SAARLAND UNIVERSITY

Faculty of Mathematics and Computer Science
Department of Computer Science
MASTER THESIS

Robust Cloud Segmentation in Satellite Images using Multi-view learning

submitted by
Rushan Mukherjee 7015520
Saarbrücken
March 2025

Advisor:
Francisco Mena
German Research Center for Artificial Intelligence
Trippstadter Str. 122
Kaiserslautern, Germany

Reviewer 1: Prof. Dr. Antonio Krüger
Reviewer 2: Dr. Marcela Charfuelan

Saarland University
Faculty MI – Mathematics and Computer Science
Department of Computer Science
Campus - Building E1.1
66123 Saarbrücken
Germany

Erklärung

Ich erkläre hiermit, dass ich die vorliegende Arbeit selbständig verfasst und keine anderen als die angegebenen Quellen und Hilfsmittel verwendet habe.

Statement

I hereby confirm that I have written this thesis on my own and that I have not used any other media or materials than the ones referred to in this thesis

Einverständniserklärung

Ich bin damit einverstanden, dass meine (bestandene) Arbeit in beiden Versionen in die Bibliothek der Informatik aufgenommen und damit veröffentlicht wird.

Declaration of Consent

I agree to make both versions of my thesis (with a passing grade) accessible to the public by having them added to the library of the Computer Science Department.

Saarbrücken, 12.03.2025 Rushan Mukherjee

(Datum/Date) (Unterschrift/Signature)

Abstract

This thesis presents a robust multi-view deep learning algorithm for cloud and cloud shadow segmentation in satellite images. The detection and removal of clouds and cloud shadows are a crucial pre-processing step in remote sensing. The interference of clouds makes it difficult to observe the earth surface for various use cases such as crop yield prediction, flood prediction and land-use land-cover classification. Traditional algorithms to remove clouds from satellite images are not robust and fail when presented with a diverse range of cloud and cloud shadow types. Therefore, a comprehensive deep learning-based model is required to accurately detect clouds in satellite images and remove them. Subsequently, a Multiview fusion model is proposed which has a wide range of input features and can precisely segment the clouds and their shadow boundary regions. A state-of-the-art global multi-view dataset namely CloudSEN12 has been used for training and testing the model. It consists of Sentinel-1 Synthetic Aperture Radar data, Sentinel-2 optical information and other auxiliary information such as Digital Elevation Maps. The proposed model performs well in hard to classify regions like polar climates and accurately differentiates between clouds and snow. Robustness experiment simulates a band dropout scenario where successive bands of Sentinel-2 stop functioning. The model is able to perform adequately in such a scenario where the full information is not available and the model has to work with limited data.

Contents

Chapter 1
Introduction

Remote sensing is the acquisition of information about the Earth's surface and atmosphere through the analysis of data collected by sensors that are not in physical contact with the object or area being observed. These sensors can be mounted on various platforms, including satellites, aircraft, and drones. Remote sensing is used in a wide range of applications, such as land use and land cover mapping, natural resource management, disaster response, and environmental monitoring. By analysing the reflected sunlight from the Earth's surface, remote sensing can provide valuable insights into the physical, biological, and chemical properties of the target area, making it a powerful tool for Earth observation and analysis. Clouds are an integral part of Earth's atmosphere and their importance in maintaining life on Earth cannot be discounted. Clouds come in various shapes and sizes and they can form a huge hindrance for scientists observing the Earth surface via satellites. This is because most satellites have optical sensors onboard which rely on light getting reflected off of the earth's surface. Clouds and cloud shadows obscure the optical image information that one might need by reflecting the sunlight before it is able to reach the Earth's surface. Zantedeschi et al [30] provide us with a comprehensive benchmarking dataset for classifying cloud types that one may use to understand the different clouds shapes and sizes. This dataset consists of 1 year of 1km resolution MODIS hyper-spectral imagery which was merged with CloudSat cloud labels. With greater availability of space launch capabilities, satellites being launched in the present carry a host of sensors and camera technology onboard with them to study the Earth. Satellites like the Sentinel-1 launched by the European Space Agency carry with them a single C-band Synthetic Aperture Radar (SAR) which can scan the earth surface in daylight or night settings. The SAR can pierce through clouds and rainfall conditions. Where as a later launched satellite Sentinel-2 has the capability to capture images of various spectra such as infrared, near infrared and visual spectrum. The wavelengths that are measured by Sentinel-2 range from 443.9 nm to detect aerosol particles to 1375.5 nm to detect Cirrus clouds. On top of these other sensors such as Cloud Displacement Index values, 360 degree Azimuth values help researchers with various use cases such as cloud segmentation, land cover land usage classification among others. This thesis is focused on demanding task of developing a robust cloud and cloud shadow segmentation that is capable of functioning in cases of satellite sensor failure. Statistical cloud detection algorithms like FMask 4.0[24] and MAJA[11] are increasing getting overshadowed by

deep neural network models such as Resnet[12] and U-net [25]. These models were originally made for image classification tasks but they can be fine-tuned or adapted for cloud and cloud shadow detection. Most major cloud detection algorithms today rely on optical satellite information that is often not enough to detect thin clouds or differentiate between cloud and snow. Li et al [17] developed a sophisticated cloud and cloud shadow detection algorithm based on a combination of UNet [25] and Residual blocks [12]. They employ a feature-level fusion [20] strategy but only use Sentinel-2 optical information as input. Therefore, their model fails in polar climates and cannot accurately distinguish between snow and clouds. Alistair Francis [10] achieves SOTA performance on their cloud and cloud shadow detection algorithm by leveraging multimodal input data i.e Sentinel-1, Sentinel-2 and Digital Elevation Maps (DEM).

The primary contributions of this thesis are as follows:

1. Propose a U-net[25] based robust multi-view fusion model for cloud and cloud shadow segmentation.
2. Used a state of the art global cloud and cloud shadow dataset CloudSen12[3] to train and test the model. Cloudsen12[3] is a multi-modal dataset comprising of Sentinel-1 and Sentinel-2 Satellite image data and other derived information such as Digital Elevation Maps and Water Vapour Occurrence [23].
3. Investigated the proposed model's performance under a variety of conditions such as Climate Zones , Cloud Coverages and Land Cover.
4. Tested the model's robustness under a simulated band dropout scenario where Sentinel-2 information is successively removed and the model's performance is tested on the remaining "active" bands.

The following sections of this thesis document are organised as follows: Section 2 presents the related work summary in the field cloud detection and multi-view image segmentation models, Section 3 is a detailed explanation of the dataset being used in the thesis and the proposed model architecture, Section 4 consists of the training procedure and obtained results, and a final conclusion is given in Section 5.

Chapter 2
Related Work

In recent years, there is been a huge increase in the usage of ML models for RS tasks. In most cases, these ML models are domain specific and vary according to the task at hand. The following section is a brief overview about the recent developments in the domain of Multi-view deep learning and cloud/cloud shadow segmentation. Before the advent of AlexNet[16] in 2012 and its successors, image segmentation tasks have been fulfilled with the use of statistical algorithms like K-means clustering [21]. Satellite image segmentation is different from the conventional image segmentation tasks due to several reasons. Compared to the typical RGB images of dogs or cats, satellite images have higher pixel resolutions. Satellites are equipped to capture a series of image bands and are not limited to RGB information. Further, satellites are typically in Low Earth Orbit(LEO) and their orbit causes them to capture images at different spatial and temporal alignments. Due to these caveats, many state of the art models do not perform adequately with satellite images and need to be adapted for the task.

