Universität Hamburg
DER FORSCHUNG | DER LEHRE | DER BILDUNG

Synthetic Aperture Radar Image Interpretation Based on Deep Learning

Dissertation
with the aim of achieving a doctoral degree at the
Faculty of Mathematics, Informatics and Natural Sciences
Department of Informatics
of Universität Hamburg

submitted by
Di Zhang
February, 2022

1st Supervisor:
Prof. Dr. Jianwei Zhang
Department of Informatics
Universität Hamburg, Germany

2nd Supervisor:
Dr. Martin Gade
Department of Oceanography
Universität Hamburg, Germany

Submitted on: February 22nd, 2022
Date of oral defence: May 18th, 2022

Abstract

Synthetic Aperture Radar (SAR) is an active radar, which can obtain high-resolution images under all-day and all-weather conditions. Because of its various advantages, SAR has been widely used in military and civil domains. SAR image interpretation is to acquire key information from SAR images. However, SAR images are strongly affected by speckle noise and geometric distortion effects. This makes the information extraction from SAR images challenging to perform.

Traditional methods for SAR image interpretation require expert knowledge and can easily cause overfitting problems. Recent advances in deep learning have opened a wide door to analyze SAR images automatically and efficiently. This thesis focuses on developing deep learning-based approaches for SAR image interpretation. Three aspects are particularly investigated, including oceanic eddy detection, intertidal sediments and habitats classification, and land cover classification. The research data sources cover Ground Range Detected (GRD), multi-band and multi-polarization SAR images, and optical data. The main scientific contributions in this thesis are summarized as follows:

Firstly, this thesis realizes automatic oceanic eddy detection on SAR images based on a novel Mask Edge Enhancement and IoU Score Region-based Convolutional Neural Network (Mask-ES-RCNN) framework. Since there are no existing SAR oceanic eddy instance segmentation datasets, we build a SAR Oceanic Eddy Detection Dataset (SOEDD) for developing deep learning-based methods. The Mask-ES-RCNN model applies implicit learning of internal texture information and adopts Mask IoU scoring to focus more on mask qualities. It outperforms a Mask-RCNN baseline in terms of Average Precision (AP).

Secondly, this thesis proposes an UNet-based semantic segmentation network with a Texture Enhancement Module (TE-UNet) for intertidal sediments and habitats classification. The application of the texture enhancement module improves the performances of the TE-UNet model by enhancing the global texture information explicitly. Apart from intensity channels of SAR data, we also use polarimetric decomposition results as inputs. Radarsat-2 (C band) and ALOS-2 (L band) SAR images are concatenated in the channel dimension to realize multi-band learning. A comparative experimental study proves the effectiveness of a multi-band and multi-polarization system for classification tasks in the intertidal zone.

Finally, this thesis presents a SAR-Optical Fusion UNet model (SOF-UNet) based on the existing largest dataset SEN12MS which provides optical and SAR pairs to realize land cover classification. The two-stream SOF-UNet consists of three parts: two encoders to extract features, a shared decoder to upsample the feature maps, and specially designed skip connections to fuse multi-modal features. The qualitative and quantitative experimental results show that SOF-UNet has a promising capability in identifying different land cover classes and can retain fine details in the prediction maps.

Zusammenfassung

Sythetic Aperture Radar (SAR) is ein aktives Radar, das unabhängig von Tageszeit und Wetterbedingungen hochauflösende Aufnahmen generieren kann. Aufgrund verschiedener Vorteile findet die Technik breite Anwendung in zivilen und militärischen Bereichen. Ziel einer SAR-Bildinterpretation ist die Extraktion anwendungsrelevanter Informationen aus den Aufnahmen. Allerdings ist SAR sehr anfällig für Aufnahmefehler wie Speckle oder geometrische Verzerrungen, und diese Effekte erschweren die Interpretation signifikant.

Klassische Methoden der SAR-Bildinterpretation setzen Expertenwissen voraus und sind sehr anfällig für eine Überanpassung auf gegebene Bilddaten. Aktuelle Fortschritte im Bereich des Deep Learning ermöglichen demgegenüber automatische und effiziente Analysen von Bilddaten. Diese Arbeit beschäftigt sich mit Ansätzen des Deep Learning zur Interpretation von SAR-Aufnahmen. Die drei hier betrachteten Anwendungsbereiche hierfür sind die Erkennung von Meereswirbeln, die Klassifikation von Sedimenten und Habitaten in Gezeitenzonen sowie, die Bestimmung von Landbedeckungklassen. Die Datensätze in diesem Forschungsfeld fallen sehr vielfältig aus und beinhalten Ground Range Detected (GRD), Multi-Band und Multi-Polarisation SAR Daten, und optische Aufnahmen. Die wissenschaftlichen Beiträge dieser Arbeit lassen sich in den folgenden drei Aspekten, analog zu den oben genannten Anwendungsbereichen, zusammenfassen:

(1) Es wird ein System (Mask-ES-RCNN) zur automatischen Erkennung von Meereswirbeln vorgestellt, das auf neuartigem Mask Edge Enhancement und IoU Score Region-based Convolutional Neural Network basiert. Da zum Training eines solchen Systems eine ausreichende Datenbasis benötigt wird, die bisher nicht existierte, wird ein neuer Datensatz (SOEDD) von Instanz-segmentierten Meereswirbeln aufgebaut. Mask-ES-RCNN nutzt das implizite Lernen von Texturinformationen und ein Mask-IoU Scoring, um die Qualität der erkannten Bildmasken zu verbessern. Das Modell erreicht eine höhere Genauigkeit als das zugrunde

gelegte Mask RCNN Modell.

(2) Die Arbeit stellt ein neuronales Netz auf Basis von UNet mit einer Erweiterung zur Verstärkung von Texturen vor, das zur semantischen Segmentierung von Gezeitenzonen und Biotopen genutzt werden kann (TE-UNet). Neben Intensitätswerten von SAR Aufnahmen, werden auch polarimetrische Zerlegungen berücksichtigt. Radarsat-2 (C-Band) und ALOS-2 (L-Band) Aufnahmen werden übereinandergelegt, um aus Mehrbanddaten zu lernen. Eine vergleichende Studie demonstriert die gesteigerte Effektivität bei Nutzung von Mehrbanddaten und verschiedenen Polarisationen für die Klassifikation von Gezeitenzonen.

(3) Zur Klassifikation von kombinieren optischen und SAR-Aufnahmen wird ein auf UNet basierendes Modell (SOF-UNet) vorgestellt, das auf den Datensatz SEN12MS angewendet wird. Das Modell nutzt zwei Inferenzpfade und besteht aus zwei Encoder-Modulen zur Featureextraktion, einem geteilten Decoder zum Upsampling der Featuremaps, sowie speziellen Direktverbindungen (skip connections) zur Verschmelzung multimodaler Features. Die qualitative wie quantitative Analyse der Klassifikationsergebnisse zeigt ein vielversprechendes Potential des Modells, verschiedene Klassen von Bodenbedeckungen zuzuordnen und Details in den Vorhersagemasken zu erhalten.

Contents

List of Figures

List of Tables

Chapter 1

Introduction

1.1 Background and Motivation

1.1.1 Synthetic Aperture Radar

Synthetic Aperture Radar (SAR) is one of the most commonly used earth observation instruments on satellites. As an active microwave sensor, SAR operates by illuminating the target area with electromagnetic microwave pulses and measures the amplitude, polarization, and phase information of the return signals. After substantial signal processing of the collected data, the final SAR product is a two-dimensional image, where each pixel in the image represents the reflectivity of a region at the transmitted frequency [1].

SAR can achieve a high spatial resolution by utilizing the coherent nature of the transmitted radar pulses. Compared with optical sensors, the relatively long wavelengths and active imaging technique make SAR acquire data under all weather and all illumination conditions. Further, due to the sensitivity of the electromagnetic wave to dielectric and geometric properties of the scatterers, polarization information of the targets is also included in the echoes of SAR systems. Polarization information is an indispensable description of targets reflecting unique properties like surface roughness, structural symmetry, and orientation, which is beneficial for the identification of targets on SAR images.

Because of these advantages, SAR imagery is helpful in almost any field that benefits from environmental monitoring. Its application areas involve geographic mapping [2], resource surveying [3], climatic change [4], autonomous driving [5], and many other civil fields. Besides, SAR is also widely used in military applications, such as battlefield situation assessment, military reconnaissance, and target detection [6].

SAR image interpretation is to acquire key information from SAR images. It is a decisive step for the successful application of in-orbit SAR satellites. The geometric and electromagnetic characteristics of SAR data could provide distinctive information for image interpretation. However, because of the side-looking imaging geometry and complex backscattering mechanisms, SAR image quality is strongly affected by geometric distortion and speckle noise. Therefore, it is a very challenging task to realize successful SAR image interpretation.

The early SAR systems only worked in a single frequency band and at a single polarization. In recent years, SAR imaging technology has been developed towards a high-resolution, multi-band, and multi-polarization direction. Consequently, the information in SAR imagery has been largely increasing with the fast development of SAR sensors. The workload of manual SAR image interpretation exceeds the limit of rapid manual judgment. The subjective and comprehension errors caused by manual interpretation are also unavoidable. Hence, the research on the automatic interpretation of SAR images is particularly important.

1.1.2 Deep Learning

Deep learning, which has attracted broad attention in recent years [7], is a powerful tool focusing on Deep Neural Network (DNN). It refers to a Neural Network (NN) involving usually more than two “hidden” layers (they are called deep for this reason). A NN creates interconnected nodes, which represent non-linear mappings connected by linear transforms. The linear transform is performed on a matrix of weights, and the non-linear mapping is referred to as the activation functions [8]. Unlike conventional algorithms, deep learning-based methods commonly employ hierarchical architectures to extract feature representations exclusively from input data. For example, a Convolutional Neural Network (CNN) is capable of learning low-level and high-level features from input images with stacks of convolutional and pooling layers [9].

DNN became an explicit research subject in the early 1990s [10], but it was ignored at that time due to the difficulty of training. In 2012, AlexNet, a CNN-based architecture, famously won the 2012 ImageNet competition overwhelmingly [11]. This was mainly because of the use of large-scale training data, Rectified Linear Units (ReLU), and Graphics Processing Units (GPU) [12]. After the ImageNet competition, deep learning techniques have been widely adopted and verified effective for different fields in artificial intelligence such as image processing, natural language processing, and robotics [13].

In the wake of this success and thanks to the increased availability of remote sensing data and computational resources, the remote-sensing community has shifted its attention to deep learning since 2014 [14]. Concerning SAR applications, deep learning technology is gradually applied in various intelligent SAR image interpretation tasks, such as SAR object detection, SAR semantic segmentation, parameter inversion, despeckling, specific applications in Interferometric SAR (InSAR), and SAR-optical fusion [9].

1.1.3 Interdisciplinary Motivation

The objective of this thesis is at the intersection of two important trends: deep learning, a driver of disruptive innovation, especially in computer vision, and exploitation of SAR technologies and analysis, which is expected to show strong growth in the remote sensing field. These two trends meet up in a field, where data is both an opportunity and a challenge. The aim of this interdisciplinary effort is to utilize deep learning technology for intelligent SAR image interpretation, even when oriented to non-SAR professionals.

Deep learning methods are considered as the main means for intelligent SAR image interpretation in the future [9]. Traditional SAR image interpretation technology is composed of multiple individual steps. Taking SAR semantic segmentation as an example, we need to perform feature extraction, feature selection, feature fusion, and classifier selection separately [15]. Such complex procedures require much time and energy, and also harm integral performance. In contrast, end-to-end deep learning algorithms can automatically learn the most discriminative information from SAR data and thus, the efficiency can be improved dramatically once a deep learning model is well trained. Moreover, deep learning-based models have advantages of good expansibility and adaptability [16], and can be easily adapted to new SAR targets and new complicated applications.

However, there are also some challenges in this interdisciplinary field. On one hand, SAR imagery suffers from a special type of deterministic and multiplicative noise called speckle noise, which is caused by incoherent imaging mechanism [17]. The geometric distortions due to side-looking imaging further distort the interpretability of SAR images [18]. These problems are even more serious on high spatial resolution SAR imagery. On the other hand, the lack of balanced and large-scale sets of SAR-derived labels further limits the accurate interpretation.

For the moment, most of the existing deep learning methods in the SAR field tailor the models designed for optical images, lacking full consideration of SAR image characteristics [9]. Therefore, the development of deep learning-based approaches specially designed for SAR image interpretation is an open problem and forms the basis for the research topic of this thesis.

1.1.4 Thesis Focus

Through addressing three specific problems, this thesis aims to develop multiple effective deep learning-based SAR interpretation models, making full use of the target properties and SAR image characteristics.

As shown in Figure 1.1, the research aspects in this thesis are determined by geographical positions. For different Earth's surface types, we selected corresponding typical SAR image interpretation tasks to accomplish. The three specific aspects are oceanic eddy detection, intertidal sediments and habitats classification, and land cover classification. They correspond to three SAR image interpretation tasks: SAR object detection, SAR semantic segmentation, and SAR-optical data fusion.

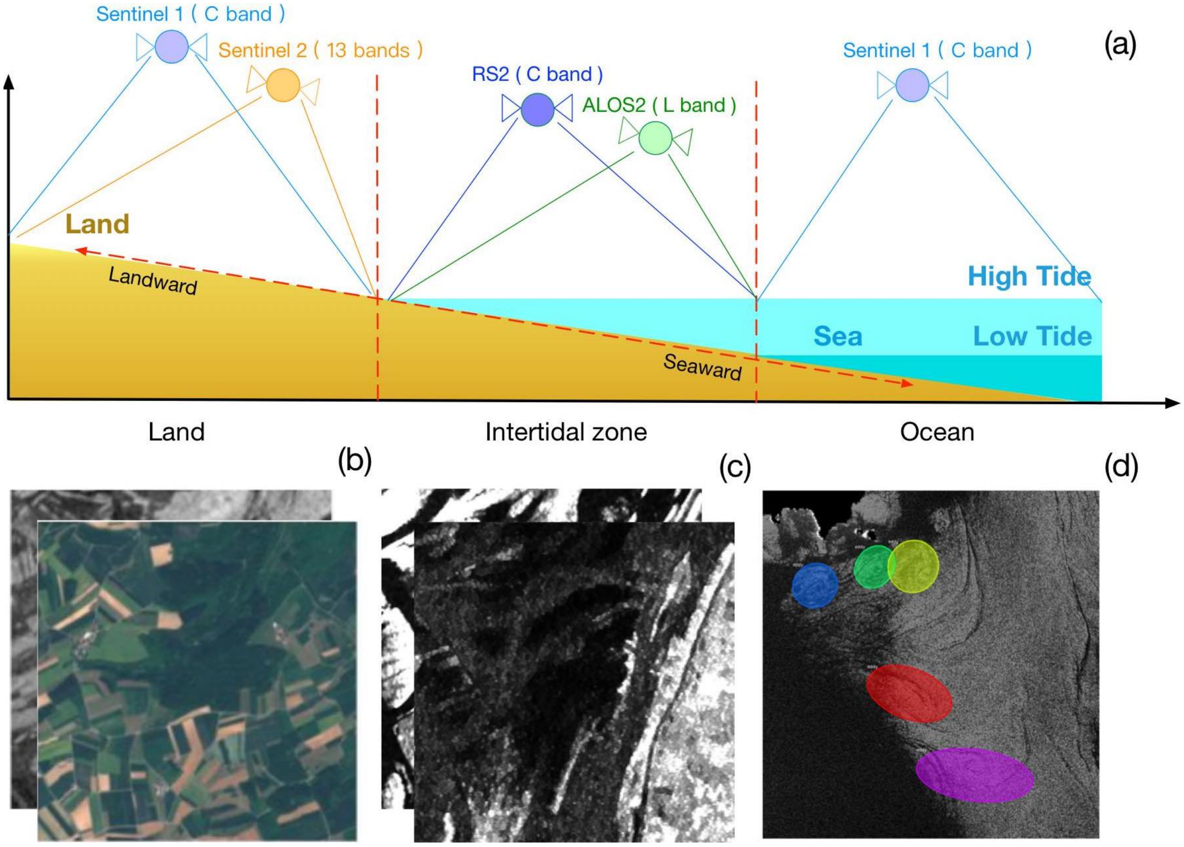


Figure 1.1: Three research aspects in the thesis. (a) presents the extent of the study area delineated by two vertical red dashed lines. (b–d) are datasets examples for land cover classification (SAR + Optical), intertidal sediments and habitats classification (multi-band multi-pol SAR), and oceanic eddy detection (single-channel SAR), respectively.

For these three aspects, the research data sources are increased from single-channel Ground Range Detected (GRD) SAR products, multi-band and multi-polarization SAR images, to multi-modal SAR and optical data. Accordingly, the complexity and difficulty of the three research aspects also increase. Specifically, we need to deal with more classification types and more complex backgrounds. Notably, considering the differences in problem complexity and data source among these three aspects, we design different specific models in this thesis, respectively. The names and band types of satellites that show in Figure 1.1 will be further introduced in Section 2.1.

1.2 Thesis Contributions

The main scientific contributions of this thesis are summarized as follows:

- Firstly, this thesis proposes a Mask Edge Enhancement and IoU Score Region-based Convolutional Neural Network (Mask-ES-RCNN) model for automatic oceanic eddy detection. We first construct a new SAR Oceanic Eddy Detection Dataset (SOEDD) and develop a Mask RCNN and Edge Enhancement model based on it. The Mask RCNN and Edge Enhancement model uses edge detection as an intuitive way to enhance texture information. However, it is not an end-to-end deep learning model and the strategy to incorporate prior knowledge is too simple. We, therefore, propose the Mask-ES-RCNN model applying implicit learning of internal texture information and scoring Mask IoU to focus more on mask quality. The performance of the Mask-ES-RCNN model outperforms the Mask RCNN baseline on the SOEDD.

- Secondly, this thesis presents a UNet-based model with a Texture Enhancement Module (TE-UNet) for the classification of sediments and habitats on the intertidal zone. The Texture Enhancement Module (TEM) helps the model to learn global texture information explicitly. Both Radarsat-2 (C band) and ALOS-2 (L band) SAR images are used as the inputs of TE-UNet. Apart from amplitude information of SAR images, we also add polarimetric information to TE-UNet by polarimetric decomposition. Extensive ablation studies are investigated for multi-band and multi-polarization configurations.

- Finally, this thesis develops a SAR-Optical Fusion UNet model (SOF-UNet) based on the SEN12MS [19] dataset which provides optical and SAR pairs for land cover classification. We first propose an initial SAR-Optical Fusion Network (SOFNet), which provides better segmentation results, compared with the methods that simply superimpose SAR and optical images as input. However, the contour lines of SOFNet predictions are very vague and lots of details are ignored by the model. We then propose the SOF-UNet model, which includes designed skip connections in the encoding and decoding phase to keep more details of predictions and extract more discriminative multi-modal features. The experiment results show that this design has a promising capability to identify different land cover classes. Symmetric Cross Entropy (SCE) loss is also verified as useful in this frame.

1.3 Thesis Structure

The thesis structure is illustrated in Figure 1.2. The whole thesis is structured into seven chapters. Chapter 1, 2 and 3 present the introduction and theory basics of the thesis topic. Chapter 4, 5, and 6 present three different SAR image interpretation applications and corresponding algorithms.

Chapter 1 presents the introduction to the thesis. Chapter 2 introduces the basics of remote sensing, and Chapter 3 summarizes the challenges and recent advances in SAR image interpretation utilizing deep learning technologies.

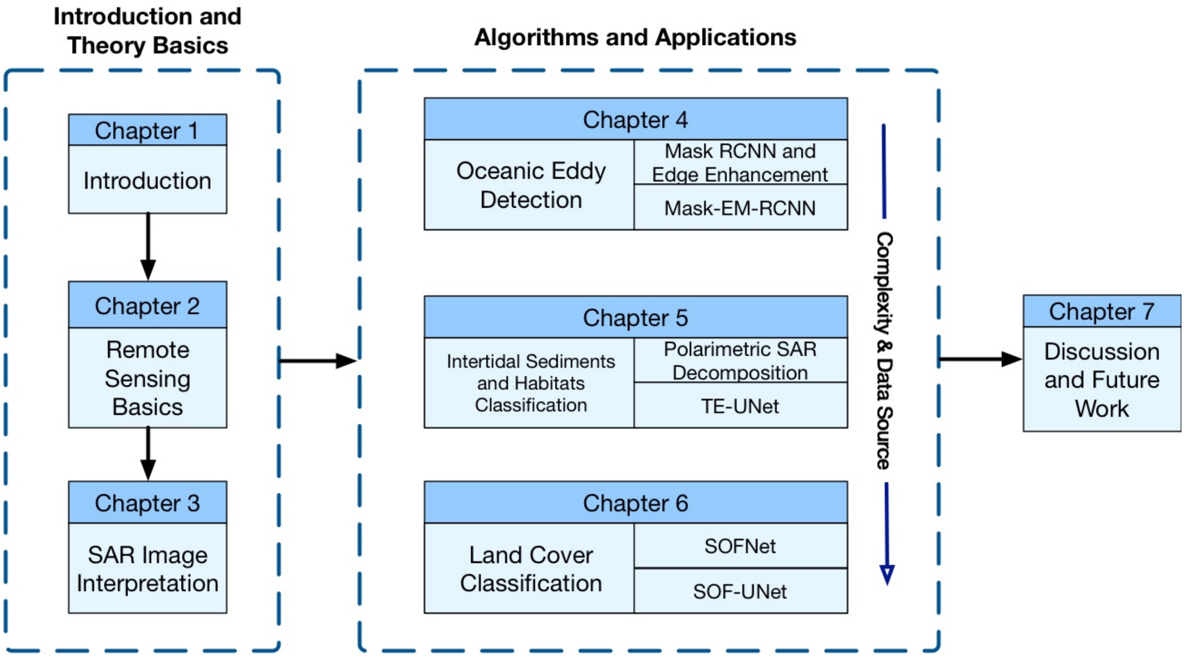

Introduction and Theory Basics
Chapter 1
Introduction
Chapter 2
Remote Sensing Basics
Chapter 3
SAR Image Interpretation
Algorithms and Applications
Chapter 4
Oceanic Eddy Detection
Mask RCNN and Edge Enhancement
Mask-EM-RCNN
Chapter 5
Intertidal Sediments and Habitats Classification
Polarimetric SAR Decomposition
TE-UNet
Chapter 6
Land Cover Classification
SOFNet
SOF-UNet
Complexity & Data Source
Chapter 7
Discussion and Future Work

Figure 1.2: Organization of the thesis.

The following three chapters present each contribution of this thesis:

- Chapter 4 introduces ocean eddy detection by two models. The Mask RCNN and Edge Enhancement model is an initial model on the SOEDD. Inspired by this model, we design a multi-task learning framework Mask-EM-RCNN.
- Chapter 5 introduces intertidal sediments and habitats classification by TE-UNet model. We first design a pre-processing procedure on SAR data, and put emphasis on the polarimetric SAR decomposition process. The processed multi-band multi-polarization SAR images are then used as the input of the TE-UNet model.
- Chapter 6 introduces land cover classification by two multi-modal fusion networks. The SOFNet is a Deeplab V3 Plus-based two-stream model. Since the SOFNet loses detailed information in prediction maps, we further develop a SOF-UNet model to utilize

more low-level features.

Finally, Chapter 7 summarizes the key ideas and approaches described in the thesis. It presents the main achievements of this thesis, concludes the inspirations and limitations of the presented work, and suggests future research directions.

Chapter 2

Basics of Remote Sensing

Remote sensing is devoted to gathering information from the Earth's surface at a distance [20], by means of electromagnetic energy or other mediums. This chapter presents some basics of remote sensing that we will use in the rest of the thesis, and puts emphasis on the basic theory of SAR. In Section 2.1, we introduce optical and radar sensors and corresponding common satellites. The working principle and theory of polarimetric SAR sensors are described in Section 2.2 and Section 2.3, respectively.

2.1 Remote Sensors

For the case of this thesis, remote sensors remotely collect data by measuring the electromagnetic (EM) radiation at specific spectral ranges (usually called bands). When they are deployed on satellites or mounted on aircraft, they are called spaceborne and airborne remote sensors, respectively. Spaceborne remote sensors orbit the Earth at heights of 500 to 800 km [21], providing data in a wide range of ground resolutions, and at various polarizations and radar bands. In this thesis, we analyze the spaceborne remote sensing data.

There are two types of remote sensors acquiring information in fundamentally different ways. Optical sensors are passively imaging systems. They can capture information on the physical and biogeochemical properties of the Earth's surface. Radar sensors actively emit EM-radiation and then measure the returning signals (also known as backscatter), which depends on the three-dimensional structure of target objects.

Figure 2.1 displays the different spectral regions of these two types of sensors within the EM-spectrum. More details in this figure on optical sensors and radar sensors are further described in Section 2.1.1 and Section 2.1.2, respectively.

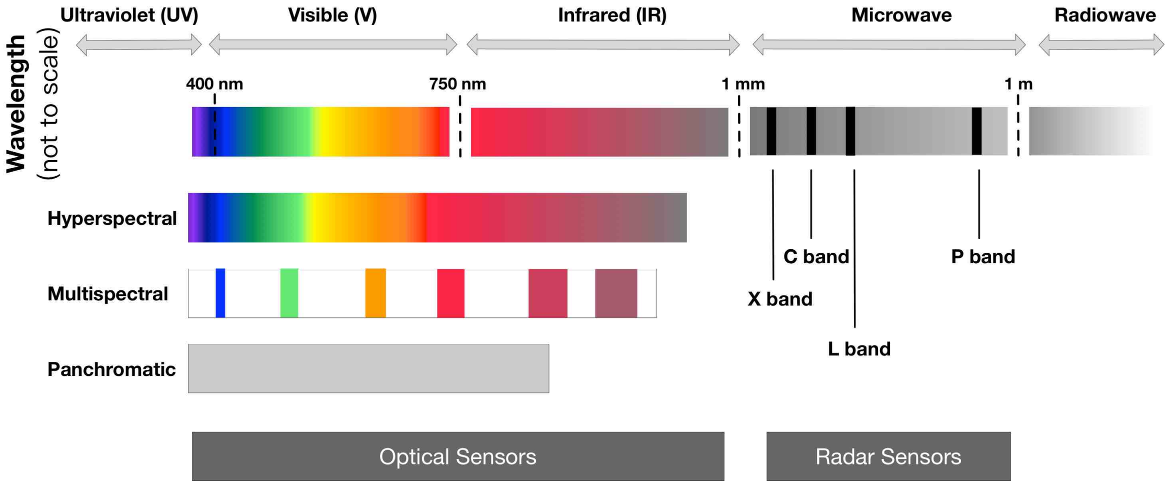


Figure 2.1: Spectral regions of optical sensors and radar sensors within the electromagnetic spectrum. Image adapted from [22].

2.1.1 Optical Sensors

Optical sensors are sensitive to a spectrum ranging from visible to infrared wavelengths, and they produce panchromatic, multispectral, or hyperspectral images. The panchromatic sensors use a monospectral channel detector to collect EM-radiation from a wide range of wavelengths, while multispectral and hyperspectral sensors collect information using multiple channels. In general, hyperspectral data contain hundreds of bands showing a high spectral resolution. Multispectral data contain fewer spectral bands than hyperspectral images, but more than panchromatic images [23].

A comparison among the common optical sensors is established in Figure 2.2. The number of bands differs from instrument to instrument, but they still have lots of intersection sets. MODIS and ASTER are both spaceborne imaging instruments on the Terra and Aqua platforms launched in December 1999. They are the result from a collaboration of the National Aeronautics and Space Administration (NASA) and Japan's Ministry of Economy Trade and Industry (METI). MODIS and ASTER require data in 36 and 14 bands, respectively. Landsat 8 data is divided into 11 bands. It is a joint effort of NASA and the United States Geological Survey (USGS), and carries the Operational Land Imager (OLI) and the Thermal Infrared Sensor (TIRS) launched in February 2013. The European Sentinel-2 carries a payload of sensors working at 13 bands, including visible, near-infrared, and shortwave infrared sensors. The Sentinel-2 satellite was launched in the frame of the European Copernicus Program in June 2015. The band designations for optical sensors in Figure 2.2 can help users to decide which spectral bands work best to identify their features of interest for image interpretation.

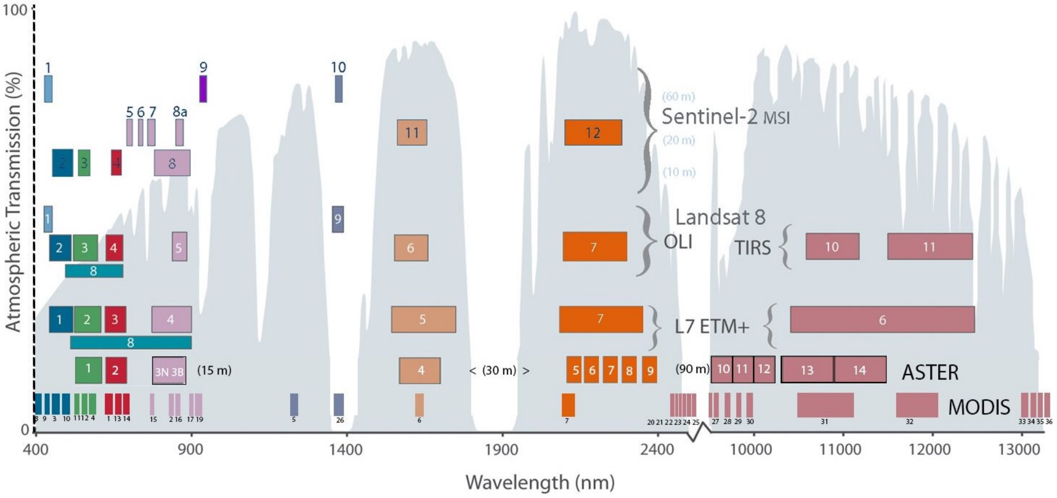


Figure 2.2: Comparison of MODIS, ASTER, Landsat 8 and Sentinel-2 bands. Image adapted from USGS Landsat Program [24].

2.1.2 Radar Sensors

Radar sensors work in the region of wavelengths ranging from 1mm to 1m. Compared with optical sensors, they utilize longer wavelengths, which allows "seeing" through clouds and rain. There are three common spaceborne microwave sensors: radar altimeter, scatterometer, and SAR. The radar altimeter and scatterometer are used to measure specific parameters like Sea Surface Height (SSH) and Sea Surface Wind (SSW). SAR sensors provide radar images of the Earth's surface, which are chosen in this thesis.

Different radar bands are marked with letters. Table 2.1 contains the band with the associated frequency and wavelength. The wavelength range is very important, since it determines the interaction between radar signals and surfaces, as well as the penetration depth of the microwave signals. The frequency bands explored in this thesis are C- and L-band.

Band	Frequency (GHz)	Wavelength (cm)
Ka	27-40	1.1-0.8
K	18-27	1.7-1.1
Ku	12-18	2.4-1.7
X	8-12	3.8-2.4
C	4-8	7.5-3.8
S	2-4	15-7.5
L	1-2	30-15
P	0.3-1	100-30

Table 2.1: Different bands in microwave remote sensing.

Table 2.2 gives some examples of commonly used spaceborne SAR sensors, which are operating at present. They are Radarsat-2 (RS2) from Canadian Space Agency (CSA), TerraSAR-X/TanDEM-X from German Aerospace Center (DLR), ALOS PALSAR-2 (ALOS2) from Japan Aerospace Exploration Agency (JAXA), Sentinel-1A/1B from European Space Agency (ESA), and Gaofen-3 from China National Space

Administration (CNSA). We use RS2, ALOS2, and Sentinel 1A/1B data in the thesis. The parameters of polarimetric modes, spatial resolutions, and coverage will be described in Section 2.2 and Section 2.3.

Mission/SAR	Agency	Launch	Band	Polarization	Resolution(m)	Coverage(km)	Revisit(days)
Radarsat-2	CSA	2007	C	Quad	9-100	25-170	24
TerraSAR-X/TanDEM-X	DLR	2007(2010)	X	Quad	0.25-40	10-150	11
ALOS PALSAR-2	JAXA	2014	L	Quad	1-100	25-490	14
Sentinel 1a/1b	ESA	2014(2016)	C	Dual	5-100	80-400	12
Gaofen-3	CNSA	2016	C	Quad	1-500	10-650	29

Table 2.2: List of operational typical spaceborne SAR systems.

2.2 Fundamentals of SAR Imaging

The principle of the active microwave imaging of SAR sensors makes it fundamentally different from optical sensors. We describe the SAR geometry and spatial resolution in Section 2.2.1. Then we introduce three typical SAR acquisition modes, which will affect the spatial resolution and coverage in Section 2.2.2. In Section 2.2.3, the scattering mechanisms of the SAR sensors are given.

2.2.1 SAR Geometry and Spatial Resolution

The geometry of a side-looking monostatic SAR system is sketched in Figure 2.3. An antenna, placed on a moving platform, transmits electromagnetic pulses in a side-looking direction towards the Earth's surface (side-looking system). The reflected signal, known as the echo, is backscattered from the surface and received by the same antenna (monostatic radar). The length and width of the antenna, indicated by A_L and A_W, determine the size of the illuminated area on the ground. A SAR sensor moves with constant speed V_s and at constant height h above the ground. The flight direction of SAR is defined as azimuth (or along-track) direction, while the direction of radar illumination is referred to as the range direction.

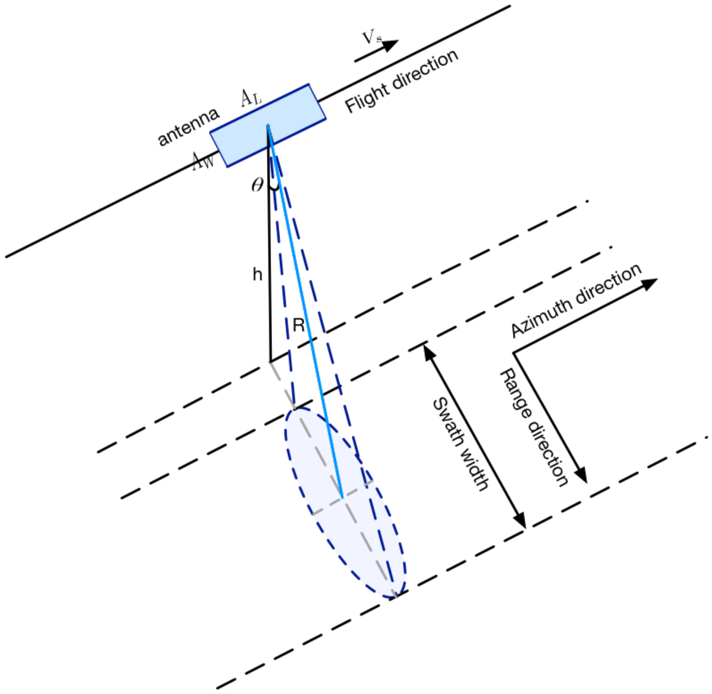


Figure 2.3: Schematic SAR acquisition geometry.

The spatial resolution in the ground range direction is define as:

$$r_{gr} = \frac{c_0 \tau}{2 \sin \theta} \tag{2.1}$$

where c_0 is the speed of light, τ is the pulse duration, θ is the incidence angle. A typical setting of these parameters ($\tau = 10\mu s, \theta = 30°$) [25] produces a r_{gr} equal to 3000m, which is not satisfactory to most SAR applications. A pulse compression method is applied to improve the resolution. The final formula of r_{gr} is expressed as:

$$r_{gr} = \frac{c}{2B \sin \theta} \tag{2.2}$$

where B is the pulse bandwidth. Typical values of r_{gr} using pulse compression are below ten meters.

