Pontifícia Universidade Católica
do Rio de Janeiro

Javier Noa Turnes

Atrous cGAN for SAR to optical image translation

Dissertação de Mestrado

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

Advisor : Prof. Raul Queiroz Feitosa
Co-advisor: Dr. Patrick Nigri Happ

Rio de Janeiro
August 2020

Pontifícia Universidade Católica
do Rio de Janeiro

Javier Noa Turnes

Atrous cGAN for SAR to optical image translation

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

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

Dr. Patrick Nigri Happ
Co-advisor
Departamento de Engenharia Elétrica – PUC-Rio

Dr. Leonardo Alfredo Forero Mendoza
Universidade do Estado do Rio de Janeiro – UERJ

Dr. Wesley Nunes Gonçalves
Universidade Federal do Mato Grosso do Sul – UFMS

Rio de Janeiro, August the 12th, 2020

Javier Noa Turnes

The author received his engineering degree in Telecommunications and Electronic Engineering at the Universidad de Oriente (UO) in 2016.

Bibliographic data

Noa Turnes, Javier

Atrous cGAN for SAR to optical image translation / Javier Noa Turnes ; advisor: Raul Queiroz Feitosa ; co-advisor: Patrick Nigri Happ. – 2020.
71 f. : il. color. ; 30 cm

Dissertação (mestrado)–Pontifícia Universidade Católica do Rio de Janeiro, Departamento de Engenharia Elétrica.
Inclui bibliografia

1. Engenharia Elétrica – Teses. 2. Redes Adversárias Generativas Condicionais;. 3. Convolusões Atrous;. 4. Radar de Abertura Sintética.. I. Feitosa, Raul Queiroz. II. Happ, Patrick Nigri. III. Pontifícia Universidade Católica do Rio de Janeiro. Departamento de Engenharia Elétrica. IV. Título.

CDD: 621.3

To my parents, brothers, and wife.

Acknowledgments

To my parents, Jorge and Josefa, and my brothers, Jorgito and Jardi; thanks for your support.

To my wife Daliana Lobo Torres, thanks for the love and comprehension.

To my advisor Raul Queiroz Feitosa, for being an excellent professor and guide.

To my co-advisor Patrick Nigri Happ and his valuable contributions.

To Pedro Soto, Brenda, and the LVC team.

To the excellent performance of the Department of Electrical Engineering of the PUC-Rio.

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

Abstract

Noa Turnes, Javier; Feitosa, Raul Queiroz (Advisor); Happ, Patrick Nigri (Co-Advisor). **Atrous cGAN for SAR to optical image translation**. Rio de Janeiro, 2020. 71p. Dissertação de mestrado – Departamento de Engenharia Elétrica, Pontifícia Universidade Católica do Rio de Janeiro.

The capture of land cover scenes with optical satellite sensors is often constrained by the presence of clouds that corrupt the collected images. Among the methods for recovering satellite optical images corrupted by clouds, several image to image translation approaches using Generative Adversarial Networks (GANs) have emerged with profitable results, managing to create realistic optical images from Synthetic Aperture Radar (SAR) data. Conditional GAN (cGAN) based methods proposed so far for SAR-to-optical image synthesis tend to produce noisy and unsharp optical outcomes. In this work, we propose the *atrous-cGAN*, a novel cGAN architecture that improves the SAR-to-optical image translation. The proposed generator and discriminator networks rely on atrous convolutions and incorporate the Atrous Spatial Pyramid Pooling (ASPP) module to enhance fine details in the generated optical image by exploiting spatial context at multiple scales. This work reports experiments carried out to assess the performance of *atrous-cGAN* for the synthesis of Landsat images from Sentinel-1A data based on four public datasets. The experimental analysis indicated that the *atrous-cGAN* overcomes the classical *pix2pix* model as a feature learning tool for semantic segmentation. The proposal also generates higher visual quality images, in general with higher similarity with the true optical image.

Keywords

Conditional Generative Adversarial Networks; Atrous Convolutions; Synthetic Aperture Radar.

Resumo

Noa Turnes, Javier; Feitosa, Raul Queiroz; Happ, Patrick Nigri. **Atrous cGAN para tradução de imagens SAR à ótica**. Rio de Janeiro, 2020. 71p. Dissertação de Mestrado – Departamento de Engenharia Elétrica, Pontifícia Universidade Católica do Rio de Janeiro.

A captura de cenas de cobertura da Terra com sensores óticos de satélite é freqüentemente limitada pela presença de nuvens que corrompem as imagens coletadas. Entre os métodos para recuperar imagens óticas de satélite corrompidas por nuvens, várias abordagens de tradução de imagem-imagem usando Redes Adversárias Generativas (GANs) têm surgido com bons resultados, conseguindo criar imagens óticas realistas a partir de imagens de Radar de Abertura Sintética (SAR). Os métodos baseados em GANs condicionais (cGAN) propostos até agora para a síntese de imagens SAR-óticas tendem a produzir imagens ruidosas e com pouca nitidez. Neste trabalho, propomos a *atrous-cGAN*, uma nova arquitetura que melhora a transformação de imagem SAR em ótica. As redes propostas para o gerador e discriminador contam com convolusões dilatadas (*atrous*) e incorporam o módulo Pirâmide Espacial Atrous Pooling (ASPP) para realçar detalhes finos na imagem ótica gerada, explorando o contexto espacial em várias escalas. Este trabalho apresenta experimentos realizados para avaliar o desempenho da *atrous-cGAN* na síntese de imagens Landsat a partir de dados Sentinel-1A, usando quatro bases de dados públicas. A análise experimental indicou que a *atrous-cGAN* supera o modelo clássico *pix2pix* como uma ferramenta de aprendizado de atributos para segmentação semântica. A proposta também gera imagens com maior qualidade visual, e em geral com maior semelhança com a verdadeira imagem ótica.

Palavras-chave

Redes Adversárias Generativas Condicionais; Convolusões Atrous; Radar de Abertura Sintética.

Table of contents

List of figures

List of tables

1
Introduction

Remote Sensing has provided for a long time regularly detailed information about land cover, which is essential to make decisions over different human activities like agriculture and urban occupation, and also to counteract deforestation and the effects of disasters such as fire, floods, and landslides. The satellite imagery has enabled the development of a large number of methods to determine what covers the Earth's surface accurately. However, the acquisition of optical images faces the uncontrollable drawback of the weather conditions, and is usually affected by the presence of clouds.

Cloud cover is a critical problem for optical remote sensing. The brightening effect of the clouds causes partial or even total loss of the optical data [1], and therefore, reduces the number of available observations [2]. Also, the clouds project regions of shadow causing biased estimations, mistakes in land cover classification, and false detection of land cover changes [3]. Approximately 55% of the land surface is typically covered by clouds[1], mainly concentrated in tropical regions [4]. This is the scenario of most part of Brazilian territory.

Several strategies have been conceived with the aim to overcome the cloud covering problem using optical sensors. Spectral-based [5], spatial-based [6], temporal-based [7], and hybrid approaches [8] have been proposed to restore the information corrupted by clouds. These methods are mainly based on image analysis techniques, and despite their progress, they still require adjustments and only can operate under limited conditions [9]. Other works like [1] and [10] use Deep Learning (DL) approaches in an effort to eliminate the noise effect of clouds using only optical data. The problem becomes especially critical when thick clouds cover large areas. In these cases, the available data free of clouds necessarily come from distant regions, little correlated with the occluded regions, which degrades the accuracy of these methods.

The Synthetic Aperture Radar (SAR) offers an alternative to optical data, as they are hardly affected by clouds. SAR sensors can operate day and night, almost regardless of weather conditions [11]. While optical images capture the physical-chemical properties of targets, SAR images capture the "roughness" and "wetness" of the Earth's surface [12]. As a consequence, SAR

[1] https://modis.gsfc.nasa.gov/

images are harder to interpret than the optical counterpart, both for human experts and for machine learning methods. For this reason, recent years have witnessed the efforts of several research groups to synthesize optical data of areas covered by clouds from SAR data of the same regions [13].

In the past decade, Deep Learning (DL) became the dominant trend in image data analysis, mostly due to the capacity of DL models to learn discriminative features directly from data [14]. The Generative Adversarial Network (GAN) [15] is a DL based method widely used in countless applications such as image inpainting [16] and domain translation [17], with outstanding results. The so-called image-to-image translation methods exploit GANs' ability to generate samples of complex probability distributions to synthesize optical images from SAR data over cloud-covered areas [11, 13, 18–20].

The authors of [21] introduced a method for image-to-image translation called *pix2pix*, which takes an image from one domain as input and produces a realistic version of that image in another related domain. The *pix2pix* relies on a GAN variant called conditional GANs (cGANs) [22]. Thanks to its success, *pix2pix* was later adapted to synthesize cloud-free multispectral optical images from a corresponding SAR image (e.g.,[13]). Later works proposed improvements to that model, like new terms in the objective loss function [11, 18], new training protocols [19], and also the use of temporal [20] and spectral [18] data. In all cases, the basic architectures comprising a generator and a discriminator network was maintained.

These methods managed to produce realistic optical images. Nevertheless, the spatial accuracy of synthesized images falls short when it comes to generate fine details. The *pix2pix* model has a U-Net [23] as generator that follows an encoder-decoder architecture. A well-known drawback of such architectures is the loss of spatial information as the input image is being processed by the encoder [24, 25]. This effect becomes more noticeable as the resolution of the synthesized images increases. To handle this problem, [26] proposes a new design for the generator and discriminator networks. The generator is split into several branches that process the input image at different scales. The results produced by these branches are concatenated at a certain point and later processed up to the output. Also, the discriminator operates at different scales. Thus, this cGAN design operates with several receptive fields, which improves the spatial consistency of the generated images.

Recently, atrous convolution was introduced in the DeepLab framework as a way to improve the spatial accuracy in semantic segmentation tasks [24, 27–29]. This operation allows to adjust the receptive field of convolutional neural networks without increasing the number of parameters and the computational cost. In [30], the atrous convolution is used in GANs for facial attribute transfer, producing results with finer details without using scaled versions of the input image. Also in [31], the contextual information is exploited in cGANs for natural image matting problems [32].

Inspired by this scenario, this dissertation proposes a new cGAN architecture for SAR-to-optical image synthesis, aiming to improve the spatial accuracy of the generated optical images. To this end, we introduce a novel cGAN architecture, from now on called *atrous-cGAN*, which incorporates atrous convolutions proposed in the DeepLabv3 [28] framework. We hypothesize that the use of information at different scales sharpens the fine details of optical images generated from SAR images by a cGAN. The resulting synthesized optical images are evaluated according to three criteria: visual quality, the similarity with real optical data, and as a feature learning tool.

We validated the proposed cGAN architecture in four sets of remote sensing data from regions in Brazil, usually covered by clouds: two of them for crop mapping and two for detecting deforestation.

1.1 Objectives

- **General Objective:**
  Design a new cGAN architecture for SAR-to-optical image translation, capable of generating fine details on the synthesized optical images.
- **Specific Objectives:**

  1. Investigate how atrous convolutions can be incorporated into the cGANs generator and discriminator network architectures for the synthesis of optical images from SAR data, aiming to improve the quality of generated images.
  2. Assess how much the exploitation of contextual information at multiple scales can improve the spatial accuracy of images generated by a cGAN for SAR-to-optical image translation.
  3. Evaluate the resulting proposed cGAN design for SAR-to-optical image translation on public datasets dedicated to crop mapping and deforestation detection.

1.2 Contributions

The contributions of this dissertation are the following:

1. A novel cGAN generator and discriminator architecture for SAR-to-optical image translation.

2. An experimental assessment of the proposed cGAN for the synthesis of Landsat images from Sentinel-1A data using four public datasets.

1.3 Organization of the manuscript

Chapter 2 makes a review of the available researches in the literature related to the cloud removal problem in remote sensing optical images.

Chapter 3 exposes the fundamentals in which the development of this dissertation is supported.

Chapter 4 explains the proposed *atrous-cGAN* method to improve the optical image synthesis incorporating multi-context information.

Chapter 5 describes the experimental methodology to evaluate the proposed method. It also presents the datasets, the implementation details, the evaluation metrics, and the results discussion.

Chapter 6 summarize the conclusions drawn from the experiments and announces future directions of the research.

2 Related Works

This chapter presents a set of researches related to the recovery of optical images affected by clouds. The different approaches are mainly oriented to solve two problems: the presence of thin clouds, which are responsible for partial data occlusion, and the presence of thick clouds, which cause total data occlusion. Thin clouds are generally easier to remove, but harder to detect. On the other hand, the approaches referred to thick clouds are more complex, and in general, also encompass the removal of thin clouds.

The thin or semitransparent clouds allow light to pass through them, producing a scattering reflection and resulting in blurring and low contrast landscape acquisitions. Regions covered by thin clouds present both atmosphere and ground information [3], which generally makes the recovery process easier, but complicates the cloud detection. Given their similarity with haze, fog, and smoke, it can be assumed that thin clouds produce a noisy effect, as shown in the Figure 2.1. Hence, some works, like [33], treat the problem using image enhancement methods such as histogram matching [34] and color constancy enhancement [35]. The problem with these approaches is that they do not consider the physical model of clouds [36], limiting their effectiveness and resulting in cloud fragments and color distortion.

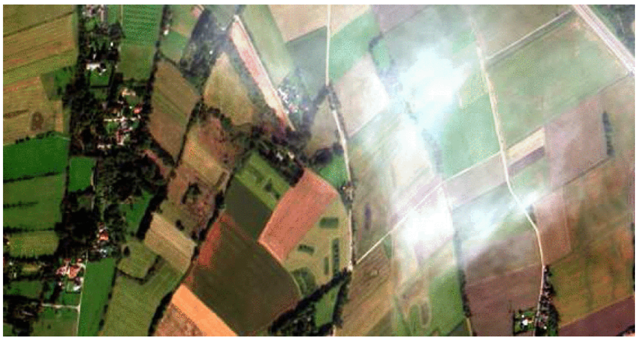


Figure 2.1: Thin cloud example, adopted from [37].

Spectral based techniques were also proposed for removing thin clouds. For example, [36] and [5] use filtering techniques considering that thin clouds occupy the low-frequency part of the image in the frequency domain. On the one hand, [36] uses a low-pass filter with a post-processing that requires previous information about the location of clouds. On the other hand, [5] uses the homomorphic filter [38], to suppress the low-frequency components, that potentially correspond to thin clouds (illumination), while amplifies the high-frequency ones, corresponding to land cover (reflectance). Some major drawbacks are that the cut-off frequency for at least one channel must be determined manually (then it is used to determine the values for the other channels semi-automatically), as well as the adjustment length related to the pixels that need to be corrected in the cloud-free image.

The thick and dense clouds are fairly easy to identify because of their high reflectance [3]. However, as Figure 2.2 shows, information of land cover can not be obtained when clouds are thick. Moreover, unlike thin clouds, the darkening effect of cloud shadows is more remarkable. In these cases, some methods take advantage of the spatial information when the image is partly contaminated by clouds. For example, [6] uses classic image inpainting techniques. Image inpainting does not consist in recovering missing regions, but in replacing them with trimmings of the unaffected neighborhood that have a close resemblance with the original image. This method assumes that the occluded regions share considerable characteristics with the clear ones. For that reason, large areas with a lot of missing information are harder to reconstruct, because the information in the clear parts of the image is not enough to infer what has been lost. Additionally, it is difficult to handle complex geometrical structures and edge continuity.

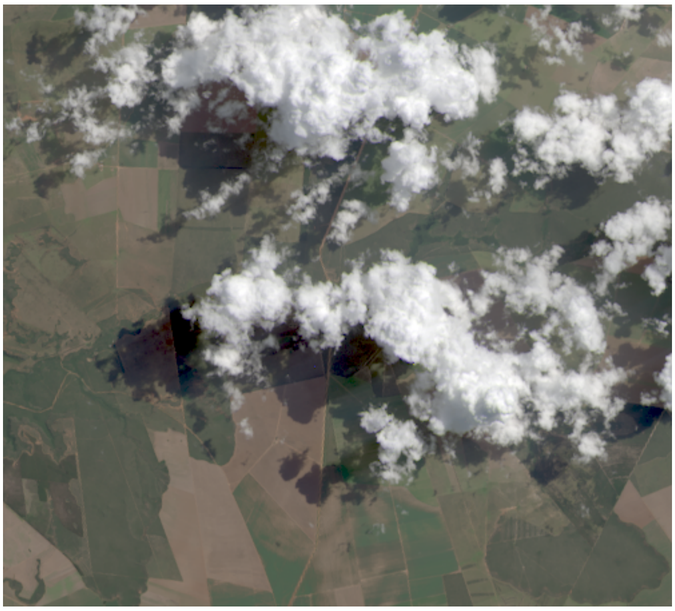


Figure 2.2: Thick cloud example.

Other approaches, like [39], use multitemporal co-registered optical images to reconstruct the corrupted pixels based on their corresponding data extracted from other dates. However, this solution presents important limitations since it is only valid under the assumption that the land cover do not change significantly in the short term and that the missing pixels are free of cloud in the other images. These methods can lead with both thin and thick clouds, but generally, they need prior knowledge of the cloud location, depending on automatic [40] or manual detection.

Recently, the popular Generative Adversarial Networks (GANs) [15] raise an unsupervised learning conception that allows to generate data from one distribution to another. Therefore, the problem of cloud cover has been widely addressed by using GANs to synthesize the missing data. In this sense, [1] proposed the Cloud-GAN method to produce cloud-free RGB images from cloudy ones. This approach is based on the Cycle-Consistent GAN (CycleGAN) [41], a particular type of GAN that learns the mapping functions $G : X \to Y$ and $F : Y \to X$ between the domains X and Y, each represented by samples of cloudy and cloud-free images respectively, that do not need to be spatially co-registered. Despite the resulting cloud-free images being quite realistic, this method is limited to thin clouds, since it requires some information of the cloudy land cover.

Other researches based on GANs use alternative data sources to synthesize optical images. The most popular is the Synthetic Aperture Radar (SAR), which is an imaging radar whose electromagnetic waves are able to pass through the clouds (both thin and thick). Therefore, the measured reflections are almost independent of weather conditions, turning SAR data into an alternative source for optical data recovery. As in [1], [42] also applies a CycleGAN architecture for cloud removal, but using SAR images as the source domain X. Thus, since SAR data do not have cloud information, the mapping function $G : X \to Y$ can produce cloud-free optical images, learning mapping functions between data from different sensors. Specifically, this method synthesizes single band panchromatic images [43], and despite the good results reported, it presents typical CycleGAN problems such as the generation of abnormal objects, specially if there are substantial geometrical changes between the two domains.