2.1 Cloud and Cloud Shadow Segmentation

Significant research has been conducted in the topic of cloud and cloud shadow segmentation. Before the advent of deep learning algorithms, statistical algorithms like FMask[24] and Sen2Cor[18] were at the forefront of detecting clouds in satellite images. With the success of convolutional neural networks like Resnet[12] , CNNs form the backbone of the state of the art cloud detection algorithms today. Most of these algorithms use optical satellite information to train their ML models. Rozenhaimer et al[26] propose an algorithm for cloud and cloud shadow detection in multi-channel satellite imagery, specifically using World-View-2 and Sentinel-2 optical satellite information. The algorithm utilizes both spectral and spatial information in the imagery to learn deep invariant features for cloud detection. The authors also propose a domain adaptation CNN-based approach that allows for better adaptation between the WV-2 and Sentinel-2 satellite platforms during the prediction step, without the need to train separately for each platform. This approach increased the prediction accuracy of both clear and cloudy pixels compared to a network trained only by WV-2. Chai et al[4] propose another CNN

model which utilizes multi-level spatial and spectral features from the entire image and all the bands to label each pixel as cloud, thin cloud, cloud shadow, or clear sky. Zhang et al[31] propose a cloud and cloud shadow segmentation network that is able to detect smaller clouds and obtain finer edges, addressing the limitations of traditional methods. They utilized ResNet-18 [12] as the backbone for feature extraction and incorporated the Multi-scale Global Attention Module to enhance channel and spatial information, improving detection accuracy. They further introduced the Strip Pyramid Channel Attention Module to learn spatial information at multiple scales, enabling better detection of small clouds. Further, Encoder-decoder type networks such as U-Net[25] have also given a huge edge in improving the accuracy of these single-view models. Ouyang and Li[22] propose a DSSN-GCN framework that combines the deep semantic segmentation network (DSSN) and the graph convolutional neural network (GCN) for remote sensing image semantic segmentation. The framework leveraged the appearance extraction ability of DSSN and the topological relationship modelling capability of GCN. They introduce a new deep learning architecture called the Attention Residual U-shaped network (AttResUNet), which uses residual blocks to encode feature maps and an attention module to refine the features. AttResUNet is trained to extract high-level features and initialize the graph nodes in the GCN.

2.2 Data Fusion and Multi-View Cloud Segmentation

Data fusion refers to the process of integrating multiple sources of data in order to get more consistent and accurate results of one's problem statement rather than using just one data source. In the case of cloud and cloud shadow segmentation, data fusion refers to the process of using instruments like Synthetic Aperture Radar(SAR) and LiDAR along with optical satellite information to achieve better results. Li et al[17] proposed a deep learning-based cloud detection method called multi-scale convolutional feature fusion (MSCFF) for remote sensing images of different sensors, achieving higher accuracy than traditional rule-based methods especially in bright surface covered areas. The MSCFF model includes a symmetric encoder-decoder module and a multi-scale feature fusion module for feature extraction and fusion. Their model was trained on LandSat-8 and Gaofen-1 [6] optical satellite datasets. One important caveat of this model is that it wrongly classifies bright snow as clouds. Hu et al [13] propose MCANet which is a multi-modal network in the sense that it has two parallel branches where the optical satellite information is processed. One branch is a normal Unet [25] where as the other branch is a Vision Transformer [8]. This network was explicitly made to distinguish between cloud and snow but it was unnecessarily complicated and only used optical satellite information as input. Xu et al[27] propose a novel global-local fusion algorithm GLF-CR that utilises the power of the SAR and fuses it with optical images for the task of cloud removal of remote sensing images. The authors took the inspiration of transformers which capture global features and applied it to a novel SAR-guided Global Context Interaction(SGCI) block. This SGCI block makes sure that the structures of the cloud removed regions are consistent with the cloud-free regions of the original optical images. Secondly, to reduce the speckle noise that plagues the SAR, authors propose another block named as SAR-guided local feature compensation(SLFC). This block generates more reliable texture data and avoids the speckle noise of SAR.

Chapter 3
Methodology

Cloud and cloud shadow segmentation is a well researched topic. It is the first data pre-processing task required before one can utilise optical satellite images for any downstream tasks. Therefore, the task of cloud segmentation has become one of the most important pre-processing tasks in geospatial data science. Deep learning algorithms have increased the accuracy at which clouds and cloud shadows are segmented. But these deep learning algorithms fail in robustness like accurately segmenting clouds in snowy regions or clouds over oceans. Most of the cloud segmentation algorithms usually use optical satellite information such as Sentinel-2 or LandSat-8. However, using just optical satellite information often leads to inaccurate segmentation results. Therefore, some deep learning algorithms fuse SAR information along with optical information to improve segmentation results. This section provides an overview of our methodology which fuses these multiple sensors for cloud and cloud shadow segmentation.

3.1 Dataset

For this thesis, we use Cloudsen12[3] dataset, which is a global dataset for cloud and cloud shadow segmentation. It consists of Sentinel-1 Synthetic Aperture Radar (SAR) information, Sentinel-2 optical information distributed across 13 bands, along with DEM (Digital Elevation Maps) and other derivative information such as Land Cover [29] , Azimuth and Water Occurrence [23]. Sentinel-2 optical satellite is a multi-spectral satellite that consists of 13 bands which range from visible spectrum aka Red, Green, Blue to Near Infrared (NIR) and Short Wave Infrared (SWIR). Sentinel-2 information can either be L1C product or L2A product. Sentinel-2 L1C product retains atmospheric influences such as Water Vapour and Aerosols. For this thesis, we have used Sentinel-2 L1C product as it consists of Top-of-Atmosphere reflectance (TOA) and retains its atmospheric influence information. L2A product gets rid of this valuable information and is therefore not used for cloud and cloud shadow detection. The Sentinel-1 Synthetic Aperture Radar information is divided into 3 bands namely Vertical Transmit/Vertical Receive (VV) Polarization, Vertical Transmit/Horizontal Receive (VH) Polarization and the Angle of Incidence. Elevation map data is saved in meters and obtained from MERIT Hydro

File/ Folder	Name	Scale	Wavelength	Description
S2L1C & S2L2A	B1	0.0001	443.9 nm (S2A)/442.3 nm (S2B)	Aerosols.
	B2	0.0001	496.6 nm (S2A)/492.1 nm (S2B)	Blue.
	B3	0.0001	560 nm (S2A)/559 nm (S2B)	Green.
	B4	0.0001	664.5 nm (S2A)/665 nm (S2B)	Red.
	B5	0.0001	703.9 nm (S2A)/703.8 nm (S2B)	Red Edge 1.
	B6	0.0001	740.2 nm (S2A)/739.1 nm (S2B)	Red Edge 2.
	B7	0.0001	782.5 nm (S2A)/779.7 nm (S2B)	Red Edge 3.
	B8	0.0001	835.1 nm (S2A)/833 nm (S2B)	NIR.
	B8A	0.0001	864.8 nm (S2A)/864 nm (S2B)	Red Edge 4.
	B9	0.0001	945 nm (S2A)/943.2 nm (S2B)	Water vapor.
	B11	0.0001	1613.7 nm (S2A)/1610.4 nm (S2B)	SWIR 1.
	B12	0.0001	2202.4 nm (S2A)/2185.7 nm (S2B)	SWIR 2.
S2L1C	B10	0.0001	1373.5 nm (S2A)/1376.9 nm (S2B)	Cirrus.
S2L2A	AOT	0.001	—	Aerosol Optical Thickness.
	WVP	0.001	—	Water Vapor Pressure.
	TCI_R	1	—	True Color Image, Red.
	TCI_G	1	—	True Color Image, Green.
	TCI_B	1	—	True Color Image, Blue.
S1	VV	1	5.405 GHz	Dual-band cross-polarization, vertical transmit/ horizontal receive.
	VH	1	5.405 GHz	Single co-polarization, vertical transmit/vertical receive.
	angle	1	—	Incidence angle generated by interpolating the 'incidenceAngle' property.
extra/	CDI	0.0001	—	Cloud Displacement Index .
	Shwdirection	0.01	—	Azimuth. Values range from 0°–360°
	elevation	1	—	Elevation in meters. Obtained from MERIT Hydro datasets
	ocurrence	1	—	JRC Global Surface Water . The frequency with which water was present.
	LC10	1	—	ESA WorldCover 10 m v100 product.