The spatial resolution in azimuth direction is defined as:

$$r_{az} = \frac{\lambda}{A_L} R \tag{2.3}$$

where λ is the radar wavelength and R is the slant range between the antenna and a ground resolution cell. A typical setting of these parameters on airborne system ($\lambda = 0.03m, A_L = 3m, R = 2000m$) [25] produces a r_{az} equal to 20m. However, for the spaceborne SAR, the large R in the space results in very coarse r_{az} (above 10km). It is not feasible to increase the size of the spaceborne antenna, so the synthetic aperture technology is used, as shown in Figure 2.4. The small aperture radar antenna is virtualized into a larger aperture radar antenna by utilizing the motion of the antenna. We can finally receive the r_{az} in the finer resolution after signal processing according to the Doppler and phase history:

$$r_{az} = \frac{A_L}{2} \tag{2.4}$$

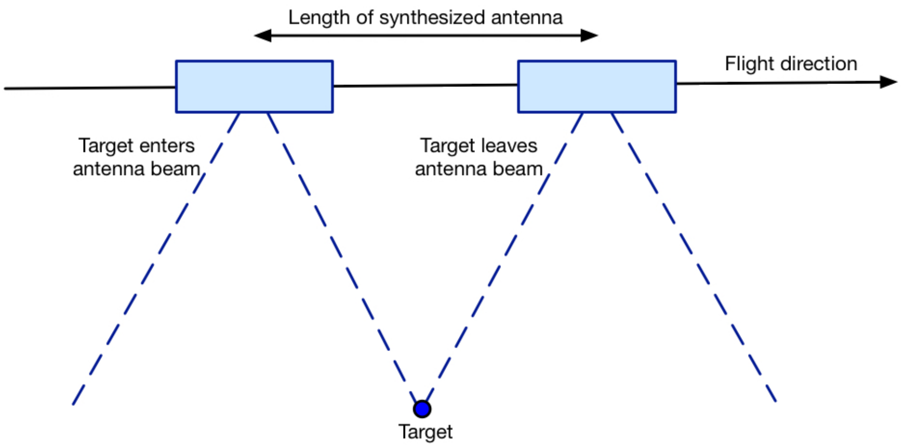

Length of synthesized antenna
Flight direction
Target enters antenna beam
Target leaves antenna beam
Target

Figure 2.4: Formation of a synthetic antenna array.

2.2.2 Acquisition Modes

Most SAR sensors can operate in different acquisition modes. The final products of different modes vary in spatial resolution, coverage, and polarimetric ways. Figure 2.5 illustrates three typical acquisition modes.

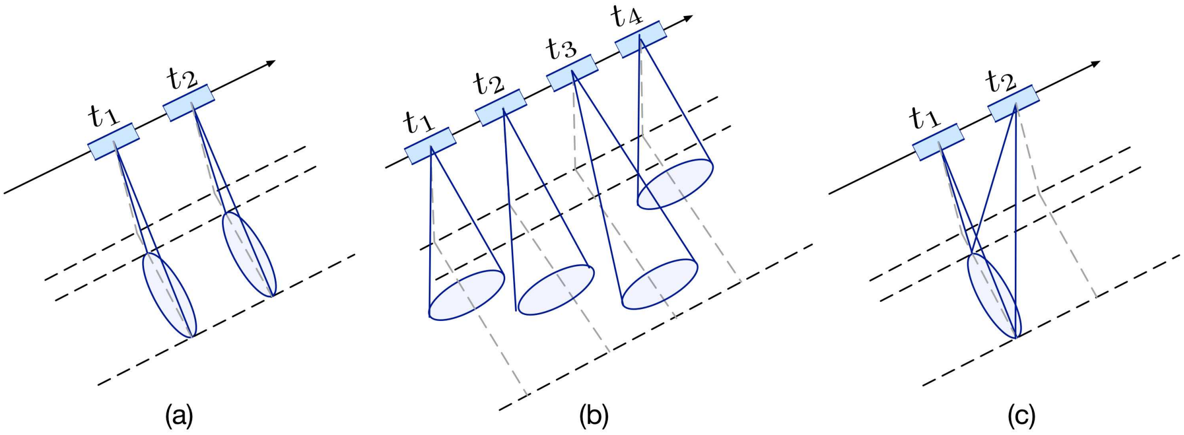


Figure 2.5: SAR acquisition modes: (a) Stripmap (b) ScanSAR (c) Spotlight.

Figure 2.5 (a) is the principle of Stripmap mode, which is also the most commonly used mode. The antenna beam is pointing to a fixed azimuth angle and then the ground swath is realized by a continuous sequence of the pulse. The antenna usually gives the flexibility to select an imaging swath by changing the incidence angle. Figure 2.5 (b) is the wider swath ScanSAR mode and Figure 2.5 (c) is the higher resolution Spotlight mode. ScanSAR mode achieves swath widening by the use of an antenna beam that is electronically steerable. Each sub-swath is illuminated by multiple pulses but in a shorter time than the Stripmap mode. However, the azimuth resolution is degraded correspondingly. For a better azimuth resolution, the Spotlight mode can be chosen. It adjusts the antenna beam to continuously illuminate the same patch for a longer time, thereby realizing a higher azimuthal resolution.

2.2.3 Scattering Mechanisms

A SAR sensor measures the electromagnetic energy that is backscattered from the targets. There are lots of factors that can affect the radar backscatter, and they can be divided into two categories.

The first category relates to the SAR parameters, such as the frequency, polarization, and incident angle. If the frequency and polarization parameters are fixed, the increasing incident angle causes the decreasing backscatter intensity from a homogeneous surface. Therefore, the intensity decreases gradually on SAR images from near range to far range. This effect must be taken into consideration during SAR image interpretation.

The second category relates to the surface parameters, such as the surface roughness, the surface geometry, and the dielectric constant of the surface. There are three basic groups of scattering mechanisms that can contribute to the returned signal: surface scattering, volume scattering, and hard target scattering, as illustrated in Figure 2.6.

In general, a rougher surface results in a stronger backscatter intensity, corresponding to a brighter area on SAR imagery. In the case of the ocean surface, a smooth ocean surface causes low (or no) radar backscatter (Figure 2.6 (a)), showing dark areas on the SAR image. For the areas of moderate ocean surface roughness (Figure 2.6 (b)), such as areas in moderate surface wind speed, they usually present as greyish areas. A rough ocean surface (Figure 2.6 (c)) caused by high surface wind speed or other factors results in bright areas.

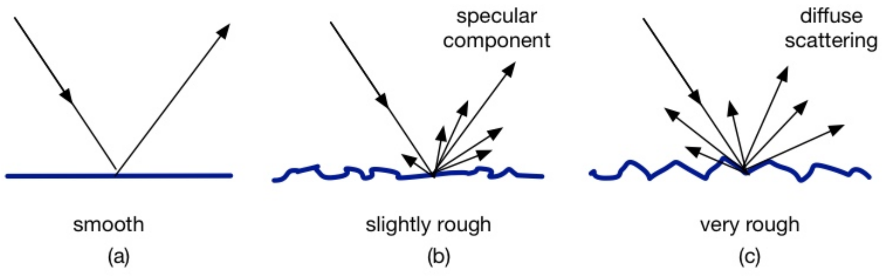


Figure 2.6: Specular and diffuse surface scattering, depending on surface roughness.

2.3 Polarimetric SAR Theory

SAR polarimetry is a widely used technique for the derivation of qualitative and quantitative physical information for different interpretation tasks. Measuring the full scattering matrix allows distinguishing different shapes, orientations, and dielectric properties of scatterers. In Section 2.3.1, we first introduce the basic concepts of electromagnetic polarization. The polarization scattering description is then shown in Section 2.3.2 to represent the polarization state. Based on the representations in Section 2.3.2, we describe the polarimetric decomposition methods in Section 2.3.3.

2.3.1 Electromagnetic Polarization

An electromagnetic wave consists of a magnetic and an electric field. These two fields are perpendicular to each other and also to the direction of wave propagation. Apart from frequency, amplitude, and phase, an electromagnetic wave also contains polarization information. Polarization is defined as the orientation of the oscillating electric field, which can be described in terms of two orthogonal basis vectors [26]. Electromagnetic waves are generally elliptically polarized, with linear or circular polarization as special cases [27]. Most of the SAR sensors use linear polarization on both the transmitter and the receiver. There are four linear polarization configurations in total:

- HH (co-polarization): horizontal transmission and horizontal reception;
- HV (cross-polarization): horizontal transmission and vertical reception;
- VH (cross-polarization): vertical transmission and horizontal reception;
- VV (co-polarization): vertical transmission and vertical reception.

Historical SAR satellites carried single-polarized sensors, which support only one linear polarization. More recent sensors provide either dual-polarization or quad-polarization capabilities. For quad-polarization SAR systems, they can transmit H- and V-polarized waveforms and receive both H and V simultaneously.

It is important to acquire the polarization information, because different polarization channels interact differently with the targets during the scattering process. For example, the echo of co-polarization is stronger than that of cross-polarization for low vegetation. Cross-polarization is more sensitive to some tiny targets like cars, compared with co-polarization.

2.3.2 Polarization Scattering Description

The basic concept of SAR polarimetry is given by the 2×2 complex scattering matrix:

$$\mathbf{E}^r = \begin{bmatrix} E_H^r \\ E_V^r \end{bmatrix} = [\mathbf{S}] \cdot \mathbf{E}^t = \frac{e^{ik_0 R}}{R} \cdot \begin{bmatrix} S_{HH} & S_{HV} \\ S_{VH} & S_{VV} \end{bmatrix} \cdot \begin{bmatrix} E_H^t \\ E_V^t \end{bmatrix} \tag{2.5}$$

where $\mathbf{E}^t$ is the two-dimensional transmitted wave vector, $\mathbf{E}^r$ is the two-dimensional received wave vector, R is the distance between the radar antenna and the ground target, k_0 is the radar wave number, $[\mathbf{S}]$ is the polarization scattering matrix that describes how the scatterers modify the incident electric field vector. The reciprocity theorem states that $S_{VH} = S_{HV}$ [25], which is adequate for SAR remote sensing from space.

Generally speaking, $[\mathbf{S}]$ is not only defined by the physical factors of the target like materials and structures, but also highly related to SAR parameters such as the relative position between radar and target and SAR frequency.

For convenience, we often need to vectorize the target's polarimetric scattering matrix. Different orthogonal basis corresponds to different polarization basis expression methods. There are two common polarimetric basis Borgeaud $\vec{k}_B$ and Pauli $\vec{k}_P$. Under the condition that the reciprocity theorem is satisfied, they are donated as:

$$\vec{k}_B = \left[S_{HH}, \sqrt{2}S_{VH}, S_{VV}\right]^T \tag{2.6}$$

$$\vec{k}_P = \frac{1}{\sqrt{2}}\left[S_{HH}+S_{VV}, S_{HH}-S_{VV}, 2S_{VH}\right]^T \tag{2.7}$$

A polarimetric scattering matrix is used for the point target case. Based on the polarimetric basis, we can formalize a 3 × 3 coherency or covariance matrix to characterize the distributed scatterers.

The coherency matrix [**C**] based on Borgeaud $\vec{k}_B$ is denoted as:

$$[\mathbf{C}] = E\left\{\vec{k}_{\mathrm{B}} \cdot \vec{k}_{\mathrm{B}}^{\dagger}\right\} = \begin{bmatrix} E\left\{\left|S_{\mathrm{HH}}\right|^2\right\} & \sqrt{2}E\left\{S_{\mathrm{HH}}S_{\mathrm{HV}}^*\right\} & E\left\{S_{\mathrm{HH}}S_{\mathrm{VV}}^*\right\} \\ \sqrt{2}E\left\{S_{\mathrm{HV}}S_{\mathrm{HH}}^*\right\} & 2E\left\{\left|S_{\mathrm{HV}}\right|^2\right\} & \sqrt{2}E\left\{S_{\mathrm{HV}}S_{\mathrm{VV}}^*\right\} \\ E\left\{S_{\mathrm{VV}}S_{\mathrm{HH}}^*\right\} & \sqrt{2}E\left\{S_{\mathrm{VV}}S_{\mathrm{HV}}^*\right\} & E\left\{\left|S_{\mathrm{VV}}\right|^2\right\} \end{bmatrix} \tag{2.8}$$

The coherency matrix [**T**] based on Pauli $\vec{k}_P$ is denoted as:

$$[\mathbf{T}] = E\left\{\vec{k}_{\mathrm{P}} \cdot \vec{k}_{\mathrm{P}}^{\dagger}\right\} = \begin{bmatrix} T_{11} & T_{12} & T_{13} \\ T_{12}^* & T_{22} & T_{23} \\ T_{13}^* & T_{23}^* & T_{33} \end{bmatrix} \tag{2.9}$$

$$T_{11} = \frac{1}{2}E\left\{\left(S_{\mathrm{HH}}+S_{\mathrm{VV}}\right)\left(S_{\mathrm{HH}}+S_{\mathrm{VV}}\right)^*\right\} \tag{2.10}$$

$$T_{12} = \frac{1}{2}E\left\{\left(S_{\mathrm{HH}}+S_{\mathrm{VV}}\right)\left(S_{\mathrm{HH}}-S_{\mathrm{VV}}\right)^*\right\} \tag{2.11}$$

$$T_{13} = E\left\{\left(S_{\mathrm{HH}}+S_{\mathrm{VV}}\right)\left(S_{\mathrm{HV}}\right)^*\right\} \tag{2.12}$$

$$T_{22} = \frac{1}{2}E\left\{\left(S_{\mathrm{HH}}-S_{\mathrm{VV}}\right)\left(S_{\mathrm{HH}}-S_{\mathrm{VV}}\right)^*\right\} \tag{2.13}$$

$$T_{23} = E\left\{\left(S_{\mathrm{HH}}-S_{\mathrm{VV}}\right)\left(S_{\mathrm{HV}}\right)^*\right\} \tag{2.14}$$

$$T_{33} = 2E\left\{\left(S_{\mathrm{HV}}\right)\left(S_{\mathrm{HV}}\right)^*\right\} \tag{2.15}$$

where $(\cdot)^\dagger$ is the conjugate transpose operation, $(\cdot)^*$ is the complex conjugate operation, $E\{\cdot\}$ is the average expectation, and $|\cdot|$ is the amplitude of echos.

2.3.3 Polarimetric Decomposition

The purpose of polarimetric decomposition is to extract the physical information of the target surface. Through polarimetric decomposition methods, the scattering process of the target is decomposed into several terms representing different scattering mechanisms, each of which corresponds to a different physical meaning. These decomposition components are further used in different SAR image interpretation applications.

The polarimetric decomposition techniques are broadly classified into two categories: Coherent Target Decomposition (CTD) and In-Coherent Target Decomposition (ICTD). In the case of CTD, The scattering target is required to be deterministic or stationary, and the scattered echoes are coherent. For example, many man-made structures belong to such pure targets. The first proposed and most common CTD method is Pauli polarimetric decomposition based on Pauli basis $\vec{k}_P$ [28].

However, in natural scenes, there are lots of distributed targets. At this point, the scattering target can be non-deterministic and the echoes are corresponding incoherent. These scatterers can be analyzed by exploiting the coherency matrix $[\mathbf{C}]$ and the coherency matrix $[\mathbf{T}]$. Freeman-Durden [29] and Cloude-Pottier [30] are two most famous and widely used ICTD methods.

Chapter 3

SAR Image Interpretation

In this chapter, two main challenges are described in Section 3.1. Based on the basic deep learning concepts and models in Section 3.2, we finally summarize the related work of deep learning in SAR image interpretation in Section 3.3.

3.1 Challenges for SAR Image Interpretation

There are many factors that can affect the interpretability of SAR images. In this section, we focus on two main challenges resulting from the side-looking and the coherent SAR imaging mechanism. The principles of geometric distortions and speckle noise are described in Section 3.1.1 and Section 3.1.2, respectively.

3.1.1 Geometric Distortions

Geometric distortions are an inherent error of SAR images caused by side-looking geometry and topographic relief [31]. These distortions can be divided into different types. Figure 3.1 shows the origins and main characteristics of most related geometric distortions: foreshortening, layover, and shadow.

Figure 3.1(a) shows the geometric background of foreshortening. For the slopes facing the SAR sensor, when the incident angle θ is larger than the local terrain slope angle β, the slopes have the shorter length in SAR images (A′ to B′) compared with real flat terrain (the slope between points A and B). The foreshortening effect has the worst result, when θ equals β.

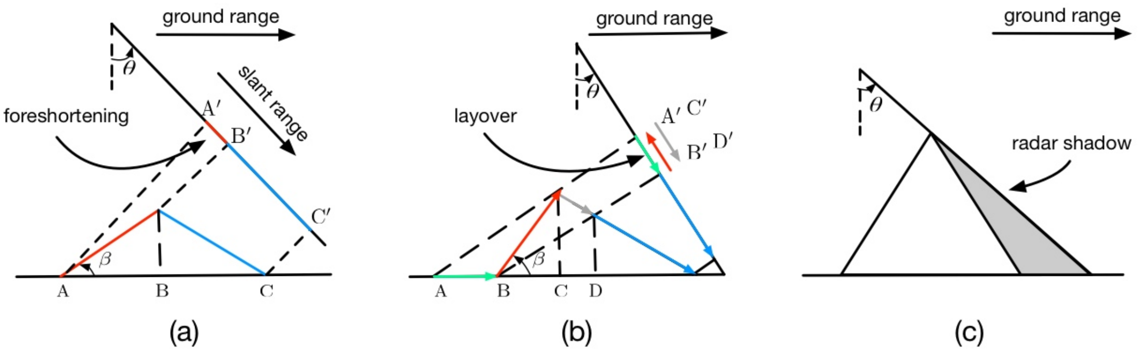


Figure 3.1: Main geometric distortions on SAR images with their dependence on acquisition geometry: (a) foreshortening, (b) layover, and (c) shadow. Image adapted from [32].

Figure 3.1(b) shows the geometric background of the layover. When the local incident angle θ is smaller than the local terrain slope angle β, the bottom and the top of such slopes are reversely imaged and their flipped backscatter will overlay in SAR images (green, red, and gray areas).

Both foreshortening and layover effects decrease with increasing incident angle θ. However, a large θ will result in a shadow problem. Figure 3.1(c) shows the geometric background of shadow. The area behind the slope is not illuminated by the radar. Therefore, geometric distortions caused by topography cannot be finally eliminated.

3.1.2 Speckle Noise

Speckle noise is formed because of the coherent imaging principle. Since many elemental scatterers are located in one resolution cell, the final scattering response from the resolution cell is the coherent sum of thousands of individual scattering events [32]. The mutual interference results in a certain fluctuation in the amplitude and phase of the synthesized EM wave vectors, which make the “salt-and-pepper” noise (also called speckle noise) appear. In the actual SAR images, speckle noise exists in the form of a multiplicative noise, which is manifested in the image as a drastic change in the image intensity.

The appearance of speckle noise degrades the SAR image quality and increases the difficulty of interpreting SAR images. Lots of effective speckle filters were developed during the past decades such as the refined Lee filter and Wiener filter [33]. However, the removal of the speckle noise means a reduction of spatial resolution to some extent. All speckle-noise reduction methods try to find a balance between them.

Figure 3.2 shows an original SAR image and corresponding speckle filter results. The original VV-polarization channel SAR image is acquired by RS2 satellite (used in Chapter 5). A refined Lee filter with a window size 7 × 7 is applied for speckle deductions. The results show that this filter obtains effective speckle noise removal, while preserving the fine edges of the original image.

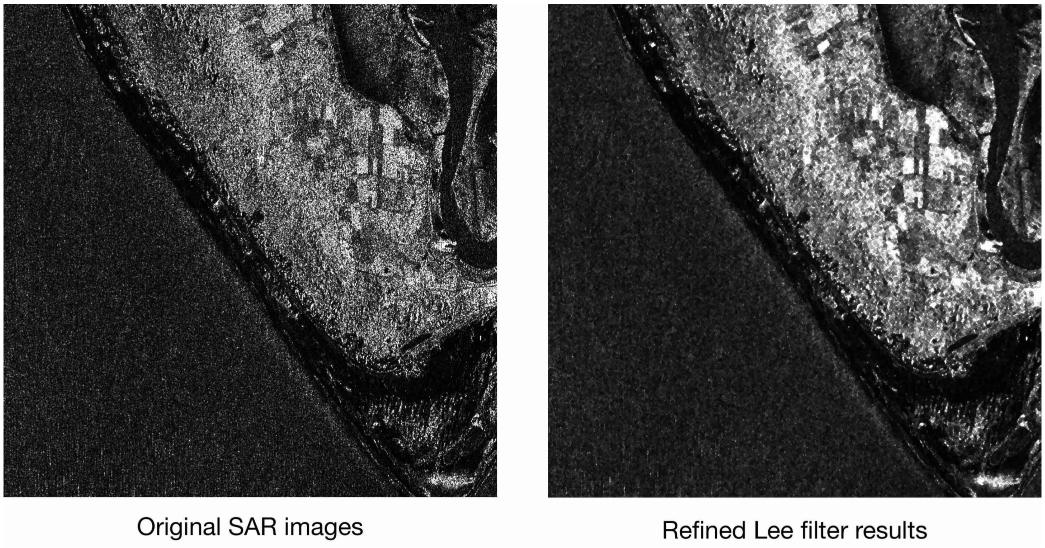


Figure 3.2: Comparison of an original SAR image (left) and corresponding speckle filter results (right). SAR image ©MacDonald, Dettwiler and Associates Ltd. 2015.

3.2 Deep Learning Models

In this section, we give a brief introduction to commonly used NN models, with emphasis on CNN and Graph Convolutional Network (GCN) models involved in this thesis. We then present some typical CNN models in Section 3.2.2.

3.2.1 Neural Networks

NN was initially inspired by the human brain for perception and cognition. It has been widely used in the computer vision field and achieved remarkable results. NN can transform the input data into a new feature space through nonlinear transformation, and can automatically learn feature representations.

Currently, many deep learning models have been proposed like CNN [34], GCN [35], Recurrent Neural Network (RNN) [36], and Generative Adversarial Network (GAN) [37]. Among them, CNN is the most widely used tool to abstract features.

Many networks employ pre-trained CNN on the ImageNet dataset as feature extractors, such as VGGNet [38], AlexNet [11], and GoogLeNet [39]. Furthermore, different variants of CNN like 3-Dimensional Convolutional Neural Network (3D-CNN) [40] and Spatial Convolutional Neural Network (SCNN) [41] have been proposed to improve the learning ability and adapt to different applications.

Recently, GCN has received increasing attention owing to its ability in performing convolutions on arbitrarily structured graphs. Figure 3.3 shows the convolution design of CNN and GCN. The biggest difference between them is that CNN performs convolutions on a Euclidean space like images, but GCN applies convolutions on a non-Euclidean space like graphs. GCN aggregates information from the neighbors of each node, which can be utilized to explore the relationship among objects.

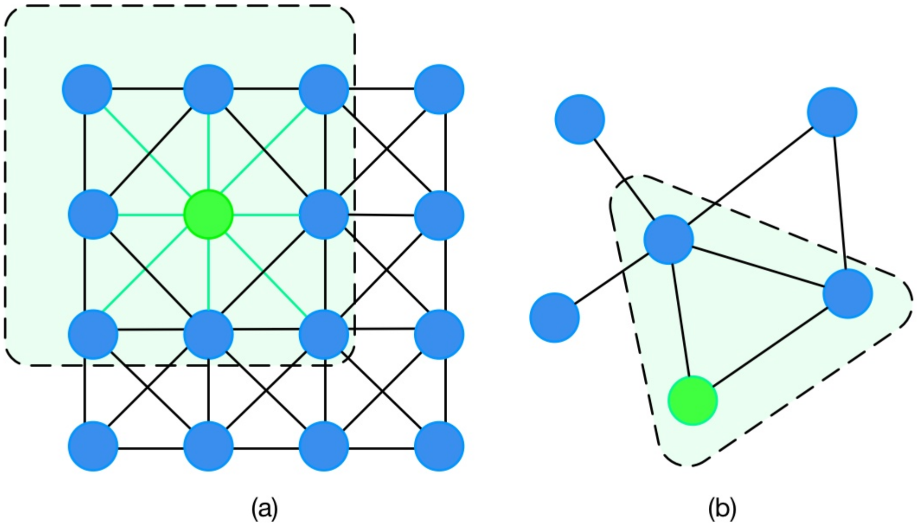


Figure 3.3: Comparison of convolution design between (a) CNN and (b) GCN. Image adapted from [42].

3.2.2 CNN Models

Object detection and instance segmentation are two basic tasks in computer vision. The former is object-level detection and the latter is pixel-level extraction. We discuss the classic CNN models of these two tasks in Section 3.2.2.1 and Section 3.2.2.2 separately.

3.2.2.1 Object Detection Models

Object detection is to determine, where objects are located (object localization), and which category each object belongs to (object classification) [43]. Deep learning has been successfully applied to object detection tasks. In 2014, Girshick et al. [44] proposed a Region Convolutional Neural Network (R-CNN) model using a selective search algorithm to extract regional candidate boxes. Based on R-CNN, many improved models have been designed, including Fast R-CNN [45], which learns classification and bounding box regression tasks at the same time, YOLO [46], which splits the input into a grid of cells to directly predict bounding box and classification, and Faster R-CNN [47], which uses a Region Proposal Network (RPN) to generate region proposals. In 2017, Mask R-CNN [48] was proposed and at that time it outperformed all existing solutions in object detection tasks.

Figure 3.4 illustrates the overall structure of Mask R-CNN, which is mainly composed of three parts: an RPN to extract the feature maps, a network head to generate the target classification and localization, and a network head for mask generation.

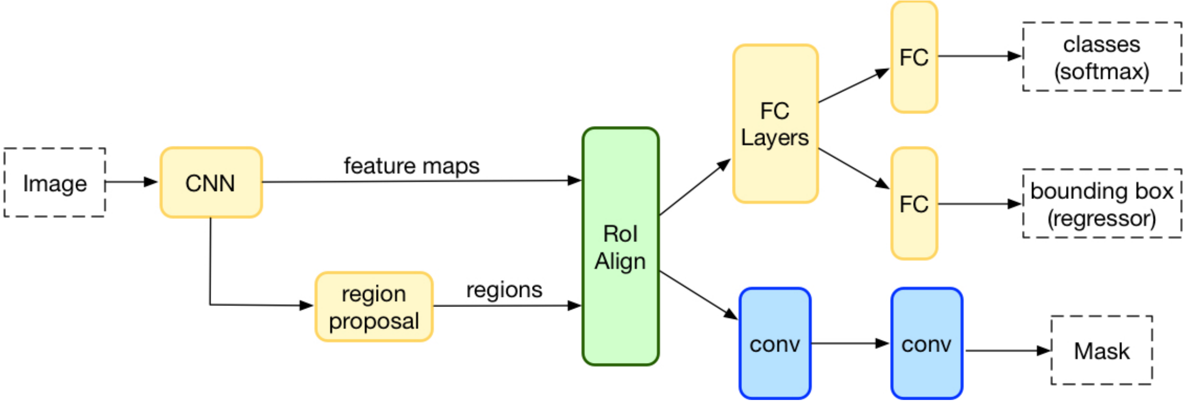


Figure 3.4: The overall structure of Mask R-CNN. Image adapted from [48].

3.2.2.2 Semantic Segmentation Models

The semantic image segmentation task is to classify each pixel of an image into a class [49]. Modern deep learning methods usually employ a Fully Convolutional Network (FCN) [50] to address this task. FCN consists of two parts: a downsampling path to extract and interpret the context, and an upsampling path for pixel localization. Following FCN, there are two main architectures for semantic segmentation, namely DilatedFCN and EncoderDecoder [51].

DilatedFCN applies dilated convolutions to capture multi-scale context information on the final feature maps. For instance, PSP-Net [52] uses pooling operations at multiple grid scales and DeepLabV3 [53] adopts Atrous Spatial Pyramid Pooling (ASPP) module.

EncoderDecoder consists of the encoding branch and a decoding branch, which gradually recover the spatial information using skip connections. UNet [54] and DeeplabV3 Plus [55] are two typical representatives of this type.

Figure 3.5 displays the overall architecture of UNet. The corresponding layers of the encoder and decoder network are connected by skip connections, prior to pooling and subsequent to a de-convolution operation, respectively.

Figure 3.6 shows the overall architecture of DeeplabV3 Plus. The encoder network abstracts the multi-scale contextual information with help of atrous convolutions, while the decoder module refines the boundaries of the segmentation results. It combines the advantages of DilatedFCN and EncoderDecoder by ASPP and one path of skip connection.

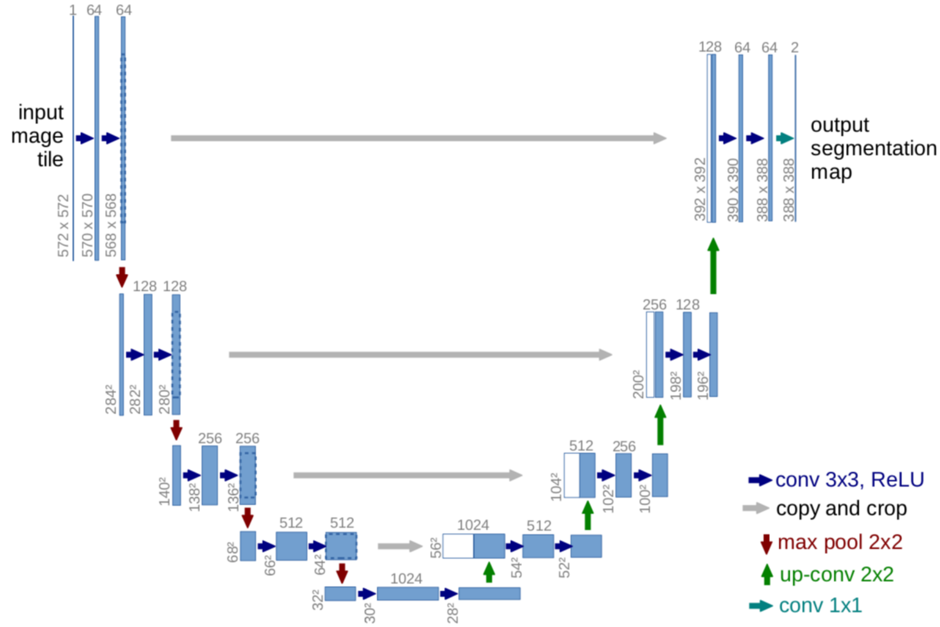


Figure 3.5: The overall structure of UNet. Image adapted from [54].

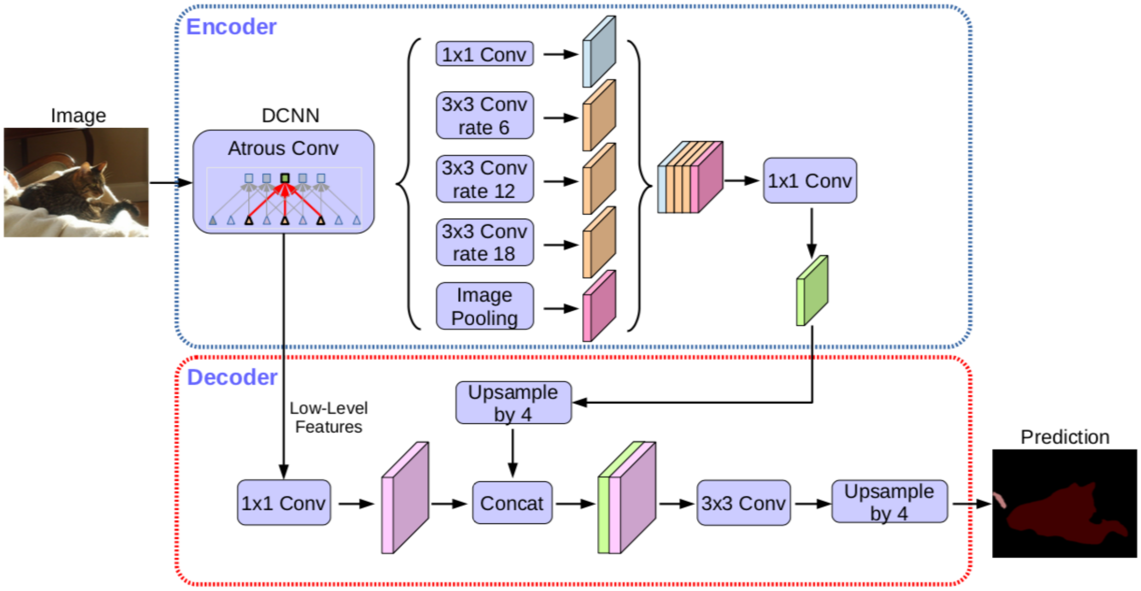


Figure 3.6: The overall structure of DeeplabV3 Plus. Image adapted from [55].

3.3 Deep Learning in SAR Image Interpretation

In this section, we choose three typical SAR image interpretation tasks to demonstrate their notable developments, which correspond to three different applications thereafter, in Chapter 4, Chapter 5, Chapter 6, respectively.

3.3.1 SAR Object Detection

Most of the earlier work on SAR object detection applied deep learning detection methods in the computer vision field with minor tweaks. The first attempt can be found in SAR military vehicle detection on the Moving and Stationary Target Acquisition and Recognition (MSTAR) dataset [56]. MSTAR is one of the earliest datasets for SAR target recognition collected by Sandia National Laboratory (SNL). This dataset is publicly available and contains 10 classes of vehicles, plus one class of simple geometric targets. In [57], Chen et al. used a single layer of convolutional neural network and trained the convolution kernel on random samples using an unsupervised sparse auto-encoder. This application shows the potential of deep learning in SAR images. After that, more deep learning-based methods were developed on MSTAR. Chen et al. [58] proposed a simple 5-layer CNN, A-ConvNets, to automatically learn vehicle features from SAR images. Wagner et al. [59] designed a network on MSTAR, combining a CNN and a Support Vector Machine (SVM) to incorporate a prior knowledge. Feng et al. [60] proposed a self-matching Class Activation Mapping (CAM) to visualize what a CNN learns from SAR images to make a decision. However, due to the limited samples and rather ideal scenarios of MSTAR, most of these methods face the overfitting problem that the test accuracies are above 99%.

Another SAR object detection application is ship detection. There are several open SAR ship detection datasets developed in the past decades, such as OpenSARShip [61], SAR-Ship-Dataset [62], and SSDD [63]. These ship detection datasets have a relatively large size of SAR images and complex ocean backgrounds, which is of benefit for the development of automatic SAR ship detection. Kang et al. [64] proposed an algorithm combining CFAR with faster R-CNN. Zhang et al. [65] applied transfer learning to the SAR ship detection field. In more recent works, Guo et al. [66] used a feature pyramids fusion module and a head enhancement module to improve ship detection performance. In [67], a two-stage detection network SCLANet is proposed for SAR ship detection based on consistency learning and adversarial learning.

For the vehicle detection and ship detection tasks, the targets manifest as prominent bright areas on SAR images, which shows the potential to make a large-scale labeled dataset. However, some of the targets on SAR images like oil spills and oceanic eddies are very difficult to be found and labeled. The lack of high-quality standardized datasets heavily constrains the development of these applications. Moreover, most of the existing

SAR detection methods have not considered the specific characteristics of SAR imagery, which shows the progressive space for improvements.

3.3.2 SAR Semantic Segmentation

The semantic segmentation of SAR images, namely the pixel-wise classification of SAR images according to ground surface types, is one of the most important SAR image interpretation applications.

