Escuela Superior Politécnica del Litoral

Faculty of Electrical and Computer Engineering

Processing and representation of multispectral images using deep learning techniques

DEGREE WORK

Prior to obtaining the title of:

DOCTOR IN APPLIED COMPUTATIONAL SCIENCES

Author:

Patricia Leonor Suárez Riofrío

Guayaquil - Ecuador

2020

Director	**Dr. Angel Sappa** Escuela Superior Politécnica del Litoral, Ecuador
Co-Director	**Dr. Boris Vintimilla** Escuela Superior Politécnica del Litoral, Ecuador
Thesis committee	**Dr. Cristian Aguilera Carrasco** Universidad Tenológica de Chile, Chile
	Dr. Federico Dominguez Escuela Superior Politécnica del Litoral, Ecuador
	Dr. Jean-Bernard Hayet Centro de Investigación en Matemáticas, México
	Dr. Miguel Realpe Escuela Superior Politécnica del Litoral, Ecuador

Be wise in the way you act toward outsiders;
make the most of every opportunity.
Let your conversation be always full of grace,
seasoned with salt, so that you may know
how to answer everyone.
Colossians 4:5-6

To Almighty God and
my family

Acknowledgements

I want to thank God for giving me the wisdom, patient, constancy, health, and strength to complete this stage of my life.

To my lovely family, who advised and supported me all the time and could understand, that I could not always be with them on all the occasions they needed me.

To my director and mentor Dr. Angel Sappa, who from the beginning of my studies was the fundamental axis to achieve all the proposed objectives, showing me the way to follow, advised me and had me patience to teach me everything necessary to fulfill the research objectives defined in my thesis. He has always been willing to help and encourage me when I could no longer, without him, I could not have successfully completed my Ph.D. program. There are no words I can write that can express how grateful I am for all that he has done and continues doing for me. God has blessed me to have him as my mentor and now my friend because he believed in me, more than I believed myself, as he has made it possible for me to achieve all of my research goals during my doctoral studies.

To my co-director Dr. Boris Vintimilla for his guidance, patience, help and advice given during my doctoral studies.

Abstract

Computer vision is a scientific discipline that has been developed in recent decades due to technological advances in acquisition devices together with the increase on computational capabilities. The reduction in prices of hardware, both acquisition and processing, allows this technology to be available to most users. Additionally, there is a technological advance that allows sensors to be sensitive to different spectra, including smart mobile devices. Computer vision is defined as a field of study that develops multiple techniques to ensure that machines can "see" and "understand" information in images or videos of any spectrum, using mathematical models that process, analyze and interpret digital information extracted from images.

With the advance of convolutional neural networks (CNN), the usage of machine learning based techniques has made great progress in recent years. Specifically, many techniques have been developed to implement a process similar to the visual reasoning of human vision, to perform tasks such as detection, recognition, segmentation, coloring, filtering, improvement, similarity, etc. using CNN. This thesis presents a series of CNN-based techniques applied to images of different spectra, especially the near infrared spectrum (NIR) and the visible spectrum. Among the techniques implemented are: perform similarity detection between images of VISIBLE and NIR spectra, colorization of NIR images, estimation of normalized difference vegetation index (NDVI) using only one band of the spectrum and eliminate the haze present in the images. It should be noted that to implement these techniques, generative adversarial models have been used in their standard, conditional, stacked and cyclic variants, which are the latest generation in these type of networks.

Keywords: Convolutional Neural Networks, Generative Adversarial Network, Infrared Imagery colorization, Haze, Normalized Difference Vegetation Index, Stacked Generative Adversarial Network.

Resumen

La visión por computadora es una disciplina científica que se ha desarrollado en las últimas décadas debido a los avances tecnológicos en los dispositivos de adquisición, así como a la reducción de los precios, que les permite estar disponibles para la mayoría de los usuarios. Además, hay un avance tecnológico que permite que los sensores sean sensibles a diferentes espectros, incluidos los dispositivos móviles inteligentes. Vision por computador se define como un campo de estudio que desarrolla múltiples técnicas para garantizar que las máquinas puedan ”ver” y ”comprender” la información de las imágenes o videos de cualquier espectro, utilizando modelos matemáticos que procesan, analizan e interpretan la información digital extraída de las imágenes.

Con el avance de las redes neuronales convolucionales (CNN) distintas técnicas basadas en aprendizaje automático han sido propuestas en los últimos años. Específicamente, se han desarrollado muchas técnicas para implementar un proceso similar al razonamiento visual de la visión humana, para realizar tareas como detección, reconocimiento, segmentación, coloración, filtrado, mejora, similitud, etc. utilizando CNN. Esta tesis presenta una serie de técnicas basadas en CNN aplicadas a imágenes de diferentes espectros, especialmente el espectro infrarrojo cercano (NIR) y el espectro visible. Entre las técnicas implementadas se encuentran: realizar detección de similitud entre imágenes de espectros VISIBLES y NIR, coloración de imágenes NIR, generación de índice de vegetación de diferencia normalizada (NDVI), usando solo una banda del espectro, y remoción de la neblina presente en las imágenes. Cabe señalar que, para implementar estas técnicas, se han utilizado modelos generativos adversariales en sus variantes estándar, condicionales, apiladas y cíclicas, que son la última generación en este tipo de redes.

Palabras clave: Redes neuronales convolucionales, Red generativa adversarial, Coloración de imágenes infrarrojas, Neblina, Indice de vegetación de diferencia normalizada, Red generativa adversarial apilada.

Contents

List of Figures

List of Tables

Chapter 1

Introduction

Ultimately, in the field of computer vision, techniques based on the use of several spectra, not only the visible spectrum, are being proposed to address problems of detection, recognition, composition of materials, surface characteristics, among others. For example, with the use of cross-spectral images (visible and near infrared), additional information is obtained for each spectrum that can be used to improve the different existing visualization techniques. With these images, great advantages can be obtained in recognition and detection tasks. To acquire this type of cross-spectral data set, multiple cameras are often used, which requires alignment or estimation of the disparity of the images. Increasingly, multi-camera cross-spectrum systems are integrated into active RGBD devices (for example, RGB-NIR cameras in Kinect and iPhone X). However, searching for the coincidences in images from different spectral bands is a challenge due to the great variations in the appearance, therefore, it is necessary to register them to obtain the parity of each scene in both spectra [127]. The use and combination of different spectral bands that determine the appropriate correspondences between them are necessary to improve the vision techniques that previously only used images of the visible spectrum.

Generally, there are some limitations with the lighting conditions, material texture or temperature that considerably affect the performance of computer vision techniques when only visible spectrum images are used. These limitations can be overcomed using cross-spectral images, to tackle the design of computer vision problems more effectively.

There are some approaches designed to use images from different spectral bands to exploit all features inherent to those bands other than visible and in that way improve the performance of the techniques. Some of these techniques can be applied in the following ways; in surveillance systems, using the dense depth information on images from the thermal and visible spectrum to determine the differences between them and obtain adequate estimates for object localization or environment navigation [78].

Another technique used to find self-correlation similarity between image regions using a descriptor could be based on multi-spectral to further improve the matching quality and

runtime efficiency [54], [3]. In biometrics, the spectral information can be used to enhance the accuracy of the matching process; for example, spectral signatures of different regions of the iris, face or fingertip can be created from multi-spectral band images, based on the estimation of the similarity between those signatures to predict a corresponding biometric representation. Some applications use multi-spectral information to detect any given material composition, based on the reflectance and electromagnetic radiation emitted, for example, to classify defects presented on white maize [92].

1.1 Motivation

Artificial intelligence (AI) has had a breakthrough in the last decade. This advance of AI is attributed to five main factors: more, cheaper and better hardware computing power, better designed for computer vision tasks, the availability of a large amount of information (big data) and hardware accelerators (graphics processing units (GPU) that were originally designed for the gaming industry, and tensor processing units (TPU). With the GPU's, now it turns out to be very useful in computing because it accelerates deep learning, analysis, and engineering applications. GPUs have unlocked the potential of neural networks using deep learning as an approach to computer vision, along with the huge amount of currently available data has led to qualitative improvements in computer vision algorithms.

Some of the big corporations, such as, NVIDIA, ORBCOMM, Google, Microsoft, Facebook, etc., are the ones that drive most research and AI-based products or services, making the technology commercially available in the form of an API (interface of application program), deep learning libraries, personal and professional agents, chatbots, robots, and many other interesting products [99].

Computer vision is one of the sciences that uses AI advances, for the treatment and analysis of images and videos, with the aim of understanding the scenes and facilitating decision making. In such a way that using the images of different spectra to the visible one, allows to improve the techniques that have been applied only with images from the visible spectrum.

The simultaneous use of images from different spectra can be helpful to improve the performance of many computer vision tasks. The core idea behind the usage of cross-spectral approaches is to take advantage of the strengths of each spectral band providing a richer representation of a scene, which cannot be obtained with just images from one spectral band.

During the last decade, computer vision has had a great advance. Four underlying reasons have driven the breakthroughs: more and better hardware computing power, cheaper, better designed for computer vision tasks, and a lot of data (Big Data). Also, the Graphics Processing Unit (GPU) chips that were originally designed for the gaming industry, now turns out to be very useful in computing because it accelerates deep learning, analysis, and engineering applications. GPUs have unlocked the potential of neural networks using deep learning as an approach to computer vision, along with the huge amount of currently available data has led to qualitative improvements in computer vision algorithms.

Recently Convolutional Neural Network (CNN) based approaches are becoming the dominant paradigm in almost every computer vision task. CNN's have shown outstanding results in various and diverse computer vision tasks such as stereo vision, metric 3D information using computer vision as a measurement device, or as a source of semantic information like objects, activities, locations, faces, gestures, motion, emotions, text/writing, scenes, etc. There are several applications in use today for 3D urban modeling like Google Maps or Photosynth by Microsoft. Also there are face or smile detection used in cameras like Cannon, Sonny, Fuji, Apple, etc. Think of what more can be done by machines when they are able to detect, recognize, describe an object in a given scene as accurate as a human eye. The human eye is a complex structure and it goes through a more complex phenomenon of understanding the environment. Similarly, making machines see things and make them capable enough to figure out what they are seeing and further categorize it, is still a pretty tough job. Powered by GPUs and tons of data, deep neural networks are driving progress today. Computer vision applications can be found in almost every domain, including topics such as medical imaging, gaming, video surveillance, multimedia, industrial applications, remote sensing, just to mention a few. In most of the cases, these applications are based on images obtained from cameras working at the visible spectrum. Nevertheless, the appealing factor of using images from different spectral bands lies on the other hand on the possibility to obtain information that cannot be seen at the visible spectrum, opening a new range of possibilities to be explored in the computer vision and pattern recognition domain.

Image acquisition devices have largely expanded in recent years, mainly due to the decrease in price of electronics together with the increase in computational power. This increase in sensor technology has resulted in a large family of images, able to capture different information (from different spectral bands) or complementary information (2D, 3D, 4D); hence, we can have: HD 2D images; video sequences at a high frame rate; panoramic 3D images; multispectral images; just to mention a few.

Although innovative approaches have been developed over the past 20 years, the power of using a multispectral image with computer vision technology remains unknown to many potential end-users, such as decision-makers, farmers, environmental watchers in both the private and governmental sectors, city planners, stockholders, and others. This is mainly because the use of multispectral sensors has a relatively high cost of its final products and on the need for professional manpower to operate the instrument and process the data.

1.2 Objectives

This thesis focuses on the exploration of the use of information from the different bands of the electromagnetic spectrum in such a way that it can be exploited to solve some problems existing in the field of computer vision, for which different deep learning architectures have been designed to be used with multispectral images.

The research carried out during my doctoral studies includes the implementation of novel

approaches to the processing and representation of multispectral images using deep learning techniques, especially using cross-spectral information with visible and near-infrared spectrum images. The approaches addressed in the thesis are detailed below:

1. Cross-Spectral Image Similarity: Determine similarity and reach a performance similar or better than other methods based on only visible spectra.
2. NIR image colorization: Implement a novel colorization process using near infrared images.
3. NDVI vegetation index estimation: Facilitate the process of analyzing the health of the vegetation, avoiding dependence on acquisition devices sensitive to the near infrared spectra.
4. Image Dehazing: Improve the quality of the image using near infrared spectra images.

1.3 Thesis Outline

The remainder of this thesis is organized as follows. Chapter 2 summarizes the work related to techniques to determine the similarity of cross-spectral image regions (VISIBLE-NIR), using traditional, deep learning techniques. Also, several techniques with generative adversarial networks to solve the colorization problem using near infrared images are reviewed. Additionally, It has been covered vegetation index estimation based also on near infrared spectra, and image dehazing methods using only visible and cross-spectral (visible and near infrared) images based on CNN, especially, conditional generative adversarial networks with multiple loss approaches. Also includes work related to techniques to remove haze from images of both the visible and near-infrared spectrum using hand-craft and deep learning approaches.

Chapter 3 presents several approaches to determine patch similarity using cross-spectral dataset from [14]. Two approaches have been proposed, the first one using CNN and the second one based on meta-learning; from the dataset [14], the categories: *country, indoor, oldbuilding and Urban* have been selected. These images are the most affected in lighting conditions and textures, which directly affect the complexity of the process of establishing their similarity through the detection of characteristic points and, therefore, are the most challenging scenarios for the training process. First, the characteristic points of patches of the visible images have been obtained using the SIFT algorithm, to search these points in their corresponding on the near infrared spectrum images. To carry out the experiments, 64×64 pixels patches have been generated centered on the previously detected points in both the visible and near infrared images, since the images are perfectly aligned. Then the corresponding patches are extracted with the previously defined size. The first approach is based on a CNN network to determine the ability to compare regions (patches) of (visible-near) images, focusing on learning similarity between cross-spectral image patches with a 2 channel convolutional neural network (CNN) model. This proposed approach is an adaptation of previous work, trying to obtain similar results than the state of the art but with low-cost

hardware. The second approach is a technique based on meta-learning, which proposes an 8-$shot$ 1-way meta-learning metric based network model. For training, a total of 16 perfectly balanced cross-spectral image patches (matched and not matched) have been prepared for each category. The results of the experiments of the two approaches presented in chapter 3 have been compared to each other and also with another similar technique presented in [3].

Chapter 4 introduces works that tackle colorization grayscale and near infrared images using non-parametric methods and convolutional neural network (CNN) approaches, like Generative Adversarial Network (GAN), with local and global priors. This chapter proposes novel approaches for colorizing near infrared (NIR) images using (GAN's) architectures. The first model proposed is based on the usage of a single level of learning to colorize the images. On the second approach, a conditional model with a triplet level architecture is proposed for learning each color channel independently, in a more homogeneous way. It allows a fast convergence during the training, obtaining a greater similarity between the colored NIR image and the corresponding ground truth. The third approach presents a modified architecture that includes in the final layer of red channel the concatenation of the infrared image to enhance the details obtained from the learning process, resulting in a sharp RGB image. The fourth approach includes a novel stack model and multiple loss functions over a conditional probabilistic generative model. The use of a variant of stacked GAN architecture is proposed to add in each layer a feature hierarchy that encourages the representation manifold of the generator to align with that of the bottom-up discriminative network, leveraging the powerful discriminative representations to guide the generative model [44]. This stacked learning model allows accelerating the diversity obtained in the multiple level of training representing each of the channels of an image of the visible spectrum (RGB). Also, the model will receive as an input a near infrared patch (NIR) fused with Gaussian noise to ensure more diversity of colors. Also, in this approach a layer of Gaussian noise has been included in each level of the triplet architecture of the generator model to reinforce the generalization and therefore be able to optimize the learning of the colorization process. The proposed approaches have been evaluated with a large dataset of NIR images and compared to each other to determine the one that gets the best results. All these approaches are based on a GAN architectures where all the color channels are obtained at the same time and also the same metrics have been used to do the comparison.

Chapter 5 tackles the estimation of the Normalized Difference Vegetation Index (NDVI), to solve problems related to plants health to automate agriculture activities to improve yields productivity. Generative adversarial networks approaches have been used to implement this solution. Specifically, three different approaches are proposed; the first one is based on the usage of a Conditional (CGAN) architecture. In the first stage, it learns how to generate the NDVI index from the given input image. Three different architectures are evaluated, flat, siamese and triplet models. In the evaluated models, the final layer of the architecture considers the infrared image to enhance the details, resulting in a sharp NVDI image. The second approach obtains the NDVI from a single spectral band using a CGAN. The architecture has been designed with a flat scheme, Gaussian noise is added to each patch to increase the variability in the learning process of the generation of the NDVI index, reducing the time of

the convergence and generalization and obtaining a good performance to solve the index estimation problem. The last approach for NDVI index estimation proposes a novel approach to obtain it, just from a RGB image. The NDVI values are obtained by using images from the visible spectral band together with a synthetic near infrared image obtained by a cycled GAN. The cycled GAN network is able to obtain a NIR image from a given grayscale image. It is trained by using an unpaired set of grayscale and NIR images by using a U-net architecture and multiple loss functions (grayscale images are obtained from the provided RGB images). Then, the NIR image estimated with the proposed cycle generative adversarial network is used to compute the NDVI index. Experimental results are provided showing the validity of the proposed approach. Additionally, comparisons with previous approaches are also provided.

Chapter 6 summarizes several techniques to remove haze from images using stacked GAN and multiple loss functions. Two approaches have been tackled to solve this problem. The first one proposes to remove haze degradations in RGB images using a stacked CGAN. It employs a triplet of GAN to remove the haze on each color channel independently. A multiple loss functions scheme, applied over a conditional probabilistic model, is proposed. The proposed GAN architecture learns to remove the haze, using as conditioned entrance, the images with haze from which the clear images will be obtained. Such formulation ensures a fast model training convergence and a homogeneous model generalization. The last approach proposed to remove haze from RGB images using near infrared images based on a dense stacked conditional Generative Adversarial Network. The architecture of the implemented deep network receives, besides the images with haze, its corresponding image in the near infrared spectrum, which serve to accelerate the learning process of the details of the characteristics of the images. The model uses a triplet layer that allows the independent learning of each channel of the visible spectrum image to remove the haze on each color channel separately. Also, a multiple loss function scheme is proposed, which ensures a balanced learning between the colors and the structure of the images. Experimental results have shown that the proposed method effectively removes the haze from the images. Additionally, the proposed approach is compared with a state of the art approach showing better results.

Finally, Chapter 7 summarizes the entire work implemented in this thesis, highlighting the principal contributions and describing future architectures, new loss functions, datasets, and better data augmentation processes to increase the effectiveness and accuracy of the proposed approaches.

Chapter 2

Related Work

This chapter summarizes work related to the research topics covered in this thesis. It starts with cross-spectral image similarity, Section 2.1, which is the first topic related with the multispectral image processing. Then, a survey of image colorization approaches is given in Section 2.2. Next, a review of vegetation indexes is presented in Section 2.3. A survey of haze removal from images is addressed in Section 2.4. Additionally, works related to CNN, in particular, Generative Adversarial Network (GAN), are reviewed in Section 2.5, since this framework is used in most of the algorithms that have been implemented in this thesis. Furthermore, a summary of work based on cyclic GAN's networks is presented in Section 2.6.

2.1 Cross-Spectral Image Similarity

Images are often represented using dense pixel-based properties or by compact region descriptors (features) often used with interest point detectors. Dense properties include raw pixel intensity or color values from image patches. There are other techniques like common compact region descriptors that include distribution based descriptors (e.g., SIFT, SURF), differential descriptors (e.g., local derivatives), shape-based descriptors using extracted edges (e.g., shape context) and others. For a comprehensive comparison of local descriptors for image matching see [69].

Although these representations and their corresponding similarity measures may vary significantly, they all share the same basic assumption, that there exists a common underlying visual property (e.g., pixels colors, intensities, edges, gradients or other filter responses), which is shared by the two image patches, and can, therefore, be extracted and compared across images/sequences, see [93]. The comparison between the representations, using the aforementioned similarity measures can be embedded into learning methods, which are able to find the non-linear relationship between the representations. These learning-based approaches generally rely on some easy-to-compute distance metric (e.g., Hinge distance) that some times correlates with the semantic similarity. Different learning approaches have been proposed in the literature. Recently, Convolutional Neural Network (CNN) based learning

techniques are among the best option producing appealing results (e.g., see [21]).

CNN's are a specific type of neural network thoroughly used in deep learning algorithms. Their convolutional kernel-based philosophy makes them easy to apply in the computer vision domain for classical problems. One of them is the extraction of interesting parts of an image, obtaining feature vectors needed for a task like object detection, classification, segmentation, etc. These techniques do not ignore the structure and compositional nature of images, so they can learn to extract features directly from raw images, eliminating the need for manual feature extraction.

Several approaches for image patch similarity have been proposed in last years, some techniques are proposed based on hand-crafted methods, or using CCN networks. In [119] a novel region-based active contour model via local patch similarity measure for image segmentation is proposed. Using the spatial constraints on local region-based models to construct a patch similarity measure, which balances the noise suppression and the image details reservation. Another approach, [20], proposes a novel deep similarity learning method that trains a binary classifier to obtain the metric of the correspondence of two image patches. The classification output is transformed into a continuous probability value, then used as the similarity score. For the comparison, two commonly used metrics are presented: normalized mutual information and local cross-correlation.

Romero et al. [89] present an unsupervised deep feature extraction for remote sensing image classification, using a greedy layerwise unsupervised pretraining coupled with a highly efficient algorithm for unsupervised learning of sparse features. The algorithm is rooted on sparse representations and enforces both population and lifetime sparsity of the extracted features, simultaneously. The proposed algorithm clearly outperforms standard principal component analysis (PCA) and its kernel counterpart (kPCA), as well as current state-of-the-art algorithms of aerial classification, while being extremely computationally efficient at learning representations of data. In [82] the authors use an off-the-shelf CNN representation named OverFeat, with simple classifiers to address different recognition tasks. Nevertheless, it showed itself to be a strong competitor to the more sophisticated and highly tuned state-of-the-art methods. The same trend was observed for various recognition tasks and different datasets, which highlights the effectiveness and generality of the learned representations.

Another research presented by Aguilera et al. [3], propose a novel approach for learning cross-spectral similarity measures, inspired on the network structure for stereo matching presented in [121], where they avoid defining a hand-made descriptor being the CNN responsible for jointly learning the representation and the measurement. This approach has been referred to as a 2 channel network (2ChNet). Patch matching has also been addressed in [37]; in this case, the authors propose a generalization of the siamese networks in order to speed up the matching process. The architecture of the network consists of two parts, firstly a network is used for describing the patches, then another network is proposed for the matching (metric network). Following the siamese architecture, [96] proposes to train a Siamese network that compares the similarity between image patches just using the L2 distance. This simple match-

ing speeds up the whole process since it is possible to use fast approximate nearest neighbor algorithms to find the correspondences and thus improve the overall matching runtime. A comparative study between 2 channel and siamese architecture has been performed in [3]. It is shown that the 2ChNet has a considerably better performance in the cross-spectral domain, which outperforms those approaches based on hand-made descriptors (e.g., [4], [6], [71]). Also, in [101] the authors present an approach to learn data representations using a novel autoencoder-based fabric defect detection method. However, the texture (non-defect) area cannot be well reconstructed, which makes the pixel-wise detection inaccurate. For this reason, they explore similarities between different patches in the whole test image. In order to maintain the texture area in the reconstructed patch, the original encoded latent variable is modified and the cross-patch similarity is introduced for determining the modification function.

In [1] the authors present a technique to perform registration of images of different nature using SAR and optical images, based on a neural network in order to build feature point descriptors; then, they use the RANSAC algorithm to align found matches. Another approach has been proposed in [76]. A deep local descriptor learning framework for cross-modality face recognition, to learn discriminant and compact local information directly from raw facial patches, is presented. It also includes a novel cross-modality enumeration loss to eliminate the modality gap on the local patch level.

Song et al. in [98] present an adversarial discriminative feature learning framework to close the sensing gap via adversarial learning on both raw-pixel space and compact feature space. This approach integrates cross-spectral face hallucination into an end-to-end adversarial network, in the feature space. Additionally, an adversarial loss and a high-order variance discrepancy loss are employed to measure the global and local discrepancy between two heterogeneous distributions respectively to enhance domain-invariant feature learning and modality independent noise removing.

In [26], the authors propose a new approach to align two images related by an unknown 2D homography, where the local descriptor is learned from scratch from the images and the homography is simultaneously estimated. This technique uses a siamese convolutional neural network to optimize by a single loss function. This method has been designed to align images of different modalities such as RGB and near infrared without using any prior labeled data.

In [46], the authors present a deep coupled learning approach to solve the problem of matching polarimetric thermal face photos against a gallery of visible spectrum faces. With the polarization state information of thermal faces it is possible to obtain the missing textural and geometric details in the thermal face imagery, which exist in the visible spectrum. A coupled deep neural network model has been designed, which leverages relatively large visible and thermal datasets to overcome the problem of overfitting. It also finds global discriminative features in a nonlinear embedding space to relate the polarimetric thermal faces to their corresponding visible faces.

2.2 Image Colorization

Another problem that has been addressed in this thesis is related with infrared image colorization. As mentioned before somehow it shares some common problems with monochromatic image colorization that has been largely studied during the last decades. Colorization techniques mostly differ in the ways they obtain and treat the data for modeling the correspondences between grayscale and color. Coarsely speaking colorization techniques can be classified into parametric and non-parametric approaches. Non-parametric methods, given an input grayscale image, firstly they define one or more color reference images (provided by a user or automatically retrieved) to be used as source data. Then, following the image analogy framework, color is transferred onto the input image from analogous regions of the reference image(s). Parametric methods, on the other hand, learn prediction functions from large datasets of color images at training time, posing the problem as either regression onto continuous color space or classification of quantized color values.

In many vision applications, including surveillance and driving assistance, RGB video sensors are preferred since depicted images are similar to the human visual perception system. Visible spectrum images are referred through this work indistinctly as a visible spectrum or RGB images have limitations related to lighting conditions and object surface color. The limitations mentioned above can be easily overcome using Near Infrared (NIR) imagery.

The NIR band of the electromagnetic spectrum is just outside the range of what humans can see and can sometimes offer clearer details than what is achievable with visible light imaging. The NIR spectrum is independent of the brightness and color of the targets, which has potential benefits, including non-visible illumination requirements. Different solutions could take advantage of this contribution. For instance, in [41] the authors propose address the task of restoring RGB images taken under low illumination, when an aligned near infrared image is available under low lighting conditions, the NIR band is less noisy than the visible and restoring the R,G, and B bands is possible based on the NIR band.

Although the problem of NIR image colorization shares some particularities with color correction/transfer (e.g., [24], [35], [72], [73]) there are some important differences. First, in the image colorization domain (grayscale image to RGB) the chrominance is the only feature that needs to be calculated, because the luminance is given by grayscale input. Secondly, in the case of color correction/transfer techniques, in general, three channels are given as input to obtain the new representation in the new three-dimensional space.

Surface reflection in the NIR spectral band is material dependent. This means that the difference in the NIR intensities is not only due to the particular color of the material but also to the absorption and reflectance of dyes. In spite of the advantages of NIR imagery, when the information needs to be shown to the people of the interest group, the visible representation (e.g., RGB) is always preferred since it allows a better appreciation and understanding of the scene and therefore it has been possible to have a better decision making. Welsh et al. [113] describe a semi-automatic technique for colorizing a grayscale image by transferring color from a reference color image. They examine the luminance values in the neighborhood of each

pixel in the target image and transfer the color from pixels with matching neighborhoods in the reference image. This technique works well on images where different colored regions give rise to distinct luminance clusters or possess distinct textures. Colorization algorithms mostly differ in the ways they obtain and treat the data for modeling the correspondences between grayscale and color. There have been several approaches, as designed by, Celebi et al. [18] to introduce a spatial and variational based frequency method, which obtain perceptually inspired color and contrast enhancement of digital images.

Also, Gavet et al. [32] present the color logarithmic image processing (CoLIP) and antagonist space, which is a framework that defines a vectorial space for color images. It illustrates the representation of the chromaticity diagram with color modification application, namely white balance correction and color transfer. Another technique is the grayscale image matting and colorization; Chen et al. [19] present a variation of a matting algorithm with the introduction of alpha's distribution and gradient into the Bayesian framework and an efficient optimization scheme. It can effectively handle objects with intricate and vision sensitive boundaries, such as hair strands or facial organs, plus they combine this algorithm with the color transferring techniques to obtain its colorization scheme. In other cases, the user must direct the search for matching pixels by specifying matches indicating corresponding regions in the two images. It is also difficult to fine-tune the outcome selectively in problematic areas. There are other approaches like colorization by example; in [47] an algorithm that colorizes one or more input grayscale images is presented. It is based on a partially segmented reference color image. By partial segmentation, they assume that one or more mutually disjoint regions in the image have been established, and each region has been assigned to a unique label.

The approaches presented above have been implemented using classical image processing techniques. However, recently convolutional neural network based approaches are becoming the dominant paradigm in almost every computer vision task. CNN's have shown outstanding results in various and diverse computer vision tasks such as stereo vision [122], image classification [100] or even difficult problems related with cross-spectral domains [3] outperforming conventional hand-made approaches. Hence, there are some recent image colorization approaches based on deep learning, that exploit to the maximum the capacities of this type of convolutional neural networks. As an example, it can be mentioned the approach presented on [124]. It proposes a fully automatic approach that produces brilliant and sharp image color. They model the unknown uncertainty of the desaturated colorization levels designing it as a classification task and use class-rebalancing at training time to augment the diversity of colors in the result.

On the contrary, [45] presents a technique that combines both global priors and local image features. Based on a CNN a fusion layer merges local information, dependent on small image patches, with global priors, computed using the entire image. The model is trained in an end-to-end fashion, so this architecture can process images of any resolution. They leverage an existing large-scale scene classification database to train the model, exploiting the class labels of the dataset to more efficiently and discriminatively learn the global priors. In [61], recent research on colorization, addressing images from the infrared spectrum, has

been presented. It uses convolutional neural networks to perform an automatic integrated colorization from a single channel NIR image to a RGB image. The approach is based on a deep multi-scale convolutional neural network to perform a direct estimation of the low RGB frequency values. Additionally, it requires a final step that filters the raw output of the CNN and transfers the details of the input image to the final output image.

2.3 Vegetation Index Estimation

Another research problem that has been tackled in this doctoral thesis corresponds to the estimation of vegetation indexes from images of just one spectral band. In this section, the works related to the estimation and use of vegetation indices are reviewed.

The agricultural industry has adopted the use of new technologies based on computer vision approaches, which has been widely applied to various tasks of agricultural development to automate all activities related to the improvement of productivity of the fields, the decrease in rising labor costs and the training given to farmers to face increasingly aggressive globalized competition. One of the approaches based on computer vision is the estimation of vegetation indexes. Such indexes enable a better identification, analysis, and management of this temporal and spatial in-field variability. Also, with the rapid and constant innovation of camera sensors and Graphics Processing Unit (GPU) [102], enormous progress has been achieved on the agricultural applications.

Nowadays, the owners of the plantations, after the automation of their agricultural processes, can rely on the early identification of diseases or the infection of insects that affect the health of the plants. These techniques, together with the use of unmanned aerial vehicles (UAVs), are used to accurately apply fungicides that use GPS coordinates on the specific areas that are present in the plantation. The use of sensors sensitive to the near infrared spectrum allows obtaining information about of plant health. On the other hand, with the visible spectrum, the exterior of the plant can be observed, at a photometric level (color of the plant), due to the transformation process produced by photosynthesis.

Additionally, with NIR sensors, all the crop captured information could be used to obtain statistical information of every year trying to predict the health of future plantations for better crop productivity. Many computer vision techniques have evolved to offer solutions for this kind of agricultural problems. These techniques came from mathematical and statistical methods to deep learning neural networks.

As mentioned above, deep learning models have obtained state-of-the-art results on most of the computer vision complex problems [53]. Nevertheless, there are many challenging problems in the agriculture still pending to be solved and deep learning-based solutions seem to be the most appropriate to be used, obviating the need for a pipeline of specialized and hand-crafted methods currently used. Some researchers have proposed deep learning-based approaches for remote sensing and agricultural applications. In [68] the authors propose to use SAR images to estimate missing spectral features through data fusion and deep learning,

exploiting both temporal and cross-sensor dependencies on Sentinel-1 and Sentinel-2 time-series, in order to obtain the Normalized Difference Vegetation Index (NDVI).

Huang et al. in [43] have proposed a novel method for effective and efficient topographic shadow detection for the images obtained from Sentinel-2A multispectral imager (MSI) by combining both the spectral and spatial information. This method uses a convolutional neural network, directly operating on indices input due to its remarkable classification performance, exploiting the spatial contextual information and spectral features for effective topographic extraction. Another approach, presented in [83], has proposed a method to predict the vegetation dynamics using Moderate Resolution Imaging Spectroradiometer (MODIS) NDVI time-series datasets and long short term memory network, an advanced technique adapted from the artificial neural network. In [62], the authors have proposed a decision-level fusion approach with a simpler architecture for the task of dense semantic labeling. This method first obtains two initial probabilistic labelings resulting from a fully-convolutional neural network and a simple classifier, e.g., logistic regression exploiting spectral channels and LIDAR data, respectively. Then a Conditional Random Field (CRF) estimates the final dense semantic labeling results. Some of these techniques are presented below.

Recently, in [7] an automated approach is proposed to detect and count individual palm trees from UAV using a combination of spectral and spatial analyses. The proposed approach comprises a step that discriminates the vegetation from the surrounding objects by applying the NDVI and another step used to detect individual palm trees using a combination of Circular Hough Transform (CHT) and the morphological operators.

In [12] the authors present a methodology to predict the NDVI by training a crop growth model with historical data. Although they use a very simple soybean growth model, the methodology could be extended to other crops and more complex models. According to [103] long-term observations of vegetation phenology can be used to monitor the response of terrestrial ecosystems to climate change. They propose a method for observing phenological events by analyzing time series of vegetation indices such as the normalized vegetation difference index to investigate the potential of a Photochemical Reflection Index (PRI) to improve the accuracy of MODIS-based phenological estimates in an evergreen coniferous forest. The results suggest that PRI can serve as an effective indicator of spring seasonal transitions, and confirm the usefulness of MODIS PRI for detecting phenology. Also, [59] presents a study to evaluate the economic benefits of greening programs (e.g., planting urban trees, adding or enhancing parks, providing incentives for green roofs) using low-cost normalized difference vegetation index data from satellite imagery, using the spatial lag-Tobit models that predict tree canopy cover from NDVI.

In another research [81], the authors focus on temporal NDVI and surface temperature, to analyse the changes of the environmental conditions related to the land surface temperature on urban areas. The research demonstrates the correlation between temporal NDVI and surface temperature exemplified with a case study conducted over two different regions, geographically as well as economically. In [116] the authors present a method to reconstruct

normalized difference vegetation index time-series datasets for monitoring long-term changes in terrestrial vegetation. This Temporal–Spatial Iteration (TSI) method was developed to estimate the NDVIs of contaminated pixels, based on reliable data. The TSI method is the most applicable when large numbers of contaminated pixels exist.

Additionally, [126] presents a local modeling technique to estimate regression models with spatially varying relationships, using Geographically Weighted Regression (GWR), to investigate the spatially nonstationary relationships between NDVI and climatic factors at multiple scales in northern China. The results indicate that all GWR models with appropriate bandwidth represent significant improvements of model performance over the Ordinary Least Squares (OLS) models. The results reveal that the ecogeographical transition zone and the GWR model can improve the model ability to address spatial, nonstationary, and scale-dependent problems in landscape ecology. Also in [29] the authors present a high-throughput phenotyping platform to dynamically monitor NDVI during the growing season for the contrasting wheat crops. The high-throughput phenotyping platform capture the variation of NDVI among crops and treatments (i.e., irrigation, nitrogen, and sowing). The high-throughput phenotyping platform can be used in agronomy, physiology, and breeding to explore the complex interaction of genotype, environment and management.

Usually, CNN approaches depend on the existence of accurately registered images (i.e., since images from different spectra are considered, they may look different, so the problem is how to find the same set of points in both spectra [86] to be used as references). However, deep learning-based approaches have been proposed to overcome this problem and to obtain correspondences in cross-spectral domains (e.g., [5]). Once corresponding points are obtained, the image registration can proceed in a single reference system and feed these pre-processed images to the learning CNN models.

Recently, some approaches based on deep learning proposing solutions to agriculture problems related to health and productivity of the plantation has been implemented, especially using generative networks, like Cyclic GAN [10]. In [79], the authors propose a novel deep learning-based generative adversarial model, RefineGAN, for fast and accurate compressed sensing for magnetic resonance imaging (CS-MRI) reconstruction. The proposed model is a variant of a fully residual convolutional autoencoder and Generative Adversarial Networks (GANs), specifically oriented for CS-MRI formulation; the architecture has been designed with a generator and a discriminator networks with a cyclic data consistency loss for a correct interpolation in the given under-sampled multiple space data. Bansal et al. [9] introduce a data-driven approach for unsupervised video retargeting that translates content from one domain to another while preserving the style native to a domain. This approach combines both spatial and temporal information along with adversarial losses for content translation and style preservation. It includes a study about the advantages of using spatiotemporal constraints over spatial constraints for effective retargeting. In the next sections, an introduction to GAN and Cyclic GAN architectures is provided since they are used in this thesis.

2.4 Image Haze Removal

The image haze removal problem has been studied for more than two decades. Some of the solutions proposed in the literature have been based on image attributes, transmission map, air light conditions, atmospheric scattering model, among others. (e.g., [87]). Most of these solutions start from the usage of images from other spectral bands to extract certain characteristics that are used to remove the haze. In [13], a non-local haze-lines for removing haze from image is proposed; this method is based on the observation that the number of distinct colors in an image is orders of magnitude smaller than the number of pixels, based on the assumption that an image can be faithfully represented with just a few hundreds of distinct colors.

Another model-based approach has been presented by [108]; this work proposes a selection of an atmospheric light value that is directly responsible for the color authenticity and contrast of the resulting image. Additionally, they propose a fast transmission estimation algorithm to be more efficient and reduce the process time. Also using a haze model, [31] presents a haze removal technique that uses a fusion-based variational remove haze method, which combines the minimized outputs of two energy functionals to produce a haze-free version. Ju et al. [51] present an improvement by addressing the weaknesses inherent to the atmospheric scattering models; the authors develop a way to remove the haze using an adaptive method for adjusting scene transmission based on the image features. The input image is partitioned into several scenes based on the haze thickness. Then, they obtain the rough scene transmission map by maximizing the contrast in each scene and then remove the haze by using the proposed adaptive method. Similarly to the previous work, Fattal et al. [30] propose to estimate the optical transmission in hazy scenes, given a single input image; the scattered light is eliminated to increase scene visibility and recover haze-free scene contrasts.

Recently, in [109], a fast algorithm for single remove haze is proposed. It is based on linear transformation, by assuming that a linear relationship exists in the minimum channel between the hazy image and the haze-free image. In [118], the authors proposed an enhanced detail and dehaze technique for haze removal based on modified channel prior scheme and combine the dehazed image with a non-sky detail layer using a method to improve the image details. After that, the recovered image contrast has been enhanced based on a histogram equalization approach. Also using a haze model, [60] proposes an improved contrast enhanced restoration. This technique is based on a quadtree subdivision searching method, where the sky area of a multi-channel polarization image is automatically extracted, and the atmospheric light and degree of polarization are calculated; then, the scene depth information of an image is calculated based on contrast enhancement method. Finally, the atmospheric intensity is thinly restored by the guided filter, and the degraded image is restored. In [31] a haze removal technique that uses a fusion-based variational method is presented, which combines the minimized outputs of two energy functionals to produce a haze-free version. The authors in [57] present a detailed survey and experimental analysis based on Dark Channel Prior (DCP) methods that explain the effectiveness of the individual step of the dehazing process and facilitate the development of advanced dehazing algorithms.

In [66] a model-based approach has been presented, which proposes an algorithm based on image filtering DCP estimations of atmospheric light to obtain an unhazed image and finally improving the local contrast.

In [117], the authors propose a method that consists of a combined algorithm based on both dark channel prior and histogram optimization, which can make the image contrast stretching, so the impact of the haze on the image can be weakened. If the obtained dehazed image cannot meet the minimum quality required, the dark channel prior can be used to estimate the haze intensity. Also, in [49] the authors propose a method for combining DCP and Bright Channel Prior (BCP) for single image dehazing. The proposed technique achieves airlight approximations by implementing numerical proximity of atmospheric light, which use the average value of the DCP and BCP.

