
The high-resolution aerial imagery used in this research is in .ecw format. The aerial imagery has a spatial resolution of 8cm. As discussed in Section 3.1.2, the aerial imagery is subsampled from 8cm to 20cm spatial resolution using Global Mapper Pro v24. This is done by loading the .ecw aerial imagery with a spatial resolution of 8cm in Global Mapper Pro and then exporting the loaded raster file as GeoTiff. Before exporting the subsampled Geotiff, the following parameters are set as followed:

- For Sampling space or scale, 20cm is used for both X-axis and Y-axis.
- The Box average resampling method is used when subsampling, this is a default resampling method used in Global Mapper when subsampling. This resampling method finds average values of the nearest 9 (for 3x3), 16 (4x4), 25 (5x5), or 49 (7x7) pixels and uses that as the value of the sample location. The method produces good results when resampling data at a lower resolution (bluemarblegeo, 2023).
- LZW Compression is used by default, the exported GeoTiff is compressed using the lossless LZW algorithm.

The resulting aerial imagery is a three bands GeoTiff with a spatial resolution of 20cm. Figure 12 below shows a sample of the downscaled aerial imagery in the GeoTiff format

Figure 11: A sample of downsampled GeoTiff aerial imagery used for training, validation, and testing of the deep learning models.

In addition, DTM, DSM, and nDSM of the Blaauwberg district are generated from the LiDAR data as per the workflow processes shown in Figure 3 in Section 3.1.2. Figure 12 below shows the extracts from the generated DTM (a), DSM (b), and nDSM (c) of the Blaauwberg district.



(a) (b) (c)

Figure 12: A sample of the three digital models with their height ranges, including (a) DTM, (b) DSM, and (c) nDSM with a spatial resolution of 20cm.

The nDSM is a normalized DSM and it is also used for training, validating, and testing the Mask R-CNN models in addition to the subsampled aerial imagery. The normalized DSM is used instead of DSM because it allows for more efficient training of the neural network, as it removes the dependency on the surface elevation, making the height range much more compact and dense. As a result, it improves the Average Precision score with fewer training samples (Esri, 2020).

4.1.2 2D Labelled Building Footprints for Model Training and Validation

For the training and validation of the deep learning, models used to extract building footprint in this research, the 3D building shapes created using stereo-images in socetGXP software are converted into 2D labelled building footprints using the "Multipatch Footprint" geoprocessing tool in ArcGIS Pro, see Section 3.1.1 and Figure 2 for the workflow processes followed to generate this. Figure 13 below shows samples of the 2D labelled building footprints used to generate the training and validation sets in conjunction with either the aerial imagery or LiDAR-derived nDSM for both formal residential, industrial area and informal settlements.

(a) **(b)** **(c)**

Figure 13: Labelled 2D Building Footprint for Training and Validation, for (a) Formal Residential, (b) Industrial Area, and (c) Informal Settlements

4.2 Training dataset

In total, 21,714 labeled building footprints for formal residential areas, 1,501 for industrial areas, and 14,217 for informal settlement areas have been prepared. These footprints are used for training and validating Mask R-CNN models. This training dataset includes both structured and unstructured buildings with varying roofing materials, shapes, and widths. The formal

residential areas are characterized by structured building roofs with different shapes, sizes, and roofing materials. Similarly, the industrial areas comprised structured building roofs with different shapes and roofing materials, but these buildings are larger and usually have brighter roofs compared to residential buildings. As a result, the trained Mask R-CNN model is capable of distinguishing between these types of buildings when segmenting their footprints. In contrast, the informal settlements consisted of unstructured shacks with bright roofs, built in close proximity to each other. Therefore, the Mask R-CNN can differentiate shacks from formal residential and industrial buildings.

In total, six (6) training datasets are generated for training Mask R-CNN models, three from 20cm aerial imagery and another three from LiDAR-derived nDSM. The training data is divided into training and validation sets. The training set accounted for 80% of the data, while the validation set made up the remaining 20%. These sets are used to train and validate the Mask R-CNN models, as outlined in Section and Figure 4.
For each type of area (formal residential, industrial, and informal settlement), the training sets from both the 20cm aerial imagery and LiDAR-derived nDSM consisted of 5,097, 2,094, and 452 image chips of size 256x256 pixels, along with their corresponding masks. Similarly, the validation sets included 1,274, 524, and 113 image chips of the same size and their respective building masks.

Figure 14 below shows an example of image chips from input 20cm aerial imagery and labels, and Figure 15 shows an example of cropped image chips from input LiDAR-derived nDSM and masks used for training and validation of the Mask R-CNN models.

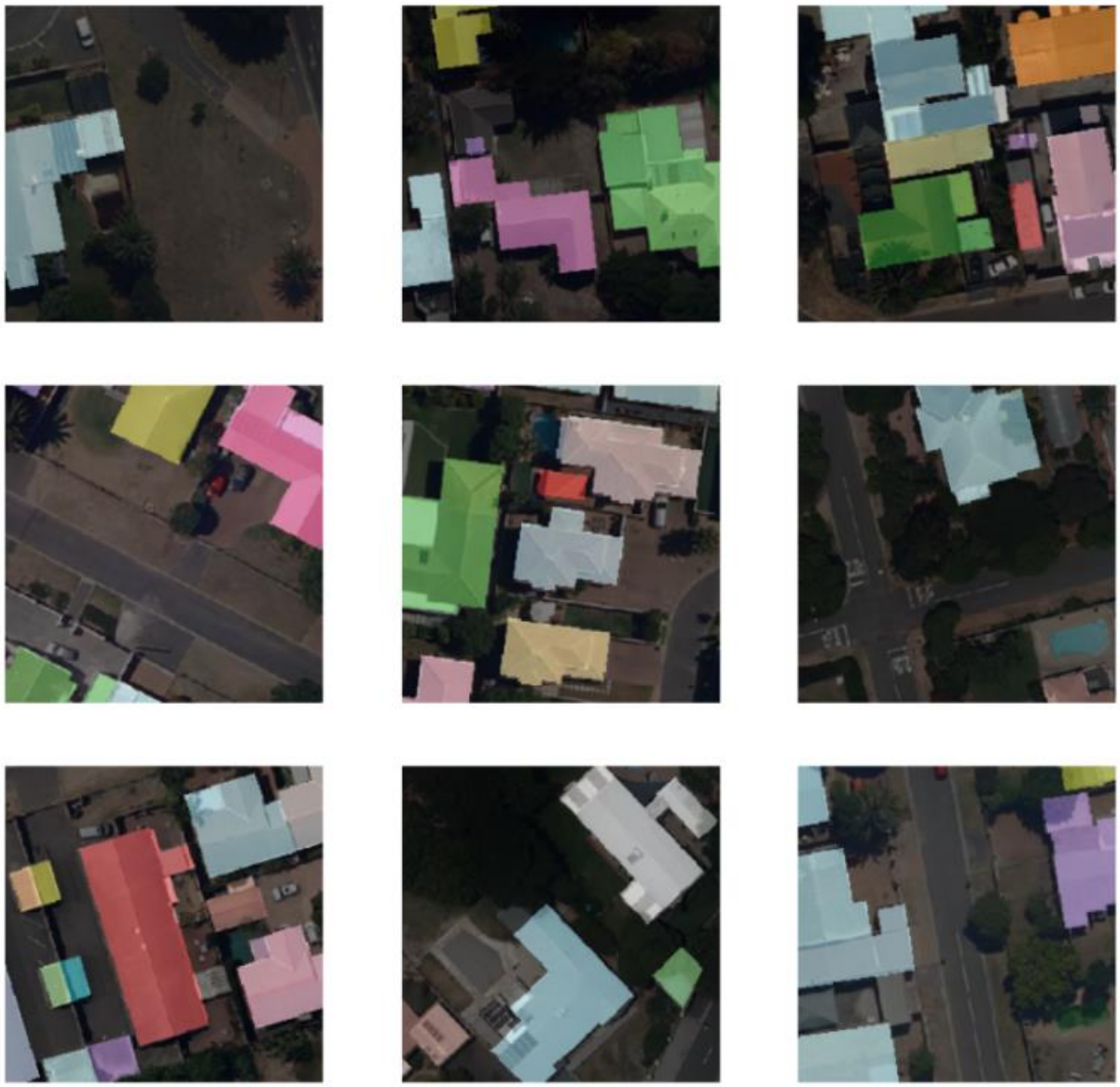


Figure 14: An example of 256 x 256 pixels image chips of input aerial imagery (background) and corresponding labelled building footprint mask (overlay).

