Masters Program in Geospatial Technologies

INFORMAL SETTLEMENT SEGMENTATION USING VHR RGB AND HEIGHT INFORMATION FROM UAV IMAGERY: A CASE STUDY OF NEPAL

Ganesh Prasad Sigdel

Dissertation submitted in partial fulfilment of the requirements
for the Degree of *Master of Science in Geospatial Technologies*

NOVA IMS Information Management School
Universitat Jaume I
Westfälische Wilhelms-Universität Münster

INFORMAL SETTLEMENT SEGMENTATION USING VHR RGB AND HEIGHT INFORMATION FROM UAV IMAGERY: A CASE STUDY OF NEPAL

Dissertation supervised by:
Filiberto Pla Bañón, PhD
Professor, Institute of New Imaging Technologies (INIT),
Universitat Jaume I (UJI),
Castellon de la Plana,
Spain

Co-supervised by:
Rubén Fernández Beltrán, PhD
Institute of New Imaging Technologies (INIT),
Universitat Jaume I (UJI),
Castellon de la Plana,
Spain

Co-supervised by:
Pedro Cabral, PhD
Associate Professor, NOVA IMS
Universidade NOVA de Lisboa(UNL)
Portugal

February 21, 2021

Acknowledgements

I express my deep gratitude to my thesis Supervisor **Dr. Filiberto Pla Banon** for providing this greatest experience ever. I feel privileged to have you as my supervisor. Learning during this period means a lot in my future career. Words are not enough to explain how excited and happy I am after working under your supervision. I am equally indebted to co-supervisor **Dr.Ruben Fernandez Beltran** , without whom I would not have been able to accomplish my thesis. You always motivated me to keep high spirit. I am quite impressed of your concept, dedication and contribution for me. Thanks beyond limit for being my on my side at every difficulties during this period. Many thanks to another co-supervisor **Dr. Pedro Cabral** for his utmost care, crucial and constructive comments and thorough review and suggestions on my work.

I will always remember **UJI** and **INIT** for letting me be a part of this wonderful and memorable moments at Geospatial Technologies course. Sharing workplace with colleagues, **Saadoon** and **Mo** and discussing while working has to do a lot with fine-tuining of my work. Thank you, Saadoon and Mo for your great company. I would like to extend my warm thanks to **Mateen** for critical questions, that made me think out of box in several occasions.

My friends at my home **Anu**, **Poshan** and **Janak** are simply awesome. Nothing can replace your role here for me. Thanks to my senior **Sanjeevan Shrestha**, who triggered this interesting thesis topic and always kept himself standby to face my queries during the thesis.

Collecting UAV images without active co-ordination of **Khimlal Gautam**, **Dinesh Lamichhane** and **Nabin Raj Bhatt** would have been impossible. Thank you so much for taking it the task as your own. Thanks to Uttam Pudasaini from **Naxa Pvt. Ltd.** for support with orthophotos from Kathmandu.

My parents have always kept trust on me and hid their pain for my happiness. Thank you for constantly pushing me ahead to keep my pace with thesis amid tough situations and your personal wish to have a good time with son. You are my greatest motivation ever.

INFORMAL SETTLEMENT SEGMENTATION USING VHR RGB AND HEIGHT INFORMATION FROM UAV IMAGERY: A CASE STUDY OF NEPAL

ABSTRACT

Informal settlement in developing countries are complex. They are contextually and radiometrically very similar to formal settlement. Resolution offered by Remote sensing is not sufficient to capture high variations and feature size in informal settlements in these situations. UAV imageries offers solution with higher resolution. Incorporating UAV image and normalized DSM obtained from UAV provides an opportunity of including information on 3D space. This can be a crucial factor for informal settlement extraction in countries like Nepal. While formal and informal settlements have similar texture, they differ significantly in height. In this regard, we propose segmentation of informal settlement of Nepal using UAV and normalized DSM, against traditional approach of orthophoto only or orthophoto and DSM. Absolute height, normalized DSM(nDSM) and vegetation index from visual band added to 8 bit RGB channels are used to locate informal settlements. Segmentation including nDSM resulted in 6 % increment in Intersection over Union for informal settlements. IoU of 85% for informal settlement is obtained using nDSM trained end to end on Resnet18 based Unet. Use of threshold value had same effect as using absolute height, meaning use of threshold does not alter result from using absolute nDSM. Integration of height as additional band showed better performance over model that trained height separately. Interestingly, benefits of vegetation index is limited to settlements with small huts partly covered with vegetation, which has no or negative effect elsewhere.

Keywords

Informal Settlement

Unmanned Aerial Vehicle (UAV)

Deep Learning

Semantic Segmentation

Normalized Digital Surface Model(nDSM)

Visible-Band Difference Vegetation Index (VDVI)

INDEX OF THE TEXT

INDEX OF FIGURES

INDEX OF TABLES

Acronyms

AFS	Average Fusion Strategy
CNN	Convolutional Neural Network
DN	Digital Number
DSM	Digital Surface Model
DTM	Digital Terrain Model
FOV	Field of View
GCP	Ground Control Point
GIS	Geographic Information System
GPS	Global Positioning System
MFS	Maximum Fusion Strategy
nDSM	normalized Digital Surface Model
ODM	Open Drone Map
RGB	Red Green Blue
SfM	Structure from Motion
SVM	Support Vector Machine
UAV	Unmanned Aerial Vehicle
UN	United Nations
UTM	Universal Transverse Mercator
VDVI	Visible-Band Difference Vegetation Index

Chapter 1

Introduction

1.1 Contextual Background

Global population distribution is dynamic [1]. Rural-urban proportion is also changing temporally, with major movements towards city. While city-centered development, better education and health facilities[2] have been major pull factor for migration, loss of jobs in primary sectors such as agriculture pushes one from the rural area to city [1]. [3] estimates that the number of people living in the urban area will exceed those in rural area by 2030. This share was only thirty percent in mid 20^{th} century. This demands rapid development in infrastructures. But, infrastructure development rate is insufficient to meet these requirement[4]. One of the adverse effect of these "unintended urbanization" is growth of unorganized settlements [5].

In developing countries, while core centres are developed, areas in the off-centers are deprived of proper housing development, creating an urban divide[3]. The poor then occupy a land, generally public or state-owned [4] and build houses. As these lands are not formally recognized by authorities, these are termed *informal settlements*.

Alternate terms as squatter settlements, slums and regional terms as *ghettos*, *Zopadpatti*, *ranchos*, *katchi abadi*, *shanty town*,etc. are also used to refer to informal settlements[5, 6], "slum" and informal settlement" being the commonly used ones. However, these term differ in micro level. While informal settlement is concerned more with right aspect, slum sees from environmental aspect. According to [7], informal settlements are the settlements developed without

proper land ownership and without regard to regulations and standards for safety, health. On the other hand, slums are settlements with poor living standards. So, informal settlements can have good housing as well, and all slums might not necessarily be informal ones[5]. But, both of these,have some common characteristics from i) lack of recognition, ii) absence of tenure security, iii) inadequate infrastructures, iv) overcrowded and sub-standard living and v) location of land less suitable for occupation [7]. Due to the global and broader scope, the term "informal settlement" is used in majority of literature to make it more general[4, 5, 7] .

Though global percent of people living in unorganized settlements is descending, absolute number is not[7]. 40 to 70 percent of urban dwellers from developing countries live in informal settlement[5] and there is no sign of improvement in trend[6]. In this regard, Agenda 11.1 of Sustainable Development Goals aims to ensure access for all to adequate, safe and affordable housing and basic services and upgrade slums by 2030 [8]. The first task of each government in this initiative is to map spatial extent of informal settlements. These tasks are challenging due to the need of collaboration between several agencies,growing pattern of informal settlements, limited or insufficient data to delineate informal settlement [9].

Informal settlements are heterogeneous from city to city[1], and also within centre and outskirt of city[10]. For instance, the informal settlements in the global south take the form of single storey buildings built from scrap materials on abandoned area[4, 5] as in Mumbai. But in Bucharest, they have pointed and hipped roofs [1]. Thus, developing a functioning approach for extraction of informal settlement is subjective to context, and a universal model for informal settlement extraction does not exist[11].

In recent years, informal settlement mapping has been benefited from medium to high resolution satellite images. Expert meeting in 2008[10] has also highlighted the potential of application of satellite images in informal settlement extraction. Methodologically, OBIA has been used dominantly in last fifteen years along with other methods [11].

Resolution available with low and medium resolution do not serve properly for informal settlements as informal settlements have small buildings and narrow roads [11, 12]. In addition, heterogeneity of roofing material yields mixed pixels, adding complexity to automatic extraction [13]. Further limitations are imposed by large extent of similarity between radiometric characteristics and context of formal and informal settlement in developing countries.

Unmanned aerial vehicle overcomes the issues with cloud cover and resolution of satellite images. High resolution cloud-free images can be taken from low height [14, 15] on any time with UAV. This ensures on-demand and fast availability of orthophotos [16] as well as other features such as DSM from the same platform. Neural network using very high resolution UAV data brings together the benefits of high resolution and deep learning. It takes advantage of spectral, textural, geometrical and contextual features on classifying image, enhancing the result[15].

However,optical band 2D orthophoto alone does not offer enough solution towards it, especially in complex situations. Furthermore, both formal and informal settlement share common type of context, posing challenge to deep learning as well.This challenge is more prominent in developing countries. In this regard,including height information in form of DSM, nDSM or point cloud have been observed to enhance segmentation task[17] and feature extraction[12, 18, 19]. Thus, UAV orthophoto along with feature height can be a key in informal settlements segmentation for Nepal,as informal buildings are usually single-storey and shorter in contrast to those in formal settlement.

Our work consists of segmentation of informal settlement using orthophoto, nDSM and vegetation index obtained from UAV, applied to developing country, Nepal. We also quantify the contribution of integrating height information and analyse the effect of using threshold height in place of absolute height in segmentation. In a nutshell, we assess the value added by SfM derived feature height and very high resolution orthophoto on extraction of informal settlement in complex scenario.

1.2 Motivation and Problem Statement

Nepal is a developing country with 34 % of multidimensional poverty. It stands on 142^{nd} position[20] in Human development index according to Human Development Report 2020. Large fraction of population here lives in informal settlement [2]. Thus the country is in need of proper method for informal settlement extraction.

Existing approaches on informal settlement mapping have used deep learning either using planimetric detail only [21], or are limited to context of developed countries only[12]. These approaches do not address the complexity of informal settlement in Nepal. In Nepal, settlements are mixed, and formal and informal buildings are very similar from radiometric and contextual

perspectives. Small buildings, partly obstructed by vegetation are not well identifiable in satellite images. UAV offers advantage of availability of DSM and 3D point cloud along with horizontal details, which is not available with satellite images[16]. So, very high resolution UAV orthophoto combined with height information is expected to provide better segmentation result for informal settlement extraction in context of Nepal. This is because, informal buildings are usually single-storey and shorter in height [22].

We develop a framework to extract informal settlement in developing country like Nepal by deep learning with UAV orthophoto, and analyse experimentally whether including nDSM as the additional channel to UAV-based very high resolution RGB orthophoto leads to better result. If it enhances, we also test a concept of nDSM threshold, as it offers elimination of irrelevant heights and representation of height over wider DN range, and might enhance segmentation accuracy.

1.3 Aims and Objectives

The main aim of the our work is to develop a framework to detect and locate informal settlement using UAV images and normalized DSM in a developing country like Nepal. The overall aim is complemented by following specific objectives:

- to assess the performance of state-of-art segmentation models to detect and locate informal settlement extraction in context of Nepal,
- to quantify the benefits of additional features such as height and vegetation index in UAV image for informal settlement extraction,
- to develop a CNN-based segmentation technique and its adequate configuration, including additional data such as height to analyse the segmentation performance.

1.4 Methodology

The methodology of our research has three major stages namely (i) UAV image acquisition and processing, (ii) Pre-processing of image and ground truth data, and (iii)Experimentation and extraction of informal settlement from very high resolution images, and analysis on the contribution of height and height integration approach(figure 1.1).

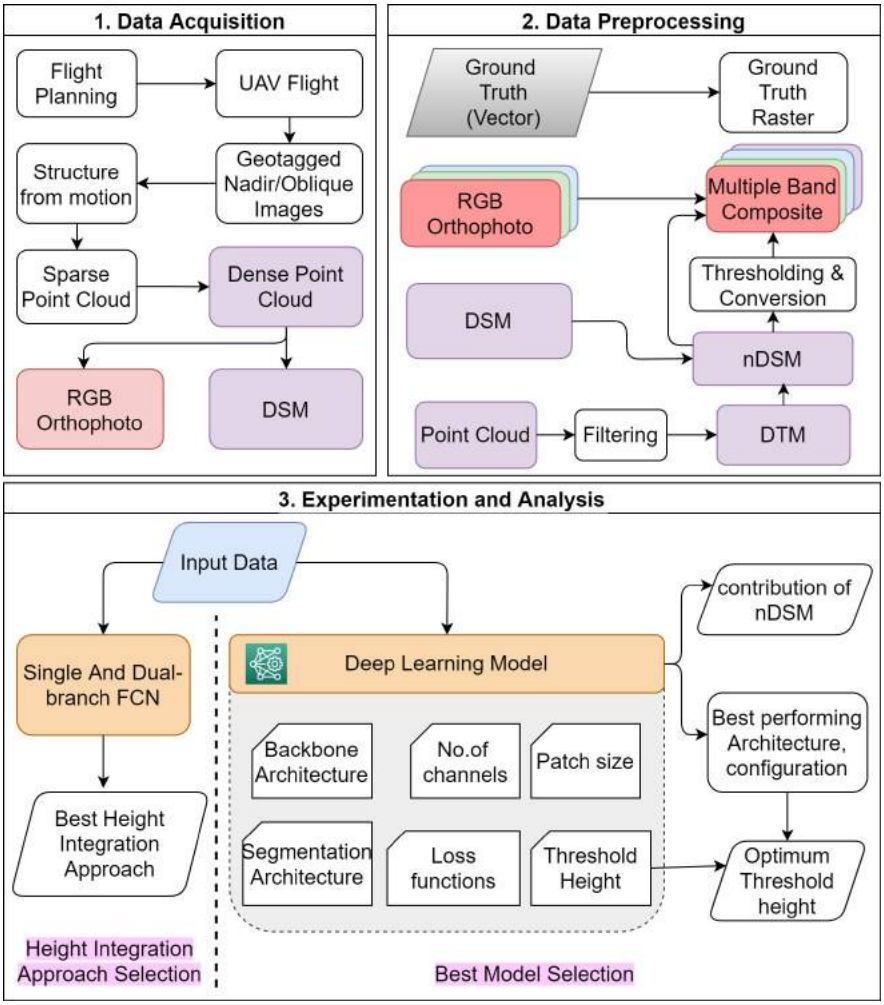


Figure 1.1: Overall methodology of Research Work

First phase of data collection and preparation deals with UAV image data acquisition from nine sites of informal settlements in Nepal, and their processing to prepare ready-to-feed image data. Ground truth cadastral data are also collected on the same phase. Images and cadastral maps are acquired in the month of September and October 2020.

In second phase, we pre-process the data for deep segmentation model. Point cloud are further processed to generate DTM representing bare earth, which are subtracted from DSM to obtain normalized DSM (nDSM). Using chosen threshold value for height, nDSM are converted to 8-bit DN range and appended to RGB orthophoto as additional channel. On the same stage, cadastral data obtained as shapefile are completed by manual digitization if necessary, and then rasterized. This stage thus prepares data to use as input for training our model.

Third step is the major step and consists of experimentation on performance of models with a combination of parameters. Eight main architectures are formed by combination of two standard encoders and four segmentation models. Performance of each of these combinations of backbone architecture and segmentation models are assessed for patch sizes from [48,64,96,128,192]. Feeding RGB data and multi-channel data respectively and training them model, the contribution of height on segmentation result are analysed at the first sub-phase of this phase.

In the later part,the best performing architecture is further chosen for the identification of contribution of height and optimum threshold height to segment informal settlement. The result are cross-validated using left-out tiles. At the end we are able to draw conclusion on whether or not integration of height enhances segmentation. If it contributes,we will answer the approach of height information incorporation and find the threshold height to effectively segment informal settlement from others.

On the same phase, the optimum method of integrating height is analyzed. Segmentation based on Fully Convolutional Network (FCN) with single and dual parallel branches are assessed to identify the optimum strategy for incorporation of height-related information.

1.5 Contribution

Our work will explore and experimentally test the feasibility on advantages of height information and vegetation index on segmenting informal settlement in complex scenarios. Following are the major contributions of the work in the scientific community:

- It will propose a methodology, including data acquisition and preprocessing techniques, CNN-based segmentation algorithms and optimally performing configurations to map informal settlements in complex scenario,

- It will also analyse the usefulness and impact of auxiliary data in addition to RGB images, such as height and their implications on the use of this data,

1.6 Thesis Structure

This thesis has been divided into seven chapters. Chapter 2 presents the literature review and related works. Theoretical concept required for reading of this thesis has been discusses on chapter 3. It is followed by details of study area, data preparation and pre-processing in chapter 4. Overall methodology has been discussed in the same chapter. Chapter 5 further describes the experimental arrangement, tests performed for finalization of model, and choice of hyperparameter. Also discussed in the same chapter is the approach used for test of contribution of height.

Major results from the experiment have been discussed in chapter 6. Some of the limitations and recommends have been included in same chapter.Final conclusion is delivered in chapter 7 respectively .

Chapter 2

Literature Review

2.1 Conventional Approaches for Informal Settlement Extraction

Expert Group Meeting on Slum Identification and Mapping held in 2008[10] has categorized existing approaches in informal settlement mapping as (i) Visual image interpretation, (ii) Object-based Image Analysis (OBIA), (iii) texture-based methods, and (iv)community-based methods. Despite being more reliable and supervised, visual image interpretation is labour-intensive [11]. This led to algorithm-based approaches. As reviewed by [11], OBIA, visual interpretation and standard pixel-based image classification has been used in 32%, 17% and 13 % respectively of the recent publications on informal settlement extraction. Later,machine learning and deep learning evolved as better alternative, and was mentioned in 14 % of the works in last fifteen years.

Traditional approaches of household survey, integrated with GIS are still popular. One of those include work carried out by [23]in Pune, India. These approach offer advantage of direct interaction with people, and also availability of attribute data as well. Recently, remote sensing community has been greatly benefited by high spatial and temporal resolution of satellite images[24]. In 2007 [25] extracted informal settlement in peri-urban region in America using existing urban boundaries, census data and orthophotos. In another scenario, [26] observed that Free Google Earth Images taken from 8000 feet and 6000 feet offered sufficient accuracy for informal settlement in Johannesburg. Also

[27] had used IKONOS images to segment informal settlement from Malaysia using spectral classification in 2013. However,spectral characteristics based approach required additional visual interpretation due to high in-class variance and context dependence of settlements[11].

OBIA overcomes limitations of spectral classification by considering context and shape in addition to spectral properties. Use of rule-set is the key for improved segmentation of features, especially in urban area[28]. In 2013,[29] used GLCM and Gabor filter and extracted informal settlement from Google Earth images. This approach achieved accuracy of 74.15%, a lot higher than 53.65 % from previous works. In alignment with this, [30] and [31] observed that including features like road and parcel boundary could segment informal settlement with classification accuracy of 87% and 93.5 %, respectively, as seen from study in India and Jamaica. However, transfer-ability of rule-set are limited due to their variation and dynamics in spatial and temporal domain [28]. Thus accuracy varies widely [11].To some extent, these rule-set can be generalized using ontology-driven model over data-driven model [24].