Lately, novel image haze removal approaches based on deep learning techniques have been proposed obtaining acceptable results. In [58] a model-based on a reformulated atmospheric scattering model is proposed, instead of estimating the transmission matrix and the atmospheric light separately. Ren et al. [85] present a multi-scale deep neural network for single-image remove haze by learning the mapping between hazy images and their corresponding transmission maps. The proposed algorithm consists of a coarse-scale net that predicts a holistic transmission map based on the entire image and a fine-scale net that refines results locally. Cai et al. [16] propose a trainable end-to-end system called DehazeNet, for medium transmission estimation. DehazeNet takes a hazy image as an input and outputs its medium transmission map that is subsequently used to recover a haze-free image via an atmospheric scattering model.

More recently the generative adversarial network framework has been used obtaining appealing results. In [123] the authors propose a unified single remove haze GAN network that jointly estimates the transmission map and performs the haze process; the network is trained using synthetic images and a two-terms loss function. The first term of the loss function is a pixel-wise Euclidean distance, while the second term considers perceptual information. In the current chapter, a loss function based on multiple terms is proposed. Additionally, in the GAN architecture, a stacking strategy is proposed to speed up the learning process.

Two basic things can be done with generative based deep learning models. One is to take a collection of points and infer a function that describes the distribution that generated them; the second one is to build a generative model, which is to take a machine that observes many samples from a distribution and can create more samples from the same distribution. They allow a network to learn to generate data with the same internal structure as other data.

According to [91], a feature matching is proposed to address the instability of a GAN network establishing a new objective for the generator that prevents it from over-training maximizing the output of the discriminator, requiring to generate data that matches the statistics of the real data. More details on GAN network architectures are provided in the next section.

2.5 Generative Adversarial Networks

Generative Adversarial Networks (GANs) are a class of neural networks that have gained popularity in recent years. They allow a network to learn to generate data with the same internal structure as some target data. GANs are powerful and flexible tools, and one of its more common applications is image generation. It is a framework presented on [34] for estimating generative models via an adversarial process, in which simultaneously two models are trained: a generative model G that captures the data distribution and a discriminative model D that estimates the probability that a sample came from the training data rather than from G. The training procedure for G is to maximize the probability of D making a mistake. This framework corresponds to a minimax two-player game. In the space of arbitrary functions G and D, a unique solution exists [34], with G recovering the training data distribution and D equal to $1/2$ everywhere.

The generator G implicitly defines a probability distribution p_g as the distribution of the samples $G(z)$ obtained when $z \sim p_z$. The Algorithm (see **Algorithm 1**) could converge to a good estimator of p_{data}, if given enough capacity and training time.

Algorithm 2.1 Minibatch stochastic gradient descend training of generative adversarial nets. The number of steps to apply to the discriminator, k, is an hyperparameter.

for `Number of training iterations` **do**
  for k `steps` **do**
    - `Sample minibatch of` m `noise samples` $z^{(1)},...,z^{(m)}$ `from noise prior` $p_g(z)$
    - `Sample minibatch of` m `samples` $x^{(1)},...,x^{(m)}$ `from generated data distribution` $p_{data}(x)$
    - `Update the discriminator by ascending its stochastic gradient:`

$$\nabla_{\theta_d} \frac{1}{m} \sum_{i=1}^{m} [\log D(x^{(i)}) + \log(1 - D(G(z^{(i)})))].$$

  end for
  - `Sample minibatch of` m `noise samples` $z^{(1)},...,z^{(m)}$ `from noise prior` $p_g(z)$
  - `Update the generator by descending its stochastic gradient:`

$$\nabla_{\theta_g} \frac{1}{m} \sum_{i=1}^{m} \log(1 - D(G(z^{(i)}))).$$

end for

According to [70], GANs have the advantages that Markov chains are never needed, only backpropagation is used to obtain gradients, no inference is required during learning, and a wide variety of factors and interactions can easily be incorporated into the model. To learn the generator's distribution p_g over data $\boldsymbol{x}$, the generator builds a mapping function from a prior noise distribution $p_z(z)$ to a data space $G(z;\theta_g)$; and the discriminator, $D(x;\theta_d)$, outputs a single scalar representing the probability that x came from training data rather than p_g. G and

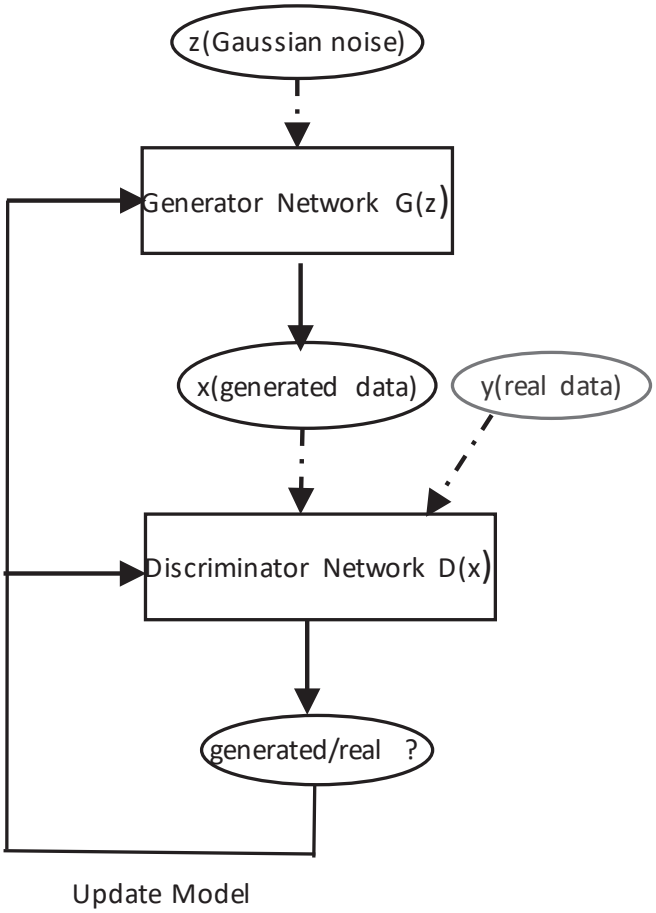


Figure 2.1 – Illustration of a generative adversarial network.

D are both trained simultaneously. The parameters model for G: θ_g (are adjusted to minimize $log(1 - D(G(z)))$ and for D: θ_d to minimize $logD(x)$ with a value function $V(G,D)$:

$$\min_G \max_D V(D,G) = \mathbb{E}_{x \sim p\ \text{data}(x)}[logD(x)] + \mathbb{E}_{z \sim p\ \text{data}(z)}[log(1 - D(G(z)))]. \tag{2.1}$$

Generative Adversarial Networks are quite useful in several computer vision problems. Figure 2.1 shows an illustration of this architecture.

According to [70], in an unconditioned generative model, there is no control on modes of the data being generated. However, by conditioning the model on additional information it is possible to direct the data generation process. Generative adversarial networks can be

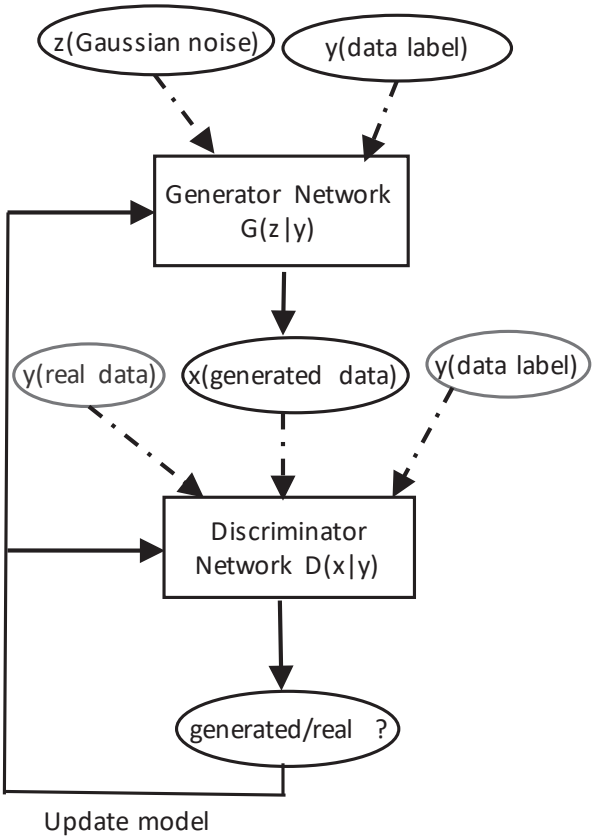


Figure 2.2 – Illustration of a conditional generative adversarial network.

extended to a conditional model if both the generator and discriminator are conditioned on some extra information y, see Fig. 2.2. This information could be any kind of auxiliary information, such as class labels or data from other modalities. The conditioning can be performed by feeding y into both discriminator and generator as an additional input layer. Now, with the addition of the conditional information, the objective function of a two-player minimax game becomes:

$$\min_G \max_D V(D,G) = \mathbb{E}_{x \sim p\ \mathrm{data}(x)}[logD(x|y)] + \mathbb{E}_{z \sim p\ \mathrm{z}(z)}[log(1-D(G(z|y)))]. \tag{2.2}$$

The discriminator performs a binary classification including the extra information fed to the network. As a result, the discriminator and generator reach more accurate gradients. Conditional GANs enhance the stability of the model, but it affects the learning of the semantic characteristics of the image samples, meaning that both the generator and discriminator are conditioned on some sort of auxiliary information such as class labels or data from other modalities.

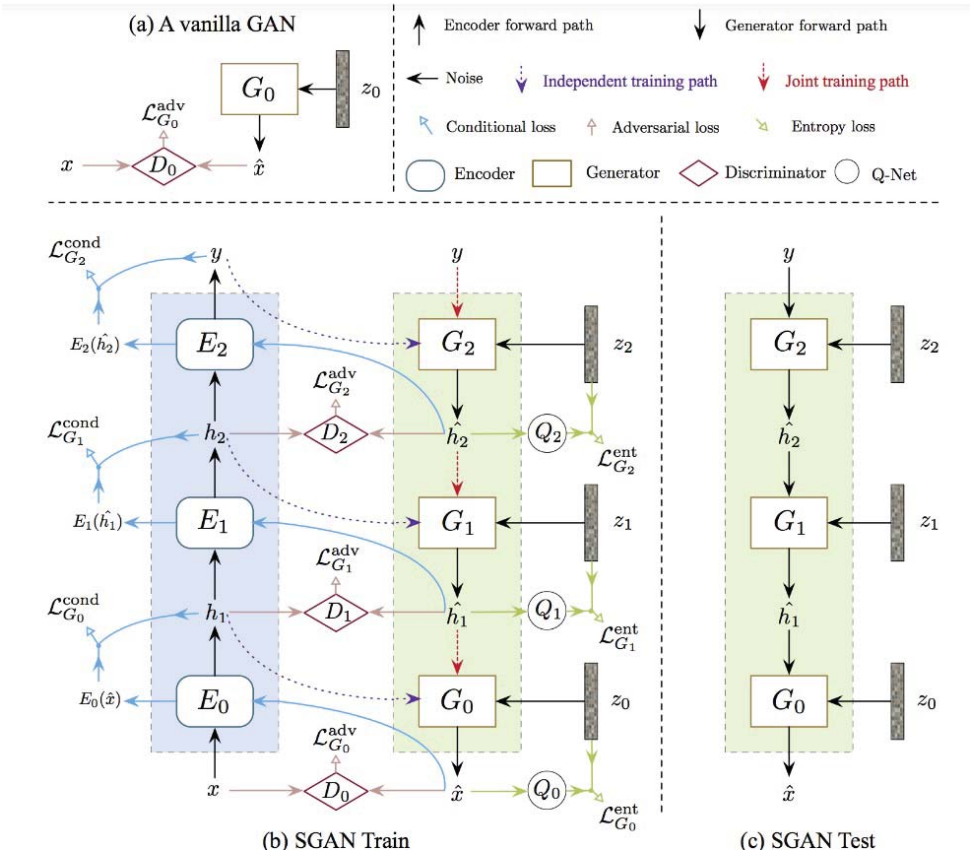


Figure 2.3 – Illustration of a stacked generative adversarial network, proposed by [44].

In [91] some techniques are presented to improve the efficiency of the generative adversarial networks. One of them, called the virtual batch normalization, allows to significantly improve the network optimization using the statistics of each set of training batches. The disadvantage is that this process is computationally expensive, because it requires running forward propagation on two minibatches of data, so it is only used in the generator network.

Continuing with the investigation of different adversarial generative models, the stacked GAN architecture is presented in [44]. It consists of a top-down stack of GANs, each designed to generate lower-level representations conditioned on higher level representations. A representation discriminator is introduced at each feature hierarchy to encourage the representation manifold of the generator to align with that of the bottom up discriminative network, leveraging the powerful discriminative representations to guide the generative model. Besides, a conditional loss is introduced to encourage the use of conditional information from the layer above, and a novel entropy loss that maximizes a variational lower bound on the conditional entropy of generator outputs, see Fig. 2.3. First, each stack is trained independently and then the training to the whole model is continued comprehensively (end to end). Unlike the original

GAN that uses only a randomized initialized noise vector to represent all possible variations, the Stacked GAN architecture decomposes variations at multiple levels and gradually solve uncertainties in the generative process from top to bottom. Based on the visual inspection, using the metric Inception Scores (IS) and the Turing visual test, it is shown that the Stacked GAN is capable of generating much higher quality images than traditional unstacked GANs and this strategy allows accelerating the learning process to generate the required output.

In this section, only the GAN architectures used on this thesis have been reviewed. According to the GAN's Zoo, every week, new papers on Generative Adversarial Networks (GAN) are coming out and it is hard to keep track of them all [1], not to mention the incredibly creative ways in which researchers are naming these GANs!, see Fig. 2.4.

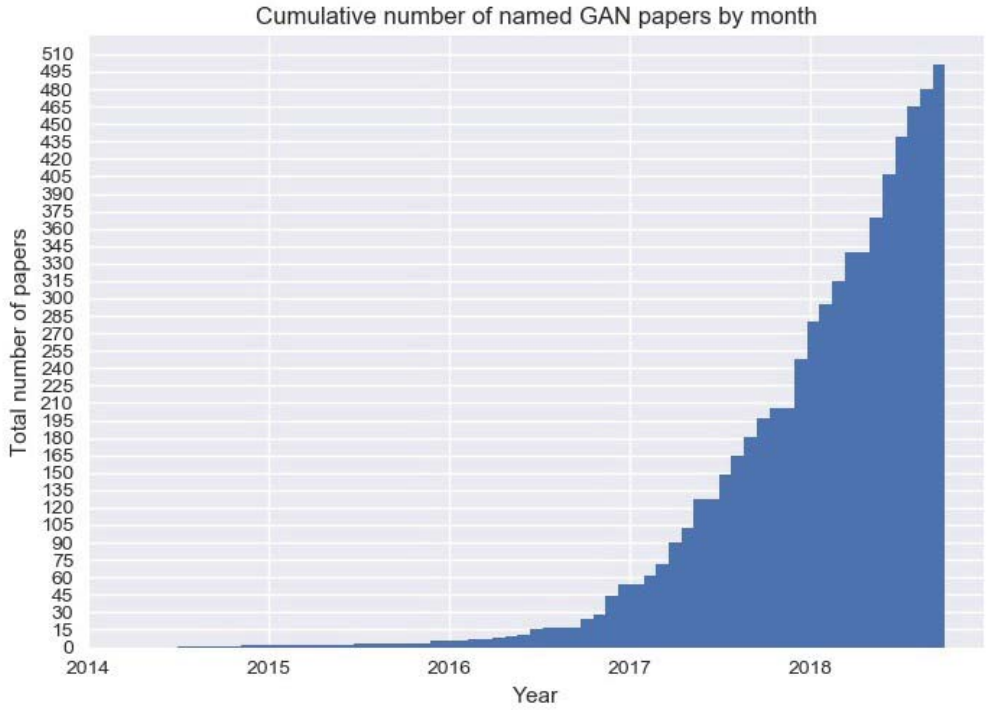


Figure 2.4 – Illustration of "All the named GAN variants cumulatively by month since 2014"; Credit: Bruno Gavranović.

2.6 Cyclic Generative Adversarial Networks

Image-to-image translation is the process of transforming an image from one domain to another, where the goal is to learn the mapping between an input image and an output image

[1] Deep Hunt Newsletter. Last updated on Feb 23, 2018.

using a training set of aligned image pairs. However, for many tasks, paired training data are hard to obtain, and to prepare them it often takes a lot of work from specialized personnel to obtain thousands of paired image datasets, especially with complex image translations. Cyclic Generative Adversarial Network (CyclicGAN) is an architecture to address this problem, because it learns to perform image translations without explicit pairs of images. No one-to-one image pairs are required. CyclicGAN learns to perform style transfer from the two sets despite every image having vastly different compositions.

The Fig. 2.5 and 2.6 depict the CycleGAN model proposed in [128]. As can be appreciated, the CyclicGAN architecture generates synthetic images through two generators (*G and F*) and two discriminators Dx, Dy. In order to generate a synthetic image, the architecture takes advantage of the joint of cycle-consistency in addition to the usual discriminator and generator losses.

According with [128] the objective of a CyclicGAN is to learn mapping functions between two domains, X and Y, given training samples ${x_i}_{i=1}^{N} \in X$ and ${y_j}_{j=1}^{M} \in Y$.

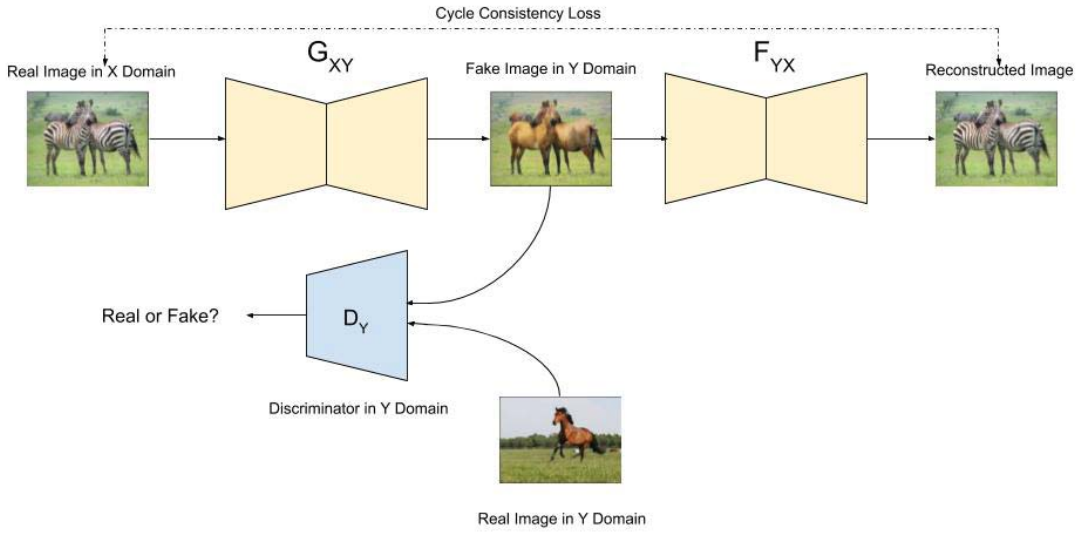


Figure 2.5 – Cycle generative adversarial model G: $X \rightarrow Y$ and its discriminator D_y.

The model includes two mappings functions $G : X \rightarrow Y$ and $F : Y \rightarrow X$. Besides, it introduces two adversarial discriminators D_x and D_y , where D_x aims to distinguish between images x and translated images $F(y)$; in the same way, D_y aims to discriminate between y and $G(x)$. Besides, the proposed approach includes two types of loss terms: adversarial losses [34] for matching the distribution of generated synthetic images to the data distribution in the target domain real images; and a cycle consistency loss to prevent the learned mappings G and F from contradicting each other. The adversarial losses defined according to [34] to both mapping functions. For the mapping function $G : X \rightarrow Y$ its discriminator D_y, is defined as:

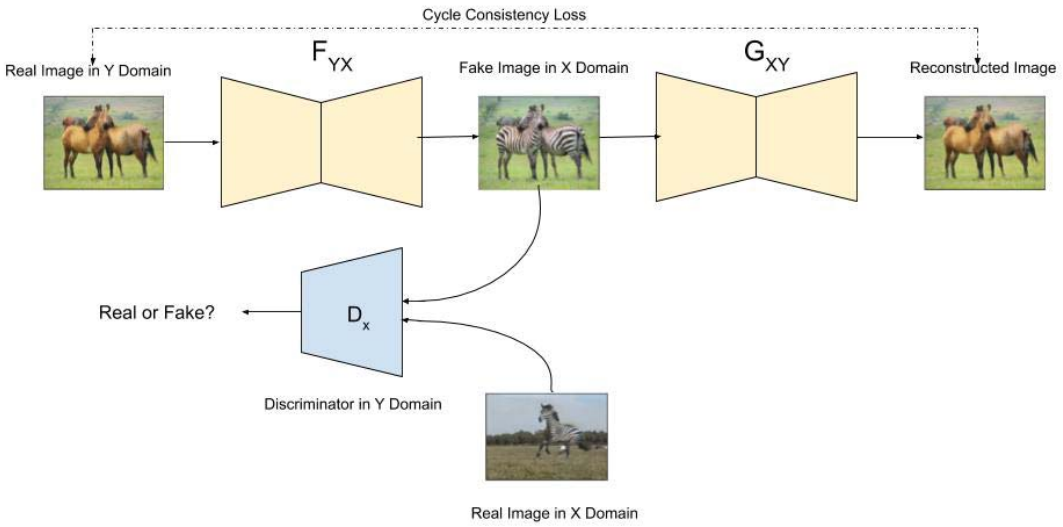


Figure 2.6 – Cycle generative adversarial model F: $Y \rightarrow X$ and its discriminator D_x.

$$\mathscr{L}_{GAN}(G, D_y, X, Y) = \mathbb{E}_{y \sim p\ \mathrm{data}_{(y)}}[log D_Y(y)] + \mathbb{E}_{x \sim p\ \mathrm{data}_{(x)}}[log(1 - D_Y(G(x)))], \tag{2.3}$$

where G tries to generate images $G(x)$ that look similar to images from domain Y, while D_y aims to distinguish between translated samples $G(x)$ and real samples y.

For the mapping function $F: Y \rightarrow X$ its discriminator D_x, is defined as:

$$\mathscr{L}_{GAN}(F, D_x, Y, X) = \mathbb{E}_{x \sim p\ \mathrm{data}_{(x)}}[log D_X(x)] + \mathbb{E}_{y \sim p\ \mathrm{data}_{(y)}}[log(1 - D_X(F(y)))], \tag{2.4}$$

where F tries to generate images $F(y)$ that look similar to images from domain X, while D_x aims to distinguish between translated samples $F(y)$ and real samples x.

According to [128], to reduce the space of possible mapping functions, the learned mapping functions should be cycle-consistent. For each image x from domain X, the image translation cycle should be able to bring x back to the original image, i.e., $x \rightarrow G(x) \rightarrow F(G(x)) \approx x$, calling this forward cycle consistency. Therefore, for each image y from domain Y, G and F should also satisfy backward cycle consistency: $y \rightarrow F(y) \rightarrow G(F(y)) \approx y$. This cycle consistency loss is defined as:

$$\mathscr{L}_{cycle}(G, F) = \mathbb{E}_{x \sim p\ \mathrm{data}_{(x)}}[\|F(G(x)) - x\|_1] + \mathbb{E}_{y \sim p\ \mathrm{data}_{(y)}}[\|G(F(y)) - y\|_1]. \tag{2.5}$$

The final objective is defined as:

$$\mathscr{L}(G,F,D_X,D_y) = \mathscr{L}_{GAN}(G,D_y,X,Y) + \mathscr{L}_{GAN}(F,D_x,Y,X) + \lambda\mathscr{L}_{Cycle}(G,F) \tag{2.6}$$

where λ controls the relative importance of the two first terms of the equation, with the aim to solve:

$$G^*,F^* = \arg\min_{G,F} \arg\max_{D_x,D_y} \mathscr{L}(G,F,D_x,D_y). \tag{2.7}$$

2.7 Instance Normalization

Much of the recent work on GANs is focused on developing techniques to stabilize training. Thus, GANs are known to be unstable (during training) and very sensitive to changes in the hyper-parameter values of the learning model. Another field of analysis has emerged around the style of an image evaluated by the statistics of convolutional neural network filters, a renewed interest in the texture generation and image stylization problems in order to obtain qualitative improvement in the generated image, discarding the instance-specific contrast information from an image during style transfer. This can be evaluated by the statistics of convolutional neural network filters.

Ulyanov et al. [104] shows that it is possible to train a generator network $g(x,z)$ that can apply to a given input image x the style of another $x0$. They introduce a method named *instance normalization* for a better stylization and texture synthesis, that derive entropy loss which improves samples diversity. This method prevents instance-specific mean and covariance shift simplifying the learning process. The instance normalization layer is applied at test time as well as at training time. According to [104] the generator network should discard contrast information in the content image to learn a highly nonlinear contrast normalization function as a combination of such layers. Let $x \in \mathbb{R}^{NxCxWxH}$ an input tensor containing a batch of N images, where C, W and H are the depth, width and high respectively of the image tensor and let x_{tijk} denote its $tijk$-th element of x image tensor, where k and j span spatial dimensions, t is the index of the image in the batch, i is the feature channel (in the case of an RGB image being used as an input, it would represent a color channel). Thus, a simple version of instance normalization is defined as:

$$y_{tijk} = \frac{x_{tijk}}{\sum_{l=1}^{W}\sum_{m=1}^{H} x_{tilm}}. \tag{2.8}$$

A small change in the stylization architecture proposed by [104] is a qualitative improvement in the generated images. The change is limited to swapping batch normalization with instance normalization, and to apply the latter both at training and testing times. The resulting method can be used to train high-performance architectures for real-time image generation.

2.8 Metalearning

The field of meta-learning has as one of its primary goals the understanding of the interaction between the mechanism of learning and the concrete contexts in which that mechanism is applicable. The field has seen a continuous growth in the past years with interesting new developments in the construction of practical model-selection assistants, task-adaptive learners, and a solid conceptual framework [106].

Meta-learning, has been proposed to simulate the human learning process. Normally, humans apply the method known as learning to learn with few samples, therefore, knowledge already acquired is always used. Meta-learning algorithms tries to implement architectures that can learn new knowledge or reinforce something already known to apply it to another situation efficiently. Without the need to rely on large data sets, such as artificial intelligence algorithms based on deep learning. There are three meta-learning models, the first one based on distance metrics, the second one known as model-based, with the use of memory and the last one named as optimization-based.

In this thesis, it has been implemented Few-shot classification which is a type of meta-learning model in the field of supervised learning. Where a dataset D is often split into two parts, a support set S for learning and a prediction set B for testing, $D = (S, B)$. These kind of models

$$\begin{aligned} \theta^* &= \arg\max_{\theta} \mathbb{E}_{(\mathbf{x},y)\in\mathscr{D}} \left[P_\theta(y|\mathbf{x}) \right] \\ \theta^* &= \arg\max_{\theta} \mathbb{E}_{B\subset\mathscr{D}} \left[\sum_{(\mathbf{x},y)\in B} P_\theta(y|\mathbf{x}) \right] \quad ; \text{ trained with mini-batches.} \end{aligned} \tag{2.9}$$

Chapter 3

Cross Spectral Image Similarity

This chapter tackles the problem of learning similarities between cross-spectral image patches using a 2 channel network (2ChNet) model and a metric based meta-learning approach.

The core idea is to propose novel approaches to learn the best representation of the image patches to determine the similarity degree between cross-spectral regions (patches) using a framework able to compare image regions and determining a similarity measure to decide if there is a similarity between the compared patches. The models have been trained end-to-end from scratch. Experimental results have shown that the proposed approaches effectively estimate the similarity of the patches and, comparing them with the state of the art approaches, better results are obtained. It should be noted that the approach based on meta-learning is the one that has shown the best results.

3.1 Introduction

One of the challenges that remain in the field of computer vision is to achieve the same effectiveness that the human eye has, to determine whether two images are similar or not. Many of the traditional techniques are based on encoding images into representation vectors, for which it is necessary to take small regions of the images to be compared and distance metrics, such as the Euclidean, are used to measure the similarity between the regions of the images.

Many factors could affect the image comparison process, such as occlusion, illumination, quality of sensors, etc. For this reason, different approaches have been developed, from hand-craft methods to deep learning-based, in order to obtain this kind of information valid for high-level vision problems.

One of the computer vision techniques that has always been in constant research is the determination of the similarity between regions of images, because it is the fundamental process of many vision tasks. The ability to compare image regions (patches) has been the basis of many approaches to core computer vision problems, including object, texture and

scene categorization. Hence, developing representations for image patches have been of interest in several works. Computer vision tackles problems related to object detection and recognition, texture classification, action recognition, segmentation, tracking, data retrieval, image alignment, etc. There are several techniques for performing these tasks, and usually based on representing an image using some global or local image properties, and comparing them using some similarity measure. Learning visual similarities has been presented recently with success on images in the mono-spectral domain [120]. Images are often represented by compact region descriptors with interest points. The main idea is to extract all possible patches no matter overlapping. These patches are usually very small compared with the original size of the image, with them we proceed with their processing to exploit interrelation between them [14].

During early 2000 different approaches based on hand-craft methods were widely used initially to obtain the representation of images using vectors. These feature descriptors such as: SIFT [63], SURF [11], KAZE [8], among the best known have been applied for the resolution of many computer vision problems based on the main characteristics of the images and had had a great impact on computer vision area.

Many researchers have been working with image patches for road detection and urban understanding, which can be used for image labeling [15]; other approaches have been proposed based on image-adaptive wavelet transform, which is tailored to sparsely represent a given image, to form a multiscale sparsifying global transform for the image in question [15]. There are some other methods based on image patch processing like a fast patch dictionary for image recovery and sparsity-based image denoising via dictionary learning and structural clustering [27], non-local means methods for image denoising [23] and image processing using the smooth ordering of its patches [80].

Approaches to perform image completion involve filling missing parts in images. In [38] the authors have proposed novel statistics of similar patches. They propose that the coincidence of similar patches in the image allows them to obtain their displacements (relative positions), and the statistics resulting from these displacements can be used to obtain reliable information to complete the image.

Another application that can be derived from good management of the regions of the images is the edition of the same ones to modify the position of objects, to make changes in the texture or any other adjustment that is required to make in an image. Also, in the analysis of medical images, techniques based on regions of images (patches) are also observed in order to compare the similarity of the images with the related databases already existing, to determine whether or not they are similar [115].

In order to overcome the aforementioned poor performance some recent approaches, based on the usage of convolutional neural networks, have been proposed with interesting results. Some times such good results are obtained using expensive dedicated GPUs. One approach is to learn a feature representation directly from image data, to obtain a general similarity function for comparing image patches. To formulate such a function, various CNN-

(*a*) (*b*)

Figure 3.1 – Cross-spectral pairs of images obtained from [14]: (*a*) Visible spectrum images; (*b*) NIR images.

based models have been designed and trained to support a wide variety of changes in image appearance [120].

The previous approaches have been developed using images or patches from the visible spectrum, in other words, monospectral approaches. These days, the coexistence of cameras working at different spectral bands have considerably increased, mainly based on recent advances in imaging devices as well as the reduction on the prices of such a technology, which makes it possible to propose new architectures based on convolutional neural network. This cross-spectral information helps to solve classical problems in poor lighting conditions or enhance visible spectrum images with information from other spectral bands (e.g., filtering [41], enhancement [125]). The current chapter is focused on the usage of images from the visible spectrum (RGB images) together with images from the near infrared spectrum (NIR images). Figure 3.1 presents two pairs of cross-spectral images used to validate the proposed approaches based on the 2 channel CNN network and based on the usage of meta-learning.

The usage of cross-spectral information, although interesting and appealing, implies new challenging and difficult problems that need to be tackled and efficiently solved. For instance, different works have been recently proposed for describing and matching feature points in cross-spectral domains based on classical approaches (e.g., [94], [4], [6], [71], to mention a few). Unfortunately, due to the natural difference between images acquired from different spectra, the obtained performance is far away from the one obtained in mono-spectral scenarios.

In the current chapter, firstly, the use of the CNN architecture presented in [3], but mod-

ifying the number of layers and reducing both the size of patches and convolution kernels is proposed, in order to use it in a low-cost hardware (about ten times cheaper than the one used in [3]). This network consists of a unified architecture that jointly learns a 2 channel deep neural network for cross-spectral patch representation (see Fig. 3.2).

Secondly, meta-learning techniques have been proposed, which allow generalizing a new model from few data. Therefore, there is no longer a dependence on large datasets for the training process. Also, meta-learning can take the advantage to use other sources of data that are not labeled but plenty available, of multimodal learning, transfer learning, and continuous learning for domain adaptation. With meta-learning, a specific transformation of a subset of features is critical for transferring the knowledge, in order to obtain a distribution of patterns in the feature space that share some characteristics that may be described by the model and easy to adapt to a new dataset of similar type to learn the representation of the features to generate a new one.

As mentioned above, the main contribution of the current chapter is to perform image matching using distance metrics and determine its similarity with low-cost hardware and reach a performance similar to the state-of-the-art techniques. The rest of the chapter is organized as follows. Section 3.2 presents the CNN architectures detailing the design and training with cross-spectral datasets. Section 3.3 depicts the experimental results and finally, conclusions are presented in Section 3.4.

3.2 Network Architecture

As mentioned above the current chapter is focused on finding correspondences between images from visible and near infrared spectra. Two different approaches have been proposed, the first network architecture proposed to find correspondences between patches from these images is similar to the one presented in [3], the 2 channel network (2ChNet) model. Figure 3.2 shows an illustration of the model. Details of the adaptation proposed to the (2ChNet) architecture are presented in Fig. 3.3; this architecture contains less layers than [3], in order to train it with a low-cost hardware. As can be appreciated in this illustration, this architecture takes as an input a pair of patches (one from each spectrum), and then a series of convolution, ReLu, and max-pooling layers are applied till the final linear layer that works as the metric network. Note that the patch from the visible spectrum (RGB image) is converted to grayscale.

The network learns the similarity by combining information from both spectra and jointly processing them through different layers. This way of processing the information has been shown as the best solution in cross-spectral domains [3]. The training process does not rely on labels assigned to each patch, but rather on pairs of patches of different spectra with similarity or non-similarity. During the training, the loss is minimized with a hinge-based term and squared $\ell 2$-norm regularization.

The model consists of different layers, like convolution, ReLU, max-pooling and a final linear layer that computes the loss of each iteration of the learning process. This last layer

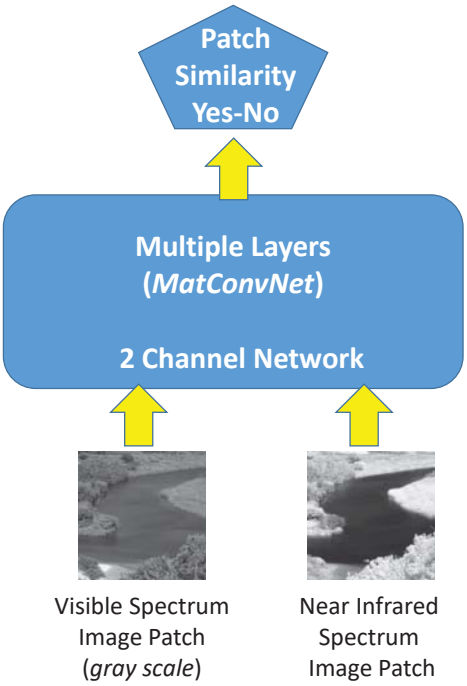


Figure 3.2 – The 2 channel network (2ChNet) model implemented on the current chapter to obtain automatic cross-spectral matchings.

acts as a metric, which permits to determine whether the pair of patches have or not correspondence. Figure 3.3 shows the adapted architecture of the model. The network architecture described above was trained in a supervised way; each layer convolves the output of the previous one, with a filter learned at each operation are followed by a non-linear activation function (ReLU). Some layers permit to change the spatial size of the output, obtaining the maximum or an average value of previous convolution layers, or the corresponding activation function. The last layers are fully connected and multiply the output obtained with a matrix of learned parameters. We use a margin criterion based on a *hinge loss* and squared *l2-norm* regularization term as in [120]:

$$\min_{w} \frac{\lambda}{2} \| w \|_2 + \sum_{i=1}^{N} max(0, 1 - y_i o_i^{net}), \tag{3.1}$$

where w is the network weight, o_i^{net} is the training output for the i-th training sample iteration and y_i is the i-th training label; the value domain is {-1,1} for a false and true similarity respectively, and λ denote the weight decay.

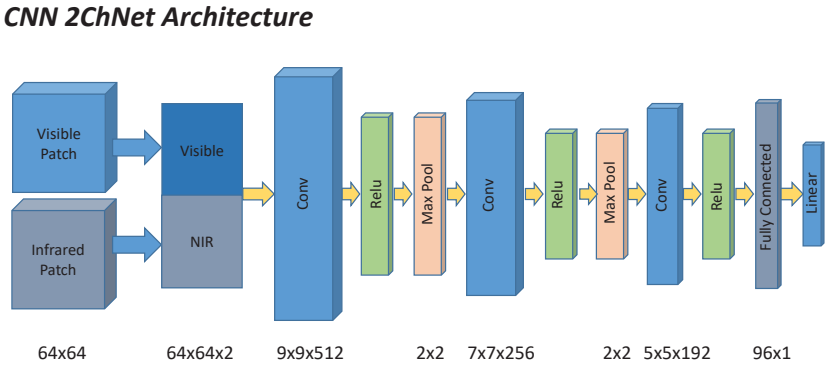


Figure 3.3 – Architecture of the 2ChNet adapted from [3]

Algorithm 3.1 Image Patch Similarity: i is the index of the image patches pairs vectors, IP_1 and IP_2 and the index of their corresponding label vector Y_i; $epochs$ is the number of iterations for training process; n is the number of images per batch; t is the number of images of the training dataset.

for `Number of image patches in the training set` (t) **do**
Determine $\overrightarrow{\text{IP}_1}$ $\overrightarrow{\text{IP}_2}$ pairs from the complete sample training set: $S_{\overrightarrow{IP}}$.
calculate the corresponding label Y_j, so that $Y_j = 0$ when $\{\overrightarrow{\text{IP}_1}\}_{i=1}^{t}$ are similar $\{\overrightarrow{\text{IP}_2}\}_{i=1}^{t}$ and otherwise $Y_i = 1$. **end for**
for $epochs$ `steps` **do**
    for n `steps` **do**
        Initialize network weights
        Instance the architecture of the siamese network
        Contrastive loss calculation, by minimization
        Neural Net Optimization
    end for
    Fine-tuning weights of the net based on error rate
end for

The second architecture implemented in this chapter to evaluate image similarity is a metric based meta-learning approach. One of the challenges raised by meta-learning techniques is the design of a deep training model that uses only a few training data and the previous experience taken from very similar learning tasks. This learning strategy is known to learn from few data shots, trying to simulate the human capacity to learn from one or a few examples. In the current chapter, an architecture that is capable of detecting the similarity of patches of cross-spectral images is proposed. In this case, the model has been designed to generate a similarity metric solely based on *K-shots in N-ways* learning in which it is given little training

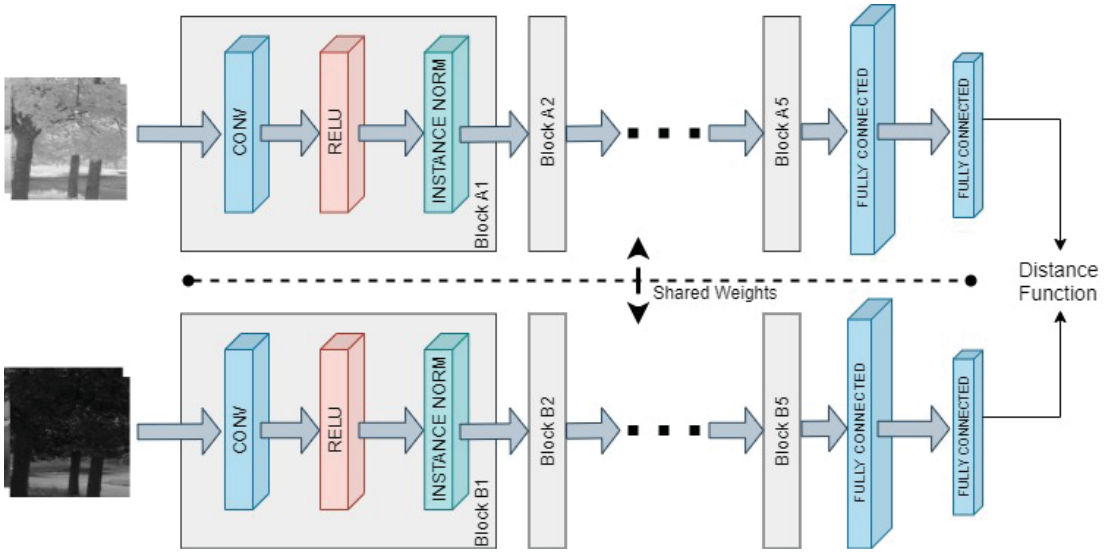


Figure 3.4 – Architecture of the siamese network for patch similarity, Blocks A_i/B_i contain the same set of operations (Conv, Relu and Instance Normalization).

data to determine whether or not similarity exists between the K classes with N data in each, (see Algorithm 3.1).

