SEGMENTATION OF SATELLITE SAR IMAGES USING SQUEEZE AND ATTENTION BASED DEEP NETWORKS

A THESIS SUBMITTED TO

THE GRADUATE SCHOOL OF ENGINEERING AND SCIENCE

OF BILKENT UNIVERSITY

IN PARTIAL FULFILLMENT OF THE REQUIREMENTS FOR

THE DEGREE OF

MASTER OF SCIENCE

IN

COMPUTER ENGINEERING

By

Elmira Khajei

September 2021

Segmentation of Satellite SAR Images Using Squeeze and Attention Based Deep Networks
By Elmira Khajei
September 2021

We certify that we have read this thesis and that in our opinion it is fully adequate, in scope and in quality, as a thesis for the degree of Master of Science.

Ibrahim Körpeoğlu(Advisor)

Sedat Ozer(Co-Advisor)

Ayşegül Dündar

Hazım Kemal Ekenel

Approved for the Graduate School of Engineering and Science:

Ezhan Karaşan
Director of the Graduate School

ABSTRACT

SEGMENTATION OF SATELLITE SAR IMAGES USING SQUEEZE AND ATTENTION BASED DEEP NETWORKS

Elmira Khajei
M.S. in Computer Engineering
Advisor: Ibrahim Körpeoğlu
Co-Advisor: Sedat Ozer
September 2021

Automatic extraction of objects of interests from high-resolution satellite images has been an active research area. Numerous recent papers have investigated on various deep learning-based semantic segmentation techniques for improved segmentation accuracy. Despite the fact that existing literature provides a wealth of information on land cover and land use (e.g., segmentation of structures, roads, and water area), the majority of them have been focused on segmentation on electro-optical-based (EO) images. A recent focus has been segmenting such objects of interest in Synthetic-Aperture-Radar-based (SAR) images to overcome the limitations of using the visible spectrum. While the optical data taken at the visible spectrum is still widely preferred and used in many aerial applications, such applications typically need a clear sky and minimal cloud cover in order to function with high accuracy. SAR imaging is particularly useful as an alternative imaging technique to alleviate such visibility-related problems such as when weather and cloud may obscure conventional optical sensors (as in during severe weather conditions and cloud cover). Recent segmentation techniques use multiple deep solutions based on U-Net. Recent attention based developments in deep learning when combined with the SAR image features, segmentation of objects of interests can be increased especially under low visibility conditions. In this thesis, a squeeze and attention based network is proposed for semantic segmentation in satellite SAR images. In particular, we show how squeeze and attention concept can be used within a U-Net based architecture for segmenting objects of interests in remote sensing images and study its performance on multiple public datasets. Our experiments demonstrate our proposed method yields superior results when compared to multiple baseline networks on all the used datasets.

Keywords: Semantic segmentation, SAR images, Attention and squueze, U-Net, EO images, Building footprints.

ÖZET

UYDU-BAZLI SAR IMGELERINDE KISIK DIKKAT ODAKLI DERIN OGRENME KULLANAN SEGMENTASYON ALGORITMASI

Elmira Khajei
Bilgisayar Mühendisliği, Yüksek Lisans
Tez Danışmanı: Ibrahim Körpeouğlu
İkinci Tez Danışmanı: Sedat Ozer
Eylul 2021

Yüksek çözünürlüklü uydu görüntülerinden ilgili nesnelerin otomatik olarak çıkarılması aktif bir araştırma alanı olmuştur. Pek çok güncel makale, gelişmiş bölümlendirime doğruluğu için derin öğrenmeye dayalı çeşitli anlamsal bölümlendirme teknikleri hakkında araştırmalar yapmaktadır. Mevcut literatür arazi örtüsü ve arazi kullanımı (örn. yapıların, yolların ve su alanının bölümlendirilmesi) ile ilgili zengin bilgiler sağlasa da, bunların çoğu elektro-optik tabanlı (EO) görüntüler üzerinde bölümlendirmeye odaklanmıştır. Güncel çalışmaların bir diğer odağı, görünür tayf kullanma sınırlamalarının üstesinden gelmek için bu tür ilgi nesnelerini Sentetik-Açıklık-Radarı tabanlı (SAR) görüntülerde bölümlere ayırmak olmuştur. Görünür tayfda alınan optik veriler hala birçok hava uygulamasında yaygın olarak tercih edilir ve kullanılırken, bu tür uygulamalar yüksek doğrulukla çalışmak için tipik olarak açık bir gökyüzüne ve minimum bulut örtüsüne ihtiyaç duyar. SAR görüntüleme, hava ve bulutun geleneksel optik sensörleri engellemesi (şiddetli hava koşulları ve bulut örtüsü sırasında olduğu gibi) gibi görünürlükle ilgili sorunları hafifletmek için alternatif bir görüntüleme tekniği olarak yararlı olmaktaktadır. Güncel segmentasyon teknikleri, U-Net'e dayalı pek çok derin öğrenme çözümleri kullanır. İlgi ağı temelli derin öğrenmedeki son gelişmeler, SAR görüntü özellikleri ile birleştirildiğinde, özellikle düşük görüş koşullarında ilgi duyulan nesnelerin bölümlendirilmesini artırılabilir. Bu tezde, uydu SAR görüntülerinde anlamsal bölümlendirme için sıkıştırma ve ilgi tabanlı bir ağ önerilmiştir. Özellikle, uzaktan algılama görüntülerindeki ilgi nesnelerini bölümlere ayırmak için U-Net tabanlı bir mimaride sıkıştırma ve ilgi kavramının nasıl kullanılabileceğini gösteriyoruz

ve çok sayıda halka açık veri kümesi üzerindeki performansını inceliyoruz. Deneylerimiz, önerilen yöntemimizin kullanılan tüm veri kümelerinde çok sayıda temel ağla karşılaştırıldığında üstün sonuçlar verdiğini göstermektedir.

Anahtar sözcükler: Semantik bölütleme, SAR görüntüleri, Dikkat ve sıkıştırma, UNet,EO görüntüleri..

Acknowledgement

The completion of this thesis would never be possible without the exceptional tolerance, magnificent knowledge on the field, and of course, cordially professional behavior of my co-advisor, Dr.Sedat Ozer, who did no less than any perfect advisor.

I would also want to extend my gratitude to my advisor Prof. Ibrahim Körpeouğlu, who immensely helped and guided me through my days of tension.

A debt of gratitude is also owed to my parents and my brothers, especially Rahim and Aydin, for always being there when I needed them the most and for always having my back.

Last but not least, I would like to thank my true friends, especially Mert, who stood by me in sickness and in health, never stopped challenging me, and encouraging me to develop my ideas.

...

Contents

List of Figures

List of Tables

Chapter 1

Introduction

1.1 Motivation

We have a sense that allows us to preserve the environment's variation modifications. This sense enables us to make suitable or desired adjustments to the environment. When we take a step back and extrapolate this basic process to a global scale, we see a pressing need to comprehend complicated phenomena such as urbanization, climate change, biodiversity research, and socioeconomic trends. This technique is called Earth observation and has various uses, including disaster assistance and resource management [3]. Earth observation data is gathered in several ways; generally classified as remote and proximal sensing. The former is used when "the distance between the item and the sensor is much greater than the sensor's linear dimensions", while the latter is used when this distance is similar to the sensor's linear dimensions [4].

Synthetic Aperture Radar (SAR) is a unique kind of radar that can penetrate clouds, gather data in all weather situations, and collect data day and night. Overhead data collected by SAR satellites may assist disaster response efforts when weather and cloud cover impede traditional electro-optical sensors. Despite

these benefits, researchers have limited access to data on the efficacy of SAR for such applications, especially at ultra-high resolutions[5].
Machine learning research is predicated on the notion that a machine may be taught to learn the same way a person does; without being explicitly programmed. Deep learning is a branch of machine learning that uses a family of algorithms known as neural networks and their variations. These techniques supply the network (or model) with a collection of labeled instances from which it may learn or train. There are numerous examples of deep learning models on computer vision like [6]. These samples may be labeled in a variety of ways. Collaborative systems like OpenStreetMap and Crowdsourcing markets are perfect for annotating pictures, and this current volume may be used immediately [7].
Segmentation of images semantically is essential for image understanding and computer vision. Semantic segmentation is interested in predicting pixel-level labels in pictures, and therefore may be seen as a density prediction problem. Significant advances in obtaining reliable results have been made possible by the advent of the Convolutional Neural Network (CNN).

1.2 Objectives and Scope

The primary objective of this thesis is to design, construct, and experimentally analyze a deep neural network for semantic segmentation on SAR images using squeeze and attention based networks. Semantic segmentation on SAR images is one of the difficult problems in computer vision.
The resulting pipeline must include image preparation algorithms capable of dealing with input pictures of variable quality, resolution, and channel count. The following steps may be established for this purpose:

1. Provide brief review of the research work on semantic segmentation.

2. Construct a functional deep network pipeline that accepts the different types of remote sensing images (active and passive) and generates semantically segmented maps on the images.

3. Compare different neural network structures specified in the existing literature and fine-tune them to the present problem.

4. Evaluate the proposed network using four distinct datasets and compare them with other networks.

1.3 Structure of Thesis

The project is presented in six consecutive chapters: the first chapter serves as an introduction and motivation for the work. The second chapter is devoted to background research on the subject. Topics of SAR images and satellite and earth observation are in chapter three. The fourth chapter discusses the model that we have implemented. Chapter 5 discusses evaluation metrics and datasets and reports our results, both qualitative and quantitative, obtained via the use of the evaluation metrics discussed earlier. Chapter six concludes this thesis.

Chapter 2

Related Work

This chapter provides a review of the recent research and studies on semantic segmentation topics. Plenty of algorithms have been proposed for semantic segmentation tasks during the past ten years. The algorithms can be sorted into two categories: traditional image processing-based and CNN-based methods. Traditional-based algorithms had feature extraction parts and incorporated characteristics features. However, the extracted features depend on illumination conditions and sensor type. Features differ under various conditions; therefore, traditional methods could only solve specific issues on specific data[8].

Simonetto et al.[9] discuss the automated extraction of three-dimensional (3-D) buildings from high-resolution stereoscopic images captured by the French Aerospace Research Center's SAR airborne RAMSES sensor (ONERA). The article presents a two-step modified processing method. The first step is to derive stereoscopic structure from L-shaped echoes. The Hough transform is used to recognize buildings in each image. Then, based on a criteria optimization, they are identified during a stereoscopic refining step. The second step is the height measurement.

By prospering Deep Convolutional Neural Network (DCNN) and because of the variety in building shape and appearance, handcrafted feature-based algorithms were replaced by learning feature-based algorithms.

He et al.[10] were able to solve the gradient explosion in the propagation problem

and make it possible to design deep convolutional networks and leverage richer semantic features.

Furukawa[11] proposed CNN end-to-end automatic target recognition network, named Verification Support Network (VersNet). It takes an arbitrary-sized SAR image with many classes and targets as input and generates a SAR ATR (Automatic Target Recognition) image, including each identified object's position, class, and pose. It consists of encoder-decoder parts. This paper achieved a mean IOU score of 0.91 in the MASTER Dataset.

The purpose of the article [12] is to solve the complicated issue of automatically identifying man-made objects, particularly buildings, in very high resolution (VHR) synthetic aperture radar (SAR) images. The article makes two significant contributions in this context: To begin, it presents a workflow for classifying spaceborne TomoSAR point clouds generated by processing VHR SAR image stacks using advanced interferometric techniques known as SAR tomography (TomoSAR) into buildings and non-building using additional information. Second, these labeled datasets (i.e., building masks) were used to construct and train Deep Fully Convolutional Neural Networks with an additional Conditional Random Field represented as a Recurrent Neural Network to detect building regions in a single VHR SAR image. Although this cascaded structure has been effectively used in computer vision and remote sensing for optical image categorization, it has not been used in SAR images. The findings of the building identification algorithm are shown and verified using a TerraSAR-X High-resolution spotlight SAR picture spanning about 39 km^2, with mean pixel accuracies of around 93.84 percent.

Unet [13] was proposed to biomedical image segmentation, and since it was a milestone in the semantic segmentation task, later, lots of papers modified Unet to get better results.

Evolving from CNN "fully convolutional" networks that take input of the arbitrary size and produce correspondingly-sized output with efficient inference and learning. [14] In this work, they clarified FCN's usage in pixel-level label prediction and showed its connection to prior models like AlexNet, the VGG net, and GoogLeNet. They adopted these classification networks and transferred them to the segmentation task. Then, they made a new architecture that achieves (20%

relative improvement to 62.2% mean IOU on 2012) segmentation of PASCAL VOC.

Another architecture is the so-called Segnet[15]. Segnet consists of an encoder and corresponding decoder parts with a pixel-wise classification layer. Although the encoder part of the segment is almost the same as the VGG16 network (13 convolutional layers), they have novelty in that the decoder upsample its lower resolution input feature maps. To demonstrate that, the decoder uses pooling indices computed in the max-pooling steps of the corresponding encoder. They compared their architecture with Fully Connected Network (FCN) and DeepLab-LargeFOV architectures. They improve mIOU from 0.53 to 0.60.

Tobias et al.[16] claim that while current networks perform well in recognition (detecting objects), they fall weak in terms of localization accuracy. Thus, the authors suggest a novel architecture that combines context at many scales with pixel-level precision. One stream transmits data at full picture resolution, allowing for exact segment boundary adherence. The other stream is subjected to a series of pooling operations to acquire robust recognition characteristics. They assessed their network using the Cityscapes Dataset and arrived at an IOU of 71.8 %.