Xie et al. [68] first used Stacked Sparse Autoencoder (SAE) as a useful strategy to classify different surface types. Their method was verified on a real Polarimetric Synthetic Aperture Radar (PolSAR) image, which covered an agricultural area in Flevoland, the Netherlands. The results showed the feasibility to represent features for surface type classification. Geng et al. [69] then designed a deep supervised and contractive neural network (DSCNN) for SAR image classification, aiming to solve the problems of speckle noise. Three SAR images acquired from TerraSAR-X (X-band), RS2 (C-band), and ALOS2 (L-band) are applied in the experiments for urban area classification. More recently, different variations of CNNs have started to be applied in SAR semantic segmentation. Wu et al. [70] used an FCN based model with transfer learning to realize PolSAR scene segmentation with small training sets. Wang et al. [71] designed a deep neural network for scene segmentation from high-resolution SAR Data. He et al. [72] embedded low-dimensional representation learned by nonlinear manifold method into Fully Convolutional Networks (FCN) model to learn deep spatial features of PolSAR imagery, and then applied SVM for classification, which proved the effectiveness on Flevoland, Foulum, and San Francisco datasets for the land cover classification task.

In general, the deep learning-based SAR semantic segmentation methods have advanced considerably in the past decade. At first, they focused on the applications of SAE and later they concentrated on CNN. Like SAR object detection tasks, more specific SAR features and their complex nature should be considered when we design the SAR semantic segmentation models.

There is an OpenSARUrban [73] dataset consisting of Sentinel-1 GRD images, which can be considered the large-scale benchmark for urban interpretation. But for other SAR semantic segmentation tasks, large SAR datasets, especially PolSAR datasets, are still urgently needed.

3.3.3 SAR-optical Data Fusion

SAR-optical data fusion is becoming one of the most promising directions of deep learning in remote sensing. There are lots of applications that combined SAR and Optical data, for example, joint analysis of SAR and optical images [74], matching SAR and optical images [75], automatic SAR colorization [76], cloud removal from optical images [77], and SAR and optical fusion semantic segmentation [78], etc.

For the SAR and optical fusion semantic segmentation task, we mainly face two big challenges at present. The first challenge is the lack of organized optical and SAR image segmentation datasets. SEN12MS is a large-scale multi-modal land cover classification dataset [19]. However, it is highly influenced by noisy data, resulting in difficulties to compare different multi-modal fusion networks. Another challenge is the lack of effective fusion approaches for SAR and optical images. Most of the SAR and optical fusion methods are still in the early stages [79, 80, 81]. Therefore, effective strategies for SAR and optical data fusion still have a lot of room to improve.

Chapter 4

Mask-ES-RCNN: Mask Edge Enhancement and IoU Score RCNN for Oceanic Eddy Detection

4.1 Introduction

Oceanic eddies are self-sustaining rotary currents that are distributed worldwide in the ocean [82]. Their horizontal spatial scales vary from several hundred meters to several hundred kilometers [83]. They can cover long distances before dissipating and play a significant role in the mixing and transport of heat, salt, and biogeochemical tracers across the global oceans [84, 85, 86]. Moreover, eddies may appear on shipping routes and in offshore regions and consequently affect human marine activity [87]. Therefore, oceanic eddy detection is of great research value. Under the influence of ocean currents, sea surface winds, and bottom topography, oceanic eddies tend to be highly variable [88]. This changeable quality of oceanic eddies makes their detection a more challenging task.

In the primary stage of their investigation, oceanic eddy data were collected by in-situ measurements [89]. With the development of satellite sensors, more and more studies have been conducted based on remote sensing data like SAR, or satellite-derived parameters such as Sea Surface Height (SSH), Sea Surface Temperature (SST), and

Ocean Color/ Chlorophyll (CHL) [90]. SSH products use large spatio-temporal interpolation between the areas crossed by satellite tracks, resulting in low-resolution fields and uncertainty in inadequately sampled areas [91]. Since many other ocean phenomena also impact the sea surface temperature and surface ocean color, SST and CHL are prone to propose false positives [92]. Therefore, SAR sensors may be most effective for the observation of oceanic eddies, due to the high spatial resolution and the sensitivity of radar signals to natural surfactants on the water surface [93].

Conventional methods of eddy detection on SAR images [86, 94] are based on visual inspection and expert knowledge. These methods heavily rely on labor, showing significant limitations in time and cost, as well as their generalization ability. Several studies have employed deep learning methods for automatic eddy detection on SAR images in recent years. Huang et al. [92, 95] proposed a deep network named DeepEddy to learn the features of ocean eddies based on the Principal Component Analysis (PCA) filter convolution neural networks. But their method only focused on the eddy classification task. It still needs a lot of time and effort to select potential candidates manually. Zhou et al. [96] used a detection network called MFNN based on ResNet-50 [97] and ASPP to detect five types of oceanic phenomena, including eddies. However, their method does not identify each instance of an eddy and lacks in an effective utilization of the eddies' specific characteristics. Furthermore, a big challenge in automatic oceanic eddy detection is that the amount of adequately labeled data is insufficient. There are no open datasets dedicated to SAR eddy detection, due to the difficulties of SAR data procurement and interpretation.

To address the above dilemmas, we first build a dataset, namely SOEDD. Then, an eddy detector based on Mask RCNN [48] and edge enhancement is applied to get initial results. Inspired by the initial model, an end-to-end detector called Mask-ES-RCNN is proposed, aiming at detecting all eddies and their corresponding locations precisely.

Existing deep learning techniques for SAR eddy detection neglect the importance of learning the internal edge information of the eddies. Instead, we enhance edge information to help our model learn features more efficiently. At the same time, we focus on the instance segmentation mask qualities to help in improving the performances. The main contributions of this chapter can be summarized as follows:

1. A SOEDD is constructed to promote the research in oceanic eddy detection on SAR images using deep learning methods. The experimental results of different deep learning methods on SOEDD prove its ability and potential to achieve acceptable eddy detection results under the condition of limited training samples;

2. A Mask RCNN and Edge Enhancement model is first proposed to detect eddies on SOEDD. It applies Canny edge detection on the input data to enhance texture information. The final performances turn out to be better than the RCNN baseline;

3. A Mask-ES-RCNN model is further designed, based on the Mask RCNN framework with two new branches. A new Edge Head is used for implicit learning internal texture information of eddy instances. A new Mask IoU Head focuses on promoting the eddy mask quality. The combination of Edge Head and Mask IoU Head works well on SOEDD using a multi-task strategy.

The remainder of this chapter is organized as follows. In Section 4.2, we introduce the data collection and the SOEDD construction process. Section 4.3 describes the Mask-RCNN and Edge Enhancement model. The Mask-ES-RCNN architecture and experimental results are shown in Section 4.4. Finally, we summarize the chapter with discussion and insights in Section 4.5.

4.2 Dataset

4.2.1 Data Collection

A SOEDD is collected from Sentinel-1A SAR data (C band) of the Western Mediterranean Sea acquired from October 2014 to January 2015, always around 06:00 UTC and 18:00 UTC. Figure 4.1 shows the location of the study area. The original SAR data were provided by ESA's Sentinels Scientific Data Hub [98] and were level 1 products. More detailed information concerning the processing level can be found in [99].

In our work, all SAR images were downloaded as GRD products, acquired in Interferometric Wide (IW) or Extra Wide (EW) swath mode. On the sea surface, the backscatter of the cross-polarized channels (HV and VH) is usually much lower than the co-polarized channels (VV and HH) [100], sometimes even close to the noise floor of the SAR system [101]. Therefore, we only used co-polarized SAR data to construct the dataset.

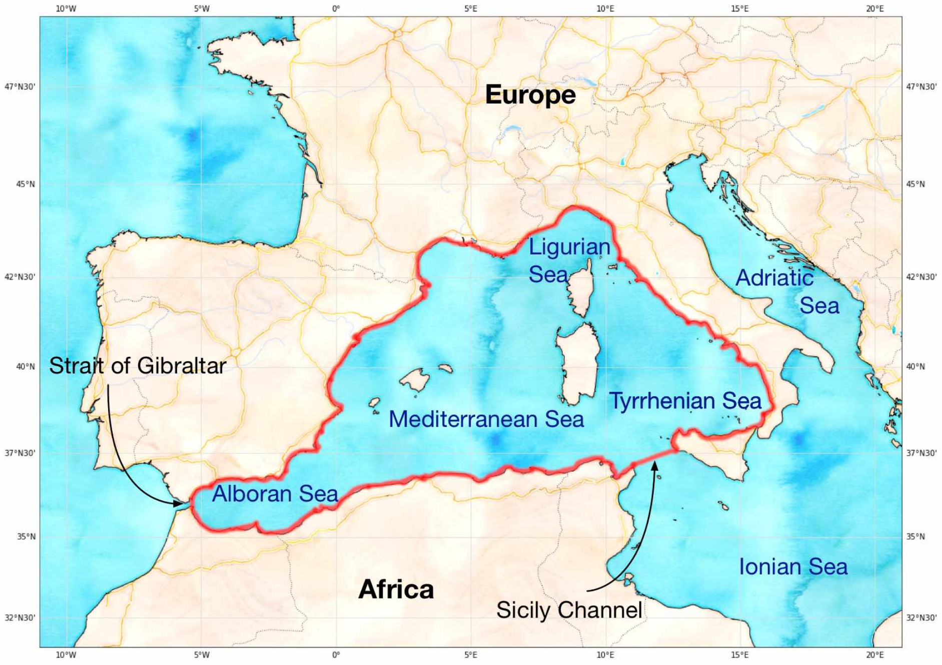


Figure 4.1: The Western Mediterranean Sea. The red line outlines the region of interest.

The eddies' manual annotations were provided by Annika Buck. We further make a correction of them by visual detection. Related eddy visual interpretation methods are detailed described in her master thesis [102]. Following [103], eddies manifest on SAR images due to two mechanisms: wave damping due to surface films [104] and surface roughening due to wave-current interaction [105]. The eddies that become visible due to surface films are called "black" eddies, and show up as dark areas or lines. The eddies that become visible due to wave-current interaction are called "white" eddies and show as bright curved lines [103]. Notably, only "black" eddy instances are included in the SOEDD.

Eddies with a diameter less than the first baroclinic Rossby radius of deformation are considered as submesoscale eddies, and more than this radius are recognized as mesoscale eddies [106]. In the SOEDD, most of them are submesoscale eddies (here we choose the radius to be 15km [107]).

The final manual annotations consist of the following key eddy parameters: positional information (the center coordinate), geometric information (one auxiliary coordinate at the outer edge, the maximum and minimum diameter), attribute information (the direction of rotation, the type "black" or "white") and the SAR imaging information (date and time).

4.2.2 Dataset Construction

Based on the collected data, we design our specific construction procedure of SOEDD. For the downloaded SAR images, all of them are preprocessed using the Sentinel Application Platform (SNAP) Toolbox developed by ESA [108]. After applying orbit file and radiometric calibration, geocoding and land masking are conducted to the SAR images. According to the eddy coordinate information obtained from manual annotation data, SAR image subsets with different numbers of eddy instances included are exported from SNAP. After that, contrast-limited adaptive histogram equalization [109] is applied to all the subsets. To keep as much information in the images as possible, we refrain from applying any speckle filters to the SAR images. For the manual annotations, we convert them automatically to COCO format [110] by self-designed python scripts. This format is adopted to store bounding box and pixel-wise classification information.

Our SOEDD is constructed for the instance segmentation task [111]. It consists of 160 training images and 40 testing images containing 260 and 62 eddy instances, respectively. In total, eddies in SOEDD with diameters ranging from 1.3 km to 15.87 km were included. The size of SAR images ranges from about 600×600 to 1200×1200 pixels. The size distribution of all eddies is shown in Figure 4.2. Most of the eddies in SOEDD appear in a near-circular shape with 100×100 to 600×600 size in pixels according to their distributions statistics.

In Figure 4.3, we display three pairs of eddy samples with their original SAR images and COCO format annotations. We assign different colors for each eddy instance. The eddy samples vary from each other in terms of shape structure, scale, and direction.

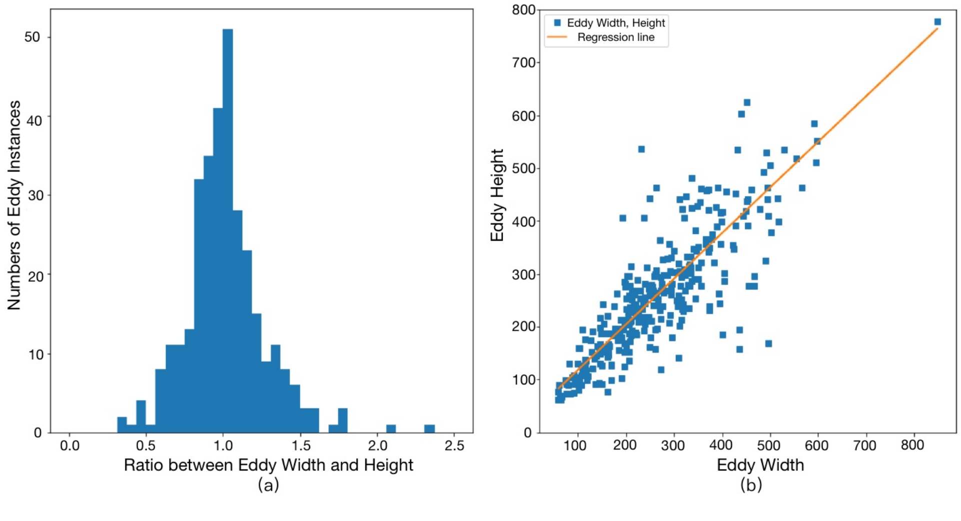
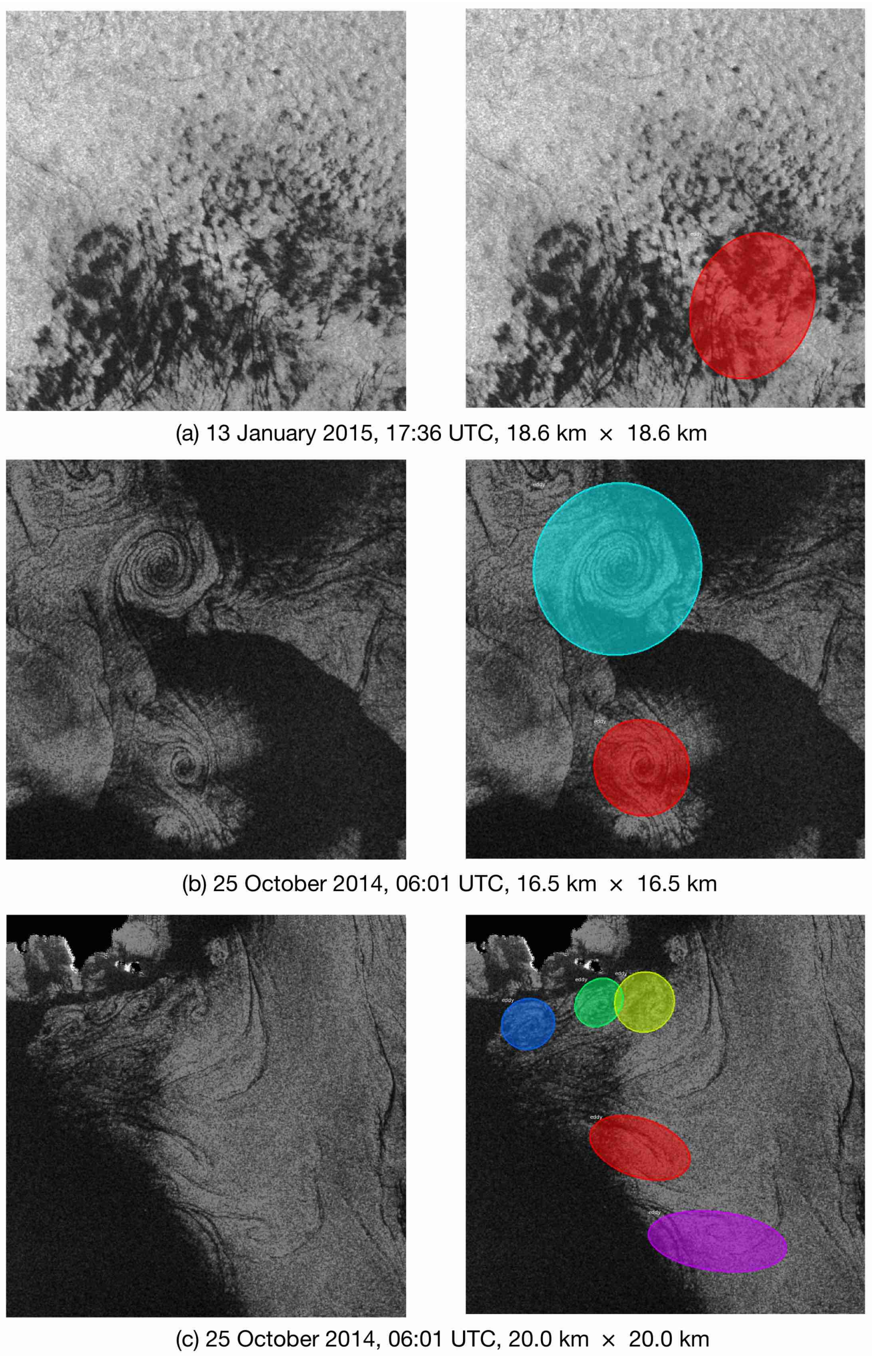


Figure 4.2: Data statistics of eddy samples in SOEDD: (a) Distribution of the ratio between the bounding box width and height; (a) distribution of the bounding box width and height.



(a) 13 January 2015, 17:36 UTC, 18.6 km × 18.6 km

(b) 25 October 2014, 06:01 UTC, 16.5 km × 16.5 km

(c) 25 October 2014, 06:01 UTC, 20.0 km × 20.0 km

Figure 4.3: Three pairs of eddy samples in SOEDD: original SAR images (left) and their annotations (right). SAR images ©ESA 2014 2015.

4.3 Mask RCNN and Edge Enhancement

4.3.1 System Architecture

The inspiration of the Mask RCNN and Edge Enhancement model is taken from the visual detection of the eddies. The "black" eddies become visible on SAR images as dark areas or lines. An expert often focuses primarily on the dark areas or lines and then gives a definitive decision according to their linear structure and morphological characters. These characters can be considered as part of texture information.

To make the networks perform in a similar way, we first extract edges in the original images, thereby enhancing the importance of texture features and filtering out irrelevant information simultaneously. This method can be regarded as a simple way to integrate prior knowledge into deep learning.

The whole training process of Mask RCNN and Edge Enhancement is divided into two steps: First, edge features are extracted from the original images, and second, both the detection results and the original images are transferred into deep learning networks for training.

4.3.1.1 Edge Detection

We choose a classic edge detection algorithm named Canny edge detection [112]. This is a multi-stage algorithm. First, a Gaussian filter is applied to the SAR images. This filter is a typical linear filtering technique, which is effective in speckle noise reduction [113]. In Figure 4.4, the middle column shows the different effects of filters of three filter sizes. A filter of 11×11 pixel size is picked in order to achieve the best results.

The filtered images are then used by the Sobel operator to derive gradients. The Sobel operator uses one filter each for horizontal and vertical directions, which is described as a first-order gradient operation.

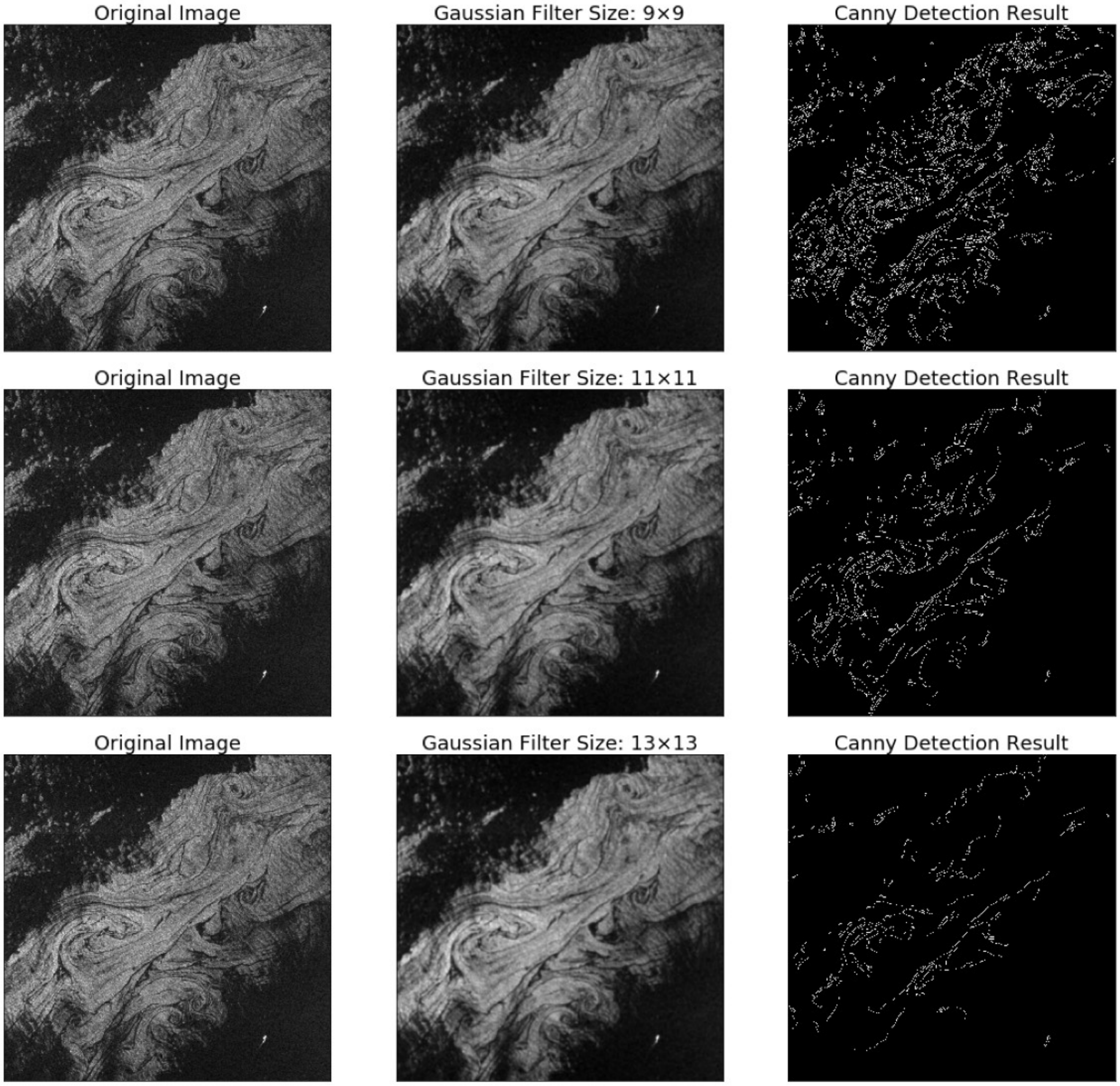


Figure 4.4: Eddy detection results with different filter sizes. SAR image acquired at 07 June, 2015, 05:36 UTC, 9.3 km × 9.3km ©ESA 2015.

The filters are represented as:

$$S_x = \begin{bmatrix} 1 & 0 & -1 \\ 2 & 0 & -2 \\ 1 & 0 & -1 \end{bmatrix}, \quad S_y = \begin{bmatrix} -1 & -2 & -1 \\ 0 & 0 & 0 \\ 1 & 2 & 1 \end{bmatrix} \tag{4.1}$$

According to gradient magnitude and direction, finally, non-maximum suppression and hysteresis thresholding are conducted. The results of eddy edge detection are illustrated in the right column of Figure 4.4.

4.3.1.2 Network Architecture

The proposed network structure is shown in Figure 4.5. The framework consists of five components: a Canny operator to extract edges, a Feature Pyramid Network (FPN) [114] as the backbone, an RPN [47] to generate proposals, a Fast RCNN for bounding box classification and regression, and a mask branch for eddy instance segmentation.

The original input data is a SAR amplitude image (GRD format) from a co-polarization channel. Doubling the original input and adding the edge detection results, we generate a three-channel input. Based on this input, the RPN generates a large number of first eddy proposals. The Region of Interest (ROI) features of the eddy proposals are then fed into the Fast RCNN and the Mask Branch to get more accurate bounding boxes and eddy segmentation maps.

Since oceanic eddies vary in size, we apply an FPN backbone with ResNet of depth 50. FPN uses the inherent multi-scale structure of ResNet networks to construct a feature pyramid that has rich semantics at all levels and facilitates the detection of eddies at different scales.

4.3.2 Experimental Setup

Experiments are conducted using an implementation of the reproduced Mask RCNN based on the Keras framework with a TensorFlow backend [115]. For the RPN part, we set five scale $\{32^2, 64^2, 128^2, 256^2, 512^2\}$

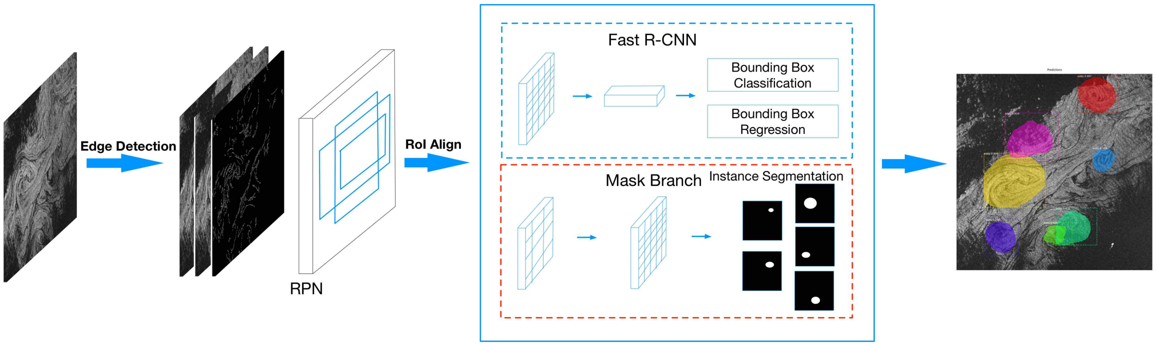


Figure 4.5: Illustration of the architecture of Mask RCNN and Edge Enhancement.

anchors at five stages $\{P_2, P_3, P_4, P_5, P_6\}$. According to the ratio statistics of the SOEDD, aspect ratios $\{0.5, 1, 2\}$ are adopted in the workflow.

All training work is carried out on an NVIDIA Pascal Titan X GPU. The model is trained until convergence by using the SGD with a momentum set as 0.9 and a weight decay set as 0.0001. All weights are initialized by a Xavier initialization. The remaining configuration for ResNet-50 was done following [48]. Under this setup, the training takes up to 2 hours. For the testing phase, we use SoftNMS [116] and retain the top-100 score detections for each image.

We adopt COCO metrics [110] for our experiments, for which we conduct training and evaluation three times. The COCO-style Average Precision (AP) score is calculated by taking the Mean AP (mAP) over 10 IoU thresholds, from 0.5 to 0.95, step 0.05 (0.5, 0.55, 0.6, 0.65, 0.7, 0.75, 0.8, 0.85, 0.9, 0.95). The APs at fixed IoU=0.5 and IoU=0.75 are also used for reporting results respectively.

4.3.3 Results

A group of comparative experiments is designed on the SOEDD to verify the effectiveness of the proposed approach, mainly focusing on evaluating the effects of edge detection inputs. Specifically, we compare the results obtained using a modified Mask RCNN framework with three different inputs: only the original SAR image (Original), only the edge detection images (Edge), and the original images and edge detection images together (Original + Edge). The corresponding results shown in Table 4.1 are averaged values. These values indicate that the Original+Edge input provided the best AP, about 2.3% higher and 14.4% higher than Original and Edge, respectively.

Input	AP	$AP_{0.5}$	$AP_{0.75}$
Original	18.7	35.6	20.1
Egde	7.6	8.8	7.9
Original+Edge	**21.0**	**38.5**	**22.2**

Table 4.1: Eddy detection results on different inputs.

4.3.4 Discussion

In our Mask RCNN and Edge Enhancement model, we apply Canny edge detection on the input data, which is a straightforward method to enhance texture information. The final performances turn out to be better than the Mask RCNN baseline in terms of all APs. This is probably because the input channel of the edge detection map forces the model to filter out useless information and to learn more texture-related features, which are used as a prior knowledge for model learning on limited SAR data.

This Mask RCNN and Edge Enhancement model is an attempt to detect eddies in the SOEDD. According to the preliminary experimental results, we can summarize the inspirations as follows:

1. Although the SOEDD has very limited training samples, we can still receive acceptable eddy detection results using deep learning-based methods on this dataset;

2. In theory, deep learning models will automatically extract all effective features without manual intervention. However, it is

difficult for deep learning models to extract essential features in small datasets, so we could provide prior knowledge to help the models learn better;

3. In the SOEDD case, combining Mask RCNN with texture features acquired by edge detection, we finally achieve better performances. This provides an idea to develop an end-to-end deep learning model on SOEDD with the help of texture features.

4.4 Mask-ES-RCNN

The Mask RCNN and Edge Enhancement model has proven effective on SOEDD. But our purpose is to realize an end-to-end deep learning model with prior knowledge. Edge detection is an intuitive way to enhance texture information. However, it can only provide a limited guidance function. The speckle noise and complex ocean background will cause the poor quality detected edges. Other phenomena like oil spills will also cause dark lines on SAR images. We need to design more flexible and effective strategies to incorporate prior knowledge.

An expert decides on the annotations of eddies based on the linear structure and morphological characters of the dark areas or lines on SAR images. However, it is extremely hard to annotate all of these dark pixels as prior knowledge to help the model learn. We observe that the internal dark areas or lines of eddies are highly related to the eddy boundaries. If we enhance the importance of the boundary pixels, we can implicitly learn the internal texture of eddies. Besides, it has been proven that focusing on mask qualities improves the performance of instance segmentation tasks [117].

Based on the multi-task learning concept [118], we propose a Mask-ES-RCNN model to learn boundary information and mask qualities simultaneously. It is inspired by the design of focusing on instance boundary in [119] and mask scoring in [117]. In the following section, we will introduce Mask-ES-RCNN in detail.

4.4.1 System Architecture

4.4.1.1 The Overall Architecture

The detailed architecture of Mask-ES-RCNN is shown in Figure 4.6, and is composed of five parts: an FPN Backbone Network, an RCNN Head, a Mask Head, an Edge Head, and a Mask IoU Head. We follow a common segmentation formula in which an object detection module is utilized before performing instance-wise segmentation on ROIs.

First, the input SAR images go forward through the FPN backbone network to extract multi-level bottom-up augmented feature maps. ROI Align is adopted for extracting the ROIs within the region proposals from RPN and multi-level feature maps. Second, we perform proposal classification, bounding box regression (using RCNN head), and mask predicting (using mask head). After that, the predicted masks are sent to both Mask IoU Head and Edge Head for predicting Mask IoU and detecting boundary edges, respectively. Finally, the predicted Mask IoU will be used in the testing phase to rescore the predicted masks.

Like Mask RCNN, Mask-ES-RCNN also adopts a multi-task learning strategy. We include two additional boundary learning and mask IoU regression tasks to help in better learning features. Following the function design in Mask R-CNN, we add the losses for two new tasks. The total loss is expressed as:

$$L_{Mask-ES-RCNN} = L_{RPN} + L_{RCNN} + L_{Mask} + \alpha L_{Edge} + \beta L_{Mask-IoU} \quad (4.2)$$

where $L_{RPN}, L_{RCNN}, L_{Mask}$ are standard losses in Mask RCNN for RPN module, RCNN Head and Mask Head, respectively. L_{Edge} is the loss for Edge Head, and $L_{Mask-IoU}$ is the loss for Mask IoU Head. In the end, we minimize the loss function consisting of these five parts to attain a good performance in oceanic eddy detection.

4.4.1.2 The Edge Head

For the edge Head, we use edge detection filters (such as Sobel and Laplacian) as identity kernel convolutions (kernel size=3).

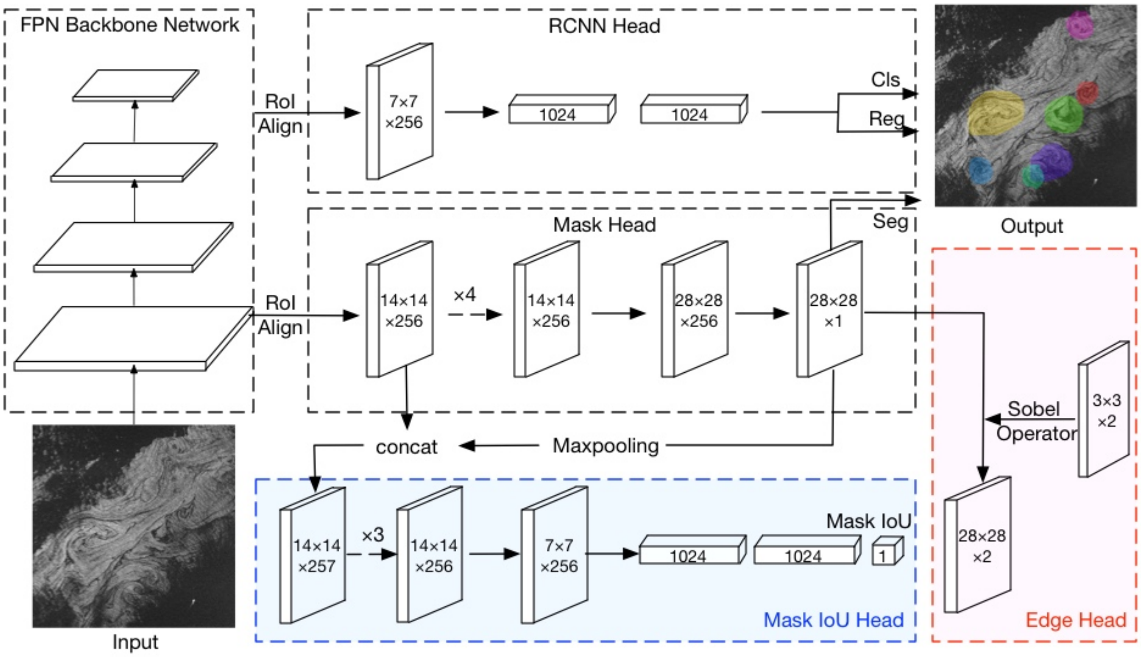


Figure 4.6: The network architecture of Mask-ES-RCNN. It consists of an FPN Backbone, an RCNN Head, a Mask Head, a Mask IoU Head, and an Edge Head.

The Edge Head loss function L_{Edge} based on the Sobel operator is then expressed as:

$$L_{Edge} = \frac{1}{n}\sum_{i=1}^{n}\left(\|\hat{m_i} * S_x - m_i * S_x\|_F^2 + \|\hat{m_i} * S_y - m_i * S_y\|_F^2\right) \qquad (4.3)$$

where n represents the number of training samples in Mask Head (with a threshold of IoU=0.5 between proposal box and the matched ground truth), $\hat{m_i}$ is the predicted mask, and m_i is the matched ground-truth, $\|\cdot\|_F$ stands for the Frobenius norm.