The CycleGANs emerged as an extension of the conditional Generative Adversarial Networks (cGANs) [22] for the image to image translation problem [21]. Unlike the CycleGANs, the cGANs perform only a mapping function $G : X \to Y$, not the other way around. To achieve this, the synthesis is made under conditions imposed by samples of the domain X. The disadvantage of the cGANs is that they require paired data between the domains, limiting the number of available samples. Nonetheless, for some problems like the SAR-optical synthesis, this approach may be desirable, whenever it is possible to obtain paired samples. For example, in [9] cGANs are used for cloud removal in multispectral images, and the pairing is guaranteed by the co-registered SAR and cloud-free optical images. The proposal of [9] is based on the *pix2pix* framework [21], a successful network for image to image translation. The resulting optical image syntheses turned out to be more representative than the directly usage of SAR images for classifying agricultural crops in images covered by thick clouds.

This work was later extended in [13], where the influence of additional temporal information is exploited as conditioning data. Specifically, the optical image is synthesized using its SAR counterpart and a SAR/cloud-free optical pair from a different date. The method is applied in regions of agricultural crops using dates of the same season, but a year apart assuming that the same classes and their seasonal characteristics are present in both dates. It was also tested to restore cloudy optical images to detect wildland fires. Effectively, the incorporation of temporal information improved the results, but a gap is still noticeable between synthesized and real optical images.

Another adaptation of the *pix2pix* framework using SAR images was presented by [19]. The approach proposes the use of cGANs in two steps. The first step consists on translating the SAR image to its cloud-free optical counterpart using a pre-trained cGAN. The second step uses this synthetic optical image and the current SAR and cloudy optical images, for training another cGAN to generate the final cloud-free optical image. The authors argue that the synthesis of a single SAR-optical cGAN presents coarse texture and low spectrum accuracy, and therefore does not achieve a satisfactory quality. Conversely, they also claim that this first synthesis is useful as an input for the second step to refine the image generation. The first cGAN is pre-trained with cloud-free data from another time in the same region, and the whole method uses both the SAR image and the cloudy image to get the cloud removal result.

Another approach for cloud removal based on cGANs and SAR-to-optical image translation is proposed by [18]. The authors state that the common presence of noise in SAR data (mainly speckle noise) can be transferred to the generated image and, to mitigate this effect, they incorporate the frequency consistency, trying to match the spectrum of fake and real optical images. The results show a slight improvement to generate line patterns, and as expected, a less noisy synthesis. However, this method was only tested using single channel SAR and optical images, so, the detail generation could be different using optical images with several spectral bands.

We can find other approaches in the literature to deal with the cloud covering problem in optical images, but most of them are closely related to the aforementioned proposals. In fact, the use of cGANs have been established as the state of the art for this application. In this work, we take the *pix2pix* based approaches as the starting point to propose a new approach to improve the SAR-optical synthesis with cGANs. Thus, in the following chapter, the foundations on which we base our study are addressed.

3 Fundamentals

This chapter presents the theory on which the proposal of this dissertation is based. Initially, an overview of Remote Sensing concepts is posed to contextualize the cloud covering problem. Next, the basics of convolutional neural networks are described, targeting the exploitation of context information. And finally, an explanation of the generative models is given focusing on the image to image translation.

3.1 Remote Sensing

Remote sensing (RS) is the process of detecting and monitoring an object, area, or phenomenon through the analysis of data captured by a remote device [44]. The RS development is mainly oriented to the Earth Observation Science (EOS), which basically collects data measuring the electromagnetic energy emanated from the Earth's surface to provide information about resources under investigation. Those measurements are made by sensors mounted on aircraft (airplanes, helicopters, drones, etc.) or spacecraft (satellites) platforms, and can be used to construct an image of the underneath landscape [45].

Considering the diverse RS platforms, their sensors have similar characteristics, but the different altitude and stability can lead to different image properties. For example, satellites are stable and have large coverage area at different resolution levels. However, they are not available permanently, but rather frequently, generally for land cover applications. Conversely, aircraft platforms allow to capture small areas, depending on the altitude, and can offer ultra high resolutions. But, sometimes, especially using drones, they can create blurred images due to instability.

RS sensors are able to measure energy using almost any wavelength in the electromagnetic spectrum. However, particularly for those operating on spacecraft altitudes, the Earth's atmosphere becomes a fundamental limitation, because the energy at some wavelengths is absorbed or scattered by the existent molecules and aerosols, like clouds, in the journey between earth and space [45]. Figure 3.1 shows the main regions of the electromagnetic spectrum and its correspondent atmospheric transmittance. Clearly, several segments of the spectrum are constrained by partial or total absorption of energy, limiting, for example, the visible and infrared (IR) regions, attenuating almost all ultraviolet (UV) wavelengths, and providing almost full transparency in radio waves.

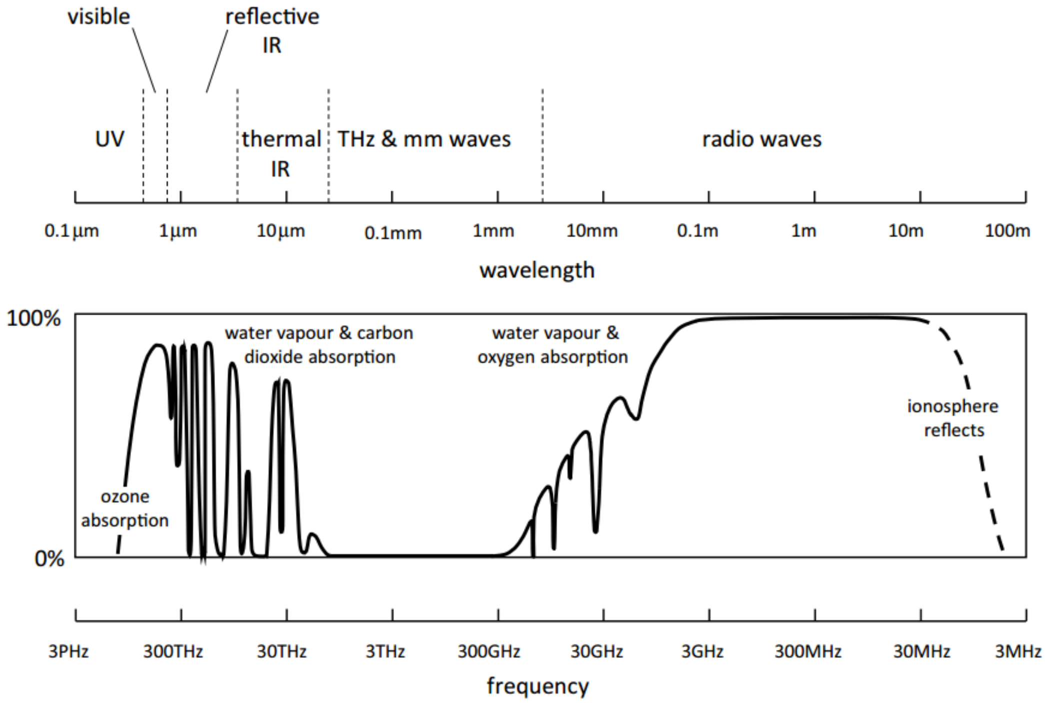


Figure 3.1: Electromagnetic spectrum and atmospheric transmittance. Adopted from [45].

To map the Earth's surface, a source of energy is needed, so that the object's reflection can be measured by the sensors. In that sense, the RS sensors are classified as *passive* or *active*. Passive sensors measure natural available energy, while active sensors supply their own source of energy to illuminate the features of interest.

3.1.1 Passive sensors

Passive sensors measure the energy emitted or reflected by an object as long as the source is not the sensor itself. The energy sources are ideally considered *blackbodies* to analyze the behavior of its emitted energy. A *blackbody* is a hypothetical physical body that spontaneously and continuously emits electromagnetic radiation according to the Planck's law [46]. Following that law, the radiation has a spectrum that is determined by the wavelength and the temperature of the body. In this way, Figure 3.2 shows the relative spectral energy measured at the surface of the Earth, taking as energy sources: i) the

Sun, and ii) the Earth itself; with about 5950K and 300K of temperature each, as principal natural energy suppliers.

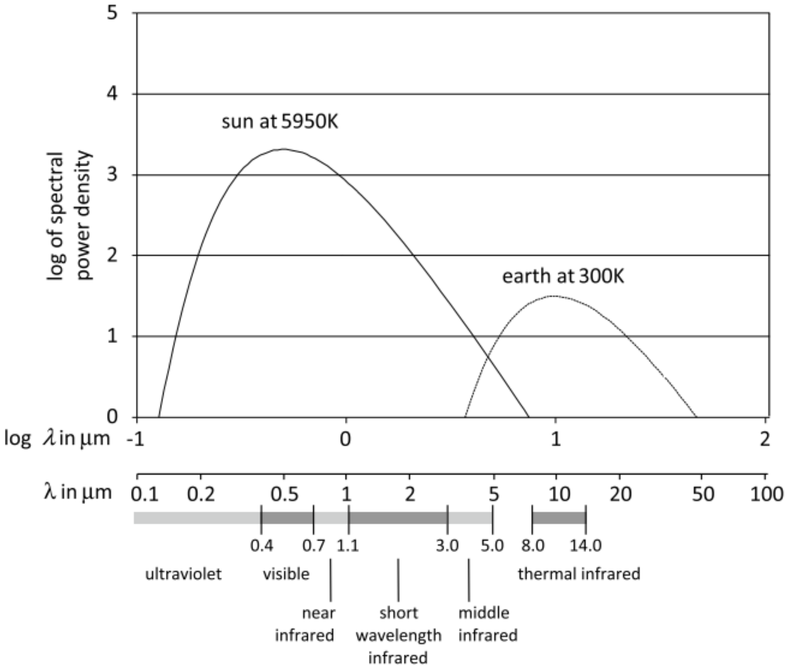


Figure 3.2: Relative levels of energy from the Sun and Earth when measured at the surface of the Earth. Modified from [45].

From Figure 3.2, we can infer that if the reflected solar radiation is measured, images can be only captured in the UV, visible and near-to-middle IR ranges of wavelengths [45]. On the other hand, if we consider the Earth as a source of energy, the acquisition can be only made in the thermal IR wavelengths. These statements justify that the passive (also called optical) RS sensors operate within the optical spectrum, that comprises wavelengths from 0.3 to 14 μm approximately. The optical images capture information about the surface of the materials [12], which is closely related to the human vision.

In the case of passive satellite sensors, almost all the measured energy is the reflected radiation of the sunlight, which practically restricts these sensors to capture images on daylight landscapes. We also know from Figure 3.1 that some wavelengths are attenuated due to atmospheric energy absorption, for example, most of the UV range. Moreover, the presence of clouds, as objects capable of reflecting the sunlight, does not allow the surface to be illuminated or to reflect enough energy; therefore, clouds severely affect the image acquisition using passive satellite sensors.

3.1.2 Active sensors

Active sensors do not need natural energy sources, instead, the energy is provided by the sensor's platform to illuminate the target object, and then, the reflected energy is measured to record the image. Recalling Figure 3.2, the energy reflected on the Earth's surface is too small (considering the logarithmic scale of the spectral power density) for wavelengths above the 50 μm, such as the microwaves or radio ranges, when the energy source is natural. In principle, only for some wavelengths above this value, it is possible to measure the energy using passive sensors. The problem is that the spatial resolution of the images tends to be low, since the pixels are necessary large to measure enough signal so that the noise does not get over the information of interest [45]. In this case, active sensors are an effective alternative for RS systems, as they can provide its own energy source and sense data at the wavelength of interest.

Examples of active sensors are the Laser Imaging Detection and Ranging (LIDAR) [47], that measures the distance between the sensor and target objects to make digital 3D representations, and the Synthetic Aperture Radar (SAR) [48], that measures physical characteristics like water content or roughness [12]. Since the purpose of this work is to contribute to the semantics of the RS images, we will approach the SAR data in more detail.

SAR uses the motion of a radar antenna to transmit successive pulses of waves over a target scene. Then, the features are captured via scattering from the surface and back to the radar instrument. The antenna produces a beam that is broad in the across-track direction, defined by the swath with, and relatively narrow in the along-track direction [45], as Figure 3.3 shows. Along the track, features are spatially located using the principle of aperture synthesis [49], which involves the use of signal processing techniques over the recorded reflections to synthesize high spatial resolution images. The electromagnetic radiation is integrated by the electric field and the magnetic field, orthogonal to each other; and the electric field direction defines the polarization of the wave. The antenna of active sensors can transmit and/or receive in horizontal (H) or vertical (V) polarization, so this is a parameter that characterizes the recorded image.

Active sensors, unlike passive sensors, can be more available for a ground trace, since they do not need daylight on the scene. The SAR sensors operate in the microwave spectrum range, allowing it to penetrate the atmosphere without the clouds reflecting the energy, so they are almost independent of weather conditions. However, SAR images are harder to interpret visually than optical images. In Figure 3.4, we can see two co-registered images captured by passive

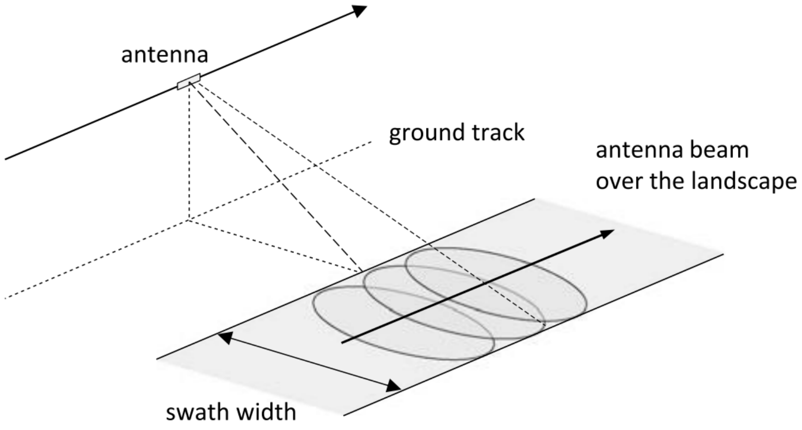


Figure 3.3: SAR system.

and active sensors respectively, in croplands of the Campo Verde municipality, Brazil. The images correspond to the same geographical area on nearby dates. Notice that both images refer to the same target region with the same objects, but in different domains.

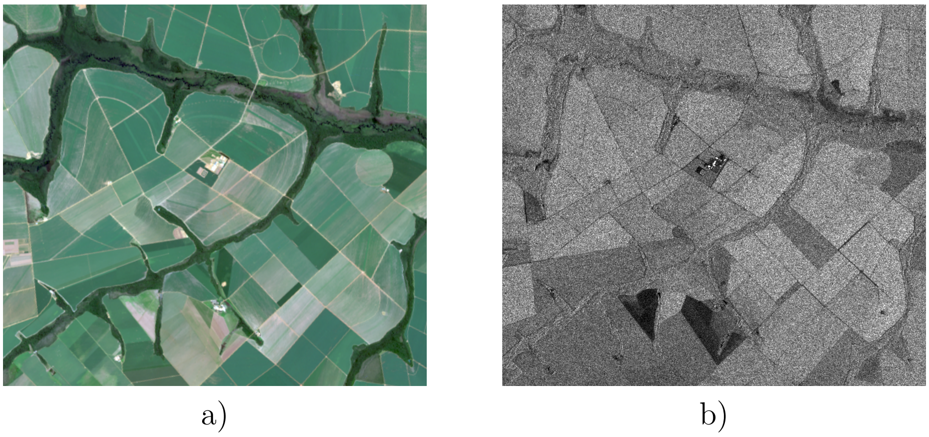

a) b)

Figure 3.4: Co-registered images of Campo Verde municipality in Mato Grosso state, Brazil. a) Passive sensor: Landsat 8 OLI, 30 m spatial resolution, RGB components. b) Active sensor: Sentinel-1A SAR (Synthetic Aperture Radar), 10 m spatial resolution, VH polarization (the first letter refers to the polarization emitted and the second to the received).

3.2 Convolutional Neural Networks

A Convolutional Neural Network (CNN) is a class of deep neural networks applied in general to analyze images or data arranged in two dimensions [50]. It was introduced in [51], and it is basically a neural network that uses convolution in at least one of its layers, see the Figure 3.5 example. The idea of the convolution was inspired by biological researches [52, 53], based on that only certain neurons in the brain are activated when the eye sees certain patterns, such as vertical or horizontal edges. In addition, the neurons are specialized and hierarchically organized so that the receptive field of single neurons contributes to the entire perception [54].

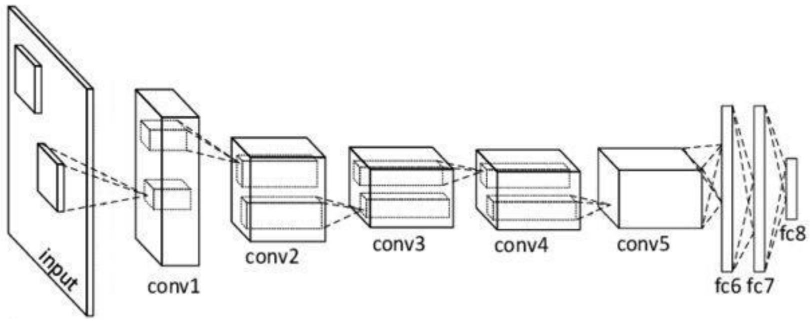
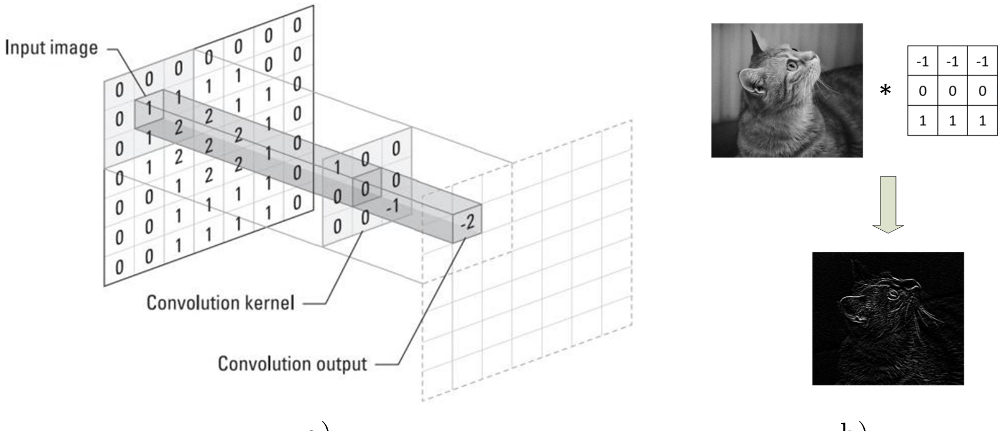


Figure 3.5: Generic scheme for a CNN composed of the input image, five convolutional layers, two fully connected layers and an output layer.

As Figure 3.6 shows, the convolution mathematically applies the aforementioned idea using linear operations. It performs a sliding dot product or cross-correlation between the input image and a small matrix called *kernel*, to extract two dimensional features in a filtering process. The convolution kernel actually has as many dimensions as the input, and operates simultaneously across the input channels to generate a feature map. The number of kernels in a layer defines the number of channels of the output feature map. When the kernel slides over the image, it can skip a certain number (called *stride*) of pixels, allowing to downsample the input.



a) b)

Figure 3.6: Convolution operation. a) Sliding dot product. b) Convolution of a grayscale image to extract horizontal edges.

