Pontifícia Universidade Católica
do Rio de Janeiro

Felipe Ferrari

Deforestation detection under diverse cloud conditions from the fusion of optical and SAR data with deep learning models

Tese de Doutorado

Thesis presented to the Programa de Pós–graduação em Engenharia Elétrica of PUC-Rio in partial fulfillment of the requirements for the degree of Doutor em Engenharia Elétrica.

Advisor : Prof. Raul Queiroz Feitosa
Co-advisor: Prof. Matheus Pinheiro Ferreira

Rio de Janeiro
September 2024

Pontifícia Universidade Católica
do Rio de Janeiro

Felipe Ferrari

Deforestation detection under diverse cloud conditions from the fusion of optical and SAR data with deep learning models

Thesis presented to the Programa de Pós–graduação em Engenharia Elétrica of PUC-Rio in partial fulfillment of the requirements for the degree of Doutor em Engenharia Elétrica. Approved by the Examination Committee.

Prof. Raul Queiroz Feitosa
Advisor
Departamento de Engenharia Elétrica – PUC-Rio

Prof. Matheus Pinheiro Ferreira
Co-advisor
Universidade de São Paulo – USP

Prof. Gilson Antonio Giraldi
Laboratório Nacional de Computação Científica – LNCC

Prof. Dário Augusto Borges Oliveira
Fundação Getulio Vargas – FGV

Prof. Raian Vargas Maretto
University of Twente – UT

Prof. Luciana Soler
Instituto Nacional de Pesquisas Espaciais – INPE

Rio de Janeiro, September the 6th, 2024

Felipe Ferrari

The author is graduated in Cartographic Engineering from the Instituto Militar de Engenharia (IME). He completed a master's degree in the Defense Engineering Program at IME. He taught programming disciplines, geographic databases, geographic information systems, project management, and entrepreneurship at IME.

Bibliographic data

Ferrari, Felipe

Deforestation detection under diverse cloud conditions from the fusion of optical and SAR data with deep learning models / Felipe Ferrari; advisor: Raul Queiroz Feitosa; co-advisor: Matheus Pinheiro Ferreira. – 2024.

v., 158 f: il. color. ; 30 cm

Tese (doutorado) - Pontifícia Universidade Católica do Rio de Janeiro, Departamento de Engenharia Elétrica.

Inclui bibliografia

1. Engenharia Elétrica – Teses. 2. Sensoriamento Remoto;. 3. Fusão de Dados;. 4. Desmatamento;. 5. Aprendizado Profundo;. 6. Nuvens.. I. Feitosa, Raul Queiroz. II. Ferreira, Matheus Pinheiro. III. Pontifícia Universidade Católica do Rio de Janeiro. Departamento de Engenharia Elétrica. IV. Título.

CDD: 621.3

Acknowledgments

I want to express my heartfelt gratitude to my thesis advisors, who have guided me throughout this journey. Their availability, expertise, patience, and dedication have been invaluable in shaping my research ideas and helping me navigate the complex academic landscape. Their constant support and encouragement have been pivotal in keeping me motivated and focused, and their insightful feedback and suggestions have enabled me to refine my work to its highest standards. I could not have accomplished this feat without their support, and I will always be grateful for their contributions to my academic and personal growth.

I also want to express my deep appreciation to my family, especially my wife, Dayana, for their unwavering support and encouragement throughout my doctoral studies. Her constant love and support have been a source of strength and inspiration, and her sacrifices have allowed me to pursue my dreams and passions. I am truly blessed to have such a wonderful family that believes in me and encourages me to reach for the stars. My success would not have been possible without my wife's constant support and encouragement. I am deeply thankful for her patience, understanding, and love throughout this journey.

I want to express my sincere appreciation to the Pontifical Catholic University of Rio de Janeiro (PUC-Rio) for awarding me the VRAc-I tuition waiver scholarship. This support has been vital in helping me focus on my studies without the added pressure of financial concerns. I am also very grateful for the university's excellent infrastructure, especially the remote access facilities, which were crucial in ensuring that I could continue my studies on schedule during the pandemic. I am deeply thankful that its assistance has been essential to my academic progress.

Finally, I would like to extend my gratitude to the Brazilian Army, which allowed me to dedicate myself entirely to concluding my PhD. Their support and understanding have been instrumental in enabling me to balance my military duties with my academic pursuits. I am deeply grateful for their unwavering support. I am proud to serve my country and thankful for the opportunities the Brazilian army has provided me.

This study was financed in part by the Coordenação de Aperfeiçoamento de Pessoal de Nível Superior - Brasil (CAPES) - Finance Code 001"

Abstract

Ferrari, Felipe; Feitosa, Raul Queiroz (Advisor); Ferreira, Matheus Pinheiro (Co-Advisor). **Deforestation detection under diverse cloud conditions from the fusion of optical and SAR data with deep learning models**. Rio de Janeiro, 2024. 158p. Tese de doutorado – Departamento de Engenharia Elétrica, Pontifícia Universidade Católica do Rio de Janeiro.

Deforestation monitoring is highly dependent on human specialists analyzing cloud-free optical images. Developing methodologies that minimize the dependency on human specialists and the availability of cloud-free optical images can contribute to environmental conservation efforts. Despite the more accessible use of optical images for deforestation detection, the presence of clouds in these images limited the operation, forcing the selection of images at specific times of the year when the presence of clouds is lower. However, even in the dryest period of the year, there are certain regions of the Brazilian Amazon Forest where the cloud presence is still high. On the other hand, the SAR images suffer less interference from clouds, but are more challenging to interpret. Aiming to take advantage of both, we investigated Deep Leaning methods of fusion of these data, especially in diverse cloud presence conditions, which is an unexplored subject, as best as we know. We proposed using a pre-training strategy from single-modality optical and SAR models. We investigated ways to combine the SAR images across the analyzed period. We also investigated Vision Transformer-based architectures. Our best results reached the same F1-Score result fusing SAR images with optical images with diverse cloud conditions and with cloud-free optical images.

Keywords

Remote Sensing; Data Fusion; Deforestation; Deep Learning; Clouds.

Resumo

Ferrari, Felipe; Feitosa, Raul Queiroz; Ferreira, Matheus Pinheiro. **Detecção de desmatamento sob condições diversas de nuvens a partir da fusão de dados ópticos e SAR com modelos de aprendizado profundo**. Rio de Janeiro, 2024. 158p. Tese de Doutorado – Departamento de Engenharia Elétrica, Pontifícia Universidade Católica do Rio de Janeiro.

O monitoramento do desmatamento é altamente dependente de especialistas humanos que analisam imagens ópticas livres de nuvens. O desenvolvimento de metodologias que minimizem a dependência de especialistas humanos bem como da disponibilidade de imagens ópticas livres de nuvens pode contribuir para os esforços de conservação ambiental. Apesar do uso mais fácil de imagens ópticas para detecção de desmatamento, a presença de nuvens nessas imagens limita a sua utilização, obrigando a seleção de imagens em épocas específicas do ano em que a presença de nuvens é menor. Porém, mesmo no período mais seco do ano, existem certas regiões da Floresta Amazônica Brasileira onde a presença de nuvens ainda é elevada. Por outro lado, as imagens SAR sofrem menos interferência das nuvens, mas são mais difíceis de interpretar. Visando aproveitar ambos, investigamos métodos de fusão desses dados, especialmente usando imagens com condições diversas de nuvens, que é um assunto inexplorado até onde sabemos. Propusemos o uso de uma estratégia de pré-treinamento a partir de modelos ópticos e SAR. Investigamos arquiteturas baseadas em Vision Transformers. Nossos melhores alcançaram o mesmo resultado de F1-Score usando a fusão de imagens SAR com imagens ópticas com condições diversas de nuvens e imagens ópticas livres de nuvens.

Palavras-chave

Sensoriamento Remoto; Fusão de Dados; Desmatamento; Aprendizado Profundo; Nuvens.

Table of contents

List of figures

List of tables

List of Abreviations

ANN – Artificial Neural Networks
ASPP – Atrous Spatial Pyramid Pooling
BAF – Brazilian Amazon Forest
BLA – Brazilian Legal Amazon
CNN – Convolutional Neural Networks
DL – Deep Learning
ESA – European Space Agency
FCN – Fully Convolution Network
GAN – Generative Adversarial Network
GELU – Gaussian Error Linear Unit
INPE – *Instituto Nacional de Pesquisas Espaciais*
LIDAR – Light Detection And Ranging
LSTM – Long Short-Term Meomory
MLP – Multilayer Perceptron
NDVI – Normalized Difference Vegetation Index
ReLU – Rectified Linear Unit
RNN – Recurrent Neural Netowrk
RS – Remote Sensing
SAR – Synthetic Aperture Radar
SGD – Stochastic Gradient Descent
SOTA – State-of-the-art
ViT – Vision Transformer

1
Introduction

Since the first aerial photographs taken by Félix Nadar from balloons in the 1850s, Remote Sensing (RS) technologies developed rapidly until the launch of modern satellites [20], these systems can board different sensors, like multispectral, hyperspectral, thermal, and radar. The main advantage of these systems is the ability to obtain data from the Earth's surface and atmosphere at any location on the planet and with regular time intervals between collections (depending on the orbit and the spatial and temporal resolutions chosen for the satellite) [22].

However, the images produced by these systems need to be interpreted to generate helpful information. When humans analyze an RS image, their brain interprets the features found in the scene (shapes, sizes, patterns, colors, textures, shadows, location, and associations) by comparing them with their personal experience [20, 22, 23]. Consequently, the manual interpretation of the images depends on qualified specialists (demanding training time) to correctly identify features in the images, limiting the productivity of generating information from the RS images.

According to the Union of Concerned Scientists, there were 1,182 satellites in Earth's orbit at the beginning of 2023, for which the purpose is Earth observation [24]. In a scenario in which an increasing number of RS imaging systems are available, visual interpretation by human experts considerably limits the potential for extracting information from the generated images, either because of the complexity involved in training these specialists or because of the natural limitations of human productivity. Automatic methods for interpreting changes in RS images, which can minimize the dependence on specialists, have been developed over the last few years, among which Deep Learning (*Deep Learning* - DL) methods represent the State-of-the-art (SOTA) [25].

The Brazilian Amazon Forest (BAF) is the largest rainforest on Earth. It is a habitat for millions of species and regulates the planet's climate [26]. Deforestation is an important driver of land use change, as this process consists of suppressing vegetation areas by anthropogenic actions [1].

Due to recent increases in deforestation rates in the BAF [27], monitoring this biome has become even more relevant. Due to the almost non-existent

transport infrastructure in native forest regions and their extensive areas, the use of RS images becomes an essential tool for detecting new areas of deforestation, especially with RS satellite systems, for which the image acquisition covers large areas and is independent of local infrastructure.

In this context, the PRODES program, conducted by the National Institute for Space Research (INPE) since 1988, detects new areas of deforestation by comparing images taken annually from optical sensors, which seek to locate changes in native vegetation produced by new deforestation areas [1]. These sensors can detect spectral responses in different wavelength ranges and produce images. To minimize cloud presence in regions with a high occurrence of clouds, PRODES employs more than one image [1]. However, there are regions in the BAF where the probability of the presence of clouds is so high that it is unlikely to obtain images with low cloud cover throughout the whole year [9].

Although the use of optical images facilitates the interpretation by human experts, they can suffer interference from atmospheric effects, especially in case of obstructions caused by the presence of clouds. When clouds are present, an alternative to optical images is using Synthetic Aperture Radar (SAR) images, which are much less affected by the presence of clouds [20]. Thus, using SAR images to monitor areas with the recurrent presence of clouds raises a practical possibility [28].

Even though DL models can employ data from each sensor individually, these data sources can also be utilized together. Optical and SAR data fusion techniques have proved helpful in computer vision tasks when applied to cloud-free RS images [15]. However, none of the works so far aimed to minimize the problem of cloud occurrence, conditioning the application of data fusion to the availability of cloud-free optical images. As cloud occurrence is frequent in the BAF region, developing new models for detecting deforestation independent of the atmospheric conditions is relevant when taking images.

In summary, this work investigates fusion models based on different base architectures, combining SAR data and optical images with diverse cloud conditions for deforestation detection. We proposed a new training strategy to minimize the optical and SAR models' convergence differences. We concluded that our methodology delivered deforestation predictions fusing SAR and cloud-diverse optical images very close to the predictions fusing SAR and cloud-free optical images.

Hypothesis

The central hypothesis of this work is that Deep Learning models, which integrate Optical and SAR data, can assess the reliability of each data source and leverage this understanding to extract the most relevant information from each.

Objective

Based on this hypothesis, this work aims to investigate Deep Learning models capable of fusing optical and SAR data under diverse cloud conditions. To evaluate these models, they will be refined to identify new deforestation areas from multitemporal Remote Sensing optical and SAR data fusion, regardless of the cloud condition of the optical data.

To test our hypothesis, we established the following specific objectives, focusing on identifying new deforestation areas from multitemporal optical and SAR images:

1. Investigate Deep Learning architectures to identify new deforestation areas from multitemporal optical and SAR images.

2. Based on these single-modality architectures, explore different methods of utilizing multitemporal optical and SAR images, selecting the one that yields the best results.

3. Investigate end-to-end DL models to fuse SAR and optical multitemporal images, using the models with cloud-free optical images as the baseline to evaluate the models using optical data with diverse cloud conditions.

Contributions

The main contributions of this work are:

1. We adapted Deep Learning models to fuse Remote Sensing images from SAR and optical sensors **regardless of the cloud presence in the optical images** to identify new deforestation areas in the Brazilian Amazon Forest biome.

2. We proposed and validated a training strategy for the fusion of optical and SAR images.

3. We investigated a different method for combining SAR images for deforestation detection, capturing the features across the analyzed period and the behavior of this data in other seasons.

4. We investigated the recent new architectures, especially those based on the Vision Transformer concept, as a substitute for convolution in DL data fusion models from different sources.

5. We evaluated the feasibility of the investigated models, especially in terms of the demand for computational resources for training and inference.

Text Organization

The rest of this work is organized as follows:

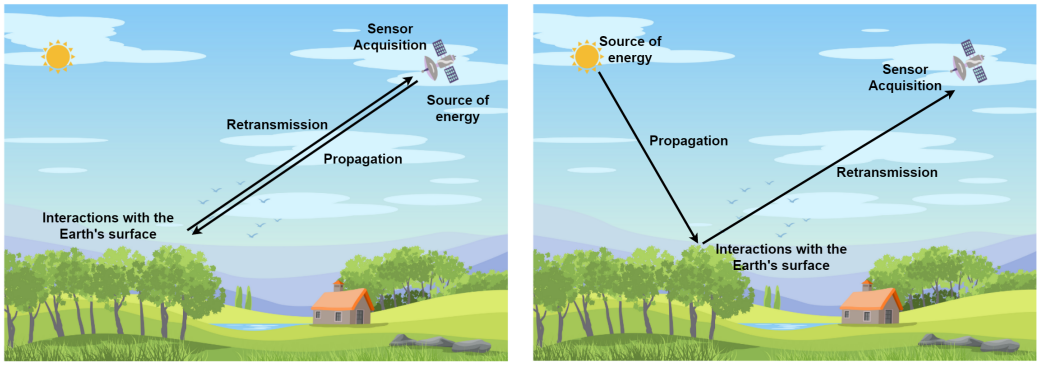
– Chapter 2 details the theoretical foundations of Remote Sensing and the models used throughout the work.
– Chapter 3 reviews prominent publications on deforestation detection, data fusion, and cloud presence (missing data) in remote sensing images.
– Chapter 4 presents the work methodology to achieve the proposed objective.
– Chapter 5 presents the experimental protocol, including the utilized data.
– Chapter 6 presents the experiments' results and the respective discussion.
– Chapter 7 concludes the work, presenting a brief discussion about the results achieved and the potential contribution for future works.

2
Theoretical Foundations

This chapter introduces the main theoretical foundations related to this work. We present the remote sensing (RS) data sources employed, including the Optical and Synthetic Aperture Radar (SAR), the PRODES project, in which data will be utilized as ground truth. Closing this chapter, we explain the Deep Learning (DL) concepts employed in this work's methodology.

2.1
Spaceborne Remote Sensing Data Sources

In the spaceborne RS systems, the satellites' sensors measure the energy reflected from the Earth's surface, transforming this measures in images [23]. The data acquisition process of these systems consists of the following elements: source of energy, propagation through the atmosphere, interactions with the Earth's surface, retransmission through the atmosphere, and satellite acquisition sensor [20]. These systems can be classified as active or passive RS depending on the energy source. In the active systems, the energy source is emitted by the system, while in the passive systems, the energy source is the environment, usually the Sun [22]. Figure 1 shows examples of these elements in the case of active and passive sensors.



1(a): Active sensors.

1(b): Passive sensors.

Figure 1: Data acquisition process of spaceborne RS systems.

These days, many nations and a broad range of corporations operate spaceborne RS systems specially designed to collect information concerning

crops, forests, water bodies, land use, cities, defense, and mineral resources from the Earth's surface [20, 22, 29]. The applications in which the images generated from these systems range from global resource monitoring to such activities as city planning, agricultural development, natural disaster response, defense activities, and forest monitoring [30–34].

The interaction between energy and Earth's surface is the information to be detected by satellite sensors. These interactions vary based on the energy wavelength [23]. The optical and radar are the most usual sensors boarded in RS spaceborne systems focused on Earth's surface observation [29].

2.1.1 Optical Data

The optical spaceborne systems operate within the optical spectrum, in which the wavelengths extend from approximately 0.3 to 14 μm. This range includes ultraviolet, visible, near-, mid-, and thermal infrared [20]. The sensed data are acquired in spectral regions, also called spectral bands, and are related to the reflectance (or the emissivity in the case of thermal sensors) of Earth's surface for each spectral band.

Civilian organizations operate many optical spaceborne systems, such as Landsat, SPOT, and Sentinel-2. The Sentinel-2 is managed by the European Space Agency (ESA) and consists of Sentinel-2A and Sentinel-2B, launched in June 2015 and March 2017, respectively. They are designed to provide continuous multispectral imagery to assist land management, agriculture, forestry, disaster response, and security programs. Sentinel-2 provides 12 bits, high-resolution multispectral imagery, including 13 spectral bands covering a 290-km swath at resolutions between 10 and 60 meters [19], as shown by Table 1.

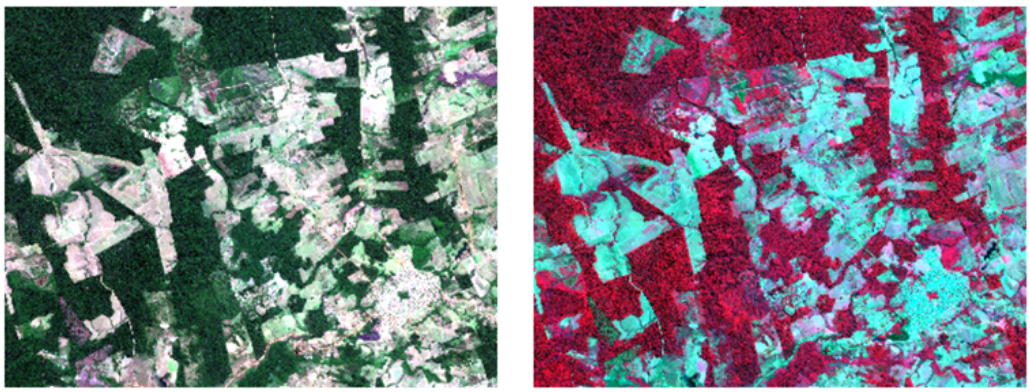
The Sentinel-2 has two products available to the users, which differ from each other by the processing level applied. The product Level-1C is radiometric and geometric corrected, including orthorectification and spatial registration), representing the Top-Of-Atmosphere reflectances. In product Level-2A, an atmospheric correction is applied in addition to the Level-1C corrections, reaching Bottom-Of-Atmosphere reflectances [19].

Remote sensing instruments detect energy across many spectral bands, including bands with wavelengths outside the visible wavelength range region. Human vision perceives the colors from the combinations of the primary colors: red, green, and blue. This limits the representation of the images by compositions of its bands, in which each band is presented by a primary color, highlighting specific bands depending on the task. Each band can represent

Band	**Central wavelength (nm)**	**Bandwidth (nm)**	**Spatial resolution (m)**
1	443	20	60
2	490	65	10
3	560	35	10
4	665	30	10
5	705	15	20
6	740	15	20
7	783	20	20
8	842	115	10
8b	865	20	20
9	945	20	60
10	1375	30	60
11	1610	90	20
12	2190	180	20

Table 1: Wavelengths and spatial resolutions of Sentinel-2 [19].

the sensed spectral band or arithmetic operations between them. The most usual is the true color composition, as shown by Figure 2(a), in which the reflectance values in the bands 4, 3, and 2 are represented by the colors red, green, and blue, respectively, which produces images similar to the human vision. However, other bands' compositions can be utilized, like the false colors, as shown by Figure 2(b), in which the reflectance values in the bands 8, 4, and 3 are represented by the colors red, green, and blue, which produces images highlighting the vegetation [23].



2(a): True colors 2(b): False colors

Figure 2: Examples of optical color compositions.

The sensed energy acquired by the satellites' sensors is affected by the atmospheric conditions during the propagation and retransmission [20], as shown by Figure 1. Some of these atmospheric effects can be corrected, but not the presence of clouds, which block the energy reflected by the Earth's surface

and produce shadows on the surface, as shown in Figure 3. The presence of clouds is common in the Brazilian Amazon Forest. In some regions, it is possible to acquire images free of clouds during some period of the year (called the dry period). However, there are other areas where clouds are constant throughout the year [9]. For these last areas, optical imagery can't be employed for tasks in which the Earth's surface is the essential data, like forest monitoring.

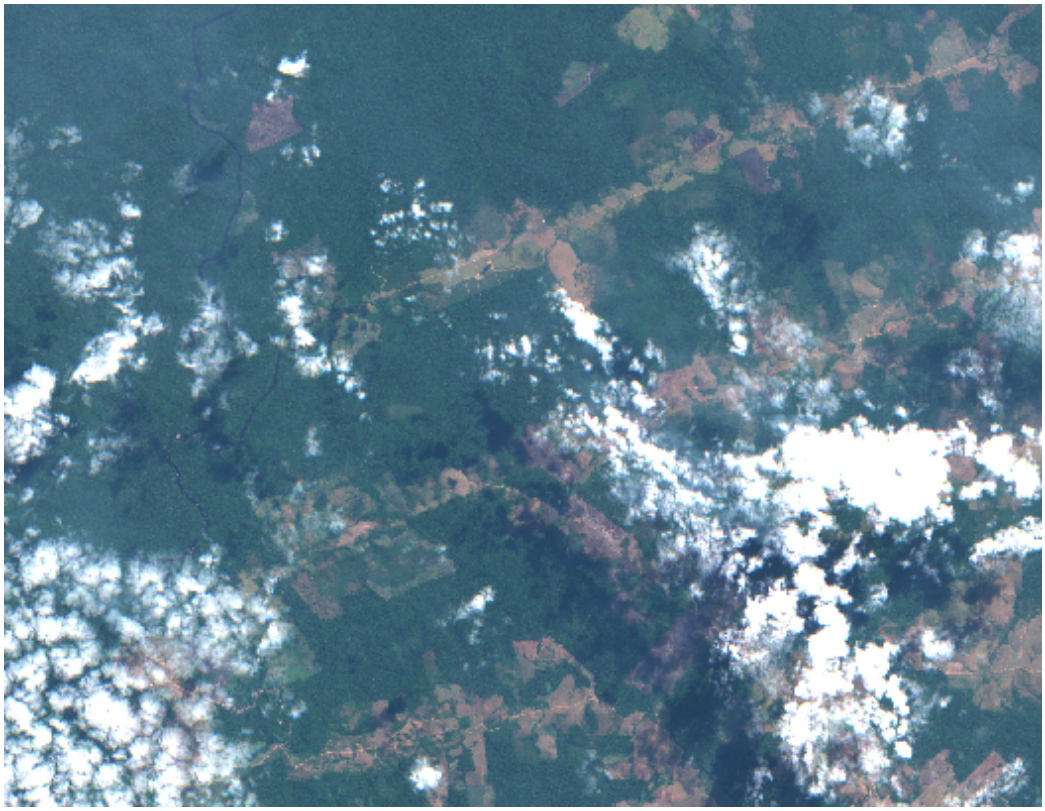


Figure 3: Sentinel-2 image with the presence of clouds.

2.1.2 Synthetic Aperture Radar Data

The active microwave sensor is an example of an active sensor that broadcasts a directed pattern of microwave energy to the Earth's surface, receiving the scattered energy back to the sensor. As optical sensors produce images from reflected solar energy, their use is constrained by weather and the time of day, as shown in section 2.1.1. On the other hand, active sensors acquire the energy they produce, and their usage is subject to fewer constraints. The main capabilities of active microwave systems include [23]:

- Less weather dependency, constrained only by extreme weather events.
- Operates free of atmospheric effects and at any time of the day.

– Record details of the emitted energy (such as wavelength, phase, and polarization), comparing them to the returned sensed signal.

– Acquire detailed imagery at great distances.

The spaceborne SAR systems employ a short physical antenna moving along its orbit, which synthesizes the effects of a very long antenna, modifying the data recording and applying processing techniques. They provide unique images that correspond to geometrical properties of the Earth's surface in diverse weather conditions [35]. Such applications include polar ice research, vegetation, biomass measurements, and soil moisture mapping [36–39].

Two main characteristics affect the signal transmission of SAR systems: the wavelength and the polarization. Table 2 presents the usual wavelength ranges used by SAR systems. Each wavelength band is represented by a letter originally designated by military security and is still utilized.

Band Designation	**Wavelength (cm)**
K_a	$0.75 - 1.1$
K	$1.1 - 1.67$
K_u	$1.67 - 2.4$
X	$2.4 - 3.75$
C	$3.75 - 7.5$
S	$7.5 - 15$
L	$15 - 30$
P	$30 - 100$

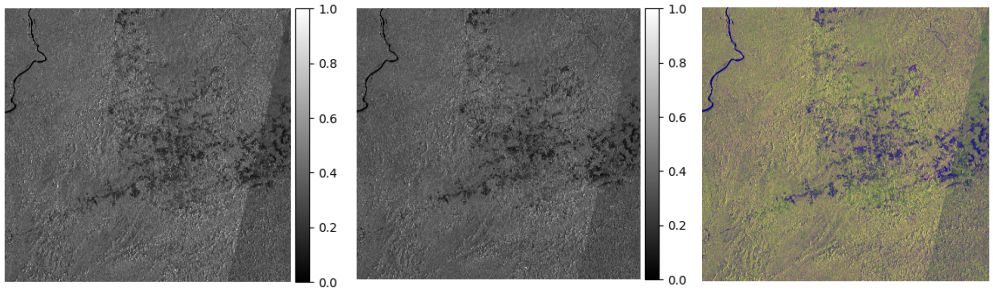
Table 2: Usual wavelengths of SAR systems and their designations [20].

The influence of the atmosphere on the radar signal is related to the signal wavelength. In general, radar signals are relatively unaffected by the presence of clouds. However, precipitation can interfere with the signal, especially in shorter wavelengths (bands K, X, and C). Another effect of weather interference in SAR images occurs in the case of rainfalls, which significantly changes the moisture of the soil or plants, affecting their signal backscatter. Spaceborne SAR systems have frequently used X, C, S, and L bands [40].

The radar signal can be emitted and sensed in different modes of polarization. The polarization of an electromagnetic wave describes the geometric plane in which the energy oscillates. Usually, spaceborne SAR systems emit energy in two polarization modes: vertical (the oscillation plane is perpendicular to the antenna) and horizontal (the oscillation plane is parallel to the antenna). As the sensor usually can emit and sense in both polarization modes, four polarization combinations are available [20]:

– HH: The energy is emitted and sensed in horizontal polarization;
– VV: The energy is emitted and sensed in vertical polarization;
– HV: The energy is emitted in horizontal polarization and sensed in vertical polarization;
– VH: The energy is emitted in vertical polarization and sensed in horizontal polarization.

The Sentinel-1 is a spaceborne SAR system operated by ESA, with a C-band sensor designed to image global landmasses, coastal zones, sea–ice, polar areas, and shipping routes at high resolution and covering the global ocean. The primary sensor can generate images with 10 m spatial resolution and VV and VH polarization modes. The system operates two satellites, which deliver images with six days of repetition [41]. Figure 4 presents an example of Sentinel-1 on VV, Figure 4(a), VH, Figure 4(b), polarization modes, and the respective colored composition in which VV, VH and the VV/VH ratio are represented by red, green and blue colors, Figure 4(c).



4(a): VV polarization. 4(b): VH polarization. 4(c): colored composition.

Figure 4: Sentinel-1 image in VV (a) and VH (b) polarization modes and VV/VH composition (c).

2.2 PRODES Program

Since 1988, the PRODES program, a part of the Program of Monitoring of Amazon and other Biomes (also called PAMZ+ and conducted by the INPE), performs the primary forest loss inventory through Earth observation satellite images. The term deforestation used by PRODES is defined as the primary vegetation suppression by anthropogenic actions. Deforestation begins with the intact forest and usually ends with the full forest conversion by other land coverage. [1].

The PRODES employs optical images generated by optical satellite systems. Currently, PRODES utilizes Landsat-8, Landsat-9, Sentinel-2, and

CBERS-4/4A images. PRODES splits the area based on the Landsat grid scenes, as shown by Figure 5 to monitor the BAF region. From each area, an annual image is selected from the dry season (between July and September) based on weather parameters. Due to the high cloud coverage probability in some BAF regions [9], more than one image can be used. However, if some regions are cloud-covered in all analyzed images in a specific year, these regions can't be inspected and are included in the cloud mask of that year [1].

Figure 5: Landsat grid coverage in the Brazilian Amazon Forest [1].

Deforestation is identified by photointerpretation of the pair of optical images, carried out by trained specialists, who delimit the deforestation polygons directly on the computer screen. These experts identify the pattern of change in forest cover based on the main elements observable in the image pair: tonality, color, shape, texture, and context. The image from the analyzed year is compared to the image from previous year, looking for differences in the elements that indicate a new area of deforestation. Cloud-free images from previous years can be used to compare when clouds cover last year's image. [1]. An example of the deforestation areas identified by PRODES is shown in Figure 6, in which the deforestation identified until 2007 and the yearly identified deforestation, including the identification year, are presented in blue and red, respectively.

The data acquired by PRODES are made publicly available through the portal Terrabrasilis [42]. Only new deforestation areas of at least 6.25 ha are

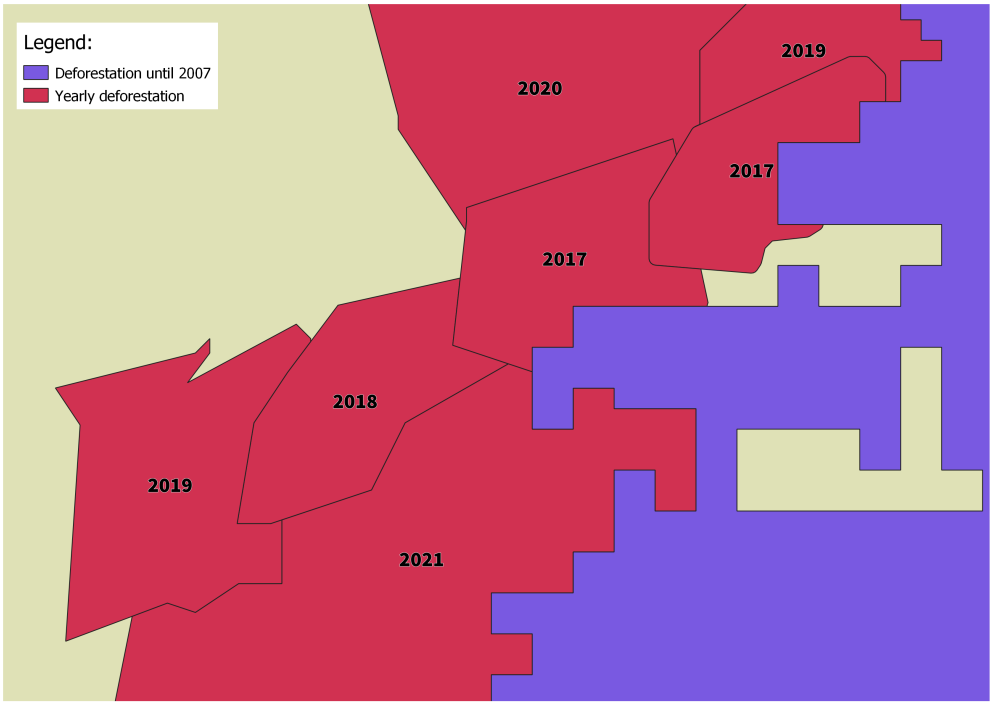


Figure 6: Deforestation areas identified by PRODES.

included in the annual deforestation increment [1]. Terrabrasilis provides the following data:

- Increment deforestation: deforestation areas identified annually from the year 2008.
- Previous deforestation: mask of deforestation areas identified before the year 2008.
- Hydrography: Mapping of water bodies (rivers, lakes, dams, and reservoirs).
- Cloud mask: Areas covered by clouds and shadows in the utilized optical image.
- No forest: Areas not included in the Forest class adopted in the mapping, which consequently, are not objects of analysis and mapping by the PRODES.
- Residual deforestation: Consists of deforested areas from previous years that, for some reason, were not observed.

Although PRODES does not carry out a field survey to inspect the actual occurrence of new deforestation areas, it can be considered a Gold-Standard reference data to assess the deforestation classification accuracy, due to the achieved accuracy higher than 93% [43, 44]. Many works aimed to identify

deforestation areas from RS images utilized PRODES data as reference [13, 45–47].

2.3 Convolutional Networks

The RS data images are typically stored in the form of 2D grid layers, in which each layer represents diverse information, like a band of Optical satellite systems or a polarization combination of SAR satellite systems. The Convolutional Network, also called Convolutional Neural Networks (CNN) [48] is a specialized neural network for processing this data.

These models are based on the convolution operation, defined formally for real-valued arguments by the Equation 2-1, in which x is the input, and w is the kernel. In image-based applications, the input and the kernel are discretized and multi-dimensional. An example of the convolution applied in two-dimensional arrays as input and kernel can be found in Equation 2-2, in which I and K are the input and kernel, respectively. Figure 7 shows an example of a convolution evaluation with a 2×2 kernel applied in 4×4 two-dimensional image, highlighting in which elements the operation was applied [49].

$$s(t) = (x * w)(t) = \int x(a)w(t-a)da \tag{2-1}$$

$$S(i,j) = (I * K)(i,j) = \sum_m \sum_n I(i+m, j+n)K(m,n) \tag{2-2}$$

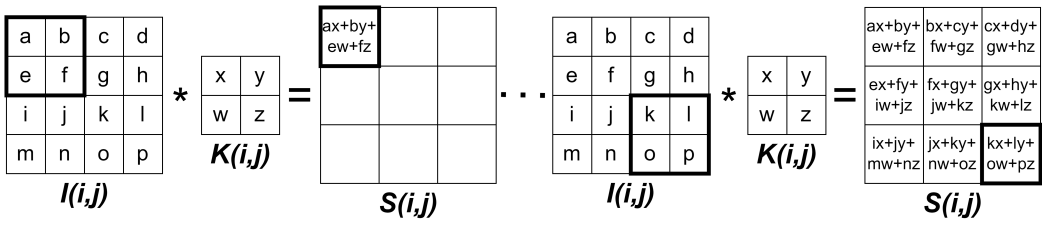


Figure 7: 2D Convolution operation example.

A typical DL model based on the convolutional operation and widely used for RS tasks is the ResUnet [2], which improved the U-Net [50], including the residual learning concept [51], minimizing the vanish gradient difficulty. Figure 8 shows the residual block (a) and the general architecture of the ResUnet.

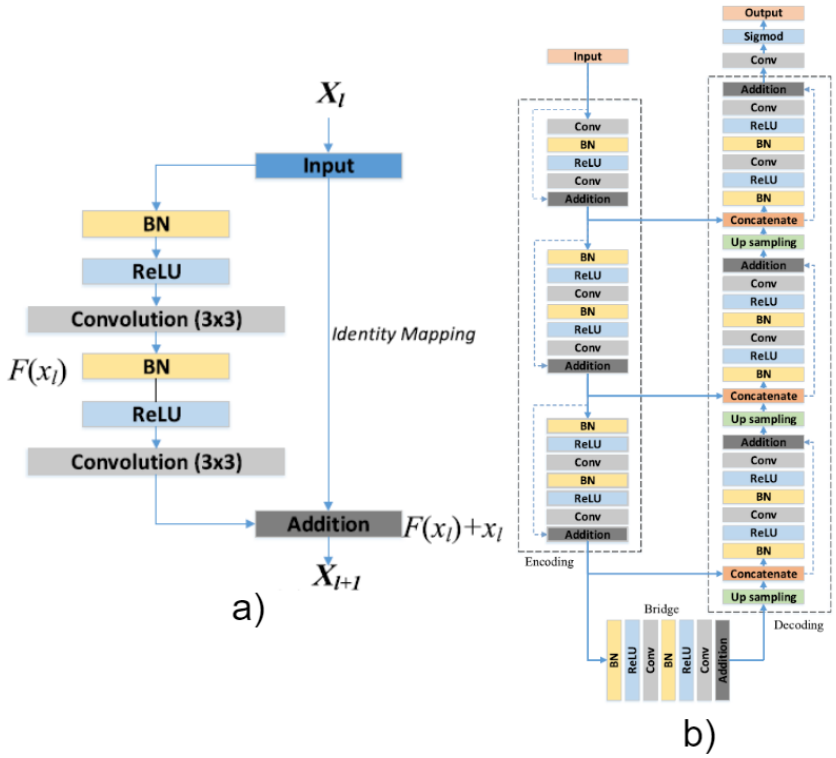


Figure 8: The residual block in detail (a) and ResUnet general architecture (b). Adapted from [2].

2.4 Vision Transformer Networks

Many DL models designed to identify deforestation areas from RS data were variations of CNN models [12, 52–58]. However, recently, these models have been superseded by models based on the Vision Transformer (ViT) model [3], which was based on the Transformer concept, based on the self-attention mechanism, initially employed for language translation task [4].

Initially, ViT was designed to classify images. Figure 9 shows the ViT model overview. The Transformer uses a latent vector with a size D, similar to a CNN feature maps' depth. In its first step, the input image is split into a sequence of two-dimensional flattened patches, in which its initial pixel values are projected to the latent size D, generating the embedded patches. A position embedding is then added to the embedded patches, and its result serves as input to the Transformer Encoder.

The Transformer Encoder consists of a sequence of L blocks, as presented by Figure 9b. The blocks have two parts. Around each block part, there is a residual connection. The first part consists of a layer normalization followed by multi-head attention, and the second one consists of a layer normalization followed by multilayer perceptrons (MLP). The MLP consists of two densely connected layers activated by Gaussian Error Linear Unit (GELU), a non-

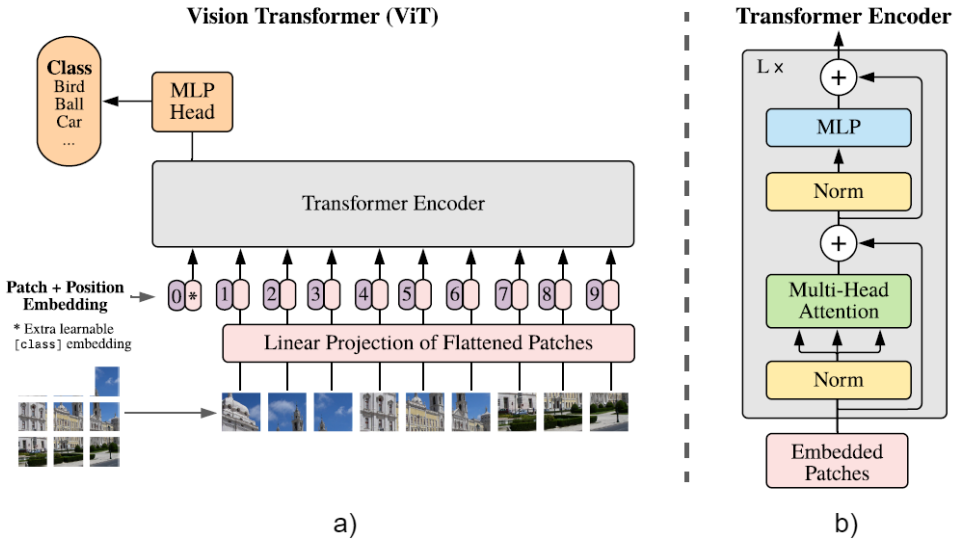


Figure 9: Vision Transformer: the model overview (a), and the Transformer Encoder (b) [3].

linear activation function [59]. Figure 10 describes the multi-head attention layer.

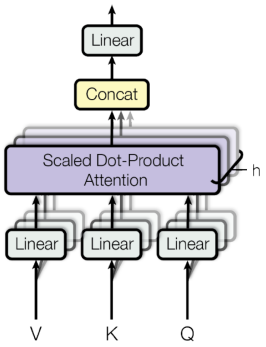


Figure 10: Multi-head attention description [4].

In each of the h attention heads, the same V, K, and Q are projected by the product with L_i^V, L_i^K, and L_i^Q, respectively, where i is the head index. Equation 2-3 describes each i-th head attention output, where d_K is the dimension related to the matrix $L_i^K K$. The multi-head attention output, described by Equation 2-4, is given by the concatenation of all attention heads followed by a projection, given by the product with L^O [4].

$$\text{Attention}_i(Q,K,V) = \text{softmax}(\frac{(L_i^Q Q)(L_i^K K)^T}{\sqrt{d_K}})(L_i^V V) \tag{2-3}$$

$$\text{Multi-head}(Q,K,V) = (\text{Concat}(\text{Attention}_i(Q,K,V)))L^O \tag{2-4}$$

The Swin Transformer [5] improved from ViT, replacing the multi-head self-attention layer with the window multi-head self-attention (W-MSA)

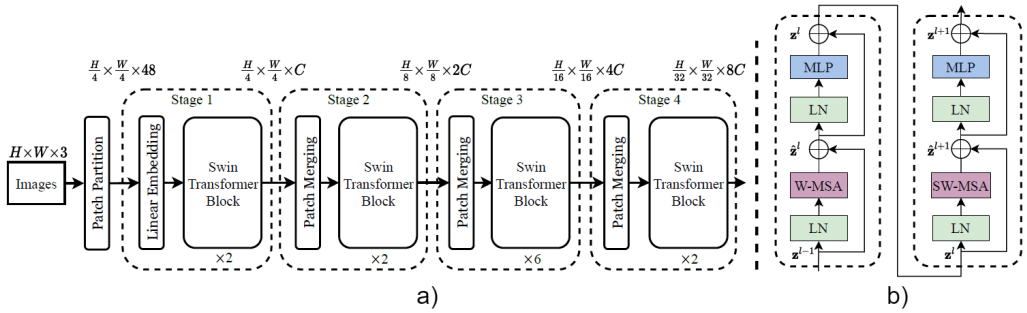


Figure 11: The Swin Transformer model: the overview (a) and two successive Swin Transformer blocks (b)[5].

and shifted window multi-head self-attention (SW-MSA), which are applied alternately in each successive Swin Transformer block. It also included the successive downsize operation performed by the Patch Merging operation. The model overview and two successive Swin Transformer blocks are shown in Figure 11.

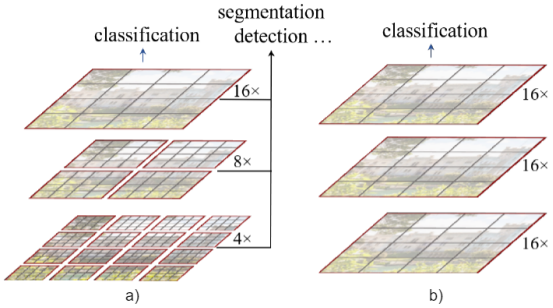


Figure 12: The Swin Transformer hierarchical feature maps (a) in comparison to ViT (b). The attention operations are limited to the red boxes [5].

In the Swin Transformer, the attention mechanism is limited to the patches inside the windows. The window size is a parameter that must be chosen. These windows limitations allow a faster computation than ViT, in which the attention must be computed globally. Figure 12 compares the feature maps distribution between the layers inside the models [5].

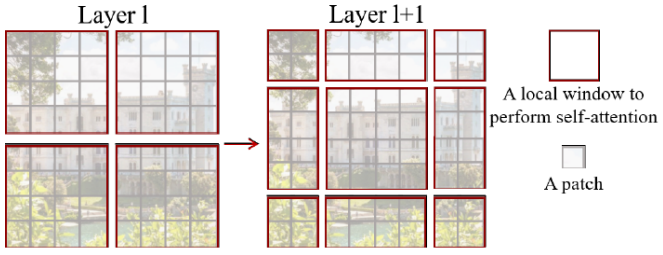


Figure 13: The Swin Transformer windows shift [5].

Despite the reduction in complexity, applying self-attention to a window area could ignore important relations between information that are not close to each other in the image. To minimize this problem, the shifted window is utilized, as shown by Figure 13, which presents how the windows are shifted between two successive layers [5].

The Swin-unet [6], based on [5, 50], was proposed for the semantic segmentation task in medical images. Figure 14 shows the Swin-Unet model architecture.

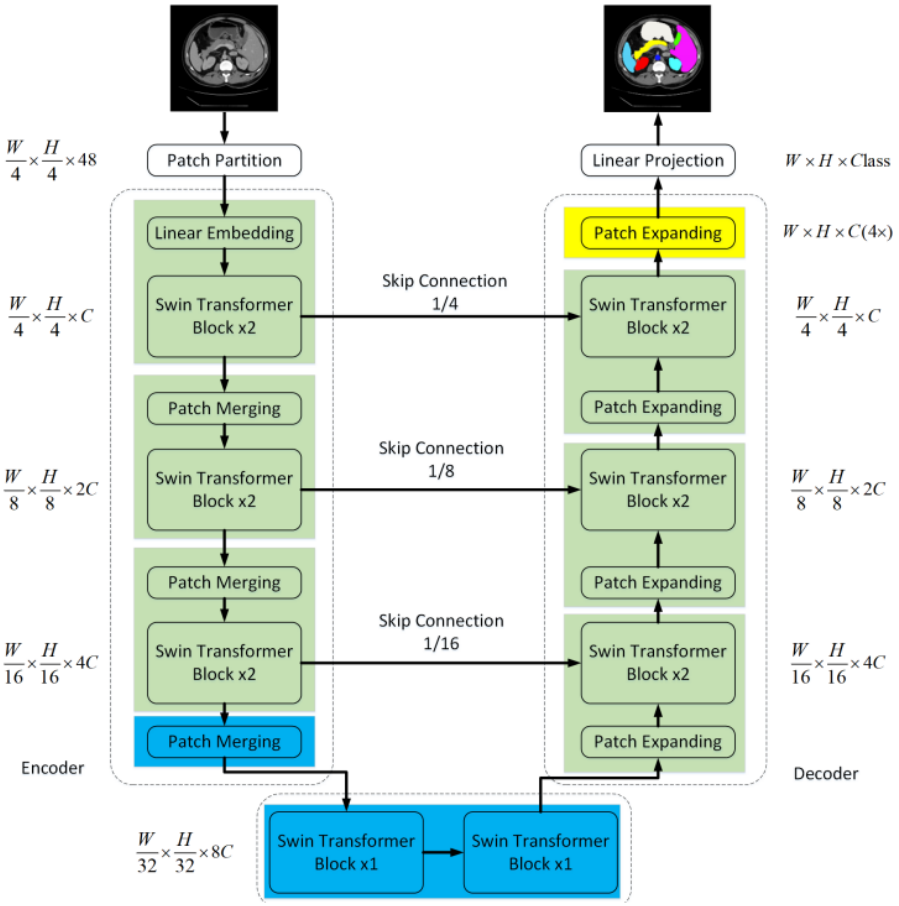


Figure 14: The Swin-Unet model architecture [6].

2.5 Change Detection

Change detection is a computer vision task that aims to identify differences in the state of an object or a phenomenon between two different instants. The Remote Sensing field is based on the premise that the difference of the state or phenomenon on the Earth's surface results in differences in the energy interaction and, consequently, in the energy sensed by the sensor [60].

In a deep learning context, change detection methods can be classified into three sets based on the training strategy. The first one employs supervised methods, where enough labeled data is available for training purposes. Another approach utilizes fully unsupervised methods and is suitable when labeled data is unavailable. The last one is used when labeled data is limited and needs to be used more to train models. They consist of the methods that apply transfer-learning techniques, pre-training the model with available labeled data, and refining it with the limited available labeled data [61].

The change detection capability strongly depends on the feature extraction. The feature extraction strategies can be classified as single-stream or multi-stream (usually dual-stream). The temporal aggregation strategy (also called fusion strategy, but this terminology will not be utilized to avoid misleading with sensor data fusion) also affects the change detection quality. The most usual temporal aggregation strategies are concatenation, difference, and summation operations performed in the images' channel dimension [62, 63]. Figures 15 and 16 present the single and multi-stream feature extraction strategies, respectively.

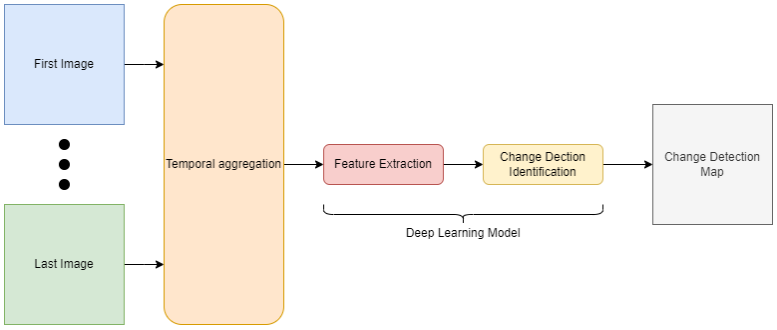


Figure 15: Single-stream feature extraction strategy.

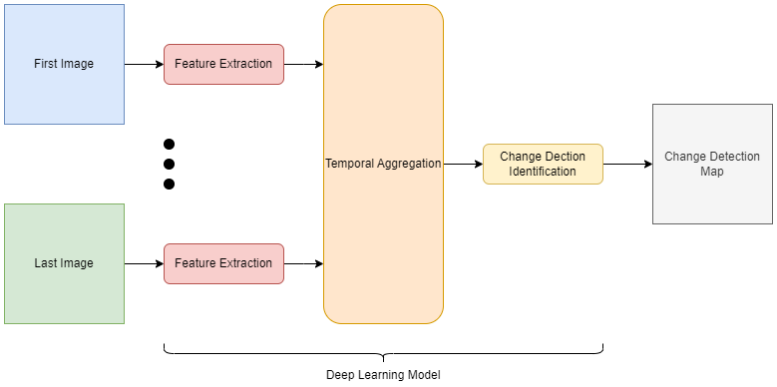


Figure 16: Multi-stream feature extraction strategy.

2.6 Data Fusion

The number of available RS systems has increased in the last few years. Multiple sensors provide a large variety of RS data. Despite each sensor type capturing diverse characteristics, the multi-sensor fusion can exploit the complementary and correlated, generating more comprehensive and accurate information [64, 65]. Figure 17 presents the data fusion workflow for RS applications [7].

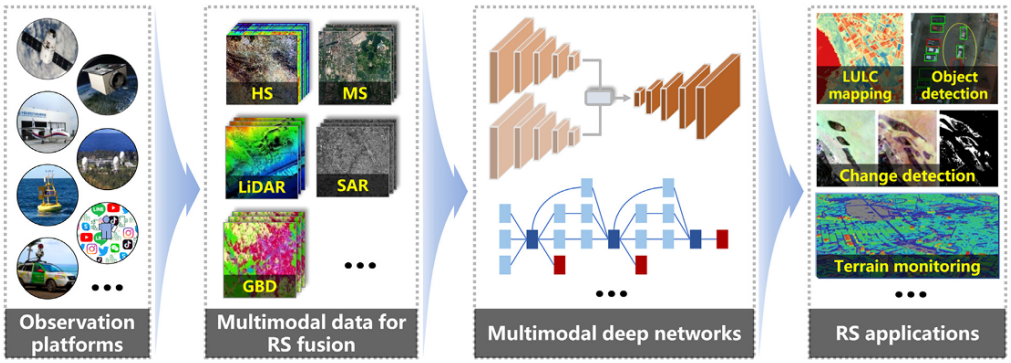


Figure 17: Data Fusion for Remote Sensing tasks [7] .