Development in UAV and deep learning techniques has taken informal settlement extraction to next level. Number of researches have focused in deep learning and UAV these days, which we discuss in section 2.2 and 2.3.

2.2 Deep Learning and UAV in Informal Settlement

UAV offer flexible alternatives to satellite imageries, providing cloud-free very high resolutions at any time[16, 32]. Additionally, availability of 3D point cloud from the same platform opens door for access to varieties of products from same mission. Research community has been using UAV in the domain of building extraction[15], crop monitoring, weed monitoring [32], informal settlement [12, 17, 21] and several other domains. However, the applicability of UAV is partly limited by availability of spectral information in only visible bands. Additionally, high resource requirement for processing on larger extent is questionable issue. But, at the same time, we can take advantage of 3D point cloud and DSM [16].

A number of authors have published their work with contribution on several aspects of informal settlement using UAV using deep learning. [21] and [9] experimented on extraction of informal settlement from low to medium

resolution images. On the other hand, [12],[19], and [18] conducted their work on enhancement of segmentation result of UAV images from deep learning by incorporation of vertical information in form of DSM, nDSM or point cloud. Applicability of transfer learning and its limitations has been analysed by [22] in 2019. They observed that, despite the availability of large number of pre-trained images and models, their transerability and generalizability is limited [5] especially from RGB domain to extended domain as SAR[22] or DEM based products.

One of the the major researches on integration of high resolution images and deep learning approach has been made by [21]in 2017. It was observed that having deeper network and larger window allowed training model for complex scene and using larger context window. In their case study from Tanzania, CNN with kernel size of 7x7, patch size of 129 among[65,99,129,165] and network of five convolution layers performed best on extraction of both informal as well as other areas. CNN with five convolution layers yield overall accuracy of 91.71% compared to baseline method of SVM +LBP with accuracy of 90.48%. Interestingly, drop in accuracy by 0.18 % was observed on increasing depth by one more layer. It is attributed by less number of samples to learn from, relative to number of parameters to be learnt.

Next year, in 2018, applicability and transfer-ability of segmentation model SegNet on extraction of buildings was assessed by [15]. The architecture was able to segment buildings with accuracy of more than 90 % in UAV datasets from two riverbank of different scenarios. Confusion between ground and buildings were however observed in small buildings. The performance of the model was verified independently in Postdam dataset and led to promising results.

As these images are expensive and resource-hungry, [9] in 2019 proposed two machine learning models called cost-effective model and cost-prohibitive model for slum extraction. While cost- effective method used Canonical Correlation Forests(CCF), to deal with low resolution image, cost prohibitive method were designed to perform segmentation on very high resolution DigitalGlobe image. Segmentation architecture of DeepLabl V3+ built on top of Xception 65 as backbone, trained and tested on 30-50 cm resolution image resulted on IoU of upto 83% compared to its counterpart CCF limited to 74.0% . This highlighted the advantage of using high resolution images as well as context information at the same time to enhance performance of the model.

However,all of these were based on 2D products, and did not consider vertical dimension.

2.3 Enhancement of Feature Extraction with Height Information

Having information on vertical dimension provides opportunity for segregating features with similar spectral characteristics, but lying on different plane [17]. Availability of 3D point cloud has been proven to greatly reduce misclassification due to shadows [12, 17, 19] and extract precise building footprints. Segmentation with region growing algorithm with 3D point cloud were able to segment roof structures of a building, which look similar in spectral domain, as observed by [17]. [19] also noticed that including DSM could help identify narrow gaps between buildings, which are not distinguishable with 2D products.

But,[18] claims that adding DSM to a RGB image as additional channel does not necessarily increase the performance of deep learning models. They suggest the better way of enhancing segmentation accuracy using nDSM is to perform a multi-stage training with DSM as a separate backend. They propose to correct errors of false top-hat and false ground on initial output of RGB segmentation through the morphological filters and use DSM as backend. This offers the advantage of removing the erroneous segmentation from the first stage in the second stage. In their research, FCN constructed on top of Resnet, with Maximum Fusion Strategy(MFS) resulted in accuracy of 90.6 % in ISPRS Vaihingen dataset, showing 2 % improvement compared to using RGB only . Interestingly, DSM added as additional channel had reduced accuracy by 0.9% . Authors claim the homogeneity in DSM values in smooth areas restrict the model from learning when the height itself has been added as channel.

[19] in 2020 tested the performance of building extraction with integration of RGB images with additional channels of DSM, Visible Band Difference Vegetation Index (VDVI). Outputs from two different sites with different characteristics showed that having higher number of indices in addition is always beneficial. However, the contribution of the indices might differ in magnitude according to site characteristics. For example, integration of DSM was more significant in the region with large buildings where VDVI didnot contribute on accuracy. In contrast, RGB with both nDSM and VDVI was observed to

perform superior in site with small buildings. The result was validated using Postdam dataset. They observed more than 4 % increment in per-class IoU for building as well as non-building features yielding overall accuracy of 97.14 % in contrary to 95.79 % with RGB image.

The work very similar to our proposed work was performed by [12] in 2017. They observed that including 3D and 2.5D information to UAV could segment informal settlement with accuracy upto 95.2%, which is limited to 73.8% for only radiometric properties. UAV-borne products were categorized into radiometric, textural, 2D segment, 2.5D topographic, 3D spatial binning,3D spatial binning and 3D point-based features. Using support vector machine, the model was trained with different combination of these features.

Results from two sites(Kigali and Maldonado) were consistent despite the difference in settlement characteristics. They however found that the major contributor might differ for features to features and scene to scene. For an instance,buildings, may it be high or low, were better segmented using RT2S3(radiometric, textural,2D segment and 3D features) and bare surface using DSM. Despite the use of these features, confusion between corrugated zinc roof and terrain were persistent in all experiments, which were obviously less in magnitude than with planimetric features only.

While normalized DSM represents the height of feature above terrain, setting a threshold and segmenting them helps limit the range of concern. While these help to exclude the irrelevant range of height out of study, it extends the region of concern to cover the available DN range of image. Thus, we have proposed the study on impact of height information on informal segmentation with variation in threshold height.

Chapter 3

Theoretical Background

This chapter delivers the theoretical concept of neural networks, CNN, semantic segmentation.Section 3.1 delivers basic concept of Artificial intelligence, machine learning and deep learning. In section 3.2, we discuss briefly the architecture of CNN. If is followed by discussion on parameters and hyperparameters associated with ML model and their optimization, in section 3.3 and 3.4. This chapter ends with discussion of accuracy metrics and architectures of semantic segmentation in section 3.6, where we discuss of architectures chosen for the research,with focus on their backbone architectures.

3.1 Artificial Intelligence, Machine Learning and Deep Learning

Artificial intelligence has been popular in recent days among analysts and data scientists due to its capacity to learn from data and process complex data as well. It mimics the performance of human brain [33]. Machine learning is known as subset of Artificial Intelligence, which learns from data themselves without being programmed to perform explicitly the particular task. Deep learning are specific type of machine learning which get their name from being deeper, and thus learning complex systems.

Building blocks of neural network are the interconnected neurons,forming the highway for flow of information from input to the output. A generalized neural network has been shown in figure 3.1. It mainly consists of an input

layer, one or more hidden layers and a output layer. While input layer is analogous to dendrites of human neuron, the output layer corresponds to axons of human system. The information between these two layers is carried by a number of hidden layers, equivalent to neuron of human brain [2].

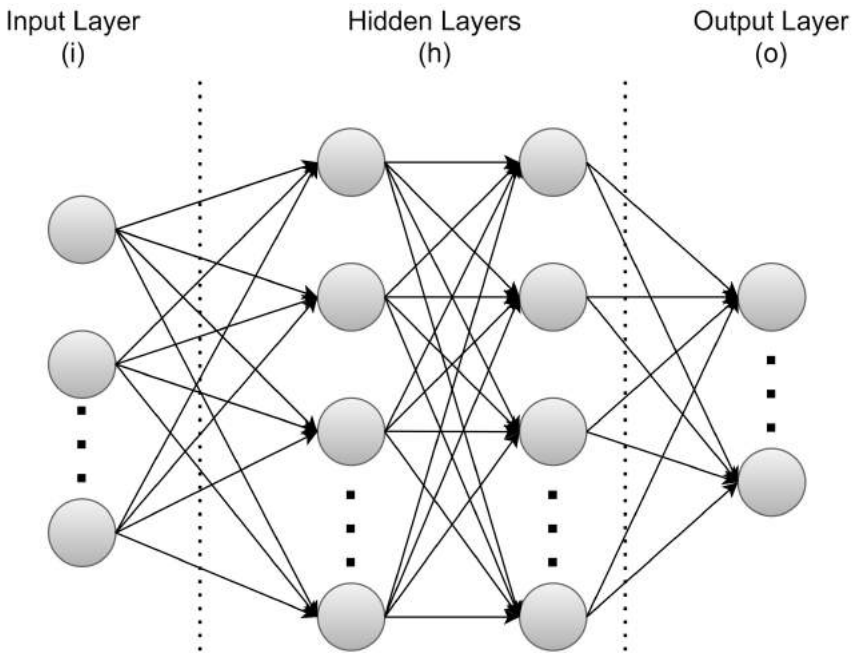


Figure 3.1: General Structure of Artificial Neural Network (ANN)

There is not a strict boundary between deep and shallow neural network however. But a universally accepted fact is deeper neural network learn complex things better. Processing and resource cost also increases with it, making it unsuitable for simpler problems.

3.2 Architecture of Convolutional Neural Network

Convolutional Neuron Network (CNN) is a type of neural network with specialized hidden layers. It gets the name from 'convolution' operation applied in hidden layers. These are specialized in a sense that these area able to detect patterns, features and objects of interest for the purpose. So, these are more suited for image data. While shallow layers of the CNN detect edges, deeper networks detect complex objects and features [34].

The convolution operation can be defined as dot product between weight matrix and part of source image. Moving a single weight kernel throughout the whole image offers computational flexibility and parameter sharing. Convolutional layers perform main operations of (i) convolution and (i)pooling,

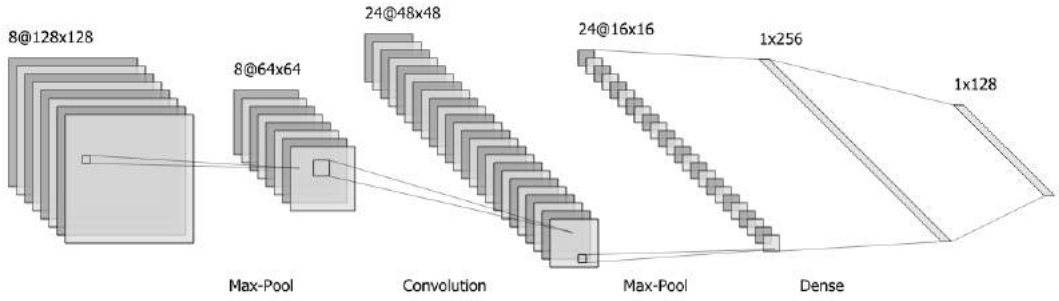


Figure 3.2: Illustration of CNN . Width and height of feature map reduces on going deeper while depth increases.

Number of filters are applied at convolution layer to change depth of feature map, which is subjected to linear or non-linear activation. The resulted feature map's dimension is further reduced using a pooling windows, usually 2x2, that calculates maximum(or average) score of the window and write to a new feature map.

At the end, the the fully connected layer comes after a series of convolution and pooling layer which calculates the class score for each pixel (figure 3.2). Following section describes details of the components of a convolutional layer

3.2.1 Convolution

Mathematically, Convolution is the dot product of two matrices, the first being the kernel with weights, and second the part of image with same dimension and overlapping with it. If k number of square filters of size l span a source image of dimension(width x height) W*H, the resultant feature map has a dimension of *(W-l+1) * (H-l+1) * k*, if the convolution is applied without jump, i.e. with stride =1. However if the stride of s is applied, the size of resultant feature map is further reduced to $(\frac{W-k}{s}+1)*(\frac{W-k}{s}+1)*k$ [35].

Convolution reduces the width and height of feature map. If required, p pixels with zero values can be padded on each side of the image, which after convolution results on a feature map of dimension $(\frac{W+2p-k}{s}+1)*(\frac{W+2p-k}{s}+1)*k$. Conversely, if the image dimension is to be preserved,and stride of 1 is used with a kernel of dimension l,$\frac{l-1}{2}$ pixels need to be added on each side of the image to preserve dimension. This approach of padding is called *same padding*.

If D is the depth of input image, then the pixel value of the (i,j) position on

the k^{th} channel of the output feature map is calculated as:

$$x_k(i,j) = \sum_{d=1}^{D} \left\{ \sum_{p=0}^{l-1} \sum_{q=0}^{l-1} x_d(i.s+p, j.s+q).h_k(p,q)) \right\} + b_c \tag{3.1}$$

where $x_d(i,j)$ is the value of input pixel at (i,j) in dth channel, s is the stride, $h_k(p,q)$ is the weight value at (p, q) at k^{th} filter, b_k is the bias parameter of k^{th} filter.

It can be inferred from previous discussions that the effective receptive field increases on going deeper, as it is the result of a number of convolutions. This is the reason for deeper layers being able to detect features instead of just edges.

3.2.2 Activation

Values from convolution need to be converted to standard range and distribution, which is obtained by activation functions. These take input from convolution and transform them step-wise, linearly or non-linearly using mathematical functions. These activation functions are linked to neurons and normalize data between range of 0 and 1 or -1 and 1. Activation functions can be broadly categorized as binary, linear and non-linear. Some commmon activation functions used in ANN are sigmoid,TanH,ReLu and Softmax [36].

Sigmoid or logistic activation function transforms input value between 0 and 1 non-linearly, and so are used on fully connected layer to predict class probabilities. Being non-linear, these are differentiable and thus can learn from back-propagation, unlike binary activation. If x_{ij} is the input value to sigmoid activation, output value after activation is calculated as:

$$A(x_{ij}) = \frac{1}{1+e^{x_{ij}}} \tag{3.2}$$

Due to flatter shape at extreme ends, it's gradient tends to be zero for very high and low input and cannot learn from input.This is termed *vanishing gradient* and is common limitations of sigmoid activation in deeper networks [36].

Another activation function very similar to sigmoid activation is hyperbolic tangent or *tanh* activation, except it returns the hyperbolic tangent as:

$$A(x_{ij}) = tanh(x_{ij}) \tag{3.3}$$

and transforms input to range of [-1,1].

Vanishing gradient problem of these saturating sigmoid functions is solved by introduction of Rectified Linear Unit (ReLU) activation. It transforms negative value to zero, preserving non-negative activations. Thus, it is monotonic

and can compute efficiently on any input value. ReLu activation is mathematically expressed as:

$$A(x_{ij}) = max(0, x_{ij}) \tag{3.4}$$

One pronounced limitation of this activation is found in data with significant proportion of negative inputs, which are immediately converted to zero and cannot learn further. This cause the neurons to be dead and never activated again. Leaky ReLU addresses this issue of *dying ReLU* by introducing a very small gradient near zero, that keeps the neurons active for learning and propagation [35].

Most commonly used activation in classification problems is softmax activation. It converts the vector of values from convolution to the corresponding vector of probabilities of same size. Thus it results in per-class probability score, between 0 and 1, one probability per class. If $(x_1(i,j), x_2(i,j), ..., x_K(i,j))$ be the vector of inputs at a pixel at position (i,j) in the input, its corresponding activation or probability at k^th channel is calculated as:

$$A(x_k(i,j)) = \frac{e^{x_k(i,j)}}{\sum_{d=1}^{K} e^{x_d(i,j)}} \tag{3.5}$$

where K is the number of channels (or number of classes). This is applied by applying exponential function to each element and normalizing each value by their exponential sum that adds up to 1. This is used in the output layer.

3.2.3 Pooling

Dimensional reduction of image after convolution is obtained using pooling. This summarizes the information available within a window of source image into a single pixel of new feature map using statistical measures. Maximum is the most commonly used statistics, making max pooling the common pooling. This reduces the size of output by the factor of size of pooling window. Pooling with window of 2 x 2 is commonly used. This increases equivalent receptive field, while preserving dominant signal or feature from the previous layer. Figure 3.3 shows a shematic representation of max pooling. The key benefit of using pooling is *local translative invariance*,which helps the network to identify the features even after translation.

One or many fully connected layers(s) follow convolution layer. While convolution layer is responsible for encoding information in compact form or feature extraction, fully connected layers compute scores for classification. The final layer feature maps are represented as vector of values passed to the fully

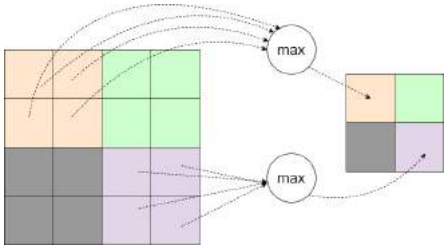


Figure 3.3: Schematic Diagram of Pooling mechanism, after [35]

connected layers. These values are then converted to prediction score, or numerical value using appropriate classifier discussed in section 3.2.2 .

Some standard architectures VGG and Alexnet have two to four fully connected layers. Attempt have been made in recent years towards decreasing computational expense on this part by use of global average pooling.

3.3 Loss Functions and Optimizers

Deep learning models are capable of learning from the label data. This means, the initial weights of a kernel are estimated randomly or by some normally distributed values. The predicted value of the label at end of iteration compared to label value quantifies error made in prediction of each pixel. These losses are expressed using mathematical functions called *loss functions*. Magnitude of loss calculated at any epoch is thus subjected to choice of loss functions.

Losses are used to evaluate the error in prediction and update the weights , using *optimizers*. Optimizers are the functions those are used to minimize the loss. These seek the new value of weight, which is propagated back through the process of *back-propagation* unless the user specified number of epochs or user-defined condition is met.

Choice of loss functions is guided by the purpose, and accuracy of a model is determined by the chosen loss function. For example, while mean squared error is used commonly in regression problem, binary cross entropy better suits binary classification. Similarly, use of cross-entropy is recommended for multi-class classification. In addition, several other loss functions such as dice loss , focal loss, etc. have been introduced for class imbalance problems and semantic segmentation. Cross-entropy is simply the sum of log of mis-classification score and is expressed as :

$$\sum_{i=1}^{c} t_i log(s_i) \tag{3.6}$$

where c is the number of classes, t_i is the ground truth value for class i and s_i is the predicted score or probability calculated for that class. If binary classification is considered, equation 3.6 reduces to the form:

$$CE = -\sum_{i=1}^{C'=2} t_i log(s_i) = -t_1 log(s_1) - (1-t_1) log(1-s_1) \tag{3.7}$$

where C_1 and C_2 are the two classes,t_1[0,1] and ,s_1 are the groundtruth and the score for C_1, and $t_2 = 1 - t_1$ and $s_2 = 1 - s_1$ are the groundtruth and the score for C_2.

Binary cross-entropy and categorical cross-entropy commonly used assume class balance. These tend to minimize overall loss by assigning equal weights to loss of each of the classes. However in imbalanced class,mis-classification of one class might be more significant than that of other class[37]. So, focal loss and dice loss perform better in class-imbalanced, where the relative importance of each of the classes is specified and loss computed accordingly.