Once the model has been trained, the similarity metric can deduce the pattern of the common characteristics that represent the images evaluated by the trained meta-learning architecture. The weight parameters of the model have been designed to be shared and they are optimized as :

$$\theta^* = \arg\min_{\theta} \mathbb{E}_{C \sim p\,(\mathrm{C})}[\mathscr{L}_\theta(C)] \tag{3.2}$$

where θ^* tries to optimize the siamese model to obtain a semantic embedding space based on few shot samples and labels $\mathbb{E}_{C \sim p}$ per category, through the learning process to generate the representation vectors and determine the patching similarity using the corresponding dataset C. The dataset C contains pairs of feature vectors and labels, $C = (ip_i, y_j)$, and each label belongs to a known label set L.

The second proposed approach is based on a cross-spectral metric meta-learning model, implemented through a siamese network, see the architecture in Fig. 3.5. As mention in chapter 2 the metalearning scheme consist of poner la frase anterior The model is capable of determining the similarity of the cross-spectral image patches of different patterns, separately categorized, with few examples in each class, having 50% of these examples with images of the visible spectrum and the rest with images of the near infrared spectrum, with a few samples on each class per spectra; from the training process and for the test process there are three

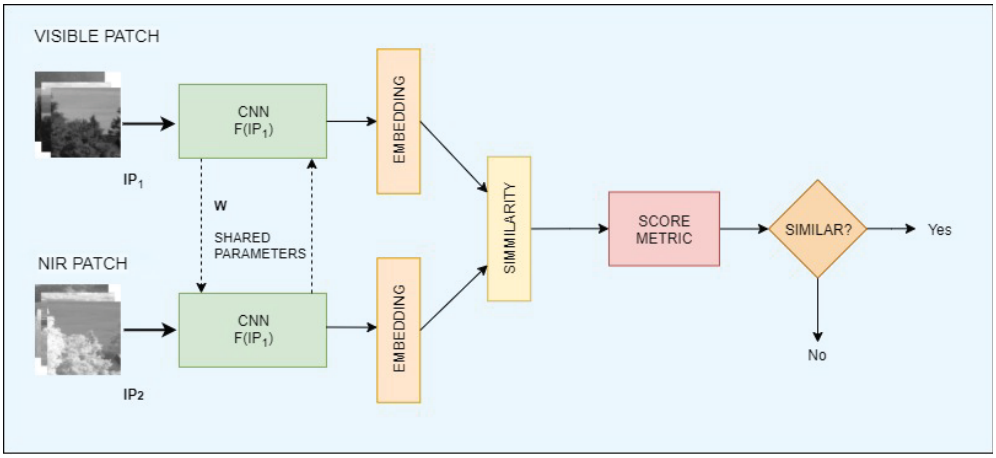


Figure 3.5 – Siamese general schema implemented on the current research.

classes not seen by the training process with few examples in each.

The meta-learning model proposed in this chapter has been designed to be trained over a variety of categories at the same time to obtain a good performance on the learning of metric similarity of the image patches. Being C the cross-spectral dataset of image patches of all categories to be considered in the training process to perform the learning similarity tasks and optimized for the best accuracy. Each task is associated with a cross-spectral categories of previously mentioned dataset C, containing both patch image representation and their corresponding labels.

Let IP_1 and IP_2 be a pair of image patches from visible and near infrared spectra respectively, and let L be their corresponding label; "0" for a similar image patch pair class and "1" for a non similar image patch pair, including cross-spectral image pairs existing in the training and test database. Let W be the shared weights in the siamese network architecture, see Fig. 3.4, which will be optimized incrementally as the proposed model is generalized. Having a generator function $G_w(ip)$ instantiated by a siamese architecture with a weight vector W. Being the siamese net instantiated $G_w(ip_1)$ and $G_w(ip_2)$ to obtain an embedding vector representation on each side of the network to measure the distance between those embeddings, and determine the similarity of the patches feed it into the network. The similarity function $G_w(IP_1, IP_2)$ is defined as:

$$G_w(IP_1, IP_2) = \|G_w(ip_1) - G_w(ip_2)\|^2. \tag{3.3}$$

Metric-based meta-learning model defines a function capable of determining the similarity between objects. Being this function a classifier ($k\theta$) that is a learner model to be trained for a given task, in this case, determine the similarity between cross-spectral image categories; the model must have also an optimizer that learns how to update the learner model's parameters

via the support set C, the corresponding dataset used for the experiments. The predicted probability over a set of known labels, in this case "0" for a similar image patch pair class and "1" for a non similar image patch pair, is a weighted sum of labels of support set C samples. The weight is generated by a kernel function $k\theta$, measuring the similarity between two data samples, it is defined as:

$$P_\theta(y|\mathbf{x}, C) = \sum_{(x_i, y_i)\in C} k\theta(x, x_i)\, y_i, \tag{3.4}$$

where C is the dataset that contains pairs of feature vectors x_i and labels y_i, and each label belongs to a known label set L. Being $k\theta$ the classifier with parameter θ that outputs a probability of a data point belonging to the class y in this case "0" or "1" when the feature vector x in this case the embedded representation of the cross-spectral image category does not belong to the class y.

The metric-based approaches learn one task invariant metric for all the tasks. Even though the metric-learning approaches allow different numbers of classes, they require the tasks all coming from a similar domain such that there exists a uniform metric that could work across tasks [107].

3.2.1 Instance Normalization

Deep learning is a technique that allows learning multiple levels of representations and abstraction to transform data to solve a specific problem. Many types of research with deep learning are focusing on developing techniques to stabilize training. Thus, some architectures are known to be unstable (during training) and very sensitive to the changes made on the model hyperparameters. For that reason, instance normalization has been implemented, see section 2.8 for a detailed information.

A small change in the stylization architecture proposed by [104] presents a qualitative improvement in the generated embedding vector. The fig. 3.4 shows the implementation of this normalization layer. The resulting method can be used to train high-performance architectures for real-time embedded vector generation. The architecture uses this normalization, applied in feed-forward style transformation, to improve the quality of the embedded feature representation generated by the model, which has been validated empirically in the experiments.

3.2.2 Contrastive Loss

To be able to differentiate if the images that are fed to the model are similar or not, an encoded representation must be obtained to conserve the structural and semantic information of images, for which the *contrastive loss* based on a maximum margin has been used.

According to [36], a meaningful mapping from high to low dimensional space maps similar input vectors to nearby points on the output manifold and dissimilar vectors to distant points. Unlike conventional learning systems where the loss function is a sum over samples, with the contrastive loss function the minimization runs over pairs of samples. Let $\overrightarrow{X_1}, \overrightarrow{X_2} \in I$ be a pair of input images shown to the system; let Y be a binary label assigned to this pair. Y = 0 if $\overrightarrow{X_1}$ and $\overrightarrow{X_2}$ are similar, and Y = 1 if they are dissimilar and P is the number of the training image pairs. Define the parameterized distance function to be learned D_W between $\overrightarrow{X_1}$, $\overrightarrow{X_2}$ as the Euclidean distance between the outputs of G_W. The distance is defined as:

$$D_w(\overrightarrow{X_1}, \overrightarrow{X_2}) = \|G_w(\overrightarrow{X_1}) - G_w(\overrightarrow{X_2})\|_2. \tag{3.5}$$

The corresponding loss function $\mathscr{L}$ is defined as:

$$\mathscr{L}_\theta(\mathscr{W}) = \sum_{i=1}^{P} L(W, (Y, \overrightarrow{X_1}, \overrightarrow{X_2})^i). \tag{3.6}$$

Applying this loss to the model, where the image pairs are the image patches of visible and near infrared spectra, such that minimizing L with respect to W results in low values of D_W for similar pairs of images and high values of D_W for dissimilar pair. As a result the above loss function can be defined as :

$$\mathscr{L}_\theta(W, (Y, \overrightarrow{X_1}, \overrightarrow{X_2})^i) = (1-Y)\frac{1}{2}(D_W)^2 + (Y)\frac{1}{2}(\max(0, m - D_W))^2, \tag{3.7}$$

where $m > 0$ is a margin. The margin defines a radius around the outputs of G_W. Dissimilar pairs contribute to the loss function only if their distance is within this radius (m), see Fig. 3.6. The first term of the equation (3.7) corresponds to the partial loss function for a pair of similar points. The second term of the equation corresponds to the partial loss function for a pair of dissimilar points. The contrastive term involving dissimilar pairs is crucial. Simply minimizing $D_W(\overrightarrow{X_1}, \overrightarrow{X_2})$ over the set of all similar pairs will usually lead to a collapsed solution, since D_W and the loss L could then be made zero by setting G_W to a constant.

3.3 Experimental Results

To test the first proposed approach the cross-spectral dataset from [14] has been used (pairs of images are presented in Fig. 3.1). This dataset consists of 477 registered images categorized in 9 groups captured in RGB (visible spectrum) and NIR (Near Infrared spectrum). To compare with the previous approach, [3], just images from the category *country* have been used for training (150 pairs of images randomly selected). These images are the most affected in

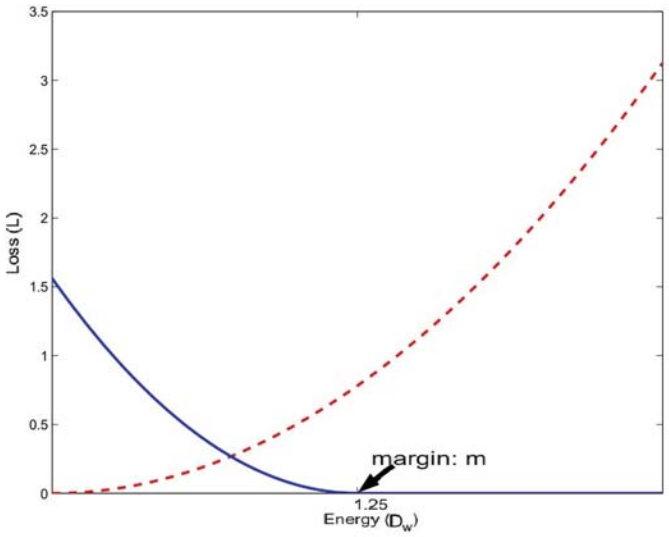


Figure 3.6 – Illustration of the loss function L against the energy D_w; the dashed (red) line is the loss function for the similar pairs and the solid (blue) line is for the dissimilar pair; illustration from [36]

conditions of varying lighting and textures, which directly affect the variability and complexity of the detection of the feature points and therefore are the most challenging scenarios for the training process. Feature points have been obtained from SIFT, applied over the visible spectrum images. Patches of 64×64 pixels have been generated centered on those points. Then, points placed in the same position than those obtained by SIFT algorithm are placed in the NIR images and the corresponding patches with the same size extracted. With this process 150.000 patches have been generated from *country* category (note that the provided dataset contains correctly registered pairs of images); the same amount of patches have been generated using random points taken from each visible and near infrared images, that is, forming unpaired image patches for the false pair dataset.

The model is trained using Stochastic Gradient Descent with a weight decay (λ) of 0.0007, a learning rate of 0.05, a momentum of 0.9 and batches size of 80 samples. All input patches were normalized by their intensity mean, previous to normalization the values of intensities must be in the $[0,1]$ range. (80%) of the dataset generated as mentioned above has been used for training, while 20% used for validation. It has been used MatconvNet toolbox for Matlab that implements Convolutional Neural Networks [105]. The 2ChNet model was trained during 6 days, on a 3.2 eight core processor with 4Gb of memory with a NVIDIA GeForce GTX970 GPU.

Once the 2ChNet has been trained with images from the *country* category it has been evaluated with other cross-spectral images from the *country* category together with other categories. Thus, 300 pairs from each of the following categories have been selected: *country*,

(*a*) (*b*)

Figure 3.7 – Cross-spectral pairs of images obtained from [14]: (*a*) Visible spectrum images; (*b*) NIR images.

Table 3.1 – Evaluations (FPR95%) on visible and near infrared image patch datasets [14] from different categories (the smaller the better, bold faces correspond to the best results in that category).

Descriptor-Network	*country*	*indoor*	*oldbuilding*	*urban*
SIFT [63]	46.6	12.4	21.3	13.27
2ch Network (from [3])	**0.23**	4.4	**2.3**	**1.58**
2ch Network (1st. Prop. App.)	0.27	**3.3**	3.4	4.6

indoor, *oldbuilding* and *urban* respectively. The results obtained from this evaluation were compared with those obtained with a classical feature descriptor (SIFT) to highlight the improvements in performance reached with the proposed approach. The FPR95% rate, which is the ratio between the number of negative coincidences wrongly categorized as positive (false positives) and the total number of actual negative coincidences (regardless of classification), is used to measure the obtained results. Additionally, these values have been compared with the ones presented in [3] using the same image categories. Table 3.1 shows the obtained performances. As expected, it can be appreciated the large improvements reached concerning SIFT. Additionally, it can be appreciated that in spite of the hardware limitations and the corresponding reduction in the size of the proposed model, the results are similar to the one presented in [3]. Actually, in one case the result is even better than the one obtained in [3].

The second proposed approach for determining patch similarity has been tested using the previous cross-spectral dataset from [14]. The proposed 8-$shot$ 1-way meta-learning metric based network model has been trained with two image categories and evaluated with four image categories, two of which have not been used in the training process. Figure 3.7 shows some examples of pairs of different categories). In order to make comparisons with similar approaches, like [3] and the first proposed approach in this chapter, *country, indoor, oldbuilding and urban* categories have been selected. The cross-spectral dataset for the experiments has been prepared: pairs of randomly selected images from visible and near infrared for each category have been collected. First, a set of characteristic points on the visible spectrum images have been obtained using the SIFT algorithm, to search these points and their corresponding ones in the near infrared spectrum images. To carry out the experiments, 64×64 pixels patches have been generated centered on the previously detected points in both the visible and near infrared spectrum images, since the images are perfectly aligned; then the corresponding patches are extracted with the previously defined size. For training, a total of 16 perfectly balanced cross-spectral image patches (matched and not matched) have been selected for each category. It has been used Adam optimizer with a learning rate of 0.0002, with stochastic gradient descent, minimizing the contrastive loss to converge the model.

Once the meta-learning metric based model has been trained, it has been evaluated with other cross-spectral image samples in order to obtain the performance of the model. Thus, new pairs from each of the categories have been selected and used to test the model. Like in the

Table 3.2 – Evaluations (FPR95%) on cross-spectral image patch datasets from different categories [14] (the smaller the better, bold faces correspond to the best results in that category).

Descriptor-Network	*country*	*indoor*	*oldbuilding*	*urban*
SIFT [63]	46.6	12.4	21.3	13.27
2ch Network (from [3])	0.23	4.4	2.3	**1.58**
2ch Network (from 1st. Prop. App.)	0.27	3.3	3.4	4.6
Metric Based Network (cross-spectral 2nd. Prop. App.)	**0.22**	**3.1**	**2.2**	1.63

Table 3.3 – Evaluations (FPR95%) on visible and near infrared image patch datasets evaluated separately from different categories, [14] (the smaller the better, bold faces correspond to the best results in that category).

Descriptor-Network	*country*	*indoor*	*oldbuilding*	*urban*
Metric Based Network 2nd. Prop. App. (only Visible spectrum images)	**0.17**	**1.6**	1.9	1.21
Metric Based Network 2nd. Prop. App. (only Near infrared images)	0.19	2.6	**1.6**	**1.43**

previous case, the results obtained from this evaluation were compared with SIFT. Additionally, these values have been compared with the ones presented in [3] and with the results from the first proposed approach. Table 3.2 shows the obtained performances. As expected, it can be appreciated the large improvements reached from SIFT. Additionally, it can be also appreciated better results than those presented in [3] and with respect to the first proposed approach, actually, only in the *urban* category previous approaches remain a bit better than the ones obtained with this second proposed approach. The model has been also evaluated when the similarity of images from the same spectra has been considered (i.e., visible spectra patches only or near infrared patches only), just to show that the proposed model can measure similarity without having to train it again showing that it is adaptable to various measurement tasks. The results are shown in the Table 3.3

3.4 Conclusions

In this chapter the challenging problem of cross-spectral image patch similarity has been tackled; firstly, by adapting a state of the art architecture [3] with low-cost hardware. The results obtained with the 2 channel approach shows that even with a low-cost hardware the obtained performance is quite similar to the state of the art, as well as it is shown that outperforms classical SIFT feature based descriptors. This implementation was the first implementation in the framework of this thesis. Hence, at that time, MatConvNet was one

of the most widely used tools to work on deep learning. Nowadays, the institution counts with more powerful computational hardware (GPU Titan XP, Titan V, among others) and with specialized libraries for deep learning (e.g., Tensorflow), that has become the standard framework on machine learning techniques developments. Hence, with these new specialized libraries and the equipment with greater computer capacity new approaches have been implemented to determine the similarity of the cross-spectral images; using much more advanced approaches, such as meta-learning metric based technique obtaining better results. In this way, the rest of the research carried out in the framework of the thesis is by using these specialized libraries, and hardware with greater computational power mentioned before.

Chapter 4

Near Infrared Image Colorization

This chapter presents novels approaches to generate RGB representations from near infrared (NIR) images by using Generative Adversarial Networks (GANs). The networks have been implemented using a single, triplet learning level architecture, applying several variations to improve the quality and performance of the network. The applied variations include a conditional, dense and a stacked model to generalize the image representation process. Finally, a model with multiple loss functions, which combine different terms to improve final results is proposed. The proposed techniques obtain satisfactory results when objects from different categories are tested. Experimental results, with a large set of real images, are provided to show the validity of the proposed approaches. Comparisons with other architectures are also provided to show the improvements reached with the proposed approaches.

4.1 Introduction

Visible spectrum images, referred through this chapter indistinctly as visible spectrum or RGB images, have limitations related to lighting conditions and object surface's color. Some of the limitations mentioned above can be easily overcome using Near Infrared (NIR) imagery. The NIR band of the electromagnetic spectrum is just outside the range of what humans can see and can sometimes offer clearer details than what is achievable with visible light imaging. In the computer vision domain, several applications take advantage of near infrared spectrum, since materials have characteristics (i.e., physical or chemical properties) that can be easily detected in the NIR spectral band. In many vision applications, including surveillance photo interpretation by imagery analysts and driving scene understanding by drivers looking at backup-aid cameras, RGB video sensors are preferred since depicted images are similar to the human visual perception system. Whereby, in spite of the advantages of NIR imagery, when the information needs to be shown to the people of the interest group the visible representation (e.g., RGB) is always preferred since it allows a better appreciation and understanding of the scene and therefore it will be possible to make better decisions.

The NIR spectrum is independent of the brightness and color of the targets, which has

potential benefits, including non-visible illumination requirements. Surface reflection in the NIR spectral band is material dependent. This means that the difference in the NIR intensities is not only due to the particular color of the material but also to the absorption and reflectance of dyes. In this context, this chapter addresses the process of colorization using images of the near infrared spectrum to obtain their representation in the visible spectrum (RGB representation). Different solutions could take advantage of this contribution, for instance, to give daylight to night images, to allow the observers can selectively attend relevant color targets and to ignore non-targets with an irrelevant color, to map cross-spectral into a three-dimensional (synthetic) color space to increase the dynamic range of a sensor system increasing detection probability [40].

For the implementation of the NIR images colorization, different GAN networks have been proposed instead of a standard CNN network. One of the principal reasons why adversarial generative networks have been selected are dependent on some advantages that those networks have to generate synthetic information, in this case, colored images. These networks apply a supervised learning method that does not require data sets to feed it as an input, which is a great advantage, additionally, GANs networks allow generating information from any domain, being these images, text, audio and video, since they are able to learn the internal representation of the data they generate as the network is trained based on examples taken from real data (ground truth). GANs can learn distributions of messy and complicated data. This can be used to solve a variety of machine learning problems, including the one that is being proposed for NIR image colorization. This framework has been presented in detail in Section 2.5.

In this chapter, four approaches for NIR image colorization are detailed, the first one, based on a standard GAN model with one leaning layer for colorization (flat), the second one with a model with multiple level of learning, one for each color channel (red, green, blue), the third one a conditional triplet level model with multiple loss functions, and the last one a stacked GAN with multi-dense connection and loss functions. Figure 4.1 illustrates the process for infrared imagery coloring using a standard conditional GAN.

The results obtained from the experiments carried out show that the best results is obtained with the approach based on a conditional stacked GAN. This stack architecture is based on the approach presented in [44]. This approach consists of a top-down GAN stack, each designed to generate lower level representations conditioned to higher level representations. This strategy allows accelerating the learning process to generate a new image representation from NIR to RGB. In addition, this stacked GAN proposed model includes multiple losses, which are continuous and differentiable with which the training process is improved. This fourth model maximizes the process of obtaining new representations from images of the near infrared spectrum so that they can be better represented in the visible spectrum. The fourth proposed approaches are detailed in Section 4.2. The multiple loss function used in the fourth model is detailed in Section 4.3 Experimental results are presented in Section 4.4. Finally, conclusions are given in Section 4.5.

4.2 Proposed Approaches

In the particular problem tackled in this chapter (NIR image colorization), as mention above, four approaches have been implemented, the first one, based on a single learning level (flat), the second one a triplet learning level model, the third one a conditional triplet model with multiple loss functions, and the last one a stacked GAN with multi-dense connection and loss functions. Figure 4.2 presents the learning architecture of a single level of learning that allows obtaining a RGB representation using a standard adversarial generative architecture, based on a Gaussian noise distribution concatenated with the NIR image.

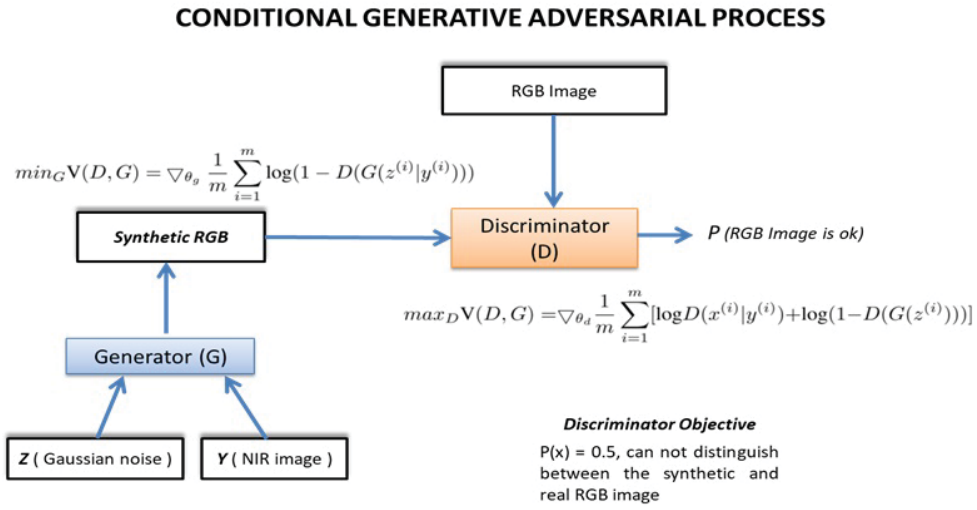


Figure 4.1 – Illustration of the process of conditional GAN network for NIR image colorization.

The Figure 4.3 shows an improved architecture a deep convolutional GAN (DCGAN), where a triple learning model is applied, that is, each channel of the RGB representation has its own learning layer, this allows improving the diversity of colorization of the network, because independent learning is maintained by each channel. Figure 4.4 shows the changes applied to the triple-level learning architecture conditional deep convolutional GAN (CDCGAN), which proposes to condition one of the learning levels of architecture, in particular, the red channel, where the output of this level of the red channel is concatenated with the corresponding NIR image, this operation is performed due to the overlap of the VISIBLE-NIR bands in the electromagnetic spectrum, and it is used to improve the details of the resulting color images (for a better understanding of the subject of the band overlap the spectral sensitivity graph is shown in Fig. 4.5). The last architecture implemented for NIR colorization is shown in Fig. 4.6 where a multiple loss function, dense connections are applied in the stacked conditional GAN (SC-GAN) has been included in the architecture to obtaining a better convergence and accuracy, enhancing also the quality of the generated images. Fig. 4.7 shows the general schema of the stacked GAN proposed in the fourth approach.

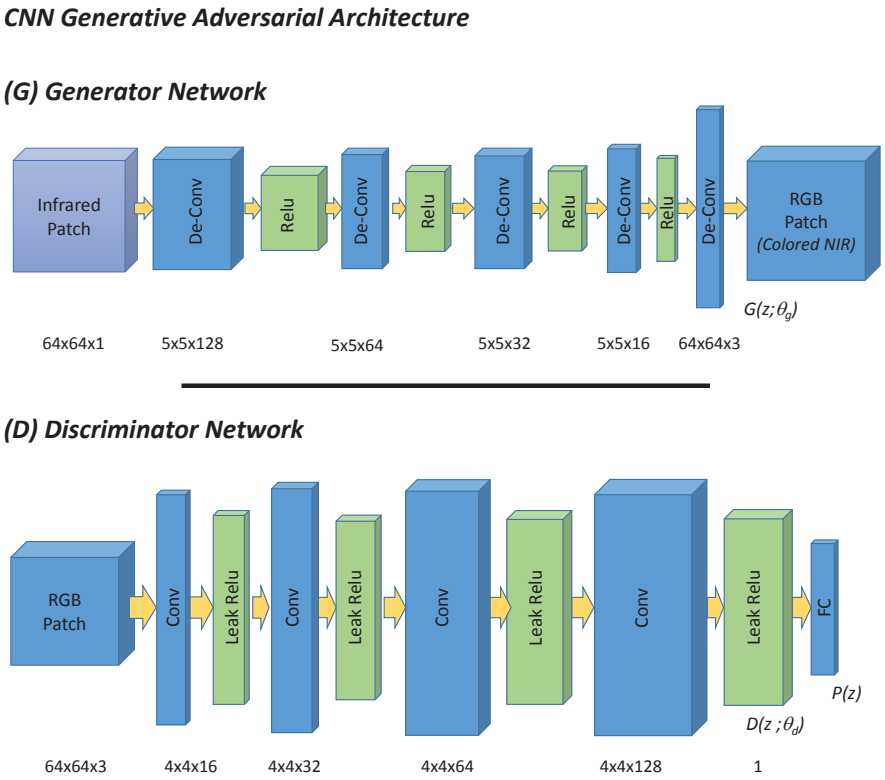


Figure 4.2 – Illustration of the flat GAN network architecture, the first approach proposed for NIR image colorization.

The proposed fourth approach builds upon the second and third approaches presented in this chapter, which consists in an adaptation of the stacked GAN architecture presented in [44]. With this stack model during the learning process, a feature hierarchy in each layer has been added to encourage the representation manifold of the generator to align with a bottom-up way to the discriminative network, leveraging the powerful discriminative representations to guide the generative model. This stacked learning model allows accelerating the diversity obtained in the multiple level of training representing each of the channels of an image of the visible spectrum (RGB). Therefore, the model will receive as input a near infrared patch (NIR) fused with Gaussian noise to ensure more diversity of colors, also in this approaches a layer of Gaussian noise has been included in each level of the triplet architecture of the generator model to reinforce the generalization and therefore be able to optimize the learning of the colorization process. An $l1$ regularization term has been added at every layer of the generator model in order to prevent the coefficients to fit so perfectly to overfit and for mitigating the Gaussian noise included in the generator model, which can reduce the time necessary to reach a generalized trained model.

In the fourth approach a stacked conditional GAN network, has been selected for these

others reasons: i) it optimizes the higher-level features resulting from the generator model; ii) the learning is conditioned on NIR images plus Gaussian noise from the source domain; iii) it has a fast convergence capability; iv) the capacity of the generator model to easily serve as a density model of the training data; and v) sampling is simple and efficient. The SC-GAN is designed to learn and generate a new sample from an unknown probability distribution. In the proposed SC-GAN framework, the generator network has been modified to use feature hierarchical representation. Additionally, to optimize the model generalization, the GAN framework is reformulated for a conditional generative image modeling tuple, see Fig 4.1. In other words, the generative model $G(z;\theta_g)$ is trained from a NIR image plus Gaussian noise, in order to produce a RGB image; the discriminative model $D(z;\theta_d)$ is trained to assign the correct label to the generated colored image, according to the provided ground truth RGB image. Variables (θ_g) and (θ_d) represent the weighting values for the generative and discriminative networks.

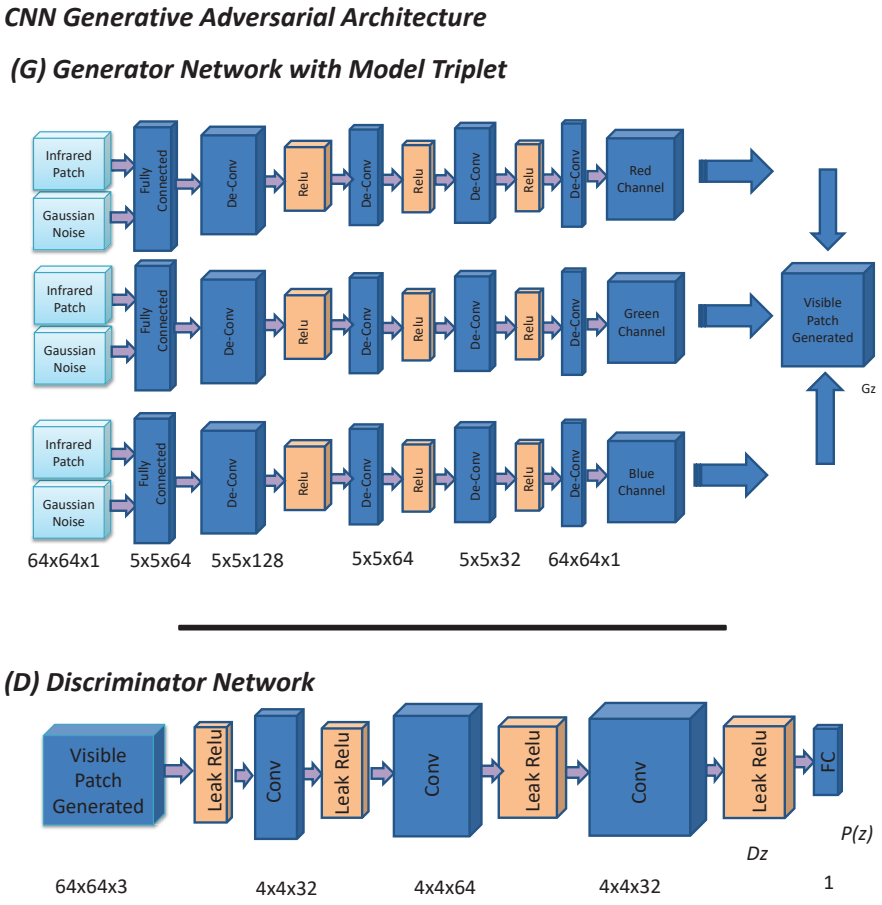


Figure 4.3 – Illustration of the triplet GAN (DCGAN) network architecture, the second approach for NIR image colorization.

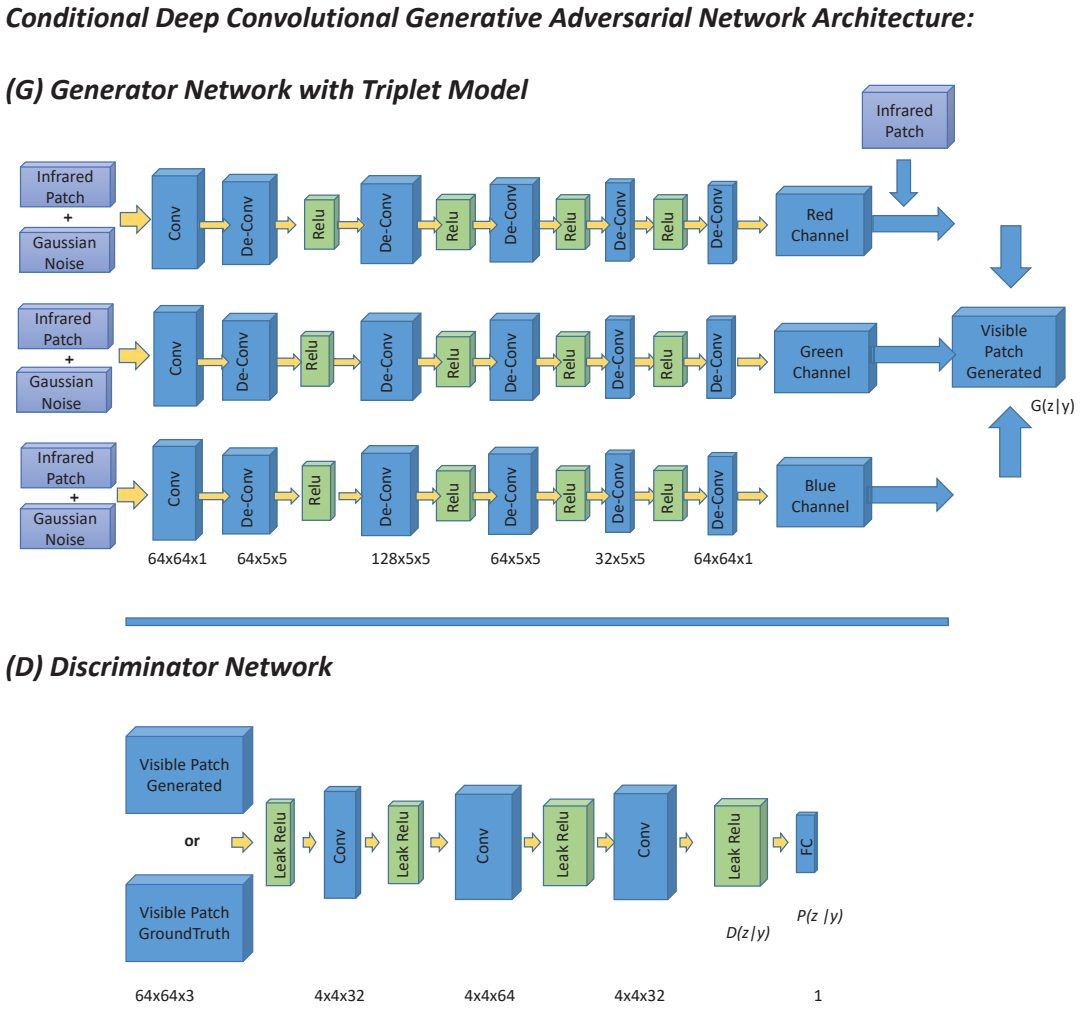


Figure 4.4 – Illustration of conditional GAN (CDCGAN) network architecture, the third proposed approach for NIR image colorization.

4.3 Multiple Loss Function

The third model has been defined with a multiple term loss function ($\mathscr{L}$) formed by the combination of the adversarial loss, plus the intensity loss (MSE) and the structural loss (SSIM). This combined loss function has been defined to avoid the usage of only a pixel-wise loss (PL) to measure the mismatch between a generated image and its corresponding ground truth image. This multi-term loss function is better designed to human perceptual criteria of image quality, which is detailed below.

The adversarial loss is designed to minimize the cross-entropy to improve the texture loss:

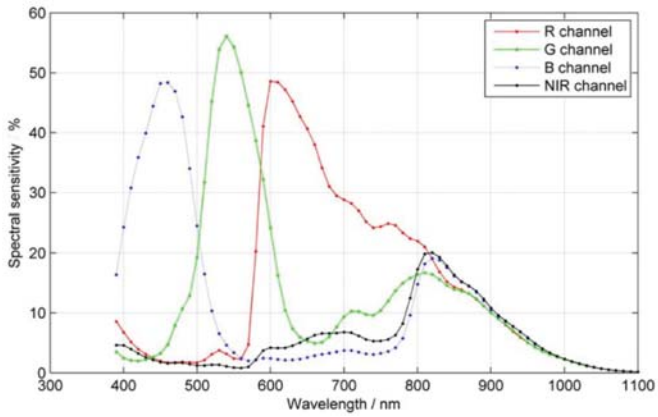


Figure 4.5 – Illustration of spectral sensitivity graph, which shows the overlap between the VISIBLE-NIR bands in a single sensor multispectral camera.

$$\mathscr{L}_{Adversarial} = -\sum_{i} logD(G_w(I_{z|y}),(I_{x|y}), \tag{4.1}$$

where D and G_w are the discriminator and generator of the real $I_{x|y}$ and generated $I_{z|y}$ images conditioned by the near image in each channel of the SC-GAN Network.

The intensity loss is defined as:

$$\mathscr{L}_{Intensity} = \frac{1}{NM}\sum_{i=1}^{N}\sum_{j=1}^{M}(RGBe_{i,j} - RGBg_{i,j})^2, \tag{4.2}$$

where $RGBe_{i,j}$ is the estimated RGB representation and $RGBg_{i,j}$ is the ground truth RGB image. This loss measures the difference in intensity of the pixels between the images without considering texture and content comparisons. This loss penalizes larger errors, but is more tolerant to small errors, without considering the specific structure in the image. To address the limitations of the simple intensity loss function, the usage of the Structural Similarity Index (SSIM) [111] is proposed; it evaluates images accounting for the fact that the human visual perception system is sensitive to changes in local structures. The idea behind this loss function is to help the learning model to produce a visually improved image. The structural loss for a pixel P is defined as:

$$\mathscr{L}_{SSIM} = \frac{1}{NM}\sum_{p=1}^{P} 1 - SSIM(p), \tag{4.3}$$

where $SSIM(p)$ is the Structural Similarity Index (see [111] for more details) centered in pixel p of the patch (P).

The final loss function ($\mathscr{L}_{final}$) used in this work is the weighted sum of the individual loss

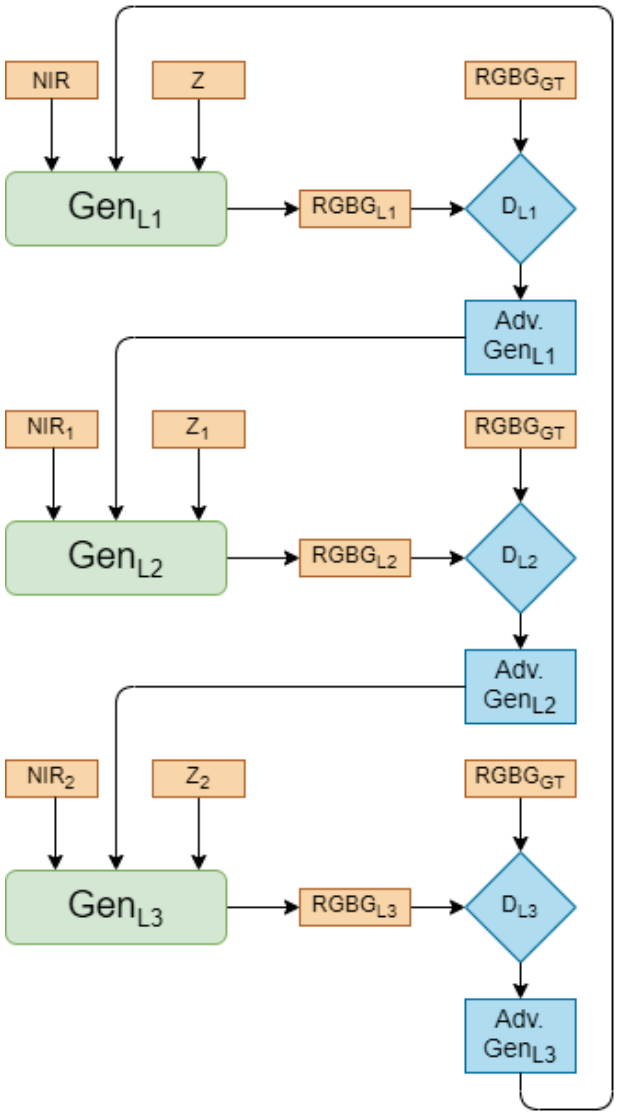


Figure 4.6 – Illustration of the fourth proposed stacked triplet GAN (SC-GAN) architecture with multiple losses proposed for NIR image colorization.

function terms:

$$\mathscr{L}_{final} = 0.65\mathscr{L}_{Adversarial} + 0.2\mathscr{L}_{Intensity} + 0.15\mathscr{L}_{SSIM}. \tag{4.4}$$

The proportion assigned to each loss has been defined based on the variability of the values obtained by each of the losses during the training process; the losses with greater fluctuation were assigned a greater proportion of impact on the optimization of the model.

