Master Thesis

**Deep Learning based Harmonization and Super-Resolution of Landsat-8 and Sentinel-2 images**

Author: Venkatesh Thirugnana Sambandham

April 19, 2022

Advisers:

**Supervisor**

Prof. Dr. rer. nat. Frank Ortmeier  
Department of Computer Science  
Otto-von-Guericke University  
Universitätsplatz 2  
39106 Magdeburg, Germany

**Supervisor**

Dr. Mojtaba Karami  
Remote Sensing Lead  
BASF Digital Farming GmbH  
Im Zollhafen 24  
50678 Köln, Germany
Thirugnana Sambandham, Venkatesh:
Deep Learning based Harmonization and Super-Resolution of Landsat-8 and Sentinel-2 images
Master Thesis, Otto-von-Guericke University
Magdeburg, 2022.
STATEMENT OF AUTHORSHIP of the Student

**Thesis:** Deep Learning based Harmonization and Super-Resolution of Landsat-8 and Sentinel-2 images

**Surname:** Thirugnana Sambandham   **Name:** Venkatesh

**DOB:** 29.08.1995   **Matriculation:** 226098

I herewith assure that I wrote the presented thesis independently, that the thesis has not been partially or fully submitted as graded academic work and that I have used no other means than the ones indicated. I have indicated all parts of the work in which sources are used according to their wording or to their meaning.

I am aware of the fact that violations of copyright can lead to injunctive relief and claims for damages of the author as well as a penalty by the law enforcement agency.

Magdeburg, April 19, 2022

[Signature]

[Signature]

[Signature]
In many industries where giant data sets simply don't exist, I think the focus has to shift from big data to good data.

Having 50 thoughtfully engineered examples can be sufficient to explain to the neural network what you want it to learn.

Andrew Ng, CEO Founder, Landing AI
Abstract

Earth Observation using remotely sensed images from satellite sensors has been a fascinating topic of study in recent days. Landsat-8 by NASA and sentinel-2 (A&B) by ESA are two very prominent multi-spectral imaging satellite projects that provide open-source data. Images from these sensors are used in monitoring vegetation changes, urban development, catastrophe management, and many cutting-edge applications. However, these multi-spectral imaging sensors work in the visible to the infrared region of the Electro-Magnetic Spectrum which could not penetrate through clouds. Since majority of the earth’s surface is consistently covered with clouds, images from these sensors over a cloud-covered region cannot be used. The combined use of multi-spectral images from multiple image sources is a viable option to increase the temporal availability of cloud-free images. However, such an approach brings in many uncertainties in the analysis due to the differences in the configurations of the imaging sensors and spatial resolution. Several types of research are already done on bringing down the differences. This thesis explores a possibility of using a Deep Learning based pipeline that brings down the spectral differences between the two image sources, and thereby improving the spatial resolution of landsat-8 image.

The dataset for this work is created using the images from both sensors that were taken on the same day over the study area. The proposed pipeline has a comprehensive preprocessing step, which also includes the development of a band-pass adjustment function that reduces the spectral differences of the common bands between the two imaging sensors. The preprocessing inputs are then upsampled using a Convolution Neural Network based super-resolution architecture. To train this architecture the high-resolution sentinel-2 images are used as ground truth, the landsat-8 images are then brought to the resolution of sentinel-2. The best architecture for this pipeline is a UNet based super-resolution model that fuses the pan-chromatic band of landsat-8 with the multi-spectral bands and thereby upsampling the multi-spectral bands.

The proposed pipeline improves the spatial details of the landsat-8 image and around 5% improvement in the SSIM metrics is observed. A significant drop in the pixel-to-pixel NRMSE metrics between the images was also observed. In addition to that significant improvements in the correlation between the derived bands of Normalized Difference Vegetation Index (NDVI) and Normalized Difference Water Index (NDWI) are also observed. The robustness of the pipeline is demonstrated by performing a use-case scenario of field observation over a period. All the tools, libraries, and the data that are used in this work are from open sources and the whole work is easily reproducible.
Acknowledgements

First and foremost, I thank my parents for all the unconditional love and support that they have given me over all these years. The decision to move across the world for a post-graduation is a big deal and I cannot thank you both enough for all the support and encouragement. I wish I had the tolerance you both have at handling all my shenanigans for the past few months. This whole work is dedicated to you both. This thesis would not have been possible without Mr. Sayan Mukhopadhaya from BASF Digital Farming GmbH GmbH who is more of a friend than a guide or a teacher. Dear Sayan, thank you so much for all the guidance over the past year. The things that I have learned at BASF Digital Farming GmbH working with you over the course of my term as well as this thesis are huge and I believe will act as a solid foundation for my upcoming career. Thank you for holding me on a leash during the course of this thesis and preventing me to go astray with all the decisions that I made along the way. Unfortunately, we could not work together in-person but I never felt like being left alone throughout the entire time that I have spent working with BASF Digital Farming GmbH.

Mr. Konstantin Kirchheim, my guide at the OVGU, thank you so much for all the inputs that you have given over the course for completion of this thesis. Your timely review and comment on each sections of the report has been very crucial. Most of all, thank you for the workplace at the university.

I would also like to thank Prof. Dr. Frank Ortmeier from OVGU, and Dr. Mojtaba Karami from Xarvio Gmbh for allowing me to work on this topic. I would like to express my gratitude to all the teachers from OVGU for all the guidance and for helping me take this path.

Last but not least my friends in Magdeburg as well as in India, especially Sidharth, thank you for all the hours of much-needed DOTA 2 diversion over the course of this work.

- Venkatesh Thirugnana Sambandham
Contents

Abstract

Abbreviations and Notations

1 Introduction and Motivation
  1.1 Introduction ................................................. 1
    1.1.1 Earth Observation (EO) ............................. 1
    1.1.2 Motivation .............................................. 4
  1.2 Research Questions ........................................ 4
  1.3 Aim of this thesis ......................................... 5
  1.4 Research Approach ....................................... 5
  1.5 Structure of this thesis .................................. 5

2 Background
  2.1 Inter-Sensor Harmonization .............................. 7
    2.1.1 Harmonized Landsat sentinel-2 Product (HLS) .... 8
  2.2 Deep Learning based Super-Resolution ................. 11
    2.2.1 Early Super-Resolution Models .................... 12
    2.2.2 Super-Resolution Architecture Developments ...... 12
  2.3 Super Resolution on Remote Sensing Images ............ 16
    2.3.1 Inter-Sensor Super-Resolution ..................... 16
    2.3.2 Pan Sharpening Techniques ......................... 17

3 Methods
  3.1 Data Acquisition ...................................... 20
    3.1.1 Study Area .......................................... 20
    3.1.2 Data Search and Acquisition .................... 21
  3.2 Preprocessing ......................................... 23
    3.2.1 Reprojection ........................................ 23
    3.2.2 Image Registration .................................. 24
    3.2.3 Cloud Masking and Filtering .................... 25
    3.2.4 Patching ............................................. 27
    3.2.5 Further Cloud Filtering ......................... 28
    3.2.6 Band-Pass adjustments ............................ 29
  3.3 Super-Resolution Models ................................. 30
    3.3.1 Deep Convolution Neural Network Architectures .. 30
    3.3.2 Data Loader Pipeline .............................. 34
    3.3.3 Training Configuration ............................. 35
  3.4 Evaluation Protocol ..................................... 36
    3.4.1 Evaluation Metrics ................................. 37
3.4.2 Derived Bands and Metrics ............................................. 38
3.4.3 Representation of Metrics - box plot ............................. 40
3.4.4 Test of Significance ...................................................... 41

4 Experiments and Evaluation
4.1 Band Pass Adjustment model experiments .......................... 43
4.2 SR DCNN results ........................................................... 45
  4.2.1 Initial Experiments ...................................................... 46
  4.2.2 Effects of Band-Pass Adjustments and High Pass Filter .... 49
4.3 Evaluation of Derived Bands ............................................ 52
  4.3.1 Derived Bands Test Metrics ........................................ 52
  4.3.2 Time Series analysis ................................................ 54

5 Conclusions
5.1 Answers to the Research Questions .................................. 59
  5.1.1 RQ1 : How to harmonize images from both the sources and
          super-resolve landsat-8 images to sentinel-2 size? .......... 59
  5.1.2 RQ2 : What are the existing super-resolution techniques that
          improves the spatial resolution of images? How are their perfor-
          mances on remote sensing dataset? ............................. 60
5.2 Summary ................................................................. 61

6 Future Work

A Detailed Results
  A.1 Significance of co-registration ..................................... 65

B List of Figures

C List of Tables

D Bibliography
## Abbreviations and Notations

<table>
<thead>
<tr>
<th>Acronym</th>
<th>Meaning</th>
</tr>
</thead>
<tbody>
<tr>
<td>EO</td>
<td>Earth Observation</td>
</tr>
<tr>
<td>NASA</td>
<td>National Aeronautics and Space Administration</td>
</tr>
<tr>
<td>OLI</td>
<td>Optical Land Imager</td>
</tr>
<tr>
<td>MS</td>
<td>Multi Spectral</td>
</tr>
<tr>
<td>PAN</td>
<td>Panchromatic</td>
</tr>
<tr>
<td>LS8</td>
<td>Landsat-8</td>
</tr>
<tr>
<td>S2</td>
<td>Sentinel-2</td>
</tr>
<tr>
<td>EO</td>
<td>Earth Observation</td>
</tr>
<tr>
<td>TIRS</td>
<td>Thermal Infrared sensors</td>
</tr>
<tr>
<td>m</td>
<td>meters</td>
</tr>
<tr>
<td>ESA</td>
<td>European Space Agency</td>
</tr>
<tr>
<td>S2A</td>
<td>Sentinel-2A</td>
</tr>
<tr>
<td>S2B</td>
<td>Sentinel-2B</td>
</tr>
<tr>
<td>EMS</td>
<td>Electro Magnetic Spectrum</td>
</tr>
<tr>
<td>MSI</td>
<td>multi-Spectral Instrument</td>
</tr>
<tr>
<td>NIR</td>
<td>Near InfraRed</td>
</tr>
<tr>
<td>SWIR</td>
<td>Short Wave InfraRed</td>
</tr>
<tr>
<td>AOI</td>
<td>Area of Interest</td>
</tr>
<tr>
<td>SSIM</td>
<td>Structural Similarity Index Measure</td>
</tr>
<tr>
<td>PSNR</td>
<td>Peak Signal to Noise Ratio</td>
</tr>
<tr>
<td>HLS</td>
<td>Harmonized Landsat Sentinel-2</td>
</tr>
<tr>
<td>SRF</td>
<td>Spectral Response Function</td>
</tr>
<tr>
<td>USGS</td>
<td>United States Geological Survey</td>
</tr>
<tr>
<td>TOA</td>
<td>Top Of the Atmosphere</td>
</tr>
<tr>
<td>BOA</td>
<td>Bottom of the Atmosphere</td>
</tr>
<tr>
<td>SuR</td>
<td>Surface Reflectance</td>
</tr>
<tr>
<td>AC</td>
<td>Atmospheric Correction</td>
</tr>
<tr>
<td>FMASK</td>
<td>Fast Masking</td>
</tr>
<tr>
<td>AROP</td>
<td>Automated Registration and Ortho Rectification Package</td>
</tr>
<tr>
<td>BRDF</td>
<td>Bidirectional Reflectance Distribution Function</td>
</tr>
<tr>
<td>NBAR</td>
<td>Nadir BRDF Adjusted Reflectance</td>
</tr>
<tr>
<td>LR</td>
<td>Low Resolution</td>
</tr>
<tr>
<td>HR</td>
<td>High Resolution</td>
</tr>
<tr>
<td>SR</td>
<td>Super Resolution</td>
</tr>
<tr>
<td>DCNN</td>
<td>Deep Convolution Neural Network</td>
</tr>
<tr>
<td>GAN</td>
<td>Generative Adversarial Neural Network</td>
</tr>
<tr>
<td>SOTA</td>
<td>State of the Art</td>
</tr>
<tr>
<td>MOS</td>
<td>Mean Opinion Score</td>
</tr>
<tr>
<td>RGB</td>
<td>Red Green Blue</td>
</tr>
<tr>
<td>MGL</td>
<td>Mixed Gradient Loss</td>
</tr>
<tr>
<td>MGRS</td>
<td>Military Grid Reference System</td>
</tr>
<tr>
<td>STAC</td>
<td>Spatio Temporal Access Catalogue</td>
</tr>
<tr>
<td>CRS</td>
<td>Coordinate Reference System</td>
</tr>
<tr>
<td>EPSG</td>
<td>European Petroleum Survey Group</td>
</tr>
<tr>
<td>AROSICS</td>
<td>Automated and Robust Open Source Image Coregistration Software</td>
</tr>
<tr>
<td>QA</td>
<td>Quality Assessment</td>
</tr>
<tr>
<td>Acronym</td>
<td>Meaning</td>
</tr>
<tr>
<td>---------</td>
<td>---------------------------------------------</td>
</tr>
<tr>
<td>SCL</td>
<td>Scene Classification Layer</td>
</tr>
<tr>
<td>ReLU</td>
<td>Rectified Linear Unit</td>
</tr>
<tr>
<td>LReLU</td>
<td>Leaky ReLU</td>
</tr>
<tr>
<td>PReLU</td>
<td>Parametric ReLU</td>
</tr>
<tr>
<td>DN</td>
<td>Digital Numbers</td>
</tr>
<tr>
<td>ADAM</td>
<td>Adaptive Moment Estimation</td>
</tr>
<tr>
<td>AMP</td>
<td>Automatic Mixed Precision</td>
</tr>
<tr>
<td>GPU</td>
<td>Graphic Processing Unit</td>
</tr>
<tr>
<td>NRMSE</td>
<td>Normalized Root Mean Squared Error</td>
</tr>
<tr>
<td>NDVI</td>
<td>Normalized Difference Vegetation Index</td>
</tr>
<tr>
<td>NDWI</td>
<td>Normalized Difference Water Index</td>
</tr>
<tr>
<td>MAE</td>
<td>Mean Absolute Error</td>
</tr>
<tr>
<td>BOAR</td>
<td>Bottom of the Atmosphere Reflectance</td>
</tr>
</tbody>
</table>
Introduction and Motivation

1.1 Introduction

1.1.1 Earth Observation (EO)

Recent developments in hardware technology have brought surreal possibilities for efficient data handling and manipulation. One such field which exploited all these development in recent years is remote sensing technology. A large number of satellite clusters are constantly detecting and monitoring the physical characteristics of the earth’s surface using active and passive electromagnetic sensors. Multi-spectral imaging satellites follow low earth, sun-synchronous near-polar orbits, meaning the satellite passes over any given point of the earth’s surface at the same local time. These satellites capture data across multiple bands of the electromagnetic spectrum. Since the launch of the landsat-1 mission in 1972 by the National Aeronautics and Space Administration (NASA) a consistent observation of the earth’s surface has been obtained [BERNSTEIN (1976)] and open-sourced. After a series of landsat missions, the most recent one is the landsat-9 which was launched in September 2021. Previous landsat instruments like landsat-7 and landsat-8 (LS8) are operational with landsat-8 still working at full capacity.