These losses are optimized by optimizers, which approximate the hyperparamters for next iteration from the current weights and assigned learning rate. Gradient Descent is one of the common optimizers used in deep learning, and is expressed as:

$$\theta = \theta - \alpha . \nabla J(\theta) \tag{3.8}$$

But this traditional optimizer is too slow and may trap at local minima. To overcome it, stochastic gradient descent(SGD) is introduced. The key difference is that SGD updates model parameters after each sample as:

$$\theta = \theta - \alpha . \nabla J(\theta; x(i); y(i)) \tag{3.9}$$

where x(i) and y(i) are the i^{th} samples of training examples. Considering every sample, it is subjected to more calculcation and more varience. A midway between gradient descent and SGD is provided by mini-batch gradient descent, that splits training sample into mini-batches of some images and update weights after each mini-batch.

Sometimes momentum is introduced in optimizer, that helps to direct the loss function to head to the relevant direction. But, choice of high momentum might impose risk of the loss function missing the minima.

All over this is Adaptive Moment Estimation (Adam), which works with momentum of first and second order. It is appreciated for its rapid convergence.

3.4 Hyperparameters

Hyperparameters are the model configuration those tune or influence the performance of the model. Some of the hyper-parameters used in deep learning has been discussed briefly here.

Learning rate is an important hyperparameter governing the rate at which the weights will be updated. With higher learning rate, the model updates the weights fast but the risk of skipping minima is always there. In contrast, despite smoother learning, lower learning rate may lead to very long training time, making it too slow that the minima might not be reached.

Number of epoch defines the number of iterations considered for optimizing weights. It is selected in synchronization with learning rate, such that the loss curve remains stable, and the model stops learning.

Mini-batch size refers to the number of training samples to be considered as one batch for updating weights. While large batch size might yield visually pleasant learning curve, it is also more generalized. But the computational cost increases as larger number of samples need to be scanned. In contrast, smaller batch size may capture too much details including noise, limiting generalisability of the model.

Activation functions influence the output of a model, as has been seen earlier. So, the choice of appropriate function is another consideration in hyperparameter tuning.

Regulariziers are used with optimizers to make the model generalized, i.e. to prevent overfitting. An overfitting model performs perfect in training data but poorly on unseen data. Common optimizers used for regularization include L1 regularizer, L2 regularizer and dropout.

- **L1 regularizer** and **L2 regularizer** penalize the possible biases due to large errors. While L1 regularizer take into consideration the magnitude of error, L2 takes the square of error and penalize for it. As square of large errors tend to be high and can be avoided, L2 regularizer is more popular in choice [38].

- Another way of ensuring generalise-ability is the use of **dropout**. As the complexity of model increases with number of neurons and connections, dropping some neurons randomly and their connection simplifies the model avoiding overfitting [33]. However,excess dropping might make the model too simple and not enough to capture complexity of the subject matter.

- **Data augmentation** comes as handy tool for overfitting when the training data volume is less. By applying flip, rotation, minor scaling, or their combination, complexity is added to the data and to make it more representative to general scenario [39].

- **Early stopping** also prevents overfitting by stopping training at the point where loss curve of validation data diverges from that of training curve. This is controlled by monitor and patience as parameters, where monitor assigns the metrics to be used to check of overfitting, and patience specifies the number of epochs to watch for.

Selection of a best performing model is quite significant. Hyperparameter tuning consists of selecting combination of hyperparamters and picking the best performing combination. **Grid search** is one of the methods for it. Using appropriate accuracy measures discusses in section 3.5, the best performing combination of configurations can be further used.

3.5 Accuracy Metrics

Accuracy metrics evaluate the performance of a model. A good model is the one yielding good result on test as well as validation and external dataset. Accuracy metrics should be chosen considering purpose and data distribution.

The simplest accuracy metrics used in classification and segmentation is overall or pixel-wise accuracy. It is expressed as a ratio of correctly mapped pixels relative to the total number of pixels. Being unable to address issue with class imbalance, it is not suitable for all cases.

So, precision and recall are commonly used in semantic segmentation as accuracy metrics. While precision indicates the fraction of true positive among positively predicted values [33], recall indicates the fraction of positive labels those were correctly predicted as positive. Thus, these are expressed as:

$$Precision = \frac{TruePositive}{TruePositive + FalsePositive} \tag{3.10}$$

$$Recall = \frac{TruePositive}{TruePositive + FalseNegative} \tag{3.11}$$

Equations 3.10 and 3.11 suggest that precision value is more influenced by false positive while false negative has impact on recall. Both of these have mathematical range from 0 to 1. Having both values higher is ideally not

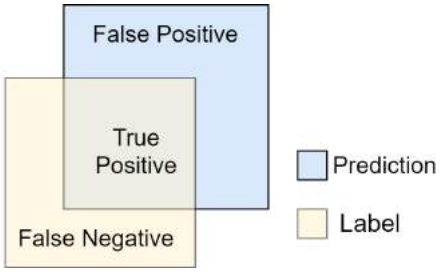


Figure 3.4: Illustration of Intersection over Union(IoU)

possible as the correct classification of one class comes with cost of incorrect classification of another one.

So , a term **F-score** or **Dice score** is introduced, that measures the weighted harmonic mean of precision and recall, with default weight being same for both classes. The weight is controlled by term β. If equal weight is assigned for precision and recall, F_1 score is calculated as

$$F1 - score = \frac{2 * Precision * Recall}{Precision + Recall} = \frac{2 * TP}{2 * TP + FP + FN} \tag{3.12}$$

where TP,FP and FN refer to True Positive, False Positive and False negative respectively. In binary classification, the converse of Precision and Recall are negative predictive value and specificity respectively,and are calculated the same way as precision and recall.

Positively correlated with Dice score is IoU (Intersection over Union), calculated class-wise from confusion matrix. It is the expressed as the ratio of area of overlap between true and predicted class normalized by total area covered by them. Mathematically, IoU for positive and negative class in binary classification is calculated as:

$$IoU_{positive\ class} = \frac{TP}{TP + FP + FN},\ IoU_{negative\ class} = \frac{TN}{TN + FN + FP} \tag{3.13}$$

This is more popular in semantic segmentation and object detection, as it offers calculation of IoU per class and also works well with class imbalance too. In our study, we will analyse accuracy using this metrics. Figure 3.4 illustrates it graphically.

3.6 Segmentation

Beyond classification problem is segmentation. Segmentation is the computer vision algorithm of assigning pixels or their continuous group to a predefined class. These deal with what is in the image, as well as where they are.

So,localization is crucial in segmentation [35]. Segmentation comes in two forms. While semantic segmentation assigns each pixel to predefined class without identifying individual object in scene, instance segmentation also identifies individual of them. Semantic segmentation architecture is composed of

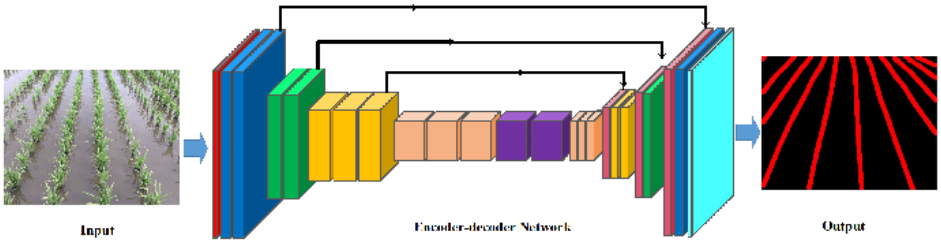


Figure 3.5: Schematic Diagram of Encoder-Decoder network in Semantic Segmentation , after [40]

encoder or the contraction path and decoder or upsampling path. Encoder path is composed of convolution and pooling layers. Going deeper into the network extracts high level features or semantics. This is followed by decoder or deconvolution path that symmetrical up-samples the output till the original image shape is reconstructed. These replace fully connected layer of CNN. Each layer of deconvolution layers receives features from its counterpart in encoder part through skip connection . These are combined in some way and in number of stages to prepare labelled output in same dimension as input image.

History of semantic segmentation dates back to the fully Convolutional Network (FCN), on which updates have been made to achieve popular models such as Unet, FPN and PSPNet. These segmentaion models have been discussed in section **??**.

3.7 Overview of Architectures used

First part of semantic segmentation model come from standard encoders those act as feature extractor. The later part or the decoder performs deconvolution to prepare segmented map.

VGG16 is the model submitted to LLSVRC-2014. It reduces the processing cost of previously existing architectures such as AlexNet by using series of smaller kernel of 3x3 instead of large-sized kernels[41]. This provides increase in receptive field, reducing number of parameters to be learnt at the same time. Convolution layers consist of maxpooling layers after each block consisting of

a number of convolutions. With progression, while the size is halved, depth of feature space is doubled. The use of three non-linear layers in the fully connected layers instead of a single makes the model more descriminative. Original architecture, trained on 224 x 224 RGB images resulted in remarkable performance with top 1 validation error of 23.7%, top 5 test and validation error of 6.8 % . Varying the number of weight layers gives rise to different variants of VGG such as VGG16 and VGG19.

Though convolution networks learn better on going deeper, the training error increases on going very deep. **Residual network or Resnet** offers solution on it by adding back *'identity block'* to the network through skip connections, as shown in figure 3.6. Adding the identity shortcut feeds low level information from shallow layers to deeper layer after some layer helping to preserve detail. Residual networks are easy to optimize despite their deeper depth. Adding

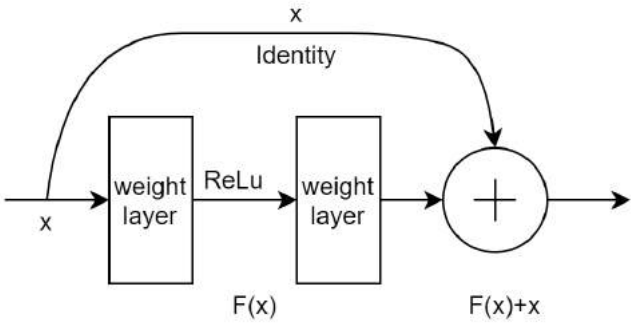


Figure 3.6: Schematic diagram of identity block of Resnet, after [42]

the identity block does not add computational complexity or number or parameters. Rather it helps to retain features. While training error increases on plain networks, residual networks rather had improved performance of deeper network. It's basic architecture is also inspired by that of VGG. When trained on 224 x 224 image randomly resized from source 256 x 480 image, it resulted in top 5 validation error of 4.49 %.

Chapter 4

Dataset Preparation

4.1 Study Area

Nine sites from three provinces of Nepal are chosen for the purpose of this thesis. These nine sites of informal settlements have their unique characteristics and vary in terms of roofing material, context, distribution and extent. Figure 4.1 shows the location of the study sites along with their orthophotos. In Kathmandu, figure shows the whole stretch of Bagmati river,containing four informal settlement blocks.

Study sites(figure 4.1) and their characteristics have been listed in table 4.2. Each of these sites contain very less percentage of informal settlement ranging between 6.5 to 15.21 percent (table 4.1). This is because the informal settlements in Nepal are small clusters distributed in different geographical region, in contrast to those in big cities such as Dharabi, Mumbai [9].We discuss briefly about each of these sites in following paragraphs.

Kathmandu, capital city of Nepal has large number of informal settlements, majority of them along the bank or Bagmati river. Our study area has four informal settlements at Balkhu, Teku, Thapathali and Shankhamul. All of these, except those at Thapathali are characterized by corrugated zinc roofed clusters along the bank of Bagmati river. Informal settlements in Thapathali has mixed roof of plastic and zinc roofs. These are separated from formal settlements by a narrow and dirty road. Though some vegetation can be seen around settlement, inner part of settlement has no vegetation at all.

Simpani is located along the bank of Seti River, Pokhara of Gandaki province.

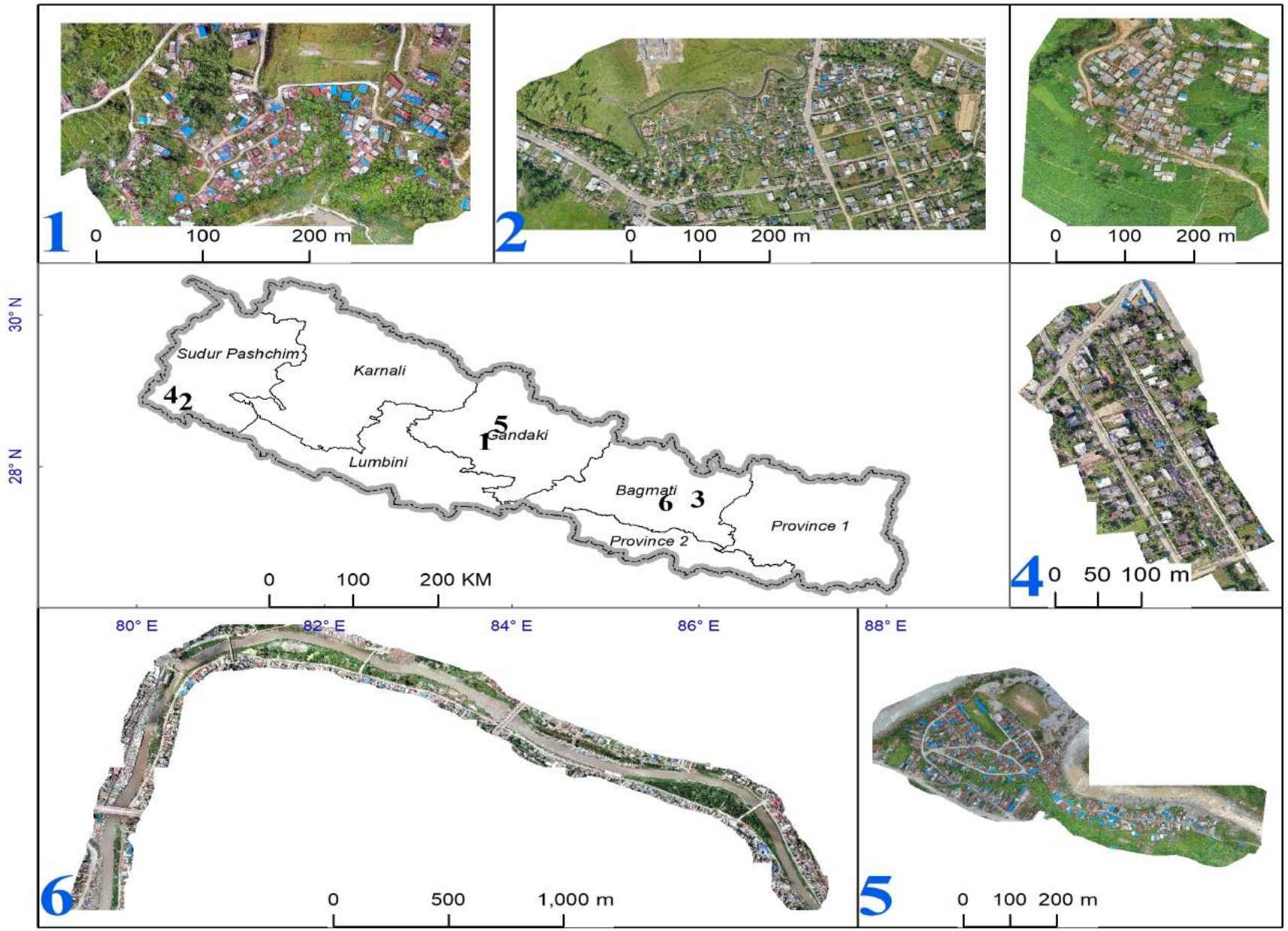


Figure 4.1: Study area: (1) Chhorepatan, (2) Saalghari, (3) Indrawati (4) Kataan (5)Simpani and (6) Kathmandu.Image from Kathmandu has four different informal settlements namely Balkhu, Teku, Thapathali and Shankhamul

Table 4.1: General Description of informal settlements in study area

Location	E-W Extent (m)	N-S Extent (m)	Total Area (Hectares)	Coverage Percentage		
				Informal Settlement	Formal Settlement	Others
Balkhu	330	510	9.31	15.21	16.9	67.88
Chhorepatan	428	311	8.13	11.99	4.53	83.48
Indrawati	223	257	3.53	15.58	0	84.42
Kataan	272	452	5.95	6.5	16.75	76.75
Shankhamul	407	428	5.92	13.56	16.8	69.64
Saalghari	680	342	17.63	7.87	8.46	83.67
Simpani	845	527	17.76	12.22	11.28	76.5
Teku	320	340	3.88	18.44	13.47	68.09
Thapathali	534	432	10.18	7.26	14.3	78.44

Table 4.2: Characteristics of Informal settlements in study areas

Location	**Terrain Type**	**Roofing Material**	**Settlement Characteristics**
Balkhu	Flat	Corrugated zinc	Densely constructed, isolated
chhorepatan	Flat to mild slope	Corrugated zinc	Densely constructed,mixed
Indrawati	Flat to mild slope	Corrugated Zinc	Spaced, isolated
Kataan	Flat	Straw	Sparse,mixed
Sankhamul	Flat	Straw	Densely constructed, isolated
Saalghari	Flat	Corrugated zinc	Sparse,mixed
Simpani	Flat to mild slope	Corrugated Zinc	Densely constructed,mixed
Teku	Flat	Corrugated Zinc	Densely constructed, isolated
Thapathali	Flat	Plastic	Densely constructed, isolated

It is characterized by mixed settlement on flat to mild slope. While all of the informal buildings are single-storey and roofed with corrugated zinc, formal buildings are mostly multi-storey and roofed with concrete or corrugated zinc. Roofs are visible in orthophoto and has very less vegetation.

Chhorepatan in Pokhara of Gandaki province is located on slope land in form of clustered buildings. Buildings are roofed with corrugated zinc and partly covered by vegetation. Settlement contains narrow and low quality road, and are separate from formal settlement.

Kataan lies in Far-western Province of the country, and is located on flat terrain. Clustered single-storey informal buildings with straw roof are surrounded by formal single to multi-storey concrete buildings. Settlement contains little vegetation, and informal building tops are partly covered with green

cultivated vegetable leaves.

Indrawati is informal settlement from modereate slope land in Sindhupalchowk district,Bagmati province. It contains a small village of spaciously constructed single-storey buildings with zinc roof.

Saalghari consists of single storey straw-roofed buildings covered partly by green vegetation. Informal settlements are separated from concrete single to multi-storey buildings by road.

4.2 Data Preparation

4.2.1 UAV Data Acquisition

We acquired very high resolution UAV images of study area except Kathmandu ourselves, using registered UAV with flight permission from authorities. DJI Mavic 2 pro equipped with 20 megapixel RGB camera with FOV of 77°and equivalent focal length of 35 mm [43] was used. Flight planning was performed using open source mobile application DroneDeploy[44]. It offers better flexibility on flight configuration and customization compared to its counterparts such as Pix4DCapture [45].

Table 4.3: UAV and flight configurations used

UAV Specifications		**Mission parameters**	
Parameter	**Value**	**Parameter**	**Value**
UAV	DJI mavic 2 Pro	Mission Type	Double grid
Camera Resolution	20 Megapixels	Flying Height	60 m above takeoff level
FOV	77 °	Foreward overlap	80 % or higher
Image size	5472 x 3648 pixels	Lateral overlap	70 % or higher
Positioning Sensors	GPS +GLONASS	Image Acquisition Date	September - October 2020
Sensor Type	RGB		

Double grid missions were used with flying height of 60 to 80 meters above takeoff level. This resulted in effective flying height of 40 to 120 meters because of terrain undulation. Unidirectional flight paths suffer hindered objects, especially on densely vegetated or builtup area. This tends to reduce density of point cloud and hence accuracy of orthophoto and DSM. We overcame this possible limitation using double grid mission having perpendicular flight lines

with minimum longitudinal overlap and lateral overlap of 80 % and 70 % respectively. Overlap was kept higher in area with low texture variation and presence of water bodies, to ensure better image matching and sufficiently dense point cloud. Figure 4.2 shows one instance of flight plan for the project. Detailed UAV specification and image acquisition configuration are listed in table 4.3.