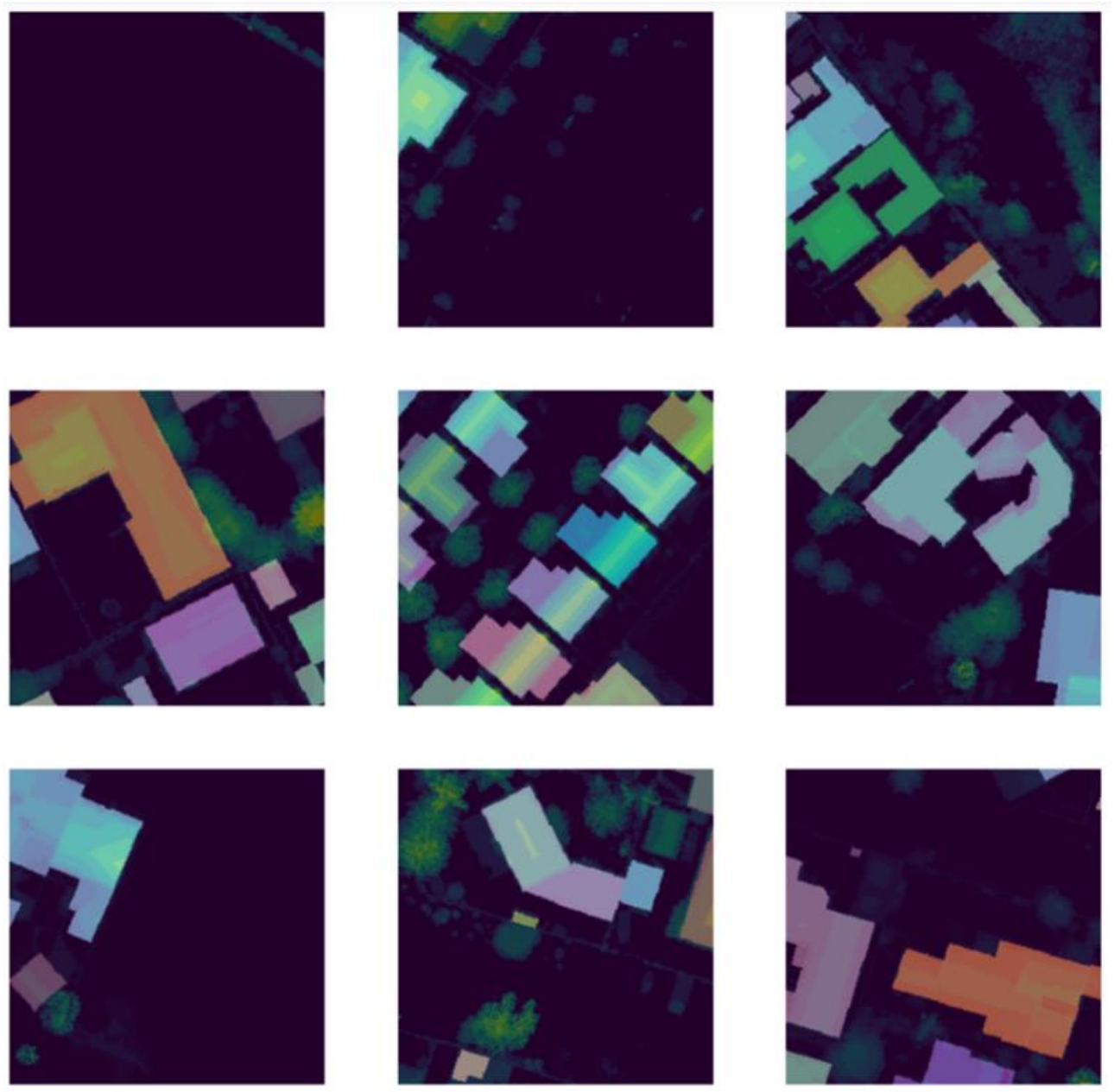


Figure 15: An example of 256 x 256 pixels image chips of input LiDAR-derived nDSM raster (background) and corresponding building footprint mask (overlay).

4.3 Building Footprint Detection Results

4.3.1 Object Instance Segmentation

In this study, Mask R-CNN is used to detect building footprints from a 20cm aerial imagery and LiDAR-derived nDSM. ResNet101 is used as the backbone of the models. Six (6) models are trained separately, two for each formal residential, industrial area, and informal settlement from aerial imagery and LiDAR-derived nDSM. For each of these models, 30 epochs have been run during the training period using the training and validation sets presented in Section 4.2. This means for each epoch, the model sees the complete training set once, and so on. A batch size of 4 has been used for training all the Mask R-CNN models. Batch size is the number of images a model will train on each step inside an epoch. Table 2 shows the time each model took to learn from the training dataset.

Table 2: Number of epochs and duration of training Mask R-CNN using (a) high-resolution aerial imagery and (b) LiDAR-derived nDSM.

(c) High-Resolution Aerial Imagery

Area Type	Formal Residential	Industrial	Informal Settlements
Number of epochs	30	30	30
Training duration	12.5 Days	7 Days	30hrs

(d) LiDAR-derived nDSM

Area Type	Formal Residential	Industrial	Informal Settlements
Number of epochs	30	30	30
Training duration	12.5 Days	7 Days	30hrs

4.3.2 Training and Validation Loss Curve

The training and validation loss for all six (6) Mask R-CNN models are calculated using the loss function discussed in Section 3.3.1 and results are presented in this section. Training and validation loss is calculated for each batch of training data passed through neural networks inside an epoch. Training loss helps to optimize the model parameters, while validation loss helps to monitor the generalization performance of the model on unseen data and prevent overfitting.

i. Training and Validation Loss Curve for Formal Residential

The training and validation curves for formal residential models trained with ResNet101 as the backbone from aerial imagery and LiDAR-derived nDSM are shown in Figure 16 and Figure 17.

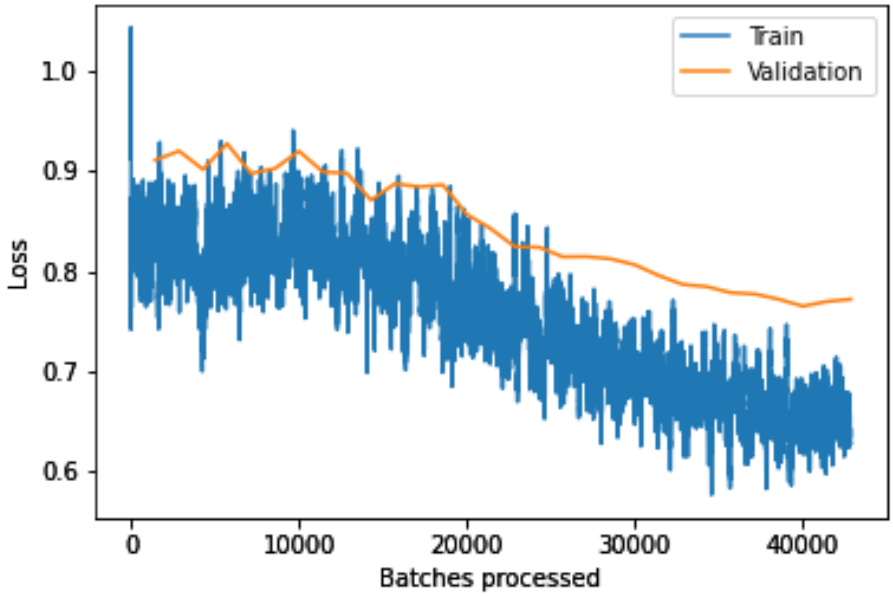


Figure 16: Training and validation loss curve for a formal residential model trained from aerial imagery and ResNet101 as the backbone.

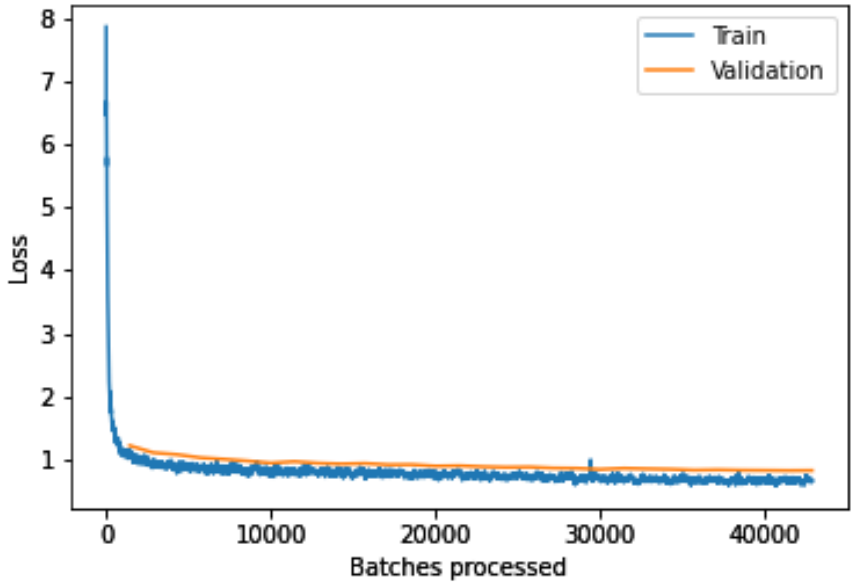


Figure 17: Training and validation loss curve for a formal residential model trained from LiDAR-derived nDSM and ResNet101 as the backbone.

During the 30 epochs, both the formal residential models are trained using LiDAR-derived nDSM and aerial imagery. Over 40,000 batches of image chips and masks are processed. The training and validation loss is plotted against the number of batches processed. As the model encountered more training sets, both the training and validation loss decreased, indicating improvement as more training datasets were seen.

After processing 40,000 batches, the validation loss in Figure 16 started to slightly increase, while the training loss continued to decrease. This suggests that the model is starting to overfit the training data. To prevent overfitting, the model training is stopped. In Figure 17, the training and validation loss exhibited a continuous decrease and the curve converged well, indicating a good fit.

ii. ***Training and Validation Loss Curve for Industrial***

The train and validation curves for the model trained from aerial imagery and LiDAR-derived nDSM are shown in Figure 18 and Figure 19 for the industrial area models trained with ResNet101 as the backbone.

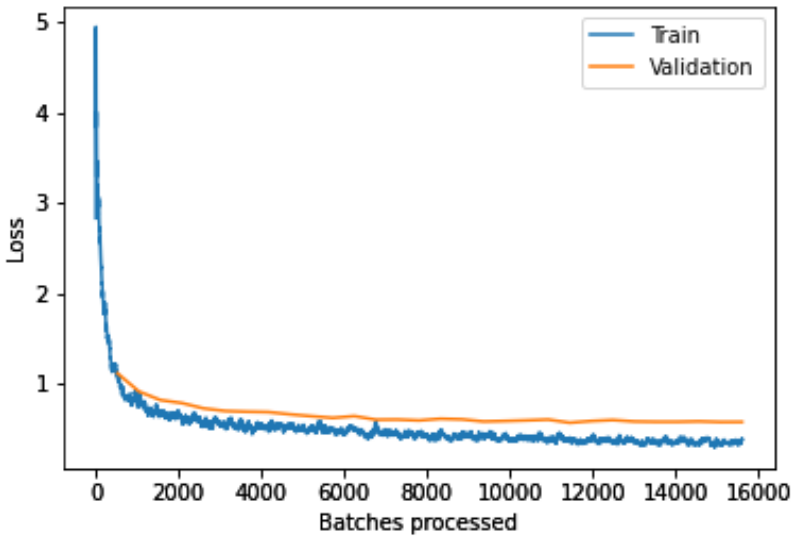


Figure 18: Training and validation loss curve for an industrial model trained from aerial imagery and ResNet101 as the backbone.

