DEEP LEARNING-BASED MULTIMODAL FUSION OF SENTINEL-1 AND SENTINEL-2 DATA FOR MAPPING DEFORESTED AREAS IN THE AMAZON RAINFOREST

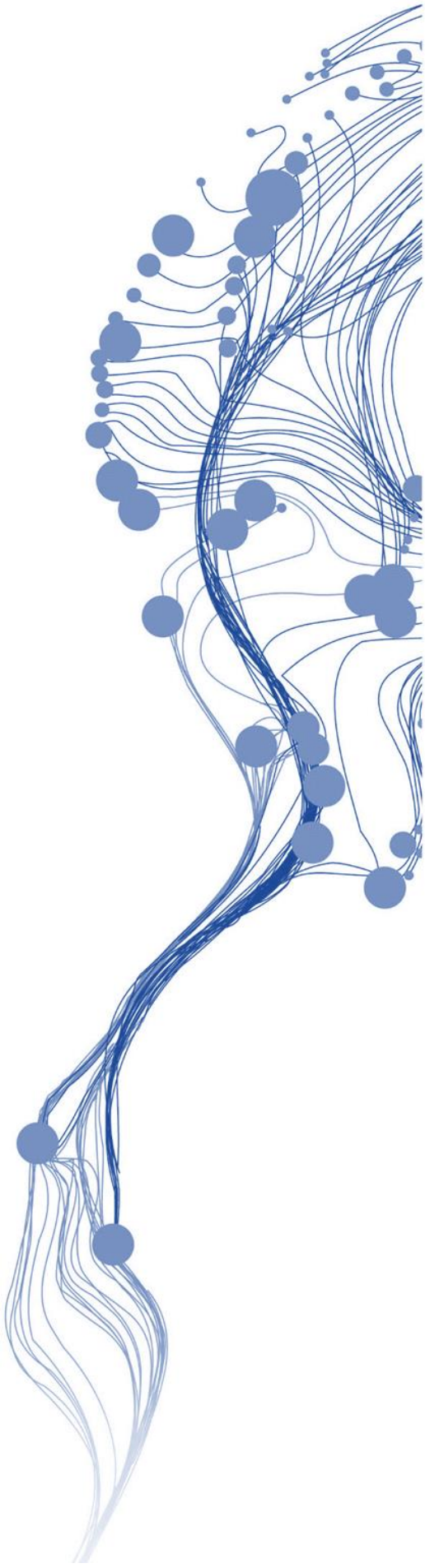
BIMAN BISWAS

June 2022

SUPERVISORS:

Dr. Raian V. Maretto

Prof. Dr. Claudio Persello



DEEP LEARNING-BASED MULTIMODAL FUSION OF SENTINEL-1 AND SENTINEL-2 DATA FOR MAPPING DEFORESTED AREAS IN THE AMAZON RAINFOREST

BIMAN BISWAS

Enschede, The Netherlands, June 2022

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

Specialization: Geoinformatics

SUPERVISORS:

dr. Raian Vargas. Maretto, Asst. Professor, Department of EOS

prof.dr.ir. Claudio Persello, Assoc. Professor, Department of EOS

THESIS ASSESSMENT BOARD (CHAIR):

prof.dr.ir. Alfred Stein, Full Professor, Department of EOS

dr. Leila Maria Garcia Fonseca, (External Examiner, National Institute for Spatial Research, Brazil)

UNIVERSITY OF TWENTE. | ITC

DISCLAIMER

ABSTRACT

As deforestation continues to increase rapidly, it is critical to map it reliably and efficiently to protect tropical rainforests and implement effective containment policies. Furthermore, it supports monitoring deforestation and assessing its effects on local and global climate and biodiversity loss. However, conventional methods to map deforestation using optical satellite imagery suffer from persistent cloud cover and are often impractical and unsuitable for complex, large-scale analysis. Synthetic aperture radar (SAR) images have the ability to penetrate the cloud cover and provide an alternative data source to monitor deforestation. Therefore, this research aims to perform a fully convolutional network (FCN) based multimodal fusion of optical and SAR data to map the accumulated deforestation regardless of any atmospheric condition. The experiments were carried out in parts of Pará state in the Brazilian Amazon. 10m Sentinel-1 (S-1) SAR data and 10m bands of Sentinel-2 (S-2) optical data were used as input data, and primary forest and non-forest data from the Brazilian Amazon Deforestation Monitoring Program (PRODES) were used as reference data. Five image pairs for five different cloud scenarios, from 0% to 100% cloud cover, were used to prepare the training, testing, and validation data. U-Net variations with early fusion, late fusion and spatial attention mechanisms were used to experiment with the two input data sets in two different scenarios of experiment setups. Scenario-1 was set up to train and test on the same image. And scenario-2 was set up to train and test images from different dates.

The results from the experiments in scenario-1 suggest that the accuracy of the standalone S-2 image outperforms every other model in the zero percent cloud scenario. The fusion based models come very close to standalone S-2 performance in this scenario but do not improve the results further. As expected, the performance degrades abruptly for standalone S-2 images when the cloud-cover increases. Results from scenario-2 suggest that with the help of fusion of S-1 and S-2 images during a cloudy scenario, the models can output impressive classification results even during an extreme cloud cover scenario. Further investigation about improving the fusion accuracy during cloud-free conditions in scenario-2 was left for future research works.

Keywords:

deforestation, deep learning, fully convolutional network, multimodal fusion, attention mechanism, sentinel-1, sentinel-2

ACKNOWLEDGEMENTS

The journey of completing this MSc research would have been impossible without my supervisor, dr. Raian V. Maretto, who guided and mentored me with his knowledge and expertise throughout the research work. I am thankful to him for showing patience, support and understanding of my struggle to learn a new topic at the beginning of my research. I would also like to thank my second supervisor, prof.dr.ir. Claudio Persello, for his valuable feedback and creative ideas for the experiments during my thesis. I would like to extend my sincere thanks to my chair prof.dr.ir. Alfred Stein and replacement chair, dr. Marian Belgiu for their insightful comments and support during various stages of my thesis work.

Additionally, I am grateful to ITC for providing me with ITC Excellence Scholarship and making it possible to start my journey at ITC. Sincere gratitude also goes to the Dutch Ministry of Education, Culture and Science for granting me a Holland Scholarship. Keep up the excellent work.

Words cannot express my gratitude to my mentor and inspiration, dr. Markand P. Oza, for making my dream of ITC journey a reality. You inspire me to be a better human being every day.

Special thanks to my internship supervisors at DLR, dr. Michael Nolde and Florian Willy Fichtner for understanding my challenges and giving me enough time whenever required to work independently on my thesis.

I would also like to thank my group of wonderful friends, Nitheshnirmal Sadhasivam, Enzo Campomanes V, Sahara Sedhain, Sachita Shahi, Lynette Dias, and Diana Collazos Cortez, who made Enschede feel like a home away from home. Thanks to you guys for keeping me on track throughout this thesis journey. Thanks for making studio 5-036 a very fun and enjoyable workplace for sharing ideas, stories and constructive feedback. Without your presence and encouragement, I would not be completing this thesis in time.

Nobody has been more important to me during the last two years than you, Miss Binita Khanal. Thank you for always being a friend, guide, critic, and partner for me throughout this thesis period. I could not have asked for a better support than you.

Last but not least, I would like to mention my family, especially Maa, Baba, and my sister Tintin, for their blessings and constant moral support during this thesis period.

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

LIST OF ABBREVIATIONS

ANN	Artificial Neural Network
AOI	Area of Interest
BCE	Binary Cross Entropy
CBAM	Convolutioinal Block Attention Module
CNN	Convolutional Neural Network
DEM	Digital Elevation Model
DL	Deep Learning
EF	Early Fusion
ESA	European Space Agency
FCN	Fully Convolutional Networks
GAN	Generative Adversarial Network
GRD	Ground Range Detected
GSD	Ground Sampling Distance
HIS	Hue-Intensity-Saturation
IoU	Intersection over Union
IW	Interferometric Wide
LF	Late Fusion
LULC	Land Use Land Cover
MSI	MultiSpectral Instrument
NPCC	National Policy on Climate Change
PCA	Principal Component Analysis
PMGI	Proportional Maintenance of Gradient and Intensity
PRODES	Brazilian Amazon Deforestation Monitoring Program
S-1	Sentinel-1
S-2	Sentinel-2
SAR	Synthetic Aperture Radar
SNAP	Sentinel Application Platform
SVM	Support Vector Machines
TAFFN	Triplet Attention Feature Fusion Network
U-Net2	U-Net with 2 blocks of Encoder
U-Net3	U-Net with 3 blocks of Encoder

Chapter 1

Introduction

This chapter provides an overview of information about the current deforestation trends in the Amazon rainforest. The past and current technological developments for mapping deforestation are also discussed.

1.1 BACKGROUND AND MOTIVATION

1.1.1 Deforestation in Amazon

The tropical forests have been cleared at an alarming rate (Hoang and Kanemoto, 2021). Loss of natural vegetation through deforestation is the second largest source of anthropogenic greenhouse gas emissions. The consequences of such phenomena include the loss of biodiversity, changes in hydrological cycles, local and global climate change, and disruption of rights and livelihood of local communities (D'Almeida et al., 2007; Giam, 2017; Hoang and Kanemoto, 2021; Houghton, 1999). Economic development, population growth, and international trade are primarily responsible for global deforestation, driven by commodity production, forestry, agriculture, and urbanisation (Rodrigues et al., 2009). The Amazon rainforest in Brazil encompasses the most extensive stretch of tropical forest in the world (spanning 6.7 million km^2, double the size of India). It is biologically the wealthiest region of our planet, hosting ≈25% of global biodiversity (Malhi et al., 2009; Nicolau et al., 2021). Despite the target set by National Policy on Climate Change (NPCC) to reduce deforestation in the Amazon, there has been an increasing trend in the rate of deforestation. The year 2019 saw an increase of 34% in deforestation, equating to an alarming rate of 10,129 km^2 of clear-cut deforestation. The Brazilian Amazon Deforestation Monitoring Program (PRODES) reported the area of accumulated deforestation for 2020 to be 10,851 km^2. This amount was 176% higher than the established NPCC target of 3925km^2 (Silva Junior et al., 2021). It corresponds to 648 TgCO2 (648 million tons of CO2) released into the atmosphere due to deforestation (Silva Junior et al., 2021).

Furthermore, PRODES estimated deforestation in 2021 to be 13,235km^2 based on 45% of the monitored area. Hence, this continued large-scale deforestation of the amazon would cause perpetual damage to the functioning and diversity of the biosphere. As deforestation increases rapidly (Werth, 2002), it is crucial to map it accurately and rapidly for managing tropical rainforests and undertaking effective containment policies (Maretto et al., 2021). In addition, it helps in monitoring deforestation and understanding its implications on local and global climate and the decline in global biodiversity (Cabral et al., 2018; de Bem et al., 2020; Werth, 2002). On a global scale, these maps help achieve the target 15.2 of sustainable development goal (SDG) number 15, which aims to promote the sustainable management of all types of forests, stop deforestation, restore the degraded forest, and significantly increase global reforestation and afforestation.

1.1.2 Mapping Deforestation Using Satellite Imagery

Remote sensing technologies have played a significant role in recent decades by providing consistent, accurate, and timely information to study our planet (Cremer et al., 2020; Maretto, 2020). Land Use and Land Cover (LULC) change detection is one of the main uses of satellite remote sensing data (Syrris et al., 2019; Treitz, 2004). It consists of analysing and quantifying the state of an object at different times (Singh, 1989) and is an essential step in understanding deforestation processes. However, traditional manual analyses to study deforestation from the imagery are expensive for complex, large-scale analysis. So, producing an accurate, automated, fast, and responsive deforestation detection system with a reasonable

accuracy has been an open challenge in the remote sensing community (Ball et al., 2017; Camps-Valls et al., 2014; Lu and Weng, 2007; Syrris et al., 2019). The presence of artefacts from cloud and cloud shadows, signal inconsistency due to varying environmental conditions, and phenological changes are a few challenges that may hinder mapping a LULC change phenomenon (Liu et al., 2020; Nguyen et al., 2020).

While optical satellite data is widely used in LULC mapping (Sefrin et al., 2020; Wang et al., 2020; Yin et al., 2018), Synthetic Aperture Radar (SAR) data is gaining popularity as data from SAR sensors become available freely. SAR sensors have enabled the ability to acquire images regardless of weather conditions. As part of the Copernicus program of the European Space Agency (ESA), the Sentinel-1 (S-1) satellite with its C-band SAR provides a revisit frequency of six days. In the interferometric wide (IW) swath mode, Nominal land acquisition provides a spatial resolution of 5 m × 20 m in dual-polarization channels in the form of phase and amplitude information. The free, full, and open data policy enables users to access extensive scale data with rich source information. These open large-scale geodata represent a huge opportunity to create an advanced innovative methodology for different LULC mapping like deforestation. However, there are only a few reliable and automated methods for detecting deforestation using these big geodata with SAR images.

Several studies have looked at the viability of SAR imagery for LULC mapping, with an emphasis on polarimetric multitemporal (Bruzzone et al., 2004) and multi-frequency SAR in the L-band, C-band, and X-band (Lonnqvist et al., 2010; Waske and Braun, 2009), as well as the combining the use of SAR and optical data (Ullmann et al., 2014). The ability of a longer wavelength (L-band) to penetrate deeper into the forest structure is more appropriate for mapping forest cover. Unfortunately, there is no free SAR. L-Band time series dataset available worldwide.

1.1.3 Deep Learning for Image Classification

In recent years Deep Learning (DL) based models have shown a remarkable feature representation capability in various fields, including image scene classification from remote sensing satellite images (Cheng et al., 2017, 2016; Hu et al., 2015; Nogueira et al., 2017; Yao et al., 2016; Zou et al., 2015). DL refers to a set of Artificial Neural Networks (ANNs) with the ability to learn a hierarchical representation of data for image classification, object detection, and many other applications (Lecun et al., 2015). DL models, especially Convolutional Neural Networks (CNNs), have achieved state-of-the-art results in the domain of remote sensing image classification (Yanfei Liu et al., 2018; Yu and Liu, 2018) and are most commonly utilised for pattern recognition from images (O'Shea and Nash, 2015).

CNNs are composed of a series of processing layers that perform three major tasks: 2D convolutions, unit-wise nonlinear activations, and spatial pooling with subsampling (Persello and Stein, 2017). Weights and biases of the convolution operations are learnt in a supervised way to reduce classification error. Standard architectures employ a sequence of convolutional layers that are flattened into a one-dimensional vector and fed to fully-connected layers. The Convolutional layers learn the spatial features, whereas fully-connected layers learn the classification rule that will be applied to the retrieved feature vector (Lecun et al., 2015). Because the network is trained from beginning to end, feature extraction and classification occur in the same framework. This method has been shown to be effective in various computer vision tasks, especially in image classification or object detection, where one label is assigned to the entire input scene. Deep CNNs have been successfully applied to image categorisation benchmarks, considerably outperforming techniques based on hand-designed features.

Furthermore, CNNs have been modified to perform pixel-wise classification, also known as semantic segmentation. The traditional patch-based method involves training the CNN to label the centre pixel of patches derived from the input picture (Bergado et al., 2016). Nevertheless, if applied to classify a large RS

image, this method will result in redundant processing and incur high computational costs. To overcome this computational issue, Fully Convolutional Networks (FCNs) (Shelhamer et al., 2014) are trained to infer the pixel-wise classification of an entire image or patch at once. In an FCN, the fully connected layers from CNNs are replaced with one or more upsampling layers that resample the feature map extracted by convolutional layers to the exact resolution of the input image (Badrinarayanan et al., 2017; Noh et al., 2015; Shelhamer et al., 2014).

The combination of diverse types of sensors like SAR and optical provides complementary information for the same target (Adrian et al., 2021). SAR sensors capture more of the structural properties from the backscatter energy of an object on the ground. However, they are more complex for interpretation due to the presence of speckle noise. On the other hand, the optical image has better spatial resolution and is easier to interpret but widely affected by atmospheric effects. A reliable approach is required to extract and fuse information from these two sensors. DL techniques have the potential to efficiently combine information from these two sensors because it has the advantage of automatically learning the hierarchical representation from the different modality of SAR and optical image. (Ramachandram and Taylor, 2017). It has gained a foothold and continues to gain rapid advancements in the field of human activity recognition Ebrahimi Kahou et al. (2015); Neverova et al. (2014); Radu et al. (2016); medical applications (Kiros et al., 2014; Tajbakhsh et al., 2017; Wu et al., 2013), and autonomous systems (Gu et al., 2016; Lenz et al., 2013). However, DL techniques have yet to be substantially investigated to fuse multimodal data from SAR and optical remote sensing sensors.

Attention mechanisms, like many other in DLbased methods, attempt to emulate how the human brain or eye processes data (Ghaffarian et al., 2021). The human visual system does not perceive the entire image simultaneously; instead, it focuses on specific parts. The focused part of the image is perceived as in "higher resolution", whereas the part out of focus is "low-resolution" (Ghaffarian et al., 2021). The main idea behind the attention mechanism is to give higher weights to the most relevant information in the network. Inspired by this process, Bahdanau et al., (2014) developed an attention mechanism for natural language processing. Gradually, attention mechanisms have also been successfully applied to semantic segmentation tasks (Khanh et al., 2020; Oktay et al., 2018; Roy et al., 2019, 2018; Vahadane et al., 2021; Zhao et al., 2020; Zhou et al., 2020). In the case of convolutional networks, a spatial attention mechanism focuses on the local region from a given set of feature maps (Woo et al., 2018). It produces a rich representation of the relevant features of interest from the local domains and cut out the irrelevant information or noises (Zhang et al., 2019).

1.2 RELATED WORKS

1.2.1 Image Fusion

Image fusion combines information from two or more images from the same or different sensors of different wavelengths of the same scene (Wang et al., 2005). Table 1.1 shows the commonly used existing image fusion techniques and their limitations.

Table 1.1 Overview of various image fusion approaches

Approach	Description	Limitation	Sample Literature
Simple average	Most basic approach for pixel-level image fusion.	No guarantee of an improved image.	(Malviya and Bhirud, 2009)
Simple Maximum	Compared to the average approach, results in a highly	Influenced by the blurring effect, which	(Malviya and Bhirud, 2009;

	focused image generated from the input image.	directly impacts the image's contrast.	Zheng et al., 2004)
PCA	PCA is a tool that transforms the number of correlated variables into the number of uncorrelated variables, which is helpful in image fusion.	Strong correlation between the input image is required, and fused image will have lesser quality than any of the input images.	(Abdikan, 2018; Sun et al., 2005; Walker et al., 2010)
DWT	The DWT fusion method may surpass PCA in reducing spectral distortion. Has a higher signal-to-noise ratio than pixel-based methods.	Output image has a lower spatial resolution.	(Desale and Verma, 2013)
Combined DWT & PCA	Multilevel fusion yields better results when the image is fused twice using an efficient fusion technique. The final image had a high spatial resolution and high spectral quality.	Complex method. For a better result, a good fusing technique is required.	(Pajares and de la Cruz, 2004)

1.2.2 Deep Learning-based Image Fusion

Multiple DL-based fusion approaches have been proposed for image fusion in different application fields. However, the multimodal fusion of SAR and optical images is still an evolving research field. Table 1.2 shows a few approaches that utilise DL-based networks for the fusion of multimodal images from various sources.

Table 1.2 Overview of DL-based image fusion approaches

Approach	Description	Sample Literature
CNN based fusion	Adopts a Siamese-based CNN to fuse images from different modalities. Fuse images in a multi-scale manner via image pyramids.	(Liu et al., 2017)
DenseFuse	DL-based multimodal fusion of infrared and visible images with encoder, decoder and a fusion block.	(Li and Wu, 2019)
PMGI	Fast unified fusion network based on proportional maintenance of gradient and intensity (PMGI). Can handle various tasks like medical image fusion, visible and infrared image fusion, multi-exposure image fusion and pan-sharpening.	(H. Zhang et al., 2020)
U2Fusion	Fusion of medical images based on an end-to-end unsupervised fusion network. Uses an information preservation degree of the extracted feature to evaluate the importance of each source image.	(Xu et al., 2022)

cGAN	Conditional Generative Adversarial Network(cGAN) based approach for fusion of multimodal SAR and optical imagery to synthesise cloud-free optical images.	(Bermudez et al., 2019; Y. Li et al., 2020)

1.2.3 Attention Mechanisms

Attention mechanisms were initially introduced for natural language processing (Bahdanau et al., 2014). Nevertheless, it has widely been used in various fields since its introduction, especially for medical image segmentation. However, limited research utilises attention mechanisms in remote sensing applications. Moreover, according to our best knowledge, only a few researches use spatial attention mechanisms to fuse SAR and optical images and potentially substitute the cloudy optical images with SAR images automatically. Table 1.3 shows the literature related to our study incorporating attention mechanisms.

Table 1.3 Overview of literature related to attention mechanisms

Approach	Description	Reference
AttentionU-Net	Uses a soft-attention gate inside a generic U-Net to produce attention maps that emphasise the location of the pancreas for medical image segmentation.	(Oktay et al., 2018)
CBAM	Convolutional Block Attention Module (CBAM) uses a channel and spatial attention module in a feed-forward CNN to refine the encoder features.	(Woo et al., 2018)
SCAU-Net	Enhances U-Net encoder and decoder framework with spatial and channel attention modules for medical image segmentation.	(Khanh et al., 2020; Zhao et al., 2020)
SCAttNet	Uses an end-to-end semantic segmentation network with a lightweight channel and spatial attention module for feature refinement in high-resolution remote sensing images.	(H. Li et al., 2021)
TAFFN.	Triplet Attention Feature Fusion Network (TAFFN) for the fusion of SAR and optical image. Uses spatial, channel and cross attention based on a self-attention mechanism to extract and integrate long-range and complementary information from the images and perform a land cover classification.	(Xu et al., 2021)

1.3 PROBLEM ANALYSIS

Most researches on mapping deforestation still rely heavily on the ability of an optical sensor to capture the phenomenon. Optical sensors have a huge advantage in terms of the interpretability of the images. However, as discussed in section 1.1.2, there is a limitation to using optical remote sensors to map deforestation due to their inability to penetrate the cloud. The presence of persistent clouds covering the Amazon rainforest season-wide hinders the ability to detect deforestation, especially during the wet season, when it is nearly impossible to get a cloud-free image over some regions (Griffiths et al., 2018).

With the advancement of SAR sensors and their ability to penetrate the clouds, monitoring the rainforest even during seasons with persistent cloud covers has presented an alternative way to monitor deforestation. However, there is always a trade-off for only using the SAR sensors to study deforestation. SAR images are complex to interpret and have a different modality than optical sensors. Therefore, in recent years, a substantial amount of research has focused on optimally fusing complementary and correlated data of multimodal sensors. Various attempts to fuse SAR and optical data include the wavelet-merging technique (Abdikan, 2018; Hong and Zhang, 2008; Lu et al., 2011), Principal Component Analysis (PCA) (Abdikan, 2018; Pereira et al., 2013; Walker et al., 2010) and intensity-hue-saturation (IHS) (Abdikan, 2018). However, these approaches fuse the image at a pixel level, which suffers from spectral distortion and fails to maintain the spatial resolution of input images (Yu Liu et al., 2018).

1.4 RESEARCH GAP AND SCIENTIFIC CONTRIBUTION

Even though most studies related to deforestation use longer wavelength L-band S.A.R. data because, they have better penetration capability into the forest (Almeida-Filho et al., 2007; Mitchard et al., 2011; Watanabe et al., 2018). However, the L-band S.A.R. data are currently not freely available. Even though C-band SAR data has a shorter wavelength than L-band, the C-band data from S-1 satellites is freely available to download. Hence, the backscatter data from a C-band SAR combined with optical data will be used in this research to assess its suitability for mapping deforestation. Furthermore, the current state-of-the-art techniques for detecting deforestation only utilise optical remote sensing data, which sometimes relies on visual analysis to extract information from the data source.

Many researchers have attempted to fuse remote sensing imagery from various sensors using DL techniques. Several studies tried to fuse medium-resolution multispectral images with a high-resolution panchromatic image to generate higher-resolution images with all the spectral information (Masi et al., 2016; Shao and Cai, 2018; Zhong et al., 2016). Using CNN-based architecture, Palsson et al. (2017) fused hyperspectral and multispectral images. Methods based on Generative Adversarial Networks (GANs) (Goodfellow et al., 2014) try to directly synthesise optical cloud-free images from an input SAR image (Bermudez et al., 2019; Gao et al., 2020; Y. Li et al., 2020). (Y. Li et al., 2021a, 2021b) use a supervised method based on CNN to fuse medical images. However, research on fusing optical and SAR data to detect deforestation are limited.

Fusion techniques with the addition of attention mechanisms have also enhanced the performance of multimodal fusion for multiple classification tasks. (X. Li et al., 2020) used a multimodal fusion network with a second-order channel attention mechanism for land cover classification. (Oktay et al., 2018) used an attention mechanism in a U-Net architecture (Ronneberger et al., 2015) to identify the pancreas from medical imagery. (Zhu et al., 2020) performed a multimodal fusion of audio, visual and language parameters with a self-attention mechanism for sentiment analysis. (H. Li et al., 2021) used a combination of spatial and channel attention mechanisms for semantic segmentation using high-resolution images for land use classification; however, their method does not use any fusion mechanism. Even though extensive research identified many advantages of using an attention mechanism in a neural network, there is no research on its usability to effectively fuse complementary information from SAR and optical data. Therefore, this research aims to effectively explore the potential of a fusion with an attention mechanism to map deforestation irrespective of weather conditions.

1.5 RESEARCH IDENTIFICATION

This section presents the research objective and sub-objectives of this research.

1.5.1 Objective and Sub-objective

This research aims to formulate a method to map deforestation in parts of the Brazilian Amazon rainforest using an FCN-based multimodal fusion of S-1 SAR and S-2 optical data. The model is expected to be robust and adaptable for mapping deforestation regardless of any atmospheric condition.

To achieve the primary goal of this research, the following sub-objectives were formulated:

1. Prepare and build datasets for designing, training, testing, and validating the model.
2. Design and implement a DL-based multimodal fusion model with a spatial attention mechanism.
3. Evaluate the performance of the fusion models in images with different levels of cloud cover and their impact on the final accuracy
4. Train, calibrate, and generalise the DL model to be applied to unseen data and validate the output quality based on the reference data.

1.5.2 Research Questions

The following research questions are derived from the study objectives and sub-objectives listed above:

1. How to prepare and build the dataset for training, testing and validation? (Objective 2)
2. What are the crucial criteria for designing the network and performing a multimodal fusion of S-1 and S-2 data with an attention mechanism? (Objective 3)
3. How to optimise the network to reduce loss during training and validation for better generalisation capability? (Objective 3)
4. How does the network perform in images with varying cloud cover? Does the fusion with the attention mechanism help to increase the classification quality? (Objective 4)
5. How does the predicted deforestation map compare with the reference data regarding different accuracy metrics and qualitative analysis? (Objective 4)

Chapter 2

Study Area and Datasets

This section highlights geographical information about the study area and the datasets used to conduct the study.

2.1 STUDY AREA

The study is conducted over a small area of Pará state, the second largest state in Brazil. With about 1.25 million square kilometres of area, the state is located in the northern part of Brazil. Since 2006 42.65% of total deforestation in Legal Amazon has occurred in the Pará state (INPE, n.d.). As the state is considerably large, a smaller study area was chosen from the southern part of the state, as depicted in Figure 2.1. The study area covers an approximate area of 18 thousand square kilometres. Tropical monsoon climate dominates the study area, with annual precipitation exceeding 2000 mm (Griffiths et al., 2018). June, July, and August compose the dry seasons. The rest of the year receives frequent rainfall accompanied by persistent cloud cover. The highest elevation lies in the southwestern corner of the study area. Its approximate elevation is 600m above sea level.