Feature pyramids are a fundamental component of object recognition algorithms that recognize things of various sizes. Pyramid representations have been avoided in some deep learning object detectors because they are computed and memory expensive. The intrinsic multi-scale, pyramidal structure of deep convolutional networks is exploited [17] to create feature pyramids for a minimal additional cost. To create high-level semantic feature maps at various sizes, a top-down architecture with lateral connections is created. This design, known as a Feature Pyramid Network (FPN), demonstrates substantial improvement as an available feature extractor in some applications. The proposed technique delivers state-of-the-art single model results on the COCO detection benchmark using FPN in a simple Faster R-CNN system.

To perform semantic segmentation in real life, networks should not have many floating-point operations. To this aim, Enet (efficient neural network) [18] proposes a network that is up to 18× faster, requires 75× fewer FLOPs, has 79× fewer parameters, and provides similar or better accuracy to existing models.

They used Cityscapes, CamVid, and SUN Dataset to evaluate their work.
Linknet[19] uses only 11.5 million parameters and 21.2 GFLOPs for processing an image of resolution $3 \times 640 \times 360$. Dataset is trained and tested on CamVid and Cityscapes datasets. It achieves an mIOU of 76.4 in the Cityscapes Dataset.

DeepLabv3+ [20] which is an extended form of DeepLabv3, leveraged from the Spatial pyramid pooling module and encoder-decoder structures together. These networks are used in DCNN for semantic segmentation tasks. While the Spatial pyramid pooling networks can encode multi-scale contextual information, encoder-decoders can capture sharper object boundaries by gradually recovering the spatial information. So, this article combines these two modules. The authors further explore the Xception model and apply the depthwise separable convolution to Atrous Spatial Pyramid Pooling and decoder modules, resulting in a faster and stronger encoder-decoder network. They evaluated their proposed model on PASCAL VOC 2012 and Cityscape datasets, achieving the test set performance of 89.0% and 82.1% without any post-processing
Due to the importance of the real-time semantic segmentation challenge, many papers focus on this task. There are many practical applications, yet there is a fundamental difficulty in reducing a large portion of computation for pixel-wise label inference.Deep learning-based approaches are used in today's state-of-the-art object tracking, identification, and segmentation algorithms. Techniques usually need a lot of computation and a lot of memory. as well as energy resources. [21] Romera et al. [22] To solve this issue, developed an image cascade network (ICNet) that integrates multi-resolution branches under proper label direction. The report introduces the cascade feature fusion unit to enable rapid segmentation of high-quality features. Moreover, Cityscapes, CamVid, and COCO-Stuff datasets have been used to demonstrate the success of their work.
Previous research focusing on high-speed inference has struggled to generate high-accuracy segmentation results. Efficient Dense modules with Asymmetric convolution (EDANet) [23] is another convolutional network that claims to be 2.7 times quicker than the previous rapid segmentation network, ICNet while achieving the same mIoU score. Dilated convolution and dense connectivity are used in EDANet's asymmetric convolution structure. They tested their network using

datasets from Cityscapes and CamVid.

As mentioned above, the importance of semantic segmentation tasks in real life and even applying its models on mobile devices has been increasing rapidly. To tackle the problem of heavy complication cost and lots of FLOPS, Yuan Lo et al.[24] proposed a new network called Context Guided Network (CGNet), which they claim as a lightweight and efficient network for semantic segmentation. The Context Guided (CG) block in CGNet learns the standard features of both the local and surrounding features and improves the common features with the global context. The number of parameters in CGNet is considerably decreased. Experiments on Cityscapes and CamVid datasets yielded a mean IoU of 64.8 percent on Cityscapes with less than 0.5 M parameters.

The other network that addresses semantic segmentation problems with a good trade-off between high quality and computational resources is the RFNet [25] network. The core of this approach is a layer that uses residual connections and factorized convolutions to remain efficient while retaining good accuracy. To show the effectiveness of the model, they evaluate the network on the cityscape Dataset. Although this network achieves an accuracy that is similar to state of the art, it is faster than the previous networks.

Even though many papers emphasize high accuracy at a low cost of computation, state-of-the-art results from networks with a large number of convolutional layers and feature channels make semantic segmentation a computationally expensive task, which is a disadvantage in a scenario with limited resources. To solve this problem, Wang et al. [26] claim to have created an efficient symmetric network named (ESNet). ESNet is an encoder-decoder network with a nearly symmetric design based on factorized convolution units (FCU) and parallel equivalents. In addition, the FCU uses a standard 1D factorized convolution in residual layers. In addition, in the design of the residual module, the parallel version uses a transform-split-transform-merge method, where the split branch uses dilated convolutions with various rates to extend the receptive field. The authors demonstrated that their method archives deliver state-of-the-art outcomes in terms of speed and accuracy trade-off for real-time semantic segmentation on

the Cityscapes Dataset.

The recent incorporation of attention mechanisms into segmentation networks enhances their representational capacities through a strong focus on more informative characteristics.

For accurately extracting coarse-to-fine building characteristics[27] developed a lightweight attention mechanism-based model — refined cross attention neural network (RCA-Net). The RCA-Net captures the long-range multi-scale context via the use of spatial and channel attention. Then, they propose an efficient attention module, the Global Attention Fuse (GAF), which fuses local and global cross-channel connections to capture essential characteristics without increasing computational complexity. Additionally, a loss function called unified loss is given that combines BCE and dice loss to address unbalanced class distribution. Their suggested approach beats the most recent method DSNet by 2.06% and 1.47 percent in IoU and 2.11 percent and 1.27 percent in F1-score, respectively, using two publically accessible datasets: the Massachusetts roads dataset and the Inria Aerial Image Labeling dataset.

The article [28] proposes a technique for detecting SAR image targets based on a visual attention model that combines bottom-up and top-down processes. At the bottom-up stage, the proposed approach modifies the conventional Itti model to account for the peculiarities of SAR images and target recognition tasks. They offer a new top-down learning method for determining the optimum weights required to build a saliency map. Their suggested approach is unique in three ways. First, the new weighting function facilitates the identification of numerous targets. Second, top-down signals are added to help the user to choose the appropriate weights for the training session. Finally, the judgment step utilizes previous information about the target, such as its area and length, as thresholds, ensuring that the final choice is trustworthy. The simulation experiments demonstrate that the suggested approach is more capable and resilient than existing state-of-the-art visual models and detection methods, such as CFAR and YOLOv2.

[29] proposes a novel Attention Graph Convolution Network (AGCN) for performing super pixel-wise segmentation in extensive SAR imagery data. An attention mechanism layer and Graph Convolution Networks make up AGCN (GCN).

By generalizing convolutions to the graph domain, GCN can operate on graph-structure data and has been successfully applied in tasks such as node classification. The attention mechanism layer is introduced to guide the graph convolution layers to focus on the most relevant nodes to make decisions by assigning different coefficients to different nodes in a neighborhood. Because the attention layer comes before the convolution layers, noisy information from nearby nodes has a lower impact on the attention coefficients. Experiments on two datasets of airborne SAR images show that the proposed method outperforms other state-of-the-art segmentation methods. It also performs significantly faster than current pixel-level semantic segmentation networks.

[30] presents a region-merging-based technique for synthetic aperture radar image segmentation, in which the merging cost is a combination of the texture pattern similarity measure (TPSM), the statistical similarity measure (SSM), and the relatively common border length penalty, among other things (RCBLP). The segmentation process is broken down into three phases. The image is first over segmented using the multi-scale Bhattacharyya distance to produce an initial partition of significant regions, after which the image is segmented again. Second, areas with sizes less than a certain threshold are required to be combined to produce a middle segmentation. In the third step, a region-merging procedure is carried out iteratively, utilizing the new merging cost to obtain the final segmentation. Because of the inclusion of the TPSM in the merging cost, the new technique eliminates the incorrect merging of neighboring areas with different textures, which would otherwise occur.

Due to the massive size of the original synthetic aperture radar image,[31] divides the input image into small slices, and then the pieces are reassembled. The picture slices are fed into an attention-based, fully convolutional network, which produces the segmentation outcomes. Finally, to improve the segmentation performance of the network, the fully connected conditional random field is used as a final step. The followings are some of the method's new features: 1) The multi-scale attention network is embedded within the attention-based fully convolutional network, which is capable of enhancing the extraction of image features through three strategies: multi-scale feature extraction, channel attention extraction, and spatial attention extraction. 2) The attention-based fully convolutional

network incorporates the multi-scale attention network, which can enhance the extraction of image features through two strategies: multi-scale feature extraction and channel attention extraction. 3) A novel loss function for the attention fully convolutional network is created by merging Lovasz-Softmax and cross-entropy losses and is applied to the network. In addition, the novel loss enables simultaneous optimization of the intersection over union and the pixel classification accuracy of the segmentation results obtained using the new method. The trials are carried out on two aircraft equipped with synthetic aperture radar.

Chapter 3

Overview

Remote sensing is an important research field, and a tremendous amount of work has been done in this field. One of the sub-research topics in remote sensing is segmenting regions of interest in the data collected by the remote sensing sensors. There are two broad categories of sensors used for remote imaging: passive and active imaging. Passive imaging techniques, such as electro-optical (EO) imaging, acts as a passive receiver while active imaging sensors typically act as a transmitter and a receiver (trans-receiver) at the same time, such as synthetic aperture radar (SAR) imaging.

While EO-based (RGB) images are commonly used in many remote sensing applications, they are limited by the visual availability of the scene. Therefore, in aerial imaging clouds, stormy or rainy atmospheric effects or time of the imaging time (night vs. day) that limit what can be seen visually may not be the best sensor types to be used in many applications. Alternative approaches use active sensor types to deal with this problem. Digital elevation models (DEM) and SAR are two types of sensors (and modality) for such active sensors operating on different wavelengths.

AI-based segmentation techniques have been used in many standard computer vision applications that typically operate on RGB images obtained by EO-based

sensors. When the goal is remote sensing in active imaging techniques, including SAR and DEM images, the images show different characteristics since they are taken at different wavelengths. Consequently, pre-trained models that are trained on standard (RGB) images would not be effective to be used for SAR and DEM images. New techniques and models are needed. Consequently, in this thesis, we study the performance of different and recent deep learning-based techniques on SAR and DEM images. We introduce a new architecture that would yield better than several benchmarking segmentation models on both SAR and DEM images in our preliminary experiments.

3.1 Satellites and Earth Observation

Some satellites have particular missions for placing them in orbit, and one such mission is Earth observation. The first satellite to take images of the Earth was Explorer 6, which was launched in 1959 [7]. Since then, the use of satellite images and the total number of remote sensing satellites have been increased to be used in various applications.

The earth is heavily packed with Low Earth Orbit (LEO) and Medium Earth Orbit (MEO) remote sensing satellites. LEO ranges from 160 to 2000km altitude, while MEO ranges from 2000km to below the geostationary orbit. EO satellites are often positioned in a sun-synchronous orbit that is optimized for their intended purpose. A sun-synchronous rotation is when the sun's location in respect to the satellite and the earth remains constant.

Earth Observation (EO) satellites have applications in a broad range of fields, including mapping, urban planning, disaster relief, real estate management, econometric/social trend research, military intelligence, and climatic studies [32, 7].

For example, with the advancement of automated drone delivery systems and self-driving cars, there is an increased need for satellite images that may be utilized as redundant data for sensor fusion in the vehicles. As a result, it is critical to comprehend urban architecture via images.

3.2 Synthetic Aperture Radar (SAR) Images

In the general public's mind, when individuals are asked to think about "satellite image," they generally imagine an optical image, although the camera obtains these images. However, optical images are not the only means for a satellite or airplane to view the earth's surface.
Synthetic Aperture Radar, or SAR, is a fundamentally different method of generating pictures than optical images, unlike EO images which are passive remote sensing. SAR is an example of active remote sensing which provide their own artificial radiant energy source for illumination. SAR images have a number of advantages as well as drawbacks. As for their advantages, we can state the following [33]:

- Nearly all weather capability
- Day or night capability
- Penetration through the vegetation canopy
- Penetration through the soil
- Minimal atmospheric effects
- Sensitivity to dielectric properties (liquid vs. frozen water)
- Sensitivity to structure

As for the disadvantage, we may state that:

- Information content is different than optical and sometimes difficult to interpret
- Speckle effects (graininess in the image)
- Effects of topography

Random constructive and destructive interference from numerous scattering returns inside a pixel cell causes speckle, which looks like an image's highly grainy salt-and-pepper texture. Even for a single surface type, a thorough examination of radar images reveals that there may be many gray level differences across neighboring resolution cells. These differences cause the graying texture of radar images. So, it is desirable to reduce speckle before interpretation and analysis.

Speckle reduction can be achieved in two ways: multiple looking and spatial filtering. The technique of multi-looking divides the radar beam into many smaller sub-beams. Each of these sub-beams have a distinct "look" and is speckled. However, all of the "looks" may be averaged, which reduces the amount of speckle in the final averaged image [34]. Spatial filtering is another technique for reducing speckles. It entails re-positioning a small window with a size of a few pixels. It iterates over each pixel in the image, mathematical calculations using the pixel values inside or under that window. The center pixel is then substituted with the new value. Thus, the window is shifted one pixel at a time in both the X and Y dimensions until the whole image is covered. A smoothing effect is produced by computing the average of small windows around each pixel, which reduces the optical impression of the speckle. Both multiple looking and spatial filtering reduces speckle at the expense of resolution [33].

Deep learning is being utilized in an increasing number of ways using SAR data, including change detection and land cover categorization. Researchers have even used a generative adversarial network (GAN) to do image "translation" — converting a single-polarization SAR image to a simulated full-color optical image of the same region [34].