The Laplacian operation is a second-order gradient operation, which detects edges in an image with zero crossings. The Laplacian $L(x,y)$ of an image with pixel values $I(x,y)$ is expressed as:

$$L(x,y) = \frac{\partial^2 I(x,y)}{\partial x^2} + \frac{\partial^2 I(x,y)}{\partial y^2} \qquad (4.4)$$

The discrete Laplacian can be given as convolution with the following kernel:

$$L = \begin{bmatrix} 0 & 1 & 0 \\ 1 & -4 & 1 \\ 0 & 1 & 0 \end{bmatrix} \tag{4.5}$$

In our experiments, we use another version of Laplacian operator with diagonal additional elements in the kernel:

$$L = \begin{bmatrix} 1 & 1 & 1 \\ 1 & -8 & 1 \\ 1 & 1 & 1 \end{bmatrix} \tag{4.6}$$

The Edge Head loss function L_{Edge} based on the Laplacian operator is then expressed as:

$$L_{Edge} = \frac{1}{n} \sum_{i=1}^{n} \left(\| \hat{m_i} * L - m_i * L \|_F^2 \right) \tag{4.7}$$

In image processing, images are usually smoothed before using detection filters. We tried Gaussian smoothing before the filters, but the results turned out to be of no help, so we dropped this design.

4.4.1.3 The Mask IoU Head

For the Mask IoU Head, we use both ROI feature maps and predicted mask as input. After 4 convolutions (kernel=3) and 3 fully connected layers, we finally get the Mask IoU values. For each instance, we use the Mask IoU between the binary mask and the matched ground truth as the Mask IoU target. The L_2 loss is adopted for regressing Mask IoU, which is defined as:

$$L_{Mask-IoU} = \frac{1}{n} \sum_{i=1}^{n} \left| mi\hat{o}u_i - miou_i \right|^2 \tag{4.8}$$

where n represents the number of training samples in the Mask Head, $mi\hat{o}u_i$ is the predicted mask IoU by Mask IoU Head and $miou_i$ is the Mask IoU target.

Following the scoring system [117], we also decompose the mask scoring tasks into mask classification and mask IoU regression, defined as:

$$score_{mask} = score_{cls} \cdot score_{iou} \tag{4.9}$$

where $score_{cls}$ is the classification score in RCNN Head and $score_{iou}$ is the Mask Iou value in Mask IoU Head. We use $score_{mask}$ as the final confidence score to rank top-k target masks in the testing phase.

4.4.2 Experimental Setup

We set hyper-parameters mainly following on the Mask R-CNN baseline. The base model is ResNet 50 and pre-trained on ImageNet, with the standard FPN [114]. We randomly initialize all new layers by drawing weights from a zero-mean Gaussian distribution with a standard deviation of 0.001. Synchronized SGD is adopted as an optimizer with momentum 0.9 and weight decay 0.0001. The training rate is 0.001 for all experiments. It takes around 2 hours to complete training the network with an NVIDIA Pascal Titan X GPU.

We resize the input images to have 512 pixels along the short axis and 1024 along the long axis for both training and testing. We use SoftNMS [116] and retain the top-100 score detections for each image.

4.4.3 Quantitative Results

A comparison of different detectors on SOEDD is shown in Table 4.2. The Mask R-CNN framework serves as a state-of-the-art baseline. The Mask-ES-RCNN achieves the best results in all AP, $AP_{0.5}$, $AP_{0.75}$ metrics. Especially for $AP_{0.5}$, when the evaluation method is not so strict, it can exceed the baseline by 12.9% and Mask RCNN and Edge Enhancement by 10%, respectively.

We conduct comparison experiments on Mask IoU Head, Edge Head, and a combination of the two Heads, to verify the effectiveness of our proposed Mask-ES-RCNN. Table 4.3 shows the ablation study results.

Method	Backbone	AP	$AP_{0.5}$	$AP_{0.75}$
Mask RCNN	ResNet-50-FPN	18.7	35.6	20.1
Mask RCNN and Edge Enhancement	ResNet-50-FPN	21.0	38.5	22.2
Mask-ES-RCNN	ResNet-50-FPN	**24.8**	**48.5**	**27.1**

Table 4.2: Eddy detection results using different methods.

Both the Mask IoU Head (including Mask re-score mechanism in the test phase) and Edge Head improve our model compared with the Mask R-CNN baseline. Further, if we combine the Mask IoU Head and Edge Head, we can obtain better results than using any one of them alone. This proves that multi-task learning help in learning useful representations from the same input images, allowing the gradient from two tasks to influence shared feature maps. Experiments with different weights of Edge Head and Mask IoU Head are also conducted. We find that the model achieves the best experiment results if we set these two Heads with equal weight.

Backbone	Mask IoU Head	Laplace Head	Sobel Head	AP	$AP_{0.5}$	$AP_{0.75}$
ResNet-50-FPN				18.7	35.6	20.1
	$\surd$			23.3	45.0	25.1
		$\surd$		22.9	44.9	24.8
			$\surd$	23.6	46.7	25.9
	$\surd$		$\surd$	**24.8**	**48.5**	**27.1**

Table 4.3: Eddy detection results on different design choices of the Mask IoU Head and Edge Head.

For the Edge Head, the influence of edge detection filters is analyzed. Sobel filter outperforms the Laplacian filter in our model with a 0.7% relative improvement in terms of AP. This situation might be explained by the two filters structure of the Sobel filter, which means the eddy orientation information can also be used during the back-propagation process.

4.4.4 Visualization Results

In addition to the accuracy evaluations, the visualization of detected eddy samples also presents an overview of the effectiveness of Mask-ES-RCNN. We set the predicted eddy mask with confidence scores above 0.9 and an NMS with a threshold of 0.1 to remove duplication.

As shown in Figure 4.7, Mask-ES-RCNN shows the strong capacity to detect SAR oceanic eddies varying in scale, rotational direction, and morphological character. The model can successfully detect eddies under the conditions of complex ocean background and indistinct texture information, which are even very hard for experts to find.

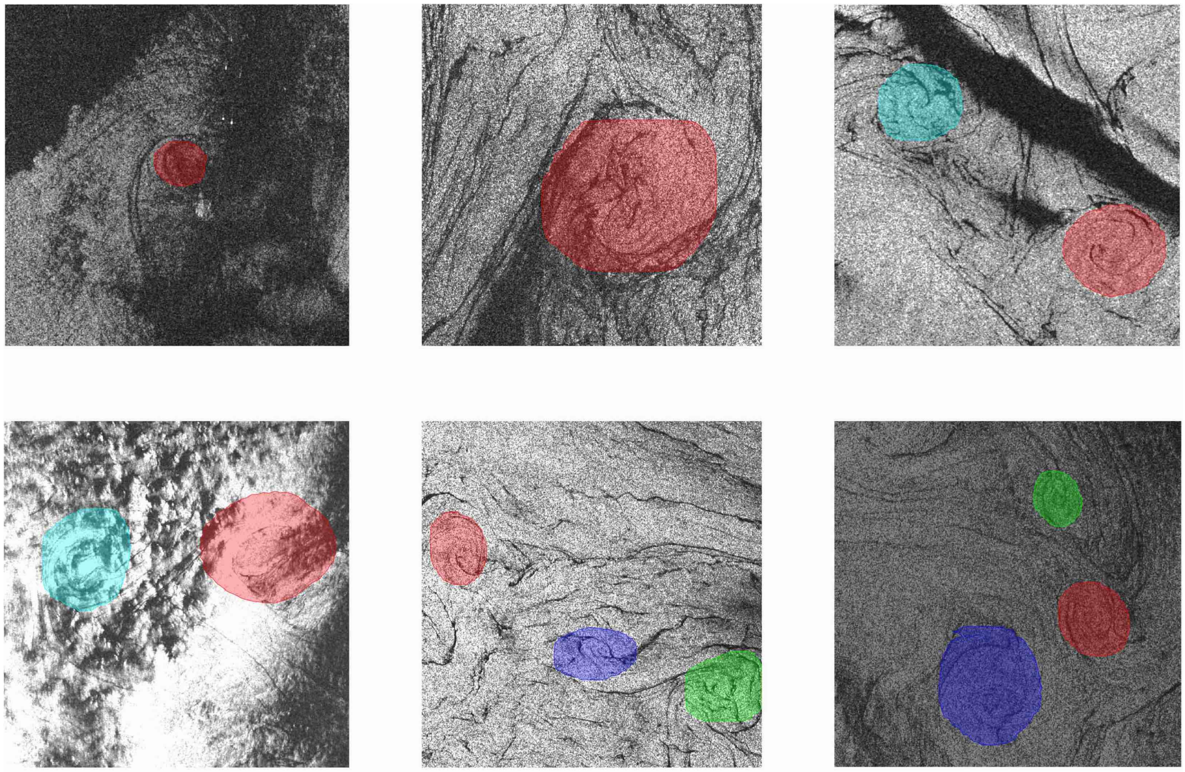


Figure 4.7: Acceptable visualization results of Mask-ES-RCNN.

Figure 4.8 displays some unacceptable prediction results of Mask-ES-RCNN to demonstrate further optimizing of the our model is still needed.

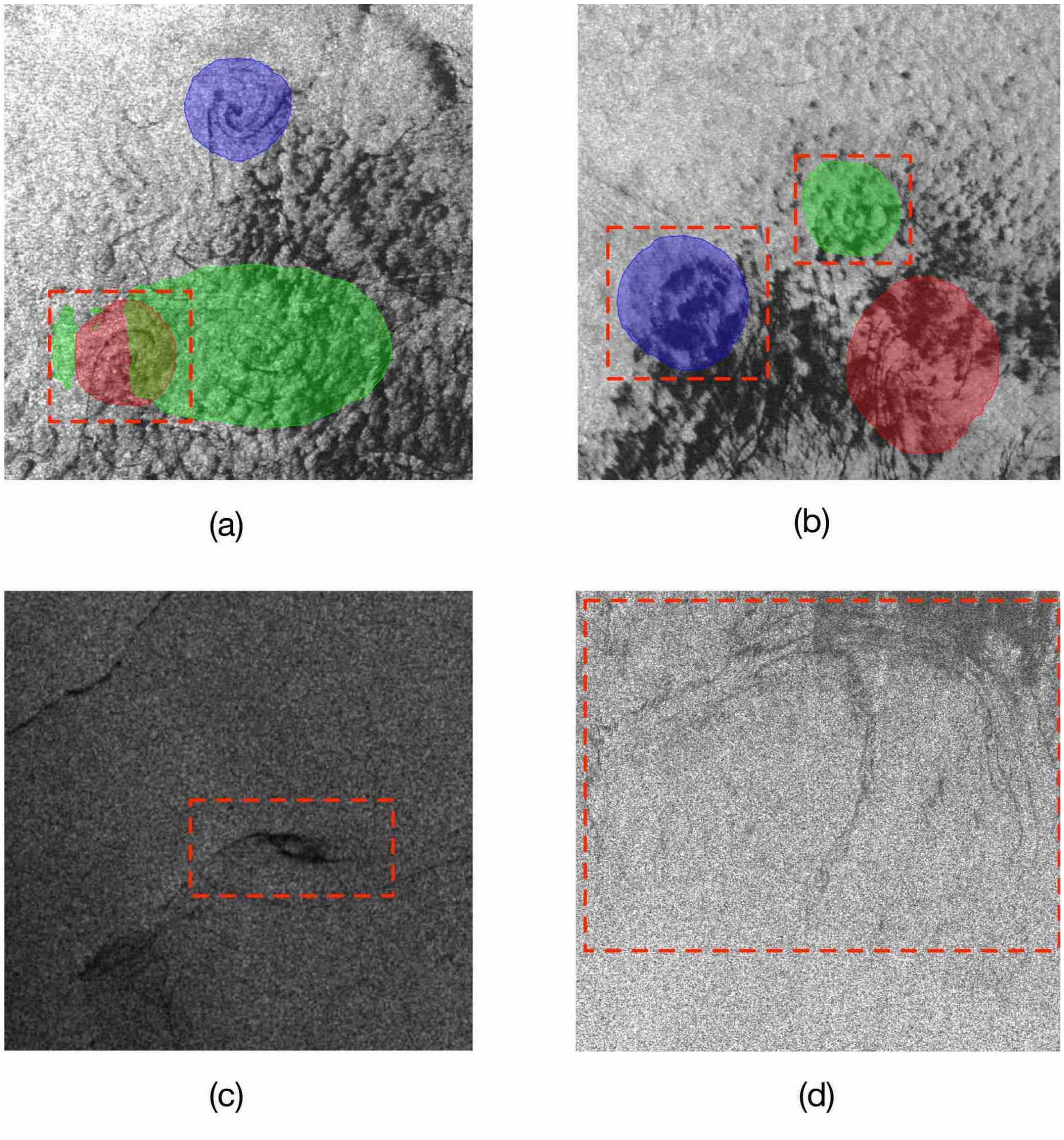


Figure 4.8: Unacceptable visualization results of Mask-ES-RCNN: (a) densely packed eddies; (b) common false alarms; (c) undetected eddy due to unusual morphological character; (d) undetected eddy due to large aspect.

For densely packed eddy instances, as shown in Figure 4.8 (a), the adjacent mask predictions will affect each other, which results in poor mask qualities. Dynamic refined network [120] and rotated bounding box

[121] methods can be adopted to tackle this problem.

Figure 4.8 (b) gives typical examples of false alarms. If we observe the inner texture features of blue and green instances, we can find the false alarms have connected black spots (which are possibly caused by wind or other natural phenomena) to confuse the deep learning model. It is therefore recommended to use additional information like wind speed together with SAR images to reduce the risk of false alarms.

In contrast, our model was unable to find eddies in some cases. In Figure 4.8 (c), an ordinary aspect ratio eddy cannot be detected, because of the unusual morphological character. As Figure 4.8 (d) illustrates, the large eddy manifests in open surface structures which are hard for current detectors to identify. For the missing detection problem, we may increase the diversity of SOEDD or use a more realistic setting like few-shot learning [122].

4.5 Summary

In this chapter, we construct the SOEDD for oceanic eddy detection on SAR images. Two deep learning methods are proposed based on SOEDD. The experimental results of the Mask RCNN and Edge Enhancement model prove the potential to predict acceptable eddy detection results from the SOEDD, as well as the effectiveness of incorporating prior knowledge on a small SAR dataset. The Mask-ES-RCNN model outperforms the Mask RCNN and Edge Enhancement model on SOEDD in terms of all APs. The combination of the Edge Head and Mask IoU Head works well on the SAR eddy detection task. The Edge Head realizes implicit learning of internal texture information and the Mask IoU Head improves the quality of the predicted mask. They can be further generalized to other target detection tasks on SAR images.

The performances will be further improved by enriching the dataset both in size and scope, and by developing more suitable detection algorithms on SAR images. We only use the GRD format of SAR images in SOEDD, but more information can be used to help the model learn. For our eddy detection case, the bounding boxes are inherently ambiguous. Therefore, in the future, we can propose a new evaluation system instead of adopting common assessment methods.

Chapter 5

TE-UNet: Texture Enhancement UNet for Intertidal Sediments and Habitats Classification

5.1 Introduction

The intertidal zone is the coastal area, where the ocean meets the land within the tidal range. Intertidal environments have a complex mosaic of surfaces, including barrier islands, sandy and mixed sediments, seagrass meadows, etc [123]. Influenced by tidal processes, the intertidal zone is exposed to highly dynamic conditions, and it shows high ecological diversity and productivity [124].

Intertidal ecosystems can provide crucial ecological services such as nutrient cycling [125], carbon storage [126], storm surge protection [127] and nursery habitats for aquatic life [128]. Moreover, the intertidal areas can also be reclaimed for commercial and recreational usage [129]. However, intertidal flats also represent a typical environmentally fragile and sensitive zone, which can be easily affected by global climate change, sea-level rise, species invasion, and human activities [130]. In recent years, the intertidal zone has been exposed to anthropogenic threats such as (over-) fishing, high nutrient loads, oil and gas production, or tourism [131], which makes it necessary to realize the intertidal cover classification and continuous monitoring.

In the early stages, the understanding of valuable coastal environments is limited by the lack of data. In-situ observations of intertidal flats are sparse since most of the areas are inaccessible [132]. Satellite remote sensing provides an important resource for monitoring coastal zones [131]. With the fast development of SAR sensors, multi-band and multi-polarization SAR data has been applied for the classification of the intertidal zone in some research [133, 130, 134, 135]. Gade et al. [136] presented a method applied to dual-frequency, co-polarized Spaceborne Imaging Radar-C/X-Band SAR (SIR-C/X-SAR) data for sediments classification. Van Beijima et al. [133] investigated the use of S-band and X-band quad-polarimetric SAR data to map natural coastal salt marsh vegetation habitats. Wang et al. [132] proposed a classification scheme for intertidal sediments and habitats and verified it for different bands of SAR data (L-band, C-band, and X-band, respectively).

However, most existing methods adopt traditional machine learning algorithms to develop classification schemes. It is difficult for them to adaptively capture features and learn classifiers from specific SAR datasets [137]. At the same time, they only design and verify their models on different bands of SAR data separately, which is not true “multi-band”. Besides, existing models divide the surfaces of the intertidal zone into very limited types. The coarse results usually cannot meet the demands of end-users.

In recent years, the framework of deep learning has improved the state-of-the-art in many computer vision tasks, as well as in remote sensing applications [138, 139, 140, 15]. Compared with pre-defined features using machine learning, the features from data-driven deep learning models prove to be more robust under various influential factors [141, 142], which is especially useful for the dynamic intertidal zone. Therefore, deep learning technology offers promise for building new data-driven models for sediments and habitats classification on intertidal flats.

Generally, several CNN-based semantic segmentation methods on SAR images have been proposed [70, 71]. Although those methods can extract classification features from SAR images, the abundant texture information can be further extracted, especially on small SAR datasets. In the intertidal zone, texture features have proved to be very useful in classifying sediments [143] and habitats [144] on intertidal flats.

Texture features are composed of local structural and global statistical properties [145]. Deep learning models are good at extracting local structural features like boundaries. However, there are no clear systems to extract and utilize global statistical texture information for CNN semantic segmentation. In [146], Zhu et al. proved that easily computable textural features have general applicability for a wide variety of image-classification computer vision tasks. Inspired by their work, we design a TE-UNet model based on the UNet [54] framework, taking statistical texture into consideration, when classifying intertidal sediments and habitats on SAR imagery. We evaluate our proposed TE-UNet model using full-polarization SAR data from Radarsat-2 (C band) and ALOS-2 (L band).

In summary, there are four main advantages of the proposed TE-UNet model:

1. Instead of only using four intensity channels (VV, VH, HV, and HH) of quad-polarization SAR data, we also combine decomposition results as TE-UNet inputs to utilize polarimetric information;

2. Radarsat-2 (C band) and ALOS-2 (L band) SAR images are concatenated in the channel dimension to realize multi-band learning using TE-UNet;

3. Global statistical texture information is explicitly learned in the TE-UNet architecture;

4. TE-UNet is used for fine-grained SAR image classification on intertidal flats: sediments are further classified into bright sands (beach) and other sediments; habitats are further classified as bivalves, seagrass, and thin coverage of vegetation or bivalves.

The remainder of this chapter is organized as follows: Section 5.2 introduces the details of the dataset, including the area of interest in Section 5.2.1 and SAR data used herein in Section 5.2.2. In Section 5.3, we describe the whole processing chain diagram and the structural details of the TE-UNet model. Verification methods and comprehensive analyses of the model results are shown in Section 5.4. Finally, Section 5.5 concludes this chapter.

5.2 Dataset

5.2.1 Study Area

As shown in Figure 5.1, the study area is located in the northern part of the German Wadden Sea, between the islands of Amrum and Föhr. This area belongs to the world's largest coherent intertidal area, the Wadden Sea, stretching over more than 500km along the North Sea coasts of the Netherlands, Denmark, and Germany [123, 147].

The sediments (pure or mixed with mud) are the dominating the surface type of this area, and their distribution strongly depends on the local hydrodynamic forces. Vegetated areas and bivalve beds (mainly Pacific oysters and cockles, but also blue mussels) are also contained in this region [135]. In order to realize the fine-grained classification, we divide surfaces of the study area into six types: land, seagrass, bivalves, bright sands (beach), water, sediments, and thin coverage of vegetation or bivalves.

Figure 5.2 is the classification map in the study area (color coding). The bright sands (beach), sediments, and bivalves classification information is derived from Landsat-8 OLI data acquired in the years 2014-2016 (© Brockmann Consult 2020). The classification of seagrass and bivalves is based on SPOT-4 data acquired in August 2015 and April 2016 (© Brockmann Consult 2020). The bivalves locations in SPOT-4 are used to adjust derived information from Landsat-8 OLI.

We split the whole classification map (1187 × 1699 pixel) into three

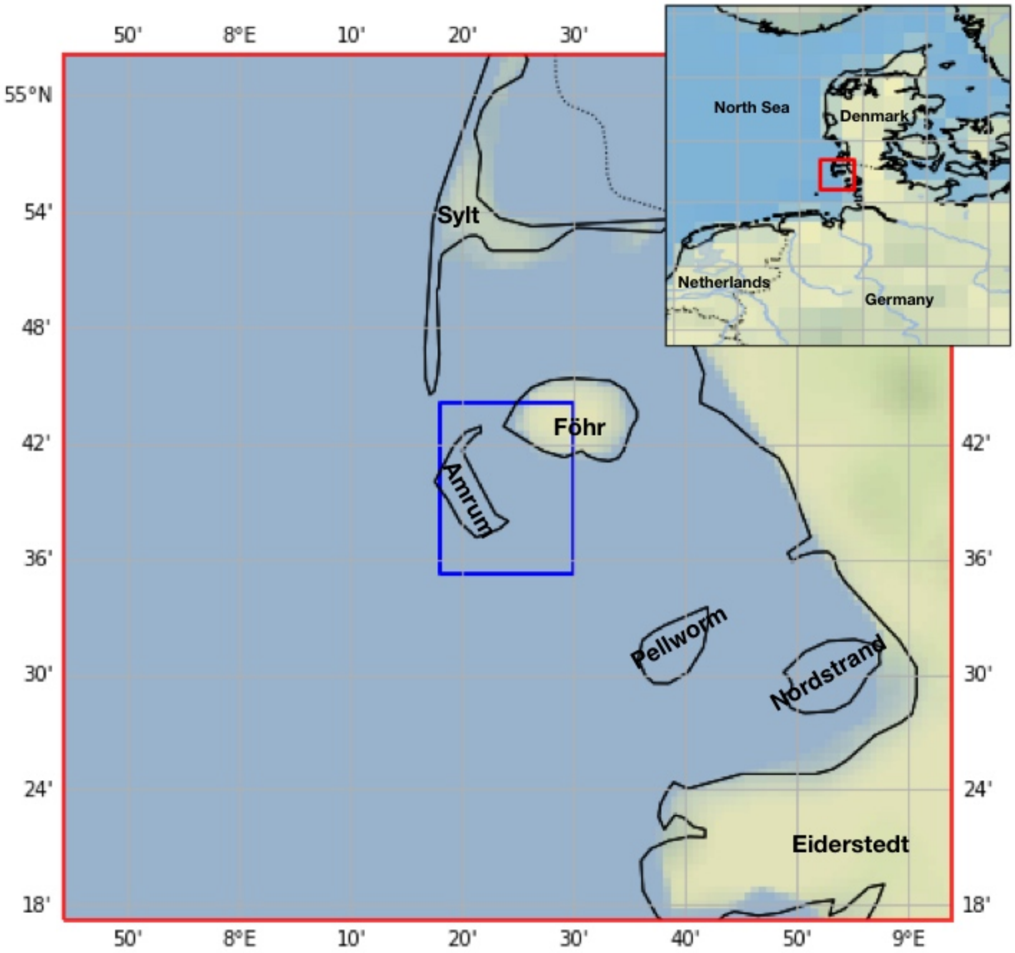
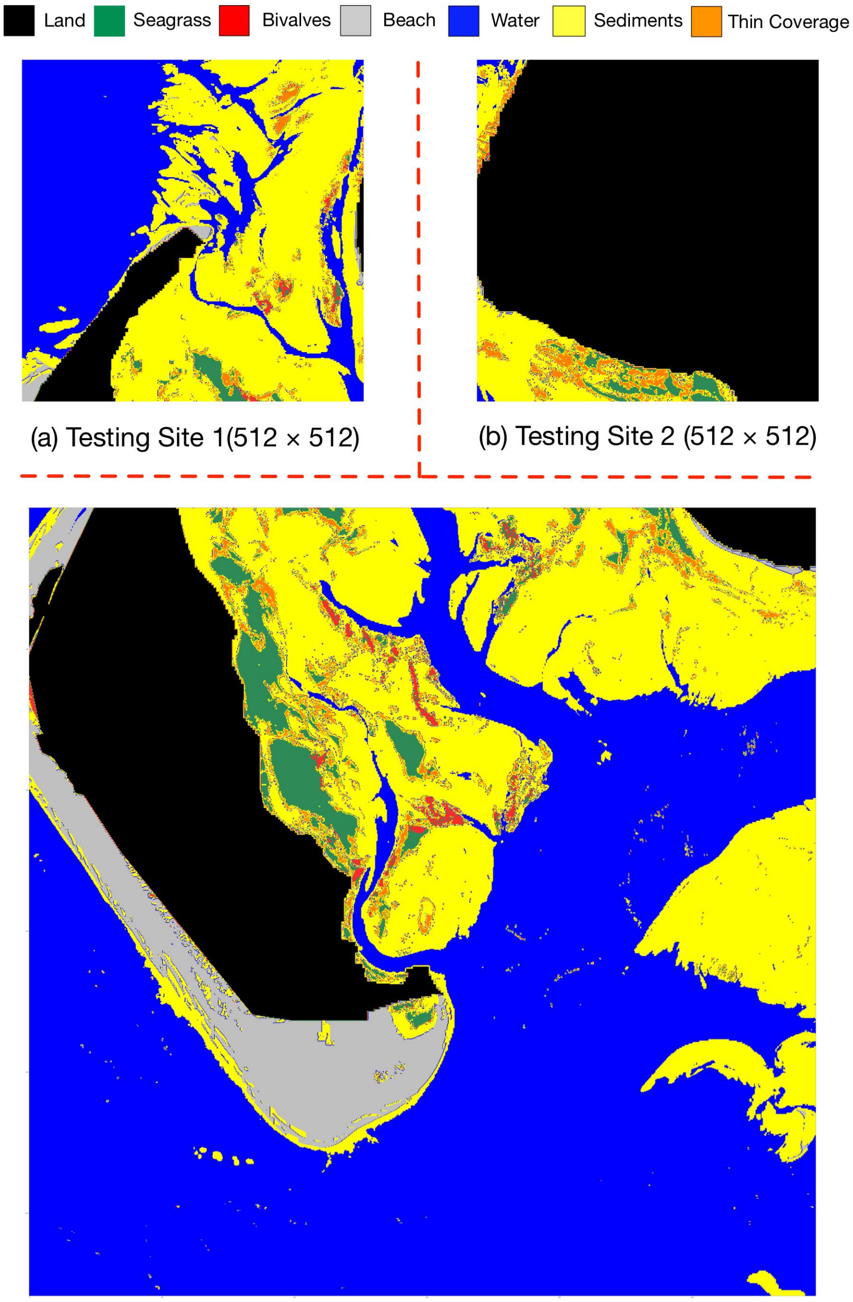


Figure 5.1: Area of interest (blue rectangle in the large map) on the German North Sea coast, east of the island of Amrum and southwest of the island of Föhr.

parts. Figure 5.2 (a), (b) are used for training, and (c) is used for testing. It indicates that we use around 70% of data for training and 30% of data for testing. So there is no data leakage in the testing process. The bivalves, seagrass, and thin coverage classes are very limited both in the training and testing sites. The beach class only has small samples in the testing sites while it accounts for a relatively high portion in the training site. In theory, these imbalanced data and the distributional differences between training and testing datasets will have an adverse effect on final classification results.



(a) Testing Site 1(512 × 512)

(b) Testing Site 2 (512 × 512)

(c) Training Site (1187 × 1187)

Figure 5.2: Classification map in the study area, which is split into three pieces: (c) is used for training, (a) and (b) are reserved for testing.

5.2.2 PolSAR Data

Two Single-Look Complex (SLC) SAR images of the study area are used. Their pixel sizes range from 1 m × 1 m to 5 m × 5 m. Figure 5.3 shows VV-polarization images of RS2 and ALOS2 acquired on 24 December 2015 and 29 February 2016, respectively. The beach and seagrass classes appear as dark patches. This may be caused by remnant water, which flattens the surface. The bivalves make the intertidal surface rougher, so they show as bright patches. More information on the SAR data can be found in Table 5.1.

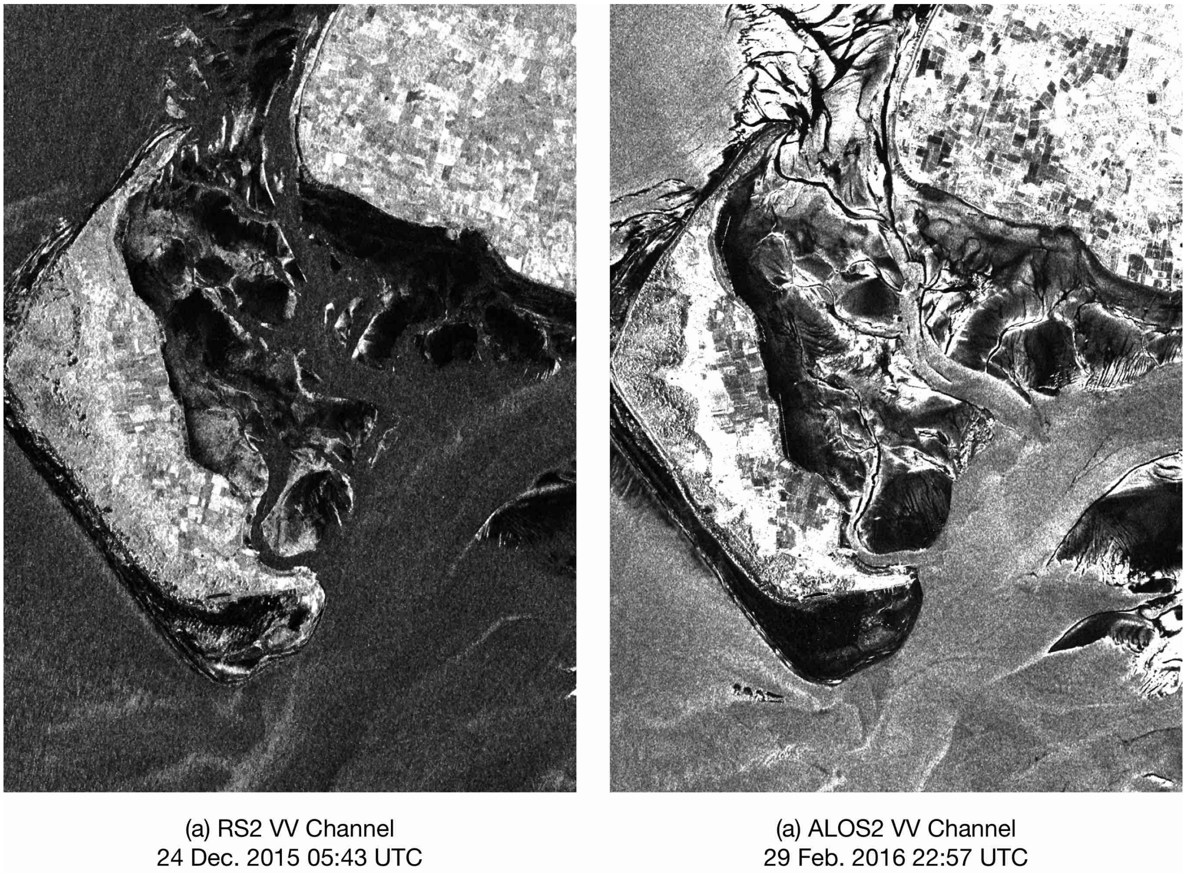


Figure 5.3: SAR images of the study area acquired shortly after low tide: (a) RS2 VV-polarization channel, (b) ALOS2 VV-polarization channel. RS2: ©MacDonald, Dettwiler and Associates Ltd. 2015; ALOS2: ©JAXA 2016.

sensor/band	date/time	low tide time/water level	water level
RS2/C	24 Dec 2015/05:43 UTC	05:25 UTC/-103 cm	-94 cm
ALOS2/L	29 Feb 2016/23:10 UTC	23:46 UTC/-176 cm	-171 cm

Table 5.1: SAR acquisition information. Water levels are measured at the tide station in Amrum, Hafen (Wittdünn).

5.3 System Architecture

The classification systems are divided into two parts. The first part is pre-processing of SAR and classification data. The second part is to use the processed results to feed the TE-UNet model. We will give detailed descriptions of these two parts in Section 5.3.1 and Section 5.3.2, respectively.

5.3.1 Pre-processing Procedure

The flow diagram in Figure 5.4 shows the data processing flow in this chapter. The pre-processing of SAR images and classification results is displayed in blue blocks and green blocks separately. The training and testing phases of TE-UNet are displayed in yellow blocks.

During the pre-processing phase of the PolSAR images, we first apply the radiometric calibration on RS2 and ALOS2 quad-polarization SAR data to convert the image pixel values from Digital Number (DN) to a standard geophysical measurement unit of radar backscatter. Next, polarization scattering matrices [**S**] are transformed to coherence matrices [**T**]. We then apply a Multi-looking operation to suppress speckle noises. After that, Slant to Ground Range conversion and Terrain Correction is also conducted to remove geometry-dependent radiometric distortions. A Refined Lee polarization filter with a window size 7 × 7 (range direction × azimuth direction) is applied for further speckle reduction. We then apply polarimetric decomposition algorithms to extract polarized information.

Rational Polynomial Coefficient (RPC) Orthorectification and resizing

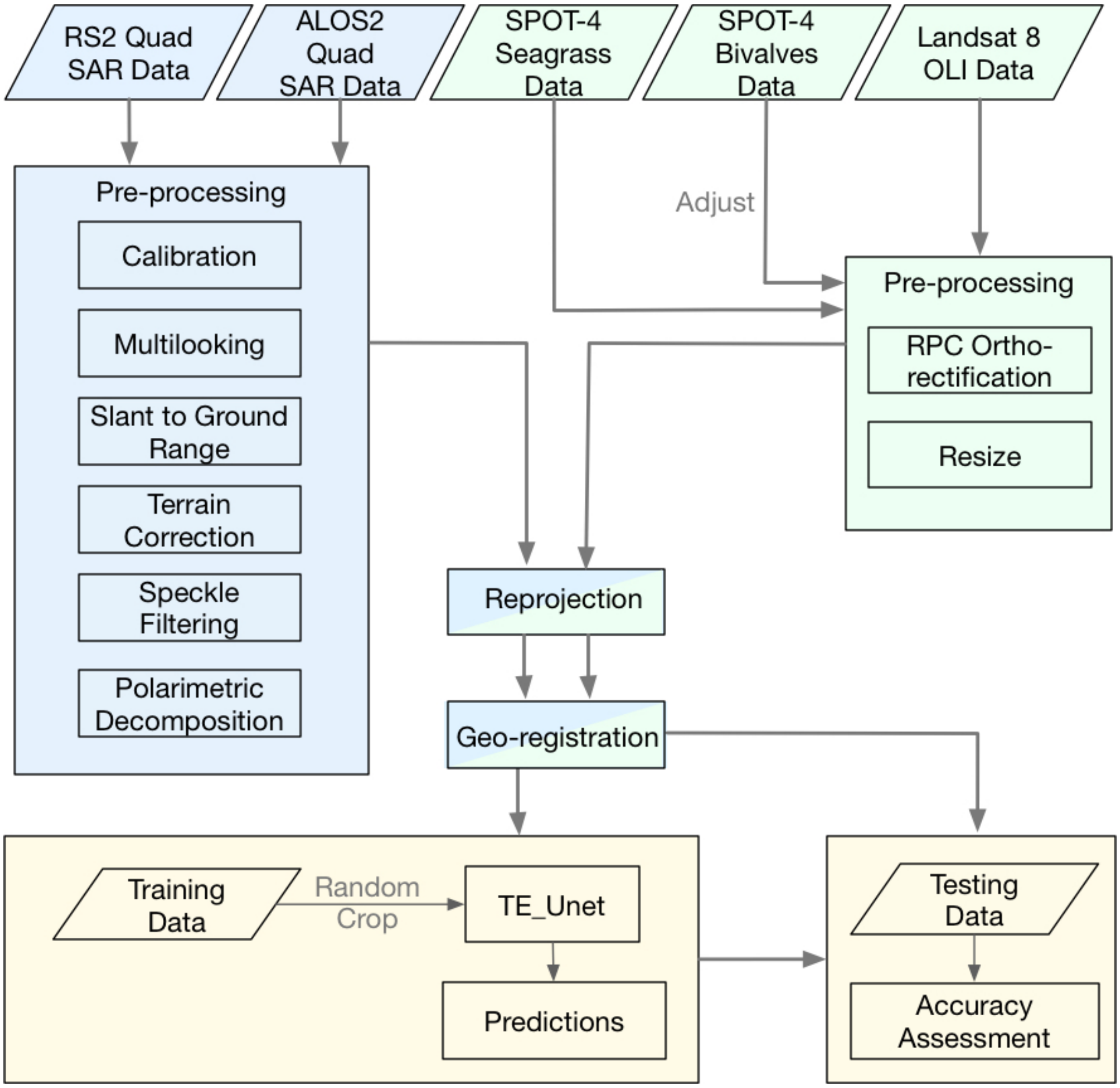


Figure 5.4: Processing diagram for TE-UNet.

operations are conducted on the classification data from SPOT-4 and Landsat 8. We finally reproject both SAR and classification data to the Universal Transverse Mercator (UTM) zone 32N system and realize geo-registration under the same geographical coordinates. Finally, the processed SAR data and classification map are divided into training sets and testing sets for the TE-UNet model.