Landsat-8 uses a multi-spectral image sensor called Operational Land Imager (OLI) to acquire scenes in the visible, near-infrared, and short wave infrared spectrum while the Thermal Infrared Sensor (TIRS) acquires images at the infra-red level to observe the surface temperature of the scene. These sensors acquire images at a medium-high resolution of 30m (meters). The availability of consistent observations of the earth at multiple spectral wavelengths opens a wide range of applications like crop yield and disease modeling in agriculture, urban development, climate change detection, land use land cover changes and so on [ROY et al. (2014), MUKHOPADHAYA (2016b)]. Not only the direct applications but also the possibilities to merge with other works like rainfall detection, ground water estimation, etc. to make the anal-
yses robust [RAIHAN et al. (2022), MUKHOPADHYAYA (2016c)]. The landsat-8 satellite has a median revisit period of 14 days which means it covers every part of the earth once in at least 14 days.

The sentinel-2 (S2) mission from European Space Agency (ESA) comprises two satellites sentinel-2A and sentinel-2B similar to that of landsat-8. These satellites are high-resolution, multi-spectral imaging satellites phased out of each other's orbit by 180 degrees. These orbital specifications made it possible for this mission to have a revisit period of 5 days and the availability of images from sentinel-2 (S2) is comparatively more frequent than landsat-8. Both the satellites, sentinel-2A (S2A) and sentinel-2B (S2B) carry a similar sensor called multi-spectral Instrument (MSI) as payload. The MSI captures scenes at 13 different spectral bands at different resolutions ranging from 10m to 60m. Apart from the similar spectral bands of landsat-8, the sentinel-2 sensors also captures images at the red-edge spectral region. The spectral information and the resolution of landsat-8 and sentinel-2 sensors can be found in table 1.1. It can be observed that both sentinel-2 and landsat-8 have 8 bands in common. Although they are similar but they are not identical due to differences in sensors and hence having differences in central wavelengths and also spatial resolutions. In almost all cases (except coastal/aerosol and cirrus) sentinel-2 has better spatial resolution comparatively.

Figure 1.1 illustrates the spectral characteristics of the sensors at the common bands covering similar wavelengths in the Electro-Magnetic Spectrum (EMS). The multi-spectral image data are represented as reflectances, it is the ratio of the amount of light leaving the target to the amount of light striking the target. Reflectances are typically a value between 0 and 1 with no units. However, if there are highly reflective surfaces the reflectance value tends to go a little higher than 1. In general reflectance values are used to determine the properties of the materials that are being observed.
1.1. INTRODUCTION

Table 1.1: Central wavelengths and resolutions of different bands from sentinel-2 and landsat-8 with common bands and highest resolution in bold.

<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>Costal/aerosol</td>
<td>0.443</td>
<td>30</td>
<td>0.442</td>
<td>60</td>
<td>0.442</td>
<td>60</td>
</tr>
<tr>
<td>Blue</td>
<td>0.492</td>
<td>30</td>
<td>0.492</td>
<td>60</td>
<td>0.492</td>
<td>60</td>
</tr>
<tr>
<td>Green</td>
<td>0.562</td>
<td>30</td>
<td>0.559</td>
<td>10</td>
<td>0.559</td>
<td>10</td>
</tr>
<tr>
<td>Red</td>
<td>0.655</td>
<td>30</td>
<td>0.664</td>
<td>10</td>
<td>0.664</td>
<td>10</td>
</tr>
<tr>
<td>Red Edge 1</td>
<td>-</td>
<td>-</td>
<td>0.704</td>
<td>20</td>
<td>0.703</td>
<td>20</td>
</tr>
<tr>
<td>Red Edge 2</td>
<td>-</td>
<td>-</td>
<td>0.74</td>
<td>20</td>
<td>0.739</td>
<td>20</td>
</tr>
<tr>
<td>Red Edge 3</td>
<td>-</td>
<td>-</td>
<td>0.782</td>
<td>20</td>
<td>0.779</td>
<td>20</td>
</tr>
<tr>
<td>NIR</td>
<td>-</td>
<td>-</td>
<td>0.832</td>
<td>10</td>
<td>0.832</td>
<td>10</td>
</tr>
<tr>
<td>Narrow NIR</td>
<td>0.865</td>
<td>30</td>
<td>0.864</td>
<td>20</td>
<td>0.864</td>
<td>20</td>
</tr>
<tr>
<td>Water vapour</td>
<td>-</td>
<td>-</td>
<td>0.945</td>
<td>60</td>
<td>0.943</td>
<td>60</td>
</tr>
<tr>
<td>cirrus</td>
<td>1.375</td>
<td>30</td>
<td>1.373</td>
<td>60</td>
<td>1.376</td>
<td>60</td>
</tr>
<tr>
<td>SWIR1</td>
<td>1.61</td>
<td>30</td>
<td>1.613</td>
<td>20</td>
<td>1.61</td>
<td>20</td>
</tr>
<tr>
<td>SWIR2</td>
<td>2.2</td>
<td>30</td>
<td>2.202</td>
<td>20</td>
<td>2.185</td>
<td>20</td>
</tr>
<tr>
<td>Pan Chromatic</td>
<td>0.59</td>
<td>15</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>TIRS-1</td>
<td>10.8</td>
<td>100</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>TIRS-1</td>
<td>12</td>
<td>100</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>Revisit(days)</td>
<td>15</td>
<td></td>
<td>5</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Figure 1.1: Spectral characteristics of the common bands from landsat-8(OLI) and sentinel-2(MSI) sensors.

[Figure 1.1 taken from https://landsat.usgs.gov/spectral-characteristics-viewer]
1.1.2 Motivation

As mentioned in section 1.1.1, Landsat-8 has a revisit period of 14 days and the Sentinel-2 pair has a 5-day revisit period. Li and Chen (2020) did a comprehensive study on the revisit time of Landsat-8 and 9 combined with Sentinel-2 A and B. The study shows that a combination of using all 4 sensors together could bring down the temporal gap between the revisits to around 2.3 days. However, it is never certain that all the images acquired are valid and usable. The prime reason for these images to become unusable is the cloud cover, where the Area of Interest (AOI) is covered by clouds. Since the multi-spectral sensors do not observe wavelengths that could penetrate the opaque clouds, any image covered by clouds is completely invalid. There could be possible cases where an image from one sensor is covered by clouds and can be replaced using a cloud-free image of another sensor. When this kind of multi-sensor data acquisition is required, a lot of factors have to be considered to combine them into a single data source.

- The spectral differences caused by the differences in the imaging sensors have to be minimized.
- The difference in the spatial resolution of the image sources has to be handled.

Spectral harmonization of multiple satellite image sources has been a wide topic of study since the availability of a large number of openly sourced remote sensing multi-spectral image data. This process greatly improves the quality of the data that could be used for other analysis that involves features derived from multi-spectral remote sensing images.

Deep learning-based approaches play a key role in a lot of applications in the field of remote sensing [Zhu et al. (2017)]. Due to the flexibility provided by the various deep-learning frameworks and the availability of state-of-the-art hardware technology, it has become relatively easier these days to set up efficient pipelines to extract features from the given data, and train models that could generate (or predict) the desired output. Extensive research has been done on the topic of deep learning-based super-resolution (SR) with some techniques showing exceptional improvements in the perceptual quality of low-resolution images [Anwar et al. (2019)].

1.2 Research Questions

Based on the motivation of this thesis, the following research questions were formulated.
1.3. AIM OF THIS THESIS

- How to Harmonize images from landsat-8 and sentinel-2 image sources and super-resolve landsat-8 images to sentinel-2 size?

- What are the existing super-resolution techniques that improve the spatial resolution of images? What are their performances on remote sensing datasets?

1.3 Aim of this thesis

The main objectives of this work can be summarised as below

- To investigate the difference in the spectral quality of images taken from sentinel-2 and landsat-8 image sources that were taken over an Area of Interest (AOI) on the same day.

- To build and evaluate a pipeline that could match the image from a low-resolution source (landsat-8) to the image from high-resolution target (sentinel-2).

1.4 Research Approach

To answer the research questions of this work, a dataset with 46 image pairs of landsat-8 and sentinel-2 image sources that were captured on the same day are gathered. The image pairs are filtered for clouds, geo-spatially registered, and are split into small patches with multiple channels from common bands across these two image sources for scalability of the training pipeline. Several deep-learning based super-resolution architectures using different loss functions were trained with the landsat-8 images as input and sentinel-2 images as ground truths. Common Multi-Spectral bands in combination with pan-chromatic bands were also used as inputs to train the super-resolution model to further enhance the spatial features of the low-resolution images. The super-resolved images generated by deep-learning models were evaluated against the ground truth images and the best model is chosen. Standard image quality metrics like Peak Signal to Noise Ratio (PSNR), Structural Similarity Index Measure (SSIM), and Root Mean Square Error were used to compute the differences between the images.

1.5 Structure of this thesis

The whole thesis is organized into 6 chapters. Chapter 1 introduces the background information about the research work, the motivation behind this work
and the approach that was followed. Chapter 2 discusses the different previously existing researches that are done similar to this work. Chapter 3 includes information about the study area that was chosen and an in-depth discussion about various materials and methods that were adopted in this work is presented. Chapter 4 presents and discusses the results from various experiments that were conducted. Chapter 5 answers the formulated research questions based on the results that were observed. The thesis concludes with discussions on a wide range of possible future research opportunities in Chapter 6.
The given input landsat-8 image has to be transformed into a sentinel-2-like image by enhancing the spatial resolution and minimizing the spectral differences. The whole system can be divided into two subsystems, one that handles the spectral differences and the other that improves the spatial resolution of the given input images. This chapter is divided into three sub-sections. In Section 2.1 different inter sensor-harmonization techniques that have been researched are discussed. Details about the existing Harmonized landsat-8 - sentinel-2(HLS) product are discussed. Section 2.2 talks about deep learning based super-resolution(SR) techniques that increase the spatial resolution of the input low-resolution(LR) images into high-resolution(HR) images. In section 2.3 we will discuss SR works that are explicitly applied to up sample multispectral remote sensing images.

2.1 Inter-Sensor Harmonization

The Spectral Response Function(SRF) characterizes the sensitivity of a sensor toward a specific spectral band, this SRF varies with sensors. Hence applications that rely on multi-sensor satellite data are known to have uncertainties due to the inclusion of data from different sources. Efforts on bringing down these differences are not a new concept. A lot of researches has been done on this topic comparing the SRF differences across multiple satellite sensors and methods for bringing down these differences are formulated, some of which are discussed below.

GONSAMO and CHEN (2013) studied the SRF differences across 21 Earth Observation satellite sensors. The work primarily focused on red, near-infrared, and short wave infrared bands of the satellite sensors which are generally used for global vegetation monitoring. The training data for this work were generated using state-of-the-art radiative transfer models [JACQUEMOUD and BARET (1990), VERHOEF (1984), VERMOTE et al. (1997)], a comprehensive tool...
that represents the scattering and absorption of radiation by the elements in
the earth’s atmosphere. The training data is simulated for different combina-
tions of atmospheric and land-cover parameters. Polynomial Regression and
spectral curve information obtained from the training data are used to esti-
mate the correction co-efficient for different pairs of sensors. However, This
work did not include parameters for landsat-8 or sentinel-2.

Scheffler et al. (2020) in their study compared the effect of spectral harmo-
nization techniques on inter-sensor reflectance consistency. Unlike the pre-
vious study where the data is being simulated, the authors of this paper used
hyper-spectral data to build a training dataset. This approach is specifically
focused on landsat-8 and sentinel-2. A cluster of Machine Learning regres-
sion models to predict matching bands across sensors was developed. In ad-
dition to this, this work also proposed an approach that could predict missing
spectral bands of a satellite source, like the Red-Edge spectrum for landsat-
8. The reflectance values from the sensors are clustered, and for each cluster,
a unique Machine Learning regression model is trained for the data points
in the cluster. The final prediction for a target is the ensemble of outputs
generated from the regression models for the closest clusters for each pixel.
However, hyperspectral images need a lot of pre-processing [Mukhopadhaya
(2016a)]

Villaescusa-Nadal et al. (2019) used spectral libraries developed by the
United States Geological Survey(USGS)[Clark et al. (2007)] to simulate re-
fectances, which were used to train models that correct the differences
caused by the variations of sensors. Swinnen et al. (2022) built an end-to-
end pipeline that also includes an atmospheric correction module. The work
came up with a unified module that harmonized images from 4 different satel-
lite sources like sentinel-2A/B, landsat-8, DEIMOS-1 and ,Proba-V. The out-
puts from the pipeline were evaluated against in-situ measurements of Bio-
Physical parameters. Though a significant improvement in accuracy could
not be obtained, the model demonstrated consistency in the reflectances
across sensors.

2.1.1 Harmonized Landsat sentinel-2 Product(HLS)

Considering a relatively lower revisit period for medium satellites and the
potential to improve the temporal resolution of available multi-spectral Im-
ages, NASA has started taking efforts to harmonize Landsat and sentinel-2
products. The result of this effort is a product called Harmonized landsat-8
sentinel-2 product(HLS). The HLS products are generated based on an end-
to-end pipeline defined in Claverie et al. (2018). The overall summary of the
HLS product generation pipeline can be found in Figure 2.1. The product
generation starts from Raw landsat-8 or sentinel-2 image at Top Of the Atmosphere (TOA) level, which means these images are not corrected for distortions and other artifacts caused by the presence of atmosphere between the spacecraft and the earth's surface. Normally, these types of TOA products are represented as Level-1 products. Both the image sources also provide Level-2 products i.e., images that are corrected for the effects of the atmosphere. These are called Surface Reflectance products or BOA products. The BOA products are then spatially co-registered and the Bidirectional Reflectance Distribution Function (BRDF) normalization is applied to convert the BOA images into Nadir BRDF Adjusted Reflectance (NBAR) images.

**Atmospheric Correction and Cloud Filtering**

Both L8 and S2 sources have their atmospheric correction algorithm to convert the TOA images to BOA. Landsat uses the LaSRC [Vermote et al. (2018)]
module to generate Level-2 Surface Reflectance (SuR) products and sentinel-2 Surface Reflectance images are generated by a module called Sen2Cor[Main-Knorn et al. (2017)]. A common atmospheric correction algorithm is used to generate the BOA images to avoid any biases caused by either of the algorithms which could affect the results further down the pipeline. In the case of the official HLS product, the LaSRC module was used to do the conversion for both image sources. However, other researches similar to this pipeline, like the work by Nguyen et al. (2020) used a Py6S Radiative transfer model-based system to perform the atmospheric correction and was able to achieve reasonable consistency between the images over time.