Figure 3.1: Table describing the modalities available in Cloudsen12 Dataset

datasets [28] . Land Cover information is obtained from ESA WorldCover [29] 10m resolution product. Azimuth values range between 0° to 360° and Water Occurrence refers to water presence frequency obtained from JRC Global Surface Water [23]. A detailed overview about the different modalities can be found in Figure 3.1 This global dataset consists of images from 1827 unique Regions of Interest (ROI)s in the training set and 173 unique ROIs in the test set spread all across the world. Each ROI was selected specifically to have a diverse set of locations and natural features. Further, for every ROI, there are 5 types of cloud coverage available in the dataset i.e. cloud-free (0%), almost-clear (0–25%), low-cloudy (25–45%), mid-cloudy (45–65%), and cloudy (>65%). Here the percentage refers to the portion of a sample image obstructed by clouds as shown in Figure 3.2 .

This ensures scene variability in the temporal domain. The number of images per cloud coverage type are equally divided in the training, validation and test sets. In the spatial domain, each ROI has the spatial resolution of 5090x5090 meters. This translates

to having 509x509 pixels per image. For our task, these images were resampled to a size of 512x512 pixels. The dataset consists of four primary output classes namely No Cloud, Cloud Shadow, Thin Clouds and Thick Clouds. No Cloud images form the majority of the dataset, followed by Thick Clouds, Cloud shadows and Thin Clouds in decreasing order. The authors further provide these labels in three forms namely high-quality annotated (with human assistance), scribble annotated and no annotation (for unsupervised learning). For this thesis, I have chosen only the high-quality annotated labels which are suitable for Supervised Deep Learning tasks.

3.2 Model Architectures

In the following subsection I give an overview on the three models that I have constructed as part of my thesis. The first set of models are multi-view cloud/cloud shadow segmentation models. They have 1 encoder for each of the views available, but are restricted to one decoder as the final output is a segmentation mask with 4 classes. Here sentinel-2 optical information is fused along with Sentinel-1 SAR information, DEM, Water Occurrence, Azimuth and Land cover information. Two such models are constructed which are based on UNet [25] architecture. They have 6 encoders and 1 decoder each. Each Encoder and Decoder consists of two sets of convolution operations. After every convolution operation, we used ReLU (Rectified Linear Unit) activation function and Batch Normalisation [14].The feature maps were downsampled using Maxpool2D and upsampled using Pytorch's upsampling operation. We also used Dropout after every convolution layer which started at 30% at the top layer, 20% in the mid layers and 10% at the bottleneck. The major difference between the two models lie in the fusion strategy used at the bottleneck level. The first fusion strategy is simple concatenation where all the modalities are stacked into a single tensor naively. This increases the size of the model significantly as each view consists of hundreds of channels at the bottleneck layer. Subsequent decoder layers also employ concatenation fusion at the skip connections. The second fusion strategy is Gated Fusion [2] where each of the modalities are dynamically combined by the model itself. The Gating function is a learnable hyperparameter that learns to give more importance to modalities that are contributing to the final output and and less importance to the ones that are not. Figure is an illustration of a gating function used in this thesis.

Figure is a detailed illustration of the proposed model with all modalities. The second class of model has Sentinel-2 optical information as input and its output is not just a segmentation mask but also DEM, Land cover Product and Sentinel-1 radar data. The objective of such a model is to demonstrate as a proof of concept that simple optical information networks have sufficient capabilities to also perform other pretext tasks. In this case, the architecture of the models are reversed i.e. there is only 1 encoder for the Sentinel-2 optical information and has a separate decoder for each of the previously mentioned tasks. The Main Loss is a weighted sum of the individual four losses from the four decoders each. In our experiment, we gave the cloud segmentation loss as the highest weight of 1.0 and the other 3 pretext losses were given 0.3 each. Figure 3. gives an illustration about the architecture of the pretext task model. In case of the individual losses, cloud segmentation and land cover labels are categorical variables. Therefore, we used Cross-Entropy Loss [19] for them. Whereas, in the cases of DEM and S1, we used Mean Squared Error Loss [1] as they are continuous variables.

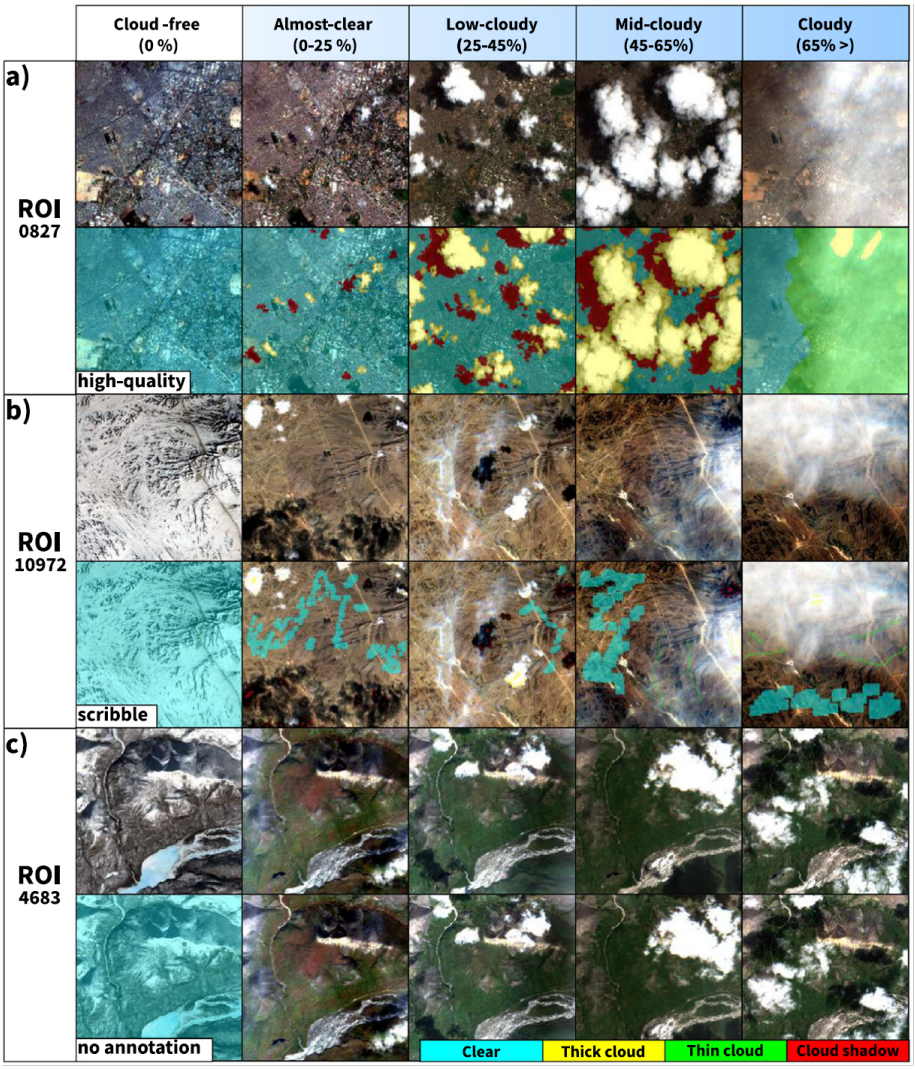

Figure 3.2: Example plot showing the 5 different types of Cloud Coverages along with 3 types of Labelled Images