The CNNs are gradient-based optimization methods that pursue to minimize the value of an objective loss function adjusting the parameters of the network. The parameters are all the neurons weights and their respective biases. Besides convolution, other operations have been developed to improve the efficiency of the networks, structuring different types of layers. The most used are described in the following:

- **Activation function.**
  As mentioned before, the convolutions perform linear operations. In order to make the network a universal function approximator [55], nonlinear functions such as *tanh*, *Rectified Linear Unit (ReLU)*, and *Leaky-ReLU* are used in between convolutional layers. In general, the activation functions must be monotonic and continuously differentiable. The *ReLU* and *Leaky-ReLU* are not continuously differentiable, but it is possible to use them thanks to frameworks that handle that issue. Other activation functions are the *sigmoid* and *softmax*, that are generally used as output layers in classification tasks.
- **Pooling.**
  The pooling allows to downsample the feature maps by combining the outputs of neuron groups at one layer into a single neuron in the next layer. The groups are defined by a window that slides upon each feature map separately applying a stride. With a stride > 1, the feature map resolution is reduced and, therefore, computational cost is saved. The most common pooling operations are Maximum Pooling (Max-Pooling) [56] and Average Pooling that summarize the maximum and average activation inside the window, respectively. According to [50], pooling also helps to make the feature representation invariant to small translations in the input.
- **Batch Normalization.**
  Batch Normalization was introduced in [57] to increase the stability of neural networks arguing that it avoids the internal covariate shift in the hidden layers. The operation relies on normalize the output of the previous activation layer by subtracting the mean and dividing it by the standard deviation. The resulting values are then denormalized using a local mean and the standard deviation learned during the training. According to [58], the batch normalization smooths the objective loss function, so it improves the gradient flow through the network during the backpropagation [59]. Thus it allows higher learning rates. It also reduces the dependence on the initialization and allows each layer to learn more independently from the others.
- **Dropout.**
  Dropout is a regularization method for reducing the overfitting in neural networks by avoiding complex co-adaptations on training data [60]. The

overfitting happens mainly due to training large networks with relatively small datasets. Applying dropout to a layer means to randomly ignore certain neurons, such that the set of parameters to update change on each iteration. The number of neurons to be dropped out in the layer is defined by a predefined fraction (*rate*). In this way, the dropout forces the network to learn redundant representations and also to make the training process noisy in order to generalize better during the test.

– **Fully Connected Layer.**
This layer connects each neuron with all the activations of the previous layer. The outputs are calculated only with a matrix multiplication and the sum of biases. It has high consumption of computation and memory.

3.2.1
Fully Convolutional Neural Networks

A Fully Convolutional Neural Network (FCN) is a CNN that does not use fully connected layers and operates with inputs of any size to produce an output of the corresponding (maybe resampled) spatial dimensions [61]. The FCNs have been mainly developed to produce pixel-wise predictions in applications like semantic segmentation or image synthesis, since the traditional CNNs, inferring over the image central pixel, are inefficient and less accurate for these tasks.

FCNs generally have a contraction stage called *encoder* that reduces the size of feature maps, and an expansion stage called *decoder* that progressively returns the features to the desired size. Figure 3.7 shows a roughly design of the U-Net [23] architecture, one of the most stable and successful FCNs, where we can see the encoder-decoder structure. The encoder reduces the features dimensionality using striding convolutions or pooling. For upsampling the features in the decoder there are several techniques such as interpolation, *unpooling* methods (e.g., Max-Unpooling [62]) and transposed convolutions. Specially the transposed convolutions [63] allow to upsample the feature maps optimally using learnable parameters.

Figure 3.7 also shows the use of the so-called *skip connections*, that consist in reusing the feature maps by concatenating them with posterior layer inputs with the same dimension. For applications that involve dense predictions there is low-level information shared between the input and output, so the skip connections pass this information directly across the net. These concatenations give the network great stability as they prevent the gradient vanishing. In the case of the U-Net, the connections are formed symmetrically: from encoder to decoder layers.

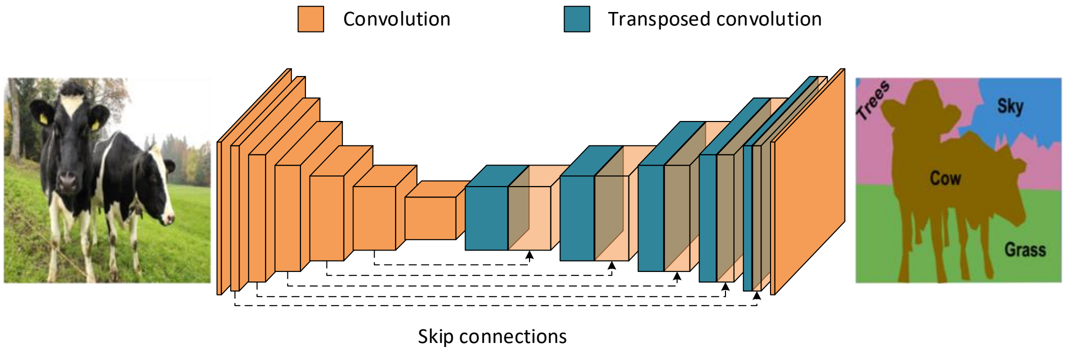


Figure 3.7: U-Net architecture for semantic segmentation.

3.2.2 Atrous Convolution and Atrous Spatial Pyramid Pooling (ASPP)

In FCNs, the encoder-decoder idea [25] reduces the computational cost and removes certain redundancy. However, the repeated combination of striding convolutions and pooling at consecutive layers in the encoder reduces significantly the spatial resolution of the feature maps [27]. This problem cannot be sufficiently avoided even using transposed convolutions with skip connections. Also, to achieve a receptive field that covers the whole image, the network depends on: i) Deeper designs with more convolutional layers, which implies reducing the resolution if we do not want to incur a high computational cost; and ii) Greater size of the convolutional kernel, with the added computation and memory costs.

To overcome those problems, the atrous convolution was introduced in [24]. This convolution uses kernels that are upsampled by inserting zeros in between the parameters, according to a given dilation rate. Equation 3-1 describes the output $Y(i,j)$ of the atrous convolution with rate r between the two-dimensional input X and a kernel W of size $K \times L$. As shown, the number of operations is invariant to r. Notice that with $rate = 1$, the atrous convolution is reduced to the standard convolution.

$$Y(i,j) = \sum_{k}^{K} \sum_{l}^{L} X(i + r \cdot k, j + r \cdot l) \cdot W(k,l) \tag{3-1}$$

Figure 3.8 illustrates different ways to reach the same receptive field for a certain output activation. For example, with 3×3 convolutional kernels (Figure 3.8-a)), two consecutive layers are necessary to reach receptive fields of 5×5 pixels. In return, if 5×5 kernels are used (Figure 3.8-b)), only a layer is required, but, at the expense of a greater number of parameters. With atrous convolutions of 3×3 kernels and $rate = 2$, this receptive field is reached in a single layer without increasing the number of parameters and the number of operations. Although the filter size increases with the rate, only the non-zero filter values are considered.

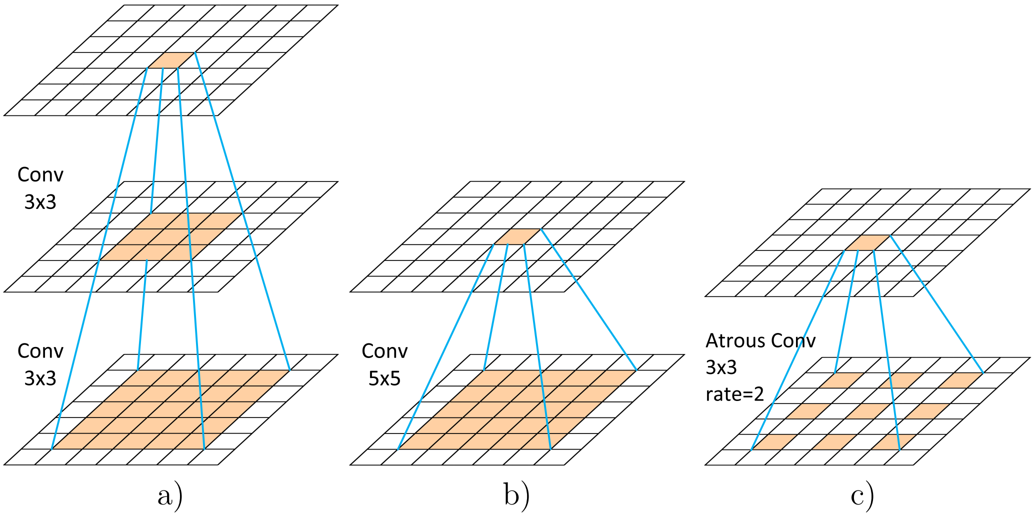


Figure 3.8: Receptive field capture. a) 3×3 convolution. b) 5×5 convolution. c) 3×3 atrous convolution with $rate = 2$.

For image analysis, both local and contextual perception contribute to infer conclusions about the content. With the atrous convolution, it is possible to easily adjust the field of view of the network and to exploit spatial context at different scales. Besides, the atrous convolutions avoid reducing the resolution of the feature maps, which helps mitigating the loss of spatial information. To extract features at multiple scales simultaneously, the DeepLab framework further proposed the Atrous Spatial Pyramid Pooling (ASPP) in [27]. ASPP performs a set of convolutions in parallel with different dilation rates to capture multi-scale information. For each rate, the features are calculated independently, and then concatenated along the channels.

The atrous convolution and the ASPP have been successfully applied for semantic segmentation problems. In this work, we take advantage of these contributions to improve the SAR-to-optical image synthesis.

3.3 Generative Adversarial Networks (GANs)

Generative models are usually known in the context of statistical classification, based for example in the maximum probability $P(X|Y = y)$ of an observation x given a target value y (maximum likelihood) [64]. However, the term is also used to describe models that generate samples of a given distribution, by processing samples from another distribution, without establishing an explicit or analytical relationship. That is the case of the Generative Adversarial Networks (GANs).

The GAN approach was proposed in [15] to generate data using neural networks. It is based on the game theory principle in which a generator network (G) must compete in an adversarial scheme with another network named discriminator (D). To learn the probability distribution over the data x, the generator builds a mapping function $G(z;\theta_G)$ from samples of a known noise distribution $p_Z(z)$. On the other hand, the discriminator $D(x;\theta_D)$ outputs the probability of x belonging to the real data rather than the fake synthesized by G. The discriminator is trained as a classifier to correctly assign the real (1) and fake (0) labels to the samples coming from $p_X(x)$ and $G(z)$ respectively, while the generator is simultaneously trained to fool the discriminator trying to synthesize realistic x samples. Formally, the learning process solves the min-max optimization problem:

$$G^* = \arg\min_{\theta_G}\max_{\theta_D}\mathcal{L}_{GAN}(G,D), \tag{3-2}$$

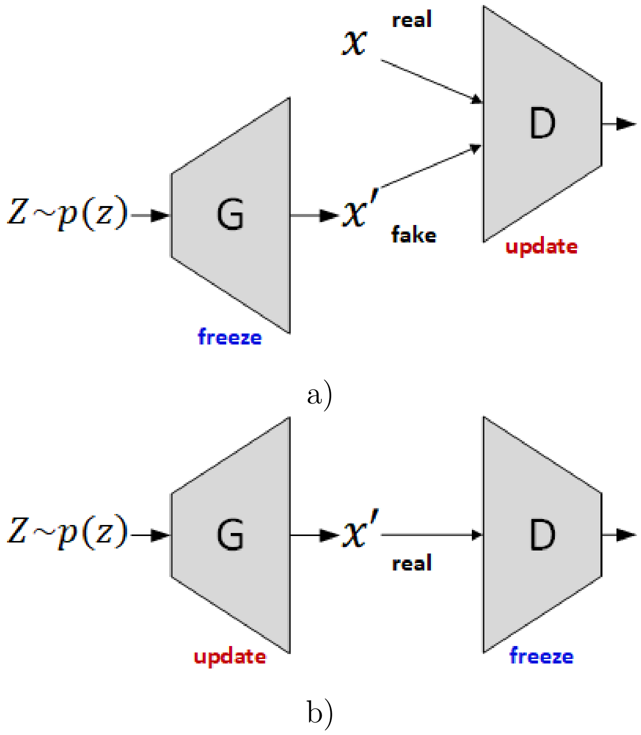
over the objective loss function:

$$\mathcal{L}_{GAN}(G,D) = \mathbb{E}_x[\log D(x)] + \mathbb{E}_z[\log(1 - D(G(z)))] \tag{3-3}$$

The optimal model is reached at a saddle point that is a local minima with respect to G and a local maxima with respect to D [50], in order to get the generator G^*. In practice, the training of G and D is alternated within each training loop. First, $G(z)$ is generated and only the parameters θ_D are updated to maximize $\mathcal{L}_{\mathcal{GAN}}(G,D)$, see Figure 3.9-a). Then, $G(z)$ is assumed to be real data, and only the parameters θ_G are updated to minimize $\log(1 - D(G(z)))$, see Figure 3.9-b). It is difficult to achieve the optimal model in GANs mainly because of convergence problems and the non convexity of the objective loss function $\mathcal{L}_{GAN}(G,D)$. However, the adversarial learning allows to reach a great approximation.

3.3.1 Paired image-to-image translation

The concept of image to image translation was introduced in [21] as a method that maps an input image to a corresponding output image. In that work, the *pix2pix* model was developed as a modification of the conditional Generative Adversarial Networks (cGANs). In cGANs [22], the generative model is conditioned on auxiliary information y that steer the synthesis of the samples x (see Figure 3.10). Besides a sample noise z, the cGAN



a)

b)

Figure 3.9: Adversarial learning. a) Discriminator training. b) Generator training.

generator takes as input a sample y to produce a realistic sample $G(y, z)$ of the distribution $p(x|y)$. The discriminator then evaluates if both samples y and x are from the real data distribution $y, x \sim p_{data}(y, x)$ or if the second one was produced by the generator. So, the cGAN objective loss function is as the Equation 3-4 shows.

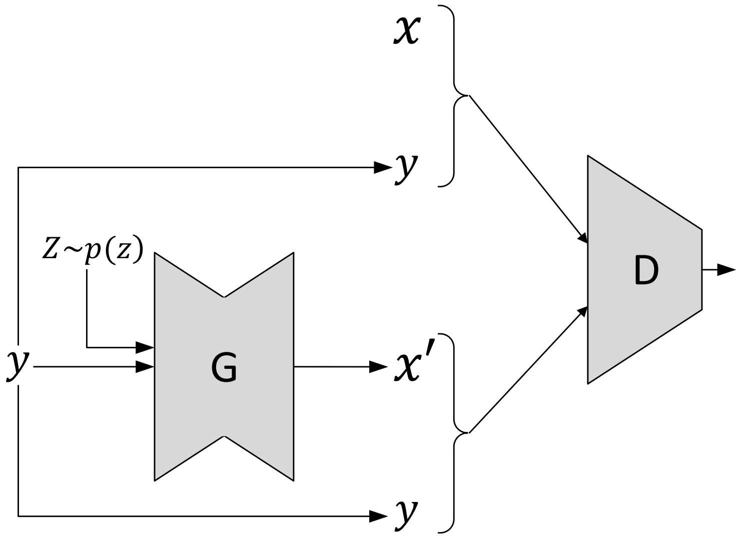


Figure 3.10: Conditional Generative Adversarial Network scheme.

$$\mathcal{L}_{cGAN}(G, D) = \mathbb{E}_{y,x}[\log D(y, x)] + \mathbb{E}_{y,z}[\log(1 - D(y, G(y, z)))] \tag{3-4}$$

To translate images y of one domain to images x of another domain, the cGAN learns a mapping function that needs paired samples (y, x). In other words, it is necessary that the input and output images share a great amount of low-level information such as geometric patterns. For example, pairs can be made up of two co-registered images of the same scene, but acquired with different sensors, or in grayscale and RGB composition, respectively. In short, G tries to transform the image y to its x counterpart. Then, two pairs are build: the real one (y, x), and the fake one $(y, G(y, z))$; so that D identifies them.

To favor the similarity between the real and generated images, [21] proposed the incorporation of the weighted L1 distance to the objective loss function, aiming to optimize:

$$G^* = \arg \min_{\theta_G} \max_{\theta_D} \mathcal{L}_{cGAN}(G, D) + \lambda \mathbb{E}_{x,y,z}[||x - G(y, z)\|_1] \tag{3-5}$$

The fundamentals presented so far will be used in the next chapter to describe our proposal for the SAR-optical synthesis.

4
Proposed Method

This chapter describes the proposal to improve the optical satellite image synthesis from SAR data. The proposal is based on conditional Generative Adversarial Networks (cGANs) to build a model that generates multispectral optical images in regions where optical sensors failed to capture the data, for any reason, in particular because of cloud covering. Specifically, our contribution refers to the generator and discriminator architectures, which are designed to take advantage of the context information using atrous convolutions to improve the quality of the synthesized optical image.

The basic cGAN design has the structure shown in Figure 4.1. The generator performs an image-to-image translation, that converts a SAR image at its input to a corresponding optical image at the output. Since the generator performs a dense prediction, Fully convolutional networks (FCN) are natural candidates for carrying out the task. According to [21], one can drop the z noise input (also shown in Figure 4.1), as long as dropout is added in several layers of the generator. Our proposal follows this strategy.

The discriminator is basically a binary classifier which mission is to discriminate between real and fake image pairs. In our application, the input to the discriminator consists of two corregistered SAR and optical images acquired at close dates that are concatenated along the third dimension.

The methods proposed thus far are built upon the *pix2pix* model [21]. The discriminator is a typical image classification network, consisting of a bunch of layers through which the input resolution of the input image decreases until the final activation layer that assigns a single label to the input. Often the generator network follows an encoder-decoder architecture like the U-Net (Figure 3.7). A downside of such architecture is the poor spatial/location accuracy caused by reduction occurring of spatial resolution through the encoder phase. As a result, fine details end up being poorly represented in the synthesized images. To improve the high-frequency information on the generated images, we propose the *atrous-cGAN* architecture, which incorporates the atrous convolutions and the Atrous Space Pyramid Pool (ASPP), both in the generator and in the discriminator networks.