Figure 4.2: Flight planning for UAV using DroneDeploy. Green lines represent flight path for the user-specified configuration

Oblique images were acqired along the peripheral flightl line in addition to regular nadir images. This ensures reliable DSM throughout the area reducing gaps and stretches on peripheral region due to no or less overlapping images.

Geotagged 8 bit RGB JPEG images, each of 5472 x 3648 pixels were obtained. As UAV is equipped with GPS and GLONASS, geolocation accuracy is expected to be precise enough for the purpose. Thus, no ground control points(GCPs) were used.

Orthophoto, DSM and dense point cloud of informal settlements along Bagmati river in Kathmandu valley was obtained from NAXA [46]. It contains four informal settlements at Shankhamul, Teku, Thapathali and Balkhu. Secondary source was preferred also because of permission issues for UAV flight in kathmandu.

4.2.2 UAV Image processing

UAV image processing was performed using open source application WebODM [47]. WebODM is popular among UAV user community due to its support on any operating system, for command line users as well as interface-users.

Built as native to Linux, it runs on docker for windows users [47]. We used a computer with 16 GB ram and 6 GB Nvidia GTX1060 graphics for processing of UAV images. Each of these processing took up to 25 hours depending on number of images.

Geotagged images were uploaded to WebODM and project configured to process on high resolution. Rest of the work was automated, as no GCPs were used. Due to lack of processing report options, outputs were assessed visually for possible distortion.

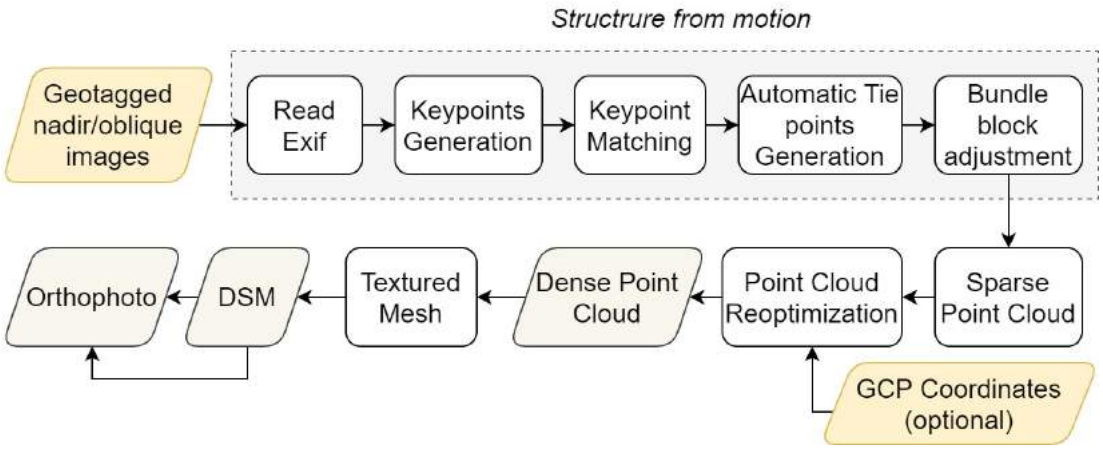


Figure 4.3: Workflow for generation of products from raw UAV images with webODM

General flow of UAV processing with webODM and other application is shown in figure 4.3. The application first reads EXIF data of each image to gather camera information such as camera position and orientation at the time of image acquisition. These quantify camera distortion. It next proceeds with OpenSfM library pipeline [32], where the EXIF data along with its image is used to extract key points from matching images. Bundle block adjustment follows next. Here the error is distributed in the model. Automatic tie-points and sparse point cloud is finally created as a preliminary model of scene [48].

On availability of GCPs, co-ordinates of GCPs are read,and point cloud is reoriented using user-entered coordinates of GCPs. The re-optimized point model is next used to extract further dense point cloud through Multi-view stereo reconstruction [49]. Density of dense point cloud is a function of surface texture variation, with denser point clouds being created in more heterogeneous scene. Next, dense point cloud constructs a three-dimensional polygonal model of surface,called textured mesh. This mesh along with dense point cloud finally form DSM and orthomosaic of the whole scene.

Successful processing of raw images from each site ended with (i) 8 bit RGB orthophoto with average resolution of 3 centimeters, (ii) DSM at same resolution as orthophoto, and (c) Dense point clouds. All of these outputs are in metric coordinate system using Universal Transverse Mercator projection. As Nepal lies in UTM zones 44 and 45, some of the products were projected in UTM44 and others on UTM45.

4.2.3 Image Data Pre-processing

One aim of our work is to see how inclusion of height and setting height threshold influences segmentation accuracy of informal settlement. As UAV image processing provides DSM with features, we started process with attempt to extract bare ground points, which could later be used to generate DTM and finally nDSM. Figure 4.4 summarizes the overall workflow. CloudCompare [50] was used for filtration of point cloud.

Cloth Simulation Filtering(CSF) plugin in CloudCompare offers segmentation of point clouds to retain only ground points or non-ground points. Segmentation algorithm is based on the work of [51]. The algorithm first inverts the point cloud vertically, and try to fit a rigid cloth above inverted cloth. The surface thus obtained by interaction of point cloud with simulated cloth defines the points lying on ground. The output of the algorithm is highly subjected

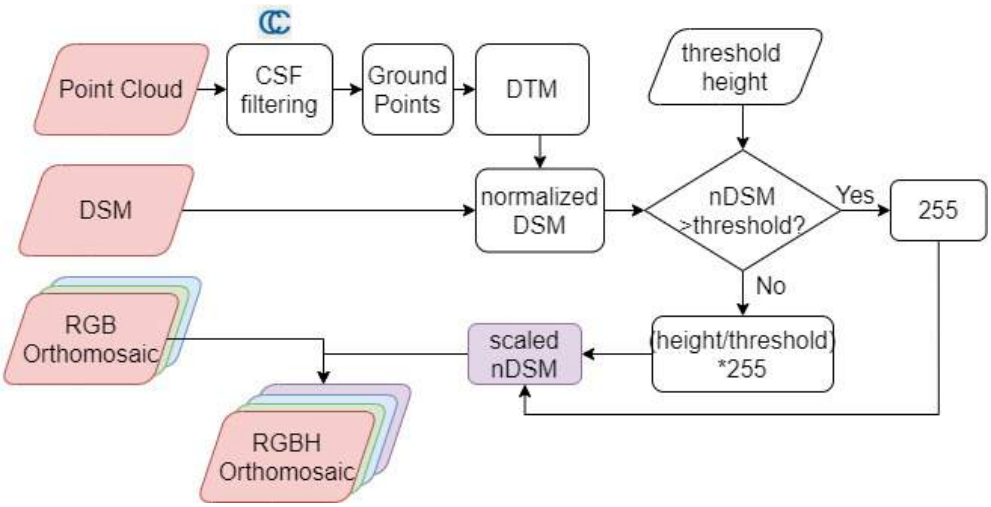


Figure 4.4: Workflow for preprocessing of image data

to choice of parameters tabulated in table 4.4. As not a single configuration works best in all scenarios, we tested the performance of each configuration by hit and trial method. Slope processing with steep scenes with cloth resolution of 2m and threshold of 0.3 m with 200 iterations offered the optimum result in

Table 4.4: Parameters governing output of CSF filtering,[52]

Parameter	Details
Scenes	Nature of Scene to be filtered (steep slope, relief or flat)
Cloth Resolution	Resolution of simulating cloth : Smaller size means higher resolution
Max iterations	No. of iterations for simulation
Classification Threshold	threshold distance to classify point as ground and non-ground

most of the cases. Very rare or no point cloud on filtered point cloud can be an issue, especially in dense settlement [12]. This situation was not observed in our study area, firstly because of smaller extent of continuous settlement and second due to the presence of open space within the settlement. We completed automatically extracted point clouds by manually segmenting point clouds on some parts, and removing left non-ground points. Digital terrain model(DTM) representing the bare earth surface at same resolution as DSM was then created interpolation using segmented ground point clouds.

Normalized DSM (nDSM) represents the height of feature above ground, and is calculated using DTM and DSM as in equation 4.1.

$$normalized\ DSM(nDSM) = DSM - DTM \tag{4.1}$$

where DSM and DTM refer to Digital Surface Model and Digital Terrain Model respectively. But one challenge with getting precise nDSM is the undulation in terrain and distance between extracted point cloud. It is not abnormal to have some negative values in nDSM due to filtering of point clouds. Small negative values were replaced with zero height, while no large negatives were observed.

As the research aims to see the effect of threshold value on segmentation accuracy, a number of threshold height were selected to threshold and normalize height data. All heights above the threshold are converted to 255 (maximum DN value for a 8 bit image) and those within threshold are stretchted to DN range of orthophoto using equation 4.2.

$$8\ bit\ nDSM = \frac{nDSM}{threshold\ nDSM} * DN_{max} \tag{4.2}$$

where,DN_{max}= maximum DN value for selected bit, corresonds to 255 in 8 bit image. The purpose os stretching is (i) to see the height difference in larger dynamic range, (ii) to make it uniform with existing RGB image for mosaiking as fourth channel. Figure 4.4 illustrates it graphically. In addition to normalized channels, nDSM in meters was also directly concatenated.

Furthermore,Visible Band Difference Vegetation Index(VDVI) was calculated using visible bands, after [15] as in equation 4.3.

$$VDVI = \frac{2\rho_{green} - (\rho_{red} + \rho_{blue})}{2\rho_{green} + (\rho_{red} + \rho_{blue})} \tag{4.3}$$

where ρ_{red}, ρ_{green}, ρ_{blue} are DN values in red, green and blue band respectively.

As the products we used up to this stage were in resolution of 2 to 4 centimeters, we downsampled the product to 25 cm resolution. In order to get rid of aliasing error because of very high downsampling ratio, downsampling was performed in multiple stages with downsampling ratio of 0.5 using cubic convolution. We ended our image data preprocessing with a set of multi-channel composite ,with channels as in table 4.5.

Table 4.5: Band Description for multi-channel orthomosaic

Band	Description
1,2,3	visible RGB
4	Absolute Height above ground(nDSM)
5 to 11	nDSM threshold and normalized by 5, 7.5, 10, 12.5, 15, 20 and 30 meters respectively
12	VDVI

4.2.4 Ground Truth Data Preparation

Cadastral data of the area are maintained by the respective cadastral survey offices. Especially for the informal settlements, these map showed only cadastral boundaries without buildings or their clusters digitized. Cadastral maps in modified UTM projection (used by Nepal for cadastral mapping) were transformed to UTM projection and overlaid over corresponding orthomosaic. In the region without buildings being mapped, buildings were digitized manually.

For the purpose of binary classification, the features were labelled as 1 and 2, 1 referring to informal settlements and 2 to all other features. We rasterized these buildings at same extent and resolution as orthomosaic. Table 4.6 shows the labelling scheme used for labelling ground truth data.

Table 4.6: Class definition for segmentation

Label	Class
1	Informal settlement
2	formal settlement and others

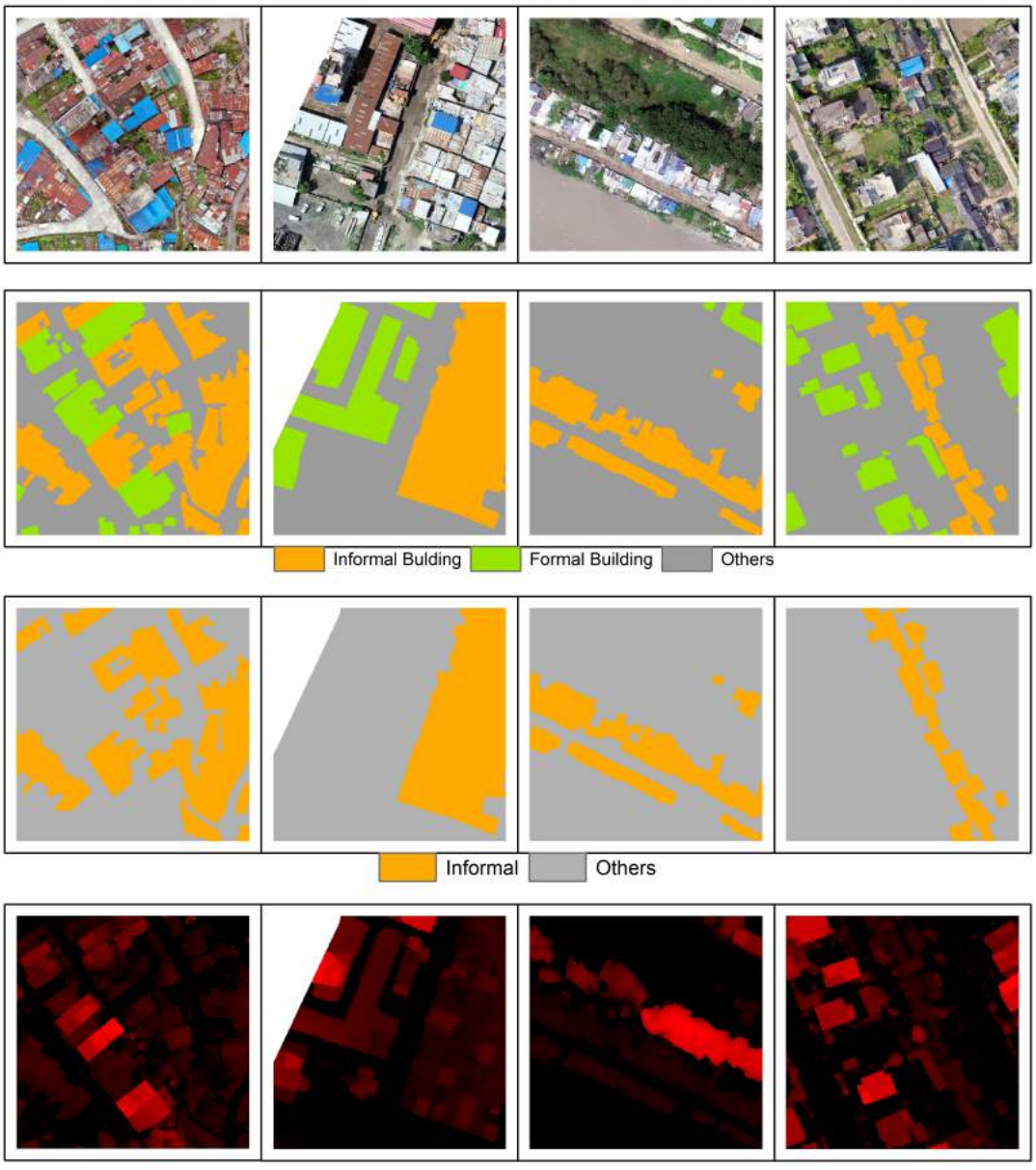


Figure 4.5: Sample data from Simpani, Balkhu, Thapathali and Kataan (left to right). First row: RGB Image, Second row: Multiclass ground truth, third row: Binary Ground truth, last row: normalized DSM in red channe. Multiclass label has not been used in study

Chapter 5

Methodological Framework

In this chapter, we discuss on the framework for selection of model and experimental arrangement. Data preparation for model has been discussed in section 5.1.In section 5.2,we briefly discuss the overall experimental framework and architectures in consideration.

It is followed by the experimentation and observations of choice of model, parameters and hyper-parameters in section 5.3, including performance analysis in section 5.7. Finally we talk of test for optimum threshold height and integration strategy in last section of this chapter.

5.1 Data Tiles Extraction

Multiple channel orthophotos(RGB, nDSM and scaled nDSM with variable threshold, VDVI) at resolution of 25 centimeters, and of variable extents were prepared from study areas mentioned in section 4.1. The label data of same extent and resolution were also prepared already (see section 4.2.4) with 1 referring to informal settlements and 2 to all other features. To account for distortions on orthophoto and class imbalance, images to the extreme edges were avoided.

A patch size of 384 x384 ,compatible with all segmentation models is adopted. Each of these represent 96 m x 96 m on ground. Considering irregular boundary of study area, tiles above size of 64 x64 were extracted during tiles extraction. Non-data part of the images are later handled during patch extraction

such that the patch were only extracted when the tile contained no data with NoDATA value.

During extraction of patch, overlap of 48 pixels was maintained between successive patches. This ensures sufficient overlap between successive patches, and also roughly balances the number of training samples taken, irrespective of the patch size.

Dataset was divided into train, test and validation dataset using *train test split* functionality of *keras* [53]. Due to variation in the scene,higher percentage of test and validation data were used. Whole data set was split into train and test data with 60% and 40 % with 30 % of train data further split as validation data.

Table 5.1 lists the number of tiles used for training, testing and validation for patch size of 128 pixels by 128 pixels.

Table 5.1: Number of image tiles used(patch size of 148. No. of images are subjected to change with patch size)

Training	Test	Validation	Out (full tile 384 x384)
1352	1289	580	7

5.1.1 Data Augmentation

Due to relatively less number of image tiles available compared to complexity of case, we performed augmentation of train images, whereas test and validation data were note augmented.

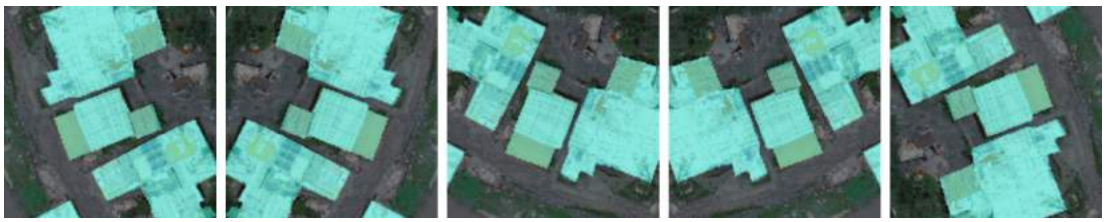


Figure 5.1: Augmented Images after five times augmentation with Positive Class Annotated

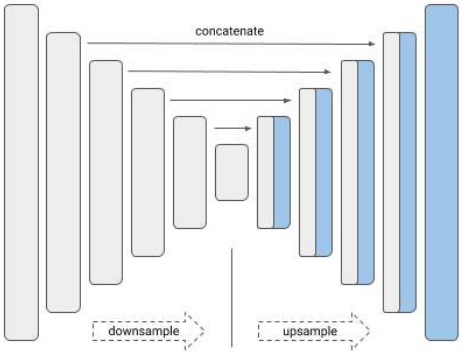
Data augmentation favors training in two ways: (i) it adds complexity to the data making it capable to know complex context, (ii) it increases the number of trainable images. In contrast, risk of unusual validation loss compared to

training loss is always there due to validation data being too simple compared to train data.

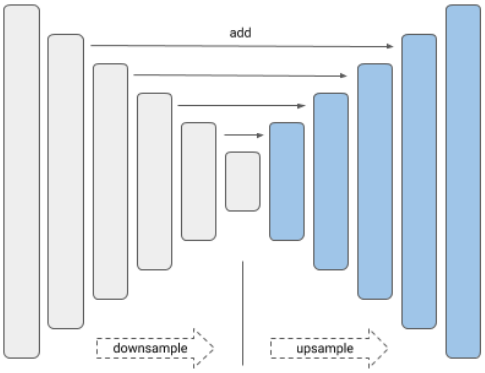
We applied five times augmentation as shown in figure 5.1. Rotation, mirroring, flipping, slight scaling and their combination were applied to generate complex images, and their corresponding label data.