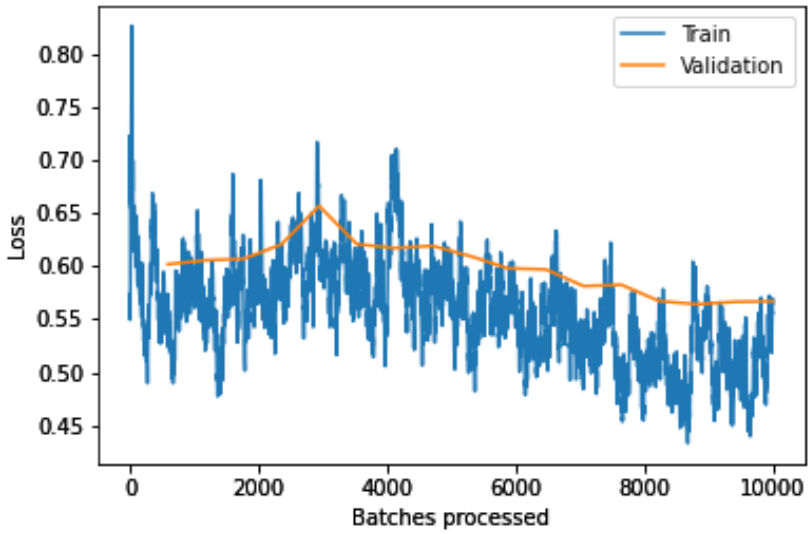


Figure 19: Training and validation loss curve for an industrial model trained from LiDAR-derived nDSM and ResNet101 as the backbone.

During 30 epochs, the industrial model is trained using aerial imagery. Over 15,000 batches of image chips and masks are processed. It is worth noting that the validation loss is higher than the training loss. This suggests that the model may be underfitting, which means it is unable to accurately represent the training data. This is typically caused by a lack of sufficient training data.

For this study, 1,501 labeled building footprints are used to train the industrial models. Interestingly, when the industrial model is trained with LiDAR-derived nDSM using the same number of labeled building footprints as the aerial imagery model, both the training and

validation losses decreased. The training and validation loss curve indicates that the model neither overfits nor underfits (see Figure 18)

iii. Training and Validation Loss Curve for Informal Settlement

The training and validation curves for informal settlement models trained with ResNet101 as the backbone from aerial imagery and LiDAR-derived nDSM are shown in Figure 20 and Figure 21.

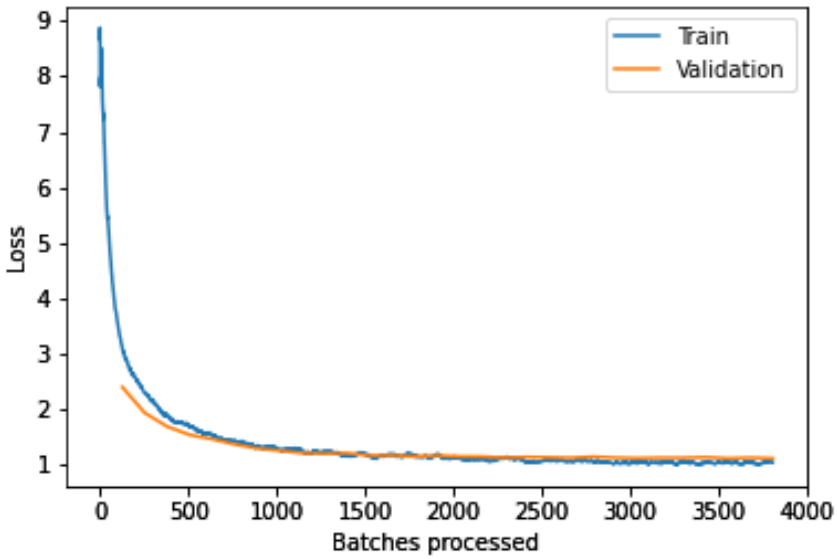


Figure 20: Training and validation loss curve for informal settlement model trained from aerial imagery and ResNet101 as the backbone.

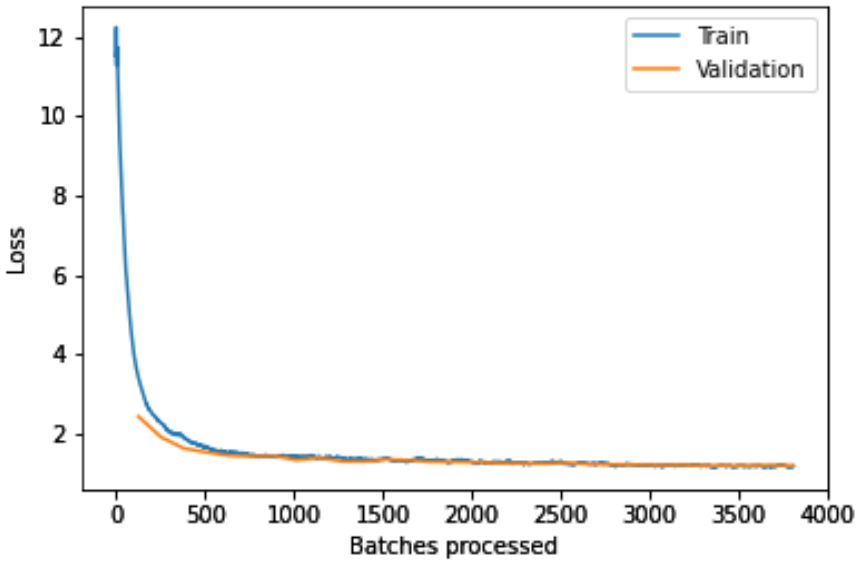


Figure 21: Training and validation loss curve for an informal settlement model trained from LiDAR-derived nDSM and ResNet101 as the backbone.

For the 30 epochs both the informal settlement models trained with LiDAR-derived nDSM and aerial imagery, over 3500 batches of image chips and masks are processed. The plotted graphs, Figure 20 and Figure 21 demonstrate that as the models are exposed to more training data, both the training and validation loss decrease. This implies that the models improve with more exposure to the training dataset. Additionally, Figure 20 and Figure 21 illustrate a consistent

decrease in both the training and validation loss. After approximately 3500 batches, the loss stabilizes, indicating an optimal fit and suggesting that the models neither overfit nor underfit.

4.3.3 Models Evaluation on Training Dataset

In this section, the results of the evaluation of Mask R-CNN models on the training datasets are presented and discussed. The AP, which can consider both incorrect detection (false positive) and missed detection (false negative), has become the key evaluation metric for Mask R-CNN (Anwar, 2022).

For each Mask R-CNN model trained on aerial imagery and LiDAR-derived nDSM for 30 epochs, the Average Precision score is measured. In addition, the ground truth building masks versus Mask –RCNN's predictions on training datasets are presented.

i. Formal Residential Results

Figure 22: Ground Truth versus Prediction for a model trained on aerial imagery for formal residential.

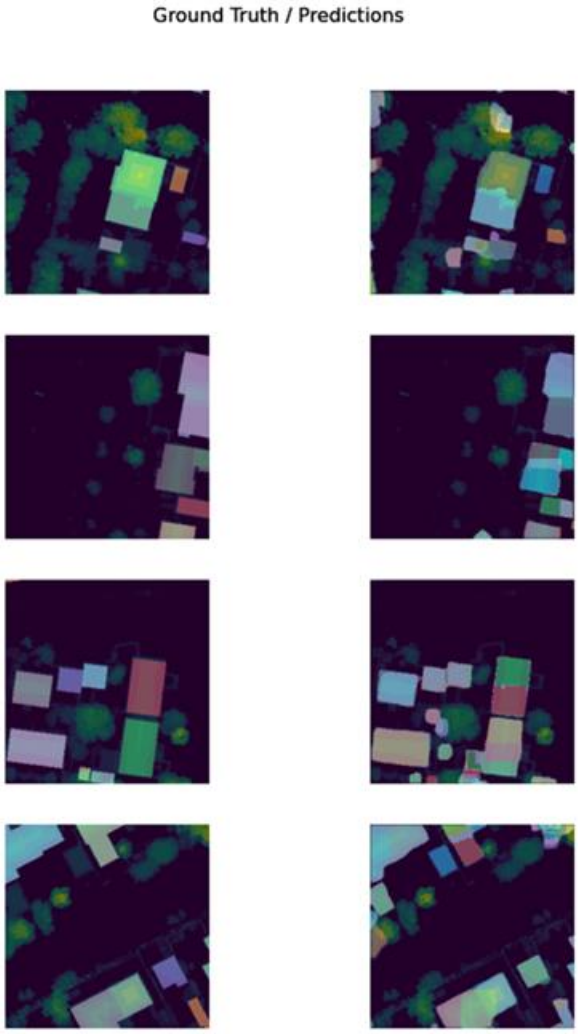


Figure 23: Ground Truth versus Prediction for a model trained on LiDAR-derived nDSM for formal residential.

Figure 22 and Figure 23 show that the Mask R-CNN model is detecting formal residential buildings well and it can also be seen that the Mask R-CNN works not only on high-resolution aerial imagery but also nDSM representing depth information can be used for formal residential building footprint extraction. The Average Precision score comparison for formal residential models from aerial imagery and LiDAR-derived nDSM is shown in Table 3.

Table 3: Average Precision Score of Aerial Imagery and LiDAR-derived nDSM Mask R-CNN models for formal residential.

Mask R-CNN	
Data Type	**Average Precision Score**
Aerial Imagery (RGB)	0.74
LiDAR-derived nDSM	0.73

The Average Precision score of the aerial imagery and LiDAR-derived nDSM models, measured from the training datasets in Table 3, is quite similar. Both scores are satisfactory, indicating high precision and recall. This suggests that these models can effectively handle false positives (incorrect detections) and true positives (correct detections). As a result, these models can

accurately identify and categorize formal residential buildings using both aerial imagery and LiDAR-derived nDSM.

ii. ***Industrial Results***

Figure 24: Ground Truth versus Prediction for a model trained on aerial imagery for industrial areas.