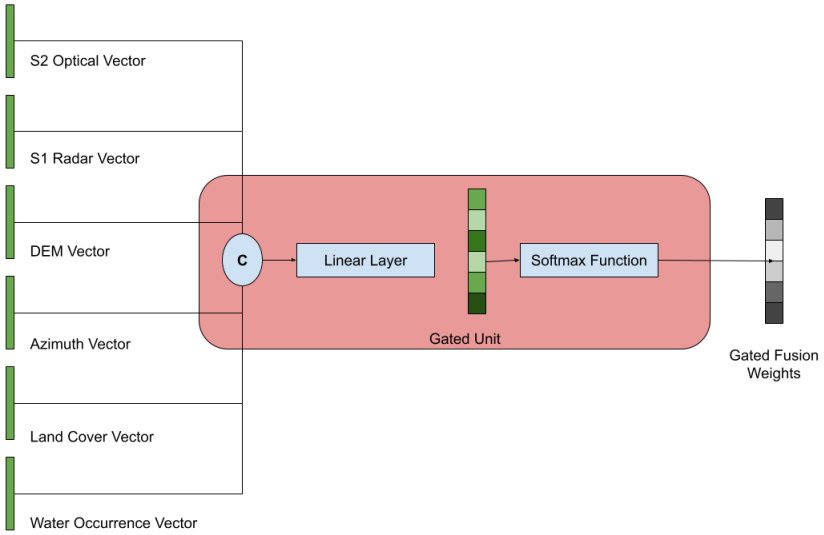

Figure 3.3: **Gated Fusion Module** with six input views.

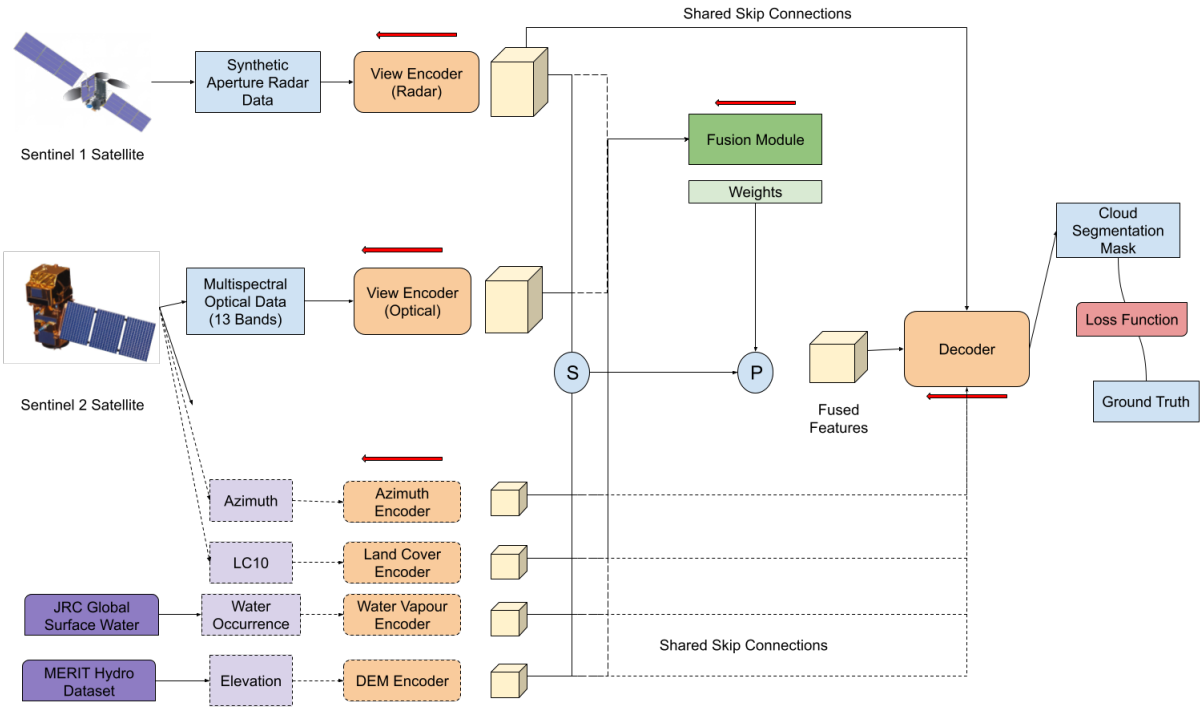

Figure 3.4: **Pretext Task Model Diagram** with weighted combined loss.

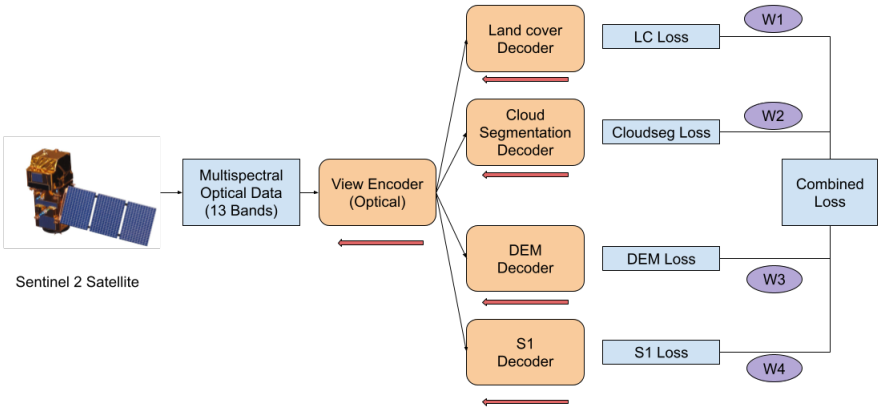

Figure 3.5: **Proposed Multi-view Model Diagram** with every modality highlighted in detail.

Chapter 4
Experimentation

4.1 Training Procedure

4.1.1 Model Framework

All the models were written in Python and uses Pytorch deep learning library. To facilitate model training, I further used Pytorch Lightning [9] modules which simplify writing down the deep learning pipeline. Pytorch Lightning has inbuilt modules and classes to setup the training, validation and testing regimen. Along with that, it also has metrics that can imported and linked with the pipeline that are calculated every epoch. All these experiments were then easily tracked by using MLFLOW[5] MLOps framework which sets up a local server to keep track of each experiment and its constituent hyperparameters.

4.1.2 Data Preprocessing

To prepare the Cloudsen12 dataset suitable for training, the Sentinel-2 and Sentinel-1 sensor information were normalised to [-1,1]. This is especially necessary for multi-view fusion because it creates a common scale for the deep learning model to integrate Sentinel-1 and Sentinel-2 information more effectively. Normalization of data also helps in faster convergence of deep learning models using gradient-based optimizations. For the specific case of Pretext Prediction task, the landcover labels were encoded into a one-hot encoded vector in order to make it easier for the model to distinguish the categorical labels and predict the correct landcover information. While training, random rotations of 90, 180 and 270 degrees were also used as a data augmentation technique. Lastly the whole dataset was split into a training, validation and test sets in the ratio of 85%, 5% and 10%. This results in 8490 images per modality in the training set, 535 in the validation set and 975 in training set respectively.

4.1.3 Training Specifications

Every model was trained for 50 epochs with Early stopping that restricted the models from training more than it is required. ADAM optimizer was used as an optimizer with a weight decay of 1e-5 to have a smoother learning procedure. The learning rate for all the models were set at 0.0001 with a batch size of 64. A higher learning rate made the validation loss jump up and down signifying that the model overfits to the training set. Since our main task was generating a segmentation mask, the loss function used was Cross-Entropy Loss [19]. Where as, in the Pretext task model I used either Cross-Entropy Loss [19] or Mean Squared Error Loss [1] depending on the output. This is because a Cloud/Cloud shadow Segmentation mask and Landcover Product are discrete values that require a categorical-based Loss but on the other hand DEM and Water Occurrence[23] are continuous outputs which require a euclidean-based Loss.