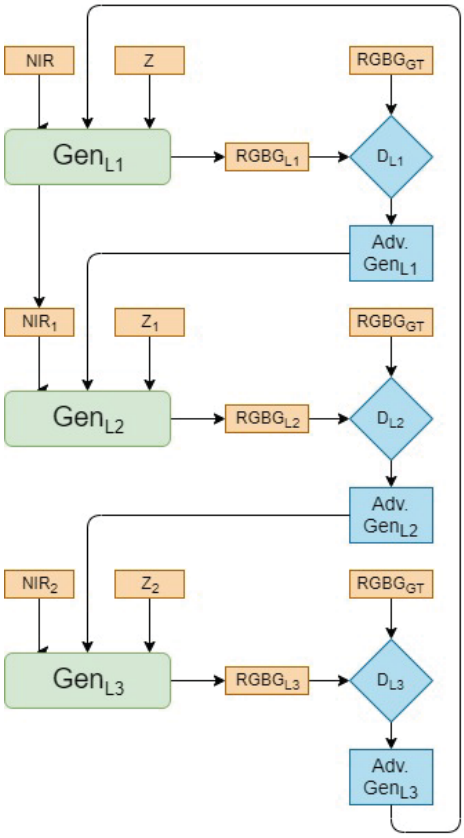


Figure 4.7 – Illustration of the proposed stacked GAN schema used NIR image colorization.

The stacked conditional GAN network has been trained using Stochastic AdamOptimizer since it prevents overfitting and leads to convergence faster. Furthermore, it is computationally efficient, has little memory requirements, is invariant to a diagonal rescaling of the gradients, and is well suited for problems that are large in terms of data and/or parameters. The image dataset was normalized in a (-1,1) range and an additive Gaussian distribution noise with a standard deviation of 0.021, 0.024, 0.026 added to each image channel of the proposed triplet model. The following hyper-parameters were used during the training process: learning rate 0.0002 for both the generator and the discriminator networks; epsilon = 1e-08; exponential decay rate for the 1^{st} moment momentum 0.5, for discriminator and 0.4 for the generator; weight initializer with a standard deviation of 0.00282; $l1$ weight regularizer; weight decay 1e-5; leak relu 0.2 and patch's size of 64×64.

The triplet architecture used in the three previous approaches maintains the same operations in the structure. The architecture is formed by convolutional, de-convolutional, relu, leak-relu, fully connected and the activation functions *tanh* and *sigmoid* for generator and discriminator networks respectively. Additionally, every layer of the model uses batch normalization for training any type of mapping that consists of multiple compositions of affine

Figure 4.8 – Pair of images (1024×680 pixels) from [14]: *urban* Category (the two images in the left side) and *oldbuilding* category (the two images in the right side): (top) NIR images to colorize; ($bottom$) RGB images used as ground truth.

transformation with element-wise nonlinearity and do not stuck on saturation mode. The spatial information in the generator model is maintained. This is achieved by dropping pooling and drop-out layers. In the experiments, a stride of 1 has been used to avoid downsizing the images.

To prevent overfitting a $l1$ regularization term (λ) has been added in the generator model, this regularization has the particularity that the weights matrix ends up using only a small subset of their most important inputs and become quite tolerant to input images noise. Park et al. [75] present a color restoration method that estimates the spectral intensity of the NIR band in each RGB color channel to effectively restores natural colors. According to the spectral sensitivity of conventional cameras with the IR cut-off filter, the contribution of the NIR spectral energy in each RGB color channel is greater in the red channel, hence the architecture adds the NIR band at the final red channel layer, this improves the details of generated images, color and hue saturation.

The generator (G) and discriminator (D) are both feedforward deep neural networks that play a min-max game between one another. The generator takes as an input a NIR image blurred concatenated with a patch of Gaussian noise with a size of 64×64 pixels and transforms it into the form of the data that are interested in imitating, in this case, a RGB image. In training (model building), the discriminator takes as an input a set of data, either real image (z) or generated image ($G(z)$), and produces a probability of that data being real ($P(z)$). The discriminator is optimized to increase the likelihood of giving a high probability to the real data (the ground truth image) and a low probability to the fake generated data (wrongly colored NIR image), as introduced in [70]. The equations are explained in Section 2.5.

4.4 Experimental Results

This section presents the results obtained after the implementation of the proposed approaches using several variations of a GAN network for NIR colorization. The proposed approaches have been evaluated using NIR images and their corresponding RGB obtained from [14]. The *urban* and *oldbuilding* categories have been considered for evaluating the performance. These categories have been selected since they look quite similar; the intention is to evaluate the capability of the network to be used in scenarios containing similar objects, which have not been used during the training stage. Figure 4.8 presents two pairs of images from each of these categories. The *urban* category contains 58 pairs of images of (1024×680 pixels), while the *oldbuilding* contains 51 pairs of images of (1024×680 pixels). From each of these categories, 280.000 pairs of patches of (64×64 pixels) have been cropped both in the NIR images as well as in the corresponding RGB images. Additionally, 5600 pairs of patches per category have been generated for validation. It should be noted that images are correctly registered so that a pixel-to-pixel correspondence is guaranteed.

The quantitative evaluation of the fourth proposed approaches, consists of measuring at every pixel the angular error between the obtained result (colorized NIR image) and the corresponding RGB image provided in the given dataset as ground truth values:

$$\text{Angular Error} = \cos^{-1}\left(\frac{\text{dot}(RGB_{NIR}, RGB_{GT})}{\text{norm}(RGB_{NIR}) * \text{norm}(RGB_{GT})}\right), \tag{4.5}$$

where RGB_{NIR} is the colorized NIR image, and RGB_{GT} is the corresponding ground truth image, both perfectly aligned. This angular error is computed over every single pixel of the whole set of images used for validation.

4.4.1 NIR Colorizing results from single and triplet GAN approaches

Four colorization models based on GAN networks have been proposed in this chapter, of which, the first three (single, triplet, conditional triplet GAN networks respectively), despite the good results obtained, have been overcome by the fourth proposed approach that is based on hierarchical learning (a stacked GAN network).

The first NIR colorization model implemented was through a GAN network with a single learning level for both the generator and discriminator networks, using *urban* and *oldbuilding* category images. The colorization results are shown in Fig. 4.9 and the quantitative results are presented in table 4.1

The second NIR colorization model implemented in the framework of this chapter was through a Conditional GAN network (DCGAN) with a triplet learning level for generator and a single level for discriminator networks, using *urban* and *oldbuilding* categories. This second proposed architecture has been evaluated using two different training schemes. Firstly, the DCGAN network has been trained with the *urban* category and evaluated with both *urban* and

Figure 4.9 – Results obtained from the first proposed approach: (*top*) Original NIR patches to be colorized (64 × 64 pixels); (*middle*) Results from the proposed approach; (*bottom*) Ground truth images.

Table 4.1 – Average angular error obtained with the proposed single level GAN architecture, for each image category.

Architecture	Angular error	
	urban	*oldbuilding*
Flat GAN from the 1st. Prop. App.	5.07	15.28

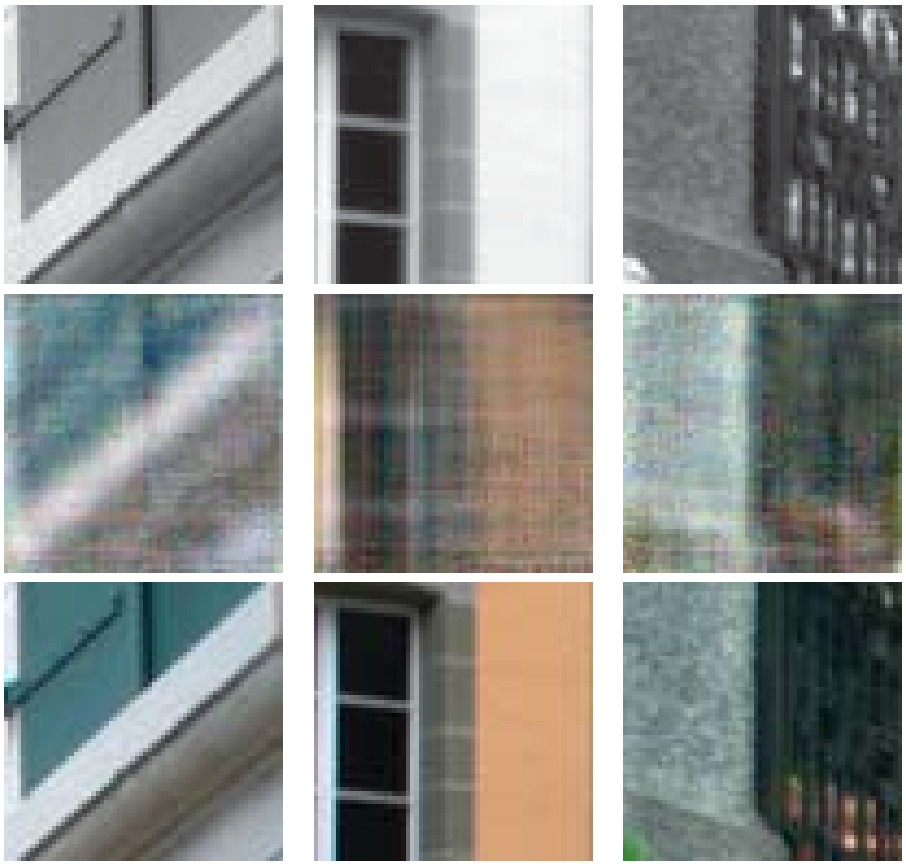


Figure 4.10 – Results obtained with the second approach: (top) NIR images from the *oldbuilding* category; ($middle$) Images colorized with the DCGAN network trained with *urban* Images; ($bottom$) Ground truth images.

oldbuilding categories, and secondly, the model has been trained with *oldbuilding* category and testing with both *urban* and *oldbuilding* categories.

Also, the DCGAN network has been trained with both datasets and evaluated independently in each of them, *urban* and *oldbuilding* categories. The colorization results for oldbuilding category are shown in Fig. 4.10 and for the urban category in Fig. 4.11. The quantitative results are presented in table 4.2:

Figure 4.11 – Results obtained with the second approach:(top) NIR images from the *urban* category; ($middle$) Images colorized with the DCGAN network trained with *urban* Images; ($bottom$) Ground truth images.

Table 4.2 – Average angular errors obtained with the second proposed approach a Triplet based DCGAN architecture.

Category	Angular error	
	urban	*oldbuilding*
urban	4.8	8.6
oldbuilding	9.8	7.1
both categories	7.4	8.2

The third NIR colorization model implemented was a variant of the previous Conditional GAN network with an inclusion of the near infrared image as a dense connection on the red layer of the network to improve the detail of the estimated colored image. The colorization results for *oldbuilding* category are shown in Fig. 4.10 and for the urban category in Fig. 4.11.

Table 4.3 – Average angular errors obtained with the second approach, a triplet based DCGAN, and with the third proposed approach, a conditional triplet based CDCGAN architecture.

Category	2nd.Prop. App. (DCGAN)	3rd. Prop. App. (CDCGAN)
urban	4.8	4.34
oldbuilding	9.8	5.71

Table 4.4 – Average angular errors (AE), mean squared error (MSE) and structural similarities (SSIM) obtained with the proposed stacked conditional GAN architecture by using different loss functions (SSIM values, the bigger the better).

Architecture	AE		MSE		SSIM	
	urban	*Oldbuilding*	*urban*	*Oldbuilding*	*urban*	*Oldbuilding*
CDCGAN from 3rd Prop. App.	5.77	5.96	18.91	18.25	0.84	0.86
Proposed SC-GAN with $\mathscr{L}_{Adversarial} + \mathscr{L}_{Intensity}$	5.43	5.21	18.74	18.11	0.86	0.89
Proposed stacked conditional GAN with $\mathscr{L}_{Adversarial} + \mathscr{L}_{SSIM}$	5.32	4.97	18.53	18.02	0.90	0.91
Proposed stacked conditional GAN with $\mathscr{L}_{final}$	5.04	4.78	17.63	17.34	0.90	0.91

The quantitative results obtained are shown in table 4.3

4.4.2 NIR colorizing with stacked conditional GAN, dense connections and multiple loss

The fourth approach of NIR colorization is the stacked conditional GAN (SC-GAN) with dense connections and multiple loss network as mention before, this approach has obtained the best colorization results. The model has been trained using a 3.2 eight-core processor with 16GB of memory with an NVIDIA TITAN XP GPU. On average every training process took about 60 hours. Results from the proposed architecture have been compared with those obtained with the Conditional GAN model presented in the third colorization approach.

The quantitative evaluation consists of measuring several metrics with the results obtained with the third colorization approach and the proposed stacked conditional GAN approach with different loss functions for each category; one of the metrics consists of measuring at every pixel the angular error (AE), see equation 4.5, between the obtained result (RGB_{NIR}) and the corresponding ground truth value (RGB_{GT}). AE is included since this measure is quite similar to the human visual perception system; some studies show the high correlation between the AE and the perception of human observer [33]. AE is probably the most widely used performance measure in color constancy research.

Quantitative evaluations for the different architectures using these three evaluation metrics (AE, MSE and SSIM), when considering the categories *urban* and *Oldbuilding*, are provided in Table 4.4. It can be observed that in all cases the results obtained with the proposed stacked

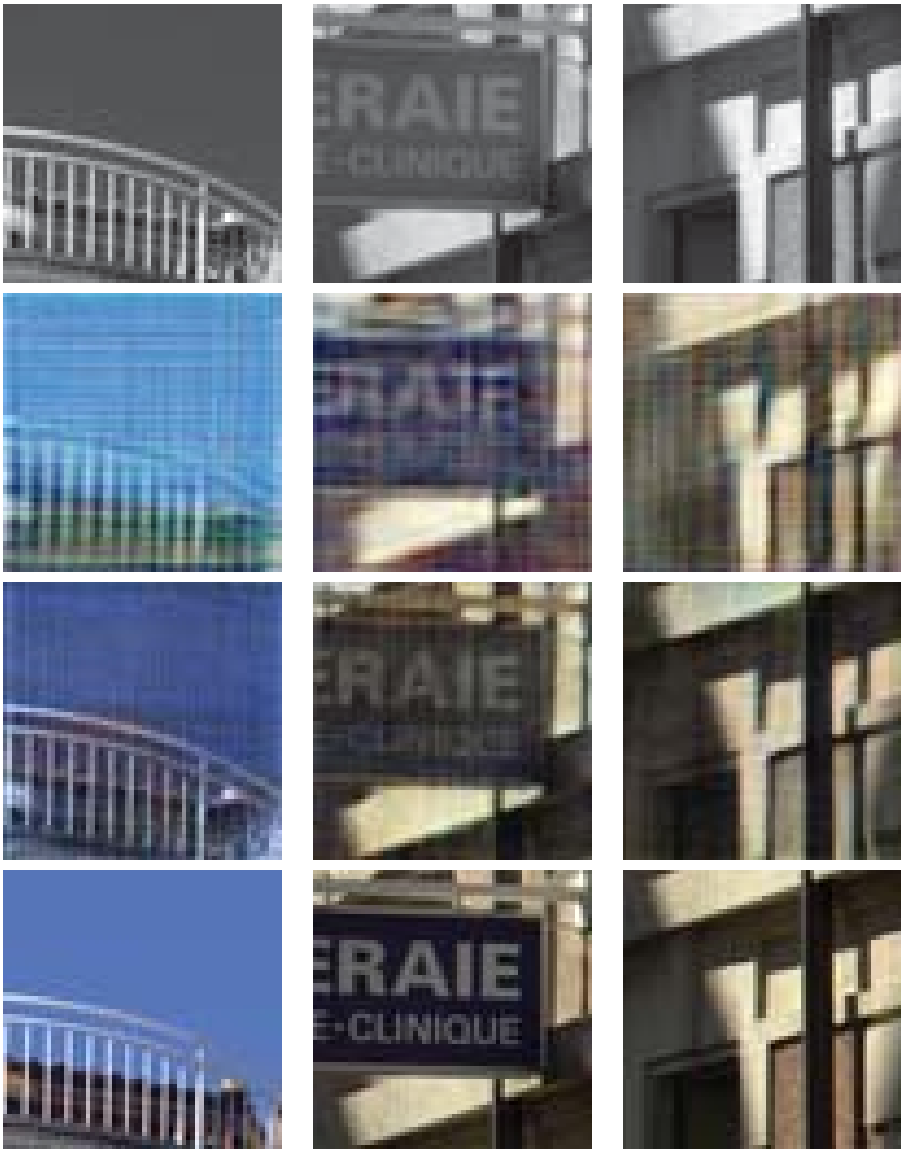


Figure 4.12 – Results of colorization: ($1st\ row$) NIR patches from the *urban category*; ($2nd\ row$) Results from the first approach (flat). ($3rd\ row$) Results from the third approach (conditional triplet) (CDCGAN network); ($4th\ row$) Ground truth images.

conditional GAN, using as a loss function the weighted sum of the three terms, are better than those obtained with the third proposed approach. Finally, some RGB images of the category *urban*, generated with the proposed stacked conditional GAN network, are shown in Fig. 4.14 for qualitative evaluation. The results of the category *oldbuilding* generated with the proposed stacked conditional GAN network are shown in Fig. 4.15 also for qualitative evaluation. As can

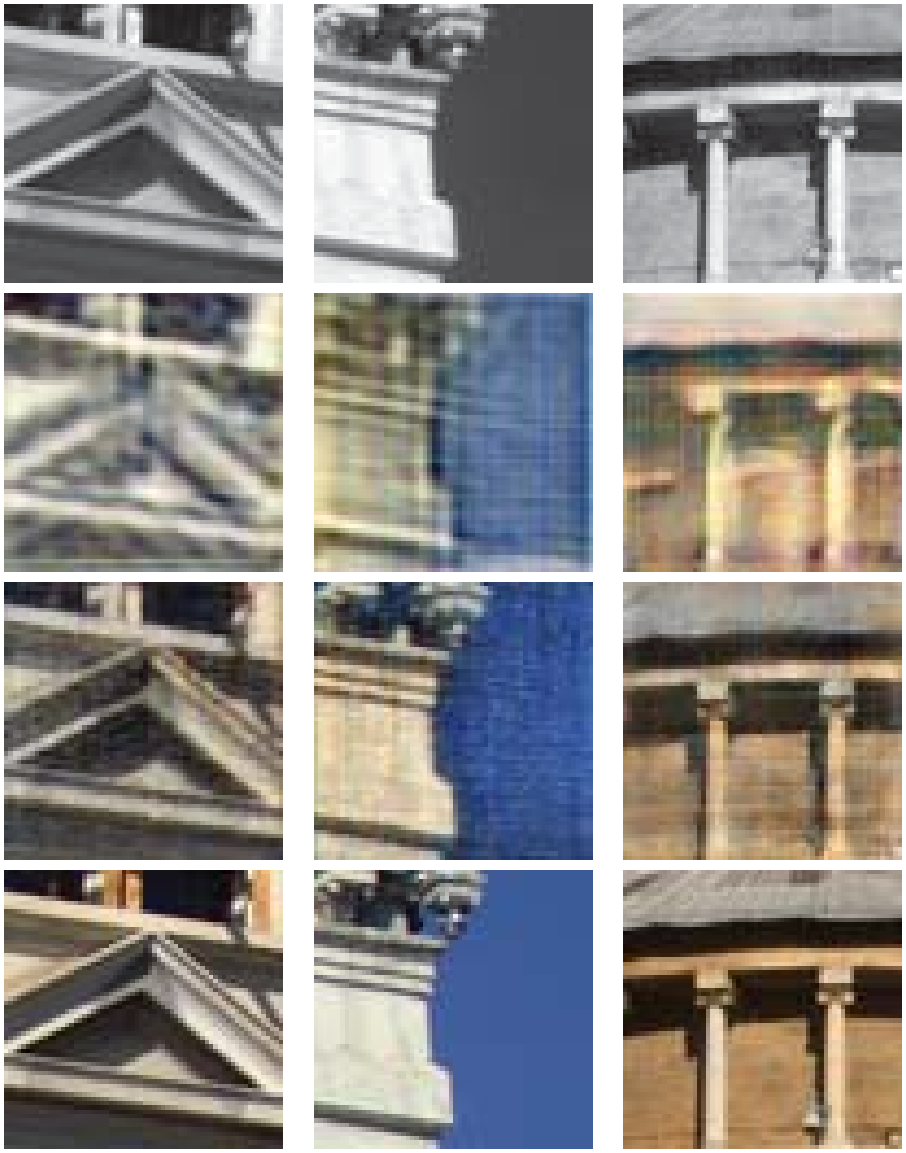


Figure 4.13 – Results of colorization: ($1st\ row$) NIR patches from the *oldbuilding category*; ($2nd\ row$) Results from the first approach flat) ($3rd\ row$) Results from the third proposed approach proposed (CDCGAN network); ($4th\ row$) Ground truth images.

be seen, in both cases, the proposed stacked conditional GAN, with a loss function based on the weighted sum of the three terms (see equation 4.4), produces images quite similar to the provided original images (*ground truth*). Despite these good results, there are some difficult cases in which the fourth proposed approach fails to color NIR images accurately. Fig. 4.16 and 4.17 show some illustrations of the cases where the desired results were not obtained. It should

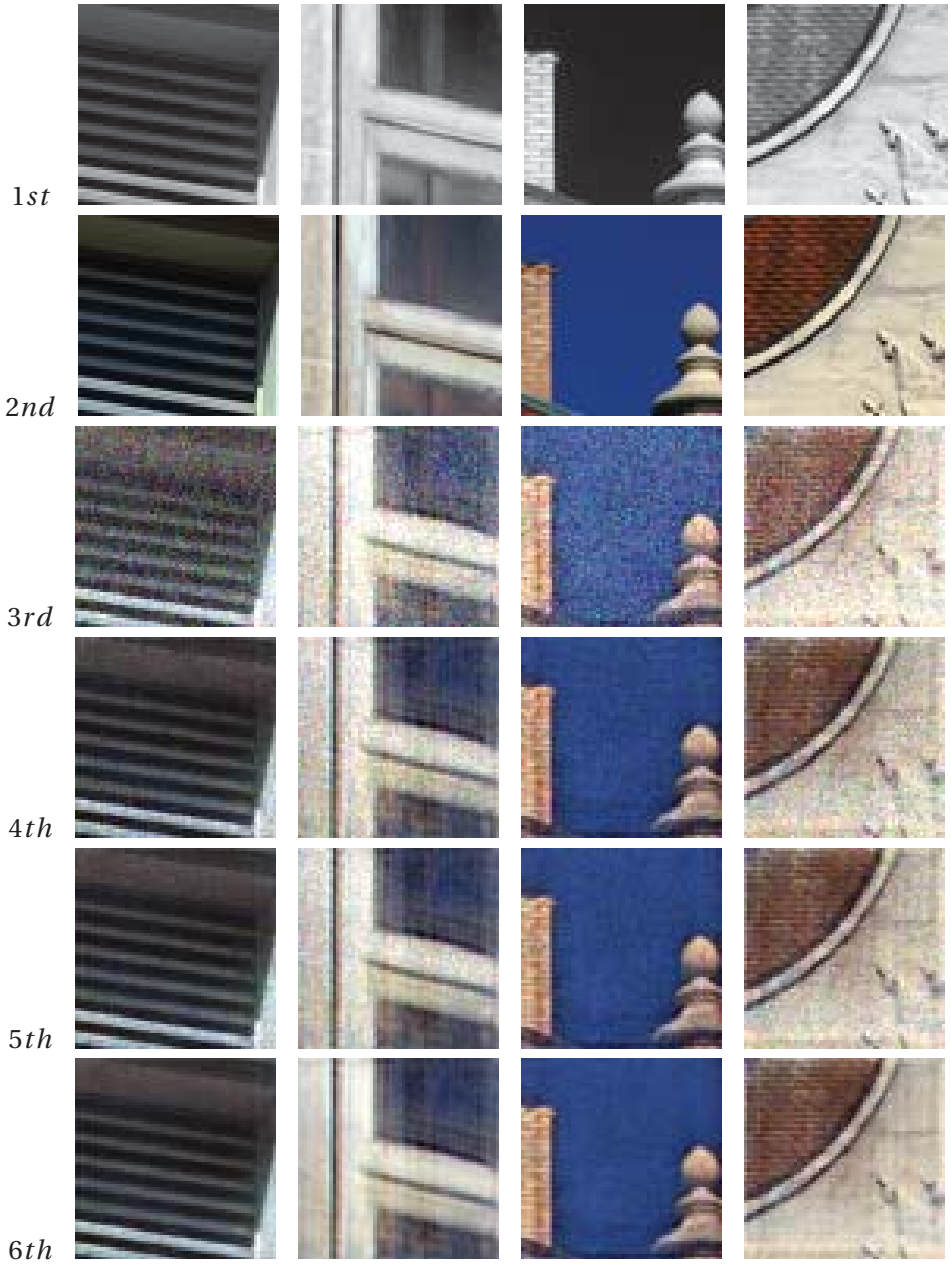


Figure 4.14 – Results from *urban*: *(1st row)* NIR patches; *(2nd row)* Ground truth images; *(3rd row)* Results from the third proposed approach (CDCGAN); *(4th row)* RGB representation obtained with the fourth proposed approach, (loss Function: $\mathscr{L}_{Adversarial} + \mathscr{L}_{Intensity}$); *(5th row)* RGB representation obtained with the fourth proposed approach, (loss function: $\mathscr{L}_{Adversarial} + \mathscr{L}_{SSIM}$); *(6th row)* RGB obtained with the fourth proposed approach, (loss function: $\mathscr{L}_{final}$).

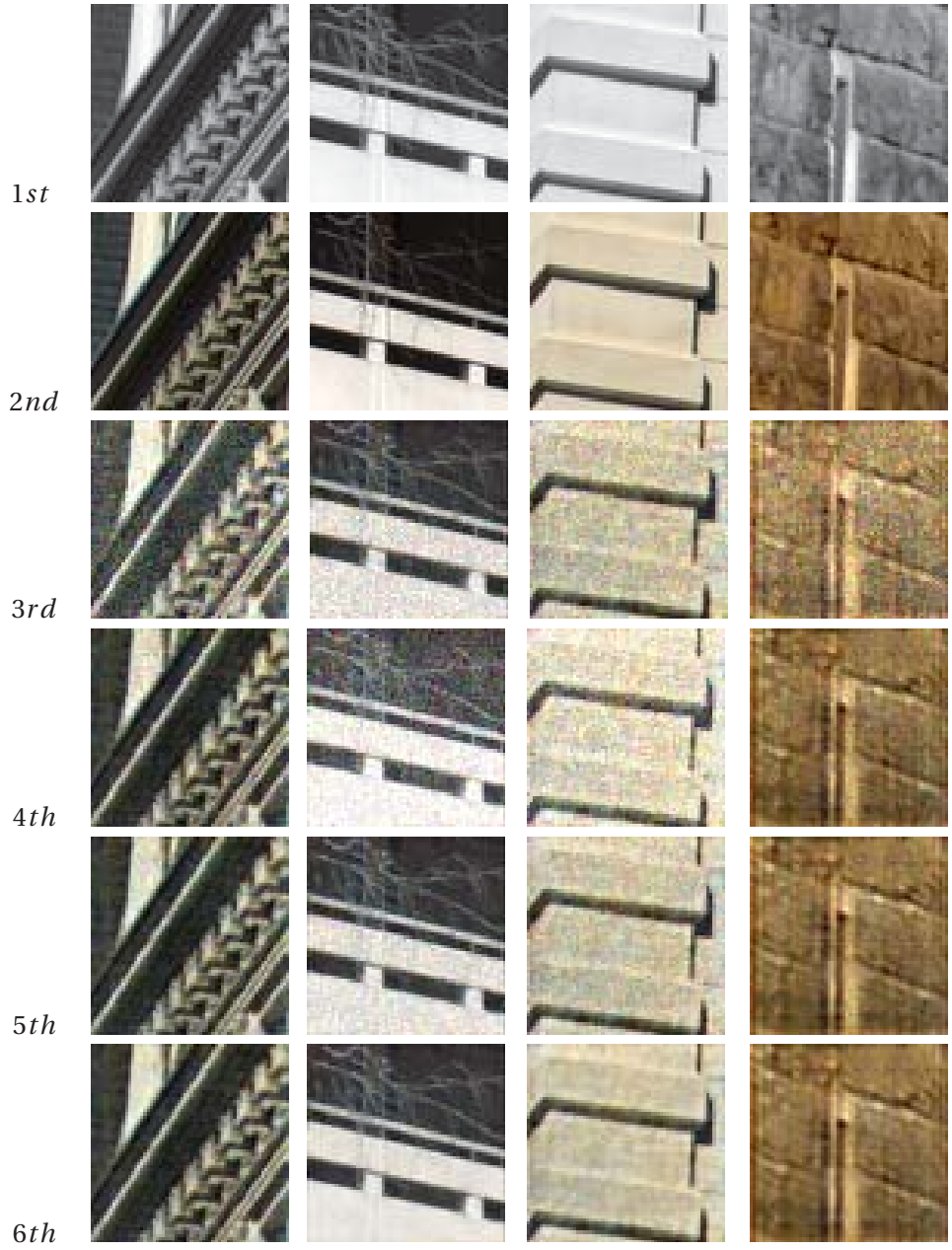


Figure 4.15 – Results from *oldbuilding*: *(1st row)* NIR image patches; *(2nd row)* Real images (*ground truth*); *(3rd row)* Results from the third proposed approach (CDCGAN); *(4th row)* RGB representations obtained with the fourth proposed approach, (loss Function: $\mathscr{L}_{Adversarial} + \mathscr{L}_{Intensity}$); *(5th row)* RGB representations obtained with the fourth proposed approach, (loss Function: $\mathscr{L}_{Adversarial} + \mathscr{L}_{SSIM}$); *(6th row)* RGB representations obtained with the fourth proposed approach, (loss Function:$\mathscr{L}_{final}$).

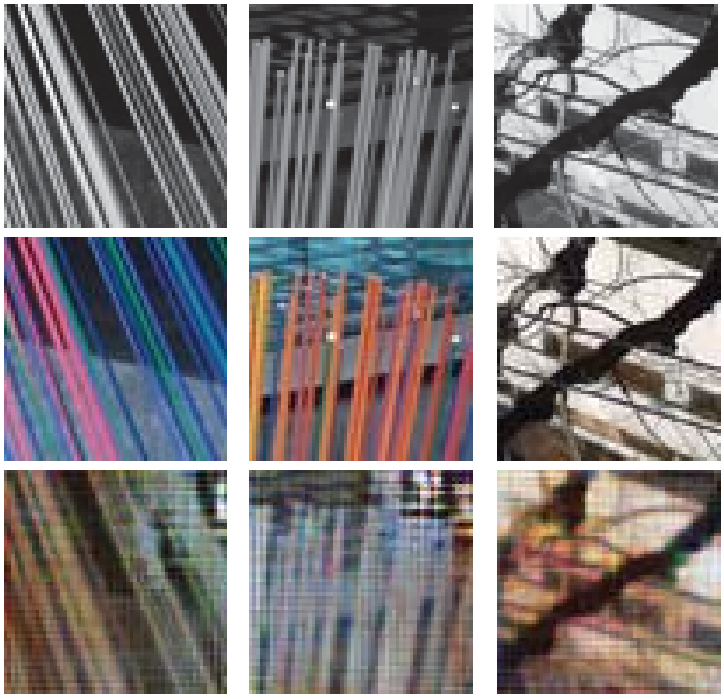


Figure 4.16 – Difficult cases in *urban* Category: *(1st row)* NIR patches; *(2nd row)* Real images (*ground truth*); *(3rd row)* Bad results obtained with the fourth proposed approach (loss Function: $\mathscr{L}_{final}$).

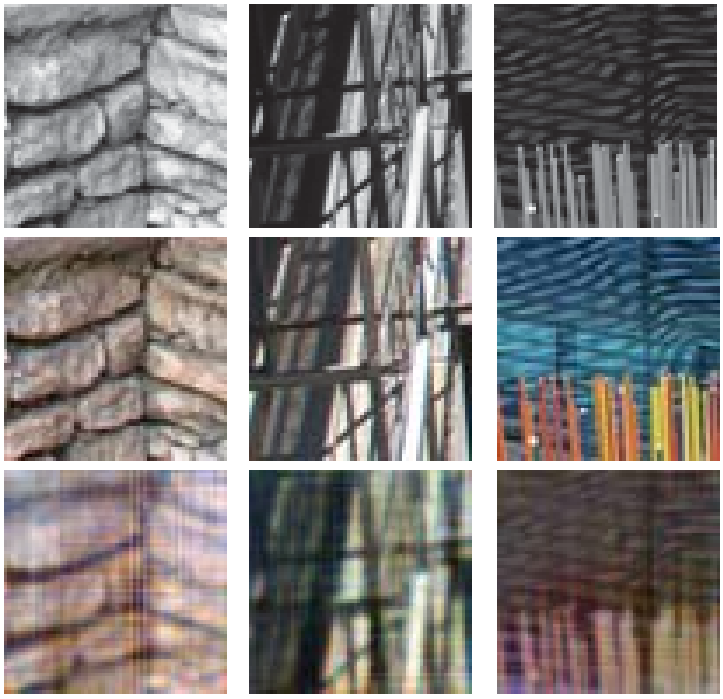


Figure 4.17 – Difficult cases in *oldbuilding* Category: *(1st row)* NIR patches; *(2nd row)* Real images (*ground truth*); *(3rd row)* Bad results obtained with the fourth proposed approach (loss Function: $\mathscr{L}_{final}$).

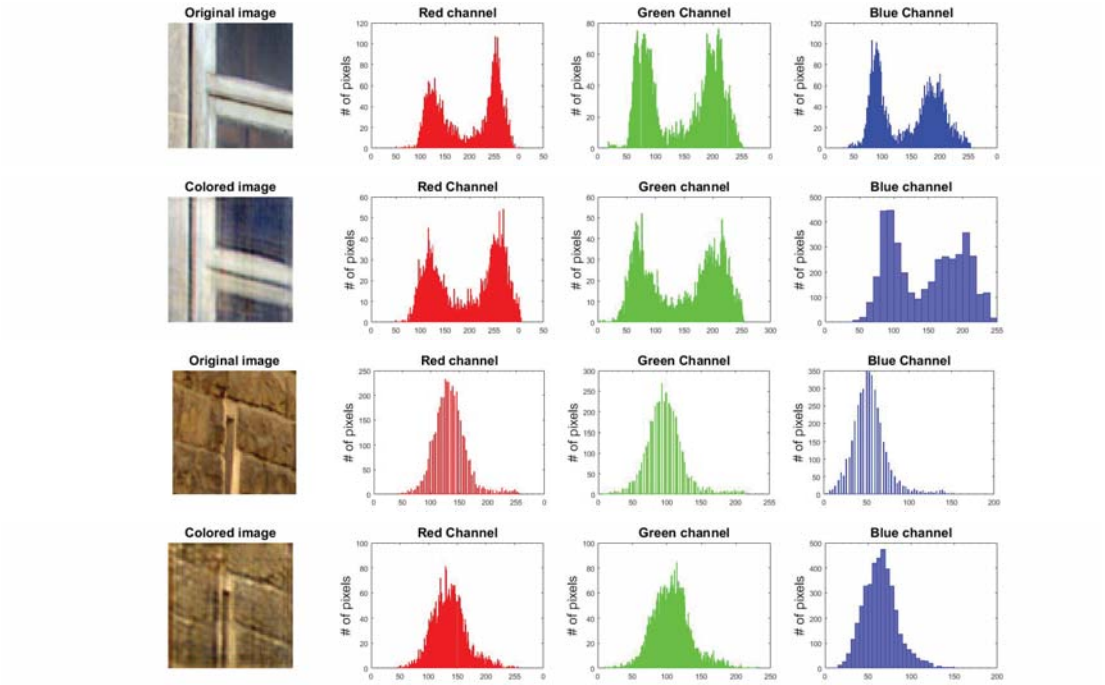
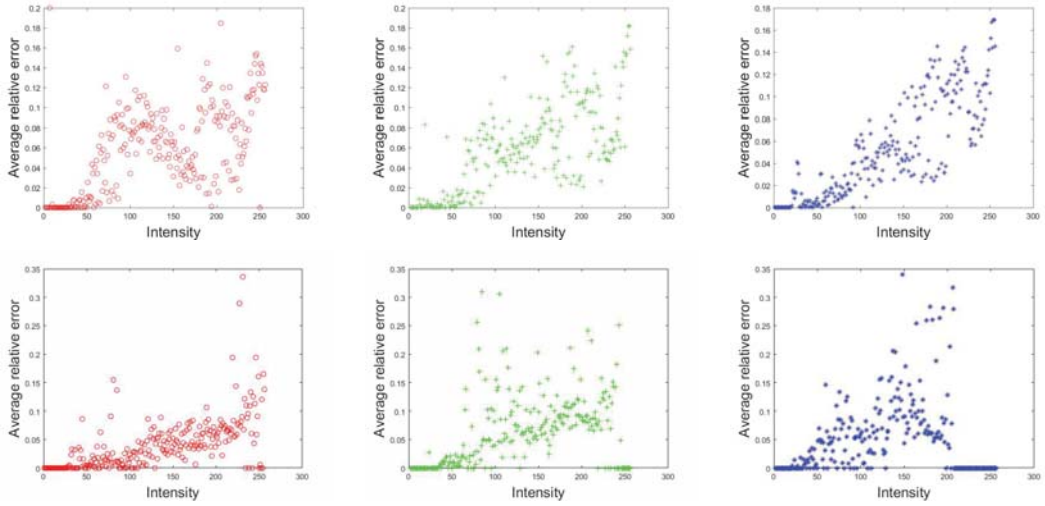


Figure 4.18 – Illustration of colorized images and their corresponding color distribution histograms: (*up*) *urban* Category; (*down*) *oldbuilding* Category.

be taken into account that in these cases there are large changes in the color of the texture, which makes the colorization of the images difficult. It should be noted that these particular cases are not common in the categories *urban* and *oldbuilding*, where the color of the surface changes smoothly. Figure 4.18 shows some examples corresponding to the colored *urban* and *oldbuilding* categories together with their corresponding color distribution histogram per channel, both for the category *urban* and for *oldbuilding*. The Figure 4.19 shows the relative error distribution presented for each channel per category (*urban* and *oldbuilding*. This error is calculated for each intensity value ([0.255]) as follows:

$$\mathscr{E}_{Ic} = \sum_{j=0}^{255} \frac{\sum_{k=0}^{Tp} |Ic_{(I_{j_{(k)}})} - I_{j_{(k)}}|}{I_j T_p}, \tag{4.6}$$

where I is the intensity value of the original image (*ground truth*), as mentioned earlier, this variable takes values in the range $j = \{0, 255\}$; Ic is the corresponding intensity value of the element $I_{j_{(k)}}$ in the colorized image; T_p is the number of times a certain intensity level I appears in the *ground truth*. The average error obtained in the numerator of Eq. (4.6) is divided by the value of I to obtain a measure relative to the intensity value being considered.



(red channel) *(green channel)* *(blue channel)*

Figure 4.19 – Illustration of the average relative error distribution of colored images: (up) *urban* Category; $(down)$ *oldbuilding Category*.

Figure 4.19 shows the average relative errors of the differences found in the process of colorization of the images of the *urban* category, it can be observed that the greatest difference has been presented in the red channel having higher average relative error values, for pixels, having a peak of up to 18.4% in pixels near zero; it is followed by the green channel with a peak of up to 18% for pixels near the pixel 255, however, presents average relative errors for each of the pixels smaller than those presented in the red channel. Finally, the blue channel has the lowest average relative error in the distribution of pixels, reaching a peak of up to 16.9%. However, if we analyze the colorization of the images belonging to the category *oldbuilding*, see Fig. 4.19, it can be seen that the greatest error has occurred in the green channel having a peak of up to 30.5% in pixels close to 100, followed by the blue channel, which despite having a peak of 34% in pixels close to 255, the distribution of average relative errors are smaller than those presented in the green channel. Finally, it is observed that the red channel has the lowest distribution of average relative errors, presenting a peak of up to 33.5%, for pixels close to pixel 255, it is observed that the values of the relative errors for each of the pixels are lower than those presented in the green and blue channels.

4.5 Conclusions

This chapter summarizes the research and work done to perform the colorization of images of the near infrared spectrum, four learning architecture schemes based on adversarial generative networks have been implemented, which have been quantitatively evaluated using different metrics. Results show that the model of Stacked Conditional Adversarial Generative Network (SC-GAN) for NIR image colorization is the best one; this fourth approach generates good

quality color images of different scenes (i.e., content, texture, shades, color intensity). Future works will be focused on the evaluation of other network architectures, like variational auto-encoders, cycle-consistent adversarial networks, metalearning and reinforcement learning, which have shown appealing results in recent works. The fourth colorization approach will be evaluated with different loss functions to improve and accelerate the training process. Finally, it will be considered to increase the number of images for the training process or the number of layers or feature maps to increase the diversity of colors.