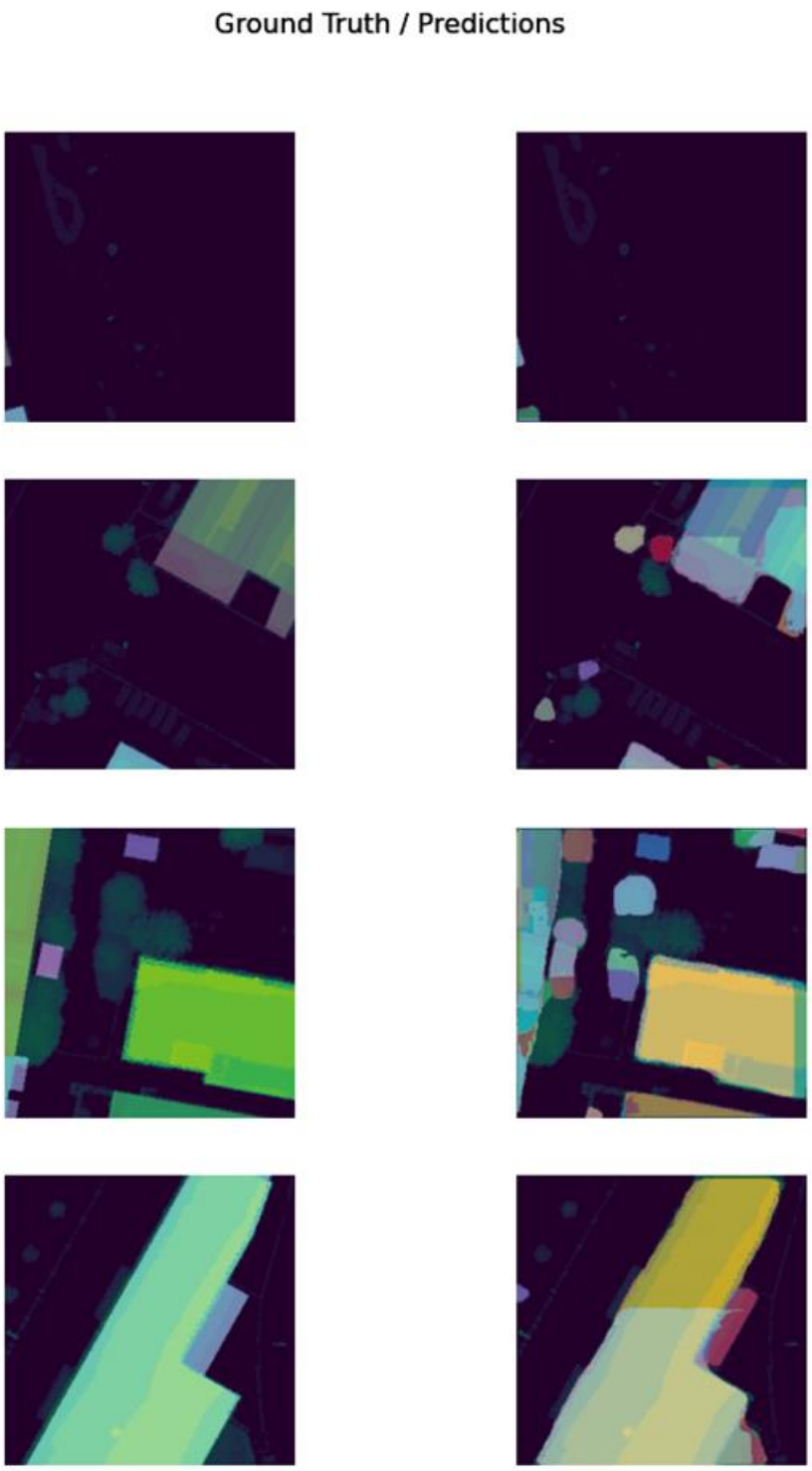


Figure 25: Ground Truth versus Prediction for a model trained on LiDAR-derived nDSM for industrial areas.

Figures 24 and 25 demonstrate the effective use of the Mask R-CNN algorithm in detecting and segmenting industrial buildings from aerial imagery and LiDAR-derived nDSM. However, both the aerial imagery and LiDAR-derived nDSM models exhibit false positives. The aerial imagery model sometimes mistakes tarred roads for buildings, while the LiDAR-derived nDSM model occasionally identifies trees as buildings. Table 4 presents a comparison of the Average Precision scores for the aerial imagery and LiDAR-derived nDSM industrial models.

Table 4: Average Precision Score of Aerial Imagery and LiDAR-derived nDSM Mask R-CNN models for industrial areas.

Mask R-CNN	
Data Type	**Average Precision Score**
Aerial Imagery (RGB)	0.79
LiDAR-derived nDSM	0.80

Based on the comparison, both the aerial imagery (RGB) and LiDAR-derived nDSM models have similar Average Precision scores obtained from the training datasets. Their AP is higher compared to the AP of formal residential models. As previously discussed, a high AP signifies high precision and recall. It also suggests that these models can effectively handle false positives (incorrect detections) and true positives (correct detections). As a result, these models can accurately detect and classify industrial buildings using both aerial imagery and LiDAR-derived nDSM data.

iii. ***Informal Settlement Results***

Figure 26: Ground Truth versus Prediction for a model trained on aerial imagery for informal settlements

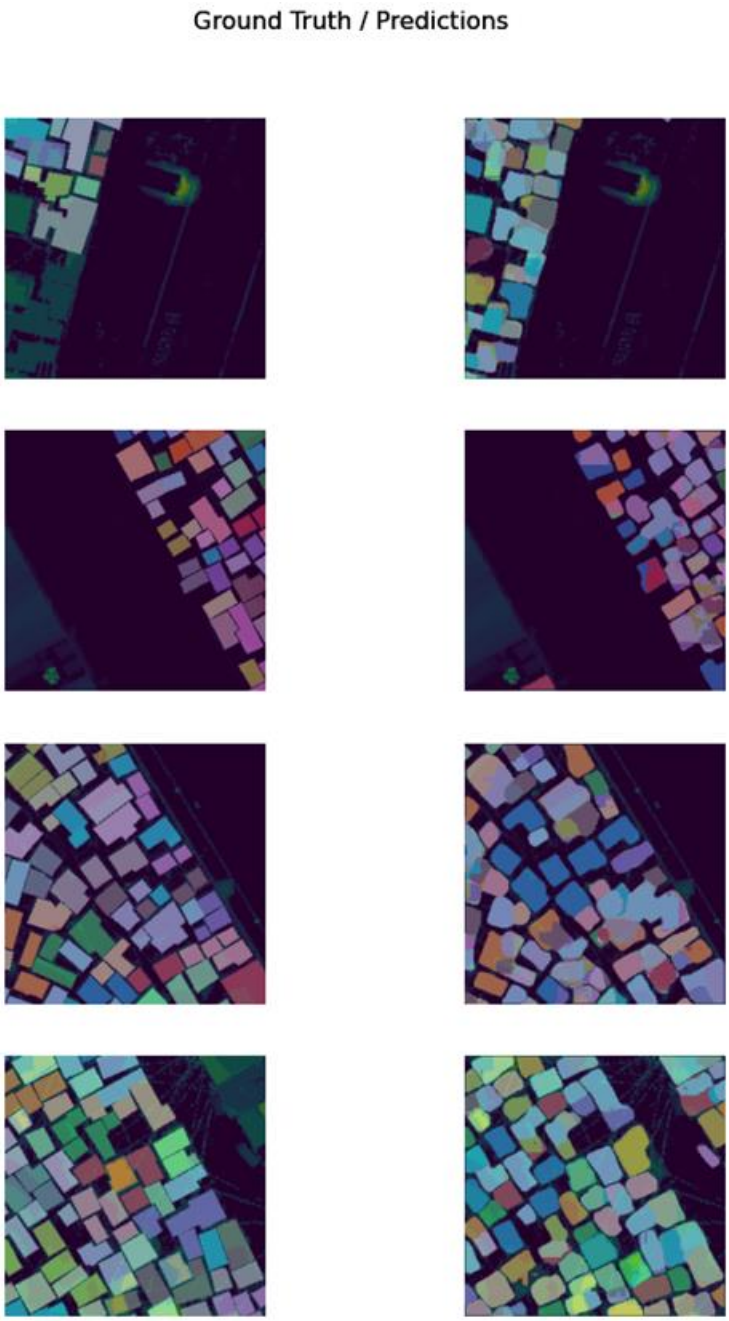


Figure 27: Ground Truth versus Prediction for a model trained on LiDAR-derived nDSM for informal settlements

Figures 26 and 27 demonstrate that the Mask R-CNN algorithm effectively detects and segments shacks in informal settlements using both aerial imagery and LiDAR-derived nDSM. Table 5 presents the comparison of Average Precision scores for the informal settlement models based on aerial imagery and LiDAR-derived nDSM.

Table 5: Average Precision score of Aerial Imagery and LiDAR-derived nDSM Mask R-CNN models for informal settlements.

Mask R-CNN	
Data Type	**Average Precision Score**
Aerial Imagery (RGB)	0.71
LiDAR-derived nDSM	0.65

From the comparison, the Average Precision score measured from the aerial imagery model is slightly higher than the one measured from the LiDAR-derive nDSM model. As discussed before, high AP indicates high precision and recall. The AP comparison in Table 5 indicates that the aerial imagery model can handle false positives and true positives better than the LiDAR-derived nDSM model.

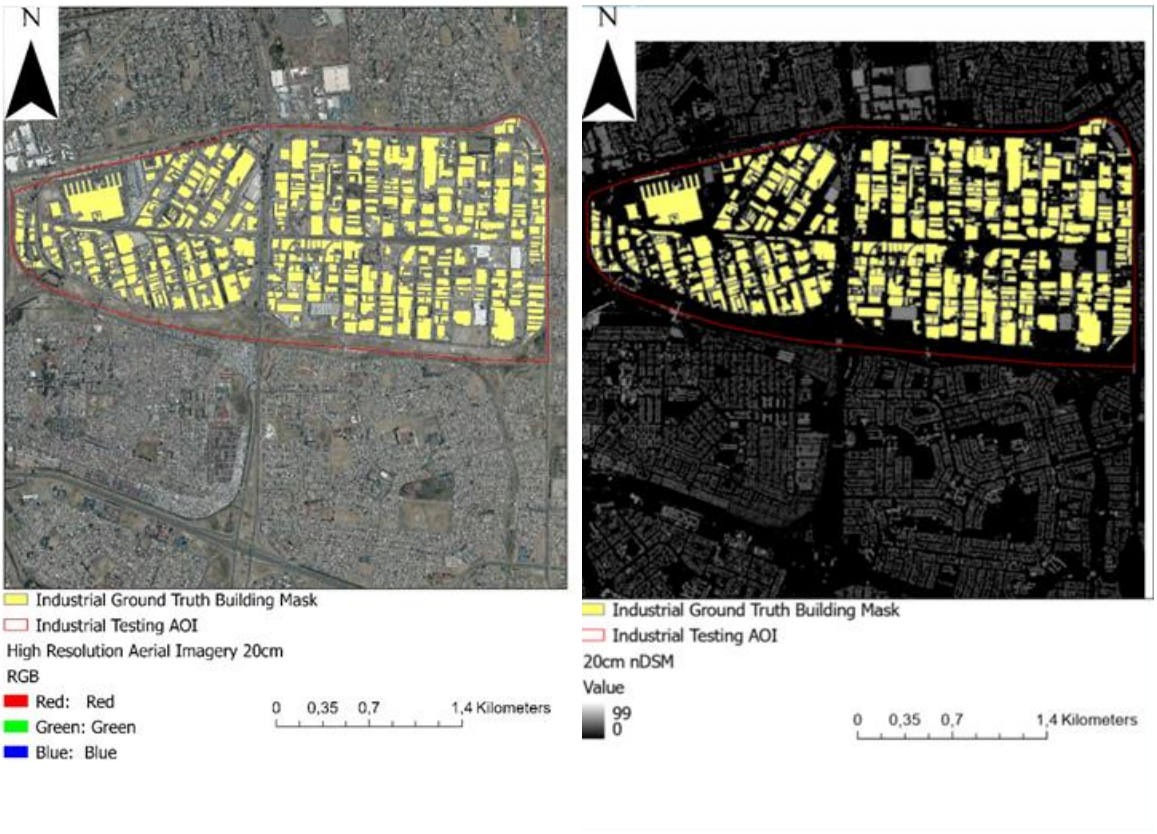
4.3.4 Model Performance Analysis

In this section, the performance of Mask R-CNN models is thoroughly analyzed. Models' ability to accurately extract building footprints from aerial imagery and LiDAR-derived nDSM. Additionally, an analysis of the factors influencing models' performance is discussed. The evaluation of the Mask R-CNN models' performance is based on metrics obtained from the testing dataset. To ensure a fair comparison, a new test dataset, unseen during training or validation, is used. The evaluation metrics used for accuracy analysis include precision, recall, F1-score, and AP score. These metrics are calculated using an IoU threshold of 0.5.