4.1.4 Evaluation Metrics

Model performances were evaluated using common metrics from literature like Accuracy, Precision, Recall and F1-Score. MeanIOU was also added as a segmentation task specific metric to check for segmentation mask overlap. All the above mentioned metrics rely on the computation of a confusion matrix which consists of the number of samples that are either True Positive(TP), False Positive(FP), True Negative(TN) and False Negative(FN). Accuracy metric in an image segmentation task is defined as the number of pixels that are correctly classified with respect to the total number of pixels.

$$\text{Accuracy} = \frac{\text{TP} + \text{TN}}{\text{TP} + \text{TN} + \text{FN} + \text{FP}} \tag{4.1}$$

Precision refers to the proportion of pixels that are correctly labelled to the the total number of predicted pixels in an image.

$$\text{Precision} = \frac{\text{TP}}{\text{TP} + \text{FP}} \tag{4.2}$$

Recall or True Positive Rate is similar to Precision in the sense that it is defined as the proportion of correctly labelled pixels to the sum of correctly predicted pixels and false negative pixels.

$$\text{Recall} = \frac{\text{TP}}{\text{TP} + \text{FN}} \tag{4.3}$$

Recall measures how many pixels were correctly labelled in an image. F1 Score or Dice coefficient is the harmonic mean of precision and recall. In segmentation tasks, it is defined as the proportion of twice the area of overlap to the total area i.e. the sum of predicted samples and ground truth samples.

$$\text{F1 Score} = 2 \cdot \frac{\text{Precision} \cdot \text{Recall}}{\text{Precision} + \text{Recall}} \tag{4.4}$$

Finally, the Intersection over Union(IOU) score measures the ratio of the Intersection of the predicted bounding box and the ground truth bounding box with respect to the Union of the two mentioned bounding boxes.

$$\text{IoU} = \frac{\text{Intersection}}{\text{Union}} = \frac{\text{TP}}{\text{TP} + \text{FP} + \text{FN}} \tag{4.5}$$

4.2 Results

4.2.1 Competing Models

The proposed model's performance was compared with baseline models. The first set of baseline models are traditional algorithm based models for cloud and cloud shadow segmentation i.e. Sen2Cor [18] and KappaMask [7]. Both of these algorithms utilize Sentinel-2 optical satellite infomation as an input for their models. Further, they were selected because their output classes correspond to the four afformentioned classes in CloudSen12[3]. Cloudsen12 already provide KappaMask[7] and Sen2Cor[18] labels which were compared with the high quality manually annotated labels for comparison. Moving on to deep learning models, two single-view UNet models were constructed for Sentinel-1 and Sentinel-2 information respectively. These models have one encoder and one decoder each. Subsequently, two Dual-view fusion models comprising of Sentinel-1 and Sentinel-2 information as modalities were constructed. Both these models have two encoders each for Sentinel-1 and Sentinel-2 information which are fused at the bottleneck layer. They further consist of one decoder each for the cloud/cloud shadow segmentation output. The only defining difference between the above two models are the fusion technique used. One model consists of concatenation fusion and the other consists of gated fusion. The four deep learning models were trained using the same training specifications given in the previous subsection. Table 4.1 illustrates the complete list of all models and their performance on the aforementioned evaluation metrics.

Model	Fusion	Accuracy	Precision	Recall	F1Score
Sen2Cor	-	0.49	0.59	0.49	0.45
KappaMask	-	0.52	0.58	0.52	0.50
Sentinel 1	-	0.25	0.13	0.25	0.17
Sentinel 2	-	0.65	0.70	0.65	0.63
2Modality	Concat	0.67	0.72	0.67	0.69
2Modality	Gated	0.66	0.72	0.66	0.66
All Modality	Concat	**0.69**	**0.77**	**0.69**	**0.72**
All Modality	Gated	0.66	0.74	0.66	0.68

Table 4.1: Proposed Model Performance as compared with competing models

4.2.2 Quantitative Results

Table 4.1 shows how the proposed model with concatenation fusion achieves the best performance across all evaluation metrics. Traditional algorithmic models i.e. KappaMask and Sen2Cor achieve around 50% performance across all metrics. Subsequently, the table shows that using just Sentinel-1 Synthetic Aperture Radar (SAR) information for a deep learning model has inferior performance than even traditional non-deep learning algorithms. Performance starts increasing drastically as we begin using Sentinel-2 optical information. Combining Sentinel-1 and Sentinel-2 information provides marginal improvement in overall performance. Where as, combining DEM [28], Landcover [29], Water Occurence [23] and Azimuth information along with Sentinel-1 and Sentinel-2 information improves cloud and cloud shadow detection capabilities by a significant margin. Concatenation Fusion also comes up as a clear winner in quantitative performance as compared to Gated Fusion.

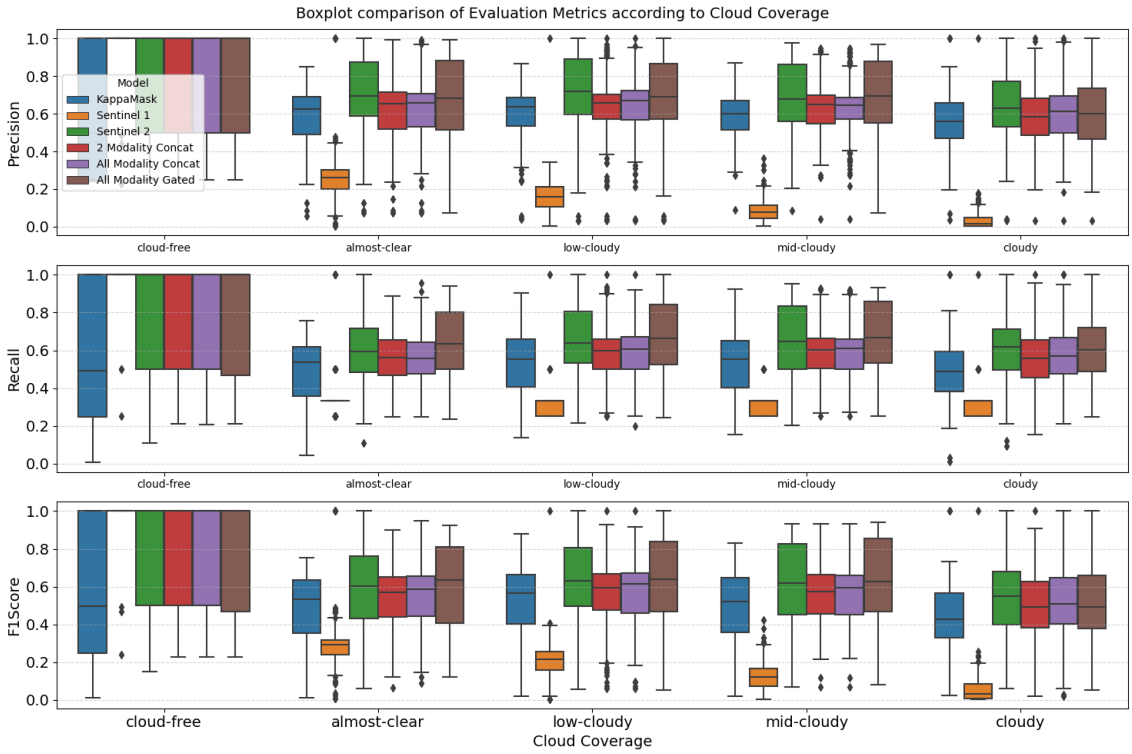

Figure 4.1: **Box plot comparison of six models' performance according to cloud coverage.** The evalution metrics are illustrated on the Y axis and the cloud coverage in ascending order on X axis.