According to Li *et al.* [7], RS data fusion can be classified into two groups based on the sensors' characteristics: homogeneous and heterogeneous fusion. Homogeneous fusion aims to increase the spectral, spatial or temporal resolution. They include the fusion of multispectral or hyperspectral (narrow wavelengths spectral bands and higher spatial resolution) with panchromatic data (broad wavelengths spectral bands and smaller spatial resolution), called MS and HS panshaperning, respectively. Spatiotemporal fusion (also a homogeneous fusion) aims to increase spatial and temporal resolutions, fusing lower spatial and higher temporal resolutions with higher spatial and lower temporal resolutions. Different from homogeneous, heterogeneous fusion aims to extract complementary information from diverse sensors. Each sensor modal has diversified capabilities. Optical and SAR fusion, e.g., can explore the Earth's surface iteration with different wavelengths and optical and Radar bands in a complementary way.

Based on which level the fusion takes place, we can organize DL data fusion into three categories: pixel-level, feature-level, and decision-level. Pixel-level, also called low level or raw data level, we generate new images (fused) from the original data. Feature-level, in which the fusion occurs on the feature extractors' outputs. Decision-level, also called high level, refers to the methods in which the data fusion is done after the individual classifications [7, 8, 66].

Fusion Class	Domain
Homogeneous	Panshaperning
	Hyperspectral Panshaperning
	Hyperspectral - Multispectral
	Spatiotemporal
Heterogeneous	Hyperspectral - LiDAR
	SAR - optical
	RS - Geospatial Big Data

Table 3: Classifications of RS data fusion classification tasks (adapted from [7]).

Figures 18, 19, and 20 present examples of the fusion on pixel-, feature-, and decision-levels, respectively.

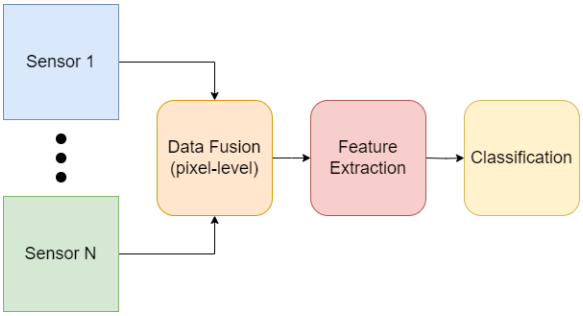


Figure 18: Pixel-level data fusion (adapted from [8]).

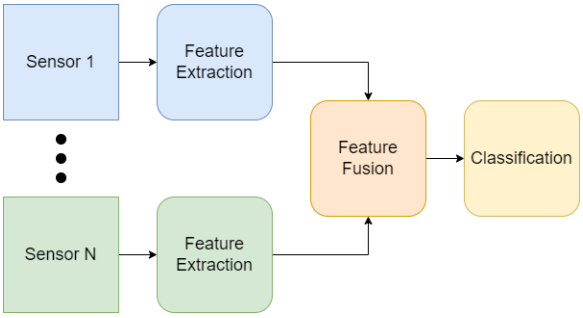


Figure 19: Feature-level data fusion (adapted from [8]).

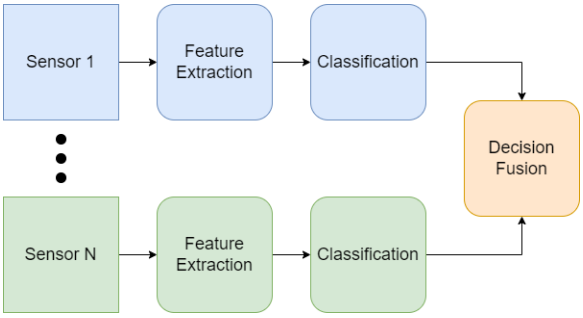


Figure 20: Decision-level data fusion (adapted from [8]).

2.7
Cloud Coverage

As shown by section 2.1.1, the surface information is occluded by clouds in images from optical sensors. Asner *et al.* [9] evaluated the monthly probability of obtaining a Landsat scene with 30% or less cloud presence in the BAF, analyzing all acquired scenes between 1984 and 1997. Figure 21 synthetizes the monthly probability month by month. From these results, one note that is challenging to have cloud-free (or with low cloud presence) from some BAF regions, especially the northern region.

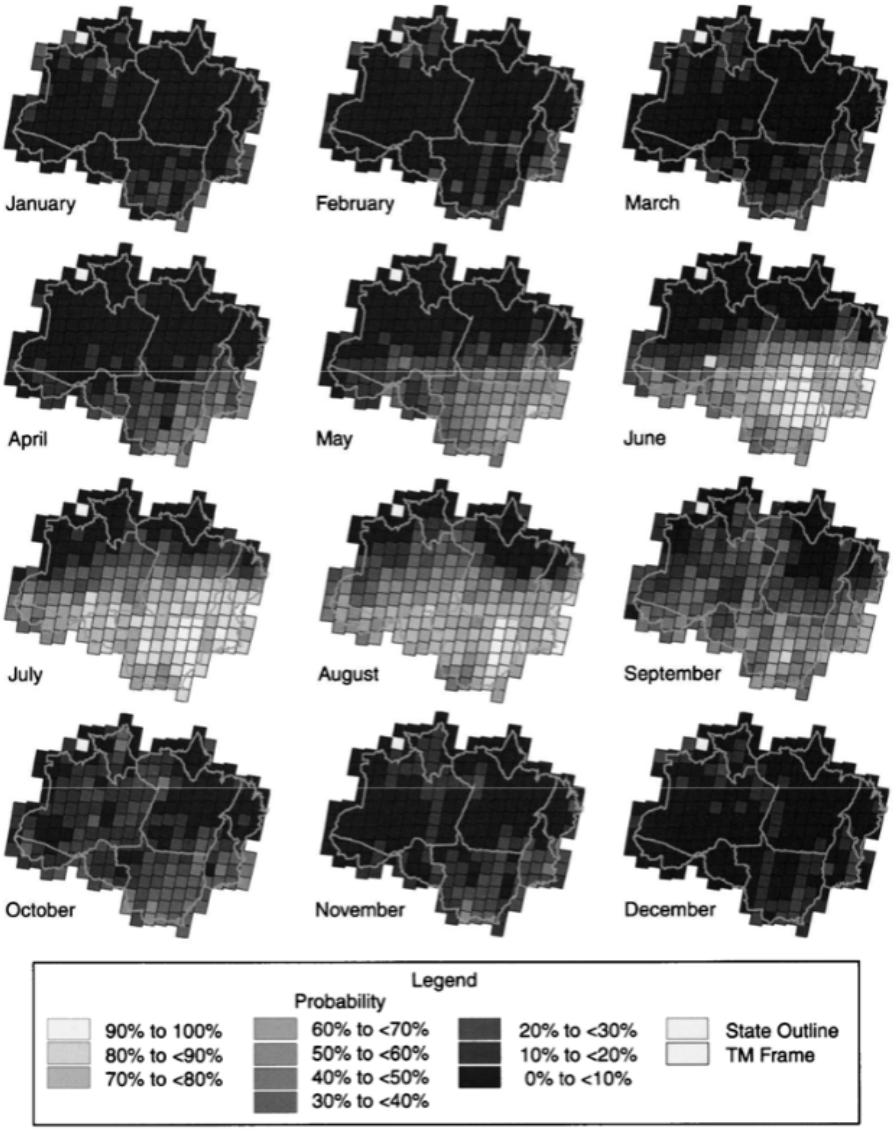


Figure 21: Monthly probability to obtain a Landsat scene with 30% or less of cloud presence in the BAF [9].

3
Related Works

This chapter presents the main published works closely related to this Thesis, especially those that aim to detect deforestation areas (as a typical change detection task) from multitemporal RS images using deep learning models. To demonstrate the gap in the knowledge covered by the contributions of this work, we highlighted those focused on data fusion and missing data by cloud coverage in Section 3.5.

3.1
Change Detection

A simple way to apply CNN architectures to detect changes from RS images is with single-stream temporal aggregation strategy, combining the temporal images before input in a traditional CNN model architecture, like [10, 67–75]. Figure 22 presents an example of a single-stream temporal aggregation strategy evaluated by [10].

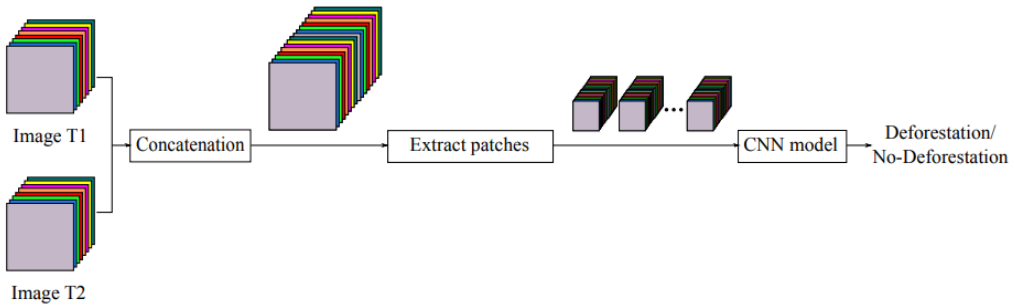


Figure 22: Single-stream model strategy evaluated by [10].

However, the most usual way to identify change detections from bi-temporal RS images is from dual-stream architectures, particularly with Siamese networks. In these architectures, the weights of the feature extractors are shared between streams [10, 76–94]. Some of these studies also included the contrastive loss function [95] to ensure that stream outputs belong to the same latent space [96–98]. These networks can perform multiple tasks simultaneously, including change detection identification [99]. Figure 23 presents

an example of a multi-stream temporal aggregation strategy, which was also evaluated by [10].

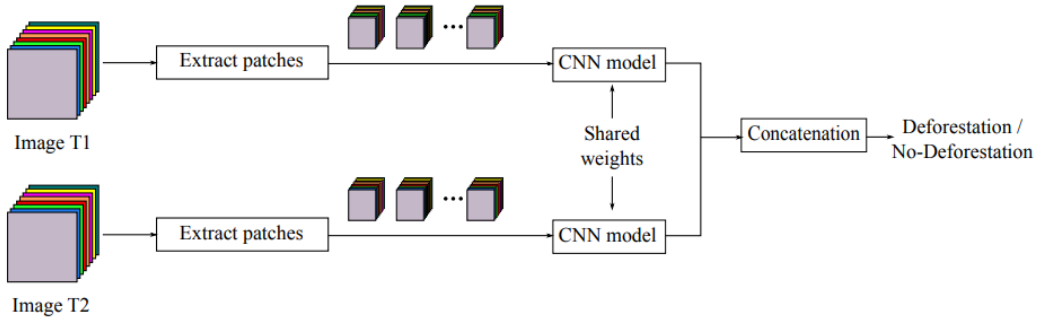


Figure 23: Multi-stream model strategy evaluated by [10].

Recurrent Neural Networks (RNN) have been utilized to handle sequential temporal data to identify its effects on the Earth's surface [100–102]. These architectures are usually based on Long Short-Term Memory (LSTM), which is a particular RRN unit that stores and forgets the temporal information [103–108].

Lately, attention-based methods have increased in popularity, including for Change Detection purposes. These methods usually apply transformer concept [3] to identify the differences between the images from different instants [109, 110]. However, other attention methods are also utilized, especially channel and spacial excitation [111], which increase the network attention in some channels or locations in the image, respectively [98, 112–114].

3.2 Deforestation Detection

Xiang *et al.* [11] investigated convolutional-based models in combination with various loss functions, using optical images (Sentinel-2) between 2017 and 2021 to identify annual forest changes in Hunan Province (center of China). In this work, they identified any changes in the forest, including deforestation and forest growth. Figure 24 presents the results from each investigated model by [11].

Other works, like [12, 52, 53, 55–58, 115–118], conducted similar research using optical images, investigating different models or optical images from other sensors. Some of these works used the single-stream temporal aggregation of the images, and others multi-stream. However, none of them explored optical images with cloud presence. Figure 25 presents examples of deforestation detection from [12].

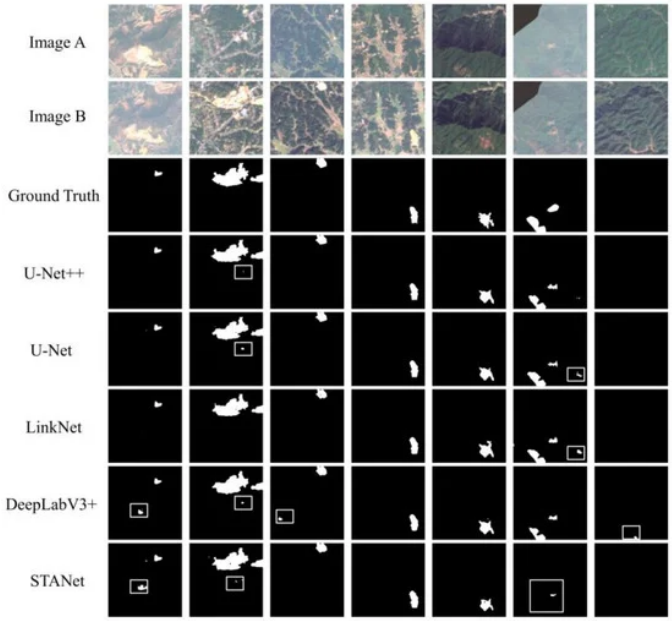


Figure 24: Comparison between results from the investigated models from [11].

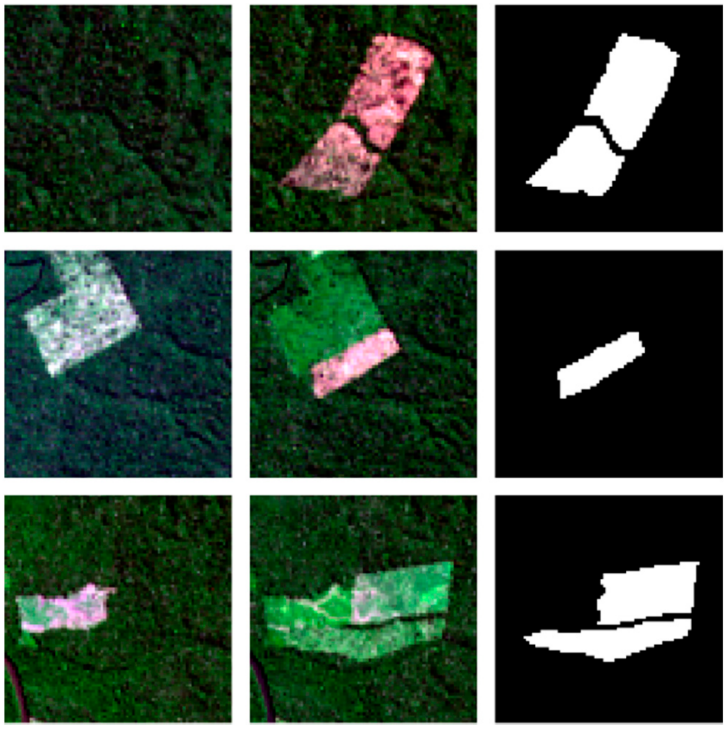


Figure 25: Examples of deforestation detection predictions from [12].

3.3 Cloud Presence

As is widely known, the presence of clouds in optical images affects any investigation using these images. We could not find any work investigating cloud presence interference in deforestation detection tasks using deep learning models.

However, as cloud coverage is an obstacle to the wide use of optical images, some alternatives emerged, like SAR data, especially in regions with high-frequency cloud presence [119]. Some works, like [54, 120], investigated deforestation detection by replacing optical images with SAR images, exploring single and multi-stream temporal aggregation strategies. Other works, like [13, 121], compared the convolutional-based networks using cloud-free optical and SAR images to monitor forests. Although these works investigated the SAR images as a replacement for optical images, they didn't compare their results with those in the presence of clouds, as shown by figure 26.

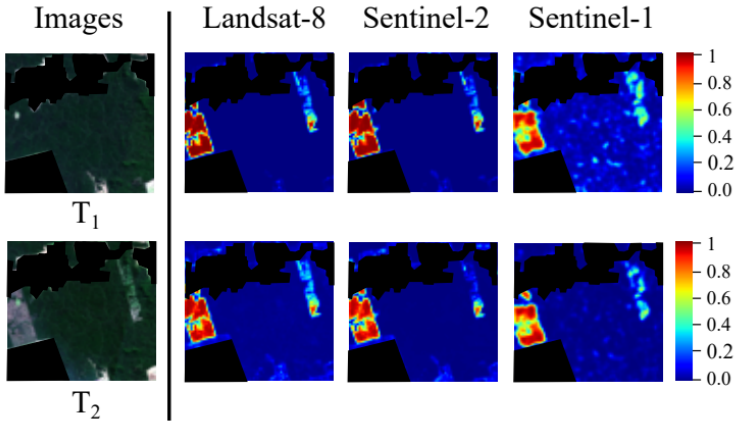


Figure 26: Sample of deforestation detection from [13].

3.4 Data Fusion

Deep Learning models can apply pixel-level fusion to generate new simulated images. For example, the fusion of panchromatic with multispectral or hyperspectral can generate a new image with a higher spatial resolution (from panchromatic) holding the spectral resolution (from multispectral or hyperspectral), also called pansharpening. A simple DL strategy for pansharpening is the End-to-end model, in which the source images are the input of a DL model, and its output is the simulated image [122–129]. The adversarial training strategy, like GAN [130], is also a usual way to pansharpening [131–135]. The previous strategies are based on supervised methods, but sometimes, the desired high-resolution data volume needs to be more available. In these cases, unsupervised methods were proposed for pansharpening tasks [136, 137].

Beyond the typical pansharpening task fusing data from two or multiple sources, other traditional machine learning tasks, like classification or regression, can also be performed by fusing data in pixel, feature [138–145], or decision levels, usually to exploit the complementary data from diverse sources.

Another usual strategy is translating the data from one source modal domain to another. These strategies usually apply GAN generating a synthetic image (target sensor domain) from a real one (source sensor domain) [146–151]. This strategy allows the creation of synthetic optical images to remove the clouds in the image and replace them with information based on other modality data, like SAR.

Other works investigated the SAR-optical translation replacing the cloud-covered pixels with synthetic data based on the SAR information in the optical data, generating a synthetic cloud-free optical image [150, 152, 153]. These synthetic images could be used to classify optical models.

Hong *et al.* [14] proposed a fusion strategy called cross-fusion, which focuses on learning representations across modalities from different subnetworks. Unlike other methods, in this approach, each modality's stream can learn specific properties from itself and also incorporate diverse information from another stream to achieve comprehensive information blending, as shown by Figure 27.

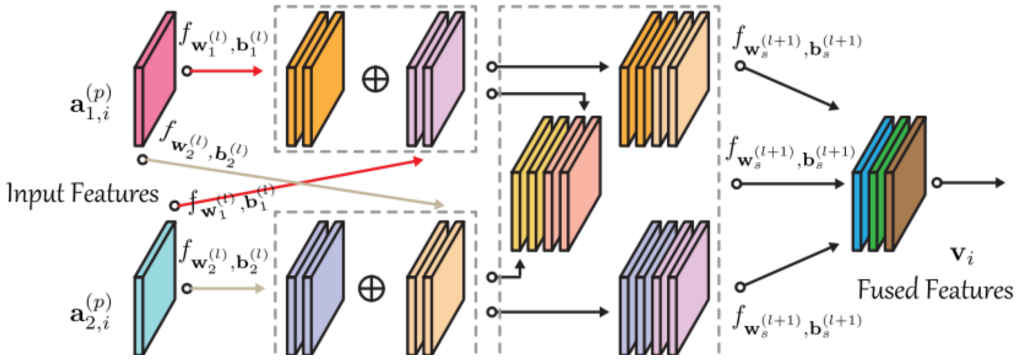


Figure 27: Cross-fusion layout [14].

Cue La Rosa *et al.* [15] investigated a multitask fusion model, using optical and SAR data to detect new deforestation areas in the BAF region. Figure 28 presents the proposed multitask model architecture. They trained the model utilizing only cloud-free optical images, concluding that the SAR classification can be used independently of the fusion and optical.

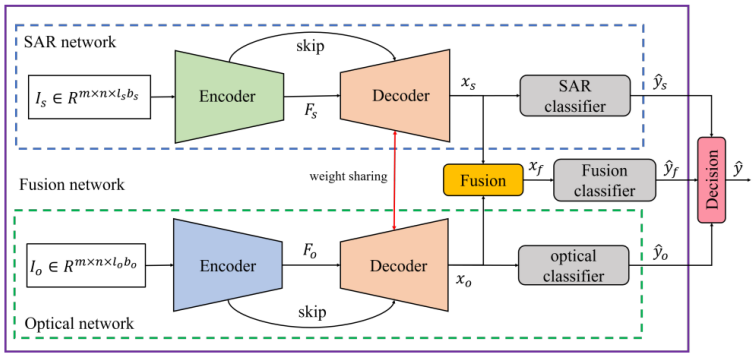


Figure 28: Multitask proposed model by [15]

3.5 Gap of Knowledge

The earlier section explored research on Change Detection, Deforestation Detection, Cloud Presence, and Data Fusion in combination with Remote Sensing data and Deep Learning context. To demonstrate the innovation of this thesis, we present a compilation of works focused on deforestation detection from multitemporal data, classifying if they addressed data fusion, cloud presence, or both. Table 4 summarizes the main works and the respective investigated issues.

Main Works	Investigated Topics	
	Data Fusion	Cloud Presence
[12] [52] [53] [55] [56] [57] [58] [10] [115] [116] [117] [118]	✗	✗
[15]	✓	✗
[13] [54] [120] [121]	✗	✓
[16] [17] [154] [18]	✓	✓

Table 4: Related works investigation topics.

Despite the previous works, which mainly investigated each topic individually, some investigated data fusion and cloud presence, like this current work.

Even with the visual quality of cloud-free images, usually generated from optical and SAR data fusion, these synthetic images are limited to the available information under cloud-covered optical regions, in this case, the information provided by the SAR sensor. The use of these synthetic images for classification purposes typically did not outperform the use of only SAR data, as in [16], who investigated deforestation detection at two sites in the BAF using synthetic

images from Sentinel-1 and Sentinel-2 and the same DL model from [13]. Figure 29 presents the results of the F1-Score from Sentinel-2 cloud-free (S2 Clear), Sentinel-1 (S1), and synthesized images, showing that the synthesized images from any method outperformed the single-modality models.

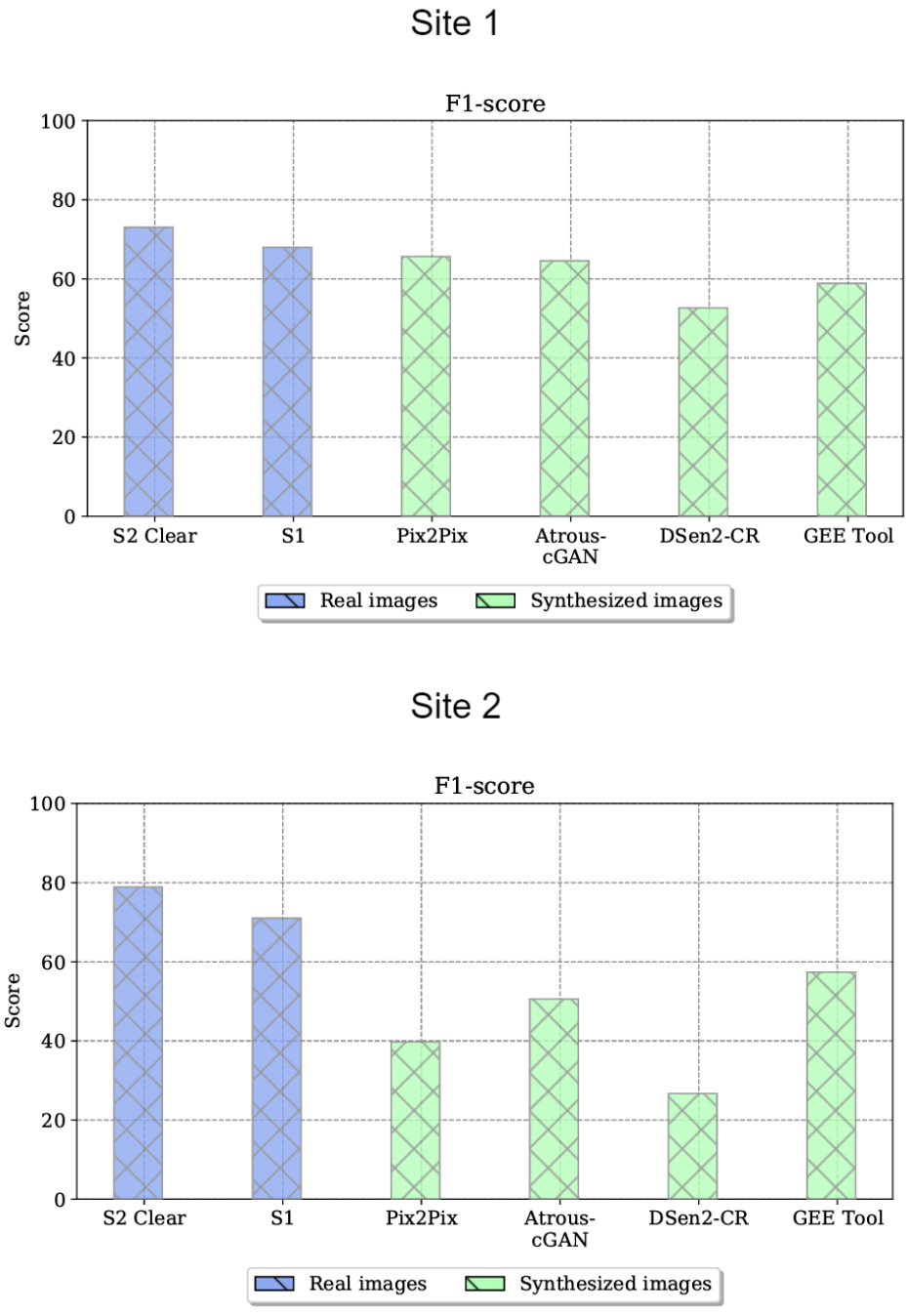


Figure 29: Comparison between deforestation detection using real cloud-free optical, SAR, and synthetic optical cloud-free images from different sites [16].

Li *et al.* [17] proposed a spatiotemporal fusion method to calculate the Normalized Difference Vegetation Index (NDVI), which can be used to estimate the vegetation density in the area, usually a proxy to monitor the deforestation

process. They used a combination of Sentinel-1 and Sentinel-2 (cloud-free or partially covered by clouds), as shown in Figure 30. Unlike our work, the optical images' parts partially covered by clouds were discarded and replaced by the information from SAR data.

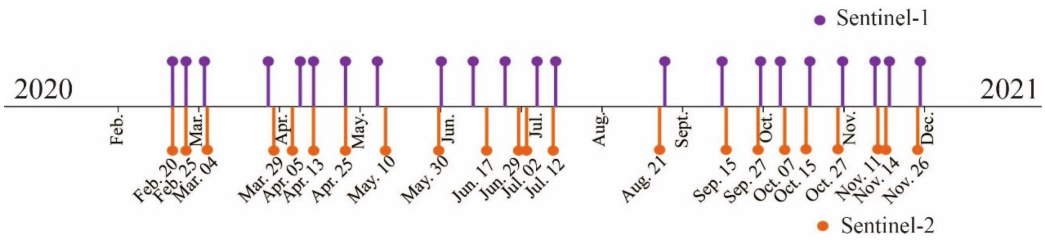


Figure 30: Timeline of the images from Sentile-1 and Sentinel-2 used by [17].

In previous works, we have investigated the fusion of SAR and optical data with diverse cloud conditions. In [154], we created multiple tiles across the BAF. We used optical with diverse cloud conditions and SAR images from two consecutive years, using some tiles for training, validation, and test purposes. In that work, the best model delivered predictions with an F1-Score value of 0.72. We also investigated this in [18], replacing the location-based training-validation-test split used in [154] with a time-based split, using images from two consecutive years (2018 and 2019) for training-validation purposes and images from another consecutive year (2019 and 2020) for testing, in which the best model delivered predictions with F1-Score values of 0.92 for all pixels, as shown by Figure 31.

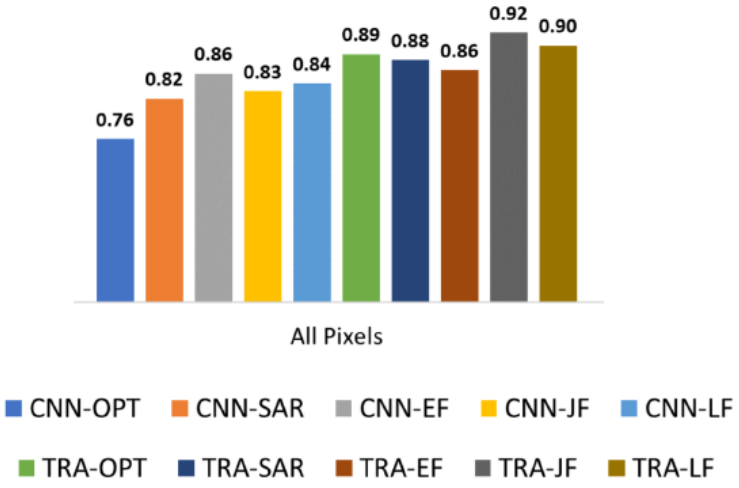


Figure 31: F1-Score from convolution (CNN-*) and transformer-based (TRA-*) models, using optical (*-OPT) and SAR (*-SAR) data and the early (*-EF), joint (*-JF) and late (*-LF) fusion of optical and SAR data evaluated by [18]

From the presented works, confirmed by Table 4, we can identify a lack of investigation into combining the data fusion of RS data with cloud presence

to deforestation detection using Deep Learning models. Despite some works investigating the fusion of SAR and optical images, they didn't explore the optical images with cloud presence.

4
Methodology

This chapter introduces the methodology utilized in this Thesis. Initially, the baseline models will be shown. Then, we present the fusion models, followed by a strategy to train them to improve their robustness.

As described in Chapter 1, we hypothesize that Deep Learning models can learn how to identify cloud-covered regions and select the trustful data source(s). To evaluate this hypothesis, we will investigate models capable of fusing data from optical and SAR satellite systems to identify new areas of deforestation from bitemporal images with an interval of one year between them.

Initially, we selected some architectures based on CNN (ResUnet [2]) and Transformer (Swin-Unet [6]), which are aimed at semantic segmentation. These models will serve as the baseline and identify the influence of the temporal aggregation strategy and the previously known information on deforestation.

4.1
Base Architectures

4.1.1
ResUnet Based Architecture

The ResUnet-based model is organized into Encoder, Decoder, and Classifier, as shown in Figure 32. The input $X \in R^{H \times W \times B}$, in which H, W, and B are the height, the width, and the number of bands in the Input Image Patch. The Encoder consists of a sequence of Residual Blocks followed by 2×2 max pooling operations (orange arrows). The first Residual Block's convolutions have D filters, and the following doubles after each max pooling operation. The Decoder comprises a nearest neighbor upsampling operation (green arrows), followed by a Residual Block, whose outcome is concatenated with the feature map produced by the corresponding Encoder until the height and width match the original image size. Finally, the Classifier consists of a Residual Block followed by a 1×1 convolution layer, with a softmax activation, producing the Patch Prediction with N classes $\in R^{H \times W \times N}$.

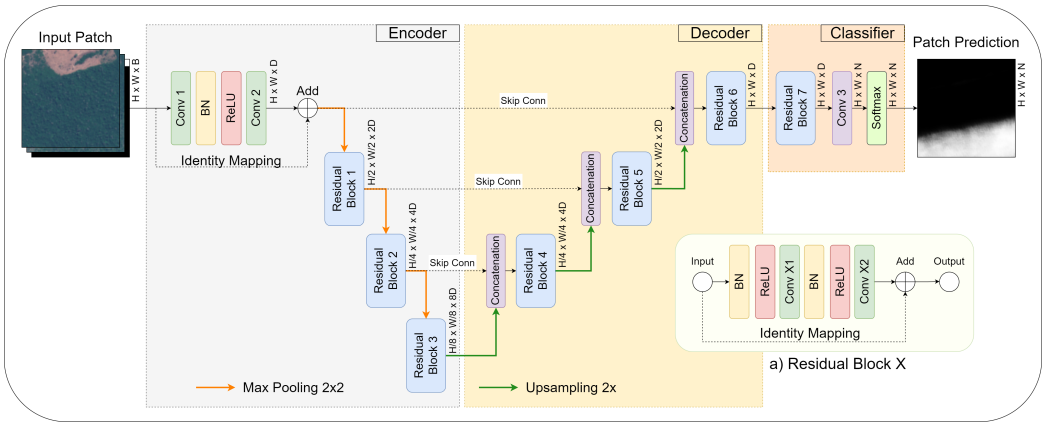


Figure 32: ResUnet-based architecture.

The Residual Block is a sequence of two 3×3 convolution layers with a dropout between them and a residual connection, in which a 3×3 convolution layer is applied to adjust the depth dimensionality. Figure 32a shows the Residual Block. After 3×3 convolution layers, a ReLU activation is applied.

4.1.2 Swin Based Architecture

Figure 33 presents the Swin-based architecture, organized into Encoder, Decoder, and Classifier. Initially, the Input Image Patch $X \in R^{H \times W \times B}$ is partitioned into smaller patches called image tokens $\in R^{4 \times 4 \times B}$, where H, W, and B are the height, the width, and the number of bands in the Input Image Patch, respectively. A Linear Embedding layer projects each feature map onto an arbitrary C-depth.

The Swin-based architecture Encoder comprises the Patch Partition and a sequence of Swin Transformer Blocks followed by a Patch Merging operation. Each Swin Transformer Block consists of a Windowed Multihead Self-Attention (W-MSA) or the Shifted W-MSA (SWMSA) operation. The W-MSA and SW-MSA apply the MSA between image tokens restricted to the same window. To ensure the connection between the windows, each W-MSA is followed by a SW-MSA, which shifts the window by a fixed number of tokens. Following the W-MSA or SW-MSA, there are two layers of Multilayer Perceptron (MLP), with Gaussian Error Linear Unit (GELU) activation modules, as presented in Section 2.4. Each module is preceded by a Layer Normalization (LN) operation and a residual connection, as shown in Figure 33(a).

Patch Merging is applied to reduce the height and width of the image tokens while expanding the depth, similar to ResUnet. Each Patch Merging operation halves the height and width while doubling the depth dimension by a linear projection of the concatenated rearranged image tokens, as presented

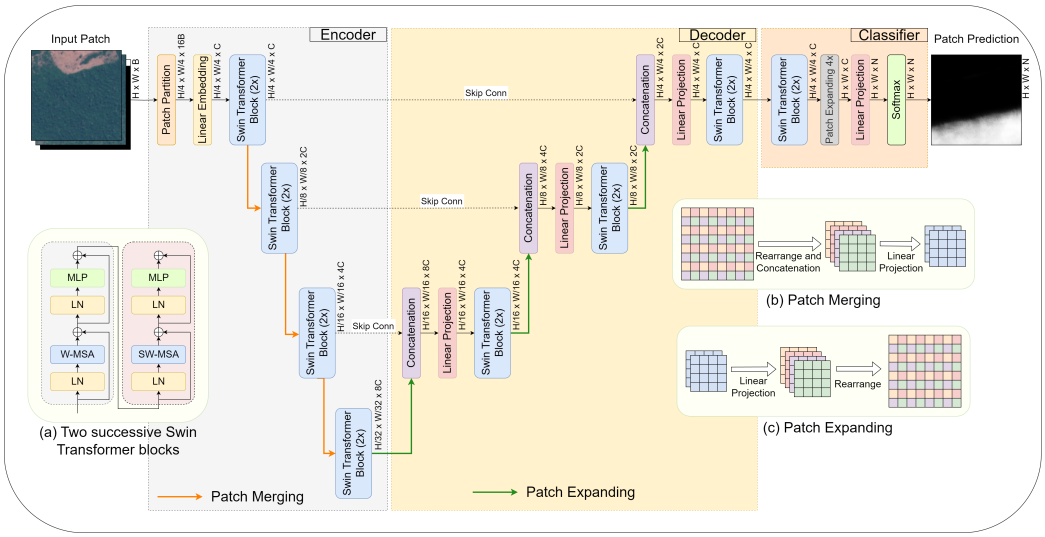


Figure 33: Swin-based architecture.

in Figure 33(b). In the rearrangement, the image tokens' colors means in which position the token will be concatenated.

The Decoder of Swin-based architecture consists of a Patch Expanding operation, a concatenation with the respective Encoder output, a linear projection to reduce the depth, and Swin Transformer Blocks applied in sequence until the tokens' height and width turn back to their original size. Patch Expanding performs the opposite operations of Patch Merging, increasing image tokens' depth with a linear projection followed by an image tokens rearrangement, increasing the height and width while reducing the depth, as presented in Figure 33(c). All Patch Expanding doubles the size while halves the depth dimension, except the last one, which quadruples the size without changing the depth dimensionality.

The Classifier of Swin-based architecture consists of a Swin Transformer Blocks followed by the last Patch Expanding, to recover the original image height and width from the image tokens' size, followed by a linear projection and softmax activation to generate the Patch Prediction with N classes, as shown in Figure 33.

4.2 Single-modality Models

Single-modality models use only one type of data (either optical or SAR) for model training and prediction. Figures 34 and 35 present how the Encoder, Decoder, and Classifier blocks are organized for single-modality optical and SAR models' architectures, respectively.

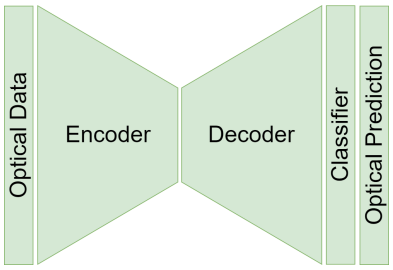


Figure 34: Optical models architecture.

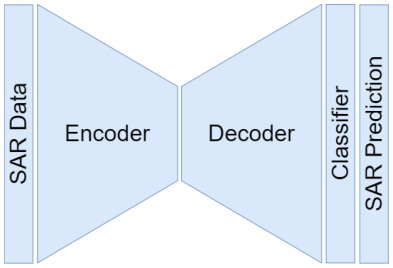


Figure 35: SAR models architecture.

4.2.1 Temporal Aggregation

As presented in Section 2.5, in change detection problems, we can perform the temporal aggregation in single-stream or multi-stream feature extraction strategies.

Figure 36 shows the single-stream temporal aggregation strategy. For optical models (Figure 36a), we just concatenated the images taken in dates (T_0 and T_1) from two consecutive years. For SAR models, we concatenated the images from dates (T_0 and T_1) from two successive years (as in the optical models) or all available data between these two dates(T_0, ..., T_N), capturing the features across the analyzed period and the behavior of this data in other seasons, depending on the SAR dataset. SAR datasets will be discussed in Section 5.3.1.2. We concatenated the images with the available auxiliary data for both models. The auxiliary data used in this work are detailed in Chapter 5.

The Siamese architecture models utilize the multi-stream temporal aggregation strategy. In this strategy, only two images were taken on two dates (T_0 and T_1) from consecutive years. The Encoder blocks related to each image input share their weights. If the model uses any auxiliary data, it is input in a sequential Maximum Pooling block, generating multiple outputs with the same size (height and width) as the related skip connections outputs from the Encoder blocks. The single Decoder block has as input the concatenated feature maps from the encoders and the Auxiliary Data Maximum Poolings outputs.

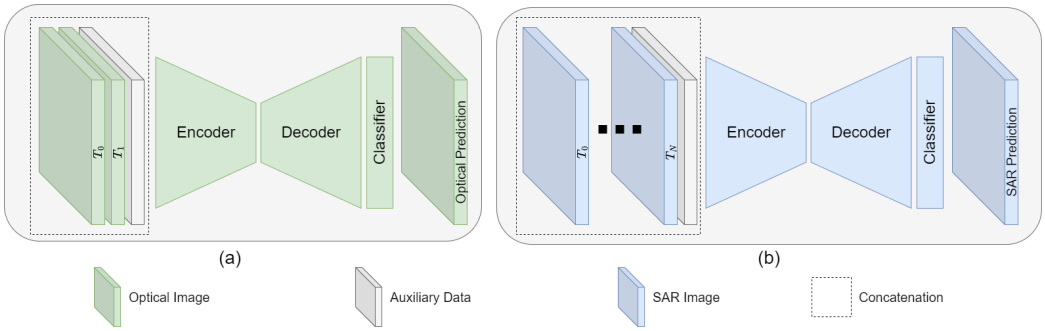


Figure 36: Single-stream temporal aggregation for optical (a) and SAR (b) models

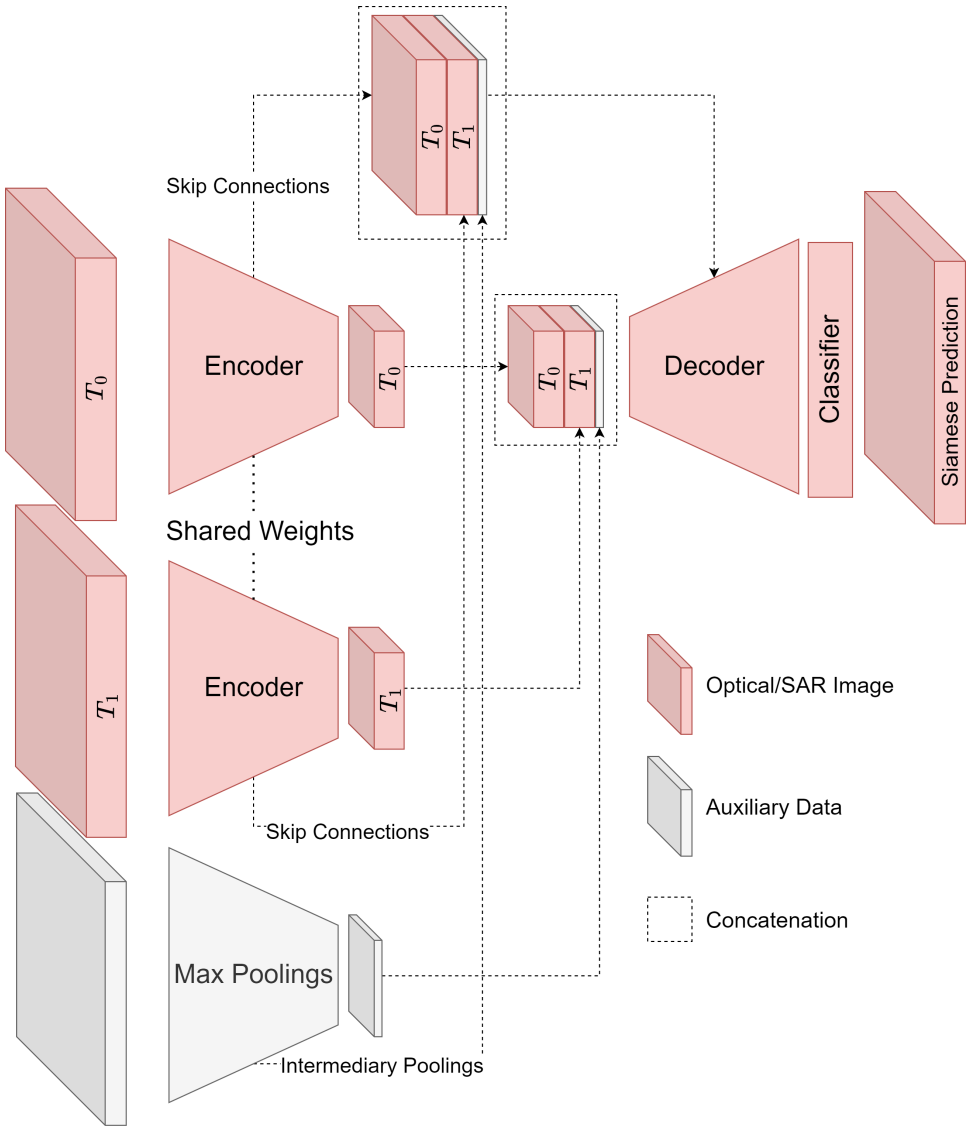


Figure 37: Multi-stream temporal aggregation

4.3 Fusion Models

We investigated three main architectures for optical and SAR data fusion: Pixel Level, Feature Level (Middle), and Feature Level (Late).

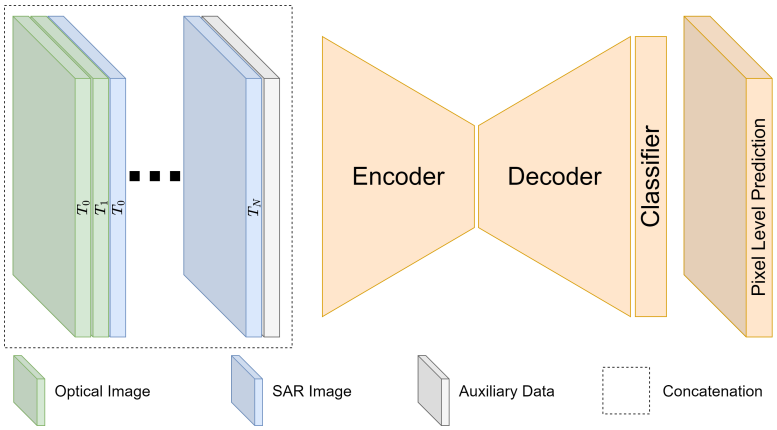


Figure 38: Pixel Level models architecture.

In the Pixel Level strategy (Figure 38), the optical, SAR, and Auxiliary Data are concatenated in the channel dimension before input into the model. The Encoder, Decoder, and Classifier setup is the same as in the single-modality models.

In the Feature Level (Middle) fusion strategy (Figure 39), the model architecture has two independent Encoder blocks. The Encoder blocks outputs (including the skip connections) are fused before input in the Decoder. This strategy uses only the concatenation-fusion strategy to fuse optical and SAR data.

In the Feature Level (Late) fusion strategy (Figure 40), the model architecture has independent Encoder and Decoder blocks. The Decoder outputs are fused before entering the Classifier. This strategy uses concatenation-fusion and cross-fusion to fuse optical and SAR data.

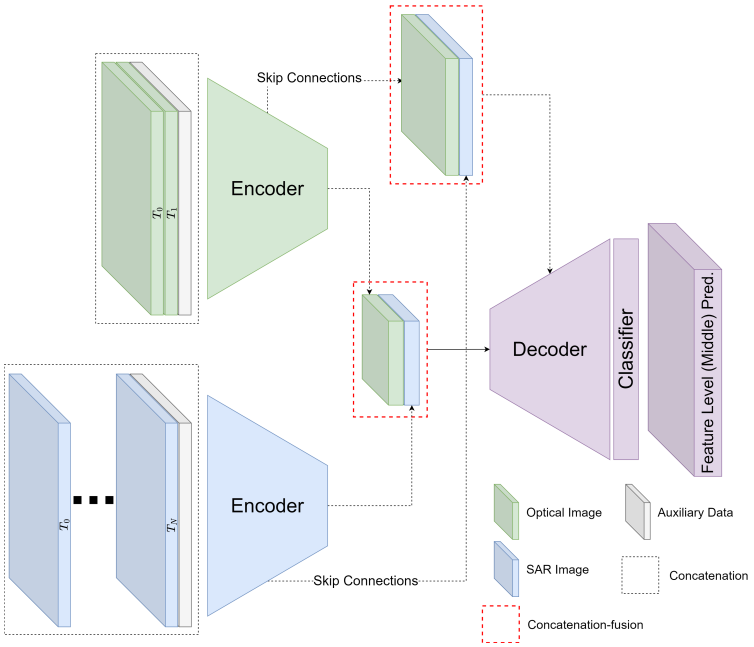


Figure 39: Feature Level (Middle) models architecture.

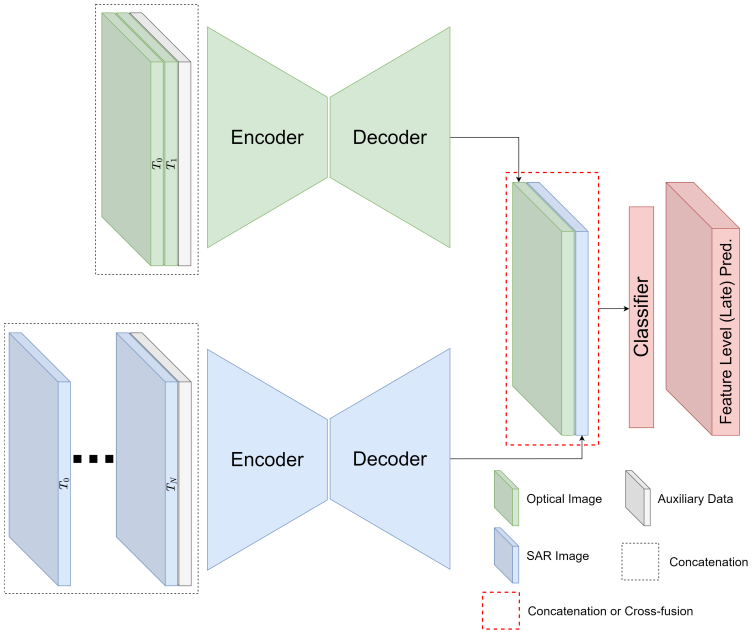


Figure 40: Feature Level (Late) models architecture.

4.3.1 Optical-SAR Fusion Methods

We investigated two ways to fuse optical-SAR data: Concatenation and Cross-fusion. The simplest way to combine data from optical and SAR is to concatenate them in the channel dimension. The Concatenation-fusion consists of the concatenation of the feature maps from each modal, optical, and SAR, followed by a 1×1 convolution with N kernels, in which N means the depth of each feature map, as shown by Figure 41.

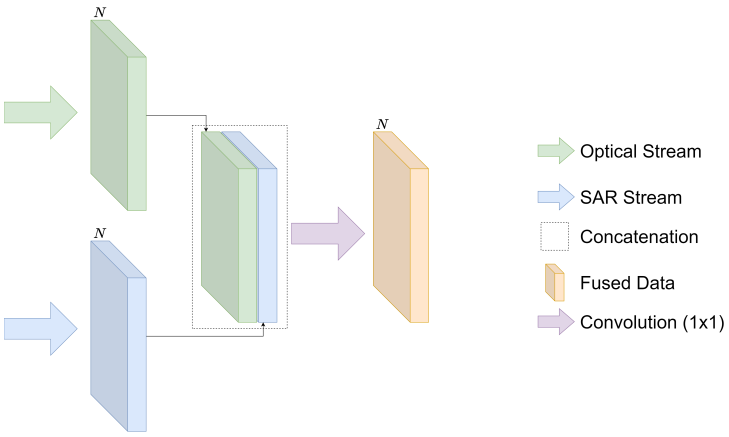


Figure 41: Concatenation-fusion.

The other fusion strategy was the Cross-fusion, based on [14]. As applied in the original work, we only used this strategy in the Feature Level (Late) models. Figure 42 shows the Cross-fusion, in which the 1×1 convolutions are presented by the solid arrows, in which the same color means they share their weights. Dashed arrows mean the feature maps flow. Red and black dotted rectangles mean the sum and concatenation operations, respectively.

4.3.2 Pre-training Strategy

Typically, DL model weights are initialized randomly. However, models trained using optical and SAR data show distinct convergence behaviors, which can impact the performance of fusion models trained from randomly initialized weights. To address this issue, we proposed a pre-training strategy to minimize the convergence differences between these data types. By aligning their convergence more closely, our approach reduces the convergence discrepancies between these data types in fusion models, enhancing the overall effectiveness of these models, especially using optical images with the presence of clouds.