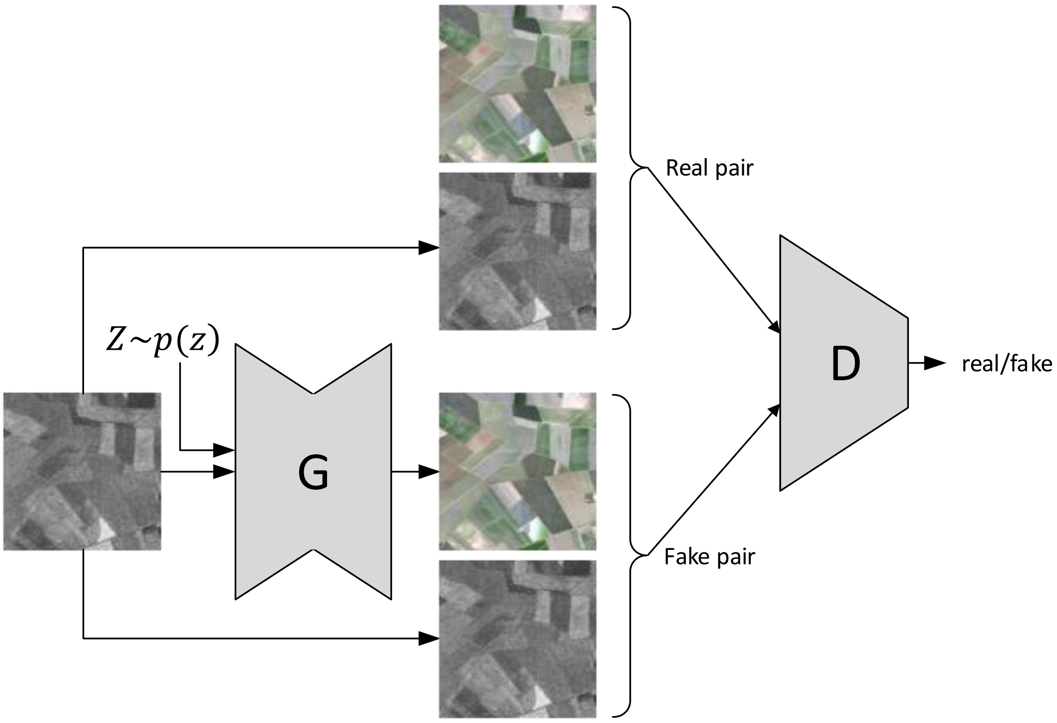


Figure 4.1: cGAN for SAR-optical synthesis.

4.1 Generator

Figure 4.2 shows the design of the proposed generator network. It is inspired on the DeepLabv3 [28] framework for semantic segmentation. As shown, the generator exhibits the structure of a FCN, with a SAR image as input and a synthesized optical image as output. Although it includes encoder and decoder stages, the spatial dimension of the feature maps is slightly reduced, always using a factor of 2 at each step. Actually, for all the experiments reported in Chapter 5, feature maps of 32×32 were guaranteed at the bottleneck. Notice that to reduce the spatial dimension of the feature maps, the architecture does not use pooling layers, but convolutions with $stride = 2$.

The encoder stage comprises a convolutional layer and four residual blocks. These residual blocks contain 3, 4, 6, and 3 residual units in cascade, respectively, similar to the ResNet-50 [65] network. As in [28], the fourth block performs atrous convolutions with rates equal to 1, 2 and 4 (each value correspond to a residual unit). The residual units are modules that learn a function $H(x) = F(x) + x$ (see Figure 4.3a)) that results from adding a non linear function $F(x)$ (residual) computed by a shallow network with convolutional and activation layers, and the input x. Since it is easy to learn a zero residual $F(x) = 0$, a residual unit can learn the identity function

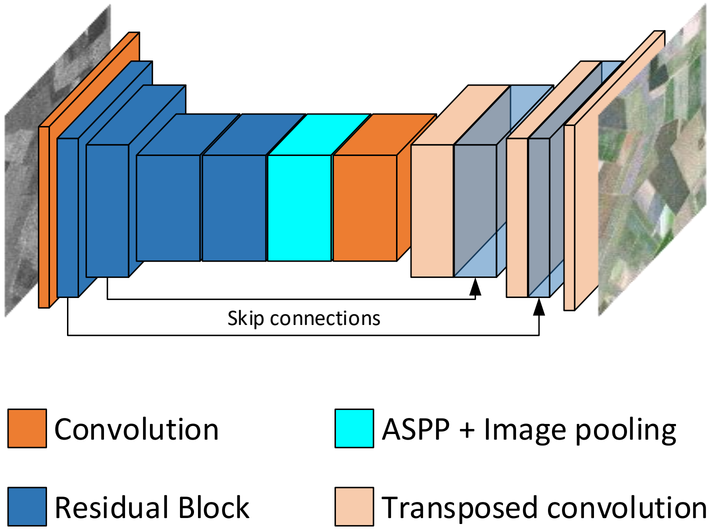


Figure 4.2: Proposed Generator.

$H(x) = x$, allowing to build very deep networks without that limiting it to learn such simple functions as the identity. Moreover, that skip connection via addition, also called *shortcut*, prevents the vanishing gradients [66] problem.

In residual units, the feature maps that represent $F(x)$ and x must have the same dimensions. For cases where the number of channels does not match, we use the projection shortcut (Figure 4.3b)) to adjust the depth of x with a 1×1 convolutional layer, the number of filters used must be equal to the third dimension of $F(x)$. On the other hand, when the width or height do not match because of downsampling, we use a shortcut with a Max-Pooling layer (see Figure 4.3c)).

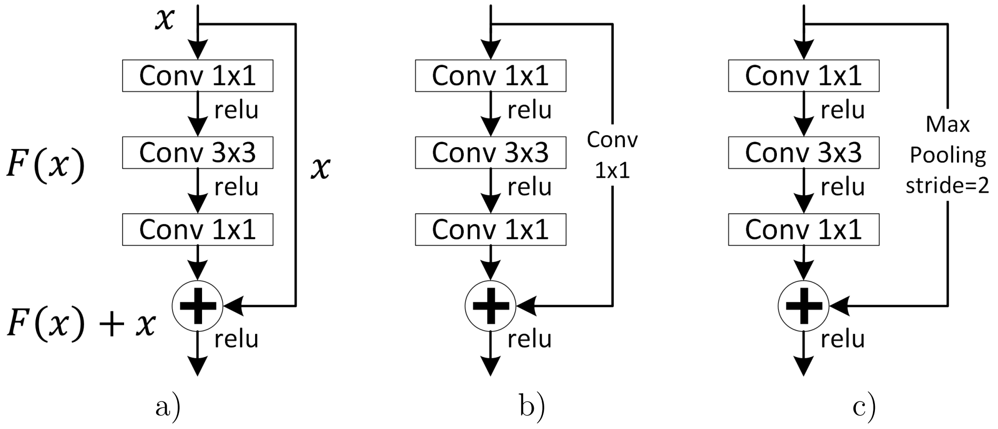


Figure 4.3: Residual Units. a) Identity shortcut. b) Projection shortcut. c) Pooling shortcut.

After the residual blocks, comes the ASPP module with a 1×1 convolution and three 3×3 atrous convolutions with rates 6, 12, and 18. As in [28], the

ASPP is equipped with global image-level features (Image Pooling), provided by a global average pooling [67] and a bilinear upsampling. The image-level features [68] are descriptors of the overall scene that capture the global context. The convolutions of the ASPP and the Image Pooling are processed in parallel and preserve the spatial dimension. The resulting feature maps are concatenated along the third dimension before being forwarded to another 1×1 convolution layer.

Finally, the resolution of the optical image is recovered by transposed convolutions, which outcomes are concatenated with feature maps from the encoder thanks via skip connections.

4.2 Discriminator

The discriminator plays a key role in improving generator's performance during the cGAN adversarial training. The discriminator guides the generator training. On the other hand, if the discriminator learns faster than the generator, it ends up rejecting generated images with a high degree of confidence, causing oscillations in the generator instead of convergence; in other words, the loss function value related to the generator will remain elevated without experiencing progress in learning. Therefore, generator and discriminator must be designed and trained in such a way that both learn at a compatible pace.

The improvement of the generator makes it difficult for the discriminator to distinguish between real from false SAR/optical pairs. So, we designed a discriminator that also takes context at multiple scales. Notice that the proposed discriminator, showed in Figure 4.4, incorporates the ASPP in between convolutional layers. Just like in the generator, the ASPP consists of four parallel modules: one 1×1 convolution and three 3×3 atrous convolutions with the resulting features concatenated and passing by another 1×1 convolution, but this time the correspondent rates were empirically set at 2, 3 and 4. The convolutional layers reduce the spatial dimension by 2 until the final layer, which outcome serve as the input feature for the final binary classification of the SAR/optical pairs.

In this chapter, it was showed the essential elements of novel designs for the generator and discriminator in cGANs. The number of parameters and hyperparameters tuning will depend on the 3D dimensions of the optical and SAR images. The next chapter describes in detail the implementation of the proposed architecture used in our experimental analysis.

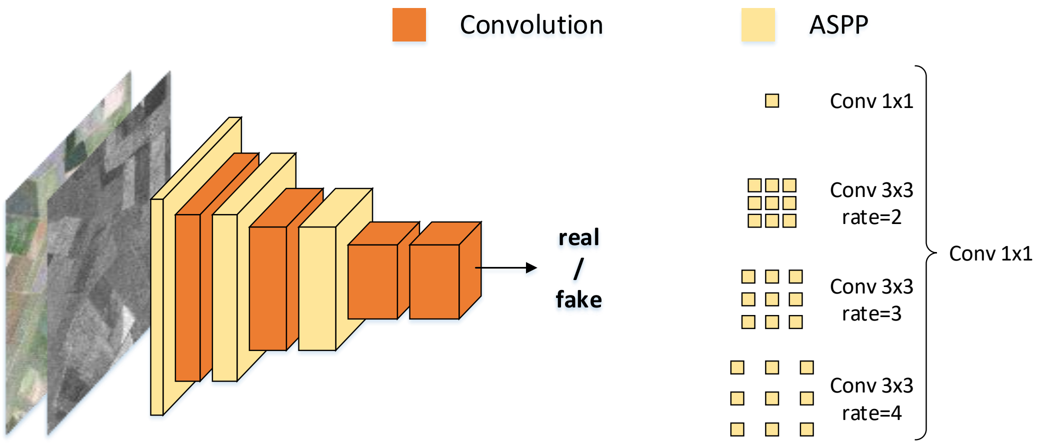


Figure 4.4: Proposed Discriminator.

5 Experiments and Results

This chapter presents a detailed description of the experiments carried out to assess the proposed *atrous-cGAN* for the SAR-optical synthesis. We use three procedures to evaluate the synthesized optical images: first, by visual inspection, second, measuring the similarity to the ground truth, and finally, by assessing the accuracy of the semantic segmentation of the synthetic optical image. Initially, we present the datasets used in the experiments. Then, we explain the experimental protocols and the implementation details. Finally, we report and discuss the results.

5.1 Datasets

The evaluation of the proposed method focused on two applications: crop recognition and deforestation detection; each with two datasets. For all datasets we have co-registered SAR/optical images that refer to land covers of the Brazilian territory. Being these tropical areas, they are very often covered by clouds that impair the acquisition of optical images.

The optical images were captured by the Landsat 8 Operational Land Imager (OLI) satellite, and comprise 7 spectral bands with 30m of spatial resolution. Table 5.1 shows the band designations. SAR images were captured by the Sentinel-1A satellite with the Interferometric Wide (IW) swath mode, at Level-1 Ground Range Detected (GRD)[1], and dual polarization VV and VH; so they have 2 bands with 10m of spatial resolution. The optical images are available in the United States Geological Survey (USGS) web service[2], and the SAR ones in the Copernicus Open Access Hub[3]. All images were captured on clear sky days.

5.1.1 Crop recognition datasets

The agriculture is responsible for most of Brazilian exports. Therefore, there is great interest in monitoring crops. Two public datasets for crop

[1] `https://sentinel.esa.int/web/sentinel/missions/sentinel-1`
[2] `https://earthexplorer.usgs.gov`
[3] `https://scihub.copernicus.eu`

Table 5.1: Optical spectral bands.

Bands	**Wavelength** (micrometers)
Band 1 - Coastal aerosol	0.43 - 0.45
Band 2 - Blue	0.45 - 0.51
Band 3 - Green	0.53 - 0.59
Band 4 - Red	0.64 - 0.67
Band 5 - Near Infrared (NIR)	0.85 - 0.88
Band 6 - Short Wave Infrared (SWIR) 1	1.57 - 1.65
Band 7 - Short Wave Infrared (SWIR) 2	2.11 - 2.29

recognition were used in our experiments:

1. **Campo Verde (CV).**

   This is a public dataset [69] that comprises images with its corresponding labels and refers to a rural region in the municipality of Campo Verde (CV), Mato Grosso State. The study area is located between the coordinates 14°57'19"S - 15°55'18"S latitude and 54°25'04"W - 55°27'02"W longitude. One pair SAR/optical with data acquisition May 8, 2016 for the SAR image, and May 5, 2016 for the optical one, was used. The size of the optical image is 2831 × 2665 pixels, and the SAR image 8493 × 7995 pixels. The site contains a total of nine land cover classes: cotton, maize, pasture, noncomercial crops (NCC), cerrado, eucalyptus, sorghum, soil, beans and turfgrass. Figure 5.1 shows the representation of these classes in the dataset.

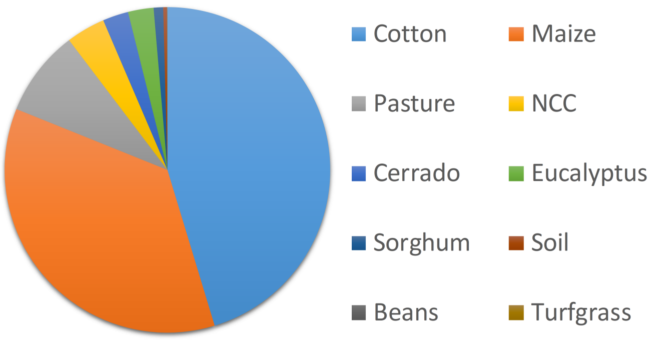


Figure 5.1: Classes representation of CV dataset.

2. **Luis Eduardo Magalhães (LEM).**

   The other dataset related to crop mapping is located in the municipality of Luis Eduardo Magalhães (LEM), Bahia State. It is between the coordinates 11°51'37"S - 12°32'16"S latitude and 45°39'26"W - 46°23'25"W longitude. The size of the optical image is 2828 × 2886 pixels, and the

SAR image 8484 × 8658 pixels. The dataset is public [70] and comprises 794 fields with their respective land use. The pair SAR/optical used was acquired on June 12, 2017 and June 15, 2017 for the SAR and optical images respectively. The region contains 13 classes, their representations are shown in Figure 5.2.

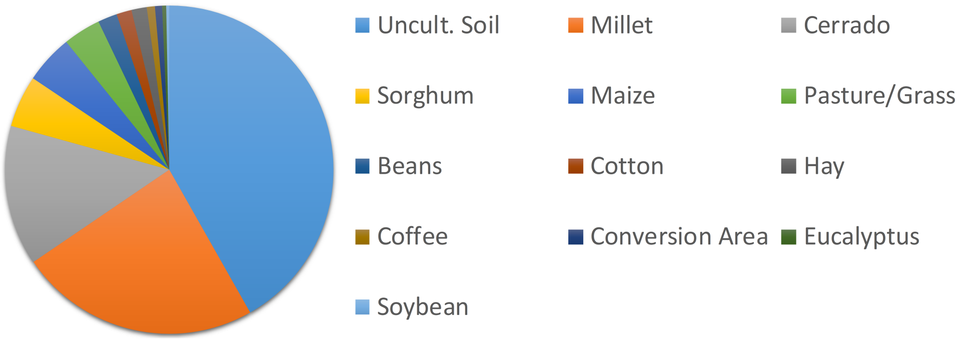


Figure 5.2: Classes representation of LEM dataset.

5.1.2 Deforestation detection datasets

Change detection on remote sensing is the process of identifying differences in the composition of land covers observed at different times. In the case of forested regions, the change is generally caused by deforestation, often illegally. Our proposal was tested on two datasets related to this application. Each dataset comprises two coregistered SAR/optical pairs of the same geographical region. We denote with t_0 and t_1 the acquisition dates of both image pairs, whereby t_0 is one year before t_1. For each dataset we have two references, the deforested areas until date t_0 and the areas deforested between t_0 and t_1. All references are available at the PRODES Deforestation Mapping Project website[4], provided and maintained by the Brazilian National Institute of Space Research (INPE). The datasets are the following:

1. **Amazon Rainforest (AR).**

   This dataset is oriented to the detection of deforested regions within a year in the Amazon Rainforest (AR). The AR is the largest tropical rainforest in the world, with special climatic importance due to its regulatory influence. Its vegetation is dense and formed by large trees. The deforestation in the Amazon is mainly encouraged by the cattle sector [71], and constitute the human activity that most threatens that region.

[4] `http://terrabrasilis.dpi.inpe.br/map/deforestation`

The data we used in our experiments represent part of the Rondônia State, between the coordinates: 09°36'51"S - 10°18'35"S latitude and 62°56'41"W - 64°20'51"W longitude. Table 5.2 shows the acquisition dates of each SAR and optical image. Notice that the pairs are one year apart from each other. The size of the optical images is 2550×5120 pixels, and the SAR images 7650×15360 pixels. In this dataset, the class imbalance is remarkable since the deforested area from time t_0 to t_1 represent approximately the 1% of the whole scene.

Table 5.2: Acquisition dates of AR dataset.

Image	Date
SAR_{t_0}	July 22, 2018
$optical_{t_0}$	July 24, 2018
SAR_{t_1}	July 22, 2019
$optical_{t_1}$	July 27, 2019

2. **Cerrado Biome (CB).**

Images of the Cerrado Biome (CB) were also used to evaluate our design. The Brazilian Cerrado is a vast tropical biome composed of savannas and grasslands amid humid and dry forests. In this region, the woodlands have scattered trees and shrubs and the deforestation generally has an agricultural purpose [72]. The deforestation patterns are different from those in the Amazon Rainforest.

The study area is located in the Maranhão State, specifically between the coordinates: 04°44'52"S - 05°12'48"S latitude and 43°37'55"W - 44°01'23"W longitude. The acquisition dates are shown in Table 5.3. The size of the optical images is 1719×1442 pixels, and the SAR images 5157×4326 pixels. As in the previous dataset, the classes are quite unbalanced; the deforestation from t_0 to t_1 dates represents approximately 3% of the target site.

Table 5.3: Acquisition dates of CB dataset.

Image	Date
SAR_{t_0}	August 23, 2017
$optical_{t_0}$	August 18, 2017
SAR_{t_1}	August 18, 2018
$optical_{t_1}$	August 21, 2018

5.2 Experimental Protocol

For the experiments, we assumed that parts of the optical images were missing. In real scenarios it could be due to sensor malfunctioning or to the presence of dense clouds. Thus, we arbitrarily divided each scene into *available* and *missing* regions to train and test the method. In the cGAN training phase, we used paired SAR/optical patches extracted from the *available* region using a sliding window procedure with a fixed stride. After training, an optical image of the *available* and *missing* regions was synthesized building a mosaic of patches predicted by the generator, using as input the corresponding SAR patches.

The aforementioned steps were carried with the proposed *atrous-cGAN*, and with the *pix2pix* architecture, which served as baseline. The *missing* regions of the optical images synthesized by both models were evaluated through the semantic segmentation. Besides comparing the result of the semantic segmentation of the synthesized optical images, we also assessed the visual quality, and their similarity with the real optical image, using standard similarity metrics.