4.2.3 Qualitative Results

Cloudsen12 provides us with metadata information for every data sample such as the annotator's names, landcover labels, dificulty scores, Sentinel-1 and Sentinel-2 acquisition date, cloud coverage, coordinates of the image in latitude and longitude among others. Difficulty scores denote the manual annotator's confidence in labelling images on a scale from 1 to 5; 1 being high confidence samples which are easy to annotate and 5 implies hard to annotate samples. Cloud coverage as mentioned in previous chapters indicates the percentage of the image covered in clouds; from cloud-free (0% clouds) to cloudy (>65% clouds). Landcover label shows the predominant land use in the image using the ESA WorldCover[29] 10m scale with categorical labels ranging from 10 to 100. To make it legible for human understanding, the categorical labels were converted to their corresponding land use labels such as Tree Cover, Shrubland, Snow and Ice. The centroid coordinates were used to find the corresponding Koeppen-Geiger Climatic Zones [15]. The KG climate zones are a system of classifying the earth surface according to relative heat and humidity it receives throughout the year. The KG climate zones are divided into 5 broad categories namely; A - Tropical , B - Arid, C - Temperate, D - Continental, E - Polar and finally water bodies are classified as 0 - Ocean.

Figure 4.1 shows how the AMG model beats other competing models in the difficult to classify cloud coverage conditions such as almost-clear and low-cloudy. These conditions have thin clouds present which make it difficult for optical satellites to detect. Figure

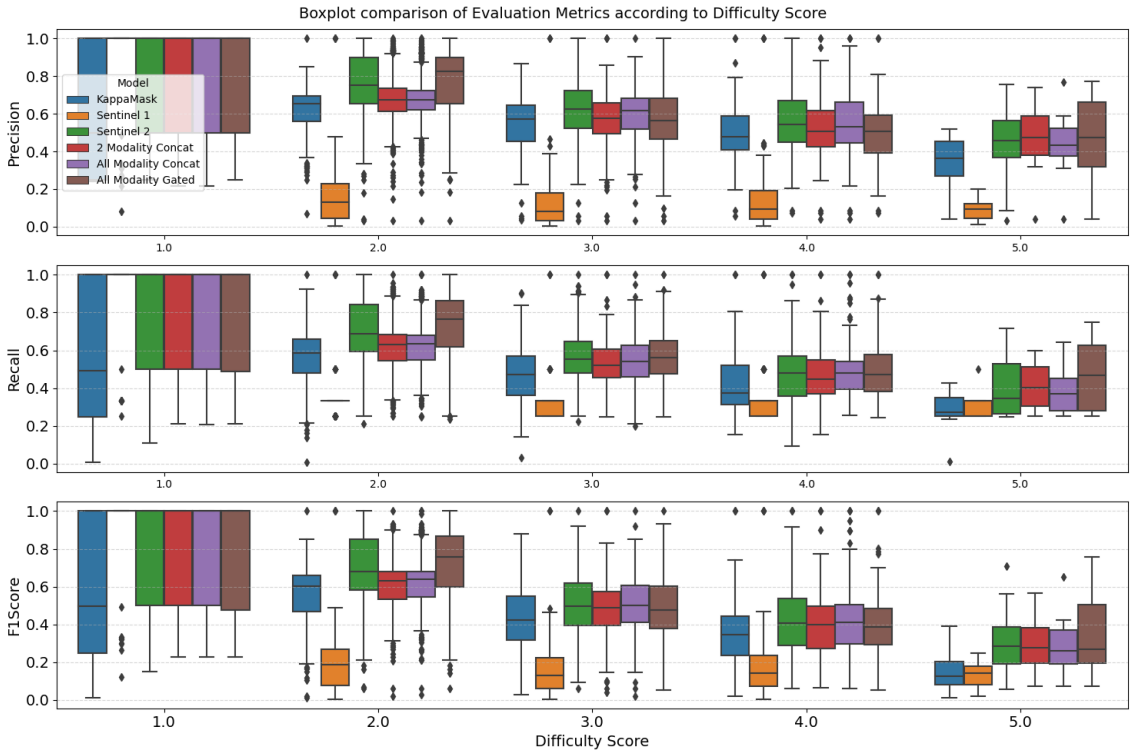

Figure 4.2: **Box plot comparison of six models' performance according to Difficulty scores.** The evalution metrics are illustrated on the Y axis and the difficulty score in ascending order on X axis.

4.2 illustrates how the different models perform on increasing order of difficulty. The proposed model with gated fusion is able to achieve significantly better results than the competing models in medium difficulty images. Various Outliers are also observed in the competing models of 2 Modality Gated and All Modality Concat. Where as in medium-high difficulty samples of 3.0 to 4.0, the proposed model achieves equal performance with competing models. Lastly, in 5.0 difficulty, the proposed model again achieves better performance.

Figure 4.3 describes the major models' F1Score distribution over different KG climate zones [15] . The proposed model (AMG) performs better in polar climates. This refers to the fact that the model is able to distinguish between clouds and snow significantly better than other competing models. Where as, in all other climate zones, the proposed model performs better than 2 Modality Concat and All Modality Concat (AMC) models but fails to beat the single-view Sentinel-2 model. Figure 4.4 illustrates the F1Score for for ten landcover types. The proposed model is able to outpeform all the competing models in snow and ice conditions. In wetland conditions, the proposed model shows much higher confidence in cloud/cloud shadow predictions than competing models. Where as, in the case of built-up areas, the proposed model shows the least amount of confidence i.e. a varied set of predicted labels. In all other landcover types, the proposed model performs better than other competing models except Sentinel-2 single-view model.

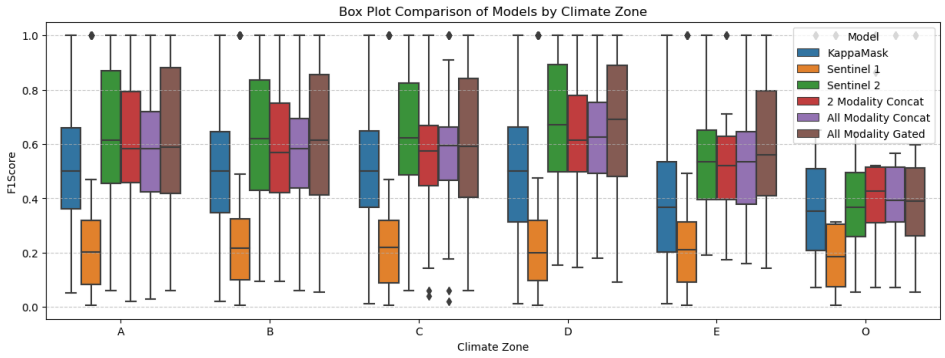

Figure 4.3: **Box plot comparison of six models' F1Score according to Koeppen-Geiger Climatic Zones.** The F1Score is illustrated on the Y axis and the climate zone on X axis.