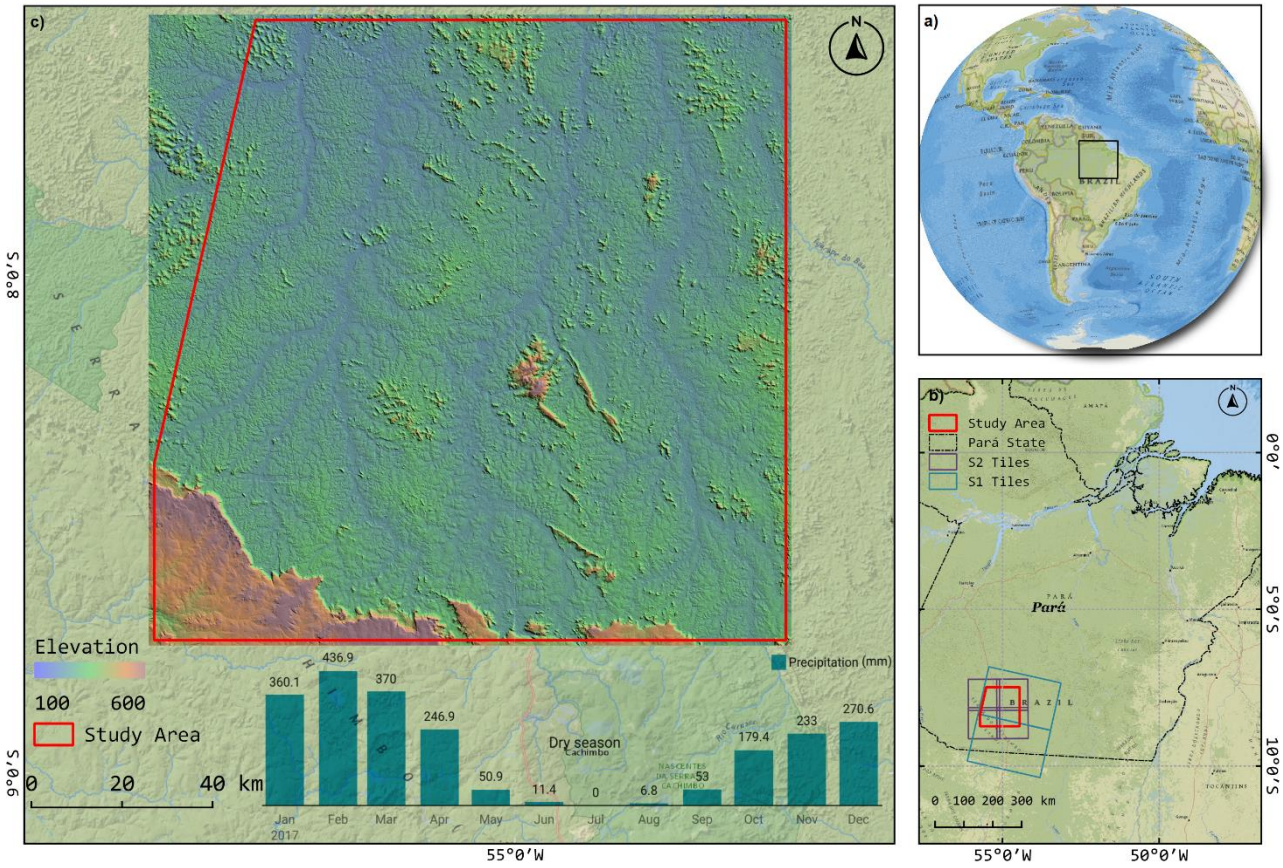


Figure 2.1 Map of the study area: a) Location of the study area in a global context; b) Location of the study area inside the Pará state along with the footprint of the Sentinel-1 and Sentinel 2 tiles over that. c) The study area map with monthly average precipitation for 2017.

2.2 DATASETS

For our study area, the datasets were acquired from 3 different sources. The optical dataset is a Level 1-C (L1C) Top of Atmosphere (TOA) reflectance from the S-2 mission, the radar data from Level-1 Ground Range Detected (GRD) products of the S-1 mission and the ground truth shapefile for 2017 from PRODES. Deforestation shapefiles are distributed by (INPE, n.d.). Since the most recently released products of both S-1 and S-2 are kept online in the official data access server for only one month, acquiring the data over one month is a considerable challenge. Therefore, for S-1, we acquired all our data from Alaska Satellite Facility[1] (ASF) mirror using a python wrapper of the ASF[2] SearchAPI[3]. Similarly, for S-2, we obtained our images from a public Google Cloud storage dataset[4] mirror of the S-2 L1C data products. Google hosts this dataset and provides public access to them through Google Cloud console, gsutil, or Cloud Storage API.

2.2.1 Optical Data

S-2 mission of the European Space Agency (ESA) consists of two constellations of earth observation satellites that provide high-resolution optical imagery covering the entire globe with a revisit time of 5 days. The L1C product from S-2 has provided Top of Atmosphere (TOA) reflectance data since June 2015 globally. Each tile of the L1C product covers an area of 100 ×100 km² in the Universal Transverse Mercator (UTM) projected coordinate system. Among all the multispectral bands shown in Table 2.1, only Near-Infrared, Red, Green, and Blue channel is distributed at a spatial resolution of 10 metres. In addition to these bands, L1C products also include qualitative information about cloud masks.

Table 2.1 Description of multispectral bands of S-2 MultiSpectral Instrument (MSI) sensor

Band	Resolution	Central Wavelength	Band Description
B1	60 m	443 nm	Ultra-blue (Coastal and Aerosol)
B2	10 m	490 nm	Blue
B3	10 m	560 nm	Green
B4	10 m	665 nm	Red
B5	20 m	705 nm	Vegetation Red Edge
B6	20 m	740 nm	Vegetation Red Edge
B7	20 m	783 nm	Vegetation Red Edge
B8	10 m	842 nm	Near Infrared (NIR)
B8a	20 m	865 nm	Vegetation Red Edge
B9	60 m	940 nm	Water vapour
B10	60 m	1375 nm	Short Wave Infrared (SWIR) - Cirrus
B11	20 m	1610 nm	Short Wave Infrared (SWIR)
B12	20 m	2190 nm	Short Wave Infrared (SWIR)

2.2.2 Radar Data

S-1 mission comprises a constellation of two polar-orbiting satellites. They provide data by capturing C-band SAR images irrespective of the time of the day or the weather. Focused SAR data is detected, multi-looked, and projected to the ground range using an Earth ellipsoid model in Level-1 GRD Products. Pixel values depict the observed magnitude while the phase information of GRD products is lost in the process. The resulting image has an approximate square resolution pixels with reduced speckle noise. GRD products

[1] https://search.asf.alaska.edu/
[2] https://cloud.google.com/storage/docs/gsutil
[3] https://github.com/asfadmin/Discovery-asf_search
[4] https://cloud.google.com/storage/docs/public-datasets/sentinel-2

at the highest available resolution of 10m (GRDH) were acquired for this study.

2.2.3 Reference Data

Primary forest and non-forest data were downloaded for the year 2017 from PRODES (INPE, n.d.). The data is available to download as a shapefile or classified raster for each year. PRODES uses satellite images from LANDSAT at 30m resolution to produce these maps. The classes in the downloaded reference data could be grouped into four classes. The first class is called 'forest', which consists of forest areas. The second class group is 'non-forest1' and 'non-forest-2', which is the non-forest class and consists of land covers like bare soil, rocks, hill tops, and vegetation cover that are not forest formation. For the deforestation class, there are two sub-categories, namely r_*yyyy* and d_*yyyy*. The r_*yyyy* is the residual deforestation. It is deforestation from an unknown previous year than *yyyy* but was only detected in the year *yyyy* because of factors like cloud coverage or unavailable data. The d_*yyyy* is the actual deforestation for the year *yyyy*. And the final class is called 'hydrography', which includes bigger water bodies and rivers.

Table 2.2 shows the required data to achieve the objectives of the research and answer all the research questions:

Table 2.2 Overview of datasets used for the study

Category	Data	Date	Provider
Optical	S-2 L1C MSI (10m bands)	2017-01-01 to 2017-12-31	(ESA, n.d.)
SAR	S-1 GRDH (10m)	2017-01-01 to 2017-12-31	(ESA, n.d.)
Reference Data	PRODES Deforestation (30m)	2017	(INPE, n.d.)

Chapter 3

Research Methodology

This chapter provides step-by-step information on transforming the raw acquired data from all the sources and making them ready to be used for generating training data.

3.1 DATA PREPROCESSING

3.1.1 Sentinel-1 Preprocessing

The standard generic workflow in Figure 3.1; was used to preprocess the S-1 GRD images. The complete workflow to preprocess the data was accomplished using pyroSAR[5], a python framework for large-scale SAR satellite data processing developed over the APIs provided by European Space Agency (ESA.) Sentinel Application Platform (SNAP) (ESA, n.d.).

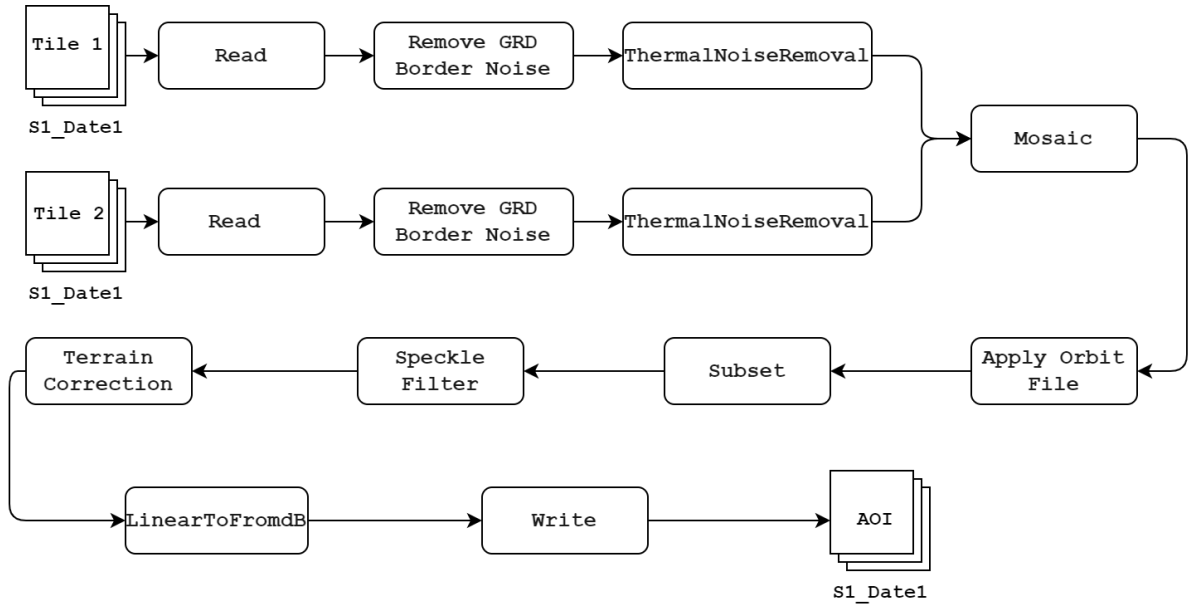


Figure 3.1 General pre-processing steps for Sentinel-1 GRD image

The workflow starts with reading multiple files from the adjoining acquisition of the same date. The border noise removal algorithm is then used to remove invalid data and low-intensity noise of the scene edges. After this, the thermal noise removal tool reduces the noise level in inter-sub-swath texture, particularly by normalizing the backscatter signal across the entire scene. With the help of the Slice Assembly operator available from the SNAP toolbox, we mosaic the border noise and thermal noise corrected image into a single scene. From this point on, the rest of the preprocessing steps are done on the mosaiced scene in sequential order.

The metadata information about the orbit state vector of raw SAR images is generally inaccurate. A precise orbit state of the satellite is sometimes determined days/weeks after acquiring the image. This precise orbit

[5] https://github.com/johntruckenbrodt/pyroSAR

was applied to generate accurate satellite position and velocity information. After this, we subset the image to our area of interest (AOI) and filter the speckle noise to improve the quality of the image. For that, the Refined Lee filter is used because of its ability to preserve edges, linear features, and texture information (Lee et al., 2009). The next step is to fix the geometric distortion of the topography. Range Doppler terrain correction (Schubert and Small, 2008) was used to rectify geometric distortions in the topography, such as foreshortening and shadows. It uses a digital elevation model (DEM) to adjust the location of each pixel. In the final phase of the preprocessing workflow, the unitless backscatter coefficient is transformed to dB using a logarithmic transformation.

3.1.2 Sentinel-2 Preprocessing

The S-2 preprocessing pipeline depicted in Figure 3.2 was built using the Python API of the Geospatial Data Abstraction Library[6] (GDAL/OGR). At first, four tiles covering the study area for a single date were read into the workflow. All the bands with a spatial resolution of 10m were extracted from these four images, which are red, green, blue, and near-infrared. Simultaneously the vector file that provides information about the cloud mask is also extracted and rasterised to stack with the 10m bands of each tile. After this, the stacked tile consisting of 10m bands and the cloud mask were mosaiced together as a virtual raster and stored in temporary storage. Finally, the virtual raster was clipped with the AOI and saved to the file system.

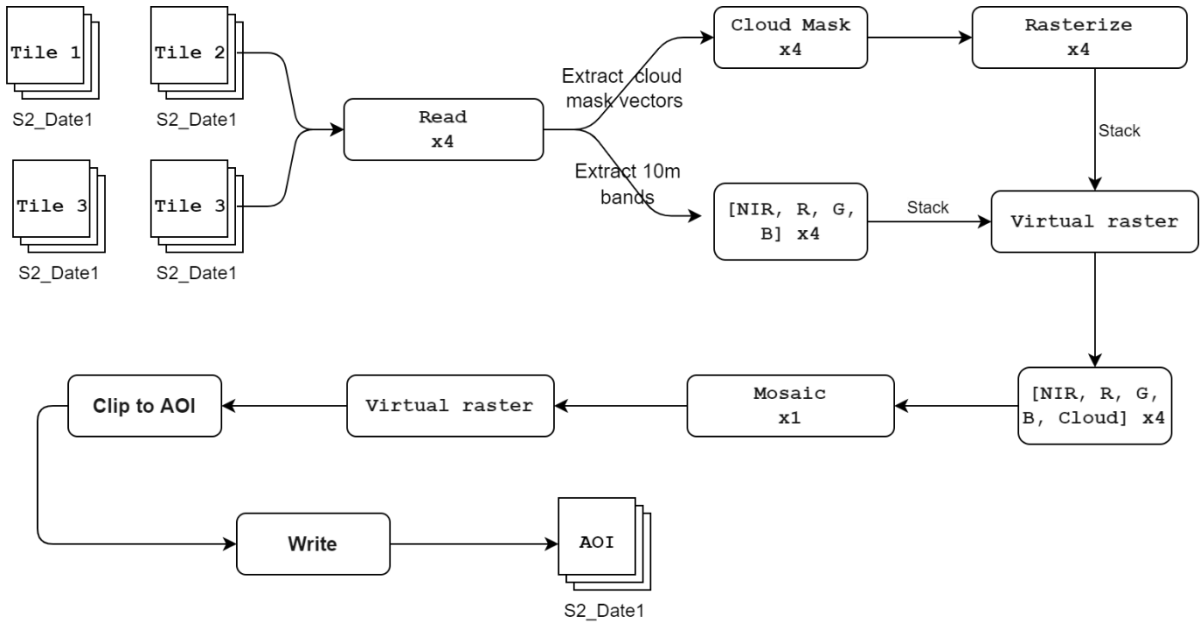


Figure 3.2 Pre-processing pipeline for Sentinel 2 L1C data

3.1.3 Reference Data Preparation

Reference data for the study was prepared by reclassifying the classes described in section 2.2.3. As seen in Figure 3.3, most of the deforested areas were recorded before 2017. For future reference, it should be noted that the area deforested before 2017 could be well vegetated with shrubs, grass, or small trees in a shorter period. However, it takes a considerable amount of time to regrow a deforested patch. Therefore, this research aims to map the accumulated deforestation, not only the deforestation from 2017. Therefore, the area deforested before and during 2017 was merged to create the deforestation reference class. The rest of

[6] https://gdal.org/python/index.html

the classes were merged together to form the not-deforested class. After completing the reclassification, 27.5% of the study area belonged to the deforested class, and 72.5% were not-deforested class.

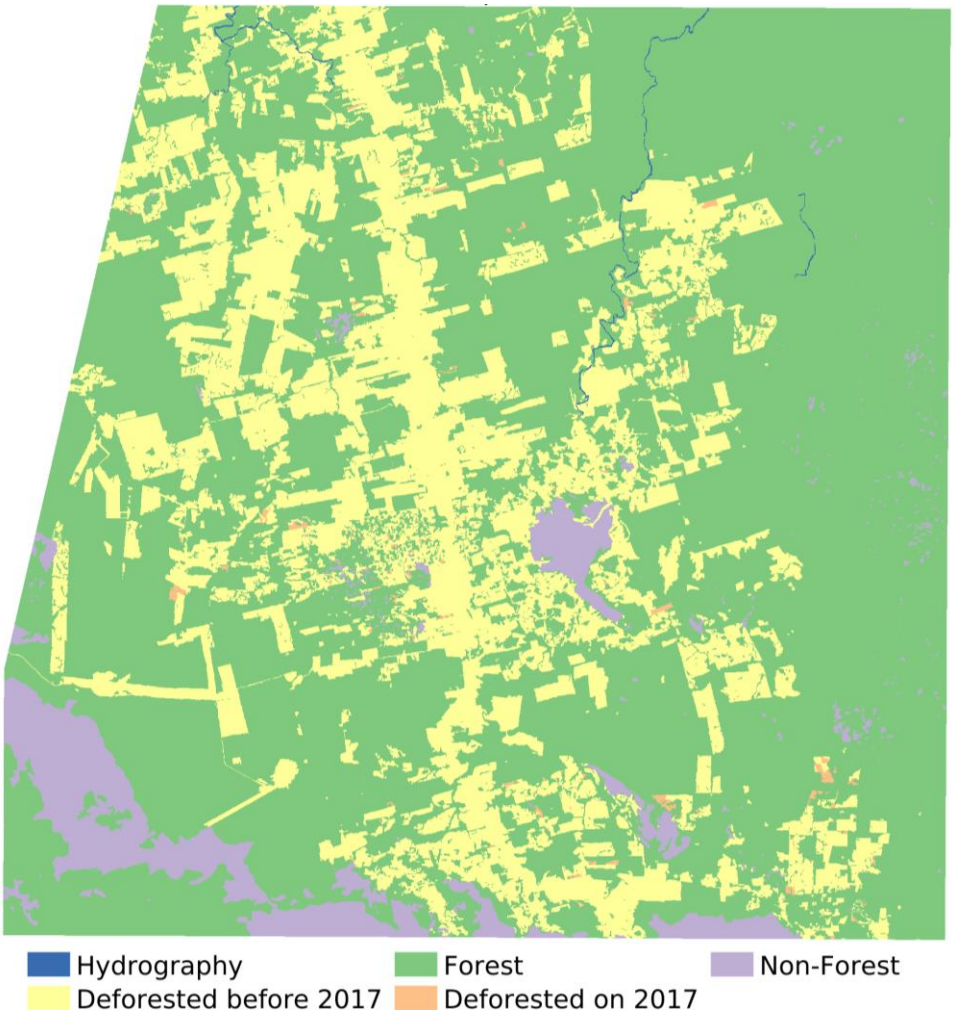


Figure 3.3 Reference map showing different classes in the study area.

3.2 MODEL DEVELOPMENT

3.2.1 U-Net

The core structure of our model was adapted from the U-Net architecture, first introduced by (Ronneberger et al., 2015) for biomedical image segmentation. The main idea of a U-Net architecture is to use skip-connection to fuse high semantic but coarse spatial features with corresponding low semantic but finer spatial features (Figure 3.4). It receives then the locational information from the encoder layers and aggregates it with the decoder features to recover the spatial information. However, the multiscale skip connection in the U-Net receives unnecessary information from the low-level encoder features (Khanh et al., 2020). Therefore, the network needs to focus on the salient low-level features of the encoder, representing rich spatial contextual information

The fully convolutional model in this research is inspired by the U-Net architecture and is explicitly designed to handle image segmentation problems. Figure 3.4 depicts the base architecture for the U-Net proposed

in this thesis. The implementation details for each of the layers in the U-Net are elaborated in-depth in Figure 3.5.

The proposed model is a variation of the original U-Net with the same structure for the encoder blocks, where we use a 2-D convolutional block followed by a max-pooling layer to encode features representing the input image at multiple levels. Each max-pooling operation increases the number of feature mappings (channels) in a typical convolutional neural network (CNN) design. However, we opted to maintain a consistent number of sixty-four feature maps across our network. Two observations influenced this decision. First, because the model has access to low-level features in the upsampling path of the decoder, we may allow the network to lose some information after the downsampling layer from the decoder. Second, because there is no notion of depth or high temporal frequencies to understand in the input satellite images, many feature maps in the upper layers may not be necessary for optimal performance.

In the encoder, padding of $(1,1)$ was used for convolution operation and padding of $(0,0)$ with stride $(2,2)$ in the MaxPool operation to down-scale the feature maps. Transposed-convolution processes are followed by concatenation and standard convolution in the decoder part of the network. In addition to the padding of (1,1), the output padding of (1,1) was used in transposed convolution to upscale the feature maps. The objective of the decoder is to semantically project the discriminative feature representations of the encoder (lower resolution) onto the pixel space (higher resolution) to obtain a rich classification.

3.2.2 U-Net with Early Fusion

U-Net with early fusion (EF) is adapted from EF implementation by Maretto et al., (2021). It follows the same architecture and implementation as the base U-Net, with only a different input size to the model. All the inputs were stacked together as one to feed to the network. The rest of the architecture is the same U-Net variation as in Figure 3.4.

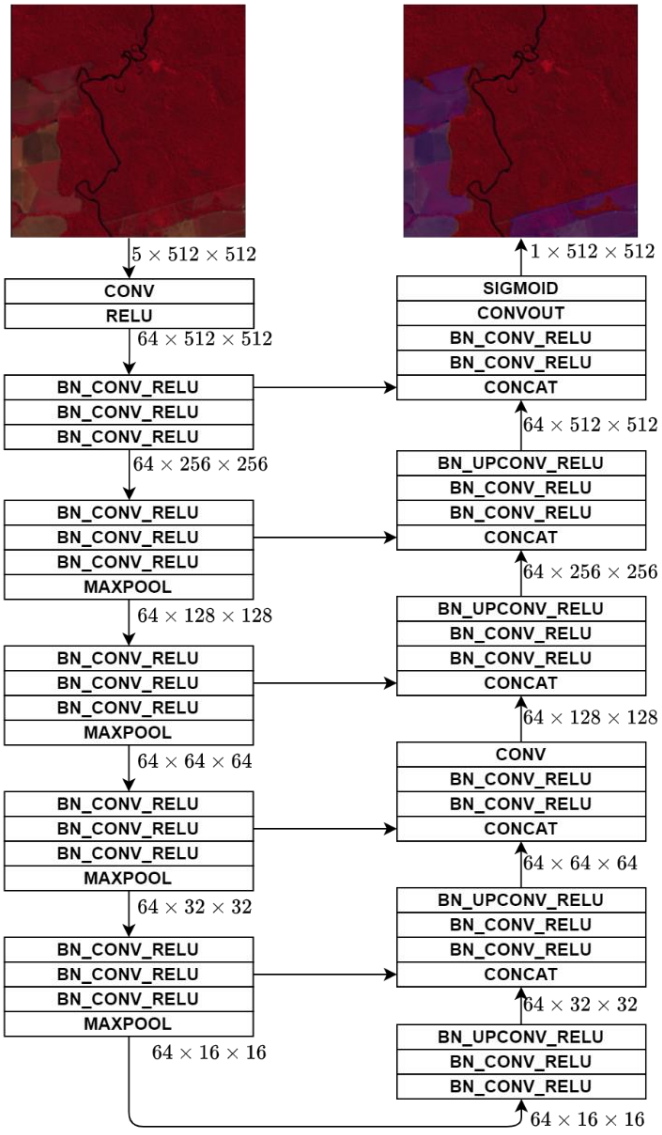


Figure 3.4 U-Net architecture

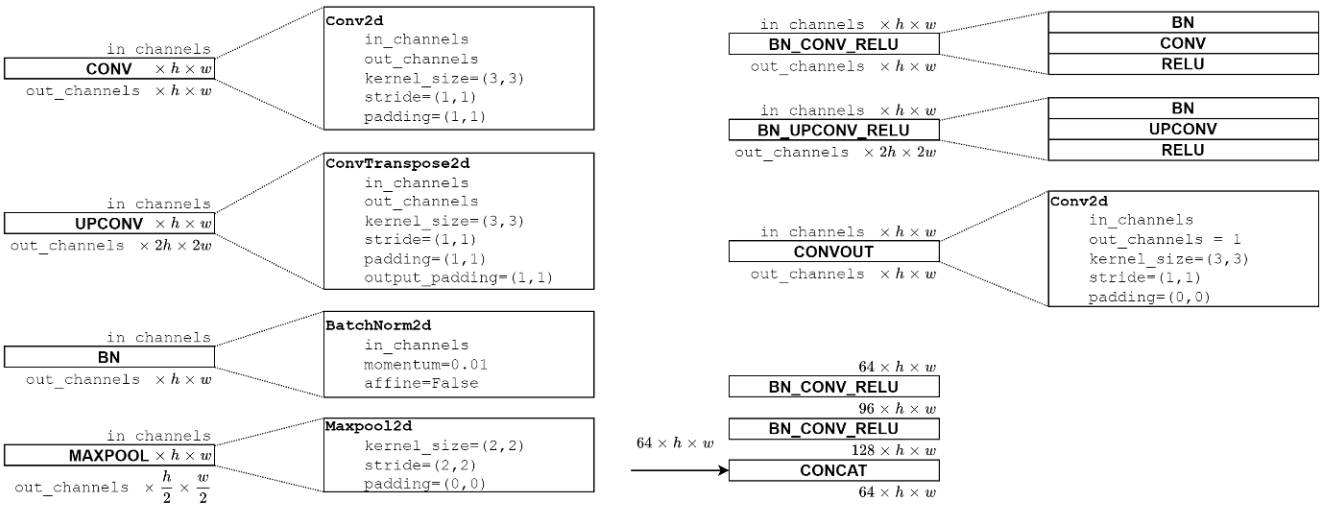


Figure 3.5 Implementation details of the U-Net

3.2.3 U-Net with Late Fusion

The implementation of multimodal feature fusion is adapted from the late fusion (LF) proposed by Maretto et al., (2021). The U-Net encoder is extended in the LF version, as shown in Figure 3.6, where each image is processed by its respective encoder. The feature maps created by both encoders were fused after each convolutional block. Then, the fused feature maps were cropped and copied to concatenate on the corresponding block of the decoder.