As mentioned in section 1.1.2, the presence of cloud cover makes these kinds of multi-spectral images completely useless. Such kind of images has to be avoided in analysis and using them will result in many uncertainties. Landsat uses the FMASK algorithm[Zhu and Woodcock (2012)], a threshold-based model to detect cloudy pixels in images. In addition to clouds, the default quality mask from Landsat also has information about land cover classification like vegetation, water, and so on. Sentinel-2 also provides a similar cloud/land cover classification mask, in addition to which an FMASK algorithm-based cloud mask is also generated in the HLS product.

Spatial Co-Registration

Mis-registration errors in these image pairs can be categorized into two types. The first one is the mis-registration of a sensor to the ground reference. At the time of writing this work, the mis-registration of sentinel-2 is less than 0.3 pixels as per the monthly Data Quality Report1. The second category is the inter-sensor mis-registration, according to Storey et al. (2016) the spatial misalignment between landsat-8 and sentinel-2 could exceed more than 38m. Nguyen et al. (2020) focused their study on Vietnam and Lebanon region, they were able to identify the inter-sensor misregistration as 10m and 32m in the respective region. The official HLS product uses the Automated Registration and Orthorectification Package (AROP) an open-source library used for processing Landsat and Landsat-like images[Gao et al. (2009)].

Bi-directional Reflectance Distribution Function (BRDF) Normalization and Band Pass Adjustments

The angle of light from the sun illuminating the surface also known as the solar illumination angle and the view angle i.e, the angle at which the craft absorbs the reflected light from the surface varies between different sensors.

---

The difference in this angle could cause some variations in the images captured by these sensors. These angle effects are corrected by the Bidirectional Reflectance Distribution Function (BRDF). The function is solved by setting the view angle at Nadir therefore the result is the simulated reflectance value as observed by the sensor from directly above the point of interest. The end product after the BRDF correction is Nadir BRDF Adjusted Reflectance (NBAR) images.

After performing the BRDF correction, variations in reflectance values caused by the differences in sensors are rectified. These corrections were done using cross-sensor transformation coefficients. These coefficients are generated similarly to the method described by Schieffer et al. (2020).

After performing all the necessary preprocessing steps the Landsat images are kept as they are at 30m resolution and the sentinel-2 is resampled to 30m resolution from the native 10m.

The Sen2Like platform described in Saunier et al. (2019) crunched down all the processes involved in generating the HLS product into a package and made it open-source. In addition to all the steps mentioned above the Sen2Like package also includes a data fusion module that takes into account the nearest available image of sentinel-2 and fuses it with a landsat-8 image to upsample the landsat image into 10m resolution. However, the atmospheric correction module and the BRDF modules require auxiliary data to do the respective processing which although open-source are not easily available.

### 2.2 Deep Learning based Super-Resolution

Super-resolution in itself is an ill-posed problem i.e., Instead of a unique solution there exists multiple solutions for the same LR image [Gonsamo and Chen (2013)]. There are traditional and deep learning methods to solve the SR problem. Traditional up-sampling methods use different interpolation techniques to fill the surrounding pixels while increasing the size of the image. The interpolation techniques include nearest neighbour interpolation, linear interpolation, cubic interpolation, and so on. Unlike traditional techniques, deep learning methods are data-driven approaches. Prior features are learned from a pool of datasets and the actual SR is performed by manipulating features generated from the LR using learned kernels from a trained deep convolution neural network.

Considering the LR images are the result of degradation of a HR image. The degradation process can be represented as:

\[ I_{LR} = F(I_{HR}; \theta) \]
where $I_{LR}$ is the LR image and $I_{HR}$ is the HR counterpart. $F$ is the degradation function and $\theta_\eta$ are the degradation parameters. In this case, the degradation parameters are the differences in the spatial resolutions, the differences in spectral response functions between landsat-8 and sentinel-2 sensors, noises caused by spectral reflectance saturation, the presence of Haze over the Area Of Interest and so on.

The objective of the SR problem is to determine the inverse of the function $F$. In other words, A function has to be derived which nullifies the degradation process and generates a HR-like image given an input LR image. Following this Equation 2.1 becomes

$$I_{SR} = F^{-1}(I_{LR}; \theta_\lambda)$$

Where $I_{SR}$ is the super-resolved image, $F^{-1}$ is the SR function which is the Deep Convolution Neural Network (DCNN) model. $\theta_\lambda$ are the parameters of the function $F^{-1}$ which are the weights and biases of the DCNN model.

### 2.2.1 Early Super-Resolution Models

There exists a wide variety of DCNN models that perform the intended SR tasks. A search in IEEE Xplore search engine with keywords “Super-Resolution” and “Deep Learning” returns around 450 Journals with applications in various fields ranging from medical images, microscopic images, and so on. The pioneer of all DCNN SR architecture is the Super-Resolution Convolution Neural Network proposed by Dong et al. (2015)(SRCNN). It is a straightforward approach that consists of a sequence of Convolution layers combined with ReLU activation for non-linearity. The LR images are upsampled to the size of the HR image and then fed into the model to generate the SR outputs. As the developments in the field of deep learning progressed, the SR architectures also improved. The Denoising Convolution Neural Network (DnCNN) by Zhang et al. (2016) and Image Restoration CNN (IRCNN) by Zhang et al. (2017) are very similar to SRCNN and are tested on SR tasks.

### 2.2.2 Super-Resolution Architecture Developments

**Late Upsampling models**

One of the problems with the early SR architectures is that the inputs are fed to the model after resampling them to the desired-high resolution size. So the entire training and inference pipeline becomes computationally heavy and provides a challenge for an application that requires very minimum latency like the SR of videos. To tackle this issue Dong et al. (2016) came up with the
Fast-Super Resolution Convolution Neural Networks (FSRCNN). The architecture is almost similar to SRCNN except the features are generated at the size of the LR images and there is a final de-convolution layer that upsamples the features to the required HR size. However, the deconvolution operation gave rise to checkerboard artifacts, due to the intersection of the kernels over the same pixel during the convolution operations [Odena et al. (2016)]. Alternatives were suggested to handle this artifact like replacing the deconvolution operation with a pair of up-sample and convolution operations. Shi et al. (2016) suggested a sub-pixel convolution or pixel shuffle operation to upsample the low resolution features on-the-fly. It was considerably faster compared to the deconvolution operation and also showed significant improvement in the performance.

Residual Networks and Generative Adversarial Network (GAN) based training

With the late up-sampling or up-sampling on-the-fly architectures greatly reducing the computational bottleneck in the feature extraction stage of SR architectures, more complex architectures are applied for the SR problem. Deeper CNN architectures with residual connections and feature concatenation from the early stages of the networks were proven to outperform many strictly sequential architectures in image recognition tasks, like the Residual Network (ResNet) architecture proposed by He et al. (2015) and DenseNet architecture by Huang et al. (2016). These architectures were modified for SR tasks and were proven to generate a more realistic super-resolved image. The SR-ResNet proposed by Ledig et al. (2016) followed the Res-Net style architecture which was considered to be the state of the art (SOTA) at that time. The reasons for these significant improvements in the image resolution include not only the aspects of ResNet architecture but also the training process itself. All the early deep learning based SR models before this were trained using the Least Absolute Deviations (L1) or the Least Square Errors (L2) loss function which calculated the pixel-wise difference between the generated SR image and the intended ground-truth HR images. They included a compilation of different loss functions to train the intended SR architecture. In addition to the regular L2 criterion, a content loss is included. Instead of directly comparing the generated SR and HR images, the content loss function compares the features generated on the SR and HR images by another known well-performing architecture like a VGG16 model [Simonyan and Zisserman (2014)] trained on an ImageNet dataset [Deng et al. (2009)] was used to calculate content loss.

Generative Adversarial Networks (GANs), proposed by Goodfellow et al. (2014) were specifically designed for generative deep learning models. Gener-
ative deep learning models are a category of DCNN that learns from a huge collection of data and generates some form of output given an input. All super-resolution architectures in a way can be considered a generative deep learning models. GANs employ a two sub-model supervised learning framework which includes a generator architecture, which in this case is the architecture that super-resolves the input LR images and a discriminator architecture. The role of the discriminator is to classify if the inputs are either real(HR) or Fake(SR). The Super Resolution Redidual Network(SR-ResNet) paper employed this kind of GAN framework to train their model and called it Super Resolution GAN(SRGAN). The SR-ResNet showed exceptional improvement in the perceptual quality of the SR images and the Mean Opinion Score(MOS) proved that it performed well in tricking a human rater into thinking that the generated SR images were the best images among the generated SR images. Although SRGAN performed well in the MOS metrics there is a significant drop in the general image comparison metrics like Structural Similarity Index Measure (SSIM) and Peak Signal to Noise Ratio(PSNR) which heavily relies on the intensity of the images. In this application, although the spatial details of the images are important, it is also necessary to preserve the spectral reflectance information. Therefore, introducing a GAN-based training framework for this task would result in uncertain predictions in terms of spectral information.

Multi-Branch and UNet architectures

The architectures discussed in previous sections are either sequential(single-stream) or skip-connection designs. Multi-Branch SR architectures are a specific type of architecture that aims to obtain features in multiple contexts in a parallel fashion. For example, the architecture defined by Liu et al. (2018) consists of multiple branches and each branch has different receptive fields of kernel sizes say 3x3, 5x5, 7x7 until 13x13. The final SR image is generated by fusing the features from all these branches. A gradient guided structure preserving SR architecture was proposed by Ma et al. (2021). The architecture has 2 branches each is a Residual in Residual Dense Block(RRDB) architecture. RRDB is inspired from DenseNet based architecture[ Huang et al. (2016)] modified for SR task. The first branch learns features from the actual LR images, and the second branch learns features from the gradient maps of LR images. Gradient maps contain high-level features specifically the prominent edges in an image. These gradient maps are generated by convolving a high pass filter over the LR images. The final SR image is generated by fusing the features generated from LR images and gradients of LR images using 1x1 convolution. 1x1 convolution is a type of feature compression mechanism first introduced in Google LeNet architecture[ Szegedy et al. (2014)] as
2.2. DEEP LEARNING BASED SUPER-RESOLUTION

a method to control the features generated by the intermediate layers to conserve memory without losing more prominent feature information. The structure of this architecture can be seen in Figure 2.2.

Figure 2.2: The Structure Preserving Super Resolution (SPSR) architecture as proposed by MA et al. (2021) containing 2 branches, the main branch works with the RGB image and the gradient branch works with the gradient maps of the RGB image.

The U-Net architecture depicted in Figure 2.3 brought a revolution in the medical imaging domain with exceedingly well performance on medical image segmentation tasks. It was first proposed by RONNEBERGER et al. (2015), and the architecture was based on the standard encoder-decoder model. The encoder path consists of blocks, each block has a pair of convolution layers followed by a non-linear activation layer. At the end of each encoder block, the features are downsampled and the subsequent encoder blocks learn convolution kernels from the downsampled feature maps. In the decoder path, each block upsamples the features until the desired output size is reached. In addition to these blocks, each encoder-decoder pair has a residual connection that concatenates features from the encoder and decoder of the same level. MAO et al. (2016) was the first to use UNet flavored architecture for single image SR tasks. LU and CHEN (2021) also worked on an UNet-based model but they introduced a new loss function called mixed gradient loss (MGL). MGL is calculated between the gradients of HR and SR images and then added with the actual loss function (L2 criterion). In their work, the gradients were generated using sobel filters [KANOPoulos et al. (1988)]. There are numerous flavors of DCNN architectures that perform SR of LR images to generate SR images that are close to HR images. In the next section, the SR works that are explicitly researched for multi-spectral remote sensing images are discussed.
2.3 Super Resolution on Remote Sensing Images

When it comes to the objective of improving the spatial resolution in remote sensing multi-spectral images, research is done with multiple categories. Inter-sensor up-sampling categories convert images from a LR source to SR images using models trained on HR images. The other categories include pan-chromatic sharpening based upsampling techniques, where the same sensor provides images at both LR and HR. These type of sensors collect data to formulate pan chromatic band which is considerably at higher resolution but at gray-scale. These pan-chromatic bands are fused with the multi-spectral bands to generate HR-like multi-spectral images.

2.3.1 Inter-Sensor Super-Resolution

These methods involve 2 different sensors, a source sensor with LR images and a target sensor that provides a HR counterpart which in our case is the landsat-8 and sentinel-2 images respectively. The objective is to train a model that could convert the input LR image to SR image like the other SR tasks mentioned in the previous sections. GALAR et al. (2019) used ResNet based architecture to upsample sentinel-2(S2) images to 5 meter rapideye(RE) image resolution. They designed a series of experiments and concluded that an
SR-ResNet model trained to directly convert S2 images to RE image’s resolution were performing significantly better. However, there weren’t many details about the spectral differences between the two images and efforts taken to bring those difference down.

LATTE and LEJEUNE (2020) did upsampling using the same ResNet architecture. sentinel-2(10m) and planet scope(2.5m) sensors were considered in this study. The proposed ResNet architecture not only achieved SR but also corrected for spectral inconsistencies. The work also focussed on inter-sensor radiometric inconsistencies caused in planet scope satellite constellations. ISA et al. (2021) proposed a DCNN architecture where the input landsat-8 images are converted to a latent space and the latent space vectors are fed to multi-branch CNN architectures each trying to reproduce sentinel-2 like images. This work also did not talk about the spectral differences in the images and methods to handle them.

2.3.2 Pan Sharpening Techniques

Pan Sharpening or Pan-Chromatic sharpening is a special type of upsampling technique that has been a topic of study since 1985 when CLICHE (1985) integrated the pan-chromatic band of SPOT imageries with the multi-spectral(MS) bands to enhance the resolution of MS images. Since then a lot of approaches were developed. WANG et al. (2005) reviewed different classical approaches which include the intensity hue saturation method, brovey transform[GILLESPIE et al. (1987)] and principal component substitution[CHAVEZ et al. (1991)]. These methods although improving the spatial resolution of the multi-spectral images but causes heavy deviation in the reflectance values. After the success of deep learning-based SR architectures in single image applications the same techniques were applied for pan-sharpening applications.