Figure 3.1: Comparison of SAR image with Electro Optic image

Chapter 4

Proposed Models

We start this chapter with a brief review of convolutional neural networks and deep learning. Then we explain our networks.

4.1 Neural Networks

A typical neural network (such as fully connected neural networks) consists of neurons. A neuron is the most basic structure of a neural network. A neuron operates on the given input vector xi to perform a linear operation on it and to yield its output y as shown below:

$$y = w^T * x + b \tag{4.1}$$

A neural network is essentially composed of a set of neurons connecting. Where y is the scalar output of the neuron, w^T is the transpose of the weight vector for the input, and the b_i is the scalar value (bias) for the neuron. In a typical neural network, once the linear output y is obtained, it also goes through one more step known as the activation function to add nonlinearity to the network. If we include the activation function $g(.)$, we can give each neuron an index i as shown below

to obtain the final nonlinear output of any given neuron in a network as follows:

$$y_i = g(w_i^T * x_i + b_i) \tag{4.2}$$

We can also write the input and output relations for a stack of neurons in a matrix form.

For an image classification task, this function effectively transforms the input x_i from image space into an image-specific score y_i where y_i can be interpreted as the classifier's confidence on the input belonging to a specific label. The problem of the image segmentation task is formulated as a pixel-based classification. Thus, the final output y_i would represent the class score assigned to a specific pixel of the given input image. To improve the performance of a neural network, multiple layers of neurons are typically stacked in order. A layer defines a set of neurons in a classical neural network. While a single-layer neural network may still produce approximate predictions, more hidden layers along with the use of activation functions can aid in optimizing and refining the network for better accuracy, [35].

There are many resources available on generic neural networks and deep learning, providing a comprehensive exposure as in [36, 37]. Therefore, rather than providing a generic background on deep learning concepts here, we will focus on our proposed deep learning-based architectures in the rest of this thesis.

4.2 Proposed Models

4.2.1 Preliminary Work

First, to segment objects of interest in SAR images accurately, we studied various recent deep architectures. Our earlier models were built on VersNet [11] for SAR images (in particular for SpaceNet6 dataset [5]), and the architecture of VersNet is given in the figure 4.1. Our additions on that model are listed below:

1. Adding drop out after each convolutional layer.
2. Applying batch normalization in the last layer.
3. Changing optimizer from SGD to Adam.
4. Modifying architecture to identify and segment only objects of interest.
5. Using binary cross-entropy as a loss since we have only one foreground class to detect in this case.

VersNet contains four convolutional blocks and two convolutional layers in the encoder section. There is a transpose convolution layer in the decoder section that increases the sampling rate 16 times more. Like VGGNet, each convolution block consists of two convolution layers of kernel size (3x3) and a max-pooling layer. Dropout occurs just in the last layer of the encoder component of the VersNet network. We used dropout as a regularizer after each convolution block and layer to deal with overfitting problems. Additionally, we used Batch Normalization (BN) after each convolution block and two final convolution layers. All layers except for the last one have rectified linear unit activation functions (ReLu). We used binary cross-entropy as the loss function. For the optimization of the loss function, we used Adam. The output dimension of this model was (832 x 832). Figure 4.2 shows the architecture of the modified network in detail. Although this network performance has a high IOU of around 0.60 during training, the results of validation IOU were about 0.30.

Then, by adjusting the final kernel size to (5 x 5), the output size was increased (900,900). Following that, both batch normalization and dropout were evaluated. Validation IOU of 0.30 did not increase with this architecture.

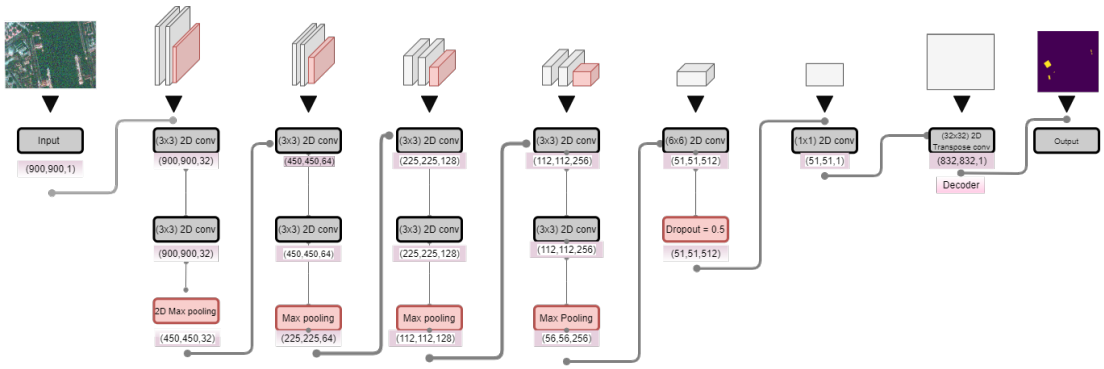


Figure 4.1: Architecture of VersNet

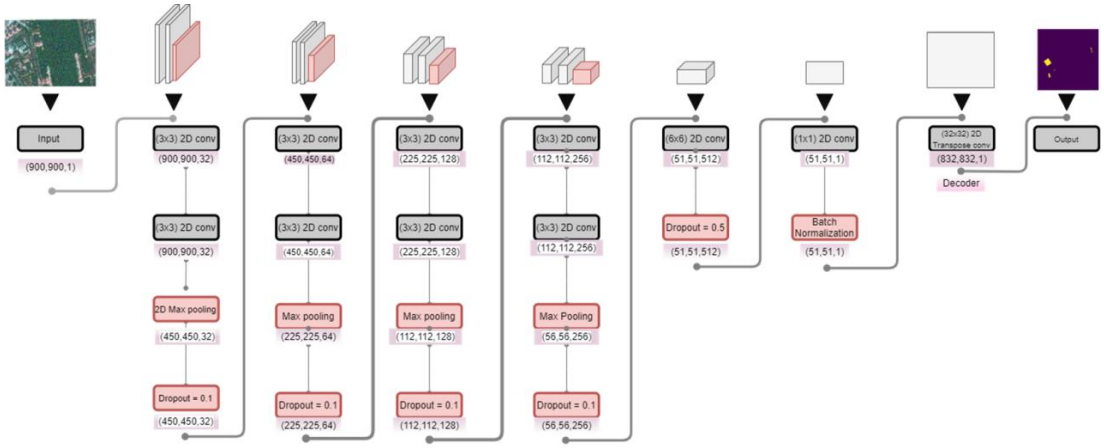


Figure 4.2: Architecture of VersNet with BN and dropouts

U-Net comprises three components: a contracting module, an encoder module, and a decoder module. First, we designed an encoder in the network based on VersNet and a decoder based on U-Net. U-Net architecture is a successfully used model for semantic segmentation in many applications; therefore, we re-designed the decoder part of VersNet based on U-Net. We name our architecture Bilnet. Our skip connection was inserted in the Bilnet design to concatenate activation between layer i in the encoder and $n - i$ in the decoder, where n is the layer number. Additionally, we evaluate the network in a number of configurations, including the following changes or additions:

1. Adding activation function after each convolutional layer,
2. using different configurations with and without batch normalization and dropout after each layer in the encoder part,
3. replacing concatenation operation with adding,
4. reducing the number of skip connections.

Additionally, we also tested Bilnet with dice core loss that surprisingly did not work well.

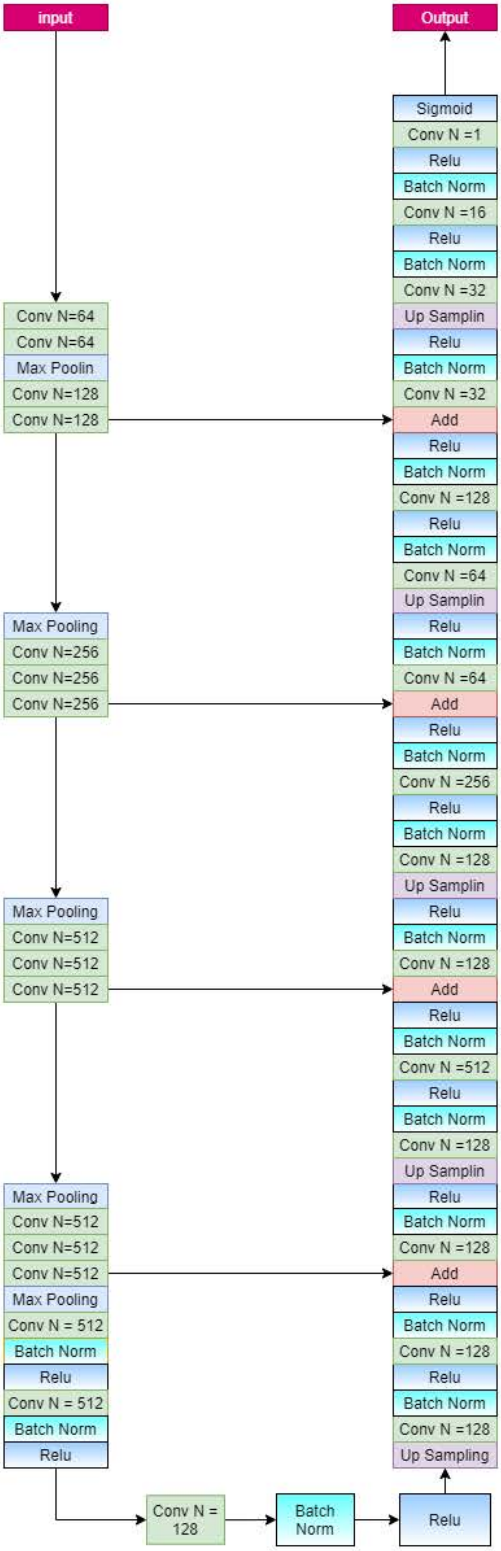


Figure 4.3: Original U-Net

Figure 4.3 demonstrates the baseline U-Net architecture. In the encoder section, there are 16 convolutional layers with (3x3) kernels. As with the decoder portion, five upsampling stages are followed by two convolutional layers using the ReLu activation function. For Bilnet1 shown in Figure 4.4 we reduced the number of convolutional layers to 10 in the encoder part. The encoder part of the network is the same as VersNet. Throughout the network, each convolution employs a 3x3 kernel, and a ReLu function follows each convolution. As upsampling, we used transposed convolution rather than bilinear upsampling. Max-pooling uses a 2x2 kernel with a stride size of 2. We chose a pooling kernel size of 2x2 since a larger number increases the danger of data loss. Dropout was not a feature of the original U-Net Architecture. In our implementation, a dropout follows each convolution block. This additional step also aims to mitigate the overfitting issue.

Figure 4.5 shows Bilnet2, in which we have used residual blocks in both encoder and decoder parts. There is a skip connection before each max pooling until the next one. Since the performance of the network was not good, we came up with Bilnet3.

In Figure 4.6, we have shown Bilnet3, which includes another interpretation of residual block. Here we have skip connection after the first max polling and add it before the max-pooling of the next block. We also have the ReLu activation function after adding skip connections.

4.2.2 Squeeze and Attention Based Segmentation

Due to their improving performance feature in many applications, residual networks (ResNets) are frequently used as the backbone in many segmentation applications. Residual networks include residual blocks, which contain small networks and skip connections. As shown in Figure 4.7a, conventional residual blocks can be formulated as:

$$X_{out} = X_{in} + X_{res} = X_{in} + F(X_{in}; \theta, \Omega)$$
$$X_{res} = F(X_{in}; \theta, \Omega)$$