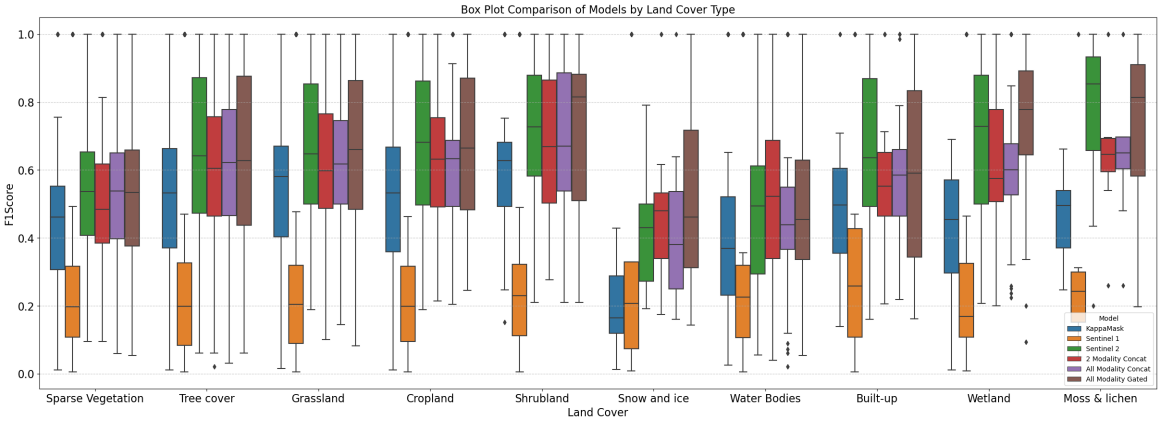

Figure 4.4: **Box plot comparison of six models' F1Score according to Land Cover.** The F1Score is illustrated on the Y axis and the landcover label on X axis.

4.2.4 Band Removal Experiment

In order to test the robustness of the proposed model AMG, we decided to simulate an information loss experiment. In such a scenario, the S-2's bands stop functioning one after the other. We replaced S-2 Band values in the Test set to 0 starting from Band 1 till Band 12. To keep the simulation simple, we started from Band 1 and consecutively increased the number of Bands with every turn. Those results are illustrated in the table 4.2 below. There is an immediate drop in performance when the fist band is removed and it remains constant in performance till 2 bands are removed. Here the bands being removed are Band 1 (Aerosols) and Band 2(Blue). The performance drops staggeringly as soon as Band 3(Green) is removed, and then remains almost stagnant till Band 5 is removed. Model Performance seems to jump back which signify that the proposed model AMG needs the combination of Blue, Green and Red together to give tangible results. If two of them are removed, then the model loses accuracy in its predictions. Performance drops again when Near Infrared band Band 8 is removed. The last sudden increase in performance happens after Band 10 removal. In such a case only Short Wave Infrared and Cirrus bands of S-2 are capable enough to give as close of a performance as when only 1 band is removed. This experiment suggests that the proposed model is only

robust in certain combinations of bands. Further experiments are needed to properly ascertain which bands of S-2 are redundant for accurate cloud and shadow segmentation and which bands are absolutely necessary for the same task.

Bands Dropped	Accuracy	F1Score	MeanIOU	Precision	Recall
0	0.66	0.68	0.41	0.74	0.66
1	0.45	0.44	0.25	0.46	0.45
2	0.45	0.44	0.25	0.46	0.45
3	0.29	0.20	0.28	0.33	0.29
4	0.25	0.11	0.28	0.32	0.25
5	0.25	0.11	0.28	0.32	0.25
6	0.45	0.40	0.22	0.43	0.45
7	0.49	0.41	0.31	0.53	0.49
8	0.34	0.27	0.28	0.32	0.34
9	0.26	0.13	0.29	0.37	0.26
10	0.43	0.32	0.31	0.36	0.43
11	0.39	0.37	0.28	0.45	0.39
12	0.25	0.12	0.28	0.34	0.25

Table 4.2: Proposed Model Performance in consecutive Band Removal Experiment

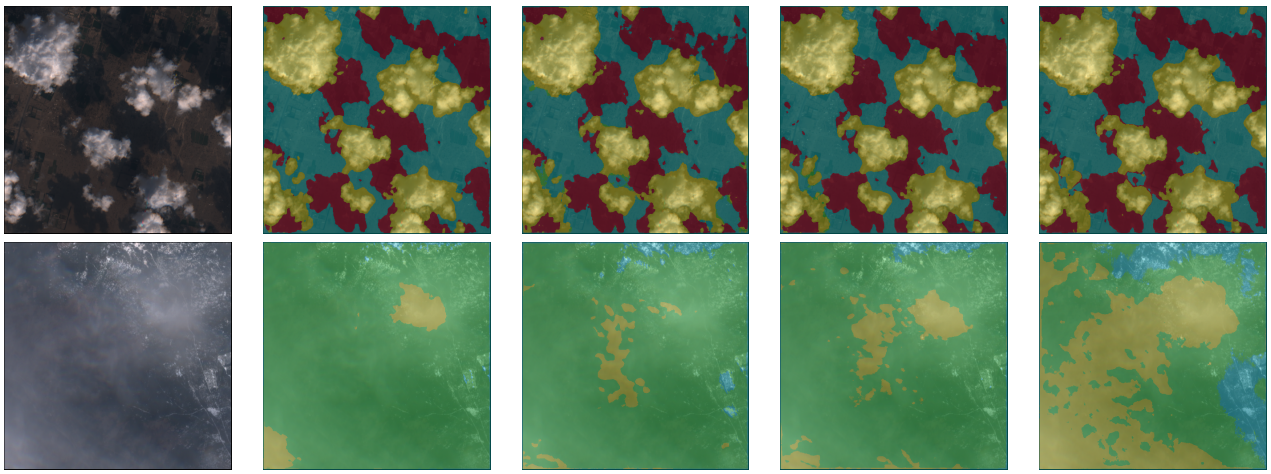

Table 4.3: Example Predictions of different models starting from left to right: RGB Ground truth, Sentinel-2 only, 2 Modality Concat, All Modality Concat, All Modality Gated. Blue signifies Background, Green is thin clouds, Yellow is thick clouds and Red signifies Cloud Shadows

Chapter 5
Conclusion

In this thesis, I present a robust cloud and cloud shadow segmentation network which is based on multi-view machine learning and feature level data fusion. This model is quite light in comparison to other state of the art models yet its able to achieve sufficient performance in metrics. My experiments show that using a multi-view machine learning makes a model robust to various climatic conditions and varying levels of clouds in images. The proposed model performs well in hard scenarios such as polar climates and is able to adequately distinguish between clouds and snow. On the other hand it is not very robust in information loss scenarios. Removing one band of optical satellite information makes my model perform worse by 20 %. But it is still able to hold on to this reduced performance even if more bands are removed. This confirms that multi-view learning and feature level data fusion has various benefits and should be investigated as an alternative to heavy networks. Future work could replace the convolutional encoders with Transformer based encoders and theoretically improve model performance. Additional temporal information can also be added as another view to make the model perform better. It is also imperative that more work needs to be done in order to make cloud and cloud shadow segmentation models robust in case of information loss. All in all, this thesis contributes towards making better data pre-processing algorithms for raw satellite images that may enrich the performance of downstream tasks.

Abbreviations

ML	Machine Learning
RS	Remote Sensing
S1	Sentinel 1
S2	Sentinel 2
DEM	Digital Elevation Maps
AMG	All Modality Gated
AMC	All Modality Concat

Bibliography

[1] 2008. *Mean Squared Error*. Springer New York, New York, NY, 337–339. DOI: http://dx.doi.org/10.1007/978-0-387-32833-1_251

[2] John Arevalo, Thamar Solorio, Manuel Montes-y Gómez, and Fabio A González. 2017. Gated multimodal units for information fusion. *arXiv preprint arXiv:1702.01992* (2017).

[3] Cesar Aybar, Luis Ysuhuaylas, Jhomira Loja, Karen Gonzales, Fernando Herrera, Lesly Bautista, Roy Yali, Angie Flores, Lissette Diaz, Nicole Cuenca, and others. 2022. CloudSEN12, a global dataset for semantic understanding of cloud and cloud shadow in Sentinel-2. *Scientific data* 9, 1 (2022), 782.

[4] Dengfeng Chai, Shawn Newsam, Hankui K Zhang, Yifan Qiu, and Jingfeng Huang. 2019. Cloud and cloud shadow detection in Landsat imagery based on deep convolutional neural networks. *Remote sensing of environment* 225 (2019), 307–316.