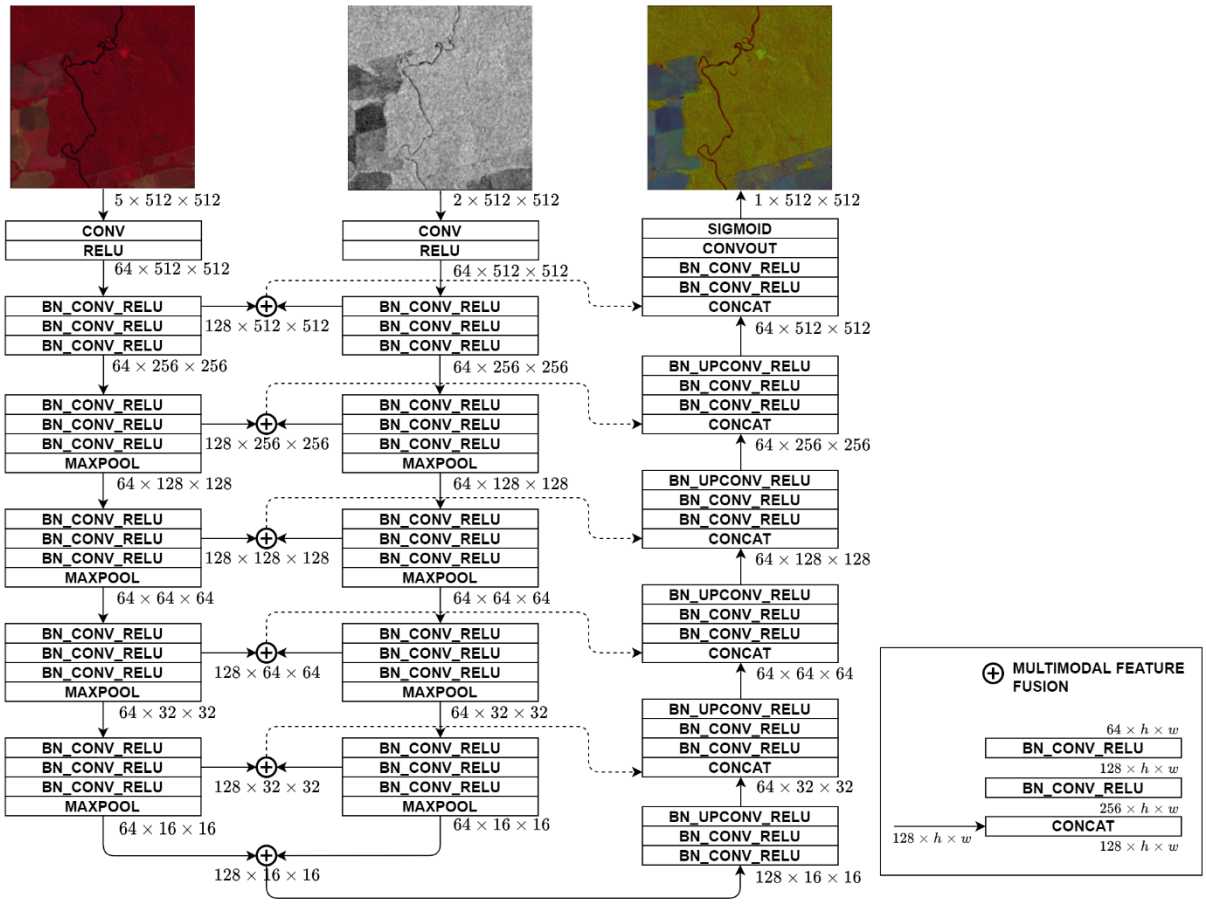


Figure 3.6 U-Net with late feature fusion

3.2.4 Late Fusion with Attention Mechanisms

In a traditional U-Net, skip-connection fuses high semantic but coarse spatial features with corresponding low semantic but finer spatial features. This way, it receives the locational information from the encoder layers and aggregates it with the decoder features to recover the spatial information. However, the multilevel skip connection in the U-Net is prone to receiving unnecessary information from the low-level encoder features (Khanh et al., 2020). Therefore, the network must focus on the low-level salient features of the encoder, representing rich spatial contextual information. Hence, an attention mechanism is used to filter out irrelevant information from the encoder features and overcome the drawback of traditional U-Net architecture.

3.2.4.1 Spatial Attention

A typical skip connection in a U-Net concatenates the encoder and decoder characteristics, wasting computational resources and producing redundant information as the model not always can recognize where an object is located. One way to overcome this issue is to use the most significant spatial features and give them more weight to determine the target object, which is the primary purpose of the spatial attention gate.

As shown in Figure 3.7, the spatial attention gate takes the input feature from the encoder. To compute the spatial attention map, average pooling $F^s_{avg} \in \mathbb{R}^{1\times H\times W}$ and max pooling $F^s_{max} \in \mathbb{R}^{1\times H\times W}$ were applied in the channel dimension of the encoder and concatenated for feature representation purposes. Pooling procedures along the channel axis have been found to help emphasize informative locations (Zagoruyko and Komodakis, 2016). Woo et al., (2018) inspired the use of average pooling and max-pooling. Zhou et al., (2015) suggests using the average-pooling to learn the target of the extent object effectively. Hu et al., (2017) used average-pooling in their attention module to calculate spatial statistics. Additionally, max-pooling was also used because Woo et al., (2018) suggested it helps to generate additional distinctive features which help in inferring finer spatial attention maps. The concatenated feature map is then passed through a large-sized 7×7 convolutional filter for capturing long-range contextual information to produce a spatial attention map $M_s(F)$ for the encoder feature. Finally, the spatial attention map $M_s(F)$ was computed by applying a sigmoid function on the sum.

$$\begin{aligned} M_s(F) &= f_1^{7\times7}([AvgPool(F), MaxPool(F)]) \\ &= \sigma\left(f^{7\times7}\left([(F)^s_{avg}, (F)^s_{max}]\right)\right) \end{aligned} \tag{3.1}$$

where, $f^{a\times a}$ denotes $a \times a$ filter size convolution operation, and σ denotes the sigmoid function.

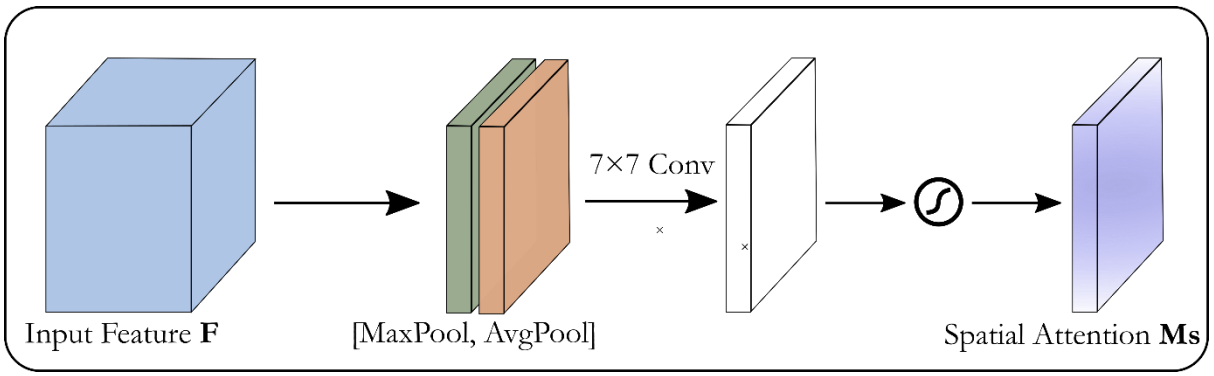


Figure 3.7 Inside the spatial attention gate.

3.2.5 Spatial Attention on Optical Image

In this variation of the LF, spatial attention was used on top of the convolution block of each encoder level (Figure 3.8). The proposed spatial attention module attempts to achieve one main target: to give more weight to deforested areas and less weight to cloudy optical images. So, the spatial attention gates feed the decoder with more relevant information instead of plain old skip connection from the low-level encoder feature with unnecessary information.

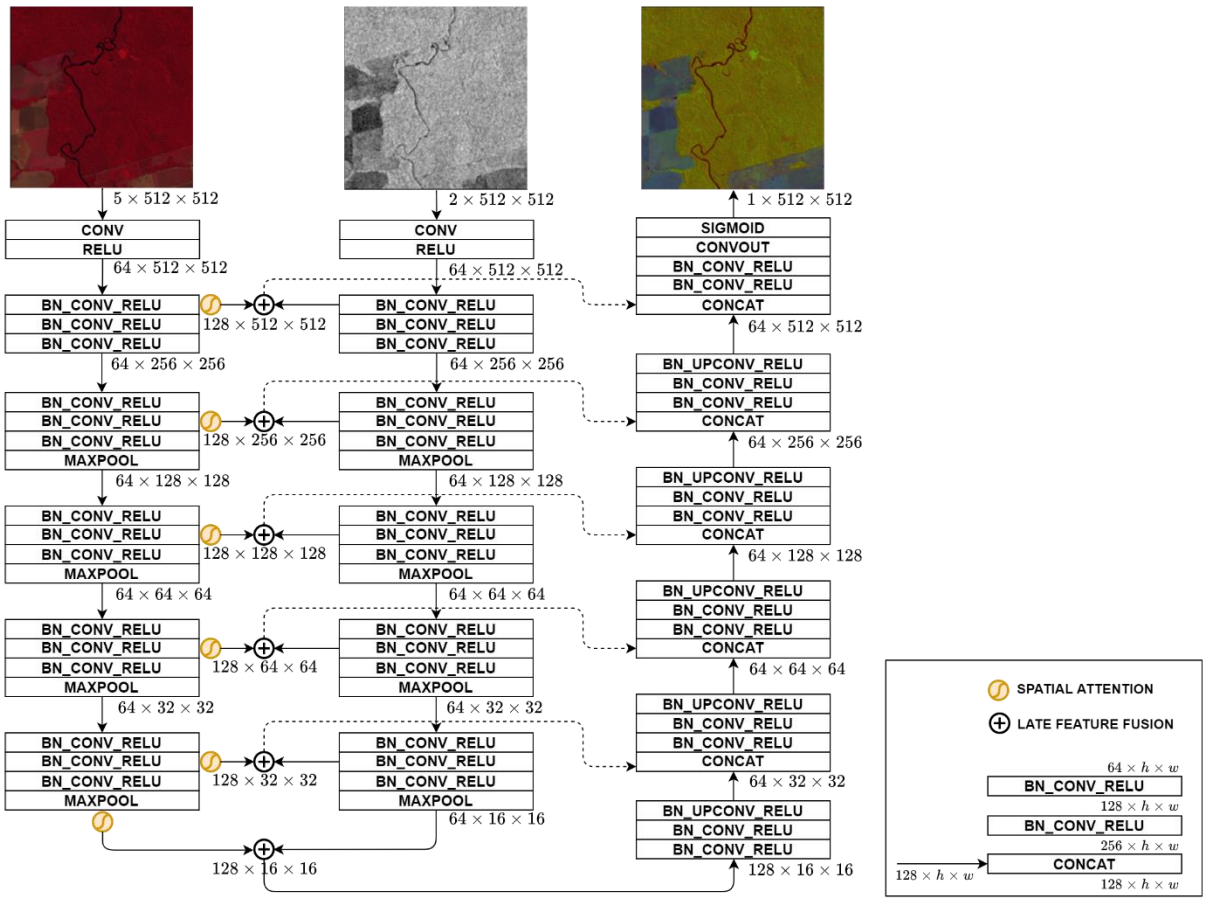


Figure 3.8 LF U-Net with spatial attention on the optical image.

3.2.6 Spatial Attention on Optical and SAR Image

In addition to using spatial attention only on the optical image, this model also adds spatial attention to the SAR image. As can be observed from the network architecture in Figure 3.9, spatial attention was used on two separate occasions. The first attention was applied to the encoder feature of the optical image and fused with the encoder feature of the SAR image. The second spatial attention was applied to these fused features and concatenated with the decoder part of the model. After figuring out a way to pay less attention to the optical images, the model also attempts to give more weight to the essential features of SAR images.

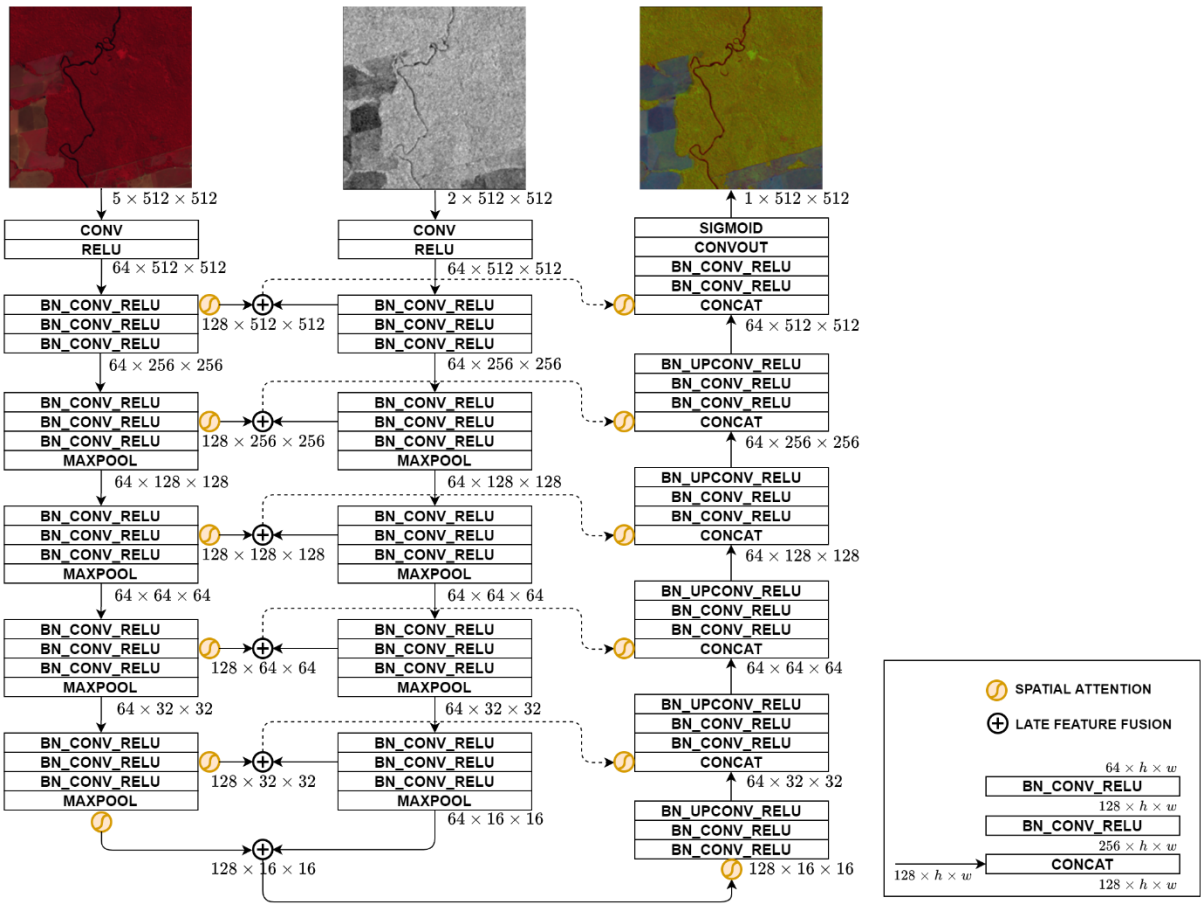


Figure 3.9 LF U-Net with spatial attention on both input data

3.3 LOSS FUNCTION

3.3.1 Binary Cross-entropy

Cross-entropy is defined as the difference between two probability distributions for a given random variable or sequence of events. It is widely used for classification objectives, and since segmentation involves pixel-level classification, it performs effectively (Jadon, 2020).

Binary Cross-Entropy is defined as:

$$L_{BCE}(y, \hat{y}) = -(y log(\hat{y}) + (1 - y) log(1 - \hat{y})) \tag{3.2}$$

where y is the actual value and $\hat{y}$ is the predicted result.

3.4 MODEL PERFORMANCE METRICS

3.4.1 Overall Accuracy

Model overall accuracy calculated using Eq. (3.1) is one of the performance metrics used to evaluate the model. The overall accuracy of a classifier was used in two different scenarios. (a) During the training of the model, monitor its performance. Furthermore, (b) during the prediction phase where the adaptability of the different trained models was evaluated.

3.4.2 Precision, Recall and F1 Score

Apart from the overall accuracy, confusion matrices resulting from all the classifier's variations were also investigated. Using Eq. (3.3), we can calculate the precision and recall of each class using the confusion matrix:

$$precision = \frac{tp}{tp+fp} \qquad recall = \frac{tp}{tp+fn} \tag{3.3}$$

$$Accuraccy = \frac{tp + fn}{tp + tn + fp + fn} \tag{3.4}$$

where, tp is the total number of true positives, fp is the number of false positives, and fn is the number of false negatives. Precision indicates the proportion of deforestation areas correctly identified by the classifier. Recall indicates the proportion of deforested areas in the reference data correctly identified by the classifier. A harmonic mean of precision and recall parameters is used to calculate the F1 score in Eq. (3.5) (Flach and Kull, 2015).

$$F1\ score = 2 \cdot \frac{precision \cdot recall}{precision + recall} \tag{3.5}$$

3.4.3 Intersection over Union (Jaccard index)

Proposed initially by Jaccard, (1901), the Jaccard index is an accuracy metric that measures the overlap between two sample sets. It is most commonly used for the task of semantic segmentation. For any two finite sets, A and B, the Jaccard index is defined as the ratio of the size of the intersection over union (IoU), as given in Eq. (3.6):

$$J(A, B) = \frac{|A \cap B|}{|A \cup B|} \tag{3.6}$$

The Jaccard index ranges from [0,1], where 0 indicates no overlap between the target and predicted sample, and 1 implies a complete overlap.

Chapter 4

Design of Experiments

This chapter provides finer details about the necessary setup to answer the research question about the need for a multimodal fusion.

The experiments were set up in two different scenarios. Scenario-1 is designed considering the sub-objective 3 in section 1.5.1 to evaluate the performance of fusion models in images with different cloud cover and their impact on the accuracy. Based on the result of scenario-1, scenario-2 was designed to achieve sub-objective 4 in section 1.5.1. The sub-objective is to train a model that can be implemented in a real-world scenario to detect deforestation irrespective of any atmospheric condition.

4.1 IMAGE PAIRS

Before setting up the experiments for each scenario, a common table containing images from both sensors was created (Table 4.1). Images from this common table were then used in a different setup for different scenarios.

The table contains images from S-1 and S-2 for the year 2017. Images from S-2 were classified into five different cloudy conditions inside the table. Cloud cover percentage information on the metadata provided with all the S-2 images was used to categorize the images. Five images for each category were chosen for each cloudy scenario. Only four images were chosen for the 60-80% cloudy scenario due to the unavailability of data during 2017. To form the image pair for a corresponding S-2 image, we choose the next closest acquisition date available for the S-1 image. A summary of all the 24 image pairs and their acquisition date is shown in Table 4.1.

Table 4.1 Image pair used for different cloudy scenarios.

Cloudy Scenarios	Image ID	S-2 Image	S-1 Image
No cloud	im1	26-06-2017	25-06-2017
	im2	06-07-2017	13-07-2017
	im3	16-07-2017	13-07-2017
	im4	21-07-2017	19-07-2017
	im5	31-07-2017	31-07-2017
20-40%	im1	28-12-2017	22-12-2017
	im2	29-10-2017	29-10-2017
	im3	03-11-2017	04-11-2017
	im4	29-09-2017	29-09-2017
	im5	08-03-2017	09-03-2017
40-60%	im1	28-03-2017	02-04-2017
	im2	07-04-2017	02-04-2017
	im3	27-05-2017	01-06-2017
	im4	24-09-2017	29-09-2017
	im5	03-12-2017	28-11-2017
60-80%	im1	13-12-2017	10-12-2017
	im2	26-02-2017	25-02-2017
	im3	06-02-2017	01-02-2017
	im4	03-12-2017	28-11-2017

	im1	07-01-2017	08-01-2017
	im2	27-01-2017	01-02-2017
Above 80%	im3	16-02-2017	13-02-2017
	im4	18-03-2017	21-03-2017
	im5	18-11-2017	16-11-2017

4.2 SCENARIO I: TRAIN & TEST ON THE SAME IMAGE

Scenario-1 aims to evaluate the performance effects of combining the two data to detect deforestation in different cloudy scenarios. The following sections explain more information about how the experiment scenario was set up.

4.2.1 Normalisation

Input images were normalised in the range of [0,1] using a minimum-maximum scaling transformation. Separate minimum and maximum values of each individual image were used to transform between zero and one.

4.2.2 Sampling

All the images were split into training, testing, and validation tiles to prepare training data for the model. First, the entire dataset of input and labels was divided into 25 larger tiles, with 2767 × 2810 pixels each (Figure 4.1). Each of the 25 tiles was then split in a ratio of 60:20:20 to get the training, testing, and validation sets. There were 15 training tiles, 5 test tiles and five validation tiles. This split was done randomly in NumPy[7] with a fixed random seed set to 4327. Different random seeds were randomly chosen before fixing it to 4327; this ensured that the training and testing samples were not spatially correlated. The spatial distribution of the train, test, and validation sets is shown by overlaying with a colour code in Figure 4.2.

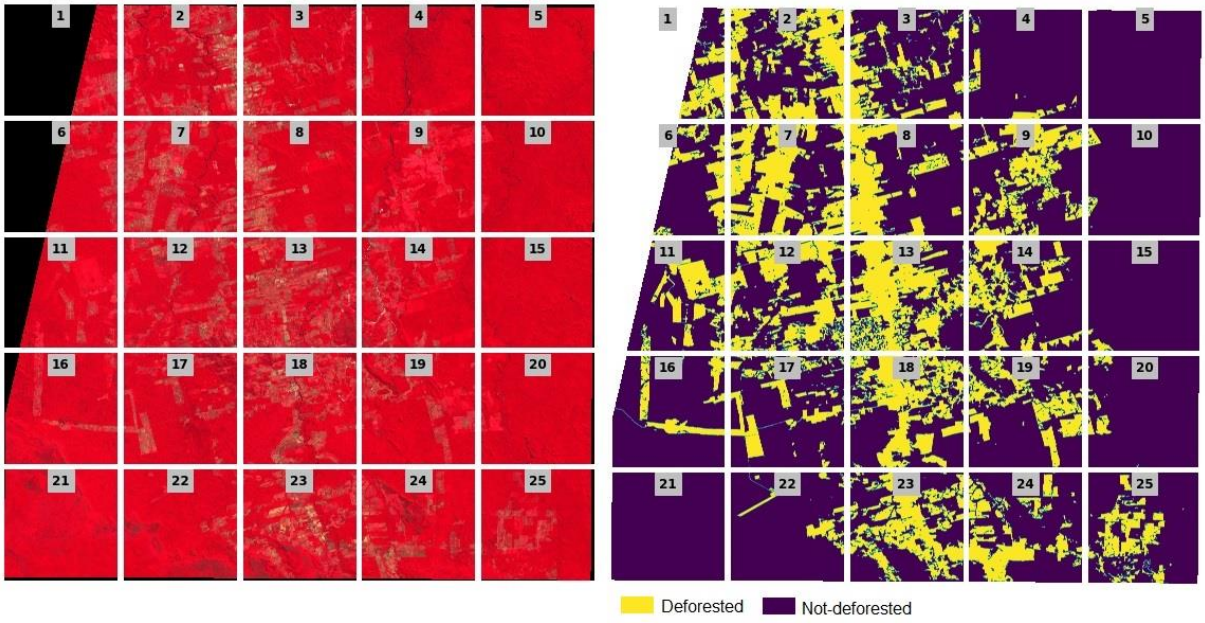


Figure 4.1 Initial splitting of the entire image into a 5×5 tile for a Sentinel-S2 image (on the left) and the corresponding label (on the right).

[7] https://numpy.org/

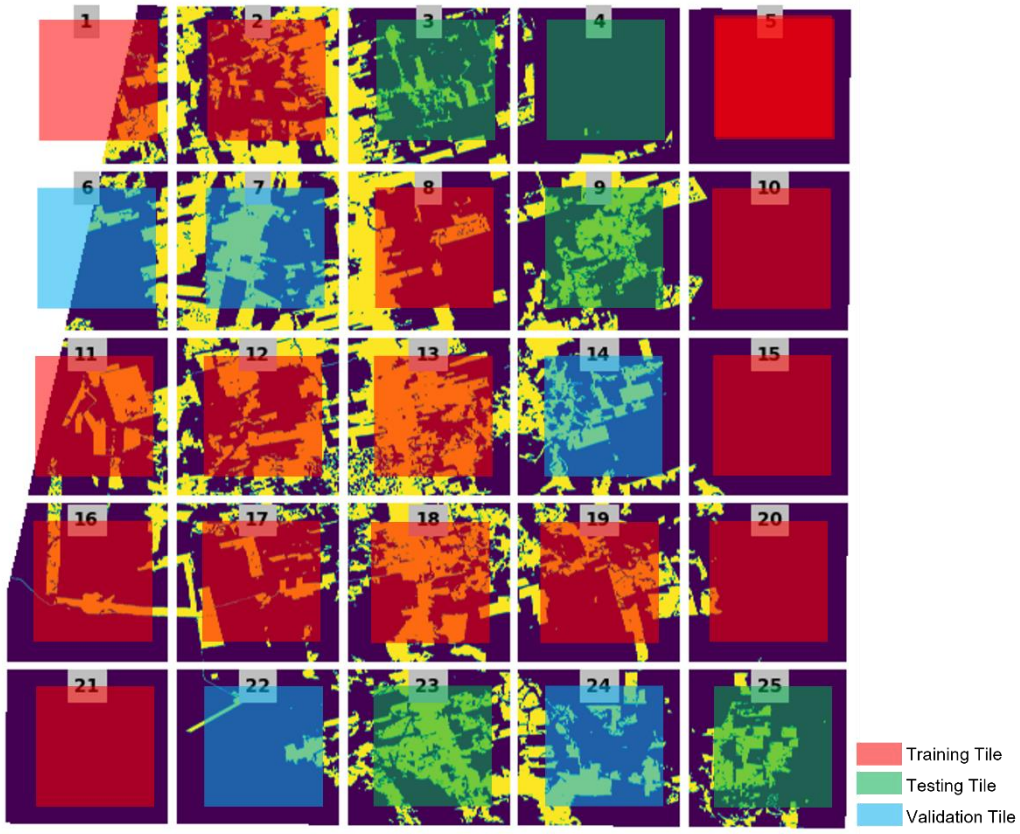


Figure 4.2 Training, testing and validation split of the entire study area

4.2.3 Generating Patches

All the variations of the model used in this study had an input size of 512×512 pixels. The training, testing and validation set tiles were further split into smaller patches of this size. While generating the patches, all patches containing NoData or null values were discarded from the input data to keep the training inputs simpler. Overlapping was used in the last row and column to avoid wasting input data information (Figure 4.3). In the end, each image pair had 448 training patches, 149 validation patches and 152 testing patches.