A pan-sharpening problem similar to a SR task aims at finding an inverse of the degradation function similar to the equation 2.2, However, in this case an additional input factor is also added which is the HR pan-chromatic image. Equation 2.2 now becomes,

\[ I_{SR} = F^{-1}(I_{MS}, I_{PAN}; \theta_x) \]  (2.3)

where \( I_{MS} \) is the LR multi-spectral(MS) image, \( I_{PAN} \) is the high resolution pan-chromatic image. \( F^{-1} \) is the DCNN that fuses the HR panchromatic images with the LR multi-spectral images using the parameters \( \theta_x \) to generate HR pansharpened MS image \( I_{SR} \)
MAKI et al. (2016) was the first to employ deep convolution neural networks for the pan-sharpening tasks. An SRCNN-like architecture was trained by concatenating the pan-chromatic band along with the multi-spectral bands up-sampled to the size of intended HR size. Experiments were done using different configurations of SRCNN and the model outperforms the conventional pan-sharpening architectures. YANG et al. (2017) proposed the PanNet architecture, which is similar to the one described by MAKI et al. (2016). The main difference is that the DCNN model learns on the high-level features or the gradients typically the edges of the multi-spectral and panchromatic bands. The actual reflectance details are added at the end of the network by introducing a residual connection. In the works defined above during the training process the input LR images are simulated by downsampling the actual LR-MS images, train the model with these downsampled images and the model is tested with the LR images at actual size. ZHOU et al. (2021) handled the pansharpening problem in a completely unsupervised fashion. This means there are no ground truths involved in training the model. They relied on a two-stream generative multi adversarial network and introduces a novel loss function to facilitate training in an unsupervised setting.

Landsat-8 products provide a pan-chromatic band at 15m resolution and multi-spectral bands at 30m resolution. Sentinel-2 provides some multi-spectral bands at 10m resolution and some bands at 20m resolution. For this objective of transforming landsat-8 images into sentinel-2-like images, a mix of standard inter-sensor SR is required. In combination with that since landsat-8 products also include panchromatic band a pan-sharpening can also be used to enhance the spatial resolution of the landsat-8 image. So to achieve the objectives mentioned in the research questions of this thesis a combination of both the architectures has to be involved. To the best of our knowledge, there exists no work that involves a deep learning based system to harmonize images from two different satellite sources that also performs a SR of a LR image to up-sample it to a HR-like image using a panchromatic band.
Methods

The task at hand is to build a data driven model that transforms input landsat-8 image into a sentinel-2 like image by improving the spatial resolution and reducing the spectral differences.

Figure 3.1: Model Training and Evaluation Pipeline

The overall pipeline for training a DCNN architecture for this task can be found in Figure 3.1. In section 3.1 we will discuss the data acquisition strategies that were followed to acquire a dataset. The section will also talk about the study area and how the study area was chosen for this analysis. In section 3.2 we will discuss all the necessary processing done on the raw image pairs and steps taken to prepare the data to train a deep learning model. The section 3.3 explains different DCNN architectures that were designed for this purpose. In addition to that some explanations about training strategies can also be found in this section. We will conclude this chapter of the thesis by discussing about different evaluation strategies that were designed to demonstrate the robustness of the trained model.
The scope of this thesis is to harmonize only the common multi-spectral bands which can be found in table 1.1. All the mentioned bands follow landsat-8 annotations as shown in table 3.1.

<table>
<thead>
<tr>
<th>Band Name</th>
<th>S2 Annotation</th>
<th>L8 Annotation</th>
<th>Annotation in this thesis</th>
</tr>
</thead>
<tbody>
<tr>
<td>Blue(B)</td>
<td>B2</td>
<td>B2</td>
<td>B2</td>
</tr>
<tr>
<td>Green(G)</td>
<td>B3</td>
<td>B3</td>
<td>B3</td>
</tr>
<tr>
<td>Red(R)</td>
<td>B4</td>
<td>B4</td>
<td>B4</td>
</tr>
<tr>
<td>Near InfraRed(NIR)</td>
<td>B8A</td>
<td>B5</td>
<td>B5</td>
</tr>
<tr>
<td>Short Wave InfraRed 1(SWIR1)</td>
<td>B11</td>
<td>B6</td>
<td>B6</td>
</tr>
<tr>
<td>Short Wave InfraRed 2(SWIR2)</td>
<td>B12</td>
<td>B7</td>
<td>B7</td>
</tr>
<tr>
<td>Pan Chromatic(Pan)</td>
<td>-</td>
<td>B8</td>
<td>B8</td>
</tr>
</tbody>
</table>

3.1 Data Acquisition

For any data driven problem data acquisition is a very crucial step. Section 3.1.1 describes the strategy followed to select the study area. Section 3.1.2 explains the actual data acquisition pipeline. Starting from data search to look for images taken over an area on the same day and strategies to choose optimal image pairs.

3.1.1 Study Area

Similar to any supervised learning approach a model is trained with some specific set of data, some portion of the dataset is held out for validation and testing the performance of the model. This hold out is done to ensure the generalizability of the model. The study area chosen for this work spreads across Germany and the regions are specifically chosen using the military grid reference system(MGRS) at 100km precision. MGRS is the standard tiling system used for sentinel-2 products and the HLS products mentioned in section 2.1.1.
3.1. DATA ACQUISITION

Figure 3.2 shows an overview of tiles selected for this project. The tiles are selected in such a way that the train and test, validation region are apart from each other. This strategy ensures there is no bias in the whole process and it can be claimed with certainty that no same area of interest is used to train and evaluate the models.

3.1.2 Data Search and Acquisition

The spatio temporal access catalogs (STAC) [HANSON (2019)] is an indexing technique developed recently. It is an open-source project which makes remote sensing image search and retrieval an easier process. A lot of open-source research has been going on around this topic and NASA themselves have a STAC API to look-up for images at a give AOI and date. This project leverages sat-search\(^1\) python library to search for optimal images. The overall data search and acquisition pipeline is summarised in Figure 3.3.

\(^1\) https://github.com/sat-utils/sat-search
The data search is initialized for landsat-8 and sentinel-2 level-2 bottom of the atmosphere (BOA) products. As mentioned in section 2.1 BOA products consist images that are corrected for the distortions created by the intermediate atmosphere between the sensor and the earth surface. In this project microsoft planetary computer STAC API\(^2\) is used to perform the meta-data search. For a given AOI all the images available between 01.01.2018 till 01.11.2021 with cloud coverage of less than 20% are queried for both landsat-8(L8) and sentinel-2(S2). The API returns metadata of all the available products that satisfy the query. The meta data for both the sources are stored in tabular format. Each records of meta-data files has details about the acquisition date, cloud coverage percentage, area of coverage and so on. The meta-data table from L8 and S2 search queries are then merged based on the intersecting dates. This gives us a list of image pairs i.e., dates when L8 and S2 sensors passed over an AOI on the same dates.

On an average there were around 14-18 unique dates over a period of 3 years, where the images from the two sensors intersect over each tile. In order to diversify the dataset it was chosen to pick 5 best image pairs from each tiles. The image pairs are prioritized by the area of intersection between the pairs. It is not ideal that all the image pairs have perfect area of intersection, therefore pairs with maximum area of intersection are prioritized. Once the final image

---

\(^2\) https://github.com/microsoft/planetary-computer-apis
pairs are chosen they are downloaded from the individual sources. Landsat-8 images are downloaded using the USGS earth explorer API\(^3\). The sentinel-2 images are downloaded from the publicly available amazon web services S3 bucket\(^4\).

### 3.2 Preprocessing

Both the sentinel-2 and landsat-8 products are structured in such a way that every spectral bands are stored as a separate raster files. Raster files are a specific categories of file format that are used predominantly to store large remote sensing images. Rasters are similar to any image format. However, in addition to the image data each raster data comprises of metadata that represents the bounds of the raster in real world co-ordinate system. The metadata also has details about the pixel sizes, the height and the width of the data. In this section of the thesis, we will discuss about different preprocessing steps that were applied on the acquired rasters before preparing them to train the DCNN architecture.

#### 3.2.1 Reprojection

The earth is ellipsoidal in shape, for easy representation and analysis the images captured over an ellipsoidal surface is re projected into a flat 2-D space. This type of reprojection has been studied for centuries and there are many efficient methods developed over time. Once the image is projected a standardized coordinate reference system(CRS) is maintained for that specific projection based on which locations of each pixels can be described and maintained. A particular CRS is referenced using EPSG code, EPSG is a structured dataset of CRS and co-ordinate transformation developed by the European Petroleum Survey Group(EPSG). In general many mapping systems like Google maps, open street maps etc., follow EPSG:4326 co-ordinate system which by default is represented in degree units. Each raster files has the CRS as a metadata value in them. In general almost all the sentinel-2 products acquired over the study region follows the same CRS (EPSG:32633) across all the raster bands which has a default unit of meters. However, Some of the landsat-8 products followed a different CRS(EPSG:32632), specially the products from the south western part of the study area. The first step of preprocessing was to transform all the landsat-8 raster files to be in the same CRS unit as that of the sentinel-2 counterpart.

\(^3\) [https://pypi.org/project/usgs-api/](https://pypi.org/project/usgs-api/)

\(^4\) [https://registry.opendata.aws/sentinel-2-l2a-cogs/](https://registry.opendata.aws/sentinel-2-l2a-cogs/)
As mentioned in Section 2.1.1 the inter-sensor mis-registration between sentinel-2 and landsat-8 varies based on location. The next step of preprocessing is to correct the mis-registration between the image pairs across all bands. A number of packages has been developed to analyse and correct the inter-sensor mis-registration as mentioned in 2.1, but the objective of the work is to build an efficient pipeline. Therefore it is necessary to include a compatible module in the pipeline that does not tamper much with the existing environment but also could achieve significant level of image to image co-registration. In this work the AROSICS (an Automated and Robust Open-Source Image Co-Registration Software for multi-sensor satellite data) package developed by Scheffler et al. (2017) is used to perform registration analysis and to perform the corresponding spatial shifting. AROSICS is a lightweight python based open-source software which performs co-registration based on pixel shift estimation in frequency domain. In other words, the images are converted into fourier space and the amount of pixel shift between two mis-registered images are calculated by a moving window manner. The sanity of the calculations are verified using various validation and quality estimation metrics. The package was evaluated on multi-sensor image pairs and it was able to achieve noticeable inter-sensor co-registration.

Figure 3.4: spatial shifts calculated by AROSICS package across all image pairs.

The intra-sensor mis-registration are ignored as no significant co-registration issues were observed after manually inspecting some image samples. However, this work includes bands from 3 different products the sentinel-2 level-2 multi spectral bands, landsat-8 level-2 multi spectral bands and landsat-8 level-1 pan-chromatic band. Noticeable mis-registrations were observed among these three products. Although the landsat-8 level-2 multi spectral bands and landsat-8 level-1 pan-chromatic bands are supposed to be iden-
tical, USGS performs a terrain correction during the level-1 to level-2 conversion on the multi spectral bands which is not done on the pan-chromatic band.

![Image of image registration before and after co-registration](image.png)

Figure 3.5: An example patch from co-registered image pair.

It was decided to use the band B4 of sentinel-2 level 2 product as the target image, and the spatial shift across all the bands from the source sensor (landsat-8) with respect to the target image were calculated. The estimated spatial shift along the vertical (Y) axis and horizontal (X) axis across all the image pairs can be observed from the histogram in Figure 3.4. It can be observed that the maximum shift along the X-axis were between 5m-6m along both the directions, considering the fact that the spatial resolution of high-resolution (HR) target image is itself 10m this can be neglected. However, along the Y-axis it varied between -14m to +14m, this is considerable amount of mis-registration assuming that these images are used for precision application. Based upon these calculations the shifts were also performed using the same AROICS library on all the landsat-8 bands. The image pairs are then manually verified for co-registration using visual analysis. Figure 3.5 shows an example patch from an example image pair which shows a significant improvement in the image registration.

### 3.2.3 Cloud Masking and Filtering

Clouds are a major issue in multi-spectral images. The sensors of interest in this study captures images between visible and infrared region of the electro-
magnetic spectrum, which cannot penetrate through cloud cover. Hence, any image infested by cloud cover is typically of no use. The raw rasters of image pairs obtained from the method described in 3.1 covers a region of 185km × 185km for landsat-8 and 290km × 290km for sentinel-2. Hence, finding a cloud free image for such large area is a tough task. Therefore, the images are coupled together with a mask that gives information on the bad pixels present in the image. These bad pixels are not only caused by the clouds but also due to the saturation of the imaging sensor at certain wavelength or no-data pixels. Both landsat-8 and sentinel-2 products themselves come with a quality assessment(QA) band that classifies each pixel in the image into multiple classes like vegetation, water, no data, cloud and so on.

<table>
<thead>
<tr>
<th>Bit</th>
<th>Flag Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td>Fill or No Data</td>
</tr>
<tr>
<td>1</td>
<td>Dilated Cloud</td>
</tr>
<tr>
<td>2</td>
<td>Cirrus</td>
</tr>
<tr>
<td>3</td>
<td>Cloud</td>
</tr>
<tr>
<td>4</td>
<td>Cloud Shadow</td>
</tr>
<tr>
<td>5</td>
<td>Snow</td>
</tr>
<tr>
<td>6</td>
<td>Clear</td>
</tr>
<tr>
<td>7</td>
<td>Water</td>
</tr>
<tr>
<td>8-9</td>
<td>Cloud Confidence</td>
</tr>
<tr>
<td>10-11</td>
<td>Cloud Shadow Confidence</td>
</tr>
<tr>
<td>12-13</td>
<td>Snow/ice Confidence</td>
</tr>
<tr>
<td>14-15</td>
<td>Cirrus Confidence</td>
</tr>
</tbody>
</table>

Table 3.2: landsat-8 Level-2 product cloud mask bit map configuration

Landsat-8 products uses CFMASK[ZHU and WOODCOCK (2012)] algorithm to generate pixelwise cloud classification mask by using thresholds across different multi-spectral and thermal bands. The QA band of landsat-8 is a raster file of 16-bit unsigned integer datatype. According to the description in the product data format control book every bits are assigned with a flag as described in the table 3.2. The images are masked for clouds, cloud shadows and snow at all confidence levels.

Sentinel-2 level-2 products also has a cloud mask band called scene classification layer(SCL). It is generated using a scene classification(SC) algorithm that involves a series of steps which includes band thresholding and index calculations. The SC generation for level-2 products are typically done using the Sen2Cor algorithm which is used to perform atmospheric correction. However, the mask generated from SCL has issues in accurately classifying

5 https://earth.esa.int/eogateway/documents/20142/0/landsat-8-9-OLI-TIRS-Collection-2-Level-2-Data-Format-Control-Book-DFCB.pdf
3.2. PREPROCESSING

the pixel\cite{BAETENS2019}. To compensate that, this study also included s2cloudless\textsuperscript{6} algorithm in the pipeline. S2cloudless is an open-sourced gradient boosting model based cloud detection algorithm that takes all the multi-spectral bands of sentinel-2 image as inputs and generates a cloud mask. Figure 3.6 represents the RGB composite of an example image pair from tile 33UUU, captured on 22-03-2020 with the cloud mask.

\textbf{Figure 3.6:} Sample RGB composite with for an image pair, the semi-transparent green overlay in the images represents the pixel mask, which denotes the presence of clouds, cloud shadow and no data.