For reference, Figure 28, Figure 29, and Figure 30 depict the test areas for formal residential, industrial, and informal settlements, respectively. These areas have known building footprints that serve as ground truth building masks to evaluate the performance of the Mask R-CNN models. In addition to using these test areas with known building footprints for performance analysis, the trained models are applied across various urban residential areas, industrial areas, and informal settlements. This broader application helps provide a better understanding of the models' extraction performance in different spatial contexts.

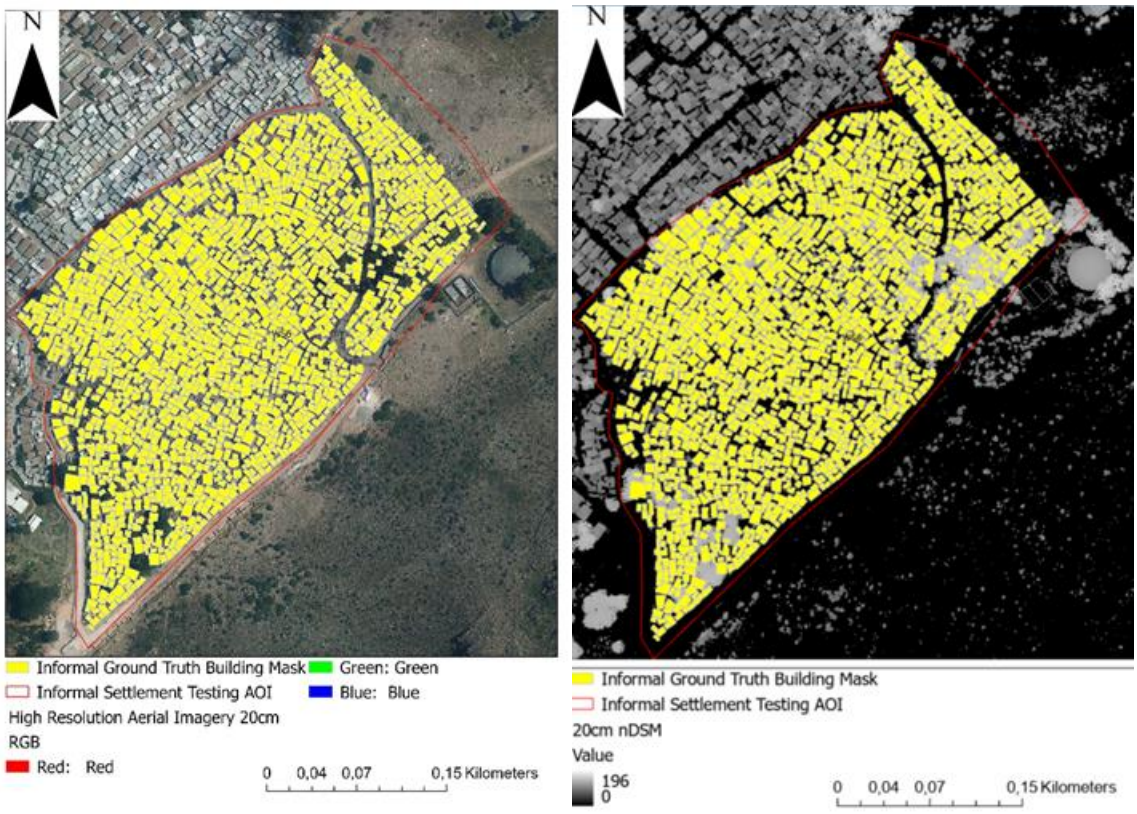
(a) (b)

Figure 28: Formal residential test area with ground truth building mask. (a) aerial imagery. (b) LIDAR-derived nDSM



(a) (b)

Figure 29: Industrial test area with ground truth building mask. (a) aerial imagery. (b) liDAR-derived nDSM



(a) (b)

Figure 30: Informal settlement test area with ground truth building mask. (a) aerial imagery. (b) liDAR-derived nDSM

i. Aerial Imagery Versus LiDAR-derived nDSM Extraction Results for Formal Residential

In this case, the two trained Mask R-CNN for aerial imagery and LiDAR-derived nDSM have been applied to the testing formal residential dataset as shown in Figure 28. The models performed well, as demonstrated in Figure 31 and Figure 32. The results indicate that the Mask R-CNN algorithm effectively extracts formal residential building footprints of various sizes and shapes.

(a) (b)

Figure 31: Formal Residential Building footprint extraction results from high-resolution aerial imagery (a) and LiDAR-derived nDSM (b).

In addition to the testing sample block, the models have been applied in various formal residential areas to gain a broader and improved understanding of their performance in different spatial contexts. The additional testing sample blocks lack the ground truth building mask, so the extracted results are compared to the background aerial imagery to evaluate their performance in those areas. Figure 32 shows results extracted across different formal residential areas within the City of Cape Town metropolitan.

Sunningdale (a) Sunningdale (b)

Milnerton (a)

Milnerton (b)

Houtbay (a)

Houtbay (b)

Figure 32: Formal Residential Building footprint extraction results from high-resolution aerial imagery (a) and LiDAR-derived nDSM(b) across Sunningdale, Milnerton, and Houtbay residential areas.

The footprints extracted across these additional testing sample areas as shown in Figure 32, show that the Mask R-CNN models perform effectively in extracting formal residential buildings from high-resolution aerial imagery and LiDAR-derived nDSM.

Aerial Imagery Versus LiDAR-derived nDSM Extraction Results for Industrial Areas

For this case, two trained models are used: Mask R-CNN for aerial imagery and LiDAR-derived nDSM. They are applied to the industrial dataset for testing, as depicted in Figure 29. The LiDAR-derived nDSM model yielded excellent results, as shown in Figure 33(b) when compared to the results obtained from aerial imagery in Figure 33(a). These results indicate that the Mask R-CNN performs well in accurately extracting industrial buildings of various sizes and shapes from LiDAR-derived nDSM. Further details and evaluation metrics for both aerial imagery and LiDAR-derived nDSM models using Mask R-CNN can be found in Table 7.

Epping Industrial (a) Epping Industrial (b)

Figure 33: Industrial Building footprint extraction results from high-resolution aerial imagery (a) and LiDAR-derived nDSM (b)

Similarly, like the formal residential areas, in addition to the testing sample block shown in Figure 29, the models have been used in different industrial areas to obtain a wider and enhanced understanding of how they perform in various spatial contexts. These additional testing sample blocks do not have the ground truth building mask, so the extracted results are compared to the high-resolution aerial imagery of the background to assess their

performance in those areas. Figure 34 shows results extracted across different industrial areas within the City of Cape Town metropolitan.

Boquinar Industrial (a)

Boquinar Industrial (b)

Parow Industrial (a)

Parow Industrial (b)

Figure 34: Industrial Building footprint extraction results from high-resolution aerial imagery (a) and LiDAR-derived nDSM(b) in Boquinar Industrial and Parow areas.

The footprints extracted from additional testing sample industrial areas, as shown in Figure 34, reveal that the Mask R-CNN model is effective in extracting industrial building footprints from LiDAR-derived nDSM than from high-resolution aerial imagery. In the Parow and Boquinar

industrial areas, the Mask R-CNN sometimes fails to detect bright industrial roofs when extracting buildings from high-resolution aerial imagery. In contrast, it performs well in extracting these buildings from LiDAR-derived nDSM.

Aerial Imagery Versus LiDAR-derived nDSM Extraction Results for Informal Settlement

In this case, the two trained Mask R-CNN models for aerial imagery and LiDAR-derived nDSM have been applied to test an informal settlement dataset, as shown in Figure 30. The models for informal settlements are capable of detecting and segmenting shacks, but with a high number of false negatives.