In this pre-training strategy, the randomly initialized weights are replaced by the weights of the same blocks from the single-modality trained models. As

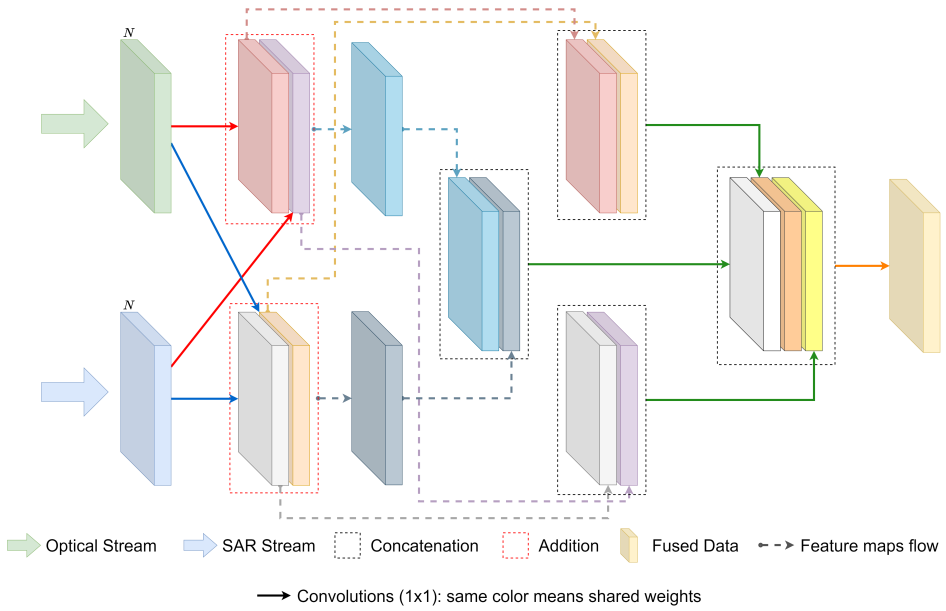


Figure 42: Cross-fusion (adapted from [14]).

the Pixel Level models don't have any stream belonging to a single model, we can't apply this strategy in these models. Figures 43 and 44 present the proposed pre-training strategy for Feature Level (middle) and Feature Level (late) models, respectively, in which the dashed arrows mean the source (single-modality) and destination (fusion models stream related to respective modality data) weights.

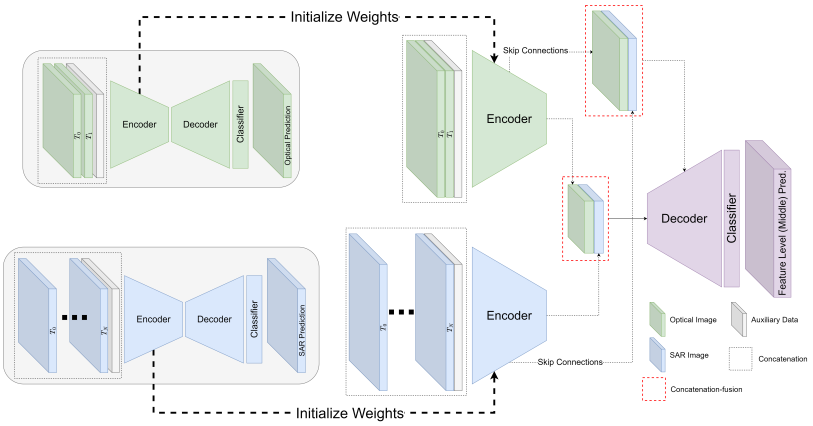


Figure 43: Pre-training strategy for Feature Level (middle) models

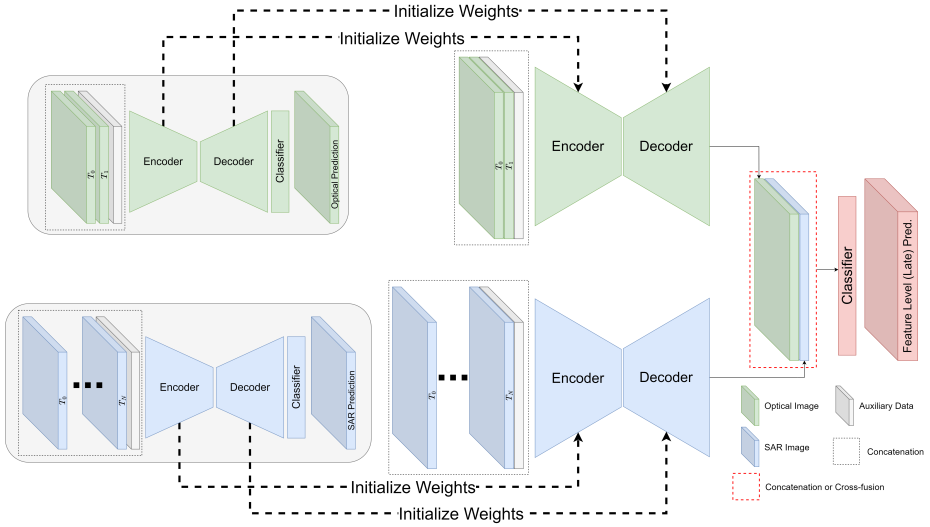


Figure 44: Pre-training strategy for Feature Level (late) models

5
Experimental Protocol

This chapter serves as a roadmap for the experiments executed in this study. It outlines a series of steps, beginning with defining the study areas (Section 5.1) and corresponding reference dates (Section 5.2). We meticulously detail the comprehensive data collection process, including Remote Sensing and PRODES (Section 5.3). Sections 5.4 and 5.5 describe how we created the Reference Data and the Previous Deforestation Map, respectively. Section 5.6 presents the evaluated models. Section 5.7 explains how we prepared this data to train the models. Section 5.8 details how the prepared data is used to train these models effectively. The trained models are then employed to predict new deforestation areas, as described in Section 5.9. Finally, Section 5.10 explains how these predictions are used to evaluate the performance of each proposed model.

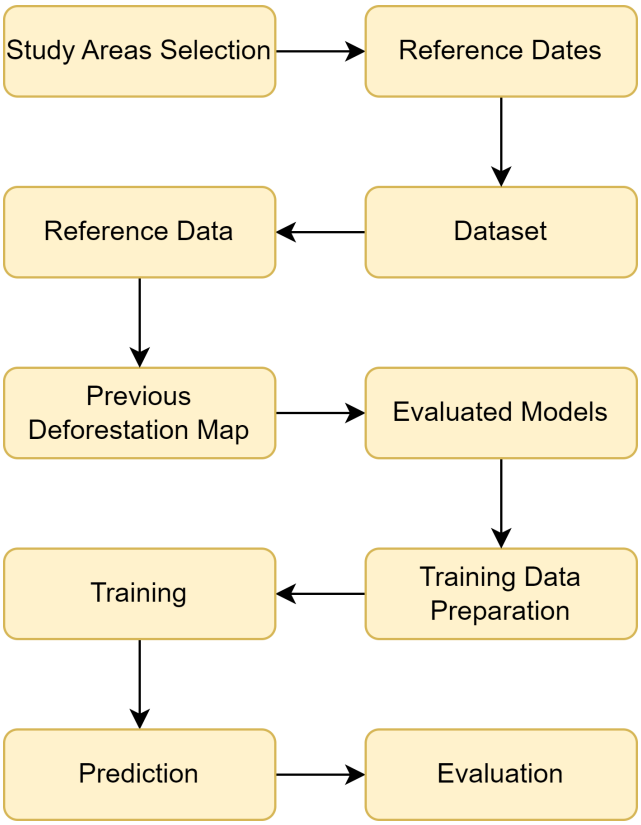


Figure 45: Experimental protocol steps flow.

5.1 Study Areas Selection

We chose our study areas based on two main factors: the presence of deforested regions suitable for training and testing and the minimal cloud coverage in the optical satellite images used by PRODES between 2019 and 2021. The training step utilized images from 2019 to 2020, while testing relied on images acquired between 2020 and 2021. Figure 46 presents the selected sites' locations, called Site 1 and 2.

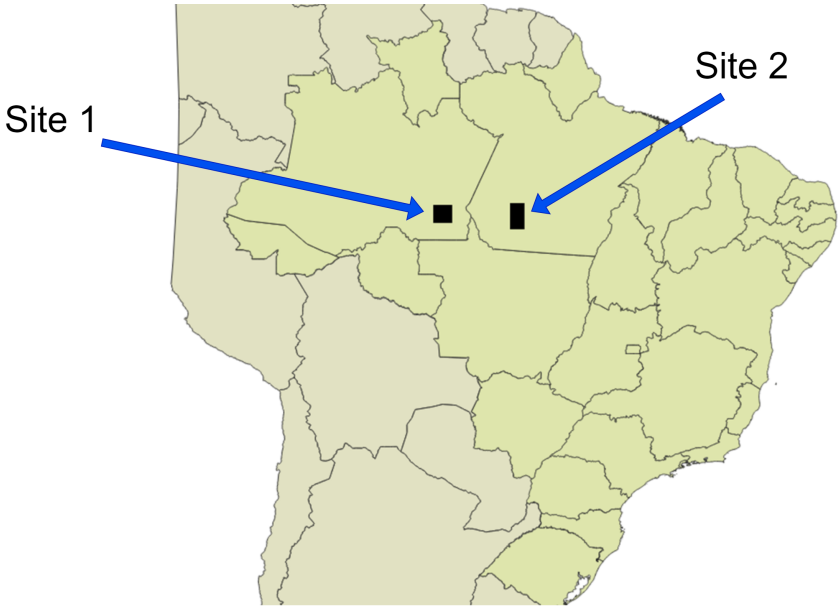
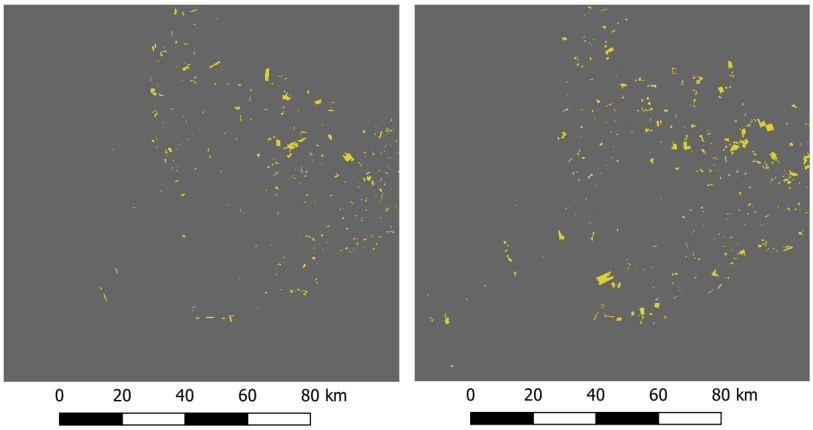


Figure 46: Study Sites.

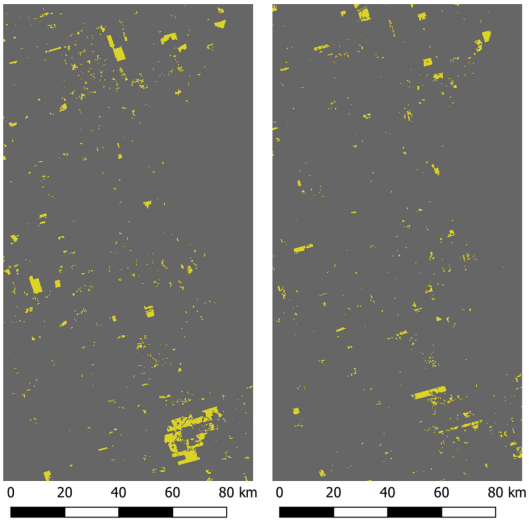
A high number of deforestation areas detected by PRODES is crucial for mitigating the imbalance issue. Even in regions where deforestation happens rapidly, the proportion of total deforested area within a year compared to the total area is low. Choosing areas with a larger deforested area aids in model training. Furthermore, to ensure an adequate number of samples for evaluation, it's also important that study areas exhibit a high quantity of deforested regions during the testing period. Figures 47 and 48 present the spatial distribution of deforestation occurrence in Site 1 and 2, respectively, and Table 5 show the proportion of deforestation occurrence in each site in each period, in comparison to total site's area.

Although this research focuses on deforestation detection independently of cloud-free optical images, the images utilized by PRODES for deforestation detection in the chosen years must have minimal cloud cover. PRODES's usage of an (almost) cloud-free image ensures that a human operator has examined (almost) the entire extent of the area, as these data will be used to generate reference data employed in model training and evaluation.



47(a): 2019-2020. 47(b): 2020-2021.

Figure 47: Deforestation areas (in yellow) identified by PRODES in Site 1.



48(a): 2019-2020. 48(b): 2020-2021.

Figure 48: Deforestation areas (in yellow) identified by PRODES in Site 2.

Site	**Total Area** (Km^2)	**Deforestation Proportion (%) (2019-2020)**	**Deforestation Proportion (%) (2020-2021)**
Site 1	15 112	0.72	1.16
Site 2	16 481	1.85	1.62

Table 5: Deforestation areas identified by PRODES in both Sites.

5.2 Reference Dates

As described in Section 2.2, the PRODES project utilizes a range of optical imagery sources to identify newly deforested areas. Each deforestation polygon is associated with the date it was first detected, corresponding to when the image capturing the deforestation was acquired. Figure 49 presents examples of PRODES deforestation polygons identified in 2020, showing their respective image acquisition dates.

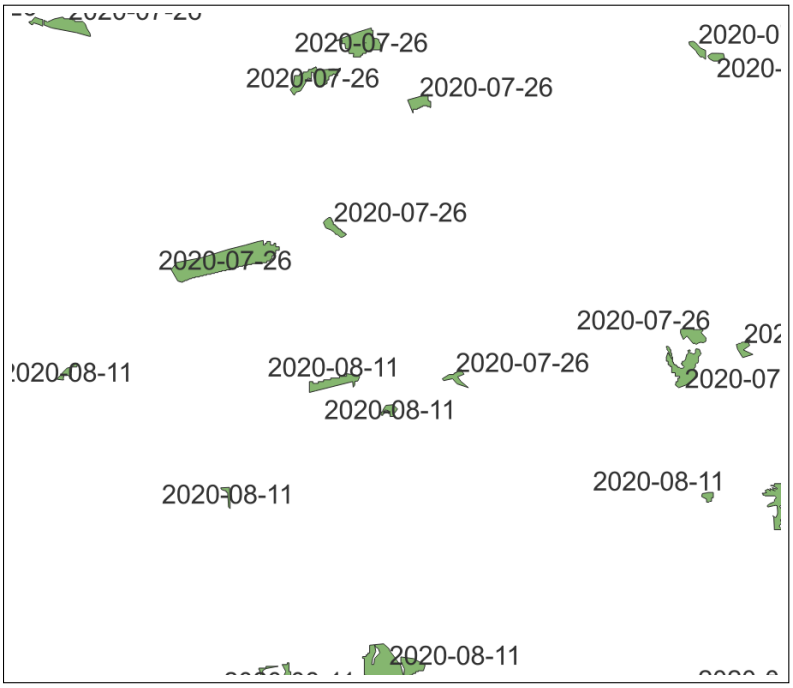


Figure 49: Reference dates examples.

We had to select a reference date for each site for each year, which would be used later to choose the images to be acquired. We attempted to select the most representative date possible, considering that the PRODES deforestation polygons from that year are present in each site. Table 6 presents the selected reference dates for each site in each year. The reference date may not coincide with the PRODES date.

Site	2019	2020	2021
Site 1	2019-08-14	2020-07-31	2021-07-02
Site 2	2019-07-24	2020-07-26	2021-07-29

Table 6: Reference dates for each site in each year.

5.3 Dataset

All data used in this work were collected from two primary sources. The imagery data, Sentinel-1 and Sentinel-2, were collected from Google Earth Engine Platform [155]. The deforestation data were acquired from TerraBrasilis [21].

5.3.1 Remote Sensing Images

5.3.1.1 Optical Images

The optical data used in this work consisted of Sentinel-2 images covering the study areas. The processing level product used was Level-2A to minimize the effect of atmospheric distortions, as described in Section 2.1.1. These images were downloaded using the Google Earth Engine platform, in which the product ID is "COPERNICUS/S2_SR_HARMONIZED", which guarantees the spectral response uniformity between the different processing procedures updates [155].

We excluded the optical image bands with 60 meters of spatial resolution and resampled the remaining bands to 10 meters using the nearest neighbor method when necessary. The optical bands used in this work and their respective spatial resolutions are presented in Table 7.

Band	Original Spatial Resolution	Final Resolution
B2	10	10
B3	10	10
B4	10	10
B5	20	10
B6	20	10
B7	20	10
B8	10	10
B8A	20	10
B11	20	10
B12	20	10

Table 7: Optical bands and spatial resolutions.

The image dates were selected from all available images within a window of ±30 days around each reference date presented in Table 6, respecting the

following other criteria:

- Three optical images were selected for each reference date;
- One image must be cloud-free;
- One image must have some thin clouds;
- One image should have the maximum cloud cover possible;
- All images must be as close to the reference date as possible.

The optical images were organized into two optical datasets called *CLOUD-FREE* and *CLOUD-DIVERSE*. The first one comprises only the optical images with cloud-free conditions, and the last one comprises all optical images, regardless of the cloud condition.

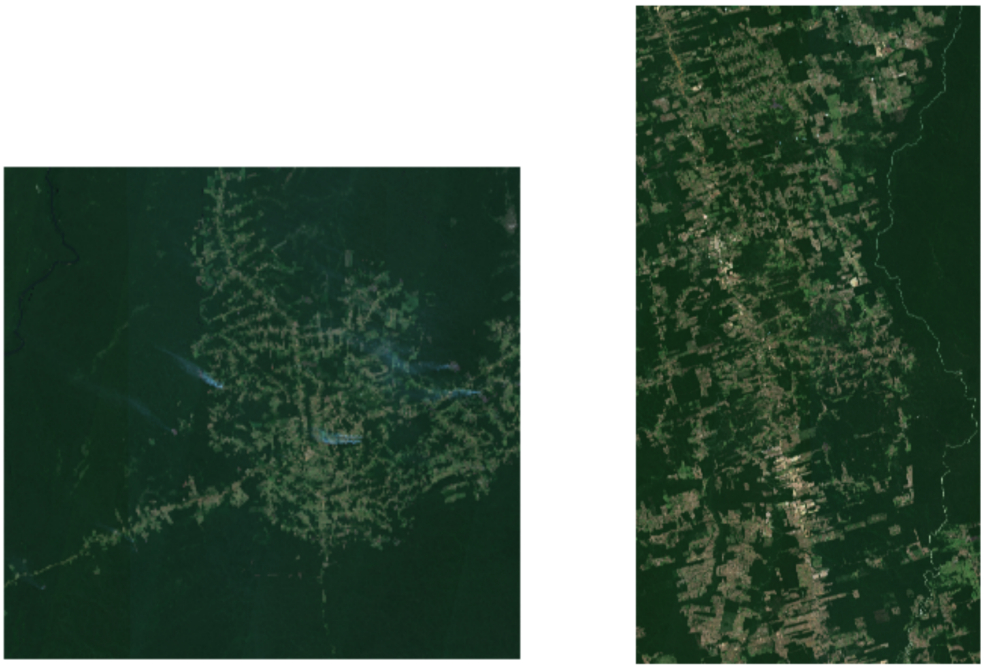


Figure 50: Samples of RGB composition from both sites.

Figure 50 shows examples of optical images from both sites. Appendix B.1 presents an overview of the optical data, including their acquisition dates.

5.3.1.2 SAR Images

The SAR images utilized in this study were sourced from Sentinel-1 images covering the study area. Specifically, the Ground Range Detected (GRD) product with VV and VH bands (available bands within the study area) was employed. Like the optical data, the SAR images were acquired via the Google Earth Engine platform using the product ID "COPERNICUS/S1_GRD_FLOAT". This product delivers GRD information in raw power values rather than in decibels [155].

The SAR data is already available at a 10-meter resolution, so there's no need for resampling. Furthermore, because the Sentinel-1 image is georeferenced using the same parameters as Sentinel-2, they are already coregistered, and there's no need for a new registration between them.

We selected two approaches for the SAR images. The first was to use a single SAR image, similar to the optical images. The second was to use all SAR images between two sequential reference dates.

Given the limitation of not being able to acquire Sentinel-1 images covering all study areas on a single day, we opted to mosaic Sentinel-1 images from multiple days. This process was carried out for each year and each site, according to the following steps, generating a dataset henceforth called *SINGLE-2*:

1. Mosaic all Sentinel-1 images belonging to the same satellite cycle;
2. Discard mosaics with corrupted data;
3. Select the mosaics acquired as close to the reference date.

Another approach aims to exploit all available SAR data between two consecutive reference dates, capturing the features across the analyzed period. To accomplish this, we organized all the days between one month before reference date 1 and reference date 2 into 12 sets. Subsequently, we identified all available Sentinel-1 images for each set and computed the average image from this selection. Figure 51 presents a schema of the construction of these average images. We generated two additional datasets from these average images, called *AVERAGE-12* and *AVERAGE-2*. The *AVERAGE-12* dataset comprises all 12 average images, while the *AVERAGE-2* dataset contains only the first and last average images. We generated average images for the reference dates from 2019 to 2020 and 2020 to 2021.

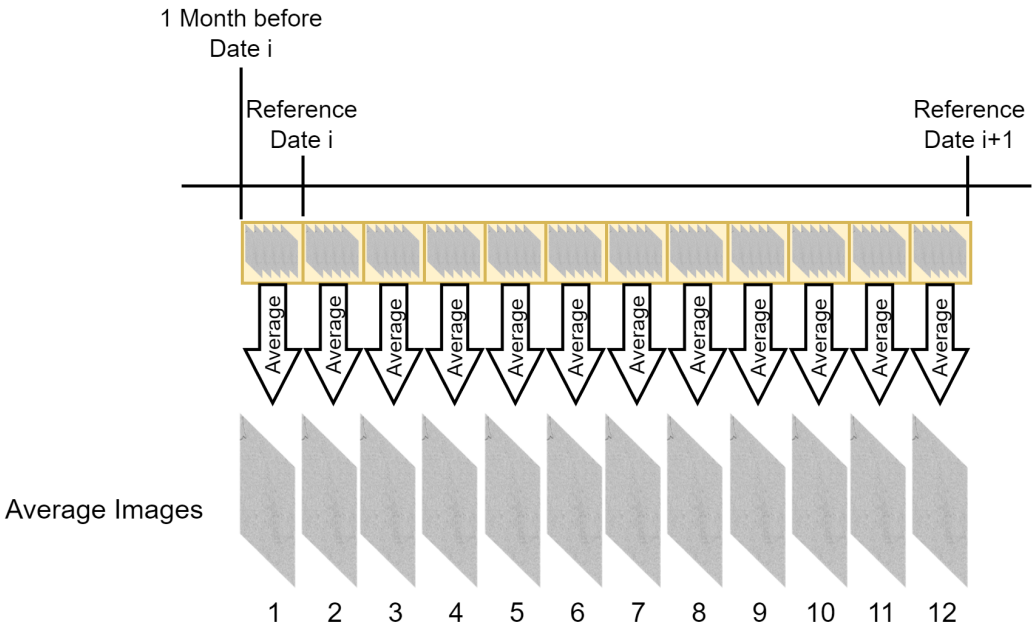


Figure 51: Average SAR.

An overview of the SAR images, including their acquisition dates, is presented in the Appendix B.2.

5.3.1.3 Cloud Map

The Level-2A of Sentinel-2 data provides cloud coverage maps and optical images. This map has a spatial resolution of 20 meters, where each pixel value indicates the percentage probability ($[0, 100]$) of cloud presence in the corresponding optical image. Figure 52 shows examples of cloud maps and the respective optical image dates.

An overview of the cloud maps, including their acquisition dates, is presented in Appendix B.3.

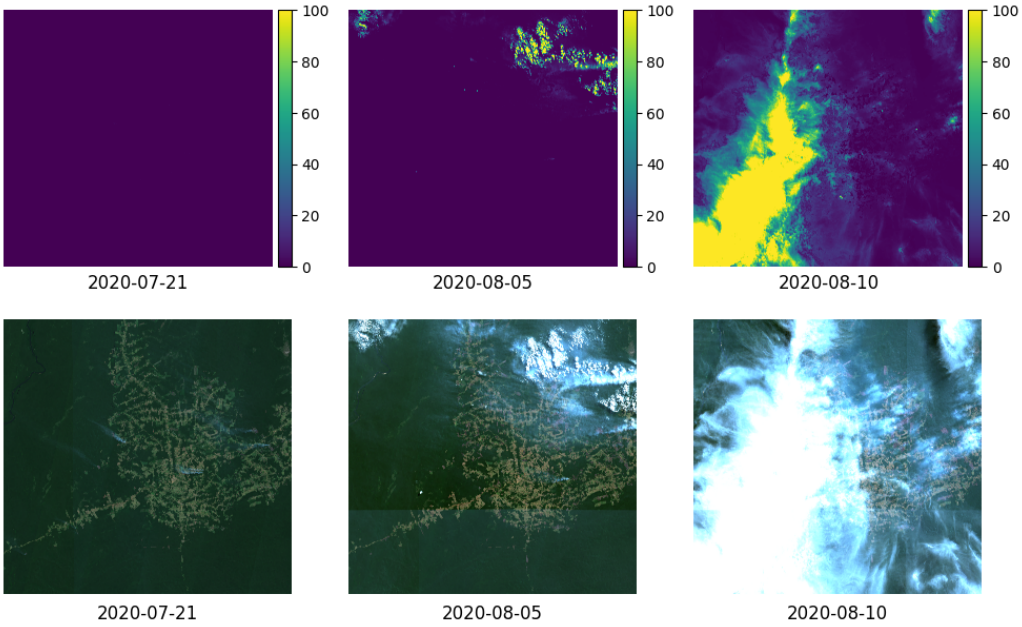


Figure 52: Cloud maps examples.

5.4 Reference Data

We require reference data to train deep learning models and evaluate their predictions. For semantic segmentation, the ideal reference data is generated by classifying individual pixels by a human specialist who views the same input images (optical and SAR). Since this is impractical, we used the PRODES project's deforestation polygons to generate the labeled pixels.

5.4.1 Deforestation Data

The deforestation data used in this study belongs to the Amazon Biome, in which Sites 1 and 2 are located. They were downloaded in a shapefile format containing polygons of each information class, as described in Table 8. In addition to deforestation data, PRODES also provides other information such as Hydrography, Non-Forest, and Cloud Mask, which were also used, as will be described in section 5.4.

PRODES Data	Data Description
Accumulated Vegetation Suppression Area Mask	Deforestation polygons identified until 2007
Hydrography	Represents the water bodies coverage.
Annual Deforestation Increment	Provides information on annual deforestation increment polygons.

PRODES Data	Data Description
Cloud Mask	Identifies cloud-covered areas in the PRODES imagery each year.
Non-Forest	Indicates non-forest land cover.
Annual Residue in Vegetation Suppression	Reviewed deforestation areas from previous years before the current mapping year

Table 8: PRODES data Description [21].

5.4.2 Classes

The primary objective of the models trained in this study was to detect newly deforested areas between two consecutive dates, with an interval of approximately one year between them. As detailed in Section 5.2, we chose one reference date per year for each site. We proposed four classes for use in this study based on two consecutive dates, denoted as D_0 and D_1. Table 9 presents each class, an acronym, and the respective description.

Class	Acronym	Description
No Deforestation	$C_{no\ def}$	No deforestation was identified until D_1.
Deforestation	C_{def}	Deforestation was identified between D_0 and D_1.
Previous Deforestation	$C_{prev\ def}$	Deforestation was identified until D_0.
Background	C_{bg}	Areas excluded from this study during training, prediction, and evaluation steps.

Table 9: Classes and descriptions.

Using the PRODES polygons, we labeled the pixels with a spatial resolution of 10 meters and a coordinate reference system of the data images (optical and SAR). In this process, usually called rasterization, we classified each pixel based on the presence of PRODES deforestation polygons covering (even partially) the pixel area.

Figure 53 shows an example of the described classes. Pixels were classified as $C_{no\ def}$ if any PRODES deforestation polygon did not cover them. If any PRODES deforestation polygon covers the pixel, it was classified based on the acquisition date from the image used in the deforestation identification

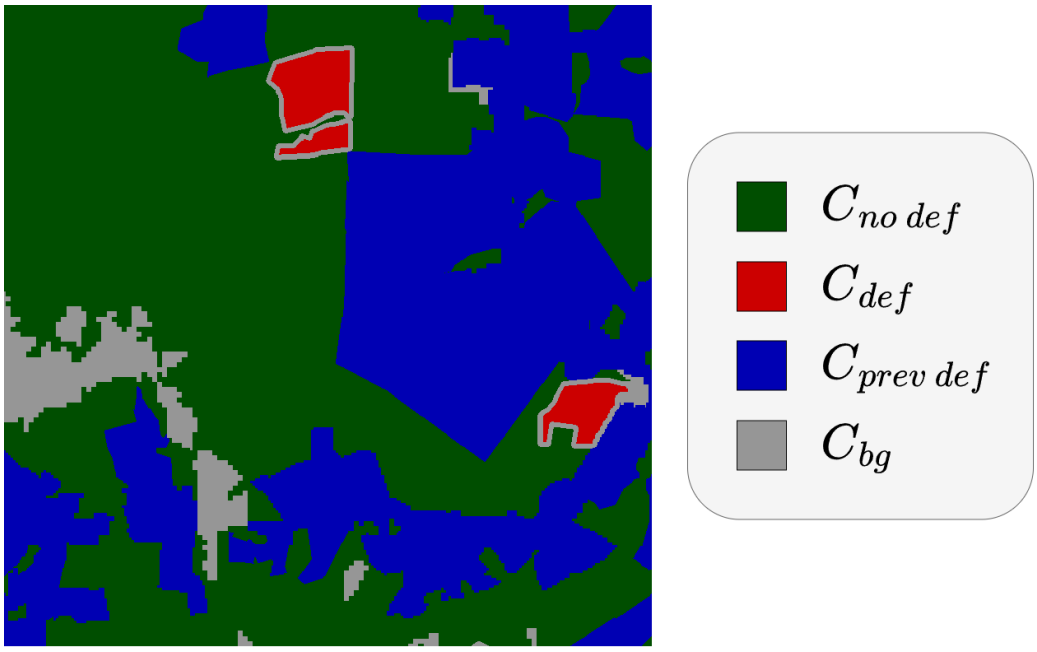


Figure 53: Examples of $C_{no\ def}$, C_{def}, $C_{prev\ def}$, and C_{bg} classes.

by PRODES. If the date was after D_1, or between D_0 and D_1, or before D_0, the pixel was classified as $C_{no\ def}$, C_{def}, and $C_{prev\ def}$, respectively. If a pixel was covered by the polygons belonging to hydrograph, no forest, or residual deforestation PRODES classes (described in Table 8), it was classified as C_{bg}. Due to the difference in the resolution of the main PRODES imagery source, which is the Landsat with 30 meters of spatial resolution, and the imagery used in this work, Sentinel with 10 meters of spatial resolution, the pixels covered by a buffer from the C_{def} polygons borders were classified as C_{bg}. This buffer was 30 meters (the equivalent of Landsat spatial resolution) and was applied in both directions, inner and outer.

5.5 Previous Deforestation Map

To identify new deforestation between consecutive PRODES dates, D_0 and D_1, we can assume we know pre-existing deforestation before D_0. We proposed using the Previous Deforestation Map for the model to see the already-known data. This map differentiates the old deforestation from recent deforestation, including the deforestation age information, setting the value 1 for the deforestation identified in D_0 and reducing this value by 0.1 for each additional year of the deforestation age, keeping a minimum value of 0.1. The areas with no deforestation until D_0 are represented by value 0. Figure 54 presents an example of the Previous Deforestation Map.

Figure 54: Previous Deforestation Map.

5.6 Evaluated Models

All proposed models were based on Swin-Unet and ResUnet networks presented in Chapter 4. Initially, we evaluated the single-modality models with all optical and SAR datasets. Table 10 lists the single-modality models. Each single-modality model has a ResUnet and Swin-Unet variation, as presented in Section 4.2. Some models incorporated the Previous Deforestation Map as auxiliary data, assessing its effectiveness in deforestation detection, while others didn't. We also investigated an alternative architecture, Multi-stream networks, similarly to [10].

Model Name [Dataset]	**Previous Deforestation Map**	**Multi-stream Architecture**
Optical[CLOUD-FREE]	Yes/No	Yes/No
Optical[CLOUD-DIVERSE]	Yes	No
SAR[AVERAGE-12]	Yes/No	No
SAR[AVERAGE-2]	Yes	Yes/No
SAR[SINGLE-2]	Yes	Yes/No

Table 10: Single-Modality models.

The evaluation of these single-modality models had the following primary objectives:

1. To be used as a baseline to be compared to the fusion models;
2. Assess the influence of the Previous Deforestation Map in detecting new deforestation areas;
3. Investigate the influence of three SAR datasets approaches;

4. Identify the influence of the Multi-stream architecture.

Based on the results of these single-modality models, which will be thoroughly discussed in Chapter 6, we decided:

1. To use the Previous Deforestation Map in all models;
2. To use the *AVERAGE-12* as SAR dataset;
3. Discard the Multi-stream architecture for temporal aggregation.

Guided by these decisions, we proposed optical and SAR fusion models based on ResUnet and Swin-Unet, using *AVERAGE-12* as the SAR dataset and including the Previous Deforestation Map as a crucial component. Table 11 outlines the Optical and SAR fusion models, each with a ResUnet and Swin-Unet variation, as detailed in Section 4.3. We also evaluated if the models used the pretraining strategy, where applicable, to ensure the integrity of our findings.

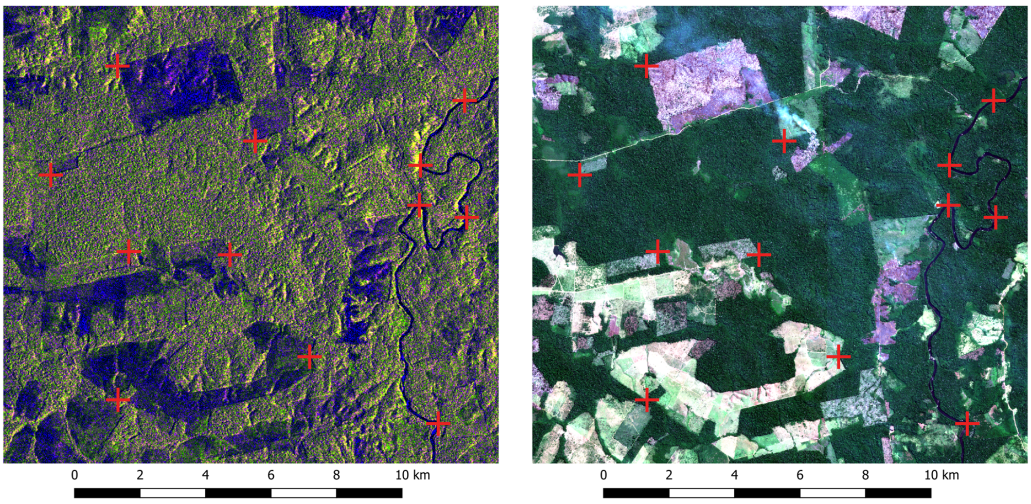
Model Name [Optical Dataset]	**Pre-trained**
Pixel Level[CLOUD-FREE]	No
Pixel Level[CLOUD-DIVERSE]	No
Feature Level (Middle Concat.)[CLOUD-FREE]	Yes/No
Feature Level (Middle Concat.)[CLOUD-DIVERSE]	Yes/No
Feature Level (Late Concat.)[CLOUD-FREE]	Yes/No
Feature Level (Late Concat.)[CLOUD-DIVERSE]	Yes/No
Feature Level (Late Cross-fusion)[CLOUD-FREE]	Yes/No
Feature Level (Late Cross-fusion)[CLOUD-DIVERSE]	Yes/No

Table 11: Optical and SAR fusion models.

5.7 Training Data Preparation

The Sentinel-1 (SAR) and Sentinel-2 (Optical) images, created using analogous algorithms, yield registered images within the same coordinate reference system and spatial resolution. This is illustrated in Figure 55, where Sentinel-1 (a) and Sentinel-2 (b) images cover the same region. The red crosses in both images mark the same sample points located in both images, showing that they are co-registered, thus eliminating the need for further registration between them.

The Sentinel-1 GRD images were already acquired as raw power powers instead of decibels, eliminating the need for conversion. The Sentinel-2 images were downloaded with processing Level-2A, which scales the reflectance in



55(a): Sentinel-1. 55(b): Sentinel-2.

Figure 55: Sample points located in Sentinel-1 and Sentinel-2.

each wavelength band by multiplying it by 10000 and rounding to an integer. Sentinel-2 data were converted to reflectance, dividing the values by 10000.

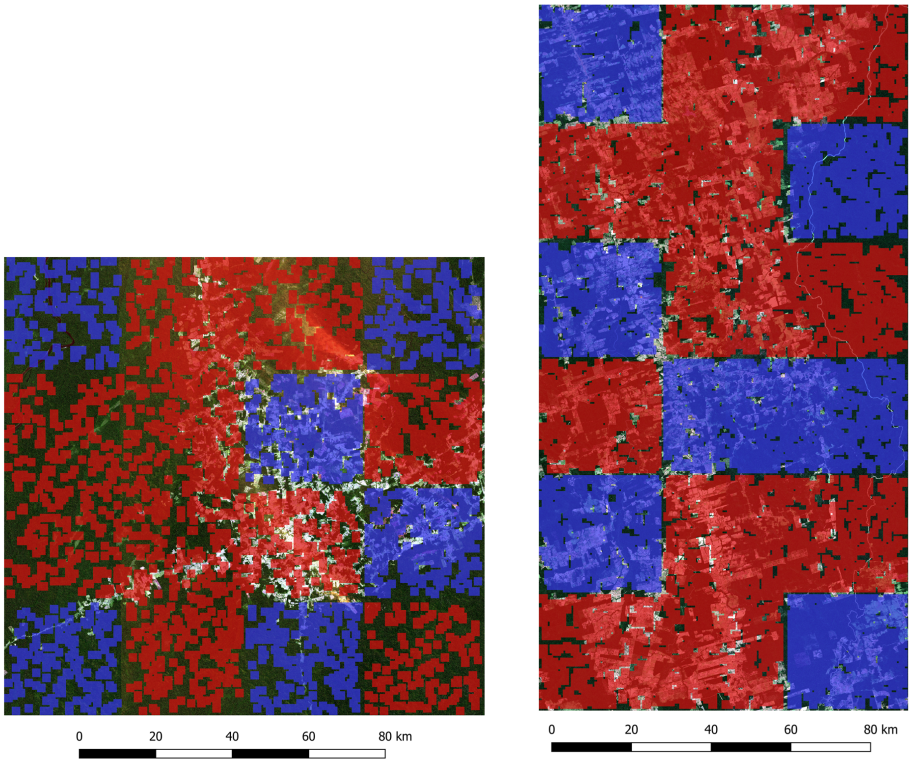
Outliers in the Sentinel-1 and Sentinel-2 images were removed to reduce the influence of extreme values on the analysis. This was done by clipping the data to values between the 0.01% and 99.99% percentiles of each image. The missing pixels, usually represented by NaN values, were then set to 0.

5.8 Training

To train the models, we generated patches with 224×224 pixels with 70% overlapping. We divided the area into training and validation tiles, from which we extracted the patches for the Training and Validation patches sub-datasets. Patches covering more than one type of tile (training or validation) were discarded to avoid data leakage, which may lead to overfitting the model's weights.

As we have an imbalanced dataset, we decided to artificially minimize it by showing more deforestation pixels to the model during the training step. We classified each patch according to the C_{def} pixels percentage. Patches with a minimum of 2% of C_{def} pixels were classified as P_{def} and the rest as $P_{no\ def}$. Due to the class imbalance, the number of $P_{no\ def}$ is much greater than P_{def}. We randomly discarded $P_{no\ def}$ patches until the numbers of $P_{no\ def}$ and P_{def} were equal. Figure 56 presents the spatial distribution of training (**red**) and validation (**blue**) patches in both study areas.

Another strategy to minimize the class imbalance was to choose Focal



56(a): Site 1.

56(b): Site 2.

Figure 56: Training (**red**) and validation (**blue**) patches.

Loss as the loss function to adjust the weights. We selected the Focal Loss parameter $\gamma = 3$ and ignored the C_{bg} pixels.

To allow the comparison of the models' convergence, we fixed the number of mini-batches for each epoch to 200 mini-batches per epoch.

As some models, especially the single-modality SAR models, presented convergence issues, we did a warming-up training for 10 epochs for all models except the pre-trained models.

We trained all models for 10 warmup epochs. Then, we monitored the F1-Score metric of the C_{def} from the training and validation patches during the training process. We trained the model for 500 epochs or until the C_{def} F1-Score, using the validation patches, didn't improve more than 10^{-3} for 10 epochs (Early Stopping). The model was discarded if the C_{def} F1-Score was at most 0.2 to avoid models that did not converge.

We trained each proposed model, presented in Section 5.6, five times, creating a committee of five independent instances for each model, using the Adam optimizer with a learning rate of $2.0 \cdot 10^{-5}$.

The datasets *CLOUD-DIVERSE* and *SINGLE-2* have three possible images for each year. The models using these datasets employed random image combinations. This strategy maximized the diversity conditions in the images

the models saw during the training step, especially the cloud diversity in the optical images from the *CLOUD-DIVERSE* dataset.

5.9 Prediction

For each model instance, we predicted the whole study area. As the patch size is much smaller than the image size, we divided the site region into patches with 224×224 pixels with an overlap of 36 pixels. From each predicted patch, border pixels of 36 pixels (from all sides) were discarded to avoid a border effect in the patches.

As in the training step, we employed all possible image combinations, which is especially important to evaluate more diverse cloud-covered optical images in the *CLOUD-DIVERSE* dataset. All assessed models applied the Softmax in the output, ensuring that all class scores totaled 1. The final prediction for each image combination is the average of the predictions generated by each model instance.

5.10 Evaluation

Before calculating the metrics, we adjusted the class scores to incorporate prior knowledge into the evaluation. Considering the score of the classes from each pixel as a vector $[P_{no\ def}, P_{def}, P_{prev\ def}, P_{bg}]$, we adjusted these scores which the labels belong to the known classes, $C_{prev\ def}$ and C_{bg}, as presented in Equation 5-1, in which $*$ is the element-wise product. Then, we determined the predicted class, identifying the maximum value of P_*.

$$Values_{adjusted} = \begin{cases} [0,0,1,0] & \text{if label is } C_{prev\ def} \\ [0,0,0,1] & \text{if label is } C_{bg} \\ [1,1,0,0] * Values_{output} & \text{otherwise} \end{cases} \tag{5-1}$$

Each predicted image was evaluated using F1-Score, Precision, and Recall metrics. Each predicted pixel classified to classes $C_{prev\ def}$ and C_{bg} were reclassified as discard class ($C_{discard}$), which weren't considered in the metrics calculation.

To ensure the compatibility between the predictions of the class C_{def} and the PRODES data, which was used to generate the labels, the contiguous predictions belonging to the class C_{def} with an area less than 6.25 hectares were reclassified to class $C_{discard}$.

The cloud presence can affect each pixel of the predictions from the optical data models. We classified each pixel as *Cloudy Pixel* or *Cloud-free Pixel* to evaluate this effect. The Cloud Maps, described in Section 5.3.1.3, related to the optical, were used for this classification. If the maximum value of the pixels from these Cloud Maps were greater than 50, the pixel was classified as *Cloudy Pixel*; otherwise, the pixel was classified as *Cloud-free Pixel*.

6
Results

This chapter presents the results of the models proposed in the previous chapters, followed by a discussion about them. Section 6.1 presents the results for Single-modality models, comparing the single-stream and multi-stream temporal aggregation strategies and the Previous Deforestation Map. Section 6.2 presents the Fusion models results, aiming to verify if the proposed methodology can generate results using *CLOUD-DIVERSE* dataset close to the results from *CLOUD-FREE* dataset.

In this chapter, we present the F1-Score metric results. However, the results from all evaluated metrics can be found in Appendix B.4

To make comprehension easier in this chapter, we created short names for some models, as shown in Table 12

Model Name	**Temporal Aggregation**	**Short Name**
Optical	Single-stream	Optical
Optical	Multi-stream	Optical Multi-Stream
SAR	Single-stream	SAR
SAR	Multi-stream	SAR Multi-Stream
Pixel Level	Single-stream	Pixel
Feature Level (Middle Concat.)	Single-stream	Feat-Mid
Feature Level (Late Concat.)	Single-stream	Feat-Late
Feature Level (Late Cross-fusion)	Single-stream	Cross-Fusion

Table 12: Models' short names.

6.1
Single-Modality Models

The first single-modality analysis aimed at assessing the influence of the Previous Deforestation Map, presented in Section 5.5, as auxiliary data. We analyzed this influence by comparing the results from the evaluation metrics from the models' predictions using or not the Previous Deforestation Map as auxiliary data.

Figures 57 and 58 compare the models' F1-Score with (**orange** bars) and without (**blue** bars) the Previous Deforestation Map as part of the input.

We considered ResUnet and Swin-based architectures with single and multi-stream temporal aggregation strategies using optical *CLOUD-FREE* dataset in Site 1 and 2, respectively.

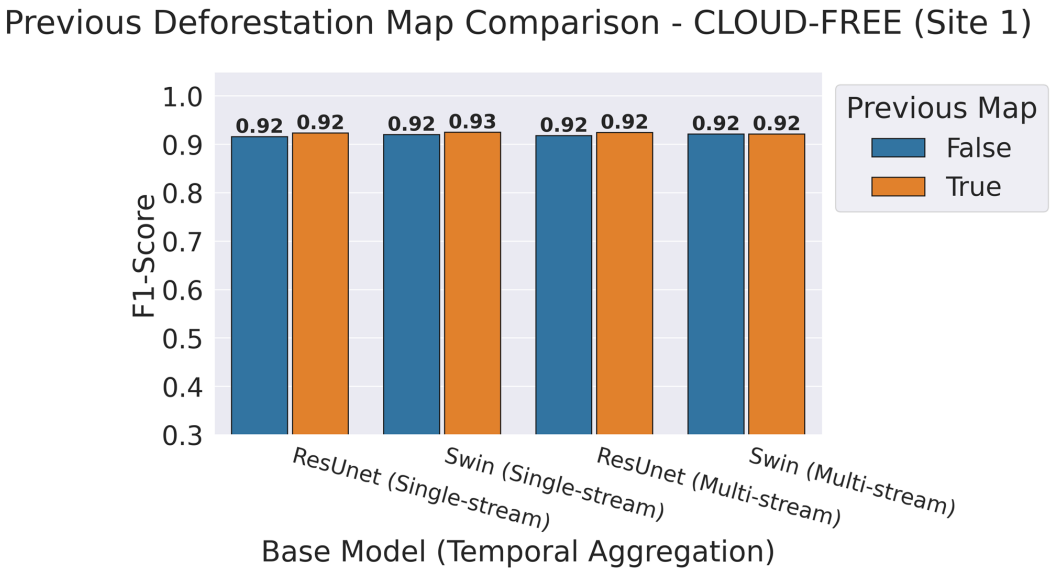


Figure 57: F1-Score Comparison of the models with (**orange**) and without (**blue**) Previous Deforestation Map (*CLOUD-FREE* from Site 1).

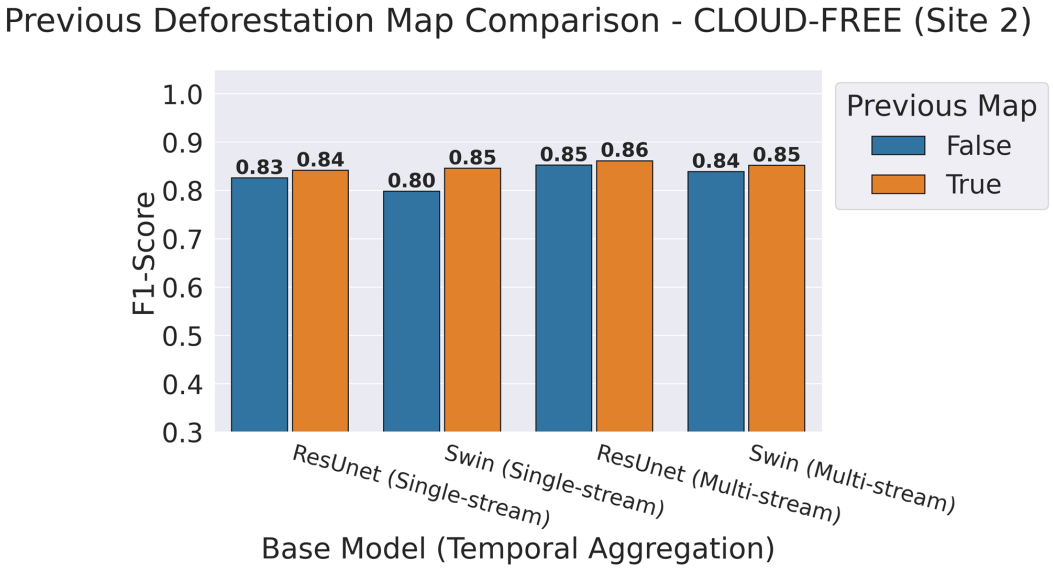


Figure 58: F1-Score Comparison of the models with (**orange**) and without (**blue**) Previous Deforestation Map (*CLOUD-FREE* from Site 2).

We also investigated the influence of the Previous Deforestation Map on the F1-score in experiments on the SAR datasets. Figures 59 and 60 compare the models with (**orange** bars) and without (**blue** bars) this auxiliary data, whereby we took the ResUnet and Swin-based architectures with single (using *AVERAGE-12* dataset) and multi-stream (using *AVERAGE-2* and *SINGLE-2* datasets) temporal aggregation strategies in Site 1 and 2, respectively.

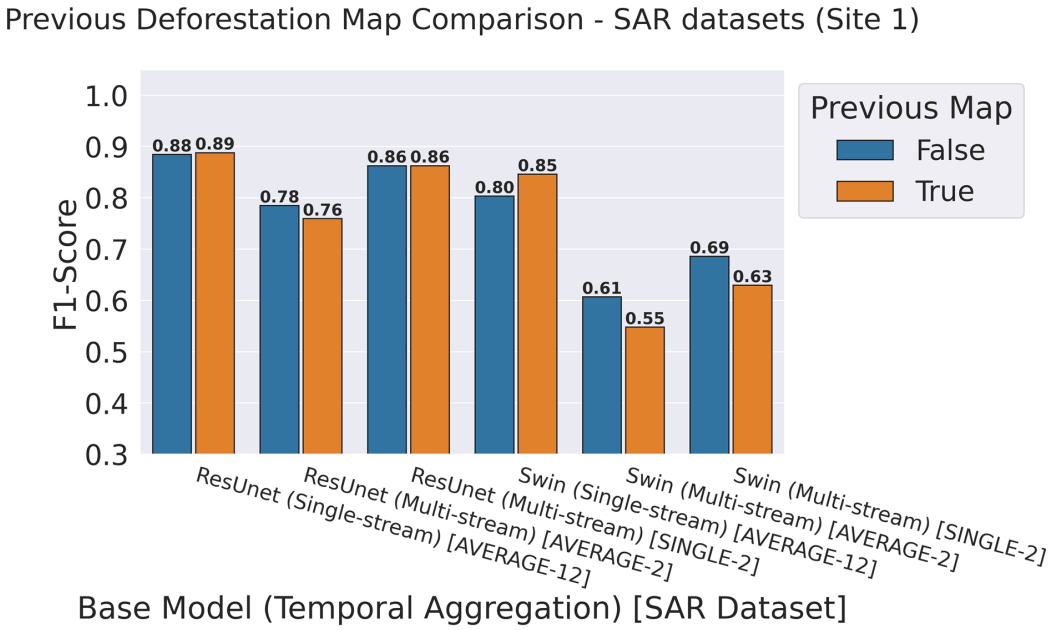


Figure 59: F1-Score Comparison of the models with (**orange**) and without (**blue**) Previous Deforestation Map (*AVERAGE-12*, *AVERAGE-2*, and *SINGLE-2* datasets from Site 1).