5.2.1 Crop recognition

Figure 5.3 shows the spatial distribution of the classes in the selected landscape of the CV dataset. They are presented from left to right and up to bottom, in decreasing order of the number of pixels. Observing Figures 5.1 and 5.3, it can be verified that most classes are weakly represented, and are concentrated in restricted regions of the target site. The most critical cases are the pixels of soil, beans, and turfgrass. Indeed, to see them it is necessary to enlarge the image. Figure 5.4 shows the spatial distribution of classes in the LEM dataset in decreasing order according to the number of pixels. Again, analyzing Figures 5.2 and 5.4, we found weakly represented and isolated classes.

It is important to highlight that classes poorly represented in the training set of the cGAN may not be well generated during the test. Furthermore, the concentration of some classes in restricted image regions makes it difficult to split the scene into *available* and *missing* regions preserving approximately the same proportion of samples for all classes in both datasets. To circumvent this difficulty, in both CV and LEM datasets, we merged the minority classes into a single class denoted hereafter *other crops*. Figures 5.5a) and 5.6a) shows the final classes considered for the semantic segmentation experiments and their correspondent percentage of pixels in CV and LEM datasets, respectively. This consideration allowed to split the scenes into approximately 50% for training

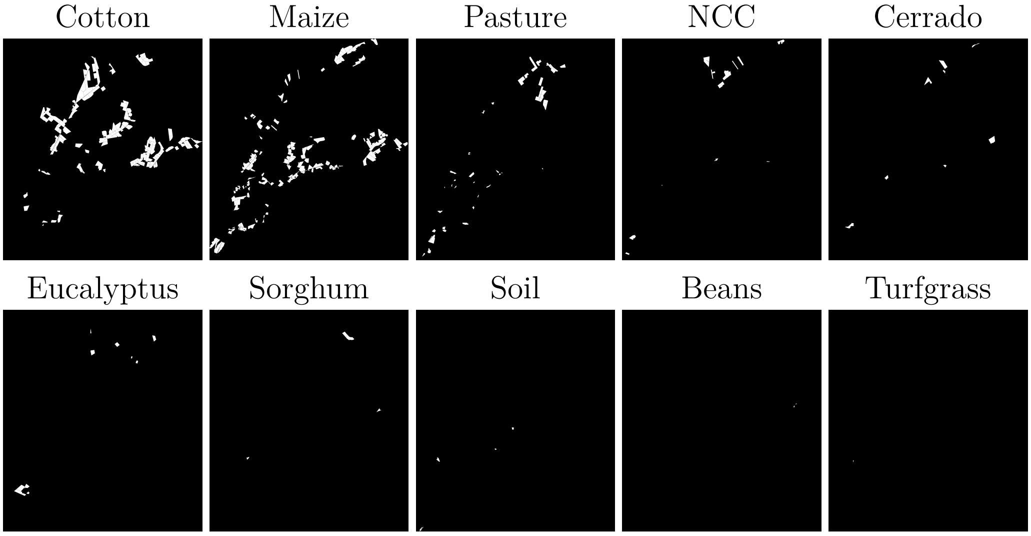


Figure 5.3: Spatial distribution of classes in CV dataset.

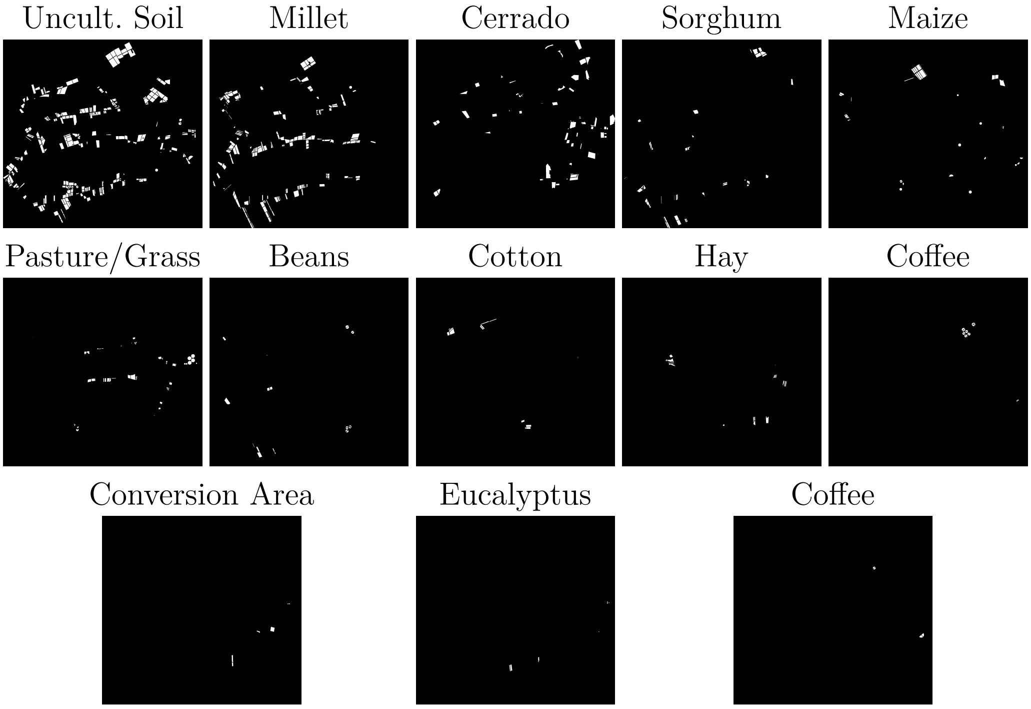


Figure 5.4: Spatial distribution of classes in LEM dataset.

and 50% for testing. Figures 5.5 b and 5.6 b show how the labeled pixels were split in CV and LEM dataset, respectively.

5.2.2 Deforestation detection

In the datasets related to deforestation detection we faced a different difficulty than that analyzed in the previous section. AR and CB datasets involve a highly unbalanced binary classification problem. However, as Figure

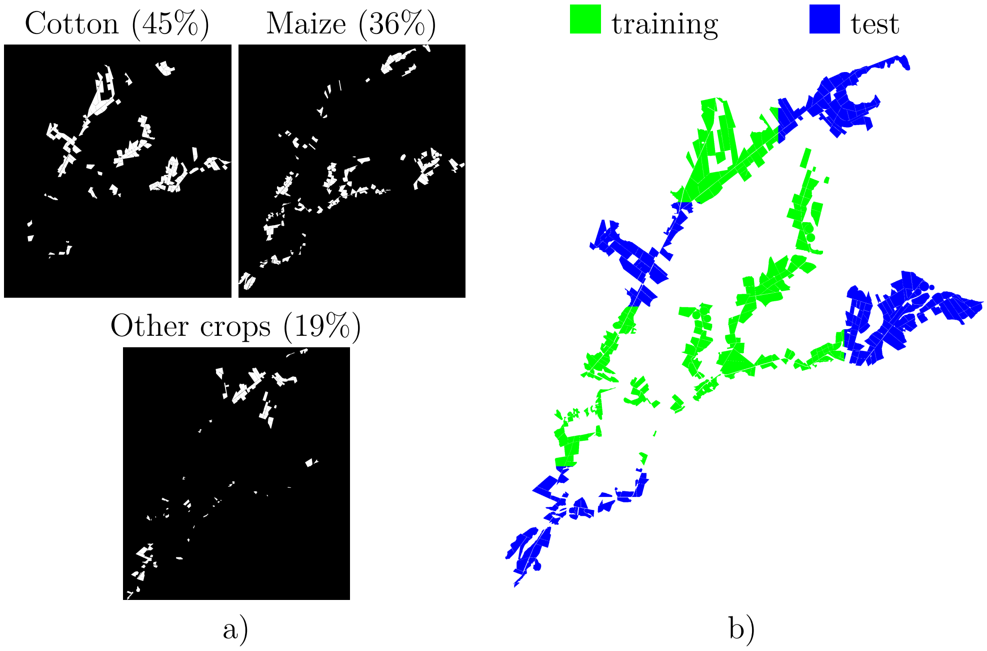


Figure 5.5: Final distribution of classes in CV dataset. a) Classes labels. b) Available (training) and missing (test) regions.

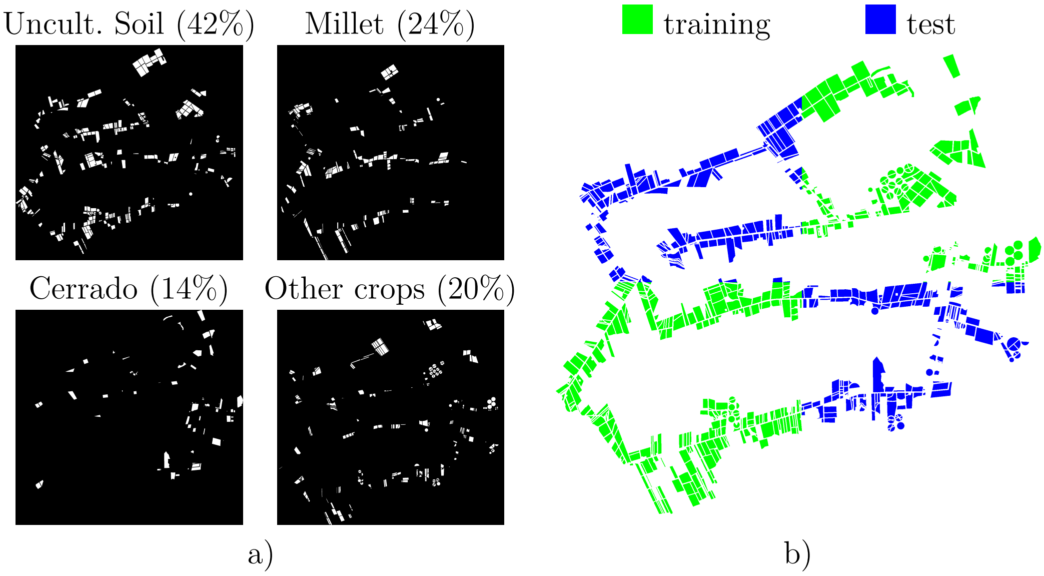
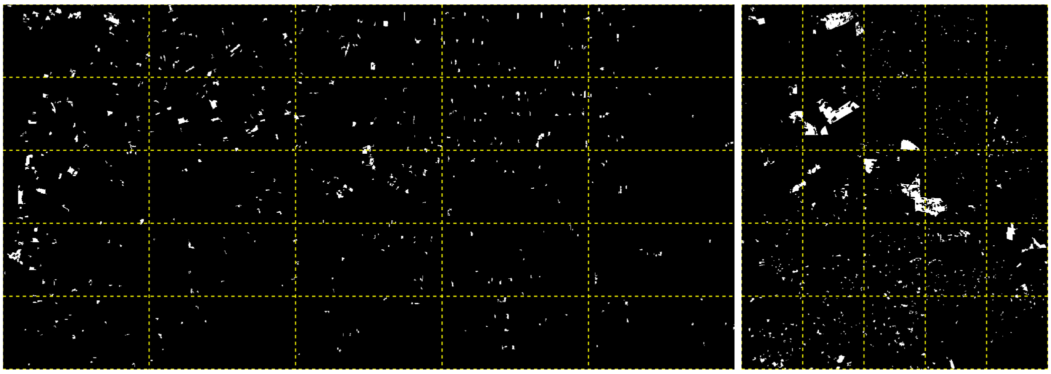


Figure 5.6: Final distribution of classes in LEM dataset. a) Classes labels. b) Available (training) and missing (test) regions.

5.7 shows, the samples of class *deforestation* are spread all over the target sites. This allowed us to evaluate the cGANs variants using a k-fold strategy with $k = 5$. To this goal, the scenes were divided into 25 disjoint and equal size tiles (also shown in Figure 5.7). Five tiles (20% of the whole scene) were randomly selected as *missing* regions for test. The *missing* regions in each fold had no overlap. The remaining 20 tiles (80% of the whole scene) were used for training.



a) b)

Figure 5.7: Spatial distribution of deforestation class and division in tiles of the images. a) AR dataset. b) CB dataset.

As Table 5.2 and 5.3 show, in the AR and CB datasets two optical images had to be synthesized, one for time t_0 and the other for time t_1. Instead of training separately for each image, we trained the cGAN upon the *available* SAR/optical patches in both times simultaneously, and then, the same model was used to synthesize the *missing* optical regions of both dates. This strategy allowed the usage of more training samples, and reduced the training effort.

5.3 Implementation details

The size of input patches extracted from the optical images in CV and LEM datasets was set to 256 × 256 pixels. Since we are interested in the synthesis of crop regions, we only extract patches that contain at least one labelled pixel, in CV and LEM datasets. On the other hand, the size of optical patches in AR and CB datasets was set to 128×128 pixels. Consequently, due to the difference of spatial resolution between Sentinel-1A and Landsat 8 sensors, the SAR patches were three times bigger than their optical counterpart, 768 × 768 pixels for CV and LEM, and 384 × 384 pixels for AR and CB. To match the resolution of both sensors, the SAR patches were downsampled using an additional convolutional layer with $stride = 3$.

The networks used in this work were all implemented using the Tensorflow framework [73]. The details of the proposed *atrous-cGAN* are described in Table 5.4. Notice that we point out some parameters and layers that are not used for the AR and CB datasets, since the input patch size is smaller. Batch normalization and the ReLU function were applied after each layer, except for the output of the generator and the discriminator. With this configuration, the discriminator contains approximately 5k parameters; and, the generator contains 58M and 56M parameters for crop recognition and deforestation detection datasets, respectively.

Table 5.4: *Atrous-cGAN* Architecture: C, T, RU, ASPP and IP, denote convolution, transposed convolution, residual unit, Atrous Spatial Pyramid Pooling, and image pooling, respectively. $(w \times w, k, s)$ denotes (kernel size, number of kernels, stride).

Generator		**Discriminator**
Encoder	Decoder	
$\mathrm{C}(5 \times 5, 16, 3)$	$\mathrm{T}(5 \times 5, 128, 2)^*$	$\mathrm{C}(5 \times 5, 2, 3)$
$\mathrm{C}(7 \times 7, 64, 2)$	$\mathrm{T}(5 \times 5, 64, 2)$	ASPP , $\mathrm{C}(1 \times 1, 128, 1)$
3×RU with $\begin{cases}(1 \times 1, 64, 1)\\(3 \times 3, 64, a)\\(1 \times 1, 256, 1)\end{cases}$	$\mathrm{T}(5 \times 5, 7, 2)$	$\mathrm{C}(5 \times 5, 64, 2)$
4×RU with $\begin{cases}(1 \times 1, 128, 1)\\(3 \times 3, 128, b)\\(1 \times 1, 512, 1)\end{cases}$		ASPP , $\mathrm{C}(1 \times 1, 64, 1)$
6×RU with $\begin{cases}(1 \times 1, 256, 1)\\(3 \times 3, 256, 1)\\(1 \times 1, 1024, 1)\end{cases}$		$\mathrm{C}(5 \times 5, 128, 2)$
3×RU with $\begin{cases}(1 \times 1, 512, 1)\\(3 \times 3, 512, 1)\\(1 \times 1, 2048, 1)\end{cases}$		ASPP , $\mathrm{C}(1 \times 1, 128, 1)$
ASPP , IP		$\mathrm{C}(5 \times 5, 256, 2)$
$\mathrm{C}(1 \times 1, 2560, 1)$		$\mathrm{C}(5 \times 5, 512, 2)$
		sigmoid()

$a = 2$ in the last residual unit and $a = 1$ otherwise; $b = 2$ in the last residual unit for CV and LEM and $b = 1$ otherwise; *layer present only for CV and LEM datasets.

On each training epoch, the patches were randomly transformed using cropping, horizontal and vertical flips, and rotations as data augmentation strategies. In all experiments, the models were trained using the Adam optimizer [74] with learning rate $\alpha = 0.0002$ and momentum $\beta_1 = 0.9$. All the code used in this work is publicly available[5], including a full description of the generator and discriminator architectures, and the adopted hyperparameter values. Tuning of hyperparameters was achieved after some tests.

Table 5.5 describes the architecture of the *pix2pix* model used as baseline; its implementation was adapted from [13][6] for SAR-to-optical synthesis. As mentioned before, the generator has a U-Net architecture (Figure 3.7) and the discriminator is a conventional convolutional classifier. Batch normalization and ReLU function are used on each convolutional layer. In the *pix2pix*, the discriminator contains approximately 4k parameters; and, the generator

[5] https://github.com/jnoat92/atrous-cGAN-for-SAR-optical-synthesis

[6] https://github.com/bermudezjose/SAR2Optical-using-cGANS

contains 85M and 65M parameters for crop recognition and deforestation detection datasets, respectively.

Table 5.5: *pix2pix* Architecture: C and T denote convolution and transposed convolution, respectively. $(w \times w, k, s)$ denotes (kernel size, number of kernels, stride).

Generator		**Discriminator**
Encoder	Decoder	
$\mathrm{C}(5 \times 5, 4, 3)$	$\mathrm{T}(5 \times 5, 512, 2)^*$	$\mathrm{C}(5 \times 5, 4, 3)$
$\mathrm{C}(5 \times 5, 64, 2)$	$\mathrm{T}(5 \times 5, 512, 2)$	$\mathrm{C}(5 \times 5, 64, 2)$
$\mathrm{C}(5 \times 5, 128, 2)$	$\mathrm{T}(5 \times 5, 512, 2)$	$\mathrm{C}(5 \times 5, 128, 2)$
$\mathrm{C}(5 \times 5, 256, 2)$	$\mathrm{T}(5 \times 5, 512, 2)$	$\mathrm{C}(5 \times 5, 256, 2)$
$\mathrm{C}(5 \times 5, 512, 2)$	$\mathrm{T}(5 \times 5, 256, 2)$	$\mathrm{C}(5 \times 5, 512, 2)$
$\mathrm{C}(5 \times 5, 512, 2)$	$\mathrm{T}(5 \times 5, 128, 2)$	sigmoid()
$\mathrm{C}(5 \times 5, 512, 2)$	$\mathrm{T}(5 \times 5, 64, 2)$	
$\mathrm{C}(5 \times 5, 512, 2)$	$\mathrm{T}(5 \times 5, 7, 2)$	
$\mathrm{C}(5 \times 5, 512, 2)^*$		

*layer present only for CV and LEM datasets.

5.4 Semantic segmentation

To evaluate the image synthesis as a representation learning tool, two FCNs were used for the semantic segmentation of the optical images. For all datasets, we trained and tested the classifiers using: i) the real optical images, ii) the images synthesized by the *atrous-cGAN*, iii) the images synthesized by the *pix2pix*, and iv) the SAR images. The training was made on the *available* regions and the test on the missing ones. This procedure was carried out five times for each input data, hence, we reported the average metrics over the five runs.

A U-Net network was a segmentation tool used in our experiment. As exposed in Chapter 3, the main contribution of this architecture is the presence of skip connections via concatenation between correspondent encoder and decoder layers. The details of the network are shown in Table 5.6. This time we followed the original proposal of the U-Net for semantic segmentation [23], using Max-Pooling layers to downsample the feature maps in the encoder, and the ReLU activation function after every convolutional layer. Different classifier configurations do not affect the objective of the evaluation, since we are interested in the relative results using the different data sources.