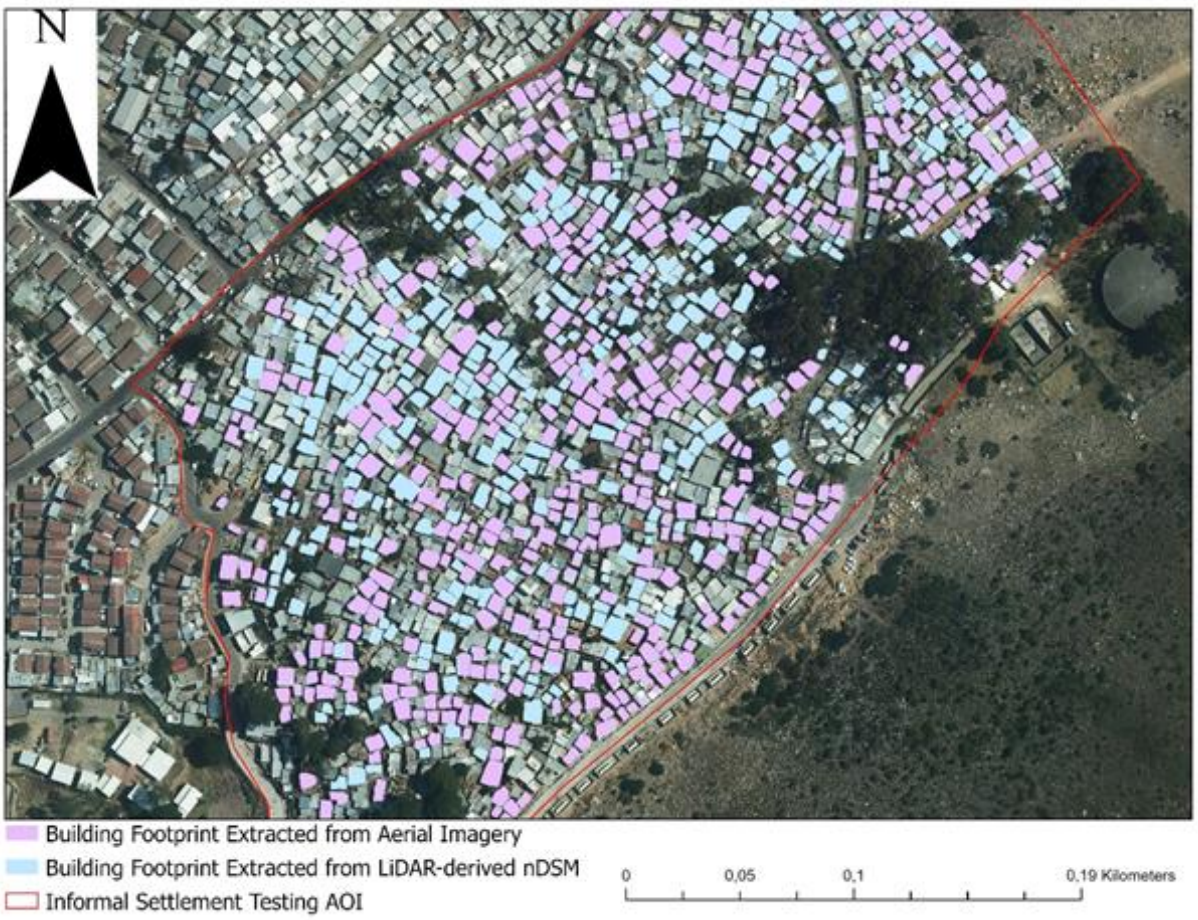


Figure 35: Informal Settlement (Shacks) footprint extraction results from high-resolution aerial imagery (blue) and LiDAR-derived nDSM (purple)

4.4 Analysis of Results

4.4.1 Effectiveness of the Mask R-CNN

To answer research question 2. How can remote sensing data such as aerial imagery and LiDAR data be effectively used to extract accurate building footprints? In this section, the calculated F1-score and Average Precision score results are presented and analyzed.

By utilizing the Mask R-CNN method for building footprint extraction, the Blaauwberg district is chosen for training and validation, with divisions made for formal residential, industrial, and informal settlements. This division helps to understand the performance of Mask R-CNN in these specific areas and accounts for computational time limitations. In Section 4.3.4, trained Mask R-CNN is used to extract building footprints from aerial imagery and LiDAR-derived nDSM in the testing areas. The Average Precision score and F1-score are calculated for both aerial imagery and LiDAR-derived nDSM in each of these areas. The Mask R-CNN performs well, with an Average Precision score ranging from 0.28 to 0.82.

In formal residential areas, the performance of building footprint extraction from LiDAR-derived nDSM is comparable to that from aerial imagery using Mask R-CNN.

Table 6: Evaluation metrics results for high-resolution aerial imagery and LiDAR-derived nDSM using Mask R-CNN in the formal residential testing area with ground truth building mask.

Data type	Precision	Recall	F1-score	AP	TP	FP	FN
Aerial Imagery	0.83	0.69	0.76	0.60	2063	420	915
LiDAR-Derived nDSM	0.85	0.70	0.77	0.61	2094	383	884

Table 6 displays information about buildings in the formal residential testing dataset. There are a total of 2978 buildings. Of these, 2063 buildings are correctly identified, 420 are mistakenly identified, and 915 buildings are missed in the aerial imagery. Based on the LiDAR-derived nDSM, 2094 buildings are correctly identified, 383 are mistakenly identified, and 884 buildings are missed. The evaluation metrics in Table 6 indicate that the Mask R-CNN algorithm performs well in accurately identifying formal residential building footprints using both LiDAR-derived nDSM and aerial imagery. The Mask R-CNN model achieves an Average Precision score of 0.61 when extracting formal building footprints from LiDAR-derived nDSM, compared to 0.60 from aerial imagery. However, the use of LiDAR-derived nDSM to extract building footprints using Mask R-CNN only slightly improves the F1-score and AP score by 1.0% in the formal residential category.

The Mask R-CNN method effectively extracts building footprints of various sizes and shapes from both aerial imagery and LiDAR-derived nDSM, yielding Average Precision scores of 0.60 and 0.61 respectively. The calculated F1-score from aerial imagery and LiDAR-derived nDSM is 0.76 and 0.77, respectively. It is important to note that the AP scores achieved in this research cannot be directly compared to the AP scores obtained in a study conducted by Esri (2018) in collaboration with Nvidia and Miami-Dade County. Both studies followed a similar workflow, but the Esri study achieved an AP score of 0.48. On the other hand, this research shows an

enhanced performance in extracting formal buildings by using Mask R-CNN. Additionally, another study conducted by Tiede et al. (2021) in Khartoum, Sudan also employed Esri's workflow to extract formal dwelling footprints from high-resolution aerial imagery, resulting in an F1-score of 0.78. In this research, the extraction of formal building footprints from high-resolution aerial imagery yielded an F1-score of 0.77. Consequently, these values can be directly compared.

In industrial areas, the accuracy of extracting building footprints is significantly better using LiDAR-derived nDSM compared to using aerial imagery.

Table 7: Evaluation metric results for aerial imagery and LiDAR-derived nDSM using Mask R-CNN in the industrial testing area (Epping industrial area).

Data type	Precision	Recall	F1-score	AP	TP	FP	FN
Aerial Imagery	0.38	0.76	0.51	0.30	434	706	134
LiDAR-Derived nDSM	0.68	0.83	0.75	0.82	471	219	97

Table 7 indicates that there are 568 buildings in the industrial testing dataset. Out of these 568 buildings, 434 are correctly identified as buildings, while 706 are erroneously identified as buildings. Additionally, there are 134 missed buildings when comparing the aerial imagery to the 471 correctly identified buildings and the LiDAR-derived nDSM reveals 219 buildings that are mistakenly identified and 97 missed buildings. The evaluation metrics in Table 7 demonstrate that the Mask R-CNN algorithm performs better in extracting industrial building footprints from the LiDAR-derived nDSM compared to aerial imagery. The Mask R-CNN model can accurately extract industrial building footprints of different shapes and sizes from the LiDAR-derived nDSM and aerial imagery, achieving an Average Precision score of 0.82 and 0.30, respectively. The calculated F1-score from aerial imagery and LiDAR-derived nDSM is 0.51 and 0.75, respectively. The AP score obtained from aerial imagery is significantly influenced by the large number of false positives (709), whereas only 219 false positives are derived from LiDAR-based nDSM. The utilization of LiDAR-derived nDSM greatly enhances the extraction of building footprints of various sizes and shapes in industrial areas, resulting in a 52% improvement in the Average Precision score and 24% in the F1-score. This improvement is due to the more efficient training of the neural network enabled by LiDAR-derived nDSM, requiring fewer training samples, as reported by Esri (2020).

Furthermore, high-resolution aerial imagery presents a challenge when distinguishing between industrial building roofs and other bright background objects, such as containers, parking lots, roads, and reflective surfaces. The similarity in brightness makes it difficult for the Mask R-CNN model to accurately extract these building footprints, resulting in a high number of false positives (erroneously extracted buildings) and false negatives (missed detection). In this research, some parking lots and roads were mistakenly extracted as industrial buildings from the aerial imagery, as shown in Figure 36. Similarly, some industrial buildings with bright roofs were not detected from the aerial imagery, as illustrated in Figure 37.

Figure 36: Visualisation results for false positive (Blue colours indicate detected parking lots and roads and Red colours indicate true positives, correctly detected industrial buildings)

Figure 37: Visualisation results for false negatives (Green colours indicate missed buildings and Red colours indicate extracted buildings)

In informal settlements, building footprint extraction performance from both aerial imagery and LiDAR-derived nDSM using Mask R-CNN is unsatisfactory.

Table 8: Evaluation metric results for aerial imagery and LiDAR-derived nDSM using Mask R-CNN in informal settlement testing area.

Data type	Precision	Recall	F1-score	AP	TP	FP	FN
Aerial Imagery	0.92	0.30	0.45	0.28	504	45	1174
LiDAR-Derived nDSM	0.90	0.33	0.49	0.31	562	60	1116

Table 8 displays the data regarding shack structures in the test dataset for informal settlements. Among the total of 1678 shack structures, only 504 are accurately identified as shacks, while 45 are mistakenly identified as such. Additionally, 1174 shack structures are missed when comparing the aerial imagery to the LiDAR-derived nDSM. In contrast, from the LiDAR-derived nDSM, 562 shack structures are correctly identified, 60 are mistakenly identified, and 1116 are missed. The evaluation metrics presented in Table 8 reveal that the Mask R-CNN algorithm performs better at extracting shacks from the LiDAR-derived nDSM compared to the aerial imagery.

The Average Precision score calculated from aerial imagery and LiDAR-derived nDSM is 0.28 and 0.30, respectively. The calculated F1-score from aerial imagery and LiDAR-derived nDSM is 0.45 and 0.49, respectively. As reported by Shoko et al. (2022), dwellers in informal settlements tend to use shack roofs as storage for objects such as scrap material, metal pieces, and other reflective materials. This introduces noise in high-resolution image processing. Moreover, the proximity and dense construction of shack structures create a highly populated area with diverse roof types and uneven textures. Consequently, the Mask R-CNN approach struggles to accurately extract and detect shacks from both aerial imagery and LiDAR-derived nDSM, as indicated by their F1-score and Average Precision being below 0.5.

Earlier in section 4.3.4, it is discussed that when extracting building footprints from aerial imagery and LiDAR-derived nDSM, evaluation metrics are calculated using an IoU of 0.5. It is found that Mask R-CNN performs better at extracting building footprints from LiDAR-derived nDSM compared to aerial imagery. This holds for various scenarios, including formal residential areas, industrial areas, and informal settlements. LiDAR-derived nDSM consistently yields higher Average Precision scores.