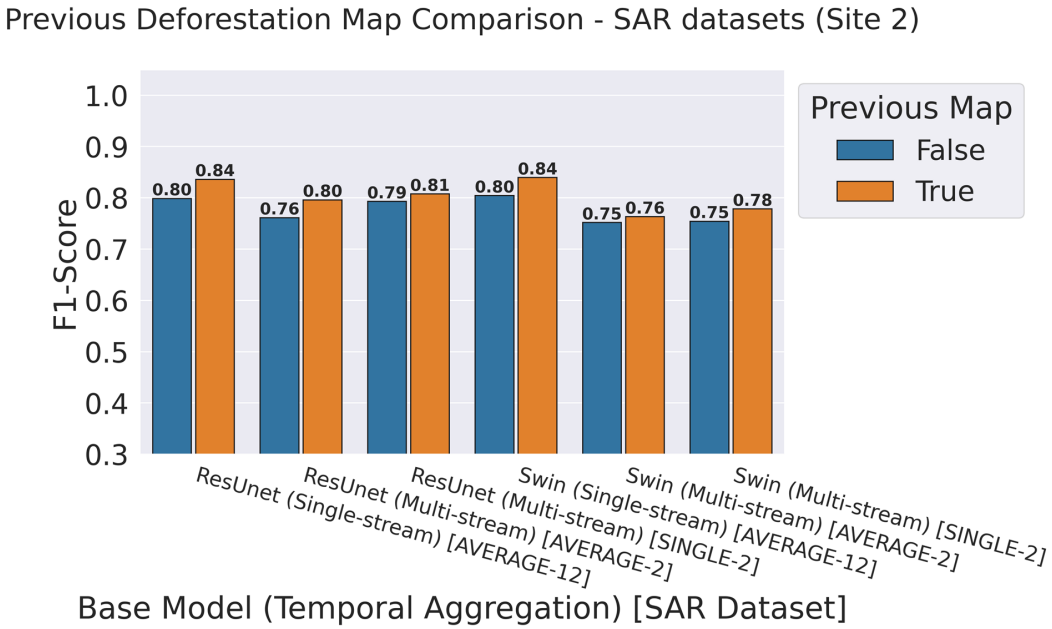


Figure 60: F1-Score Comparison of the models with (**orange**) and without (**blue**) Previous Deforestation Map (*AVERAGE-12*, *AVERAGE-2*, and *SINGLE-2* datasets from Site 2).

Figures 61 and 62 present the training and prediction average times and Standard Deviation (**gray** vertical lines) from ResUnet and Swin-based architectures using the Previous Deforestation Map (**blue** bars) and not (**orange** bars) as auxiliary data in *CLOUD-FREE* and SAR datasets, respectively.

We also calculate the number of trainable parameters of each model.

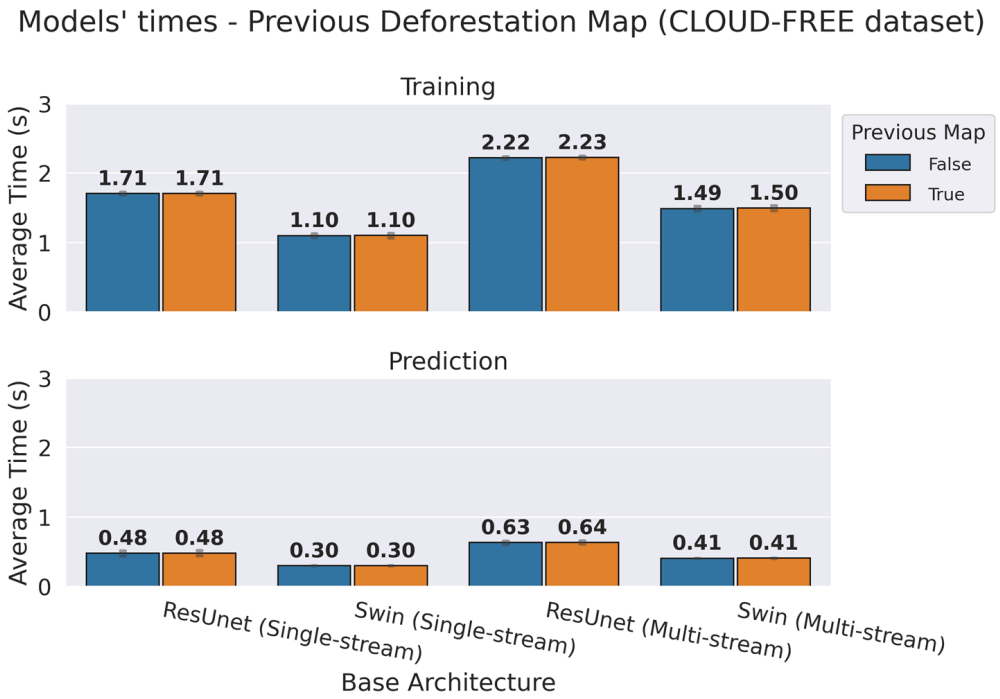


Figure 61: Training and Prediction times for Previous Deforestation Map (*CLOUD-FREE* dataset)

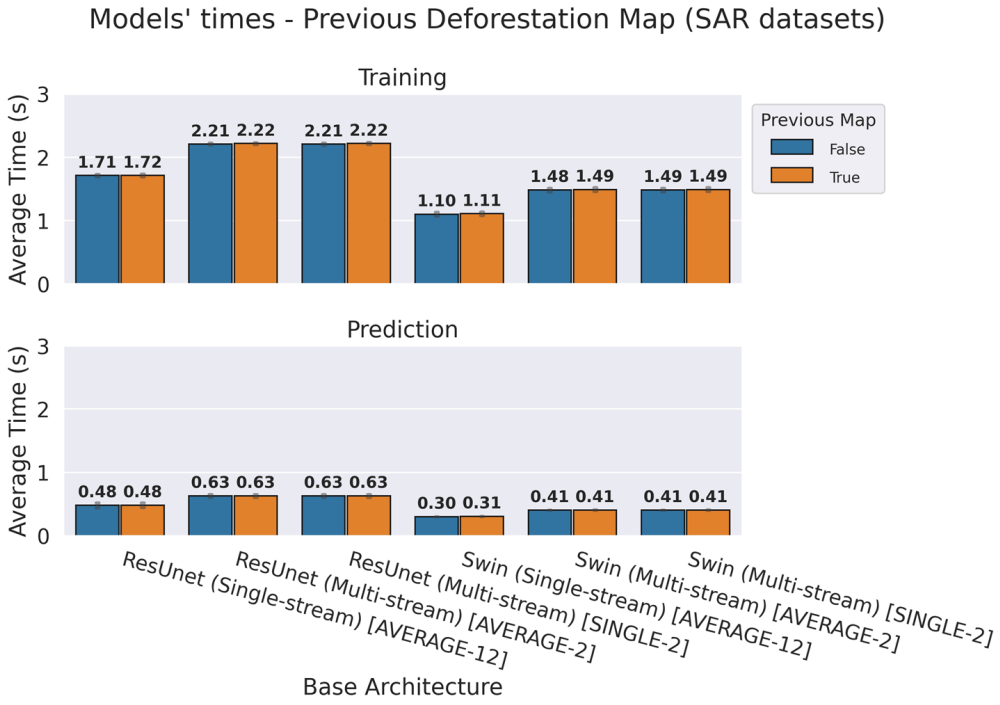


Figure 62: Training and Prediction times for Previous Deforestation Map (SAR datasets)

Figures 63 and 64 present the number of trainable parameters (Millions) from ResUnet and Swin-based architectures using the Previous Deforestation Map (**blue** bars) and not (**orange** bars) as auxiliary data in optical and SAR

datasets, respectively.

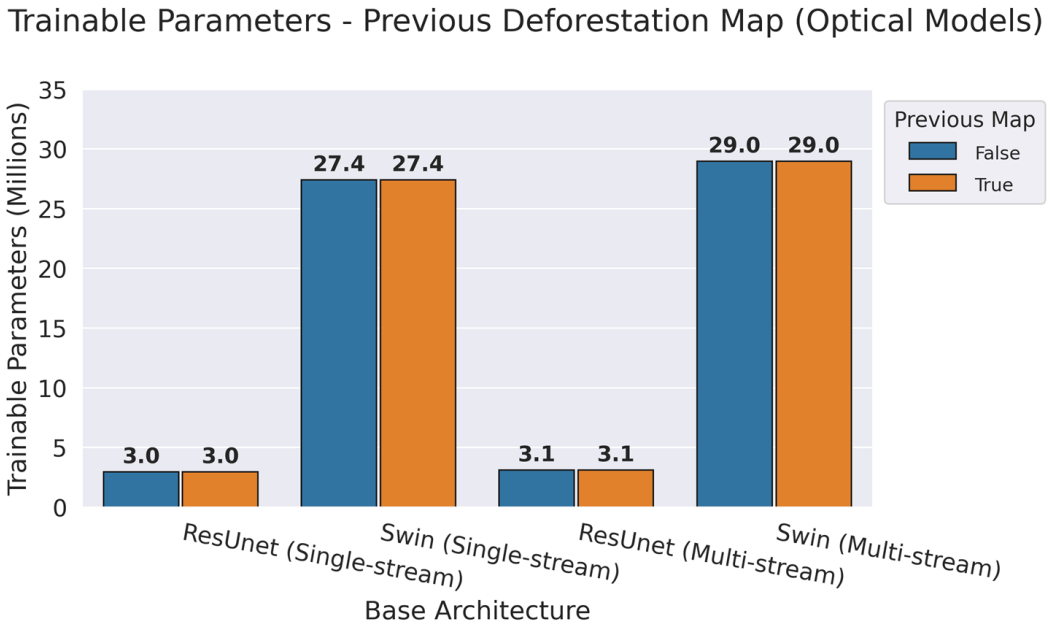


Figure 63: Trainable Parameters (Millions) comparison - Previous Deforestation Map (Optical datasets)

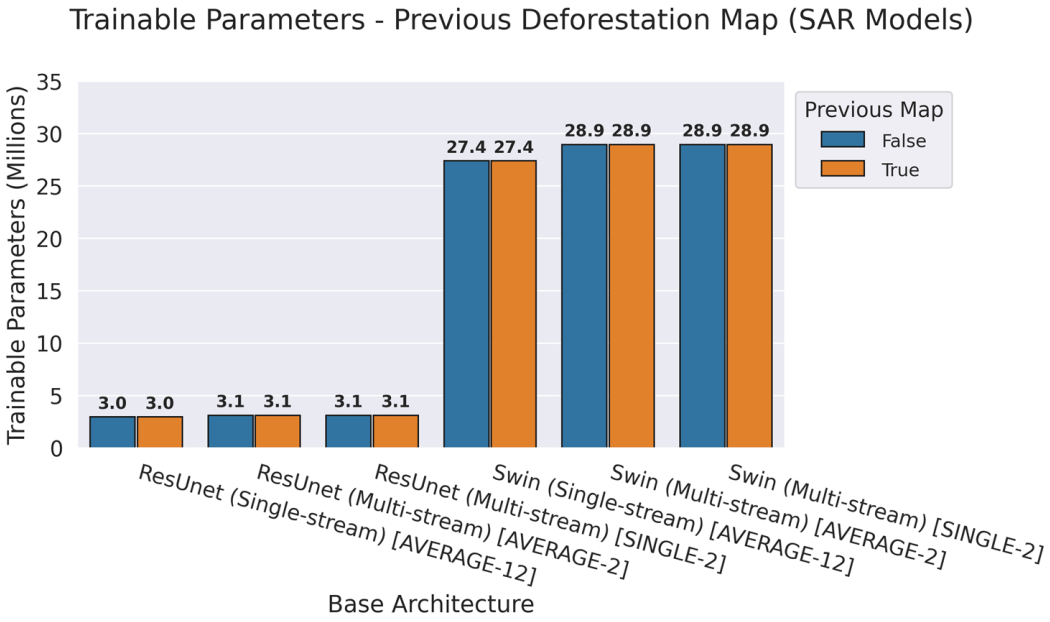


Figure 64: Trainable Parameters (Millions) comparison - Previous Deforestation Map (SAR datasets)

We also investigated the temporal aggregation strategies. In Section 4.2.1, we presented the single-stream temporal strategy, in which the temporal aggregation is done by concatenating the data from different times before input in the model. We also presented the multi-stream temporal aggregation. This analysis was realized by comparing the results from the evaluation metrics

from the models' predictions using single and multi-stream architectures. All models in this analysis used the Previous Deforestation Map as auxiliary data.

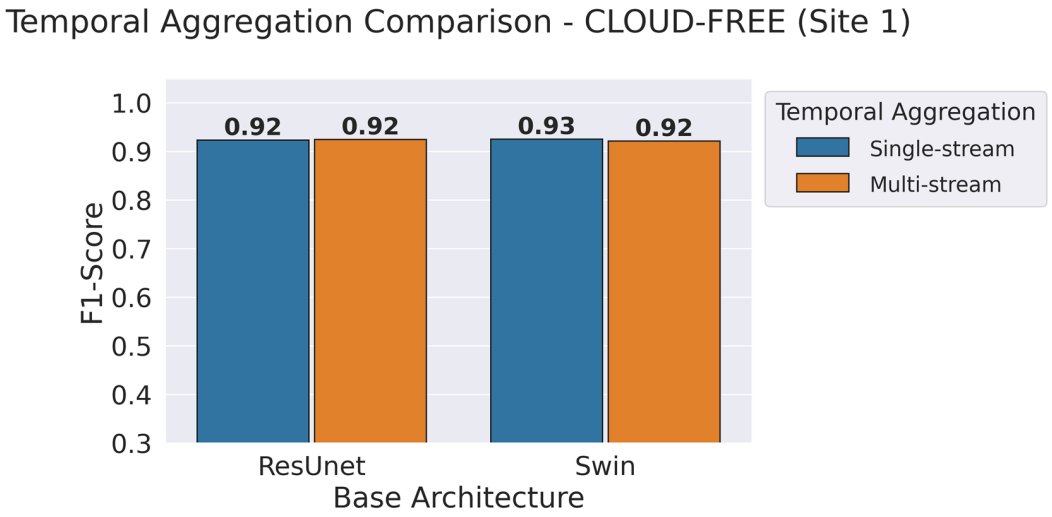


Figure 65: Temporal aggregation comparison (F1-Score) in *CLOUD-FREE* dataset (Site 1).

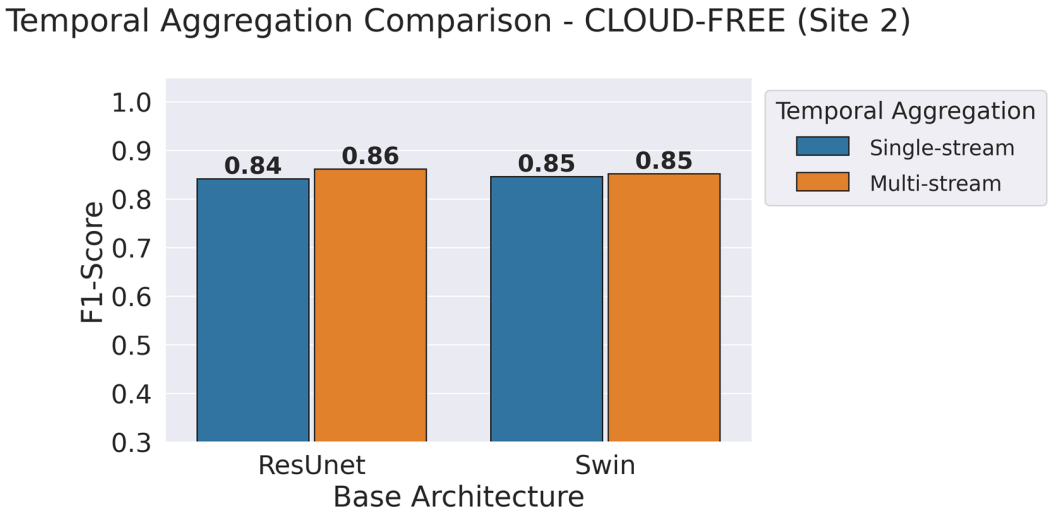


Figure 66: Temporal aggregation comparison (F1-Score) in *CLOUD-FREE* dataset (Site 2).

The temporal aggregation analysis in the optical data used only the *CLOUD-FREE* dataset. Figures 65 and 66 compare the single (**blue** bars) and multi-stream (**orange** bars) temporal aggregation strategies by the F1-Score metric, from ResUnet and Swin-based architectures with optical *CLOUD-FREE* dataset in Site 1 and 2, respectively.

We also analyzed the temporal aggregation using the SAR datasets. Figures 67 and 68 compare the single (**blue**, **orange**, and **green** bars) and multi-stream temporal (**red** and **purple** bars) aggregation strategies by the

F1-Score metric, from ResUnet and Swin-based architectures with *AVERAGE-12*, *AVERAGE-2*, and *SINGLE-2* datasets in the Site 1 and 2, respectively.

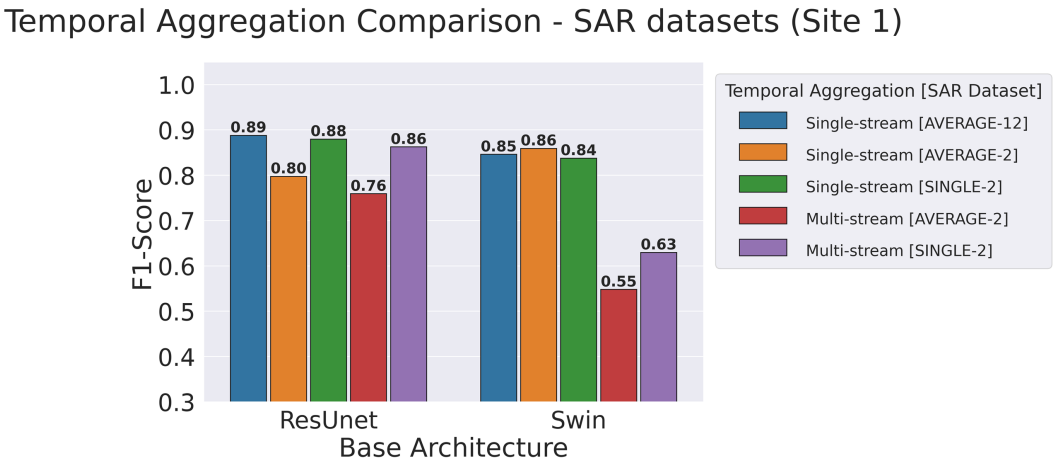


Figure 67: Temporal aggregation comparison (F1-Score) in SAR datasets (Site 1).

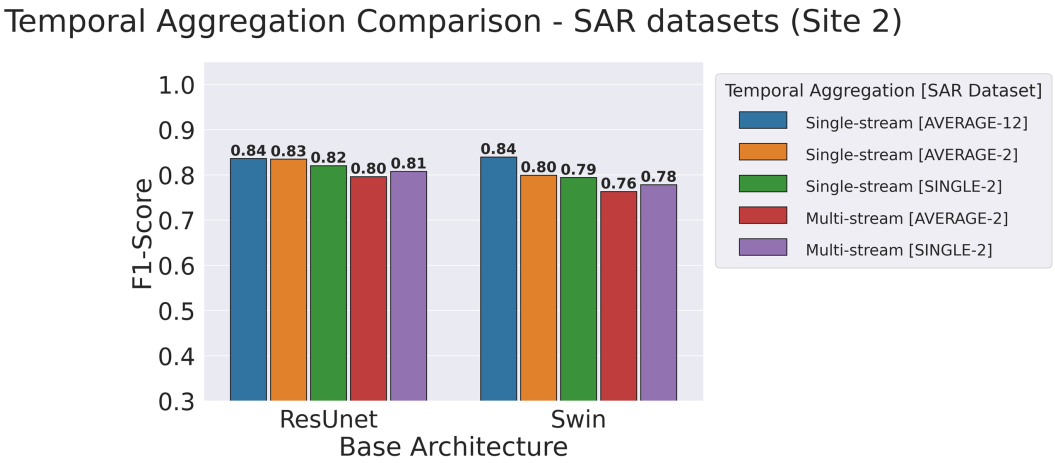


Figure 68: Temporal aggregation comparison (F1-Score) in SAR datasets (Site 2).

Figure 69 presents the training and prediction average times and Standard Deviation (**gray** vertical lines) from ResUnet and Swin-based architectures using Single-Stream (**blue** bars) and Multi-Stream (**orange** bars) in optical datasets.

Figure 70 presents the training and prediction average times and Standard Deviation (**gray** vertical lines) from ResUnet and Swin-based architectures using single (**blue**, **orange**, and **green** bars) and multi-stream temporal (**red** and **purple** bars) aggregation strategies in SAR datasets.

We also calculate the number of trainable parameters of each model. Figure 71 presents the number of trainable parameters (Millions) from ResUnet

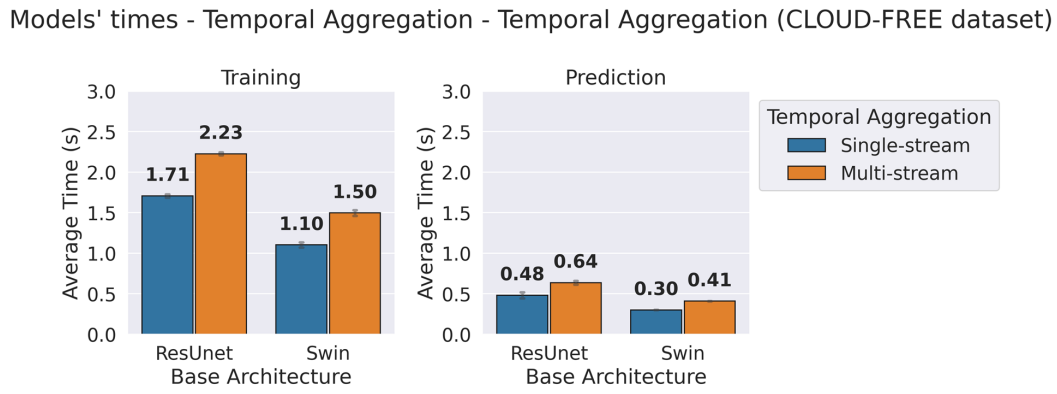


Figure 69: Training and Prediction times for Temporal Aggregation's strategies (*CLOUD-FREE* dataset)

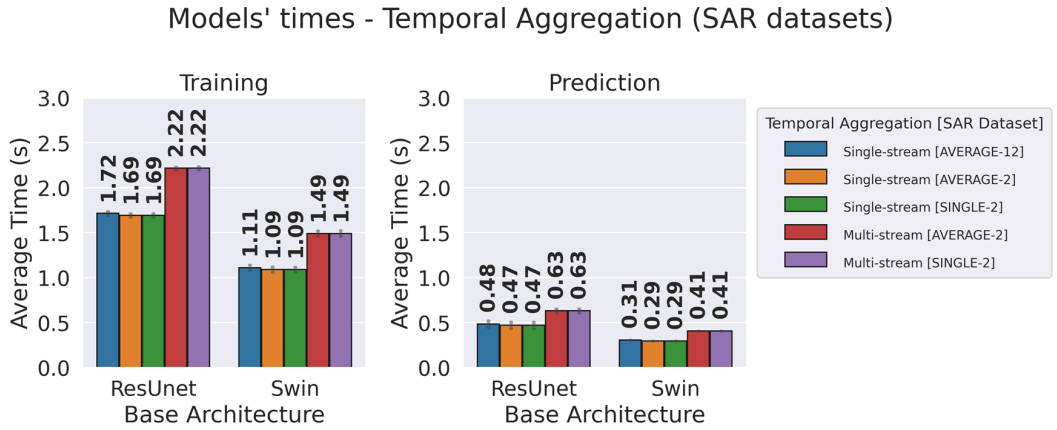


Figure 70: Training and Prediction times for Temporal Aggregation's strategies (SAR datasets)

and Swin-based architectures using Single-Stream (**blue** bars) and Multi-Stream (**orange** bars) in optical datasets.

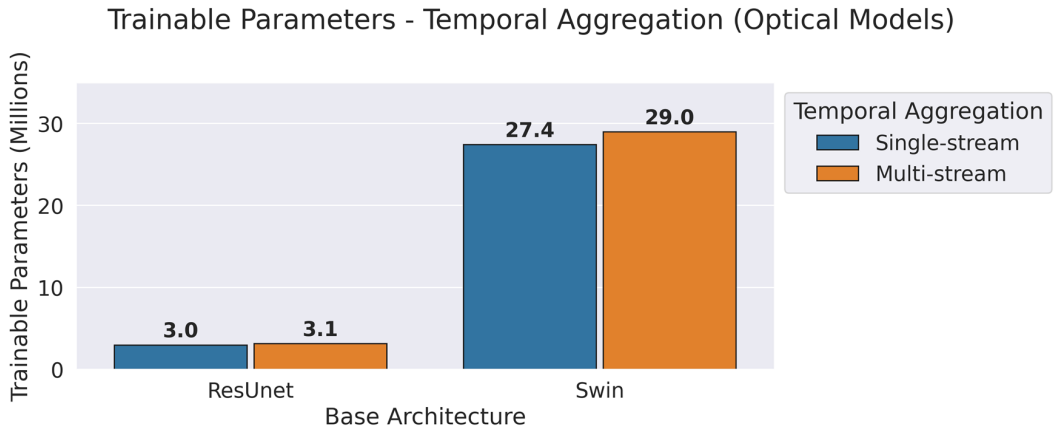


Figure 71: Trainable Parameters (Millions) comparison - Temporal Aggregation (Optical datasets)

Figure 72 presents the number of trainable parameters (Millions) from

ResUnet and Swin-based architectures using single (**blue**, **orange**, and **green** bars) and multi-stream temporal (**red** and **purple** bars) aggregation strategies in SAR datasets.

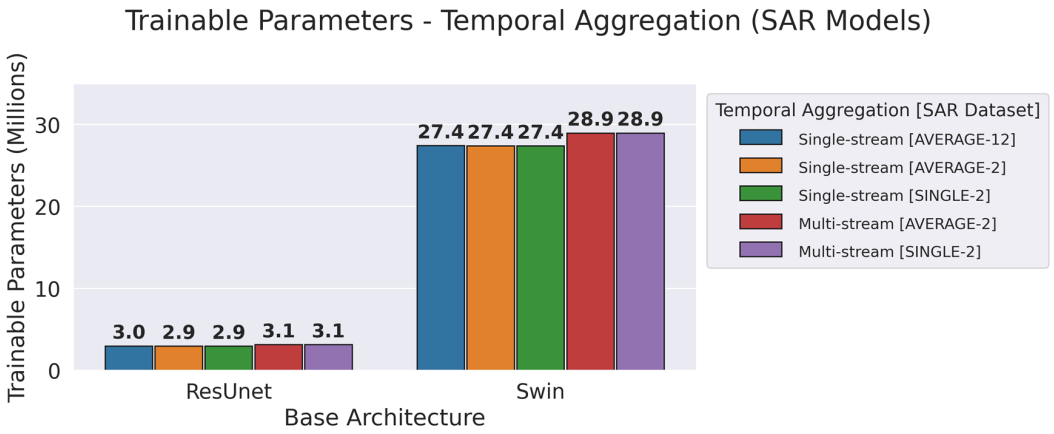
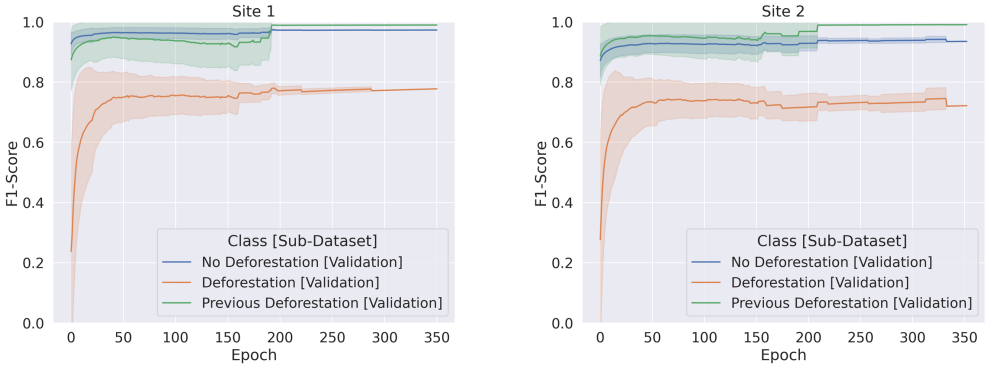


Figure 72: Trainable Parameters (Millions) comparison - Temporal Aggregation (SAR datasets)

6.1.1 Single-modality Models Discussion

Analyzing the single-modality models, we can identify that the classes $C_{no\ def}$ and $C_{prev\ def}$ converge faster than C_{def}. Figure 73 presents the F1-Score for Validation sub-datasets from each class (including the respective Standard Deviation colored bands) evaluated during the training of all single-modality models. The imbalance between the classes' pixels can explain this difference. As the C_{def}'s convergence is the hardest, all the convergence analysis was focused on this class.



73(a): Site 1.

73(b): Site 2.

Figure 73: F1-Score for Validation sub-dataset patches from classes C_{nodef} (**blue** lines), C_{def} (**orange** lines), and $C_{prev\ def}$ **green** lines), and the respective Standard Deviation (colored bands), for all single-modality models.

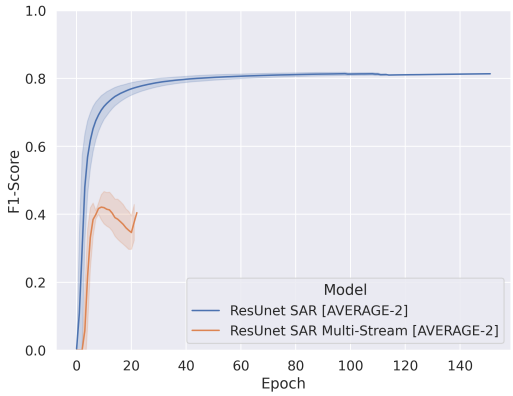
Based on the comparison results of the Previous Deforestation Map, this additional data improved the deforestation detection capabilities of almost all the proposed models without significant computational cost impact. The optical models using the Previous Deforestation Map always produced greater or equal results than the respective models without this data. In Site 2 (Figure 58), the improvement was more significant than in Site 1 (Figure 57), probably because the results in Site 1 were already very high and close to the reference data, which makes any improvement harder.

The SAR models incorporating the Previous Deforestation Map generally yielded results equal to or better than their counterparts, which lacked this data across most models. Notable exceptions were observed in the multi-stream models utilizing the *AVERAGE-2* (ResUnet and Swin-based) and *SINGLE-2* (solely Swin-based) datasets from Site 1 (Figure 59). An examination of the F1-Score for C_{def} within the Validation sub-dataset patches over the training epochs (Figure 74) reveals that these multi-stream models (shown in Subfigures 74(a), 74(c), and 74(e)) encountered convergence issues, even when excluding models that failed to achieve the minimum F1-Score. Since this issue was exclusive to the images from Site 1, a decision was made not to retrain these models.

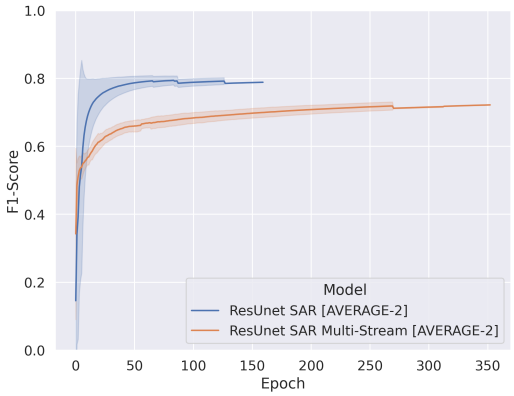
Based on the optical models' results, the multi-stream strategy did not significantly improve overall. When examining the results from Site 1 (Figure 65), the single-stream strategy performed slightly better than the multi-stream strategy. However, the multi-stream temporal aggregation strategy required more substantial training and prediction times, with increases of 33% and 35% on average, respectively, compared to the single-stream approach (Figure 69). This discrepancy highlights the variability in performance between different sites and suggests that the effectiveness of the multi-stream strategy may depend on specific site characteristics.

Unlike the optical models, where the single and multi-stream strategies produced similar results, the SAR models showed more varied outcomes. After analyzing each base architecture and site combination, the best model always used single-stream temporal aggregation. The *AVERAGE-12* dataset, which was only used in single-stream models, could better extract the deforestation features in SAR images because it could capture the information throughout the year, allowing the identification of deforestation occurrences where vegetation has regenerated and is not identifiable in the last image, as shown by the sample in Figure 75.

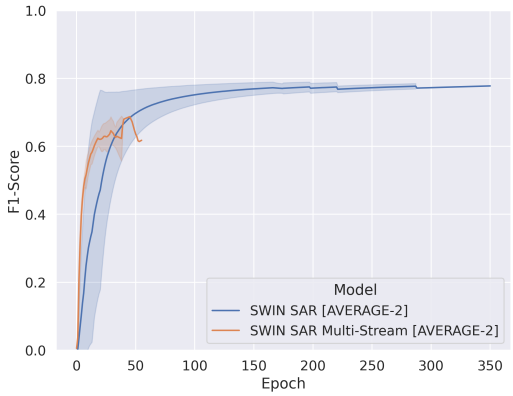
Despite the significant difference in the number of trainable parameters between ResUnet and Swin-based models (Swin-based models have approxi-



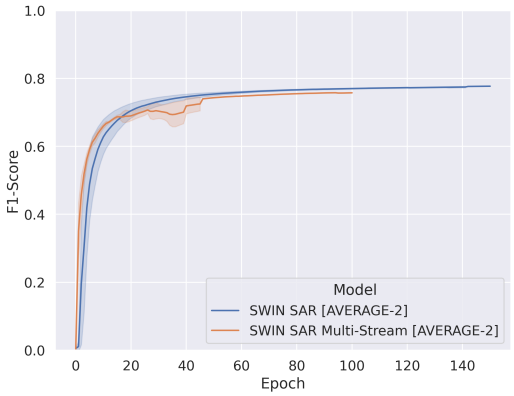
74(a): ResUnet (*AVERAGE-2*) - Site 1.



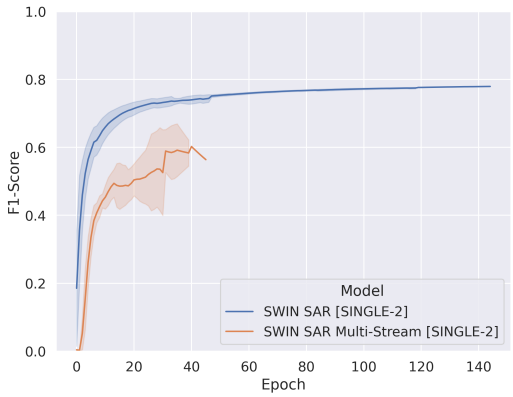
74(b): ResUnet (*AVERAGE-2*) - Site 2.



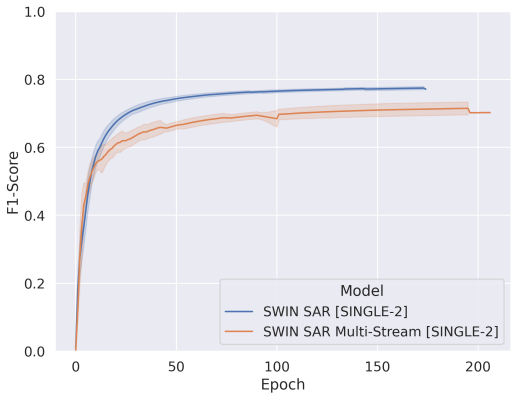
74(c): Swin (*AVERAGE-2*) - Site 1.



74(d): Swin (*AVERAGE-2*) - Site 2.



74(e): Swin (*SINGLE-2*) - Site 1



74(f): Swin (*SINGLE-2*) - Site 2

Figure 74: F1-Score for C_{def} in Validation sub-dataset patches from training epochs in SAR models using single (**blue** lines) and multi-stream (**orange** lines) temporal aggregation strategies.

mately nine times more trainable parameters), this discrepancy did not affect the training and prediction times (Figures 63, 64, 71, 72, 61, 62, 69, and 70). This can be attributed to the convolution operation, which reduces the number of trainable parameters but applies them multiple times in a sliding window operation across the entire input feature maps (Section 2.3). In contrast, the

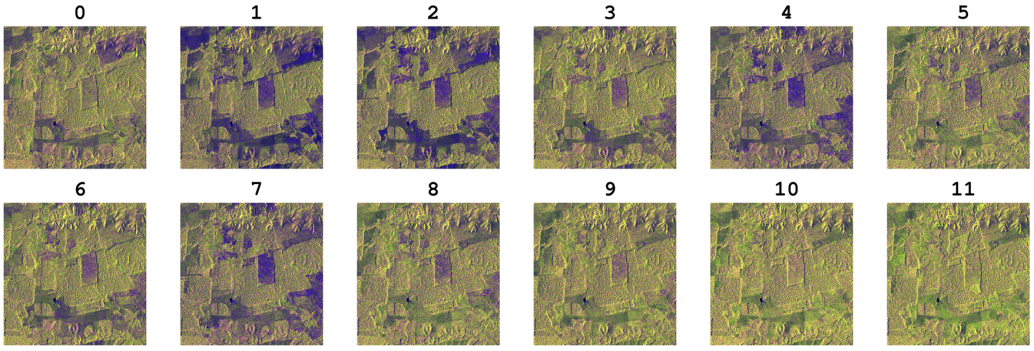


Figure 75: AVERAGE-12 SAR dataset sample (12 monthly average images).

Vision Transformer concept, which is used in the Swin-based models, repeats the operation for each window much less frequently than the sliding window operation (Section 2.4), and were the same results from [18].

Based on the presented Single-modality models' results, we made the following decisions regarding the fusion models:

1. Use the single-stream strategy for temporal aggregation across all models. The multi-stream strategy did not show significant improvements and required more computational resources.

2. Use the *AVERAGE-12* dataset for all models. This dataset consistently delivers the best, or nearly the best, results in all scenarios, especially when combined with the previously decided strategy.

3. Use the Previous Deforestation Map as auxiliary data for all models. This additional data improved the results for all scenarios using the single-stream temporal aggregation strategy, with an insignificant impact on training and prediction times and the trainable parameters.

6.2 Fusion Models

We compared the results from the models using the optical *CLOUD-DIVERSE* (with and without pre-training strategy) and the *CLOUD-FREE* dataset.

Figures 76 and 77 present the F1-Score metric results from Optical (**blue** bars), Pixel Level Fusion (**orange** bars), Feature Level (Middle) Fusion (**green** bars), Feature Level (Late) Fusion (**red** bars), Feature Level (Late Cross-Fusion) Fusion (**purple** bars), and SAR (**brown** bars) models, using ResUnet and Swin-based architectures in Site 1, respectively. These models were also organized based on the dataset used. The results from the models

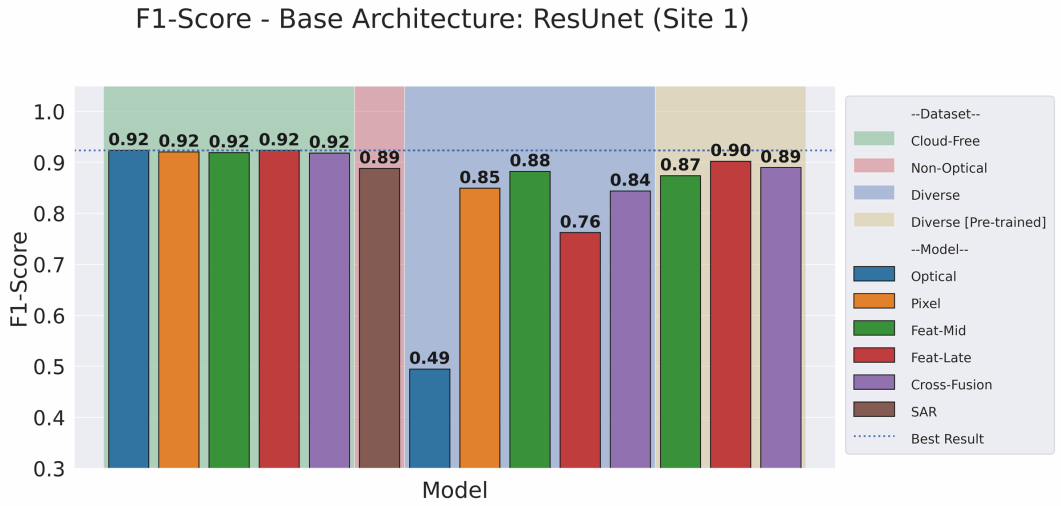


Figure 76: F1-Score for ResUnet-based models' comparison (Site 1)

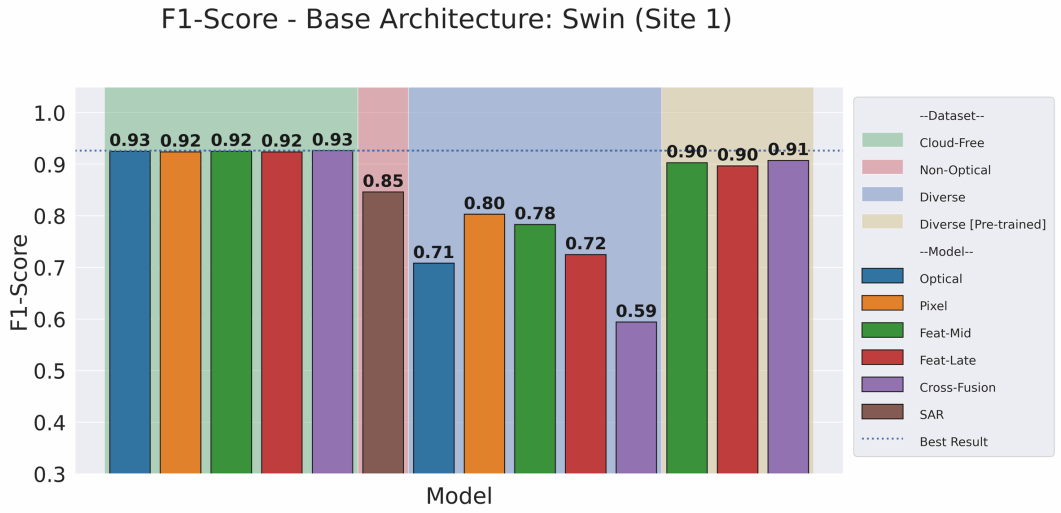


Figure 77: F1-Score for Swin-based models' comparison (Site 1)

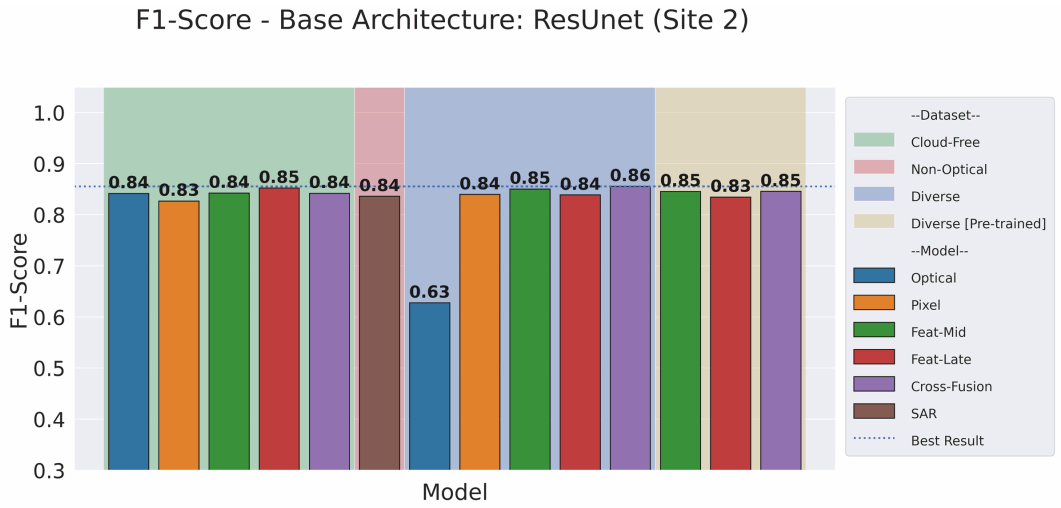


Figure 78: F1-Score for ResUnet-based models' comparison (Site 2)

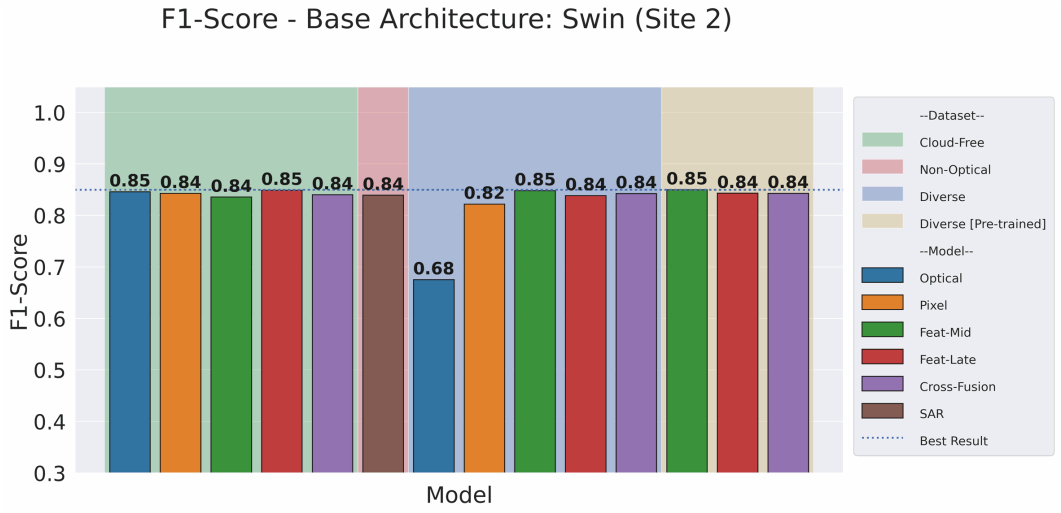


Figure 79: F1-Score for Swin-based models' comparison (Site 2)

using *CLOUD-FREE* optical dataset are grouped in the **green** background. The SAR model didn't use any optical dataset and is presented with red background. The models using *CLOUD-DIVERSE* optical dataset are grouped in the **blue** (without pre-training strategy) and in the yellow (using the pre-training strategy) background. The best result from each base architecture is presented as a dotted **blue** line, allowing the comparison between the different models (single-modality or fusion), training strategies (from scratch or pre-training), and datasets (*CLOUD-FREE*, *CLOUD-DIVERSE*, or non-optical). Figures 78 and 79 present the same results for Site 2.

We detailed the cloud influence, analyzing the results in different cloud conditions. To accomplish this goal, we classified each pixel based on the maximum value of the Cloud Maps (presented in Section 5.3.1.3) related to the optical images used by the model. If the Cloud Map's pixel from any optical image was greater than 50%, that pixel was classified as *Cloudy Pixel*; otherwise, the pixel was classified as *Cloud-free Pixel*. These models were also organized based on the dataset used.

Figures 80 and 81 present the F1-Score metric results from each model, based on the cloud condition, using ResUnet and Swin-based architectures in Site 1, respectively. *All Pixels* (**blue** bars) presents the results considering all pixels, independent of the cloud condition. The models using only *CLOUD-FREE* or SAR datasets present only this result. *Cloudy Pixels* (**orange** bars) presents the pixels affected by clouds in any optical image. Finally, *Cloud-free Pixels* (**green** bars) presents the pixel less affected by clouds. The results from the models using *CLOUD-FREE* optical dataset are grouped in the **green** background. The SAR model didn't use any optical dataset and is presented with red background. The models using *CLOUD-DIVERSE* optical

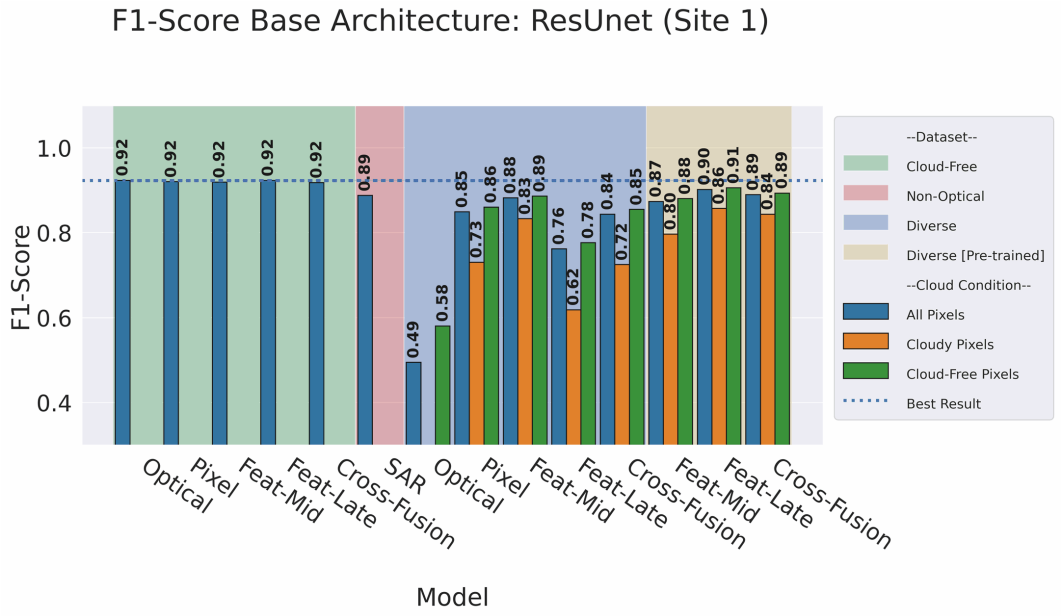


Figure 80: F1-Score for ResUnet-based models' cloud effect comparison (Site 1)

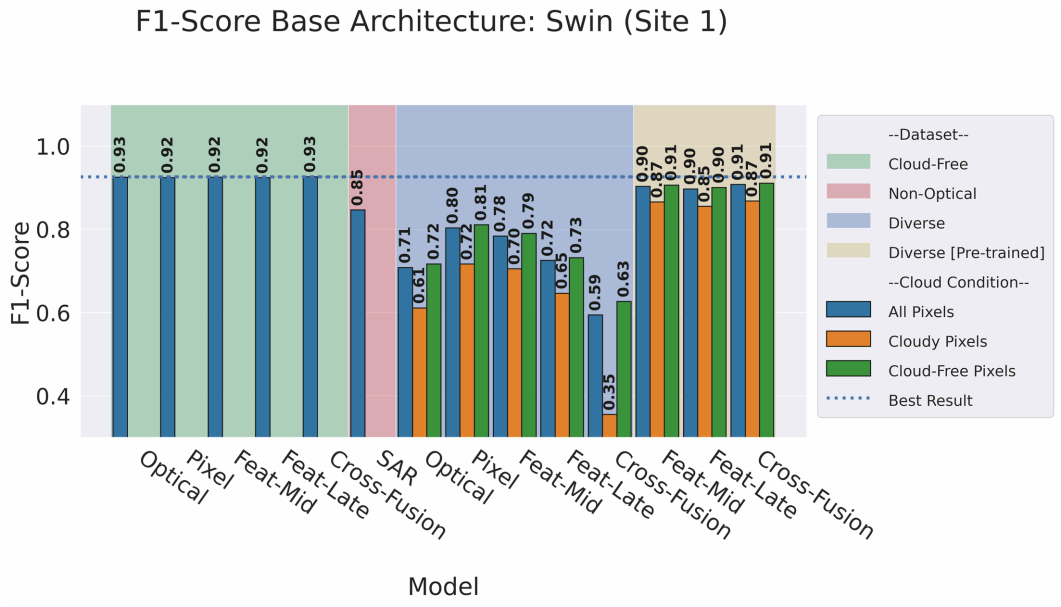


Figure 81: F1-Score for Swin-based models' cloud effect comparison (Site 1)

dataset are grouped in the blue (without pre-training strategy) and in the yellow (using the pre-training strategy) background. The best result from each base architecture is presented as a dotted **blue** line, allowing the comparison between the different models. Figures 82 and 83 present the same results for Site 2.