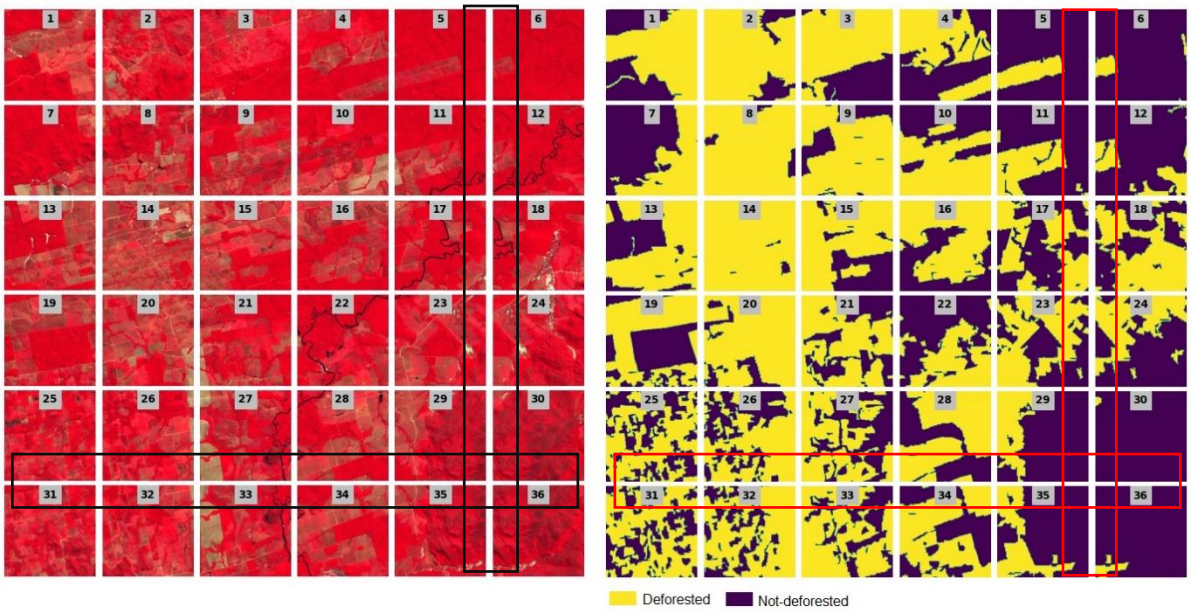


Figure 4.3 Generated patch (512×512) from an example tile (2767× 2810) with the highlighted box showing the overlapping part.

4.2.4 Experiments

Initial experiments were performed by training the U-Net with individual S-1 and S-2 images as input. U-Net with three layers of encoder and decoder block (U-Net3) and U-Net with two layers of encoder and decoder block (U-Net2) were experimented with. A total of 24 S-1 images and 24 S-2 images were used to train the network. The next step was to combine the two input images through the EF U-Net model discussed in section 3.2.2. Therefore, the stacked input of the S1+S2 ([VV, VH, NIR, R, G, B, Cloud Mask]) image was used as the input to this model. EF of the two images uses the same underlying architecture as U-Net with only a difference in the input shape of combined channels from the S-1 and S-2 images.

The next iteration of the experiment used three different variations of LF architecture on the 24-image pairs, resulting in 72 runs. Except for the first LF architecture, the other two used spatial attention mechanisms to combine the input data. An overview of all these experiments is summarized in Table 4.2.

Table 4.2 Setup of image pairs as input to all different variations of the model.

Model:	**U-Net3**			**U-Net3 LF**	**U-Net3 LF Spatial attention on S-2**	**U-Net3 LF Spatial attention on both**
Input	**S-1**	**S-2**	**Stacked(S1+S2)**	**[S-1, S-2]**	**[S-1, S-2]**	**[S-1, S-2]**
S2-Cloud %	**Number of runs per cloud scenario**					
0	5	5	5	5	5	5
20-40	5	5	5	5	5	5
40-60	5	5	5	5	5	5
60-80	4	4	4	4	4	4
>80	5	5	5	5	5	5
Total	**144**					

4.2.5 Implementation Details

All the experiments in this research were implemented using the TensorFlow-Keras[8] framework. A 16-gigabyte NVIDIA RTX A4000 graphics card hosted on the geospatial computing platform[9] (CRIB) servers was used for GPU-assisted computing. A summary of the technical specification for the used server is shown in Table 4.3.

Table 4.3 Specification of the server unit used for model training

Unit	PowerEdge R730
Architecture	Intel x86-64
CPU	E5-2695 v4
Max Speed (GHz)	3.3
Cores	2 × 18
Thread	72
Memory (GB)	768
GPU	NVIDIA RTX A4000 (CC 8.6)

[8] https://keras.io/
[9] https://crib.utwente.nl/

Some initial experiments were performed with one single cloud-free image pair as an input to optimize the hyperparameters of the models. The initial experiment was conducted with different variations of network depth and different loss functions. In this phase, there were minor improvements in the overall accuracy when using U-Net3 over U-Net2. Therefore, only U-Net3 were chosen to move forward with the training.

A summary of all the hyper-parameters is presented in above Table 4.4. All U-Net variations used in this scenario were trained with a U-Net3. Because of limited GPU memory, it was impossible to use a batch size larger than two. Adaptive moment estimation (Adam) optimiser (Kingma and Ba, 2014) with a learning rate of 0.00007 was used to train the network. The maximum number of epochs for each model was set to 10000, and model callbacks like early-stopping and model checkpoint were used to train and restore the best model weights. The callbacks monitor the validation IoU with the patience of 30 epochs before early-stopping the training and restoring the model with the best weights. In this context, the patience of 30 epochs implies that the model training will terminate only if the chosen performance (Validation IoU in this study) metrics do not improve for 30 epochs in a row.

Table 4.4 Summary of hyperparameters used.

Hyper parameters	Value
Network depth	3
Optimizer	Adam
Learning rate	0.00007
Early stopping patience	30
Max number of epochs	10000
Batch size	2
Loss function	BCE

Total trainable parameters and time for each model to complete each training epoch are presented in Table 4.5. As the number of parameters increases, the model takes more time to complete each epoch of training, consequently increasing the total training time.

Table 4.5 Model trainable parameters and training time for one epoch during scenario-1

Model	No. of trainable params	Train time per epoch
U-Net3 S-1	1,297,282	40 $\pm$ 3 seconds
U-Net3 S-2	1,299,010	43 $\pm$ 3 seconds
U-Net3 EF	1,300,162	47 $\pm$ 2 seconds
U-Net3 LF	2,372,610	55 $\pm$ 3 seconds
Spatial attention on optical	2,373,100	68 $\pm$ 3 seconds
Spatial attention on both	2,373,590	83 $\pm$ 4 seconds

4.3 SCENARIO II: TRAIN & TEST ON DIFFERENT IMAGES

In this experiment scenario, the main objective was to develop a model that can be used in a real-world situation, where the model should be able to generalise unseen data from any date or under any cloudy conditions. Therefore, the training data preparation and implementation details differ from that used in the previous scenario.

4.3.1 Image Pair

Out of 24 image pairs, ten were used for training, 5 for validation and 5 for testing. Image pairs with id 'im2' (Table 4.1) from each of the five cloudy scenarios were used for training. Image pair id 'im4' were used for testing, and image pair id 'im3' of the 60-80% cloud category plus 'im5' of the rest four cloud categories were used for validation.

4.3.2 Normalization

Global minimum and maximum values for optical data and another global minimum and maximum values for SAR data were used separately to normalize all the training, testing, and validation image in the range of [0,1]. This normalization is done to keep the same transformation across all the images. Otherwise, it may confuse the model with a different range of numbers for each image and make it unable to converge (Singh and Singh, 2020). Figure 4.4 shows the normalized histograms of optical data in different cloud scenarios. And an example of a normalized histogram of the S-1 image is shown in Figure 4.5.

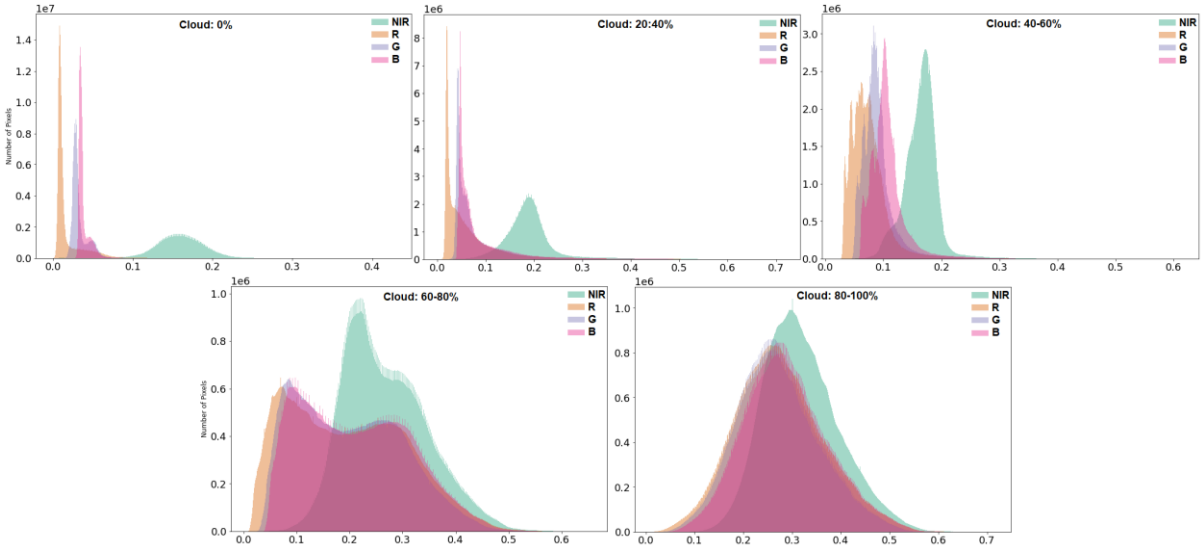


Figure 4.4 Normalized histogram of S-2 images at different cloudy conditions

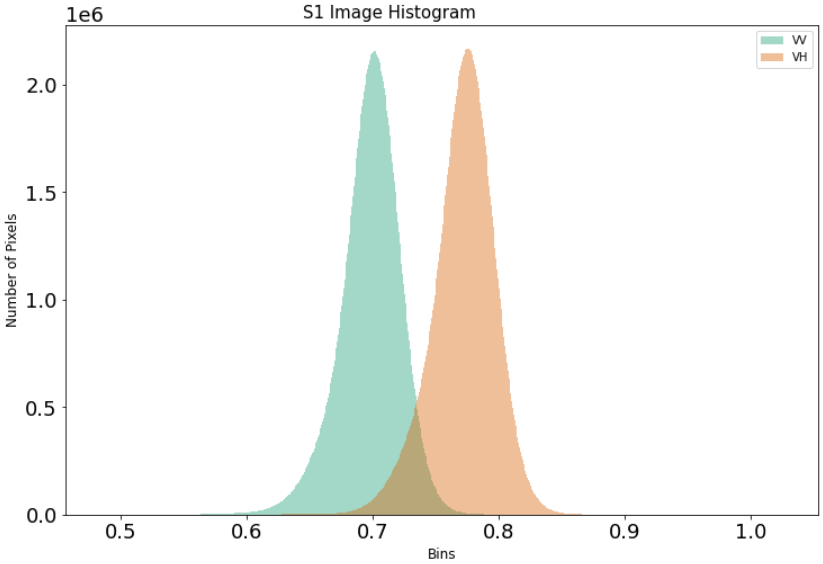


Figure 4.5 Example of a normalized histogram of S-1 image.

4.3.3 Generating Patches

Before start generating the patches, the whole image was trimmed down from all four sides to eliminate all the NoData values. After that, training and validation patches with 512×512 pixels were extracted. The whole spatial extent was used to extract patches as training, and testing data were split over different image pairs. Patches from the all the images were concatenated on the batch axis to create a total of 6210 training patches with a size of 512×512×5 for optical data and 512×512×2 for SAR data. For the validation image, there was a total of 3105 patches. Patches from the test image pair were extracted on the fly to perform prediction and then reconstructed to create a large classification map for the entire extent.

Training and validation samples were wrapped in a data generator using the 'Sequence' class of Keras to avoid loading all the images into GPU memory. In this way, the network only loads the specified number of batch sizes at a time onto the GPU memory. The sequence of the patches was shuffled at the end of each epoch to ensure that the model does not receive the batches in a similar pattern and hence avoid overfitting.

4.3.4 Experiments and Implementation

Due to time constraints and the large training and validation data size, a lighter variation of U-Net2 architecture was used to carry out the experiments. Optical and SAR data were trained as standalone input and combined. The combined experiments use the simpler U-Net2 LF. Adam optimiser with a learning rate scheduler that reduces at an exponential rate was used. The initial learning rate was set to 0.01 with a decay rate of 0.9 per epoch.

A summary of hyperparameters used for the LF U-Net2 and base U-Net2 is presented in Table 4.6, and the time taken for each epoch and number of trainable parameters is presented in Table 4.7

Table 4.6 Summary of hyperparameters for U-Net2 and U-Net2 with LF.

Hyper parameters	Value
Network depth	2
Optimizer	Adam
Learning rate	0.01
Decay rate	0.9
Early stopping patience	35
Max number of epochs	10000
Batch size	2
Loss function	BCE

Table 4.7 Model trainable parameters and training time for one epoch for scenario-II

Model	Input	No. of trainable params	Train time per epoch
U-Net2	S1	985,570	550 ± 5 seconds
U-Net2	S2	985,570	560 ± 5 seconds
U-Net2 LF	[S1, S2]	1,822,338	890 ± 6 seconds

Chapter 5

Result

5.1 SCENARIO I: TRAIN & TEST ON THE SAME IMAGE

Scenario-1 presents the performance metrics and classified map combining the two data to detect deforestation in different cloudy scenarios. The models in this scenario were trained and evaluated using the different spatial locations of the same image.

5.1.1 Individual Image & Early Fusion

The individual image of an image pair and their EF were trained with a base U-Net3 to assess the performance in different cloudy scenarios.

5.1.1.1 Performance Metrics

The F1 score in all the tables presented is the accuracy for only the deforested class and not a combined overall F1 score of all the classes. The F1 and IoU values in Table 5.1 indicate an average of all the image pairs for each cloudy scenario. And σF1 and σIoU are the standard deviations of the performance metrics, which indicates how the result within a cloudy scenario the result varies from the mean.

It is noticeable from Table 5.1 that an individual S-2 image has the best and most consistent F1 and IoU scores in a cloudless condition. This consistency is because a standalone optical image has superior feature representation capability compared to a radar image. It also performs better than the early fusion because the simpler data of the optical sensor make it easy for the model to extract features.

Table 5.1 Mean F1 and IoU score of deforested class with the different input images to a U-Net3. The bold numbers indicate the highest accuracy for each cloud scenario among the three inputs.

	S-1				**S-2**				**[S1+S2] Stacked**			
Cloud	F1	IoU	σF1	σIoU	F1	IoU	σ F1	σIoU	F1	IoU	σ F1	σIoU
0%	76.74	62.28	1.66	2.17	**87.75**	**78.18**	0.28	0.44	87.42	77.62	0.42	0.73
20-40%	80.56	67.49	1.97	2.73	72.55	57.42	7.44	8.47	**80.72**	**67.91**	4.59	6.23
40-60%	79.67	66.27	2.39	3.31	71.53	55.87	4.54	5.39	**80.52**	**67.44**	2.06	2.90
60-80%	79.65	66.23	2.11	2.87	41.26	28.32	20.28	18.31	**80.79**	**67.79**	1.36	1.91
>80%	80.24	67.13	0.63	0.87	16.09	9.63	16.12	10.19	**81.26**	**68.50**	2.27	3.22

5.1.1.2 Qualitative Analysis

When we see the classified map in Figure 5.1, the optical image was able to classify the deforested areas much better than the standalone S-1 image. The S-1 image mostly under-estimated the deforested areas. This underestimation is due to the inability of a radar image to distinguish between a backscatter intensity of forest and deforested area with regrowth of smaller vegetation (Durieux et al., 2019). Also, as discussed in section 3.1.3, most of the deforested areas were mapped in the earlier years and, therefore, could undergo an initial reforestation phase, making it difficult for the model to differentiate. An example of this phenomenon is discussed below in section 5.2.2.

The accuracy of the classified map dramatically changes when clouds are present in the image. However, the result from the radar image remains stable since those are not affected by clouds. Figure 5.2 shows the result of the classified map using different inputs at 40-60% cloud cover. The poor performance of using an optical image with 100% cloud is also evident in Figure 5.3.

When the input was changed from individual images to a stacked input of image pair, it was noticed that the model could avoid clouds from the S-2 image and extract representative features from the S-1 image. The accuracy score when using stacked input reflects that the model maintained a consistent result irrespective of the cloud cover in the S-2 image.

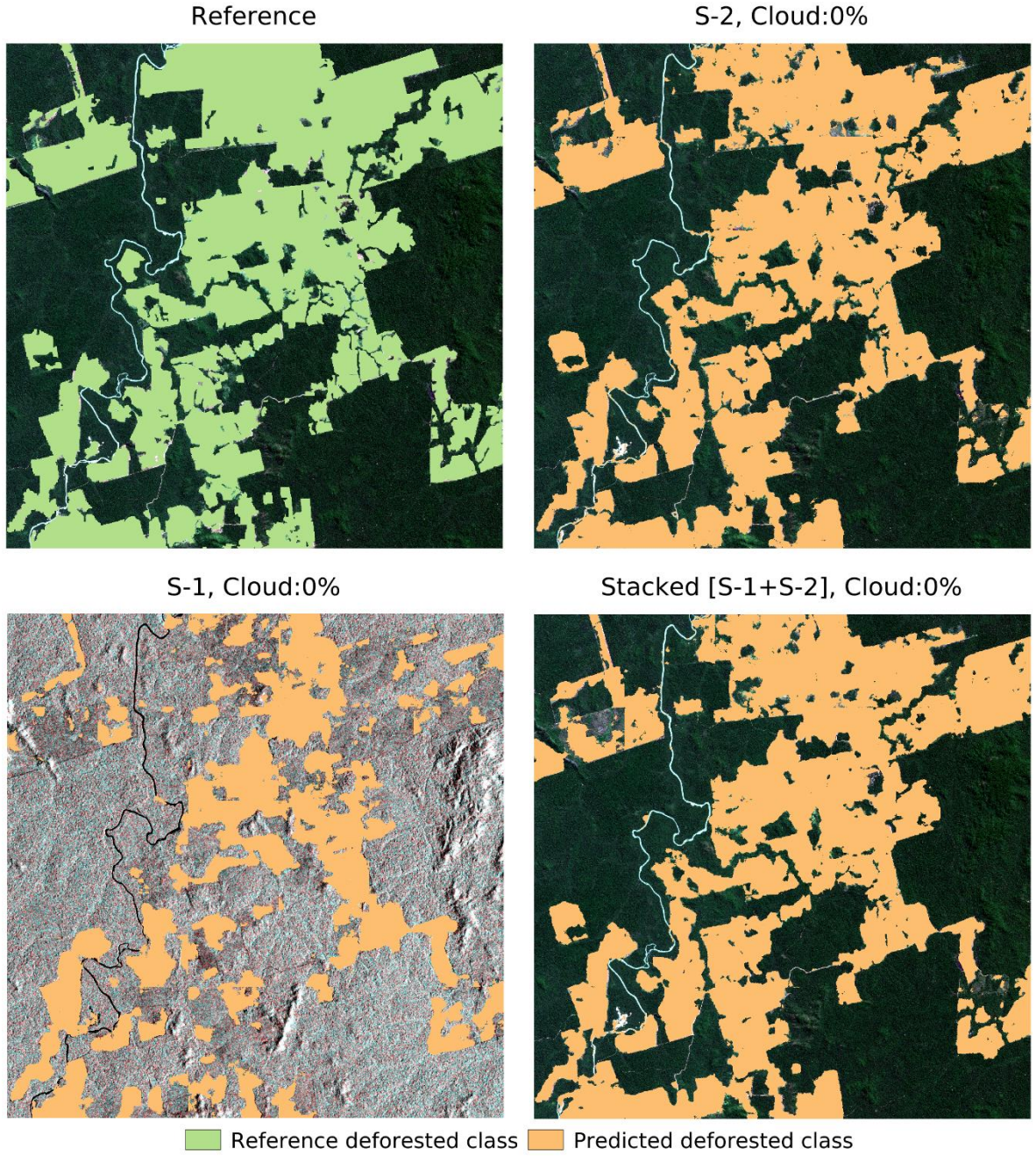


Figure 5.1 Classified map of test tile 9 using U-Net3. Top-left image shows the reference label superimposed on S-2 image, and the others show the classification maps with different input images at zero percent cloud scenario overlaid on their input S-1 and S-2 image.

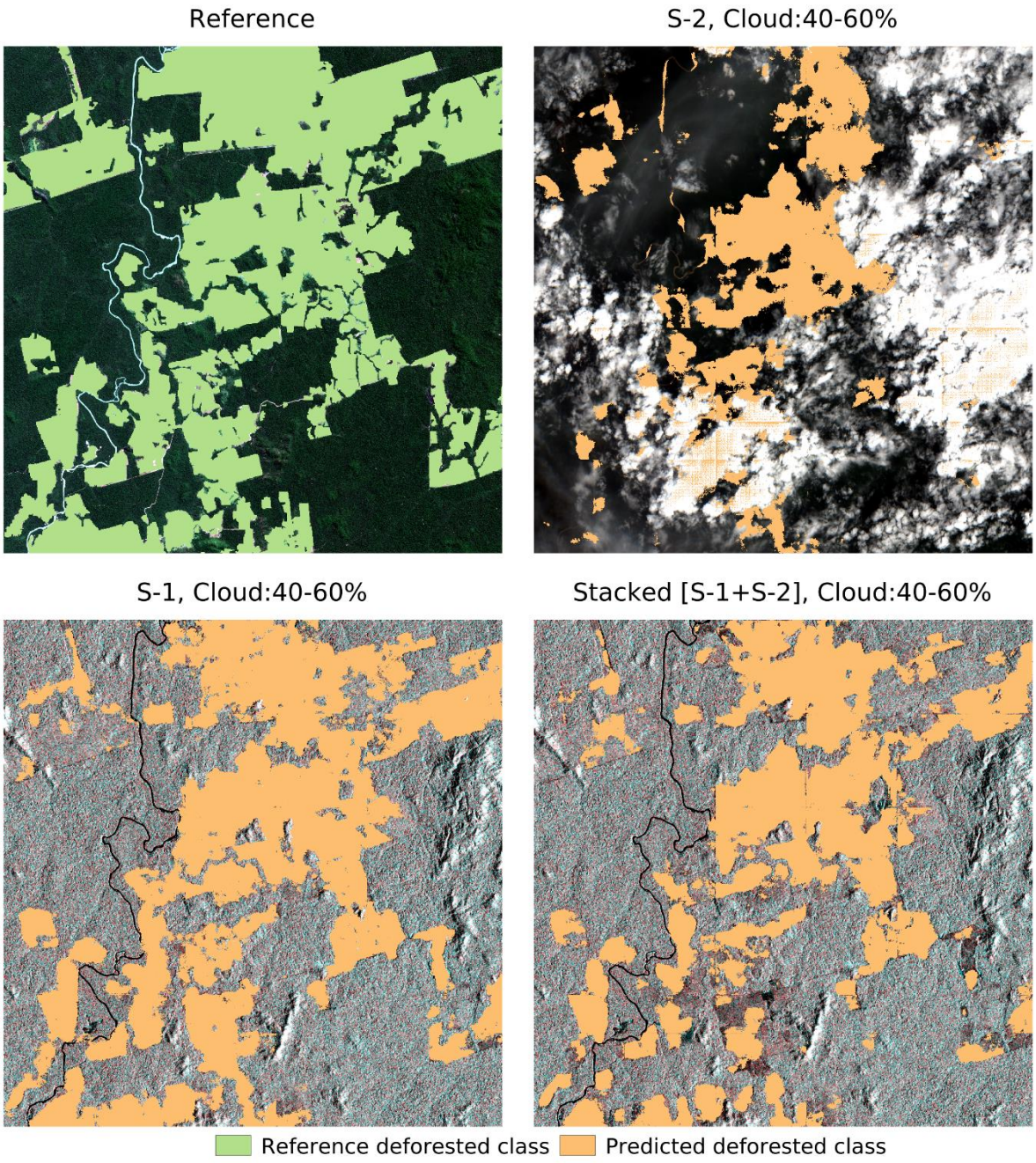


Figure 5.2 Classified map of test tile 9 using U-Net3. Top-left image shows the reference label superimposed on S-2 image, and the rest shows the classification map using different input images at 40-60% cloud cover overlaid on their input S-1 and S-2 image.

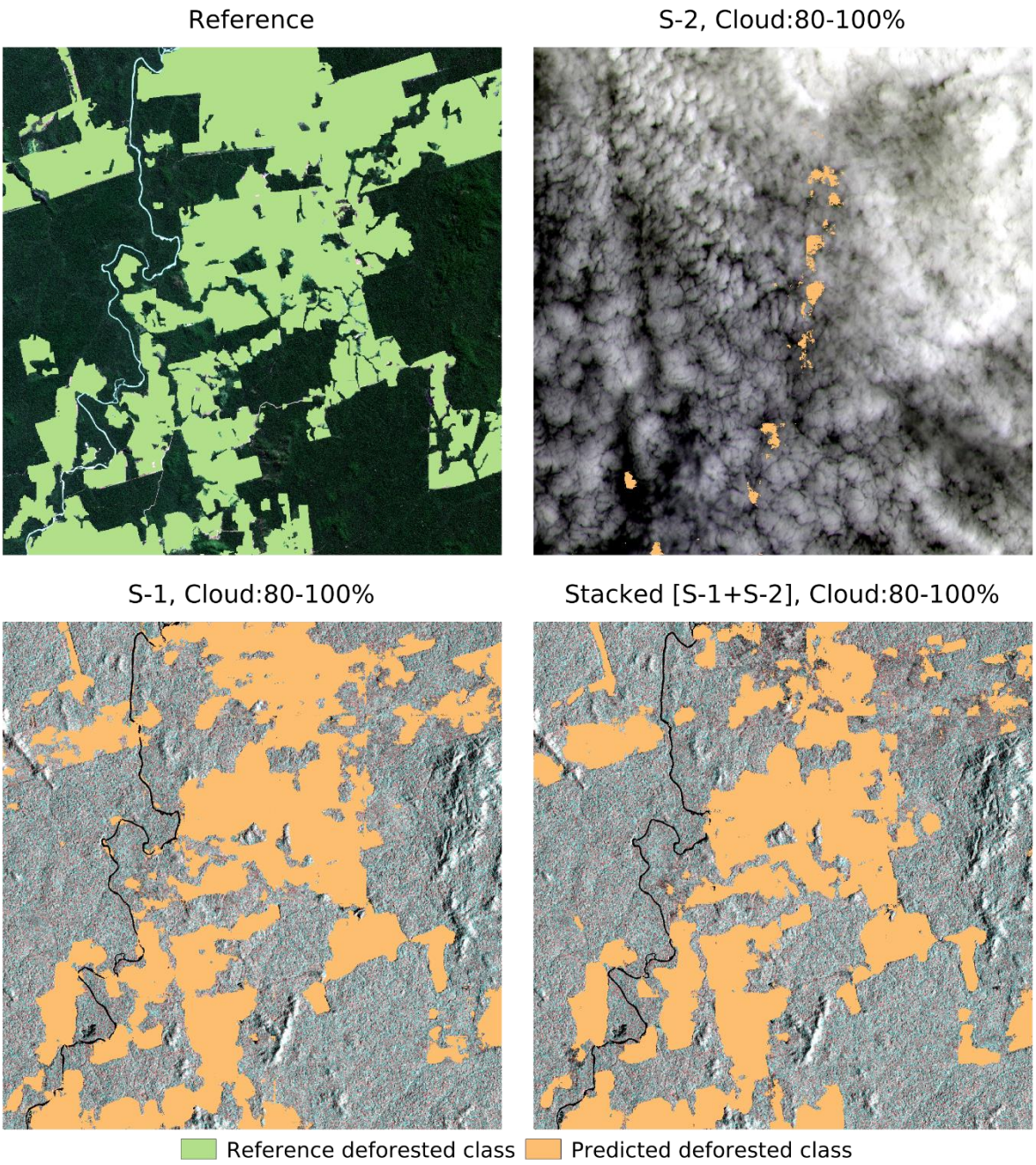


Figure 5.3 Classified map of test tile 9 using U-Net3. Top-left image shows the reference label superimposed on a clear S-2 image, and the rest shows the classification map using different input images at 80-100% cloud cover overlaid on their input S-1 and S-2 image.