3.2.4 Patching

Until this point of pre-processing all the processing is done at tile level. The data of each bands of landsat-8 product at 30m resolutions are represented as image matrices of size $8131 \times 8051$. Using the data of this size for training a DCNN will result in memory overflow issues. Hence the intersecting area across each patches are then subdivided into smaller geometries of size $2560 \times 2560$ with an intersection of 60m in all the directions. Once these geometries are generated, they are filtered for clouds. The cloud filtering is done in such a way that none of the intended patch contains any clouds, cloud shadows or snow. All the valid geometries are filtered and then the patches are generated by cropping the raster from both the products. Each patches from a multi-spectral bands of landsat-8 and sentinel-2 are 6-channel arrays, consisting of common bands B2,B3,B4,B5,B6,B7 as mentioned in table 3.1. The

\textsuperscript{6}https://github.com/sentinel-hub/sentinel2-cloud-detector
height and width of each patch is 256x256 pixels for sentinel-2 and 86x86 pixels for landsat-8. Consecutively, the pan-chromatic band is a single channel array with 192x192 pixels as height to width. Once these patches were extracted they are stored as hdf5 file. After this step there were,

- 16100 patches from training pairs.
- 6440 patches from testing pairs.
- 3040 patches from validation pairs.

### 3.2.5 Further Cloud Filtering

After analysing the generated patches some cloudy patches were identified even after applying the generated cloud masks. This could be caused by the inaccurate false negative classification by the respective cloud mask generation algorithm. The presence of these cloud patches could affect the model training. The following method was developed to identify patches with clouds. The sentinel-2 images were first down-sampled to the size of the landsat-8 counterpart. The S2 and L8 image pairs are then compared against each other using the peak signal-noise ratio (PSNR) and structural similarity index metrics (SSIM) which are described in section 4.2. These two metrics were plotted against each other. In Figure 3.7 it can be observed that there were outliers. Investigating these outliers revealed that the reasons for these are due to the fact that either one of the image pair is affected by cloud or pixel saturation in one of the sensor. The presence of cloud deteriorates the structural information in the patch which resulted in very low SSIM. Similarly clouds also introduces noise in the image thereby resulting in very low PSNR. Based on this analysis a minimum threshold of 22 PSNR and 0.80 SSIM metrics were manually defined to filter out the patch pairs with clouds. After this filtering there were

- 15935 patches for training
- 5586 patches for testing
- 2977 patches for validation
3.2. PREPROCESSING

3.2.6 Band-Pass adjustments

Band-pass adjustment is a process to correct small spectral differences between sensors. As mentioned in Section 2.1, a number of efforts were already made to correct for these differences. The general procedure is to find a linear function that brings the inter-sensor reflectance values closer as described in section 2.1. These linear functions are generated either by directly co-relating the sensors, or by including a hyper-spectral sensor to derive a ground truth and then to convert the reflectances from both sensors to the ground truth. In this work, a direct sensor to sensor mapping approach is employed i.e., regression functions were developed to convert landsat-8 reflectances to sentinel-2. The sentinel-2 images were downsampled to landsat-8 resolution. By doing this each pixel of the downsampled sentinel-2 image corresponds to an approximate representative of the surrounding 8 pixels of the higher resolution sentinel image i.e., one pixel of down sampled sentinel is of area 900 sq. meters (30m × 30m). A dataset to generate the band-pass adjustment functions was created from the train, test and validation by randomly sampling pixels from image pairs and performing a point-to-point mapping between the image pairs. After point to point mapping there were

Figure 3.7: Cloud Filtering using cross-comparison of quality metrics.
• 1,331,280 datapoints for training.

• 369,800 datapoints for testing

• 147,920 datapoints for validating

These data points are organized into a tabular structure and then stored as a parquet file [Vohra (2016)]. A pool of regression models were trained and tested on the datapoints mentioned above. The best model was chosen from these experiments based on the final mean absolute difference, root mean squared error metrics and the prediction time. More details about the experiments are discussed in Section 4.1. The final band-pass function which converts the input landsat-8 reflectance to sentinel-2 counterparts are a series of polynomial equation as shown in table 3.3.

<table>
<thead>
<tr>
<th>Bands</th>
<th>Band-Pass Functions</th>
</tr>
</thead>
<tbody>
<tr>
<td>B2</td>
<td>( y = -0.6593x^3 + 0.2387x^2 + 1.0633x )</td>
</tr>
<tr>
<td>B3</td>
<td>( y = -0.3053x^3 - 0.1678x^2 + 1.0474x )</td>
</tr>
<tr>
<td>B4</td>
<td>( y = -0.6076x^3 + 0.2855x^2 + 1.0035x )</td>
</tr>
<tr>
<td>B5</td>
<td>( y = -0.5295x^3 + 0.4631x^2 + 0.8159x )</td>
</tr>
<tr>
<td>B6</td>
<td>( y = -0.8926x^3 + 0.3253x^2 + 0.8926x )</td>
</tr>
<tr>
<td>B7</td>
<td>( y = -0.5009x^3 + 0.3507x^2 + 0.9636x )</td>
</tr>
</tbody>
</table>

Table 3.3: Band-Pass polynomial equations

### 3.3 Super-Resolution Models

The data is now prepared to train a deep learning architecture that brings the input landsat-8 image as close as possible to sentinel-2 counterpart. This section of the thesis discusses the specifications of intended DCNN architectures that were trained and tested on the preprocessed dataset. In sub-section 3.3.2 we will discuss how the dataloader class of pytorch framework is exploited to achieve lazy loading of the inputs to train the deep learning model. In addition to that, the configurations and the settings for the actual training procedures are discussed in the final sub section 3.3.3.

#### 3.3.1 Deep Convolution Neural Network Architectures

It has been extensively discussed in Sections 2.2 and 2.3 about how different flavours of deep learning based architectures have been developed and modified overtime that were able to achieve unrealistic performances in terms of
perceptual quality of the generated super-resolved image. In this work two
flavours of DCNN architectures were chosen to be tested on the dataset that
has been prepared due to their flexibility and reported superior performances
in various literature [MAO et al. (2016), MA et al. (2021)]. The first model is a
simple U-Net based architecture as described in section 2.2.2 with some mod-
ifications to accommodate the structure of the data in scope. Followed by a
multi-branch RRDB flavoured architecture inspired from the work by MA et al.
(2021) as described in section 2.2.2.

UNet Architecture (UNet-Pan)

U-Net is derived from standard encoder-decoder module structure with resid-
ual connections that shares features from each level of encoder block to the
equivalent decoder block. The proposed pipeline for the prediction of super-
resolved image using the UNet architecture is illustrated in Figure 3.9.

![UNet Architecture](image)

Figure 3.8: Proposed U-Net flavoured architecture to up-sample the input landsat-8
multi-spectral bands by fusing the pan-chromatic band.

Before feeding the multi-spectral and panchromatic inputs, they are resized
to the intended shape of the output super-resolved image. In this case the
intended upsampling is $\times 3$ times along both the dimensions with respect to
the multi-spectral inputs. The pan-chromatic bands are also upsampled to
the same dimensions, bi-cubic interpolation technique was used to perform
this up-sampling. After upsampling both the images to the same shape, they
are then concatenated together. The final input to the U-Net model is a 2-
D 7 channel image. The ordinality of the channels are blue, green, red, NIR,
SWIR1, SWIR2 and pan.

The structured input is then fed into the encoder part of the UNet architecture.
The objective of the encoder block is to convert the input image into a latent
space representation. This is achieved by a series of convolution and down-
sampling steps.

![ReLU and Leaky ReLU activation functions](image)

**Figure 3.9:** An illustration of the ReLU and Leaky ReLU activation functions where \( y \) is the layer inputs, \( f \) is the activation function and \( a \) is the slope of in the negative direction. In case of Leaky ReLU the value of \( a \) is fixed, while for Parametric ReLU it is a trainable parameter.

The basic components of the encoder block consists of convolution layers with a receptive field of \( 3 \times 3 \) and a stride of 1 along all direction. Each convolution layer is followed by a Parametric ReLU (PReLU) activation, which is a modification of Leaky ReLU where rather than having a hard threshold of replacing all values less than zero a slope is introduced that allows soft limit for values less than zero. In Leaky ReLU the slope is fixed and in case of Parametric Relu the slope is considered as a trainable parameter of the model. The slope of the PReLU is similar to the weights of the rest of the convolution blocks which are determined by back-propagation. It is proven to have improved performances in the super-resolution (SR) works by MA et al. (2021), LEDIG et al. (2016) and MAO et al. (2016). Each downsampling block consists of two layers of these convolution PReLU layers, followed by an average pooling of size 2x2 along both the dimensions, that downsamples the features generated from the convolution \( 3 \times 3 + \) PReLU pair. The depth of the proposed UNet architecture is 3 so this conv-downsampling operation is done twice and finally a latent space representation of the final image is obtained.

The objective of decoder block of the architecture is to up sample the latent space representation of the image to reconstruct the image to its original dimension. This is achieved by performing convolution transpose operation followed by a series of \( 3 \times 3 \) convolution and PReLU operations. At each up-sampling block in the decoder section, features from the encoder section counter part is concatenated. This ensures the presence of high level features
generated at the early convolutions in reconstructing the SR image from the latent space representation. The final up sampling block outputs a total of 64 features, a $1 \times 1$ convolution layer is then added at the end that does the feature fusion which finally outputs a 6 channel super resolved landsat-8 image.

**Multi-Branch RRDB architecture (MB-RRDB)**

Ma et al. (2021) in their work proposed the Structure Preserving Super Resolution (SPSR) which consists of 2 branches of Residual in Residue Dense Block (RRDB) architectures. In the main SR branch the image is directly fed and in the secondary (Gradient) branch learns features from the same image but in the gradient space. This kind of architecture has been modified for this work and the SR image prediction pipeline is illustrated in Figure 4.1.

![Diagram](image)

**Figure 3.10: Proposed 2-branch RRDB architecture to up-sample the input landsat-8 Multi Spectral bands**

The basic component of the proposed architecture is the Residual Dense Block (RDB), each RDB consists of a series of convolution layer with kernel size of $3 \times 3$ followed by a PReLU activation layer. Each RDB has 6 such layers and a growth rate of 12 i.e., with each layers 12 features are generated. The features generated in each layers are cascaded with the subsequent layers at the end of each dense block a $1 \times 1$ convolution is used to do a feature compression, in the end each RDB generates 6 feature maps. Similarly, there are 3 RDB in each branch, in the end the features from all the RDB from multi-spectral branch and pan-branch are concatenated together. A final $1 \times 1$ Convolution the combines all the concatenated features into the final intended 6-channel SR outputs.


3.3.2 Data Loader Pipeline

As mentioned in Section 3.2.4 the patches are generated from the original raster files and each patch is stored as a HDF5 file. In total there were 24,498 patches for training, testing and validation, each patch file containing a combination of landsat-8 multi-spectral, sentinel-2 multi spectral and landsat-8 pan-chromatic patch. In addition to the image array it also consists of other meta data like the AOI, date of acquisition and tile information. All the patches has an overall size of 23.4GB, due to resource limitations all these patches cannot be held in the memory at once to train the model. To solve this issue a lazy loading strategy is handled. A pytorch dataset\(^7\) class has been implemented to read the data from the respective patch files. The Dataset class reads a CSV file containing the meta data about each patch files, which also contains the path to a specific patch. The dataset class has other methods that does the necessary transformations, some transformation methods in the dataset class are,

- **DN to BOA reflectance transform**: As mentioned in Section 3.2.4 the raw raster files are huge, for storage efficiency the reflectance values of each pixels in the rasters are encoded and stored as integer data types. These values are called as Digital Numbers (or DNs), DNs as themselves do not represent the actual reflectance informations. To be used for analysis the DNs are converted to Bottom Of Atmosphere (BOA) reflectances. Each satellite data provider has different methods of converting the DNs to BOA reflectance (BOAR). In case of sentinel-2 the BOAR can be determined by dividing the DNs by 10000, for landsat-8 products the DNs are converted using the following equation

\[
BOAR = DN \times OFFSET + BIAS
\]  

(3.1)

the OFFSET and the BIAS factors can be retrieved from the metadata of every landsat-8 product. The patches are initially stored as DNs and are converted to BOAR by the dataset class.

- **Band-Pass Adjustment**: The Band-Pass adjustment equations mentioned in the table 3.3 are applied in this method. Experiments were done by training a model with and without band-pass adjustments.

- **High-Pass Filter**: ZHOU et al. (2021) and YANG et al. (2017) worked on a deep learning based pan-sharpening method in which the pan-chromatic bands are fed into the model after convoluting a high pass

\(^7\) https://pytorch.org/docs/stable/data.html\#torch.utils.data.Dataset
3.3. SUPER-RESOLUTION MODELS

filter through it. High-pass filter emphasizes the fine details in an image like edges. In this work the following high pass filter is used,

\[
\begin{bmatrix}
1 & 1 & 1 \\
1 & -8 & 1 \\
1 & 1 & 1 \\
\end{bmatrix}
\]

The high-pass filter enhances the high-frequency spatial features, the result of applying such a high pass filter is illustrated in the Figure 3.11

Figure 3.11: (a) is an example patch of the pan-chromatic image and (b) represents the extracted high-level feature map generated by passing the High Pass Filter over the image patch.

3.3.3 Training Configuration

As mentioned before the models and data loader pipelines were developed using PyTorch framework (Paszke et al. (2019)). The absolute error (L1) criterion and the squared error (L2) criterion are the most commonly used loss functions to train the SR architectures. Anagun et al. (2019) did a comprehensive study of different loss functions on various DCNN based SR models. In addition to the point to point criterion like L1 and L2 metrics, they also tested SSIM and perceptual loss function on a Single-Image Super Resolution task. The study concluded that usage of L1 based loss function for a SR task is more robust and computationally inexpensive. L2 based loss function resulted in blurry images on the SR outputs. In the end the L1 criterion with a learning rate of \(1 \times 10^{-5}\) on an ADAM optimizer (Kingma and Ba (2014)) is used to train all the architectures. In order to ensure the repeatability of experiments a constant random seed value of 42 is maintained for all the experiments.

The models are trained for a maximum of 100 epochs with a batch size of 10, considering the 15935 image patches for training, each model weights are optimized for 1593 times for each epochs. The models are validated at the end
of each epoch and the weights of the model at the end of best epoch is saved. A learning rate scheduler is used which reduces the learning rate by a factor of 0.5 if the validation loss does not drop by a significant value for over 5 continuous epochs. In addition to all these the model is trained with Automatic Mixed Precision(AMP) where some section of the models are handled as a Float16 datatype and some are handled as Float32 datatype. This efficiently handles the memory and speeds up the training process. The models are trained using an 8GB NVIDIA Tesla M60 GPU. The entire training process is inferred and monitored using Weights and Biases[BIWALD (2020)] tool.