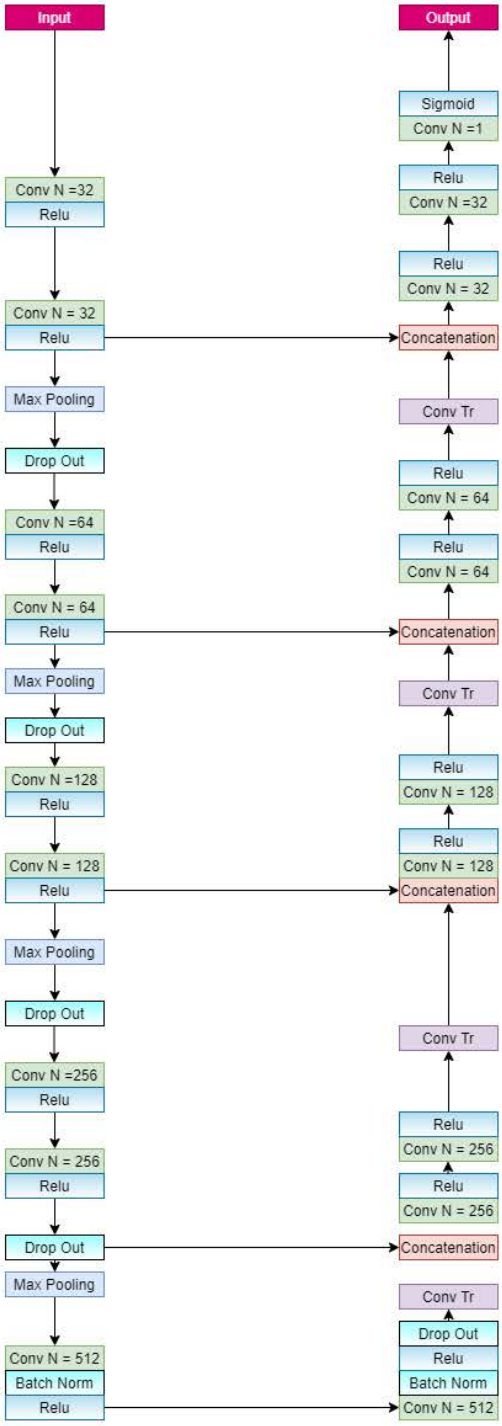


Figure 4.4: Bilnet1

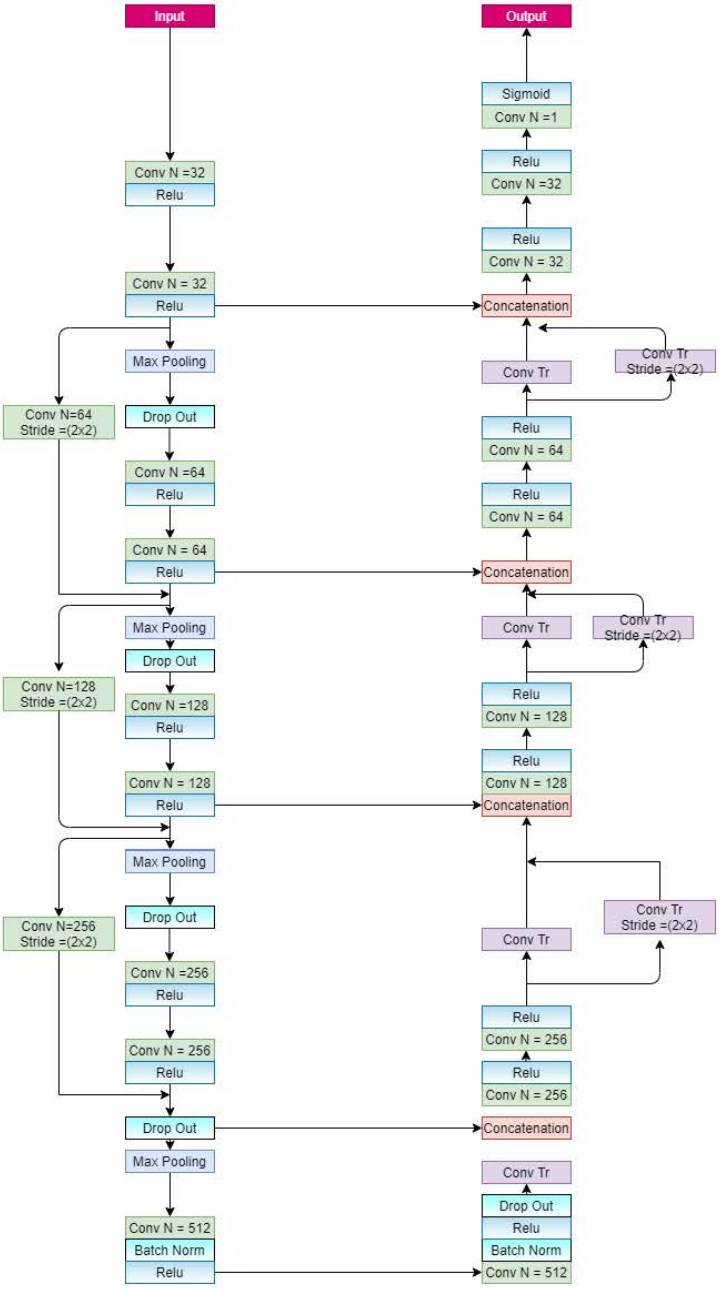


Figure 4.5: Bilnet2

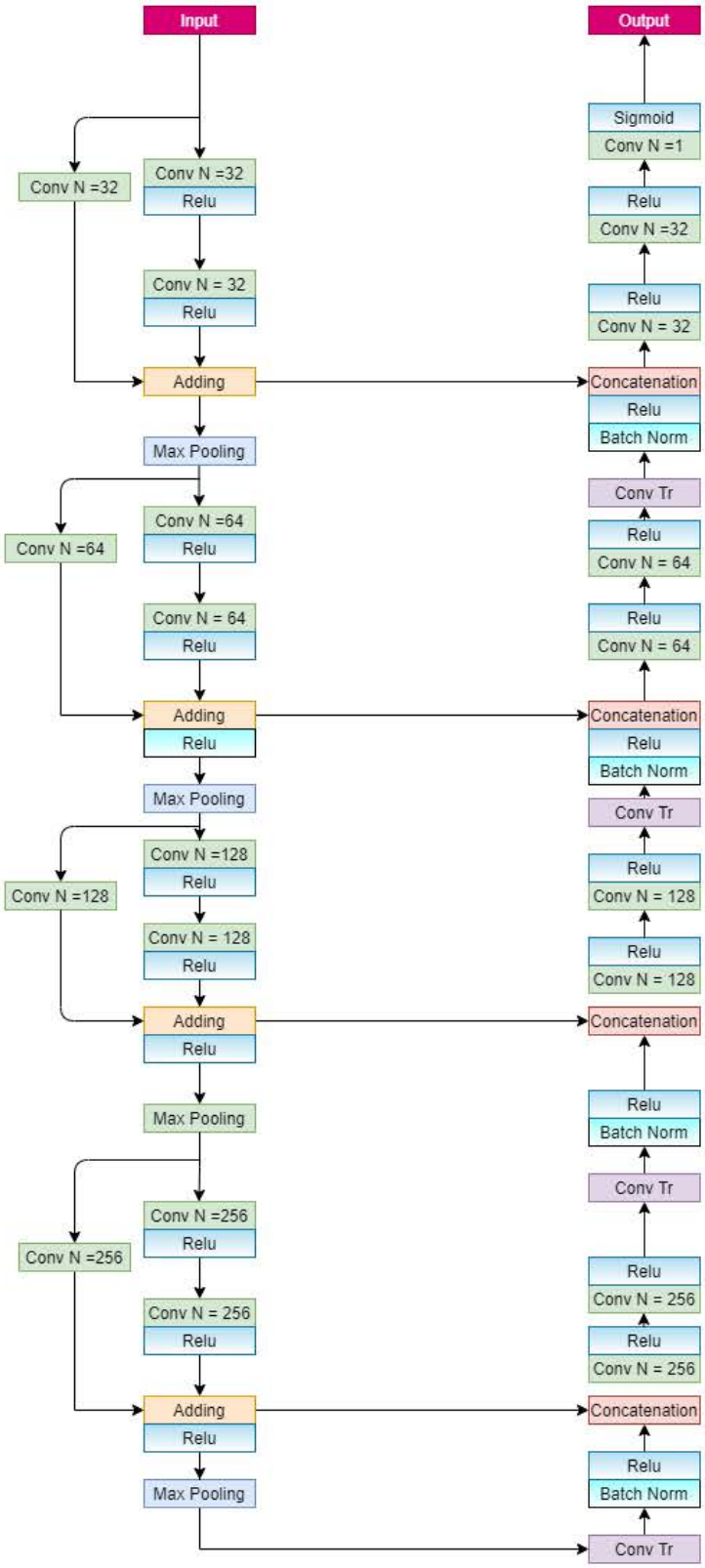


Figure 4.6: Bilnet3

where $F(.)$ represents the residual function, which contains two consecutive blocks where each block contains one convolution layer followed by a batch normalization and ReLu, and parameterized by θ. Ω denotes the architecture of two consecutive blocks. The squeeze and excitation [38] module improves the residual block by re-calibrating feature map channels [38].As shown in Figure 4.7b, squeeze and excitation (SE) blocks can be formulated as:

$$X_{out} = W * X_{in} + F(X_{in}; \theta, \Omega) \tag{4.3}$$

where the learned weights w for re-calibrating the channels of input feature map X_{in} is calculated as:

$$W = \Phi(W_2 * \sigma(W_1 * APool(X_{in}))) \tag{4.4}$$

where Φ indicates the sigmoid function and σ the ReLU activation function, respectively. First, an average pooling which is shown by APool(.), layer is used to 'squeeze' the input feature map X_{in}. Then, two fully connected layers parameterized by W1 and W2 are used to get the "excitation" weights. The SE module effectively improves the representational capacity of residual blocks by including this simple re-weighting technique [1].

The recent incorporation of attention mechanisms into segmentation networks enhances their representational capacities through a strong focus on more informative characteristics. In this work, we incorporated the squeeze-and-attention (SA) [1] module on row U-Net, which is shown in Figure 4.8 to account for two distinctive characteristics of segmentation: i) pixel-group attention and ii) pixel-wise prediction. Specifically, the SA modules impose pixel-group attention on conventional convolution by introducing an 'attention' convolutional channel, thus, efficiently considering spatial-channel inter-dependencies [1].

At both the global and local levels of an image, useful representations for semantic segmentation emerge. Convolution layers produce feature maps conditional on local information at the pixel level since convolution is performed locally around each pixel. Convolution at the pixel level serves as the basis for all semantic segmentation modules, and increasing the receptive field of convolution

layers in a variety of ways improves segmentation performance, demonstrating that a wider context is beneficial for semantic segmentation [1, 39, 40].

Context may be used to determine which portions of feature maps are active at the global image-level since contextual features suggest which classes are likely to appear together in the image [1]. Additionally, [40] demonstrates that the global context offers a more expansive field of vision, which is advantageous for semantic segmentation. Global context features encode these regions holistically, rather than learning a re-weighting for each image segment separately. However, little research has been conducted on encoding context at a finer granularity, which is necessary since various parts of the same image may include completely different surroundings.

To do this, the squeeze-and-attention (SA) module is implemented, and we utiliezed it to acquire more representative features for the task of semantic segmentation through a re-weighting method that considers both local and global factors. As shown in Figure 4.7c a simple squeeze-attention module can be formulated as:

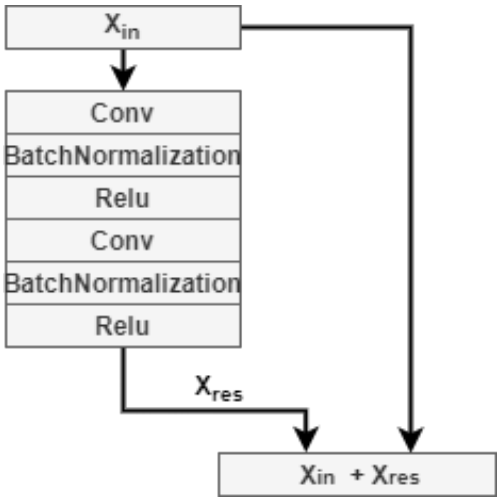
$$X_{out} = X_{attn} * X_{res} + X_{attn} \tag{4.5}$$

where $X_{attn} = Up(\sigma(\widehat{X}_{attn}))$ and $Up(.)$ is a up-sampled function to expand the output of the attention channel:

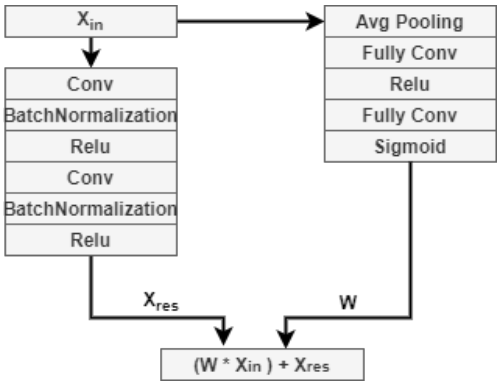
$$\widehat{X}_{attn} = F_{attn}(APool(X_{in}); \theta_{attn}, \Omega_{attn}) \tag{4.6}$$

where $\widehat{X}_{attn}$ represents the output of the attention convolution channel $F_{attn}(.)$, which is parameterized by θ_{attn} and the structure of attention convolution layers Ω_{attn}. An average pooling layer $APool(.)$ is used to perform the not fully-squeezed operation, and then the output of the attention channel X_{attn} is up-sampled to match the output of the main convolution channel X_{unet}

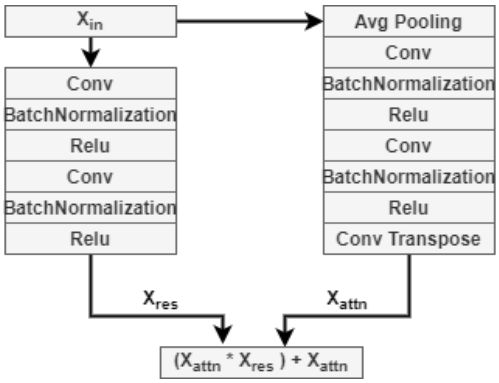
Figure 4.8 illustrate the network's structure in more depth.