We also carried out experiment using the SegNet network [62] for the semantic segmentation. Similar to the U-Net, it comprises encoder and decoder stages, but apply no feature reusability with skip connections. Instead, the

Table 5.6: U-Net Classifier: $(w \times w, k, s)$ denotes (kernel size, number of kernels, stride) for convolutional layers. In MaxPooling layers only the kernel size and the stride are specified.

Encoder	Decoder
2×C$(3 \times 3, 32, 1)$	T$(3 \times 3, 128, 2$
MaxPool$(2 \times 2, 2)$	2×C$(3 \times 3, 128, 1)$
2×C$(3 \times 3, 64, 1)$	T$(3 \times 3, 64, 2)$
MaxPool$(2 \times 2, 2)$	2×C$(3 \times 3, 64, 1)$
2×C$(3 \times 3, 128, 1)$	T$(3 \times 3, 32, 2)$
MaxPool$(2 \times 2, 2)$	2×C$(3 \times 3, 32, 1)$
2×C$(3 \times 3, 128, 1)$	C$(1 \times 1, n_classes, 1)$
	softmax()

maximum pool indices computed in the encoder stage are stored and later used in the upsampling stage for positioning the values delivered by the preceding layer. The upsampling operation that recall these indices is called Max-Unpooling. This process involves associating encoder Max-Pooling layers with decoder Max-Unpooling layers. Table 5.7 describes in detail the parameter configuration of these layers. As in [62], batch normalization and ReLU function were applied after each convolutional layer.

Table 5.7: SegNet Classifier: $(w \times w, k, s)$ denotes (kernel size, number of kernels, stride) for convolutional layers. In MaxPooling layers only the kernel size and the stride are specified.

Encoder	Decoder
2×C$(3 \times 3, 32, 1)$	MaxUnpool
MaxPool$(2 \times 2, 2)$	2×C$(3 \times 3, 128, 1)$
2×C$(3 \times 3, 64, 1)$	C$(3 \times 3, 64, 1)$
MaxPool$(2 \times 2, 2)$	MaxUnpool
3×C$(3 \times 3, 128, 1)$	C$(3 \times 3, 64, 1)$
MaxPool$(2 \times 2, 2)$	C$(3 \times 3, 32, 1)$
	MaxUnpool
	C$(3 \times 3, 32, 1)$
	C$(1 \times 1, n_classes, 1)$
	softmax()

We used the Adam optimizer with learning rate $\alpha = 0.0007$ and momentum $\beta_1 = 0.9$. Early stopping was used as stop criteria after 10 epochs with no improvement over validation patches extracted from the *available* region. The size of the patches entering the network was set to 128 × 128 pixels. To reduce the impact of class imbalance, the weighted categorical cross entropy was used as objective loss function. The weight predefined for each class was inversely proportional to the number of training pixels of that class in the respective dataset. For CV and LEM datasets, the unlabeled pixels were ignored setting a zero weight to them in the loss function. In the case of AR and CB datasets, the pixels that indicate deforested areas before t_0 were similarly ignored to focus the deforestation occurred between t_0 and t_1.

5.4.1 Specifications for deforestation detection

For deforestation detection we adopted the Early Fusion (EF) approach [75]. As Figure 5.8 shows, in EF the input to the FCN (either U-Net or Segnet) is formed by concatenating the images acquired in t_0 and t_1 along the third dimension.

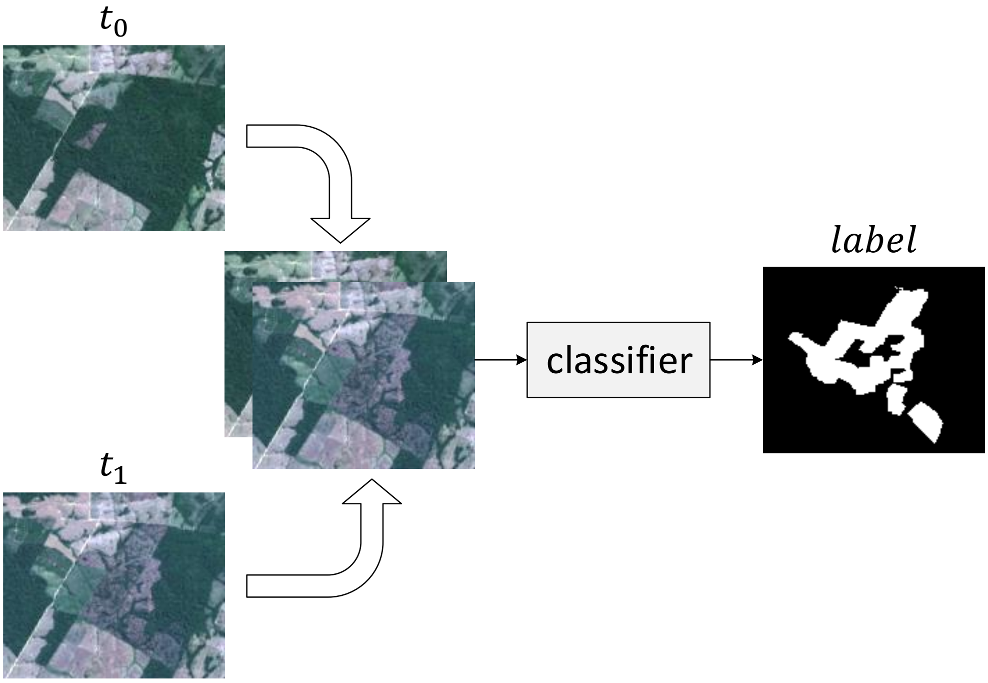


Figure 5.8: Early fusion structure. The illustrated input corresponds to the RGB composition of the real optical data.

In the experiments for deforestation detection using optical images, before the concatenation we stacked the Normalized Difference Vegetation Index (NDVI) as an 8th band on each image. The NDVI, calculated using the Equation 5-1, is an indicator that heightens the live green plant canopies in multispectral remote sensing images [76].

$$NDVI = \frac{NIR - Red}{NIR + Red}, \tag{5-1}$$

where NIR and Red stand for the *near infrared* and *red* bands, respectively.

5.5 Evaluation Metrics

Several metrics were used in this work to quantify the methods' performance. For semantic segmentation we adopted the Overall Accuracy (OA) and F1-Score, as given in the following.

- **Overall Accuracy (OA)**:

  In semantic segmentation the OA indicates the probability of correct class assignement. Mathematically, OA is the ratio between the amount of pixels correctly classified and the total number of pixels in the test set.

$$OA = \frac{number\ of\ correctly\ classified\ pixels}{total\ number\ of\ pixels} \tag{5-2}$$

- **F1-Score**:

  It is the harmonic mean between the *Precision* and the *Recall* (Equation 5-5). For one class, *Precision* is given by the ratio of true positives (tp) and the total number of positive predictions ($tp + fp$). *Recall* refers to the proportion the samples that belongs to the target class ($tp + fn$) and are predicted as such (tp). The number of true positives (tp), false positives (fp), and false negatives (fn) are extracted from the confusion matrix. Thus, the F1-Score is calculated for each class.

$$Precision = \frac{tp}{tp + fp} \tag{5-3}$$

$$Recall = \frac{tp}{tp + fn} \tag{5-4}$$

$$F1 = 2 \cdot \frac{Precision \cdot Recall}{Precision + Recall} \tag{5-5}$$

In the case of multi-class classification, as is the case of CV and LEM datasets, the F1-score reported was the mean over the considered classes.

To evaluate the similarity between synthetic and real optical images, we used the Peak Signal to Noise Ratio (PSNR) and the Spectral Angle Mapper (SAM) as defined below.

- **Peak Signal to Noise Ratio (PSNR)**:

  The PSNR is the ratio between the maximum power of a reference image and the power of noise that affect it. For measuring the similarity between the real (I) and synthetic ($\hat{I}$) optical images, the maximum intensity value of the reference image is $I_{MAX} = 2^{16} - 1$, since the images are represented in 16 bits. The noise, in this case, is calculated via Root Mean Square Error (RMSE) (Equation 5-6) between the images. The PSNR (Equation 5-7) is expressed in dB, and indicates greater similarity between the images as its value increases.

$$RMSE = \sqrt{\frac{1}{mn}\sum_{i=0}^{m-1}\sum_{j=0}^{n-1}[I(i,j) - \hat{I}(i,j)]^2} \tag{5-6}$$

$$PSNR = 20 \cdot \log_{10}\left(\frac{I_{MAX}}{RMSE}\right), \tag{5-7}$$

  where m and n are the dimensions of the images.

  Equation 5-7 refers to the calculus of the PSNR considering a single band image. We actually calculate the metric for all spectral bands and report the average over pixels and bands.

- **Spectral Angle Mapper (SAM)**:

  SAM expresses the difference between colors, based on the angle defined by their vector representations in a given color space [77]. It is calculated for each pixel according to the Equation 5-8. The lower the SAM, the greater the similarity between the images. The reported values of this metric is the average among all pixels.

$$\theta(x,y) = cos^{-1}\left(\frac{\sum_{i=1}^{nb} x_i y_i}{(\sum_{i=1}^{nb} x_i^2)^{\frac{1}{2}} \cdot (\sum_{i=1}^{nb} y_i^2)^{\frac{1}{2}}}\right), \tag{5-8}$$

  where nb represents the number of bands of the images, and x and y refers to a pixel from images I and $\hat{I}$ respectively, at the same location.

5.6 Results

5.6.1 Semantic Segmentation

5.6.1.1 Experiments for crop recognition

Figures 5.9 and 5.10 show the results of the semantic segmentation in CV and LEM datasets respectively, using the U-Net and the SegNet. Each group of bars refers to the *missing* areas of the original optical image, the optical synthesized by the proposed *atrous-cGAN*, the optical synthesized by the *pix2pix* model, and the SAR one. Overall Accuracy and the average F1-Score over all considered classes are reported in each case.

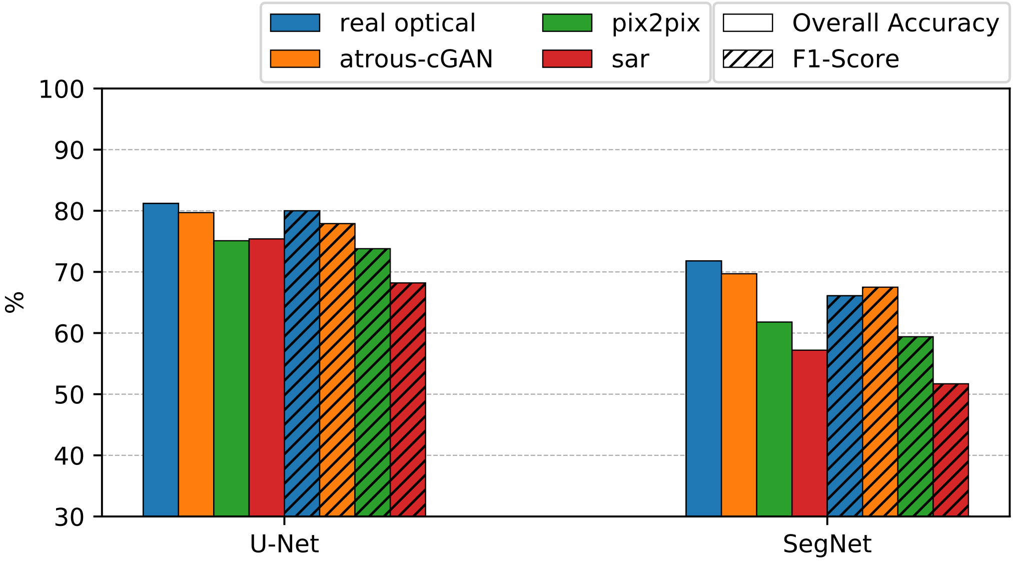


Figure 5.9: Performance of the Semantic Segmentation in terms of OA and F1 on Campo Verde dataset.

Not surprisingly, in both datasets, and with both segmentation networks the highest accuracy was obtained with the real optical image. This is regarded as the best achievable results or the reference in these experiments. On the other hand, the segmentation of the SAR data delivered in most cases the lowest scores, indicating that the SAR imagery is less descriptive than the optical counterparts.

The results obtained with the optical image synthesized by the *atrous-cGAN* were close to the performance observed for the real optical image, achieving even the better result in terms of average F1-Score with the SegNet.

The results obtained on images produced by the *pix2pix* were in most cases superior to what was achieved on SAR images. Exceptions were recorded

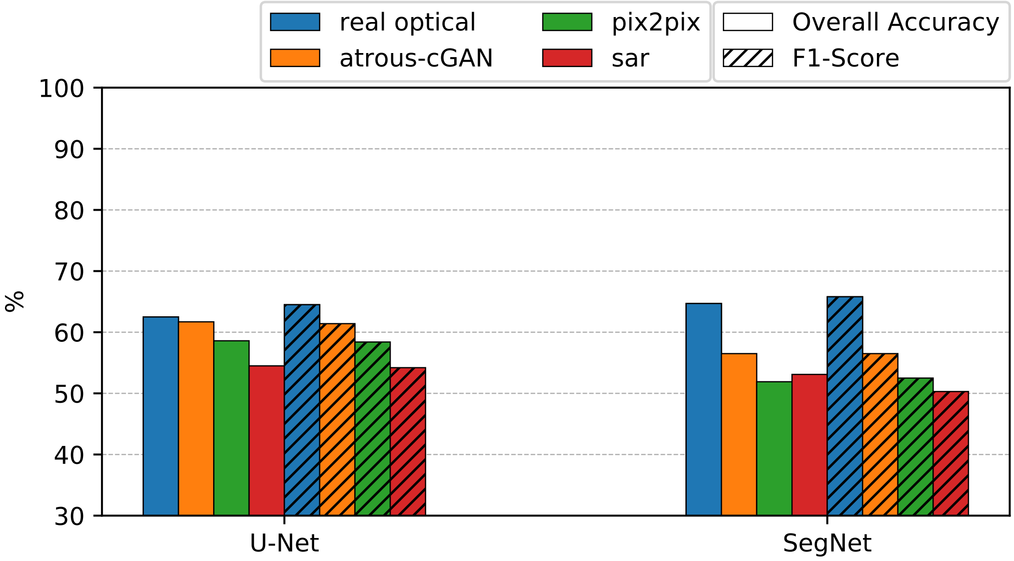


Figure 5.10: Performance of the Semantic Segmentation in terms of OA and F1 on Luis Eduardo Magalhães dataset.

for the CV dataset using the U-Net, and for the LEM dataset using the SegNet. However, in these cases the overall accuracy was slightly worse.

The *atrous-cGAN* consistently outperformed the *pix2pix* baseline and the raw SAR data by a large margin both in terms of OA and F1 score. In sum, the experiments for the semantic segmentation of both datasets devoted to crop mapping indicated a clear superiority of the *atrous-cGAN* design over the *pix2pix*.

5.6.1.2 Experiments for deforestation detection

The plots on Figures 5.11 and 5.12 show the results for deforestation detection in AR and CB datasets, respectively. As described before, we followed a k-fold validation procedure in these experiments, where the *missing* regions had no overlap among folds. The results in the plots are the average over the $k = 5$ folds. The datasets in this application are characterized by high class imbalance, which affects the average F1-score. So, we decided to report for these experiments the F1-Score of the class *deforestation*.

For both datasets, in all cases, the overall accuracies were close to 100%. This result was not unexpected, considering that most samples are concentrated in one class. Another consequence is that the absolute differences among the approaches is very small and probably not statistically significant. The F1-score metric is in this case more appropriate to compare the methods.

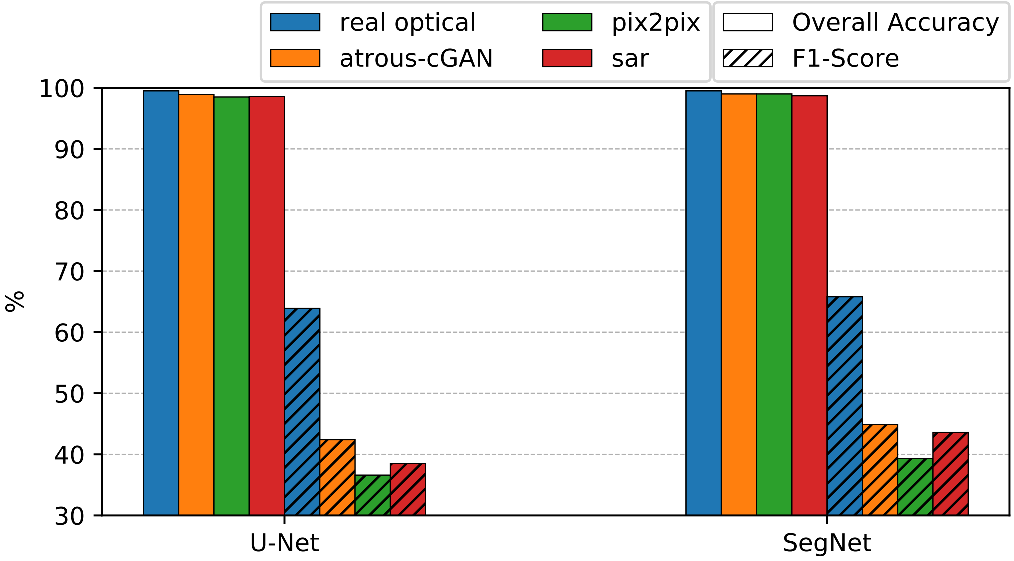


Figure 5.11: Performance of the Semantic Segmentation in terms of OA and F1 on Amazon Rainforest dataset.

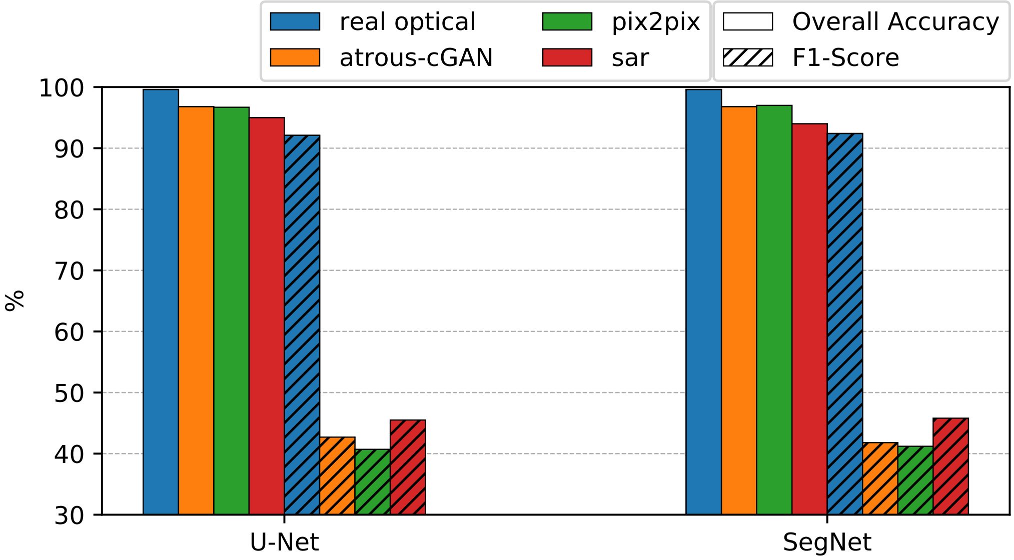


Figure 5.12: Performance of the Semantic Segmentation in terms of OA and F1 on Cerrado Biome dataset.