Specifically, Mask R-CNN demonstrates satisfactory performance in extracting building footprints from both aerial imagery and LiDAR-derived nDSM in formal residential areas, achieving Average Precision scores of 0.60 and 0.61, respectively. However, for industrial buildings, Mask R-CNN performs satisfactorily only with LiDAR-derived nDSM, scoring an Average Precision of 0.82. In Figure 35 of section 4.3.4, it can be seen that the Mask R-CNN's unsatisfactory performance in detecting shacks in informal settlements using both aerial imagery and LiDAR-derived nDSM. Nevertheless, by combining footprints extracted from LiDAR-derived nDSM and high-resolution aerial imagery, the Average Precision Score improves to 0.52. However, this AP score is not comparable to the values achieved in a study conducted by Mohamed et al. (2020) in a populated rural area of Maghagha City, Egypt. The study extended the workflow employed in this research by combining two Mask R-CNN ResNet backbones (34, 101) and implemented a post-processing phase to enhance the extracted building footprints. The study achieved a combined F1-score of 0.88 and an AP score of 0.95. These values cannot be directly compared to this research's values since Mask R-CNN models were trained using ResNet34 and ResNet101 specifically to efficiently extract informal building footprints. In this research, ResNet101 is solely used as the backbone.

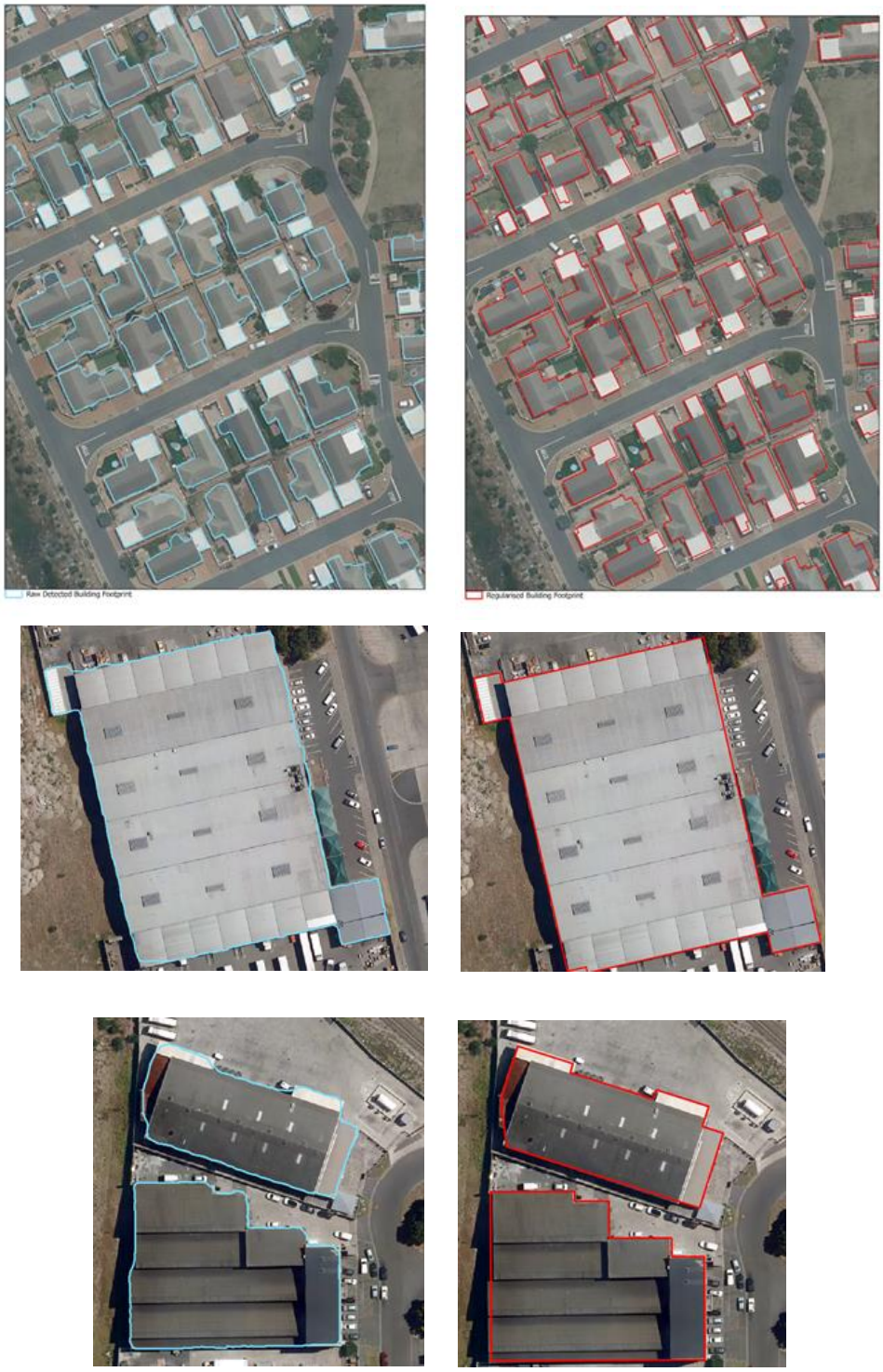
This research shows that larger, more standardized, and well-separated features result in higher Average Precision scores. Therefore, Mask R-CNN achieves higher scores in formal

residential and industrial areas compared to informal settlements since the structures of shacks in informal settlements are closely built together.

Research question 2 explores the effective use of remote sensing data, such as aerial imagery and LiDAR data, for accurate extraction of building footprints. The analysis in section 4.3.4 shows that aerial imagery and LiDAR are highly effective for extracting building footprints in formal residential areas using the Mask R-CNN method. In industrial areas, LiDAR-derived nDSM also performs well in accurately extracting industrial building footprints. For informal settlements, combining footprints extracted from aerial imagery with LiDAR-derived nDSM produces satisfactory results.

4.4.2 Boundary Regularization

The extracted building footprints from either aerial imagery or LiDAR-derived nDSM show irregular and noisy outlines due to the locality of pixel-wise labelling conducted by Mask R-CNN as stated in section 3.4.1. To improve the extracted building polygons and make them regular, an advanced "Regularize Building Footprints" geoprocessing tool in ArcGIS is utilized as per the detailed workflow processes shown in section 3.4.1, Figure 10. The resulting regularized building footprints are shown in Figure 39 below.



(a) (b)

Figure 38: Visualisation results for a) irregular buildings (raw detected building footprints) and b) regularized building footprints

5. Conclusions and Remarks

Chapter 4, presented, discussed, and analyzed the results of the research that aimed to answer the research question 2: How can remote sensing data, such as aerial imagery and LiDAR data, be effectively used to accurately extract building footprints?

The Mask R-CNN algorithm showed excellent performance in extracting formal residential building footprints from both high-resolution aerial images and LiDAR-derived nDSM. In industrial areas, the algorithm performed well in extracting footprints from LiDAR-derived nDSM. However, extracting footprints in dense informal settlements presented challenges due to the proximity and varying roof textures.

Based on the results and analysis presented in Chapter 4, this chapter presents the research conclusion and provides recommendations for future work.

5.1 Conclusion

Building footprints are one of the noticeable characteristics of urban areas. As advanced aerial imagery and LiDAR data become more accessible, the approach to extracting urban features has evolved. Instead of traditional methods, researchers now utilize neural networks like Convolutional Neural Networks (CNNs) for semantic and instance segmentation. In this study, the goal was to automatically extract building footprints from remote sensing data in the City of Cape Town, South Africa. To achieve this, a literature review was conducted to find a suitable spatial analysis method that could accurately and consistently extract building footprints from aerial imagery and LiDAR data. The review revealed that Mask R-CNN, an algorithm known for its effectiveness in instance segmentation and object extraction, was the most suitable choice. Its remarkable performance in extracting building footprints from high-resolution images and LiDAR-derived nDSM (normalized Digital Surface Model) made it the preferred option over Unet. This answered research question 1 and addressed research objective 1.

For the training of the Mask R-CNN algorithm, the study focused on the Blaauwberg district within the City of Cape Town, which includes formal residential areas, industrial zones, and informal settlements. To evaluate and analyze the effectiveness of the model separately for each area, two Mask R-CNN models were trained: one using aerial imagery and the other using LiDAR-derived nDSM. This addressed research objective 2.

To compare the performance of these models effectively, a new test dataset was used, which had not been seen during training or validation. Evaluation metrics such as precision, recall, F1-score, and Average Precision were calculated to assess accuracy. The Mask R-CNN algorithm exhibited excellent performance in extracting formal residential building footprints from both high-resolution aerial images and LiDAR-derived nDSM, with satisfactory precision, recall, F1-score, and Average Precision. However, when it came to industrial areas, the algorithm performed well in extracting building footprints from LiDAR-derived nDSM but showed unsatisfactory results with high-resolution aerial imagery. Extraction of shacks in dense settlements proved challenging due to the proximity and varying roof textures, resulting in a

highly populated area. Consequently, the calculated F1-score and Average Precision values were less than 0.5. However, the fusion of footprints extracted from aerial imagery and LiDAR-derived nDSM improved the Average Precision Score to above 0.5. Therefore, for informal settlements, the fusion footprints generated from both aerial imagery and LiDAR-derived nDSM can be utilized to produce footprints for all informal settlements in the City of Cape Town. Additionally, the trained Mask R-CNN models for extracting formal residential building footprints from high-resolution aerial images and LiDAR-derived nDSM, along with the model for extracting industrial building footprints from LiDAR-derived nDSM, can be used to extract building footprints across the entire City of Cape Town. Thus, this answered both research questions 2 and 3 and addressed research objectives 2 and 3.

Conclusively, the Mask R-CNN models used in this research have great potential to solve the problem of extracting building footprints from remote sensing data on a large scale. The resulting footprints can be combined with the existing 2D building footprints layer to fill in any gap. It is important to note that the trained Mask R-CNN models are scalable to extract building footprints across different South African' Metropolitans, as formal and informal zones co-exist in these areas and they have similar environmental settings. The building footprints results from the formal residential and industrial models are of great quality for use within the City of Cape Town. This data plays a vital role in various infrastructure planning initiatives in the City of Cape Town, including stormwater and sewer networks, electrification, substation planning, road and MyCity Bus route and stops planning. Moreover, building footprint data is essential for applications such as change detection, service delivery planning, household counting (census estimates), cadastral policy formulation, humanitarian interventions, and land use planning.

5.2 Future Works

The work presented in this research can still be developed further, considering the limited training datasets for industrial areas and informal settlements as well as the computational limitation to efficiently train the Mask R-CNN models on more training datasets.