5.1.2 Late Fusion

5.1.2.1 Performance Metrics

Compared to the performance metrics using a standalone S-2 image under cloud-free conditions, the accuracy of all the variations of LF U-Net3 was almost identical (Table 5.2). During a cloudy scenario, the LF helped increase the accuracy by using both S-2 and S-1 input. This increase was more pronounced if the result was compared with a standalone S-2 image. Therefore, the result from the LF implies that fusion is necessary to ignore the cloud in the S-2 image automatically.

Table 5.2 Mean F1 and IoU score of deforested class with the input of S-1 and S-2 images to different LF U-Net3

	LF				**Spatial attention on S-2**				**Spatial attention on both**			
Cloud	F1	IoU	σF1	σIoU	F1	IoU	σ F1	σIoU	F1	IoU	σ F1	σIoU
0%	86.15	75.70	1.40	2.14	86.46	76.16	0.70	1.09	**87.17**	**77.26**	0.59	0.93
20-40%	81.54	68.76	1.98	2.97	**81.73**	**69.11**	0.49	0.69	81.63	69.01	1.80	2.57
40-60%	80.82	67.91	2.79	3.87	**81.79**	**69.19**	0.56	0.80	81.20	68.37	0.93	1.32
60-80%	79.50	66.10	3.35	4.49	78.79	65.09	2.74	3.76	**82.32**	**69.97**	1.00	1.45
>80%	79.48	66.18	4.51	6.09	**80.55**	**67.50**	2.54	3.55	79.63	66.22	2.44	3.36

The performance of two spatial attention variations was better in most scenarios than using a standard LF by a small margin. The most notable improvement of using the attention mechanism was for cloud cover between 60-80%, where spatial attention on both inputs improves the F1 score by almost 2%.

An overall summary of all the results comparing fusion models with individual and early fusion is demonstrated in Table 5.3. The comparison shows that the fusion with attention mechanism has the best performance metrics in most cloudy scenarios. And therefore, the attention mechanism has the potential to improve further if the architecture is modified a little and trained with larger training samples.

Table 5.3 Summary of results from all experiments in scenario-1

	U-Net				**U-Net EF**		**LF**		**Att. on S-2**		**Att. on both**	
Input	S-1		S-2		[S-1,S-2]		[S-1,S-2]		[S-1,S-2]		[S-1,S-2]	
Cloud	F1	IoU	F1	IoU	F1	IoU	F1	IoU	F1	IoU	F1	IoU
0%	76.74	62.28	**87.75**	**78.18**	87.42	77.62	86.15	75.70	86.46	76.16	87.17	77.26
20-40%	80.56	67.49	72.55	57.42	80.72	67.91	81.54	68.76	**81.73**	**69.11**	81.63	69.01
40-60%	79.67	66.27	71.53	55.87	80.52	67.44	80.82	67.91	**81.79**	**69.19**	81.20	68.37
60-80%	79.65	66.23	41.26	28.32	80.79	67.79	79.50	66.10	78.79	65.09	**82.32**	**69.97**
>80%	80.24	67.13	16.09	9.63	**81.26**	**68.50**	79.48	66.18	80.55	67.50	79.63	66.22

5.1.2.2 Qualitative Analysis

The classification map in Figure 5.4 and Figure 5.5 displays the output of the LF models in two different cloudy scenarios. At zero percent cloud condition (Figure 5.4), all the models could consistently delineate between the deforested and not-deforested classes. The edges were well-preserved, and most of the deforestation patches were extracted according to the reference data.

Compared to the predicted map from zero per cent cloud, the quality of the classification map reduced slightly during a more cloudy situation (Figure 5.5: Cloud cover: 40-60% and Figure 5.6: Cloud cover 80-100%). In cloudy scenarios, the performance of the LF models emulates the performance of a standalone S-1 image. The reason is that out of an S-1 image and a cloudy S-2 image, the most learnable feature for an

LF model was S-1; therefore, the result looks like the model has automatically output a map that resembles closely like it was classified by just using S-1 image. Hence, all the models output a consistent result using an S-1 radar image during a cloudy scenario. Moreover, if the image is clear, it gives a more accurate result using the feature of the S-2 image.

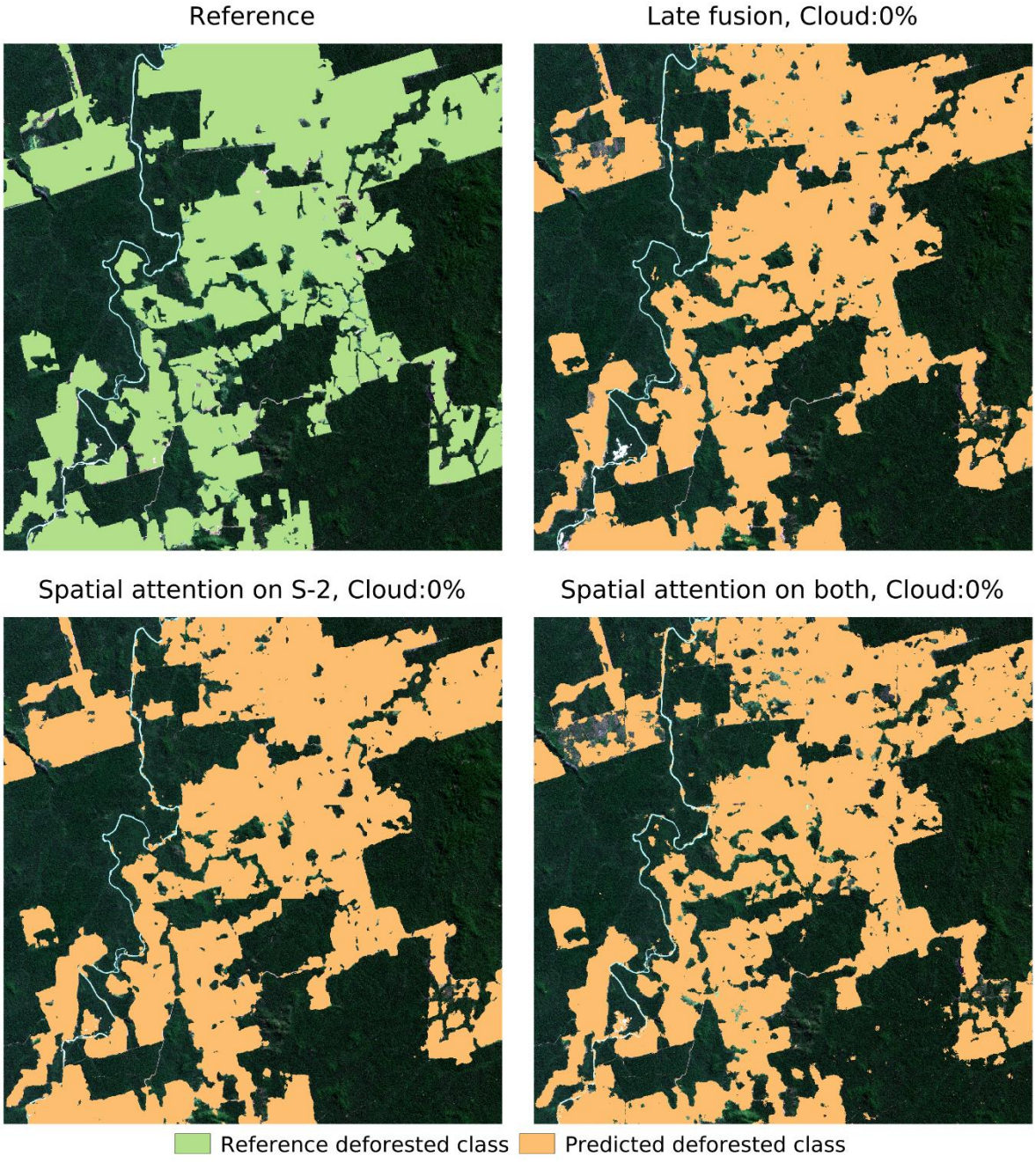


Figure 5.4 Classified map of test tile 9 using different LF U-Net3. Top-left image shows the reference label superimposed on a clear S-2 image, and the rest shows the classification map using different input image pairs at 0% cloud cover overlaid on their input S-2 image.

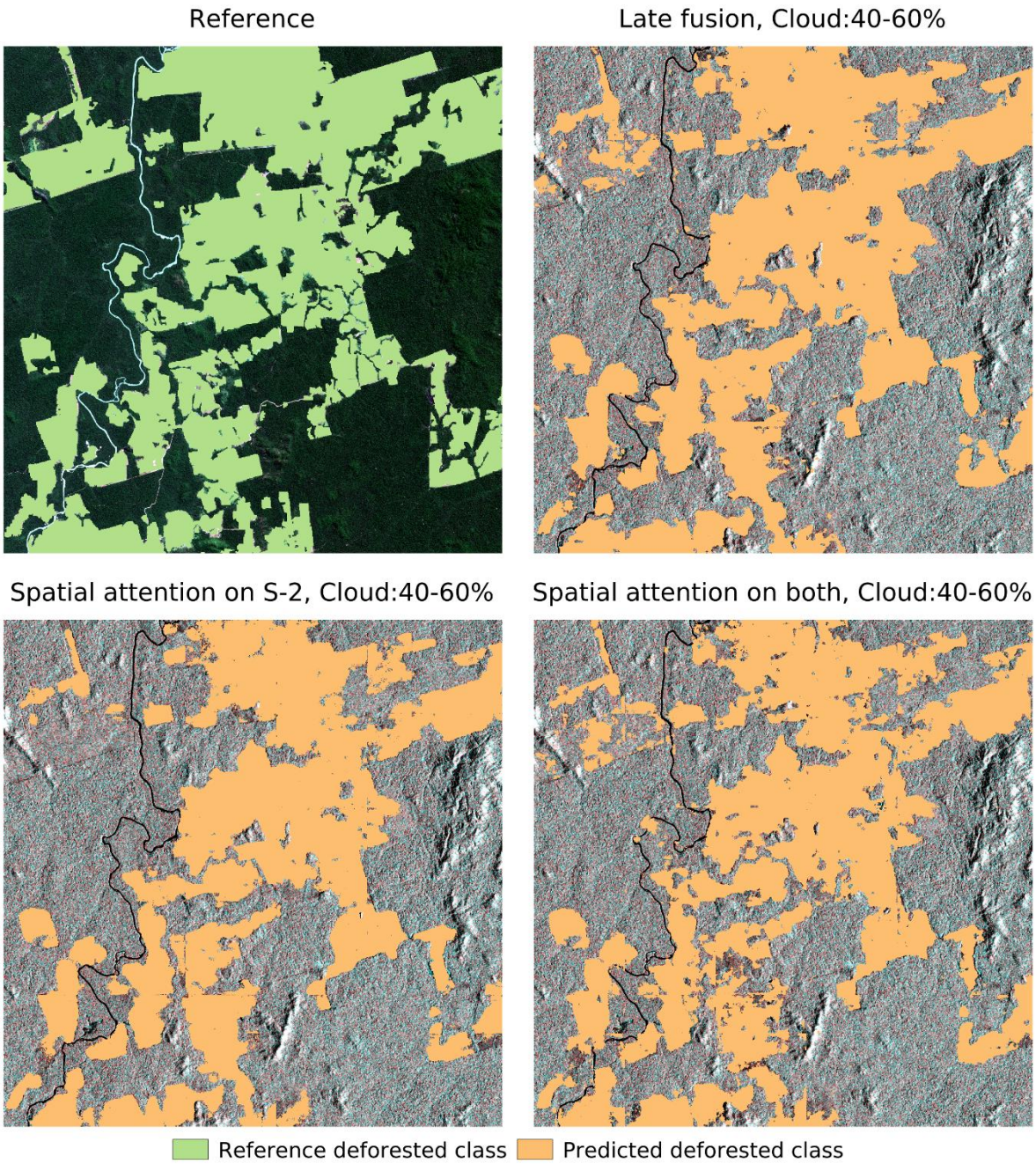


Figure 5.5 Classified map of test tile 9 using different LF U-Net3. Top-left image shows the reference label superimposed on a clear S-2 image, and the rest shows the classification map using different input image pairs at 40-60% cloud cover overlaid on their input S-1 and S-2 image.

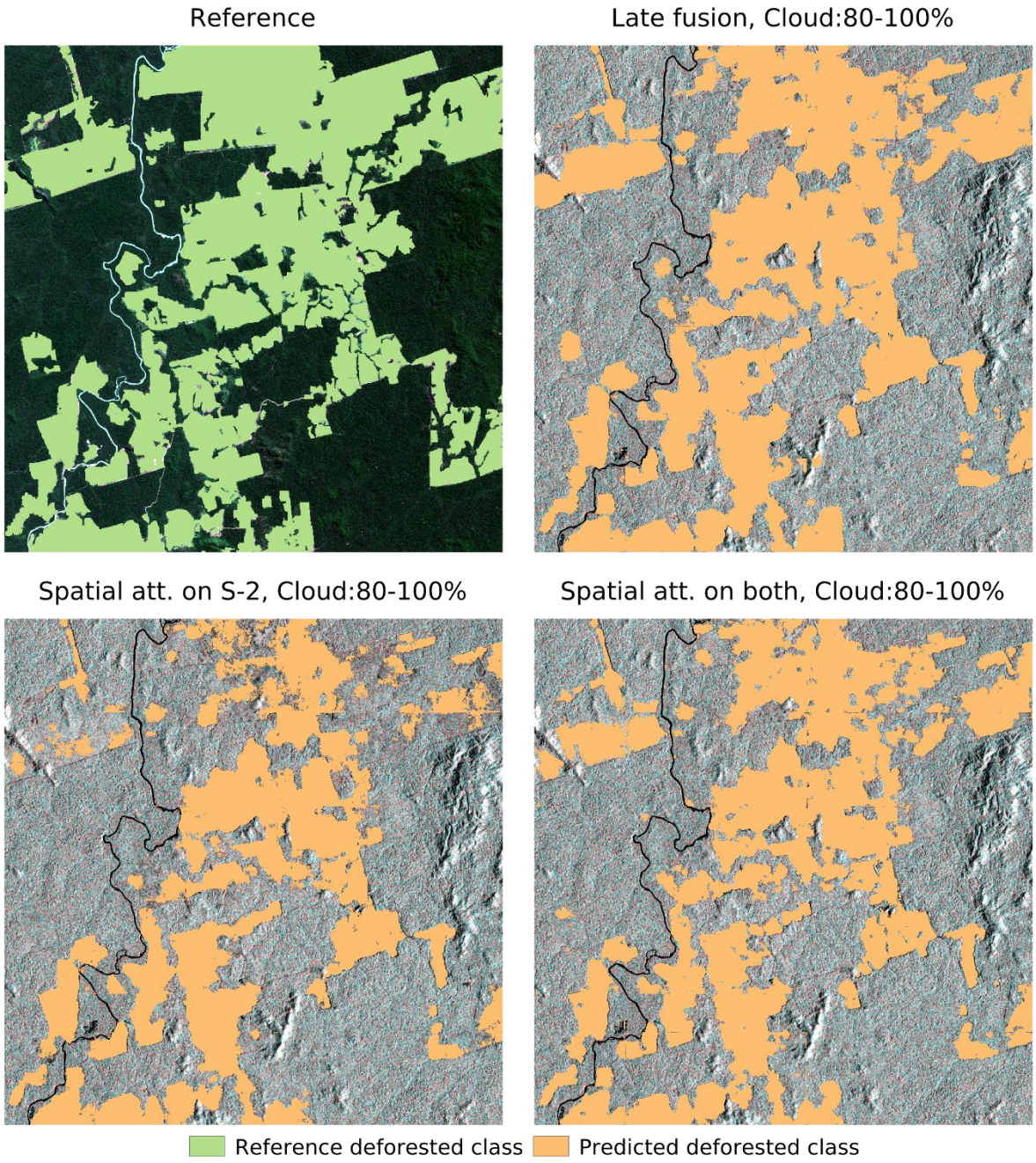


Figure 5.6 Classified map of test tile 9 using different LF U-Net3. Top-left image shows the reference label superimposed on a clear S-2 image, and the rest shows the classification map using different input image pairs at 80-100% cloud cover overlaid on their input S-1 and S-2 image.

5.2 SCENARIO II: TRAIN & TEST ON DIFFERENT IMAGE

This section presents the result from scenario 2 with the primary objective of developing a model that can be used in a real-world situation and be able to generalise unseen data from any date or under any cloudy conditions.

5.2.1 Performance Metrics

A summary of all the results from this scenario is presented in Table 5.4. A reduction in the F1 and IoU scores can be noticed when the results of using individual S-2 images are compared with results from scenario-1 in section 5.1. This reduction is because only 20% of the training images were cloud-free, and 80% were cloudy. As the amount of training data for S-2 was imbalanced with more cloudy images, the model training process for S-2 data could not find an optimal set of weights to generalise both cloudy and cloud-free images.

On the other hand, the results from the experiment using the S-1 input image improved from 80% F1 score in scenario-1 to 85% F1 score in the current scenario. The higher accuracy of S-1 for scenario-2 is mainly because the model had ten times more training samples to learn from and, therefore, could perform better during inference. Therefore, it is evident that the accuracy can be improved by increasing the training samples.

The LF result during the zero per cent cloud scenario did not improve compared to scenario-1. Scenario-1 had an average accuracy of 87.17% F1 score at cloudless condition, whereas it is 85.46 for LF U-Net2 for scenario-2. The leading cause goes back to the same sample imbalance problem discussed above. Another reason could be that the LF needs much more samples to train and generalize than the standalone sensors. The result remains constant for all the other conditions when there was more than 20% of cloud coverage in the S-2 image. In those cases, the model also learned to ignore the clouds in the S-2 image automatically and focus on the S-1 image to classify deforested areas.

Table 5.4 Summary of results using U-Net2 with and without fusion.

	S-2 with U-Net2		**S-1 with U-Net2**		**[S-1, S-2] with LF U-Net2**	
Cloud	F1	IoU	F1	IoU	F1	IoU
No cloud	78.92	65.19	**85.52**	**74.71**	85.46	74.61
20-40%	66.41	49.72	**85.51**	**74.68**	85.50	**74.68**
40-60%	3.90	2.00	84.00	72.41	**84.91**	**73.77**
60-80%	33.77	20.31	**85.85**	**75.20**	85.18	74.19
>80%	3.46	1.76	**85.65**	**74.91**	84.74	73.53

5.2.2 Qualitative Analysis

A close view of the classification quality generated from LF U-Net2 is displayed in Figure 5.7. The output in the figure was generated using a test image pair with 40-60% cloud cover. A more zoomed-in view from Figure 5.8 reveals the superior edge detection capability of the model. The reference image has a spatial resolution of 30m and does not perfectly align if superimposed on top of an input S-2 image. However, the prediction from the model gives a better representation of the deforestation patch than the reference data. It can segment the deforested patches of land from the not-deforested class very well.

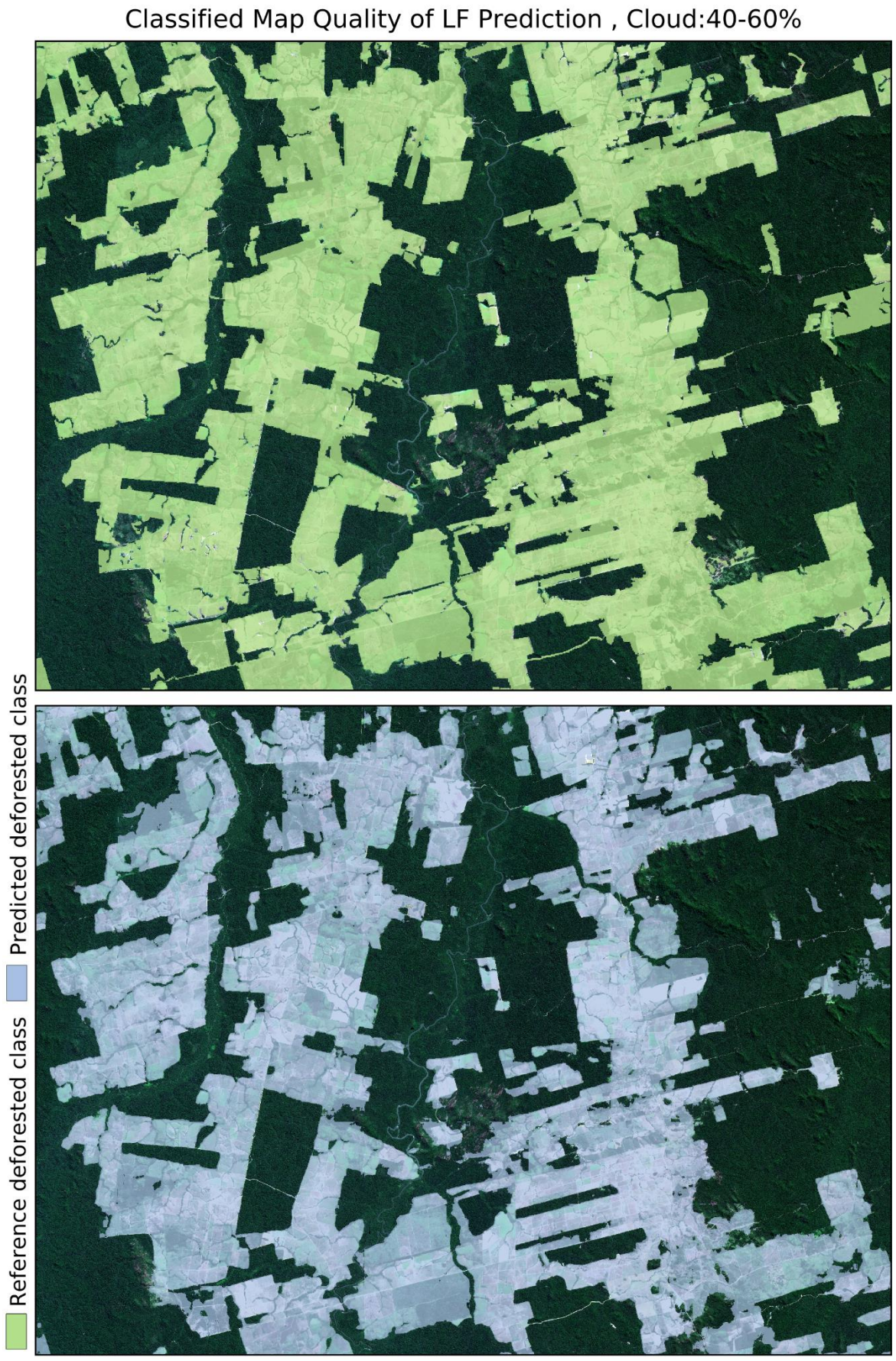


Figure 5.7 Close view of a classified map from LF U-Net2 compared with reference data

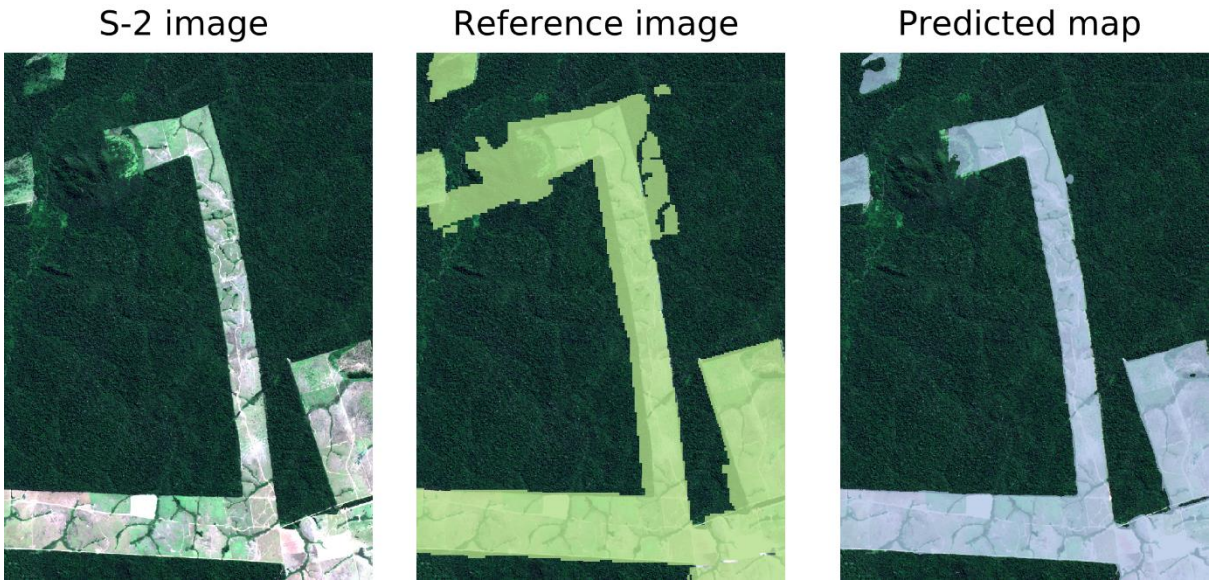


Figure 5.8 Map showing the quality of predicted result compared to reference data when overlayed on top of an input S-2 image.

Even though the model performs adequately in most scenarios, it also has limitations. An example of such shortcomings is depicted in Figure 5.9. The map from the Figure 5.9 shows two types of error that most of the classified map from U-Net2 LF and S-1 with the U-Net2 model produces. The error map is calculated by subtracting the prediction map from the reference map and portrays the distribution of false positives and false negatives over the whole output area. A false positive is where the model incorrectly predicted the region as deforested and is highlighted in pink. False negative is the deforested areas in the reference data that the model could not identify and is highlighted in orange.

The bottom left image, highlighted with a red frame, shows an example of false-negative areas overlaid with orange. According to the reference data, the orange area indicates that it should be a deforested area, but the model could not predict it as deforestation. The areas that could not identify as deforested are mainly due to two underlying reasons. 1) The model struggles to identify an area as deforested if there has been a regrowth of vegetation (further discussed in section 3.1.3). This similarity makes it too difficult for the model to identify these regrowth areas. Although they are reforested, they are not primary forests anymore and, for that reason, are labelled as deforestation in the reference data. 2) The edges of a deforestation patch in the reference data do not align perfectly with the input data. This alignment issue is mainly because of the resolution difference, since the reference data was generated from a satellite image with 30m spatial resolution compared to the 10m spatial resolution of the input data.

The second type of error highlighted inside the blue frame is false positives areas overlaid with pink. These are the areas that the model predicted as deforestation, but they are not deforested in the reference data. Rocky outcrops, bare soil or sparsely vegetated hilltops are some examples of ambiguous areas that the model classifies as deforestation. None of the models experimented with in this study could distinguish between such an ambiguous class that resembles a deforested land. However, this can probably be fixed with the inclusion of a digital elevation model (DEM). The DEM can be integrated as a separate channel to the model's input.

Figure 5.9 False positives and false negatives of the predicted map overlaid on top of input S-2 image.

5.3 FINAL RECAP

The result of each experiment and their performance analysis is presented in this section. For scenario-1, it was noticed that the performance of EF U-Net3 and base U-Net3 with standalone input has the best performance with zero-cloud cover images. Moreover, the performance of the S-2 individual image degrades rapidly with the increase in the cloud cover. As expected, since SAR data is not affected by clouds, the performance of the S-1 images remains stable irrespective of the cloud cover scenario. The same stability is also observed in the EF scenarios. Even though this stable performance was not perfect, it still somehow avoids features from the S-2 cloudy image and uses the features from the S-1 image. Nevertheless, the prediction from EF still had room to improve as it fell short of the majority of the LF models in terms of accuracy.