In the AR dataset, the F1-scores for U-Net and SegNet were very similar. In terms of F1-score, the results for *atrous-cGAN* was superior to the ones obtained on the optical image synthesized by *pix2pix* and on the SAR image with both segmentation networks. The results on the images synthesized by the *atrous-cGAN*, although quite inferior, again achieved the closest performance when compared to the original optical images (recall that for this application two images t_0 and t_1 built the input for the segmentation networks). On the other hand, the F1-Scores recorded on images synthesized by the *pix2pix* model were even inferior to the results on the SAR data.

Figure 5.13 shows the difference of F1-scores ($F1_{atrouscGAN}$-$F1_{pix2pix}$) between results recorded on image generated by the *atrous-cGAN* and *pix2pix* on each fold. The difference was positive in all folds for both networks. These results are an evidence of the superiority of the proposed method over the baseline for the AR dataset.

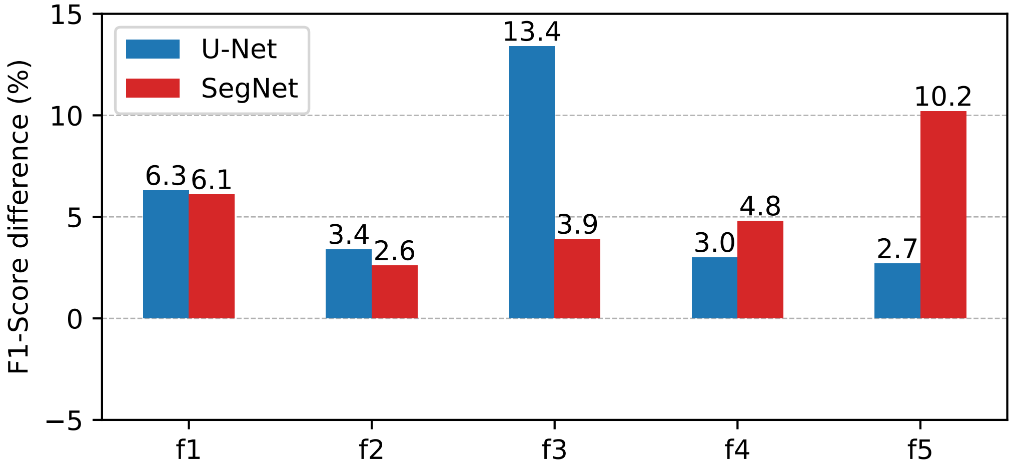


Figure 5.13: F1-Score difference *atrous-cGAN−pix2pix* for each fold in Amazon Rainforest dataset.

In the CB dataset, the F1-Score computed on the semantic segmentation of SAR images was the closest one to what has been obtained on the real optical. The optical images synthesized by both methods were behind the SAR image in this case. It is noteworthy that the F1-Scores recorded for the synthetic and SAR image were considerably lower than what has been reached using the real optical image. The semantic segmentation using the original optical images seems almost unaffected by the class imbalance, since the F1-Score is comparable with the accuracy values.

Even without being a good result, there is a small superiority of the *atrous cGAN* synthesis over the *pix2pix*. We also computed the F1-Score difference ($F1_{atrouscGAN}$-$F1_{pix2pix}$) for the CB dataset in each fold. As Figure 5.14 shows, the difference was positive in four out of five folds, but small in most cases.

The main conclusion that can be drawn from these experiments is that the SAR data used in the configuration adopted in our experiments were poorly suited for detecting deforestation, at least at our target sites. Therefore, the synthesis of optical images from SAR data was not shown to be an adequate

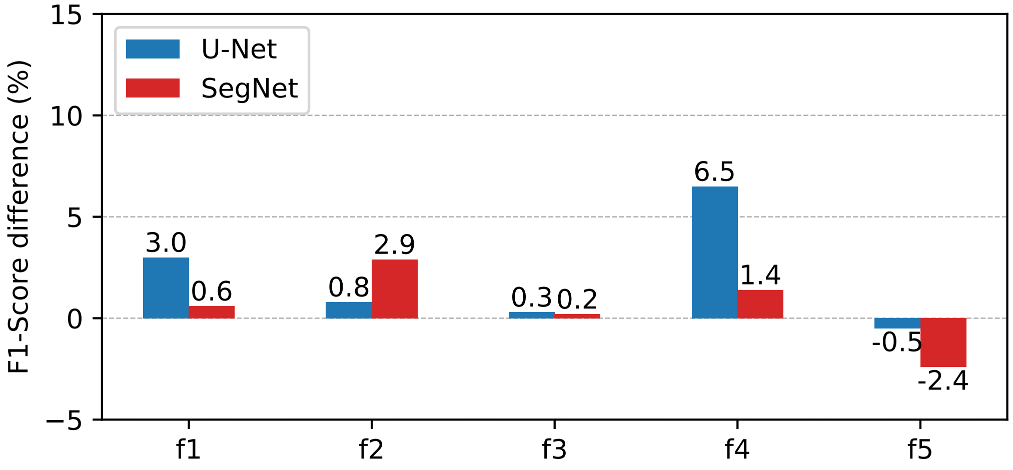


Figure 5.14: F1-Score difference *atrous-cGAN*−*pix2pix* for each fold in Cerrado Biome dataset.

method of feature learning for deforestation detection. The main evidence in favor of this conclusion is the marked difference between the results obtained from the real optical image, and from the SAR images as well as the optical images derived from them. Even in such a scenario, the proposed *atrous-cGAN* still performed better than the *pix2pix*.

5.6.2 Visual Analysis

With the aim of evaluating the visual quality of the synthesized images, snips of the *missing* regions are shown on Figures 5.15 and 5.16, namely the real optical image, the optical images synthesized by the *atrous-cGAN* and *pix2pix*, and the SAR image. For each dataset, the figures show images of different sources for the same region, pointing out some details for a better comparison. The optical images, either real or fake, are composed of the visible RGB bands. To enhance the visualization, they were processed using the same contrast stretching transformation for each dataset. The illustrated SAR images are represented by the grayscale VH polarized band.

The results for CV and LEM datasets (Figure 5.15) show clearly that the images synthesized by the *atrous-cGAN* look more similar to the real optical image. Compared to the *pix2pix*, the proposed *atrous-cGAN* produced sharper images with better delineated edges and less noise over smooth areas. The same behavior can be observed on images corresponding to the AR and CB datasets (Figure 5.16). Especially in the CB images, although the proposed *atrous-cGAN* delivered a better result than *pix2pix*, the synthesized snips deviated

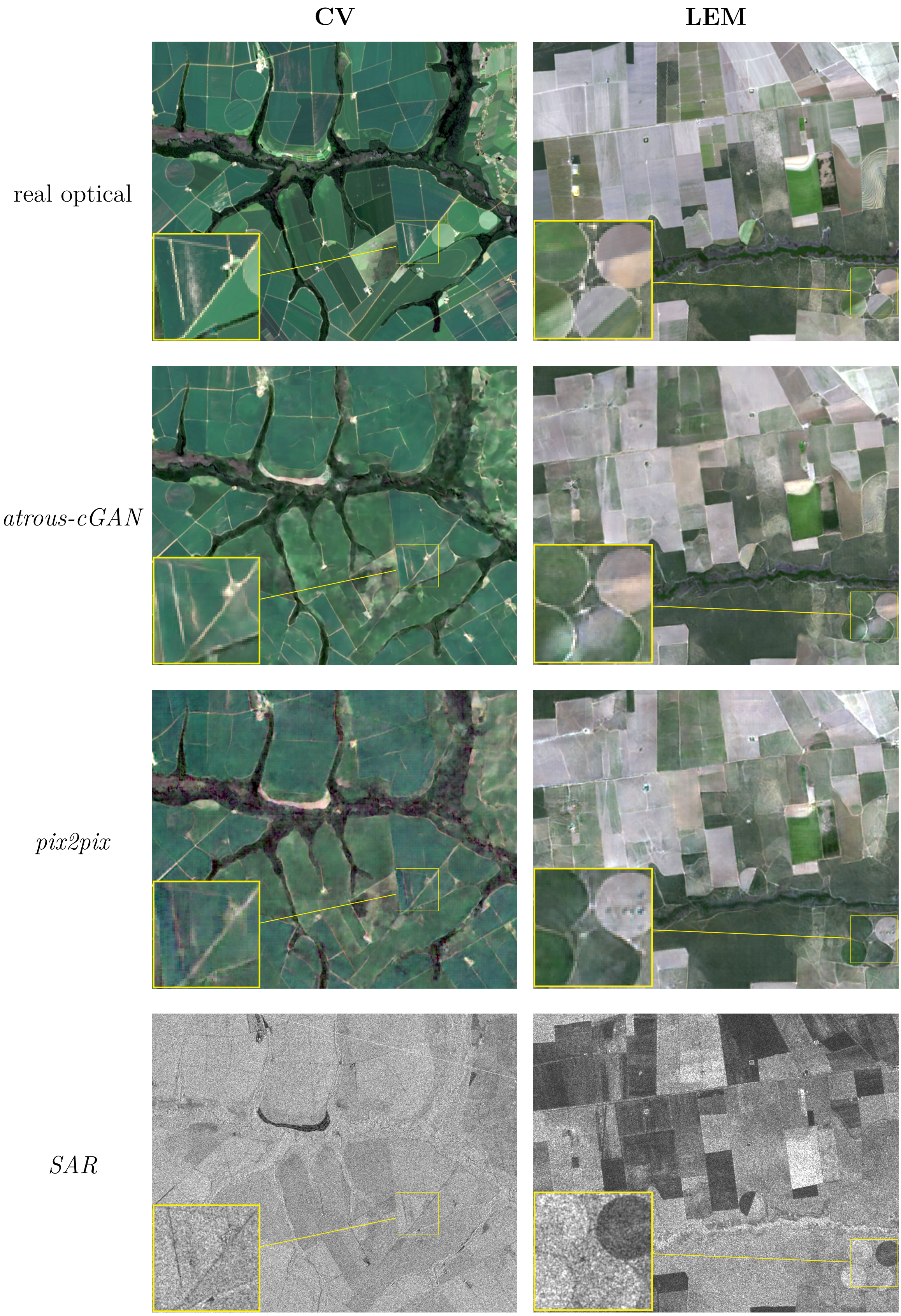


Figure 5.15: Snips of *missing* regions in CV and LEM datasets.

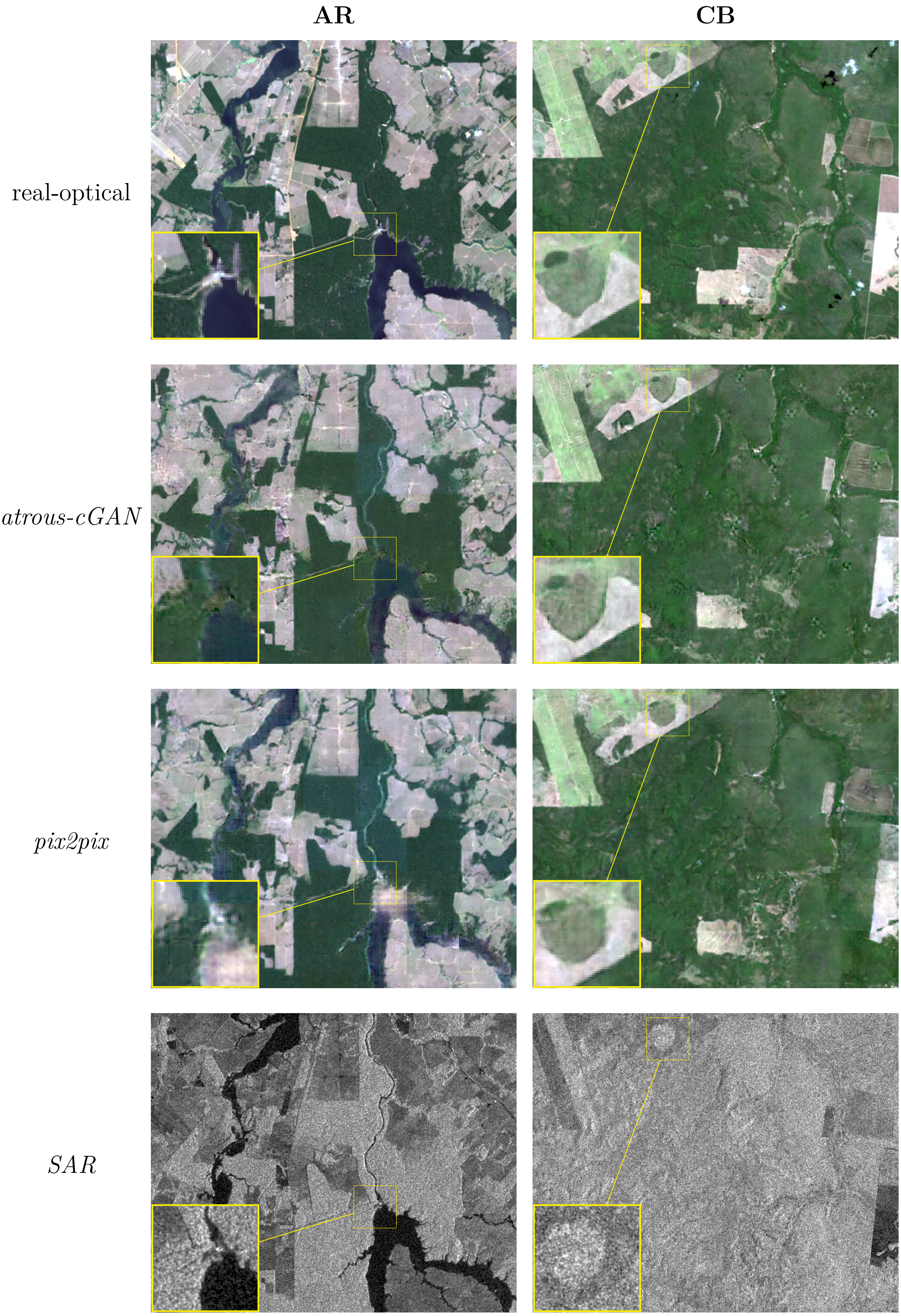


Figure 5.16: Snips of *missing* regions in AR and CB datasets.

significantly from the real optical images in some spots. This is consistent with the results recorded in the previous subsection for the semantic segmentation. As expected, in all datasets, the SAR images are less interpretable for the human eye.

5.6.3 Similarity

We also assessed the quality of cGANs outcomes by measuring the similarity between synthetic and real optical images in terms of Peak Signal to Noise Ratio (PSNR) and Spectral Angle Mapper (SAM). The values are shown in Table 5.8.

Table 5.8: Image similarity metrics.

		PSNR (db)	**SAM**
CV	*atrous-cGAN*	**35.4**	**2.2**
	pix2pix	34.9	2.4
LEM	*atrous-cGAN*	**46.2**	**2.8**
	pix2pix	45.8	3.0
AR	*atrous-cGAN*	**36.2**	**1.8**
	pix2pix	35.8	1.9
CB	*atrous-cGAN*	35.0	2.0
	pix2pix	**35.3**	**1.9**

The *atrous-cGAN* delivered slightly better scores than the *pix2pix* for CV, LEM and AR datasets. For the CB dataset, the *pix2pix* performed slightly better. The small difference expressed by these similarity metrics for CV, LEM, and AR datasets contrasts with the remarkable visual superiority of the *atrous-cGAN* images.

In some regions, the images synthesized by both methods do not even look like the real optical image. Such deviations, even if bound to a small portion of the image, may cancel out gains over other regions. Therefore, these results prevent claiming clear superiority of our proposal over the *pix2pix* in terms of PSNR and SAM.

5.6.4 Processing time

For the experiments made on this work, we used a computer equipped with an Intel(R) Core(TM) i7-8700K microprocessor at 3.70 GHz, and a graphic card NVIDIA GeForce RTX 2080 Ti. Table 5.9 shows the time consumption for the training of the proposed method and the baseline, on each dataset. Noticeable, the *atrous-cGAN* is the slowest model, given its greater depth, even containing considerably fewer parameters (see section 5.3).

Table 5.9: Training times (hours).

	CV	LEM	AR	CB
atrous-cGAN	11.7	18.1	84.3	5.5
pix2pix	6.7	10.1	45.9	4.5

The training times depend on the computational cost of the networks, the amount of training data, and the number of training epochs. Hence, to have a more precise difference between the methods, we also calculated the time consumed translating a single input patch SAR to the optical counterpart. Table 5.10 shows these results reaffirming the already established conclusions. Crop recognition datasets share the same inference time, since the same parameters and input size are used. The same occurs with deforestation detection datasets.

Table 5.10: Inference times for one input patch (seconds).

	CV and LEM	AR and CB
atrous-cGAN	3.0	2.8
pix2pix	2.1	1.9

6
Conclusions

In this work, we proposed a novel conditional Generative Adversarial Networks (cGANs) for SAR-to-optical image translation. The method has two main applications: firstly as a tool to synthesize missing optical data due to cloud covering, and secondly as a way to improve the comparatively poor interpretability of SAR images. Specifically, our design incorporates atrous convolutions and Atrous Spatial Pyramid Pooling (ASPP), which exploit the spatial context information at multiple scales.

The performance of the proposed *atrous-cGAN* was compared with the classical *pix2pix* approach. The synthesised images were evaluated in terms of the accuracy of the semantic segmentation and also considering their similarity with respect to the real optical images. Four public datasets from the Brazilian territory were used in our experiments, two of them related to crop recognition, and the other two, to deforestation detection.

In three of the four datasets, the semantic segmentation of the images synthesized by the *atrous-cGAN* achieved the closest scores to those of the optical images. Only for one dataset, our design was outperformed by the direct segmentation of the SAR image. Nevertheless, in all cases, the classification of the proposed synthesis was consistently more accurate than that achieved using the established *pix2pix* baseline.

In terms of visual quality, the *atrous-cGAN* produced less noisy optical images with finer and sharper details when compared with the *pix2pix*. Thus, the experimental analysis provided evidence that the proposed *atrous-cGAN* may also be a valuable aid to human analysts in the task of visual interpretation of SAR data. In terms of similarity metrics, the superiority of the proposed method was rather moderate.

The initial hypothesis that exploiting context at multiple scales as well as reducing downsampling would allow synthesizing sharper and less noisy optical images from their SAR counterpart than the traditional pix2pix was effectively confirmed by the experiments.