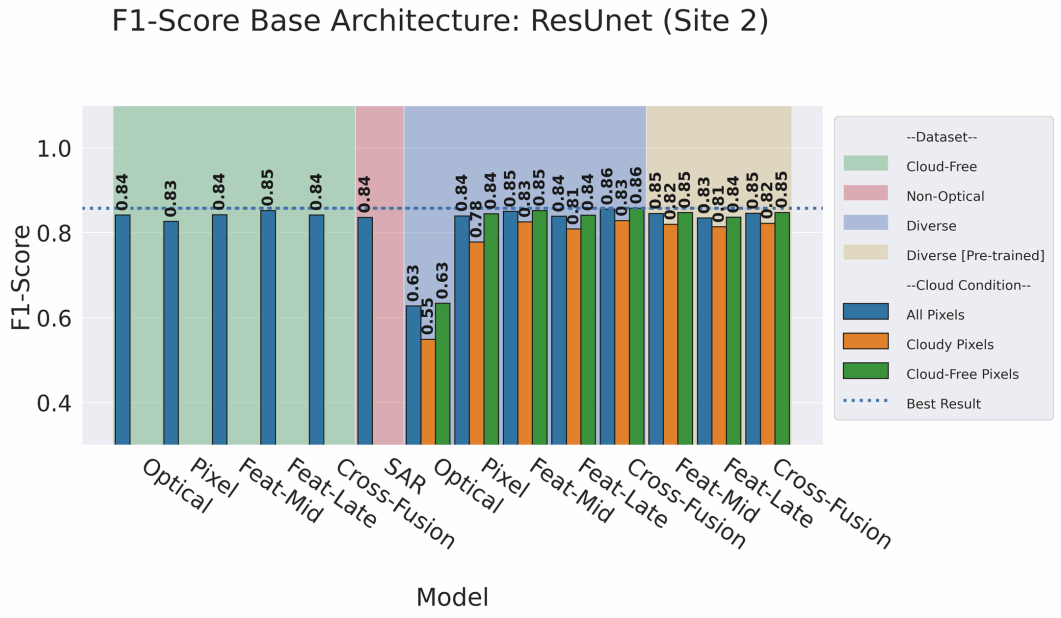


Figure 82: F1-Score for ResUnet-based models' cloud effect comparison (Site 2)

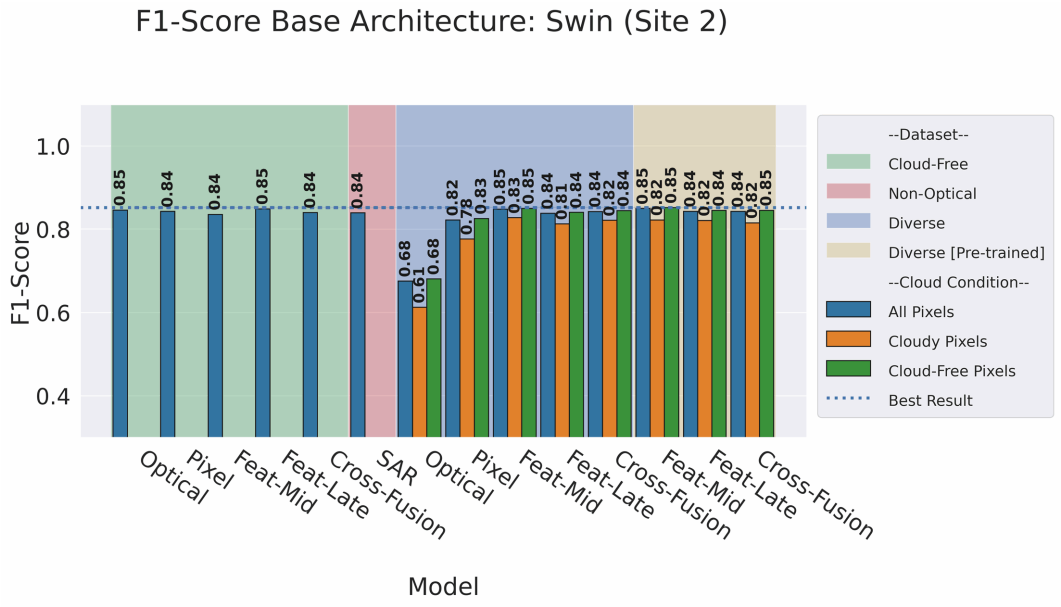
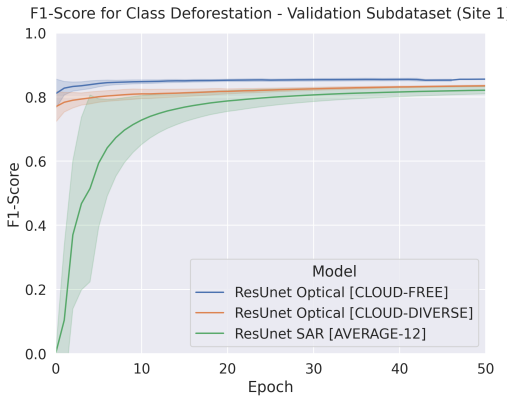


Figure 83: F1-Score for Swin-based models' cloud effect comparison (Site 2)

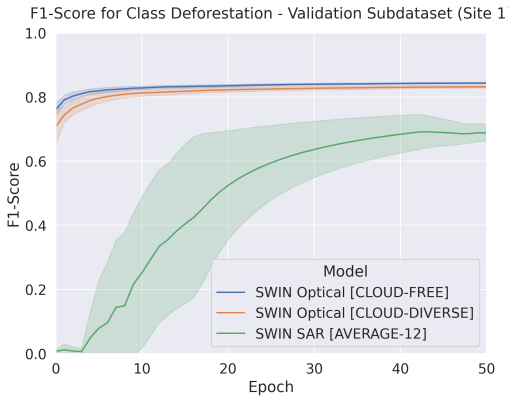
6.2.1 Pre-training Strategy Discussion

A notable difference in convergence was observed in the F-1 Score metric between models using SAR and optical datasets during the training step for single-modality models (Figure 84).

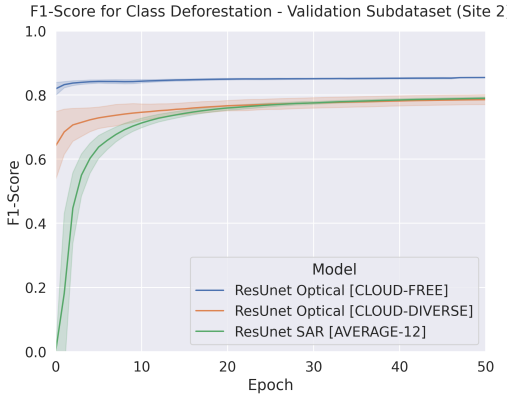
These aspects, such as the difference in the number of available patches, their cloud presence, and the convergence of the Optical and SAR models, could affect the models from Site 1. To minimize this problem, we proposed the pre-training strategy for fusion models using *CLOUD-DIVERSE* dataset.



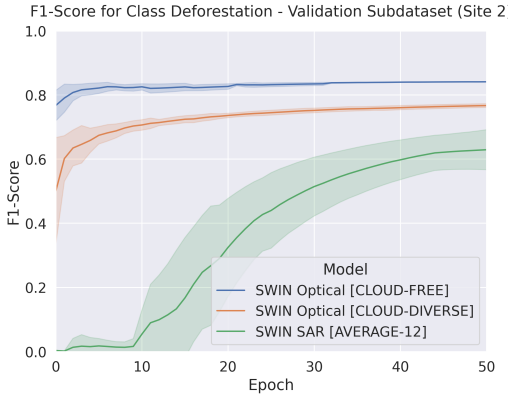
84(a): ResUnet-based - Site 1.



84(b): Swin-based - Site 1.



84(c): ResUnet-based - Site 2.



84(d): Swin-based - Site 2.

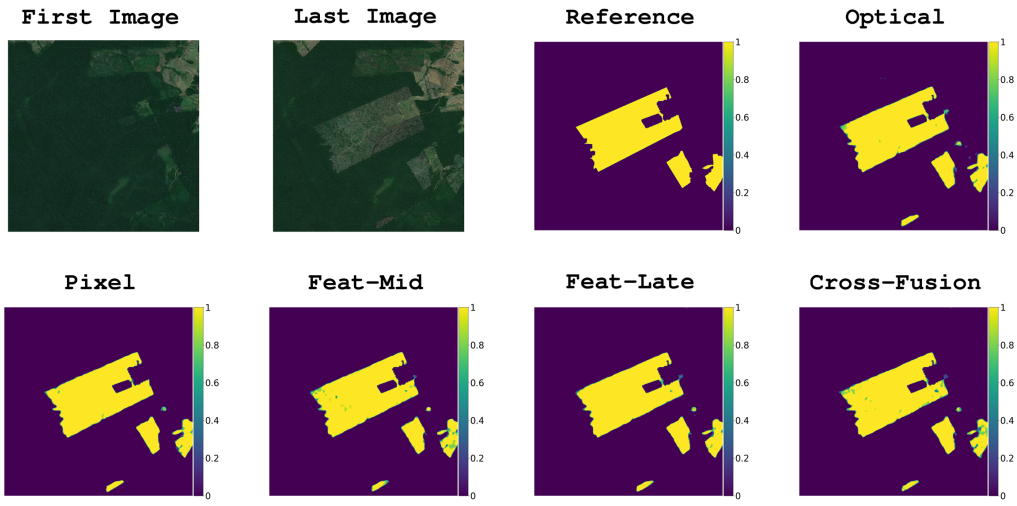
Figure 84: F1-Score for C_{def} in Validation sub-dataset patches from training epochs in all single-modality models for Site 1 and 2.

The exception was the Pixel Level model, in which the fusion occurs before input in the model, and we don't have weights related to a specific modal.

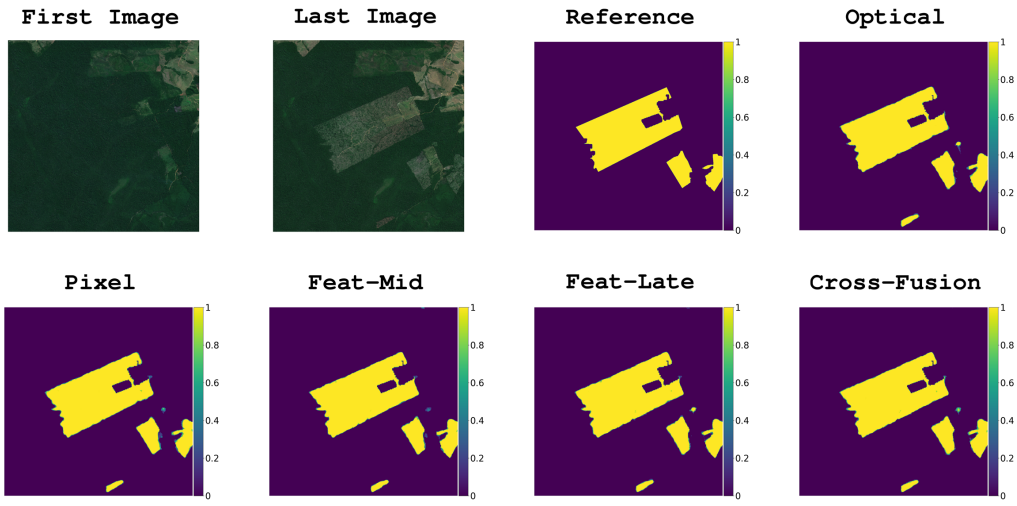
Initially, we tried initializing the weights with those from the respective single-modality models trained on the *CLOUD-DIVERSE* dataset. However, this did not improve the results, likely because the optical model trained with *CLOUD-DIVERSE* patches was affected by previously mentioned issues. Therefore, we used weights from optical models trained on the *CLOUD-FREE* dataset to initialize the fusion models and then trained these models using the *CLOUD-DIVERSE* dataset. This strategy showed the best results and is presented by the bars in the yellow background in Figures 76, 77, 78, 79, 80, 81, 82, and 83.

6.2.2 Cloud Presence Discussion

Analyzing the predictions and the metrics' results, we can infer that, using the *CLOUD-FREE* dataset, the optical-SAR data fusion didn't significantly improve the deforestation detection capabilities. Figure 85 presents examples from the input optical images, the reference data, and the predictions from Optical, Pixel, Feat-Mid, Feat-Late, and Cross-Fusion models based on ResUnet and Swin, from the same area.



85(a): ResUnet-based.



85(b): Swin-based.

Figure 85: Predictions from optical and fusion models using *CLOUD-FREE* dataset.

As expected, the predictions from the optical model using *CLOUD-DIVERSE* dataset (**blue** bars in blue background from Figures 76, 77, 78, 79) dropped in comparison to the respective model using *CLOUD-FREE* dataset.

However, this difference was more evident in the ResUnet-based models than in the Swin-based models, especially from Site 1. Figure 86 compares the predictions from ResUnet and Swin-based Optical models using *CLOUD-DIVERSE* dataset from a region from Site 1. Further experiments are needed to determine the reason for the discrepancy between ResUNet and Swin-based models. However, the difference in model parameters and the learning strategies inherent to each architecture may offer some justification.

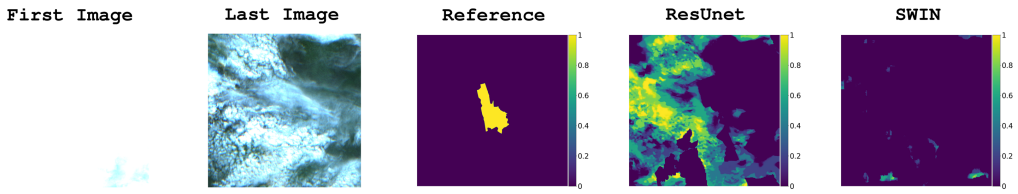


Figure 86: Predictions from ResUnet and Swin-based optical models using *CLOUD-DIVERSE* dataset.

The results from the fusion models using *CLOUD-DIVERSE* dataset (blue background from Figures 76, 77, 78, 79) presented a diverse behaviour between the sites. In the models from Site 2, all fusion models delivered results very close to the models using *CLOUD-FREE* dataset, especially the Feat-Mid, Feat-Late, and Cross-Fusion models. However, the fusion models from Site 1 presented a more varied performance, especially the Swin-based models.

In analyzing the cloud effect comparison (Figures 80, 81, 82, 83), it is evident that the fusion models from Site 1 performed significantly worse on *Cloudy Pixels* compared to their performance on *Cloud-free Pixels*. This disparity was not as apparent in the models from Site 2. Although all models showed reduced performance on *Cloudy Pixels*, this effect was more pronounced in the fusion models from Site 1. Table 13 shows the probability of a cloudy patch and the number of patches in each sub-dataset for each site. In this case, a cloudy patch was defined as a patch in which at least 50% of its pixels can be classified as *Cloudy Pixel* (as described in the Section 5.10).

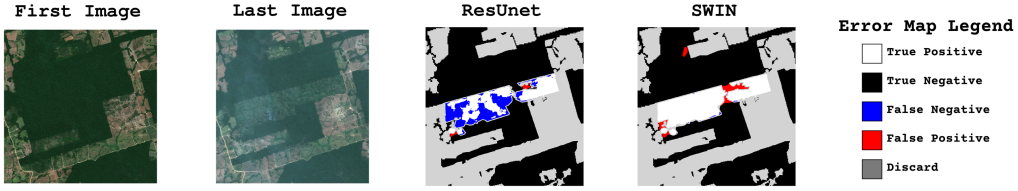
Site	**Sub-dataset**	**Cloudy Patch Probability**	**Amount of Patches**
1	Training	12.3%	25 416
1	Validation	6.9%	16 002
2	Training	8.4%	59 418
2	Validation	7.3%	43 074

Table 13: Average cloud probability in Training and Validation sub-datasets.

We hypothesized that differences in cloud presence and the number of patches between the training and validation sub-datasets could impact the

training process.

87(a): Sample 1.



87(b): Sample 2.

87(c): Sample 3.

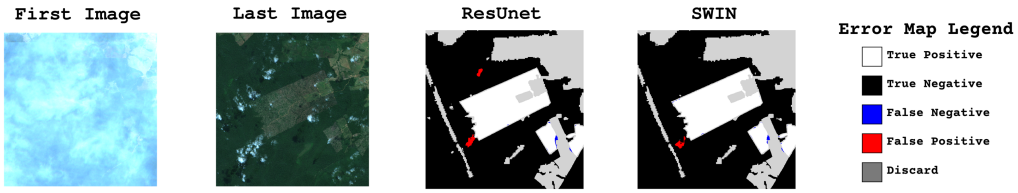
Figure 87: Error maps from ResUnet and Swin-based Optical models using *CLOUD-DIVERSE* dataset in the same area.

Figure 87 presents the error maps from ResUnet and Swin-based Optical models using *CLOUD-DIVERSE* dataset using different input image combinations from the same area. In the error map, the True Positive, True Negative, False Negative, and False Positive are presented by **white**, **black**, **blue**, and **red** pixels, respectively. As the pixels from $C_{discard}$ weren't considered during the evaluation step, they are shown as **gray**. These error maps confirm the metrics results in which the Swin-based models using *CLOUD-DIVERSE* dataset were less affected by clouds than the respective ResUnet-based. The ResUnet-based model prediction presented much more False Negative pixels than the Swin-based one using the same input optical images.

However, this effect is only observed when the clouds are present in the last optical image. Figure 88 presents the same error maps from another area, in which only the first optical image is affected by cloud presence. We can note that both models could identify deforestation well despite the first optical image being fully covered by clouds. This effect can be explained by the training protocol that ignored the C_{bg} pixels. During the training step, the models might

prioritize the last optical image, focusing on identifying the presence or absence of deforestation phenomena in this image.

88(a): Sample 1.

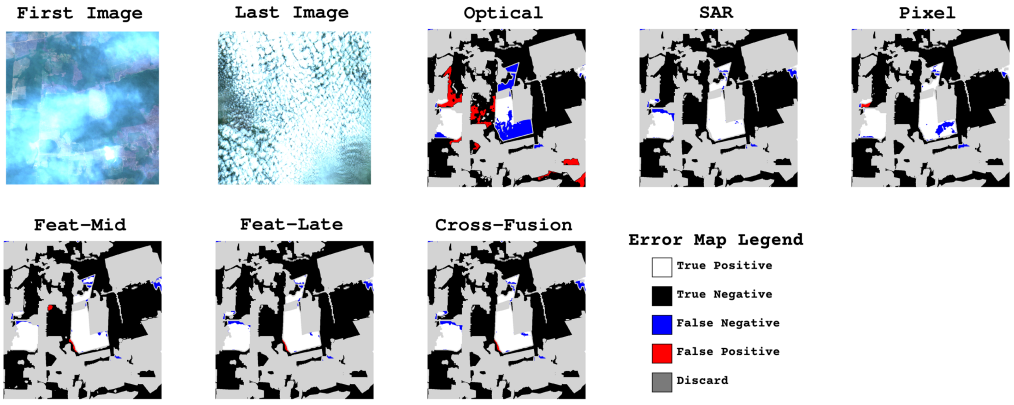


88(b): Sample 2.

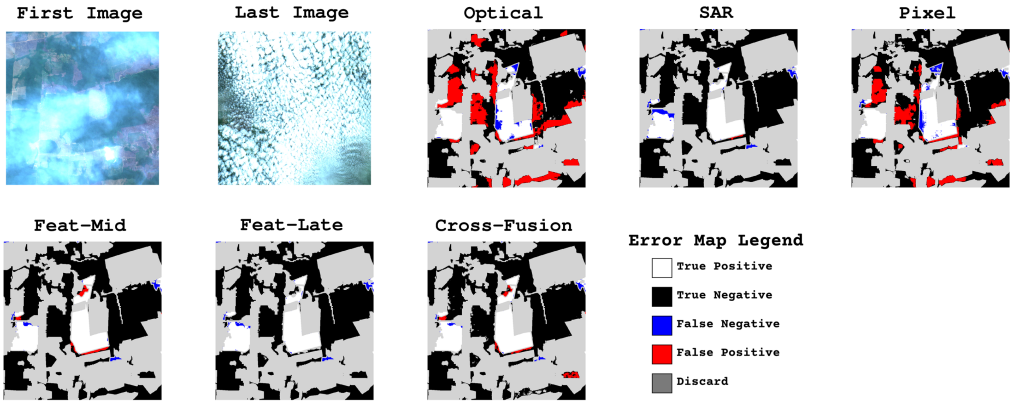
Figure 88: Error maps from ResUnet and Swin-based Optical models using *CLOUD-DIVERSE* dataset in the same area.

When we examine the error maps from the fusion models using the *CLOUD-DIVERSE* dataset and a pre-training strategy (except in the Pixel Level model), as shown in Figure 89, we notice a clear pattern: the Pixel Level model is more adversely affected by clouds in the optical data compared to the other fusion strategies. This indicates that the Pixel Level model is less robust to cloud interference.

From these error maps, we can observe that when clouds impacted the optical model, the fusion models (except the Pixel Level model) produced errors similar to those of the SAR model. This is evident because the errors (False Positives or False Negatives) predicted by the SAR model appeared in similar locations within the fusion models' predictions.



89(a): ResUnet-based.



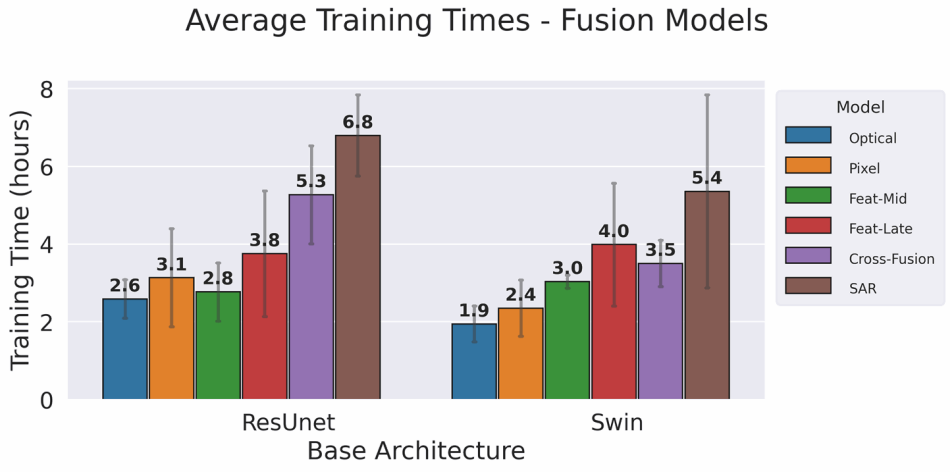
89(b): Swin-based.

Figure 89: Error maps from fusion and single-modality models using *CLOUD-DIVERSE* dataset in the same area.

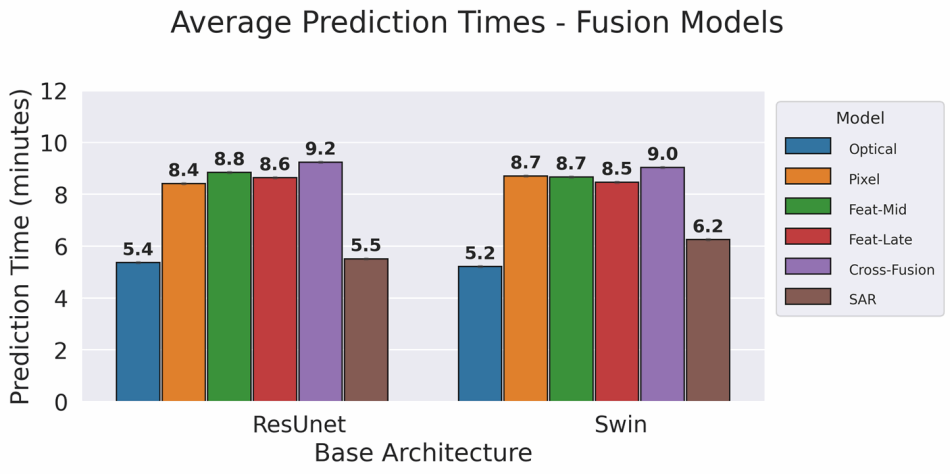
6.2.3 Models Computational Complexity Analysis

Figures 90(a) and 90(b) present the total training and prediction times (considering the five independent instances for each model), including the respective Standard Deviation (**gray** vertical lines) from ResUnet and Swin-based architectures using Optical (**blue** bars), Pixel Level Fusion (**orange** bars), Feature Level (Middle) Fusion (**green** bars), Feature Level (Late) Fusion (**red** bars), Feature Level (Late Cross-Fusion) Fusion (**purple** bars), and SAR (**brown** bars) models. Considering the total area of the BAF and the areas of the sites used in this work, we can estimate the prediction time of the whole Brazilian Amazon Forest using the Cross-Fusion models is 47 hours and the total training time of 678 days, using one single computer with GPU, considering the five models committee and the pre-training strategy. If the same optical and SAR models are used in the pre-training strategy, the total training time for all BAF can be reduced to 332 days. Appendix A describes

the computer setup used in this estimation.



90(a): Average training times.



90(b): Average Prediction Times.

Figure 90: Training and prediction average times.

We calculate the number of trainable parameters of each model. Figure 91 presents the number of trainable parameters (Millions) from ResUnet and Swin-based architectures using Optical (**blue** bars), Pixel Level Fusion (**orange** bars), Feature Level (Middle) Fusion (**green** bars), Feature Level (Late) Fusion (**red** bars), Feature Level (Late Cross-Fusion) Fusion (**purple** bars), and SAR (**brown** bars) models.

Examining the training and prediction times and the number of trainable parameters of the fusion models (Figures 90 and 91, respectively), we observed a significant difference in the number of trainable parameters between the ResUnet and Swin-based models. However, this difference did not affect their training and prediction times.

In addition, within each base architecture, we observed that the later the fusion occurs, the more parameters the models have, leading to more extended

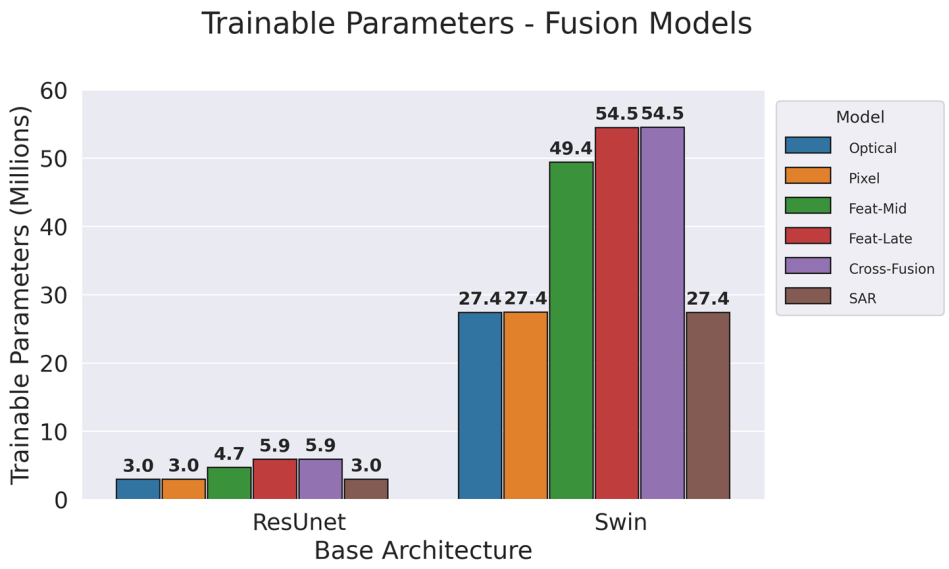


Figure 91: Trainable Parameters (Millions) comparison - Fusion Models

average training and prediction times.

7
Conclusion

We explored our central hypothesis that Deep Learning models, which combine Optical and SAR data, can assess the reliability of each data source. By doing so, these models can effectively extract the most relevant information from each type of data, leveraging their strengths to improve overall performance.

The following points summarize the main findings from the results of this research:

1. Despite the traditional usage of multi-stream as the temporal aggregation strategy, especially using the Siamese architectures, in change detection tasks, this strategy didn't improve the models' results and increased the training and prediction times. In addition to requiring less computational resources, the single-stream temporal aggregation strategy allowed using the *AVERAGE-12* dataset, which can capture deforestation throughout the year.

2. SAR models converged significantly slower than optical models. The noise inherent to SAR images could make learning difficult for these models. The higher number of available bands in the used optical images, which have ten wavelength bands, compared to two polarizations available in SAR images, could also affect the convergence.

3. The predictions from optical models when clouds are present in the first image show that these clouds are affected much less than the predictions from the same models when the cloud occurs in the last image. Utilizing the Previous Deforestation Map could minimize the importance of the first image.

4. The fusion models using the *CLOUD-FREE* dataset didn't significantly improve the deforestation detection capabilities of the models compared to the optical one. However, the optical-SAR data fusion proved helpful when the optical data can have the presence of clouds, like the *CLOUD-DIVERSE* dataset. In Site 1, the fusion only proved beneficial with the pre-training strategy.

5. Despite the significantly greater number of trainable parameters of Swin-based models, they presented lower training and prediction times than the ResUnet-based. The nature of the convolution operation, which is applied multiple times in a sliding window sequence, in opposition to the Vision Transformer concept used in a non-overlap in 4×4 patches, can explain this difference. However, both base architectures presented similar computer resource demands. Even the models that take more time to predict presented feasible values.

While this research delivered valuable findings, it is important to recognize the limitations that may have influenced the results. The following points outline the main limitations of this study:

1. In this research, we provided insights into deforestation detection in specific regions of the BAF. Although this work can be replicated in other areas, different environmental conditions, land use patterns, and deforestation dynamics can limit the applicability of the utilized method.

2. This research combined optical images taken under various cloud conditions with SAR images. To achieve multiple cloud conditions in the optical images, we selected images from different days, sometimes only a few days apart from those used as references by PRODES. As reflected in the results, this could lead to discrepancies between the reference images and the ground truth.

Despite this research's significant contributions, there are opportunities for further exploration. The following points outline the suggestions for future research:

1. This work focused on deforestation detection in the BAF biome. Future research can adapt the methodology used to identify deforestation with different environmental conditions, land use patterns, and deforestation dynamics in other geographical locations.

2. We used sensor data with identical spatial resolution and coordinate reference system, Sentinel-1 for SAR and Sentinel-2 for optical data. Future works can explore advanced fusion techniques that integrate optical and SAR data from more diverse sources, considering different spatial resolutions and data pre-processing methods. Investigating the use of deep learning architectures specifically designed for more varied multi-modal data fusion can improve the robustness of the delivered models and exploit the diversity of multiple models.

3. Our results using optical images with diverse cloud conditions showed less dependency on the cloud coverage in the first image. Future research can investigate the need for the first optical image, especially using the previous deforestation information.

4. Apply transfer learning techniques to adapt deforestation detection models trained in one region to other geographically distinct areas. Investigate domain adaptation methods to improve model generalization, reducing the need for extensive ground truth data in the new areas or multiple single-modality model training to use the pre-training strategy.

5. In this research, we don't investigate a shorter temporal window, which can be helpful in early deforestation detection efforts. Further investigations can aim to identify deforestation in shorter periods, like one month. However, the difficulty of this research comes up against the problem of difficulty in obtaining reliable data with this temporal resolution.

6. To achieve cloud diversity in the optical images, we must select images from multiple dates, trying to choose images as close as possible to the acquisition dates of the images used by PRODES. Depending on the environmental conditions, the cloud-covered images may not be available on dates close to the reference data dates. Further research can explore synthetic cloud image generation using a single cloud-free image acquired very close to the date from PRODES's images, exploring multiple cloud forms and thicknesses.

7. We used the GRD product from the Sentinel-1 satellite as SAR data. However, other information can be extracted from SAR data, such as the phase of the electromagnetic response. Investigate using this data and architectures that explore their new information can improve the deforestation detection capabilities of fusion models.

8. In this research, we identified differences in the results from the investigated sites. We proposed that this difference occurred due to the patch's availability and the cloud presence in these patches, which were mitigated through the pre-training strategy. However, other approaches to minimize this problem can be investigated, such as controlling the patches and ensuring the proportion of cloudy and cloud-free patches seen by the models during the training.

9. We proposed the pre-training strategy to minimize the convergence difference between optical and SAR models. Future works can investigate the influence of pre-training on other modality fusions, like LIDAR.

10. Future research can explore other computer vision tasks besides semantic segmentation, as this work did. Tasks like object detection or image classification can also be improved by the fusion of multiple modals when some are suffering an obstruction, like the cloud presence in optical images.

In conclusion, this study significantly advances deforestation detection using Remote Sensing and Deep Learning techniques. This research has demonstrated robust methodologies for predicting deforestation areas by integrating optical (independent of the cloud coverage) and SAR data and employing state-of-the-art models. The findings emphasize the importance of multi-modal data fusion to enhance the accuracy and reliability of deforestation detection models.

However, the research has limitations. Challenges such as the minimum deforestation area, the limited geographical location, and the data availability deserve further investigation. Addressing these limitations through continued research and collaboration will be essential for refining model robustness and applicability across diverse landscapes and socio-ecological contexts.

Future research should focus on enhancing fusion strategies, integrating additional variables, and developing new models' architectures that capture temporal and spatial dynamics in deforestation processes. Such efforts are critical for advancing the field and supporting conservation policies to safeguard global biodiversity and mitigate climate change impacts.

In summary, this thesis significantly advances our understanding of and capabilities for monitoring deforestation, refining and analyzing Deep Learning models fusing SAR and optical multitemporal data, regardless of cloud coverage in the optical images. Using SOTA architectures and proposing a training strategy for end-to-end DL models to identify new deforestation areas, fusing these data sources delivered robust results. Using the pre-training strategy, we reached the best F1-Score of 0.91, using optical images with diverse cloud conditions, which was very close to the best F1-Score of 0.93 using optical cloud-free images. This research can contribute to more effective and sustainable environmental conservation efforts in the future and expand the boundaries of scientific knowledge in the field.

References

1 INPE. **Metodologia utilizada nos sistemas PRODES e DETER**. http://mtc-m21d.sid.inpe.br/ibi/8JMKD3MGP3W34T/47GAF6S, 2022.

2 ZHANG, Z.; LIU, Q. ; WANG, Y.. **Road extraction by deep residual u-net**. IEEE Geoscience and Remote Sensing Letters, 15:749–753, 5 2018.

3 DOSOVITSKIY, A.; BEYER, L.; KOLESNIKOV, A.; WEISSENBORN, D.; ZHAI, X.; UNTERTHINER, T.; DEHGHANI, M.; MINDERER, M.; HEIGOLD, G.; GELLY, S.; USZKOREIT, J. ; HOULSBY, N.. **An image is worth 16x16 words: Transformers for image recognition at scale**. ArXiv, abs/2010.1, 10 2020.

4 VASWANI, A.; SHAZEER, N.; PARMAR, N.; USZKOREIT, J.; JONES, L.; GOMEZ, A. N.; KAISER, L. ; POLOSUKHIN, I.. **Attention is all you need**. ArXiv, 6 2017.

5 LIU, Z.; LIN, Y.; CAO, Y.; HU, H.; WEI, Y.; ZHANG, Z.; LIN, S. ; GUO, B.. **Swin transformer: Hierarchical vision transformer using shifted windows**. In: 2021 IEEE/CVF INTERNATIONAL CONFERENCE ON COMPUTER VISION (ICCV), p. 9992–10002. IEEE, 10 2021.

6 CAO, H.; WANG, Y.; CHEN, J.; JIANG, D.; ZHANG, X.; TIAN, Q. ; WANG, M.. **Swin-unet: Unet-like pure transformer for medical image segmentation**. ArXiv, 5 2021.

7 LI, J.; HONG, D.; GAO, L.; YAO, J.; ZHENG, K.; ZHANG, B. ; CHANUSSOT, J.. **Deep learning in multimodal remote sensing data fusion: A comprehensive review**. International Journal of Applied Earth Observation and Geoinformation, 112:102926, 8 2022.

8 GHASSEMIAN, H.. **A review of remote sensing image fusion methods**. Information Fusion, 32:75–89, 2016.

9 ASNER, G. P.. **Cloud cover in landsat observations of the brazilian amazon**. International Journal of Remote Sensing, 22:3855–3862, 1 2001.

10 ORTEGA ADARME, M.; QUEIROZ FEITOSA, R.; NIGRI HAPP, P.; APARECIDO DE ALMEIDA, C. ; RODRIGUES GOMES, A.. **Evaluation**

of deep learning techniques for deforestation detection in the brazilian amazon and cerrado biomes from remote sensing imagery. Remote Sensing, 12(6), 2020.

11 XIANG, J.; XING, Y.; WEI, W.; YAN, E.; JIANG, J. ; MO, D.. **Dynamic detection of forest change in hunan province based on sentinel-2 images and deep learning**. Remote Sensing, 15, 2 2023.

12 DE BEM, P.; DE CARVALHO JUNIOR, O.; GUIMARÃES, R. F. ; GOMES, R. T.. **Change detection of deforestation in the brazilian amazon using landsat data and convolutional neural networks**. Remote Sensing, 12:901, 3 2020.

13 ORTEGA, M. X.; FEITOSA, R. Q.; BERMUDEZ, J. D.; HAPP, P. N. ; ALMEIDA, C. A. D.. **Comparison of optical and sar data for deforestation mapping in the amazon rainforest with fully convolutional networks**. In: 2021 IEEE INTERNATIONAL GEOSCIENCE AND REMOTE SENSING SYMPOSIUM IGARSS, p. 3769–3772. IEEE, 7 2021.

14 HONG, D.; GAO, L.; YOKOYA, N.; YAO, J.; CHANUSSOT, J.; DU, Q. ; ZHANG, B.. **More diverse means better: Multimodal deep learning meets remote-sensing imagery classification**. IEEE Transactions on Geoscience and Remote Sensing, 59:4340–4354, 5 2021.

15 ROSA, L. E. C. L.; OLIVEIRA, D. A. B. ; FEITOSA, R. Q.. **Investigating fusion strategies on encoder-decoder networks for crop segmentation using sar and optical image sequences**. In: 2021 IEEE INTERNATIONAL GEOSCIENCE AND REMOTE SENSING SYMPOSIUM IGARSS, p. 2405–2408. IEEE, 7 2021.

16 MARTINEZ, J. A. C.; ADARME, M. X. O.; TURNES, J. N.; COSTA, G. A. O. P.; DE ALMEIDA, C. A. ; FEITOSA, R. Q.. **A comparison of cloud removal methods for deforestation monitoring in amazon rainforest**. The International Archives of the Photogrammetry, Remote Sensing and Spatial Information Sciences, XLIII-B3-2022:665–671, 2022.

17 LI, J.; LI, C.; XU, W.; FENG, H.; ZHAO, F.; LONG, H.; MENG, Y.; CHEN, W.; YANG, H. ; YANG, G.. **Fusion of optical and sar images based on deep learning to reconstruct vegetation ndvi time series in cloud-prone regions**. International Journal of Applied Earth Observation and Geoinformation, 112:102818, 2022.

18 FERRARI, F.; FERREIRA, M. P. ; FEITOSA, R. Q.. **Fusing sentinel-1 and sentinel-2 images with transformer-based network for deforestation detection in the brazilian amazon under diverse cloud conditions**. In: ISPRS ANNALS OF THE PHOTOGRAMMETRY, REMOTE SENSING AND SPATIAL INFORMATION SCIENCES, volumen 10, p. 999–1006. Copernicus Publications, 12 2023.

19 ESA. **Sentinel-2 user handbook**. Technical report, ESA, 7 2015.

20 LILLESAND, T. M.; KIEFER, R. W. ; CHIPMAN, J. W.. **Remote Sensing and Image Interpretation**. John Wiley & Sons, 7th edition, 2015.

21 INPE. **TerraBrasilis**. `https://terrabrasilis.dpi.inpe.br/`, 4 2024.

22 CAMPBELL, J. B.; WYNNE, R. H. ; THOMAS, V. A.. **Introduction To Remote Sensing**. The Guilford Press, 6th edition, 2023.

23 RICHARDS, J. A.. **Remote Sensing Digital Image Analysis**. Springer International Publishing, 2022.

24 UNION OF CONCERNED SCIENTISTS. **UCS Satellite Database**. `https://www.ucsusa.org/resources/satellite-database`, 01 2023.

25 MA, L.; LIU, Y.; ZHANG, X.; YE, Y.; YIN, G. ; JOHNSON, B. A.. **Deep learning in remote sensing applications: A meta-analysis and review**. ISPRS Journal of Photogrammetry and Remote Sensing, 152:166–177, 6 2019.

26 STRAND, J.; SOARES-FILHO, B.; COSTA, M. H.; OLIVEIRA, U.; RIBEIRO, S. C.; PIRES, G. F.; OLIVEIRA, A.; RAJAO, R.; MAY, P.; VAN DER HOFF, R. ; OTHERS. **Spatially explicit valuation of the brazilian amazon forest's ecosystem services**. Nature Sustainability, 1(11):657–664, 2018.

27 INPE. **Monitoramento do desmatamento da floresta amazônica brasileira por satélite**. `http://www.obt.inpe.br/OBT/assuntos/programas/amazonia/prodes`, 2022. Acesso em: Setembro de 2022.

28 THIEL, C.; DREZET, P.; WEISE, C.; QUEGAN, S. ; SCHMULLIUS, C.. **Radar remote sensing for the delineation of forest cover maps and the detection of deforestation**. Forestry, 79:589–597, 12 2006.

29 Bossler, J. D.; Campbell, J. B.; McMaster, R. B. ; Rizos, C., editors. **Manual of Geospatial Science and Technology**. CRC Press, 3 2010.

30 FRIEDL, M.; MCIVER, D.; HODGES, J.; ZHANG, X.; MUCHONEY, D.; STRAHLER, A.; WOODCOCK, C.; GOPAL, S.; SCHNEIDER, A.; COOPER, A.; BACCINI, A.; GAO, F. ; SCHAAF, C.. **Global land cover mapping from modis: algorithms and early results**. Remote Sensing of Environment, 83:287–302, 11 2002.

31 JAT, M. K.; GARG, P. ; KHARE, D.. **Monitoring and modelling of urban sprawl using remote sensing and gis techniques**. International Journal of Applied Earth Observation and Geoinformation, 10:26–43, 2 2008.

32 RHEE, J.; IM, J. ; CARBONE, G. J.. **Monitoring agricultural drought for arid and humid regions using multi-sensor remote sensing data**. Remote Sensing of Environment, 114:2875–2887, 12 2010.

33 CHANG, A.; EO, Y.; KIM, S.; KIM, Y. ; KIM, Y.. **Canopy-cover thematic-map generation for military map products using remote sensing data in inaccessible areas**. Landscape and Ecological Engineering, 7:263–274, 7 2011.

34 GOETZ, S.; DUBAYAH, R.. **Advances in remote sensing technology and implications for measuring and monitoring forest carbon stocks and change**. Carbon Management, 2:231–244, 6 2011.

35 ROSEN, P.; HENSLEY, S.; JOUGHIN, I.; LI, F.; MADSEN, S.; RODRIGUEZ, E. ; GOLDSTEIN, R.. **Synthetic aperture radar interferometry**. Proceedings of the IEEE, 88:333–382, 3 2000.

36 RIGNOT, E.; MOUGINOT, J. ; SCHEUCHL, B.. **Ice flow of the antarctic ice sheet**. Science, 333:1427–1430, 9 2011.

37 SHIMADA, M.; ITOH, T.; MOTOOKA, T.; WATANABE, M.; SHIRAISHI, T.; THAPA, R. ; LUCAS, R.. **New global forest/non-forest maps from alos palsar data (2007–2010)**. Remote Sensing of Environment, 155:13–31, 12 2014.

38 TOAN, T. L.; QUEGAN, S.; DAVIDSON, M.; BALZTER, H.; PAILLOU, P.; PAPATHANASSIOU, K.; PLUMMER, S.; ROCCA, F.; SAATCHI, S.; SHUGART, H. ; ULANDER, L.. **The biomass mission: Mapping global forest biomass to better understand the terrestrial carbon cycle**. Remote Sensing of Environment, 115:2850–2860, 11 2011.

39 PALOSCIA, S.; PETTINATO, S.; SANTI, E.; NOTARNICOLA, C.; PASOLLI, L. ; REPPUCCI, A.. **Soil moisture mapping using sentinel-1 images:**

Algorithm and preliminary validation. Remote Sensing of Environment, 134:234–248, 7 2013.

40 PAEK, S. W.; BALASUBRAMANIAN, S.; KIM, S. ; DE WECK, O.. **Small-satellite synthetic aperture radar for continuous global biospheric monitoring: A review**. Remote Sensing, 12:2546, 8 2020.

41 TORRES, R.; SNOEIJ, P.; GEUDTNER, D.; BIBBY, D.; DAVIDSON, M.; ATTEMA, E.; POTIN, P.; ROMMEN, B.; FLOURY, N.; BROWN, M.; TRAVER, I. N.; DEGHAYE, P.; DUESMANN, B.; ROSICH, B.; MIRANDA, N.; BRUNO, C.; L'ABBATE, M.; CROCI, R.; PIETROPAOLO, A.; HUCHLER, M. ; ROSTAN, F.. **Gmes sentinel-1 mission**. Remote Sensing of Environment, 120:9–24, 5 2012.

42 ASSIS, L. F. F. G.; FERREIRA, K. R.; VINHAS, L.; MAURANO, L.; ALMEIDA, C.; CARVALHO, A.; RODRIGUES, J.; MACIEL, A. ; CAMARGO, C.. **Terrabrasilis: A spatial data analytics infrastructure for large-scale thematic mapping**. ISPRS International Journal of Geo-Information, 8:513, 11 2019.

43 MAURANO, L. E. P.; ESCADA, M. I. S. ; RENNO, C. D.. **Padrões espaciais de desmatamento e a estimativa da exatidão dos mapas do prodes para amazônia legal brasileira**. Ciência Florestal, 29:1763–1775, 12 2019.

44 PARENTE, L.; NOGUEIRA, S.; BAUMANN, L.; ALMEIDA, C.; MAURANO, L.; AFFONSO, A. G. ; FERREIRA, L.. **Quality assessment of the prodes cerrado deforestation data**. Remote Sensing Applications: Society and Environment, 21, 1 2021.

45 WALKER, W. S.; STICKLER, C. M.; KELLNDORFER, J. M.; KIRSCH, K. M. ; NEPSTAD, D. C.. **Large-area classification and mapping of forest and land cover in the brazilian amazon: A comparative analysis of alos/palsar and landsat data sources**. IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing, 3:594–604, 12 2010.

46 CABRAL, A. I.; SAITO, C.; PEREIRA, H. ; LAQUES, A. E.. **Deforestation pattern dynamics in protected areas of the brazilian legal amazon using remote sensing data**. Applied Geography, 100:101–115, 11 2018.

47 VAN MARLE, M. J. E.; VAN DER WERF, G. R.; DE JEU, R. A. M. ; LIU, Y. Y.. **Annual south american forest loss estimates based on passive microwave remote sensing (1990–2010)**. Biogeosciences, 13:609–624, 2 2016.

48 LECUN, Y.; BOSER, B. E.; DENKER, J. S.; HENDERSON, D.; HOWARD, R. E.; HUBBARD, W. E. ; JACKEL, L. D.. **Backpropagation applied to handwritten zip code recognition**. Neural Computation, 1:541–551, 1989.

49 GOODFELLOW, I.; BENGIO, Y. ; COURVILLE, A.. **Deep Learning**. MIT Press, 2016. `http://www.deeplearningbook.org`.

50 RONNEBERGER, O.; FISCHER, P. ; BROX, T.. **U-net: Convolutional networks for biomedical image segmentation**. Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics), 9351:234–241, 2015. cited By 30698.

51 HE, K.; ZHANG, X.; REN, S. ; SUN, J.. **Deep residual learning for image recognition**. In: 2016 IEEE CONFERENCE ON COMPUTER VISION AND PATTERN RECOGNITION (CVPR), volumen 2016-December, p. 770–778. IEEE, 6 2016.

52 ADARME, M. O.; FEITOSA, R. Q.; HAPP, P. N.; ALMEIDA, C. A. D. ; GOMES, A. R.. **Evaluation of deep learning techniques for deforestation detection in the brazilian amazon and cerrado biomes from remote sensing imagery**. Remote Sensing, 12:910, 3 2020.

53 GALLWEY, J.; ROBIATI, C.; COGGAN, J.; VOGT, D. ; EYRE, M.. **A sentinel-2 based multispectral convolutional neural network for detecting artisanal small-scale mining in ghana: Applying deep learning to shallow mining**. Remote Sensing of Environment, 248:111970, 10 2020.

54 ZHAO, F.; SUN, R.; ZHONG, L.; MENG, R.; HUANG, C.; ZENG, X.; WANG, M.; LI, Y. ; WANG, Z.. **Monthly mapping of forest harvesting using dense time series sentinel-1 sar imagery and deep learning**. Remote Sensing of Environment, 269:112822, 2 2022.

55 HARKAT, H.; NASCIMENTO, J. M. P. ; BERNARDINO, A.. **Fire segmentation using a deeplabv3+ architecture**. In: Notarnicola, C.; Bovenga,

F.; Bruzzone, L.; Bovolo, F.; Benediktsson, J. A.; Santi, E. ; Pierdicca, N., editors, IMAGE AND SIGNAL PROCESSING FOR REMOTE SENSING XXVI, p. 20. SPIE, 9 2020.

56 ORTEGA, M. X.; BERMUDEZ, J. D.; HAPP, P. N.; GOMES, A. ; FEITOSA, R. Q.. **Evaluation of deep learning techniques for deforestation detection in the amazon forest**. ISPRS Annals of the Photogrammetry, Remote Sensing and Spatial Information Sciences, IV-2/W7:121–128, 9 2019.

57 TORRES, D. L.; TURNES, J. N.; VEGA, P. J. S.; FEITOSA, R. Q.; SILVA, D. E.; JUNIOR, J. M. ; ALMEIDA, C.. **Deforestation detection with fully convolutional networks in the amazon forest from landsat-8 and sentinel-2 images**. Remote Sensing, 13, 12 2021.

58 SEYDI, S. T.; SAEIDI, V.; KALANTAR, B.; UEDA, N. ; HALIN, A. A.. **Fire-net: A deep learning framework for active forest fire detection**. Journal of Sensors, 2022:1–14, 2 2022.

59 HENDRYCKS, D.; GIMPEL, K.. **Gaussian error linear units (gelus)**. ArXiv, 6 2016.

60 SINGH, A.. **Review article digital change detection techniques using remotely-sensed data**. International Journal of Remote Sensing, 10(6):989–1003, 1989.

61 KHELIFI, L.; MIGNOTTE, M.. **Deep learning for change detection in remote sensing images: Comprehensive review and meta-analysis**. IEEE Access, 8:126385–126400, 2020.

62 SHI, W.; ZHANG, M.; ZHANG, R.; CHEN, S. ; ZHAN, Z.. **Change detection based on artificial intelligence: State-of-the-art and challenges**. Remote Sensing, 12(10), 2020.

63 JIANG, H.; PENG, M.; ZHONG, Y.; XIE, H.; HAO, Z.; LIN, J.; MA, X. ; HU, X.. **A survey on deep learning-based change detection from high-resolution remote sensing images**. Remote Sensing, 14(7), 2022.

64 XIN ZHANG, Y. Z.; LUO, J.. **Deep learning for processing and analysis of remote sensing big data: a technical review**. Big Earth Data, 6(4):527–560, 2022.

65 SCHMITT, M.; TUPIN, F. ; ZHU, X. X.. **Fusion of sar and optical remote sensing data — challenges and recent trends**. In: 2017 IEEE

INTERNATIONAL GEOSCIENCE AND REMOTE SENSING SYMPOSIUM (IGARSS), p. 5458–5461. IEEE, 7 2017.

66 MURA, M. D.; PRASAD, S.; PACIFICI, F.; GAMBA, P.; CHANUSSOT, J. ; BENEDIKTSSON, J. A.. **Challenges and opportunities of multimodality and data fusion in remote sensing**. Proceedings of the IEEE, 103:1585–1601, 9 2015.

67 LV, Z.; HUANG, H.; GAO, L.; BENEDIKTSSON, J. A.; ZHAO, M. ; SHI, C.. **Simple multiscale unet for change detection with heterogeneous remote sensing images**. IEEE Geoscience and Remote Sensing Letters, 19:1–5, 2022.

68 LAN, L.; WU, D. ; CHI, M.. **Multi-temporal change detection based on deep semantic segmentation networks**. In: 2019 10TH INTERNATIONAL WORKSHOP ON THE ANALYSIS OF MULTITEMPORAL REMOTE SENSING IMAGES (MULTITEMP), p. 1–4, 2019.

69 PENG, D.; ZHANG, Y. ; GUAN, H.. **End-to-end change detection for high resolution satellite images using improved unet++**. Remote Sensing, 11(11), 2019.

70 JATURAPITPORNCHAI, R.; MATSUOKA, M.; KANEMOTO, N.; KUZUOKA, S.; ITO, R. ; NAKAMURA, R.. **Newly built construction detection in sar images using deep learning**. Remote Sensing, 11(12), 2019.

71 GAO, Y.; GAO, F.; DONG, J. ; WANG, S.. **Change detection from synthetic aperture radar images based on channel weighting-based deep cascade network**. IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing, 12(11):4517–4529, 2019.

72 WANG, Q.; YUAN, Z.; DU, Q. ; LI, X.. **Getnet: A general end-to-end 2-d cnn framework for hyperspectral image change detection**. IEEE Transactions on Geoscience and Remote Sensing, 57(1):3–13, 2019.

73 ZHENG, Z.; WAN, Y.; ZHANG, Y.; XIANG, S.; PENG, D. ; ZHANG, B.. **Clnet: Cross-layer convolutional neural network for change detection in optical remote sensing imagery**. ISPRS Journal of Photogrammetry and Remote Sensing, 175:247–267, 2021.