Scenario-1, with different LF and attention mechanisms, has overall results that outperform the EF results. The result from LF models had a similar level of accuracy to that of a standalone S-2 image when the image has no clouds. However, the LF and the LF with attention mechanisms outperformed the others when the input image had clouds. The visual quality of the result from LF models was better in delineating the edge of deforestation patches. The main limitation of all the LF and EF was that the result of fusion does not significantly improve the performance compared to performance from standalone S-2 cloud-free inputs.

In scenario-2, a more realistic experiment was set up. The model was trained with much more training samples, and the prediction was performed on unseen images of the same area. Due to time constraints, only one variation of U-Net2 with LF was performed. The results suggest that the model can predict deforested areas regardless of any cloud cover situation. The prediction is also stable with increasing cloud cover in the input images. This stability suggests that the model was able to avoid any amount of cloud in the input S-2 image and extract features from only the S-1 image during a cloudy scenario. However, the major drawback of this experiment is that the fusion of a cloud-free image pair yields in similar result to that of an image pair with a cloudy scenario. However, this result may be explained by the imbalance of cloudy and non-cloudy images in the training samples. Nevertheless, none of the experiments performed in this research could differentiate between the bare rock on a hilltop with deforested areas.

Chapter 6

Discussion

In this section, a more comprehensive scope of the study is discussed by comparing the result of this research with other similar studies and discussing the implication of this study in a broader scope.

6.1 COMPARISON WITH OTHER STUDIES

It is challenging to compare this research with other studies as most of the studies use their datasets in different study areas and applied different models to tackle a specific challenge. No studies set up their experiments based on cloudy images to train their model; hence, comparing the accuracy and other performance metrics would be inappropriate. However, multiple studies utilise cloud-free optical images to study deforestation and other LULC mapping methods. Ortega et al., (2021) used different FCN-based architectures to compare optical and SAR data for deforestation mapping. Their experiment found a mean average precision of 93.65 when using S-2 data and 85.05 when using S-1 data. Nevertheless, the accuracy attained by them is similar to this research when it is compared with non-cloudy scenarios.

John and Zhang, (2022) used attention-based U-Net for detecting deforestation using satellite imagery, which partially resembles the one used in this research. The attention mechanism is similar to this research but not entirely the same. They perform their experiment using a three-band RGB dataset of the Amazon Rainforest (Bragagnolo et al., 2019) and 4-band RGB+NIR multispectral data for the Atlantic Forest (Bragagnolo et al., 2021). They achieved an IoU of 91.99%. However, the semantics of their classification was much more straightforward than in this study. They split their binary class into forest and non-forest. The non-forest class includes everything else than forest, making it a more straightforward classification problem than this study. Also, they produced their label using a modified k-means clustering algorithm on the input dataset. Therefore, they had the exact matching resolution for input and target data, which helped them achieve good accuracy.

Fusion of optical and SAR images using SVM for LULC classification was performed by Gibril et al., (2016); R. Zhang et al., (2020) in a cloud-free scenario yielded a lower accuracy than this research. This lower accuracy is because the neural network used in this research is superior in learning more complex features over an SVM (Support Vector Machines).

6.2 RESEARCH IMPLICATIONS

In the domain of image classification from remotely sensed imagery, most researches rely heavily on the availability of cloud-free optical images. The main advantage of this study is that it assessed the performance of the fusion models with cloudy images. Very few studies use cloudy images as part of their training process. A few variations of GANs (Gao et al., 2020b; Grohnfeldt et al., 2018) try to remove clouds from the optical image. However, the resulting image from GAN-based architectures is often artificial and can create undesirable artefacts. Several studies fuse cloud-free optical and SAR images at a pixel level to enhance the output image. However, they cannot deal with the situation when there are noises in either image.

Nevertheless, to the best knowledge, during this research, zero studies assessed the performance of their models in a combination of cloudy optical images and SAR data to identify deforestation or other

segmentation problem in the field of image classification from satellite imagery. Furthermore, the fusion models can output impressive classification results even during an extreme cloud cover scenario. Therefore, the primary contribution of this research is a cloud-independent deforestation mapping application which works in any season and weather condition. This research also demonstrates that the scientific methods used for this study could be easily modified and adapted to study different remote sensing segmentation problems that are held back due to persistent cloud cover. Examples of such studies include change detection, wetland monitoring (Montgomery et al., 2019), flood mapping (D'Addabbo et al., 2016), and real-time wild-fire monitoring (Zhang et al., 2021), disaster mitigation, and many others.

Chapter 7

Conclusion and Future Developments

7.1 CONCLUSIONS

This study explored a deep learning-based multimodal fusion of S-1 and S-2 data for mapping deforestation. Initial experiments were performed to understand the ability of each sensor to map deforestation in different cloud-cover scenarios. The initial experiments were performed with a twofold objective. The first objective was to identify how the model performs on each individual image of the image pairs with different cloudy scenarios. The second objective was to explore different fusion mechanisms that automatically exclude features from cloudy optical images in cases of high obstruction by clouds and rely more on radar images. Different performance analysis metrics – F1 score, IoU score, and qualitative map analysis- show that the fusion improves the consistency of mapping approaches, allowing it to generate maps in any conditions. The results from LF architectures perform similarly to each other in different cloud scenarios. Using an attention mechanism in the LF increases the classification accuracy only by a couple of fractions, showing the need for further improvements in the attention component. Therefore, plain LF was used in the final set of experiments. For the final experiment, 10 out of 24 image pair was used to generate the training sample. The number of images for this experiment was significantly more prominent than in all the previous experiments. Therefore, a lighter architecture with only two encoder blocks and two decoder blocks of U-Net was used to perform the LF. The model of the final experiments was used to predict an unseen image pair. The result outperforms the initial experiments for several cloudy scenarios. One major drawback of this experiment is that the accuracy does not improve when cloud-free images are in available the optical input. This model behaviour could be attributed to the imbalance between the amount of cloudy and non-cloudy training samples. The model also lacks on robustness to differentiate between bare rocks on a hilltop and a deforested patch.

This section also includes the answer to the research questions that were presented in Chapter 1:

1. How to prepare and build the dataset for training, testing and validation?

   Prior to preparing the dataset, a general preprocessing workflow was developed to convert the raw downloaded files to the AOI of the study. Section 3.1 demonstrate different preprocessing steps that each dataset went through before preparing them for training, testing and validation. After preprocessing, train test and validation patches were generated for two different scenarios. In scenario-1, the same image was divided to split train, test and validation patches; in scenario-2, the training, testing, and validation data were chosen from different images.

2. What are the crucial criteria for designing the network and performing a multimodal fusion of S-1 and S-2 data with an attention mechanism?

   In total, there were four different variations of U-Net architecture that were implemented in this research. The base U-Net3 and U-Net2 architecture take individual input to assess their performance at different cloud covers in the optical images. The U-Net3 was also used for performing the EF of the stacked input images. The rest three architectures were developed to perform the LF. All of the designed LF architecture had a dual encoder which is concatenated at each encoder level to perform the multimodal fusion. Two different variations of spatial attention were also used in the LF U-Net. Spatial attention on the optical data was designed to give less weight to the cloudy pixels in the optical input features. Spatial attention on both the input was

designed such that, in addition to less attention to the cloudy optical image, it should also give more attention to the representative features from the SAR image.

3. How to optimise the network to better optimize the loss during training and validation for better generalisation capability?

   As demonstrated in section 4.2.4, initial experiments were performed with different hyperparameters of the learning rate, the number of encoder blocks in the network, and batch size. After the result from these experiments, the hyperparameters with the best overall accuracy were chosen as a starting point for all the 144 experiments in Scenario-1. The hyperparameters used in Scenario-1 were used as a reference point to further tune the hyperparameters in Scenario-2. Apart from the hyperparameters, different callbacks were used throughout the experiments to monitor the model performance during the training process, stop it from overfitting, and give a generalized prediction.

4. How does the network perform in images with varying cloud cover? Does the fusion with the attention mechanism help to increase the classification quality?

   The performance metrics of Scenario-1 discussed in section 5.1 suggest that performing a multimodal fusion is necessary to create a robust fusion mechanism that does not relies on the availability of a cloud-free optical image. The result suggests that a multimodal fusion model could identify a deforested patch of land regardless of cloud presence in the optical image. The same conclusion can be drawn from the discussion in section 5.2 that the fusion automatically helps avoid cloudy image pixels and map deforestation using S-1 data. However, the fusion does not necessarily improve the accuracy even when the optical images are cloud-free. The fusion automatically detects features from the most representative features out of the input images and performs prediction. Using an attention mechanism in the fusion results in a similar performance to that of an LF. Therefore further studies regarding the attention mechanism are necessary to improve the result.

5. How does the predicted deforestation map compare with the reference data regarding different accuracy metrics and qualitative analysis?

   The predicted map's qualitative and quantitative analysis is discussed for each scenario in sections 5.1.1.2 and 5.2.2. The result suggests that the LF U-Net model is capable of predicting impressive classification results when there is zero per cent cloud in the optical image. Moreover, in addition to that, the model can also generate good quality results in a 100% cloud cover scenario.

7.2 FUTURE DEVELOPMENTS

Finally, for future developments, the following steps are recommended:

- Reference data quality: The reference data used in this study is generated from various LANDSAT images with a GSD of 30m. Therefore, the patches of deforestation do not exactly overlay on top of the input image of S-1 and S-2 (shown in Figure 5.8). This mismatch is due to the spatial resolution of 10m in the input image. Because of this mismatch, different performance metrics were underestimated by some amount. For some instances, the output predicted map overlaid better to a patch of deforested land than the reference image. Therefore, using higher-resolution reference data is recommended to get a more accurate estimate of the model performance.

- Explore different base model architectures other than U-Net.

- A different attention mechanism could be explored, which takes both encoder and decoder features to refine the attention map. This approach is successfully implemented in medical imaging segmentation (Oktay et al., 2018).

  The weights of the attention map could be initialised with the provided cloud mask with the S-2 image. From this point, the attention map could be further tuned during model training.

- The fusion accuracy in cloud-free images could be improved by taking a balanced set of training samples from non-cloudy and cloudy images. This imbalance can be tackled by oversampling the non-cloudy images or under-sampling the cloudy images.

- The fusion could also be improved by using a transfer learning procedure. The initial weights could be set by only training with cloud-free images, and then that weight could be the starting point for training with cloudy images.

- Regarding the ambiguity of deforestation class and bare rock or soil on a hilltop, a Digital Elevation Model (DEM) could be used in the training process as an additional input for the model.

LIST OF REFERENCES

Abdikan, S., 2018. Exploring image fusion of ALOS/PALSAR data and LANDSAT data to differentiate forest area. Geocarto International 33, 21–37. https://doi.org/10.1080/10106049.2016.1222635

Adrian, J., Sagan, V., Maimaitijiang, M., 2021. Sentinel SAR-optical fusion for crop type mapping using deep learning and Google Earth Engine. ISPRS Journal of Photogrammetry and Remote Sensing 175, 215–235. https://doi.org/10.1016/J.ISPRSJPRS.2021.02.018

Almeida-Filho, R., Rosenqvist, A., Shimabukuro, Y.E., Silva-Gomez, R., 2007. Detecting deforestation with multitemporal L-band SAR imagery: a case study in western Brazilian Amazônia. International Journal of Remote Sensing 28, 1383–1390. https://doi.org/10.1080/01431160600754591

Badrinarayanan, V., Kendall, A., Cipolla, R., 2017. SegNet: A Deep Convolutional Encoder-Decoder Architecture for Image Segmentation. IEEE Transactions on Pattern Analysis and Machine Intelligence 39, 2481–2495. https://doi.org/10.1109/TPAMI.2016.2644615

Bahdanau, D., Cho, K.H., Bengio, Y., 2014. Neural Machine Translation by Jointly Learning to Align and Translate. 3rd International Conference on Learning Representations, ICLR 2015 - Conference Track Proceedings.

Ball, J.E., Anderson, D.T., Chan, C.S., 2017. Comprehensive survey of deep learning in remote sensing: theories, tools, and challenges for the community. Journal of Applied Remote Sensing 11, 1. https://doi.org/10.1117/1.JRS.11.042609

Bergado, J.R., Persello, C., Gevaert, C., 2016. A deep learning approach to the classification of sub-decimetre resolution aerial images, in: 2016 IEEE International Geoscience and Remote Sensing Symposium (IGARSS). IEEE, pp. 1516–1519. https://doi.org/10.1109/IGARSS.2016.7729387

Bermudez, J.D., Happ, P.N., Feitosa, R.Q., Oliveira, D.A.B., 2019. Synthesis of Multispectral Optical Images from SAR/Optical Multitemporal Data Using Conditional Generative Adversarial Networks. IEEE Geoscience and Remote Sensing Letters 16, 1220–1224. https://doi.org/10.1109/LGRS.2019.2894734

Bragagnolo, L., da Silva, R.V., Grzybowski, J.M.V., 2021. Amazon and Atlantic Forest image datasets for semantic segmentation. https://doi.org/10.5281/ZENODO.4498086

Bragagnolo, L., da Silva, R.V., Grzybowski, J.M.V., 2019. Amazon Rainforest dataset for semantic segmentation. https://doi.org/10.5281/ZENODO.3233081

Bruzzone, L., Marconcini, M., Wegmüller, U., Wiesmann, A., 2004. An advanced system for the automatic classification of multitemporal SAR images. IEEE Transactions on Geoscience and Remote Sensing 42, 1321–1334. https://doi.org/10.1109/TGRS.2004.826821

Cabral, A.I.R., Saito, C., Pereira, H., Laques, A.E., 2018. Deforestation pattern dynamics in protected areas of the Brazilian Legal Amazon using remote sensing data. Applied Geography 100, 101–115. https://doi.org/10.1016/j.apgeog.2018.10.003

Camps-Valls, G., Tuia, D., Bruzzone, L., Benediktsson, J.A., 2014. Advances in hyperspectral image classification: Earth monitoring with statistical learning methods. IEEE Signal Processing Magazine 31, 45–54. https://doi.org/10.1109/MSP.2013.2279179

Cheng, G., Han, J., Lu, X., 2017. Remote Sensing Image Scene Classification: Benchmark and State of the Art. Proceedings of the IEEE 105, 1865–1883. https://doi.org/10.1109/JPROC.2017.2675998

Cheng, G., Zhou, P., Han, J., 2016. Learning Rotation-Invariant Convolutional Neural Networks for Object Detection in VHR Optical Remote Sensing Images. IEEE Transactions on Geoscience and Remote Sensing 54, 7405–7415. https://doi.org/10.1109/TGRS.2016.2601622

Cremer, F., Urbazaev, M., Cortes, J., Truckenbrodt, J., Schmullius, C., Thiel, C., 2020. Potential of Recurrence Metrics from Sentinel-1 Time Series for Deforestation Mapping. IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing 13, 5233–5240. https://doi.org/10.1109/JSTARS.2020.3019333

D'Addabbo, A., Refice, A., Pasquariello, G., Lovergine, F., 2016. SAR/optical data fusion for flood detection. International Geoscience and Remote Sensing Symposium (IGARSS) 2016-November, 7631–7634. https://doi.org/10.1109/IGARSS.2016.7730990

D'Almeida, C., Vörösmarty, C.J., Hurtt, G.C., Marengo, J.A., Dingman, S.L., Keim, B.D., 2007. The effects of deforestation on the hydrological cycle in Amazonia: A review on scale and resolution. International Journal of Climatology 27, 633–647. https://doi.org/10.1002/joc.1475

de Bem, P., de Carvalho Junior, O., Fontes Guimarães, R., Trancoso Gomes, R., 2020. Change Detection of Deforestation in the Brazilian Amazon Using Landsat Data and Convolutional Neural Networks. Remote Sensing 12, 901. https://doi.org/10.3390/rs12060901

Desale, R.P., Verma, S. v., 2013. Study and analysis of PCA, DCT & DWT based image fusion techniques. International Conference on Signal Processing, Image Processing and Pattern Recognition 2013, ICSIPR 2013 1, 66–69. https://doi.org/10.1109/ICSIPR.2013.6497960

Durieux, A.M., Calef, M.T., Arko, S., Chartrand, R., Kontgis, C., Keisler, R., Warren, M.S., 2019. Monitoring forest disturbance using change detection on synthetic aperture radar imagery, in: Zelinski, M.E., Taha, T.M., Howe, J., Awwal, A.A., Iftekharuddin, K.M. (Eds.), Applications of Machine Learning. SPIE, p. 39. https://doi.org/10.1117/12.2528945

Ebrahimi Kahou, S., Bouthillier, X., Lamblin, P., Gulcehre, C., Michalski, V., Konda, K., Jean, S., Froumenty, P., Dauphin, Y., Boulanger-Lewandowski, N., Chandias Ferrari, R., Mirza, M., Warde-Farley, D., Courville, A., Vincent, P., Memisevic, R., Pal, C., Bengio, Y., 2015. EmoNets: Multimodal deep learning approaches for emotion recognition in video. Journal on Multimodal User Interfaces 10, 99–111. https://doi.org/10.1007/s12193-015-0195-2

ESA, n.d. Copernicus Sentinel data 2018-2021 for Sentinel data [WWW Document]. URL https://sentinel.esa.int/web/sentinel (accessed 12.6.21a).

ESA, n.d. STEP – Science Toolbox Exploitation Platform [WWW Document]. URL http://step.esa.int/main/ (accessed 5.16.22b).

Flach, P., Kull, M., 2015. Precision-Recall-Gain Curves: PR Analysis Done Right, in: Cortes, C., Lawrence, N., Lee, D., Sugiyama, M., Garnett, R. (Eds.), Advances in Neural Information Processing Systems. Curran Associates, Inc.

Gao, J., Yuan, Q., Li, J., Zhang, H., Su, X., 2020. Cloud Removal with Fusion of High Resolution Optical and SAR Images Using Generative Adversarial Networks. Remote Sensing 12, 191. https://doi.org/10.3390/rs12010191

Ghaffarian, S., Valente, J., van der Voort, M., Tekinerdogan, B., 2021. Effect of attention mechanism in deep learning-based remote sensing image processing: A systematic literature review. Remote Sensing 13, 2965. https://doi.org/10.3390/RS13152965/S1

Giam, X., 2017. Global biodiversity loss from tropical deforestation. Proc Natl Acad Sci U S A 114, 5775–5777. https://doi.org/10.1073/pnas.1706264114

Gibril, M.B.A., Bakar, S.A., Yao, K., Idrees, M.O., Pradhan, B., 2017. Fusion of RADARSAT-2 and multispectral optical remote sensing data for LULC extraction in a tropical agricultural area. Geocarto International 32, 735–748. https://doi.org/10.1080/10106049.2016.1170893

Goodfellow, I., Pouget-Abadie, J., Mirza, M., Xu, B., Warde-Farley, D., Ozair, S., Courville, A., Bengio, Y., 2014. Generative Adversarial Networks. Commun ACM 63, 139–144. https://doi.org/10.1145/3422622

Griffiths, P., Jakimow, B., Hostert, P., 2018. Reconstructing long term annual deforestation dynamics in Pará and Mato Grosso using the Landsat archive. Remote Sensing of Environment 216, 497–513. https://doi.org/10.1016/J.RSE.2018.07.010

Gu, S., Holly, E., Lillicrap, T., Levine, S., 2016. Deep Reinforcement Learning for Robotic Manipulation with Asynchronous Off-Policy Updates. Proceedings - IEEE International Conference on Robotics and Automation 3389–3396. https://doi.org/10.1109/ICRA.2017.7989385

Hoang, N.T., Kanemoto, K., 2021. Mapping the deforestation footprint of nations reveals growing threat to tropical forests. Nature Ecology & Evolution 5, 845–853. https://doi.org/10.1038/s41559-021-01417-z

Hong, G., Zhang, Y., 2008. Comparison and improvement of wavelet-based image fusion. International Journal of Remote Sensing 29, 673–691. https://doi.org/10.1080/01431160701313826

Houghton, R.A., 1999. The annual net flux of carbon to the atmosphere from changes in land use 1850-1990. Tellus, Series B: Chemical and Physical Meteorology 51, 298–313. https://doi.org/10.3402/tellusb.v51i2.16288

Hu, F., Xia, G.S., Hu, J., Zhang, L., 2015. Transferring deep convolutional neural networks for the scene classification of high-resolution remote sensing imagery. Remote Sensing 7, 14680–14707. https://doi.org/10.3390/rs71114680

Hu, J., Shen, L., Albanie, S., Sun, G., Wu, E., 2017. Squeeze-and-Excitation Networks. IEEE Transactions on Pattern Analysis and Machine Intelligence 42, 2011–2023. https://doi.org/10.1109/TPAMI.2019.2913372

INPE, n.d. Terrabrasilis – Geographic Data Platform [WWW Document]. URL http://terrabrasilis.dpi.inpe.br/en/home-page/ (accessed 1.6.22).

Jaccard, P., 1901. Étude comparative de la distribution florale dans une portion des Alpes et du Jura. https://doi.org/10.5169/SEALS-266450

Jadon, S., 2020. A survey of loss functions for semantic segmentation. 2020 IEEE Conference on Computational Intelligence in Bioinformatics and Computational Biology, CIBCB 2020. https://doi.org/10.1109/CIBCB48159.2020.9277638

John, D., Zhang, C., 2022. An attention-based U-Net for detecting deforestation within satellite sensor imagery. International Journal of Applied Earth Observation and Geoinformation 107, 102685. https://doi.org/10.1016/J.JAG.2022.102685

Khanh, T.L.B., Dao, D.-P., Ho, N.-H., Yang, H.-J., Baek, E.-T., Lee, G., Kim, S.-H., Yoo, S.B., 2020. Enhancing U-Net with Spatial-Channel Attention Gate for Abnormal Tissue Segmentation in Medical Imaging. Applied Sciences 10, 5729. https://doi.org/10.3390/app10175729

Kingma, D.P., Ba, J.L., 2014. Adam: A Method for Stochastic Optimization. 3rd International Conference on Learning Representations, ICLR 2015 - Conference Track Proceedings. https://doi.org/10.48550/arxiv.1412.6980

Kiros, R., Popuri, K., Cobzas, D., Jagersand, M., 2014. Stacked Multiscale Feature Learning for Domain Independent Medical Image Segmentation. Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics) 8679, 25–32. https://doi.org/10.1007/978-3-319-10581-9_4

Lecun, Y., Bengio, Y., Hinton, G., 2015. Deep learning. Nature 521, 436–444. https://doi.org/10.1038/nature14539

Lee, J.S., Jurkevich, I., Dewaele, P., Wambacq, P., Oosterlinck, A., 2009. Speckle filtering of synthetic aperture radar images: A review. http://dx.doi.org/10.1080/02757259409532206 8, 313–340. https://doi.org/10.1080/02757259409532206

Lenz, I., Lee, H., Saxena, A., 2013. Deep Learning for Detecting Robotic Grasps. International Journal of Robotics Research 34, 705–724. https://doi.org/10.1177/0278364914549607

Li, H., Qiu, K., Chen, L., Mei, X., Hong, L., Tao, C., 2021. SCAttNet: Semantic Segmentation Network with Spatial and Channel Attention Mechanism for High-Resolution Remote Sensing Images. IEEE Geoscience and Remote Sensing Letters 18, 905–909. https://doi.org/10.1109/LGRS.2020.2988294

Li, H., Wu, X.J., 2019. DenseFuse: A fusion approach to infrared and visible images. IEEE Transactions on Image Processing 28, 2614–2623. https://doi.org/10.1109/TIP.2018.2887342

Li, X., Lei, L., Sun, Y., Li, M., Kuang, G., 2020. Multimodal bilinear fusion network with second-order attention-based channel selection for land cover classification. IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing 13, 1011–1026. https://doi.org/10.1109/JSTARS.2020.2975252

Li, Y., Fu, R., Meng, X., Jin, W., Shao, F., 2020. A SAR-to-Optical Image Translation Method Based on Conditional Generation Adversarial Network (cGAN). IEEE Access 8, 60338–60343. https://doi.org/10.1109/ACCESS.2020.2977103

Li, Y., Zhao, J., Lv, Z., Li, J., 2021a. Medical image fusion method by deep learning. International Journal of Cognitive Computing in Engineering 2, 21–29. https://doi.org/10.1016/J.IJCCE.2020.12.004

Li, Y., Zhao, J., Lv, Z., Pan, Z., 2021b. Multimodal Medical Supervised Image Fusion Method by CNN. Frontiers in Neuroscience 15, 303. https://doi.org/https://doi.org/10.3389/fnins.2021.638976

Liu, C., Li, W., Zhu, G., Zhou, H., Yan, H., Xue, P., 2020. Land Use/Land Cover Changes and Their Driving Factors in the Northeastern Tibetan Plateau Based on Geographical Detectors and Google Earth Engine: A Case Study in Gannan Prefecture. Remote Sensing 2020, Vol. 12, Page 3139 12, 3139. https://doi.org/10.3390/RS12193139

Liu, Y., Chen, X., Cheng, J., Peng, H., 2017. A medical image fusion method based on convolutional neural networks. 20th International Conference on Information Fusion, Fusion 2017 - Proceedings. https://doi.org/10.23919/ICIF.2017.8009769

Liu, Yu, Chen, X., Wang, Z., Wang, Z.J., Ward, R.K., Wang, X., 2018. Deep learning for pixel-level image fusion: Recent advances and future prospects. Information Fusion 42, 158–173. https://doi.org/10.1016/J.INFFUS.2017.10.007

Liu, Yanfei, Zhong, Y., Fei, F., Zhu, Q., Qin, Q., 2018. Scene classification based on a deep random-scale stretched convolutional neural network. Remote Sensing 10, 444. https://doi.org/10.3390/rs10030444

Lonnqvist, A., Rauste, Y., Molinier, M., Hame, T., 2010. Polarimetric SAR Data in Land Cover Mapping in Boreal Zone. IEEE Transactions on Geoscience and Remote Sensing 48. https://doi.org/10.1109/TGRS.2010.2048115

Lu, D., Li, G., Moran, E., Batistella, M., Freitas, C.C., 2011. Mapping impervious surfaces with the integrated use of Landsat Thematic Mapper and radar data: A case study in an urban–rural landscape in the Brazilian Amazon. ISPRS Journal of Photogrammetry and Remote Sensing 66, 798–808. https://doi.org/10.1016/J.ISPRSJPRS.2011.08.004

Lu, D., Weng, Q., 2007. A survey of image classification methods and techniques for improving classification performance. International Journal of Remote Sensing 28, 823–870. https://doi.org/10.1080/01431160600746456

Malhi, Y., Aragão, L.E.O.C., Galbraith, D., Huntingford, C., Fisher, R., Zelazowski, P., Sitch, S., McSweeney, C., Meir, P., 2009. Exploring the likelihood and mechanism of a climate-change-induced dieback of the Amazon rainforest. Proc Natl Acad Sci U S A 106, 20610–20615. https://doi.org/10.1073/pnas.0804619106

Malviya, A., Bhirud, S.G., 2009. Image Fusion of Digital Images. International Journal of Recent Trends in Engineering 2, 146.

Maretto, R.V., 2020. Automating Land Cover Change Detection: A Deep Learning Based Approach to Map Deforested Areas (Doctoral Dissertation). Instituto Nacional de Pesquisas Espaciais - INPE, São José dos Campos.