5.2 Architecture

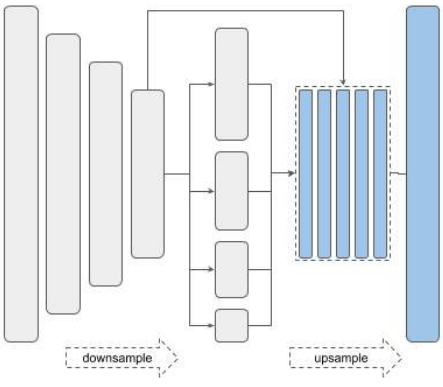
For the first phase of experimentation, two well-known architectures: Resnet18 and VGG16 with pretrained weights are chosen as backbone, on top of which, four segmentation models UNet, LinkNet, PSPNet and FPN are built.



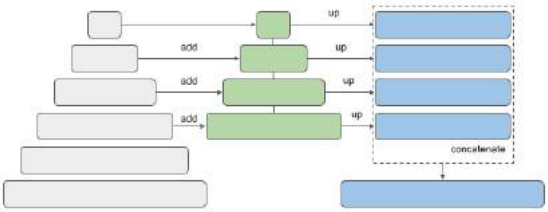
(a) Unet



(b) Linknet



(c) PSPnet



(d) Feature Pyramid Network

Figure 5.2: Segmentation models used (after [54]). Gray blocks represent feature space from downsampling path, blue from upsampling path

Figure 5.2 shows overview of architectures of these segmentation models. **Unet** (figure 5.2a), gets it's name from U-shaped architecture in terms of encoder-decoder path. Upsampling or deconvolution path are symmetric to downsampling path. At each level, details is regained from contraction path

and concatenated to double the number of channels. This is then subjected to convolution and upsampling until the same size as input is reconstructed[55].

Very similar to Unet is **Linknet**(see figure 5.2b). The major difference is that instead of appending extra channels from encoder path, the feature space is directly added to output from decoder path. This has an additional benefit of using identity blocks in the encoder path of Resnet encoder, enabling better remembering details from encoder[56] .

Another segmentation architecture, **Pyramid Scene Parsing Network** (PSPnet), takes benefit of use of pooling layer with multiple filters of varying size. Pooling layer consists of multiple filters of size 1x1, 2x2, 3x3 and 6x6, applied to the output of final encoder layer from resnet backbone (see figure 5.2c).This use of multiple size filters, followed by upsampling and concatenation to input helps preserve global to local context at same time [57]. This segmentation model requires input image with size multiple of 48 pixels on width and height.

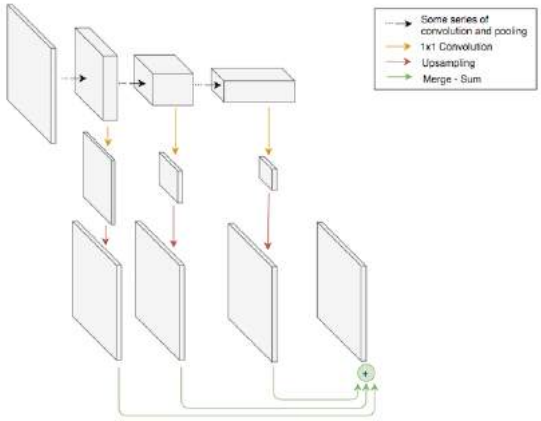
Feature Pyramid Network(FPN) is composed of bottom up path for downsampling, top-down path for upsampling and lateral connection between these [58]. Bottom up path reduces the size by factor of 2, doubling depth of feature map that forms pyramid shape. At every stage of top-down path, feature map is formed by merging 2x upsampled output from immediate higher level of pyramid, merged with 1x1 convolved output obtained from same level of bottom-up path. This way, both semantically strong part and spatially strong part are merged together to form better output.

For the second phase for the choice of integration strategy, a customized Fully convolutional network (FCN) with 1, 2 and three branches, as shown in figure 5.3 is used. Model is trained and tested in three different ways as follows:

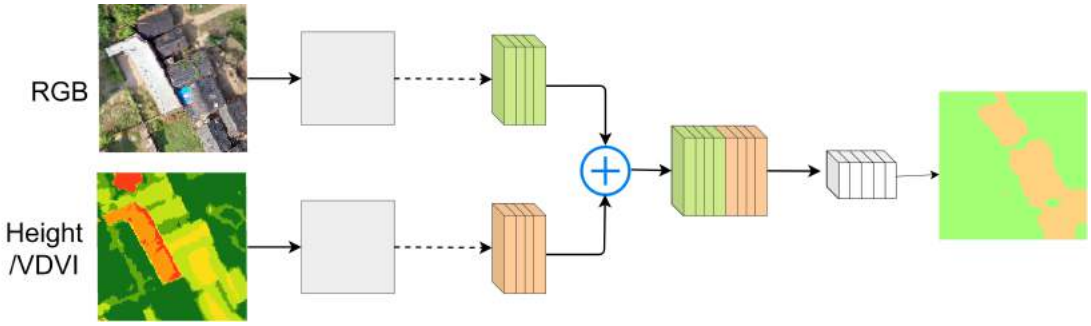
- **Single branch model**: All channels passed to a single branch
- **Dual branch model**: Two identical branches, first learning from RGB input, and second learning from height and VDVI. Outputs are merged after final convolution of encoder.

5.3 Test for Architecture

One of our aims is to select the best performing architecture from the list of available options. We performed number of trials on selecting the best model. General layout of experimental setup is shown in figure 5.4.



(a) Fully Convolutional Network



(b) Multi-branch model

Figure 5.3: Dual-branch Fully Convolutional Network for Parallel Learning. This model is used for test of best height integration approach

The experiment was accomplished in two phases, (i) first phase on selection of model parameters, hyperparameters and tuning, that ended with the optimum model identification and (ii) height threshold and integration approach test, where the optimum limiting height and integration method for extraction of informal settlement was tested.

Table 5.2 lists the combinations of parameters used for testing. In order to limit processing time and resource, test criteria is narrowed down through multi-stage test, reducing the number of combinations, starting from input patch size to then loss function and then model(see figure 5.4. Hyperparameters testing was then done on the selected model composed of selected backbone architecture, segmentation model and loss function respectively. In the second stage, selected model is utilized to perform height related experiments, with additional verification from custom model.

Table 5.2: Combinations tested for Network Configuration

Parameter	Selected Values
	Model Parameters
Backbone Architecture	VGG16, Resnet18
Segmentation Model	UNET, FPN, Linknet, PSPNet
Loss Functions	Categorical Cross Entropy, Categorical Cross Entropy + Dice Loss, Categorical Cross Entropy+ Focal Loss
	Hyperparameters
Learning Rate	0.0001, 0.001,0.005,0.01
Optimizers	Stochastic Gradient Descent, Adam
Number of epochs	Variable
Mini batch size	8,16,32
Final Activation	Sigmoid, Softmax
	Additional Parameters
Channels	as listed in 5.7
Patch size	64,96,128,192

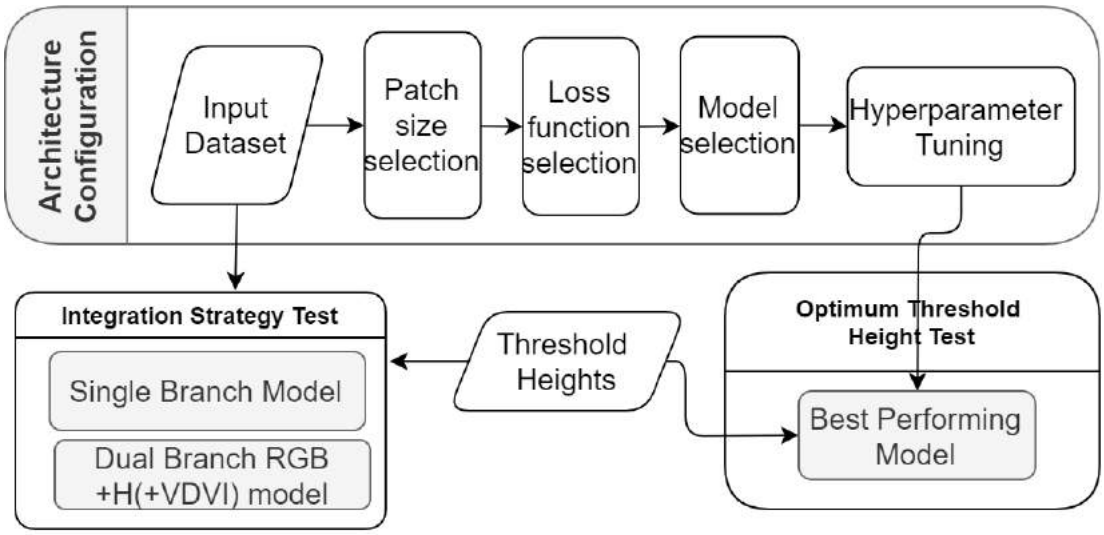


Figure 5.4: Experimental Setup: In first phase, optimum model is selected, and in second, optimum height and best height integration method is tested.

5.3.1 Test for Patch size

Among four segmentation models in use, three of them are compatible with patch size multiple of 32 while that of PSPNet is compatible with multiple of 48. So, patch size of 64, 96 128 and 192 pixels are considered for test, where PSPNet is tested at patch size of 96 and 192 only, while all three networks are

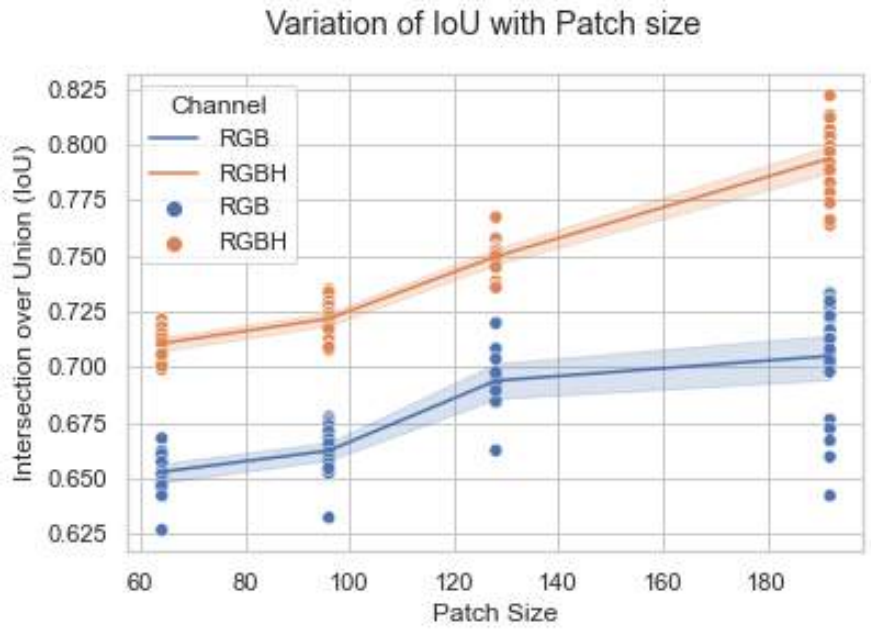


Figure 5.5: Model performance at various batch size: Larger patch size performed better. Performance was better with RGBH than RGB

tested on all patch sizes.

Running a combination of each segmentation model, with each backbone architecture, and loss function gives theoretically 108 combinations each for RGB and RGBH channel respectively, making total 216 combinations. Better mean IoU were observed with increasing patch size. The observation is valid for both RGB as well as RGBH data (see table 5.3). However, agreement between observations was poorer at larger patch size. Similar trend was observed by [21]. While larger patch size allowed to learn from larger context, higher variance comes from reduced number of images relative to number of parameters to be learnt. Increasing range of 95% confidence interval in figure 5.5 can be seen graphically in figure 5.5.

Table 5.3: Models Performances Summary relative to patch size. Larger patch size showed higher accurarcy with low consistency.

Patch size	**RGB**		**RGBH**	
	mean IoU	Standard Deviation	mean IoU	Standard Deviation
64	0.653	0.010	0.711	0.007
96	0.662	0.010	0.722	0.008
128	0.694	0.014	0.750	0.008
192	0.705	0.025	0.794	0.016

Ensuring generalizability of the model is essential. It was assessed qualitatively that the performance was observed better on larger patch size on left out tiles as well. Batch size of 128 is finally chosen considering (i) Image context, as it provides larger mid-range context, (ii) Number of samples available, it is able to generate sufficiently enough samples. Further experimentation and analysis is based on patch size of 128 pixels.

5.3.2 Test for Loss Function

As found in the literature, categorical cross-entropy is used commonly. But for the cases of semantic segmentation to address extreme class imbalance, we compared the performance of these loss functions(Dice Loss, Focal Loss and Categorical Crossentropy) over all selected combinations of patch size, backbone, architecture and channels. Both of these loss functions yield better

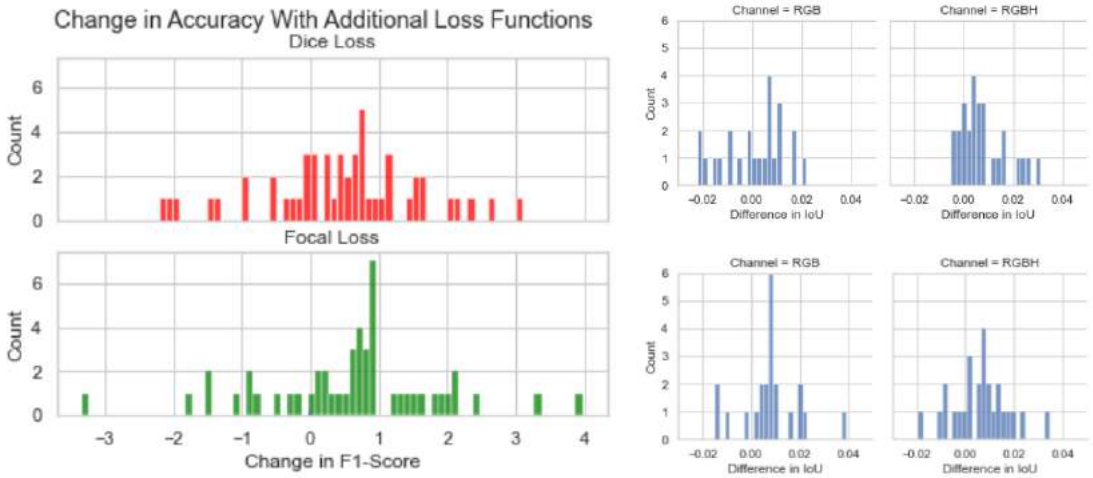


Figure 5.6: (Left)Changes in F1-score with Dice Loss(top) and focal loss(bottom) compared to Categorical Cross-entropy.Positive indicates the chosen loss performed better, and negative means categorical cross-entropy provided better result.(Right) Channel wise difference by corresponding loss functions

result in most of the cases, and majority of the differences were less than 1 % .While differences in RGB channel only was more or less symmetric, that on RGBH channel displayed right-skewed distribution with peak around 0.5%.

From figure 5.4 and table 5.6 we can infer that both of these functions have improvement in both cases, with more contribution while using RGBH channel. Due to slightly better performance on RGBH channel, focal loss is adopted for further experimentation.

Table 5.4: Summary of difference in IoU by Dice Loss and Focal Loss with categorical cross-entropy as reference

Loss Function	Average Increment	Standard Deviation of Increment
Dice Loss	0.44	1.133
Focal Loss	0.61	1.27

5.3.3 Test for Backbone and Segmentation Architectures

As we had two backbones and four segmentation models, eight different type of combinations are available. In the preliminary stage, each of these combination is compared in terms of IoU and F1-score by each of them under different patch size, from the experiment mentioned in section 5.3.1 Findings from table 5.5

Table 5.5: IoU for combination of backbone architecture and segmentation model(patch size=64. The best-perfroming loss function for each configuration has been highlighted)
CCE= Categorical Cross Entropy, Dice Loss = CCE + Dice Loss, Focal Loss = CCE + Focal Loss

Model	**RGB**			**RGBH**		
	CCE	Dice Loss	Focal Loss	CCE	Dice Loss	Focal Loss
Resnet18+FPN	0.652	0.657	**0.660**	0.706	0.706	**0.713**
Resnet18+LINKNET	0.646	0.627	**0.662**	0.713	0.720	**0.723**
Resnet18+UNET	0.642	**0.659**	0.652	0.703	0.715	**0.721**
VGG16+FPN	0.645	0.648	**0.648**	0.699	**0.702**	0.700
VGG16+LINKNET	0.662	0.648	**0.668**	0.718	0.718	0.710
VGG16+UNET	0.653	0.661	**0.662**	0.713	0.715	**0.715**

provide favor for the use of focal loss in futher experimentation. Remaining experiment is conducted with Resnet18 as backbone and Unet as segmentation model.

5.3.4 Hyperparameter Tuning

Hyperparameter tuning was limited to top two selected models due to two reasons: (i) the primary goal of research is to see the difference due to incorporating height, and (ii) There are large number of combinations possible, and each of these combination took hours to train model, running out of time.

Table 5.6 lists the hyperparameters selected from hyperparameter tuning. After all trials, following combination was chosen for further work.

Table 5.6: Parameters and Hyperparameters chosen for final model

SN	Parameter/hyperparameter	Selected Value
1	Encoder Architecture	Resnet18
2	Segmentation Architecture	Unet
3	Loss Function	Categorical cross-entropy+focal loss
4	Activation Function on Dense Layer	Softmax
5	Optimizer	Adam
6	Learning Rate	0.0005

5.4 Test for Effect of Height

Second aim of calculating the optimum height for delineating informal settlement from others was performed on the selected model only. This was further verified with experimentation on a custom FCN-8s as in figure 5.3. The model gets input of 3 or more number of bands from source image, which is fed to single branch for single branch model and two parallel branches in dual-branch model. Channels and their description has been listed in table 4.5.

Combinations of channels as shown in 5.7 are input to the model to see the impact of absolute height, height threshold and vegetation index from visible band. Impact of height and indices were tested using combination listed in table 5.7. Performance analysis were based on metrics and process described in section 5.7.

Table 5.7: Channel names used in analysis

Dataset	**Channels**
RGB	RGB
RGBHAbs	RGB, Absolute Height
RGBH050	RGB, Height threshold by 5 m
RGBH075	RGB, Height threshold by 7.5 m
RGBH100	RGB, Height threshold by 10 m
RGBH125	RGB, Height threshold by 12.5m
RGBH150	RGB, Height threshold by 15 m
RGBH200	RGB, Height threshold by 20 m
RGBH300	RGB, Height threshold by 30 m
RGBVDVI	RGB, VDVI

5.5 Test for Height Integration Approach

Literature from [59] and [18] suggest that the quantitative contribution of height is subjective to the method of integration. Both of these noticed better performance on training height along separate branches followed by concatenation of result in multiclass classification, compared to concatenation of height as additional channel in image.

In our case,this is tested using Fully convolutional network (FCN8s) structure as shown in figure 5.3. One to three identical Resnet18 branches are trained in parallel, with their outputs added, convolved and resampled as in FCN8s model.