(a) Residual Block

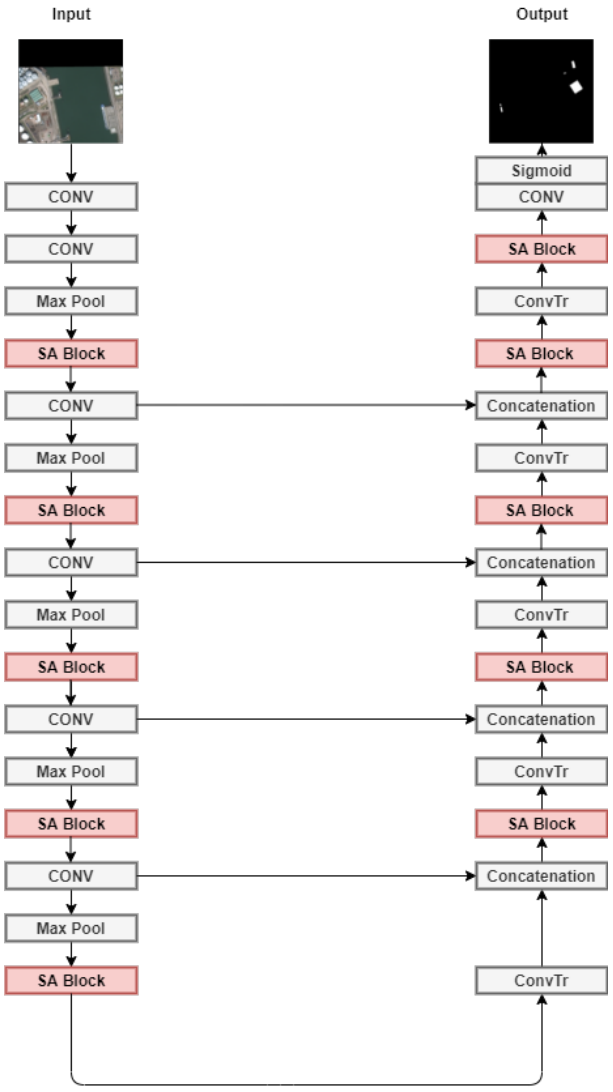


(b) Squeeze-and-excitation(SE)

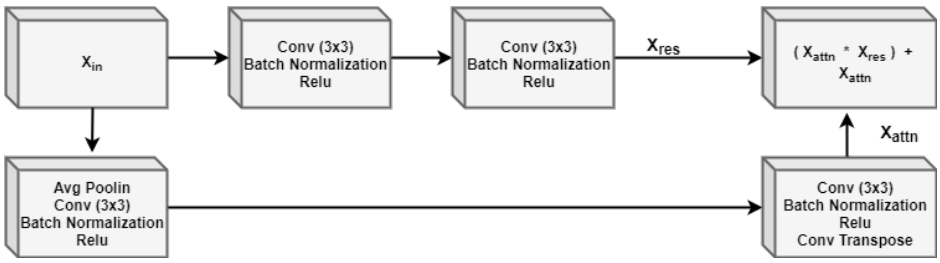


(c) Squeeze-and-attention(SA)

Figure 4.7: The SA block has a similar structure as the SE block that contains an additional path to learn weights for re-calibrating channels of output feature maps X_{out}. The difference lies in that the attention channel of SA modules uses average pooling to down sample feature maps but not fully squeeze as in the SE block [1].



(a) Architecture of our model



(b) This figure illustrates in more detail each Squeeze-and-attention block of our network.

Figure 4.8: Our Network‘s Structure

4.3 Implementation

We implement our model and perform ablation experiments using TensorFlow. All convolutions have a kernel size of (3x3), and the number of filters in the decoder part of the first convolution is 64, increasing by a power of two. We have ten SA blocks in total, and after each of them, we have a convolutional layer. We have used a poly learning schedule that decreases the rate of learning. For the Spacenet6 dataset, the initial learning rate is 0.001. We utilized Adaptive Moment Estimator (Adam) as the optimizer. We used the Binary Cross-Entropy loss function for SpaceNet6 and Massachusetts datasets, which we may define as [41]:

$$L(y, \widehat{y}) = -(y * \log(\widehat{y}_i) + (1 - y) * \log(1 - \widehat{y}_i)) \tag{4.7}$$

Where y is true value and $\widehat{y}$ is the predicted outcome.
The loss function for the challenge7 dataset is a combination of BCE and dice loss to address unbalanced class distribution.

The definition of Dice Loss is [41]:

$$L(y, \widehat{y}) = 1 - \frac{2 * y * \widehat{y} + 1}{y + \widehat{y} + 1} \tag{4.8}$$

Here, 1 is added in numerator and denominator to ensure that the function is not undefined in edge case scenarios such as when $y = \widehat{p} = 0$.
For GeoNRW, since we had ten targets, we could not use BCE, so we used the Cross-Entropy loss function.

For the Spacenet6 dataset, we train our model for 100 epochs. As for Spacenet7, Massachuset, and GeoNRW datasets, we train the model for 80 epochs.

Chapter 5

Experiments

5.1 Evaluation Metrics

A binary classifier predicts one of two outcomes given any input: positive or negative. For the sake of this pixel classification issue, we consider building pixels to be positive and background pixels to be negative. The classifier's output may be summarized as follows:

- True Positives (TP): samples correctly classified as positive.
- True Negatives (TN): samples correctly classified as negative.
- False Positives (FP): samples incorrectly classified as positive.
- False Negatives (FN): samples incorrectly classified as negative.

These numbers are often represented in a confusion matrix, as seen below, and the standard metrics for classifier evaluation are generated as different ratios from it. These metrics are then used to assess the performance of models on the initially isolated test data. Metrics provide a proxy for the model's generalizability.

Two metrics relevant here are Precision and Recall, which are defined as:

	Actually Positive (1)	Actually Negative (0)
Predicted Positive (1)	True Positives (TPs)	False Positives (FPs)
Predicted Negative (0)	False Negatives (FNs)	True Negatives (TNs)

Figure 5.1: Confusion matrix

$$Precision = (\frac{TP}{TP + FP})$$

$$Recall = (\frac{TP}{TP + FN})$$

These were selected as the primary measures due to the data's extreme class imbalance: buildings are often scarce compared to the backdrop, which occupies a far larger geographical area in each picture. True Positive Rate (TPR) and False Positive Rate (FPR) metrics are often deceptive in such situations and, therefore, omitted. Precision (correctness) may be thought of as the percentage of predicted building pixels that are actual buildings, while recall (sensitivity) is the fraction of predicted building pixels. Finally, the F1-score and Intersection Over Union determine the top-performing models for each dataset (IOU). These are defined in the following manner:

$$F1 = (\frac{2 * Precision * recall}{Precision + recall})$$

$$IOU = (\frac{Intersection}{union})$$

F1 score is the harmonic mean of Precision and Recall [42], while IOU is the intersection of actual and predicted values divided by the union of this set for a particular class. IOU is an extremely useful statistic for evaluating segmentation performance since it compares the overlap between prediction and labels (intersection) to the entire area. If forecasts are accurate, the IOU score will be high [7].

5.2 Datasets

The SpaceNet datasets are available as a public dataset on Amazon Web Services (AWS). It includes 67,000 square kilometers of very high-resolution images, more than 11 million building footprints, and 20,000 kilometers of road markings [43].

5.2.1 SpaceNet 6: Multi-Sensor All Weather Mapping

Synthetic Aperture Radar (SAR) is a novel kind of radar that can penetrate clouds, collect data in all weather situations, and record data day and night. SAR satellite overhead data gathering may be especially beneficial in assisting disaster relief operations when weather and cloud cover obstructs conventional electro-optical sensors. Despite these benefits, researchers have little accessible data on the efficacy of SAR for such applications, particularly at ultra-high resolutions.

This open-source dataset combines half-meter Synthetic Aperture Radar (SAR) images from Capella Space with half-meter electro-optical (EO) imagery from Maxar's WorldView 2 satellite. This dataset's objective is to extract building footprints automatically utilizing computer vision and artificial intelligence (AI) techniques in conjunction with SAR and electro-optical image data. The dataset's geographic emphasis was on Europe's biggest port, Rotterdam, the Netherlands. Thousands of buildings, cars, and boats of all sizes are located in this region, providing an excellent testing ground for SAR and integrating these two kinds of data.

The training dataset contained both SAR and EO images in this work, but the test and evaluation datasets contained only SAR data. The dataset was structured to mimic real-world scenarios where historical EO data are available, but simultaneous EO acquisition with SAR is often not possible due to mismatched sensor orbits or cloud cover, rendering EO data unusable [5].

The data set consists of 3401 SAR images and 3041 EO images in total. We

shuffle the images and pick ten percentage of SAR images for test ten percent of that for validation, and the rest of them for training as shown in Table 5.1. All images in this data set are 900 x 900 pixels with resolutions of 1.2m/pixel. Moreover, quad-pol SAR imagery contains four polarization channels.

Their intended usage differentiates the testing set and validation set. A model is trained on a training set and assessed on the test set in a typical machine learning assignment. To guarantee that the best model is produced, a validation set is utilized. This is accomplished by regularly assessing the validation set to adjust the model's parameters during training. For example, examining the trends in the training set and validation set accuracy combined may indicate if the model is overfitting or underfitting to the training set. This enables the regularization strength and other hyperparameters to be tuned intelligently.

Training	Validation	Test
5442	340	340

Table 5.1: – Table presenting randomly split sets of the Spacenet6 dataset

5.2.1.1 Pre-processing

Due to computational constraints, the 900 $\times$ 900 images from Table 5.1 were cropped into segments of size 256 $\times$ 256. During training, the images were fed into the models in a minibatch of size 16. The training set size is unfortunately not large enough to guarantee good production-level performance. Representative samples from the Spacenet6 dataset can be seen in Figure5.2

We take advantage of the optical imagery through transfer learning. Training the model on the optical imagery first and the SAR imagery second leads to higher performance than training the model on the SAR imagery alone. The pan-sharpened RGB optical images are transformed to four-channel pictures and used instead of the SAR images to facilitate transfer learning. (Red is used for HH polarization, green is used for VV polarization, and blue is used for HV and VH polarization.) Thus, the model may be trained on optical pictures and

Figure 5.2: Dataset examples: Top row shows SAR images. The bottom left shows optical images of the same tile, and the bottom right shows ground truth.

subsequently on SAR data without requiring the layers to be replaced [5].

5.2.2 SpaceNet 7:Multi Temporal Urban Development

The dataset is based on a newly released open-source dataset of Planet satellite imagery mosaics, which for each area of intention (AOI) has 24 pictures (one each month) covering a total of 100 distinct geographies. The dataset will include images covering over 40,000 square kilometers and comprehensive polygon labeling for building footprints in the picture, totaling over 10 million unique annotations.

The SpaceNet7 problem has far-reaching consequences for disaster preparation, environmental protection, infrastructure development, and disease control. Apart from the humanitarian implications, SpaceNet7 presents a unique computer vision problem because of the tiny pixel area of each item and the high object density in pictures [44]. Table 5.2 gives more detail about dataset.

Category	Value
Num AOIs	101
Num Observations	2389
Num Buildings	11,080,000
Total Observed Area (km ^{2})	41,000
Mean Buildings per Observation	4,700
Mean Building Area (km ^{2})	190
Mean GSD (m)	4.0

Table 5.2: The Data $\sim$ 100 locations, spread out across the globe

RGBA images of SpaceNet7 datasets have a resolution of 1024 $\times$ 1024 pixels and four channels (RGBA). There were 1260 pictures divided into three groups: ten percent for validation, ten percent for the test, and the remaining for the training dataset. Before beginning, we translated GeoJSON labels from a random coordinate reference system (CRS) into pixel coordinates or picture masks, after which we divided each image into 128 x 128 pixels to serve as pre-processing information. Examples of the dataset can be seen in Figure 5.8

5.2.3 Massachusetts Roads Dataset

This dataset included 1500x1500 pixel images of the city of Massachusetts released by the state at a resolution of 1 meter per pixel. Target maps used on these images were also readily available in the rasterized format. Each image in the original dataset comes at a 1500x1500 resolution, and the dataset is split randomly into training, validation, and test datasets as in the table below.

Training	Testing	validation
640	80	80

Table 5.3: Table presenting randomly split sets of the Massachusetts Roads dataset

Due to computational constraints, the 1500 $\times$ 1500 images from Table 5.3 were cropped into non-overlapping segments of size 256 $\times$ 256. This training included a wide range of urban, suburban, and coastal regions. During training, the images were fed into the models in a minibatch of size 16. Representative samples from the Massachusetts Roads dataset can be seen in Figure 5.4

5.2.4 GeoNRW

This dataset comprises orthorectified aerial photos, LiDAR-derived digital elevation models, and ten-class segmentation maps obtained via the German state of North Rhine-open Westphalia's data initiative and enhanced using OpenStreetMap [45]. Pre-processing includes resampling the 0.1m resolution photos to 1m, averaging the initial LiDAR return inside a $1m^2$ to arrive at the exact resolution as the photographs, and rasterizing vector files of the land cover data. PEG2000 files are used for aerial photos, whereas GeoTIFF files are used for land cover maps and digital elevation models [45]. The dataset contains 1029 unique DEM pictures and RGB image versions of the same DEM images with 1000x1000 pixels. The training dataset contained both DEM and RGB images in this work, but the test and evaluation datasets contained only DEM data.

Training	Testing	validation
1650	204	204

Table 5.4: Table presenting randomly split sets of the GeoNRW dataset

Due to computational constraints, the 1000 $\times$ 1000 images from Table 5.9 first had resized to 1024 x 1024 then cropped into non-overlapping segments of size 256 $\times$ 256. Representative samples from the GeoNRW can be seen in Figure 5.5

5.3 Results

5.3.1 Results on SpaceNet6

Building footprint extraction results of the proposed method are shown in Table 5.5. They are evaluated on the test dataset based on IOU, precision, recall, and the F1-score. Our approach achieves these results without additional processing steps and pre-training. The proposed method achieves a total F1-score of 0.84, which is an improvement of 0.16% compared with the standard U-Net–based method and also improved IOU from 0.96 to 0.98.

	Test IOU	Test recall	Test precision	Test F1-score
U-Net [13]	0.96	0.68	0.79	0.73
FPN [17]	0.95	0.65	0.75	0.69
linknet [19]	0.97	0.60	0.74	0.64
Our SA model	**0.98**	**0.80**	**0.87**	**0.84**

Table 5.5: Comparison of different networks [2] with SpaceNet6 dataset

The findings of Bilnets, are summarized in the Table 5.6.