74 LIU, R.; JIANG, D.; ZHANG, L. ; ZHANG, Z.. **Deep depthwise separable convolutional network for change detection in optical aerial**

images. IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing, 13:1109–1118, 2020.

75 LI, X.; HE, M.; LI, H. ; SHEN, H.. **A combined loss-based multiscale fully convolutional network for high-resolution remote sensing image change detection**. IEEE Geoscience and Remote Sensing Letters, 19:1–5, 2022.

76 YANG, L.; CHEN, Y.; SONG, S.; LI, F. ; HUANG, G.. **Deep siamese networks based change detection with remote sensing images**. Remote Sensing, 13(17), 2021.

77 ARABI, M. E. A.; KAROUI, M. S. ; DJERRIRI, K.. **Optical remote sensing change detection through deep siamese network**. In: IGARSS 2018 - 2018 IEEE INTERNATIONAL GEOSCIENCE AND REMOTE SENSING SYMPOSIUM, p. 5041–5044, 2018.

78 CHEN, P.; GUO, L.; ZHANG, X.; QIN, K.; MA, W. ; JIAO, L.. **Attention-guided siamese fusion network for change detection of remote sensing images**. Remote Sensing, 13(22), 2021.

79 YIN, H.; MA, C.; WENG, L.; XIA, M. ; LIN, H.. **Bitemporal remote sensing image change detection network based on siamese-attention feedback architecture**. Remote Sensing, 15(17), 2023.

80 ZHANG, X.; HE, L.; QIN, K.; DANG, Q.; SI, H.; TANG, X. ; JIAO, L.. **Smd-net: Siamese multi-scale difference-enhancement network for change detection in remote sensing**. Remote Sensing, 14(7), 2022.

81 ZHU, Q.; GUO, X.; DENG, W.; SHI, S.; GUAN, Q.; ZHONG, Y.; ZHANG, L. ; LI, D.. **Land-use/land-cover change detection based on a siamese global learning framework for high spatial resolution remote sensing imagery**. ISPRS Journal of Photogrammetry and Remote Sensing, 184:63–78, 2022.

82 ZHANG, M.; LIU, Z.; FENG, J.; LIU, L. ; JIAO, L.. **Remote sensing image change detection based on deep multi-scale multi-attention siamese transformer network**. Remote Sensing, 15(3), 2023.

83 FANG, S.; LI, K.; SHAO, J. ; LI, Z.. **Snunet-cd: A densely connected siamese network for change detection of vhr images**. IEEE Geoscience and Remote Sensing Letters, 19:1–5, 2022.

84 ZHANG, M.; XU, G.; CHEN, K.; YAN, M. ; SUN, X.. **Triplet-based semantic relation learning for aerial remote sensing image change detection**. IEEE Geoscience and Remote Sensing Letters, 16(2):266–270, 2019.

85 ZHANG, C.; YUE, P.; TAPETE, D.; JIANG, L.; SHANGGUAN, B.; HUANG, L. ; LIU, G.. **A deeply supervised image fusion network for change detection in high resolution bi-temporal remote sensing images**. ISPRS Journal of Photogrammetry and Remote Sensing, 166:183–200, 2020.

86 LIU, Y.; PANG, C.; ZHAN, Z.; ZHANG, X. ; YANG, X.. **Building change detection for remote sensing images using a dual-task constrained deep siamese convolutional network model**. IEEE Geoscience and Remote Sensing Letters, 18(5):811–815, 2021.

87 ZHANG, M.; SHI, W.. **A feature difference convolutional neural network-based change detection method**. IEEE Transactions on Geoscience and Remote Sensing, 58(10):7232–7246, 2020.

88 JIANG, H.; HU, X.; LI, K.; ZHANG, J.; GONG, J. ; ZHANG, M.. **Pga-siamnet: Pyramid feature-based attention-guided siamese network for remote sensing orthoimagery building change detection**. Remote Sensing, 12(3), 2020.

89 SONG, A.; CHOI, J.; HAN, Y. ; KIM, Y.. **Change detection in hyperspectral images using recurrent 3d fully convolutional networks**. Remote Sensing, 10(11), 2018.

90 JI, S.; SHEN, Y.; LU, M. ; ZHANG, Y.. **Building instance change detection from large-scale aerial images using convolutional neural networks and simulated samples**. Remote Sensing, 11(11), 2019.

91 ZHENG, Z.; ZHONG, Y.; WANG, J.; MA, A. ; ZHANG, L.. **Building damage assessment for rapid disaster response with a deep object-based semantic change detection framework: From natural disasters to man-made disasters**. Remote Sensing of Environment, 265:112636, 2021.

92 LI, X.; YUAN, Z. ; WANG, Q.. **Unsupervised deep noise modeling for hyperspectral image change detection**. Remote Sensing, 11(3), 2019.

93 SEYDI, S. T.; HASANLOU, M. ; AMANI, M.. **A new end-to-end multi-dimensional cnn framework for land cover/land use change detection in multi-source remote sensing datasets**. Remote Sensing, 12(12), 2020.

94 LIU, J.; CHEN, K.; XU, G.; SUN, X.; YAN, M.; DIAO, W. ; HAN, H.. **Convolutional neural network-based transfer learning for optical aerial images change detection**. IEEE Geoscience and Remote Sensing Letters, 17(1):127–131, 2020.

95 HADSELL, R.; CHOPRA, S. ; LECUN, Y.. **Dimensionality reduction by learning an invariant mapping**. In: 2006 IEEE COMPUTER SOCIETY CONFERENCE ON COMPUTER VISION AND PATTERN RECOGNITION (CVPR'06), volumen 2, p. 1735–1742, 2006.

96 ZHAN, Y.; FU, K.; YAN, M.; SUN, X.; WANG, H. ; QIU, X.. **Change detection based on deep siamese convolutional network for optical aerial images**. IEEE Geoscience and Remote Sensing Letters, 14(10):1845–1849, 2017.

97 FANG, B.; PAN, L. ; KOU, R.. **Dual learning-based siamese framework for change detection using bi-temporal vhr optical remote sensing images**. Remote Sensing, 11(11), 2019.

98 CHEN, J.; YUAN, Z.; PENG, J.; CHEN, L.; HUANG, H.; ZHU, J.; LIU, Y. ; LI, H.. **Dasnet: Dual attentive fully convolutional siamese networks for change detection in high-resolution satellite images**. IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing, 14:1194–1206, 2021.

99 CAYE DAUDT, R.; LE SAUX, B.; BOULCH, A. ; GOUSSEAU, Y.. **Multi-task learning for large-scale semantic change detection**. Computer Vision and Image Understanding, 187:102783, 2019.

100 ZHONG, L.; HU, L. ; ZHOU, H.. **Deep learning based multi-temporal crop classification**. Remote Sensing of Environment, 221:430–443, 2 2019.

101 MOU, L.; GHAMISI, P. ; ZHU, X. X.. **Deep recurrent neural networks for hyperspectral image classification**. IEEE Transactions on Geoscience and Remote Sensing, 55(7):3639–3655, 2017.

102 WANG, Q.; LIU, S.; CHANUSSOT, J. ; LI, X.. **Scene classification with recurrent attention of vhr remote sensing images**. IEEE Transactions on Geoscience and Remote Sensing, 57(2):1155–1167, 2019.

103 HOCHREITER, S.; SCHMIDHUBER, J.. **Long Short-Term Memory**. Neural Computation, 9(8):1735–1780, 11 1997.

104 LYU, H.; LU, H. ; MOU, L.. **Learning a transferable change rule from a recurrent neural network for land cover change detection**. Remote Sensing, 8(6), 2016.

105 PAPADOMANOLAKI, M.; VERMA, S.; VAKALOPOULOU, M.; GUPTA, S. ; KARANTZALOS, K.. **Detecting urban changes with recurrent neural networks from multitemporal sentinel-2 data**. In: IGARSS 2019 - 2019 IEEE INTERNATIONAL GEOSCIENCE AND REMOTE SENSING SYMPOSIUM, p. 214–217, 2019.

106 SEFRIN, O.; RIESE, F. M. ; KELLER, S.. **Deep learning for land cover change detection**. Remote Sensing, 13(1), 2021.

107 MOU, L.; BRUZZONE, L. ; ZHU, X. X.. **Learning spectral-spatial-temporal features via a recurrent convolutional neural network for change detection in multispectral imagery**. IEEE Transactions on Geoscience and Remote Sensing, 57(2):924–935, 2019.

108 CHEN, H.; WU, C.; DU, B.; ZHANG, L. ; WANG, L.. **Change detection in multisource vhr images via deep siamese convolutional multiple-layers recurrent neural network**. IEEE Transactions on Geoscience and Remote Sensing, 58(4):2848–2864, 2020.

109 CHEN, H.; QI, Z. ; SHI, Z.. **Remote sensing image change detection with transformers**. IEEE Transactions on Geoscience and Remote Sensing, 60:1–14, 2022.

110 ZHANG, C.; WANG, L.; CHENG, S. ; LI, Y.. **Swinsunet: Pure transformer network for remote sensing image change detection**. IEEE Transactions on Geoscience and Remote Sensing, 60:1–13, 2022.

111 HU, J.; SHEN, L.; ALBANIE, S.; SUN, G. ; WU, E.. **Squeeze-and-excitation networks**. IEEE Transactions on Pattern Analysis and Machine Intelligence, 42(8):2011–2023, 2020.

112 SHI, Q.; LIU, M.; LI, S.; LIU, X.; WANG, F. ; ZHANG, L.. **A deeply supervised attention metric-based network and an open aerial image**

dataset for remote sensing change detection. IEEE Transactions on Geoscience and Remote Sensing, 60:1–16, 2022.

113 SONG, L.; XIA, M.; JIN, J.; QIAN, M. ; ZHANG, Y.. **Suacdnet: Attentional change detection network based on siamese u-shaped structure**. International Journal of Applied Earth Observation and Geoinformation, 105:102597, 2021.

114 WANG, D.; CHEN, X.; JIANG, M.; DU, S.; XU, B. ; WANG, J.. **Adsnet:an attention-based deeply supervised network for remote sensing image change detection**. International Journal of Applied Earth Observation and Geoinformation, 101:102348, 2021.

115 JAVED, A.; KIM, T.; LEE, C.; OH, J. ; HAN, Y.. **Deep learning-based detection of urban forest cover change along with overall urban changes using very-high-resolution satellite images**. Remote Sensing, 15, 9 2023.

116 WYNIAWSKYJ, N. S.; NAPIORKOWSKA, M.; PETIT, D.; PODDER, P. ; MARTI, P.. **Forest monitoring in guatemala using satellite imagery and deep learning**. In: IGARSS 2019 - 2019 IEEE INTERNATIONAL GEOSCIENCE AND REMOTE SENSING SYMPOSIUM, p. 6598–6601. IEEE, 7 2019.

117 ISAIENKOV, K.; YUSHCHUK, M.; KHRAMTSOV, V. ; SELIVERSTOV, O.. **Deep learning for regular change detection in ukrainian forest ecosystem with sentinel-2**. IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing, 14:364–376, 2021.

118 MARTINEZ, J. A. C.; DA COSTA, G. A. P.; MESSIAS, C. G.; DE SOUZA SOLER, L.; DE ALMEIDA, C. A. ; FEITOSA, R. Q.. **Enhancing deforestation monitoring in the brazilian amazon: A semi-automatic approach leveraging uncertainty estimation**. ISPRS Journal of Photogrammetry and Remote Sensing, 210:110–127, 4 2024.

119 PRUDENTE, V. H. R.; SANCHES, I. D.; ADAMI, M.; SKAKUN, S.; OLDONI, L. V.; XAUD, H. A. M.; XAUD, M. R. ; ZHANG, Y.. **Sar data for land use land cover classification in a tropical region with frequent cloud cover**. In: IGARSS 2020 - 2020 IEEE INTERNATIONAL GEOSCIENCE AND REMOTE SENSING SYMPOSIUM, p. 4100–4103, 2020.

120 ADARME, M. O.; PRIETO, J. D.; FEITOSA, R. Q. ; ALMEIDA, C. A. D.. **Improving deforestation detection on tropical rainforests using sentinel-1 data and convolutional neural networks**. Remote Sensing, 14:3290, 7 2022.

121 KUZU, R. S.; ANTROPOV, O.; MOLINIER, M.; DUMITRU, C. O.; SAHA, S. ; ZHU, X. X.. **Forest disturbance detection via self-supervised and transfer learning with sentinel-12 images**. IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing, 17:4751–4767, 2024.

122 WEI, Y.; YUAN, Q.; SHEN, H. ; ZHANG, L.. **Boosting the accuracy of multispectral image pansharpening by learning a deep residual network**. IEEE Geoscience and Remote Sensing Letters, 14(10):1795–1799, 2017.

123 HUANG, W.; XIAO, L.; WEI, Z.; LIU, H. ; TANG, S.. **A new pansharpening method with deep neural networks**. IEEE Geoscience and Remote Sensing Letters, 12(5):1037–1041, 2015.

124 DENG, L.-J.; VIVONE, G.; JIN, C. ; CHANUSSOT, J.. **Detail injection-based deep convolutional neural networks for pansharpening**. IEEE Transactions on Geoscience and Remote Sensing, 59(8):6995–7010, 2021.

125 FU, X.; WANG, W.; HUANG, Y.; DING, X. ; PAISLEY, J.. **Deep multiscale detail networks for multiband spectral image sharpening**. IEEE Transactions on Neural Networks and Learning Systems, 32(5):2090–2104, 2021.

126 AZARANG, A.; MANOOCHEHRI, H. E. ; KEHTARNAVAZ, N.. **Convolutional autoencoder-based multispectral image fusion**. IEEE Access, 7:35673–35683, 2019.

127 ZHENG, Y.; LI, J.; LI, Y.; GUO, J.; WU, X. ; CHANUSSOT, J.. **Hyperspectral pansharpening using deep prior and dual attention residual network**. IEEE Transactions on Geoscience and Remote Sensing, 58(11):8059–8076, 2020.

128 JIANG, M.; SHEN, H.; LI, J.; YUAN, Q. ; ZHANG, L.. **A differential information residual convolutional neural network for pansharpening**. ISPRS Journal of Photogrammetry and Remote Sensing, 163:257–271, 2020.

129 GARGIULO, M.; MAZZA, A.; GAETANO, R.; RUELLO, G. ; SCARPA, G.. **Fast super-resolution of 20 m sentinel-2 bands using convolutional neural networks**. Remote Sensing, 11(22), 2019.

130 GOODFELLOW, I. J.; POUGET-ABADIE, J.; MIRZA, M.; XU, B.; WARDE-FARLEY, D.; OZAIR, S.; COURVILLE, A. ; BENGIO, Y.. **Generative adversarial nets**. volumen 3, p. 2672 – 2680, 2014. Cited by: 37477.

131 LIU, Q.; ZHOU, H.; XU, Q.; LIU, X. ; WANG, Y.. **Psgan: A generative adversarial network for remote sensing image pan-sharpening**. IEEE Transactions on Geoscience and Remote Sensing, 59(12):10227–10242, 2021.

132 OZCELIK, F.; ALGANCI, U.; SERTEL, E. ; UNAL, G.. **Rethinking cnn-based pansharpening: Guided colorization of panchromatic images via gans**. IEEE Transactions on Geoscience and Remote Sensing, 59(4):3486–3501, 2021.

133 SHAO, Z.; LU, Z.; RAN, M.; FANG, L.; ZHOU, J. ; ZHANG, Y.. **Residual encoder–decoder conditional generative adversarial network for pansharpening**. IEEE Geoscience and Remote Sensing Letters, 17(9):1573–1577, 2020.

134 XIE, W.; CUI, Y.; LI, Y.; LEI, J.; DU, Q. ; LI, J.. **Hpgan: Hyperspectral pansharpening using 3-d generative adversarial networks**. IEEE Transactions on Geoscience and Remote Sensing, 59(1):463–477, 2021.

135 DONG, W.; HOU, S.; XIAO, S.; QU, J.; DU, Q. ; LI, Y.. **Generative dual-adversarial network with spectral fidelity and spatial enhancement for hyperspectral pansharpening**. IEEE Transactions on Neural Networks and Learning Systems, 33(12):7303–7317, 2022.

136 QU, Y.; BAGHBADERANI, R. K.; QI, H. ; KWAN, C.. **Unsupervised pansharpening based on self-attention mechanism**. IEEE Transactions on Geoscience and Remote Sensing, 59(4):3192–3208, 2021.

137 ZHOU, H.; LIU, Q. ; WANG, Y.. **Pgman: An unsupervised generative multiadversarial network for pansharpening**. IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing, 14:6316–6327, 2021.

138 CHEN, Y.; BRUZZONE, L.. **Self-supervised sar-optical data fusion of sentinel-1/-2 images**. IEEE Transactions on Geoscience and Remote Sensing, 60:1–11, 2022.

139 HONG, D.; YOKOYA, N.; XIA, G.-S.; CHANUSSOT, J. ; ZHU, X. X.. **X-modalnet: A semi-supervised deep cross-modal network for classification of remote sensing data**. ISPRS Journal of Photogrammetry and Remote Sensing, 167:12–23, 2020.

140 WU, X.; HONG, D. ; CHANUSSOT, J.. **Convolutional neural networks for multimodal remote sensing data classification**. IEEE Transactions on Geoscience and Remote Sensing, 60:1–10, 2022.

141 LI, Q.; WONG, F. K. K. ; FUNG, T.. **Mapping multi-layered mangroves from multispectral, hyperspectral, and lidar data**. Remote Sensing of Environment, 258:112403, 2021.

142 ZHANG, L.; XIA, J.. **Flood detection using multiple chinese satellite datasets during 2020 china summer floods**. Remote Sensing, 14(1), 2022.

143 YAO, J.; ZHANG, B.; LI, C.; HONG, D. ; CHANUSSOT, J.. **Extended vision transformer (exvit) for land use and land cover classification: A multimodal deep learning framework**. IEEE Transactions on Geoscience and Remote Sensing, 61:1–15, 2023.

144 HONG, D.; YAO, J.; MENG, D.; XU, Z. ; CHANUSSOT, J.. **Multimodal gans: Toward crossmodal hyperspectral–multispectral image segmentation**. IEEE Transactions on Geoscience and Remote Sensing, 59(6):5103–5113, 2021.

145 DU, X.; ZHENG, X.; LU, X. ; DOUDKIN, A. A.. **Multisource remote sensing data classification with graph fusion network**. IEEE Transactions on Geoscience and Remote Sensing, 59(12):10062–10072, 2021.

146 SCARPA, G.; GARGIULO, M.; MAZZA, A. ; GAETANO, R.. **A cnn-based fusion method for feature extraction from sentinel data**. Remote Sensing, 10(2), 2018.

147 LIAO, W.; VAN COILLIE, F.; GAO, L.; LI, L.; ZHANG, B. ; CHANUSSOT, J.. **Deep learning for fusion of apex hyperspectral and full-waveform lidar remote sensing data for tree species mapping**. IEEE Access, 6:68716–68729, 2018.

148 SAHA, S.; BOVOLO, F. ; BRUZZONE, L.. **Building change detection in vhr sar images via unsupervised deep transcoding**. IEEE Transactions on Geoscience and Remote Sensing, 59(3):1917–1929, 2021.

149 FUENTES REYES, M.; AUER, S.; MERKLE, N.; HENRY, C. ; SCHMITT, M.. **Sar-to-optical image translation based on conditional generative adversarial networks—optimization, opportunities and limits**. Remote Sensing, 11(17), 2019.

150 WANG, L.; XU, X.; YU, Y.; YANG, R.; GUI, R.; XU, Z. ; PU, F.. **Sar-to-optical image translation using supervised cycle-consistent adversarial networks**. IEEE Access, 7:129136–129149, 2019.

151 HE, W.; YOKOYA, N.. **Multi-temporal sentinel-1 and -2 data fusion for optical image simulation**. ISPRS International Journal of Geo-Information, 7(10), 2018.

152 MERANER, A.; EBEL, P.; ZHU, X. X. ; SCHMITT, M.. **Cloud removal in sentinel-2 imagery using a deep residual neural network and sar-optical data fusion**. ISPRS Journal of Photogrammetry and Remote Sensing, 166:333–346, 2020.

153 DARBAGHSHAHI, F. N.; MOHAMMADI, M. R. ; SORYANI, M.. **Cloud removal in remote sensing images using generative adversarial networks and sar-to-optical image translation**. IEEE Transactions on Geoscience and Remote Sensing, 60:1–9, 2022.

154 FERRARI, F.; FERREIRA, M. P.; ALMEIDA, C. A. ; FEITOSA, R. Q.. **Fusing sentinel-1 and sentinel-2 images for deforestation detection in the brazilian amazon under diverse cloud conditions**. IEEE Geoscience and Remote Sensing Letters, 20:1–5, 2023.

155 GOOGLE. **Google Earth Engine**. `https://earthengine.google.com/`, 12 2022.

A
Computer Setup

The computer setup used in the experiments is:

- CPU: Intel Core i9-14900KF
- RAM: 128 GB
- GPU: RTX 4090
- GPU Memory: 24 GB

B
Utilized images

B.1
Optical Images

Table 14 presents the acquisition dates of the optical images used in this work, including the cloud condition of each image.

Site	Cloud Condition	Image dates
Site 1	Cloud-free	2019-07-27
	Partially covered	2019-08-11
	Fully Covered	2019-08-16
	Cloud-free	2020-07-21
	Partially covered	2020-08-05
	Fully Covered	2020-08-10
	Cloud-free	2021-07-21
	Partially covered	2021-06-26
	Fully Covered	2021-07-01
Site 2	Cloud-free	2019-07-16
	Partially covered	2019-07-31
	Fully Covered	2019-08-05
	Cloud-free	2020-07-30
	Partially covered	2020-08-19
	Fully Covered	2020-06-30
	Cloud-free	2021-07-25
	Partially covered	2021-07-30
	Fully Covered	2021-08-14

Table 14: Acquisition dates of the optical images and the respective cloud condition.

Figures 92, 93, and 94 present the optical images acquired from Site 1 for the years 2019, 2020, and 2021, respectively. Similarly, Figures 95, 96, and 97 illustrate images obtained from Site 2 for the same years.

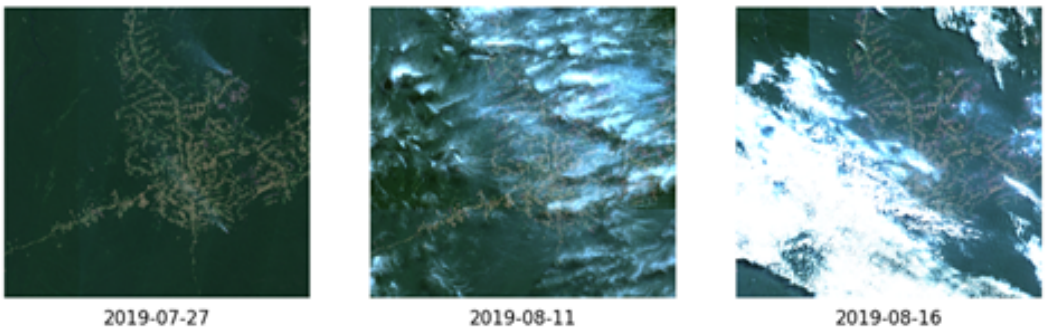


Figure 92: Optical Images from Site 1 (2019).

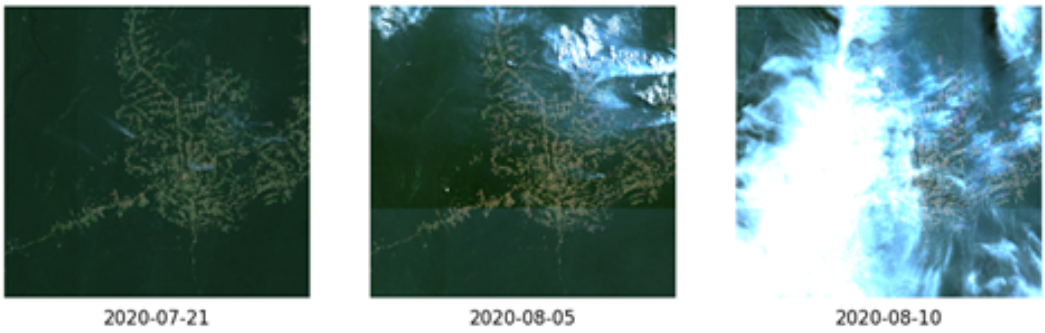


Figure 93: Optical Images from Site 1 (2020).

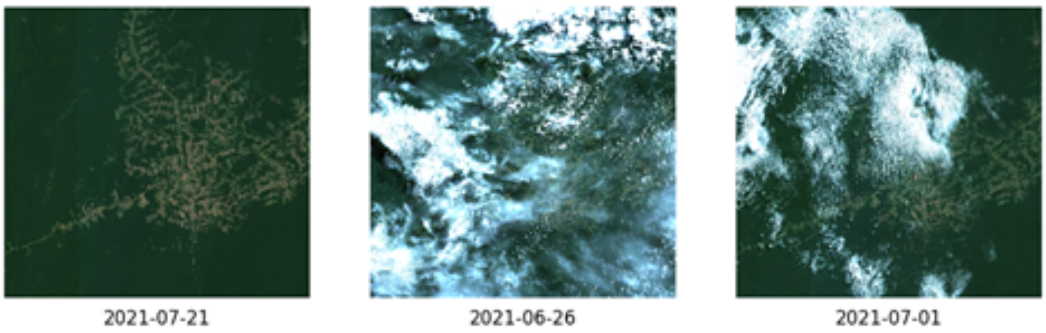


Figure 94: Optical Images from Site 1 (2021).

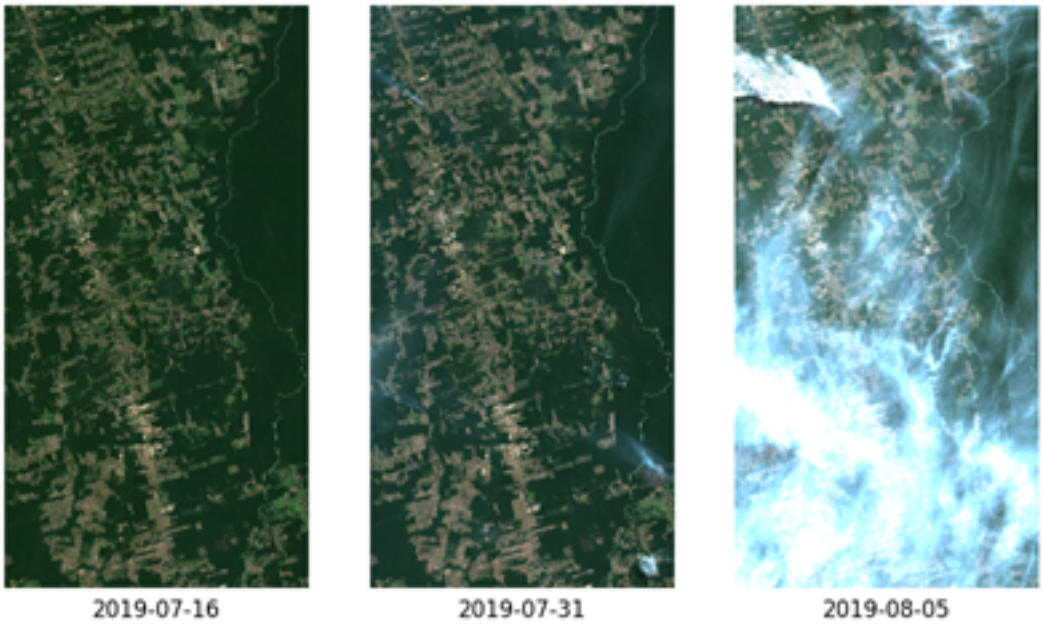


Figure 95: Optical Images from Site 2 (2019).

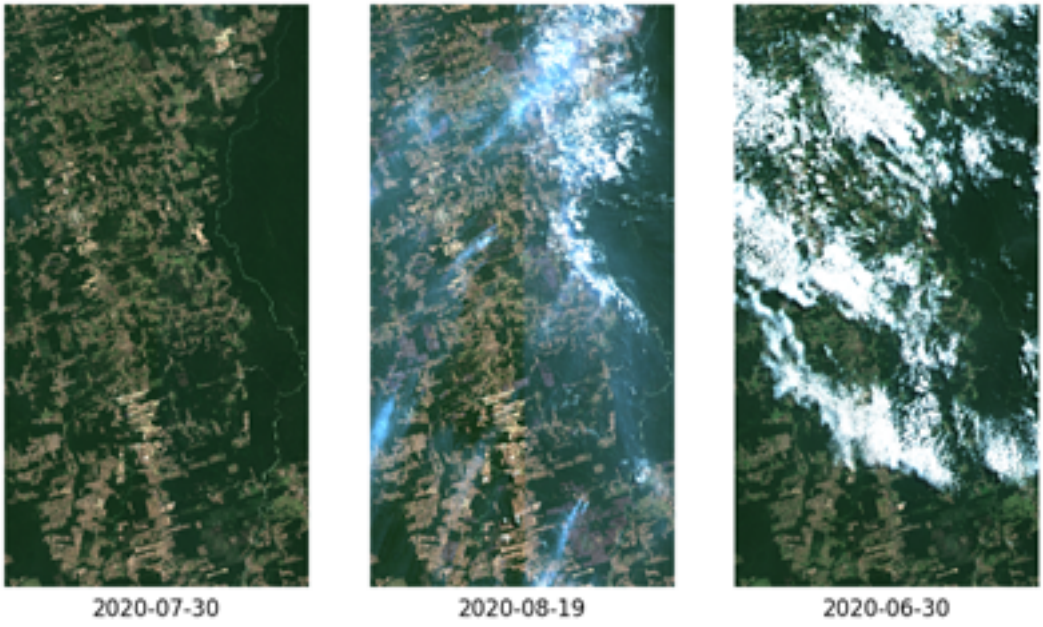


Figure 96: Optical Images from Site 2 (2020).

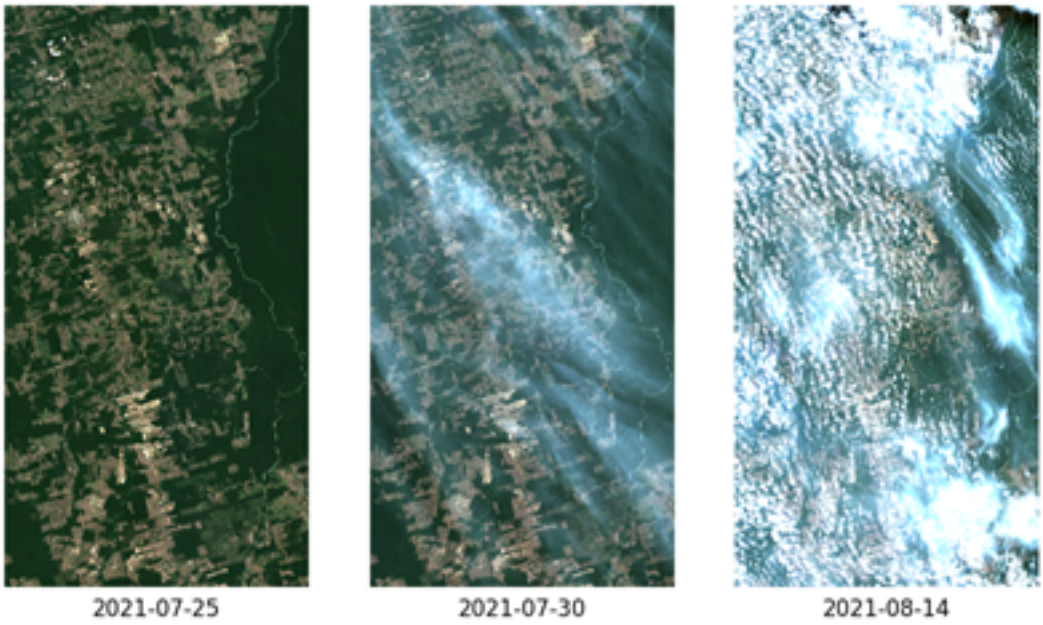


Figure 97: Optical Images from Site 2 (2021).

B.2 SAR Images

B.2.1 SAR Single Images

Table 15 presents the acquisition dates of the Sentinel-1 images used to create each mosaic. This SAR dataset is henceforth called *SINGLE*.

Site	Mosaic dates
Site 1	2019-08-05:2019-08-12
	2019-08-17:2019-08-24
	2019-08-29:2019-09-05
	2020-07-18:2020-07-25
	2020-07-30:2020-08-06
	2020-08-11:2020-08-18
	2021-06-19:2021-06-26
	2021-07-01:2021-07-08
	2021-07-13:2021-07-20
Site 2	2019-07-14:2019-07-12
	2019-07-26:2019-08-02
	2019-08-07:2019-08-14
	2020-07-08:2020-07-15
	2020-07-20:2020-07-27
	2020-08-01:2020-08-08
	2021-07-03:2021-07-10
	2021-07-15:2021-07-22
	2021-07-27:2021-08-03

Table 15: Mosaic acquisition dates of the SAR images.

Figures 98, 99, and 100 present the Sentinel-1 images from Site 1 for the years 2019, 2020, and 2021, respectively. Similarly, Figures 101, 102, and 103 illustrate images obtained from Site 2 for the same years. All these images are presented with bands VV, VH, and the bands' ratios VV/VH in the channels red, green, and blue, respectively.

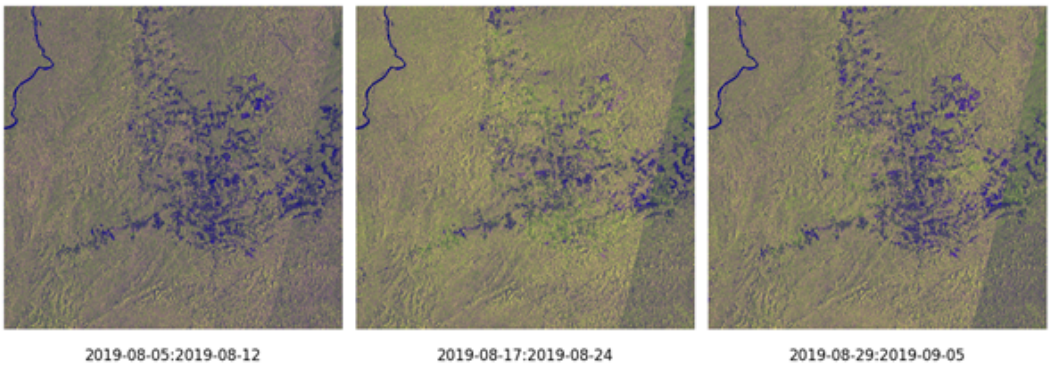


Figure 98: SAR Mosaic images from Site 1 (2019).

Figure 99: SAR Mosaic images from Site 1 (2020).

Figure 100: SAR Mosaic images from Site 1 (2021).

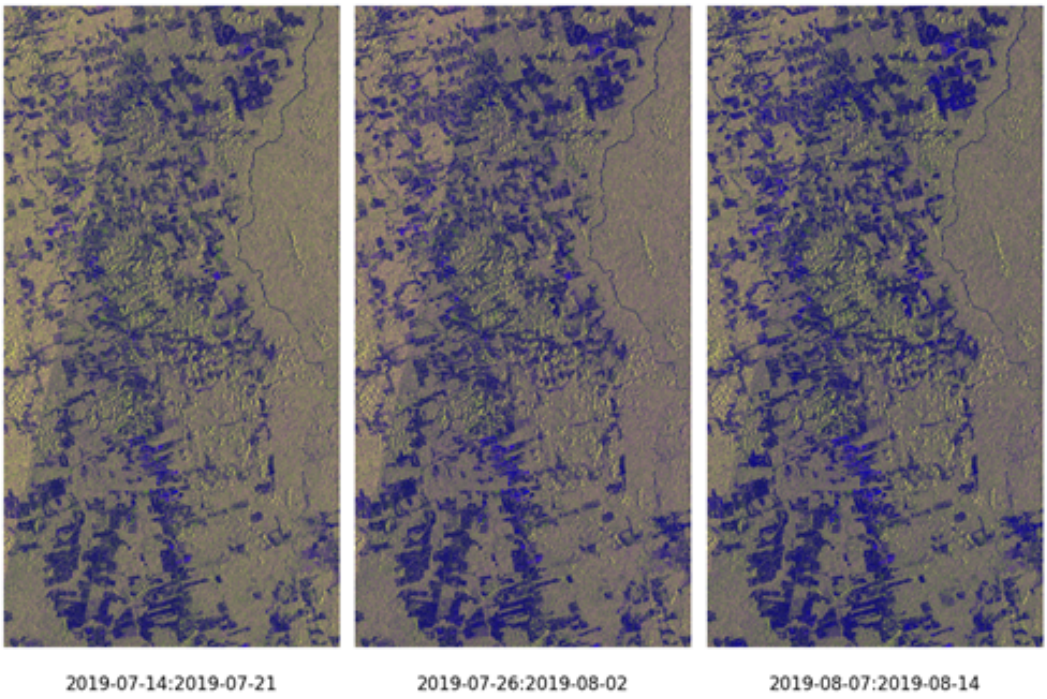


Figure 101: SAR Mosaic images from Site 2 (2019).

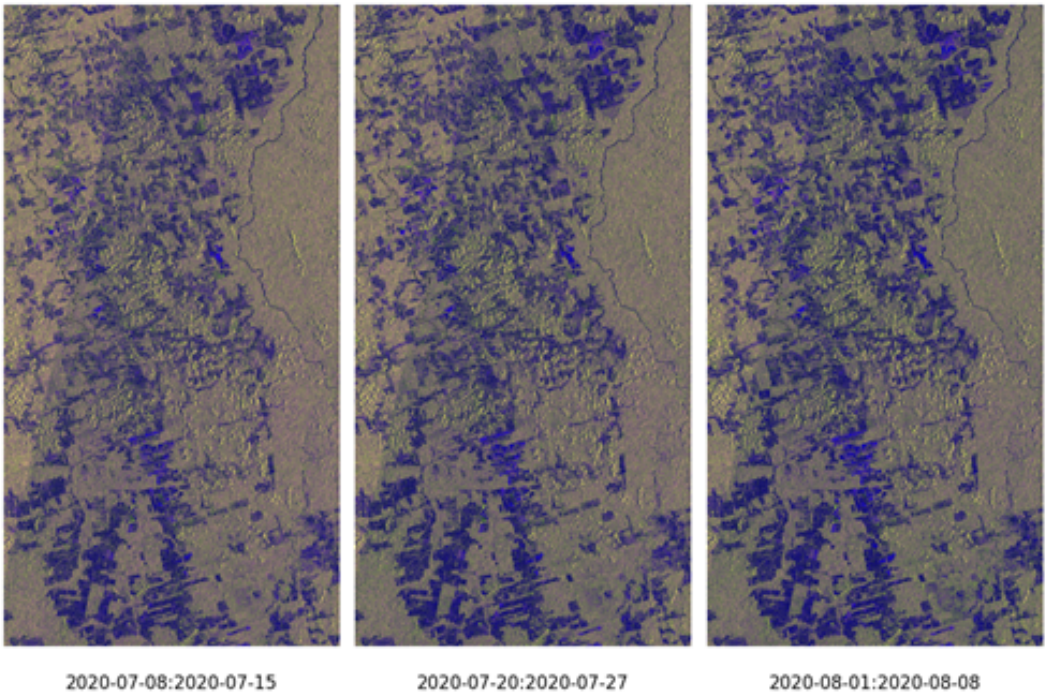


Figure 102: SAR Mosaic images from Site 2 (2020).

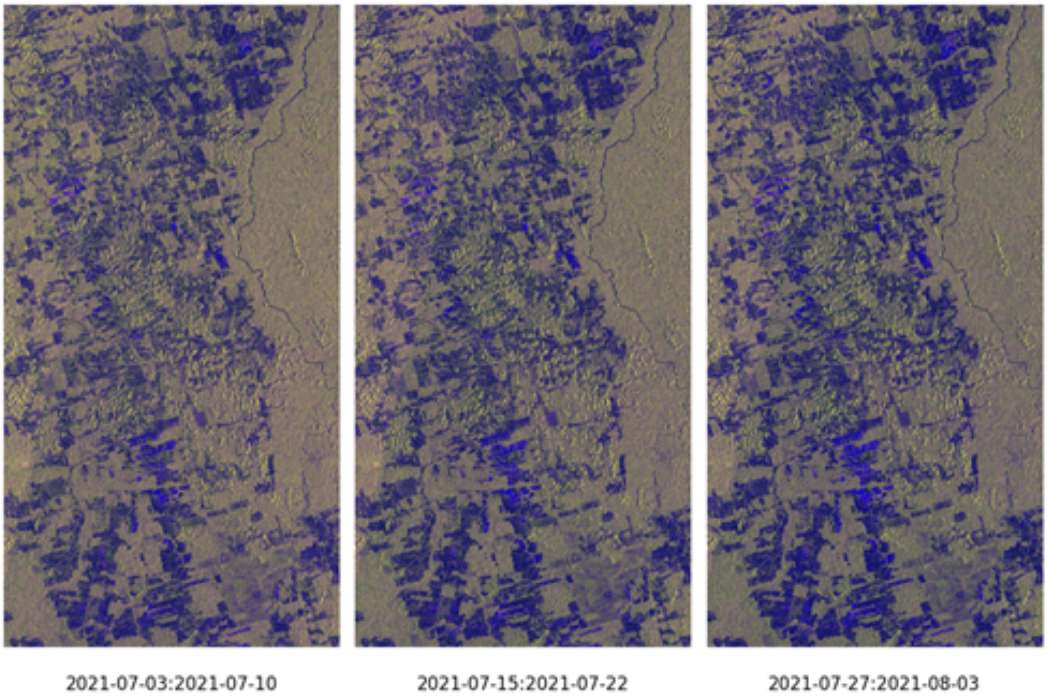


Figure 103: SAR Mosaic images from Site 2 (2021).

B.2.2 SAR Average GRD Images

Table 16 presents the acquisition dates of the Sentinel-1 images used to create each average image for each pair of consecutive years.

Site	Years	Average image dates
Site 1	2019-2020	2019-07-14:2019-08-14
		2019-08-15:2019-09-15
		2019-09-16:2019-10-17
		2019-10-18:2019-11-18
		2019-11-19:2019-12-20
		2019-12-21:2020-01-21
		2020-01-22:2020-02-22
		2020-02-23:2020-03-25
		2020-03-26:2020-04-26
		2020-04-27:2020-05-28
		2020-05-29:2020-06-29
		2020-06-30:2020-07-30
	2020-2021	2020-06-30:2020-07-30
		2020-07-31:2020-08-30
		2020-08-31:2020-09-30
		2020-10-01:2020-10-31
		2020-11-01:2020-12-01
		2020-12-02:2021-01-01
		2021-01-02:2021-02-01
		2021-02-02:2021-03-03
		2021-03-04:2021-04-02
		2021-04-03:2021-05-02
		2021-05-03:2021-06-01
		2021-06-02:2021-07-01
Site 2	2019-2020	2019-06-23:2019-07-26
		2019-07-27:2019-08-29
		2019-08-30:2019-10-02
		2019-10-03:2019-11-04
		2019-11-05:2019-12-07
		2019-12-08:2020-01-09
		2020-01-10:2020-02-11
		2020-02-12:2020-03-15
		2020-03-16:2020-04-17

Site	**Years**	**Average image dates**
Site 2	2019-2020	2020-04-18:2020-05-20
		2020-05-21:2020-06-22
		2020-06-23:2020-07-25
	2020-2021	2020-06-25:2020-07-28
		2020-07-29:2020-08-31
		2020-09-01:2020-10-04
		2020-10-05:2020-11-06
		2020-11-07:2020-12-09
		2020-12-10:2021-01-11
		2021-01-12:2021-02-13
		2021-02-14:2021-03-18
		2021-03-19:2021-04-20
		2021-04-21:2021-05-23
		2021-05-24:2021-06-25
		2021-06-26:2021-07-28

Table 16: Average SAR images' dates.

Figures 104 and 105 present the Average GRD images for the pairs of years 2019-2020 and 2020-2021 from Site 1, respectively. Likewise, Figures 106 and 106, illustrate the images obtained from Site 2 for the same pairs of years. All these images are presented with bands VV, VH, and the bands' ratios VV/VH in the channels red, green, and blue, respectively.

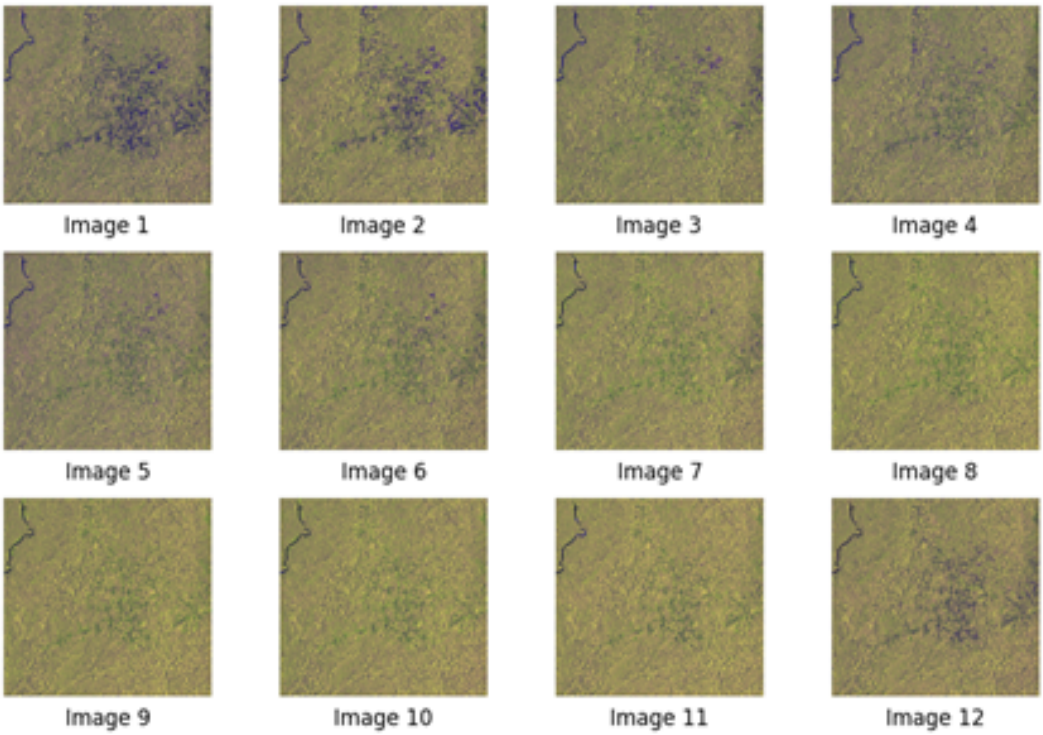


Figure 104: Average SAR images from Site 1 (2019-2020).

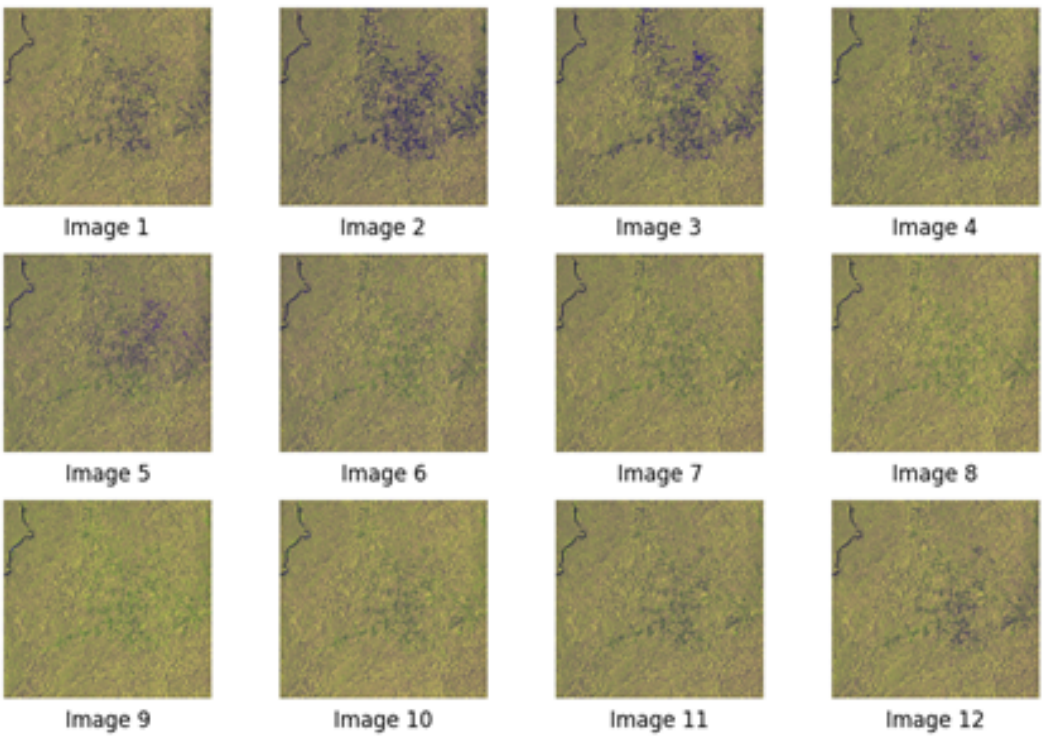


Figure 105: Average SAR images from Site 1 (2020-2021).

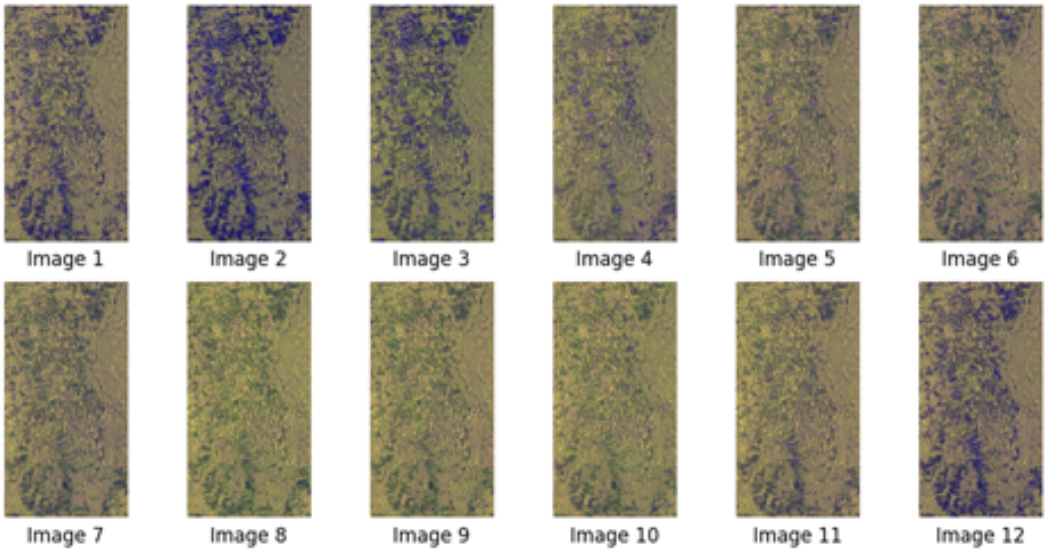


Figure 106: Average SAR images from Site 1 (2019-2020).

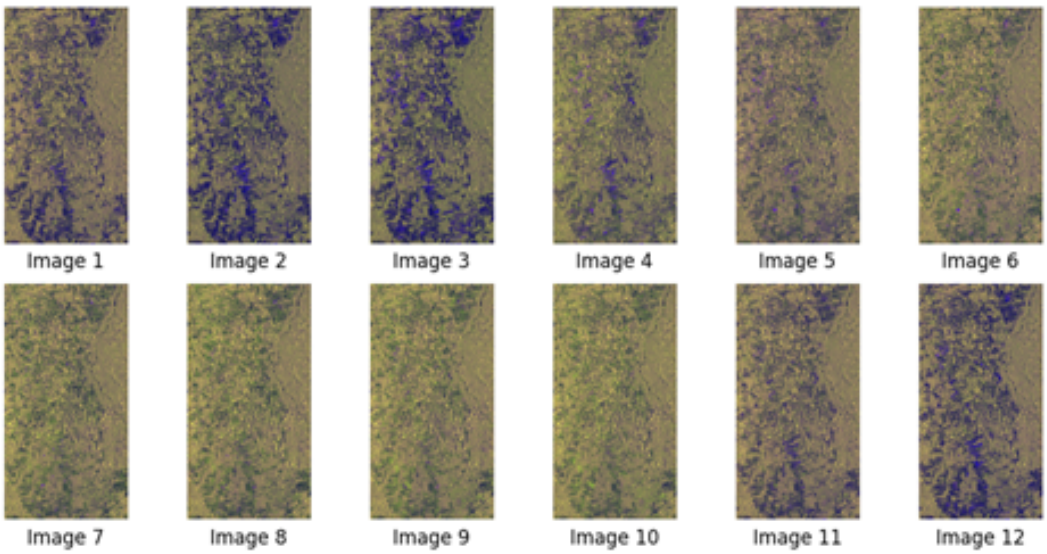


Figure 107: Average SAR images from Site 1 (2020-2021).