Single branch model takes only one input with RGB and other additional bands. Dual branch model takes two inputs: first branch training on RGB input, and second branch taking height (and VDVI) as input. Similarly, triple branch model has third channel that trains on VDVI separately.

In dual branch model, features maps from same level of each of these branches are added before upsampling. In all of these models, outputs after third, fourth and fifth pooling are upsampled to original image size using cubic convolution,and are concatenated. Final outputs come as result of additional 1*1 convolution with softmax activation at end.

5.6 Model Generalisability Assessment

We performed 5-fold cross-validation to ensure the model is generalised enough. Five fold was chosen in contrary to usually used ten fold, due to limited data volume. On using higher number of folds, it was likely to reduce the volume of test data, reducing reliability of metrics.Table 5.8 shows the outputs from five-fold cross validation. Four out of five folds had pretty consistent result.

5.7 Evalution Metrics

For each of the model, overall and per-class metrics are computed. Execution of model writes four text files, each of these containing information on:

- Overall statistics of model: Overall Precision, Overall, Recall, F1-score and mean IoU.
- Confusion matrix for test data

Table 5.8: Results from five-fold cross validation of custom dual-branch FCN8s. Four out of five models performed significantly consistent, while one saw slight deviation numerically. On visual inspection, it was consistent with others despite slight variation in numerical value.

Fold	Informal Settlement			Others		
	Precision (%)	Recall (%)	IoU (%)	Precision (%)	Recall (%)	IoU (%)
1	91.3	91.3	**84.0**	99.1	99.1	98.2
2	88.3	82.9	**74.7**	98.1	98.8	96.9
3	91.6	93.8	**86.4**	99.3	99.1	98.4
4	93.1	93.3	**87.3**	99.3	99.3	98.6
5	86.7	95.1	**83.0**	99.4	98.3	97.7

- Confusion matrix for left-out tiles
- spreadsheet with elements of confusion matrix, computed per image for left-out tiles. This is later used to assess generalisability of model

Corresponding heatmap for confusion matrices are also generated.Precision and recallcan also be used,but these cannot take into account false negative(type II error) and false positive(type I error) respectively and both need to be considered. So, using confusion matrix, we calculated class-wise IoU as optimum accuracy metric. Per class IoU, and mean IoU is further supported with precision and recall whenever necessary.Mathematical expression of these metrics has been included in section 3.5.

Chapter 6

Results and Discussion

This chapter discusses the observations found from the experimentation on segmentation and impact of including height and vegetation index in semantic segmentation of informal settlements. We start with general findings on 6.1. It is followed by comparitive study on absolute and relatie height, in section 6.2. We then proceed to the difference observed in results on use of absolute and relative height in section 6.3, where we also analyse the optimum height threshold in overall and scene specific context. Finally, we end this chapter with analysis of the effect of vegetation index in 6.5.

6.1 General observations

In overall, larger patch size led to better segmentation, by offering extraction of contextual information from larger area. Including height always enhanced performance of segmentation, in any scene or patch size. Both absolute and threshold height yield equivalent result from qualitative and quantitative aspect.

Furthermore, using vegetation indices might not always be beneficial, as it helps to differentiate vegetation from buildings, while the mis-classification in informal settlement extraction is between formal and informal buildings. It was found to increase the accuracy of semantic segmentation in some specific situations only. But when height has already been included, it cannot make further contribution. Table 6.1 summarizes general findings frome experiment.

Table 6.1: class-wise IoU for various schemes(Test Data)

Dataset	IoU Informal (%)	IoU Others (%)
RGB	77.92	97.47
RGBHAbs	83.71	98.10
RGBH050	84.67	98.27
RGBH075	84.34	98.19
RGBH100	83.89	98.17
RGBH125	85.51	98.33
RGBH150	85.62	98.39
RGBH200	85.99	98.44
RGBH300	84.86	98.26
RGBVDVI	73.75	96.85

6.2 Contribution of Absolute Height

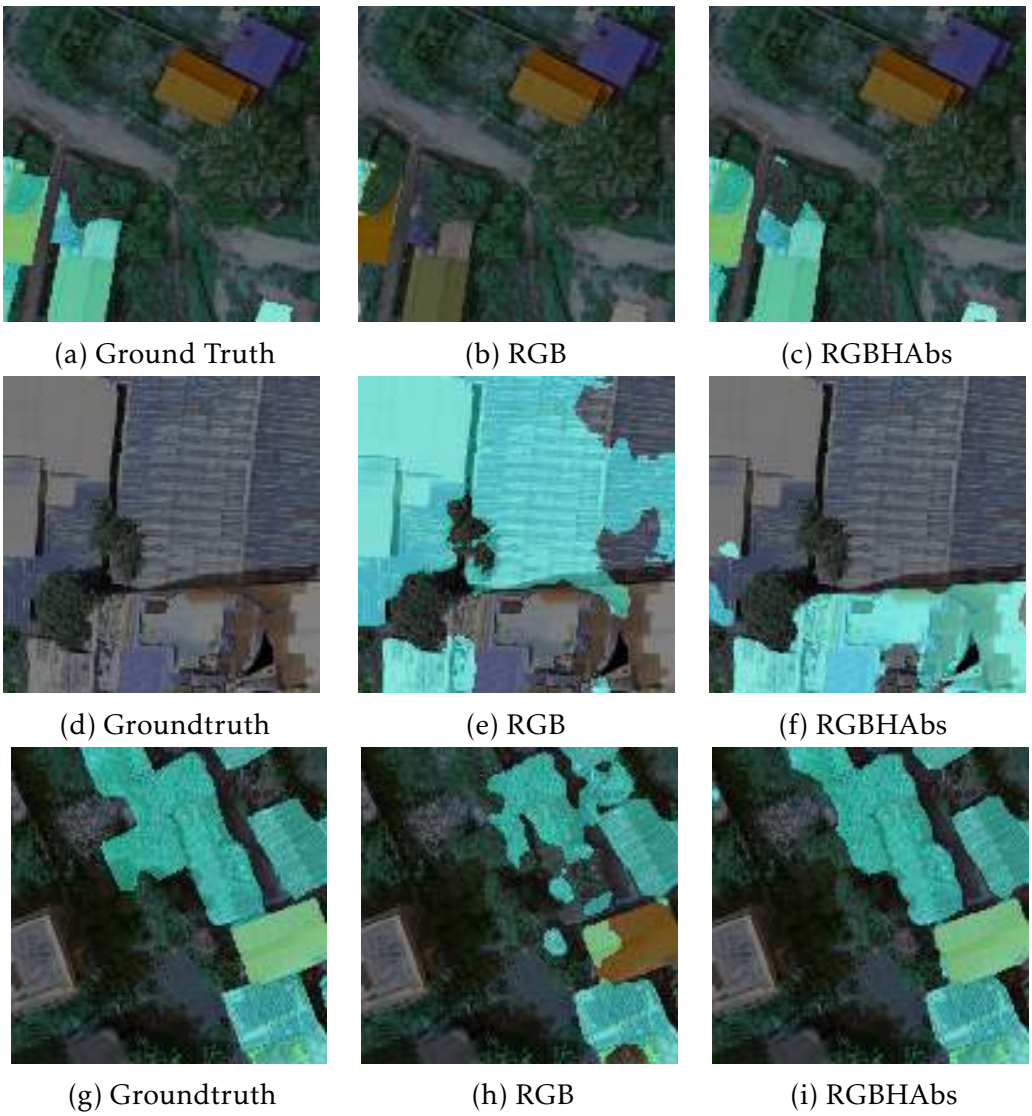
Having information regarding feature height in any form, absolute or relative has been observed to increase segmentation performance, as in table 6.1. In terms of absolute height, around 6 % increase in IoU for informal settlement was obtained by introducing absolute height.

The nature of error with RGB data is mixed type. Though in most of the cases, informal settlements are over-predicted, cases of incomplete segments and voids within segment were also observed. Thorough inspection of test images and predicted outputs showed that, using RGB had underprediction on informal settlement with small buildings with straw roofing, while they overpredicted in dense area with zinc-roofed buildings (figure 6.1).

But the rectification is not complete even with addition of height. While majority of underprediction on the straw-roofed settlements were filled, over-predictions on the top of formal buildings were persistent in many cases (figure 6.1e and 6.1f). This can be attributed to spectral similarity between roofing material, and DTM error caused due to point cloud filtering.

Aforementioned observations are confirmed visually based on their prediction on the left out tile the model has not seen. Figure 6.2 shows predicted map from RGB and RGBHAbs from two sites. Greater difference can be observed between 6.2b and 6.2c, on a settlement with small scattered huts. While 6.2b contained large number of small patches but omitted some whole buildings, these have been fixed in figure 6.2c. But, in the mixed settlement with zinc-roofed building, change is not very much significant.

Statistically, table 6.2 shows a significant rise in recall value from RGB to



(a) Ground Truth (b) RGB (c) RGBHAbs

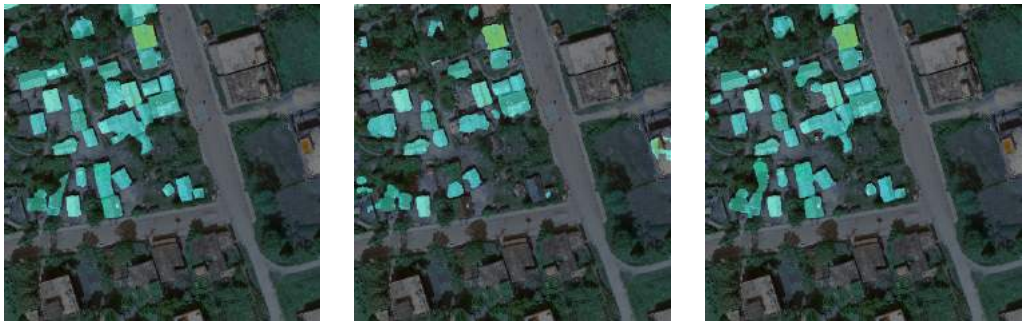
(d) Groundtruth (e) RGB (f) RGBHAbs

(g) Groundtruth (h) RGB (i) RGBHAbs

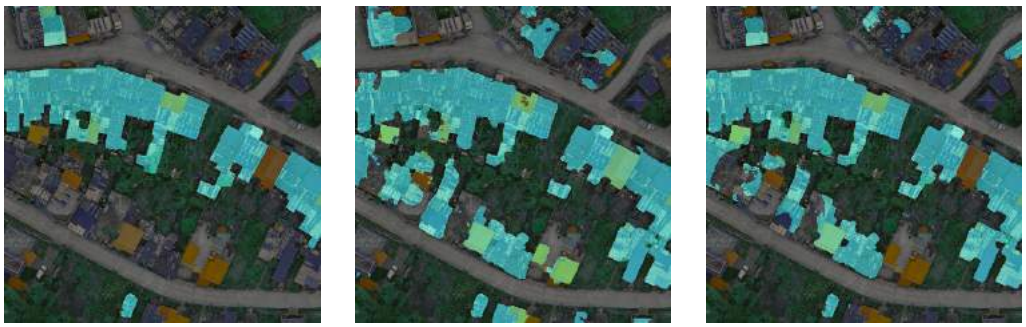
Figure 6.1: Comparison of RGB vs RGBHAbs data on test data: RGBHAbs rectified underprediction on low buildings, while overpredictions on zinc-roofed buildings are preserved

Table 6.2: Comparision of class-wise Precision, Recall and IoU for RGB and RGBHAbs. Both test and out data saw improvement in IoU of informal settlement by at least 6 % from RGB to RGBHAbs.

Dataset	**Informal**			**Others**		
	Precision (%)	Recall (%)	IoU (%)	Precision (%)	Recall (%)	IoU (%)
RGB Test	89.423	85.830	77.919	98.510	98.928	97.470
RGBHAbs Test	88.993	93.383	83.714	99.298	98.780	98.095
RGB Out	69.719	76.643	57.502	96.760	95.446	92.490
RGBHAbs Out	76.189	93.152	72.148	99.034	96.017	95.126



(a) Groundtruth (b) RGB (c) RGBHAbs



(d) Groundtruth (e) RGB (f) RGBHAbs

Figure 6.2: Prediction map from model trained on RGB vs RGBH data on external tile. (c),Omitted small buildings has been included by RGBHAbs, (f) overprediction of informal settlements corrected by RGBHAbs in dense area reduced by RGBHAbs

RGBHAbs for both test as well as out data. This is an indicator of reduced false negatives,leading to reduced under-prediction on using RGBHAbs.

The concluding point on impact of absolute height is significant improvement on segmentation result, that counts to 6 % in our case. The trend is however dependent on site.

6.3 Absolute Versus Relative Height

At a first glance, table 6.1 shows no significant difference on using absolute height or threshold height. If very small difference are not considered, a constant improvement of around 6 % was observed for each of the threshold height. As this is the overall summary of the model, and dataset includes heterogeneous scenes, scene specific difference needs to be further analysed, that has been covered in section 6.4.1.

From this results, key conclusion to draw is that setting a threshold and normalizing data can definitely improve over RGB data , but the results are almost identical numerically for all of these cases. Table 6.3 shows that both

Table 6.3: Accuracy metrics for Informal Settlement for variable threshold height on test and out data

Dataset	Test Data			Out Tile		
	Precision (%)	Recall (%)	IoU (%)	Precision (%)	Recall (%)	IoU (%)
RGB	89.4	85.8	77.9	69.7	76.6	57.5
RGBHAbs	89.0	93.4	83.7	76.2	93.2	72.1
RGBH050	92.1	91.3	84.7	85.0	90.2	77.8
RGBH075	90.1	93.0	84.3	81.4	92.7	76.4
RGBH100	91.2	91.3	83.9	84.9	91.3	78.5
RGBH125	90.7	93.7	85.5	80.3	92.8	75.6
RGBH150	93.1	91.5	85.6	85.4	88.6	76.9
RGBH200	93.4	91.5	86.0	84.5	89.7	77.0
RGBH300	90.5	93.2	84.9	82.3	91.1	76.1

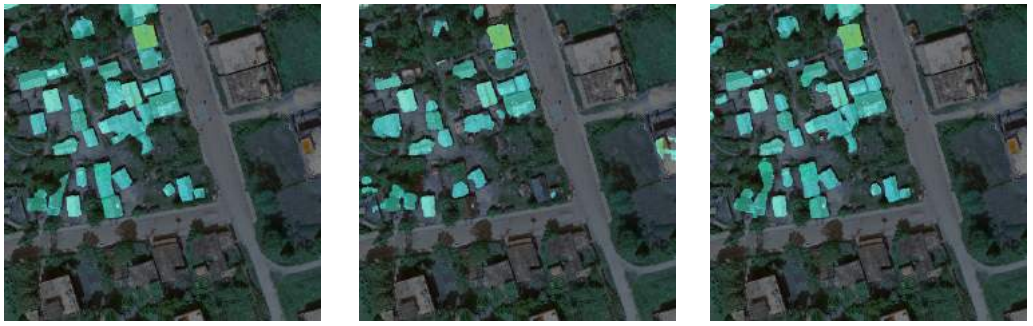
test and out data showed similar trend. Larger variation and inconsistency in out data is however due to smaller volume of external data. As class imbalance exists in our data,a small variation in majority class influences minority class more, that introduced larger difference.

Visual inspection also shows no remarkable difference on output images with different threshold value, and absolute feature height as well.

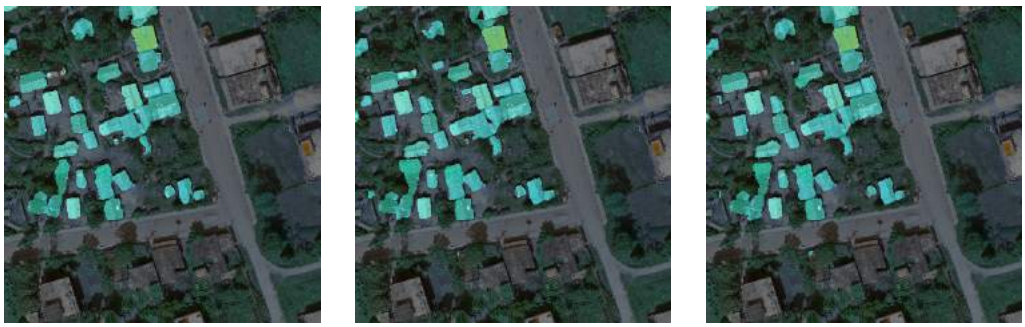
6.4 Optimum Threshold Height

Table 6.3 suggests uniform trend on precision and recall for all of the height types.Comparable value of precision and recall for positive class also confirms similar trend of mis-prediction for all of the chosen threshold height as well as absolute height. As it is the observations from larger data over heterogeneous area, it would be meaningful to see trend with nature of settlement.This has been covered in section 6.4.1.

On visual inspection, all of the combinations are observed to suffer mis-predictions, but with different trend. Predicted output map from Saalghari (figure 6.3), an area characterized by small isolated buildings shows large number of small segments and void within segment with RGB channel. Missing small buildings as a whole were also observed on prediction using RGB channel.These are improved with height,with smaller noisy segments omitted, missing informal buildings partly covered. With use of change of threshold height and increasing them,some minor noises were observed to re-appear.



(a) Ground truth (b) RGB (c) RGBHAbs



(d) RGBH050 (e) RGBH150 (f) RGBH300

Figure 6.3: Ground truth and predicted outputs on low-height scattered informal settlement using different channels. RGB suffers under-prediction of small buildings and large number of noisy segments, while others predict it well with similar visual outputs

.

6.4.1 Accuracy as Function of Settlement Nature

With assumption that the optimally performing threshold might be subjected to texture and relative difference in height betweeen formal and informal settlements, the study area has been split into three categories as in table 6.4 and trend analysed for each type.

Table 6.4: Informal Settlement Type Classification

Type	Characteristics	Out Tiles
Type A	Separated, Dense, High height difference between formal and informal	Balkhu, Thapathali, Simpani
Type B	Mixed formal and informal settlement, Comparable Heights, Zinc roof	Simpani
Type C	Separate formal and informal, sparsely constructed, straw roof	Saalghari, Kataan

Table 6.5 shows that, irrespective of threshold or number of channels, informal settlements in the city, separated from concrete formal buildings were

Table 6.5: Accuracy of Informal Settlement of types from table 6.4. Type A settlement are segmented with better accuracy. Type C got better advantage from height information, while those of type B are more challenging to segment

Dataset	Type A			Type B			Type C		
	Precision	Recall	IoU	Precision	Recall	IoU	Precision	Recall	IoU
RGB	82.2	87.7	73.7	56.0	83.8	50.5	85.7	44.9	41.8
RGBHAbs	87.0	95.5	83.6	69.1	95.5	66.9	91.5	85.5	79.1
RGBH050	92.0	92.4	85.6	77.7	93.6	73.8	92.3	81.5	76.3
RGBH075	85.9	95.2	82.3	74.9	94.9	72.0	91.3	84.8	78.4
RGBH100	91.4	94.6	86.9	79.7	93.6	75.5	93.1	81.7	77.1
RGBH125	89.8	96.2	86.7	76.2	94.0	72.7	89.2	84.1	76.4
RGBH150	88.8	89.6	80.5	87.0	92.2	81.1	91.8	81.9	76.4
RGBH200	92.9	91.8	85.8	85.1	92.9	80.0	90.3	81.4	74.9
RGBH300	90.1	92.3	83.8	80.0	95.2	76.9	89.7	83.6	76.2

distinguished with highest accuracy, whereas mixed results were obtained between mixed(type B) and sparse settlement. The building with mixed settlement is found to be most challenging, probably due to presence of buildings with similar nature mixed to each other.