Maretto, R. v., Fonseca, L.M.G., Jacobs, N., Korting, T.S., Bendini, H.N., Parente, L.L., 2021. Spatio-Temporal Deep Learning Approach to Map Deforestation in Amazon Rainforest. IEEE Geoscience and Remote Sensing Letters 18, 771–775. https://doi.org/10.1109/LGRS.2020.2986407

Masi, G., Cozzolino, D., Verdoliva, L., Scarpa, G., 2016. Pansharpening by Convolutional Neural Networks. Remote Sensing 2016, Vol. 8, Page 594 8, 594. https://doi.org/10.3390/RS8070594

Mitchard, E.T.A., Saatchi, S.S., Lewis, S.L., Feldpausch, T.R., Woodhouse, I.H., Sonké, B., Rowland, C., Meir, P., 2011. Measuring biomass changes due to woody encroachment and deforestation/degradation in a forest–savanna boundary region of central Africa using multi-temporal L-band radar backscatter. Remote Sensing of Environment 115, 2861–2873. https://doi.org/10.1016/J.RSE.2010.02.022

Montgomery, J., Brisco, B., Chasmer, L., Devito, K., Cobbaert, D., Hopkinson, C., 2019. SAR and Lidar Temporal Data Fusion Approaches to Boreal Wetland Ecosystem Monitoring. Remote Sensing 2019, Vol. 11, Page 161 11, 161. https://doi.org/10.3390/RS11020161

Neverova, N., Wolf, C., Taylor, G., Nebout, F., 2014. ModDrop: adaptive multi-modal gesture recognition. IEEE Transactions on Pattern Analysis and Machine Intelligence 38, 1692–1706. https://doi.org/10.1109/TPAMI.2015.2461544

Nguyen, H.T.T., Doan, T.M., Tomppo, E., McRoberts, R.E., 2020. Land Use/Land Cover Mapping Using Multitemporal Sentinel-2 Imagery and Four Classification Methods—A Case Study from Dak Nong, Vietnam. Remote Sensing 2020, Vol. 12, Page 1367 12, 1367. https://doi.org/10.3390/RS12091367

Nicolau, A.P., Flores-Anderson, A., Griffin, R., Herndon, K., Meyer, F.J., 2021. Assessing SAR C-band data to effectively distinguish modified land uses in a heavily disturbed Amazon forest. International Journal of Applied Earth Observation and Geoinformation 94, 102214. https://doi.org/10.1016/j.jag.2020.102214

Nogueira, K., Penatti, O.A.B., dos Santos, J.A., 2017. Towards better exploiting convolutional neural networks for remote sensing scene classification. Pattern Recognition 61, 539–556. https://doi.org/10.1016/j.patcog.2016.07.001

Noh, H., Hong, S., Han, B., 2015. Learning Deconvolution Network for Semantic Segmentation. Proceedings of the IEEE International Conference on Computer Vision 1520–1528.

Oktay, O., Schlemper, J., Folgoc, L. le, Lee, M., Heinrich, M., Misawa, K., Mori, K., McDonagh, S., Hammerla, N.Y., Kainz, B., Glocker, B., Rueckert, D., 2018. Attention U-Net: Learning Where to Look for the Pancreas.

Ortega, M.X., Feitosa, R.Q., Bermudez, J.D., Happ, P.N., de Almeida, C.A., 2021. Comparison of Optical and SAR Data for Deforestation Mapping in the Amazon Rainforest with Fully Convolutional Networks. International Geoscience and Remote Sensing Symposium (IGARSS) 3769–3772. https://doi.org/10.1109/IGARSS47720.2021.9554970

O'Shea, K., Nash, R., 2015. An Introduction to Convolutional Neural Networks.

Pajares, G., de la Cruz, J.M., 2004. A wavelet-based image fusion tutorial. Pattern Recognition 37, 1855–1872. https://doi.org/10.1016/J.PATCOG.2004.03.010

Pereira, L. de O., Freitas, C. da C., Sant´Anna, S.J.S., Lu, D., Moran, E.F., 2013. Optical and radar data integration for land use and land cover mapping in the Brazilian Amazon. GIScience & Remote Sensing 50, 301–321. https://doi.org/10.1080/15481603.2013.805589

Persello, C., Stein, A., 2017. Deep Fully Convolutional Networks for the Detection of Informal Settlements in VHR Images. IEEE Geoscience and Remote Sensing Letters 14, 2325–2329. https://doi.org/10.1109/LGRS.2017.2763738

Radu, V., Lane, N.D., Bhattacharya, S., Mascolo, C., Marina, M.K., Kawsar, F., 2016. Towards multimodal deep learning for activity recognition on mobile devices. UbiComp 2016 Adjunct - Proceedings of the 2016 ACM International Joint Conference on Pervasive and Ubiquitous Computing 185–188. https://doi.org/10.1145/2968219.2971461

Ramachandram, D., Taylor, G.W., 2017. Deep multimodal learning: A survey on recent advances and trends. IEEE Signal Processing Magazine 34, 96–108. https://doi.org/10.1109/MSP.2017.2738401

Rodrigues, A.S.L., Ewers, R.M., Parry, L., Souza, C., Veríssimo, A., Balmford, A., 2009. Boom-and-bust development patterns across the amazon deforestation frontier. Science (1979) 324, 1435–1437. https://doi.org/10.1126/science.1174002

Ronneberger, O., Fischer, P., Brox, T., 2015. U-Net: Convolutional Networks for Biomedical Image Segmentation. Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics) 9351, 234–241.

Roy, A.G., Navab, N., Wachinger, C., 2019. Recalibrating Fully Convolutional Networks With Spatial and Channel "Squeeze and Excitation" Blocks. IEEE Transactions on Medical Imaging 38, 540–549. https://doi.org/10.1109/TMI.2018.2867261

Roy, A.G., Navab, N., Wachinger, C., 2018. Concurrent Spatial and Channel 'Squeeze & Excitation' in Fully Convolutional Networks. Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics) 11070 LNCS, 421–429. https://doi.org/10.1007/978-3-030-00928-1_48

Schubert, A., Small, D., 2008. Guide to ASAR Geocoding. Zurich, Switzerland.

Sefrin, O., Riese, F.M., Keller, S., 2020. Deep Learning for Land Cover Change Detection. Remote Sensing 13, 78. https://doi.org/10.3390/rs13010078

Shao, Z., Cai, J., 2018. Remote Sensing Image Fusion with Deep Convolutional Neural Network. IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing 11, 1656–1669. https://doi.org/10.1109/JSTARS.2018.2805923

Shelhamer, E., Long, J., Darrell, T., 2014. Fully Convolutional Networks for Semantic Segmentation. IEEE Transactions on Pattern Analysis and Machine Intelligence 39, 640–651. https://doi.org/10.48550/arxiv.1411.4038

Silva Junior, C.H.L., Pessôa, A.C.M., Carvalho, N.S., Reis, J.B.C., Anderson, L.O., Aragão, L.E.O.C., 2021. The Brazilian Amazon deforestation rate in 2020 is the greatest of the decade. Nature Ecology and Evolution 5, 144–145. https://doi.org/10.1038/s41559-020-01368-x

Singh, A., 1989. Review Articlel: Digital change detection techniques using remotely-sensed data. International Journal of Remote Sensing 10, 989–1003. https://doi.org/10.1080/01431168908903939

Singh, D., Singh, B., 2020. Investigating the impact of data normalization on classification performance. Applied Soft Computing 97, 105524. https://doi.org/10.1016/J.ASOC.2019.105524

Sun, J., Jiang, Y., Zeng, S., 2005. A study of PCA image fusion techniques on remote sensing, in: Wang, C., Zhong, S., Hu, X. (Eds.), International Conference on Space Information Technology. SPIE, p. 59853X. https://doi.org/10.1117/12.658216

Syrris, V., Hasenohr, P., Delipetrev, B., Kotsev, A., Kempeneers, P., Soille, P., 2019. Evaluation of the Potential of Convolutional Neural Networks and Random Forests for Multi-Class Segmentation of Sentinel-2 Imagery. Remote Sensing 11, 907. https://doi.org/10.3390/rs11080907

Tajbakhsh, N., Shin, J.Y., Gurudu, S.R., Hurst, R.T., Kendall, C.B., Gotway, M.B., Liang, J., 2017. Convolutional Neural Networks for Medical Image Analysis: Full Training or Fine Tuning? IEEE Transactions on Medical Imaging 35, 1299–1312. https://doi.org/10.1109/TMI.2016.2535302

Treitz, P., 2004. Remote sensing for mapping and monitoring land-cover and land-use change. Progress in Planning 61, 267. https://doi.org/10.1016/S0305-9006(03)00062-X

Ullmann, T., Schmitt, A., Roth, A., Duffe, J., Dech, S., Hubberten, H.-W., Baumhauer, R., 2014. Land Cover Characterization and Classification of Arctic Tundra Environments by Means of Polarized Synthetic Aperture X- and C-Band Radar (PolSAR) and Landsat 8 Multispectral Imagery — Richards Island, Canada. Remote Sensing 2014, Vol. 6, Pages 8565-8593 6, 8565–8593. https://doi.org/10.3390/RS6098565

Vahadane, A., B, A., Majumdar, S., 2021. Dual Encoder Attention U-net for Nuclei Segmentation. Annu Int Conf IEEE Eng Med Biol Soc 2021, 3205–3208. https://doi.org/10.1109/EMBC46164.2021.9630037

Walker, W.S., Kellndorfer, J.M., Kirsch, K.M., Stickler, C.M., Nepstad, D.C., Stickler, C.M., Nepstad, D.C., 2010. Large-Area Classification and Mapping of Forest and Land Cover in the Brazilian Amazon: A Comparative Analysis of ALOS/PALSAR and Landsat Data Sources. IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing 3, 594–604. https://doi.org/10.1109/JSTARS.2010.2076398

Wang, S.W., Gebru, B.M., Lamchin, M., Kayastha, R.B., Lee, W.-K., 2020. Land Use and Land Cover Change Detection and Prediction in the Kathmandu District of Nepal Using Remote Sensing and GIS. Sustainability 12, 3925. https://doi.org/10.3390/su12093925

Wang, Z., Ziou, D., Armenakis, C., Li, D., Li, Q., 2005. A comparative analysis of image fusion methods. IEEE Transactions on Geoscience and Remote Sensing 43, 1391–1402. https://doi.org/10.1109/TGRS.2005.846874

Waske, B., Braun, M., 2009. Classifier ensembles for land cover mapping using multitemporal SAR imagery. ISPRS Journal of Photogrammetry and Remote Sensing 64, 450–457. https://doi.org/10.1016/J.ISPRSJPRS.2009.01.003

Watanabe, M., Koyama, C.N., Hayashi, M., Nagatani, I., Shimada, M., 2018. Early-stage deforestation detection in the tropics with L-band SAR. IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing 11, 2127–2133. https://doi.org/10.1109/JSTARS.2018.2810857

Werth, D., 2002. The local and global effects of Amazon deforestation. Journal of Geophysical Research 107, 8087. https://doi.org/10.1029/2001JD000717

Woo, S., Park, J., Lee, J.Y., Kweon, I.S., 2018. CBAM: Convolutional Block Attention Module. Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics) 11211 LNCS, 3–19. https://doi.org/10.1007/978-3-030-01234-2_1

Wu, P., Hoi, S.C.H., Xia, H., Zhao, P., Wang, D., Miao, C., 2013. Online multimodal deep similarity learning with application to image retrieval, in: Proceedings of the 21st ACM International Conference on Multimedia - MM '13. ACM Press, New York, New York, USA, pp. 153–162. https://doi.org/10.1145/2502081.2502112

Xu, H., Ma, J., Jiang, J., Guo, X., Ling, H., 2022. U2Fusion: A Unified Unsupervised Image Fusion Network. IEEE Transactions on Pattern Analysis and Machine Intelligence 44, 502–518. https://doi.org/10.1109/TPAMI.2020.3012548

Xu, Z., Zhu, J., Geng, J., Deng, X., Jiang, W., 2021. Triplet Attention Feature Fusion Network for SAR and Optical Image Land Cover Classification, in: 2021 IEEE International Geoscience and Remote Sensing Symposium IGARSS. IEEE, pp. 4256–4259. https://doi.org/10.1109/IGARSS47720.2021.9555126

Yao, X., Han, J., Cheng, G., Qian, X., Guo, L., 2016. Semantic Annotation of High-Resolution Satellite Images via Weakly Supervised Learning. IEEE Transactions on Geoscience and Remote Sensing 54, 3660–3671. https://doi.org/10.1109/TGRS.2016.2523563

Yin, H., Pflugmacher, D., Li, A., Li, Z., Hostert, P., 2018. Land use and land cover change in Inner Mongolia - understanding the effects of China's re-vegetation programs. Remote Sensing of Environment 204, 918–930. https://doi.org/10.1016/J.RSE.2017.08.030

Yu, Y., Liu, F., 2018. A Two-Stream Deep Fusion Framework for High-Resolution Aerial Scene Classification. Computational Intelligence and Neuroscience 2018, 1–13. https://doi.org/10.1155/2018/8639367

Zagoruyko, S., Komodakis, N., 2016. Paying More Attention to Attention: Improving the Performance of Convolutional Neural Networks via Attention Transfer. 5th International Conference on Learning Representations, ICLR 2017 - Conference Track Proceedings.

Zhang, H., Xu, H., Xiao, Y., Guo, X., Ma, J., 2020. Rethinking the Image Fusion: A Fast Unified Image Fusion Network based on Proportional Maintenance of Gradient and Intensity. Proceedings of the AAAI Conference on Artificial Intelligence 34, 12797–12804. https://doi.org/10.1609/AAAI.V34I07.6975

Zhang, P., Ban, Y., Nascetti, A., 2021. Learning U-Net without forgetting for near real-time wildfire monitoring by the fusion of SAR and optical time series. Remote Sensing of Environment 261, 112467. https://doi.org/10.1016/J.RSE.2021.112467

Zhang, R., Tang, X., You, S., Duan, K., Xiang, H., Luo, H., 2020. A Novel Feature-Level Fusion Framework Using Optical and SAR Remote Sensing Images for Land Use/Land Cover (LULC) Classification in Cloudy Mountainous Area. Applied Sciences 2020, Vol. 10, Page 2928 10, 2928. https://doi.org/10.3390/APP10082928

Zhang, Y., Fang, M., Wang, N., 2019. Channel-spatial attention network for fewshot classification. PLOS ONE 14, e0225426. https://doi.org/10.1371/JOURNAL.PONE.0225426

Zhao, P., Zhang, J., Fang, W., Deng, S., 2020. SCAU-Net: Spatial-Channel Attention U-Net for Gland Segmentation. Frontiers in Bioengineering and Biotechnology 8, 670. https://doi.org/10.3389/FBIOE.2020.00670/BIBTEX

Zheng, Y., Essock, E.A., Hansen, B.C., 2004. An advanced image fusion algorithm based on wavelet transform: incorporation with PCA and morphological processing, in: Dougherty, E.R., Astola, J.T., Egiazarian, K.O. (Eds.), Image Processing: Algorithms and Systems III. SPIE, p. 177. https://doi.org/10.1117/12.523966

Zhong, J., Yang, B., Huang, G., Zhong, F., Chen, Z., 2016. Remote Sensing Image Fusion with Convolutional Neural Network. Sensing and Imaging 17, 10. https://doi.org/10.1007/s11220-016-0135-6

Zhou, B., Khosla, A., Lapedriza, A., Oliva, A., Torralba, A., 2015. Learning Deep Features for Discriminative Localization. Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition 2016-December, 2921–2929. https://doi.org/10.1109/CVPR.2016.319

Zhou, T., Canu, S., Ruan, S., 2020. Fusion based on attention mechanism and context constraint for multi-modal brain tumor segmentation. Computerized Medical Imaging and Graphics 86, 101811. https://doi.org/10.1016/J.COMPMEDIMAG.2020.101811

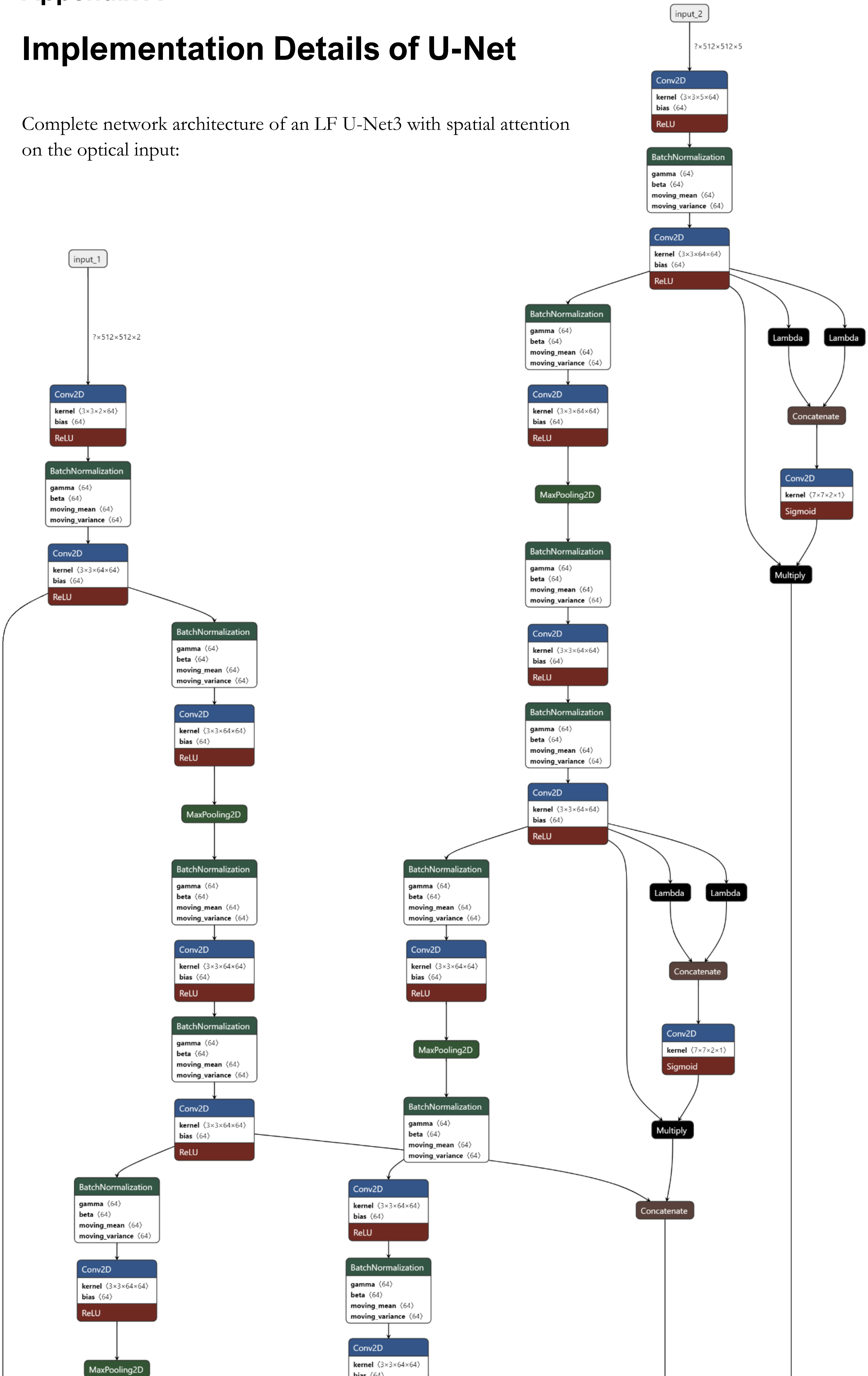
Zhu, H., Wang, Z., Shi, Y., Hua, Y., Xu, G., Deng, L., 2020. Multimodal Fusion Method Based on Self-Attention Mechanism. Wireless Communications and Mobile Computing 2020, 1–8. https://doi.org/10.1155/2020/8843186

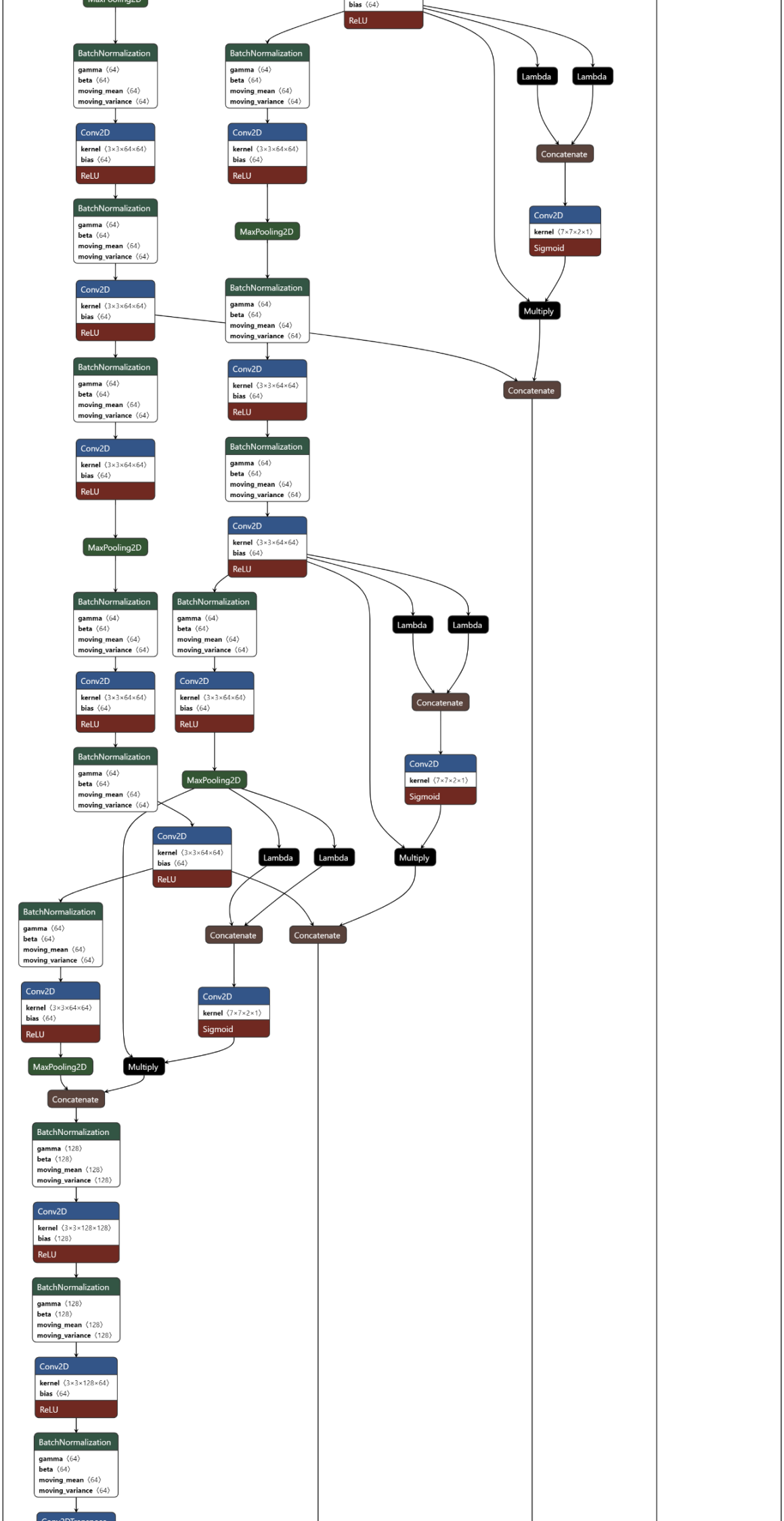
Zou, Q., Ni, L., Zhang, T., Wang, Q., 2015. Deep Learning Based Feature Selection for Remote Sensing Scene Classification. IEEE Geoscience and Remote Sensing Letters 12, 2321–2325. https://doi.org/10.1109/LGRS.2015.2475299

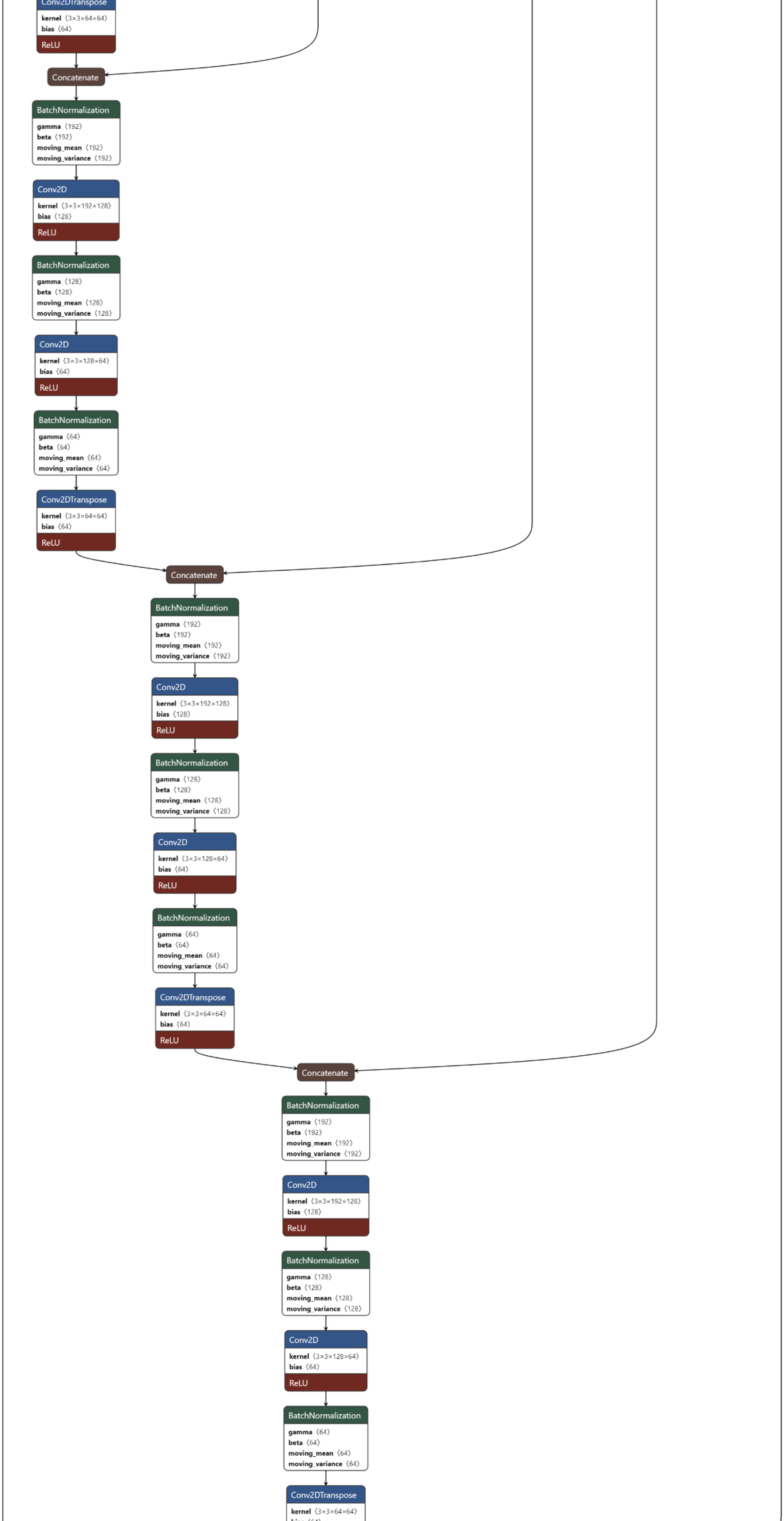
Appendix A

Implementation Details of U-Net

Complete network architecture of an LF U-Net3 with spatial attention on the optical input:




bias (64)
ReLU
BatchNormalization
gamma (64)
beta (64)
moving_mean (64)
moving_variance (64)
Conv2D
kernel (3×3×64×64)
bias (64)
ReLU
Lambda
Concatenate
MaxPooling2D
kernel (7×7×2×1)
Sigmoid
Multiply
BatchNormalization
gamma (128)
beta (128)
moving_mean (128)
moving_variance (128)
Conv2D
kernel (3×3×128×128)
bias (128)
ReLU
kernel (3×3×128×64)
bias (64)
ReLU
BatchNormalization
gamma (64)
beta (64)
moving_mean (64)
moving_variance (64)