<table>
<thead>
<tr>
<th>Hardware</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>Machine</td>
<td>Amazon EC2 instance - g3.4xlarge</td>
</tr>
<tr>
<td>CPU Mem</td>
<td>122 GB</td>
</tr>
<tr>
<td>GPU</td>
<td>NVIDIA Tesla M60</td>
</tr>
<tr>
<td>GPU Mem</td>
<td>8GB</td>
</tr>
</tbody>
</table>

| Framework      | Deep Learning Framework Pytorch                  |

<table>
<thead>
<tr>
<th>Hyper-Parameters</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>Criterion</td>
<td>L1</td>
</tr>
<tr>
<td>Optimizer</td>
<td>Adam optimizer</td>
</tr>
<tr>
<td>Learning Rate(LR)</td>
<td>$1 \times 10^{-5}$</td>
</tr>
<tr>
<td>LR Scheduler(LRS)</td>
<td>Reduce on validation loss plateau</td>
</tr>
<tr>
<td>LRS patience</td>
<td>5 epochs</td>
</tr>
<tr>
<td>LRS reduction factor</td>
<td>0.5</td>
</tr>
<tr>
<td>Max. Epochs</td>
<td>100</td>
</tr>
<tr>
<td>Model Saving Strategy</td>
<td>save best validation loss</td>
</tr>
</tbody>
</table>

### 3.4 Evaluation Protocol

The SR images generated by the model are evaluated against the ground truth images. In this section we will discuss about the evaluation strategies to test the robustness and the generalizing capability of the DCNN architecture. In subsection 3.4.1 different image quality metrics that are used to evaluate the generated images against the ground truths are described. Raw reflectance values of remote sensing images are significant in various remote sensing image analysis [MUKHOPADHAYA et al. (2018)]. But, nowadays many scientific works utilize the derived bands that provides better insights on the properties of the surface, subsection 3.4.2 discusses about these derived bands and the metrics that are used to evaluate the derived bands that are calculated from the model predicted images. All the metrics in this work are calculated using the skimage [VAN DER WALT et al. (2014)] implementation.
3.4. EVALUATION PROTOCOL

3.4.1 Evaluation Metrics

The spatial upsampling feature of the architecture improves the perceptual quality of the image by enhancing the structural components, however the intensity of the images i.e., the reflectance values has to be maintained. Therefore it is necessary to evaluate the output SR images on a pixel to pixel quality metrics as well as the perception based metrics. To achieve this three commonly used image quality metrics namely, Normalized Root Mean Squared Error (NRMSE), Peak Signal Noise Ratio (PSNR) and Structural Similarity Index Measure (SSIM) are calculated for the image. The metrics are calculated by considering the image patches as whole (6-channel) similar to a Single Image Super Resolution problem. In addition to that the metrics are calculated for individual channels to evaluate the performance of the model on each channel.

The description about all the three evaluation metrics are as follows:

- **Normalized Root Mean Squared Error (NRMSE)** - Mean Squared Error (MSE) is the most commonly used quality metrics to compare two entities. MSE is calculated by taking the average of square of the difference between the original and predicted values of the data. Root Mean Square Error (RMSE) is calculated similar to MSE but instead of taking the average of squares, the average of square roots are considered, to keep the metrics closer to the scale of the variables that are compared. NRMSE is an extension of RMSE where the variables to be compared are standardized into a same scale, there are multiple strategies to normalize the RMSE like using euclidean distance, mean, standard deviation or min-max scaling. In this study RMSE, normalized using min-max scaling is used. Since the coverage of the metrics has to span for the entire range of reflectances, including outlier pixels. The NRMSE of the image is calculated using the following formula,

  \[
  \text{RMSE} = \sqrt{\frac{1}{n} \sum_{i=1}^{n} (I_{HR} - I_{SR})^2} \quad (3.2)
  \]

  \[
  \text{NRMSE} = \frac{\text{RMSE}}{\max(I_{HR}) - \min(I_{HR})} \quad (3.3)
  \]

  where \( I_{HR} \) is the ground truth sentinel-2 High Resolution image, \( I_{SR} \) is the model generated Super-Resolved image from Low-Resolution landsat-8 input and \( n \) is the number of pixels. The lower the NRMSE the more closer is the generated SR image to the Ground truth.

- **Peak Signal Noise Ratio (PSNR)** - PSNR is also based on MSE, it was designed to determine the efficiency of signal compression algorithms. It
is a common metric of comparison for all the SR literatures mentioned in section Anwar et al. (2019). PSNR is the ratio between the maximum possible power of an image and the power of corrupting noise that affects the quality of representation. PSNR is calculated using the following equation,

$$PSNR = 10 \times \log_{10}\left(\frac{(\max(I_{HR}))^2}{MSE}\right)$$

(3.4)

The higher the PSNR the better is the SR image reconstruction with respect to the HR.

- **Structural Similarity Index Measure (SSIM)** - NRMSE is a pixel to pixel loss function, PSNR is derived from MSE which is also a pixel to pixel loss function. In addition to measuring the pixel to pixel difference, a structural metrics should also be measured in order to estimate the efficiency of up-sampling which the model does. SSIM was first introduced in Wang et al. (2004), the objective is to compare the given images and quantify the similarity with a score. The SSIM function returns a score between -1 to 1 with 1 representing same source and target image, -1 representing the source image being the inverse of the target image. This scoring is achieved by a combination of luminance, contrast and structural comparison functions. The SSIM between two images is calculated using the function in Equation 3.5, which is derived by combining the three comparison functions mentioned before.

$$SSIM(x, y) = \frac{(2\mu_x\mu_y + C_1)(2\sigma_{xy} + C_2)}{\left(\mu_x^2 + \mu_y^2 + C_1\right)\left(\sigma_x^2 + \sigma_y^2 + C_2\right)}$$

(3.5)

where $x$ is the ground truth sentinel-2 image($I_{HR}$), $y$ is the model generated SR image($I_{SR}$), $\mu$ is the mean of the respective image and $\sigma$ is the standard deviations of the images. $C_1$ and $C_2$ are arbitrary constants that are used for stability in cases where all the means and standard deviations becomes zero.

### 3.4.2 Derived Bands and Metrics

In general it is difficult to interpret the surface features like vegetation by looking at the raw reflectance values from these remote sensing sources individually. Various indices are calculated by combining different bands of remote sensing sources, to get a high level understanding of the surface features and its characteristics. Normalized Difference Vegetation Index (NDVI) is one such commonly used derived indices calculated from Red and NIR bands of the remote sensing images. NDVI has a value between -1 to 1 without any unit,
where values >0.6 represents a surface with very high vegetation density and values between 0.1-0.2 represents the presence of soil. Similar to NDVI, Normalized Difference Water Index (NDWI) is also a derived index calculated from the the NIR and SWIR bands. NDWI represents the leaf moisture content of the vegetation. Since, the outputs generated by the DCNN architecture will be used to calculate these indices it would be ideal to evaluate the effect of the model on derived indices. NDVI and NDWI are the only indices compared in this study since the calculations does not involve any other hyper parameters. The descriptions, characteristics and formula for calculating these indices are given in table 3.5.

<table>
<thead>
<tr>
<th>Index</th>
<th>Description</th>
<th>Formula</th>
<th>Range</th>
</tr>
</thead>
<tbody>
<tr>
<td>NDVI</td>
<td>Normalized Difference Vegetation Index</td>
<td>$\text{NDVI} = \frac{\text{NIR} - \text{Red}}{\text{NIR} + \text{Red}}$</td>
<td>- Range : -1 to 1</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>- &lt;0.1 – Outliers(clouds, water)</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>- 0.2-0.5 Sparse Vegetation</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>- &gt;0.6 Dense Vegetation</td>
</tr>
<tr>
<td>NDWI</td>
<td>Normalized Difference Water Index</td>
<td>$\text{NDWI} = \frac{\text{NIR} - \text{SWIR}}{\text{NIR} + \text{SWIR}}$</td>
<td>- Range : -1 to 1</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>- -1 to 0 – No Vegetation or No water content</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>- +1 Very High Leaf moisture content</td>
</tr>
</tbody>
</table>

The Derived indices mapping from the generated SR image patch is evaluated against the sentinel-2 HR patch using the Mean Absolute Error (MAE), NRMSE and Pearson Co-Relation, the descriptions of MAE and Pearson Co-Relation index are as follows,

- **Mean Absolute Error (MAE)** - MAE is the average of absolute difference between two variables. The signs of the differences are ignored, the end result is the absolute error between the predicted image and the ground truth. The scale of the error is the same as that of the scale of the intensities of the images that are compared. It is a commonly used metric to evaluate regression problem, in this case the MAE will describe the deviation of the the derived band calculated from the generated SR image from the sentinel - 2 HR image. The formula to calculate the MAE is as follows,

$$\text{MAE} = \sum_{n=1}^{n} \frac{|I_{HR} - I_{SR}|}{n}$$ \hspace{1cm} (3.6)

- **Pearson’s Correlation Coefficient (r)** - Pearson’s Correlation Coefficient (r) estimation technique has been the goto method to determine the linear correlationship between two variables. The $r$ value ranges between -1 to
1 with 1 stating that the points of variable $Y$ is in absolute co-relationship with variable $X$, while a value of -1 states that the variables are at an inverse correlation. The formula to estimate the $r$ value is

$$r = \frac{\sum_{i=1}^{n}(X_i - \mu_X)(Y_i - \mu_Y)}{\sqrt{\sum_{i=1}^{n}(X_i - \mu_X)^2} \sqrt{\sum_{i=1}^{n}(Y_i - \mu_Y)^2}}$$  \hspace{1cm} (3.7)$$

Where $X$ is the derived index mapping calculated from HR images, $Y$ is the derived index mapping calculated from the model generated SR image, $\mu$ is the mean of the respective mapping and $n$ represents the number of pixels.

### 3.4.3 Representation of Metrics - box plot

The distribution of the metrics calculated on the test dataset as mentioned in sections 3.4.2 and 3.4.1 are represented as box plots [McGill et al. (1978)]. A box plot is an efficient way of visualizing a distribution. In addition to displaying the minimum, maximum and median values, the inter quartile range of the distribution can also be visualized. This makes it possible to detect the outliers in the results.

![Example box plot representation](image)

Figure 3.12: Example box plot representation

In this work the distribution of the metrics are represented similar to the box plot in the figure 3.12, the percentage values in the plot represents the quantile values of the sorted metrics. The box between quartile Q1(25%) and quartile Q3(75%) represents a section where majority of the metrics are distributed. The line at quartile Q2(50%) represents the median value of the distribution. Any values beyond the 1% - 99% quantile range are considered outliers.
3.4.4 Test of Significance

Comparing the means of metrics calculated from different methods to estimate the best method is the strategy followed in this work. In cases where the means are very close it is necessary to prove the significance of difference of the means. For this an Independent-Samples t test strategy as described by Ross and Willson (2017) is followed. It is a test for null hypothesis that two independent samples have identical mean value. The p-value is calculated for this test that quantifies the probability that the means of two variable are same. A p-value of 1 represents that the tested variable have an identical mean, thereby proving the determined null hypothesis is true. In cases where the significance of difference has to be determined a threshold of 5% is generally set for the p-value. If the estimated p-value is less than the threshold, it can be proven with certainty that the difference in the means of distributions are significant. In this work, a python based independent t-test function implemented by scipy library [Virtanen et al. (2020)] is used to perform the analysis and to calculate the p-value.
Evaluation of the generated model is a necessary step in any data driven task. It is done to evaluate the performance and the robustness of the architecture in making predictions on data samples that are unseen during the training phase. This chapter is divided into three sections; Section 4.1 discusses the different experiments that are done to generate the Band Pass Adjustment Functions. The results of the trained DCNN architectures are discussed in Section 4.2, finally in Section 4.3, the effects of the DCNN architectures on the derived bands are discussed.

4.1 Band Pass Adjustment model experiments

As mentioned in Subsection 3.2.6 the objective of the Band-Pass Adjustment function is to take a landsat-8 reflectance as input and bring it as close as possible to the sentinel-2 reflectance. To achieve this a training dataset is prepared as mentioned before and experiments are done using different regression functions. In the end the best function is a polynomial regression function for each band as mentioned in 3.3.

The datapoints are taken and then tested on a pool of regression architectures. The pool of architectures includes conventional linear regression (LR) function, a 3$^{rd}$ degree polynomial regression function (PR) and also tree based algorithms. Random Forest Regression (RF) [Ho (1995)] technique which is a type of tree ensemble algorithm is used. Gradient Boosting framework based regression architecture called XG Boost (XGB) regression algorithm [Mitchell et al. (2018)] is also tested out for this task. In addition to these functions the linear transformation functions determined by Zhang et al. (2018) (LR-2) for the transformation of L8 BOAR to S2 BOAR were also tested for the dataset.

The best function is determined by comparing the outputs of the regression models and finding the MAE and the RMSE with respect to the actual reflectance value without any transformation. The final metrics across all bands...
before and after applying the band-pass adjustment functions derived from different regression algorithms on the test datapoints are shown in table 4.1. For reference the 1% to 99% quantile range or reflectances for each bands are also provided in the table.