Interestingly, settlement with straw-roofed building (Type C) saw great improvement in IoU with height information. One instance of this type of settlement has been shown visually in figure 6.3 and had better result at low threshold.

While informal settlements of Type A are better segmented around mid-range threshold of around 10 meters, those in mixed settlement were better extracted at higher threshold of 15 meters. In contrast, scattered settlement with straw roofs are observed to be better segmented using relatively lower threshold height. The results form the base for interpretation that segmentation results are function of chosen threshold, configuration and nature of settlement.

6.5 Effect of Vegetation Index

It was observed by [15] that using visual band difference vegetation index (VDVI) calculated as equation 4.3 enhanced segmentation of building in vegetated area. The test in our case yields no or limited advantage of using VDVI in segmentation of informal settlement (table 6.6).

On visual assessment of predicted map with and without using VDVI, majority of the misclassifation was found not between building to vegetation. Rather

it was among informal and formal buildings. As these two have similar VDVI, use of VDVI would not contribute to it.

Table 6.6: Comparison of IoU for informal settlement with and without VDVI. No systematic or significant change is observed with use of VDVI

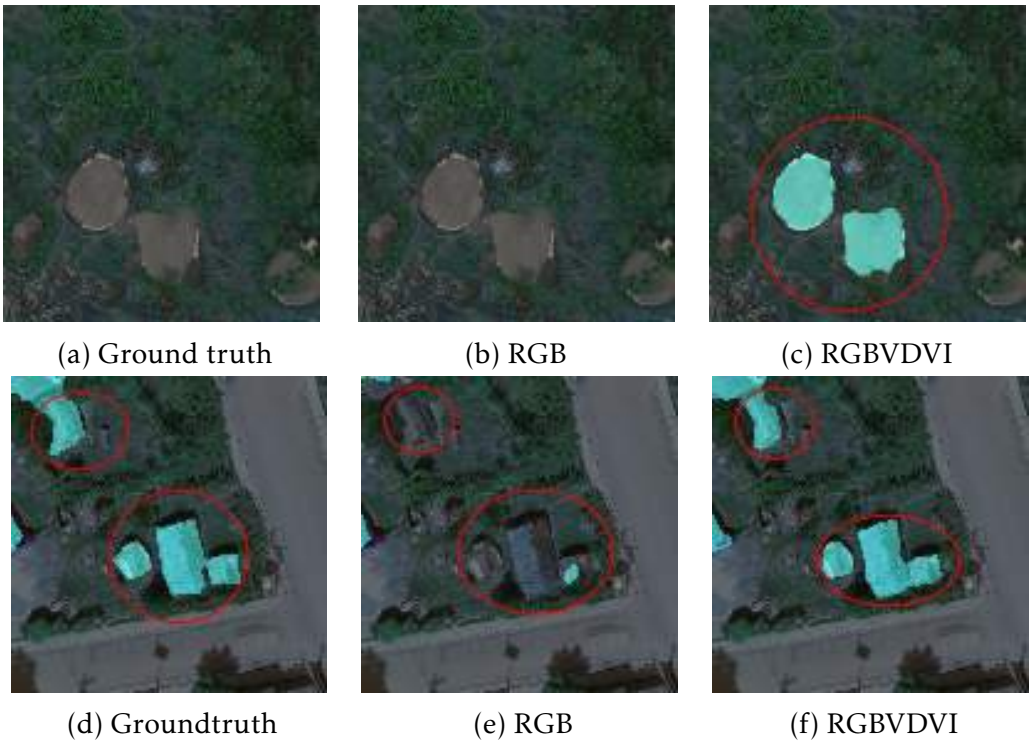
Channel	IoU Informal Settlement	
	Without VDVI	with VDVI
RGB	77.919	73.748
RGBHAbs	83.714	83.357
RGBH050	84.673	85.461
RGBH075	84.343	84.500
RGBH100	83.886	84.994
RGBH125	85.513	85.744
RGBH150	85.623	85.446
RGBH200	85.994	85.330
RGBH300	84.862	85.135

But in some specific cases, especially in scenario of Type C(see table 6.4) settlement, slight improvement was observed with VDVI. The main reason might be that these settlements have vegetation partly covering the building roof. Using VDVI at those places must have helped correct reclassification on those part. Instead of improving the segmentation, cases were found where sand piles and boulders on river were classified as informal settlement while using VDVI(figure 6.4). This leads us to the conclusion that using VDVI in segmentation of informal settlement are always not a good choice.

6.6 Best Height Integration Approach

In contrary to the observations made by previous authors, we observed that integration of height as additional band outperformed training along multiple branches followed by fusion(figure 6.5) . We tested this using custom FCN8s with Resnet18 as encoder. Around 5% better result was obtained with single branch model compared to its counterpart on dual channel model (table 6.7).

However accuracy did not vary within a single model for absolute or threshold heights(see table 6.7), in agreement with findings from section 6.4. The outputs from dual branch model, in fact contained more mis-prediction than RGB image trained on a single branch.



(a) Ground truth (b) RGB (c) RGBVDVI

(d) Groundtruth (e) RGB (f) RGBVDVI

Figure 6.4: Cases of mis-prediction and prediction with VDVI. (top row)a pile of sand mis-classified as informal settlement, (bottom row) missing building with straw roof completed with VDVI. VDVI suffered more erroneous over-predictions on large sand piles, rooftops, but rectified missing small buildings in vegetated area

.

Table 6.7: Accuracy Metrics of Informal Settlement for Single Branch and Dual Branch FC8s Model. Single Branch Model performed better in all band combinations.

Dataset	Single Branch Model			Dual Branch Model		
	Precision (%)	Recall (%)	IoU (%)	Precision (%)	Recall (%)	IoU (%)
RGB	84.3	86.2	74.3	-	-	-
RGBHAbs	90.2	88.5	80.7	82.7	91.0	76.4
RGBH050	89.4	92.3	83.2	88.1	85.8	76.9
RGBH075	90.2	90.5	82.4	83.7	91.3	77.5
RGBH100	90.0	91.6	83.1	84.3	91.1	77.9
RGBH125	91.4	89.3	82.4	85.8	88.3	77.0
RGBH150	88.3	92.8	82.6	84.6	90.5	77.7
RGBH200	87.3	93.2	82.1	83.6	89.5	76.1
RGBH300	85.8	92.9	79.0	87.0	88.0	77.8
RGBVDVI	80.5	87.0	71.9	81.7	84.2	70.8

Activation maps of dual-branch model further illustrated the probable cause for mis-prediction. Activation for a particular location in input image is different for two branches, which are then added at the end. Addition of activation maps from two branches led to higher value at regions moderately activated in both branches, which necessarily would not be informal settlement. Unnecessary patches especially on top of moderate height zinc-roofed building were thus generated, reducing the accuracy of model.

Consequently, the prediction on the external tiles with dual branch saw over predictions in high magnitude in tiles containing zinc-roofed moderate height formal settlement. In the locations with different materials or larger difference in height, predictions are quite similar with single branch as well as dual branch model.

Thus, concatenating height as additional band and training along a single branch had advantage of multiple bands input,and learning from each of these bands simultaneously. This also reduced the need for merging of outputs from two branches, eliminating risk of higher cumulative sum on moderately activated zones.

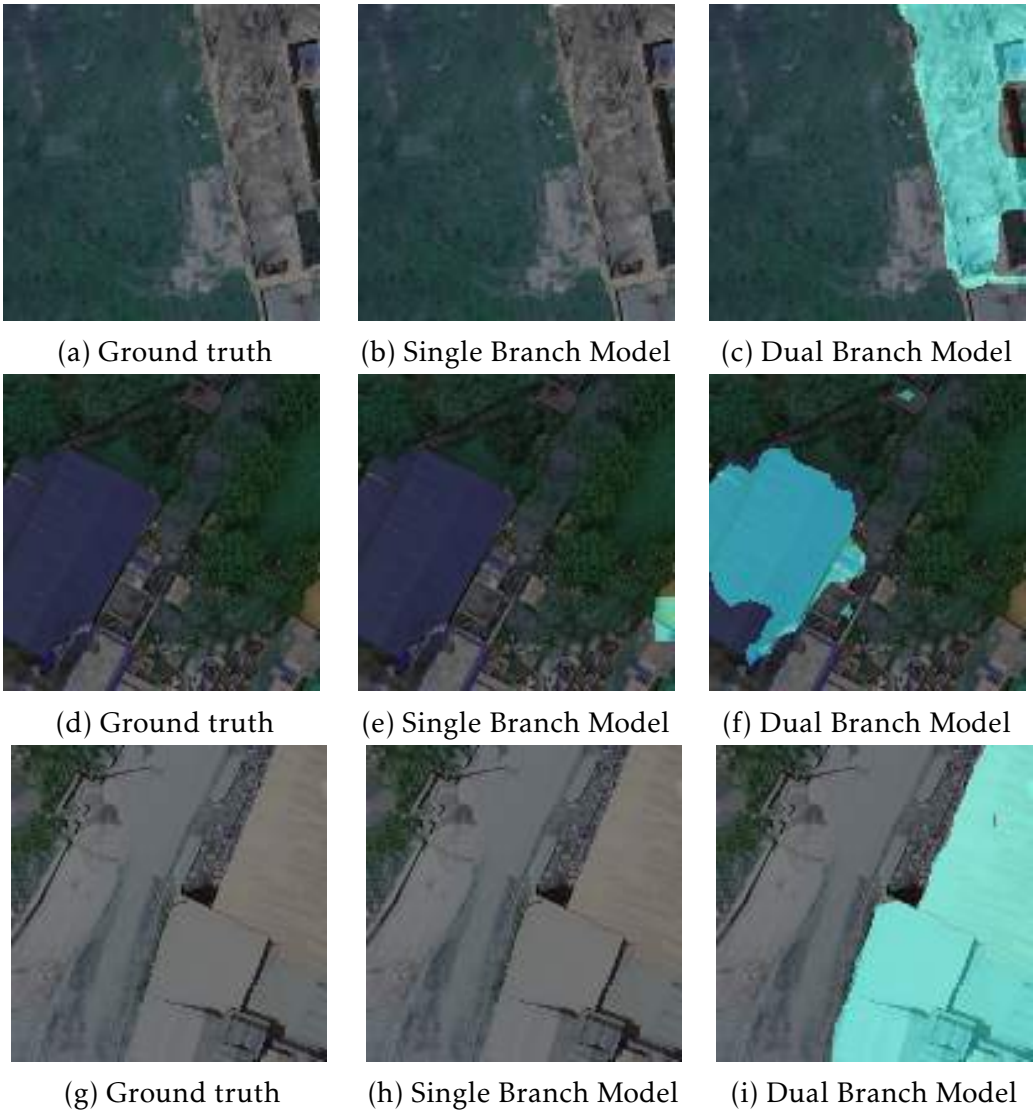
6.7 Comparison With Other Works

A number of works were found from literature review which have worked in multi-class classification including building, vegetation, etc. However, no binary classification for informal settlement, and that using nDSM and threshold concept was found to be worked in the past. So, we were unable to make comparison with previous works.

Despite this limitation,our model performance has been cross-verified by 5-fold cross-validation, multiple models on same data and prediction on external tiles. Through consistent trend and results being in agreement to each other,this provides us a confidence that our model is generalized and transferable to other developing countries as well. Also because the major contribution of our work is on use of additional UAV outputs, it has been verified experimentally.

6.8 Limitations

This work has been conducted to perform segmentation of informal settlement in complex scenario of Nepal using existing architectures. We have been able



(a) Ground truth (b) Single Branch Model (c) Dual Branch Model

(d) Ground truth (e) Single Branch Model (f) Dual Branch Model

(g) Ground truth (h) Single Branch Model (i) Dual Branch Model

Figure 6.5: Predictions on test data(RGBHAbs) by single and dual branch FCN. Dual branch model suffered overprediction and contained noisy patches, mainly on top of modereate height zinc-roofed buildings

.

to answer our research question on application of deep learning for complex scenes,and have tested a proposed approach of setting threshold height as additional channel in informal settlement segmentation.

However, due to wide range of heterogeneity among informal settlements,our model, as any other models, is not applicable to all types of informal settlement. It has not be tested and is more likely to be unable to segment particular type of informal settlements, which just lack tenure right, but are not different from formal settlement in any other aspect.

The reason behind this is, those settlements, like formal settlements have

big multi-storey buildings, with proper space and context perfectly matching formal settlements. Only the way they are informal is they are not registered. So, the dataset, nature and model we are using wont be able to segment those.

Additionally, we have used nDSM instead of DSM, that needs point cloud cloud filtering. One possible challenge, and also experienced fact in this approach, is lack of terrain point clouds in densely constructed area, which sometimes tend to mis-interpolate terrain.Fortunately, it has not been experienced in our case, due to open space within settlement.

Chapter 7

Conclusion

This chapter summerizes the findings and conclusion from our study.

7.1 Conclusion

Working on segmentation of informal settlement, and using different combination of dataset, we come to an conclusion that, standard segmentation models trained end-to-end can be applied to segment out informal settlement from complex and unique scenarios like Nepal,and can provide promising results.

Unet on top of Resnet,including height channel and trained end to end is found to be the best among tested model for extraction of informal settlement from our experiment. Despite smooth learning curve, transfer learning with pre-trained weights suffers random noises and produces numerous tiny patches. On this basis,transfer learning has not been found to offer any advantage in this specific case of informal settlement extraction in our study area.

In terms of standard encoders, network with Resnet as the backbone trained faster with less noisy output compared to their counterpart using VGG as backbone. On the choice of decoders,Unet outperformed FPN, PSPNet and Linknet by a narrow margin, despite no clearly distinguishable difference in statistical results.

IoU of 77.9% for informal settlement using RGB channel, in the complex scene we are applying is a remarkable result,despite dominance of negative class in scene. In this context,advantages of very high resolution UAV orthophoto has been proven, enabling detection of features finely. Nevertheless,

noisy outputs and misclassifications of mainly the zinc-roofed formal settlements as informal ones while using RGB channels provides favor on use of additional outputs from UAV.

We obtained a 6% gain on IoU of building using height information. Using nDSM , may it be absolute height above ground or its value normalized using a threshold showed improved accuracy in more or less same magnitude. Major improvements are observed in rectification of misclassified rooftop pixels.

Training RGB image and other information(height, VDVI) along separate branches followed by merging of outputs did not enhance segmentation of informal settlement. Rather, it tends to overestimate informal settlements due to different trend of activation in separate branches. So, concatenating height as additional band is recommended for noise-free and precise extraction of informal settlement. This also reduces the processing time and resource due to less number of parameters. So, point cloud or its equivalent on vertical dimension can be a valuable asset available without additional need of resources.

Vegetation indices derived from visible band does not boost the segmentation accuracy, with exception to sites with straw-roofed small buildings in vegetated area. This is firstly because misclassification is among buildings and not vegetation, where VDVI cannot contribute. Secondly, the model is already using its version of NDVI in some way already, leaving no or less space for improvement with use of VDVI. In majority of the scenarios, vegetation index reduced the segmentation accuracy instead of increasing it.

Our initial assumption of further enhancement of IoU with threshold height was not proven to be valid, as no difference were observed in result with absolute and normalized height, and also with varying threshold height.

Segmentation accuracy, and optimally performing threshold value are however subjected to settlement type. While isolated densely constructed informal settlement are segmented with better accuracy, those on mixed settlement or partly covered by vegetation suffer larger prediction error. In terms of threshold height, choice of lower threshold value of 5 to 7.5 meters performed better for isolated small buildings. In contrast, mixed and core urban informal settlement are slightly better segmented at higher threshold around 20 meters.

So, the optimum model and hyperparameters vary from site to site. Thus, slight modifications on model parameters to fit the scene under consideration is always necessary to achieve the best result.

In nutshell, UAV orthophoto and it's products on vertical dimension is

found to contribute on better segmentation of informal settlements in complex scenes from Nepal, and it is recommended to utilize height information as additional band from UAV for better segmentation accuracy.

Our framework is ready to implement by government agencies and development organizations working in informal settlement upgrading, as it has been trained and tested on diverse type of data including samples from most of the prevalent informal settlement types found in Nepal. Considering smaller extent of informal settlements in Nepal, our model is capable to extract informal settlement efficiently in new area. This keeps to potential to substitute ground-based methods, provided it is obstruction-free.

Bibliography

[1] T. Stark. "Using Deep Convolutional Neural Networks for the Identification of Informal Settlements to Improve a Sustainable Development in Urban Environments." In: *Technische Universität München* (2018). URL: https://elib.dlr.de/119019/1/Stark{_}MA.pdf.

[2] A. K. Shrestha, P. Nepali, U. S. Panday, and R. Shrestha. "Livelihoods of Squatter Settlements : Analysis from Tenure Perspective." In: *Fic Working Week 2017*. 8536. Helsinki, Finland, 2017.

[3] Un-Habitat. *State of the World's Cities 2010/11*. 2010. ISBN: 9781849711760. DOI: 10.4324/9781849774864.

[4] K. D. Willis. "Squatter Settlements." In: *International Encyclopedia of Human Geography* (2009), pp. 403–408. ISSN: 0031-0360. DOI: 10.1016/B978-008044910-4.00122-X.

[5] M. Mohanty. "Squatter Settlements and Slums and Sustainable Development." In: *Sustainable Cities and Communities*. Ed. by W. Leal Filho, A. M. Azul, L. Brandli, P. G. Özuyar, and T. Wall. Cham: Springer International Publishing, 2019, pp. 1–12. ISBN: 978-3-319-71061-7. DOI: 10.1007/978-3-319-71061-7_49-1. URL: https://doi.org/10.1007/978-3-319-71061-7_49-1.

[6] UN-HABITAT. *Slums of the World: The Face of Urban Poverty in the New Millennium*. Nairobi, Kenya: UN-HABITAT, 2003, p. 90. ISBN: 92-1-131683-9.

[7] UN-HABITAT. *Addressing the most Vulnerable First: Pro-poor climate action in informal settlements*. Tech. rep. Nairobi: UN-HABITAT, 2018. URL: https://unhabitat.org/sites/default/files/2019/05/pro-poor_climate_action_in_informal_settlements-.pdf.

[8] United Nations. *Transforming our world: the 2030 agenda for sustainable development*. Tech. rep.

[9] B. J. Gram-Hansen, F. Azam, P. Helber, A. Coca-Castro, P. Bilinski, I. Varatharajan, and V. Kopackova. "Mapping informal settlements in developing countries using machine learning and low resolution multi-spectral data." In: *AIES 2019 - Proceedings of the 2019 AAAI/ACM Conference on AI, Ethics, and Society* (2019), pp. 361–368. DOI: 10.1145/3306618.3314253. arXiv: 1901.00861.

[10] R. Sliuzas, G. Mboup, and A. de Sherbinin. "Report of the expert group meeting on slum identification and mapping." In: *Report by CIESIN, UN-Habitat, ITC* (2008), p. 36.

[11] M. Kuffer, K. Pfeffer, and R. Sliuzas. "Slums from Space — 15 Years of Slum Mapping Using Remote Sensing." In: *Remote Sensing* 8.455 (2016). DOI: 10.3390/rs8060455.

[12] C. M. Gevaert, C. Persello, R. Sliuzas, and G. Vosselman. "Informal settlement classification using point-cloud and image-based features from UAV data." In: *ISPRS Journal of Photogrammetry and Remote Sensing* 125 (2017), pp. 225–236. ISSN: 09242716. DOI: 10.1016/j.isprsjprs.2017.01.017. URL: http://dx.doi.org/10.1016/j.isprsjprs.2017.01.017.