Conv2DTranspose
kernel ⟨3×3×64×64⟩
bias ⟨64⟩
ReLU
Concatenate
BatchNormalization
gamma ⟨192⟩
beta ⟨192⟩
moving_mean ⟨192⟩
moving_variance ⟨192⟩
Conv2D
kernel ⟨3×3×192×128⟩
bias ⟨128⟩
ReLU
BatchNormalization
gamma ⟨128⟩
beta ⟨128⟩
moving_mean ⟨128⟩
moving_variance ⟨128⟩
Conv2D
kernel ⟨3×3×128×64⟩
bias ⟨64⟩
ReLU
BatchNormalization
gamma ⟨64⟩
beta ⟨64⟩
moving_mean ⟨64⟩
moving_variance ⟨64⟩
Conv2DTranspose
kernel ⟨3×3×64×64⟩
bias ⟨64⟩
ReLU
Concatenate
BatchNormalization
gamma ⟨192⟩
beta ⟨192⟩
moving_mean ⟨192⟩
moving_variance ⟨192⟩
Conv2D
kernel ⟨3×3×192×128⟩
bias ⟨128⟩
ReLU
BatchNormalization
gamma ⟨128⟩
beta ⟨128⟩
moving_mean ⟨128⟩
moving_variance ⟨128⟩
Conv2D
kernel ⟨3×3×128×64⟩
bias ⟨64⟩
ReLU
BatchNormalization
gamma ⟨64⟩
beta ⟨64⟩
moving_mean ⟨64⟩
moving_variance ⟨64⟩
Conv2DTranspose
kernel ⟨3×3×64×64⟩
bias ⟨64⟩
ReLU
Concatenate
BatchNormalization
gamma ⟨192⟩
beta ⟨192⟩
moving_mean ⟨192⟩
moving_variance ⟨192⟩
Conv2D
kernel ⟨3×3×192×128⟩
bias ⟨128⟩
ReLU
BatchNormalization
gamma ⟨128⟩
beta ⟨128⟩
moving_mean ⟨128⟩
moving_variance ⟨128⟩
Conv2D
kernel ⟨3×3×128×64⟩
bias ⟨64⟩
ReLU
BatchNormalization
gamma ⟨64⟩
beta ⟨64⟩
moving_mean ⟨64⟩
moving_variance ⟨64⟩
Conv2DTranspose
kernel ⟨3×3×64×64⟩

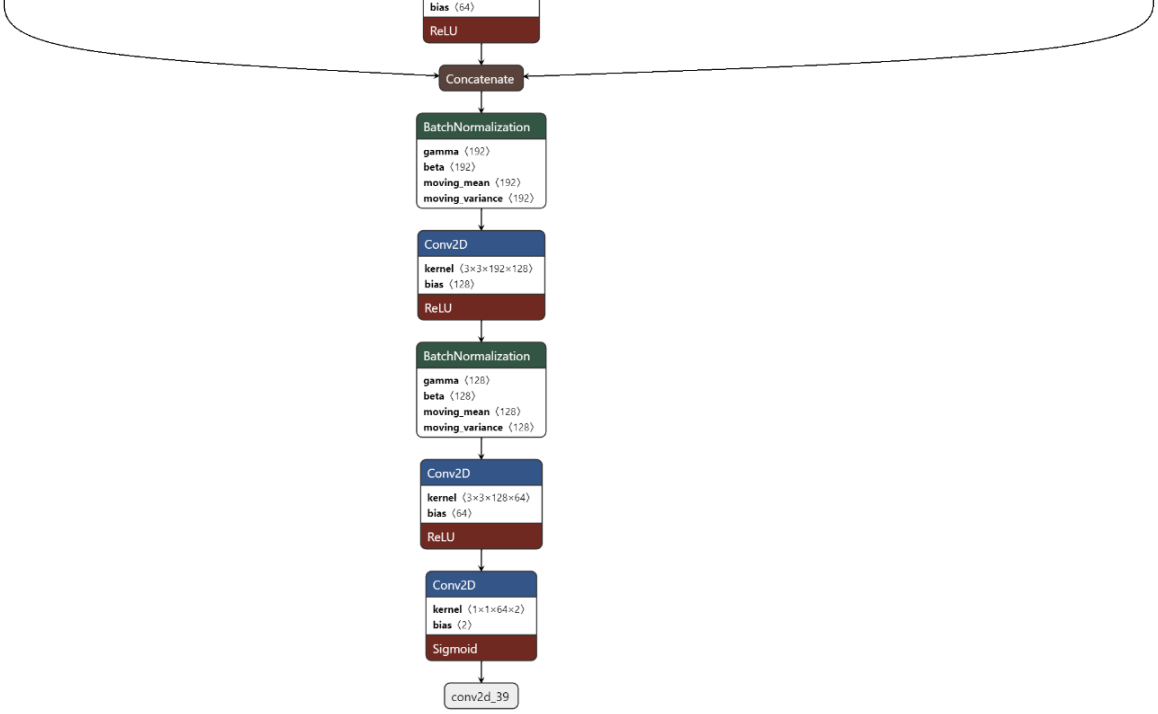


Figure A.1 Network architecture of LF U-Net3 with spatial attention.

Appendix B

Training Curves

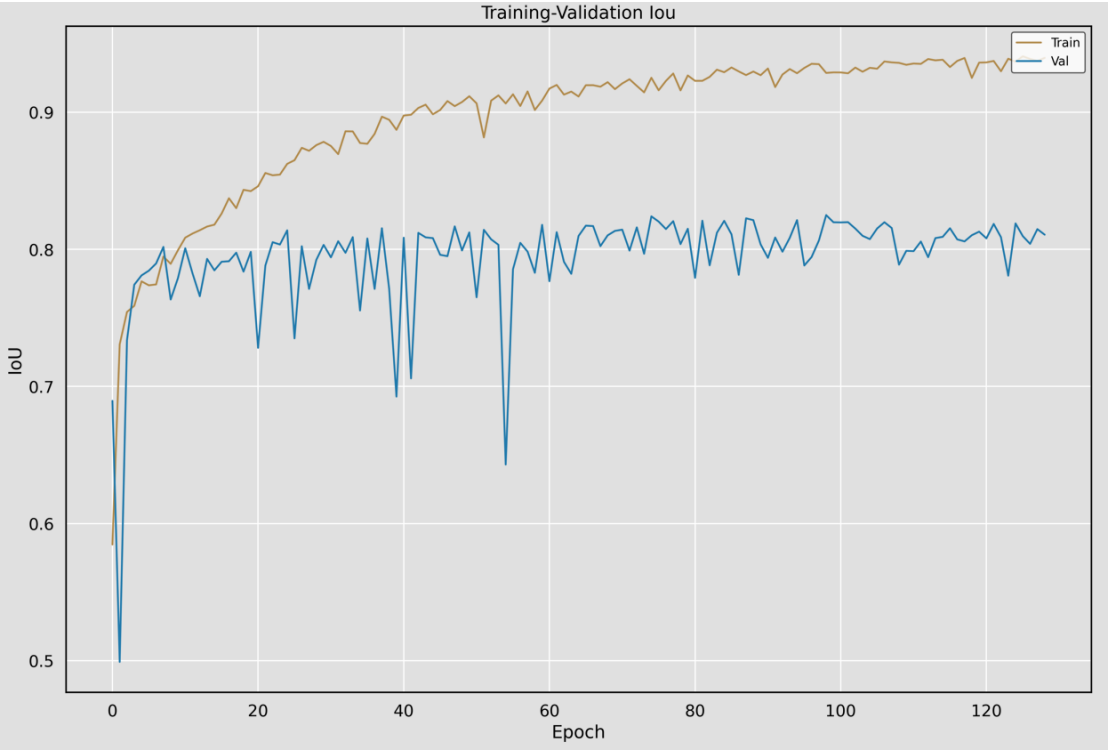


Figure B.1 Training and validation IoU of a U-Net3 with S-1 image

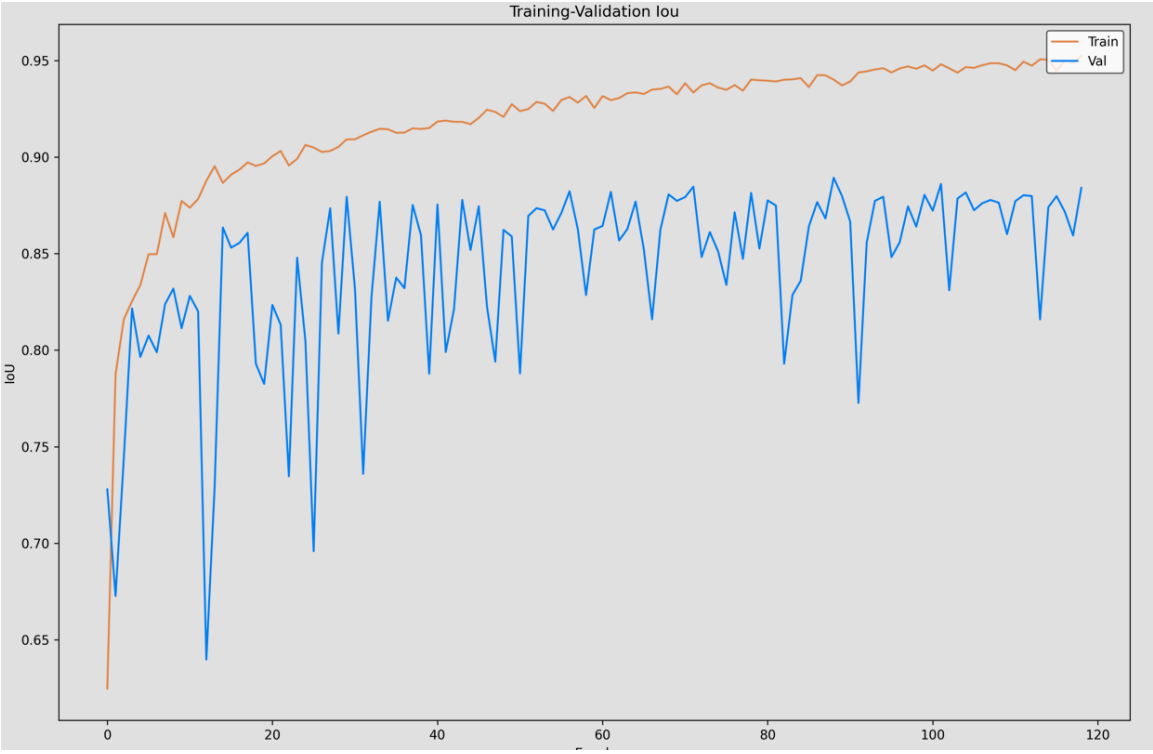


Figure B.2 Training and validation IoU of a U-Net3 with S-2 image

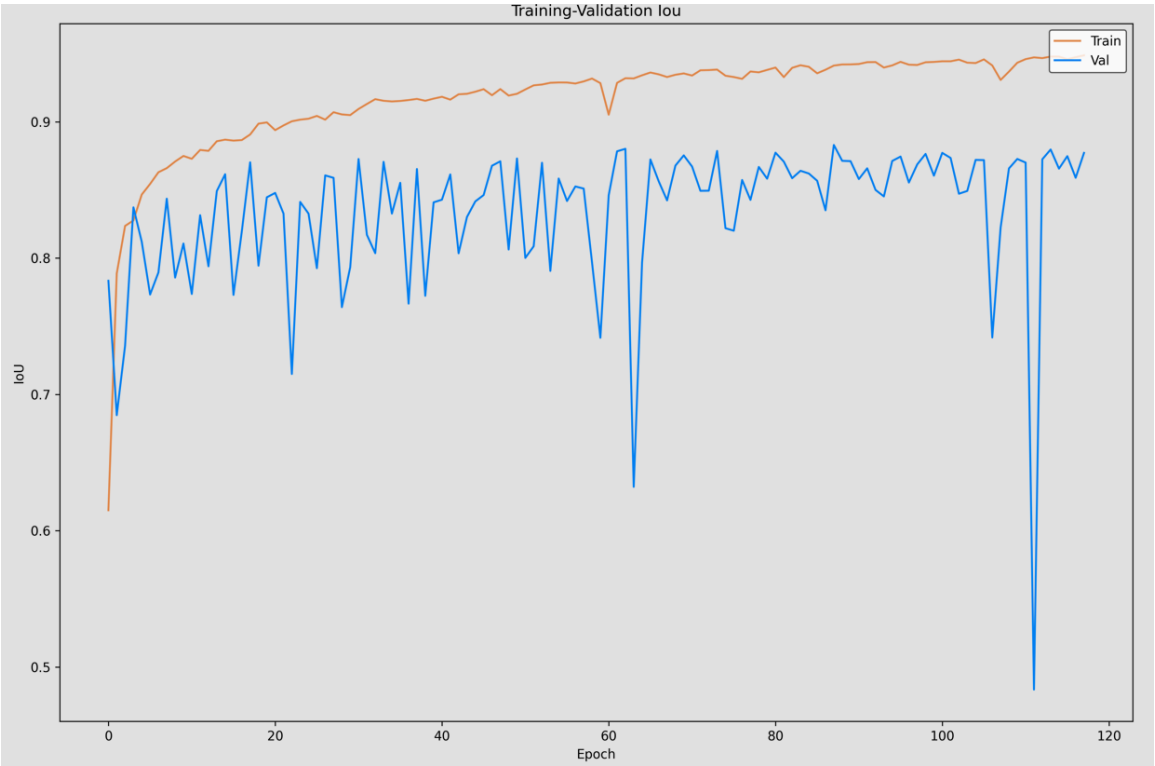


Figure B.3 Training and validation IoU of a U-Net3 EF

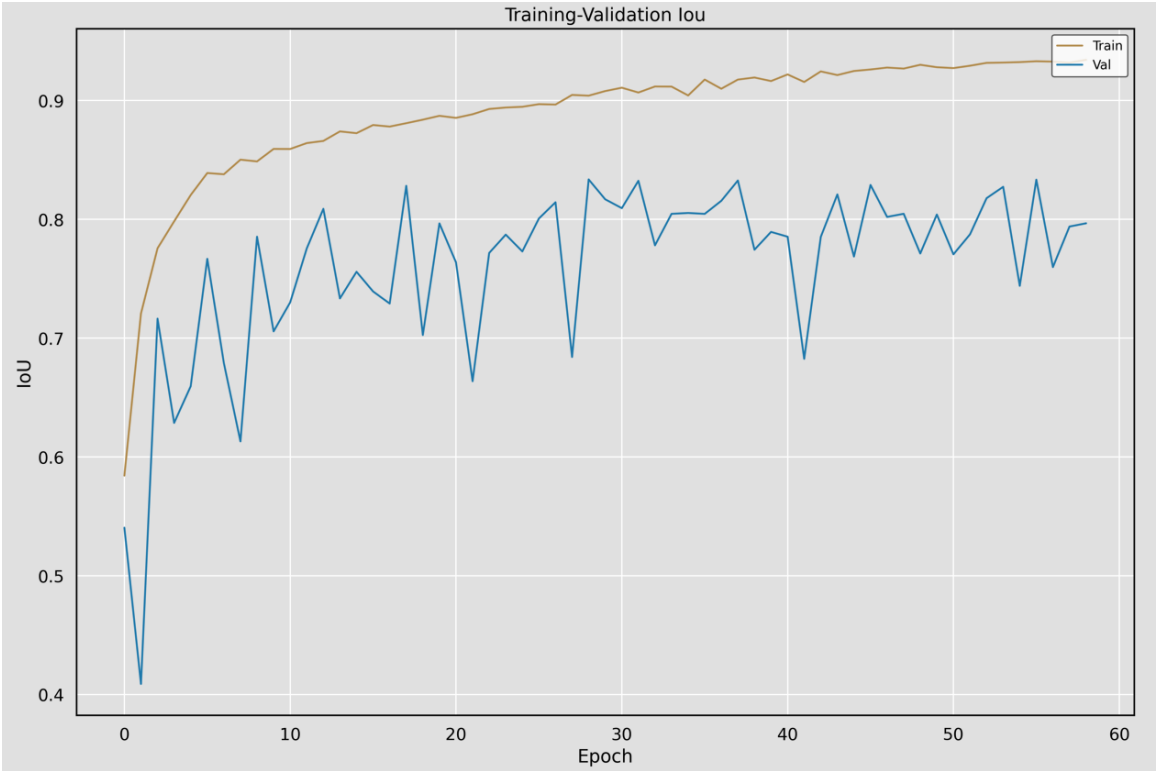


Figure B.4 Training and validation IoU of a U-Net3 LF

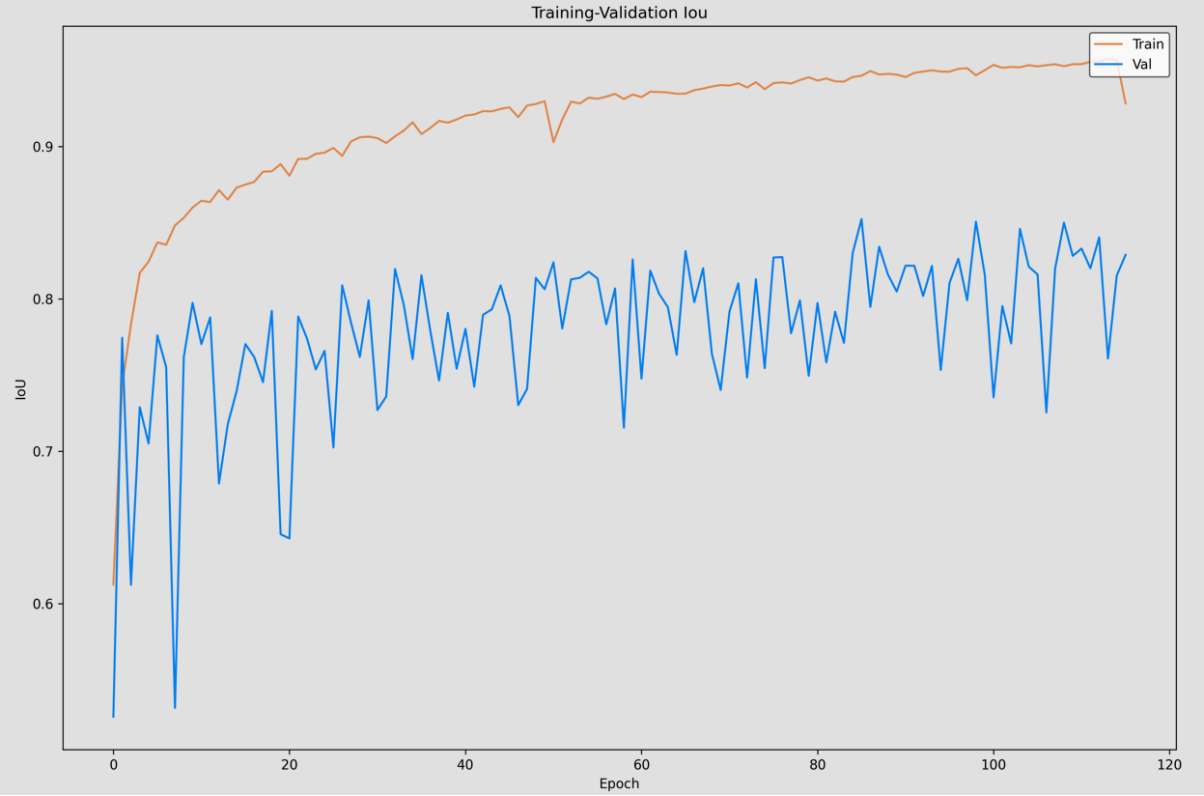


Figure B.5 Training and validation IoU of U-Net3 LF with spatial attention on optical input

Figure B.6 Training and validation IoU of U-Net3 LF with spatial attention on both input

Appendix C

Feature Map Visualization

Extracted features maps and attention maps from a single patch of cloudy image using U-Net3 LF spatial attention on both inputs is visualized here.

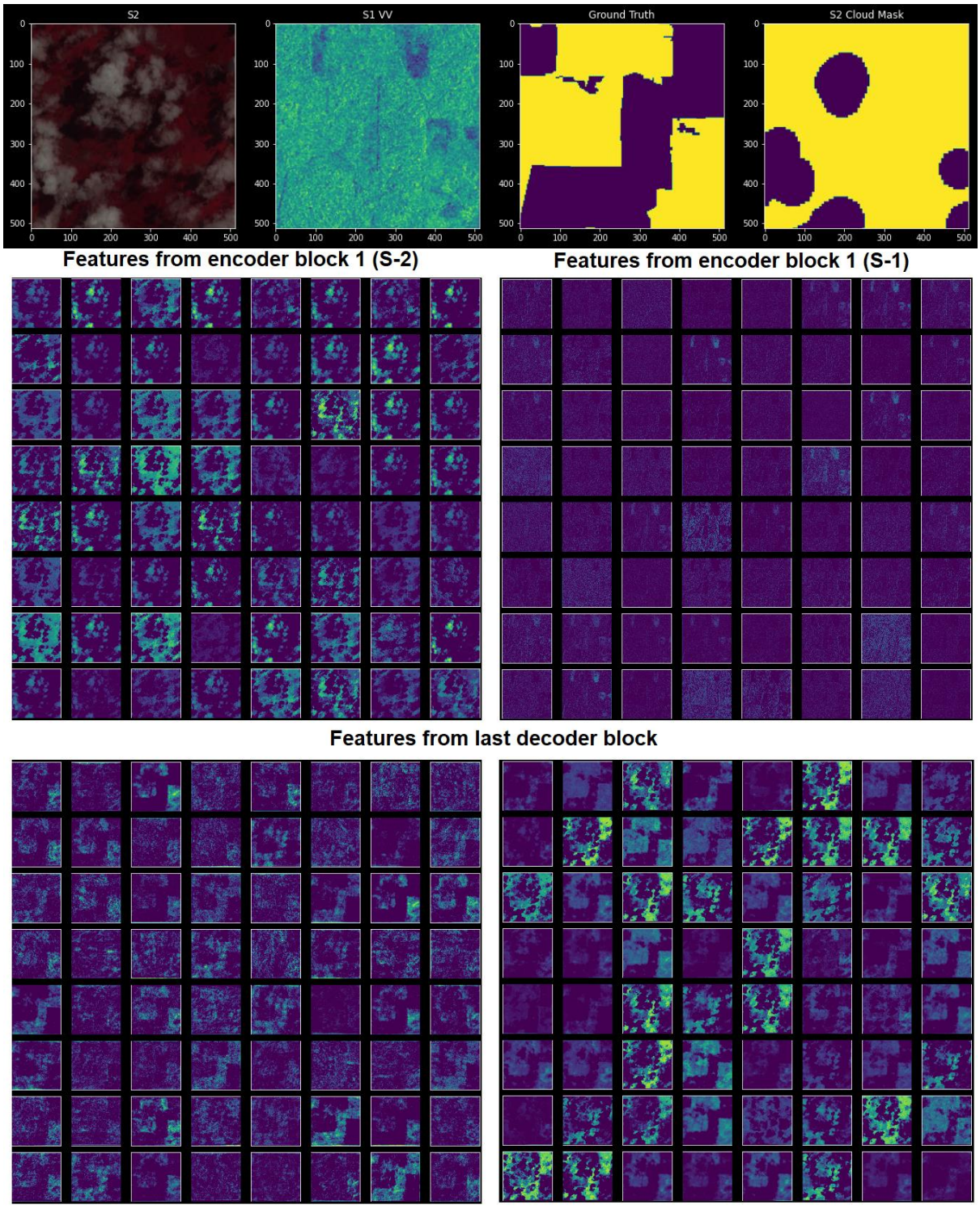


Figure C.1 Example of Input S-2 image patch, S-1 image patch, reference data and S-2 cloud mask (from left to right) and extracted feature maps from first encoder block and last decoder block from the input image

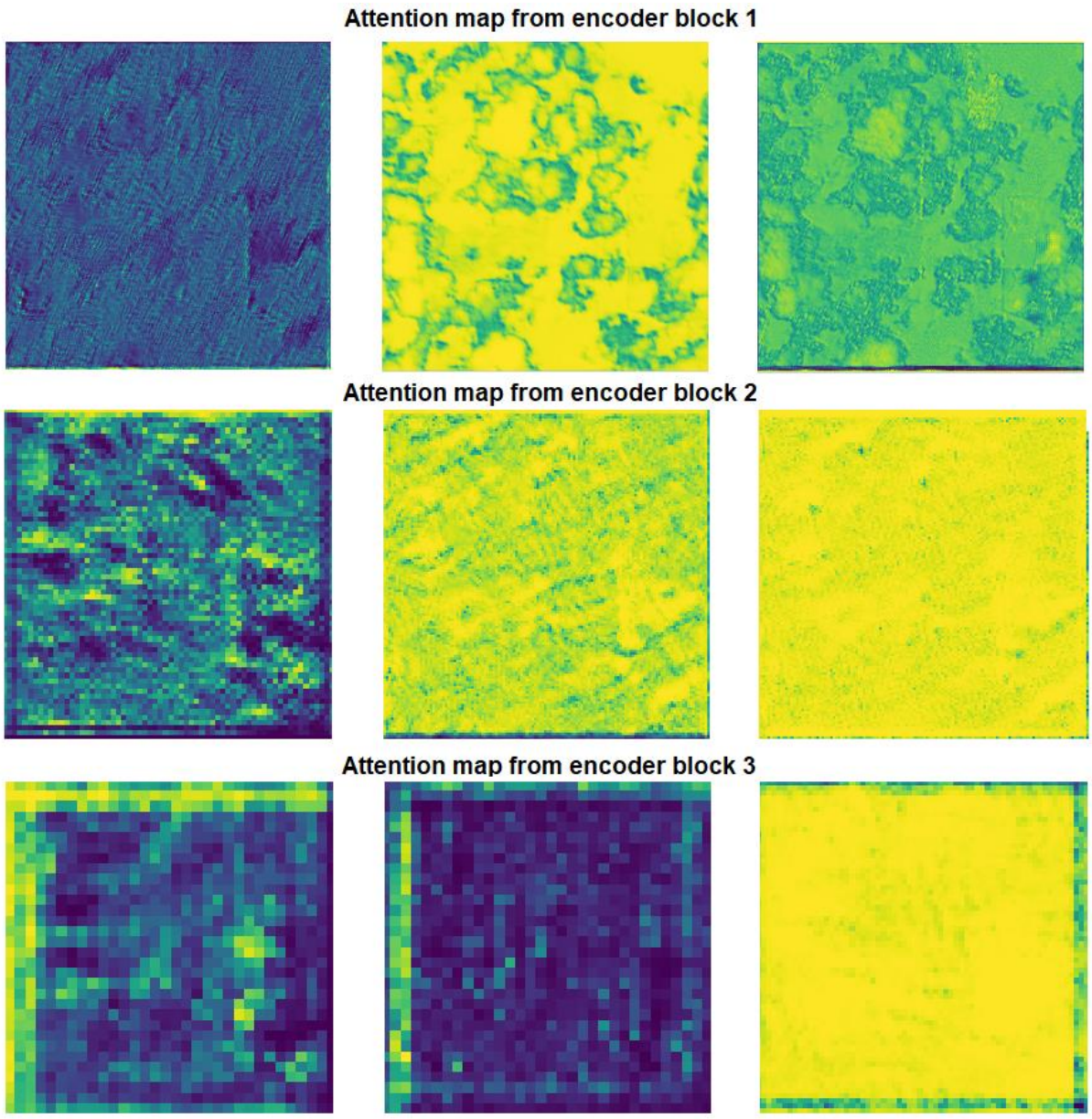


Figure C.2 Attention map at different level of encoder block