Chapter 5

Normalized Difference Vegetation Index Estimation

This chapter proposes novel approaches to estimate the Normalized Difference Vegetation Index (NDVI) using Generative Adversarial Network architectures. The first scheme is based on a supervised learning architecture, where a synthetic NDVI index is obtained from NIR images paired with their corresponding ground truth using a standard GAN network. Two different approaches have been implemented, a single and a triplet learning architecture. The second scheme is an unsupervised model based on an image to image translation applying CyclicGAN. Two approaches have been designed, the first one is an image translation between a RGB (grayscale) and NIR unpaired images to obtain a synthetic NIR. Once this NIR image is generated, the same image is used to obtain the NDVI index. The second unsupervised approach proposed in this chapter, using also a CyclicGAN, is an architecture where a synthetic NDVI index is obtained through an image translation between a red channel and NDVI unpaired images. The experimental results obtained with the last unsupervised scheme show the validity of the implemented model. Additionally, this second proposed unsupervised approach has been compared with the state of the art, showing better results.

5.1 Introduction

Computer vision is a technology that, combined with machine learning and remote sensing, allows computers to understand and estimate the quantity, quality and condition of crops. These estimations can be made based on the intensity of radiation reflected by certain bands of the electromagnetic spectrum. This information can be captured by using remote sensing technology, planes or unmanned aerial vehicles, which can incorporate sensors sensitive to near infrared (NIR), in addition to the visible spectrum, in order to simultaneously acquire images of the same scene from different spectra. Then, with such cross-spectral information, more efficient solutions can be implemented to help farmers with their crops to apply more efficient growth methods, increase yields and profits [25], [28].

A vegetation index is a single value that quantifies vegetation biomass and/or the plant health for each pixel in a remote sensing image. The index could be computed using several

spectral bands that are sensitive to the plant biomass and health. The math associated with calculating a vegetation index is derived from the physics of light reflection and absorption across bands. For instance, it is known that healthy vegetation reflects light strongly in the near infrared band and less strongly in the visible portion of the spectrum. Thus, a high ratio between the light reflected in the near infrared and light reflected in the visible spectrum will represent areas that potentially have healthy vegetation. The more a plant absorbs visible sunlight (during the growing season), the more photosynthesis and more productive it is. Conversely, the less sunlight absorbs the plant, the less photosynthesis and less productive it is. Higher-end image processing techniques are proposed by [74], to investigate the strength of key spectral vegetation indexes for agricultural crop yield prediction using neural network in order to increase agricultural production [22].

Among the different indexes proposed in the literature, the Normalized Difference Vegetation Index (NDVI) is the most widely used [90]; NDVI is often used to monitor drought, forecast agricultural production, assist in forecasting fire zones and desert maps [97]. NDVI is preferable for global vegetation monitoring since it helps to compensate for changes in lighting conditions, surface slope exposure, and other external factors. In general, it is used to determine the condition, developmental stages and biomass of cultivated plants and to forecast their yields. This index is calculated as the ratio between the difference and sum of the reflectance in NIR and red regions:

$$NVDI = \frac{R_{\text{NIR}} - R_{\text{RED}}}{R_{\text{NIR}} + R_{\text{RED}}}, \tag{5.1}$$

where R_{NIR} is the reflectance of NIR radiation and R_{RED} is the reflectance of red channel radiation.

This index defines values from -1.0 to 1.0, basically representing greens, where negative values are mainly formed from clouds, water and snow, and values close to zero are primarily formed from rocks and bare soil. Very small values (0.1 or less) of the NDVI function correspond to empty areas of rocks, sand or snow. Moderate values (from 0.2 to 0.3) represent shrubs and meadows, while large values (from 0.6 to 0.8) indicate temperate and tropical forests [17] [52].

Another application of NDVI is to track changes to an ecosystem over time. Measuring the impacts of forestry on an ecosystem by calculating the change in vegetation index over time [77]. This is valuable for understanding the impacts of climate change at varying scales. Changes to the vegetation index will vary locally and regionally. This could direct conservation efforts to areas that are subject to greater changes, or aid in management and planning efforts.

Overall, NDVI is a standardized way to measure health in vegetation based on how the plant reflects the light at certain frequencies (some waves are absorbed and others are reflected). The pigment in plant leaves, chlorophyll (a health indicator) strongly absorbs visible light, and the cellular structure of the leaves strongly reflects near-infrared light. When the plant becomes dehydrated, sick, affected with disease, etc., the spongy layer deteriorates, and the plant

absorbs more of the near-infrared light, rather than reflecting it. Thus, observing how NIR changes compared to red light provides an accurate indication of the presence of chlorophyll, which correlates with plant health. Recent studies have demonstrated the usefulness of optical indexes from hyperspectral or cross-spectral remote sensing in the assessment of vegetation biophysical variables both in forestry and agriculture [114], [2].

In order to obtain the NDVI vegetation index, registered images of the visible and infrared spectrum are needed to compute the value according to eq. (5.1). In other words, sensors from both spectra are needed to acquire the images at the same time from the same scene. This requirement makes any solution that deals with the use of this vegetation index more challenging, because depends on whether aligned image pairs are available for training, testing and validation. Cross/multi-spectral computer vision approaches provide solutions to multiple complex problems. However, as mentioned above, different preprocessing steps need to be implemented before computing these solutions or it is necessary to invest in cameras sensitive to the near infrared spectrum.

This chapter presents two schemes, the first one a supervised approach, where two architectures have been implemented, requiring both models to work with paired images, using a standard GAN network. The first model uses a triple level of learning using as an input a NIR image concatenated with some Gaussian noise. The second supervised model was designed just using information from a single spectral band, the red channel of an image concatenated with some Gaussian noise, using a single level of learning. Although interesting results have been obtained, the weakness point of these approaches lies in the need of having NIR images for the training process, which are not that much common like visible spectrum images and also the limitation that models can only be trained with paired images.

To overcome these limitations, a second scheme based on an unsupervised approach has been implemented to perform an image-to-image translation between unpaired images. Two architectures have been implemented using a CyclicGan. The first unsupervised approach has been recently presented in where a synthetic near infrared image is generated from a CyclicGAN and then this synthetic NIR image has been used to calculate the vegetation index to avoid the dependence on near infrared image. The second unsupervised approach implemented in this chapter does not depend on paired images for the training and testing process. The approach proposed is a learning model based on a Cyclic Generative Adversarial Network trained with a large dataset, one for visible spectrum and the other is the computed NDVI image. Each one is fed into a CyclicGAN to perform the image domain translation. Additionally, a modified residual network (RESNET) architecture is used to go deeper without degradation in the accuracy and error rate. A least square GAN loss function is also included, to help to stabilize the training, and to preserve details of the estimated NDVI index. The chapter is organized as follows. The proposed approaches are detailed in Section 5.2. The experimental results with a set of real images are presented in Section 5.3. Finally, conclusions are given in Section 5.4.

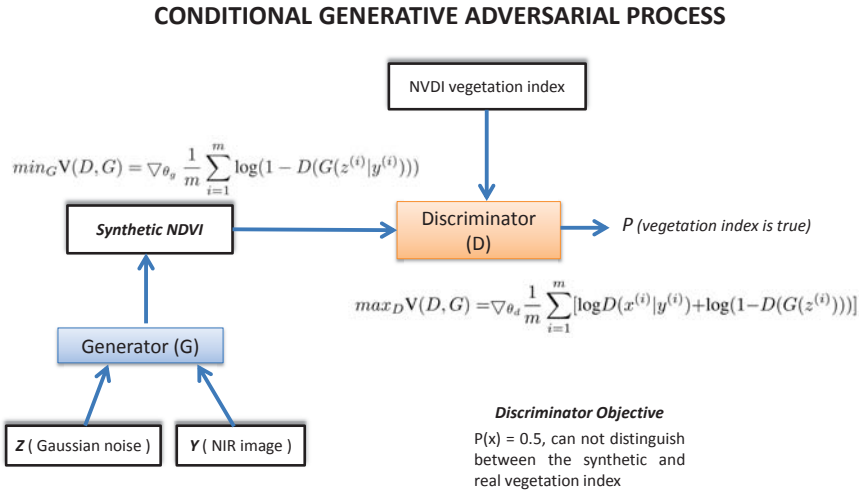


Figure 5.1 – Conditional generative adversarial process implemented on the current chapter to estimate NDVI Vegetation Index.

5.2 Proposed Approaches NDVI Vegetation Index Estimation

This section introduces the approaches implemented to estimate the NDVI vegetation index, using supervised and unsupervised schemes. For simplicity, NDVI indexes are represented as image values, so here in after the terms NDVI indexes and NDVI images will be indistintly used. The details of the implemented architectures in each scheme are given below.

5.2.1 Supervised Approaches

Two supervised approaches based on GAN networks have been designed for NDVI vegetation indices estimation. The first one is a triplet level of learning model with only adversarial loss using as an input near infrared images while the second approach is a single level learning model with multiple loss functions using the red channel of a RGB image, the approaches will be explained in detail below.

5.2.1.1 Proposed Architectures

The first approach proposed for NDVI vegetation index estimation uses a similar architecture like the one presented in the previous chapter for NIR colorization, where the usage of a conditional adversarial generative learning network has been proposed. A traditional scheme of layers in a deep network is used. In the current chapter, the usage of a Conditional GAN model is evaluated in three different schemes: Flat, Siamese and Triplet. These models have presented good performance to solve problems like colorization, segmentation, classification,

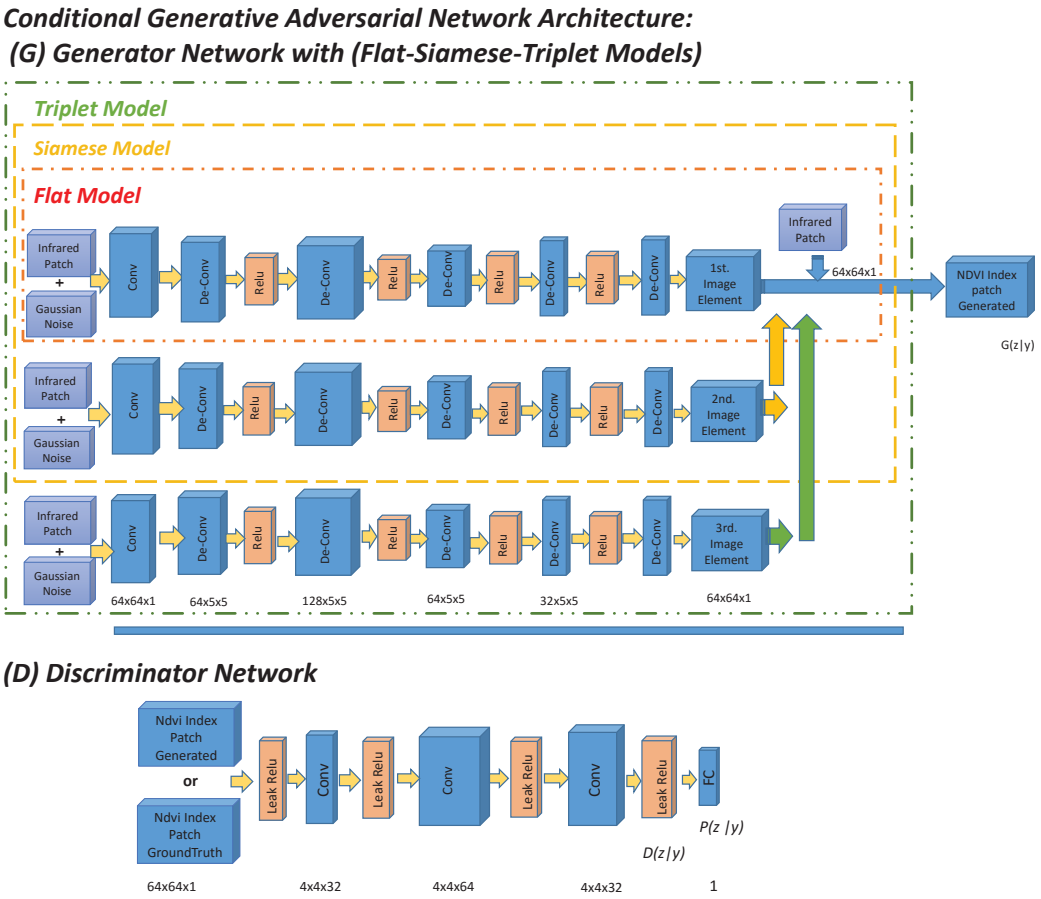


Figure 5.2 – GAN architecture of the first supervised approach for NDVI vegetation index estimation; (*top*) The three models (flat, siamese and triplet) evaluated as generator networks; (*bottom*) The discriminator network.

similarity learning, object recognition, etc. Based on the results that have been obtained on this type of solutions, where improvements in accuracy and performance have been obtained, the usage of a learning model that allows the mapping representation of a vegetation index based on cross-spectral images has been proposed.

Therefore, the model will receive as an input a near infrared patch with a Gaussian noise added in each element of the learning model to generate the necessary variability of the vegetation index patches, to be able to generalize the learning process. A $L1$ regularization term has been added on a single layer in order to prevent the coefficients to fit so perfectly to overfit, which can improve the generalization capability of the model. Figure 5.1 depicts the Conditional GAN model proposed in the current chapter. See Section 2.5 for more detail about GAN networks.

Like in the NIR image colorization problem, a conditional triplet GAN network based architecture is selected due to several reasons: i) the learning process is conditioned on NIR images from the source domain; ii) its fast generalization of the model; iii) reduces the probability of getting stuck in local minimum; and iv) learn a loss that adapts to the data. The network is intended to learn to generate new samples from an unknown probability distribution. As mentioned above, in this case, the generator network has been implemented in three different schemes: Flat, Siamese and Triplet, which are evaluated in the experimental result section. Figure 5.2 presents an illustration of the GAN network with the three generator schemes. In all the cases, at the output of the generator network, the vegetation index is obtained. This vegetation index will be validated by the discriminative network, which will evaluate the probability that the generated image (vegetation index in grayscale), is similar to the real one that is used as ground truth.

Additionally, in the generator model, in order to obtain a better image representation, the CGAN framework is reformulated for a conditional generative image modeling tuple. In other words, the generative model $G(z;\theta_g)$ is trained from a near infrared image plus Gaussian noise, in order to produce a NDVI vegetation index image; additionally, a discriminative model $D(z;\theta_d)$ is trained to assign the correct label to the generated NDVI image, according to the provided real NDVI image, which is used as a ground truth. Variables (θ_g) and (θ_d) represent the weighting values for the generative and discriminative networks.

The second approach proposed for NDVI index vegetation estimation uses a similar architecture like the first approach proposed in the NIR image colorization chapter, where a single layer adversarial generative learning network has been proposed. In this case, the network is fed with the red channel of the given RGB image, instead of the NIR image into the Conditional GAN flat model. This network model has been used, because it presented good performance to solve problems like colorization, dehazing, enhancement, object recognition, etc. Figure 5.3 presents an illustration of the GAN network used in this approach. In all the cases, at the output of the generator network, the vegetation index is obtained. This vegetation index will be validated by the discriminative network, which will evaluate the probability that the generated image (vegetation index), is similar to the real one used as a ground truth.

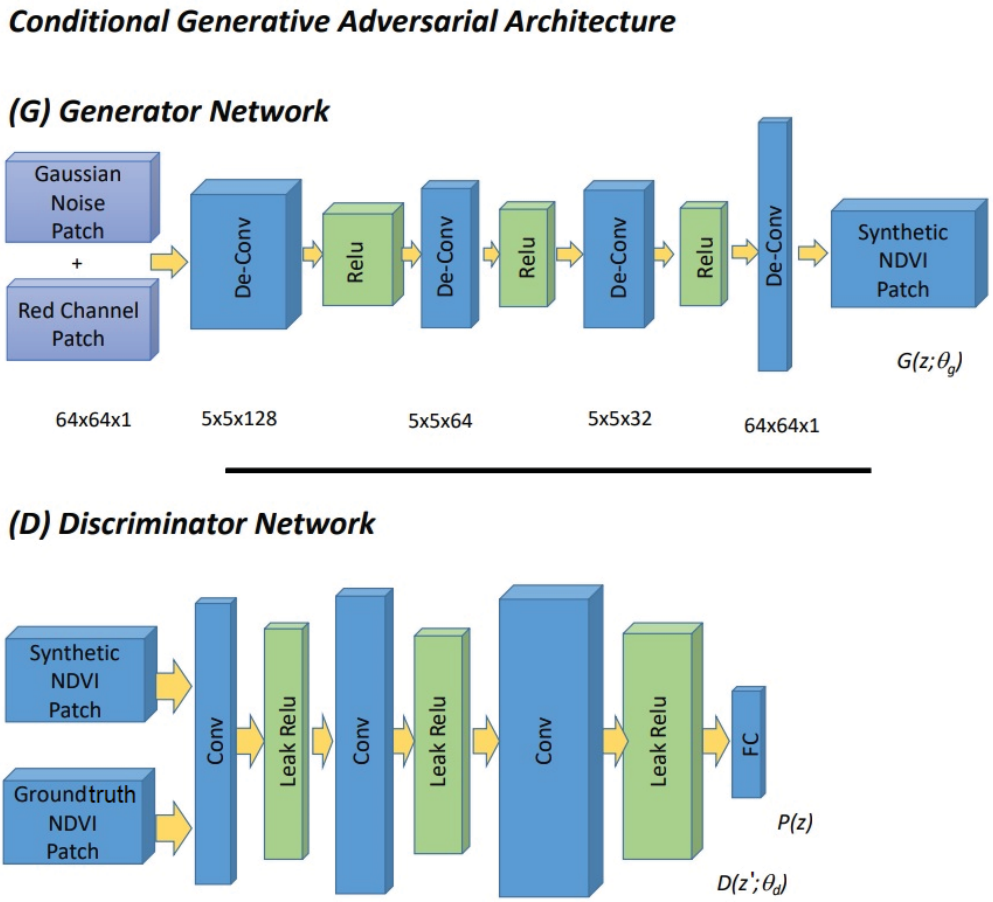


Figure 5.3 – GAN architecture of the second supervised approach for NDVI vegetation index estimation. A single level layer model (flat) evaluated as generator network; bottom the discriminator network.

Additionally, in the generator model, in order to obtain a better image representation, the GAN framework is reformulated for a conditional generative image modeling tuple. In other words, the generative model $G(z|nir;\theta_g)$ is trained from a red channel of a RGB image plus Gaussian noise, in order to produce a NDVI vegetation index image; additionally, a discriminative model $D(z';\theta_d)$ is trained to assign the correct label to the generated NDVI image, according to the provided real NDVI image, which is used as a ground truth. Variables (θ_g) and (θ_d) represent the weighting values for the generative and discriminative networks.

Based on the results that have been obtained with the first supervised approach, the usage of a learning model that allows the generation of a vegetation index based on a single channel of RBG images (the red channel) has been proposed. It should be highlighted that, on the contrary to the previous approach that estimates NDVI from NIR information, with the second supervised approach the NDVI is estimated only from the visible spectrum information (the red channel) in order to obtain improvements in the accuracy and performance of the model. Visible spectrum information (RGB images) is more accessible than NIR information. As mentioned before, the model will receive as an input a patch corresponding to the red channel of a RGB image. Gaussian noise is added to each patch of the learning architecture to increase the variability of the generated images during the training process, decreasing the time of the convergence and generalization. A $L1$ regularization term has been added on each layer of the model in order to prevent the coefficients to overfit, which make the network learns small weights to minimize the loss, and maximize the distribution of model outputs, and improve the generalization capability of the model.

5.2.1.2 Loss Functions

The first supervised model has been defined with an Adversarial loss defined as:

$$\mathscr{L}_{Adversarial_{NIR}} = -\sum_i logD(G_w(I_{NDVI_{est}}|NIR),(I_{NDVI_{real}}|NIR)), \tag{5.2}$$

where D and G_w are the discriminator and generator of the real NDVI $I_{NDVI_{real}}$ and estimated NDVI vegetation index $I_{NDVI_{est}}$ images conditioned by the near infrared image NIR in GAN network.

The second supervised model has been defined with a multi-term loss $(\mathscr{L})$ formed by the combination of the adversarial loss plus the intensity loss (MSE) and the structural loss (SSIM). These combined losses have been defined to avoid the usage of only a pixel-wise loss (PL) to measure the mismatch between a generated image and its corresponding ground truth image. This multi-term loss function is better designed to human perceptual criteria of image quality, which is detailed next. The Adversarial loss is designed to minimize the cross-entropy to improve the texture loss :

$$\mathscr{L}_{Adversarial_{red-channel}} = -\sum_i logD(G_w(I_{NDVI_{est}}|red-channel),(I_{NDVI_{real}}|red-channel), \tag{5.3}$$

where D and G_w are the discriminator and generator of the real NDVI $I_{NDVI_{real}}$ and estimated NDVI vegetation index $I_{NDVI_{est}}$ images conditioned by the red channel of a RGB image in GAN network.

The Intensity loss is defined as:

$$\mathscr{L}_{Intensity} = \frac{1}{NM}\sum_{i=1}^{N}\sum_{j=1}^{M}(NDVIe_{i,j} - NDVIg_{i,j})^2, \tag{5.4}$$

where $NDVIe_{i,j}$ is the vegetation index estimated by the network and $NDVIg_{i,j}$ is the ground truth vegetation index and $N \times M$ is the size of the patches. This loss measures the difference in intensity of the pixels between the images without considering texture and content comparisons. Additionally, this loss penalizes larger errors, but is more tolerant of small errors, without considering the specific structure in the image.

To address the limitations of the simple intensity loss function, the usage of a reference-based measure is proposed. One of the reference-based indexes is the Structural Similarity Index (SSIM) [111], which evaluates images accounting for the fact that the human visual perception system is sensitive to changes in the local structure; the idea behind this loss function is to help the learning model to produce a visually improved image. The Structural loss for a patch P is defined as:

$$\mathscr{L}_{SSIM}(P) = \frac{1}{N}\sum_{p=1}^{P} 1 - SSIM(p), \tag{5.5}$$

where SSIM(p) is the Structural Similarity Index (see [111] for more details) centered in pixel p of the patch P. The Final loss ($\mathscr{L}$) used in this second supervised model is the accumulative sum of the individual Adversarial, Intensity and Structural loss functions:

$$\mathscr{L}_{Final} = 0.60\mathscr{L}_{Adversarial_{red-channel}} + 0.25\mathscr{L}_{Intensity} + 0.15\mathscr{L}_{SSIM}. \tag{5.6}$$

5.2.2 Unsupervised Approach

On the contrary to the previous approaches, in this section, NDVI vegetation index is estimated using the image translation technique, through an unsupervised approach i.e, cyclic generative adversarial network (CyclicGAN). The architectures proposed in this section are based on the work presented on [128], a previous work that presents an unpaired image to image translation, through a CyclicGAN. This type of network permits domain style transfer which is a convenient method for image-to-image translation problems. It is not necessary to have a set of input images that capture the scene at the same time and place from different spectra, being this advantage, which is applied in the unsupervised learning models proposed in the present

chapter. Obtain this kind of set of images could be time-consuming and quite difficult based on what type of domain style the image dataset you are trying to translate between. In [48] the authors present a general-purpose image-to-image translation model in a supervised manner by using conditional adversarial; these networks not only learn the mapping from an input image to output image, also learn a loss function to train the corresponding mapping. However, the aforementioned researches are limited and still dependent on some kind of correlated labeling. Two approaches using CyclicGAN have been proposed, the first one is a domain translation from grayscale image to NIR, to generate synthetic NIR image and then use it to compute the NDVI vegetation index; and the second approach is image translation between the red channel of a RGB image to NDVI vegetation index. See Section 2.5 for more details of CyclicGAN.

5.2.2.1 Instance Normalization

GANs are a framework in which two networks compete with each other. The two networks, the generator G and the discriminator D, are both represented by function approximators. Moreover, given a training data, the generator creates samples as an attempt to mimic the ones from the same probability distribution as the training set. The discriminator, on the other hand, is a common binary classifier. It has two main objectives. First, it categorizes whether its received input comes from the true data distribution (ground truth) or the generator distribution. See Section 2.5 for more details.

During training, D receives half of the time images from the training set, and the other half, images from the generator network to maximize the probability of assigning the correct class label to both: real images (ground truth) and synthetic samples (images from the generator). To reach that the training process between the generator G and discriminator D finds an equilibrium - the Nash equilibrium [91] [34]. These GAN networks can be unstable during the training process. For that reason, to reduce instability, instance normalization has been implemented, see section 2.8 for a detailed information.

The architecture uses this normalization, applied in feed-forward style transfer, to improve the quality of the NDVI generated vegetation index.

5.2.2.2 Residual Networks (ResNET)

Deep neural networks have evolved from simple to very complex architectures depending on the type of problem to be solved, whether these are classification, segmentation, recognition, identification, etc. According to [39], deep networks naturally integrate low/mid/high-level features and classifiers in an end-to-end multilayer way, and the “levels” of features can be enriched by the number of stacked layers (depth). When deeper networks are able to start converging, a degradation problem could appear. With the network depth increasing, accuracy gets saturated and then degrades rapidly, this behavior of degradation indicates that every neural model is unique and not easy to optimize. There exists a solution by construction to

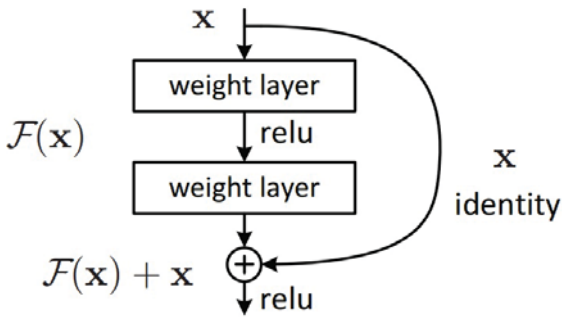


Figure 5.4 – Residual block from [39].

the deeper model: the added layers are identity mapping, and the other layers are copied from the learned shallower model. The existence of this constructed solution indicates that a deeper model should produce no higher training error than its shallower counterpart. In [39] it is presented a deep residual learning framework, where instead of waiting for the stacked layers to fit directly to a desired underlying mapping, these layers are allowed to fit the residual mapping. Formally, denoting the desired underlying mapping as $H(x)$. It is allowed that the stacked nonlinear layers fit another mapping of $F(x) := H(x) - x$. The original mapping is recast into $F(x) + x$. The authors hypothesize that it is easier to optimize the residual mapping than to optimize the original, unreferenced mapping. To the extreme, if an identity mapping were optimal, it would be easier to push the residual to zero than to fit an identity mapping by a stack of nonlinear layers.

The formulation of *F(x) + x* can be realized by feedforward neural networks with "shortcut connections", to perform identity mapping, and their outputs are added to the outputs of the stacked layers, see Fig. 5.4. Also an identity shortcut connection add neither extra parameter nor computational complexity. The entire network can still be trained end-to-end by stochastic gradient descent (SGD) with backpropagation.

5.2.2.3 Proposed Architecture using RGB: grayscale - NIR Translation

This section presents the approach proposed for NDVI vegetation index estimation just with a single image from the visible spectrum. As mentioned above, it uses a similar architecture like the one proposed on [128], a recent work for unpaired image to image translation, where the usage of a cycle generative adversarial network has been proposed. CyclicGANs is a convenient method for image-to-image translation problems, such as style transfer, because it just relies on an unconstrained input set and output set rather than specific corresponding input/output pairs. This could be time-consuming, unfeasible, or even impossible based on what two image types are trying to translate between. Another approach presented in [48] has shown results synthesizing photos from label maps, reconstructing objects from edge maps, but still dependent on some kind of correlated labeling.

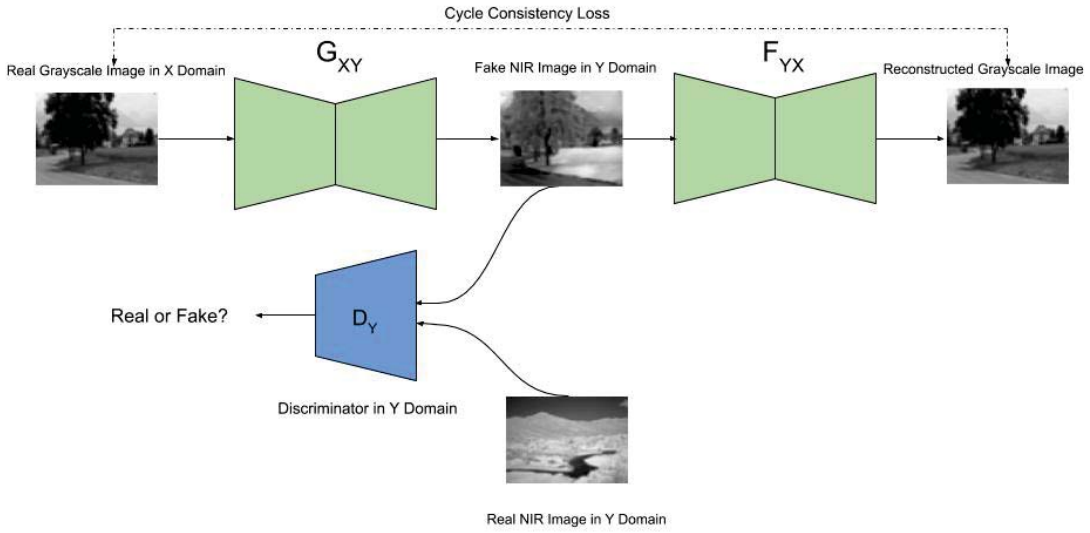


Figure 5.5 – Cycle generative adversarial model G: $X(grayscale) \rightarrow Y(NIR)$ and its discriminator D_y.

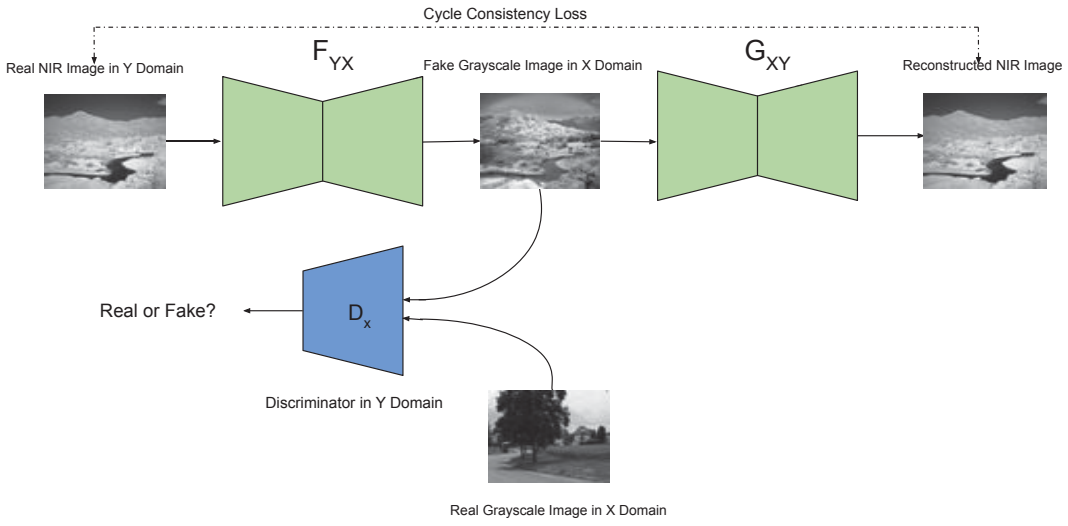


Figure 5.6 – Cycle generative adversarial model F: $Y(NIR) \rightarrow X(grayscale)$ and its discriminator D_x.

The proposed architecture is based on the approach presented in [128] about cycle consistent learning and loss functions; in the proposed work it is used to estimate the synthetic NIR images. The proposed model can learn to translate the images between the visible spectrum to their corresponding NIR spectrum, without the need to have accurately registered RGB/NIR pairs. This allows us to use these NIR synthetic images in the calculation of the NDVI vegetation index and to be able to use them in solutions oriented to solve problems related to the state of the crops and their corresponding level of productivity in the crops. Another advantage of being able to count on the synthetic images of the NIR spectrum is that, undoubtedly, the costs of the solutions are decreased since there is no need to buy acquisition devices sensitive to that electromagnetic spectrum. Additionally, the proposed architecture uses Residual Network (ResNET) [39] to perform the image transformation from one spectrum to another. It avoids the vanishing gradient problem, as the gradient is back-propagated to earlier layers, repeated multiplication may make the gradient infinitely small. As a result, as the network goes deeper, its performance can get saturated or even starts degrading rapidly. To avoid all these problems, the generator and discriminator have been implemented to propagate larger gradients to initial layers and these layers also could learn as fast as the final layers, giving us the ability to train deeper networks, more details of ResNet are given in Section 5.2.2.2

The "identity shortcut connection" enable skips of one or more layers, to ensure properties of NIR images of previous layers are available for later layers as well, so that their outputs do not deviate much from original grayscale image used as an input, otherwise the characteristics of original images will not be retained in the output and results will be very unreal. Figure 5.5 and 5.6 depicts the CyclicGAN model proposed in the current chapter. As can be appreciated in the illustrations, the CyclicGan architecture to generate NIR synthetic images is composed of two generators G, F and two discriminators D_x, D_y. In order to generate a NIR synthetic image, the architecture takes advantage of the joint of cycle-consistency and least square losses [67] in addition to the usual discriminator and generator losses. The results of the experiments have shown that these loss functions demand that the model maintains textural information of the visible and NIR images and generate uniform synthetic outputs, see Section 5.3.4 for details on results.

5.2.2.4 Proposed Architecture using Red Channel - NDVI Translation

This section focuses on the estimation of NDVI vegetation index using the red channel of RGB images. The proposed model can learn to translate the images between the visible spectrum (red channel) to their corresponding NDVI indexes, using an unpaired dataset. This allows us to transform from RGB to NDVI and to be able to use them in solutions oriented to solve problems related to the state of the crops and their corresponding level of productivity. Another advantage of the proposed network to generate the synthetic NDVI images is that, undoubtedly, the costs of agricultural solutions may decrease, since there would be no need to acquire sensors sensitive to NIR spectra together with all the cost associated with the synchronized image acquisition and registration.

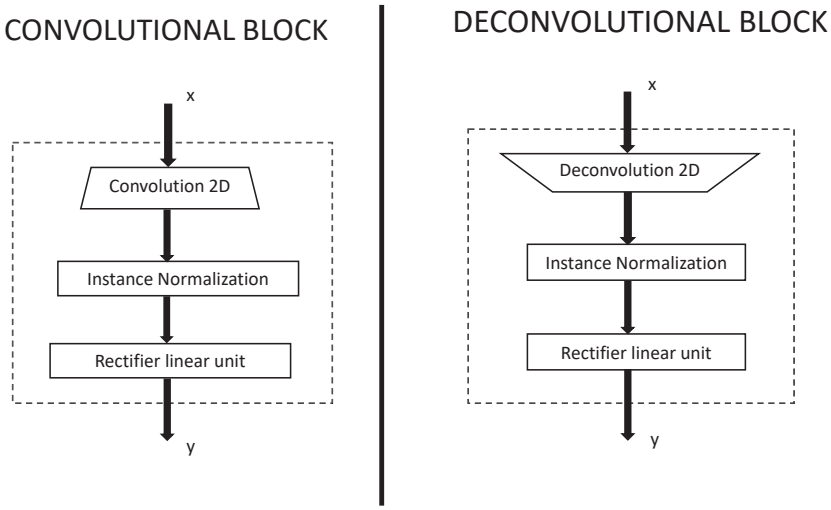


Figure 5.7 – Convolutional and deconvolutional blocks modified to included instance normalization from [39].

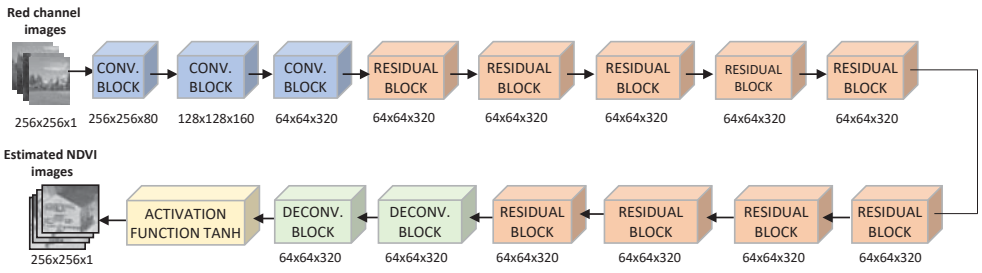


Figure 5.8 – Cyclic generative adversarial network detailed architecture [39] with the proposed changes.

The model includes two mappings functions G : *grayscale image* → *NIR image* and F : *NIR image* → *grayscale image*. In addition, it introduces two adversarial discriminators D_x and D_y , where D_x aims to distinguish between *real grayscale* images and translated images *(synthetic grayscale)*; in the same way, D_y aims to discriminate between *real NIR* and *(synthetic NIR)*. Besides, the proposed approach includes two types of loss terms: adversarial losses [34] for matching the distribution of generated synthetic NIR images to the data distribution in the target domain real NIR images; and a cycle consistency loss to prevent the learned mappings G and F from contradicting each other. Once the NIR has been estimated, it is used to obtain the NDVI vegetation index from eq. (5.1).

Additionally, the proposed architecture uses a modified residual block from (ResNET) [39] to perform the image transformation from the visible spectrum (red channel) to vegetation index and viceversa. In order to avoid the vanishing gradient problem, as the gradient is back-propagated to earlier layers, repeated multiplication may make the gradient infinitely small. See Section 5.2.2.2 for more details information on ResNet.

The "skip connections" from ResNet ensure that the properties of NDVI images of previous layers are available for later layers as well, so that their output do not deviate much from original RGB input (red channel), otherwise the characteristics of original images will not be retained in the output and results will be very unreal. In this chapter a modification of the original residual block is introduced to improve the quality of the NDVI image obtained by the network. For this purpose an instance normalization has been added, see Fig. 5.7, to calculate the μ and σ along the (C, H, W) axes for each sample and channel. The detailed architecture used to generate NDVI synthetic vegetation index is complex as can be appreciated in Fig. 5.8.

Figures 5.9 and 5.10 depict the CyclicGAN model proposed in the current section. It is composed of two generators (G, F) and two discriminators (Dx, Dy). In order to obtain a synthetic NDVI vegetation index, the architecture takes advantage of the joint of cycle-consistency and least square losses [67] in addition to the usual discriminator and generator losses. The results of the experiments have shown that these loss functions demand that the model maintain textural information of the visible (corresponding red channel) and NDVI images and generate uniform synthetic outputs.

5.2.2.5 Loss Functions

According to [128] the objective of a CyclicGAN is to learn two mapping functions between two domains *X(red channel)* and $Y(NDVI)$ given training samples of each image category $\{x_i\}_{i=1}^{N}$ where $x_i \in$ *X(red channel)* and $\{y_j\}_{j=1}^{N}$ where $y_j \in Y(NDVI)$. The data distribution is denoted by $x \sim p_{\text{data}}(x)$ and $y \sim p_{\text{data}}(y)$. The adversarial losses, according to [34], are applied to both mapping functions.

Also, according to [128], to reduce the space of possible mapping functions, the learned mapping functions should be cycle-consistent; for each image x (red channel) from domain X, the image translation cycle should be able to bring x back to the original image, i.e., $x \rightarrow$

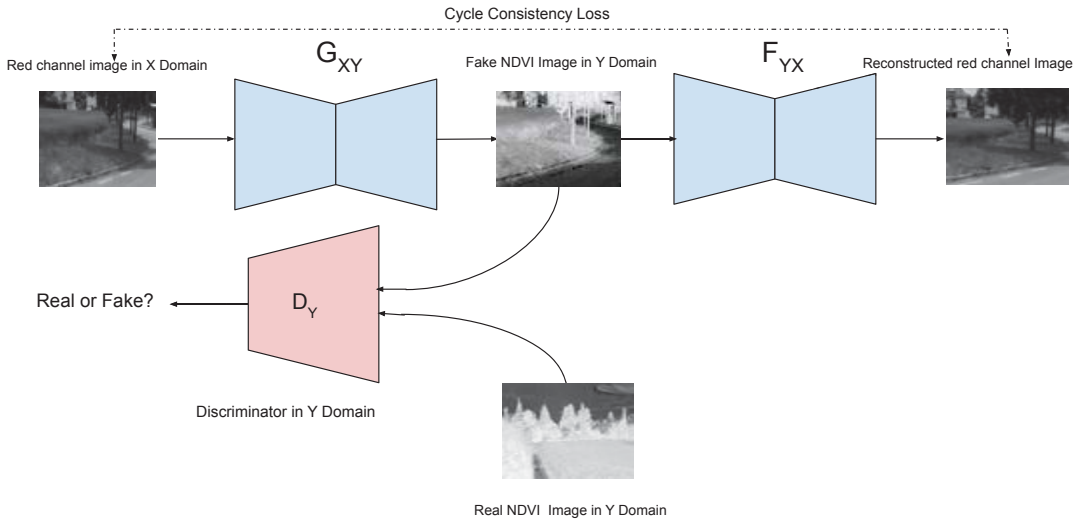


Figure 5.9 – Cycle generative adversarial model *G*: *X(red channel)* → $Y\,(NDVI)$ and its discriminator D_y.