	Test IOU	Test recall	Test precision	Test F1 score
Bilnet1	0.95	0.36	0.66	0.45
Bilnet2	0.95	0.35	0.65	0.43
Bilnet3	0.95	0.50	0.67	0.56

Table 5.6: Results of Bilnets on SpaceNet6 dataset

Our SA model's predictions are depicted in Figure 5.6. We have compared them to the predictions of row U-Net, FPN, and Linknet and achieved better results in comparison to them. As mentioned earlier, we trained the network with both EO and SAR images but tested them just with SAR images. However, to show our model's result with EO images as well, we have shown in Figure 5.7 sample results of predicted EO images and compared them with U-Net.

5.3.2 Results on SpaceNet7

The results of SpaceNet7's area of interest segmentation are presented in Table 5.7. They are assessed in terms of IOU by the test dataset. Our proposed network earned an IOU score of 0.83, outperforming comparable state-of-the-art networks. In Figure 5.8, we have also included samples of our networks' predictions from the test dataset. Although the buildings were too tiny, our network was able to locate and segment them effectively.

Model	Backbone	Loss	IOU
DeepLabv3+	MobileNet	Dice Loss	0.520
DeepLabv3+	ResNet50	Dice Loss	0.525
DeepLabv3+	ResNet101	BCE	0.521
DeepLabv3+	ResNet101	Dice Loss	0.534
SegNet	-	Dice Loss	0.463
Full resolution residual network (A)	-	Dice Loss	0.418
Full resolution residual network (B)	-	Dice Loss	0.423
FCN32	-	Dice Loss	0.426
ICNet	-	Multi scale cross entropy	0.450
ENet	-	Dice Loss	0.600
Ede Net	-	Dice Loss	0.566
ERFNet	-	Dice Loss	0.602
ESNet	-	Dice Loss	0.610
CGNet	-	Dice Loss	0.573
EF u-Net	-	Dice Loss	0.590
Our SA Model	-	**BCE + Dice Loss**	**0.83**

Table 5.7: Comparison of different networks with SpaceNet7 dataset

5.3.3 Results on Massachusetts Dataset

In order to check the performance of our network on non-SAR images, we evaluated our network with the Massachusetts dataset which has EO images. We compare it with U-Net, FPN, and Linknet. The results are shown in Table5.8 and their prediction samples are shown in Figure 5.9

	Test IOU	Test recall	Test precision	Test F1-score
U-Net	0.98	0.67	0.76	0.71
FPN	0.98	0.58	0.80	0.67
Linknet	0.98	0.61	0.78	0.68
Our SA model	**0.98**	**0.67**	**0.80**	**0.72**

Table 5.8: Comparison of different networks with Massachusetts dataset

5.3.4 Results on GeoNRW

We also have checked our network‘s performance on DEM images, and have compared it with other networks. The results of GeoNRW's area of interest segmentation are presented in the Table 5.9, and their prediction samples are shown in Figure 5.10. We have trained the model with both DEM and RGB images, however, we have tested it with just DEM images.

	Test IOU	Test recall	Test precision	Test F1 score
U-Net	0.66	0.61	0.81	0.69
FPN	0.71	0.66	0.78	0.71
Linknet	0.74	0.71	0.82	0.76
Our SA model	**0.76**	**0.72**	**0.82**	**0.77**

Table 5.9: Comparison of different networks with GeoNRW dataset

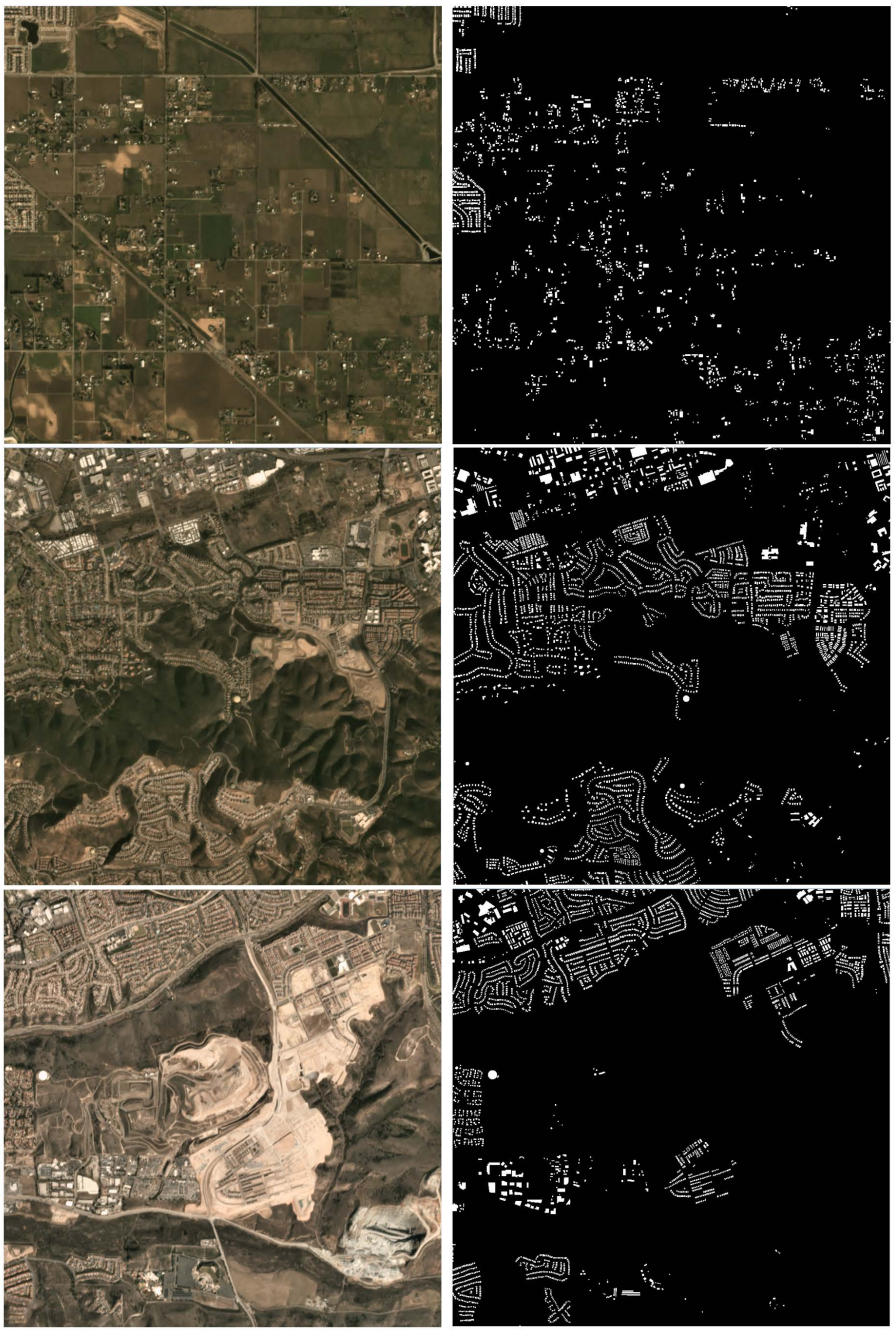


Figure 5.3: Examples of SpaceNet7 dataset. First column shows images and second one shows labels

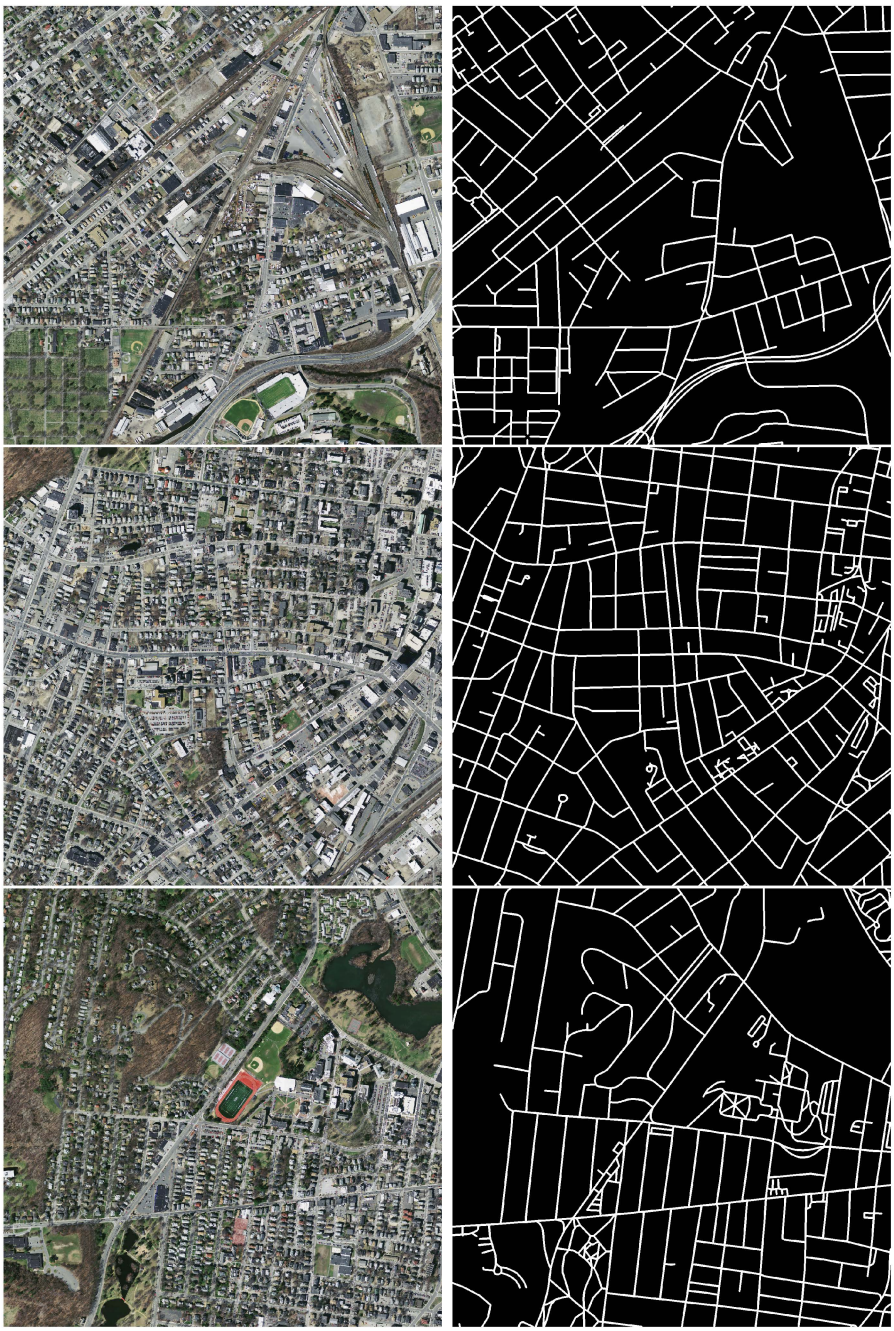


Figure 5.4: Examples of Massachusetts dataset. First column shows images and second one shows labels

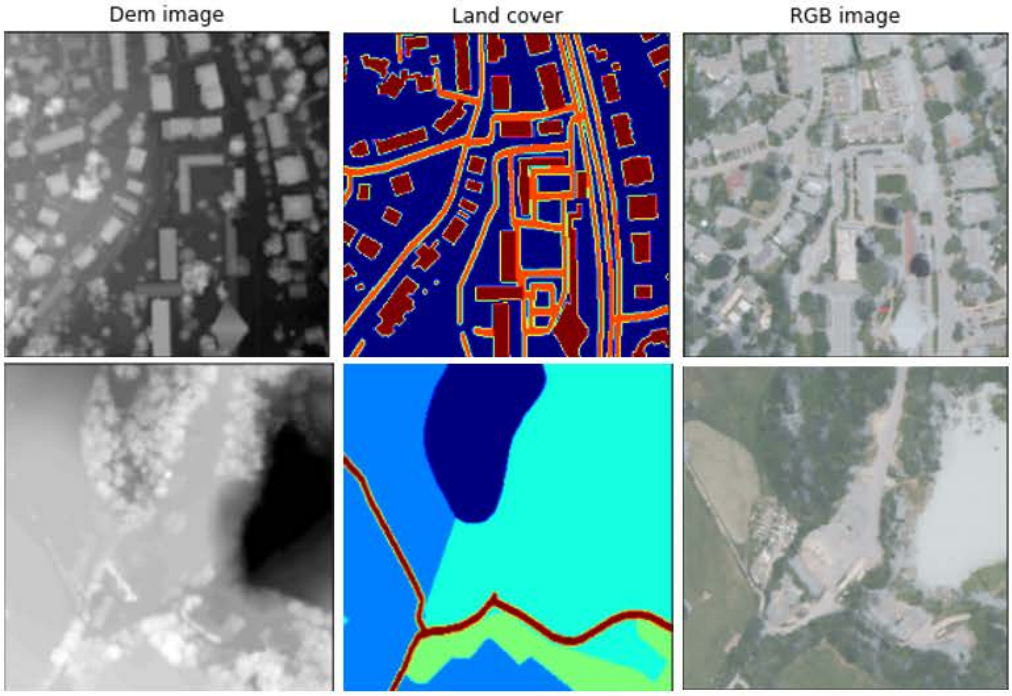


Figure 5.5: Examples of GeoNRW dataset. First column shows DEM images, second one shows labels, and third column shown RGB images