B.3
Cloud maps

Figures 108, 109, 110 present the cloud maps related to the optical images from Site 1 acquired in the years 2019, 2020, and 2021. Figures 111, 112, 113 present the same cloud maps from Site 2 acquired in the same years.

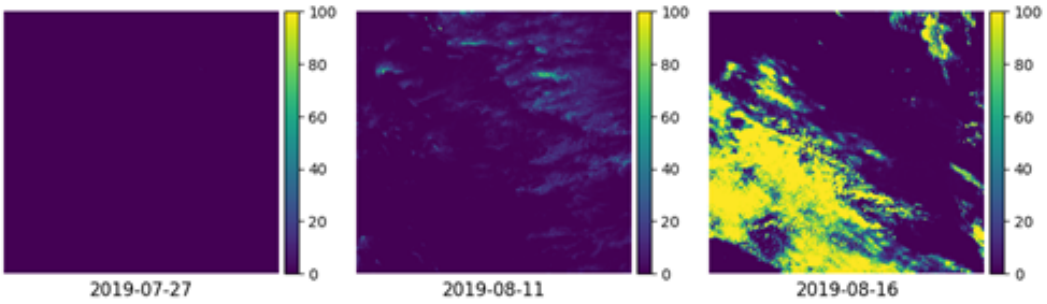


Figure 108: Cloud maps from optical data (Site 1 - 2019).

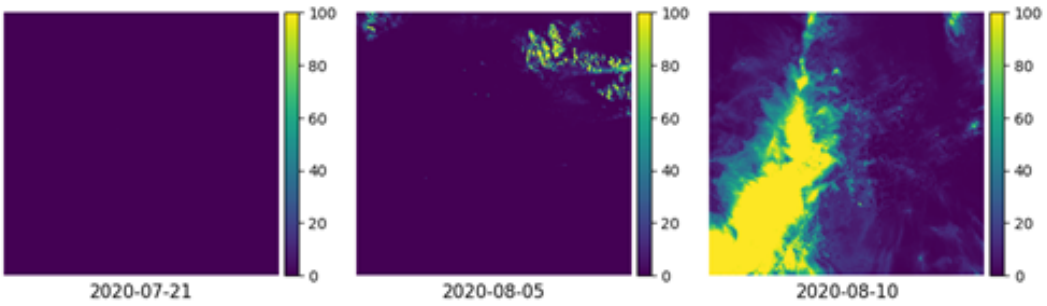


Figure 109: Cloud maps from optical data (Site 1 - 2020).

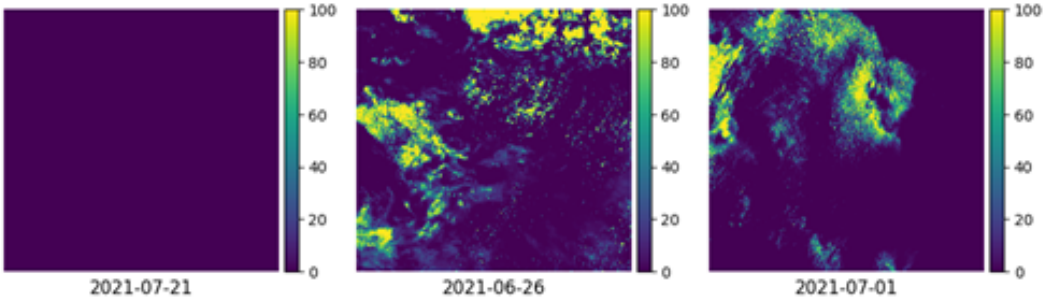


Figure 110: Cloud maps from optical data (Site 1 - 2021).

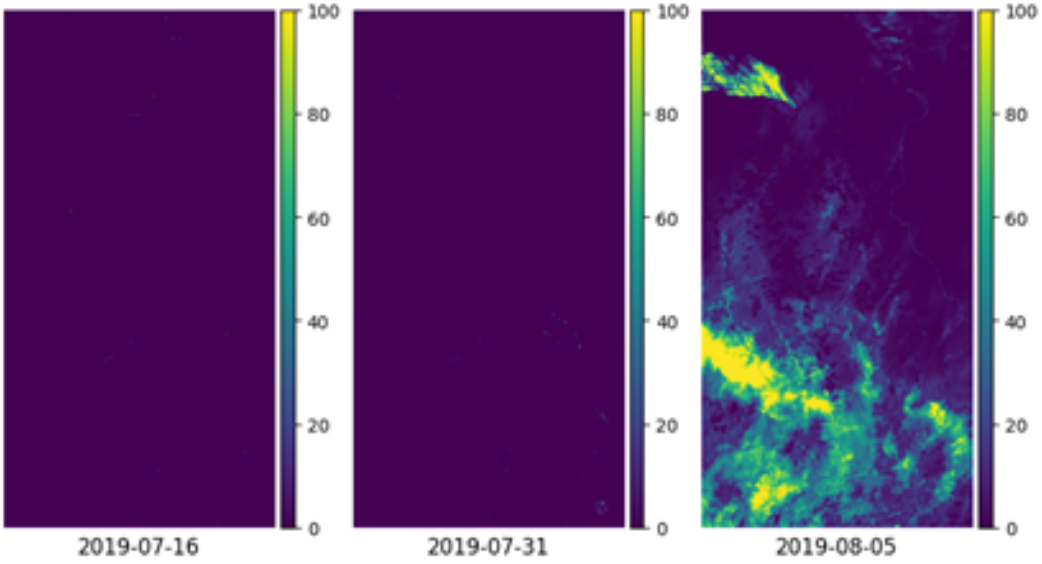


Figure 111: Cloud maps from optical data (Site 2 - 2019).

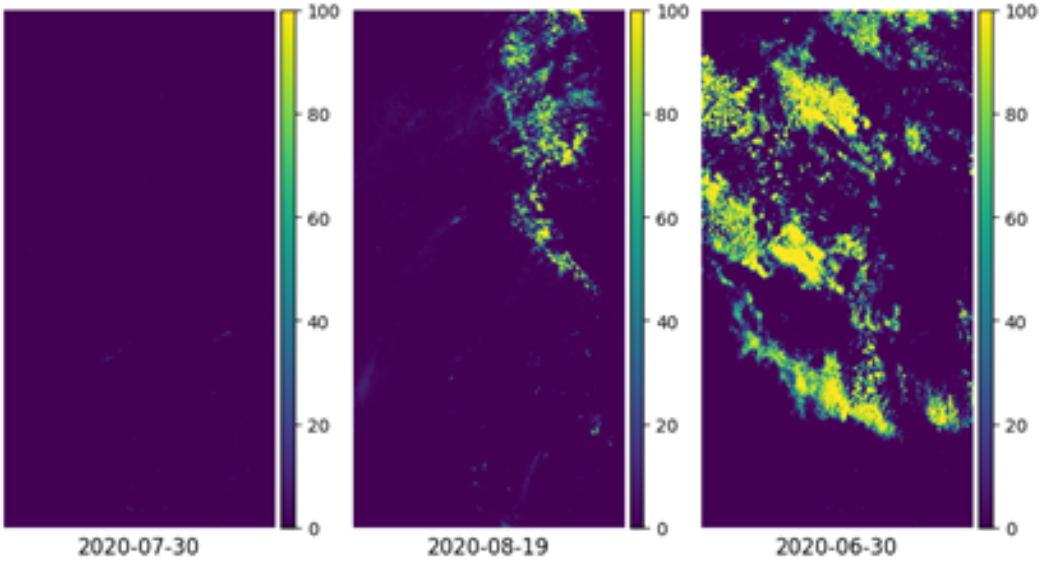


Figure 112: Cloud maps from optical data (Site 2 - 2020).

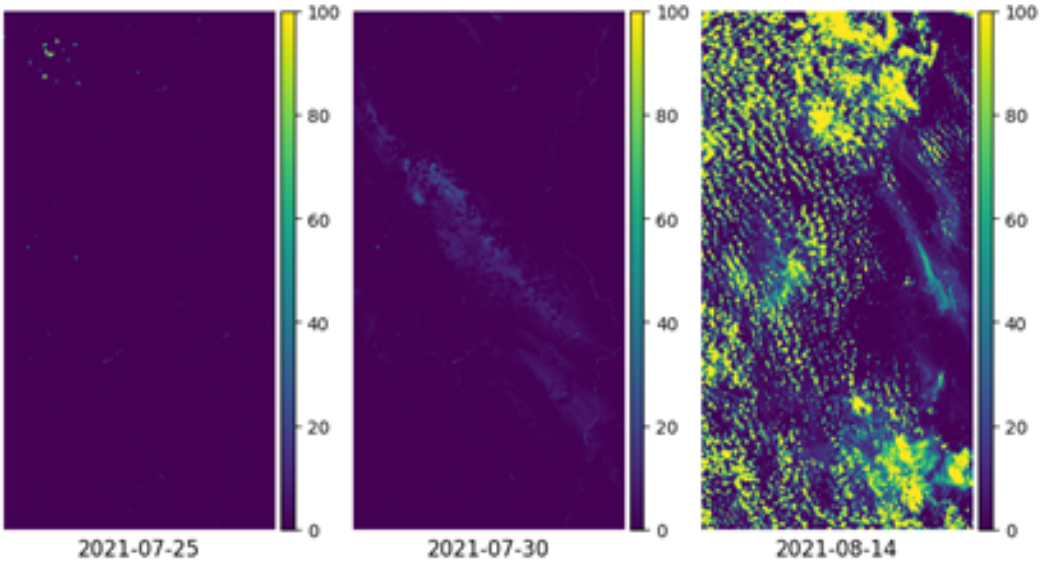


Figure 113: Cloud maps from optical data (Site 2 - 2021).

B.4 Results

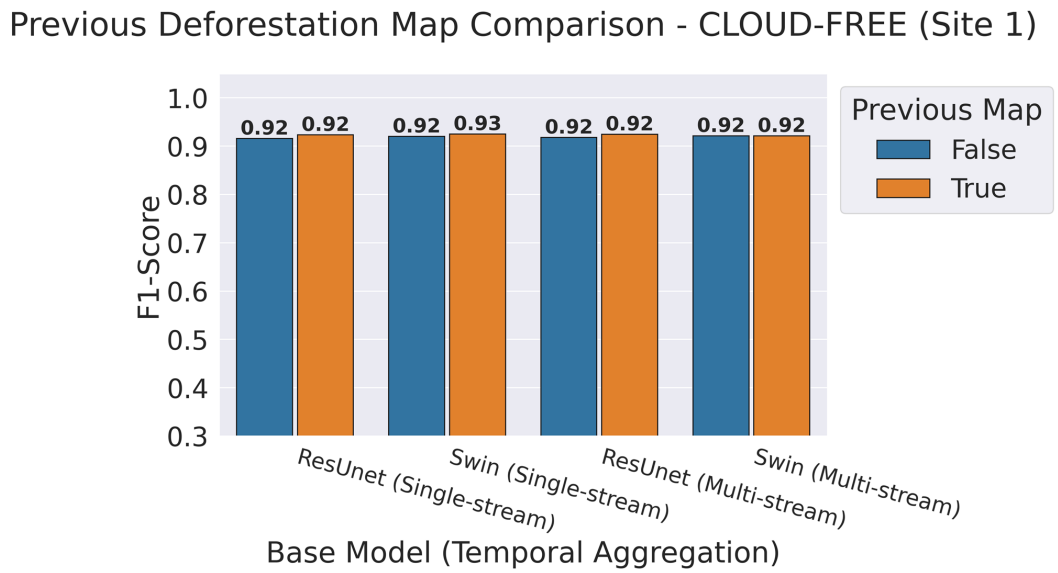


Figure 114: F1-Score Comparison of the models with (**orange**) and without (**blue**) Previous Deforestation Map (*CLOUD-FREE* from Site 1).

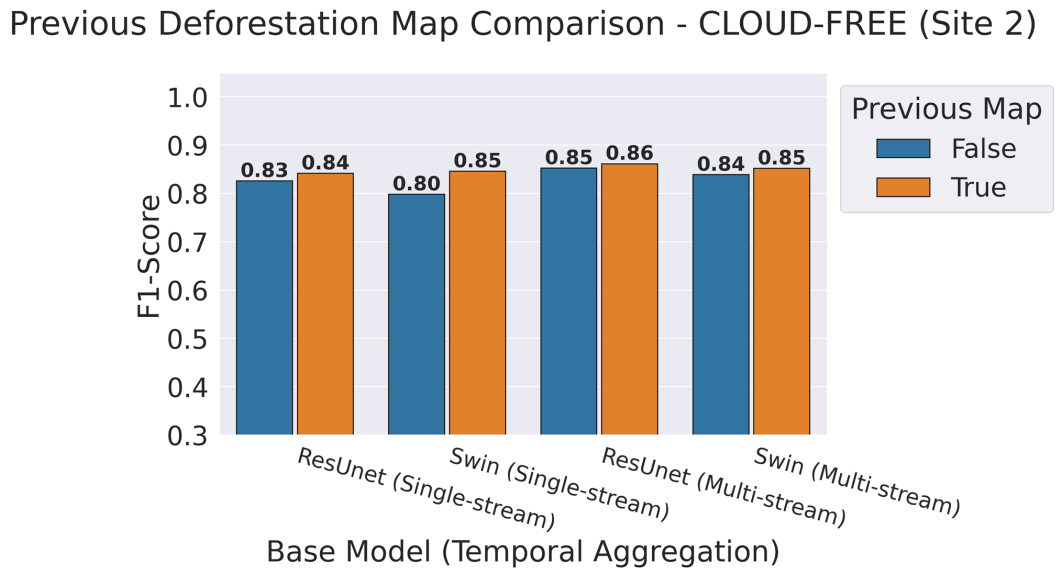


Figure 115: F1-Score Comparison of the models with (**orange**) and without (**blue**) Previous Deforestation Map (*CLOUD-FREE* from Site 2).

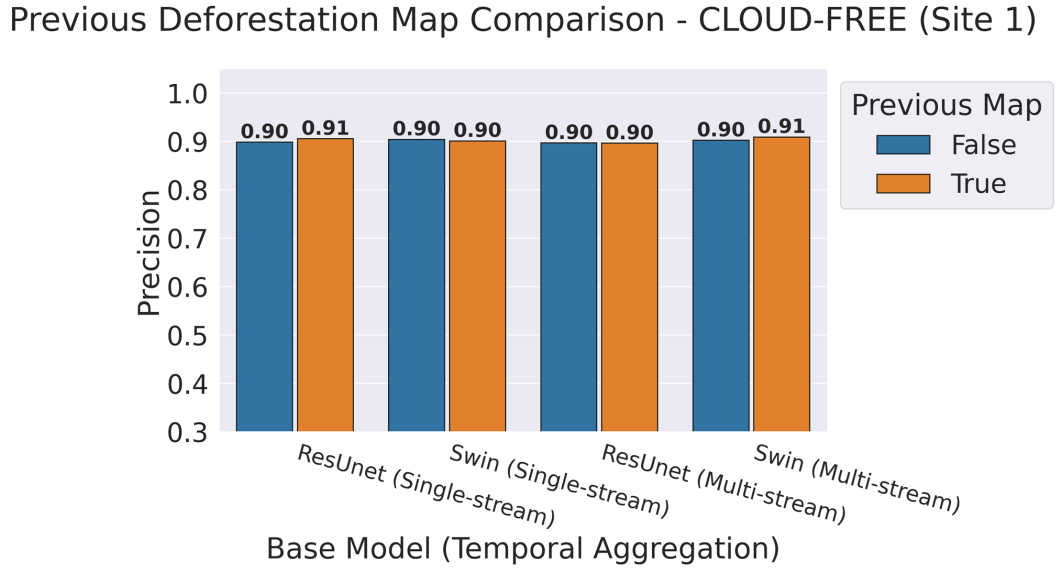


Figure 116: Precision Comparison of the models with (**orange**) and without (**blue**) Previous Deforestation Map (*CLOUD-FREE* from Site 1).

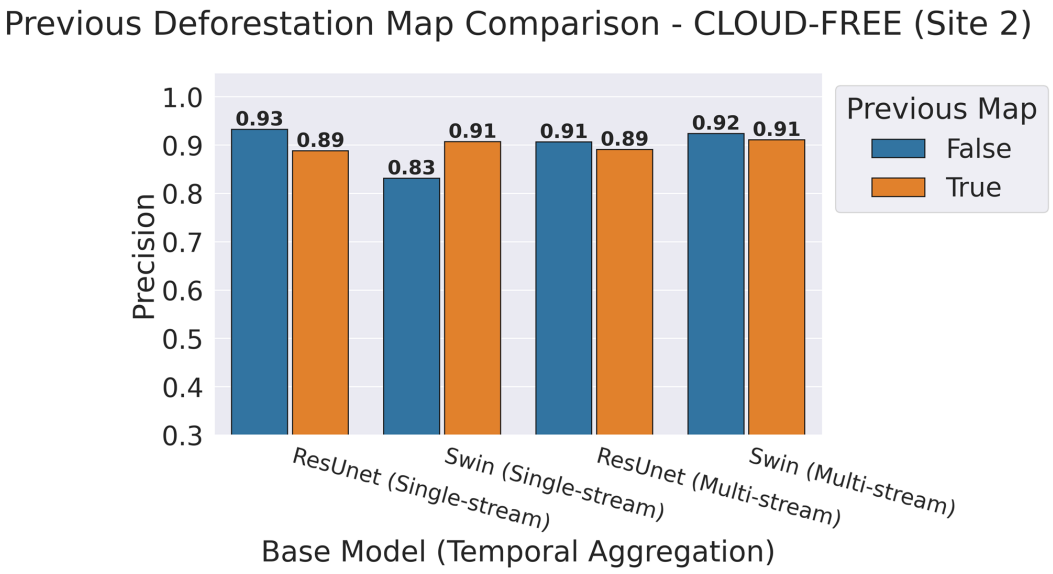


Figure 117: Precision Comparison of the models with (**orange**) and without (**blue**) Previous Deforestation Map (*CLOUD-FREE* from Site 2).

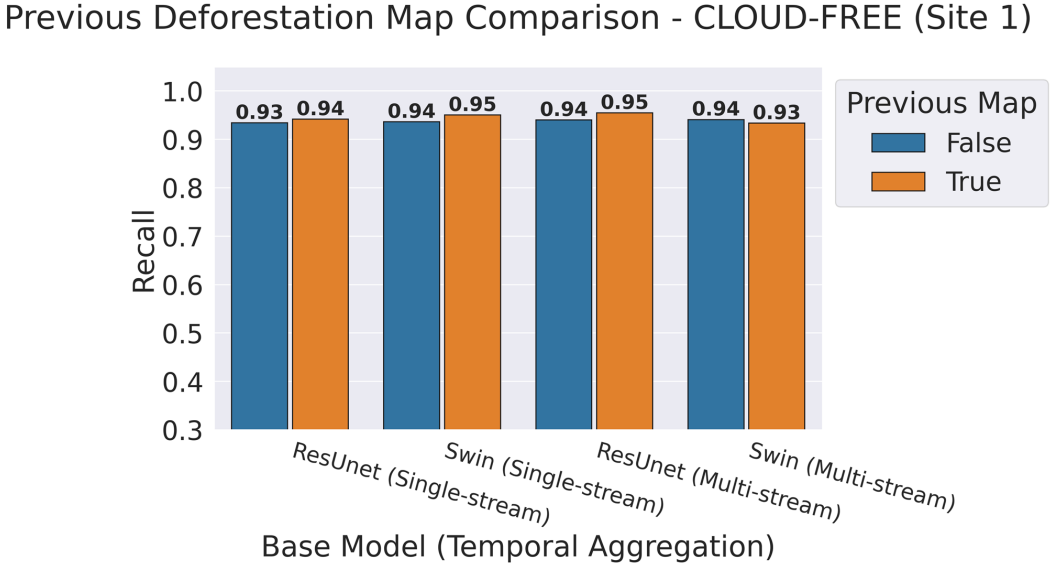


Figure 118: Recall Comparison of the models with (**orange**) and without (**blue**) Previous Deforestation Map (*CLOUD-FREE* from Site 1).

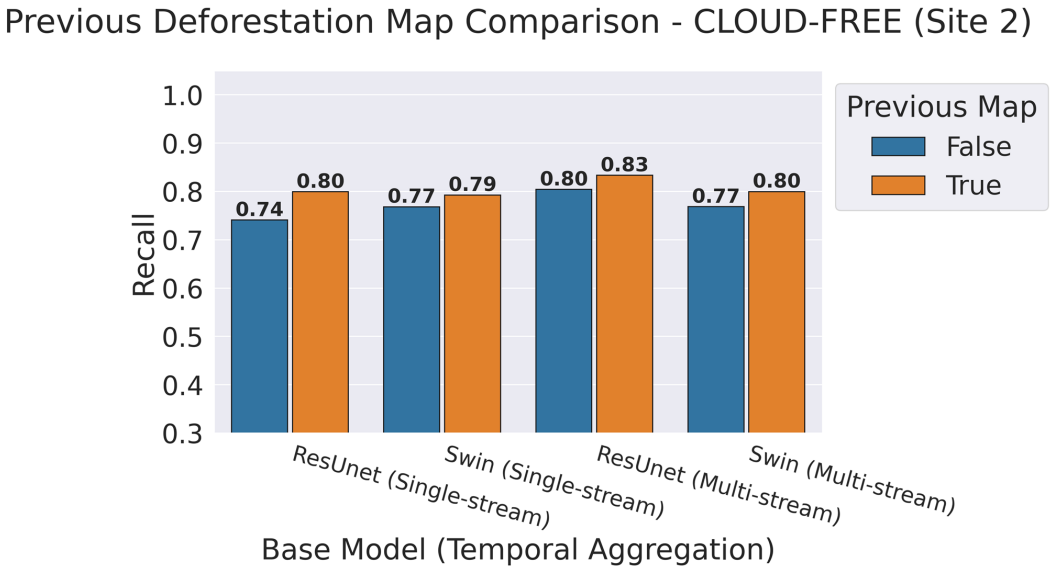


Figure 119: Recall Comparison of the models with (**orange**) and without (**blue**) Previous Deforestation Map (*CLOUD-FREE* from Site 2).

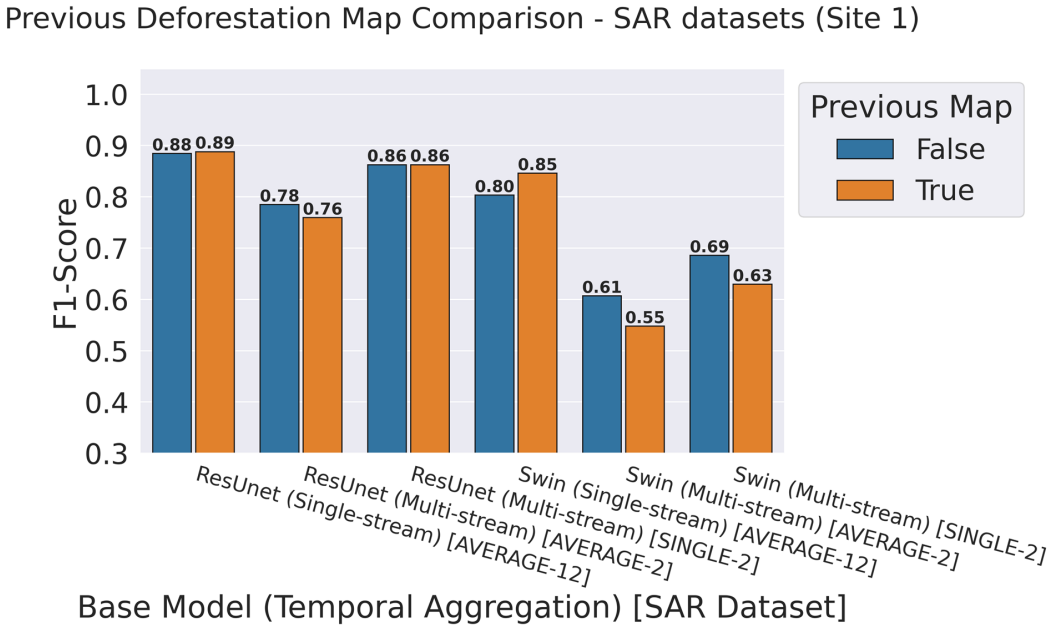


Figure 120: F1-Score Comparison of the models with (**orange**) and without (**blue**) Previous Deforestation Map (SAR datasets from Site 1).

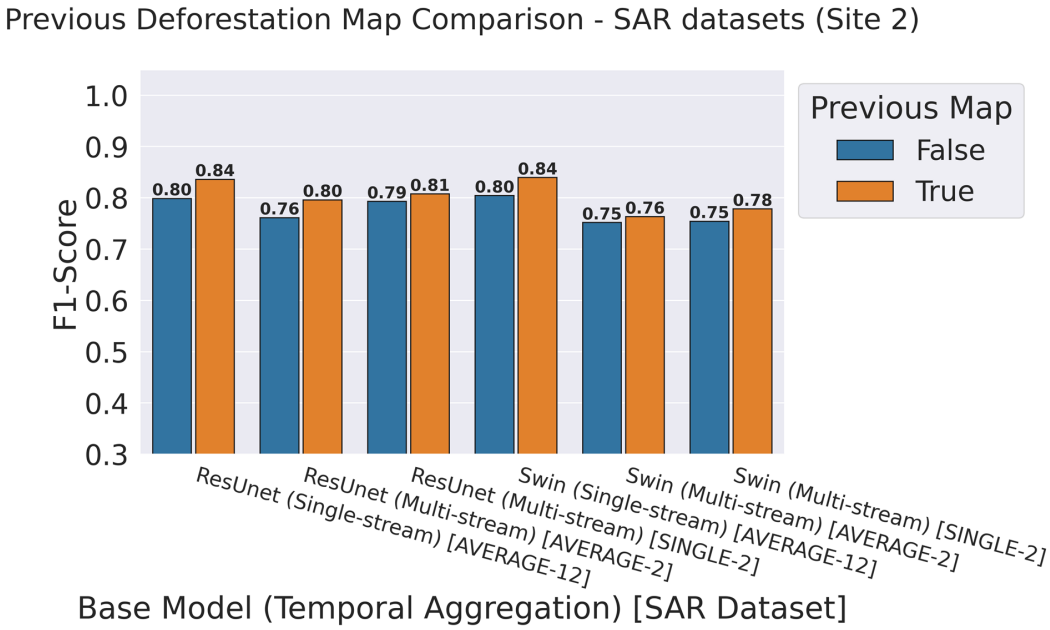


Figure 121: F1-Score Comparison of the models with (**orange**) and without (**blue**) Previous Deforestation Map (SAR datasets from Site 2).

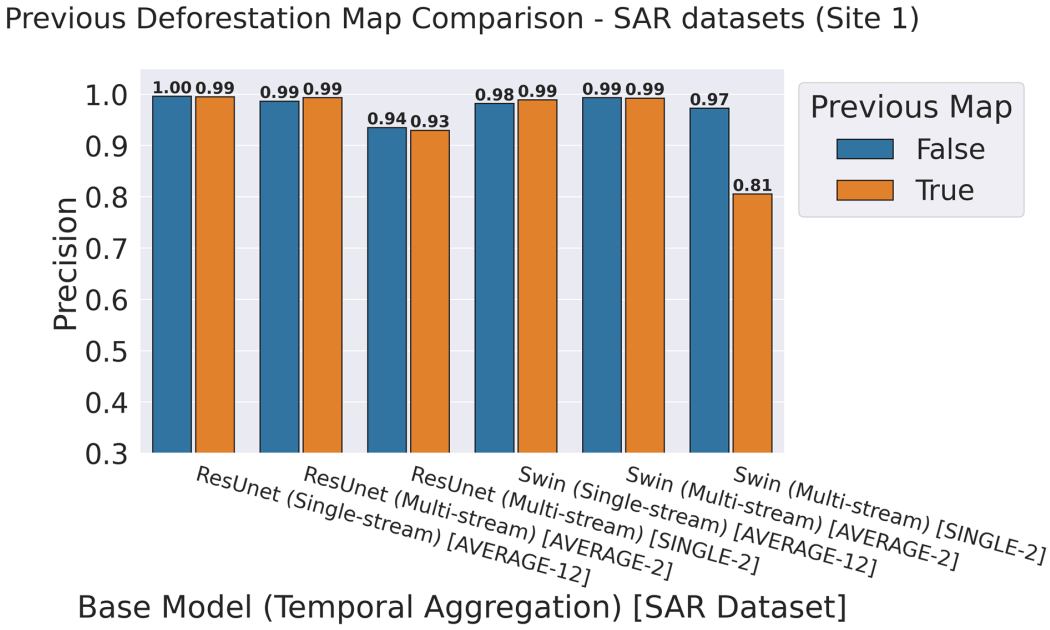


Figure 122: Precision Comparison of the models with (**orange**) and without (**blue**) Previous Deforestation Map (SAR datasets from Site 1).

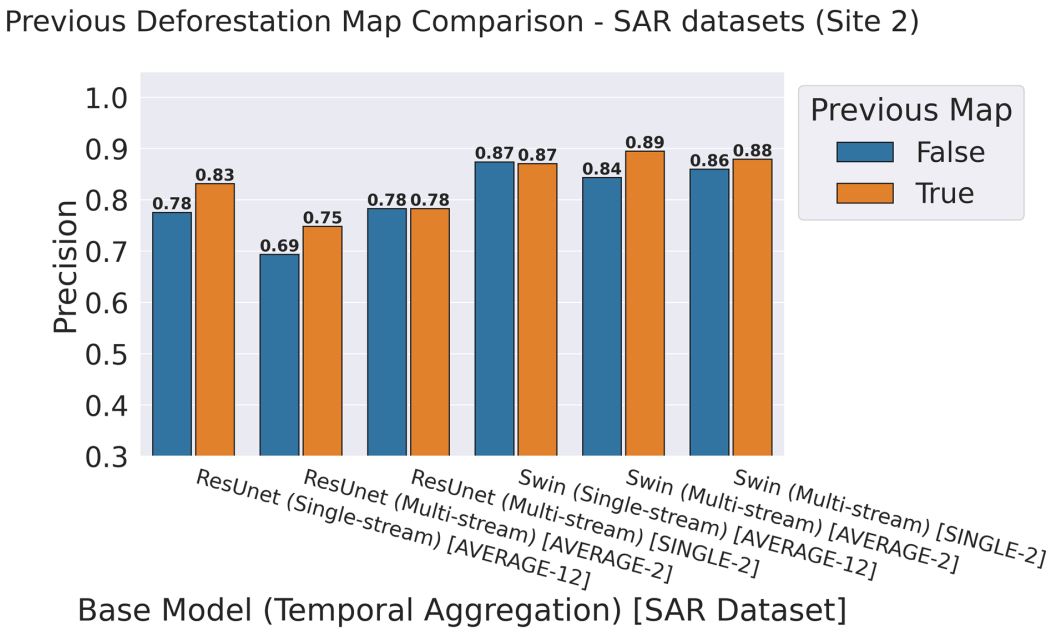


Figure 123: Precision Comparison of the models with (**orange**) and without (**blue**) Previous Deforestation Map (SAR datasets from Site 2).

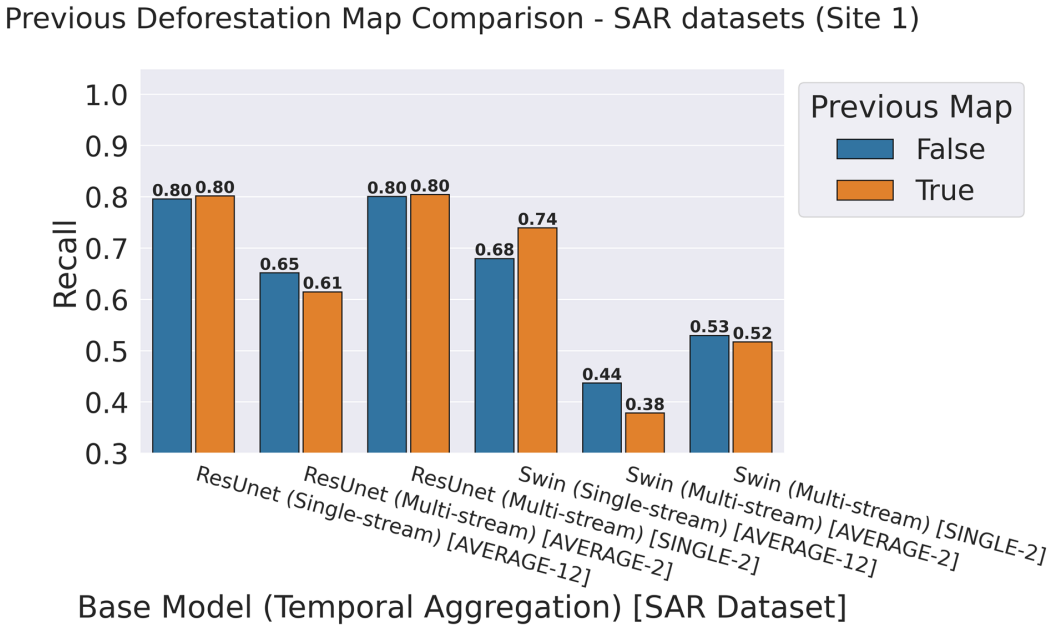


Figure 124: Recall Comparison of the models with (**orange**) and without (**blue**) Previous Deforestation Map (SAR datasets from Site 1).

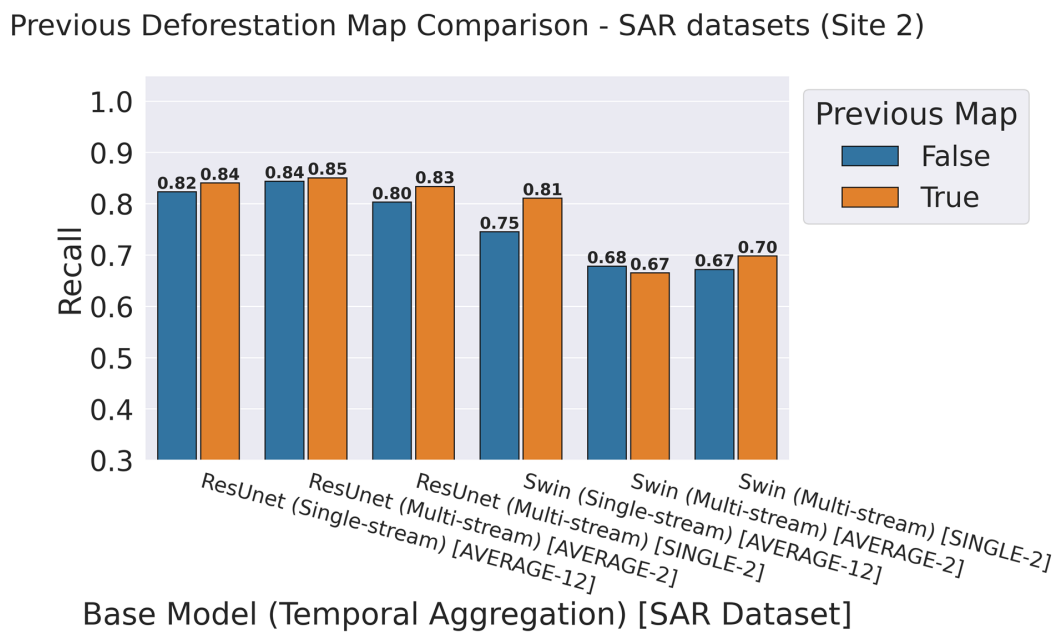


Figure 125: Recall Comparison of the models with (**orange**) and without (**blue**) Previous Deforestation Map (SAR datasets from Site 2).

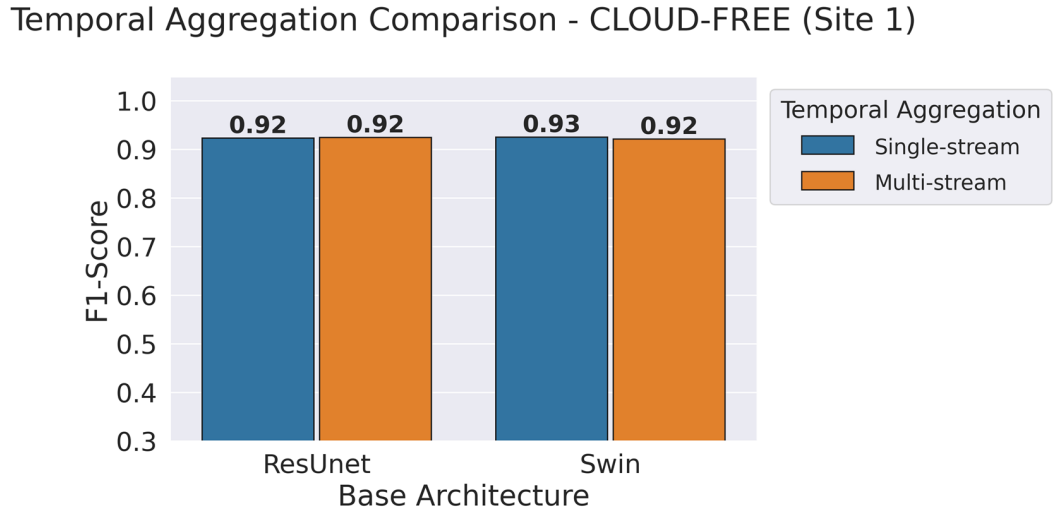


Figure 126: Temporal aggregation comparison (F1-Score) in *CLOUD-FREE* dataset (Site 1).

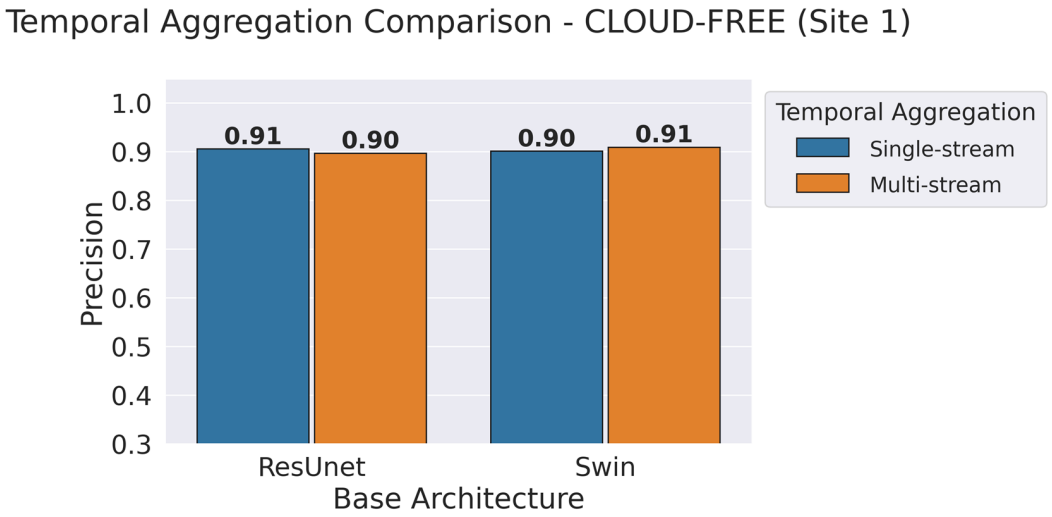


Figure 127: Temporal aggregation comparison (Precision) in *CLOUD-FREE* dataset (Site 1).

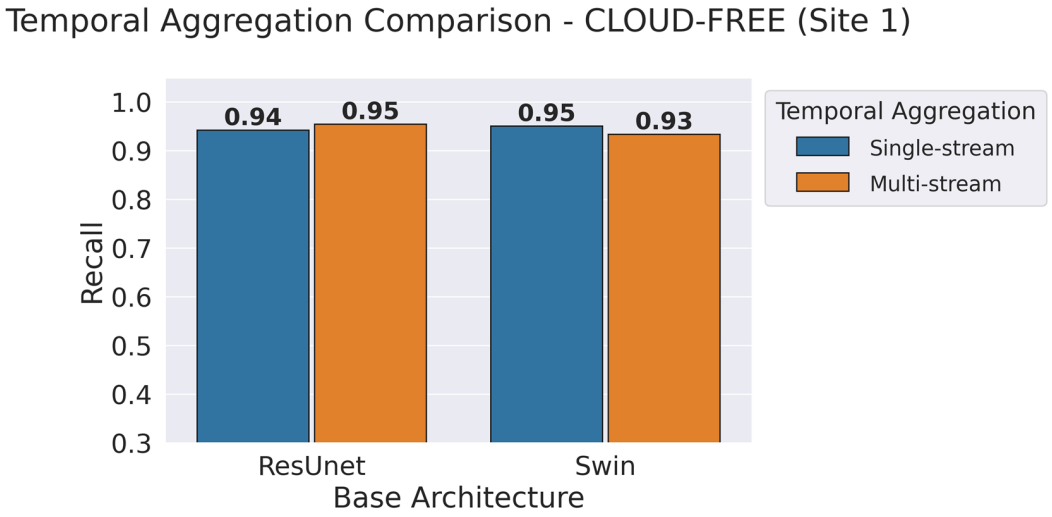


Figure 128: Temporal aggregation comparison (Recall) in *CLOUD-FREE* dataset (Site 1).

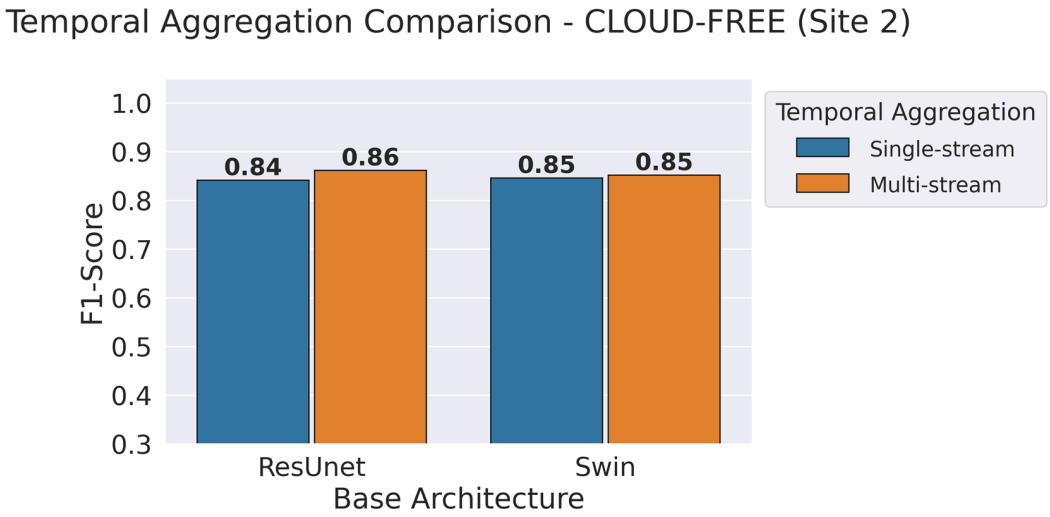


Figure 129: Temporal aggregation comparison (F1-Score) in *CLOUD-FREE* dataset (Site 2).

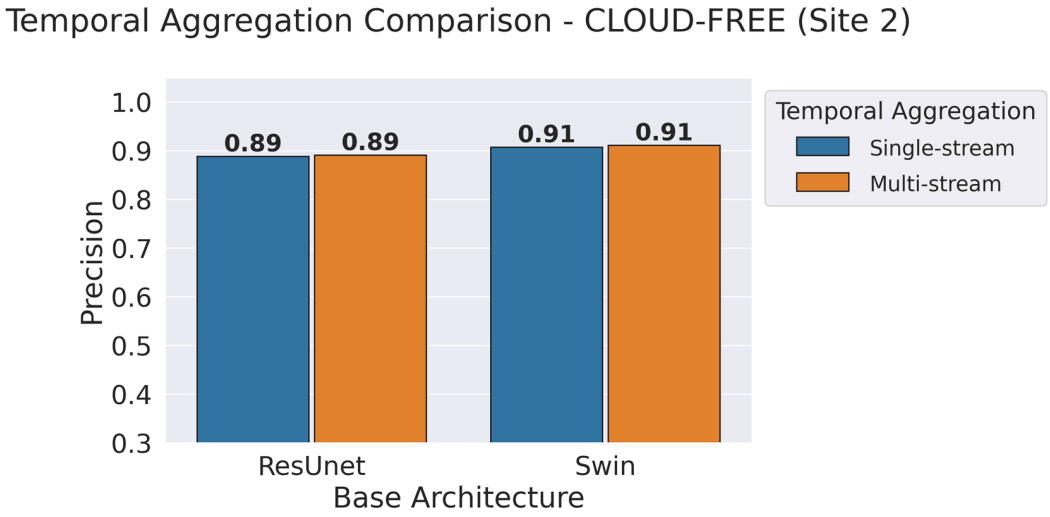


Figure 130: Temporal aggregation comparison (Precision) in *CLOUD-FREE* dataset (Site 2).

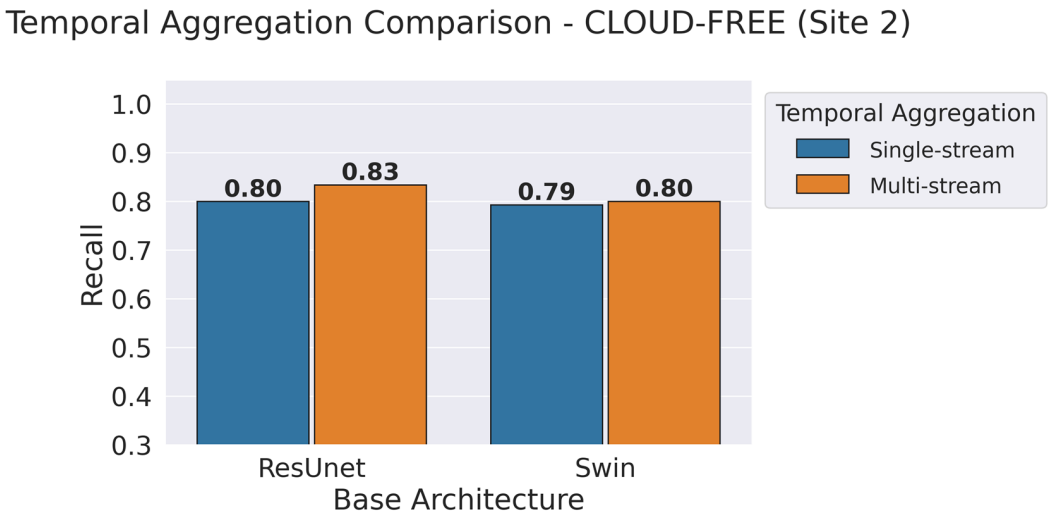


Figure 131: Temporal aggregation comparison (Recall) in *CLOUD-FREE* dataset (Site 2).

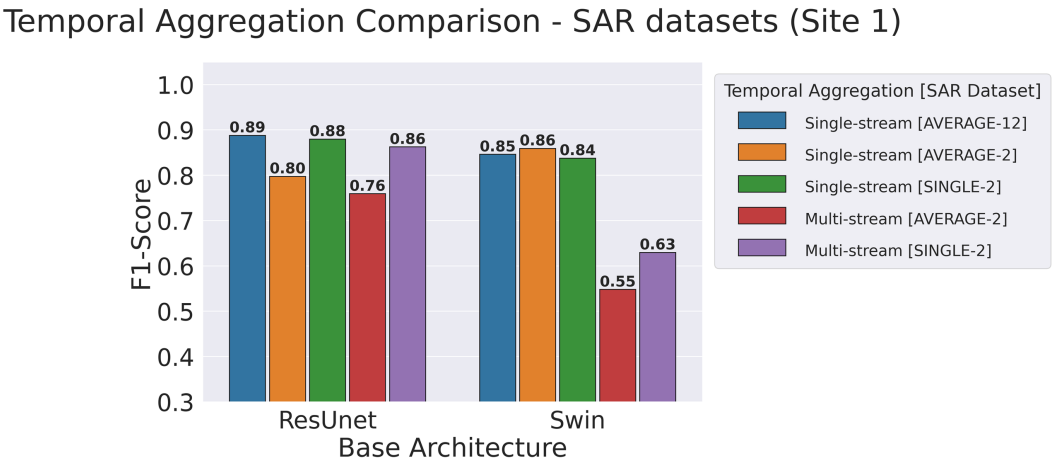


Figure 132: Temporal aggregation comparison (F1-Score) in SAR datasets (Site 1).

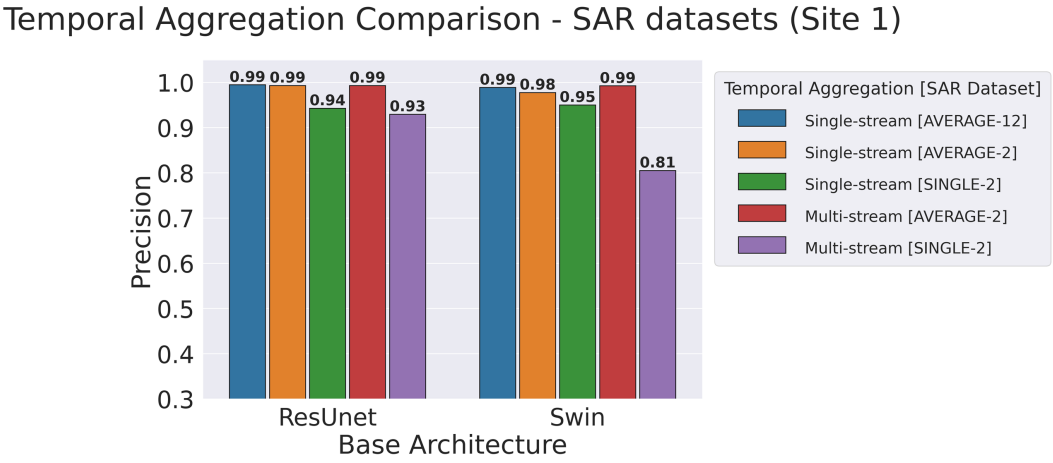


Figure 133: Temporal aggregation comparison (Precision) in SAR datasets (Site 1).

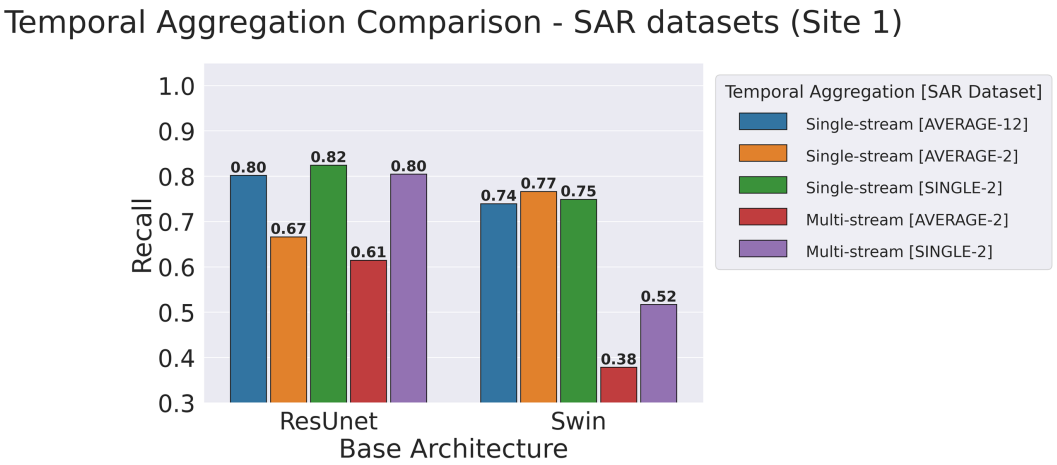


Figure 134: Temporal aggregation comparison (Recall) in SAR datasets (Site 1).