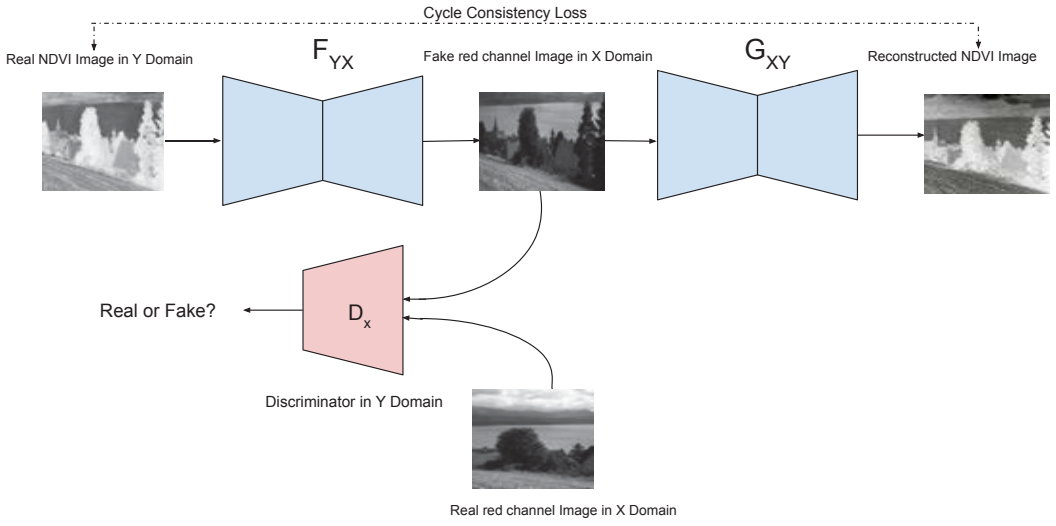


Figure 5.10 – Cycle generative adversarial model F:$Y\,(NDVI)$ → *X(red channel)* and its discriminator D_x.

$G(x) \rightarrow F(G(x)) \approx x$, calling this forward cycle consistency. Therefore, for each image y (NDVI) from domain Y, G and F should also satisfy backward cycle consistency: $y \rightarrow F(y) \rightarrow G(F(y))$ $\approx y$. This cycle consistency loss is defined as :

$$\mathscr{L}_{CYCLE}(G,F) = \mathbb{E}_{x} \sim_{p\ \mathrm{data}(x)} [\|F(G(x)) - x\|_1] + \mathbb{E}_{y} \sim_{p\ \mathrm{data}(y)} [\left\|G(F(y)) - y\right\|_1]. \tag{5.7}$$

5.2.2.6 Least Squares GAN's Loss

Recently, Generative Adversarial Networks [34] have demonstrated impressive performance for unsupervised learning tasks. Unlike other deep generative models, which usually adopt approximation methods for inference, GANs do not require any approximation and can be trained end-to-end through the differentiable networks. In spite of the great progress for GANs in image generation, the quality of generated images by GANs is still limited for some realistic tasks. Regular GANs adopt the sigmoid cross entropy loss function for the discriminator [67]. This loss function, however, will lead to the problem of vanishing gradients when updating the generator using the fake samples that are on the correct side of the decision boundary but are still far from the real data.

The implementation of the Least Squared GAN (LSGAN) can bring two principal advantages. First, unlike regular GANs, which cause almost no loss for samples that are very close to the groundtruth, this proposed loss LSGAN penalize the same samples even though they are correctly classified at the moment of generator updates parameters. On the other hand, the parameters of the discriminator are fixed, i.e., the decision boundary is fixed. As a result, the penalization makes the generator to generate samples toward the decision boundary. Additionally, the decision boundary should go across the manifold of real data for a successful GAN learning. Otherwise, the learning process will be saturated.

In the current paper a least square loss has been implemented [67] to accelerate the training process. This loss is able to move the fake samples toward the decision boundary, in other words, generate samples that are closer to real data, in this case the synthetic NDVI image. The experiments performed with this loss instead of negative log likelihood shown better results. The standard adversarial GAN loss function are replaced with the least square losses, which are defined as :

$$\mathscr{L}_{LSGAN}(G,D_y,X,Y) = \mathbb{E}_{y} \sim_{p\ \mathrm{data}(y)} [(D_Y(y)-1)^2] + \mathbb{E}_{x} \sim_{p\ \mathrm{data}(x)} [D_Y(G(x))^2], \tag{5.8}$$

$$\mathscr{L}_{LSGAN}(F,D_x,Y,X) = \mathbb{E}_{x} \sim_{p\ \mathrm{data}(x)} [(D_X(x)-1)^2] + \mathbb{E}_{y} \sim_{p\ \mathrm{data}(y)} [(D_X(F(y))^2)]. \tag{5.9}$$

5.2.2.7 Multiple Losses

For the supervised scheme, the first approach of the supervised scheme has only used the adversarial loss of the standard GAN. On the contrary, the model of the second supervised

approach has been defined with a multi-term loss ($\mathscr{L}$) formed by the combination of the adversarial loss plus the intensity loss (MSE) and the structural similarity loss (SSIM). This combined loss has been defined to avoid the usage of only a Pixel-wise Loss (PL) to measure the mismatch between a generated image and its corresponding ground-truth image. This multi-term loss function is better designed to human perceptual criteria of image quality, which is detailed next. The adversarial loss is designed to minimize the cross-entropy to improve the texture loss :

$$\mathscr{L}_{Adversarial} = -\sum_{i} logD(G_w(I_{z|y}), (I_{x|y})), \tag{5.10}$$

where D and G_w are the discriminator and generator of the real $I_{x|y}$ and generated $I_{z|y}$ images conditioned by the red channel of the RGB of the GAN network.

The intensity loss is defined as:

$$\mathscr{L}_{Intensity} = \frac{1}{NM}\sum_{i=1}^{N}\sum_{j=1}^{M}(NDVIe_{i,j} - NDVIg_{i,j})^2, \tag{5.11}$$

where $NDVIe_{i,j}$ is the vegetation index estimated by the network and $NDVIg_{i,j}$ is the ground-truth vegetation index and $N \times M$ is the size of the patches. This loss measures the difference in intensity of the pixels between the images without considering texture and content comparisons. Additionally, this loss penalizes larger errors, but is more tolerant to small errors, without considering the specific structure in the image.

To address the limitations of the simple intensity loss function, the usage of a reference-based measure is proposed. One of the reference-based index is the structural similarity index [111], which evaluates images accounting for the fact that the human visual perception system is sensitive to changes in the local structure; the idea behind this loss function is to help the learning model to produce a visually improved image. The Structural loss for a patch P is defined as:

$$\mathscr{L}_{SSIM} = \frac{1}{N}\sum_{p=1}^{P} 1 - SSIM(p), \tag{5.12}$$

where SSIM(p) is the Structural Similarity Index (see [111] for more details) centered in pixel p of the patch P.

For the two unsupervised approaches, the $\mathscr{L}_{GAN}$: (standard GAN adversarial loss), $\mathscr{L}_{CYCLE}$: (cycle-consistent loss) and $\mathscr{L}_{LSGAN}$: (least square loss), have been used. Each of the schemes has its own definition of the distribution of the weights of each component of the loss function. For the first unsupervised approach, the weighted sum of the individual loss function terms

designed to obtain the best results, is defined as:

$$\mathscr{L}_{FINAL-SNTH_{CYCLE-LSGAN}} = 0.45\ \mathscr{L}_{LSGAN} + 0.55\ \mathscr{L}_{CYCLE} \tag{5.13}$$

For the second unsupervised approach, two losses have been used, the first one is the standard adversarial loss.

And the second loss evaluated in the second unsupervised approach is the LSGAN loss where the weighted sum of the individual loss function terms defined as:

$$\mathscr{L}_{FINAL-RED_{CYCLE-LSGAN}} = 0.65\ \mathscr{L}_{LSGAN} + 0.35\ \mathscr{L}_{CYCLE} \tag{5.14}$$

The combination of the weights associated with each loss function is focused on improving the quality of the images for human perception and at the same time, they are used as regularization terms that determine which loss function is the most significant in the optimization of the model for the generation of the synthetic vegetation index. An inappropriate weights balance increases the risk that the model generates synthetic indexes with too many artifacts and that it cannot generalize properly.

5.3 Results and Discussions

This section presents the data augmentation process applied to the ground truth images used for the experiments and the results obtained with the different implemented approaches.

5.3.1 Datasets for Training and Testing

In this chapter two schemes have been implemented, the first one corresponds to supervised approaches and the second one corresponds to unsupervised approaches. Both schemes have been trained and tested receiving as an input the cross-spectral dataset from [14]. From the aforementioned dataset the *country*, *mountain* and *field* categories have been considered for evaluating the performance of the proposed approaches. For the supervised approaches, examples of the patches from this dataset are shown in Fig. 5.11, for the unsupervised approaches, some examples images resized Fig. 5.12. The cross-spectral dataset consists of 477 registered images categorized in 9 groups captured in RGB (visible spectrum) and NIR (near infrared spectrum). The *country* category contains 52 pairs of images of (1024×680 pixels), while the *field* contains 51 pairs of images of (1024×680 pixels). In order to train the supervised network schemes, implemented to generate vegetation index, from each of these categories 380.000 pairs of patches of (64×64 pixels) have been cropped both, in the RGB images as well as in the corresponding NIR and NDVI images. Additionally, 3800 pairs of patches, per category, of (64×64 pixels) have been also generated for validation. It should be noted that images are correctly registered, so that a pixel-to-pixel correspondence is guaranteed. On the

contrary, to train the schemes of unsupervised networks, from each of these categories 3500 pairs of images of (256×256 pixels) have been generated both, in the red channel of the RGB images as well as in the corresponding NIR and NDVI images (the NIR images are used to compute the ground truth NVDI indexes, which are represented as images). Additionally, 1000 pairs of images, per category, of (256×256 pixels), have been also generated for testing and 100 pairs of images per category for validation. It should be noted that the images used during the training process are not paired; in other words, they do not correspond to the same scene, because there is no need to have correspondence to be used as input for the model.

5.3.2 Data Augmentation

In this section, the data augmentation process applied to the unsupervised approach uses a cross-spectral dataset from [14]. For the second unsupervised approach, the network receives as an input an unpaired dataset of red channel images together with their corresponding NDVI images; as mentioned above, this NDVI is computed from NIR and red images. In order to enlarge the size of the training dataset, an automatic data augmentation process has been implemented to create a modified version of images in the dataset of red channel and NDVI by taking random crops with a parameterized size, randomly selecting the coordinates in the image to crop the region before the training phase. After the creation of multiple variations of the images, that can improve the performance and the ability of the fit models to generalize what they have learned to new images. The data augmentation process executed for the unsupervised approaches (see Algorithm **??**) has generated a total of 70 different variations with a size of (256×256 pixels) for each image per category existing in the dataset, which can be used to feed the learning network to synthesize vegetation indices to increase the performance and accelerate the generalization of the model. For each category 1600 pairs of images from visible and NIR spectrum have been generated. Additionally, for each category, 1200 pairs of images for testing and 400 pairs of images for validation from visible and NIR spectrum have been used.

5.3.3 Evaluation Metrics

Digital images resulting from an artificial intelligence process, such as deep neural networks, are subject to a wide variety of distortions, which may result in a degradation of visual quality. Quality is a very important parameter for all objects and their functionalities. The importance of research in the objective evaluation of image quality is to develop measures that can automatically predict the perceived image quality. In an image based technique, image quality is a prime criterion. Commonly, for a good image quality evaluation, an evaluation with complete reference metrics is applied, like MSE (Mean Square Error), one of the most used image quality metrics. The MSE metric measures the average of the squares of the errors or deviations between the generated and the original image. This error (MSE), does not match with human visual perception. In contrast to MSE, SSIM is also evaluated to measure image quality level.

Figure 5.11 – Pairs of patches (64×64) from *country* Category (two-left columns) and *field* Category (two-right columns) [14]: (top) RGB image; ($middle$) Red channel of the given RGB image; ($bottom$) NDVI vegetation index computed from RGB images and the corresponding NIR images.

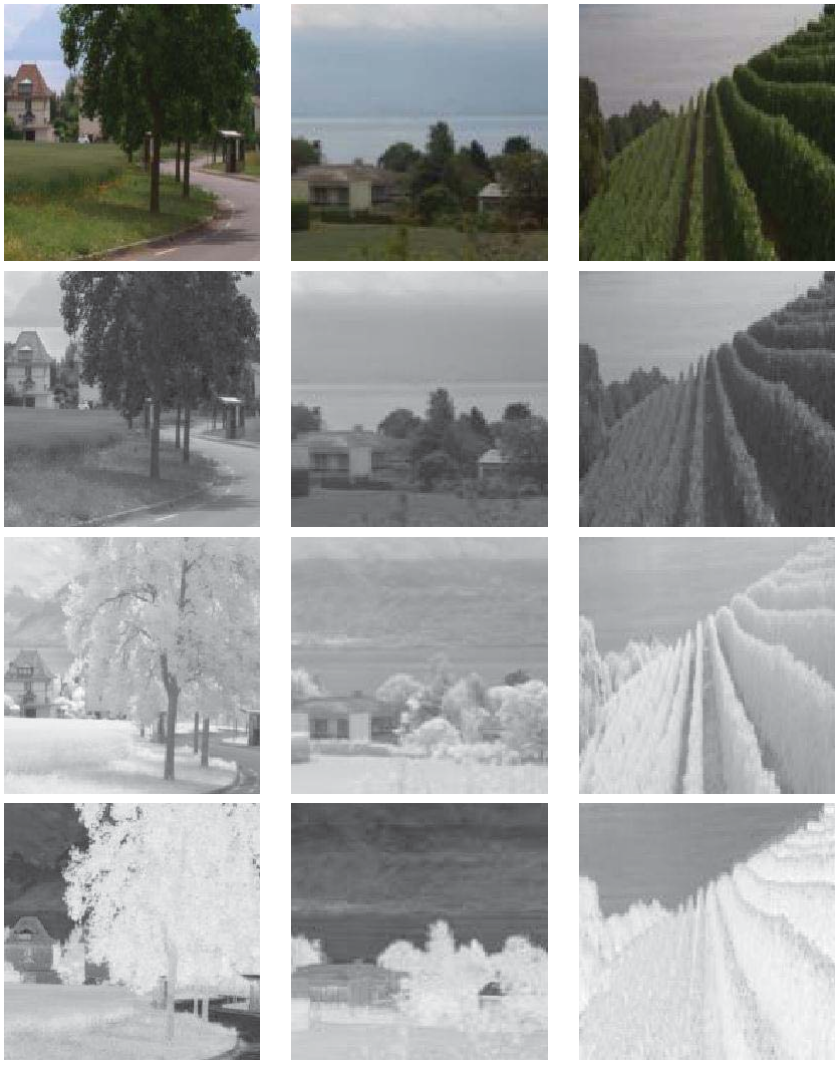


Figure 5.12 – Cross-spectral images: (*1st row*) RGB images; (*2nd row*) Red channel images used as input into the CyclicGAN; (*3rd row*) Correspondig NIR images; (*4th row*) Ground truth NDVI images. Images from [14], *country*, *field* and *mountain* categories (from left to right).

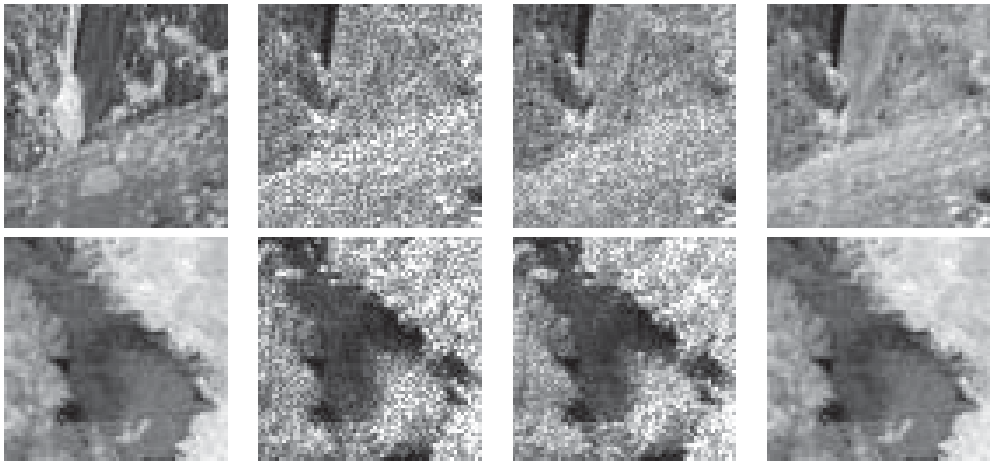


Figure 5.13 – Images of NDVI vegetation indexes obtained with the first supervised approach implemented in this chapter: (*1st col*) NDVI index as ground truth images from the *country* category; (*2nd col*) NDVI index results from flat GAN network; (*3rd col*) NDVI index obtained with the siamese GAN network); (*4th col*) NDVI index obtained with the triplet GAN network.

In the current chapter, for all the supervised and unsupervised implemented approaches, MSE and SSIM metrics have been used, with which it was possible to measure the quality of the results coming from the experiments and determine the validity of the proposed approach being evaluated. However, MSE is an absolute value of the representation perspective, instead of SSIM which is normalized. Additionally, from a semantic perspective, SSIM gives better results to measure over MSE error. Also, SSIM performs well to obtain perception and saliency-based errors. According to [111], SSIM, evaluates images accounting for the fact that the human visual perception system is sensitive to changes in the local structure.

5.3.4 Experimental Results

The first supervised proposed approach has been evaluated using NIR images and their corresponding NVDI vegetation index, obtained from the equation presented above, in which the RGB was used; this cross-spectral dataset came from [14]. The *country* and *field* categories have been considered for evaluating the performance of the proposed approach. Examples of this dataset are presented in Fig. 5.12.

In order to train the network to generate vegetation index from each of these categories 280.000 pairs of patches of (64×64 pixels) have been cropped both, in the NIR images as well as in the corresponding NVDI images. Additionally, 2800 pairs of patches, per category, of (64×64 pixels) have been also generated for validation. It should be noted that images are correctly registered, so that a pixel-to-pixel correspondence is guaranteed.

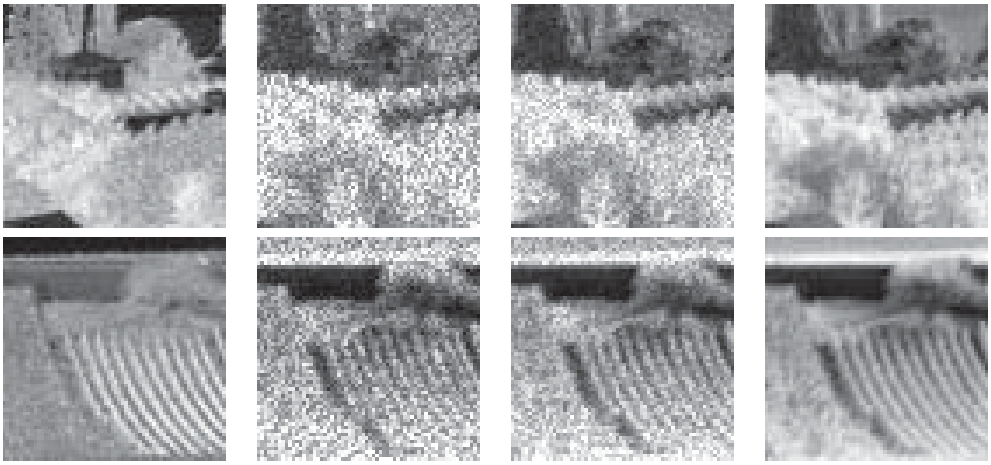


Figure 5.14 – Images of NDVI vegetation indexes obtained with the first supervised approach implemented in this chapter: (*1st col*) NDVI index as ground truth images from the *field* category; (*2nd Col*) NDVI index results from the FLAT GAN Network;(*3rd col*) NDVI index obtained with the siamese GAN network); (*4th col*) NDVI index obtained with the triplet GAN network.

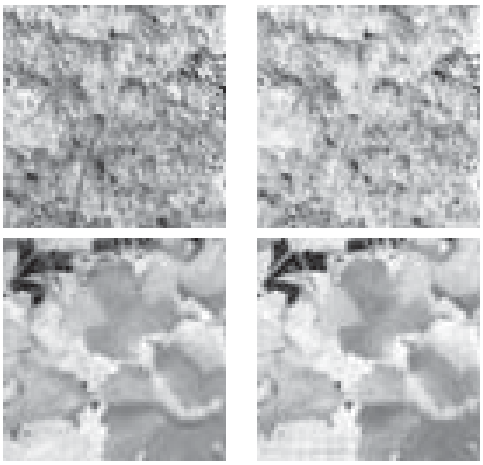


Figure 5.15 – Images of NDVI vegetation indexes obtained with the second supervised approach implemented in this chapter: (*1st col*) Ground truth NDVI index from the *country* category; (*2nd col*) NDVI index obtained with the proposed GAN architecture with $\mathscr{L}_{Final}$ loss function.

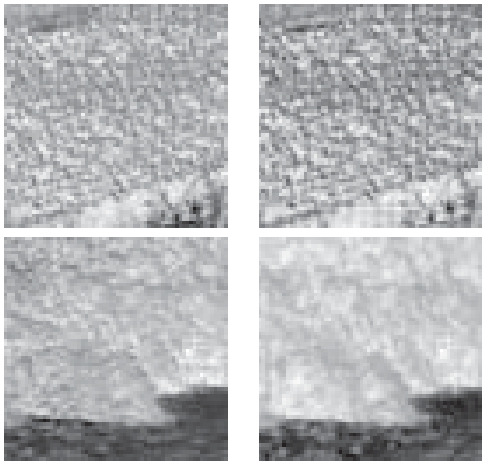


Figure 5.16 – Images of NDVI vegetation indexes obtained with the second supervised approach implemented in this chapter: (*1st col*) Ground truth NDVI index from the *field* category; (*2nd col*) NDVI index obtained with the proposed GAN architecture with $\mathscr{L}_{Final}$ loss function.

Training	RMSE		SSIM	
	country	*field*	*country*	*field*
GAN with $\mathscr{L}_{Adversarial} + \mathscr{L}_{Intensity}$	3.93	4.12	0.86	0.83
GAN with $\mathscr{L}_{Adversarial} + \mathscr{L}_{SSIM}$	3.81	3.96	0.91	0.89
GAN with $\mathscr{L}_{Final}$	3.53	3.70	0.94	0.91

Table 5.1 – Root mean squared errors (RMSE) and structural similarities (SSIM) obtained with the second supervised proposed GAN architecture by using different loss functions (SSIM the bigger the better).

The Conditional Generative Adversarial networks from the first approach evaluated in the current chapter *(Generator: Flat, Siamese and Triplet)* for NDVI vegetation index estimation, see Section 5.2.1 have been trained using a 3.2 eight core processor with 16GB of memory with a NVIDIA GeForce GTX970 GPU. Qualitative results are presented in Fig. 5.13 and Fig. 5.14. Figure 5.13 shows NDVI vegetation index images from the *country* category generated with the flat, siamese and triplet proposed GAN network. Additionally, Fig. 5.14 shows NDVI vegetation index images from the *field* category generated with the flat, siamese and triplet proposed GAN network. Quantitative evaluations for the different architectures have been obtained and provided below. Up to my humble knowledge, there is not previous work tackling the vegetation index estimation from just one spectral band. Hence, the only way to evaluate the results is by comparing the Root Mean Square Error (RMSE) of each approach. The RMSE measures the similarity between the estimated NDVI concerning the ground truth, which is the standard deviation of the residuals. Residuals are a measure of how distant are the images compared to each other.

Figure 5.17 – Illustration of NIR images obtained by the first unsupervised proposed CyclicGAN, which are later on used to estimate the corresponding NDVI indexes: (*1st row*) RGB images; (*2nd row*) Gray scale image used as input into the CyclicGAN; (*3rd row*) Estimated NIR images; (*4th row*) Ground truth NIR images. Images from [14], *country*, *field* and *mountain* categories.

Table 5.2 – Average Root Mean Squared Errors (RMSE) and Structural Similarities (SSIM) obtained from estimated NDVI vegetation index from the first unsupervised proposed approach and the real one computed from eq. (5.1) (the bigger SSIM, the better). Note that NDVI values are scaled up to a range of [0-255] since they are depicted as images as shown in Fig. 5.18.

NDVI estimation	RMSE			SSIM		
	country	*field*	*mountain*	*country*	*field*	*mountain*
Results from 2nd. Sup. App.	3.53	3.70	–	0.94	0.91	–
Results from 1st Unsup. App. $\mathscr{L}_{FINAL-SNTH_{CYCLE-GAN}}$	3.42	3.64	3.63	0.94	0.91	0.86
Results from 1st Unsup. App. $\mathscr{L}_{FINAL-SNTH_{CYCLE-LSGAN}}$	3.39	3.56	3.81	0.94	0.92	0.89

Table 5.3 – Average root mean squared errors (RMSE) and structural similarities (SSIM) obtained from the NDVI vegetation index estimated from the second unsupervised approach of the current chapter and the real one computed from eq. (5.1) (the bigger SSIM, the better). Note that NDVI values are scaled up to a range of [0-255].

NDVI estimation index	RMSE			SSIM		
	country	*field*	*mountain*	*country*	*field*	*mountain*
GAN results from 2nd. Sup. App.	3.53	3.70	–	0.94	0.91	–
CyclicGAN results from [128]	3.46	3.53	3.82	0.93	0.90	0.88
CyclicGAN results from 1st. Unsup. App. $\mathscr{L}_{FINAL-SNTH_{CYCLE-GAN}}$	3.42	3.64	3.63	0.94	0.91	0.86
CyclicGAN results from 1st. Unsup. App. $\mathscr{L}_{FINAL-SNTH_{CYCLE-LSGAN}}$	3.39	3.56	3.81	0.94	0.91	0.89
CyclicGAN results from 2nd. Unsup. App. $\mathscr{L}_{FINAL-RED_{CYCLE-GAN}}$	3.39	3.50	3.72	0.94	0.92	0.90
CyclicGAN results from 2nd. Unsup. App. $\mathscr{L}_{FINAL-RED_{CYCLE-GAN}}$	3.15	3.11	3.20	0.94	0.92	0.90

The second supervised approach to estimate NDVI vegetation index uses a Conditional Generative Adversarial network which has been designed as a Flat (single layer) of learning level using a single channel (red) to perform the estimation. It has been trained using a 3.4 four core processor with 16Gb. of memory with a NVIDIA Titan XP GPU. Qualitative results are presented in Fig. 5.15 and Fig. 5.16. Figure 5.15 shows NDVI vegetation index images from the *country* category generated with the proposed Flat GAN network with multiple loss functions, see Section 5.2.1.2. Table 5.1 presents the average root mean square errors (RSME) and the standard deviation obtained with the Flat architecture evaluated in the proposed work for the two categories. It can be appreciated that the single learning layer using as an input a red channel image reaches the best result of the supervised approaches.

Table 5.4 – Error of NDVI for *mountain* category from the second unsupervised proposed approach with LSGAN loss.

Vegetation health classes using NDVI	Evaluation Error	
	Mean	*Standard Deviation*
[-1;0.20): Barren areas	0.051	0.0472
[0.20;0.38): Crop, grass	0.041	0.0362
[0.38;0.6): Agroforestry	0.029	0.0193
[0.6;1]: Forestry	0.031	0.0256

Table 5.5 – Error of NDVI for *country* category from the second unsupervised proposed approach with LSGAN loss.

Vegetation health classes using NDVI	Evaluation error	
	Mean	*Standard deviation*
[-1;0.20): Barren areas	0.033	0.0291
[0.20;0.38): Crop, grass	0.021	0.0211
[0.38;0.60): Agroforestry	0.019	0.0102
[0.6;1]: Forestry	0.023	0.0114

Table 5.6 – Error of NDVI for *field* category from the second unsupervised proposed approach with LSGAN loss.

Vegetation health classes using NDVI	Evaluation error	
	Mean	*Standard deviation*
[-1;0.20): Barren areas	0.041	0.0349
[0.20;0.38): Crop, grass	0.032	0.0321
[0.38;0.60): Agroforestry	0.023	0.0149
[0.60;1]: Forestry	0.028	0.0176

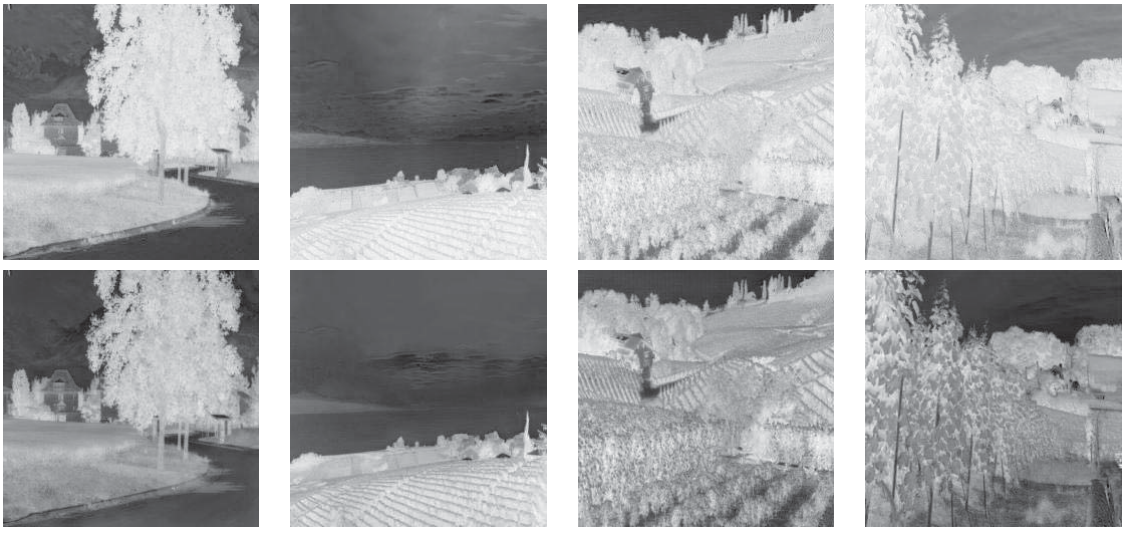


Figure 5.18 – Images of NDVI vegetation indexes obtained with the synthetic NIR generated by the proposed CyclicGAN: (*top*) Ground truth NDVI vegetation index images; (*bottom*) Estimated NDVI vegetation indexes. Images from [14], *country*, *field* and *mountain* categories.

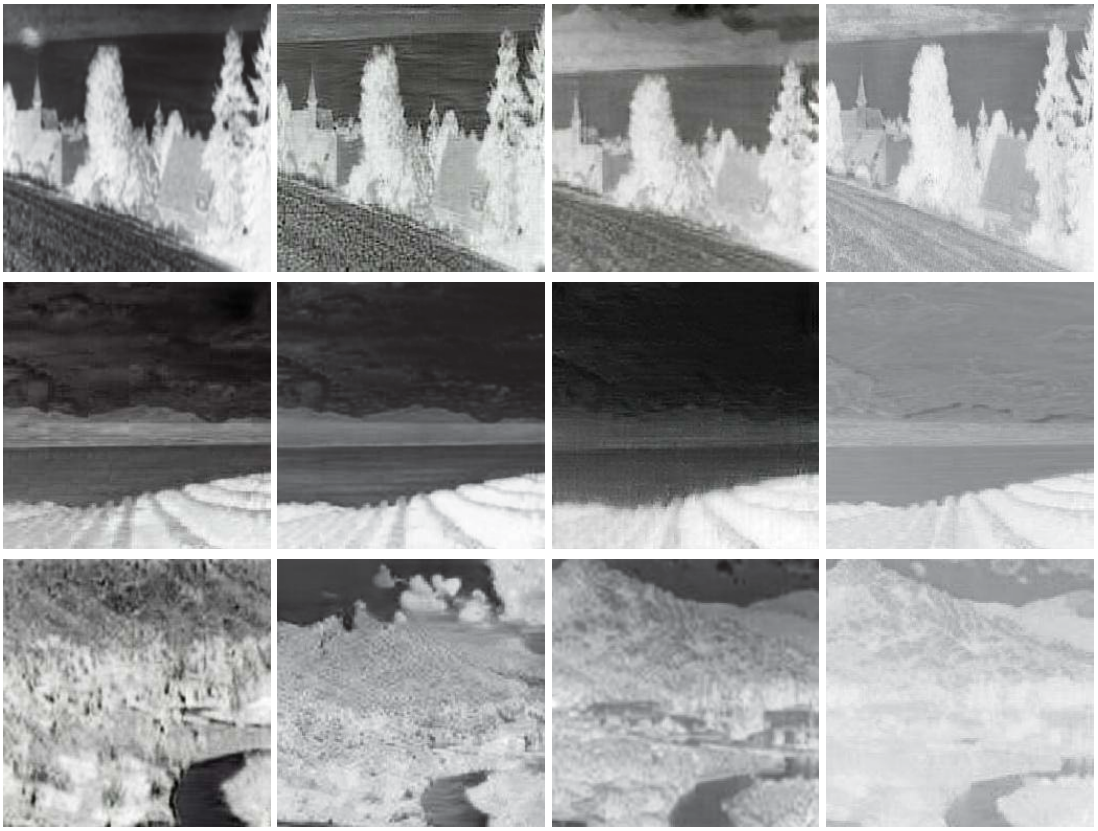


Figure 5.19 – Images of NDVI vegetation indexes obtained with the second unsupervised approach implemented in this chapter: (*1st col*) NDVI estimated with [128]; (*2nd col*) NDVI estimated by the proposed CyclicGAN; (*3rd col*) NDVI estimated by the proposed CyclicGAN with LSGAN; (*4th col*) Ground truth NDVI vegetation index. Images from [14], *mountain* category.

Results from the first unsupervised approach using a Cyclic Generative Adversarial network to generate (synthetic NIR images) in order to avoid the dependence on near infrared (NIR) image obtain better results than the two supervised approaches previously presented. Some examples of the dataset used in this approach are presented in Fig. 5.17. These synthetic NIR images are then used for estimating the NDVI indexes. Figure 5.18 presents some illustrations of NDVI indexes estimated from these NIR images and the ground truth ones computed from eq. (5.1). Quantitative evaluations are presented in Table 5.2. In this table average root mean square error (RMSE) and structural similarity index metric (SSIM) computed over the validation set are depicted, when different combinations of the proposed loss functions where considered.

The standard loss function for GANs has been implemented in the model, which is based on negative log likelihood and also uses the least square loss, which obtain better quantitative

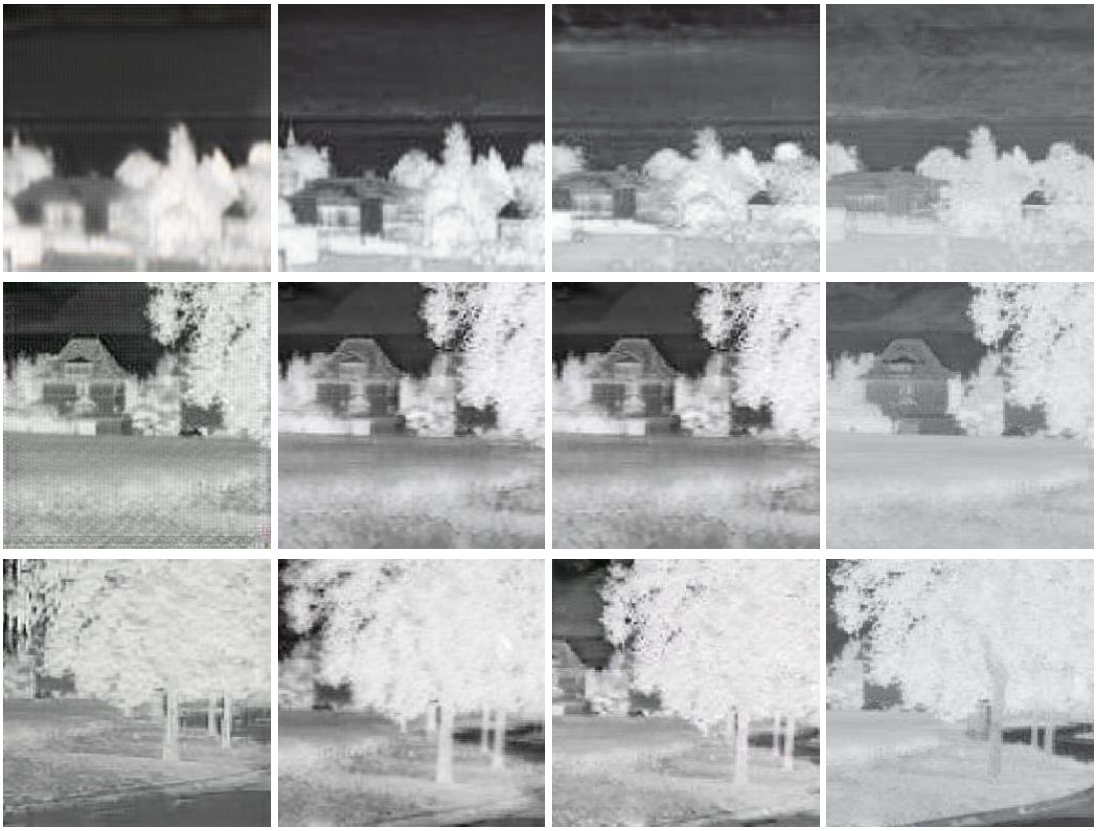


Figure 5.20 – Images of NDVI vegetation indexes obtained with the second unsupervised approach implemented in this chapter: (*1st col*) NDVI estimated with [128]; (*2nd col*) NDVI estimated by the proposed CyclicGAN; (*3rd col*) NDVI estimated by the proposed CyclicGAN with LSGAN; (*4th col*) Ground truth NDVI vegetation index. Images from [14], *field* category.

results and avoid the vanishing gradient problem, where a deep feed-forward network is unable to propagate valid gradient information from the output back to the first layer of the model. A least square loss has been implemented to accelerate and maintain stable the training process. Additionally, in this Table 5.2, results from second supervised approach are presented for comparison. It can be appreciated that in all the cases the results obtained with the least square loss in the first unsupervised proposed CyclicGAN are better than those obtained with the second supervised approach. It should be mentioned that the least square loss permits to accelerate the network convergence in less number of epochs, allowing a better optimization of the network.

Hence, discriminator becomes more stable and the network converge faster in less number of epochs

A novel second unsupervised approach implemented to estimate the NDVI vegetation index based also on a the Cyclic Generative Adversarial Network (synthetic NDVI images)

with some variations in the architecture has obtained outstanding results comparing with previous approaches, the two supervised approaches and the first unsupervised approach already reviewed, which can be used to evaluate the health status of plants and take actions depending on it. Figures 5.19, 5.20, 5.21 present some illustrations of NDVI indexes estimated from the proposed architecture using as a ground truth the NDVI index computed from eq. (5.1). Quantitative evaluations are presented in Table 5.3. In this table average root mean square errors (RMSE) (note the NDVI values are scaled up to a range of [0-255]) and structural similarity index metric (SSIM) computed over the validation set are depicted, when different combinations of the proposed loss functions were considered see (Section 5.2.2.5). In the experiments, the standard loss function which is based on negative log likelihood and also the least square loss was used for Cyclic GANs, which obtain better quantitative results and avoid the vanishing gradient problem. A deep feed-forward network is unable to propagate valid gradient information from the output back to the first layer of the model. Least square loss function has been implemented to accelerate the network convergence, allowing a better optimization of the network and maintain stable the training process. Additionally, in this table, results from standard CyclicGAN [128], second supervised and unsupervised approach are presented. It can be appreciated that in all the cases the results obtained with the least square loss in the proposed CyclicGAN, the second unsupervised approach, are better than those obtained with the others approaches.