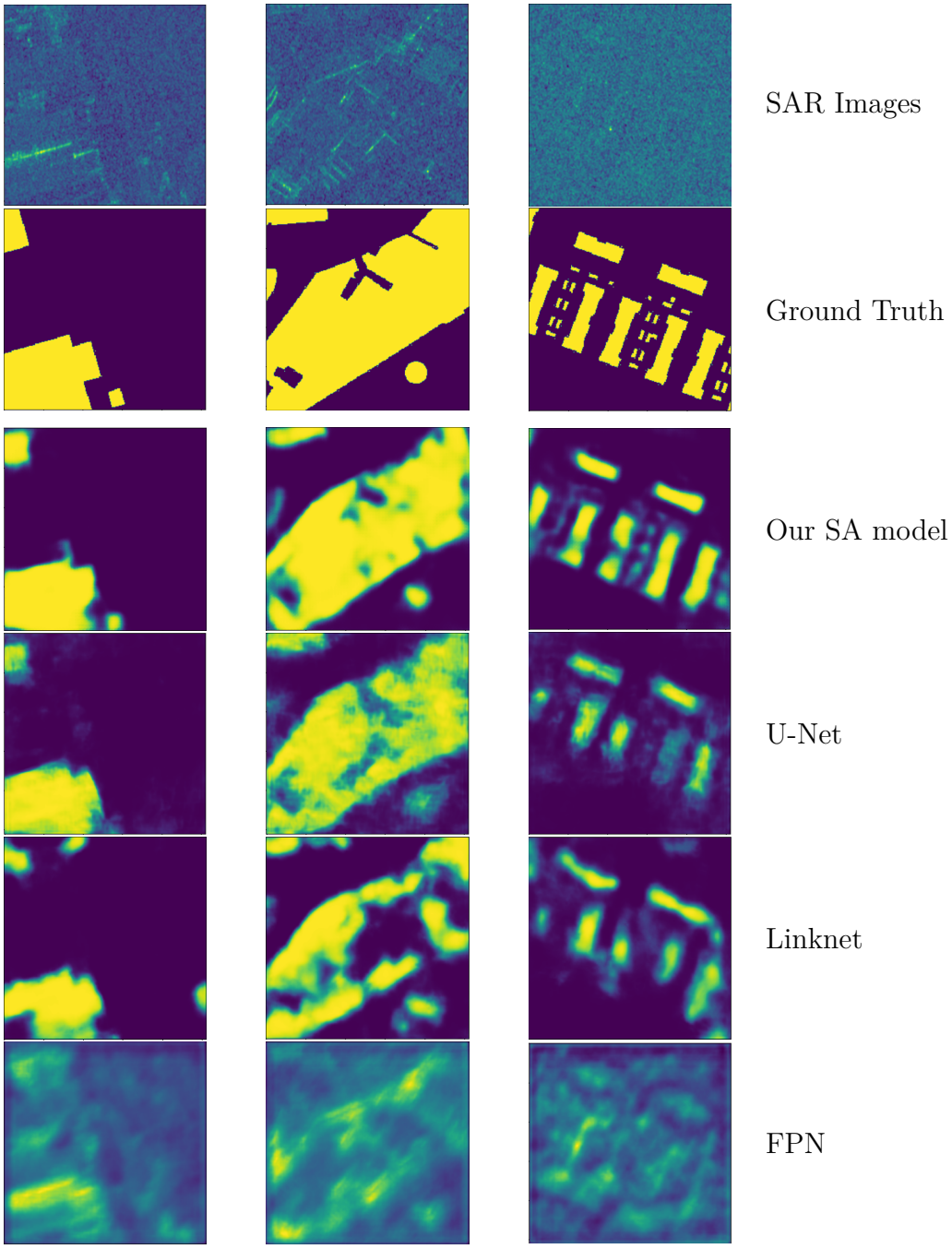


Figure 5.6: Comparative analysis of predicted samples from various networks.