[5] Andrew Chen, Andy Chow, Aaron Davidson, Arjun DCunha, Ali Ghodsi, Sue Hong, Andy Konwinski, Clemens Mewald, Siddharth Murching, Tomas Nykodym, Paul Ogilvie, Mani Parkhe, Avesh Singh, Fen Xie, Matei Zaharia, Richard Zang, Juntai Zheng, and Corey Zumar. 2020. Developments in MLflow: A System to Accelerate the Machine Learning Lifecycle. 1–4. DOI:http://dx.doi.org/10.1145/3399579.3399867

[6] Liangfu Chen, Husi Letu, Meng Fan, Huazhe Shang, Jinhua Tao, Laixiong Wu, Ying Zhang, Chao Yu, Jianbin Gu, Ning Zhang, and others. 2022. An introduction to the Chinese high-resolution Earth observation system: Gaofen-1˜ 7 civilian satellites. *Journal of Remote Sensing* (2022).

[7] Marharyta Domnich, Indrek Sünter, Heido Trofimov, Olga Wold, Fariha Harun, Anton Kostiukhin, Mihkel Järveoja, Mihkel Veske, Tanel Tamm, Kaupo Voormansik, and others. 2021. KappaMask: AI-based cloudmask processor for Sentinel-2. *Remote Sensing* 13, 20 (2021), 4100.

[8] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, Jakob Uszkoreit, and Neil Houlsby. 2021. An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale. (2021). https://arxiv.org/abs/2010.11929

[9] William Falcon and The PyTorch Lightning team. 2019. PyTorch Lightning. (March 2019). DOI:http://dx.doi.org/10.5281/zenodo.3828935

[10] Alistair Francis. 2024. Sensor Independent Cloud and Shadow Masking With Partial Labels and Multimodal Inputs. *IEEE Transactions on Geoscience and Remote Sensing* 62 (2024), 1–18. DOI:http://dx.doi.org/10.1109/TGRS.2024.3391625

[11] Olivier Hagolle, Mireille Huc, Camille Desjardins, Stefan Auer, and Rudolf Richter. 2017. MAJA Algorithm Theoretical Basis Document. (Dec. 2017). DOI:http://dx.doi.org/10.5281/zenodo.1209633

[12] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. 2015. Deep Residual Learning for Image Recognition. (2015). https://arxiv.org/abs/1512.03385

[13] Kai Hu, Enwei Zhang, Min Xia, Liguo Weng, and Haifeng Lin. 2023. Mcanet: A multi-branch network for cloud/snow segmentation in high-resolution remote sensing images. *Remote Sensing* 15, 4 (2023), 1055.

[14] Sergey Ioffe and Christian Szegedy. 2015. Batch Normalization: Accelerating Deep Network Training by Reducing Internal Covariate Shift. (2015). https://arxiv.org/abs/1502.03167

[15] Markus Kottek, Jürgen Grieser, Christoph Beck, Bruno Rudolf, and Franz Rubel. 2006. World Map of the Köppen-Geiger Climate Classification Updated. *Meteorologische Zeitschrift* 15 (05 2006), 259–263. DOI:http://dx.doi.org/10.1127/0941-2948/2006/0130

[16] Alex Krizhevsky, Ilya Sutskever, and Geoffrey E Hinton. 2012. Imagenet classification with deep convolutional neural networks. *Advances in neural information processing systems* 25 (2012).

[17] Zhiwei Li, Huanfeng Shen, Qing Cheng, Yuhao Liu, Shucheng You, and Zongyi He. 2019. Deep learning based cloud detection for medium and high resolution remote sensing images of different sensors. *ISPRS Journal of Photogrammetry and Remote Sensing* 150 (2019), 197–212. DOI:http://dx.doi.org/https://doi.org/10.1016/j.isprsjprs.2019.02.017

[18] Magdalena Main-Knorn, Bringfried Pflug, Jerome Louis, Vincent Debaecker, Uwe Müller-Wilm, and Ferran Gascon. 2017. Sen2Cor for sentinel-2. In *Image and signal processing for remote sensing XXIII*, Vol. 10427. SPIE, 37–48.

[19] Anqi Mao, Mehryar Mohri, and Yutao Zhong. 2023. Cross-entropy loss functions: Theoretical analysis and applications. In *International conference on Machine learning*. PMLR, 23803–23828.

[20] Francisco Mena, Diego Arenas, Marlon Nuske, and Andreas Dengel. 2024. Common Practices and Taxonomy in Deep Multiview Fusion for Remote Sensing Applications. *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing* 17 (2024), 4797–4818. DOI:http://dx.doi.org/10.1109/JSTARS.2024.3361556

[21] Shi Na, Liu Xumin, and Guan Yong. 2010. Research on k-means clustering algorithm: An improved k-means clustering algorithm. In *2010 Third International Symposium on intelligent information technology and security informatics*. Ieee, 63–67.

[22] Song Ouyang and Yansheng Li. 2020. Combining deep semantic segmentation network and graph convolutional neural network for semantic segmentation of remote sensing imagery. *Remote Sensing* 13, 1 (2020), 119.

[23] Jean-François Pekel, Andrew Cottam, Noel Gorelick, and Alan S Belward. 2016. High-resolution mapping of global surface water and its long-term changes. *Nature* 540, 7633 (2016), 418–422.

[24] Shi Qiu, Zhe Zhu, and Binbin He. 2019. Fmask 4.0: Improved cloud and cloud shadow detection in Landsats 4–8 and Sentinel-2 imagery. *Remote Sensing of Environment* 231 (2019), 111205. DOI:http://dx.doi.org/https://doi.org/10.1016/j.rse.2019.05.024

[25] Olaf Ronneberger, Philipp Fischer, and Thomas Brox. 2015. U-Net: Convolutional Networks for Biomedical Image Segmentation. (2015). https://arxiv.org/abs/1505.04597

[26] Michal Segal-Rozenhaimer, Alan Li, Kamalika Das, and Ved Chirayath. 2020. Cloud detection algorithm for multi-modal satellite imagery using convolutional neural-networks (CNN). *Remote Sensing of Environment* 237 (2020), 111446.

[27] Fang Xu, Yilei Shi, Patrick Ebel, Lei Yu, Gui-Song Xia, Wen Yang, and Xiao Xiang Zhu. 2022. GLF-CR: SAR-enhanced cloud removal with global–local fusion. *ISPRS Journal of Photogrammetry and Remote Sensing* 192 (2022), 268–278.

[28] Dai Yamazaki, Daiki Ikeshima, Jeison Sosa, Paul D Bates, George H Allen, and Tamlin M Pavelsky. 2019. MERIT Hydro: A high-resolution global hydrography map based on latest topography dataset. *Water Resources Research* 55, 6 (2019), 5053–5073.

[29] Daniele Zanaga, Ruben Van De Kerchove, Dirk Daems, Wanda De Keersmaecker, Carsten Brockmann, Grit Kirches, Jan Wevers, Oliver Cartus, Maurizio Santoro, Steffen Fritz, and others. 2022. ESA WorldCover 10 m 2021 v200. (2022).

[30] Valentina Zantedeschi, Fabrizio Falasca, Alyson Douglas, Richard Strange, Matt Kusner, and Duncan Watson-Parris. 2019. Cumulo: A Dataset for Learning Cloud Classes. (11 2019). DOI:http://dx.doi.org/10.48550/arXiv.1911.04227

[31] Chao Zhang, Liguo Weng, Li Ding, Min Xia, and Haifeng Lin. 2023. CRSNet: Cloud and cloud shadow refinement segmentation networks for remote sensing imagery. *Remote Sensing* 15, 6 (2023), 1664.