5.3.2 SAR Polarimetric Decomposition

During the classification process, two classic polarimetric decomposition algorithms are involved: Cloude-Pottier (CP) and Freeman-Durden (FD). The rationale of polarimetric decomposition techniques is to extract important information about the structure of the ground target, the scattering mechanism of the return signals, and the apparent shift in the phase of the signal from the target [148].

The CP polarimetric decomposition uses eigenvalues and eigenvectors to describe the dominant scattering mechanisms of each target [30]. We can use three parameters with physical meanings for CP decomposition analysis:

1. **Entropy** to characterize scattering randomness (represented by H);

2. **Anisotropy** to measure the contribution of one scattering mechanism to total scattering power (represented by AN);

3. **Alpha angle** to provide information, which scatter mechanism dominates, ranging between 0 and 90 degrees (represented by α).

The FD polarimetric decomposition is a physically-based model, which can be used to describe the polarimetric backscatter from naturally occurring scatterers [29]. We can extract three orthogonal scattering components of the FD model:

1. **Volume scatter** from a cloud of randomly oriented dipoles (represented by vol);

2. **Double-bounce scatter** from a pair of orthogonal surfaces with different dielectric constants (represented by dbl);

3. **First-order Bragg scatter** from a moderately rough surface (represented by odd).

5.3.3 Texture Enhancement UNet

5.3.3.1 The Whole Architecture

Figure 5.5 illustrates the whole network architecture of TE-UNet, which is based on the state-of-the-art model UNet [54]. This network is inspired by the Statistical Texture Learning Network (STL-Net) for semantic segmentation [146].

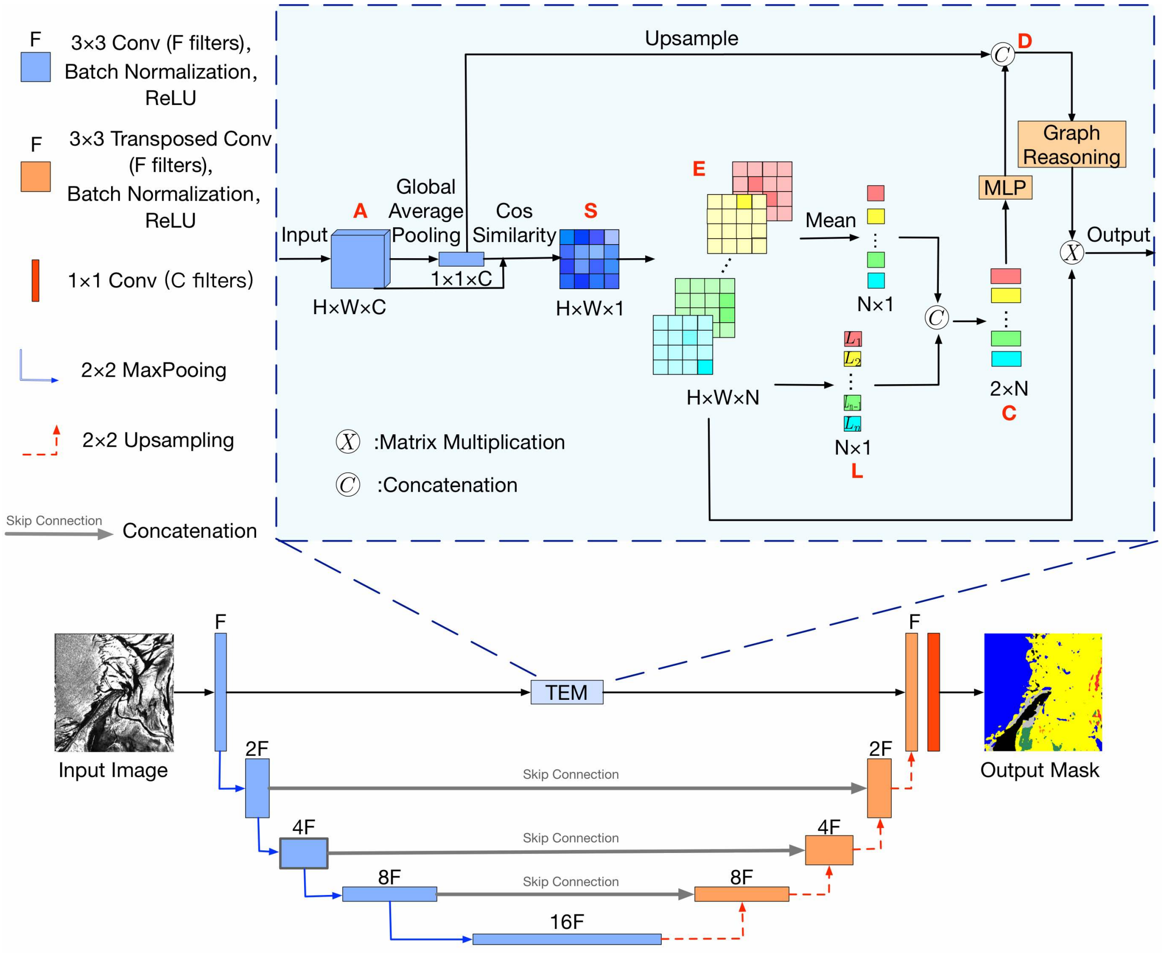


Figure 5.5: Illustration of the overall architecture of TE-UNet.

Recent deep learning-based semantic segmentation models focus on learning high-level features. The abundant details and structural information in PolSAR images would be ignored in this framework. Researchers started to use skip connections to preserve the low-level features. However, the extracted low-level features are often of low quality, especially for SAR images with speckle noise. Moreover, the ambiguous texture details will be gradually ignored in the training process of the model.

When we use the traditional methods to realize a SAR image classification in the intertidal zone, global textural information like intensity column diagrams will commonly be considered. Based on this idea, TE-UNet is proposed by using TEM to preserve and transfer better low-level features.

The TE-UNet consists of three parts: an encoder to extract features, a decoder to generate a semantic segmentation mask, and a TEM module to replace plain skip connection in the first layer to incorporate global textual information.

There are four convolution layers in the encoding phase and four corresponding deconvolution layers in the decoding phase, followed by a 1 × 1 convolution to output the prediction mask. The feature maps from the same scale encoder layer are directly received in the decoder except for the first layer. We set F in Figure 5.5 to be 64. The details of TEM will be introduced in the next part of this section.

5.3.3.2 Texture Enhancement Module

The TEM is used to enhance texture details by learning the global distribution of low-level features. The inputs are the feature maps from the first encoding layer. The outputs are directly sent to the corresponding decoding layer.

We tried to concatenate the TEM output and the encoding feature maps and then sent them to the decoding layer, but the results were worse, even compared with UNet. The reason may be that we conduct a TEM operation on encoding feature maps that are filled with local textural information. The output of TEM has already captured enough low-level details. Only using the TEM output can keep the learned textural features without adding noise, which is more effective in our limited SAR dataset. Therefore, the plain skip connection is not used in TEM.

The input feature maps of TEM are denoted as $\mathbf{A} \in \mathbb{R}^{H\times W\times C}$. H, W, C refer to the height, width, and channel numbers of feature maps, respectively. After applying global average pooling on each feature map, we get the global average feature map $g \in \mathbb{R}^{1\times 1\times C}$. The cosine similarity between g and A is calculated on every pixel of the feature map and is denoted as:

$$\mathbf{S}_{i,j} = \frac{g \cdot \mathbf{A}_{i,j}}{\|g\|_2 \cdot \|\mathbf{A}_{i,j}\|_2} \tag{5.1}$$

where $\mathbf{A}_{i,j}$ represents each pixel in the feature maps $\mathbf{A}_{i,j}(i \in [1, W], j \in [1, H])$, and $\mathbf{S}_{i,j}(i \in [1, W], j \in [1, H])$ represents the cosine similarity of each pixel.

The $\mathbf{S}$ is devided into $\mathbf{N}$ parts equally: $L = [L_1, L_2, \ldots, L_N]$. $\mathbf{S}_{i,j}(i \in [1, W], j \in [1, H])$ is then quantized with $\mathbf{N}$ functions to get the presentation of the quantization encoding matrix, $\mathbf{E} \in \mathbb{R}^{H\times W\times N}$:

$$\mathbf{E}_{i,n} = \begin{cases} 1 - |\mathbf{L}_n - \mathbf{S}_i| & \text{if} \quad -\frac{0.5}{N} \leq \mathbf{L}_n - \mathbf{S}_i < \frac{0.5}{N} \\ 0 & \text{else} \end{cases} \tag{5.2}$$

where n is the nth level of $\mathbf{L}_n$. Then we concatenate $\mathbf{L}$ and the average $\mathbf{E}$ on each feature map to get the quantization counting map, $\mathbf{C} \in \mathbb{R}^{N\times 2}$:

$$\mathbf{C} = \text{Concat}\left(\mathbf{L}, \frac{\sum_{i=1}^{HW} \mathbf{E}_{i,n}}{\sum_{n=1}^{N} \sum_{i=1}^{HW} \mathbf{E}_{i,n}}\right) \tag{5.3}$$

where $Concat$ means concatenate operation in the channel dimension.

Since $\mathbf{C}$ encodes the relative statistics of $\mathbf{A}$, the global average feature map g is then concatenated to get the absolute relative statistics of $\mathbf{D}$.The g is upsampled to $\mathbb{R}^{N\times C}$ and the $\mathbf{C}$ are sent to MLP to increase dimension. The final $\mathbf{D}$ can be expressed as:

$$\mathbf{D} = \text{Concat}(MLP(\mathbf{C}), g) \tag{5.4}$$

Like traditional histogram equalization methods, the statistics map $\mathbf{D}$ needs to reconstruct quantization levels. This is realized by a Graph Reasoning Module. We perform the relation reasoning via a simple GCN in the interaction space:

$$\mathbf{X} = \mathrm{Softmax}\left(\phi_1\left(\mathbf{D}\right)^T \cdot \phi_2\left(\mathbf{D}\right)\right) \tag{5.5}$$

$$\mathbf{L}' = \phi_3(\mathbf{D}) \cdot \mathbf{X} \tag{5.6}$$

where $\mathbf{L}'$ is the resonstructed level, and ϕ_1, ϕ_2, ϕ_3 operations are conducted by 1×1 convolution. The final output is the quantization encoding map $\mathbf{E}$ on $\mathbf{L}'$ realized by matrix multiplication:

$$\mathbf{R} = \mathbf{L}' \cdot \mathbf{E} \tag{5.7}$$

The reshaped $\mathbf{R}^{C_2 \times H \times W}$ is the final output of the TEM.

5.4 Experiments

5.4.1 Experimental Setup

All experiments are implemented in PyTorch and conducted on a GeForce RTX 2080 Ti GPU. We train the models by using the Adam optimizer with a momentum set as 0.9 and a weight decay set as 0.0005. An initial learning rate of 0.0001 is used. We use the common cross entropy loss for the multi-class segmentation task.

During the training process, we randomly crop the training site in Figure 5.3 (c) into patches of size 512 × 512. Random flipping augmentation is performed on the training dataset. The models are trained with a batch size of 4 and roughly 1×10^5 steps.

5.4.2 Evaluation Metrics

We evaluate the semantic segmentation results of different methods based on five metrics: per-class F1 score, Mean F1 score (mF1), Mean

Intersection over Union (mIoU), Average Accuracy (AA), and Overall Accuracy (OA).

Comparing the reference classifications and prediction results, we can get the confusion matrix:

$$P = \{p_{ij}\} \in \mathbb{N}^{k \times k} \tag{5.8}$$

where p_{ij} represents the number of pixels that belong to class i and are identified as class j, k is the number of classes, which is equal to seven in our work. Especially, p_{ii} is the number of pixels that are classified correctly. The corresponding average precision (P) and recall (R) can be denoted as:

$$P = \frac{1}{k}\sum_{i=1}^{k} \frac{p_{ii}}{\sum_{j=1}^{k} p_{ji}}, R = \frac{1}{k}\sum_{i=1}^{k} \frac{p_{ii}}{\sum_{j=1}^{k} p_{ij}} \tag{5.9}$$

The F1-score is a harmonic mean between P and R, which is useful for imbalanced classes. We calculate it by:

$$mF1 = 2 \times \frac{P \times R}{P + R} \tag{5.10}$$

The mIoU is one of the stringent metrics used for evaluation of image segmentation that takes every pixel into account, and is expressed as :

$$\text{mIoU} = \frac{1}{k}\sum_{i=1}^{k} \frac{p_{ii}}{\sum_{j=1}^{k} p_{ij} + \sum_{j=1}^{k} p_{ji} - p_{ii}} \tag{5.11}$$

The OA is the ratio between the number of correctly classified pixels to the total number of pixels in the testing set, and it is calculated as:

$$\text{OA} = \frac{\sum_{i=1}^{k} p_{ii}}{\sum_{i=1}^{k} \sum_{j=1}^{k} p_{ij}} \tag{5.12}$$

The AA refers to the average result of accuracies in all classes:

$$\text{AA} = \frac{1}{k}\sum_{i=1}^{k} \frac{p_{ii}}{\sum_{j=1, j \neq i}^{k} p_{ij} + p_{ii}} \tag{5.13}$$

5.4.3 Comparison Results

We compare our TE-UNet model with two popular state-of-the-art semantic segmentation models, Deeplab V3 Plus [55] and UNet [54]. Two frameworks specially designed for SAR images semantic segmentation like HR-SARNet [71] and TL-FCN [70] are also used as control groups. For a comparison, we keep the same hyper-parameters of TE-UNet like learning rate, batch size, optimizer, etc.

Note that we use the 14-channel input for all models: 3 CP components channels (H, AN, α) and 4 intensity channels (HH, HV, VH, and VV) for both RS2 and ALOS2 SAR data. We choose this setting according to the ablation study results in Section 5.4.4.

Table 5.2 summarizes the quantitative results of different models. In general, we observe that our TE-UNet obtains the best results in $mF1, mIoU$ and AA, and is only slightly (0.30%) below Deeplab V3 Plus in terms of OA. Since we have very limited training and testing pixel samples of classes seagrass, bivalves, beach, and thin coverage, their metrics are not as good as those of land, water, and sediments. Also note that the features for the low-accuracy classes are also more difficult to learn. Considering our sediments and habitats classification as an extreme sample imbalance task, the AA metric can better reflect the model capabilities, as compared to OA. In fact, the visualization results in Figure 5.6 show that Deeplab V3 Plus is overfitting to classes land, water, and sediments.

Model	F1(%)							mF1(%)	mIoU(%)	AA(%)	OA(%)
	land	Seagrass	Bivalves	Beach	Water	Sediments	Thin Coverage				
DeeplabV3 Plus	**97.49**	18.09	0.28	3.37	79.73	**78.74**	0.00	39.67	34.02	40.21	**84.25**
UNet	96.39	13.83	3.18	15.09	79.65	77.23	**3.09**	41.21	34.41	42.87	83.04
HR-SARNet	96.31	18.39	**10.05**	3.99	78.91	78.32	0.00	40.85	34.27	41.80	83.14
TL-FCN	95.82	9.05	9.30	16.08	**80.17**	77.25	0.00	41.09	34.31	42.52	83.01
TE-UNet	97.11	**18.87**	2.30	**18.49**	79.63	77.75	1.49	**42.23**	**35.43**	**43.69**	83.95

Table 5.2: Comparison of quantitative results of different instance segmentation models for intertidal sediments and habitats classification.

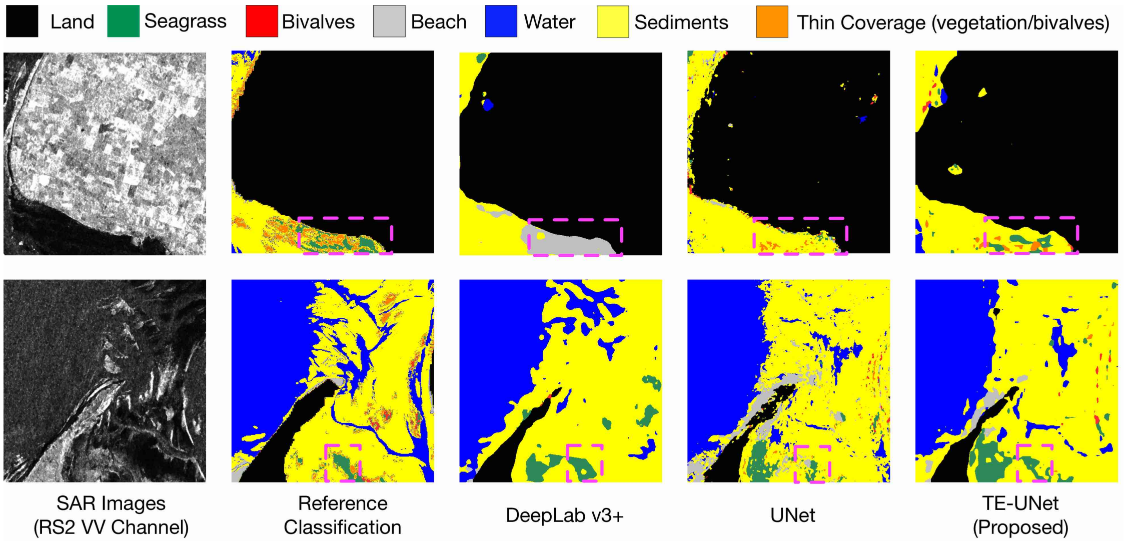


Figure 5.6: Comparison of segmented maps obtained by different models on testing areas. Pink rectangles highlight some locations, where the proposed TE-UNet model produces finer segmentation predictions.

Compared with the basic framework UNet, our proposed TE-UNet increases the mean metrics $mF1, mIoU, AA$ and OA by 1.02%, 1.02%, 0.82%, and 0.91%, respectively, which proves the effectiveness of the texture enhancement module. We also note that the basic UNet outperforms the complicated HR-SARNet and TL-FCN in terms of $mF1, mIoU$, and AA. Our application is maybe not content with the very high-resolution conditions in HR-SARNet. The new branches for FCN in TL-FCN have also been invalid in our dataset.

In order to intuitively display the superiority of TE-UNet, we show the prediction masks of the testing sites in Figure 5.6. Compared with the UNet baseline, our proposed model seems to be capable of predicting relatively smoothed but precise locations, which is likely an effect of TEM.

It is noteworthy that RS2 and ALOS2 images are acquired at different times in different tide cycles, resulting in different environmental backgrounds (e.g. wind speed, weather conditions) and water levels in the same area. Apart from the frequency, these factors will also influence the radar backscatter recorded by RS2 and ALOS2 sensors.

5.4.4 Ablation Study

We comprehensively evaluate our TE-UNet model on different input sources. For each band, the CPI (H, AN, α, HH, HV, VH, and VV) data is adopted as input. We use the same hyper-parameters in all experiments.

5.4.4.1 Multi-band Input

We first study the design choices of the RS2 (C band) and ALOS2 (L band) multi-band input. There are five design choices in Table 5.3 and Figure 5.7. The RS2 and ALOS2 are different 7-channel SAR data, the "Training" keyword means that we use this kind of data as TE-UNet input, and the "Testing" keyword means that we use this kind of data for performance testing. The "+" operation between RS2 and ALOS2 means that we concatenate them in the channel dimension.

Training Dataset	Testing Dataset	F1(%)							mF1(%)	mIoU(%)	AA(%)	OA(%)
		land	Seagrass	Bivalves	Beach	Water	Sediments	Thin Coverage				
RS2	RS2	93.01	15.28	0.11	11.22	76.82	73.65	0.72	38.69	31.75	40.53	79.93
ALOS2	RS2	24.32	1.13	0.00	0.09	65.93	35.33	0.00	18.11	12.16	22.81	39.92
RS2	ALOS2	87.21	0.94	0.00	2.53	0.42	8.18	0.00	14.18	11.94	23.03	47.30
ALOS2	ALOS2	95.61	17.90	**5.70**	18.15	77.77	**77.87**	**1.91**	42.13	34.48	42.29	82.73
RS2+ALOS2	RS2+ALOS2	**97.11**	**18.87**	2.30	**18.49**	**79.63**	77.75	1.49	**42.23**	**35.43**	**43.69**	**83.95**

Table 5.3: Comparison of quantitative results of different train and test dataset inputs for intertidal sediments and habitats classification.

The quantitative results are shown in Table 5.3. The "Training: RS2 + ALOS2, Testing: RS2 + ALOS2" setting obtains the best results in all average metrics, which proves that a combination of SAR data from different bands indeed helps the model in learning. The ALOS2 data behaves much better compared with the RS2 data. A possible explanation may be that the ALOS2 data (L band) emits longer waves whose penetration ability is stronger. This ability is very important to classify cover types in the intertidal zone like thin coverage.

The results of the "Training: ALOS2, Testing: ALOS2" setting are very close to our best results. Adding RS2 data makes the performance of seagrass, bivalves, and thin coverage classes drop slightly. More complex multi-band fusion networks can be designed for better feature extraction. When the training and testing data are from different bands, a dramatic drop happens in all the metrics. This result shows that SAR image features of different bands reflect different backscatter mechanisms.

The visualization results of the five design choices are shown in Figure 5.7. The seagrass, bivalves, and thin coverage classes nearly disappear under training and testing for the RS2 setting. But the classification of training and testing on the ALOS2 setting gets the prediction maps closer to our proposed method. The models trained and tested from different bands data confuse the dominating pixels and lose the power of detecting classes bivalves and thin coverage.

5.4.4.2 Multi-polarization Input

We studied the design choices of a combination of multi-polarization SAR input. There are seven design choices in Table 5.4 and Figure 5.8. The "I or FD or CP" keyword means we only use either four intensity channels (HH, HV, VH, and VV) or three FD components or three CP components as the TE-UNet input. The FD and CP components are combined using band concatenation as "FDCP" or unified with four intensity channels separately as "FDI" or "CPI". In the last control group, we used intensity channels, FD components, and CP components together as "FDCPI".

Table 5.4 displays quantitative results of seven comparison groups. The "CPI" setting achieves the best metrics of all the average performance among them. In theory, the "FDCPI" setting contains the most information compared with other settings, but it only slightly (0.29%) improves the sediment class metrics compared with "CPI". One possible reason is that the concatenation method is too simple for fusion

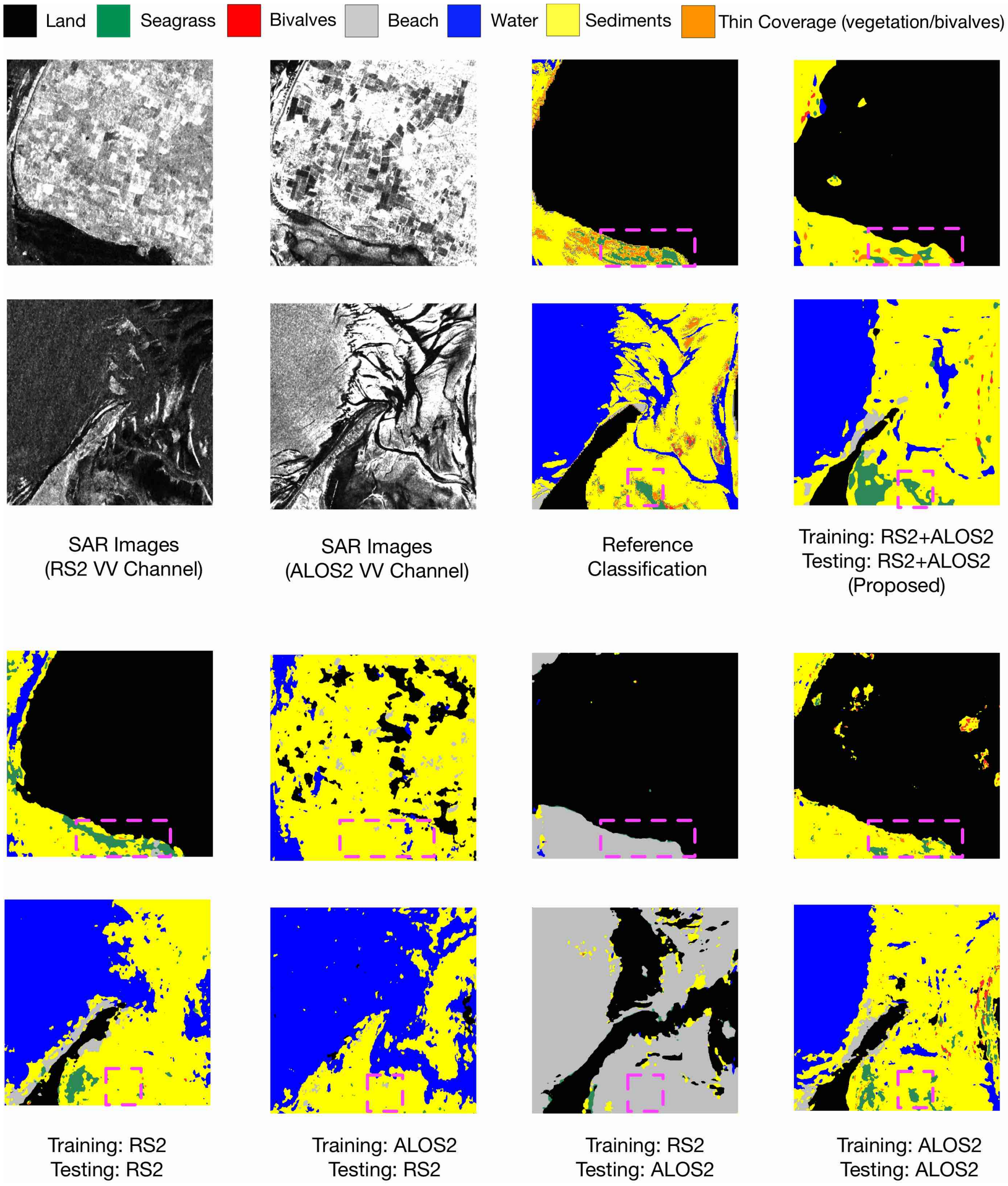


Figure 5.7: Comparison of segmented maps obtained by different training datasets and testing datasets on testing areas. Pink rectangles highlight some locations where the proposed method (Training: RS2+ALOS2, Testing: RS2+ALOS2) produces finer segmentation predictions.

intensity and polarimetric decomposition information. This information may interfere with each other during feature extractions.

In general, the different combinations of FD components, CP components, and intensity channels improve the final average metrics compared with using them alone.

Input	F1(%)							mF1(%)	mIoU(%)	AA(%)	OA(%)
	land	Seagrass	Bivalves	Beach	Water	Sediments	Thin Coverage				
I	95.65	9.39	7.36	13.31	78.31	76.93	**2.94**	40.56	33.69	41.02	82.48
FD	96.25	10.73	5.40	8.09	77.18	77.36	2.63	39.66	33.24	39.64	82.93
CP	96.34	12.35	1.56	14.94	77.82	77.89	0.00	40.13	33.70	40.76	83.51
FDI	96.14	11.25	6.69	13.58	79.41	77.43	1.71	40.89	34.17	41.97	83.06
CPI	**97.11**	**18.87**	2.30	**18.49**	**79.63**	77.75	1.49	**42.23**	**35.43**	**43.69**	**83.95**
FDCP	94.47	18.16	**11.86**	14.18	66.03	72.94	0.26	39.70	31.47	42.74	78.07
FDCPI	96.34	14.44	3.09	16.13	78.97	**78.04**	1.03	41.15	34.40	41.99	83.55

Table 5.4: Comparison of quantitative results of different input combinations of intensity channels (I), FD components (FD) and CP components (CP) for intertidal sediments and habitats classification.

Figure 5.8 is the visualization results of the seven design choices. The "CP" setting is good at identifying the sediment class, but fails to detect thin coverage class, which is in complete agreement with the quantitative results. The "FDCP" setting helps a lot to find the bivalve areas. A possible interpretation is that polarization characteristics can better reflect the features of specific classes like bivalves compared with the intensity information. This result also proves the necessity of adding different polarimetric decomposition components as the model inputs.

5.5 Summary

In this chapter, we propose a TE-UNet model for classification of sediments and habitats in the intertidal zone. The experimental results demonstrate that our model provides high-quality semantic segmentation on multi-band and multi-polarization SAR images.

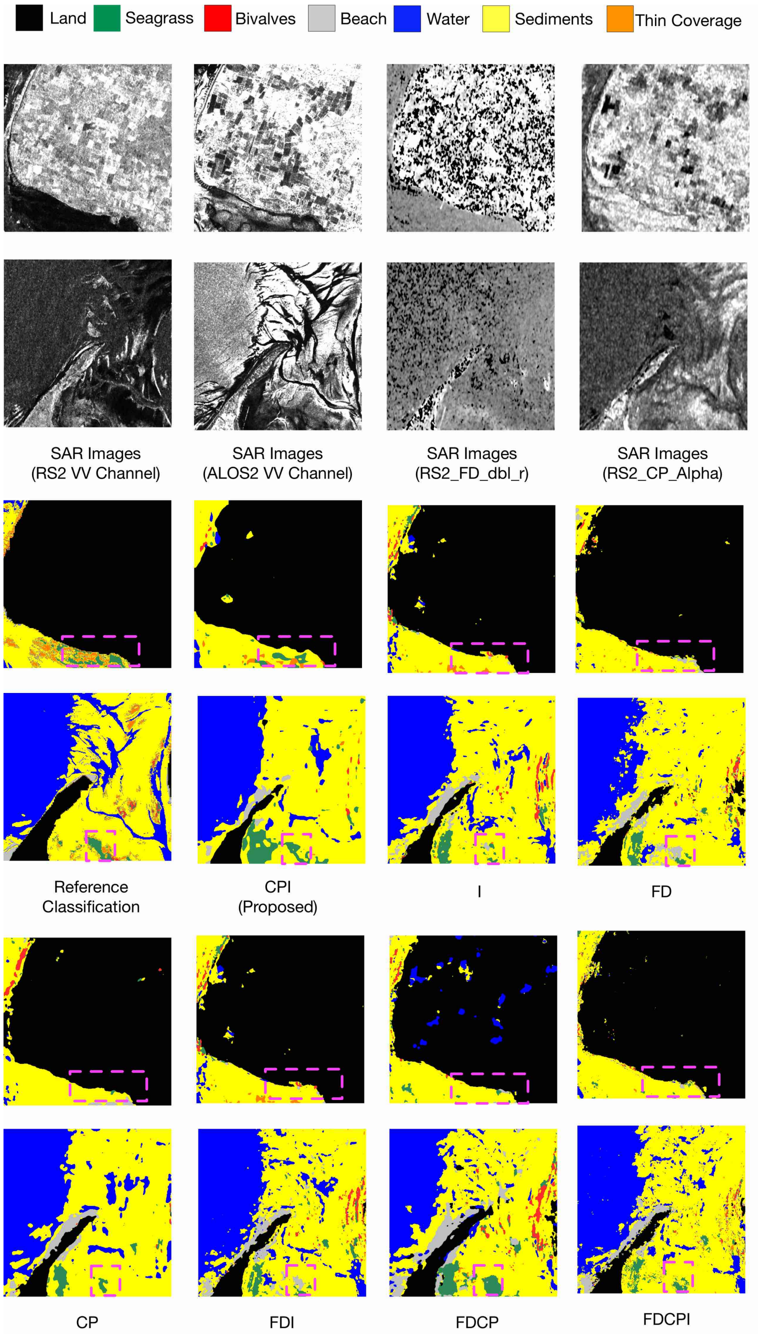


Figure 5.8: Comparison of segmented maps obtained by different input channels on testing areas. Pink rectangles highlight some locations where the proposed method (Input: CPI channels) produces finer segmentation predictions.

The application of the texture enhancement module improves the performance by explicitly enhancing the global texture information learning. Compared with the UNet baseline, TE-UNet improves the classification accuracies in terms of all average metrics. The visualization results also prove that TE-UNet can provide finer resolutions.

The comparative experimental study proves the effectiveness and potential of the multi-band and multi-polarization system for classification tasks in the intertidal zone. For the multi-band input, TE-UNet learns different features on SAR images from different bands. In general, we can obtain better classification results on ALOS2 (L-band) data compared with RS2 (C-band) data, due to the longer wavelength. But if we combine these two bands, we can obtain the best results. For the multi-frequency input, the CPI setting achieves the best average performance among the seven comparison groups. CP components are more effective than FD components in our work. CP components and intensity channels contain different features of the surface types of the intertidal zone. Therefore, the combination of two of them finally improves the classification performance compared with using them alone.

In future work, we will continue to carry out the design of the texture enhancement. Besides, more suitable fusion mechanisms for multi-band and multi-polarization SAR data can be designed. We can add more polarimetric decomposition components like the DERD parameter [134] to distinguish sediments and habitats on intertidal flats. For future intertidal monitoring, the combination of multi-sensor (e.g., SAR and optical data) shows potential to improve the model performance significantly.

Chapter 6

SOF-UNet: SAR and Optical Fusion UNet for Land Cover Classification

6.1 Introduction

Land cover is the natural and man-made characteristics of the Earth's surface, including forest, grassland, water, human infrastructure and etc [149]. Classification of land cover is often helpful in various applications like land use planning, climate change detection, natural disasters prediction, and environmental protection [150, 151, 152]. Nowadays, the application of remote sensing data for land cover classification has received widespread attention [153]. With the increasing amount of large-scale multi-sensor remote sensing imagery, it is quite urgent to realize automatic land cover classification.