<table>
<thead>
<tr>
<th>Value Range (1% - 99%)</th>
<th>Metrics</th>
<th>No Correction</th>
<th>LR</th>
<th>LR-2</th>
<th>PR</th>
<th>RF</th>
<th>XGB</th>
</tr>
</thead>
<tbody>
<tr>
<td>B2(B) 0.0058 - 0.0853</td>
<td>MAE</td>
<td>0.00795</td>
<td>0.00758</td>
<td>0.00750</td>
<td>0.00737</td>
<td>0.00736</td>
<td>0.00736</td>
</tr>
<tr>
<td></td>
<td>RMSE</td>
<td>0.01191</td>
<td>0.01137</td>
<td>0.01130</td>
<td>0.01124</td>
<td>0.01120</td>
<td>0.01128</td>
</tr>
<tr>
<td>B3(G) 0.015 - 0.129</td>
<td>MAE</td>
<td>0.00902</td>
<td>0.00905</td>
<td>0.00905</td>
<td>0.00905</td>
<td>0.00905</td>
<td>0.00905</td>
</tr>
<tr>
<td></td>
<td>RMSE</td>
<td>0.01334</td>
<td>0.01311</td>
<td>0.01416</td>
<td>0.01314</td>
<td>0.01316</td>
<td>0.01320</td>
</tr>
<tr>
<td>B4(R) 0.0096 - 0.1651</td>
<td>MAE</td>
<td>0.01002</td>
<td>0.01007</td>
<td>0.01006</td>
<td>0.00983</td>
<td>0.00977</td>
<td>0.00976</td>
</tr>
<tr>
<td></td>
<td>RMSE</td>
<td>0.01589</td>
<td>0.01577</td>
<td>0.01593</td>
<td>0.01575</td>
<td>0.01574</td>
<td>0.01573</td>
</tr>
<tr>
<td>B5(N) 0.0947-0.4787</td>
<td>MAE</td>
<td>0.03585</td>
<td>0.03563</td>
<td>0.03611</td>
<td>0.03522</td>
<td>0.03511</td>
<td>0.03511</td>
</tr>
<tr>
<td></td>
<td>RMSE</td>
<td>0.04523</td>
<td>0.04955</td>
<td>0.04925</td>
<td>0.04826</td>
<td>0.04807</td>
<td>0.04806</td>
</tr>
<tr>
<td>B6(SWIR1) 0.038-0.3664</td>
<td>MAE</td>
<td>0.02556</td>
<td>0.02390</td>
<td>0.02330</td>
<td>0.02322</td>
<td>0.02319</td>
<td>0.02318</td>
</tr>
<tr>
<td></td>
<td>RMSE</td>
<td>0.03480</td>
<td>0.03351</td>
<td>0.03259</td>
<td>0.03248</td>
<td>0.03233</td>
<td>0.03232</td>
</tr>
<tr>
<td>B7(SWIR2) 0.0167-0.2675</td>
<td>MAE</td>
<td>0.01821</td>
<td>0.01541</td>
<td>0.01614</td>
<td>0.01524</td>
<td>0.01524</td>
<td>0.01524</td>
</tr>
<tr>
<td></td>
<td>RMSE</td>
<td>0.02535</td>
<td>0.02303</td>
<td>0.02343</td>
<td>0.02258</td>
<td>0.02244</td>
<td>0.02244</td>
</tr>
<tr>
<td>Prediction Time(Sec)</td>
<td></td>
<td>0.003</td>
<td>0.003</td>
<td>0.006</td>
<td>3.137</td>
<td>0.085</td>
<td></td>
</tr>
</tbody>
</table>
4.2 SR DCNN results

As mentioned in section 3.2.4 around 5586 patch pairs were held out from the training process which accounts to about 23% of the entire dataset. These images in test set will be used to test the robustness of the model in predicting an unseen data. In this work the SR images generated using the DCNN architectures mentioned in subsection 3.3.1 are evaluated against baseline upsampling techniques like Bicubic and Bilinear interpolation. During the initial experiments the DCNN architectures were trained on input image without applying the bandpass functions. The best architecture among the two proposed are tested on the held out test dataset. Further experiments were conducted using the best DCNN architecture which includes the evaluation of performance differences caused by the band-pass adjustments and the usage of high-pass features of pan-chromatic bands.
4.2.1 Initial Experiments

For the initial experiments, the UNet-Pan and the MB-RRDB architectures are trained based on the pipeline mentioned in sub-section 3.3.3. The NRMSE, SSIM and PSNR between the outputs of the models with respect to the sentinel-2 ground truths are calculated and they are compared against the bicubic and bilinear interpolation techniques. The calculated metrics for different methods can be found in Table 4.2 and the distributions across the test patches can be seen in the box plots in Figure 4.2.

Table 4.2: Mean metrics on the test set patches with the standard deviations.

<table>
<thead>
<tr>
<th>Method</th>
<th>SSIM</th>
<th>PSNR</th>
<th>NRMSE</th>
</tr>
</thead>
<tbody>
<tr>
<td>Binlinear</td>
<td>0.8756 ± 0.050</td>
<td>31.39 ± 1.81</td>
<td>0.178 ± 0.05</td>
</tr>
<tr>
<td>Bicubic</td>
<td>0.8776 ± 0.033</td>
<td>31.33 ± 1.85</td>
<td>0.179 ± 0.05</td>
</tr>
<tr>
<td>UNet-pan</td>
<td>0.9211 ± 0.018</td>
<td>33.27 ± 1.85</td>
<td>0.144 ± 0.04</td>
</tr>
<tr>
<td>MB-RRDB</td>
<td>0.9016 ± 0.022</td>
<td>32.27 ± 1.74</td>
<td>0.161 ± 0.04</td>
</tr>
</tbody>
</table>

Based on the metrics in Table 4.2 and the distributions shown by box plots in Figure 4.2 it is clearly evident that the performance of the UNet-Pan architecture is superior compared to the other methods with almost 5% improve-
ment in the mean SSIM on the overall test dataset compared to the baseline bicubic interpolation technique. The MB-RRDB architecture outperforms the baseline techniques but still performs relatively worse than the UNet-Pan architecture. The plots shown above represents the metrics that are calculated by considering the image patches as whole (6-channels). However, a deeper channel-wise analysis also shows similar results, which are shown in Figure 4.3.

In order to confirm this observation the RGB composites from the patches with best NRMSE, median NRMSE and the worst NRMSE from the UNet-Pan based upsampling method are plotted. The RGB composites are generated from bands the red, green and blue bands of both the image sources. Each bands are normalized using a min-max scaler where the min and max values are the 1 and 99 percentiles from the sentinel-2 image. It can be observed
in Figure 4.4 that in all three cases both the DCNN based architectures effectively upsampled the landsat-8 MS image by fusing it with the panchromatic band. Significant improvement in the structural metrics as well as the pixel-wise NRMSE were observed. The UNet-Pan architecture performed best in almost all cases, except for the worst NRMSE case where the SSIM of the image generated by UNet-Pan model is better than MB-RRDB architecture while the NRMSE is relatively poor.

In order to get a deeper in-sight on the predictions generated by the model, all the bands of a patch has to be visualized. Therefore the median NRMSE patch is selected and a deeper dissection of the generated image across the individual bands are plotted in Figure 4.5. The upsampled median NRMSE patch using the bicubic interpolation technique and UNet-Pan architectures are compared. For better understanding the absolute difference maps of the upsampled images with respect to the ground truth sentinel-2 images are plotted along the side. It can be observed in Figure 4.5 that the UNet-Pan based upsampling technique significantly reduces the absolute difference between the upsampled image and the ground truth than the baseline bicubic interpolation.

Based on all the observations provided it can be concluded that the UNet-Pan based upsampling method overall performs the best in almost all the cases.
Considering all the fact, the UNet-Pan architecture is considered the best working model and further experiments and evaluations are done by considering it as base.

Figure 4.5: Bandwise tiles from the patch with median NRMSE, (a) - Ground Truth sentinel-2 patch, (b) - Bicubically upsampled landsat-8 patch, (d) - landsat-8 patch upsampled using UNet-Pan architecture, the differences column represents the absolute difference mapping of the respective images.

### 4.2.2 Effects of Band-Pass Adjustments and High Pass Filter

In the previous section the best architecture is determined by evaluating the patches in the test dataset without performing any transformations. For the next phase of experimentation the band-pass adjustment functions and the highpass filters are applied to the bands and the pan-chromatic bands respectively. The experiments are done in a sequential manner as mentioned in table 4.3, for the first experiment the band-pass functions are applied to the input
image at Low Resolution and are upsampled using bicubic interpolation. An improvement in the NRMSE and SSIM values are observed (B) compared to directly upsampling the LR image (A).

Table 4.3: Metrics from Experiments done with High-Pass filter and Band-Pass Adjustments.

<table>
<thead>
<tr>
<th></th>
<th>NRMSE</th>
<th>SSIM</th>
<th>PSNR</th>
</tr>
</thead>
<tbody>
<tr>
<td>A</td>
<td>Bicubic</td>
<td>0.1796 ± 0.051</td>
<td>0.8776 ± 0.033</td>
</tr>
<tr>
<td>B</td>
<td>Bandpass + Bicubic</td>
<td>0.1724 ± 0.050</td>
<td>0.8794 ± 0.032</td>
</tr>
<tr>
<td>C</td>
<td>UNet-Pan</td>
<td>0.1445 ± 0.044</td>
<td>0.9211 ± 0.018</td>
</tr>
<tr>
<td>D</td>
<td>Bandpass + UNet-Pan</td>
<td>0.1498 ± 0.043</td>
<td>0.9162 ± 0.018</td>
</tr>
<tr>
<td>E</td>
<td>Bandpass + UNet-Pan - Retrain</td>
<td>0.1393 ± 0.042</td>
<td>0.9249 ± 0.017</td>
</tr>
<tr>
<td>F</td>
<td>High-pass(pan) + UNet-Pan - Retrain</td>
<td>0.1474 ± 0.047</td>
<td>0.9184 ± 0.022</td>
</tr>
</tbody>
</table>

In the next step the band-pass functions are applied before feeding the LR images into the UNet-Pan architecture (D), a minor drop in the model performances are observed when compared to directly predicting the image using the UNet-Pan architecture (C). Therefore the architecture is then retrained by applying the band-pass function during the training phase itself (E), after retraining the performance of the model improved. In addition to that the metrics of the outlier patches improved as observed in figure 4.6 For the final experiment the UNet-Pan architecture is then retrained by passing a high-pass filter to the pan-chromatic inputs(F) during the training phase as mentioned in section 3.3.2. This reduced the model performance and therefore the UNet-pan architecture that has been retrained with band-pass function transformed multi-spectral inputs and the pan-chromatic band reflectance model is considered the best architecture and is used for further evaluation. However, all the experiments that involved a UNet-Pan based Upsampling technique performed better than the baseline bicubic interpolation technique.

The difference across the experiments that involved the UNet-Pan methods mentioned in table 4.3 are very less(< 1% SSIM) therefore an independent t-test as mentioned in 3.4.4 was done for all combinations of these results and the p-value of all the tests were very much less than 5%. Therefore, the difference in metrics generated by all the experiments mentioned in the table are all significant.
Figure 4.6: Distributions of the NRMSE and SSIM metrics calculated from different experiments shown in table 4.3, the annotations in the X-axis follow the first column of table 4.3.
4.3 Evaluation of Derived Bands

The best performing pipeline involves the transformation of LR images using the band-pass functions and then upsampling the LR image using the UNet-Pan architecture as described by the results in section 4.2.2. In this section the robustness of the above mentioned pipeline is tested by evaluating the derived bands (3.4.2) from model generated images. This section is divided into two parts, in section 4.3.1 the derived bands calculated from the test patches after upsampling the LR images are compared. Followed by which in section 4.3.2 a use case of remote sensing based field observation method is tested by including the pipeline mentioned above.

4.3.1 Derived Bands Test Metrics

For this evaluation, the LR MS images from the test dataset are upsampled and then NDVI and NDWI vegetation indices are calculated using the formula given in table 3.5. The metrics defined in section 3.4.2 are calculated for the generated mappings and the mappings from the ground truth sentinel-2 images. The results can be found in table 4.4 and the box plots in figure 4.7.

Table 4.4: Evaluation metrics on the test image patches stating the difference between vegetation indices generated from sentinel-2 and landsat-8 patches after upsampling.

<table>
<thead>
<tr>
<th>Method</th>
<th>NDVI MAE</th>
<th>NDVI Pearson Correlation</th>
<th>NDVI MAE</th>
<th>NDVI Pearson Correlation</th>
</tr>
</thead>
<tbody>
<tr>
<td>Bicubic</td>
<td>0.0489 ± 0.015</td>
<td>0.895 ± 0.071</td>
<td>0.0412 ± 0.0102</td>
<td>0.9282 ± 0.043</td>
</tr>
<tr>
<td>Bilinear</td>
<td>0.0504 ± 0.015</td>
<td>0.895 ± 0.069</td>
<td>0.0411 ± 0.0104</td>
<td>0.9231 ± 0.046</td>
</tr>
<tr>
<td>Bandpass + UNet-Pan - Retrain</td>
<td>0.0371 ± 0.010</td>
<td>0.942 ± 0.043</td>
<td>0.0307 ± 0.0072</td>
<td>0.9561 ± 0.032</td>
</tr>
</tbody>
</table>
4.3. EVALUATION OF DERIVED BANDS

According to the distributions in the boxplots in figure 4.7 the table 4.4, the proposed UNet based pipeline significantly reduces the differences and improves the co-relation between the vegetation indices mapping calculated from the sentinel-2 and upsampled landsat-8 images. Figure 4.8 shows the NDVI and the NDWI mapping from an example patch from the test dataset. The plots also show a significant improvement in the spatial resolutions of both the index mapping. It can also be observed that the spatial details in the NDWI maps are relatively poor on comparison to the NDVI maps, even in the ground truth sentinel-2 image. This is due to the fact that the NDVI is calculated from bands Red and NIR in which NIR is captured at 10m resolution and B5 at 20m resolution in terms of sentinel-2. However, the NDWI is calculated using bands NIR and SWIR both of which are acquired at 20m spatial resolution. Since the architecture is trained with a reference of this type, the NDWI mapping of the model generated image also follows the same trend.
Figure 4.8: The distribution of test metrics calculated between ground truth sentinel-2 and landsat-8 images upsampled using different methods.

### 4.3.2 Time Series analysis

The general use-case of these derived bands in field observation involves the consistent monitoring of the indices overtime. This gives an overview of the field behaviour due to seasonal changes like crop growth, flowering, senescence or any abnormalities caused by plant stress. In this analysis, the proposed pipeline is tested on such a field observation scenario. For this two fields were randomly chosen within Germany, out of the study area to demonstrate the robustness of the pipeline. The locations of the chosen fields can be found in figure 4.9.
Once the AOI of the fields are determined, a search query was executed over this zone for a time period of around two years starting from April, 2019 till July, 2021. The search query is executed for images from both landsat-8 and sentinel-2 image sources. The details about the number of cloud-free images and the descriptions about the chosen fields can be found in table 4.5. In Field 1 over Heidelberg, Germany, there were 46 images from sentinel-2 and 30 images from landsat-8. The cumulative count of images are calculated by considering the sentinel-2 image as a selected image source for days where image from both L8 and S2 sources co-exist. So there are a total of 76 cloudfree images for Field 1 out of which there are 4 days where the L8 and S2 image co-exist, which makes the cumulative count of 72 images on 72 unique dates across the given time frame. Similarly for field 2 there are 33 S2 images and 26 L8 images and the cumulative count is 51. Based on this observations, it is quite evident that including landsat-8 images in the analysis pipeline increases the temporal availability of observations from a remotely sensed source. Infact, in both the cases the cumulative count increases by a factor of more than 60% in comparison to using only the S2 images.

The mis-registration in the acquired landsat-8 images are then corrected, the band pass functions are applied and then upsampled using the UNet-Pan architecture. Once all this are done the NDVI of the fields are then calculated,
the mean NDVI or the field is then plotted as a time series. The NDVI value increases during the growing season of the crop and then drops typically when the crop matures and are ready to be harvested. Similar patterns are observed in both the fields considered and almost two complete crop life cycle can be observed on the two fields across the span. The time series plots of the fields can be seen in Figure 4.10. The gaps in the time series are because of the presence of clouds or snow. From these plots it is evident that introducing the proposed model based pre-processing pipeline significantly reduces the jumps in the observed NDVI values and synchronised the values over the time series. This phenomenon is very evident in images with low NDVI i.e., images with relatively less vegetation.