[13] Y. H. Tsai, D. Stow, and J. Weeks. "Comparison of object-based image analysis approaches to mapping new buildings in Accra, Ghana using multi-temporal quickbird satellite imagery." In: *Remote Sensing* 3.12 (2011), pp. 2707–2726. ISSN: 20724292. DOI: 10.3390/rs3122707.

[14] M. Kuffer, J. Barros, and R. V. Sliuzas. "The development of a morphological unplanned settlement index using very-high-resolution (VHR) imagery." In: *Computers, Environment and Urban Systems* 48 (2014), pp. 138–152. ISSN: 0198-9715. DOI: 10.1016/j.compenvurbsys.2014.07.012. URL: http://dx.doi.org/10.1016/j.compenvurbsys.2014.07.012.

[15] W. Boonpook, Y. Tan, Y. Ye, P. Torteeka, K. Torsri, and S. Dong. "A deep learning approach on building detection from unmanned aerial vehicle-based images in riverbank monitoring." In: *Sensors (Switzerland)* 18.11 (2018). ISSN: 14248220. DOI: 10.3390/s18113921.

[16] C. Gevaert, R. Sliuzas, C. Persello, and G. Vosselman. "Opportunities for UAV mapping to support unplanned settlement upgrading." In: *Rwanda Journal* 1.1S (2017). ISSN: 2305-2678. DOI: 10.4314/rj.v1i2s.4d.

[17] A. Vetrivel, M. Gerke, N. Kerle, and G. Vosselman. "Segmentation of UAV-based images incorporating 3D point cloud information." In: *International Archives of the Photogrammetry, Remote Sensing and Spatial Information Sciences - ISPRS Archives* 40.3W2 (2015), pp. 261–268. ISSN: 16821750. DOI: 10.5194/isprsarchives-XL-3-W2-261-2015.

[18] W. Sun and R. Wang. "Fully Convolutional Networks for Semantic Segmentation of Very High Resolution Remotely Sensed Images Combined with DSM." In: *IEEE Geoscience and Remote Sensing Letters* 15.3 (2018), pp. 474–478. ISSN: 15580571. DOI: 10.1109/LGRS.2018.2795531.

[19] W. Boonpook, Y. Tan, and B. Xu. "Deep learning-based multi-feature semantic segmentation in building extraction from images of UAV photogrammetry." In: *International Journal of Remote Sensing* 42.1 (2020), pp. 1–19. ISSN: 13665901. DOI: 10.1080/01431161.2020.1788742.

[20] United Nations Development Programme (UNDP). *Human Development Report 2020The next frontier Human development and the Anthropocene Human Development Report 2020*. Tech. rep. 2020, p. 349. URL: http://hdr.undp.org..

[21] N. Mboga, C. Persello, J. R. Bergado, and A. Stein. "Detection of informal settlements from VHR images using convolutional neural networks." In: *Remote Sensing* 9.11 (2017). ISSN: 20724292. DOI: 10.3390/rs9111106.

[22] M. Wurm, T. Stark, X. X. Zhu, M. Weigand, and H. Taubenböck. "Semantic segmentation of slums in satellite images using transfer learning on fully convolutional neural networks." In: *ISPRS Journal of Photogrammetry and Remote Sensing* 150.May 2018 (2019), pp. 59–69. ISSN: 09242716. DOI: 10.1016/j.isprsjprs.2019.02.006. URL: https://doi.org/10.1016/j.isprsjprs.2019.02.006.

[23] P. Joshi, S. Sen, and J. Hobson. "Experiences with surveying and mapping Pune and Sangli slums on a geographical information system (GIS)." In: *Environment and Urbanization* 14.2 (2002), pp. 225–240. ISSN: 09562478. DOI: 10.1177/095624780201400218.

[24] P. Hofmann, J. Strobl, T. Blaschke, and H. Kux. "Detecting informal settlements from QuickBird data in Rio de Janeiro using an object-based approach." In: *Lecture Notes in Geoinformation and Cartography* 0.9783540770572 (2008), pp. 531–553. ISSN: 18632351. DOI: 10.1007/978-3-540-77058-9_29.

[25] P. M. Ward and P. A. Peters. "Self-help housing and informal homesteading in peri-urban America : Settlement identification using digital imagery and GIS $." In: 31 (2007), pp. 205–218. DOI: 10.1016/j.habitatint.2007.02.001.

[26] A. Gunter. "Getting it for free: Using Google earth™ and IL WIS to map squatter settlements in Johannesburg." In: *2009 IEEE International Geoscience and Remote Sensing Symposium* 3 (2009), pp. III–388–III–391.

[27] N. Sahriman, M. Z. Z. Abiden, A. R. A. Rasam, A. M. Samad, and N. M. Tarmizi. "Urban poverty area identification using high resolution satellite imagery: A preliminary correlation study." In: *Proceedings - 2013 IEEE International Conference on Control System, Computing and Engineering, ICCSCE 2013* (2013), pp. 430–434. DOI: 10.1109/ICCSCE.2013.6720003.

[28] J. Pratomo, M. Kuffer, D. Kohli, and J. Martinez. "Application of the trajectory error matrix for assessing the temporal transferability of OBIA for slum detection Application of the trajectory error matrix for assessing the temporal transferability of OBIA for slum detection." In: *European Journal of Remote Sensing* 51.1 (2018), pp. 838–849. DOI: 10.1080/22797254.2018.1496798. URL: https://doi.org/10.1080/22797254.2018.1496798.

[29] N. H. Praptono, P. Sirait, M. I. Fanany, and A. M. Arymurthy. "An automatic detection method for high density slums based on regularity pattern of housing using Gabor filter and GINI index." In: *2013 International Conference on Advanced Computer Science and Information Systems (ICACSIS)* (2013), pp. 347–351.

[30] S. Shekhar. "Detecting Slums From Quick Bird Data in Pune Using an Object Oriented Approach." In: *ISPRS - International Archives of the Photogrammetry, Remote Sensing and Spatial Information Sciences* XXXIX-B8.June (2012), pp. 519–524. ISSN: 1682-1750. DOI: 10.5194/isprsarchives-xxxix-b8-519-2012.

[31] T. Williams, T. Wei, and X. Zhu. "Mapping Urban Slum Settlements Using Very High-Resolution Imagery and Land Boundary Data." In: *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing* PP (Dec. 2019), pp. 1–12. DOI: 10.1109/JSTARS.2019.2954407.

[32] O. H. Y. Lam, M. Dogotari, M. Prüm, H. N. Vithlani, C. Roers, B. Melville, F. Zimmer, and R. Becker. "An open source workflow for weed mapping in native grassland using unmanned aerial vehicle: using Rumex obtusifolius as a case study." In: *European Journal of Remote Sensing* 00.00 (2020), pp. 1–18. ISSN: 22797254. DOI: 10.1080/22797254.2020.1793687. URL: https://doi.org/10.1080/22797254.2020.1793687.

[33] V. Deparday, C. Gevaert, G. Molinario, R. Soden, and S. Balog-Way. *Machine Learning for Disaster Risk Management*. English. World Bank, Jan. 2019.

[34] Z. Alom, T. M. Taha, C. Yakopcic, S. Westberg, P. Sidike, M. S. Nasrin, M. Hasan, B. C. V. Essen, A. A. S. Awwal, and V. K. Asari. "A State-of-the-Art Survey on Deep Learning Theory and Architectures." In: *Electronics* 8.292 (2019), pp. 1–67. DOI: 10.3390/electronics8030292.

[35] S. Shrestha. "IMPROVED FULLY CONVOLUTIONAL NETWORK WITH CONDITIONAL RANDOM FIELD FOR BUILDING EXTRACTION." Doctoral dissertation. Universidade Nova de Lisboa, 2017.

[36] *7 Types of Activation Functions in Neural Networks: How to Choose?* URL: https://missinglink.ai/guides/neural-network-concepts/7-types-neural-network-activation-functions-right/ (visited on 02/21/2021).

[37] H. Small and Brown. "Handling Unbalanced Data in Deep Image Segmentation." In: 2017.

[38] *Regularization in Machine Learning | by Prashant Gupta | Towards Data Science*. URL: https://towardsdatascience.com/regularization-in-machine-learning-76441ddcf99a (visited on 02/21/2021).

[39] L. Thapa. "Ship recognition on the sea surface using aerial images taken by Uav: a deep learning approach." Doctoral dissertation. Universidade Nova de Lisboa, 2018, pp. 1–83. URL: http://hdl.handle.net/10362/63805.

[40] S. P. Adhikari, G. Kim, and H. Kim. "Deep neural network-based system for autonomous navigation in paddy field." In: *IEEE Access* 8 (2020), pp. 71272–71278. ISSN: 21693536. DOI: 10.1109/ACCESS.2020.2987642.

[41] K. Simonyan and A. Zisserman. "Very deep convolutional networks for large-scale image recognition." In: *3rd International Conference on Learning Representations, ICLR 2015 - Conference Track Proceedings* (2015), pp. 1–14. arXiv: 1409.1556.

[42] K. He, X. Zhang, S. Ren, and J. Sun. "Deep residual learning for image recognition." In: *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition* 2016-Decem (2016), pp. 770–778. ISSN: 10636919. DOI: 10.1109/CVPR.2016.90. arXiv: 1512.03385.

[43] *Mavic 2 -Product Information*. URL: https://www.dji.com/es/mavic-2/info (visited on 01/03/2021).

[44] *Drone Deploy Home page*. URL: https://www.dronedeploy.com (visited on 01/03/2021).

[45] *Pix4Dcapture: Free drone flight planning mobile app | Pix4D*. URL: https://www.pix4d.com/product/pix4dcapture (visited on 01/03/2021).

[46] *NAXA - Location Matters*. URL: https://naxa.com.np/ (visited on 01/03/2021).

[47] OpenDroneMap. *Drone Mapping Software - OpenDroneMap*. 2020. URL: https://www.opendronemap.org/ (visited on 01/05/2021).

[48] *Dataset Structure — OpenSfM 0.4.0 documentation*. URL: https://www.opensfm.org/docs/dataset.html{\#}reconstruction-file-format (visited on 01/05/2021).

[49] S. Shen. "Accurate Multiple View 3D Reconstruction Using Patch-Based Stereo for Large-Scale Scenes." In: *IEEE transactions on image processing : a publication of the IEEE Signal Processing Society* 22 (Jan. 2013). DOI: 10.1109/TIP.2013.2237921.

[50] *Cloudcompare*. URL: http://www.cloudcompare.org.

[51] W. Zhang, J. Qi, W. Peng, H. Wang, D. Xie, X. Wang, and G. Yan. "An Easy-to-Use Airborne LiDAR Data Filtering Method Based on Cloth Simulation." In: *Remote Sensing* 8 (June 2016), p. 501. DOI: 10.3390/rs8060501.

[52] *CSF (plugin) - CloudCompareWiki*. URL: http://www.cloudcompare.org/doc/wiki/index.php?title=CSF_plugin (visited on 01/06/2021).

[53] *Introduction to Keras for Engineers*. URL: https://keras.io/getting_started/intro_to_keras_for_engineers/ (visited on 12/21/2020).

[54] P. Yakubovskiy. *Segmentation Models*. https://github.com/qubvel/segmentation_models. 2019.

[55] O. Ronneberger, P. Fischer, and T. Brox. "U-net: Convolutional networks for biomedical image segmentation." In: *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)* 9351 (2015), pp. 234–241. ISSN: 16113349. DOI: 10.1007/978-3-319-24574-4_28. arXiv: 1505.04597.

[56] A. Chaurasia and E. Culurciello. "LinkNet: Exploiting encoder representations for efficient semantic segmentation." In: *2017 IEEE Visual Communications and Image Processing (VCIP)*. 2017, pp. 1–4. DOI: 10.1109/VCIP.2017.8305148.

[57] H. Zhao, J. Shi, X. Qi, X. Wang, and J. Jia. *Pyramid Scene Parsing Network*. 2017. arXiv: 1612.01105 [cs.CV].

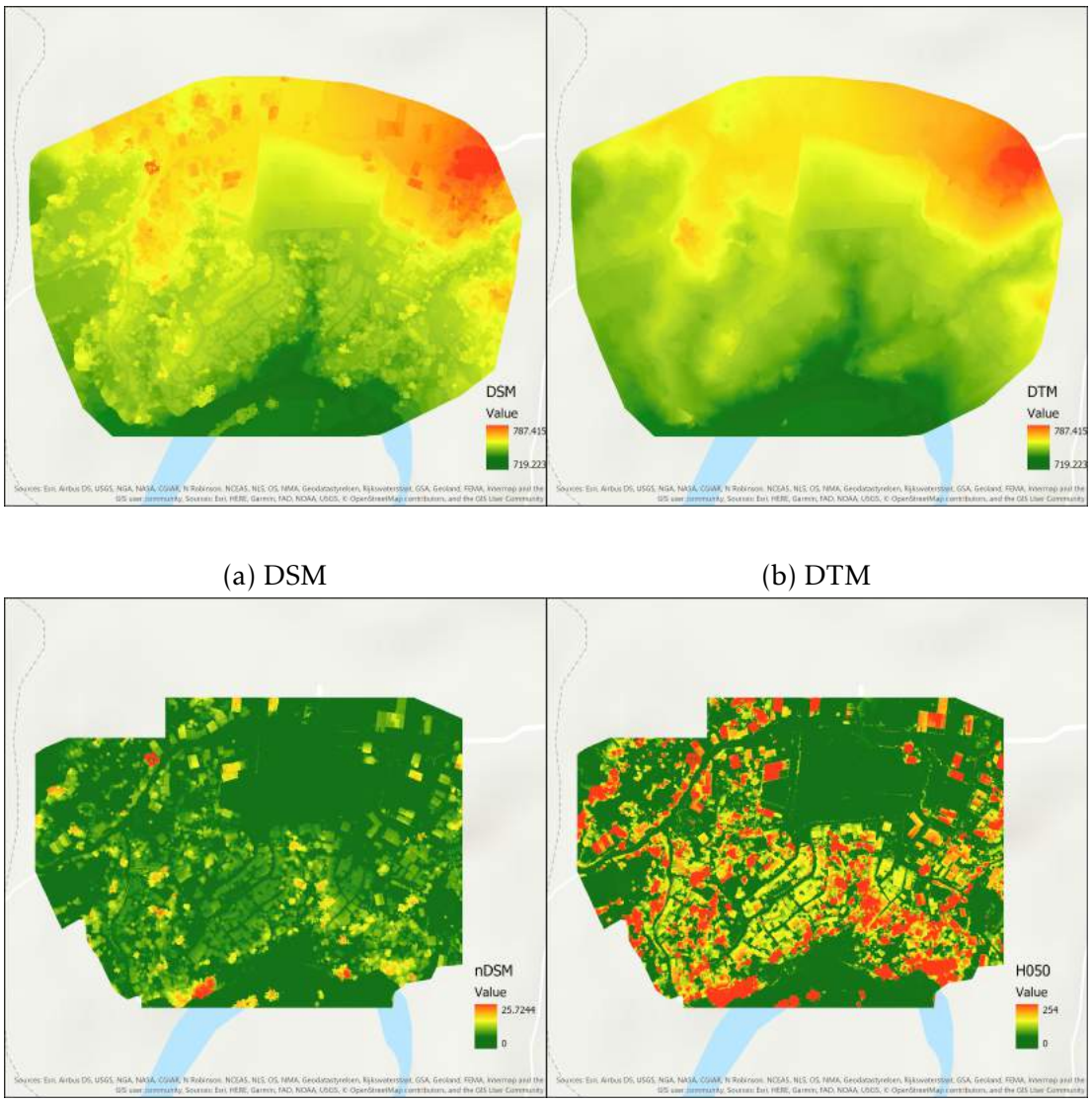
[58] T.-Y. Lin, P. Dollár, R. Girshick, K. He, B. Hariharan, and S. Belongie. *Feature Pyramid Networks for Object Detection*. 2017. arXiv: 1612.03144 [cs.CV].

[59] *Transfer Learning from RGB to Multi-band Imagery | Azavea*. URL: https://www.azavea.com/blog/2019/08/30/transfer-learning-from-rgb-to-multi-band-imagery/ (visited on 02/18/2021).

Appendix A

Point Cloud Filtering and Feature Height Derivation

Figure A.1: Sample Point cloud before and after filtering. Top: Original Point Cloud, Bottom: Filtered point cloud



(a) DSM (b) DTM

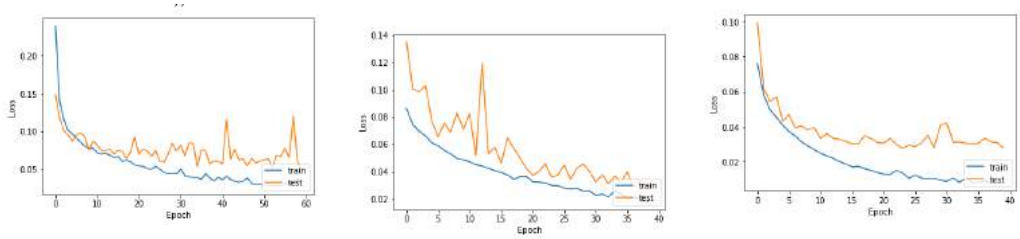
(c) nDSM (d) nDSM threshold by 5 meters

Figure A.2: Sample Height maps from Chhorepatan: DSM, DTM, nDSM and Feature Height Normalized by 5 meters. All heights above 5 meters are converted to 254 and all other values are scaled proportionally in subfigure ??

.

Appendix B

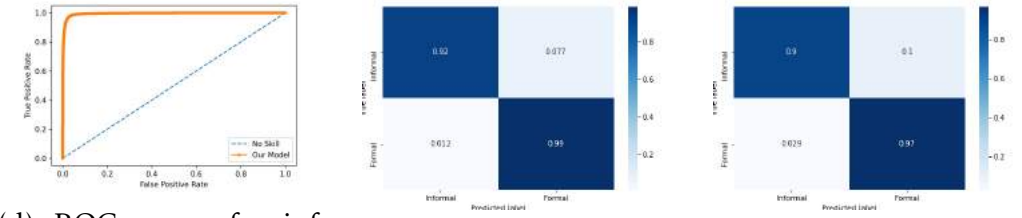
Model Performance



(a) Sample Training and Validation Loss for dual-branch FCN8s(lr=0.0001)

(b) Loss curve for a model trained end to end

(c) Loss curve for same model trained with pre-trained weights



(d) ROC curve for informal settlements from dual-branch RGBH050 model

(e) Confusion Matrix for Test Data

(f) Heatmap representation for confusion matrix

Figure B.1: Sample model performance visualisation for dual-branch RGH050 model. Using pretrained weights from imagenet had smooth loss curve. However, the visual outputs were noisy and statistically poorer

Masters Program in Geospatial Technologies

INFORMAL SETTLEMENT SEGMENTATION USING VHR RGB AND HEIGHT INFORMATION FROM UAV IMAGERY: A CASE STUDY OF NEPAL

Ganesh Prasad Sigdel

Dissertation submitted in partial fulfilment of the requirements
for the Degree of *Master of Science in Geospatial Technologies*

NOVA IMS Information Management School
Universitat Jaume I
Westfälische Wilhelms-Universität Münster

Masters Program in **Geospatial Technologies**

Supported by:
Education and Culture
ERASMUS MUNDUS