CNNs [34] have achieved tremendous success in the automatic image segmentation task. Some classic segmentation architectures like UNet [54] and DeeplabV3 Plus [55] are successfully applied in various fields for accurate pixel-wise classification. Recently, more and more deep learning-based methods have been developed for land cover classification using remote sensing data [154, 155, 156, 157]. However, most of these methods only use uni-modal data (optical or SAR). Extensive research shows that the fusion of SAR data helps to discriminate different types of land cover classes, which are indistinguishable in optical data, due to similar spectral characteristics of land features [158]. Hence, synergistically combining these two types of data is an effective way to realize better land cover classification.

Many existing multi-modal fusion strategies for land cover classification remain in the phase of adding and concatenation input bands [159], which lacks the effective use of multi-modal information. Meanwhile, most researchers usually only verify their methods on limited datasets, lacking generalization ability [152]. Therefore, we design our fusion model based on the largest existing dataset, SEN12MS [19], which provides optical and SAR pairs to realize the land cover classification task. There are several studies on this public dataset [160, 161, 162, 163], but to the best of our knowledge, this is the first work to apply multi-modal technology on the SEN12MS dataset. It is noteworthy that in [163], the UNet and Deeplab V3 Plus baselines won the traditional methods RF and k-means in terms of AA by a comfortable margin on the SEN12MS dataset.

In this chapter, a novel SOFNet model based on the DeeplabV3 Plus framework is proposed for multi-modal land cover classification. Considering the low-resolution and noisy labels of SEN12MS, we also apply the custom SCE loss function to alleviate this problem. The SOFNet work provides better segmentation results compared with the methods that simply superimpose SAR and optical images as the input. However, based on the visualization results, we find that many fine-grained details are lost in the SOFNet prediction maps. For this reason, a new multi-modal framework called SOF-UNet is further put forward. Compared with SOFNet, the SOF-UNet model preserves more fine details and receives better classification results in general. In conclusion, the following points show our main contributions:

1. A two-stream SOFNet model for multi-modal land cover classification is proposed, which has two asymmetrically encoding branches and a sharing decoding branch. ASPP modules are adopted

to obtain multi-scale context information for better segmentation results;

2. A novel two-stream SOF-UNet model is designed for finer multi-modal land cover classification. The utilization of fusion skip connections in the encoding and decoding phase improves the feature fusion representational power of the network;

3. To deal with the problem of noisy labels, a custom SCE loss function is proved to be useful in both SOFNet and SOF-UNet models on the SEN12MS dataset.

The remainder of this chapter is organized as follows. In Section 6.2, we give a detailed description of the training dataset, SEN12MS, and the testing dataset, DFC2020. Section 6.3 and Section 6.4 introduce the architecture designs of SOFNet and SOF-UNet, separately. Both qualitative and quantitative results are provided to verify the proposed methods. Conclusions and future work are drawn in Section 6.5.

6.2 Datasets

6.2.1 SEN12MS

The SEN12MS dataset was proposed by Schmitt et al. [19] to promote research in SAR-optical image fusion using deep learning methods for land cover classification. In total, it consists of 180662 patch triplets (collected throughout 4 seasons and 252 scenes across the globe), in which 162556 patches are dedicated for training and 18106 patches used for validation. Every triplet contains:

- Dual-polarized (VV and VH) Sentinel-1 SAR images;

- Multi-spectral (13 bands including RGB, infrared, etc.) Sentinel-2 optical images;

- MODIS land cover maps (simplified IGBP scheme).

The locations of the regions of interest are displayed in Figure 6.1. In the training process, Savanna labels are ignored for a specific geographical reason. More details can be found in [163]. The size of all images is cropped to 256 × 256 pixels. All the input images have a resolution of 10 m per pixel. MODIS labels natively have a resolution of 500 m per pixel, but they are upsampled to a 10 m resolution. This resolution difference also results in the severe noisy label problem.

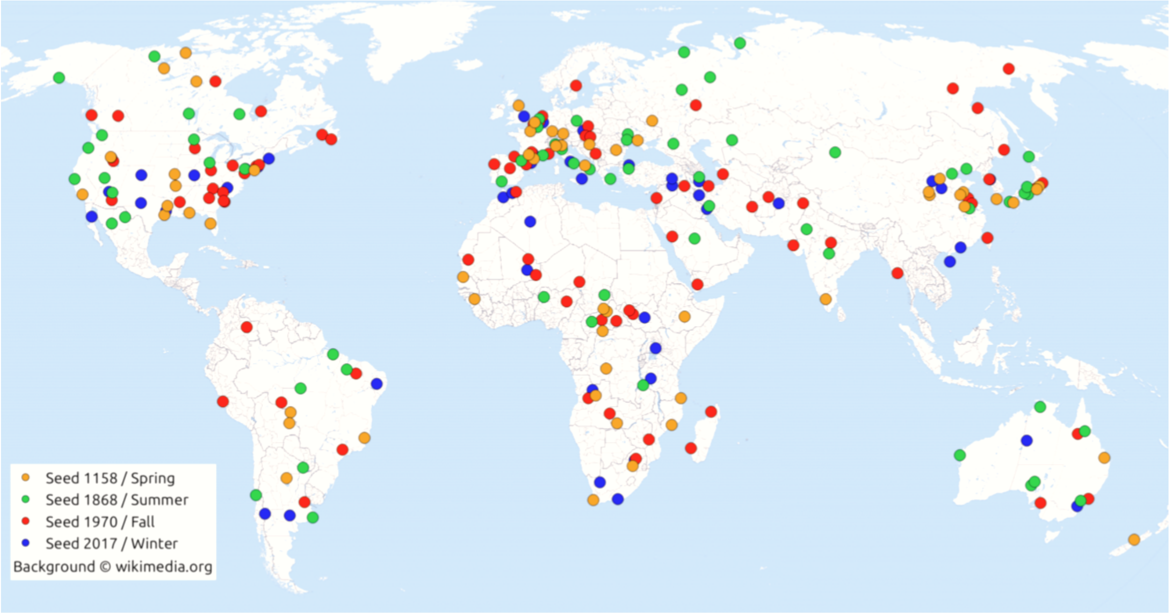


Figure 6.1: Regions of interests of the SEN12MS dataset. Image adapted from [19].

To conclude, there are mainly three challenges for land cover classification on the SEN12MS dataset. The first challenge is the peculiarity of the Savanna class. This class cannot directly be used for training, which results in incomplete supervision. The second challenge is inaccurate and inexact supervision. On the one hand, only coarse-grained MODIS label information is available. On the other hand, the existing labels show the class confusion problem, especially for the classes barren, grassland, wetland, and shrubland. The last challenge is the multi-modal fusion task for SAR and optical input, which is also the key problem we want to solve in this chapter.

6.2.2 DFC2020

The 2020 IEEE-GRSS Data Fusion Contest (DFC2020) dataset [164] is used to test models in our work. It consists of 986 patch quadruplets and 5128 patch quadruplets, with a size of 256 × 256 pixels, for validation and testing, respectively. Basically, DFC2020 shares the attributes with the SEN12MS dataset, but it adds additional high-resolution semi-manually labeled land cover maps (10 m per pixel) with the help of Google Earth aerial imagery.

Figure 6.2 displays the 7 regions of interest of the DFC2020 dataset. It can be found that the geolocation scenes in DFC2020 are not contained in SEN12MS. In our experiments, only the DFC2020 validation dataset is used to test.

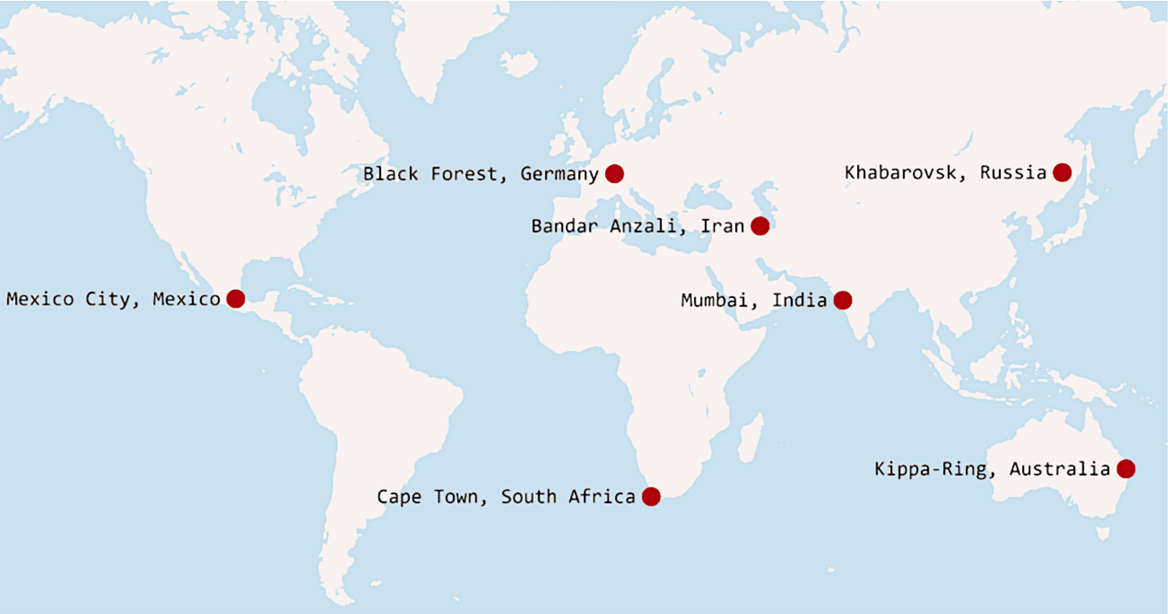


Figure 6.2: Regions of interests of the DFC2020 dataset. Image adapted from [164].

6.3 SAR-Optical Fusion Network

6.3.1 System Architecture

6.3.1.1 The Overall Architecture

The architecture of SOFNet is illustrated in Figure 6.3. It adopts the EncoderDecoder segmentation structure, as described in Section 3.2.2. Two different encoders are designed to extract optical and SAR features, respectively. And then, a corresponding decoder is used for up-sampling the feature maps. Inspired by the DeepLabV3 Plus network [55], ASPP modules are adopted in the final layers of the backbones. Skip concatenation operations of the feature maps are used in both encoder and decoder processes to combine different levels of information. In this chapter, all the Sentinel-1 SAR images are transformed into pseudo color images for better visualization.

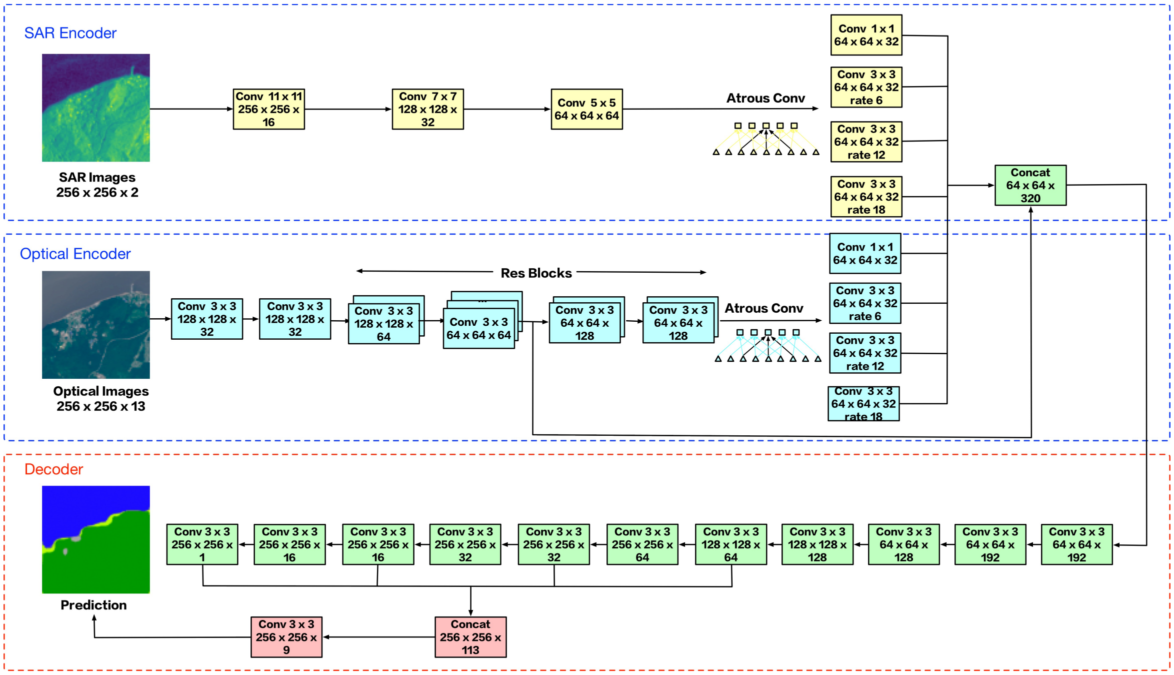


Figure 6.3: Illustration of the overall architecture of SOFNet.

6.3.1.2 The Encoders

The two encoders in SOFNet are designed asymmetrically, considering the differences between the two modalities. For the dual-polarized SAR images, we design a simpler feature extraction backbone. There are only three convolution layers in total in the feature extracting process, but the kernel sizes are larger than the optical convolution layers. This design is based on the consideration that learning useful features from 2-band SAR images is relatively easier compared to 13-band multi-spectral images. Therefore, SAR images are used as the more complete information with respect to optical images. For the optical images, after two convolution operations, we adopt residual blocks to extract more information. Then, the final features of the two branches are put into ASPP modules to learn local-to-global context information. Note that batch normalization and ReLU operations are always applied after each convolution.

We concatenate the outputs of ASPP modules from SAR and optical branches one by one. Then, we add the features of the last but one layer of the optical branch to add more low-level information. It is worth noting that the concatenate operations, not the pixel-wise add operations, are adopted on feature maps, because they can get better results in our experiments.

6.3.1.3 The Decoder

After the fusion convolution operation of encoding outputs, 320 channels feature maps are obtained in total. Then, they are put into transposed convolution layers, to gradually restore the resolution of feature maps to the original size. Only the last convolution layer does not use batch normalization and ReLU operations.

Finally, four different layers of decoding are concatenated to process together and put into the final convolution operation. We tried to use low-level layers of the optical encoding branch, like DeeplabV3 Plus, but it did not contribute to the improvements of the final results, so we deleted this skip connection in our model.

6.3.2 SCE Loss Function

The goal is to provide a loss function that finds the balance between sufficient classification learning and noise tolerance. The categorical Cross Entropy (CE) loss is the most famous loss function for classification problems. It can converge rapidly and generalize to various problems. However, in the case of noisy labels, lots of classes will be mixed together on the feature distributions of CE loss, resulting in a decline in performance. In [165], SCE loss was first proposed to solve noisy-labeled classification problems. We modified it to be suitable for our pixel-level land cover classification task. The CE loss can be designed formally as:

$$l_{ce} = -\sum_{k=1}^{K} q(k \mid x) \log p(k \mid x) \tag{6.1}$$

where $q(k \mid x)$ and $p(k \mid x)$ stand for the ground truth and the prediction classification distributions on sample x of class k. Since the ground truth distribution is not so reliable due to noise, a noise tolerance term is added:

$$l_{rce} = -\sum_{k=1}^{K} p(k \mid x) \log q(k \mid x) \tag{6.2}$$

The final loss function, l_{sce}, consists of weighted cross entropy loss and weighted reverse cross entropy loss:

$$l_{sce} = \alpha l_{ce} + \beta l_{rce} \tag{6.3}$$

l_{rce} is noise tolerant, while l_{ce} is sensitive to noisy labels. But l_{ce} plays a driving effect on the convergence of the model. Notably, $\log q(k \mid x)$ might cause the $\log 0$ problem in the image segmentation task, so we let $\log 0$ equals to a constant A. The hyperparameters A, α and β are also correspondingly changed, according to the experiment results.

6.3.3 Experiments

6.3.3.1 Implementation Details

We conduct our experiments on an NVIDIA Pascal Titan X GPU. The initial learning rate is 0.001 and the training batch size is 16. We train our model until convergence by using the Adam optimizer with a momentum set as 0.9 and a weight decay set as 0.0005. All weights are initialized by a Xavier initialization. For the SCE loss, the parameters A, α, and β are assigned to be 0.1, 6, 3, respectively. Under this setup, the training process takes up to nearly 3 hours and 20 minutes for one epoch. We set the largest epoch as 20 and choose the best model according to the results on the validation dataset.

6.3.3.2 Evaluation Metrics

We choose AA metric for evaluation considering the heavy-imbalanced-classes problem in SEN12MS and DFC2020. AA is the average of each accuracy per class, which is computed in the following formulas:

$$acc_i = \frac{\sum_{k=1}^{K} \theta_{ii}^{k}}{\sum_{k=1}^{K} \theta_{ii}^{k} + \sum_{k=1}^{K} \sum_{j=1, j \neq i}^{N} \theta_{ij}^{k}} \tag{6.4}$$

$$AA = \frac{1}{N} \sum_{i=1}^{N} acc_i \tag{6.5}$$

where acc_i represents the accuracy for class i, θ_{ii}^{k} represents the number of pixels that class i correctly classified as class i in the k-th image, θ_{ij}^{k} represents the number of pixels that class i wrongly classified as class j in the k-th image, K and N are the numbers of test images and classes, and are set as 986 and 8 in this chapter, respectively.

6.3.3.3 Results

A group of comparative experiments is designed to verify the effectiveness of SOFNet. The first step is to compare the segmentation results with state-of-the-art UNet and DeeplabV3 Plus image segmentation networks. We apply the band concatenation operation of SAR and optical images and then put the 15 channels input into UNet and DeeplabV3 Plus separately to get the classification results. We also use the SCE loss on these two models for a fair comparison.

As for the SOFNet architecture, we apply two experiments. Specifically, we delete the ASPP modules and only use the cross-entropy loss to verify the functions of ASPP modules and SCE loss. All the detailed results have been shown in Table 6.1.

Methods	Forest	Shrubland	Grassland	Wetland	Cropland	Urban	Barren	Water	AA(%)
UNet (Band Concatenation)[54]	58.53	**6.25**	61.13	3.87	54.25	78.98	0.00	94.58	44.70
DeeplabV3+ (Band Concatenation)[55]	54.92	2.29	33.62	**7.59**	**82.31**	64.63	**5.45**	96.12	43.38
SOFNet (Without ASPP)	74.03	0.07	65.09	0.00	68.79	61.16	0.03	92.84	45.25
SOFNet (Cross Entropy)	71.84	0.08	66.31	0.00	68.15	72.33	0.03	94.56	46.66
SOFNet (Symmetric Cross Esntropy)	**84.83**	0.00	**70.90**	0.00	40.12	**82.31**	0.00	**98.04**	**47.03**

Table 6.1: Average accuracies on the DFC 2020 validation dataset for different methods (Proposed: SOFNet).

These values indicate that SOFNet provides the best AA among all the networks, about 2.33% higher and 3.65% higher than UNet and DeeplabV3 networks, respectively. One possible explanation is that the simple concatenation of the SAR and optical images as the input of the network does not make full use of the relationship between multi-modal data, and it may introduce redundant features during the training phase.

Without ASPP modules, the final AA result is 1.78% lower. This verifies that the ASPP module learns spatial context information of feature maps in the specific layer, which can help in the following decoding process. Meanwhile, SCE loss works better than CE loss, which proves that the SCE loss is more suitable in noisy label cases.

Example visualization results of SOFNet are shown in Figure 6.4. In general, we can see that SOFNet is able to predict acceptable land cover classification results. Sometimes the MODIS labels are very coarse, or even missing, on target areas, but SOFNet is still able to generate fine prediction maps. This provides a possible application that SOFNet can be used as a tool to complement and adjust the automatically derived MODIS labels for land cover classification.

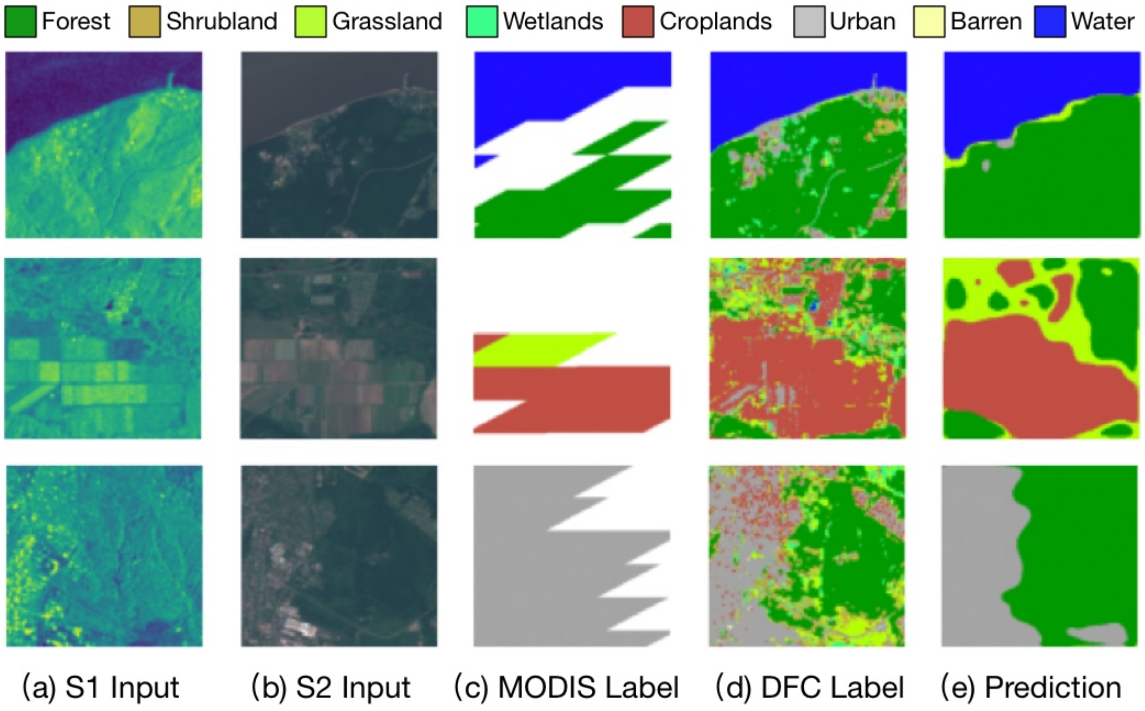


Figure 6.4: Visualization examples of SOFNet.

6.3.4 Discussion

The experiment results of SOFNet validate the effectiveness and potential of the multi-modal deep learning model in land cover classification. SCE loss also proved effective in this framework. However, comparing the SOFNet predictions with DFC2020 high-resolution labels in Figure 6.4, we can find that the contour lines of predictions are very vague, and that many details are ignored by the model. One possible solution is to supplement detailed information to the decoding branch through skip connections. For the SEN12MS dataset, under the premise of low-resolution and rough MODIS labels, rich texture information extracted from SAR and optical high-resolution input images become even more important. If we discard the low-level features learned from input images, we can no longer learn such information from corresponding low-resolution labels.

For the SOFNet model, we try to fuse feature maps of the first and second convolution layers (from SAR encoding branch, optical encoding branch, or both of them) before decoding, but it is of no help to the final performance. This may be because the semantic information obtained through the ResNet backbone and ASPP modules are relatively abstract. At the same time, adding very low-level features will not only bring little help but will introduce a lot of noise. Therefore, we consider another state-of-the-art segmentation network UNet. The UNet model has a simple structure and takes into account both low-level and high-level information. Hence, we then design a new UNet-based model SOF-UNet for land cover classification.

6.4 SAR-Optical Fusion UNet

6.4.1 System Architecture

6.4.1.1 The Overall Architecture

The architecture of SOF-UNet is illustrated in Figure 6.5. It also belongs to the EncoderDecoder segmentation structure. The network consists of three parts: two encoders to extract features, one decoder to restore the resolution, and specially designed skip connections for cross-modal feature extraction and aggregation. The design of feature fusion skip connections is inspired by FuseSeg [166] and RFNet [167]. Both of them are used for road driving segmentation applications fusing LiDAR and RGB data.

6.4.1.2 The Encoders

The optical and SAR encoder branches are almost the same except for the input dimension (13 channels and 2 channels, respectively). The low-level

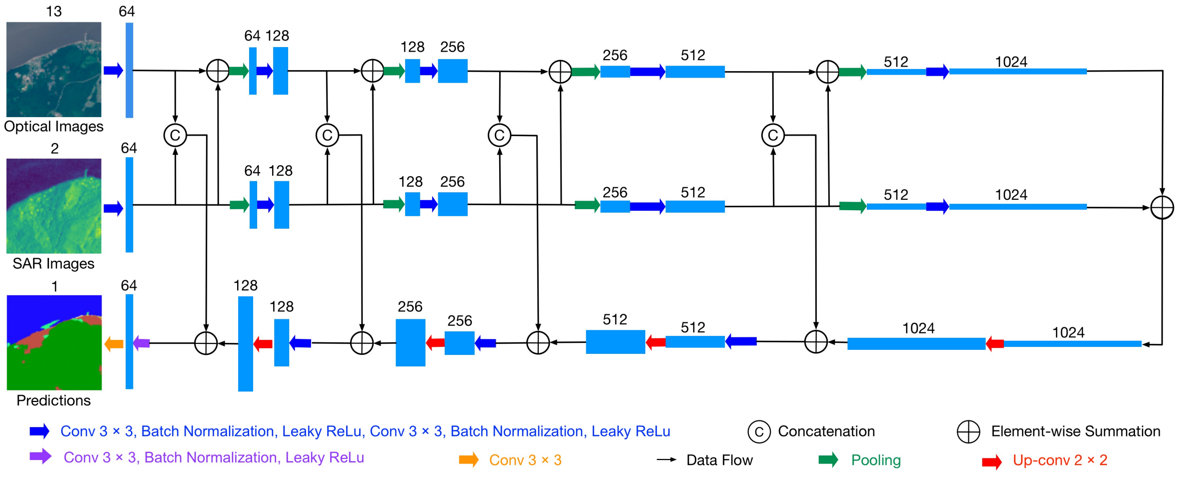


Figure 6.5: Illustration of the overall architecture of SOF-UNet.

SAR feature maps contain rich contour and location information. We add them to the RGB encoding branch to make the network focus on learning more complementary features. After fusing feature maps five times, we get high-level semantic information for decoding. In each encoding module, every convolution layer is followed by batch normalization and the nonlinear activation function Leaky-Rectified Linear Unit (Leaky ReLU).

6.4.1.3 The Decoder

Four simple upsampling modules with skip connections are adopted in the decoding phase. The ladder-style up-sampling modules have two inputs: the concatenated SAR and low-resolution optical features and the semantic features from an earlier layer of the encoder. Pixel-wise addition fusion methods are adopted here to fuse these two inputs. Finally, we blend them with two 3×3 convolutions. We also tried to concatenate SAR low-level features, low-level optical features, and upsampled semantic features, but the results turned out to be worse than the element-wise summation, so we dropped this setting.

6.4.2 Experiments

6.4.2.1 Implementation Details

We conduct our experiments on a GeForce GTX 1080 Ti GPU. In the training stage, we set the batch size to 8 and the initial learning rate to 0.0001. For the other configurations, we keep the same with SOFNet including SCE parameters. It takes around 3 hours and 30 minutes to finish one epoch under this setting. We run the model for 20 epochs and choose the best model on the validation dataset to test.

6.4.2.2 Quantitative Results

For a fair comparison, we use the same SCE loss function and evaluation metric with SOFNet. Table 6.2 displays the quantitative results for the comparison. SOF-UNet yields the best results among the different segmentation networks. We can see that our SOF-UNet outperforms SOFNet by around 1.20% in terms of AA, although the performance on the forest, grassland, and urban class drops a bit. The SOF-UNet tends to be not so "biased" because of the "zero" accuracy for shrubland, wetland, and barren disappearances. One possible explanation is that low-level texture information learned in SOF-UNet is very important to distinguish these three classes. The accuracy of water classification increases to 98.89%. This is probably the reason why the water class does not need so much semantic information as the urban class. More basic textural information in SOF-UNet could be helpful.

For the ablation of our two-stream architecture, we also compare SOF-UNet with UNet. We test UNet with three different inputs: only 2-channel Sentinel-1 SAR images (Only S1), only 13-channel Sentinel-2 multi-spectral images (Only S2), and 15-channel multi-modal concatenation images (Band Concatenation). Initially, we found that the Band Concatenation even results in a worse performance than Only S2. But the SOF-UNet outperforms the Only S2 result, which verifies the effectiveness of the two-stream architecture.

Methods	Forest	Shrubland	Grassland	Wetland	Cropland	Urban	Barren	Water	AA(%)
Unet (Only S1)	90.68	0.00	0.17	0.00	77.37	71.45	2.15	92.65	41.81
Unet (Only S2)	73.01	0.82	**77.03**	4.92	44.38	82.19	0.00	90.05	46.55
Unet (Band Concatenation)	58.53	**6.25**	61.13	3.87	54.25	78.98	0.00	94.58	44.70
DeeplabV3 Plus (Band Concatenation)	54.92	2.29	33.62	**7.59**	**82.31**	64.63	5.45	96.12	43.38
SOFNet (Symmetric Cross Entropy)	**84.83**	0.00	70.90	0.00	40.12	**82.31**	0.00	98.04	47.03
SOF-UNet (Without Encoding Skip)	70.51	0.81	64.05	0.17	50.61	70.81	11.43	97.68	45.76
SOF-UNet (Without Decoding Skip)	65.23	1.56	66.91	0.23	46.78	70.96	12.06	97.94	45.21
SOF-UNet (Cross Entropy)	73.92	0.71	67.67	0.25	49.92	74.75	11.51	97.97	47.09
SOF-UNet (Symmetric Cross Entropy)	75.52	0.97	69.14	0.19	50.38	77.64	**13.10**	**98.89**	**48.23**

Table 6.2: Average accuracies on the DFC 2020 validation dataset for different methods (Proposed: SOF-UNet).

For the ablation of the fusion strategy, we delete the fusion skip connections in the encoding and decoding phases separately. Both of the two variants provide lower performance. This indicates that multi-modal fusion plays a significant role in our model. Meanwhile, SCE loss also shows to be a superior choice here with regard to SOFNet.

6.4.2.3 Visualization Results

We show some acceptable results of SOFNet and SOF-UNet in Figure 6.6. In general, SOF-UNet predictions retain more texture information like edge and boundary compared with SOFNet. They are closer to DFC ground truth images to some extent. As shown in the last row, SOF-UNet has a better ability to identify the barren class. Sometimes this class is even identified as grassland in low-resolution training labels.

However, both SOFNet and SOF-UNet provide bad predictions in some cases. Some examples are given in Figure 6.7. For wetland, barren, and cropland pixels, most of them tend to be classified as grassland and shrubland in SOFNet and SOF-UNet, respectively. For the SOFNet model, it may be due to the overfitting of noisy MODIS labels. For the SOF-UNet model, a possible reason could be the high textural similarity between wetland, barren, cropland, and shrubland classes.

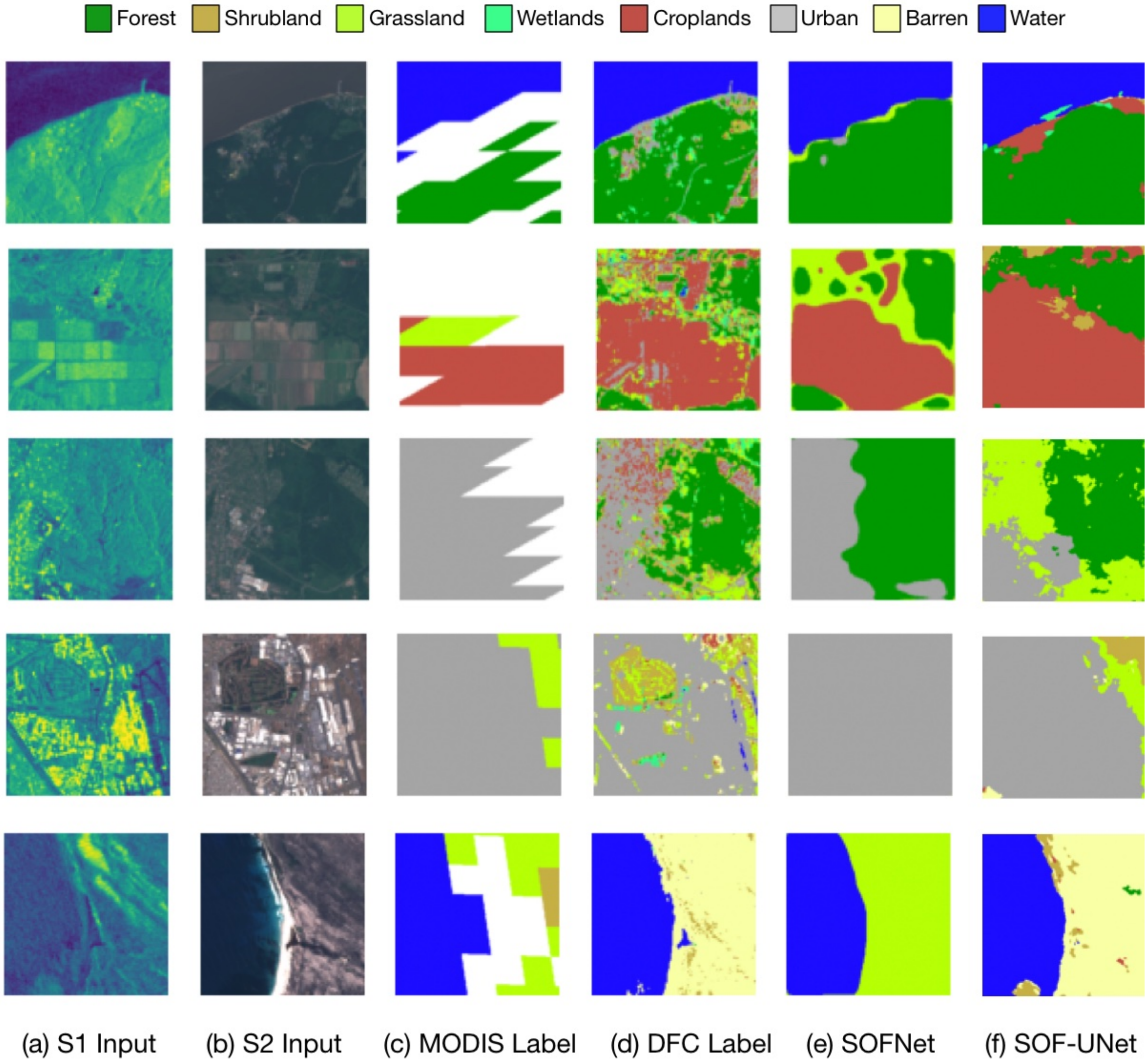
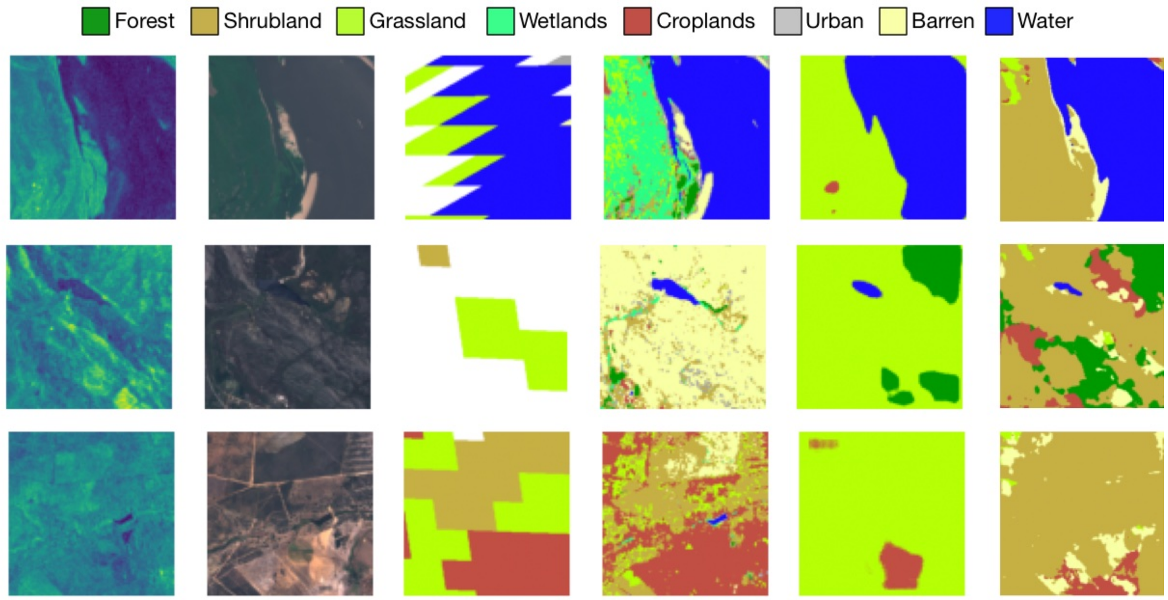


Figure 6.6: Some acceptable prediction examples of SOFNet and SOF-UNet on the DFC2020 validation dataset.