The NDVI error distribution values per category *country*, *field* and *mountain* computed from the experimental results for each multiple loss implemented in the second unsupervised approach are shown in Fig. 5.22, 5.23 and 5.24 from which it can be determined that the results of the *mountain* category with the different multiple losses is the one with an average error distribution slightly smaller than the other categories *country* and *field*. To increase the cyclic loss effect over the network $L1$ regularization has been used. The proposed CyclicGAN network has been trained using Stochastic AdamOptimizer since it is well suited for problems with deep network, large datasets and avoid overfitting. The image dataset was normalized in a (-1,1) range. The following hyper-parameters were used during the training process: learning rate 0.0003; epsilon = 1e-08; exponential decay rate for the 1st moment of momentum 0.5; $L1$ 9.5; weight decay 1e-2; leak relu 0.18.

The results obtained with the unsupervised technique, lead us to perform an analysis to determine the accuracy of the estimated indexes. This study is carried out over each category, according to the literature the NDVI range from [112] and [90], and the index values are in the range of (-1,1). The quantitative values obtained with the statistical analysis of the estimated NDVI are presented in Tables 5.4, 5.5 and 5.6. These tables shows that the similarity reached compared with the ground truth NDVI indexes, given the mean and standard deviation results obtained per category and level of NDVI classes, specially Agroforestry and Forestry, are less than 0.05 and 0.04 for mean and standard deviation respectively in the worst case *mountain category*.

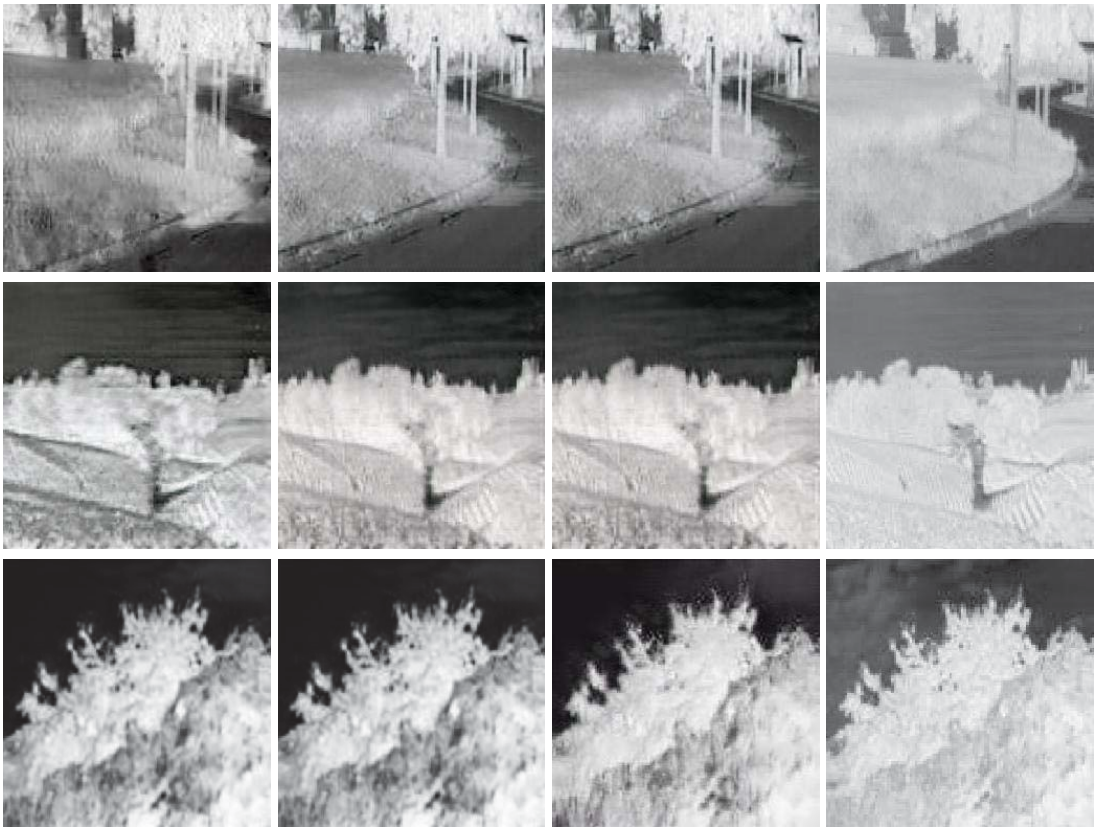


Figure 5.21 – Images of NDVI vegetation indexes obtained with the second unsupervised approach implemented in this chapter: (*1st col*) NDVI estimated with [128]; (*2nd col*) NDVI estimated by the proposed CyclicGAN; (*3rd col*) NDVI estimated by the proposed CyclicGAN with LSGAN; (*4th col*) Ground truth NDVI vegetation index. Images from [14], *country* category.

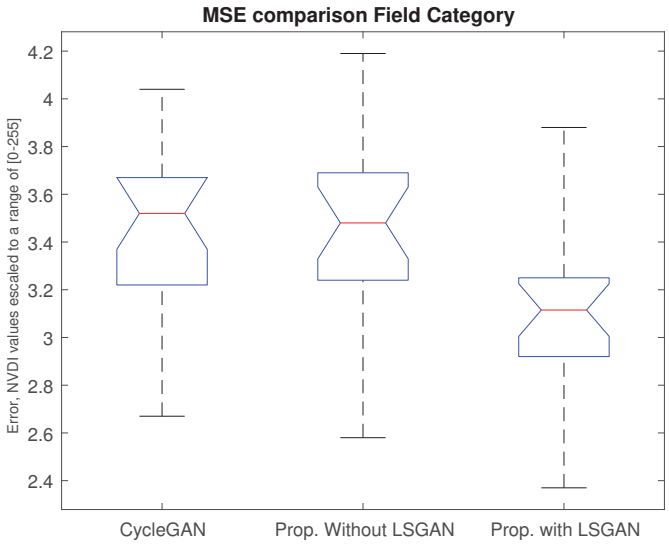


Figure 5.22 – Error distribution for each loss of the second unsupervised approach, *field* category.

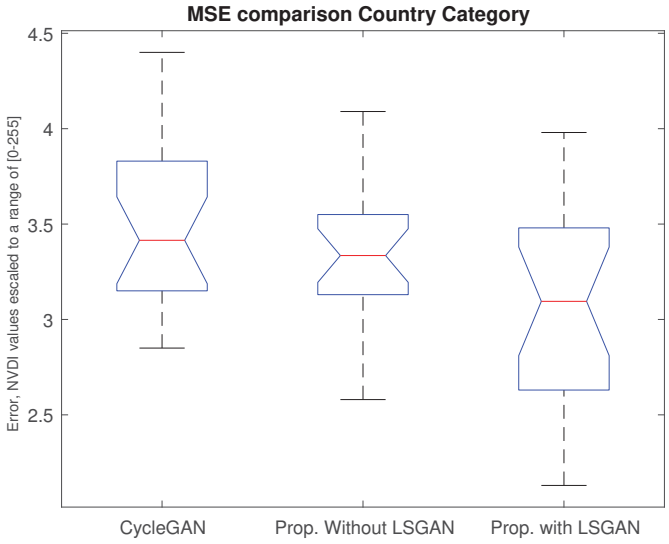


Figure 5.23 – Error distribution for each loss of the second unsupervised approach, *country* category.

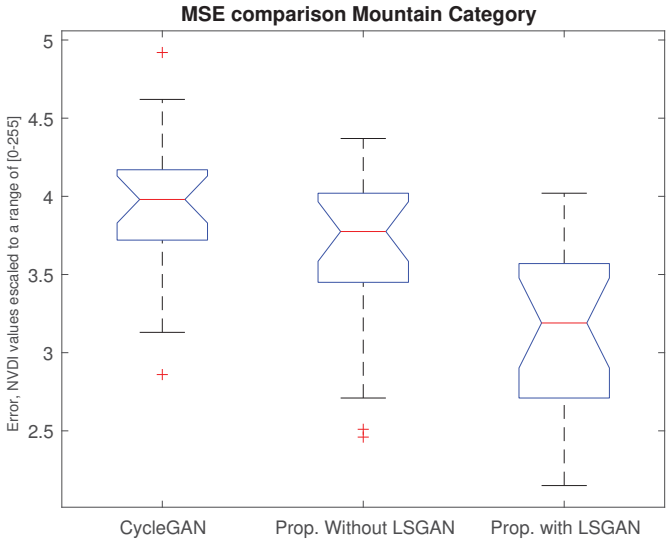


Figure 5.24 – Error distribution for each loss of the second unsupervised approach, *mountain* category.

5.4 Conclusions

This chapter proposes novel supervised and unsupervised architectures to obtain NDVI vegetation indexes using a generative adversarial and a cycle-consistent adversarial network, both using multiple losses. All the approaches are designed in order to avoid the dependence on NIR sensors to compute the mentioned index. The CycleGAN with the least square loss is the novel approach that better tackles the challenging problem of synthesizing NDVI images from a single channel (red) of a RGB representation. Experimental results have shown that the NDVI images estimated with the proposed approach are better than those obtained with the supervised scheme and also the standard unsupervised CyclicGAN model; the quantitative values presented also show better results than previous supervised and unsupervised approaches presented in this chapter for NDVI index estimation.

Chapter 6

Image Dehazing

This chapter proposes a novel approach to remove haze from RGB images using near infrared images based on a dense stacked conditional Generative Adversarial Network. The architecture of the implemented deep network receives, besides the images with haze, its corresponding image in the near infrared spectrum, which serve to accelerate the learning process of the details of the characteristics of the images. The model uses a triplet layer that allows the independent learning of each channel of the visible spectrum image to remove the haze on each color channel separately. A multiple loss function scheme is proposed, which ensures balanced learning between the colors and the structure of the images. Experimental results have shown that the proposed method effectively removes the haze from the images.

6.1 Introduction

The images can be seriously affected by different causes, one of the most common is the natural phenomena that occur when fog, dust, rain or snow are present in the scene. This considerably reduces the visibility of the objects in the images, thus affecting the understanding of the scene. Therefore, processes such as object detection, segmentation or recognition, among others, will not be able to obtain results that meet the required objectives.

Outdoor scenes usually suffer mainly from low contrast and poor visibility due to the adverse weather conditions that cause airborne particles to scatter the light present in the atmosphere. One of the atmospheric effects that occur is the mist, which is independent of the brightness of the scene and generates attenuation effects. It is affected by ambient light at the time of image acquisition. It is necessary to consider that at a greater distance from the camera focus the image more diffuse becomes. The effect of haze on image quality is as a result of a random scattering of light and hence affects all pixels of the image.

Improving the quality of images has been one of the problems that computer vision has sought to solve, several approaches have been proposed, especially aimed at removing climatic effects such as haze; some traditional techniques were focused for the elimination of the haze

presented on images using the characteristics present on them, including those that specialize in removing fog; some focus on working with depth or with multiple views of the same image as presented by [95] where a method based on generic regularity in natural images is presented where the pixels of small image patches usually exhibit a 1D distribution in the RGB color space, known as color lines. This method derives a local training model that explains the color lines in the context of fuzzy scenes and uses it to recover scene transmission based on the displacement of the lines from the origin. In [55] a technique to improve casual outdoor photographs by combining them with existing georeferenced digital terrain and urban models is proposed. This approach uses a registration process to align a photograph with that model.

These methods typically involve multi-step approaches that use depth information for removal of those degradation effects. Most methods for removing haze on images only consider using hard threshold assumptions or user input to estimate atmospheric light. Artificial lighting or applied adaptive filters [64] are also considered in some methods to remove the haze in the images. However, the error estimation of atmospheric light may affect the results of the removal process.

In recent years, deep learning has been extensively used in a wide range of fields. In deep learning, Convolutional Neural Networks are found to give the most accurate results in solving real-world problems. Among the different networks architectures, Generative Adversarial Networks (GANs) have obtained outstanding results to solve problems like the four colorization approaches (two supervised and two unsupervised) presented in chapter 4, face generation, super-resolution [88], [56], text-to-image synthesis [84], cross-spectral similarity approaches presented in chapter 3 or NVDI vegetation index generation approaches presented in chapter 5. Some of these approaches have used NIR images to improve the results obtained by the networks.

One of the advantages of using the properties of the near infrared spectrum image is to enhance the techniques that allow improving the quality of the images without sharpness, because the information that can be recovered can serve to understand a scene more quickly and facilitate making a more accurate decision to solve any kind of problems.

In the particular problem tackled in this chapter to remove the haze to obtain a clear RGB representation, the usage of a GAN architecture is proposed. Two supervised approaches have been implemented the first one used a stacked conditional GAN and the second one that includes two variations, a multi-dense connection and the use of a near infrared image; in this second approach, every channel is mapped into a three-dimensional space, using a stacked dense Conditional GAN model to accelerate convergence and improve the accuracy and efficiency of training. Also, multiple loss functions are used in both supervised approaches. The chapter is organized as follows. The proposed approaches are detailed in Section 6.2. Experimental results with a set of real images are presented in Section 6.3. Finally, conclusions are given in Section 6.4.

6.2 Proposed Approaches

The proposed supervised approaches are based on a generative adversarial model, which take a collection of haze patches and form some image representation without the haze. Generative adversarial networks help in this kind of problems, because the model is capable of generating a new image, rather than finding a function; based on this principle a stacked network architecture with a multiple loss to improve the generalization learning is proposed, this model allows accelerate the diversity obtained in the multiple levels of training. A $l1$ regularization term has been added at every layer of the generator network in order to prevent the coefficients to fit so perfectly to overfit and to introduce more robustness to the generalization of the model; additionally, it helps to reduce the time to reach a well trained network. The architecture of the first approach is presented in Fig. 6.1. Basically in the architecture, a new layer of learning was added, as well as the depth of the learning layers was increased—the learning model is conformed by convolutional, de-convolutional, relu, leak-relu, fully connected and activation function tanh and sigmoid for generator and discriminator networks respectively. Additionally, every layer of the model uses batch normalization for training any type of mapping that consists of multiple compositions of affine transformation with element-wise nonlinearity and do not stuck on saturation mode. It is very important to maintain the spatial information in the generator model, there is not pooling and drop-out layers and only the stride of 1 is used to avoid downsize the image shape.

To improve the first proposed approach some variations have been included to enhance the quality of the haze removal process; a multi dense network stacked three times to accelerate the training process has been proposed. This model, unlike the first proposed approach were, to remove the haze the network receives as an input only RGB image, in the novel proposed approach the network receive in addition to the RGB image, its corresponding image from the near infrared spectrum to obtain images with greater clarity in its details. The GAN generates the image without haze starting from the images of both RGB and NIR concatenated spectra, this architecture with the stacked scheme uses a multiple loss to learn more efficiently and to improve the convergence of the model, which allows to accelerate the obtained diversity and to generalize the learning model. The deep dense connection applied to triplet architecture used for the second proposed approach is shown in Fig. 6.2, it maintains a similar structure than the first proposed approach presented in this chapter. Basically, in the architecture a layer of learning was suppressed, as well as the depth of the learning layers was decreased, because of the concatenation of the NIR with the haze image.

Based on the results obtained in the experiments, it was determined that the second approach implemented is the one that best removes the haze from the images.

The network of the second approach has been designed to learn how to generate new images without haze from a conditional latent distribution. In this case, the generator network has been modified to use feature hierarchical representation; three levels of dense stacking conditional learning process have been included. Additionally, the model has been designed to receive cross-spectral concatenated images as input and use a multiple loss functions. To

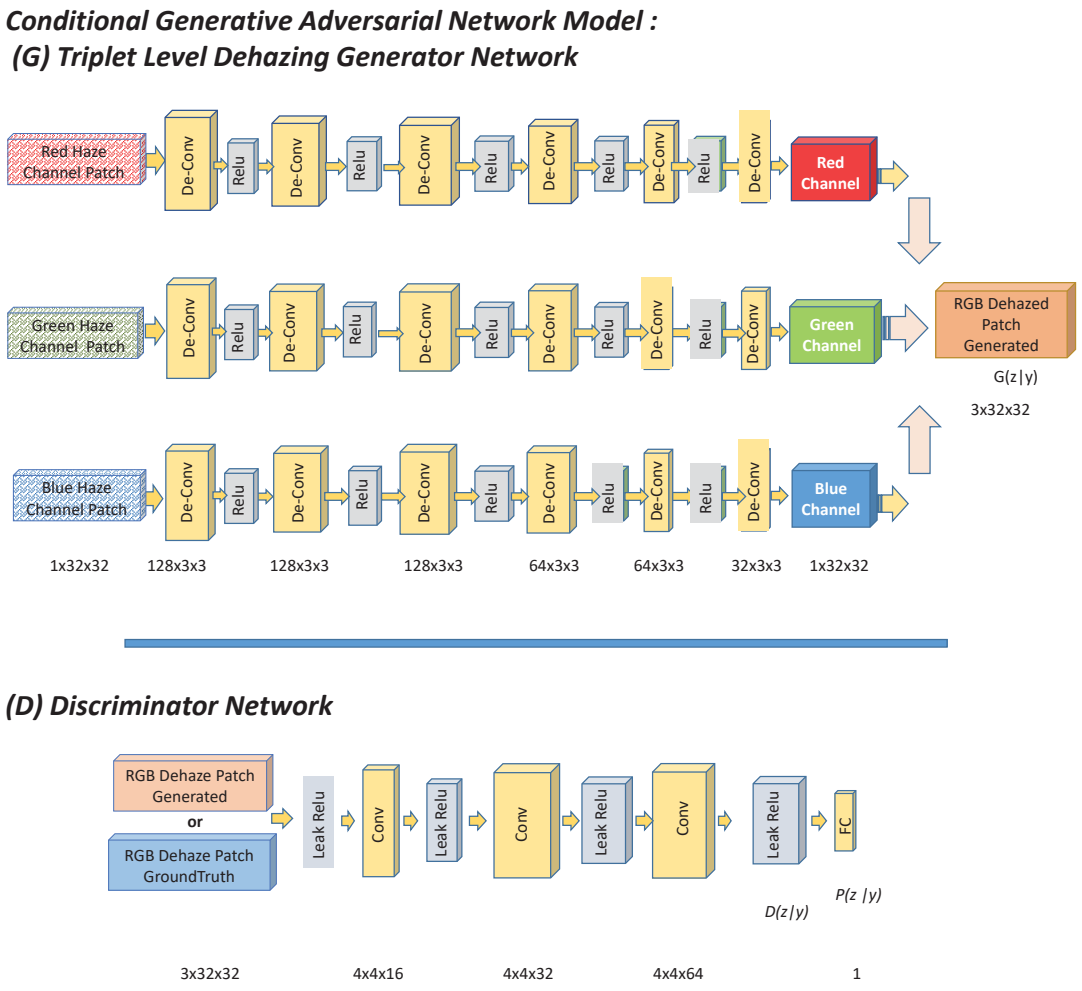


Figure 6.1 – Illustration of the first proposed approach based on triplet dense CGAN architecture used for image dehazing.

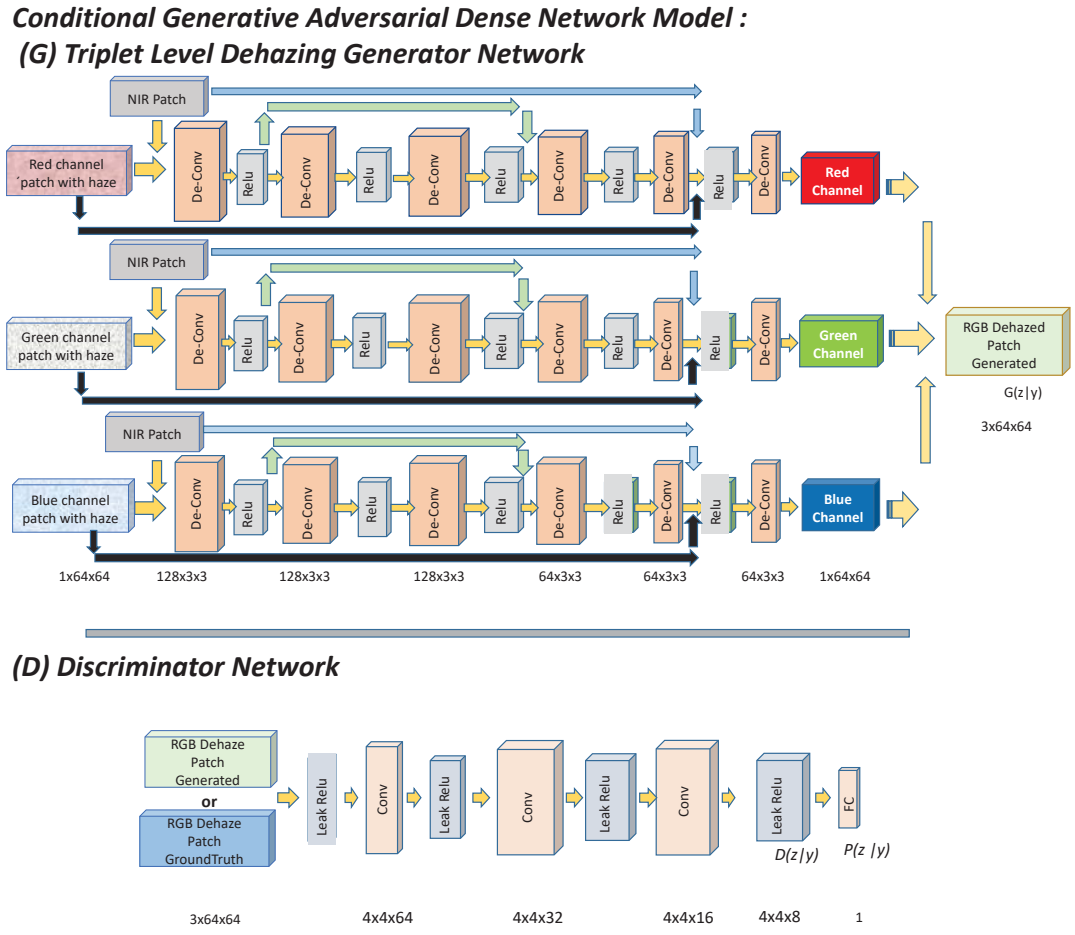


Figure 6.2 – Illustration of the second proposed approach based on triplet cross-spectral dense CGAN architecture used for image dehazing.

optimize the model generalization, the GAN framework is reformulated for a conditional generative image modeling tuple. In other words, the generative model $G(z;\theta_g)$ is trained from a haze and an infrared concatenated image and contrary to the original GAN model formulation, the random noise z is not used; with the assumption that the randomness has already been preserved by the conditioning variables provided by the images with haze, in order to produce a clear RGB image. The discriminative model $D(z;\theta_d)$ is trained to assign the correct label to the generated clear RGB image, according to the provided original color image, which is used as a ground truth. Variables (θ_g) and (θ_d) represent the weighting values for the generative and discriminative networks, (for more details on GAN architecture see Section 2.5.

Densely connected CCN has several compelling advantages: they alleviate the vanishing gradient problem, strengthen feature propagation, encourage feature reuse, and substantially

reduce the number of parameters. Applying this kind of model of connectivity between the layers achieved as a direct consequence of the input concatenation of RGB and NIR image at any level of the learning layers permits that all the feature maps learned by any of the dense net layers can be accessed by all subsequent layers. This encourages feature reuse throughout the network and leads to more compact models.

The second proposed approach introduces a dense connection between layers on the architecture, according to [42], an approach that implements shorter connections generally at the beginning and the end of the learning layers in the model is proposed, this gives to the network the capacity to train more rapidly using fewer layers.

Also, a multiple loss function ($\mathscr{L}$) has been implemented, which was formed by the combination of the adversarial loss plus the intensity loss (MSE), the structural loss (SSIM) and the image quality loss (IQ). This combined loss function has been defined to avoid the usage of only a pixel-wise loss to measure the mismatch between a generated image and its corresponding ground truth image. This multi-term loss function is better designed to human perceptual criteria of image quality, which is detailed below.

The adversarial loss is designed to minimize the cross-entropy to improve the texture loss:

$$\mathscr{L}_{Adversarial} = -\sum_{i} logD(G_w(I_{z|y}),(I_{x|y}), \tag{6.1}$$

where D and G_w are the discriminator and generator of the real $I_{x|y}$ and generated $I_{z|y}$ images conditioned by the haze and near infrared image fed in each channel of the Stacked Gan Network.

The intensity loss is defined as:

$$\mathscr{L}_{Intensity} = \frac{1}{NM}\sum_{i=1}^{N}\sum_{j=1}^{M}(RGBe_{i,j} - RGBg_{i,j})^2, \tag{6.2}$$

where $RGBe_{i,j}$ is the estimated RGB representation and $RGBg_{i,j}$ is the ground truth RGB image. This loss measures the difference in intensity of the pixels between the images without considering the texture and content comparisons. This loss penalizes larger errors, but is more tolerant of small errors, without considering the specific structure in the image.

To address the limitations of the simple intensity loss function, the usage of a reference-based measure is proposed. One of the reference-based indexes is the Structural Similarity Index (SSIM) [111], which evaluates images accounting for the fact that the human visual perception system is sensitive to changes in local structure; the purpose of using this index defines the structural information in an image as those attributes that represent the structure of objects in the scene, independent of the average luminance and contrast. The structural

loss for a pixel p is defined as:

$$\mathscr{L}_{SSIM} = \frac{1}{NM}\sum_{p=1}^{P} 1 - SSIM(p), \tag{6.3}$$

where $SSIM(p)$ is the Structural Similarity Index (see [111] for more details) centered in pixel p of the patch P.

Another loss function considered in this approach is based on the universal image quality index; the method proposed by [110] and it was designed to model any image distortion via a combination of three factors: loss of correlation, luminance distortion, and contrast distortion.

The main reason to use this quality index as a loss function is its strong ability to measure the structural distortions existing in the images with haze. It is important to bear in mind that because the signals of the images are non-stationary it is preferable to evaluate the quality of the images by locally measuring their statistical characteristics and then combine them all collectively in a single measurement of image quality.

If there are a total of M steps, at the j-th step the local quality index Q_j is computed, then the overall quality index is given by :

$$Q = \frac{1}{M}\sum_{j=1}^{M} Q_j. \tag{6.4}$$

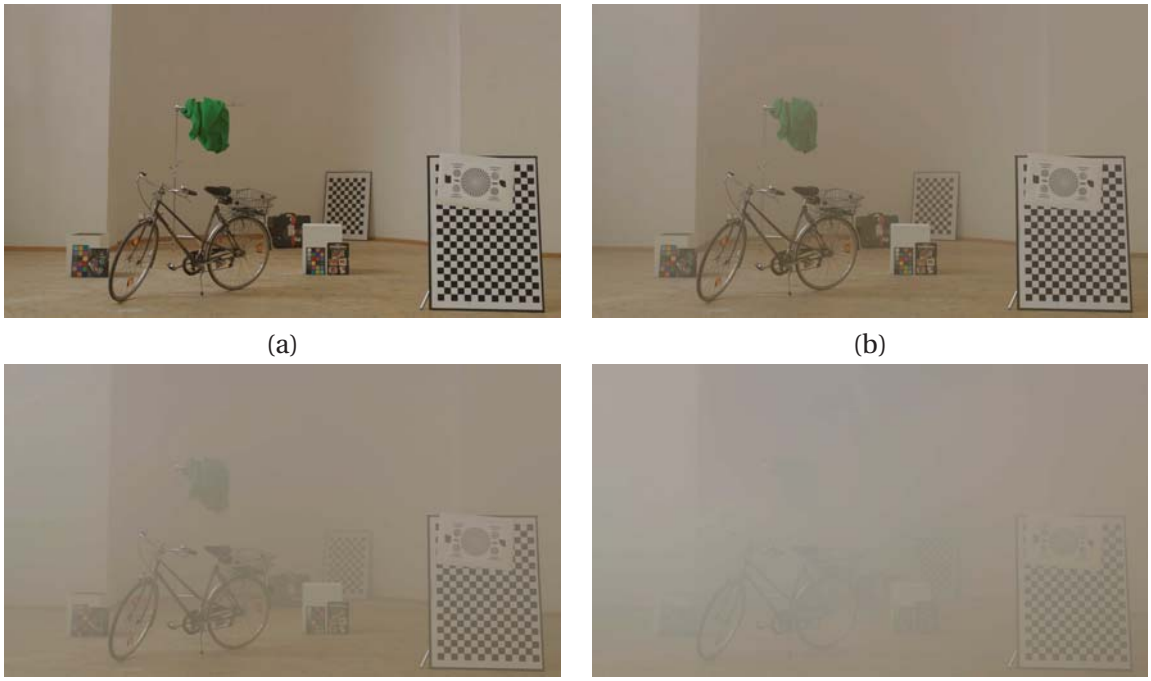
Hence, we can formulate the quality loss function as:

$$\mathscr{L}_{Q} = \frac{1}{M}\sum_{j=1}^{M} (1 - Q_j). \tag{6.5}$$

The final loss function ($\mathscr{L}_{final}$) used in this work is the accumulative weighted sum of the individual adversarial, intensity, structural and quality loss functions:

$$\mathscr{L}_{final} = 0.40\mathscr{L}_{Adversarial} + 0.25\mathscr{L}_{Intensity} + 0.20\mathscr{L}_{SSIM} + 0.15\mathscr{L}_{Q}. \tag{6.6}$$

The proportion assigned to each loss has been defined based on the variability of the values obtained by each of the losses during the training process.



(a) (b)

(c) (d)

Figure 6.3 – Set of RGB images from an indoor environment: (*a*) Ground truth image; (*b*), (*c*) and (*d*) Real images with different haze levels.

Table 6.1 – Angular Errors (AE), Mean Squared Errors (MSE), Structural Similarities (SSIM) and Image Quality Index(Q Index) obtained with the first proposed Stacked Conditional GAN architecture by using different loss functions (SSIM and Q index values, the bigger the better).

Training	AE		MSE		SSIM		Q Index	
	Light Haze	*Dense Haze*	*Urban*	*Dense Haze*	*Light Haze*	*Dense Haze*	*Light Haze*	*Dense Haze*
Proposed 1st. Prop. App. with $\mathscr{L}_{Adversarial} + \mathscr{L}_{Intensity}$	7.18	7.11	21.96	23.75	0.72	0.69	0.62	0.59
Proposed 1st. Prop. App. with $\mathscr{L}_{Adversarial} + \mathscr{L}_{SSIM}$	7.12	7.03	20.97	20.74	0.78	0.72	0.64	0.61
Proposed 1st. Prop. App. with $\mathscr{L}_{Adversarial} + \mathscr{L}_{Intensity} + \mathscr{L}_{SSIM}$	6.32	6.24	19.65	20.08	0.80	0.77	0.68	0.66
Proposed 1st. Prop. App. with $\mathscr{L}_{final}$	5.95	6.12	18.74	19.21	0.84	0.80	0.71	0.68

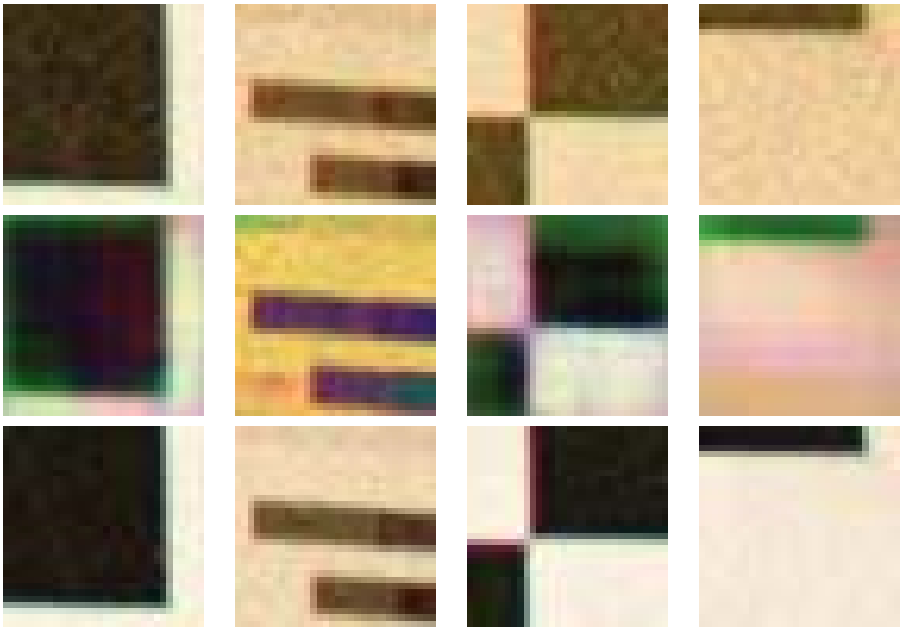


Figure 6.4 – Results from light hazed category with the first proposed approach: *(1st row)* Haze patches; *(2nd row)* Unhaze patches (Loss Function: $\mathscr{L}_{final}$); *(3rd row)* Ground truth images.

The dense stacked conditional GAN network proposed has been trained using Stochastic AdamOptimazer since it is well suited for problems that are large in terms of data and/or parameters, very appropriate for non-stationary objectives and problems with very noisy/or sparse gradients. Also, the Hyper-parameters have intuitive interpretation and typically require less tuning, prevents overfitting and leads to convergence faster. Furthermore, it is computationally efficient, has little memory requirements, is invariant to the diagonal rescaling of the gradients. The image dataset was normalized in a (-1,1) range. The following hyper-parameters were used during the training process: learning rate 0.00004 for the generator and 0.00003 for the discriminator networks respectively; epsilon = 1e-08; exponential decay rate for the 1st moment momentum 0.4 for discriminator and 0.3 for the generator; weight initializer with a standard deviation of 0.04582; $l1$ weight regularizer; weight decay 1e-2; leak relu 0.21 and patch's size of 64×64.

The learning architecture is conformed by convolutional, de-convolutional, relu, leak-relu, fully connected and activation functions $tanh$ and $sigmoid$ for generator and discriminator networks respectively. Additionally, every layer of the model uses batch normalization for training any type of mapping to prevent underfitting. It is very important to maintain the spatial information in the generator model, there is not pooling and drop-out layers and only the stride of 1 is used to avoid downsize the image shape. To prevent overfitting a $l1$ regularization term (λ) has been added in the generator model, this regularization has the particularity that the weights matrix ends up using only a small subset of their most important

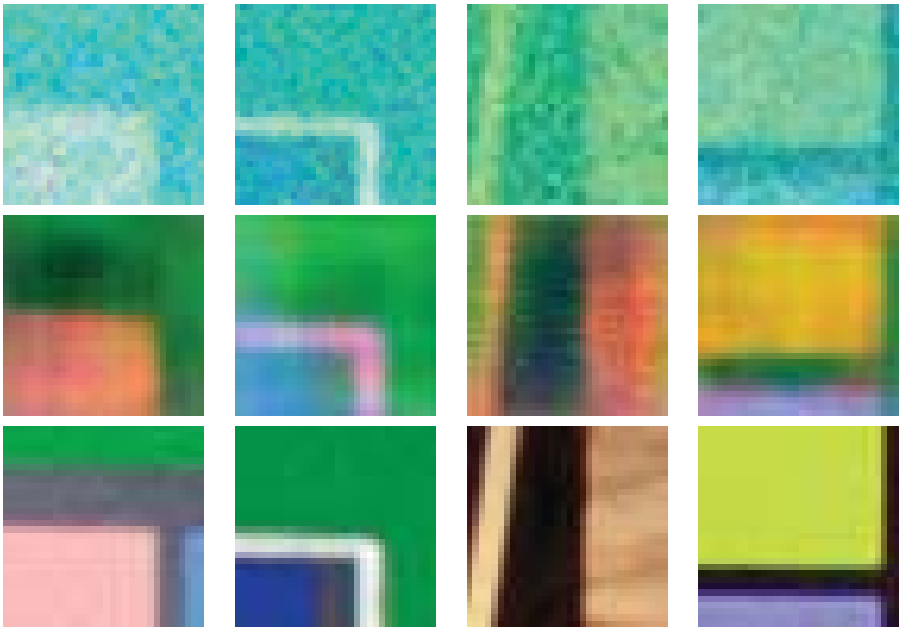


Figure 6.5 – Results from dense hazed category with the first proposed approach: *(1st row)* Haze patches; *(2nd row)* Unhaze patches (Loss Function: $\mathscr{L}_{final}$); *(3rd row)* Ground truth images.

inputs and become quite resistant to noise in the inputs. Additionally, the architecture includes a dense model implemented by the first and bottom layers in the model to increase the generalization and obtain more optimization of the learning process.

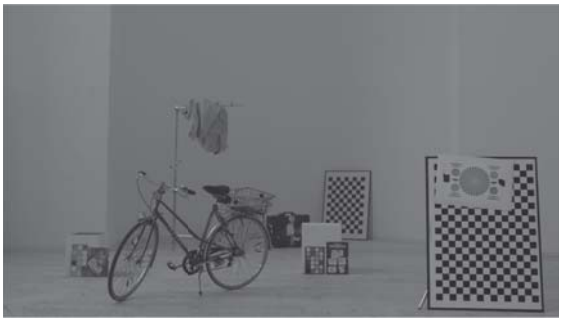
The generator (G) and discriminator (D) are both feedforward deep neural networks that play a min-max game between one another. The generator takes as input on each channel the hazy and NIR image and it is transformed into the form of the data we are interested in imitating, in this case a RGB clear image. The discriminator takes as an input a set of data, either real image (z) or generated image ($G(z)$), and produces a probability of that data being real ($P(z)$). The discriminator is optimized in order to increase the likelihood of giving a high probability to the real data (the ground truth given image) and a low probability to the fake generated data (wrongly clarified haze image), as introduced in [70]; thus, the dense conditional discriminator network is updated as follow:

$$\nabla\theta_d \frac{1}{m}\sum_{i=1}^{m}[\log D(x^{(i)}) + \log(1 - D(G(y^{(i)}, z^{(i)})))], \tag{6.7}$$

where m is the number of patches in each batch, x is the ground truth image, y is the image without haze (RGB) generated by the network and z is the random Gaussian sampled noise. The weights of the discriminator network (D) are updated by ascending its stochastic gradient. On the other hand, the generator is then optimized in order to increase the probability of the generated data being highly rated, it is updated as follow:

$$\nabla\theta_g \frac{1}{m}\sum_{i=1}^{m}\log(1 - D(G(y^{(i)}, z^{(i)}))), \tag{6.8}$$

where m is the number of samples in each batch, y is the image without haze (RGB) generated by the network and z is the random Gaussian sampled noise. Like in the previous case, the weights of the generator network (G) are updated by descending its stochastic gradient.



(a)

(b)

(c)

Figure 6.6 – Set of images from [65]: (*a*) NIR image; (*b*) Hazed image; and (*c*) Ground truth image.

Table 6.2 – Angular errors (AE), mean squared errors (MSE), structural similarities (SSIM) and image quality index (Q Index) obtained with the second proposed multi dense stacked conditional GAN architecture by using cross-spectral images (SSIM and QIndex values, the bigger the better) and the approach in the first proposed approach.

Approach	AE		MSE		SSIM		Q Index	
	Light Haze	*Dense Haze*	*Light Haze*	*Dense Haze*	*Light Haze*	*Dense Haze*	*Light Haze*	*Dense Haze*
Stacked CGAN presented in 1st. Prop. App. with $\mathcal{L}_{Adversarial} + \mathcal{L}_{Intensity}$	7.18	7.11	21.96	23.75	0.72	0.69	0.62	0.59
2nd. Prop. App. cross-spectral dense stacked CGAN with $\mathcal{L}_{Adversarial} + \mathcal{L}_{Intensity}$	6.46	6.11	19.86	21.74	0.80	0.73	0.68	0.64
2nd. Prop. App. cross-spectral dense stacked CGAN with $\mathcal{L}_{Adversarial} + \mathcal{L}_{SSIM}$	7.12	7.03	20.97	20.74	0.78	0.72	0.64	0.61
2nd. Prop. App. cross-spectral dense stacked CGAN with $\mathcal{L}_{Adversarial} + \mathcal{L}_{SSIM}$	6.28	5.97	19.19	20.33	0.86	0.84	0.76	0.69
2nd. Prop. App. cross-spectral dense stacked CGAN with $\mathcal{L}_{Adversarial} + \mathcal{L}_{Intensity} + \mathcal{L}_{SSIM}$	6.32	6.24	19.65	20.08	0.80	0.77	0.68	0.66
Proposed cross-spectral dense stacked CGAN with $\mathcal{L}_{Adversarial} + \mathcal{L}_{Intensity} + \mathcal{L}_{SSIM}$	6.13	5.93	18.47	19.85	0.89	0.81	0.78	0.71
2nd. Prop. App. cross-spectral dense stacked CGAN with $\mathcal{L}_{final}$	5.95	6.12	18.74	19.21	0.84	0.80	0.71	0.68
2nd. Prop. App. cross-spectral dense stacked CGAN with $\mathcal{L}_{final}$	5.10	5.65	17.92	18.74	0.92	0.86	0.82	0.77

6.3 Experimental Results

To evaluate both haze removal proposals, the dataset obtained from [65] has been used. For the experiments, from all these images 85000 pairs of patches of (32×32 pixels) have been cropped both, in the hazed images as well as in the corresponding clear RGB and NIR images. Additionally, 8500 pairs of patches have been also generated for validation. On average, every training process took about 60 hours using a 3.2 eight-core processor with 16GB of memory with an NVIDIA TITAN V GPU.