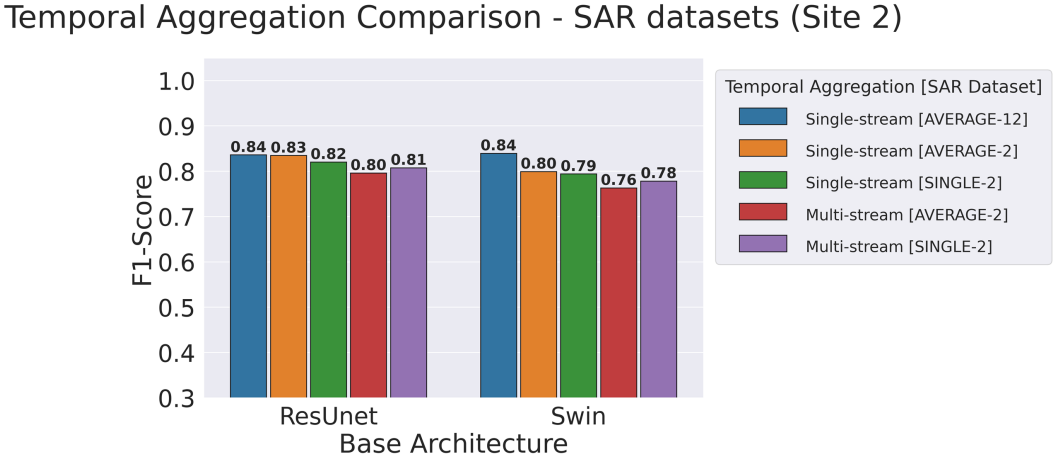


Figure 135: Temporal aggregation comparison (F1-Score) in SAR datasets (Site 2).

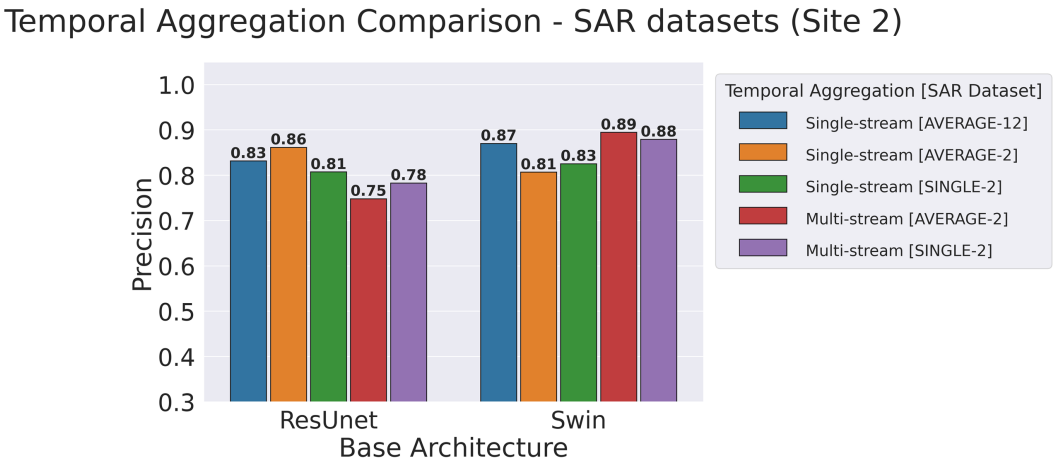


Figure 136: Temporal aggregation comparison (Precision) in SAR datasets (Site 2).

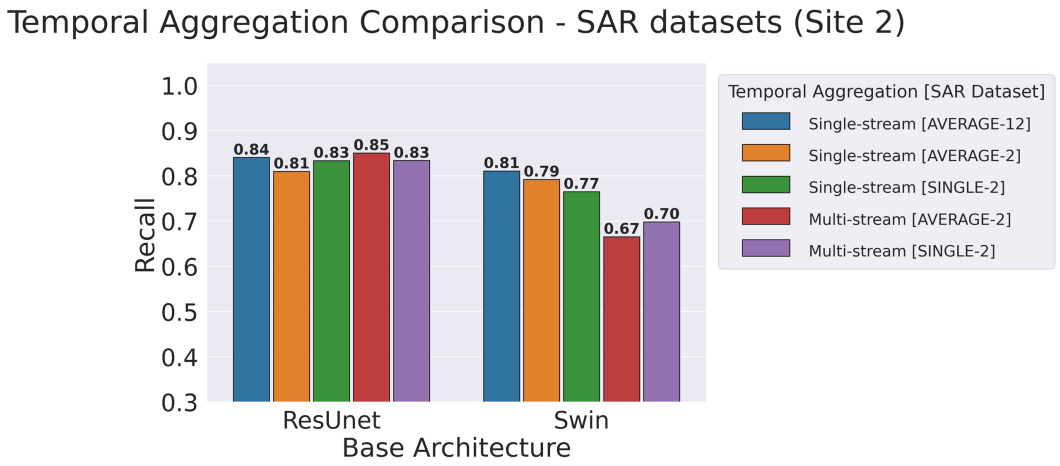


Figure 137: Temporal aggregation comparison (Recall) in SAR datasets (Site 2).

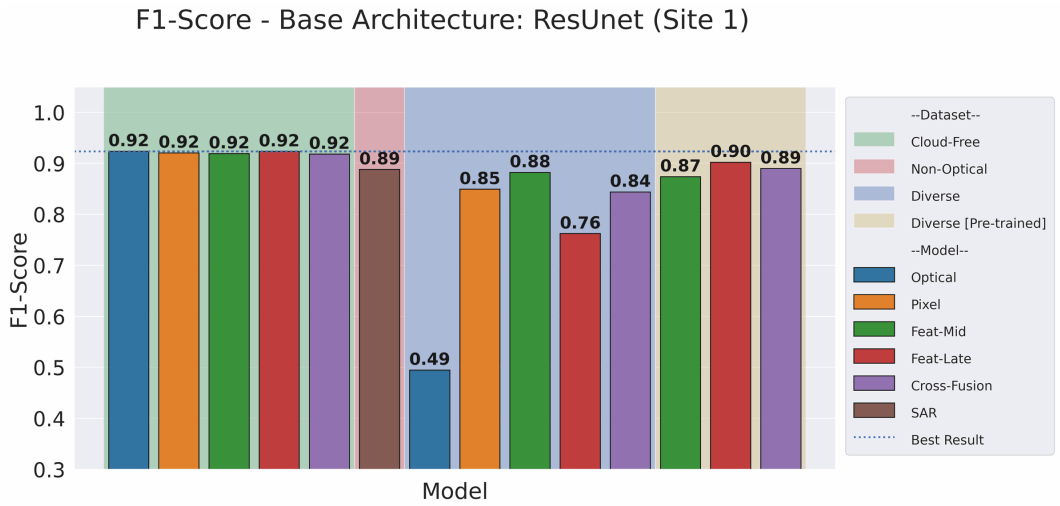


Figure 138: F1-Score for ResUnet-based models' comparison (Site 1)

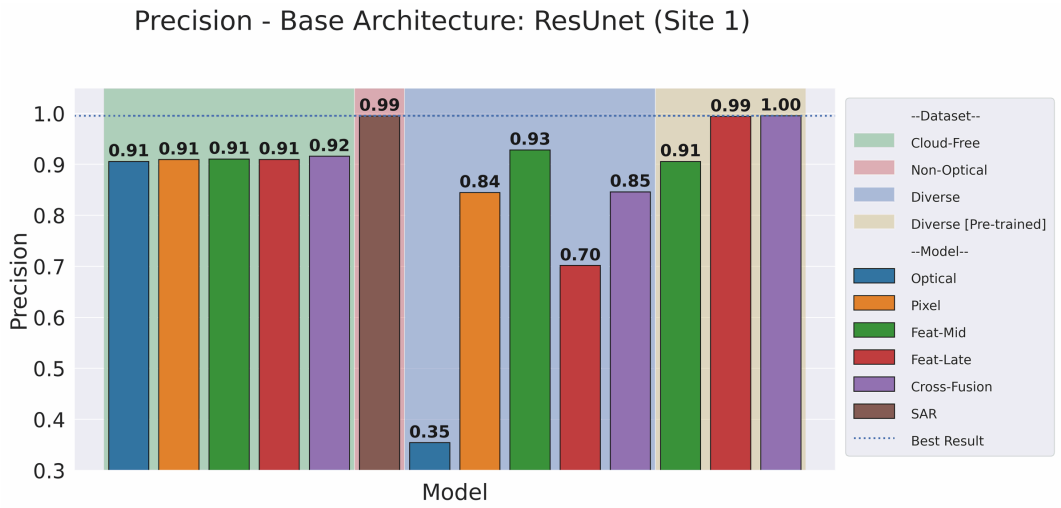


Figure 139: Precision for ResUnet-based models' comparison (Site 1)

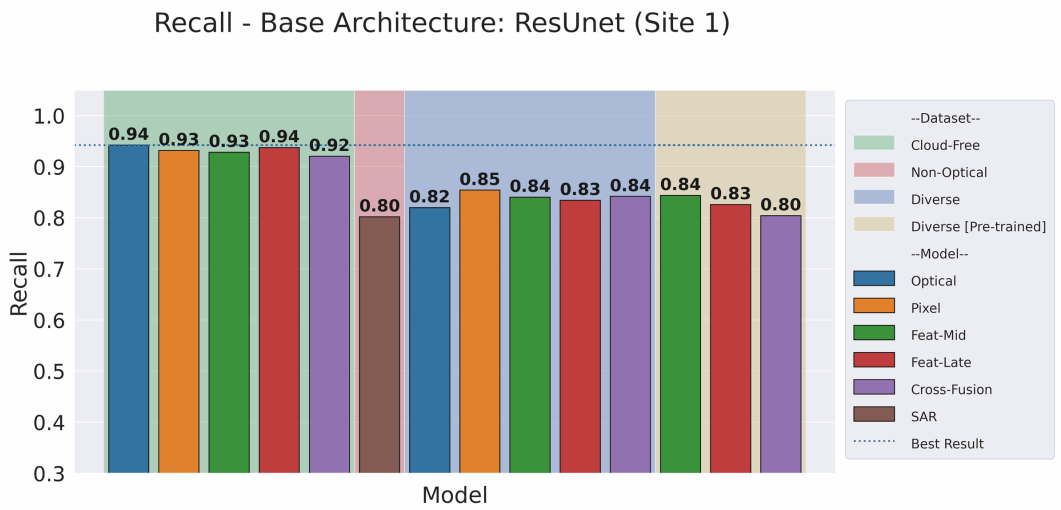


Figure 140: Recall for ResUnet-based models' comparison (Site 1)

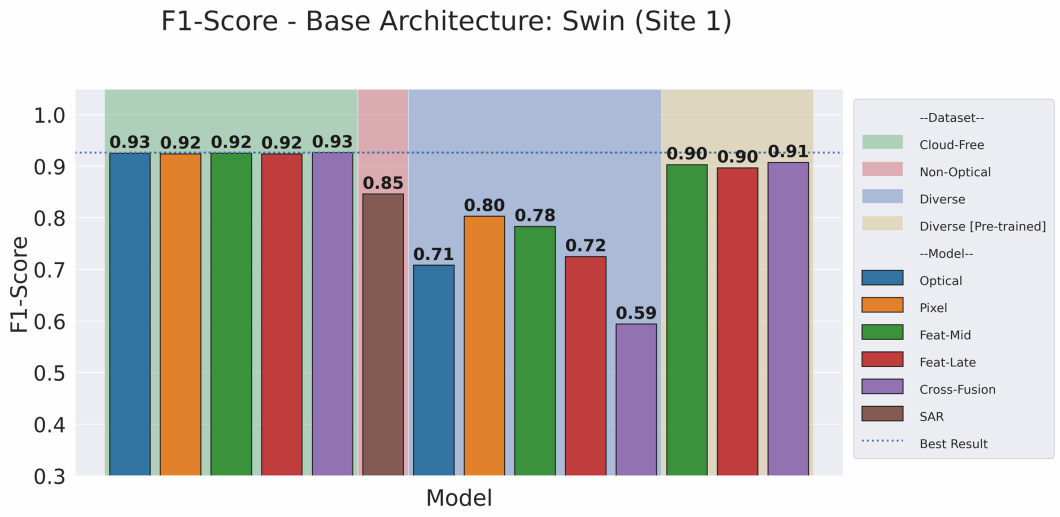


Figure 141: F1-Score for Swin-based models' comparison (Site 1)

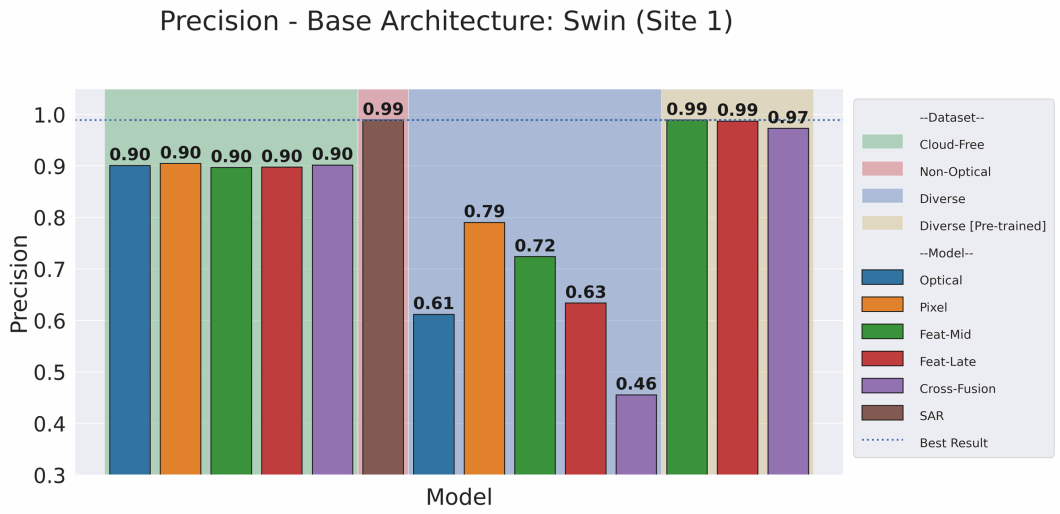


Figure 142: Precision for Swin-based models' comparison (Site 1)

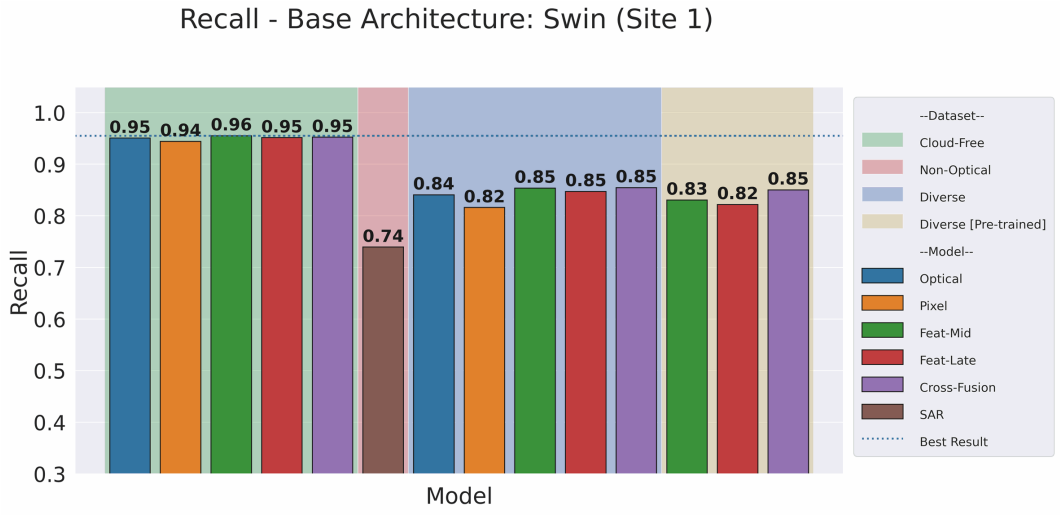


Figure 143: Recall for Swin-based models' comparison (Site 1)

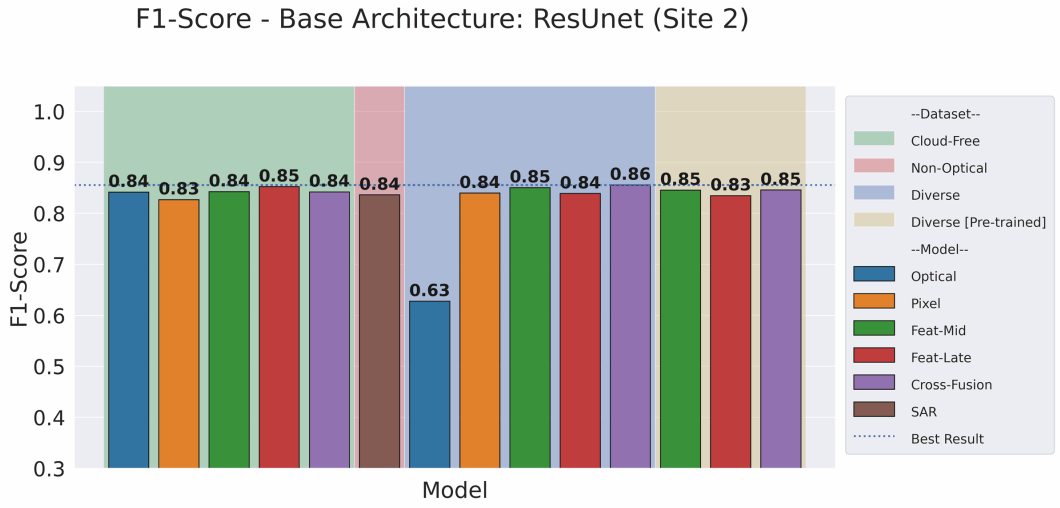


Figure 144: F1-Score for ResUnet-based models' comparison (Site 2)

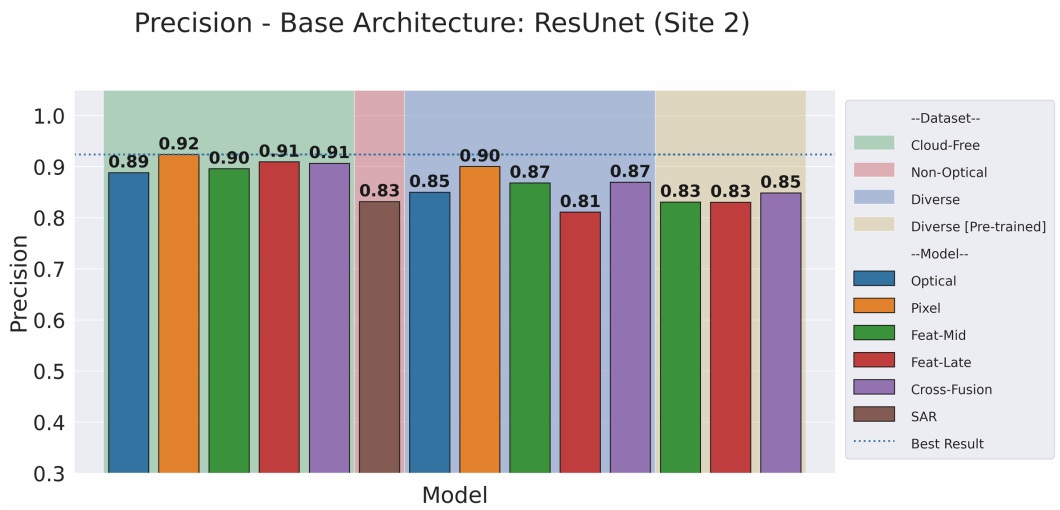


Figure 145: Precision for ResUnet-based models' comparison (Site 2)

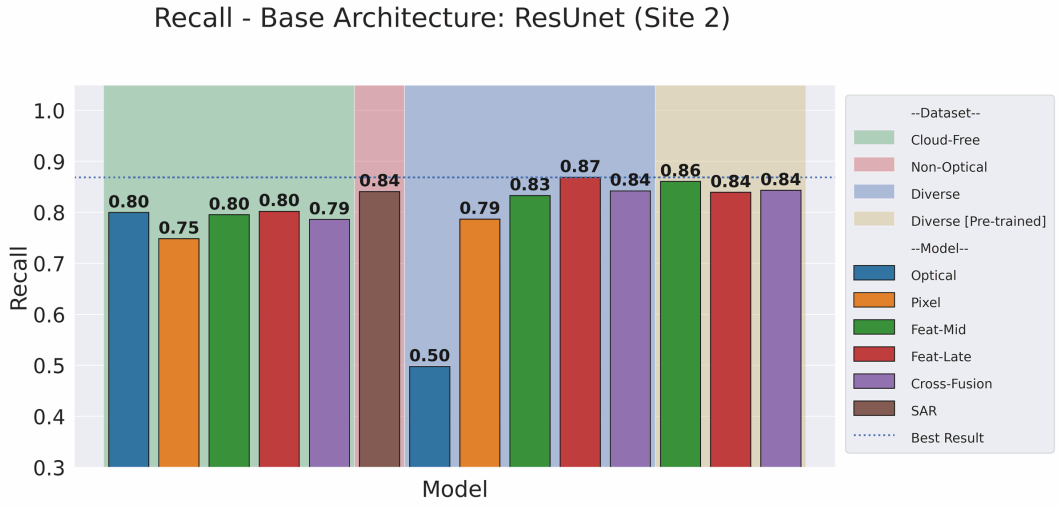


Figure 146: Recall for ResUnet-based models' comparison (Site 2)

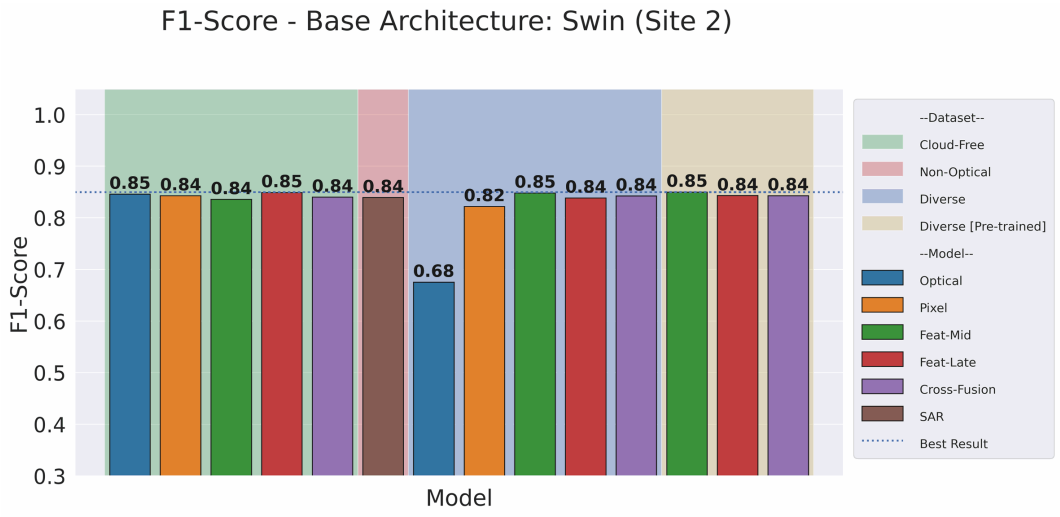


Figure 147: F1-Score for Swin-based models' comparison (Site 2)

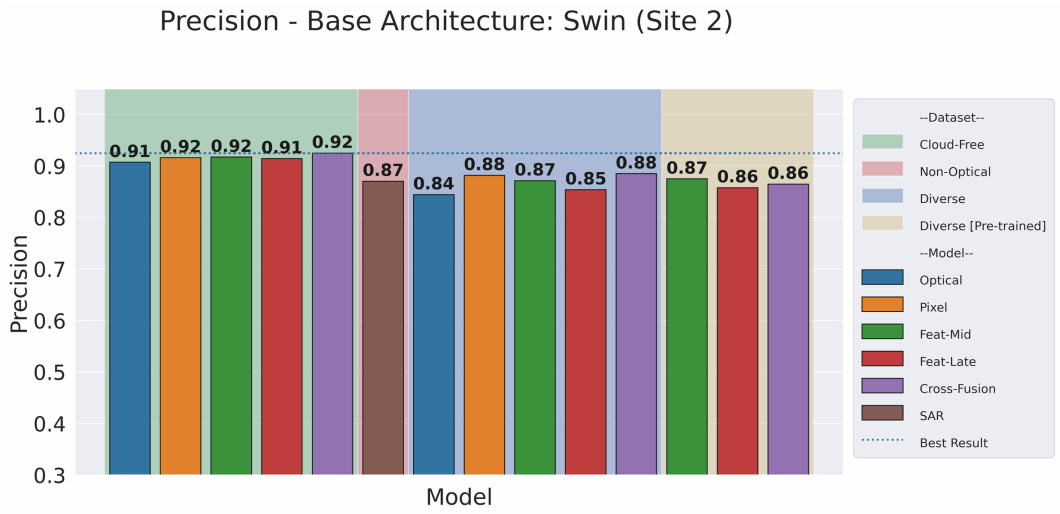


Figure 148: Precision for Swin-based models' comparison (Site 2)

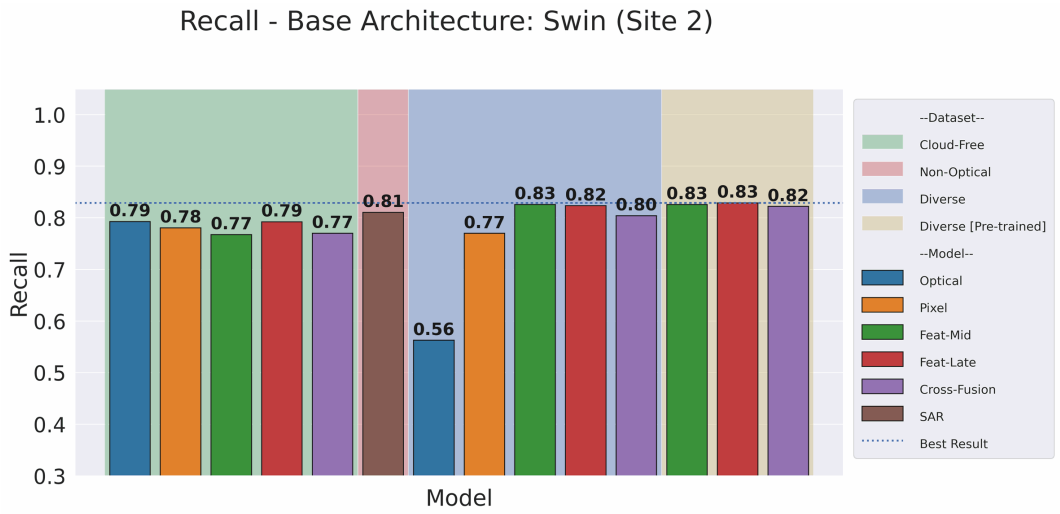


Figure 149: Recall for Swin-based models' comparison (Site 2)

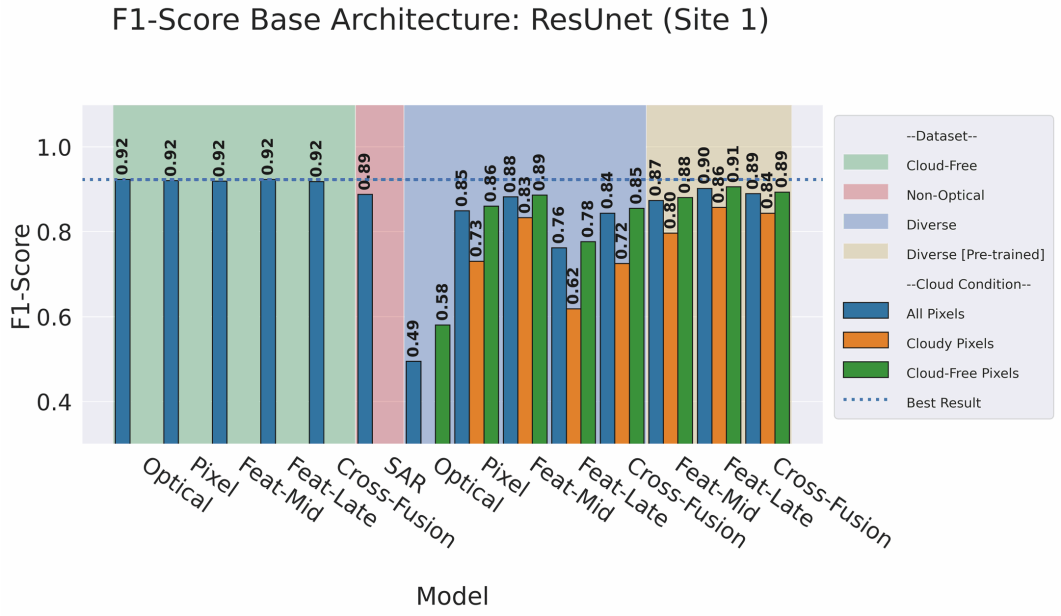


Figure 150: F1-Score for ResUnet-based models' cloud effect comparison (Site 1)

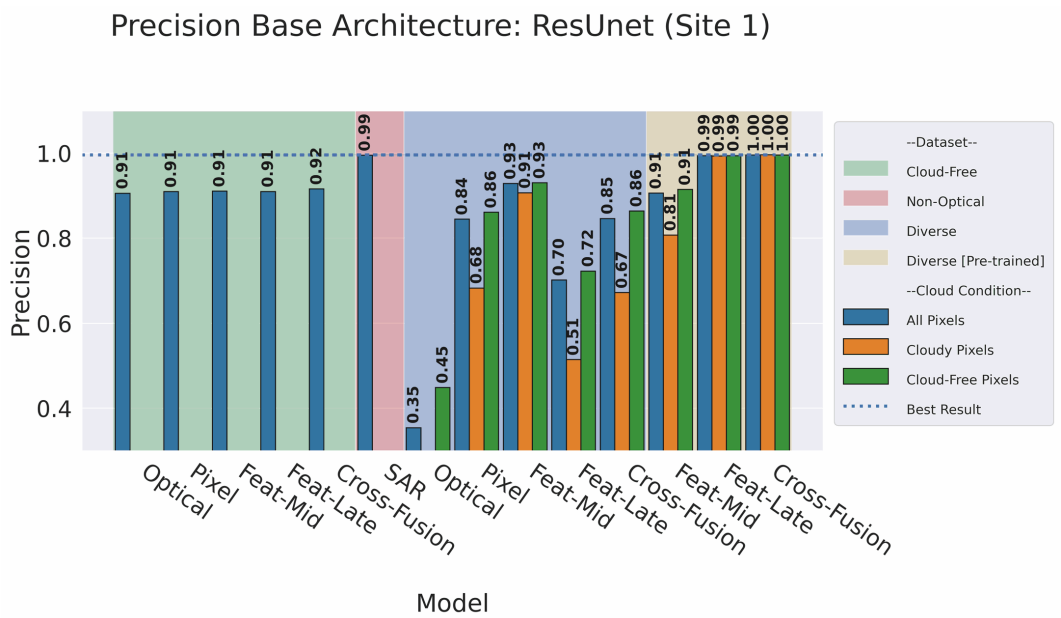


Figure 151: Precision for ResUnet-based models' cloud effect comparison (Site 1)

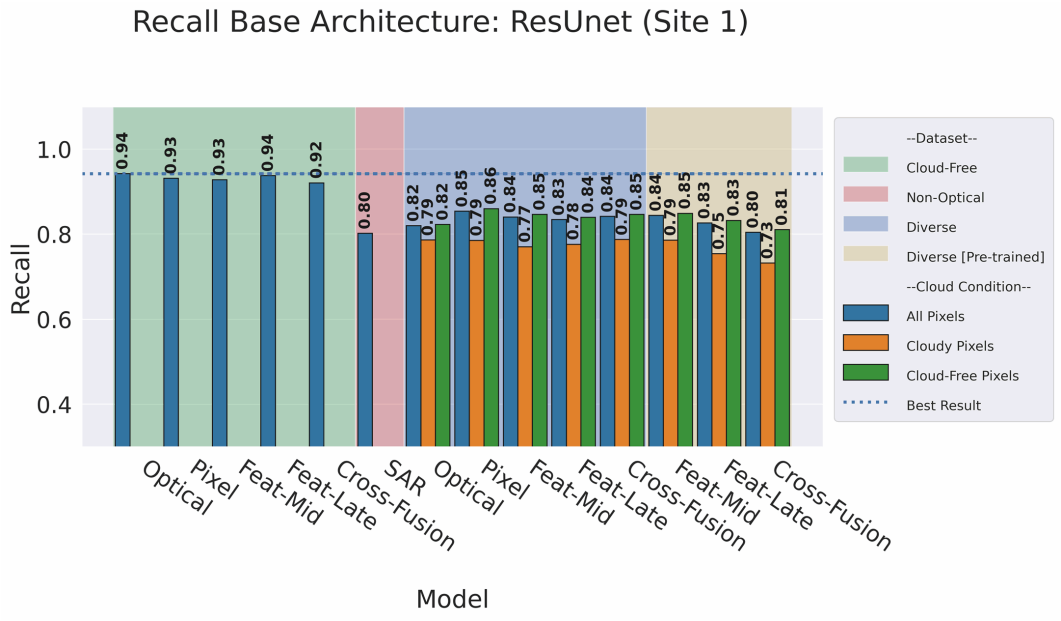


Figure 152: Recall for ResUnet-based models' cloud effect comparison (Site 1)

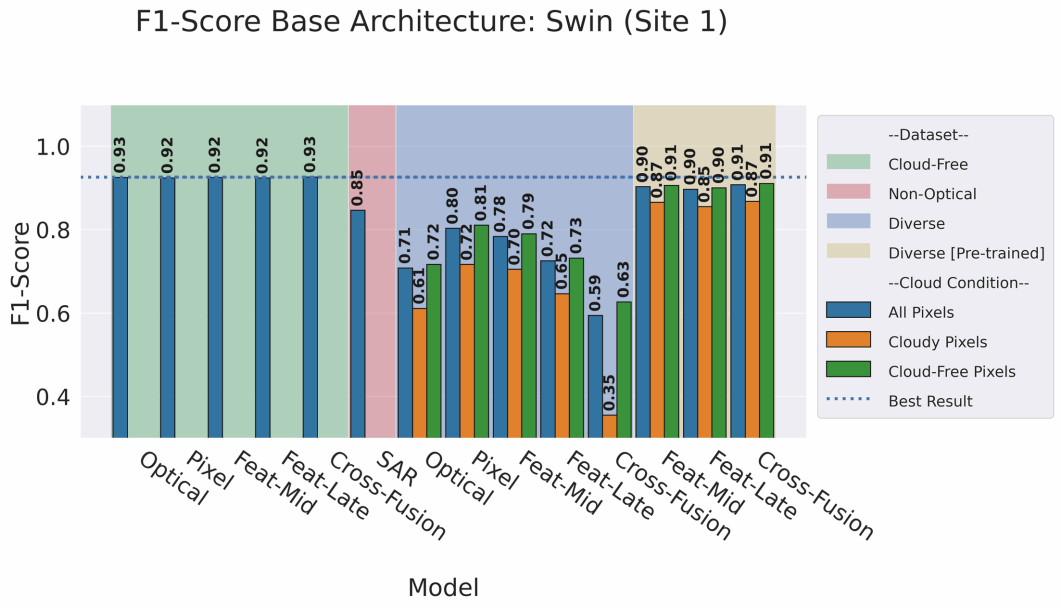


Figure 153: F1-Score for Swin-based models' cloud effect comparison (Site 1)

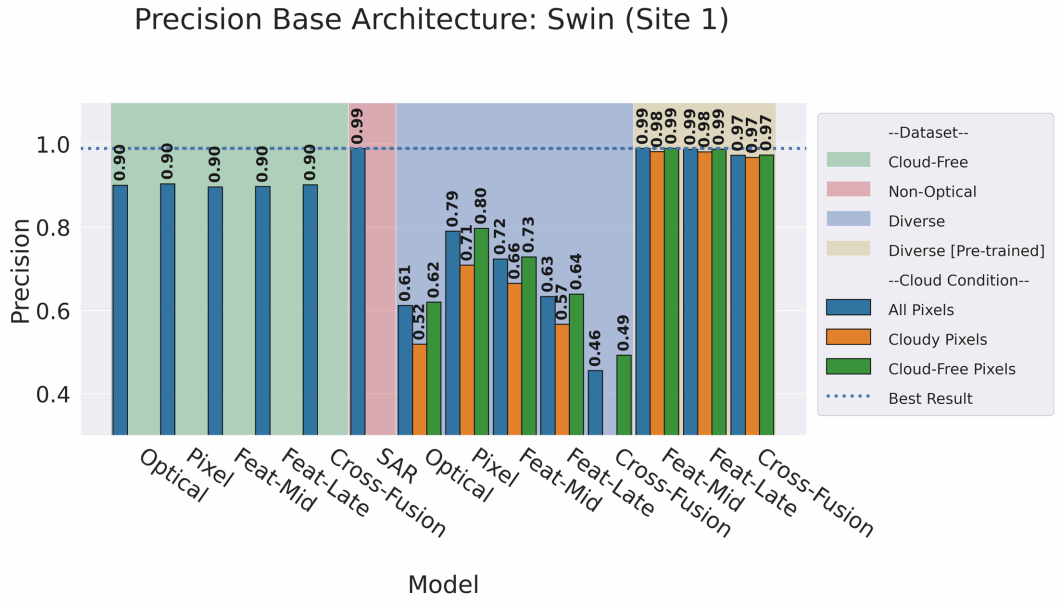


Figure 154: Precision for Swin-based models' cloud effect comparison (Site 1)

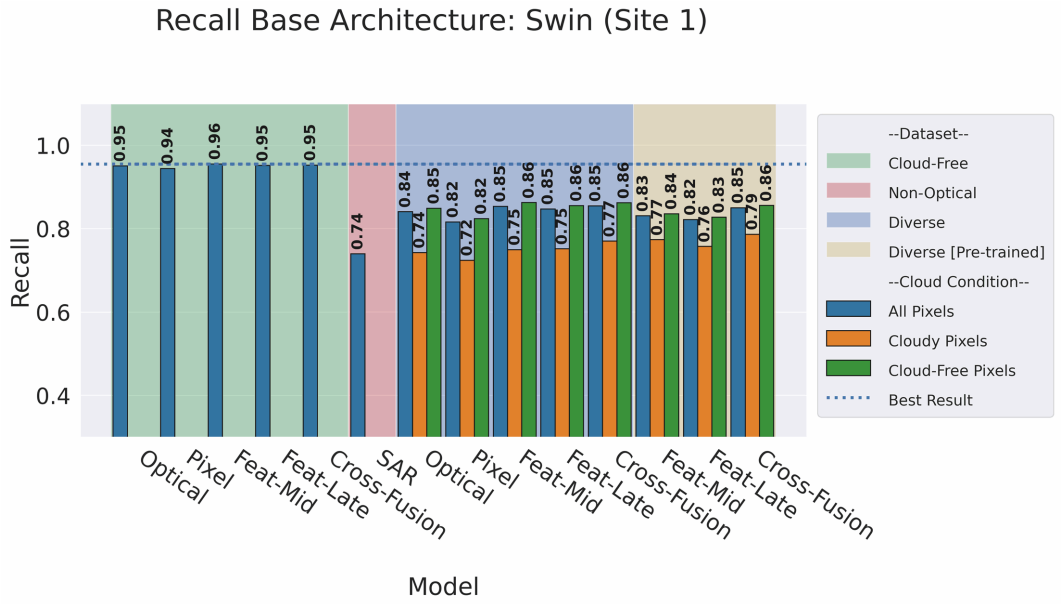


Figure 155: Recall for Swin-based models' cloud effect comparison (Site 1)

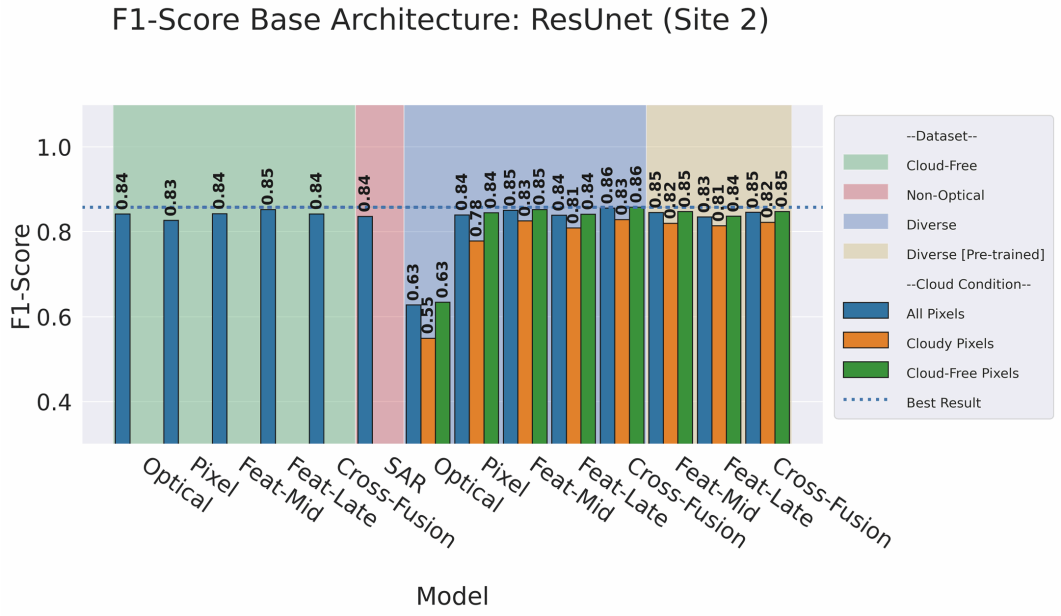


Figure 156: F1-Score for ResUnet-based models' cloud effect comparison (Site 2)

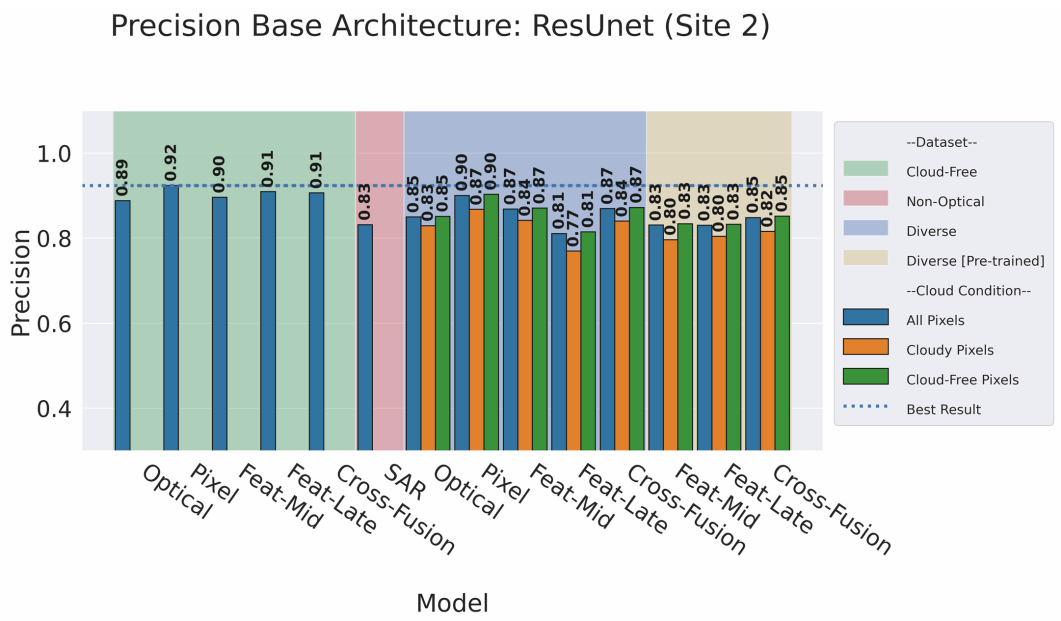


Figure 157: Precision for ResUnet-based models' cloud effect comparison (Site 2)

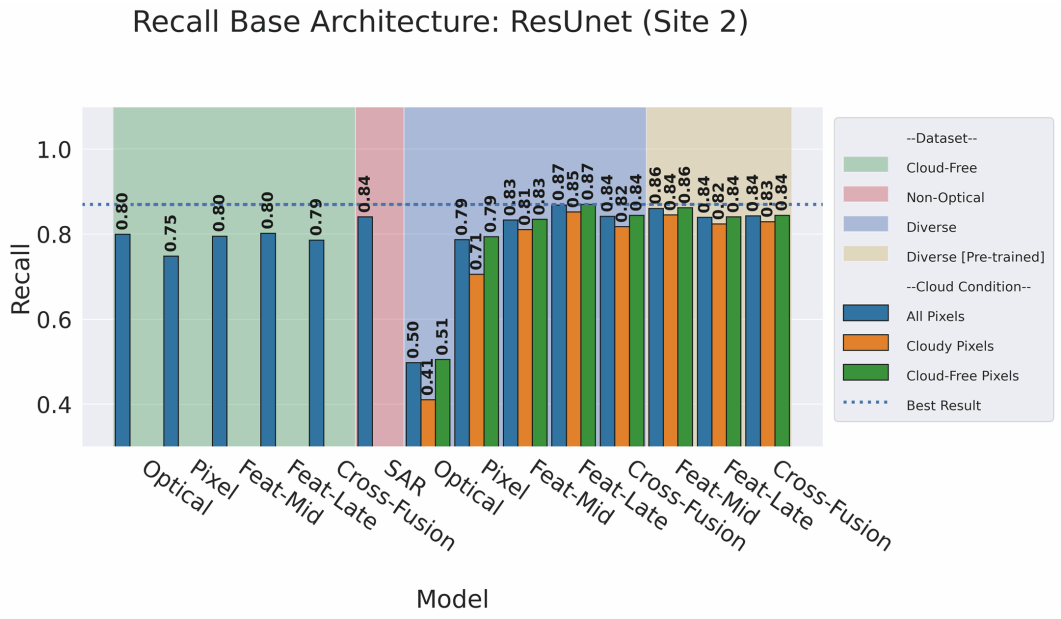


Figure 158: Recall for ResUnet-based models' cloud effect comparison (Site 2)

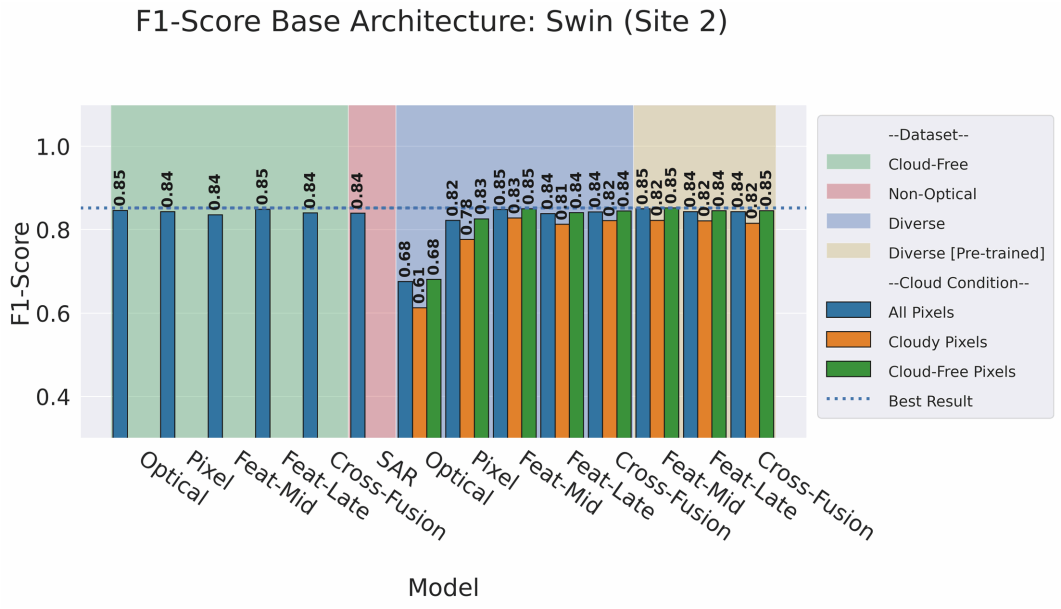


Figure 159: F1-Score for Swin-based models' cloud effect comparison (Site 2)

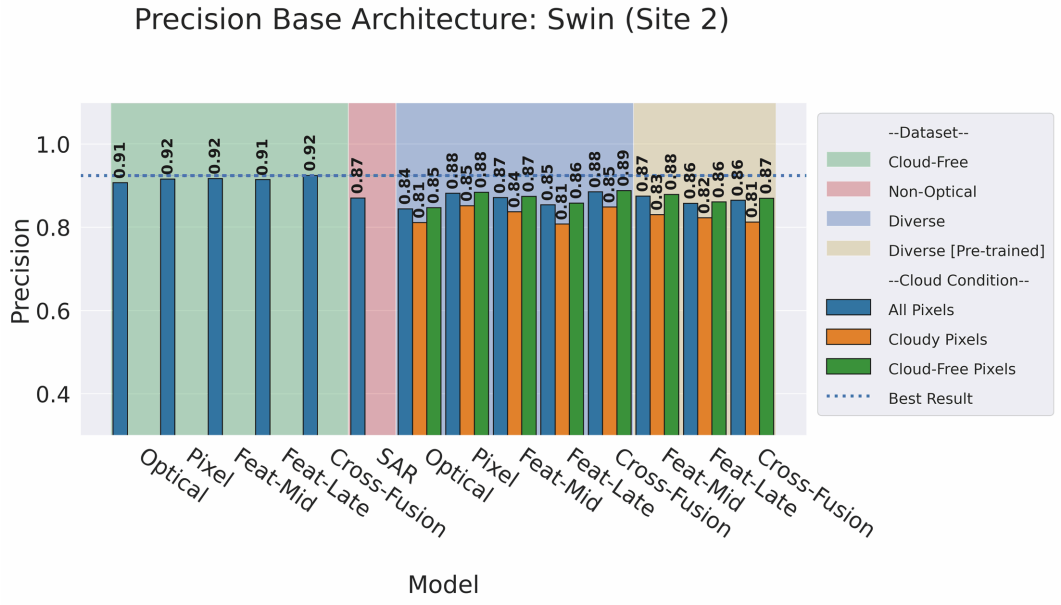


Figure 160: Precision for Swin-based models' cloud effect comparison (Site 2)

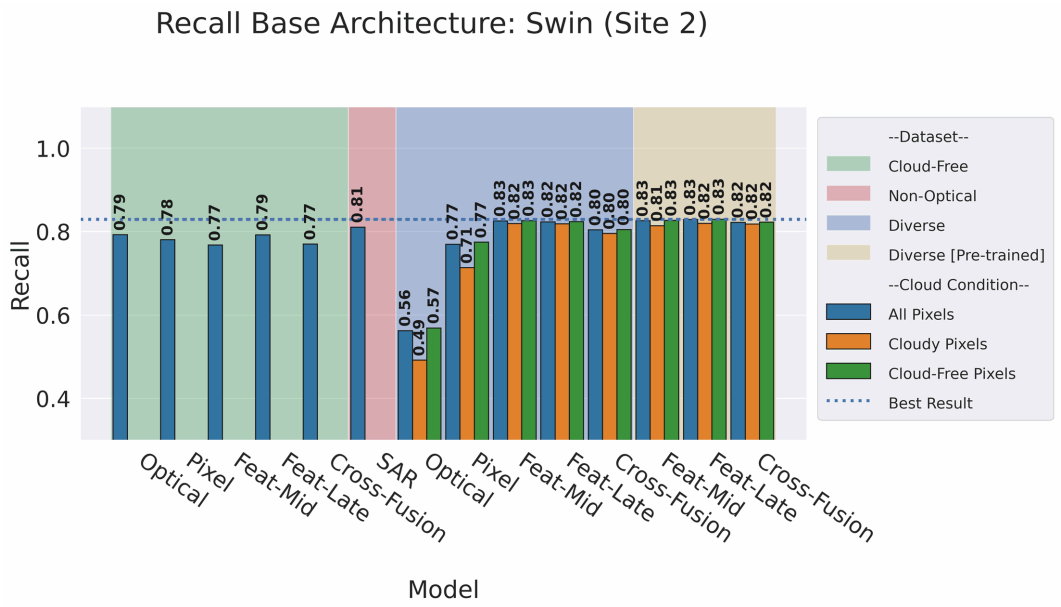


Figure 161: Recall for Swin-based models' cloud effect comparison (Site 2)