The reported results encourage the refinement of the proposed cGAN design in future works. The incorporation of data from multiple sensors and multitemporal as conditioning data seems promising. We also intend to test the method to synthesize images of other optical sensors, such as Sentinel-2A. We would like to compare the proposal also with other methods of the state of the art. Finally, the evaluation of the proposal in applications related to domain adaptation tasks is planed in the continuation of this research.

Bibliography

1 SINGH, P.; KOMODAKIS, N.. **Cloud-gan: Cloud removal for sentinel-2 imagery using a cyclic consistent generative adversarial networks.** In: IGARSS 2018-2018 IEEE INTERNATIONAL GEOSCIENCE AND REMOTE SENSING SYMPOSIUM, p. 1772–1775. IEEE, 2018.

2 EBERHARDT, I. D. R.; SCHULTZ, B.; RIZZI, R.; SANCHES, I. D.; FORMAGGIO, A. R.; ATZBERGER, C.; MELLO, M. P.; IMMITZER, M.; TRABAQUINI, K.; FOSCHIERA, W. ; OTHERS. **Cloud cover assessment for operational crop monitoring systems in tropical areas.** Remote Sensing, 8(3):219, 2016.

3 ZHU, Z.; WOODCOCK, C. E.. **Object-based cloud and cloud shadow detection in landsat imagery.** Remote sensing of environment, 118:83–94, 2012.

4 KING, M. D.; PLATNICK, S.; MENZEL, W. P.; ACKERMAN, S. A. ; HUBANKS, P. A.. **Spatial and temporal distribution of clouds observed by modis onboard the terra and aqua satellites.** IEEE Transactions on Geoscience and Remote Sensing, 51(7):3826–3852, 2013.

5 SHEN, H.; LI, H.; QIAN, Y.; ZHANG, L. ; YUAN, Q.. **An effective thin cloud removal procedure for visible remote sensing images.** ISPRS Journal of Photogrammetry and Remote Sensing, 96:224–235, 2014.

6 SOPHIA, D. L.; LALITHA, K. ; CHANDAR, J. P.. **Reconstruction of cloud contaminated remote sensing images using inpainting strategy.** International Journal of Electronics Communication and Computer Technology, 3(3):407–411, 2013.

7 CHENG, Q.; SHEN, H.; ZHANG, L.; YUAN, Q. ; ZENG, C.. **Cloud removal for remotely sensed images by similar pixel replacement guided with a spatio-temporal mrf model.** ISPRS journal of photogrammetry and remote sensing, 92:54–68, 2014.

8 HUANG, Y.; LIU, H.; YU, B.; WU, J.; KANG, E. L.; XU, M.; WANG, S.; KLEIN, A. ; CHEN, Y.. **Improving modis snow products with a**

hmrf-based spatio-temporal modeling technique in the upper rio grande basin. Remote Sensing of Environment, 204:568–582, 2018.

9 BERMUDEZ, J.; HAPP, P.; OLIVEIRA, D. ; FEITOSA, R.. **Sar to optical image synthesis for cloud removal with generative adversarial networks**. ISPRS Annals of Photogrammetry, Remote Sensing & Spatial Information Sciences, 4(1), 2018.

10 CHEN, Y.; TANG, L.; YANG, X.; FAN, R.; BILAL, M. ; LI, Q.. **Thick clouds removal from multitemporal zy-3 satellite images using deep learning**. IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing, 2019.

11 LI, Y.; FU, R.; MENG, X.; JIN, W. ; SHAO, F.. **A sar-to-optical image translation method based on conditional generation adversarial network (cgan)**. IEEE Access, 2020.

12 HUGHES, L. H.; MERKLE, N.; BÜRGMANN, T.; AUER, S. ; SCHMITT, M.. **Deep learning for sar-optical image matching**. In: IGARSS 2019-2019 IEEE INTERNATIONAL GEOSCIENCE AND REMOTE SENSING SYMPOSIUM, p. 4877–4880. IEEE, 2019.

13 BERMUDEZ, J. D.; HAPP, P. N.; FEITOSA, R. Q. ; OLIVEIRA, D. A.. **Synthesis of multispectral optical images from sar/optical multitemporal data using conditional generative adversarial networks**. IEEE Geoscience and Remote Sensing Letters, 16(8):1220–1224, 2019.

14 BALL, J. E.; ANDERSON, D. T. ; CHAN, C. S.. **Comprehensive survey of deep learning in remote sensing: theories, tools, and challenges for the community**. Journal of Applied Remote Sensing, 11(4):042609, 2017.

15 GOODFELLOW, I.; POUGET-ABADIE, J.; MIRZA, M.; XU, B.; WARDE-FARLEY, D.; OZAIR, S.; COURVILLE, A. ; BENGIO, Y.. **Generative adversarial nets**. In: ADVANCES IN NEURAL INFORMATION PROCESSING SYSTEMS, p. 2672–2680, 2014.

16 BERTALMIO, M.; SAPIRO, G.; CASELLES, V. ; BALLESTER, C.. **Image inpainting**. In: PROCEEDINGS OF THE 27TH ANNUAL CONFERENCE ON COMPUTER GRAPHICS AND INTERACTIVE TECHNIQUES, p. 417–424, 2000.

17 ANOOSHEH, A.; AGUSTSSON, E.; TIMOFTE, R. ; VAN GOOL, L.. **Combogan: Unrestrained scalability for image domain translation**. In:

PROCEEDINGS OF THE IEEE CONFERENCE ON COMPUTER VISION AND PATTERN RECOGNITION WORKSHOPS, p. 783–790, 2018.

18 LU, X.; ZHANG, J. ; ZHOU, J.. **Remote sensing image translation using spatial-frequency consistency gan.** In: 2019 12TH INTERNATIONAL CONGRESS ON IMAGE AND SIGNAL PROCESSING, BIOMEDICAL ENGINEERING AND INFORMATICS (CISP-BMEI), p. 1–6. IEEE, 2019.

19 GAO, J.; YUAN, Q.; LI, J.; ZHANG, H. ; SU, X.. **Cloud removal with fusion of high resolution optical and sar images using generative adversarial networks.** Remote Sensing, 12(1):191, 2020.

20 XIA, Y.; ZHANG, H.; ZHANG, L. ; FAN, Z.. **Cloud removal of optical remote sensing imagery with multitemporal sar-optical data using x-mtgan.** In: IGARSS 2019-2019 IEEE INTERNATIONAL GEOSCIENCE AND REMOTE SENSING SYMPOSIUM, p. 3396–3399. IEEE, 2019.

21 ISOLA, P.; ZHU, J. Y.; ZHOU, T. ; EFROS, A. A.. **Image-to-image translation with conditional adversarial networks.** In: PROCEEDINGS - 30TH IEEE CONFERENCE ON COMPUTER VISION AND PATTERN RECOGNITION, CVPR 2017, volumen 2017-Janua, p. 5967–5976, 2017.

22 MIRZA, M.; OSINDERO, S.. **Conditional generative adversarial nets.** arXiv preprint arXiv:1411.1784, 2014.

23 RONNEBERGER, O.; FISCHER, P. ; BROX, T.. **U-net: Convolutional networks for biomedical image segmentation.** In: INTERNATIONAL CONFERENCE ON MEDICAL IMAGE COMPUTING AND COMPUTER-ASSISTED INTERVENTION, p. 234–241. Springer, 2015.

24 CHEN, L.-C.; PAPANDREOU, G.; KOKKINOS, I.; MURPHY, K. ; YUILLE, A. L.. **Semantic image segmentation with deep convolutional nets and fully connected crfs.** arXiv preprint arXiv:1412.7062, 2014.

25 HINTON, G. E.; SALAKHUTDINOV, R. R.. **Reducing the dimensionality of data with neural networks.** science, 313(5786):504–507, 2006.

26 WANG, T.-C.; LIU, M.-Y.; ZHU, J.-Y.; TAO, A.; KAUTZ, J. ; CATANZARO, B.. **High-resolution image synthesis and semantic manipulation with conditional gans.** In: PROCEEDINGS OF THE IEEE CONFERENCE ON COMPUTER VISION AND PATTERN RECOGNITION, p. 8798–8807, 2018.

27 CHEN, L.-C.; PAPANDREOU, G.; KOKKINOS, I.; MURPHY, K. ; YUILLE, A. L.. **Deeplab: Semantic image segmentation with deep convolutional nets, atrous convolution, and fully connected crfs**. IEEE transactions on pattern analysis and machine intelligence, 40(4):834–848, 2017.

28 CHEN, L.-C.; PAPANDREOU, G.; SCHROFF, F. ; ADAM, H.. **Rethinking atrous convolution for semantic image segmentation**. arXiv preprint arXiv:1706.05587, 2017.

29 CHEN, L.-C.; ZHU, Y.; PAPANDREOU, G.; SCHROFF, F. ; ADAM, H.. **Encoder-decoder with atrous separable convolution for semantic image segmentation**. In: PROCEEDINGS OF THE EUROPEAN CONFERENCE ON COMPUTER VISION (ECCV), p. 801–818, 2018.

30 ZHANG, Y.; SONG, L.; XIE, R. ; ZHANG, W.. **Multi-scale generative adversarial learning for facial attribute transfer**. In: INTERNATIONAL FORUM ON DIGITAL TV AND WIRELESS MULTIMEDIA COMMUNICATIONS, p. 91–102. Springer, 2019.

31 LUTZ, S.; AMPLIANITIS, K. ; SMOLIC, A.. **AlphaGAN: Generative adversarial networks for natural image matting**. 2018.

32 WANG, J.; COHEN, M. F.. **Image and video matting: a survey**. Now Publishers Inc, 2008.

33 HELMER, E. H.; RUEFENACHT, B.. **Cloud-free satellite image mosaics with regression trees and histogram matching**. Photogrammetric Engineering & Remote Sensing, 71(9):1079–1089, 2005.

34 GONZALES, R. C.; WOODS, R. E.. **Digital image processing**, 2002.

35 WANG, M.; ZHENG, X. ; FENG, C.. **Color constancy enhancement for multi-spectral remote sensing images**. In: 2013 IEEE INTERNATIONAL GEOSCIENCE AND REMOTE SENSING SYMPOSIUM-IGARSS, p. 864–867. IEEE, 2013.

36 LIU, J.; WANG, X.; CHEN, M.; LIU, S.; ZHOU, X.; SHAO, Z. ; LIU, P.. **Thin cloud removal from single satellite images**. Optics express, 22(1):618–632, 2014.

37 MAKARAU, A.; RICHTER, R.; MÜLLER, R. ; REINARTZ, P.. **Haze detection and removal in remotely sensed multispectral imagery**.

IEEE Transactions on Geoscience and Remote Sensing, 52(9):5895–5905, 2014.

38 PONOMAREV, V. I.; POGREBNYAK, O. B.. **Image enhancement by homomorphic filters**. In: APPLICATIONS OF DIGITAL IMAGE PROCESSING XVIII, volumen 2564, p. 153–159. International Society for Optics and Photonics, 1995.

39 LIN, C.-H.; TSAI, P.-H.; LAI, K.-H. ; CHEN, J.-Y.. **Cloud removal from multitemporal satellite images using information cloning**. IEEE transactions on geoscience and remote sensing, 51(1):232–241, 2012.

40 HUANG, C.; THOMAS, N.; GOWARD, S. N.; MASEK, J. G.; ZHU, Z.; TOWNSHEND, J. R. ; VOGELMANN, J. E.. **Automated masking of cloud and cloud shadow for forest change analysis using landsat images**. International Journal of Remote Sensing, 31(20):5449–5464, 2010.

41 ZHU, J.-Y.; PARK, T.; ISOLA, P. ; EFROS, A. A.. **Unpaired image-to-image translation using cycle-consistent adversarial networks**. In: PROCEEDINGS OF THE IEEE INTERNATIONAL CONFERENCE ON COMPUTER VISION, p. 2223–2232, 2017.

42 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):2067, 2019.

43 CANTY, M. J.. **Image analysis, classification and change detection in remote sensing: with algorithms for ENVI/IDL and Python**. Crc Press, 2014.

44 LILLESAND, T.; KIEFER, R. W. ; CHIPMAN, J.. **Remote sensing and image interpretation**. John Wiley & Sons, 2015.

45 RICHARDS, J. A.; RICHARDS, J.. **Remote sensing digital image analysis**, volumen 3. Springer, 1999.

46 **Planck's radiation law**, 2009.

47 SHAN, J.; TOTH, C. K.. **Topographic laser ranging and scanning: principles and processing**. CRC press, 2018.

48 RICHARDS, J. A.; OTHERS. **Remote sensing with imaging radar**, volumen 1. Springer, 2009.

49 SAHA, S. K.. **Aperture synthesis: methods and applications to optical astronomy**. Springer Science & Business Media, 2010.

50 GOODFELLOW, I.; BENGIO, Y. ; COURVILLE, A.. **Deep learning**. MIT press, 2016.

51 LECUN, Y.; BOTTOU, L.; BENGIO, Y. ; HAFFNER, P.. **Gradient-based learning applied to document recognition**. Proceedings of the IEEE, 86(11):2278–2324, 1998.

52 HUBEL, D. H.; WIESEL, T. N.. **Receptive fields and functional architecture of monkey striate cortex**. The Journal of physiology, 195(1):215–243, 1968.

53 FUKUSHIMA, K.; MIYAKE, S. ; ITO, T.. **Neocognitron: A neural network model for a mechanism of visual pattern recognition**. IEEE transactions on systems, man, and cybernetics, (5):826–834, 1983.

54 MUELLER, J. P.; MASSARON, L.. **Data Science Programming All-in-One For Dummies**, volumen 1. John Wiley Sons, 2019.

55 CSÁJI, B. C.; OTHERS. **Approximation with artificial neural networks**. Faculty of Sciences, Etvs Lornd University, Hungary, 24(48):7, 2001.

56 YAMAGUCHI, K.; SAKAMOTO, K.; AKABANE, T. ; FUJIMOTO, Y.. **A neural network for speaker-independent isolated word recognition**. In: FIRST INTERNATIONAL CONFERENCE ON SPOKEN LANGUAGE PROCESSING, 1990.

57 IOFFE, S.; SZEGEDY, C.. **Batch normalization: Accelerating deep network training by reducing internal covariate shift**. arXiv preprint arXiv:1502.03167, 2015.

58 SANTURKAR, S.; TSIPRAS, D.; ILYAS, A. ; MADRY, A.. **How does batch normalization help optimization?** In: ADVANCES IN NEURAL INFORMATION PROCESSING SYSTEMS, p. 2483–2493, 2018.

59 WERBOS, P.. **Beyond regression:" new tools for prediction and analysis in the behavioral sciences**. Ph. D. dissertation, Harvard University, 1974.

60 HINTON, G. E.; SRIVASTAVA, N.; KRIZHEVSKY, A.; SUTSKEVER, I. ; SALAKHUTDINOV, R. R.. **Improving neural networks by preventing co-adaptation of feature detectors**. arXiv preprint arXiv:1207.0580, 2012.

61 LONG, J.; SHELHAMER, E. ; DARRELL, T.. **Fully convolutional networks for semantic segmentation**. In: PROCEEDINGS OF THE IEEE CONFERENCE ON COMPUTER VISION AND PATTERN RECOGNITION, p. 3431–3440, 2015.

62 BADRINARAYANAN, V.; KENDALL, A. ; CIPOLLA, R.. **Segnet: A deep convolutional encoder-decoder architecture for image segmentation**. IEEE transactions on pattern analysis and machine intelligence, 39(12):2481–2495, 2017.

63 DUMOULIN, V.; VISIN, F.. **A guide to convolution arithmetic for deep learning**. arXiv preprint arXiv:1603.07285, 2016.

64 MITCHELL, T. M.. **Generative and discriminative classifiers: Naive bayes and logistic regression**. Machine learning, p. 1–17, 2010.

65 HE, K.; ZHANG, X.; REN, S. ; SUN, J.. **Deep residual learning for image recognition**. In: PROCEEDINGS OF THE IEEE CONFERENCE ON COMPUTER VISION AND PATTERN RECOGNITION, p. 770–778, 2016.

66 GLOROT, X.; BENGIO, Y.. **Understanding the difficulty of training deep feedforward neural networks**. In: PROCEEDINGS OF THE THIRTEENTH INTERNATIONAL CONFERENCE ON ARTIFICIAL INTELLIGENCE AND STATISTICS, p. 249–256, 2010.

67 LIN, M.; CHEN, Q. ; YAN, S.. **Network in network**. arXiv preprint arXiv:1312.4400, 2013.

68 LIU, W.; RABINOVICH, A. ; BERG, A. C.. **Parsenet: Looking wider to see better**. arXiv preprint arXiv:1506.04579, 2015.

69 SANCHES, I. D.; FEITOSA, R. Q.; DIAZ, P. M. A.; SOARES, M. D.; LUIZ, A. J. B.; SCHULTZ, B. ; MAURANO, L. E. P.. **Campo verde database: Seeking to improve agricultural remote sensing of tropical areas**. IEEE Geoscience and Remote Sensing Letters, 15(3):369–373, 2018.

70 SANCHES, I. D.; FEITOSA, R. Q.; ACHANCCARAY, P.; MONTIBELLER, B.; LUIZ, A. J. B.; SOARES, M. D.; PRUDENTE, V. H. R.; VIEIRA, D. C. ; MAURANO, L. E. P.. **Lem benchmark database for tropical agricultural remote sensing application**. ISPRS - International Archives of the Photogrammetry, Remote Sensing and Spatial Information Sciences, XLII-1:387–392, 2018.

71 MUCHAGATA, M.; BROWN, K.. **Cows, colonists and trees: rethinking cattle and environmental degradation in brazilian amazonia**. Agricultural systems, 76(3):797–816, 2003.

72 MARRIS, E.. **The forgotten ecosystem**, 2005.

73 ABADI, M.; BARHAM, P.; CHEN, J.; CHEN, Z.; DAVIS, A.; DEAN, J.; DEVIN, M.; GHEMAWAT, S.; IRVING, G.; ISARD, M. ; OTHERS. **Tensorflow: A system for large-scale machine learning**. In: 12TH {USENIX} SYMPOSIUM ON OPERATING SYSTEMS DESIGN AND IMPLEMENTATION ({OSDI} 16), p. 265–283, 2016.

74 KINGMA, D. P.; BA, J.. **Adam: A method for stochastic optimization**. arXiv preprint arXiv:1412.6980, 2014.

75 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):910, 2020.

76 GOWARD, S. N.; MARKHAM, B.; DYE, D. G.; DULANEY, W. ; YANG, J.. **Normalized difference vegetation index measurements from the advanced very high resolution radiometer**. Remote sensing of environment, 35(2-3):257–277, 1991.

77 KRUSE, F. A.; LEFKOFF, A.; BOARDMAN, J.; HEIDEBRECHT, K.; SHAPIRO, A.; BARLOON, P. ; GOETZ, A.. **The spectral image processing system (sips)-interactive visualization and analysis of imaging spectrometer data**. In: AIP CONFERENCE PROCEEDINGS, volumen 283, p. 192–201. American Institute of Physics, 1993.