- Adding more informal settlement and industrial training datasets with sufficient roof variability and fine-tuning the Mask R-CNN models to ensure that the developed method can sufficiently learn and accurately extract shacks and industrial building footprints.
- Additional preprocessing steps to enhance the contrast of roofs and reduce background complexity in informal settlements and industrial areas and fine-tune the Mask R-CNN models to improve the detection of bright shacks and industrial buildings.
- Experiments with pre-processed training datasets with sufficient roof variability and enough model training duration should be done on a machine with high computational power and RAM storage.
- Only ResNet101 as the Mask R-CNN backbone has been considered in this research but the developed method can be extended further to compare the building footprint results obtained from ResNet50 and ResNet152.
- Training a single model with both aerial imagery and LiDAR-derived nDSM integrated.

List of References

Abdollahi, A., Pradhan, B., Gite, S., Alamri, A., 2020. Building footprint extraction from high-resolution aerial images using generative adversarial network (GAN) architecture. IEEE Access, vol. 8, pp. 209517209527.

Alsabhan, W and Alotaiby, T., 2022. Automatic Building Extraction on Satellite Images Using Unet and ResNet50. Available: https://doi.org/10.1155/2022/5008854

Anwa, A., 2022. What is Average Precision in Object Detection & Localization Algorithms and how to calculate it? Available: https://towardsdatascience.com/what-is-average-precision-in-object-detection-localization-algorithms-and-how-to-calculate-it-3f330efe697b#:~:text=Average%20precision%20is%20the%20area,is%20between%200%20to%201.

Aryal, J and Neupane, B., 2023. Multi-Scale Feature Map Aggregation and Supervised Domain Adaptation of Fully Convolutional Networks for Urban Building Footprint Extraction. Remote Sensing 15(2), 488. Available: https://doi.org/10.3390/rs15020488

Audebert, N.; Le Saux, B.; Lefèvre, S., 2017. Segment-before-detect: Vehicle detection and classification through semantic segmentation of aerial images. Remote Sensing 9(4), 368. Available: https://www.mdpi.com/2072-4292/9/4/368

Biljecki, F., Stoter, J., Ledoux, H., Zlatanova, S. & Çöltekin, A. 2015a. Applications of 3D City Models: State of the Art Review. ISPRS International Journal of Geo-Information. 2015(4):2842–2889.

Bluemarblegeo 2023. GeoTIFF. Available: https://www.bluemarblegeo.com/knowledgebase/global-mapper-19/Formats/GeoTIFF.htm

Chitturi, G. 2020. Building Detection in Deformed Satellite Images Using Mask R-CNN. Available: https://www.diva-portal.org/smash/get/diva2:1413658/FULLTEXT02

Dmitry, K., Daniel, H., Omar, M., 2018. Reconstructing 3D buildings from Aerial LiDAR with Deep Learning. Available: https://medium.com/geoai/reconstructing-3d-buildings-from-aerial-lidar-with-ai-details-6a81cb3079c0

Esri South Africa 2022. Detecting Informal Dwellings in Cape Town using Deep Learning tools in ArcGIS Pro. Available: https://www.esri-southafrica.com/deep-learning-with-arcgis-pro/ [2022, May 11].

Esri, 2023. How U-net works? Available: How U-net works? | ArcGIS API for Python

Gilania, S. A. N., Awrangjebb, M. and Lub, G., 2015. Fusion of LiDAR Data and Multispectral imagery for Effective Building Detection based on Graph and Connected Component Analysis. International Archives of the Photogrammetry, Remote Sensing, and Spatial Information Sciences pp. 65 – 72

Haithcoat, T. L., Song, W., and Hipple, J. D., 2001. Building footprint extraction and 3-D reconstruction from LIDAR data. IEEE/ISPRS Joint Workshop on Remote Sensing and Data Fusion over Urban Areas, Rome, Italy.

Hafidz, Z., 2018. Understanding Learning Rates and How It Improves Performance in Deep Learning. Available: https://towardsdatascience.com/understanding-learning-rates-and-how-it-improves-performance-in-deep-learning-d0d4059c1c10

He, Kaiming et al, 2020. Mask R-CNN. IEEE Transactions on Pattern Analysis and Machine Intelligence 42(2), pp.386–397.

He, K., Gkioxari, G., Dollár, P., Girshick, R. 2017. Mask R-CNN. IEEE International Conference on computer vision, pp. 2961-2969.

Isikdag, U., Horhammer, M., Zlatanova, S., Kathmann, R. & Van Oosterom, P.J.M. 2015. Utilizing 3D building and 3D cadastre geometries for better valuation of existing real estate. In FIG Working Week 2015: From the wisdom of the ages to the challenges of modern world. Sofia, Bulgaria: 17-21 May 2015.

Ji, S., Wei, S., M. Lu, 2019. A scale robust convolutional neural network for automatic building extraction from aerial and satellite imagery. Int. J. Remote Sens., vol. 40, pp. 3308–3322.

Kingma, D., and Ba, J., 2014. Adam: A method for stochastic optimization. Available: arXiv:1412.6980

K.Bittner, F.Adam, S. Cui, M. Körner, and P. Reinartz, 2018, March. Building footprint extraction from VHR remote sensing images combined with normalized DSMs using fused fully convolutional networks. IEEE J. Sel. Top. Applied Earth Observations and Remote Sensing 11(8), pp. 2615–2629.

Krizhevsky, A., Sutskever, I., & Hinton, G. E. (2012). Imagenet classification with deep convolutional neural networks. In Advances in neural information processing systems (pp. 1097-1105).

Lee, D. S., Shan, J., and Bethel, J.S., 2003. Class-guided building extraction from Ikonos imagery. Photogramm Engineering and Remote Sensing 69(2), 143–150.

Li, H., et al., 2013. New methodologies for precise building boundary extraction from LiDAR data and high-resolution images. Sensor Review 33(2), pp. 157- 165.

Li, Y. and Wu, H., 2013. An improved building boundary extraction algorithm based on fusion of optical imagery and LIDAR data. Optik 124(22), pp. 5357 – 5362.

Liu, Y., Zhou, J., Qi, W., Li, X., Gross, L., Shao, Q., Zhao, Z., Ni, L., Fan, X., and Li, Z. 2020. ARC-Net: An efficient network for building extraction from high-resolution aerial images. IEEE Access, vol. 8, pp. 154997155010.

Maggiori,E., Tarabalka, Y., Charpiat, G., Alliez, P., 2017, December. Can semantic labeling methods generalize to any city? The inria aerial image labelling benchmark. IEEE International Geoscience and Remote Sensing Symposium, pp. 3226–3229.

Mohamed, S., Mohmoud, A., Moustafa, M., Helmy, A., Nasr, A. 2022. Building Footprint Extraction in Dense Area from LiDAR Data using Mask R-CNN. International Journal of Advanced Computer Science and Application 12(6), pp.346 – 353.

Rajabifard, A., Atazadeh, B., Kalantari, M. & Williamson, I. 2018a. A New Method for Integrating 3D Spatial Information about Vertically Stratified Ownership Properties into the Property Map Base. In FIG Congress 2018: Embracing our smart world where the continents connect: enhancing the geospatial maturity of societies. Istanbul, Turkey: 6–11 May 2018.

Ronneberger, O., Fischer, P., & Brox, T. (2015). U-Net: Convolutional networks for biomedical image segmentation. In International Conference on Medical image computing and computer-assisted intervention (pp. 234-241).

Siddiqui et al., 2016. A Robust Gradient Based Method for Building Extraction from LiDAR and Photogrammetric Imagery. MDPI, pp.1-24.

Sohn, G. and Dowman, I., 2007. Data fusion of high-resolution satellite imagery and LIDAR data for automatic building extraction. ISPRS Journal of Photogrammetry and Remote Sensing 62(1), pp. 43–63.

Tarantino, E., and Figorito, B., 2011. Extracting buildings from true color stereo aerial images using a decision making strategy. Remote Sensing (Basel) 3, 1553–1567.

Tiede, D., Schwendemann, G., Alobaii, A., Wendt, L., Lang, Stefan. 2021. Mask R-CNN-based building extraction from VHR satellite data in operational humanitarian action: An example related to Covid-19 response in Khartoum, Sudan. Transaction in GIS(25), pp.1213 – 1227. Available: https://doi.org/10.1111/tgis.12766

Tiwari, A., 2022. Artificial Intelligence and Machine Learning for EDGE Computing. Science Direct, pp. 23 – 32. Available: https://doi.org/10.1016/B978-0-12-824054-0.00026-5

Pan, Z.; Xu, J.; Guo, Y.; Hu, Y.; Wang, G.,2020. Deep Learning Segmentation and Classification for Urban Village Using a Worldview Satellite Image Based on Unet. Remote Sensing 12(10), 1574. Available: https://doi.org/10.3390/rs12101574

Partovi, T., et al., 2017. Building outline extraction using a heuristic approach based on generalization of line segments. IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing, 10 (3), 933–947. doi:10.1109/JSTARS.2016.2611861

Wu, Q., Feng, D., Cao, C., Zeng, X., Feng, Z., Wu, J. and Huang, Z.2021. Improved Mask R-CNN for Aircraft Detection in Remote Sensing Images. Sensors, vol. 21, no. 8, p. 2618.

Wei, S., Ji, S., Lu, M., 2020, March. Toward Automatic Building Footprint Delineation From Aerial Images Using CNN and Regularization. IEEE Transactions on Geoscience and Remote Sensing 58(3), pp. 2178-2189.

Xu, Y.; Xie, Z.; Feng, Y.; Chen, Z., 2018. Road Extraction from High-Resolution Remote Sensing Imagery Using Deep Learning. Remote Sensing 10(9), 1461. Available: https://www.mdpi.com/2072-4292/10/9/1461

Zhang, K., Yan, J., and Chen, S., 2006. Automatic construction of building footprints from airborne LiDAR data. IEEE Geoscience and Remote Sensing Magazine 44(9), 2523–2533.

Zhang, P., He, H., Wang, Y., Liu, Y., Lin, H., Guo, L. and Yang, W., 2022. 3D Urban Buildings Extraction Based on Airborne LiDAR and Photogrammetric Point Cloud Fusion According to U Net Deep Learning Model Segmentation. IEEE Access, vol. 10, pp. 20889 - 20897.

Zhang, S., Han. F., Bogus, SM., 2020, March. Building Footprint and Height Information Extraction from Airborn LiDAR and Aerial Imagery. In. Construction Research Congress, pp. 326 – 335.

Zhao, K., Kang, J., Jung, J., Sohn, G., 2018, June. Building Extraction From Satellite Images Using Mask R-CNN With Building Boundary Regularization. In: CVPR Workshops, pp. 247–251.