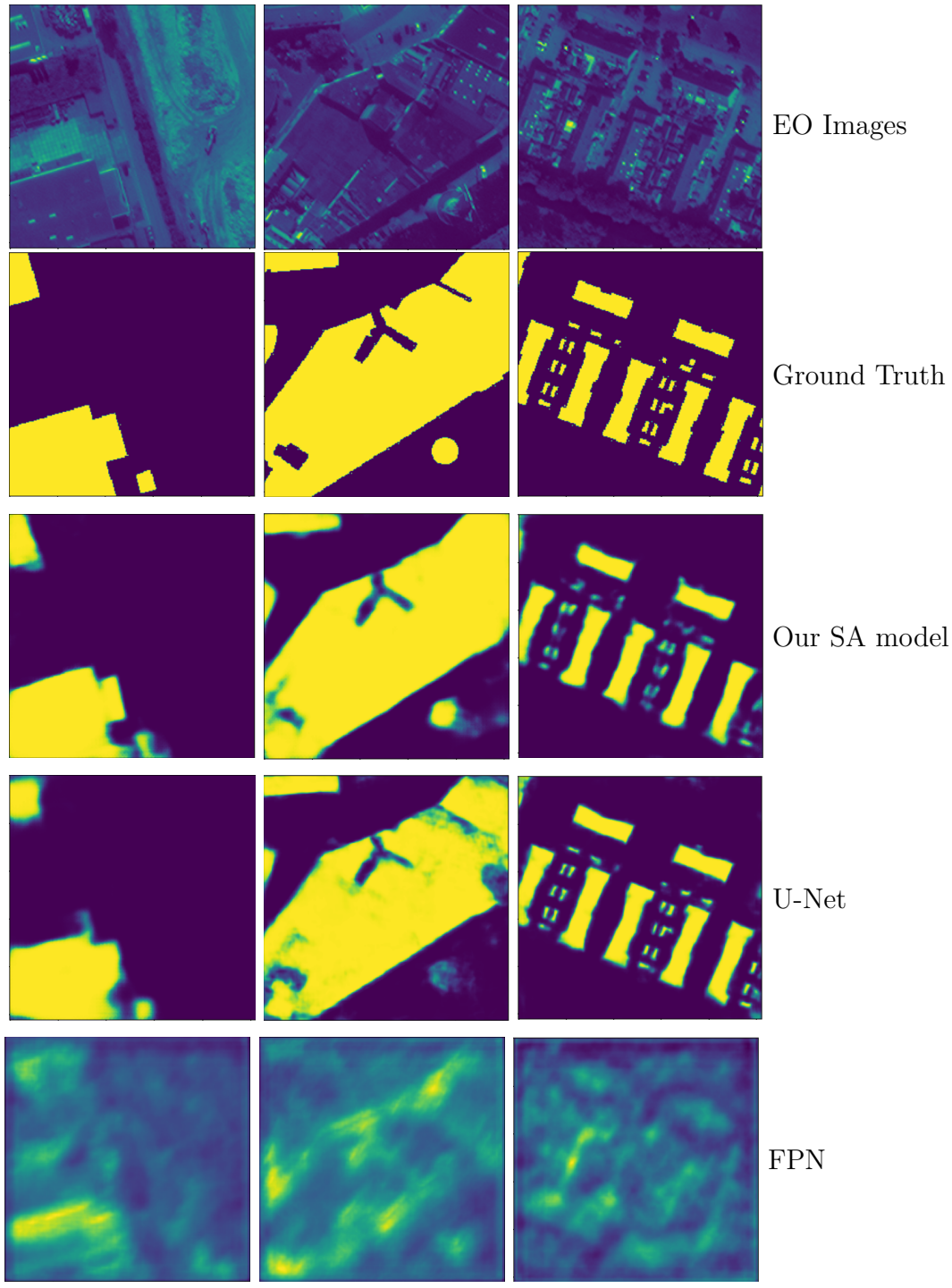


Figure 5.7: Prediction of proposed model vs U-Net model with EO images

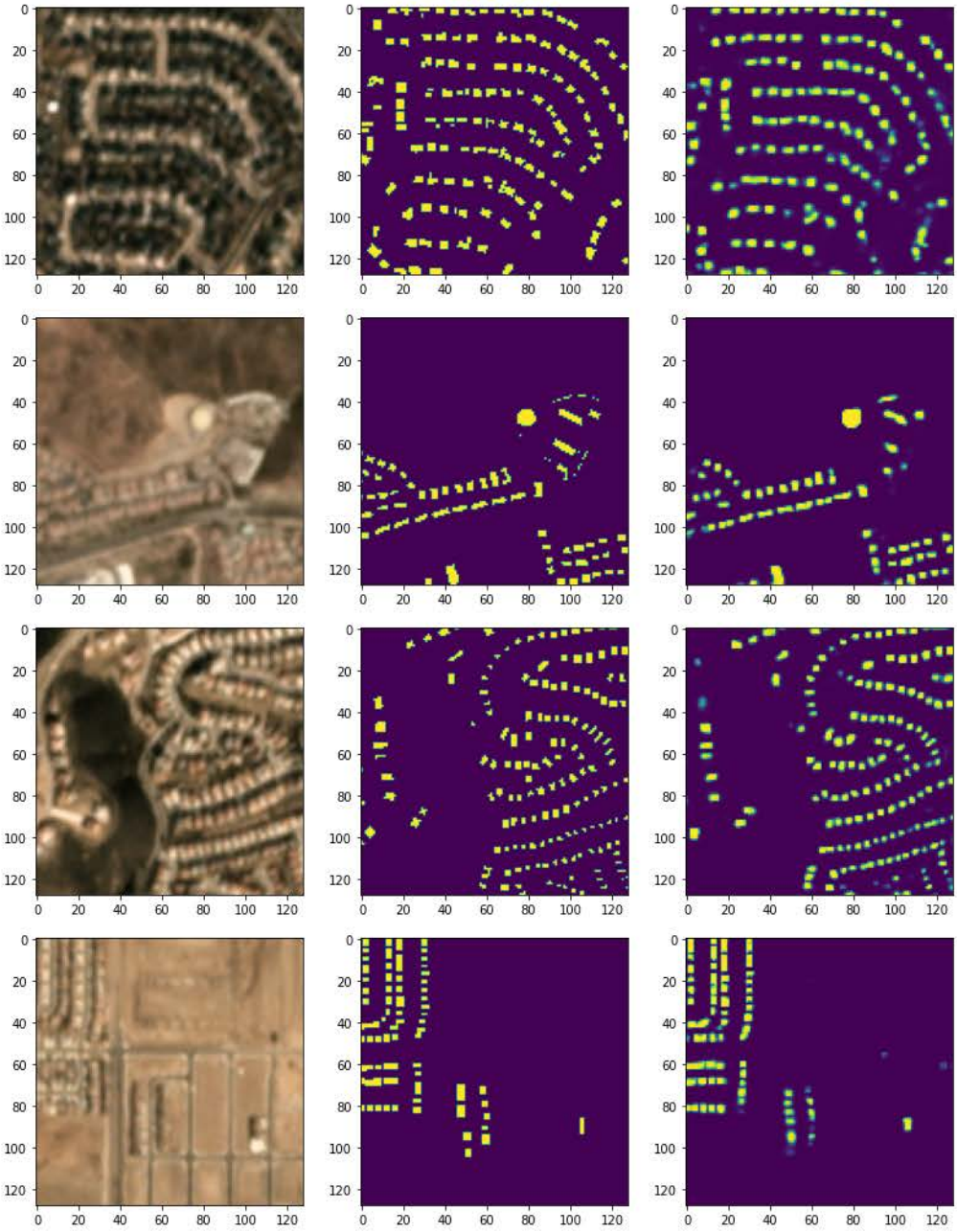


Figure 5.8: Prediction of proposed model On SpaceNet7

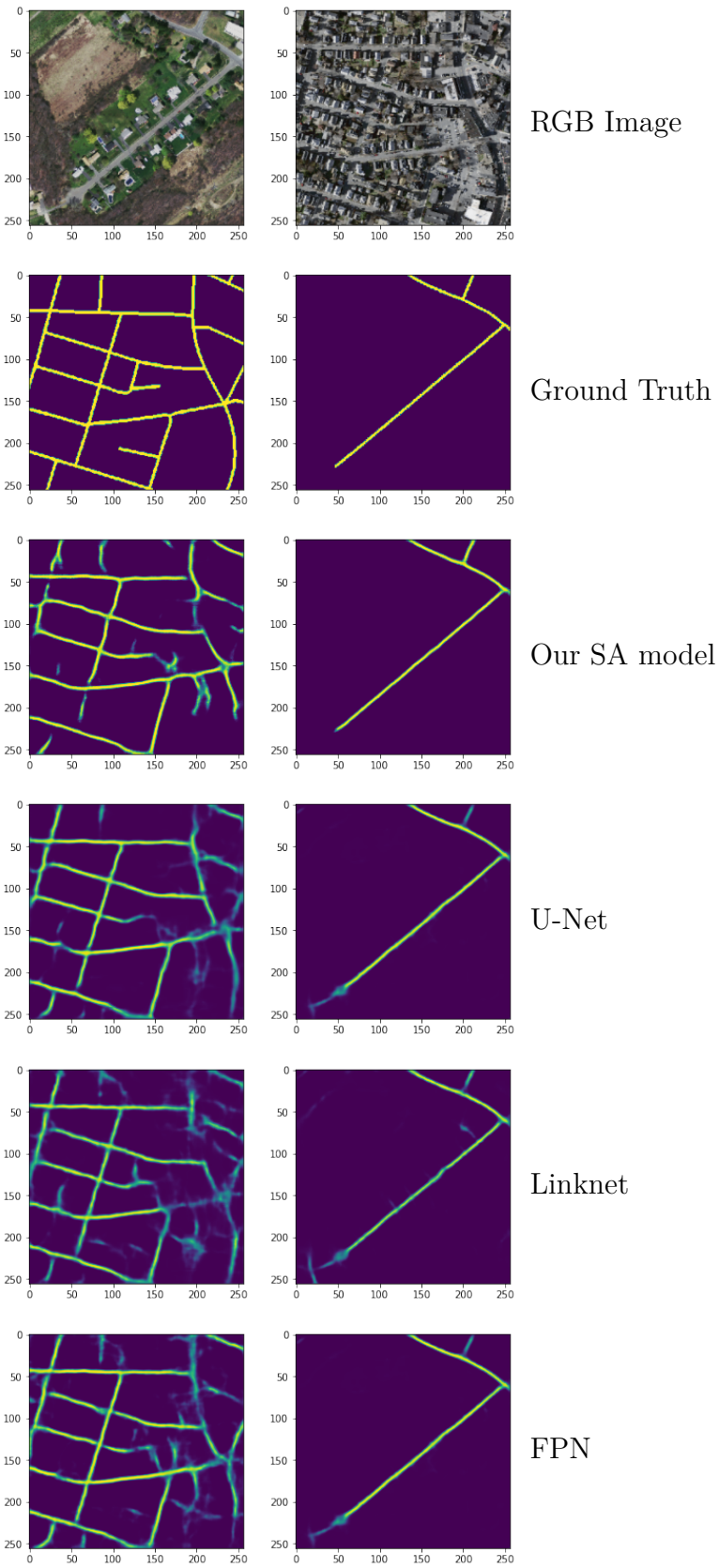


Figure 5.9: Comparison qualitative results in Massachusetts dataset

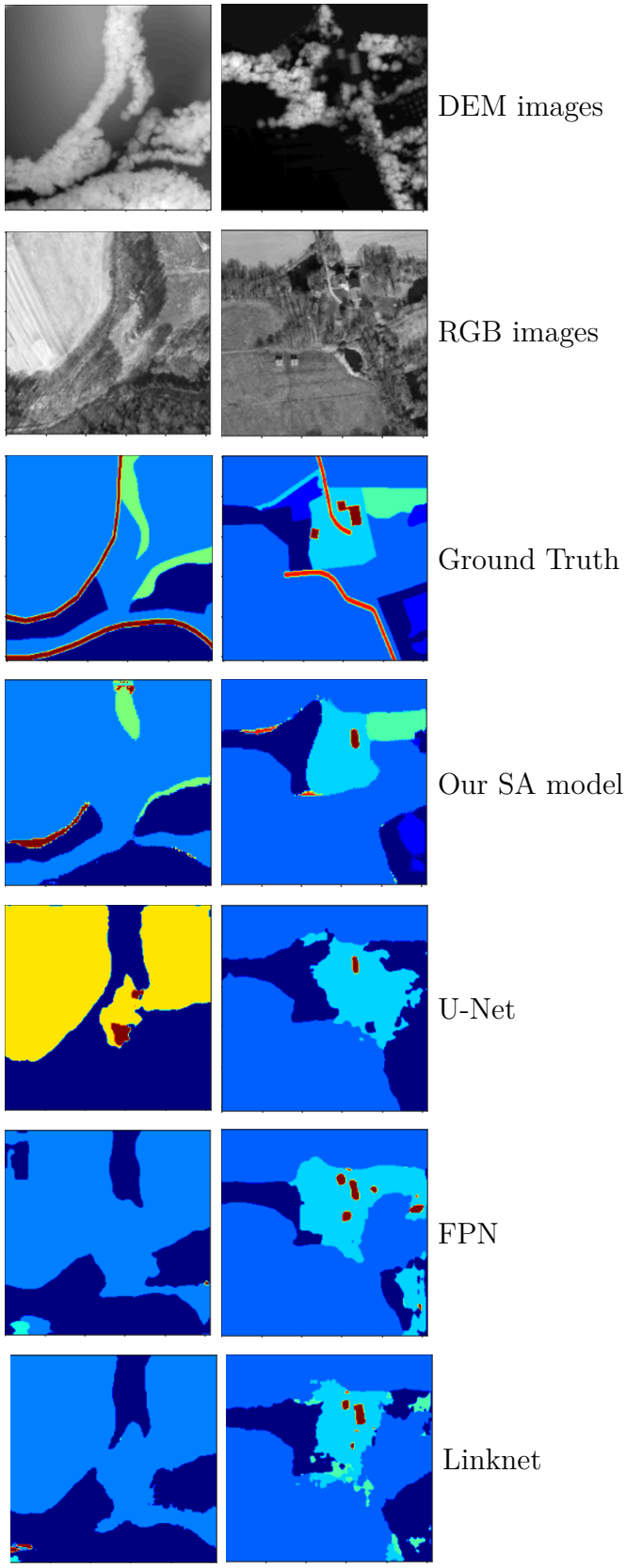


Figure 5.10: Comparison qualitative results in GeoNRW dataset

Chapter 6

Conclusion

An ongoing study topic has been the automatic extraction of objects of interest from high-resolution satellite images. An essential field in remote sensing is segmentation in Synthetic-Aperture-Radar-based (SAR) images to circumvent visible images' constraints. While working on data obtained from the visible spectrum is still commonly desired and used in many aerial applications, such applications generally require a clear sky and little cloud cover to work with high precision. When rain and clouds block traditional optical sensors, SAR imaging is particularly beneficial as an alternative approach to address visibility-related difficulties.

Due to its superior performance in object segmentation applications in computer vision, we focus on the performance of recent deep learning-based solutions in this thesis for remote sensing. We introduce an architecture based on squeeze and attention blocks to segment objects of interest in remote sensing images and compare its performance to various baseline networks on different datasets. Our experiments run on SAR images and non-SAR images, including electro-optic (EO) and digital elevation model (DEM) images. Our preliminary experiments show that our suggested architecture yields superior results than multiple baseline networks on all of those datasets.

Bibliography

[1] Z. Zhong, Z. Lin, R. Bidart, X. Hu, I. B. Daya, J. Li, and A. Wong, "Squeeze-and-attention networks for semantic segmentation," *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 13062–13071, 2020.

[2] P. Yakubovskiy, "Segmentation models," 2019.

[3] D. G¨ozen and S. Ozer, "Visual object tracking in drone images with deep reinforcement learning," in *2020 25th International Conference on Pattern Recognition (ICPR)*, pp. 10082–10089, 2021.

[4] P. M. Teillet, R. P. Gauthier, A. Chichagov, and G. Fedosejevs, "Towards integrated earth sensing: Advanced technologies for in situ sensing in the context of earth observation," *Canadian Journal of Remote Sensing*, vol. 28, no. 6, pp. 713–718, 2002.

[5] H. D. B. J. E. A. W. N. P. F. H. R. B. A. S. S. B. T. L. R. Shermeyer, J., "Spacenet 6: Multi-sensor all weather mapping dataset.," *Conference on Computer Vision and Pattern Recognition Workshop(CVPRW)*, pp. 768–777, 2020.

[6] O. Sahin and S. Ozer, "Yolodrone: Improved yolo architecture for object detection in drone images," in *2021 44th International Conference on Telecommunications and Signal Processing (TSP)*, pp. 361–365, 2021.

[7] S. Muruganandham, "Semantic segmentation of satellite images using deep learning," 2016.

[8] R. T. Deepika Adlakha, Devender Adlakha, "Analytical comparison between sobel and prewitt edge detection techniques," *International Journal of Scientific Engineering Research*, January-2016.

[9] E. Simonetto, H. Oriot, and R. Garello, "Rectangular building extraction from stereoscopic airborne radar images," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 43, no. 10, pp. 2386–2395, 2005.

[10] K. He, X. Zhang, S. Ren, and J. Sun, "Identity mappings in deep residual networks," in *Computer Vision – ECCV 2016* (B. Leibe, J. Matas, N. Sebe, and M. Welling, eds.), (Cham), pp. 630–645, Springer International Publishing, 2016.

[11] H. Furukawa, "Deep learning for end-to-end automatic target recognition from synthetic aperture radar imagery," *ArXiv*, vol. abs/1801.08558, 2018.

[12] M. Shahzad, M. Maurer, F. Fraundorfer, Y. Wang, and X. X. Zhu, "Buildings detection in vhr sar images using fully convolution neural networks," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 57, no. 2, pp. 1100–1116, 2019.

[13] O. Ronneberger, P. Fischer, and T. Brox, "U-net: Convolutional networks for biomedical image segmentation," in *Medical Image Computing and Computer-Assisted Intervention – MICCAI 2015* (N. Navab, J. Hornegger, W. M. Wells, and A. F. Frangi, eds.), (Cham), pp. 234–241, Springer International Publishing, 2015.

[14] J. Long, E. Shelhamer, and T. Darrell, "Fully convolutional networks for semantic segmentation," in *2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 3431–3440, 2015.

[15] V. Badrinarayanan, A. Kendall, and R. Cipolla, "Segnet: A deep convolutional encoder-decoder architecture for image segmentation," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 39, no. 12, pp. 2481–2495, 2017.

[16] T. Pohlen, A. Hermans, M. Mathias, and B. Leibe, "Full-resolution residual networks for semantic segmentation in street scenes," *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 3309–3318, 2017.

[17] T.-Y. Lin, P. Dollár, R. Girshick, K. He, B. Hariharan, and S. Belongie, "Feature pyramid networks for object detection," in *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 936–944, 2017.

[18] A. Paszke, A. Chaurasia, S. Kim, and E. Culurciello, "Enet: A deep neural network architecture for real-time semantic segmentation," 06 2016.

[19] A. Chaurasia and E. Culurciello, "Linknet: Exploiting encoder representations for efficient semantic segmentation," *2017 IEEE Visual Communications and Image Processing (VCIP)*, pp. 1–4, 2017.

[20] L.-C. Chen, Y. Zhu, G. Papandreou, F. Schroff, and H. Adam, "Encoder-decoder with atrous separable convolution for semantic image segmentation," in *ECCV*, 2018.

[21] H. E. Ilhan, S. Ozer, G. K. Kurt, and H. Ali Cirpan, "Offloading deep learning empowered image segmentation from uav to edge server," in *2021 44th International Conference on Telecommunications and Signal Processing (TSP)*, pp. 296–300, 2021.

[22] H. Zhao, X. Qi, X. Shen, J. Shi, and J. Jia, "Icnet for real-time semantic segmentation on high-resolution images," *ArXiv*, vol. abs/1704.08545, 2018.

[23] S.-Y. Lo, H.-M. Hang, S.-W. Chan, and L. Jing Jhih, "Efficient dense modules of asymmetric convolution for real-time semantic segmentation," 09 2018.

[24] T. Wu, S. Tang, R. Zhang, and Y. Zhang, "Cgnet: A light-weight context guided network for semantic segmentation," 11 2018.

[25] E. Romera, J. M. Álvarez, L. M. Bergasa, and R. Arroyo, "Erfnet: Efficient residual factorized convnet for real-time semantic segmentation," *IEEE Transactions on Intelligent Transportation Systems*, vol. 19, no. 1, pp. 263–272, 2018.

[26] Y. Wang, Q. Zhou, and X. Wu, "Esnet: An efficient symmetric network for real-time semantic segmentation," 06 2019.

[27] P. Das and S. Chand, "Extracting building footprints from high-resolution aerial imagery using refined cross attentionnet," *IETE Technical Review*, vol. 0, no. 0, pp. 1–12, 2021.

[28] F. GAO, A. LIU, K. LIU, E. YANG, and A. HUSSAIN, "A novel visual attention method for target detection from sar images," *Chinese Journal of Aeronautics*, vol. 32, no. 8, pp. 1946–1958, 2019.

[29] F. Ma, F. Gao, J. Sun, H. Zhou, and A. Hussain, "Attention graph convolution network for image segmentation in big sar imagery data," *Remote. Sens.*, vol. 11, p. 2586, 2019.

[30] S. Fan, Y. Sun, and P. Shui, "Region-merging method with texture pattern attention for sar image segmentation," *IEEE Geoscience and Remote Sensing Letters*, vol. 18, no. 1, pp. 112–116, 2021.

[31] Z. Yue, F. Gao, Q. Xiong, J. Wang, A. Hussain, and H. Zhou, "A novel attention fully convolutional network method for synthetic aperture radar image segmentation," *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, vol. 13, pp. 4585–4598, 2020.

[32] P. Kansakar and F. Hossain, "A review of applications of satellite earth observation data for global societal benefit and stewardship of planet earth," *Space Policy*, pp. 46–54, 2016.

[33] NASA Video", "NASA ARSET: Basics of Synthetic Aperture Radar (SAR)," 08 2018.

[34] D. Hogan, "SAR 201: An Introduction to Synthetic Aperture Radar, Part 2," 02 2020.

[35] I. C. Education, "What is deep learning?," ibm," 2019.

[36] J. J. Moolayil, "A layman's guide to deep neural networks.," 30 May 2020.

[37] B. M. Albaba and S. Ozer, "Synet: An ensemble network for object detection in uav images," in *2020 25th International Conference on Pattern Recognition (ICPR)*, pp. 10227–10234, 2021.

[38] J. Hu, L. Shen, and G. Sun, "Squeeze-and-excitation networks," in *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 7132–7141, 2018.

[39] H. Zhao, J. Shi, X. Qi, X. Wang, and J. Jia, "Pyramid scene parsing network," *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 6230–6239, 2017.

[40] H. Zhang, K. Dana, J. Shi, Z. Zhang, X. Wang, A. Tyagi, and A. Agrawal, "Context encoding for semantic segmentation," *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 7151–7160, 2018.

[41] S. Jadon, "A survey of loss functions for semantic segmentation," 06 2020.

[42] GreekDataGuy, "Evaluating ml models: Precision, recall, f1 and accuracy. medium."

[43] "Spacenet on amazon web services (aws)."

[44] "Sn7: Multi-temporal urban development challenge."

[45] G. Baier, A. Deschemps, M. Schmitt, and N. Yokoya, "Geonrw," 2020.