<table>
<thead>
<tr>
<th>ID</th>
<th>Location</th>
<th>Field Area $(km^2)$</th>
<th>Number of Cloud Free Images</th>
<th>Cummulative count</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td></td>
<td>Sentinel-2</td>
<td>Landsat-8</td>
</tr>
<tr>
<td>field 1</td>
<td>Heidelberg, Germany</td>
<td>0.44934</td>
<td>46</td>
<td>30</td>
</tr>
<tr>
<td>field 2</td>
<td>Hessen, Germany</td>
<td>0.136133</td>
<td>33</td>
<td>26</td>
</tr>
</tbody>
</table>

Figure 4.11 Represents a sequence of NDVI mapping accompanied by the RGB mosaic observing the growth of crops in Field 1. The L8 images are replaced by the upsampled L8 images and the mappings are also calculated from these upsampled L8 image. The L8 images are interpolated using Nearest interpolation technique for representation, however the calculation of mean NDVI of a Field is done without any resizing or interpolation. The spatial resolution of the generated NDVI maps are significantly improved in addition to that the consistancy of the maps with respective to the sentinel-2 images are maintained by applying the proposed pre-processing pipeline to the LR landsat-8 images.
4.3. EVALUATION OF DERIVED BANDS

Figure 4.10: NDVI time series for field 1 and field 2, the NDVIs are generated from sentinel-2(S2-NDVI) and the upsampled landsat-8(L8UNet-NDVI) images, for reference the NDVI calculated from the raw landsat-8 images without any upsampling(L8-NDVI) are also included.
Figure 4.11: Time Series of NDVI mapping from field 1.
This chapter summarises and concludes this thesis work, the research questions for this thesis work are answered based on all the results that were presented across the previous chapters.

5.1 Answers to the Research Questions

5.1.1 RQ1: How to harmonize images from both the sources and super-resolve Landsat-8 images to Sentinel-2 size?

An extensive literature survey has been done on previous works that harmonises images from multiple remote sensing image sources. A comprehensive pre-processing pipeline has been employed in this work which includes correcting the inter-sensor registration errors, cloud masking, patching to make the training process scalable and Band-Pass adjustments that brings the spectral reflectances of the images closer. Unlike starting from a level-1 (Top the Atmosphere) source as described in all the previous study, this work explores the possibility to harmonize the common bands of the level-2 (Bottom of the Atmosphere) products from both the image sources. In this work, a deep learning based approach was tested that reduces the degradations in the landsat-8 image caused by differences in the spectral response functions and spatial resolutions of the sensor with respect to the sentinel-2 image. To train the intended Deep Convolution Neural Network (DCNN), the images that were taken on the same day over the AOI are exploited. The sentinel-2 image taken over the AOI on a day is considered as ground truth and a landsat-8 image with comparatively lower resolution is upsampled with the help of the DCNN architecture. Significant improvements in the spatial resolution as well as the reduction in the radiometric difference between the landsat-8 and sentinel-2 images were observed.
5.1.2 RQ2: What are the existing super-resolution techniques that improves the spatial resolution of images? How are their performances on remote sensing dataset?

The literature survey shows a number of DCNN architectures that improves the quality of the LR images. There are many recent developments in GAN based architectures that improves the perceptual quality of the images but the results from these methods fails in traditional image comparison metrics that heavily relies on the intensity of the images, for this reason a GAN based approach is avoided and two well established generator DCNN architectures were experimented for this application. The dataset for this specific task is generated from open-source resources and the scope of the dataset covers a wide region in Germany. The generated dataset consists of area with a wide range of features like agricultural fields, vegetation, urban region, water bodies and so on. The UNet based architecture and the Multi-Branch Residual in Residual Dense Block architectures were modified specifically for this task. The architectures were trained from scratch and rather than just upsampling the multi spectral (MS) bands of the LR images, the SR images are generated by fusing the LR MS images with a medium resolution pan-chromatic band. The architectures that were trained improved the perceptual as well as the pixel-wise quality of the images in the test dataset. The UNet based framework improved the structural metric (SSIM) by about 5% in comparison to the the baseline up-sampling techniques. Visually inspecting the RGB composites of the generated images and comparing the image differences also supports the obtained results. Experiments were conducted by using the High-Pass filters on the pan-chromatic images however minor drop in the final evaluation metrics were observed as shown in 4.2.2. The performance of this architecture is tested by a series of evaluations, done on the multi-spectral bands as well as the derived bands of NDVI and NDWI, calculated from these multi-spectral bands which are discussed in 4.3. A real time use-case of the proposed pipeline was also tested, by observing the evolution of vegetation over time on two randomly chosen field outside the scope of the dataset in which the whole DCNN training process is also tested in section 4.3.2 worked on. All the results demonstrate that including the DCNN based pre-processing pipeline before including a LR landsat-8 image for an analysis along with a HR sentinel-2 image improves the harmonization between the images from the two different sensors. However, the time complexity this brings in significantly increases the overall processing time. Therefore it is recommended to use this pipeline on-demand, where patches of the images can be upsampled and then used for analysis. Unlike many other inter-sensor harmonization works like SWINEN et al. (2022), CLAVERIE et al. (2018), NGUYEN et al. (2020) and so-on, this work proposes a deep learning based pipeline. This opens up a wide range of flex-
ibility offered by DCNN models, like transfer learning the same architecture for a different satellite pair.

5.2 Summary

In this thesis, the possibility of reducing the difference between the landsat-8 image with respect to the sentinel-2 images were explored. The effectiveness of this pipeline has been demonstrated by evaluating it with some baseline techniques. The evaluation metrics shows significant reduction in the spectral and spatial difference. The robustness of performance of the pipeline is also demonstrated by performing a use case analysis, where the vegetation of a field is monitored over time. The pipeline reduced the temporal variation in the derived bands caused by the differences in the images. The proposed framework is scalable, completely open-sourced and could easily be modified for any other satellite sensor pairs.
Future Work

The performance improvements demonstrated by the DCNN architectures opens up a wide variety of possibilities that can be explored using this approach. The results from the models and also the overall image processing pipeline can be improved by considering the following ideas:

• The results from the Band-Pass functions experiments in section 4.1 makes it highly evident that the direct Band-Pass function to transform the reflectances from one sensor to the other is ill-posed and the solution could vary based on the terrain difference, seasonal differences and so on. Using reflectances from hyper-spectral images as a base and transforming the landsat-8 and sentinel-2 images as close to the base as possible could be an effective technique and has been done in many recent works[CLAYERIE et al. (2018), SCHEFFLER et al. (2020)]. However, no significant amount of work has been done with the Surface Reflectance products from both the image sources. Hence, the transformation co-efficients for this task from such kind of experiments could not be obtained.

• It would be interesting to look at the generalizing capability of the proposed pipeline on datasets with wider scope. The inclusion of more data and increasing the parameters of the DCNN architecture in combination with more efficient hyper-parameter tuning could further improve the results.

• This work only included images from landsat-8 sensor, however landsat-9 has been recently commissioned. The specifications OLI-2 Multi Spectral imaging sensor of landsat-9 is very similar to that of landsat-8. It would be very interesting to see the performance of this framework with landsat-9 images and if required prior transfer learning has to be performed on the DCNN architecture.
• In this work only the L1 criterion was used to train the DCNN architecture. There are many other hybrid loss functions which might improve the metrics even more. One such is the perceptual Loss Function where in addition to comparing the super-resolved image with high-resolution image directly the features generated by a secondary architecture on these images are compared. This secondary architecture that works on only the common bands of landsat-8 and sentinel-2 sensors are not publicly available to the best of our knowledge.

• Explainable AI (XAI) is the recent trend in the Artificial Intelligence domain. The objective of these work includes building a framework that not only generate predictions from a model, but also explains the reasons for prediction. By doing this the intentions of the models in generating a prediction can be monitored and measures can be taken at timely fashion if the model is not working as desired. This work could use such kind of framework that builds the confidence of the user about the model predictions.
A.1 Significance of co-registration

It has been discussed earlier about the existing mis-registration between the landsat-8 multi-spectral bands and the pan-chromatic band. This is caused due to the levels of these two bands, the MS bands are from level-2 (L2) and the pan band is from level-1. The mis-registration occurs due to the backend processing that are being done to generate the L2 MS bands. However, USGS does not do L2 processing on the pan-chromatic band. To prove the significance of the co-registration between these two images an analysis was done, where a random patch from the test dataset was processed through the pipeline. In the first case, the co-registration between the MS and Pan bands are skipped and in the next step the co-registration processing was included. The results from the these two test cases can be seen in figure A.1. Upsampling the image without co-registering the L8 MS bands and L8 Pan band resulted in a ghosting effect on the predicted images especially on the field boundaries where the significant changes in the surface properties occur. This phenomenon is observed in many other test image patches. Hence it is highly necessary to co-register the MS bands and the pan-chromatic band, either with each other or with one of the closest sentinel-2 image when using this system for time series analysis.
Figure A.1: upsampled L8 image with and without co-registration pre processing step.
<table>
<thead>
<tr>
<th>Figure</th>
<th>Description</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>1.1</td>
<td>Spectral characteristics of the common bands from landsat-8(OLI) and sentinel-2(MSI) sensors.</td>
<td>3</td>
</tr>
<tr>
<td>2.1</td>
<td>USGS HLS product generation pipeline</td>
<td>9</td>
</tr>
<tr>
<td>2.2</td>
<td>The Structure Preserving Super Resolution (SPSR) architecture as proposed by Ma et al. (2021) containing 2 branches, the main branch works with the RGB image and the gradient branch works with the gradient maps of the RGB image.</td>
<td>15</td>
</tr>
<tr>
<td>2.3</td>
<td>The UNet architecture first described by Ronneberger et al. (2015) for biomedical segmentation task.</td>
<td>16</td>
</tr>
<tr>
<td>3.1</td>
<td>Model Training and Evaluation Pipeline</td>
<td>19</td>
</tr>
<tr>
<td>3.2</td>
<td>Train, Test and Validation tiles with corresponding tile name</td>
<td>21</td>
</tr>
<tr>
<td>3.3</td>
<td>Data Search Pipeline Overview</td>
<td>22</td>
</tr>
<tr>
<td>3.4</td>
<td>spatial shifts calculated by AROSICS package across all image pairs.</td>
<td>24</td>
</tr>
<tr>
<td>3.5</td>
<td>An example patch from co-registered image pair.</td>
<td>25</td>
</tr>
<tr>
<td>3.6</td>
<td>Sample RGB composite with for an image pair, the semi-transparent green overlay in the images represents the pixel mask, which denotes the presence of clouds, cloud shadow and no data.</td>
<td>27</td>
</tr>
<tr>
<td>3.7</td>
<td>Cloud Filtering using cross-comparison of quality metrics.</td>
<td>29</td>
</tr>
<tr>
<td>3.8</td>
<td>Proposed U-Net flavoured architecture to up-sample the input landsat-8 multi-spectral bands by fusing the pan-chromatic band.</td>
<td>31</td>
</tr>
<tr>
<td>3.9</td>
<td>An illustration of the ReLU and Leaky ReLU activation functions where $y$ is the layer inputs, $f$ is the activation function and $a$ is the slope of in the negative direction. In case of Leaky ReLU the value of $a$ is fixed, while for Parametric ReLU it is a trainable parameter.</td>
<td>32</td>
</tr>
</tbody>
</table>
APPENDIX B. LIST OF FIGURES

3.10 Proposed 2-branch RRDB architecture to up-sample the input landsat-8 Multi Spectral bands ................................................. 33

3.11 (a) is an example patch of the pan-chromatic image and (b) represents the extracted high-level feature map generated by passing the High Pass Filter over the image patch. .................................. 35

3.12 Example box plot representation ........................................ 40

4.1 Point to point mapping of test datapoints before and after transforming the landsat-8 reflectance values using the 3rd degree polynomial functions described in table 3.3. ........................................... 45

4.2 The distribution of test metrics calculated between ground truth sentinel-2 and landsat-8 images upsampled using different methods. ................................................................. 46

4.3 The NRMSE calculated between ground truth sentinel-2 and landsat-8 images upsampled using different methods on all 6 bands individually. .................................. 47

4.4 RGB composites of the patches that were generated by different upampling techniques .................................................. 48

4.5 Bandwise tiles from the patch with median NRMSE, (a) - Ground Truth sentinel-2 patch, (b) - Bicubically upsampled landsat-8 patch, (d) - landsat-8 patch upsampled using UNet-Pan architecture, the differences column represents the absolute difference mapping of the respective images. ................................................. 49

4.6 Distributions of the NRMSE and SSIM metrics calculated from different experiments shown in table 4.3, the annotations in the X-axis follow the first column of table 4.3. ......................... 51

4.7 The Mean Absolute Error(MAE) and the Pearson Correlation(Pearson C) between the NDVI and NDWI maps generated by baseline image interpolation techniques and UNet-PAN(E) base upsampling techniques. ........................................... 53

4.8 The distribution of test metrics calculated between ground truth sentinel-2 and landsat-8 images upsampled using different methods. .................................................. 54

4.9 An extension of figure 3.2 with the locations of the fields chosen for this analysis. .................................................. 55
4.10 NDVI time series for field 1 and field 2, the NDVI's are generated from sentinel-2(S2-NDVI) and the upsampled landsat-8(L8UNet-NDVI) images, for reference the NDVI calculated from the raw landsat-8 images without any upsampling(L8-NDVI) are also included.  

4.11 Time Series of NDVI mapping from field 1.  

A.1 upsampled L8 image with and without co-registration pre processing step.
List of Tables

1.1 Central wavelengths and resolutions of different bands from sentinel-2 and landsat-8 with common bands and highest resolution in bold. ................................................................. 3

3.1 Bands Annotations followed in this thesis L8 - landsat-8, S2 - sentinel-2 ................................................................. 20

3.2 landsat-8 Level-2 product cloud mask bit map configuration . . 26

3.3 Band-Pass polynomial equations ........................................ 30

3.4 Training configurations and hyper parameters ................. 36

3.5 Derived Indices Description .............................................. 39

4.1 Evaluation metrics on the test datapoints stating the difference between sentinel-2 and landsat-8 reflectances of using different transformation functions. ........................................ 44

4.2 Mean metrics on the test set patches with the standard deviations. 46

4.3 Metrics from Experiments done with High-Pass filter and Band-Pass Adjustments. ............................................. 50

4.4 Evaluation metrics on the test image patches stating the difference between vegetation indices generated from sentinel-2 and landsat-8 patches after upsampling. ....................... 52

4.5 Details about the selected fields ....................................... 56


[BIEWALD] **Experiment Tracking with Weights and Biases.** Software available from wandb.com.