6.5 Summary

In this chapter, we present a SOF-UNet framework for land cover classification of SAR and Optical data. SOF-UNet adopts a two-stream encoder-decoder deep learning segmentation framework. The experiment results show that this design has a promising capability to identify



(a) S1 Input (b) S2 Input (c) MODIS Label (d) DFC Label (e) SOFNet (f) SOF-UNet

Figure 6.7: Some unacceptable prediction examples of SOFNet and SOF-UNet on the DFC2020 validation dataset.

different land cover classes. The SCE loss is also verified useful in this frame. Since both SAR and optical images contain rich textural features, it is often helpful to use the skip connection operation to preserve structure and boundary information in very low-level features.

Our results show clear improvements over the baseline approaches. However, some weaknesses remain, since we mainly focus on multi-modal fusion and haven't applied any label refinement on noisy labels using traditional methods or state-of-the-art weakly-supervised semantic segmentation methods. These methods have been proved useful in the 2020 IEEE GRSS Data Fusion Contest [164]. Future experiments can be conducted to combine our multi-modal fusion network and weakly supervised semantic segmentation methods together to achieve better performance.

Chapter 7

Discussion and Future Work

In recent years, the research and development of SAR image interpretation based on deep learning technology have received wide attention. This thesis concentrates on designing deep learning models applied to three specific SAR image interpretation tasks. They are oceanic eddy detection, intertidal sediments and habitats classification, and land cover classification. In this chapter, the main contributions in this thesis are summarized in Section 9.1. We conclude the inspirations and limitations in Section 9.2. Finally, three future research directions for SAR image interpretation are given in Section 9.3.

7.1 Summary

This thesis has investigated the applicability of deep learning for SAR image interpretation. To this end, three end-to-end deep learning models are proposed to address three specific applications in the wide framework of SAR image interpretation. These three models have both theoretical significance and practical value in the SAR interpretation field. The main contributions are listed as below:

1. The first contribution is the Mask-ES-RCNN model for automatic SAR oceanic eddy detection. We first constructed the SOEDD to develop deep learning-based SAR eddy detection methods. SOEDD

was firstly verified with respect to its ability and potential to achieve acceptable eddy detection results using our Mask RCNN and Edge Enhancement model. Inspired by this model, a Mask-ES-RCNN framework with two additional Edge Head and Mask IoU Heads was proposed. This multi-task learning strategy made the model focus on internal texture information of eddy instances and the qualities of predicted masks at the same time. Mask-ES-RCNN outperformed the Mask RCNN baseline on SOEDD in terms of all APs. The experimental results verified the importance to incorporate prior knowledge in deep learning models, especially for small-scale SAR datasets.

2. The second contribution is the TE-UNet model for SAR intertidal sediments and habitats classification. We proposed a UNet based model with a TEM to explicitly learn global texture information from SAR images. The experimental results proved the superiority of the TE-UNet model compared with the state-of-art semantic segmentation models. Meanwhile, comprehensive ablation studies verified the effectiveness and necessity to utilize multiple frequencies and polarimetric information for classification tasks in the intertidal zone.

3. The third contribution is the SOF-UNet model for land cover classification of SAR and optical data. Based on the existing largest dataset SEN12MS for SAR-optical fusion land cover classification, the first proposed SOFNet model obtained preliminary classification results. Inspired by the SOFNet model, we further developed a SOF-UNet to utilize more low-level features. SOF-UNet used a two-stream encoder-decoder framework with specially designed skip connections for multi-modal features extraction and aggregation. Compared with the SOFNet, SOF-UNet could retain more details in the prediction maps and obtain better results in terms of AA. The SCE loss was also verified useful in this frame.

7.2 Conclusion

A distinct advantage of end-to-end deep learning-based methods is that they can obtain SAR image interpretation results automatically and efficiently in comparison with the traditional methods. Concerning our proposed deep learning-based models in this thesis, we mainly get the following inspirations:

1. Multi-source inputs always tend to be helpful for final SAR image interpretation results, such as multi-band SAR images, PolSAR images, and SAR-optical pairs. Specific SAR image interpretation models should be designed according to different data resources.

2. It is worth noting that we consider incorporating the texture information of SAR images into proposed deep learning models implicitly or explicitly. In particular, we adopt Edge Head for Mask-ES-RCNN, TEM for TE-UNet, and designed skip connections for SOF-UNet. This design concept, which is verified to be useful in our proposed models, can be further adopted for other SAR image interpretation applications.

However, the existing models are not always very reliable on processing tasks. There are also some limitations of our systems:

1. The first limitation is the small-scale SAR image datasets. For instance, we train and test our Mask-ES-RCNN model on SOEDD which only contains 200 images with 322 eddy instances. Even if we could collect large-scale original SAR images, it is still challenging to label all of them, because of their complex properties. The small-scale size of the labeled SAR dataset limits to some extent the accuracy of the interpretation results.

2. Another problem is that we utilize limited information of SAR image characteristics in our models. For example, we use CP polarimetric decomposition components as input in TE-UNet to express polarimetric information of SAR images. This transformation process will

lose some information compared with the original coherence matrix of SAR data.

3. For our models, we do not make full use of the temporal nature of SAR images. This information is very important for SAR image interpretation tasks, especially in highly dynamic environments. Let's take the intertidal zone as an example. Different times correspond to different tides and thus present very different surface types on SAR images [130]. Hence, the model's performance can be damaged if we ignore the surface changes on SAR images at different times.

7.3 Future Work

Based on the results of the present research, there are many potential ways to improve SAR image interpretation. We mainly discuss three future research directions in this section: Complex-Valued Convolutional Neural Network (CV-CNN), spatio-temporal combination, and multimodal fusion of more modalities. Detailed descriptions are found in the following sections.

7.3.1 Complex-Valued Convolutional Neural Networks

The first direction is using CV-CNN [168] to deal with the complex-valued coherence matrix of SAR data. Unlike the optical remote sensing images, SAR images contain both amplitude and phase information. In Section 5.3.2 we use polarimetric decomposition algorithms to extract information from pre-processed SAR images. This is still not a complete end-to-end method and some information may be lost during the preprocessing progress. CV-CNN can take advantage of phase information and realize SAR image interpretation tasks with less manual intervention. Notably, specific SAR image characteristics like speckle noise and abundant texture features could be taken into consideration when we design the CV-CNN-based interpretation models.

The CV-CNN replaces each convolutional layer in the network with four real-valued convolutions that correspond to the individual real and imaginary parts of the complex tensor and weight kernel [168]. Figure 7.1 illustrates the complex-valued convolutions structure.

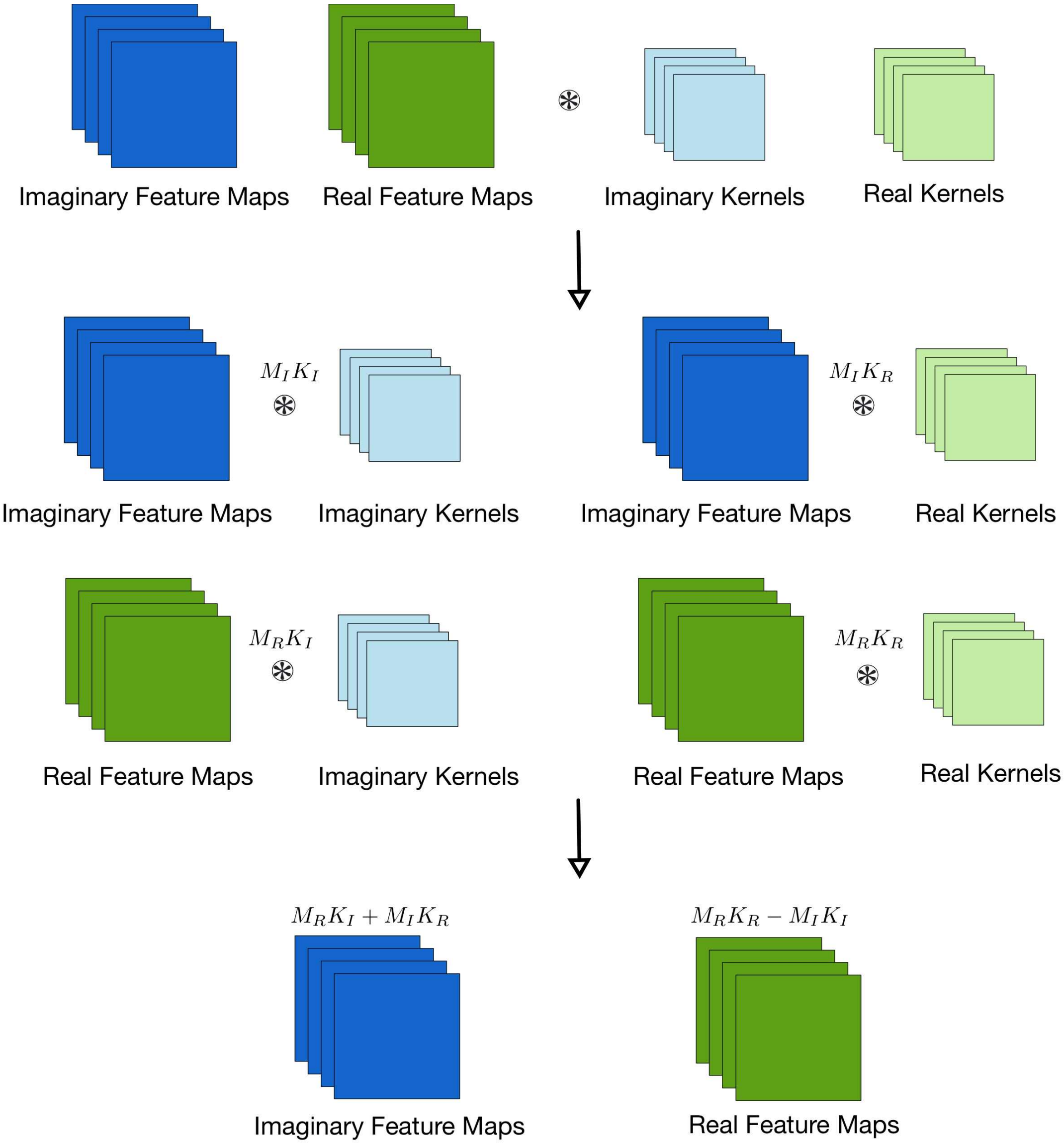


Figure 7.1: Illustration of complex-valued convolutions. Image adapted from [168].

If a complex kernel matrix $K = K_R + iK_I$ is convolved with complex data $M = M_R + iM_I$, the result can be expressed as:

$$\begin{aligned} K * M &= (K_R + iK_I) * (M_R + iM_I) \\ &= K_R * M_R + i^2 K_I * M_I + K_R * iM_I + iK_I * M_R \\ &= K_R * M_R - K_I * M_I + i\,(K_R * M_I + K_I * M_R) \end{aligned} \tag{7.1}$$

where $*$ stands for the convolution operation, and K_R, K_I, M_R, and M_I are real-valued matrices. Correspondingly, other mathematical operations of CV-CNN including upsampling, magnitude operation, and softmax operation are also defined within the complex-valued domain.

7.3.2 Spatio-Temporal Combination

The second direction is the combination of spatial and temporal information of SAR images. As previously mentioned, SAR is an all-day all-weather sensor, capable of capturing spatial information through temporal backscatter. This could give the potential of fostering a deeper understanding of how factors such as seasonality and environmental changes can influence the targets. Moreover, the effective use of frequency domain information may help obtain better SAR image interpretation results.

Three-dimensional 3D-CNN is a meaningful alternative to the 2D-CNN in spatio-temporal analyses without collapsing the temporal dimension. Figure 7.2 shows the comparison of 2D-CNN and 3D-CNN convolutions. Compared with 2D-CNN, the input of the 3D-CNN is a cube stacked with multiple feature maps, which can extract features at three scales at the same time [41]. Formally, the computation of 3D convolution is described as:

$$O^{(x,y,z)} = \sum_{c=1}^{C} \sum_{k_t=1}^{K_t} \sum_{k_x=1}^{K_x} \sum_{k_y=1}^{K_y} W_m^{(t,k_x,k_y)} * I_c^{(z+t,x+k_x,y+k_y)} \tag{7.2}$$

where x, y and z are the spatial coordinates in an output cube. The

k_x, k_y and k_t are the kernel size in width, height and temporal dimensions, respectively. C is the number of total channels.

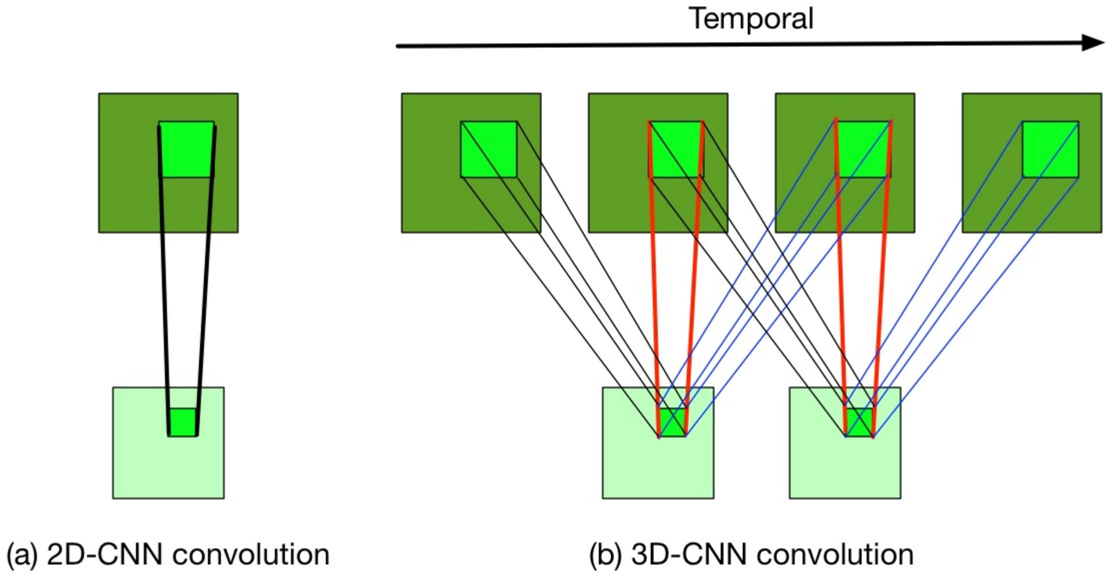


Figure 7.2: Comparison of (a) 2D-CNN and (b) 3D-CNN convolution. In (b), the size of the convolution kernel in the temporal dimension is 3, and the sets of connections are color-coded, so that the shared weights are in the same color. Image adapted from [41].

7.3.3 Multi-Modal Fusion of More Modalities

The third direction is multi-modal fusion of more modalities. We used a two-stream way with designed skip connections to fuse SAR and optical data in Section 6.4. More complex fusion strategies can be further developed on existing multi-modal data, incorporating multi-band and multi-polarization SAR images. The multi-modal fusion models can also be applied in the framework of CV-CNN and 3D-CNN.

Besides, we can also introduce more modalities and integrate them with SAR images based on specific applications. For example, looking back on our oceanic eddy detection task, surface wind speed is one of the key variables to find the eddies. Too low local surface wind speed values lead to low radar backscatter of SAR images, which makes “black” eddies less detectable. Too high wind speed values also cause the disappearance of “black” eddies, because the surface films disappear from the sea surfaces. Thus, the wind speed information can help detect oceanic eddies faster and more accurately on SAR images.

In general, if the aforementioned methods are integrated into our deep learning models, it is not too early to expect they can gradually improve the SAR image interpretation results. Eventually, with the ongoing development of this interdisciplinary research, we can bring the manner of global monitoring to the next level, where, possibly by augmenting with other real-time analyzing techniques with respect to climate change, animal migration, natural disaster warning and etc., the earth’s resources are better coordinated in terms of utilization as well as protection.

Appendix A

Nomenclature

Abbreviations

3D-CNN	3-Dimensional Convolutional Neural Network
2D-CNN	2-Dimensional Convolutional Neural Network
AA	Average Precision
ALOS2	ALOS PALSAR-2
ASPP	Atrous Spatial Pyramid Pooling
CAM	Class Activation Mapping
CHL	Chlorophyll
CNN	Convolutional Neural Network
CNSA	China National Space Administration
CP	Cloude Pottier
CSA	Canadian Space Agency
CTD	Coherent Target Decomposition
CV-CNN	Complex-Valued Convolutional Neural Network
DLR	German Aerospace Center
DN	Digital Number
DNN	Deep Neural Network

DSCNN	Deep Supervised and Contractive Neural Network
EM	Electromagnetic
ESA	European Space Agency
FCN	Fully Convolutional Network
FD	Freeman Durden
FPN	Feature Pyramid Network
GAN	Generative Adversarial Network
GCN	Graph Convolutional Network
GRD	Ground Range Detected
Leaky ReLU	Leaky-Rectified Linear Unit
NN	Neural Network
OA	Overall Accuracy
OLI	Operational Land Imager
PCA	Principal Component Analysis
RNN	Recurrent Neural Network
PolSAR	Polarimetric Synthetic Aperture Radar
SAR	Synthetic Aperture Radar
SCE	Symmetric Cross Entropy
SNAP	Sentinel Application Platform
SOEDD	SAR Oceanic Eddy Detection Dataset
TIRS	Thermal Infrared Sensor
ICTD	In-Coherent Target Decomposition
InSAR	Interferometric Synthetic Aperture Radar
IoU	Intersection over Union
IW	Interferometric Wide
JAXA	Japan Aerospace Exploration Agency
mAP	Mean Average Precision

mF1	Mean F1 score
mIoU	Mean Intersection over Union
METI	Ministry of Economy Trade and Industry
NASA	National Aeronautics and Space Administration
RCNN	Region Convolutional Neural Network
RF	Random Forest
ROI	Region of Interest
RPN	Region Proposal Network
RS2	Radarsat-2
SAE	Stacked Sparse Autoencoder
SCNN	Spatial Convolutional Neural Network
SLC	Single-Look Complex
SNL	Sandia National Laboratory
SST	Sea Surface Temperature
SSH	Sea Surface Height
SSW	Sea Surface Wind
SOEDD	SAR Oceanic Eddy Detection Dataset
SVM	Support Vector Machine
SIR-C/X-SAR	Spaceborne Imaging Radar-C/X SAR
TEM	Texture Enhance Module
USGS	United States Geological Survey
UTC	Universal Time Coordinated

Appendix B

Publications Originating from this Thesis

B.1 Conferences

Di Zhang, Martin Gade, Jianwei Zhang. Eddy Detection on SAR Images Based on Faster R-CNN, TerraSAR-X/TanDEM-X Science Team Meeting, 2019 (Published, Poster)

Di Zhang, Martin Gade, Jianwei Zhang. SAR Eddy Detection Using Mask-RCNN and Edge Enhancement, IEEE International Geoscience and Remote Sensing Symposium (IGARSS), 2020 (Published, Oral)

Di Zhang, Martin Gade, Jianwei Zhang. SOFNet: SAR-Optical Fusion Network for Land Cover Classification, IEEE International Geoscience and Remote Sensing Symposium (IGARSS), 2021 (Published, Oral)

Di Zhang, Martin Gade, Jianwei Zhang. SOF-UNet: SAR and Optical Fusion UNet for Land Cover Classification, IEEE International Geoscience and Remote Sensing Symposium (IGARSS), 2022 (Submitted)

B.2 Journal Articles

Di Zhang, Martin Gade, Jianwei Zhang. Mask-ES-RCNN: Mask Edge Enhancement and IoU Score RCNN for Oceanic Eddy Detection on SAR Images, Remote Sensing, 2022 (Submitted)

Di Zhang, Martin Gade, Jianwei Zhang. TE-UNet: Texture Enhancement UNet on Multi-band PolSAR Images for Intertidal Sediments and Habitats Classification, Remote Sensing, 2022 (Submitted)

Appendix C

Acknowledgements

This thesis is the outcome of my research during the last four years at the Technical Aspects of Multimodal Systems (TAMS) group. Along this journey, I was accompanied by excellent people to whom I would like to express my sincere gratitude.

First, I would like to thank my supervisors, Prof. Dr. Jianwei Zhang and Dr. Martin Gade, who provided me with the opportunity to start an interdisciplinary doctoral program. Jianwei has great personal charisma and was always ready to help, whose expertise was invaluable in formulating the research questions and methodology. I heartily appreciate your guidance, patience, and encouragement during my whole Ph.D. period. I am also grateful for Martin's efforts, time, and valuable comments. He helped and witnessed the entire process of my academic growth. Thank you for always being there when I face a dilemma in life and research. Thanks to Prof. Dr. Junqiang Song and Prof. Dr. Kaijun Ren for introducing this exciting topic and trusting my abilities to contribute to this field. I am also grateful to Prof. Dr. Simone Frintrop, who kindly reviewed this thesis.

Second, I want to thank Dr. Norman Hendrich helped me with constructive suggestions. I was impressed by his bright scientific mind and huge enthusiasm for research. I would like to thank Dr. Chao Zeng, Shuang Li, and Michael Görner for helping me proofread my thesis and providing valuable feedback. Special thanks to Tatjana Lu Tetsis for

enlightening me when I fell in the blues. You were like a warm family member accompanied by my side. I want to thank other colleagues in the TAMS group: Hongzhuo Liang, Jianzhi Lyu, Lin Cong, Yuyang Tu, Wenkai Chen, Shangqing Liu, Chaobin Zou, Yunlei Shi, Dr. Andreas Mäder, Dr. Florens Wasserfall, Philipp Ruppel, Niklas Fiedler,Yannick Jonetzko, Marc Bestmann, Bernd Schütz, Daniel Ahlers, Dr. Yihong Hu, Wiebke Noeske, German Junca, Dr. Dan Liu, Dr. Ge Gao etc. Your existence shined my work and life in Hamburg.

Third, I am grateful to Rudolf-Laun-Haus roommates for making my daily life more colorful. Special thanks to Melina Eva Joana Hessel. The time spent with you and your family would be one of my most precious memories in Germany.

Last but not least, I want to thank my beloved parents. Although you are far away in China, you have always given me the greatest understanding and support. To be your daughter is the luckiest and happiest thing in my life. I also want to thank my grandfather in heaven. I could not see you for the last time because of the pandemic. But I will live with your endless love and support.

References

[1] Armin W. Doerry and Fred M. Dickey. “Synthetic Aperture Radar”. In: *Optics and Photonics News* 15.11 (Nov. 1, 2004), p. 28. ISSN: 1047-6938, 1541-3721. DOI: 10.1364/OPN.15.11.000028.

[2] Jixian Zhang et al. “SAR Mapping Technology and Its Application in Difficulty Terrain Area”. In: *IEEE International Geoscience and Remote Sensing Symposium (IGARSS)*. IEEE. 2010, pp. 3608–3611.

[3] Tazio Strozzi et al. “JERS SAR Interferometry for Land Subsidence Monitoring”. In: *IEEE Transactions on Geoscience and Remote Sensing* 41.7 (2003), pp. 1702–1708.

[4] Ayman Abdel-Hamid et al. “Assessing the Impact of Drought Stress on Grasslands Using Multi-Temporal SAR Data of Sentinel-1: A Case Study in Eastern Cape, South Africa”. In: *European Journal of Remote Sensing* (July 13, 2020), pp. 3–16. ISSN: 2279-7254. DOI: 10.1080/22797254.2020.1762514.

[5] Tang Kan et al. “Implementation of Real-Time Automotive SAR Imaging”. In: *IEEE Sensor Array and Multichannel Signal Processing Workshop (SAM)*. 2020, pp. 1–4.

[6] Patrick Berens. *Introduction to Synthetic Aperture Radar (SAR)*. NATO OTAN, 2006.

[7] Ian Goodfellow, Yoshua Bengio, and Aaron Courville. *Deep Learning*. MIT Press, 2016.

[8] Eric Mason, Bariscan Yonel, and Birsen Yazici. “Deep Learning for SAR Image Formation”. In: *Algorithms for Synthetic Aperture Radar Imagery XXIV*. Ed. by Edmund Zelnio and Frederick D. Garber. Vol. 10201. 2017, p. 4. DOI: 10.1117/12.2267831.

[9] Xiaoxiang Zhu et al. “Deep Learning Meets SAR: Concepts, Models, Pitfalls, and Perspectives”. In: *IEEE Geoscience and Remote Sensing Magazine* (2021). DOI: 10.1109/MGRS.2020.3046356.

[10] Sepp Hochreiter and Jürgen Schmidhuber. “Long Short-Term Memory”. In: *Neural Computation* 9.8 (Nov. 1, 1997), pp. 1735–1780. DOI: 10.1162/neco.1997.9.8.1735.

[11] Alex Krizhevsky, Ilya Sutskever, and Geoffrey E Hinton. “ImageNet Classification with Deep Convolutional Neural Networks”. In: *Advances in Neural Information Processing Systems* 25 (2012).

[12] Lei Ma et al. “Deep Learning in Remote Sensing Applications: A Meta-Analysis and Review”. In: *ISPRS Journal of Photogrammetry and Remote Sensing* 152 (June 2019), pp. 166–177. ISSN: 0924-2716. DOI: 10.1016/j.isprsjprs.2019.04.015.

[13] Pramila P. Shinde and Seema Shah. “A Review of Machine Learning and Deep Learning Applications”. In: *International Conference on Computing Communication Control and Automation (ICCUBEA).* Aug. 2018, pp. 1–6. ISBN: 978-1-5386-5257-2. DOI: 10.1109/ICCUBEA.2018.8697857.

[14] Christina Corbane et al. “Convolutional Neural Networks for Global Human Settlements Mapping from Sentinel-2 Satellite Imagery”. In: *Neural Computing and Applications* 33.12 (June 2021), pp. 6697–6720. ISSN: 0941-0643. DOI: 10.1007/s00521-020-05449-7.

[15] Rajat Garg et al. “Semantic Segmentation of PolSAR Image Data Using Advanced Deep Learning Model”. In: *Scientific Reports* 11.1 (Dec. 2021), p. 15365. ISSN: 2045-2322. DOI: 10.1038/s41598-021-94422-y.

[16] Jonathan Schmidt et al. “Recent Advances and Applications of Machine Learning in Solid-State Materials Science”. In: *npj Computational Materials* 5.1 (Dec. 2019), p. 83. ISSN: 2057-3960. DOI: 10.1038/s41524-019-0221-0.

[17] Francesco Lattari et al. “Deep Learning for SAR Image Despeckling”. In: *Remote Sensing* 11.13 (June 28, 2019), p. 1532. ISSN: 2072-4292. DOI: 10.3390/rs11131532.

[18] Ravi P. Gupta. “Interpretation of SAR Imagery”. In: *Remote Sensing Geology.* Springer, 2018, pp. 235–252. ISBN: 978-3-662-55874-4. DOI: 10.1007/978-3-662-55876-8_16.

[19] M. Schmitt et al. “SEN12MS – A Curated Dataset of Georeferenced Multi-Spectral Sentinel-1/2 Imagery for Deep Learning and Data Fusion”. In: *ISPRS Annals of the Photogrammetry, Remote Sensing and Spatial Information Sciences* IV-2/W7 (Sept. 16, 2019), pp. 153–160. ISSN: 2194-9050. DOI: 10.5194/isprs-annals-IV-2-W7-153-2019.

[20] James B Campbell and Randolph H Wynne. *Introduction to Remote Sensing.* Guilford Press, 2011.

[21] J Everaerts. "PEGASUS—Bridging the Gap between Airborne and Spaceborne Remote Sensing". In: *New Strategies For European Remote Sensing* (2005), pp. 395–401.

[22] Lloyd Haydn Hughes. "Deep Learning for Matching High-Resolution SAR and Optical Imagery". The Technical University of Munich, 2020.

[23] Dioline Sara et al. "Hyperspectral and Multispectral Image Fusion Techniques for High Resolution Applications: A Review". In: *Earth Science Informatics* 14.4 (Dec. 2021), pp. 1685–1705. ISSN: 1865-0473. DOI: 10.1007/s12145-021-00621-6.

[24] *USGS Landsat Program. Approximate Spectral Bands Comparison Chart for Landsat, MODIS, ASTER and Sentinel2.* URL: https://twitter.com/usgslandsat/status/773939936755982336.

[25] J. A. Richards. *Remote Sensing with Imaging Radar.* Vol. 1. Signals and Communication Technology. Springer, 2009. ISBN: 978-3-642-02020-9.

[26] Charles Elachi and Jakob Van Zyl. *Introduction to the Physics and Techniques of Remote Sensing.* Wiley-Interscience, 2006. ISBN: 978-0-471-78339-8.

[27] Jong-Sen Lee and Eric Pottier. *Polarimetric Radar Imaging: From Basics to Applications.* 142. CRC Press, 2009. ISBN: 978-1-4200-5497-2.

[28] S.R. Cloude and E. Pottier. "A Review of Target Decomposition Theorems in Radar Polarimetry". In: *IEEE Transactions on Geoscience and Remote Sensing* 34.2 (Mar. 1996), pp. 498–518. ISSN: 0196-2892. DOI: 10.1109/36.485127.

[29] A. Freeman and S.L. Durden. "A Three-Component Scattering Model for Polarimetric SAR Data". In: *IEEE Transactions on Geoscience and Remote Sensing* 36.3 (May 1998), pp. 963–973. ISSN: 0196-2892. DOI: 10.1109/36.673687.

[30] S.R. Cloude and E. Pottier. "An Entropy Based Classification Scheme for Land Applications of Polarimetric SAR". In: *IEEE Transactions on Geoscience and Remote Sensing* 35.1 (Jan. 1997), pp. 68–78. ISSN: 0196-2892. DOI: 10.1109/36.551935.

[31] Xiaohong Chen, Qian Sun, and Jun Hu. "Generation of Complete SAR Geometric Distortion Maps Based on DEM and Neighbor Gradient Algorithm". In: *Applied Sciences* 8.11 (Nov. 9, 2018), p. 2206. ISSN: 2076-3417. DOI: 10.3390/app8112206.

[32] Africa Flores et al. *Synthetic Aperture Radar (SAR) Handbook: Comprehensive Methodologies for Forest Monitoring and Biomass Estimation.* NASA, 2019. URL: https://gis1.servirglobal.net/TrainingMaterials/SAR/SARHB_FullRes.pdf.

[33] Alenrex Maity et al. “A Comparative Study on Approaches to Speckle Noise Reduction in Images”. In: *International Conference on Computational Intelligence and Networks (CINE)*. Jan. 2015, pp. 148–155. ISBN: 978-1-4799-7548-8. DOI: 10.1109/CINE.2015.36.

[34] Y. LeCun et al. “Backpropagation Applied to Handwritten Zip Code Recognition”. In: *Neural Computation* 1.4 (Dec. 1989), pp. 541–551. ISSN: 0899-7667, 1530-888X. DOI: 10.1162/neco.1989.1.4.541.

[35] Thomas N. Kipf and Max Welling. “Semi-Supervised Classification with Graph Convolutional Networks”. In: *International Conference on Learning Representations (ICLR)* (2017).

[36] Alex Sherstinsky. “Fundamentals of Recurrent Neural Network (RNN) and Long Short-Term Memory (LSTM) Network”. In: *Physica D: Nonlinear Phenomena* 404 (Mar. 2020), p. 132306. ISSN: 0167-2789. DOI: 10.1016/j.physd.2019.132306.

[37] Ian J. Goodfellow et al. “Generative Adversarial Networks”. In: *Association for Computing Machinery* (2020), pp. 139–144. ISSN: 0001-0782. DOI: 10.1145/3422622.

[38] Karen Simonyan and Andrew Zisserman. “Very Deep Convolutional Networks for Large-Scale Image Recognition”. In: *International Conference on Learning Representations (ICLR)* (2015).

[39] Christian Szegedy et al. “Going Deeper with Convolutions”. In: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. June 2015, pp. 1–9. ISBN: 978-1-4673-6964-0. DOI: 10.1109/CVPR.2015.7298594.

[40] Shuiwang Ji et al. “3D Convolutional Neural Networks for Human Action Recognition”. In: *IEEE Transactions on Pattern Analysis and Machine Intelligence* 35.1 (Jan. 2013), pp. 221–231. ISSN: 0162-8828. DOI: 10.1109/TPAMI.2012.59.

[41] Xingang Pan et al. “Spatial as Deep: Spatial CNN for Traffic Scene Understanding”. In: *Proceedings of the AAAI Conference on Artificial Intelligence (AAAI)*. 2018, p. 32.

[42] Yinsheng Su et al. “Multi-Task Transient Contingency Screening with Temporal Graph Convolutional Network in Power Systems”. In: *Journal of Physics: Conference Series* 2095.1 (Nov. 1, 2021), p. 012027. ISSN: 1742-6588. DOI: 10.1088/1742-6596/2095/1/012027.

[43] Zhong-Qiu Zhao et al. “Object Detection With Deep Learning: A Review”. In: *IEEE Transactions on Neural Networks and Learning Systems* 30.11 (Nov. 2019), pp. 3212–3232. ISSN: 2162-237X. DOI: 10.1109/TNNLS.2018.2876865.

[44] Ross Girshick et al. “Rich Feature Hierarchies for Accurate Object Detection and Semantic Segmentation”. In: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. 2014, pp. 580–587.

[45] Ross Girshick. “Fast R-CNN”. In: *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*. Sept. 27, 2015, pp. 1440–1448.

[46] Joseph Redmon et al. “You Only Look Once: Unified, Real-Time Object Detection”. In: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. June 2016, pp. 779–788. ISBN: 978-1-4673-8851-1. DOI: 10.1109/CVPR.2016.91.

[47] Shaoqing Ren et al. “Faster R-CNN: Towards Real-Time Object Detection with Region Proposal Networks”. In: *IEEE Transactions on Pattern Analysis and Machine Intelligence* 39.6 (June 1, 2017), pp. 1137–1149. ISSN: 0162-8828, 2160-9292. DOI: 10.1109/TPAMI.2016.2577031.

[48] Kaiming He et al. “Mask R-CNN”. In: *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*. 2017, pp. 2961–2969.

[49] Saeid Asgari Taghanaki et al. “Deep Semantic Segmentation of Natural and Medical Images: A Review”. In: *Artificial Intelligence Review* 54.1 (Jan. 2021), pp. 137–178. ISSN: 0269-2821, 1573-7462. DOI: 10.1007/s10462-020-09854-1.

[50] Jonathan Long, Evan Shelhamer, and Trevor Darrell. “Fully Convolutional Networks for Semantic Segmentation”. In: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. 2015, pp. 3431–3440.

[51] Huikai Wu et al. *FastFCN: Rethinking Dilated Convolution in the Backbone for Semantic Segmentation*. 2019. arXiv: 1903.11816 [cs]. URL: http://arxiv.org/abs/1903.11816.

[52] Hengshuang Zhao et al. “Pyramid Scene Parsing Network”. In: *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. July 2017, pp. 6230–6239. ISBN: 978-1-5386-0457-1. DOI: 10.1109/CVPR.2017.660.

[53] Liang-Chieh Chen et al. *Rethinking Atrous Convolution for Semantic Image Segmentation*. 2017. arXiv: 1706.05587 [cs]. URL: http://arxiv.org/abs/1706.05587.

[54] Olaf Ronneberger, Philipp Fischer, and Thomas Brox. “U-Net: Convolutional Networks for Biomedical Image Segmentation”. In: *International Conference on Medical Image Computing and Computer-Assisted Intervention (MICCAI)*. 2015, pp. 234–241.

[55] Liang-Chieh Chen et al. “Encoder-Decoder with Atrous Separable Convolution for Semantic Image Segmentation”. In: *Proceedings of the European conference on computer vision (ECCV)* (2018), pp. 801–818.

[56] Timothy D. Ross et al. "Standard SAR ATR Evaluation Experiments Using the MSTAR Public Release Data Set". In: Aerospace/Defense Sensing and Controls. Ed. by Edmund G. Zelnio. Sept. 15, 1998, pp. 566–573. DOI: 10.1117/12.321859.

[57] Sizhe Chen and Haipeng Wang. "SAR Target Recognition Based on Deep Learning". In: *International Conference on Data Science and Advanced Analytics (DSAA)*. Shanghai, China: IEEE, Oct. 2014, pp. 541–547. ISBN: 978-1-4799-6991-3. DOI: 10.1109/DSAA.2014.7058124.

[58] Haipeng Wang et al. "Application of Deep-Learning Algorithms to MSTAR Data". In: *IEEE International Geoscience and Remote Sensing Symposium (IGARSS)*. Milan, Italy: IEEE, July 2015, pp. 3743–3745. ISBN: 978-1-4799-7929-5. DOI: 10.1109/IGARSS.2015.7326637.

[59] Simon A. Wagner. "SAR ATR by a Combination of Convolutional Neural Network and Support Vector Machines". In: *IEEE Transactions on Aerospace and Electronic Systems* 52.6 (Dec. 2016), pp. 2861–2872. ISSN: 0018-9251. DOI: 10.1109/TAES.2016.160061.

[60] Zhenpeng Feng et al. "Self-Matching CAM: A Novel Accurate Visual Explanation of CNNs for SAR Image Interpretation". In: *Remote Sensing* 13.9 (May 1, 2021), p. 1772. ISSN: 2072-4292. DOI: 10.3390/rs13091772.

[61] Boying Li et al. "OpenSARShip 2.0: A Large-Volume Dataset for Deeper Interpretation of Ship Targets in Sentinel-1 Imagery". In: *SAR in Big Data Era: Models, Methods and Applications (BIGSARDATA)*. Beijing: IEEE, Nov. 2017, pp. 1–5. ISBN: 978-1-5386-4519-2. DOI: 10.1109/BIGSARDATA.2017.8124929.

[62] Yuanyuan Wang et al. "A SAR Dataset of Ship Detection for Deep Learning under Complex Backgrounds". In: *Remote Sensing* 11.7 (Mar. 29, 2019), p. 765. ISSN: 2072-4292. DOI: 10.3390/rs11070765.

[63] Tianwen Zhang et al. "SAR Ship Detection Dataset (SSDD): Official Release and Comprehensive Data Analysis". In: *Remote Sensing* 13.18 (Sept. 15, 2021), p. 3690. ISSN: 2072-4292. DOI: 10.3390/rs13183690.

[64] Miao Kang et al. "A Modified Faster R-CNN Based on CFAR Algorithm for SAR Ship Detection". In: *International Workshop on Remote Sensing with Intelligent Processing (RSIP)*. Shanghai, China: IEEE, May 2017, pp. 1–4. ISBN: 978-1-5386-1990-2. DOI: 10.1109/RSIP.2017.7958815.

[65] Di Zhang et al. "Transfer Learning with Convolutional Neural Networks for SAR Ship Recognition". In: *IOP Conference Series: Materials Science and Engineering* 322 (Mar. 2018), p. 072001. ISSN: 1757-8981, 1757-899X. DOI: 10.1088/1757-899X/322/7/072001.

[66] Haoyuan Guo et al. "A CenterNet++ Model for Ship Detection in SAR Images". In: *Pattern Recognition* 112 (Apr. 2021), p. 107787. ISSN: 00313203. DOI: 10.1016/j.patcog.2020.107787.

[67] Biao Hou et al. "A Neural Network Based on Consistency Learning and Adversarial Learning for Semi-supervised Synthetic Aperture Radar Ship Detection". In: *IEEE Transactions on Geoscience and Remote Sensing* (2022). ISSN: 0196-2892, 1558-0644. DOI: 10.1109/TGRS.2022.3142017.

[68] Huiming Xie et al. "Multilayer Feature Learning for Polarimetric Synthetic Radar Data Classification". In: *IEEE Geoscience and Remote Sensing Symposium (IGARSS)*. Quebec City, QC: IEEE, July 2014, pp. 2818–2821. ISBN: 978-1-4799-5775-0. DOI: 10.1109/IGARSS.2014.6947062.

[69] Jie Geng et al. "Deep Supervised and Contractive Neural Network for SAR Image Classification". In: *IEEE Transactions on Geoscience and Remote Sensing* 55.4 (Apr. 2017), pp. 2442–2459. ISSN: 0196-2892, 1558-0644. DOI: 10.1109/TGRS.2016.2645226.

[70] Wenjin Wu et al. "PolSAR Image Semantic Segmentation Based on Deep Transfer Learning—Realizing Smooth Classification With Small Training Sets". In: *IEEE Geoscience and Remote Sensing Letters* 16.6 (June 2019), pp. 977–981. ISSN: 1545-598X, 1558-0571. DOI: 10.1109/LGRS.2018.2886559.

[71] Xiaying Wang et al. "HR-SAR-Net: A Deep Neural Network for Urban Scene Segmentation from High-Resolution SAR Data". In: *IEEE Sensors Applications Symposium (SAS)*. Kuala Lumpur, Malaysia: IEEE, Mar. 2020, pp. 1–6. ISBN: 978-1-72814-842-7. DOI: 10.1109/SAS48726.2020.9220068.

[72] Chu He et al. "Nonlinear Manifold Learning Integrated with Fully Convolutional Networks for PolSAR Image Classification". In: *Remote Sensing* 12.4 (Feb. 17, 2020), p. 655. ISSN: 2072-4292. DOI: 10.3390/rs12040655.

[73] Juanping Zhao et al. "OpenSARUrban: A Sentinel-1 SAR Image Dataset for Urban Interpretation". In: *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing* 13 (2020), pp. 187–203. ISSN: 1939-1404, 2151-1535. DOI: 10.1109/JSTARS.2019.2954850.

[74] Lichao Mou et al. "A CNN for the Identification of Corresponding Patches in SAR and Optical Imagery of Urban Scenes". In: *Joint Urban Remote Sensing Event (JURSE)*. Dubai, United Arab Emirates: IEEE, Mar. 2017, pp. 1–4. ISBN: 978-1-5090-5808-2. DOI: 10.1109/JURSE.2017.7924548.

[75] Min Chen et al. "Robust Feature Matching Method for SAR and Optical Images by Using Gaussian-Gamma-Shaped Bi-Windows-Based Descriptor and Geometric Constraint". In: *Remote Sensing* 9.9 (Aug. 25, 2017), p. 882. ISSN: 2072-4292. DOI: 10.3390/rs9090882.

[76] M. Schmitt et al. “Colorizing Sentinel-1 SAR Images Using a Variational Autoencoder Conditioned on Sentinel-2 Imagery”. In: *The International Archives of the Photogrammetry, Remote Sensing and Spatial Information Sciences* XLII-2 (May 30, 2018), pp. 1045–1051. ISSN: 2194-9034. DOI: 10.5194/isprs-archives-XLII-2-1045-2018.

[77] Patrick Ebel, Michael Schmitt, and Xiao Xiang Zhu. “Cloud Removal in Unpaired Sentinel-2 Imagery Using Cycle-Consistent GAN and SAR-Optical Data Fusion”. In: *IEEE International Geoscience and Remote Sensing Symposium (IGARSS)*. Sept. 26, 2020, pp. 2065–2068. ISBN: 978-1-72816-374-1. DOI: 10.1109/IGARSS39084.2020.9324060.

[78] Xue Li et al. “MCANet: A Joint Semantic Segmentation Framework of Optical and SAR Images for Land Use Classification”. In: *International Journal of Applied Earth Observation and Geoinformation* 106 (Feb. 2022), p. 102638. ISSN: 0303-2434. DOI: 10.1016/j.jag.2021.102638.

[79] Caner Hazirbas et al. “FuseNet: Incorporating Depth into Semantic Segmentation via Fusion-Based CNN Architecture”. In: *Asian Conference on Computer Vision (ACCV)*. Ed. by Shang-Hong Lai et al. Vol. 10111. 2017, pp. 213–228. ISBN: 978-3-319-54180-8. DOI: 10.1007/978-3-319-54181-5_14.

[80] Eunbyung Park et al. “Combining Multiple Sources of Knowledge in Deep CNNs for Action Recognition”. In: *IEEE Winter Conference on Applications of Computer Vision (WACV)*. Mar. 2016, pp. 1–8. ISBN: 978-1-5090-0641-0. DOI: 10.1109/WACV.2016.7477589.

[81] Xiaodong Xu et al. “Multisource Remote Sensing Data Classification Based on Convolutional Neural Network”. In: *IEEE Transactions on Geoscience and Remote Sensing* 56.2 (Feb. 2018), pp. 937–949. ISSN: 0196-2892, 1558-0644. DOI: 10.1109/TGRS.2017.2756851.

[82] Walter Munk et al. “Spirals on the Sea”. In: *Proceedings of the Royal Society of London. Series A: Mathematical, Physical and Engineering Sciences* 456.1997 (May 8, 2000), pp. 1217–1280. ISSN: 1364-5021, 1471-2946. DOI: 10.1098/rspa.2000.0560.

[83] Annika Stuhlmacher and Martin Gade. “Statistical Analyses of Eddies in the Western Mediterranean Sea Based on Synthetic Aperture Radar Imagery”. In: *Remote Sensing of Environment* 250 (Dec. 2020), p. 112023. ISSN: 0034-4257. DOI: 10.1016/j.rse.2020.112023.

[84] Dudley B. Chelton, Michael G. Schlax, and Roger M. Samelson. “Global Observations of Nonlinear Mesoscale Eddies”. In: *Progress in Oceanography* 91.2 (Oct. 2011), pp. 167–216. ISSN: 0079-6611. DOI: 10.1016/j.pocean.2011.01.002.

[85] James H. Faghmous et al. "A Daily Global Mesoscale Ocean Eddy Dataset from Satellite Altimetry". In: *Scientific Data* 2.1 (Dec. 2015), p. 150028. ISSN: 2052-4463. DOI: 10.1038/sdata.2015.28.

[86] Svetlana Karimova and Martin Gade. "Improved Statistics of Sub-Mesoscale Eddies in the Baltic Sea Retrieved from SAR Imagery". In: *International Journal of Remote Sensing* 37.10 (May 18, 2016), pp. 2394–2414. ISSN: 0143-1161, 1366-5901. DOI: 10.1080/01431161.2016.1145367.

[87] Fangyuan Liu et al. "Cross-Domain Submesoscale Eddy Detection Neural Network for HF Radar". In: *Remote Sensing* 13.13 (June 22, 2021), p. 2441. ISSN: 2072-4292. DOI: 10.3390/rs13132441.

[88] James H. Faghmous et al. "A Parameter-Free Spatio-Temporal Pattern Mining Model to Catalog Global Ocean Dynamics". In: *IEEE International Conference on Data Mining (ICDM).* Dec. 2013, pp. 151–160. ISBN: 978-0-7695-5108-1. DOI: 10.1109/ICDM.2013.162.

[89] C. Boller, Fu-Kuo Chang, and Yōzō Fujino, eds. *Encyclopedia of Structural Health Monitoring.* Chichester, West Sussex, U.K: John Wiley, 2009. 5 pp. ISBN: 978-0-470-05822-0.

[90] Xin Sun et al. "A Deep Framework for Eddy Detection and Tracking From Satellite Sea Surface Height Data". In: *IEEE Transactions on Geoscience and Remote Sensing* 59.9 (Sept. 2021), pp. 7224–7234. ISSN: 0196-2892, 1558-0644. DOI: 10.1109/TGRS.2020.3032523.

[91] Evangelos Moschos et al. "Deep-SST-Eddies: A Deep Learning Framework to Detect Oceanic Eddies in Sea Surface Temperature Images". In: *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP).* May 2020, pp. 4307–4311. ISBN: 978-1-5090-6631-5. DOI: 10.1109/ICASSP40776.2020.9053909.

[92] Dongmei Huang et al. "DeepEddy: A Simple Deep Architecture for Mesoscale Oceanic Eddy Detection in SAR Images". In: *IEEE International Conference on Networking, Sensing and Control (ICNSC).* Calabria, Italy, May 2017, pp. 673–678. ISBN: 978-1-5090-4429-0. DOI: 10.1109/ICNSC.2017.8000171.

[93] Martin Gade, Svetlana Karimova, and Annika Buck. "Mediterranean Eddy Statistics Based on Multiple SAR Imagery". In: *Advances in SAR Remote Sensing of Oceans.* Ed. by Xiaofeng Li et al. 1st ed. CRC Press, Oct. 12, 2018, pp. 257–270. ISBN: 978-1-351-23582-2. DOI: 10.1201/9781351235822-15.

[94] Andrei Yu Ivanov and Anna I. Ginzburg. "Oceanic Eddies in Synthetic Aperture Radar Images". In: *Journal of Earth System Science* 111.3 (Sept. 2002), pp. 281–295. ISSN: 0253-4126, 0973-774X. DOI: 10.1007/BF02701974.

[95] Yanling Du et al. "Deep Learning with Multi-Scale Feature Fusion in Remote Sensing for Automatic Oceanic Eddy Detection". In: *Information Fusion* 49 (Sept. 2019), pp. 89–99. ISSN: 15662535. DOI: 10.1016/j.inffus.2018.09.006.

[96] Zhuofan Yan et al. "Multifeature Fusion Neural Network for Oceanic Phenomena Detection in SAR Images". In: *Sensors* 20.1 (Dec. 30, 2019), p. 210. ISSN: 1424-8220. DOI: 10.3390/s20010210.

[97] Kaiming He et al. "Deep Residual Learning for Image Recognition". In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. Las Vegas, NV, USA: IEEE, June 2016, pp. 770–778. ISBN: 978-1-4673-8851-1. DOI: 10.1109/CVPR.2016.90.

[98] *Sentinel Online, Website*. Sentinel Online, website. URL: https://sentinel.esa.int/web/sentinel/.

[99] "Introduction". In: Ian S. Robinson. *Discovering the Ocean from Space*. Springer, 2010, pp. 1–6. ISBN: 978-3-540-24430-1. DOI: 10.1007/978-3-540-68322-3_1.

[100] Stine Skrunes, Camilla Brekke, and Torbjorn Eltoft. "Characterization of Marine Surface Slicks by Radarsat-2 Multipolarization Features". In: *IEEE Transactions on Geoscience and Remote Sensing* 52.9 (Sept. 2014), pp. 5302–5319. ISSN: 0196-2892, 1558-0644. DOI: 10.1109/TGRS.2013.2287916.

[101] Xiaofeng Li et al. *Advances in SAR Remote Sensing of Oceans*. CRC Press, Oct. 2018. ISBN: 978-0-367-57084-2.

[102] Annika Buck. "Statistical Analyses of Eddies in the Western Mediterranean Sea Based on Synthetic Aperture Radar Imagery". University of Hamurg, 2018.

[103] Svetlana Karimova. "Spiral Eddies in the Baltic, Black and Caspian Seas as Seen by Satellite Radar Data". In: *Advances in Space Research* 50.8 (Oct. 2012), pp. 1107–1124. ISSN: 0273-1177. DOI: 10.1016/j.asr.2011.10.027.

[104] Martin Gade et al. "Slicks as Indicators for Marine Processes". In: *Oceanography* 26.2 (June 1, 2013). ISSN: 1042-8275. DOI: 10.5670/oceanog.2013.39.

[105] J. A. Johannessen et al. "Coastal Ocean Fronts and Eddies Imaged with ERS 1 Synthetic Aperture Radar". In: *Journal of Geophysical Research: Oceans* 101.C3 (Mar. 15, 1996), pp. 6651–6667. ISSN: 0148-0227. DOI: 10.1029/95JC02962.

[106] Svetlana Karimova and Martin Gade. "Submesoscale Eddies Seen by Spaceborne Radar". In: *Proc. EMEC 10-MEDCOAST*. Vol. 1. 2013, pp. 665–676.

[107] Martin Gade, Annika Buck, and Svetlana Karimova. "Statistical Analysis of Eddies in the Western Mediterranean Based on Multiple SAR Imagery". In: *IEEE International Geoscience and Remote Sensing Symposium (IGARSS)*. July 2018, pp. 1481–1484. ISBN: 978-1-5386-7150-4. DOI: 10.1109/IGARSS.2018.8518805.

[108] *Earth Online, Website.* URL: https://earth.esa.int/eogateway/tools/snap.

[109] Ali M Reza. “Realization of the Contrast Limited Adaptive Histogram Equalization (CLAHE) for Real-Time Image Enhancement”. In: *Journal of VLSI signal processing systems for signal, image and video technology* 38.1 (2004), pp. 35–44.

[110] Tsung-Yi Lin et al. “Microsoft COCO: Common Objects in Context”. In: *Proceedings of the European Conference on Computer Vision (ECCV).* 2014, pp. 740–755.

[111] Abdul Mueed Hafiz and Ghulam Mohiuddin Bhat. “A Survey on Instance Segmentation: State of the Art”. In: *International Journal of Multimedia Information Retrieval* 9.3 (Sept. 2020), pp. 171–189. ISSN: 2192-6611, 2192-662X. DOI: 10.1007/s13735-020-00195-x.

[112] John Canny. “A Computational Approach to Edge Detection”. In: *IEEE Transactions on Pattern Analysis and Machine Intelligence* PAMI-8.6 (Nov. 1986), pp. 679–698. ISSN: 0162-8828. DOI: 10.1109/TPAMI.1986.4767851.

[113] Hyunho Choi and Jechang Jeong. “Speckle Noise Reduction Technique for SAR Images Using Statistical Characteristics of Speckle Noise and Discrete Wavelet Transform”. In: *Remote Sensing* 11.10 (May 18, 2019), p. 1184. ISSN: 2072-4292. DOI: 10.3390/rs11101184.

[114] Tsung-Yi Lin et al. “Feature Pyramid Networks for Object Detection”. In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR).* Honolulu, HI: IEEE, July 2017, pp. 936–944. ISBN: 978-1-5386-0457-1. DOI: 10.1109/CVPR.2017.106.

[115] Waleed Abdulla. *Mask R-CNN for Object Detection and Instance Segmentation on Keras and TensorFlow.* 2017. URL: https://github.com/matterport/Mask_RCNN.

[116] Navaneeth Bodla et al. “Soft-NMS – Improving Object Detection With One Line of Code”. In: *Proceedings of the IEEE International Conference on Computer Vision (ICCV).* 2017, pp. 5561–5569.

[117] Zhaojin Huang et al. “Mask Scoring R-CNN”. In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR).* 2019, pp. 6409–6418.

[118] Sebastian Ruder. *An Overview of Multi-Task Learning in Deep Neural Networks.* 2017. arXiv: 1706.05098 [cs, stat]. URL: http://arxiv.org/abs/1706.05098.

[119] Roland S. Zimmermann and Julien N. Siems. “Faster Training of Mask R-CNN by Focusing on Instance Boundaries”. In: *Computer Vision and Image Understanding* 188 (Nov. 2019), p. 102795. ISSN: 1077-3142. DOI: 10.1016/j.cviu.2019.102795.

[120] Xingjia Pan et al. “Dynamic Refinement Network for Oriented and Densely Packed Object Detection”. In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)* (2020), pp. 11207–11216.

[121] Xue Yang et al. “SCRDet: Towards More Robust Detection for Small, Cluttered and Rotated Objects”. In: *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)* (2019), pp. 8232–8241.

[122] Yaqing Wang et al. “Generalizing from a Few Examples: A Survey on Few-Shot Learning”. In: *ACM Computing Surveys (CSUR)*. 2020, pp. 1–34.

[123] Martin Gade et al. “Multi-Frequency SAR Data Help Improving the Monitoring of Intertidal Flats on the German North Sea Coast”. In: *Estuarine, Coastal and Shelf Science* 140 (Mar. 2014), pp. 32–42. DOI: 10.1016/j.ecss.2014.01.007.

[124] Robbi Bishop-Taylor et al. “Between the Tides: Modelling the Elevation of Australia’s Exposed Intertidal Zone at Continental Scale”. In: *Estuarine, Coastal and Shelf Science* 223 (July 2019), pp. 115–128. DOI: 10.1016/j.ecss.2019.03.006.

[125] M Billerbeck et al. “Nutrient Release from an Exposed Intertidal Sand Flat”. In: *Marine Ecology Progress Series* 316 (July 3, 2006), pp. 35–51. ISSN: 0171-8630, 1616-1599. DOI: 10.3354/meps316035.

[126] Gail L. Chmura et al. “Global Carbon Sequestration in Tidal, Saline Wetland Soils”. In: *Global Biogeochemical Cycles* 17.4 (Dec. 2003), pp. 1111–1123. ISSN: 0886-6236. DOI: 10.1029/2002GB001917.

[127] S. Smolders et al. “Role of Intertidal Wetlands for Tidal and Storm Tide Attenuation along a Confined Estuary: A Model Study”. In: *Natural Hazards and Earth System Sciences* 15.7 (July 30, 2015), pp. 1659–1675. ISSN: 1684-9981. DOI: 10.5194/nhess-15-1659-2015.

[128] Ying Chen et al. “Land Claim and Loss of Tidal Flats in the Yangtze Estuary”. In: *Scientific Reports* 6.1 (July 2016), p. 24018. ISSN: 2045-2322. DOI: 10.1038/srep24018.

[129] Juan G. Navedo and Alejandro G. Herrera. “Effects of Recreational Disturbance on Tidal Wetlands: Supporting the Importance of Undisturbed Roosting Sites for Waterbird Conservation”. In: *Journal of Coastal Conservation* 16.3 (Sept. 2012), pp. 373–381. ISSN: 1400-0350, 1874-7841. DOI: 10.1007/s11852-012-0208-1.

[130] Winny Adolph et al. “Remote Sensing Intertidal Flats with TerraSAR-X. A SAR Perspective of the Structural Elements of a Tidal Basin for Monitoring the Wadden Sea”. In: *Remote Sensing* 10.7 (July 7, 2018), p. 1085. ISSN: 2072-4292. DOI: 10.3390/rs10071085.

[131] Martin Gade, Wensheng Wang, and Linnea Kemme. “On the Imaging of Exposed Intertidal Flats by Single- and Dual-Co-Polarization Synthetic Aperture Radar”. In: *Remote Sensing of Environment* 205 (Feb. 2018), pp. 315–328. ISSN: 0034-4257. DOI: 10.1016/j.rse.2017.12.004.

[132] Wensheng Wang et al. “A Classification Scheme for Sediments and Habitats on Exposed Intertidal Flats with Multi-Frequency Polarimetric SAR”. In: *Remote Sensing* 13.3 (Jan. 21, 2021), p. 360. ISSN: 2072-4292. DOI: 10.3390/rs13030360.

[133] Sybrand van Beijma, Alexis Comber, and Alistair Lamb. “Random Forest Classification of Salt Marsh Vegetation Habitats Using Quad-Polarimetric Airborne SAR, Elevation and Optical RS Data”. In: *Remote Sensing of Environment* 149 (June 2014), pp. 118–129. ISSN: 0034-4257. DOI: 10.1016/j.rse.2014.04.010.

[134] Wensheng Wang et al. “A Fully Polarimetric SAR Imagery Classification Scheme for Mud and Sand Flats in Intertidal Zones”. In: *IEEE Transactions on Geoscience and Remote Sensing* 55.3 (Mar. 2017), pp. 1734–1742. ISSN: 0196-2892, 1558-0644. DOI: 10.1109/TGRS.2016.2631632.

[135] Wensheng Wang, Martin Gade, and Xiaofeng Yang. “Detection of Bivalve Beds on Exposed Intertidal Flats Using Polarimetric SAR Indicators”. In: *Remote Sensing* 9.10 (Oct. 13, 2017), p. 1047. ISSN: 2072-4292. DOI: 10.3390/rs9101047.

[136] M Gade et al. “Classification of Sediments on Exposed Tidal Flats in the German Bight Using Multi-Frequency Radar Data”. In: *Remote Sensing of Environment* 112.4 (Apr. 15, 2008), pp. 1603–1613. DOI: 10.1016/j.rse.2007.08.015.

[137] Dongling Xiao et al. *PolSAR Image Classification Based on Dilated Convolution and Pixel-Refining Parallel Mapping Network in the Complex Domain*. 2019. arXiv: 1909.10783. URL: https://arxiv.org/abs/1909.10783.

[138] Chun Liu Chun Liu, Junjun Yin Junjun Yin, and Jian Yang Jian Yang. “Application of Deep Learning to Polarimetric SAR Classification”. In: *IET International Radar Conference*. 2015, pp. 1–4. ISBN: 978-1-78561-038-7. DOI: 10.1049/cp.2015.1182.

[139] Liangpei Zhang, Lefei Zhang, and Bo Du. “Deep Learning for Remote Sensing Data: A Technical Tutorial on the State of the Art”. In: *IEEE Geoscience and Remote Sensing Magazine* 4.2 (June 2016), pp. 22–40. ISSN: 2168-6831, 2473-2397. DOI: 10.1109/MGRS.2016.2540798.

[140] Xu Liu et al. “Polarimetric Convolutional Network for PolSAR Image Classification”. In: *IEEE Transactions on Geoscience and Remote Sensing* 57.5 (May 2019), pp. 3040–3054. ISSN: 0196-2892, 1558-0644. DOI: 10.1109/TGRS.2018.2879984.

[141] Xiaofeng Li et al. “Deep-Learning-Based Information Mining from Ocean Remote-Sensing Imagery”. In: *National Science Review* 7.10 (Oct. 13, 2020), pp. 1584–1605. ISSN: 2095-5138, 2053-714X. DOI: 10.1093/nsr/nwaa047.

[142] Gang Zheng et al. “Purely Satellite Data–Driven Deep Learning Forecast of Complicated Tropical Instability Waves”. In: *Science Advances* 6.29 (July 17, 2020), eaba1482. ISSN: 2375-2548. DOI: 10.1126/sciadv.aba1482.

[143] Daphne van der Wal, Peter M.J. Herman, and Annette Wielemaker-van den Dool. “Characterisation of Surface Roughness and Sediment Texture of Intertidal Flats Using ERS SAR Imagery”. In: *Remote Sensing of Environment* 98.1 (Sept. 2005), pp. 96–109. ISSN: 0034-4257. DOI: 10.1016/j.rse.2005.06.004.

[144] O. Regniers et al. “Classification of Oyster Habitats by Combining Wavelet-Based Texture Features and Polarimetric SAR Descriptors”. In: *2015 IEEE International Geoscience and Remote Sensing Symposium (IGARSS)*. Milan, Italy: IEEE, July 2015, pp. 3890–3893. ISBN: 978-1-4799-7929-5. DOI: 10.1109/IGARSS.2015.7326674.

[145] Robert M. Haralick, K. Shanmugam, and Its’Hak Dinstein. “Textural Features for Image Classification”. In: *IEEE Transactions on Systems, Man, and Cybernetics* SMC-3.6 (Nov. 1973), pp. 610–621. ISSN: 0018-9472, 2168-2909. DOI: 10.1109/TSMC.1973.4309314.

[146] Lanyun Zhu et al. “Learning Statistical Texture for Semantic Segmentation”. In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)* (2021), pp. 12537–12546.

[147] Martin Gade and Sabrina Melchionna. “Joint Use of Multiple Synthetic Aperture Radar Imagery for the Detection of Bivalve Beds and Morphological Changes on Intertidal Flats”. In: *Estuarine, Coastal and Shelf Science* 171 (Mar. 2016), pp. 1–10. ISSN: 0272-7714. DOI: 10.1016/j.ecss.2016.01.025.

[148] W. Wang et al. “Random Forest Classification of Sediments on Exposed Intertidal Flats Using ALOS-2 Quad-Polarimetric SAR Data”. In: *ISPRS - International Archives of the Photogrammetry, Remote Sensing and Spatial Information Sciences* XLI-B8 (June 24, 2016), pp. 1191–1194. ISSN: 2194-9034. DOI: 10.5194/isprsarchives-XLI-B8-1191-2016.

[149] Xin Zhang et al. "How Well Do Deep Learning-Based Methods for Land Cover Classification and Object Detection Perform on High Resolution Remote Sensing Imagery?" In: *Remote Sensing* 12.3 (Jan. 28, 2020), p. 417. ISSN: 2072-4292. DOI: 10.3390/rs12030417.

[150] Renaud Mathieu, Claire Freeman, and Jagannath Aryal. "Mapping Private Gardens in Urban Areas Using Object-Oriented Techniques and Very High-Resolution Satellite Imagery". In: *Landscape and Urban Planning* 81.3 (June 2007), pp. 179–192. ISSN: 01692046. DOI: 10.1016/j.landurbplan.2006.11.009.

[151] Johannes J. Feddema et al. "The Importance of Land-Cover Change in Simulating Future Climates". In: *Science* 310.5754 (Dec. 9, 2005), pp. 1674–1678. ISSN: 0036-8075, 1095-9203. DOI: 10.1126/science.1118160.

[152] Sam Navin MohanRajan, Agilandeeswari Loganathan, and Prabukumar Manoharan. "Survey on Land Use/Land Cover (LU/LC) Change Analysis in Remote Sensing and GIS Environment: Techniques and Challenges". In: *Environmental Science and Pollution Research* 27.24 (Aug. 2020), pp. 29900–29926. ISSN: 0944-1344, 1614-7499. DOI: 10.1007/s11356-020-09091-7.

[153] Ava Vali, Sara Comai, and Matteo Matteucci. "Deep Learning for Land Use and Land Cover Classification Based on Hyperspectral and Multispectral Earth Observation Data: A Review". In: *Remote Sensing* 12.15 (Aug. 3, 2020), p. 2495. ISSN: 2072-4292. DOI: 10.3390/rs12152495.

[154] Fariba Mohammadimanesh et al. "A New Fully Convolutional Neural Network for Semantic Segmentation of Polarimetric SAR Imagery in Complex Land Cover Ecosystem". In: *ISPRS Journal of Photogrammetry and Remote Sensing* 151 (May 2019), pp. 223–236. ISSN: 0924-2716. DOI: 10.1016/j.isprsjprs.2019.03.015.

[155] Lin Zhu et al. "Generative Adversarial Networks for Hyperspectral Image Classification". In: *IEEE Transactions on Geoscience and Remote Sensing* 56.9 (Sept. 2018), pp. 5046–5063. ISSN: 0196-2892, 1558-0644. DOI: 10.1109/TGRS.2018.2805286.

[156] Diego Marcos et al. "Land Cover Mapping at Very High Resolution with Rotation Equivariant CNNs: Towards Small yet Accurate Models". In: *ISPRS Journal of Photogrammetry and Remote Sensing* 145 (Nov. 2018), pp. 96–107. DOI: 10.1016/j.isprsjprs.2018.01.021.

[157] Wei Zhang et al. "WTS: A Weakly towards Strongly Supervised Learning Framework for Remote Sensing Land Cover Classification Using Segmentation Models". In: *Remote Sensing* 13.3 (Jan. 23, 2021), p. 394. ISSN: 2072-4292. DOI: 10.3390/rs13030394.

[158] Samadhan C. Kulkarni and Priti P. Rege. “Pixel Level Fusion Techniques for SAR and Optical Images: A Review”. In: *Information Fusion* 59 (July 2020), pp. 13–29. ISSN: 15662535. DOI: 10.1016/j.inffus.2020.01.003.

[159] Tao Lei et al. “Multi-Modality and Multi-Scale Attention Fusion Network for Land Cover Classification from VHR Remote Sensing Images”. In: *Remote Sensing* 13.18 (Sept. 20, 2021), p. 3771. ISSN: 2072-4292. DOI: 10.3390/rs13183771.

[160] Marc Ruswurm et al. “Meta-Learning for Few-Shot Land Cover Classification”. In: *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*. Seattle, WA, USA: IEEE, June 2020, pp. 788–796. ISBN: 978-1-72819-360-1. DOI: 10.1109/CVPRW50498.2020.00108.

[161] Qiutong Yu et al. “Spatial Resolution Enhancement for Large-Scale Land Cover Mapping via Weakly Supervised Deep Learning”. In: *Photogrammetric Engineering & Remote Sensing* 87.6 (June 1, 2021), pp. 405–412. ISSN: 0099-1112. DOI: 10.14358/PERS.87.6.405.

[162] Lucas Hu, Caleb Robinson, and Bistra Dilkina. “Model Generalization in Deep Learning Applications for Land Cover Mapping”. In: *International Conference on Knowledge Discovery and Data Mining Workshop (KDDW)*. 2021, p. 9.

[163] Michael Schmitt et al. “Weakly Supervised Semantic Segmentation of Satellite Images for Land Cover Mapping – Challenges and Opportunities”. In: *ISPRS Annals of the Photogrammetry, Remote Sensing and Spatial Information Sciences*. Vol. 3. 2020, pp. 795–802.

[164] Caleb Robinson et al. “Global Land-Cover Mapping With Weak Supervision: Outcome of the 2020 IEEE GRSS Data Fusion Contest”. In: *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing* 14 (2021), pp. 3185–3199. ISSN: 1939-1404, 2151-1535. DOI: 10.1109/JSTARS.2021.3063849.

[165] Yisen Wang et al. “Symmetric Cross Entropy for Robust Learning With Noisy Labels”. In: *International Conference on Computer Vision (ICCV)*. Oct. 2019, pp. 322–330. ISBN: 978-1-72814-803-8. DOI: 10.1109/ICCV.2019.00041.

[166] Georg Krispel et al. “FuseSeg: LiDAR Point Cloud Segmentation Fusing Multi-Modal Data”. In: *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*. 2020, pp. 1874–1883.

[167] Lei Sun et al. “Real-Time Fusion Network for RGB-D Semantic Segmentation Incorporating Unexpected Obstacle Detection for Road-driving Images”. In: *IEEE Robotics and Automation Letters* 5 (2020), pp. 5558–5565.

[168] Chiheb Trabelsi et al. “Deep Complex Networks”. In: *International Conference on Learning Representations (ICLR)*. 2018, p. 19.

Eidesstattliche Versicherung

Hiermit erkläre ich an Eides statt, dass ich die vorliegende Dissertationsschrift selbst verfasst und keine anderen als die angegebenen Quellen und Hilfsmittel benutzt habe.

Ort, Datum Unterschrift