The quantitative evaluation of both proposed approaches, consists of measuring several metrics with the results obtained with the first and second proposed Stacked GAN approaches when different combinations of the proposed loss functions where considered; one of the metrics consists of measuring at every pixel the angular error (AE) between the obtained result ($RGBo_{i,j}$) and the corresponding ground truth value ($RGBg_{i,j}$). AE is included since this measure is quite similar to the human visual perception system, [33]—AE is probably the most widely used performance measure in color constancy research. Additionally, the Mean Squared Error (MSE), the Quality Index (QIndex) and the Structural Similarity (SSIM) metrics are also considered in this quantitative evaluation. On the contrary to AE and MSE, which can be considered as pixel level evaluation metrics, the SSIM and QIndex are methods for evaluating the perceived quality of the results. The SSIM provides a measurement of local image quality over space while QIndex models the image distortion relative to the reference image as a combination of three factors: loss of correlation, luminance distortion, and contrast distortion. These metrics have a high degree of sensitivity to measure to image degradations, therefore, they are the more appropriate to this type of quantitative evaluation.

The first proposed approach for remove haze from RGB image has been evaluated using real hazed images and their corresponding clear RGB representations obtained from [65]. Figure 6.3 presents four images from this dataset, where ground truth image can be appreciated on (*a*) while different real hazed images are depicted on (*b*), (*c*) and (*d*). See more details about data set generation in [65].

Some patches, with the corresponding result obtained with the first proposed approach, are depicted in Fig. 6.4 and Fig. 6.5; just for making easier the evaluation of results from the proposed approach patches have been split up into *Light Haze* and *Dense Haze*.

With the metrics mentioned above combinations of the different loss functions are evaluated, results are provided in Table 6.1. It can be appreciated that in all the cases the results obtained with the final loss proposed with Stacked Conditional GAN are better than those obtained with the other combination of losses, because are not based solely on the difference of the information of the pixels, they are based on the high-level characteristics of the images for which they are able to reconstruct better the fine details in comparison with the methods trained only by distance value of pixels. Just as illustrations, a few RGB images from *Light Haze* and *Dense Haze* categories, generated with the proposed Stacked GAN network, are depicted in Fig. 6.4 and Fig. 6.5 for qualitative evaluation.

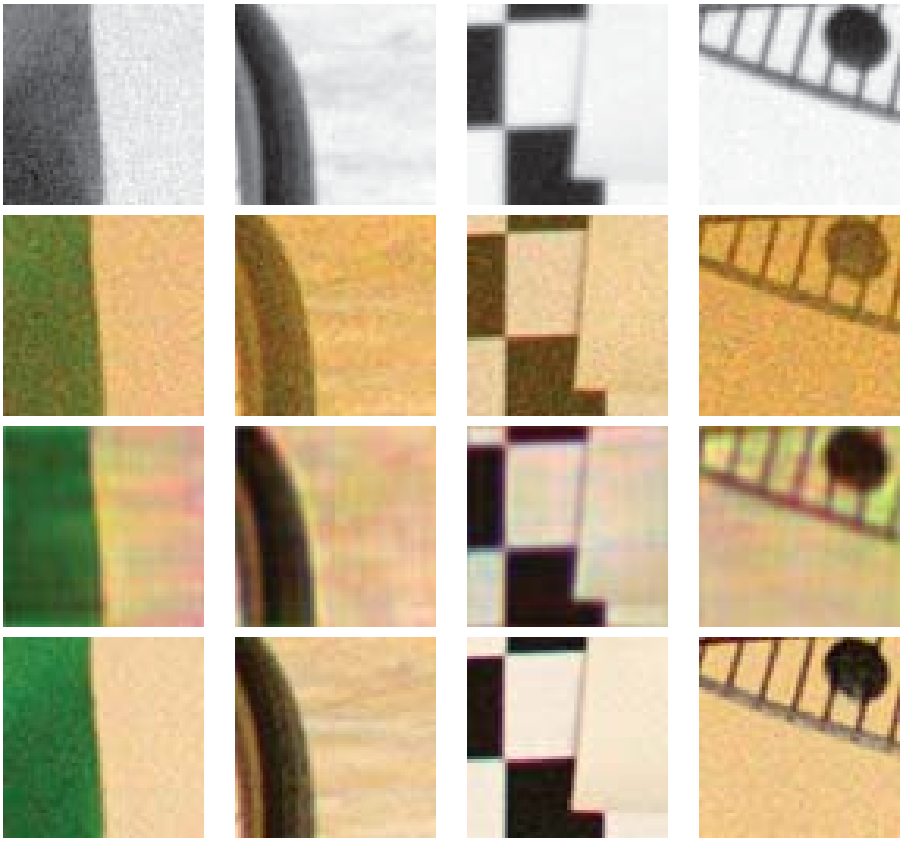


Figure 6.7 – *(1st row)* NIR patches; *(2nd row)* Light hazed patches; *(3rd row)* Results from the second proposed approach, (Loss function: $\mathscr{L}_{final}$); *(4th row)* Ground truth images.

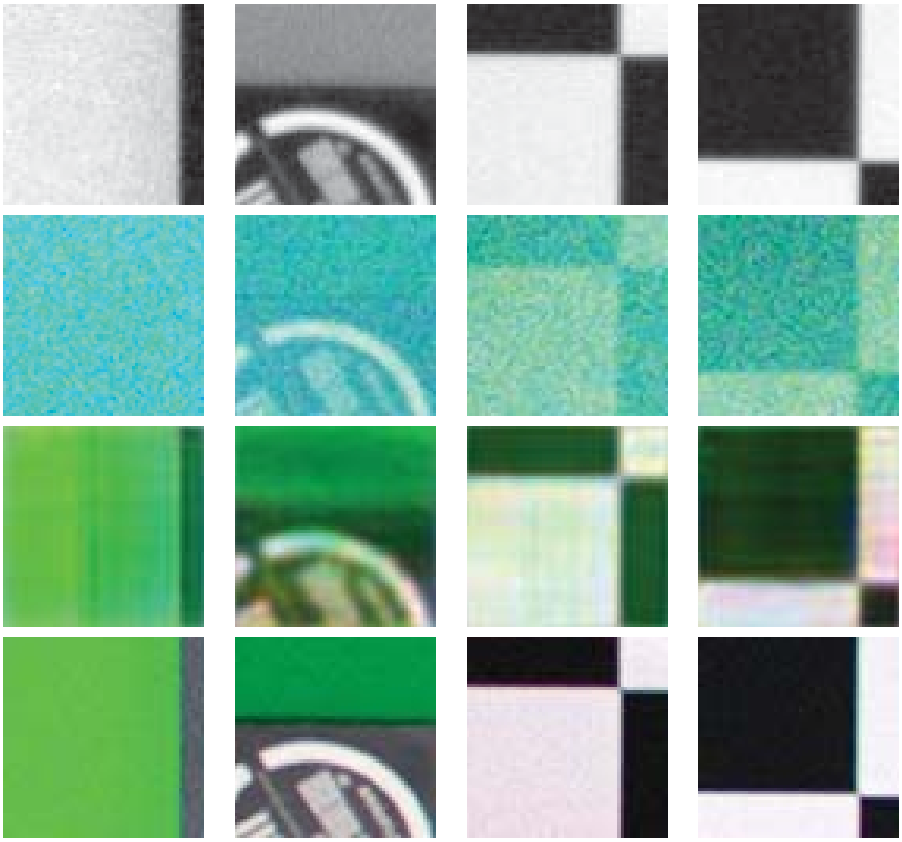


Figure 6.8 – Results from dense hazed category: (*1st row*) NIR patches; (*2nd row*) Haze patches; (*3rd row*) Results with second proposed approach. (Loss function: $\mathscr{L}_{final}$); (*4th row*) Ground truth images.

The second proposed architecture has been evaluated using real hazed images and their corresponding clear RGB and NIR infrared representations obtained from [65]. Figure 6.6 presents a set of images from this dataset. The corresponding results obtained with the proposed approach are depicted in Figs. 6.7 and 6.8; just for making easier the evaluation of results from the second proposed approach patches have been split up into *Light Haze* and *Dense Haze* categories.

With the metrics mentioned above combinations of the different loss functions are evaluated, results are provided in Table 6.2. It can be appreciated that in all the cases the results obtained with the final loss proposed with dense Stacked Conditional GAN are better than those obtained with the first proposed approach. Also, these losses, being perfectly differentiable, allow for better optimization of the network, thus accelerating the convergence process.

6.4 Conclusion

This chapter tackles the challenging problem of generating clear RGB representations from hazed images and their corresponding NIR images. In the current work two supervised approaches have been designed by using a novel dense stacked cross-spectral conditional generative adversarial network, both approaches use the stacked GAN. However, the second proposed approach has obtained the best results, after including multiple dense connections between the layers in the architecture, to facilitate the propagation of the characteristics and minimize the number of model parameters. The results have shown that in most cases the network can obtain reliable clear RGB representations. The proposed approaches have as a limitation the need of having ground truth images without haze for training.

Chapter 7

Conclusions and Future Work

7.1 Conclusions

This thesis has [50] implemented innovative techniques in the field of computer vision using images of the visible and near-infrared spectrum, using deep learning through convolutional networks, especially generative adversarial networks, which are specialists in creating information through the antagonistic game of two convolutional networks that interact with each other. In this dissertation, with this type of convolutional networks, different techniques have been created to solve challenging problems, like detect the similarity of patches of different spectra (visible-infrared), colorize images of the near infrared spectrum, estimation of vegetation index (NDVI) and the haze removal present on RGB images using NIR images. For all these techniques different variants of the GAN's networks, such as standard, conditional, stacked and cyclic have been used. It should be mentioned that together with the implementation of adversarial network models, the use of multiple loss functions has been proposed to improve the generalization of the models and increase the effectiveness of the model.

Chapter 2 summarizes the extensive literature review of the different techniques that different authors have developed to address the issues raised in this thesis. This review has allowed us to analyze the different approaches presented in the most relevant conferences and determine their advantages and limitations, so that this knowledge is used and new methods or techniques can be formulated that solve the same problem addressed in the thesis in a better way and thus contribute to the state of the art.

Chapter 3 addresses the problem of determining the similarity of the images of different spectra firstly, a supervised scheme has been proposed where the images have to be paired. This approach has obtained better results than traditional methods such as SIFT, SURF, HARRIS, etc. However, to overcome the limitation of having the images aligned, that is, that they are from the same scene, a new approach based on meta-learning has been proposed in this thesis, it is based on distance metrics, the model is capable to learn object categories from a few examples, and do not require high computational power. The designed architecture can synthesize an effective embedded representation that allows the model to learn new classes

from existing information about different previously learned classes.

In Chapter 4 the colorization of NIR images using CCN has been tackled. Different models have been proposed. One of the most elaborated model includes a novel Stacked Dense Conditional Generative Adversarial Network. The experimental results have shown that the proposed approach generates good quality colored images of different scenes (i.e., content). The principal reason to use images from the NIR spectrum is because of its independence from the brightness and color of the targets, which has potential benefits, including non-visible illumination requirements. Another contribution of the proposed model is the implementation of multiple loss functions that not only focus on the intensity of the pixels, but also on the texture, since the similarity of the structures present in the images is measured, which allows obtaining more clear colorized images, in addition, to preserve the color tones correctly.

Chapter 5 presents supervised and unsupervised techniques to estimate the NDVI vegetation index. Within the supervised scheme two approaches are proposed; the first approach that learns from NIR images with a triple learning layer, the second one is an approach that learns from a grayscale image with a single level of learning layer, both approaches are based on GAN networks. These models have the limitation of relying on registered images, that is, the model must be feed with only paired image patches, although good estimates of the NDVI index have been obtained. To overcome the limitations of relying on registered images with paired images, working with patches and relying on sensors sensitive to the near infrared spectrum, an unsupervised approach has been proposed. It works with visible spectrum images (red channel) and NDVI index images, this approach performs translation (mapping) of the images (visible-NDVI) and is based on a Cyclic GAN network with multiple loss functions for an efficient and fast generalization of the model. The results obtained by this unsupervised approach are better than those obtained with the supervised schemes previously implemented. This NDVI index estimation makes it possible the monitoring of the states of vegetation and helps to determine the best action to be performed to monitor the stages of development and biomass of crops and plants, to forecast and improve their yields.

Chapter 6, addresses the problem of removing the haze from the images, through the implementation of two supervised techniques, the approach that showed the best results was the conditional stacked GAN, densely connected, optimized through multiple loss functions. This approach has as a limitation the need of having ground truth paired images without haze for the training process. The network is fed with the hazed and its corresponding NIR image, which is concatenated at the beginning and end of the architecture layers and serves to improve the quality of the obtained clear image. With this technique, valuable information on the objects present in the hazed images can be obtained, which allows a better scene understanding. Therefore, processes such as detection, segmentation or recognition of objects, among others, will benefit from the reduction of the atmosphere effect that reduce the visibility.

7.2 Future Work

This thesis tackles challenging problems related to cross-spectral image processing, like patching similarity, near infrared image colorization, image haze removal, NDVI estimated index, for which several deep learning models have been implemented, mainly focusing on adversarial generative models. The results obtained with the model of cross-spectral image similarity using meta-learning have shown that it is possible to even with a few shot samples to obtain a performance quite similar to the state of the art, as well as it is shown that outperforms classical SIFT feature based descriptors.

As future work, other architectures based on reinforced continual learning, meta-learning, like learning models based on Model-Agnostic Meta-Learning (MAML) to enable the model to be capable to update their weights to perform fast adaptation, new function losses, and normalization techniques will be considered to improve the results already obtained.

For NIR colorization, future work will be focused on evaluating other network architectures, like variational auto-encoders and cycle-consistent adversarial networks, which have shown appealing results in recent works. Additionally, these models will be evaluated with other loss functions to improve and accelerate the training process. Finally, increasing the number of images to train will be considered in order to increase the diversity of colors.

Estimation of vegetation indexes is crucial to help farmers to improve the quality of their crops, making possible the early identification of diseases or the infection of insects that affect the health of the plants. These techniques, together with the use of unmanned aerial vehicles (UAVs), are used to accurately apply the better solutions. Future approaches addressed reinforced learning techniques so that the model is able to learn to estimate different types of vegetation indices in a progressive manner, so that it re-uses prior learning for the new tasks applied to the model, in this case, estimate new kind of vegetation indexes.

Haze removal is a challenging problem focused to improve the scene understanding. As a future work, actually, as work in progress, a cyclic GAN architecture will be proposed, but feed it with unpaired images (RGB hazed and clear images) to perform the mapping between them in order to remove the haze to obtain the corresponding clear image; also, new loss function will be implemented to improve the optimization and convergence of the model.

The results obtained in the experiments in each of the techniques implemented with deep learning in this thesis show that it is possible to use convolutional networks, especially generative adversarial networks to create information in the (visible) spectrum from the information of another spectrum (near infrared) or to make comparisons of images of different spectra. Therefore it is verified that the characteristics of the images of different spectra can be used to complement the missing information (merge) or to generate new information (hallucinate) and thus be able to apply these techniques to solve problems proposed in the field of computer vision. However, as mentioned above, new architectures based on continual learning or meta-learning can be proposed to address the problems previously described, in order to obtain better results. Also, new loss functions can be implemented in previously

challenge problems to obtain more generalized models, accelerate the convergence and obtain more accurate results. Another alternative may be to implement new data augmentation schemes to improve training processes and results.

Bibliography

[1] Dmitry Abulkhanov, Ivan Konovalenko, Dmitry Nikolaev, Alexey Savchik, Evgeny Shvets, and Dmitry Sidorchuk. Neural network-based feature point descriptors for registration of optical and sar images. In *10th International Conference on Machine Vision*, volume 10696, page 106960L. International Society for Optics and Photonics, 2018.

[2] Telmo Adão, Jonáš Hruška, Luís Pádua, José Bessa, Emanuel Peres, Raul Morais, and Joaquim Sousa. Hyperspectral imaging: A review on uav-based sensors, data processing and applications for agriculture and forestry. *Remote Sensing*, 9(11):1110, 2017.

[3] Cristhian A Aguilera, Francisco J Aguilera, Angel D Sappa, Cristhian Aguilera, and Ricardo Toledo. Learning cross-spectral similarity measures with deep convolutional neural networks. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops*, pages 1–9, 2016.

[4] Cristhian A. Aguilera, Fernando Barrera, Felipe Lumbreras, Angel D Sappa, and Ricardo Toledo. Multispectral image feature points. *Sensors*, 12(9):12661–72, January 2012.

[5] Cristhian A. Aguilera, Angel D. Sappa, Cristhian Aguilera, and Ricardo Toledo. Cross-spectral local descriptors via quadruplet network. *Sensors*, 17(4):873, 2017.

[6] Cristhian. A. Aguilera, Angel. D. Sappa, and Ricardo Toledo. LGHD: A feature descriptor for matching across non-linear intensity variations. In *Image Processing, 2015 IEEE International Conference*, pages 178–181, September 2015.

[7] Saeed Al Mansoori, Alavi Kunhu, and Hussain Al Ahmad. Automatic palm trees detection from multispectral uav data using normalized difference vegetation index and circular hough transform. In *High-Performance Computing in Geoscience and Remote Sensing VIII*, volume 10792, page 1079203. International Society for Optics and Photonics, 2018.

[8] Pablo Fernández Alcantarilla, Adrien Bartoli, and Andrew J. Davison. Kaze features. In *Proceedings of the 12th European Conference on Computer Vision - Volume Part VI*, European Conference on Computer Vision, pages 214–227, 2012.

[9] Aayush Bansal, Shugao Ma, Deva Ramanan, and Yaser Sheikh. Recycle-gan: Unsupervised video retargeting. In *The European Conference on Computer Vision*, September 2018.

[10] Ruud Barth, Jochen Hemming, and Eldert J van Henten. Improved part segmentation performance by optimising realism of synthetic images using cycle generative adversarial networks. *arXiv preprint arXiv:1803.06301*, 2018.

[11] Herbert Bay, Andreas Ess, Tinne Tuytelaars, and Luc Van Gool. Speeded-up robust features (surf). *Comput. Vis. Image Underst.*, 110(3):346–359, June 2008.

[12] Andrés Berger, Guillermo Ettlin, Christopher Quincke, and Pablo Rodríguez-Bocca. Predicting the normalized difference vegetation index (ndvi) by training a crop growth model with historical data. *Computers and Electronics in Agriculture*, 161:305–311, 2019.

[13] Dana Berman, Shai Avidan, et al. Non-local image dehazing. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 1674–1682, 2016.

[14] Matthew Brown and Sabine Süsstrunk. Multi-spectral SIFT for scene category recognition. In *Computer Vision and Pattern Recognition (CVPR), 2011 IEEE Conference on*, pages 177–184. IEEE, 2011.

[15] Clemens-Alexander Brust, Sven Sickert, Marcel Simon, Erik Rodner, and Joachim Denzler. Convolutional patch networks with spatial prior for road detection and urban scene understanding. *arXiv preprint arXiv:1502.06344*, 2015.

[16] Bolun Cai, Xiangmin Xu, Kui Jia, Chunmei Qing, and Dacheng Tao. Dehazenet: An end-to-end system for single image haze removal. *IEEE Transactions on Image Processing*, 25(11):5187–5198, 2016.

[17] Toby Carlson and David A Ripley. On the relation between ndvi, fractional vegetation cover, and leaf area index. *Remote sensing of Environment*, 62(3):241–252, 1997.

[18] Emre Celebi, Michela Lecca, and Bogdan Smolka. *Color Image and Video Enhancement*, volume 4. Springer, 2015.

[19] Tongbo Chen, Yan Wang, Volker Schillings, and Christoph Meinel. Grayscale image matting and colorization. In *Asian Conference on Computer Vision*, November 2004.

[20] Xi Cheng, Li Zhang, and Yefeng Zheng. Deep similarity learning for multimodal medical images. *Computer Methods in Biomechanics and Biomedical Engineering: Imaging & Visualization*, 6(3):248–252, 2018.

[21] Sumit Chopra, Raia Hadsell, and Yann LeCun. Learning a similarity metric discriminatively, with application to face verification. In *IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, volume 1, pages 539–546. IEEE, 2005.

[22] Snehal S Dahikar and Sandeep V Rode. Agricultural crop yield prediction using artificial neural network approach. *International Journal of Innovative Research in Electrical, Electronics, Instrumentation and Control Engineering*, 2(1):683–686, 2014.

[23] Shaobo Dang, Yanning Zhang, and Dong Gong. A patch-based non-local means method for image denoising. In *International Conference on Intelligent Science and Intelligent Data Engineering*, pages 582–589. Springer, 2012.

[24] Aditya Deshpande, Jiajun Lu, Mao-Chuang Yeh, and David Forsyth. Learning diverse image colorization. *arXiv preprint arXiv:1612.01958*, 2016.

[25] Salvatore Filippo Di Gennaro, Fulvia Rizza, Franz Werner Badeck, Andrea Berton, Stefano Delbono, Beniamino Gioli, Piero Toscano, Alessandro Zaldei, and Alessandro Matese. Uav-based high-throughput phenotyping to discriminate barley vigour with visible and near-infrared vegetation indices. *International journal of remote sensing*, 39(15-16):5330–5344, 2018.

[26] Jing Dong, Byron Boots, Frank Dellaert, Ranveer Chandra, and Sudipta Sinha. Learning to align images using weak geometric supervision. In *International Conference on 3D Vision (3DV)*, pages 700–709. IEEE, 2018.

[27] Weisheng Dong, Xin Li, Lei Zhang, and Guangming Shi. Sparsity-based image denoising via dictionary learning and structural clustering. In *Computer Vision and Pattern Recognition, IEEE Conference*, pages 457–464. IEEE, 2011.

[28] M. Fernanda Dreccer, Gemma Molero, Carolina Rivera-Amado, Carus John-Bejai, and Zoe Wilson. Yielding to the image: How phenotyping reproductive growth can assist crop improvement and production. *Plant science*, 282:73–82, 2019.

[29] T Duan, SC Chapman, Y Guo, and B Zheng. Dynamic monitoring of ndvi in wheat agronomy and breeding trials using an unmanned aerial vehicle. *Field Crops Research*, 210:71–80, 2017.

[30] Raanan Fattal. Single image dehazing. *ACM transactions on graphics (TOG)*, 27(3):72, 2008.

[31] Adrian Galdran, Javier Vazquez-Corral, David Pardo, and Marcelo Bertalmío. Fusion-based variational image dehazing. *IEEE Signal Processing Letters*, 24(2):151–155, 2017.

[32] Yann Gavet, Johan Debayle, and Jean-Charles Pinoli. The color logarithmic image processing (colip) antagonist space. In *Color Image and Video Enhancement*, pages 155–182. Springer, 2015.

[33] Arjan Gijsenij, Theo Gevers, and Marcel P Lucassen. A perceptual comparison of distance measures for color constancy algorithms. In *European Conference on Computer Vision*, pages 208–221. Springer, 2008.

[34] Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. Generative adversarial nets. In *Advances in neural information processing systems*, pages 2672–2680, 2014.

[35] Sergio Guadarrama, Ryan Dahl, David Bieber, Mohammad Norouzi, Jonathon Shlens, and Kevin Murphy. Pixcolor: Pixel recursive colorization. *arXiv preprint arXiv:1705.07208*, 2017.

[36] Raia Hadsell, Sumit Chopra, and Yann LeCun. Dimensionality reduction by learning an invariant mapping. In *IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, volume 2, pages 1735–1742. IEEE, 2006.

[37] Xufeng Han, Thomas Leung, Yangqing Jia, Rahul Sukthankar, and Alexander. C. Berg. Matchnet: Unifying feature and metric learning for patch-based matching. In *IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, 2015.

[38] Kaiming He and Jian Sun. Image completion approaches using the statistics of similar patches. *IEEE transactions on pattern analysis and machine intelligence*, 36(12):2423–2435, 2014.

[39] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778, 2016.

[40] Maarten Hogervorst and Alexander Toet. Improved color mapping methods for multi-band nighttime image fusion. *Journal of Imaging*, 3(3):36, 2017.

[41] Hiroto Honda, Radu Timofte, and Luc Van Gool. Make my day-high-fidelity color denoising with near-infrared. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops*, pages 82–90, 2015.

[42] Gao Huang, Zhuang Liu, and Kilian Q. Weinberger. Densely connected convolutional networks. *IEEE Conference on Computer Vision and Pattern Recognition*, pages 2261–2269, 2017.

[43] Hui Huang, Genyun Sun, Jinchang Ren, Jun Rang, Aizhu Zhang, and Yanling Hao. Spectral-spatial topographic shadow detection from sentinel-2a msi imagery via convolutional neural networks. In *IGARSS IEEE International Geoscience and Remote Sensing Symposium*, pages 661–664. IEEE, 2018.

[44] Xun Huang, Yixuan Li, Omid Poursaeed, John Hopcroft, and Serge Belongie. Stacked generative adversarial networks. *arXiv preprint arXiv:1612.04357*, 2016.

[45] Satoshi Iizuka, Edgar Simo-Serra, and Hiroshi Ishikawa. Let there be Color!: Joint End-to-end Learning of Global and Local Image Priors for Automatic Image Colorization with Simultaneous Classification. *ACM Transactions on Graphics (Proc. of SIGGRAPH)*, 35(4), 2016.

[46] Seyed Mehdi Iranmanesh, Ali Dabouei, Hadi Kazemi, and Nasser M Nasrabadi. Deep cross polarimetric thermal-to-visible face recognition. In *International Conference on Biometrics*, pages 166–173. IEEE, 2018.

[47] Revital Ironi, Daniel Cohen-Or, and Dani Lischinski. Colorization by example. In *Rendering Techniques*, pages 201–210, 2005.

[48] Phillip Isola, Jun-Yan Zhu, Tinghui Zhou, and Alexei A Efros. Image-to-image translation with conditional adversarial networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1125–1134, 2017.

[49] Jehoiada Jackson, Oluwasanmi Ariyo, Kingsley Acheampong, Maxwell Boakye, Enoch Frimpong, Eric Ashalley, and Yunbo Rao. Hybrid single image dehazing with bright channel and dark channel priors. In *Image, Vision and Computing, 2nd International Conference on*, pages 381–385. IEEE, 2017.

[50] Muhammad Abdullah Jamal and Guo-Jun Qi. Task agnostic meta-learning for few-shot learning. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 11719–11727, 2019.

[51] Mingye Ju, Dengyin Zhang, and Xuemei Wang. Single image dehazing via an improved atmospheric scattering model. *The Visual Computer*, 33(12):1613–1625, 2017.

[52] Amanda Heemann Junges, Denise Cybis Fontana, and Cristian Scalvi Lampugnani. Relationship between the normalized difference vegetation index and leaf area in vineyards. *Bragantia*, 78(2):297–305, 2019.

[53] Andreas Kamilaris and Francesc X Prenafeta-Boldú. Deep learning in agriculture: A survey. *Computers and electronics in agriculture*, 147:70–90, 2018.

[54] Seungryong Kim, Dongbo Min, Bumsub Ham, Seungchul Ryu, Minh N Do, and Kwanghoon Sohn. Dasc: Dense adaptive self-correlation descriptor for multi-modal and multi-spectral correspondence. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2103–2112, 2015.

[55] Johannes Kopf, Boris Neubert, Billy Chen, Michael Cohen, Daniel Cohen-Or, Oliver Deussen, Matt Uyttendaele, and Dani Lischinski. Deep photo: model-based photograph enhancement and viewing. *ACM transactions on graphics*, 27(5), 2008.

[56] Christian Ledig, Lucas Theis, Ferenc Huszár, Jose Caballero, Andrew Cunningham, Alejandro Acosta, Andrew Aitken, Alykhan Tejani, Johannes Totz, Zehan Wang, et al. Photo-realistic single image super-resolution using a generative adversarial network. *arXiv preprint*, 2016.

[57] Sungmin Lee, Seokmin Yun, Ju-Hun Nam, Chee Sun Won, and Seung-Won Jung. A review on dark channel prior based image dehazing algorithms. *EURASIP Journal on Image and Video Processing*, 2016(1):4, 2016.

[58] Boyi Li, Xiulian Peng, Zhangyang Wang, Jizheng Xu, and Dan Feng. An all-in-one network for dehazing and beyond. *arXiv preprint arXiv:1707.06543*, 2017.

[59] Wei Li, Jean-Daniel M Saphores, and Thomas W Gillespie. A comparison of the economic benefits of urban green spaces estimated with ndvi and with high-resolution land cover data. *Landscape and Urban Planning*, 133:105–117, 2015.

[60] Xiaoming Li and Qinchao Huang. Polarization filtering for automatic image dehazing based on contrast enhancement. In *Communication Software and Networks (ICCSN),IEEE 9th International Conference on*, pages 1266–1271. IEEE, 2017.

[61] Matthias Limmer and Hendrik Lensch. Infrared colorization using deep convolutional neural networks. *arXiv preprint arXiv:1604.02245*, 2016.

[62] Yansong Liu, Sankaranarayanan Piramanayagam, Sildomar T Monteiro, and Eli Saber. Dense semantic labeling of very-high-resolution aerial imagery and lidar with fully-convolutional neural networks and higher-order crfs. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops*, pages 76–85, 2017.

[63] David G. Lowe. Distinctive image features from scale-invariant keypoints. *Int. J. Comput. Vision*, 60(2):91–110, November 2004.

[64] Huimin Lu, Yujie Li, Shota Nakashima, and Seiichi Serikawa. Single image dehazing through improved atmospheric light estimation. *Multimedia Tools and Applications*, 75(24):17081–17096, 2016.

[65] Julia Lüthen, Julian Wörmann, Martin Kleinsteuber, and Johannes Steurer. A rgb/nir data set for evaluating dehazing algorithms. *Electronic Imaging*, 2017(12):79–87, 2017.

[66] Hussein Mahdi, NIDHAL EL ABBADI, and HIND RUSTUM. Single image de-hazing through improved dark channel prior and atmospheric light estimation. *Journal of Theoretical & Applied Information Technology*, 95(15), 2017.

[67] Xudong Mao, Qing Li, Haoran Xie, Raymond YK Lau, Zhen Wang, and Stephen Paul Smolley. Least squares generative adversarial networks. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 2794–2802, 2017.

[68] Antonio Marra, Massimiliano Gargiulo, Giuseppe Scarpa, and Raffaele Gaetano. Estimating the ndvi from sar by convolutional neural networks. In *IGARSS IEEE International Geoscience and Remote Sensing Symposium*, pages 1954–1957. IEEE, 2018.

[69] Krystian Mikolajczyk and Cordelia Schmid. A performance evaluation of local descriptors. *IEEE transactions on pattern analysis and machine intelligence*, 27(10):1615–1630, 2005.

[70] Mehdi Mirza and Simon Osindero. Conditional generative adversarial nets. *ArXiv*, abs-1411-1784, 2014.

[71] T. Mouats, N. Aouf, A.D. Sappa, C. Aguilera, and R. Toledo. Multispectral stereo odometry. *ITS*, PP(99):1–15, Sep 2014.

[72] Miguel Oliveira, Angel D Sappa, and Vitor Santos. Unsupervised local color correction for coarsely registered images. In *Computer Vision and Pattern Recognition*, pages 201–208. IEEE, June 2011.

[73] Miguel Oliveira, Angel Domingo Sappa, and Vitor Santos. A probabilistic approach for color correction in image mosaicking applications. *IEEE Transactions on Image Processing*, 24(2):508–523, 2015.

[74] Sudhanshu Sekhar Panda, Daniel P Ames, and Suranjan Panigrahi. Application of vegetation indices for agricultural crop yield prediction using neural network techniques. *Remote Sensing*, 2(3):673–696, 2010.

[75] Chulhee Park and Moon Gi Kang. Color restoration of rgbn multispectral filter array sensor images based on spectral decomposition. *Sensors*, 16(5):719, 2016.

[76] Chunlei Peng, Nannan Wang, Jie Li, and Xinbo Gao. Dlface: Deep local descriptor for cross-modality face recognition. *Pattern Recognition*, 90:161–171, 2019.

[77] Nathalie Pettorelli, Aliénor LM Chauvenet, James P Duffy, William A Cornforth, Alizée Meillere, and Jonathan EM Baillie. Tracking the effect of climate change on ecosystem functioning using protected areas: Africa as a case study. *Ecological Indicators*, 20:269–276, 2012.

[78] Peter Pinggera, Toby Breckon, and Horst Bischof. On cross-spectral stereo matching using dense gradient features. In *Computer Vision and Pattern Recognition, IEEE Conference on*, 2012.

[79] Tran Minh Quan, Thanh Nguyen-Duc, and Won-Ki Jeong. Compressed sensing mri reconstruction using a generative adversarial network with a cyclic loss. *IEEE transactions on medical imaging*, 37(6):1488–1497, 2018.

[80] Idan Ram, Michael Elad, and Israel Cohen. Image processing using smooth ordering of its patches. *IEEE transactions on image processing*, 22(7):2764–2774, 2013.

[81] Meenu Rani, Pavan Kumar, Prem Chandra Pandey, Prashant K Srivastava, BS Chaudhary, Vandana Tomar, and Vinay Prasad Mandal. Multi-temporal ndvi and surface temperature analysis for urban heat island inbuilt surrounding of sub-humid region: A case study of two geographical regions. *Remote Sensing Applications: Society and Environment*, 10:163–172, 2018.

[82] Ali Sharif Razavian, Hossein Azizpour, Josephine Sullivan, and Stefan Carlsson. Cnn features off-the-shelf: an astounding baseline for recognition. In *Computer Vision and Pattern Recognition Workshops,IEEE Conference on*, pages 512–519. IEEE, 2014.

[83] D Sushma Reddy and P Rama Chandra Prasad. Prediction of vegetation dynamics using ndvi time series data and lstm. *Modeling Earth Systems and Environment*, 4(1):409–419, 2018.

[84] Scott Reed, Zeynep Akata, Xinchen Yan, Lajanugen Logeswaran, Bernt Schiele, and Honglak Lee. Generative adversarial text to image synthesis. *arXiv preprint arXiv:1605.05396*, 2016.

[85] Wenqi Ren, Si Liu, Hua Zhang, Jinshan Pan, Xiaochun Cao, and Ming-Hsuan Yang. Single image dehazing via multi-scale convolutional neural networks. In *European conference on computer vision*, pages 154–169. Springer, 2016.

[86] Pablo Ricaurte, Carmen Chilán, Cristhian A Aguilera-Carrasco, Boris X Vintimilla, and Angel D Sappa. Feature point descriptors: Infrared and visible spectra. *Sensors*, 14(2):3690–3701, 2014.

[87] Rudolf Richter. Atmospheric correction of satellite data with haze removal including a haze/clear transition region. *Computers & Geosciences*, 22(6):675–681, 1996.

[88] Rafael E Rivadeneira, Patricia L Suárez, Angel D Sappa, and Boris X Vintimilla. Thermal image superresolution through deep convolutional neural network. In *International Conference on Image Analysis and Recognition*, pages 417–426. Springer, 2019.

[89] Adriana Romero, Carlo Gatta, and Gustau Camps-Valls. Unsupervised deep feature extraction for remote sensing image classification. *IEEE Transactions on Geoscience and Remote Sensing*, 54(3):1349–1362, 2016.

[90] J_W Rouse Jr, RH Haas, JA Schell, and DW Deering. Monitoring vegetation systems in the great plains with erts. *NASA SP-351*, 1974.

[91] Tim Salimans, Ian Goodfellow, Wojciech Zaremba, Vicki Cheung, Alec Radford, and Xi Chen. Improved techniques for training gans. In *Advances in neural information processing systems*, pages 2234–2242, 2016.

[92] Kate Sendin, Marena Manley, and Paul J Williams. Classification of white maize defects with multispectral imaging. *Food chemistry*, 243:311–318, 2018.

[93] Eli Shechtman and Michal Irani. Matching local self-similarities across images and videos. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 1–8. IEEE, 2007.

[94] X. Shen, L. Xu, Q. Zhang, and J. Jia. Multi-modal and Multi-spectral Registration for Natural Images. In *ECCV*, pages 309–324, Zurich, Switzerland, Sep 2014.

[95] Sarit Shwartz, Einav Namer, and Yoav Y Schechner. Blind haze separation. In *Computer Vision and Pattern Recognition, 2006 IEEE Computer Society Conference on*, volume 2, pages 1984–1991. IEEE, 2006.

[96] Edgar Simo-Serra, Eduard Trulls, Luis Ferraz, Iasonas Kokkinos, Pascal Fua, and Francesc Moreno-Noguer. Discriminative Learning of Deep Convolutional Feature Point Descriptors. In *Proceedings of the International Conference on Computer Vision*, Dec 2015.

[97] Sergii Skakun, Christopher O Justice, Eric Vermote, and Jean-Claude Roger. Transitioning from modis to viirs: an analysis of inter-consistency of ndvi data sets for agricultural monitoring. *International journal of remote sensing*, 39(4):971–992, 2018.

[98] Lingxiao Song, Man Zhang, Xiang Wu, and Ran He. Adversarial discriminative heterogeneous face recognition. In *Thirty-Second AAAI Conference on Artificial Intelligence*, 2018.

[99] Neha Soni, Enakshi Khular Sharma, Narotam Singh, and Amita Kapoor. Impact of artificial intelligence on businesses: from research, innovation, market deployment to future shifts in business models. *arXiv preprint arXiv:1905.02092*, 2019.

[100] Christian Szegedy, Wei Liu, Yangqing Jia, Pierre Sermanet, Scott Reed, Dragomir Anguelov, Dumitru Erhan, Vincent Vanhoucke, and Andrew Rabinovich. Going deeper with convolutions. *CoRR*, abs/1409.4842, 2014.

[101] Hu Tian and Fei Li. Autoencoder-based fabric defect detection with cross-patch similarity. In *16th International Conference on Machine Vision Applications (MVA)*, pages 1–6. IEEE, 2019.

[102] Paolo Tripicchio, Massimo Satler, Giacomo Dabisias, Emanuele Ruffaldi, and Carlo Alberto Avizzano. Towards smart farming and sustainable agriculture with drones. In *International Conference on Intelligent Environments*, pages 140–143. IEEE, 2015.

[103] Laura Ulsig, Caroline Nichol, Karl Huemmrich, David Landis, Elizabeth Middleton, Alexei Lyapustin, Ivan Mammarella, Janne Levula, and Albert Porcar-Castell. Detecting inter-annual variations in the phenology of evergreen conifers using long-term modis vegetation index time series. *Remote sensing*, 9(1):49, 2017.

[104] Dmitry Ulyanov, Andrea Vedaldi, and Victor Lempitsky. Improved texture networks: Maximizing quality and diversity in feed-forward stylization and texture synthesis. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 6924–6932, 2017.

[105] Andrea Vedaldi and Karel Lenc. Matconvnet: Convolutional neural networks for matlab. In *Proceedings of the 23rd ACM international conference on Multimedia*, pages 689–692. ACM, 2015.

[106] Ricardo Vilalta, Christophe Giraud-Carrier, and Pavel Brazdil. Meta-learning-concepts and techniques. In *Data mining and knowledge discovery handbook*, pages 717–731. Springer, 2009.

[107] Duo Wang, Yu Cheng, Mo Yu, Xiaoxiao Guo, and Tao Zhang. A hybrid approach with optimization-based and metric-based meta-learner for few-shot learning. *Neurocomputing*, 349:202–211, 2019.

[108] Jin-Bao Wang, Ning He, Lu-Lu Zhang, and Ke Lu. Single image dehazing with a physical model and dark channel prior. *Neurocomputing*, 149:718–728, 2015.

[109] Wencheng Wang, Xiaohui Yuan, Xiaojin Wu, and Yunlong Liu. Fast image dehazing method based on linear transformation. *IEEE Transactions on Multimedia*, 19(6):1142–1155, 2017.

[110] Zhou Wang and Alan Bovik. A universal image quality index". *In IEEE Signal Processing Letters*, 9(3):81–84, 2002.

[111] Zhou Wang, Alan C Bovik, Hamid R Sheikh, and Eero P Simoncelli. Image quality assessment: from error visibility to structural similarity. *IEEE transactions on image processing*, 13(4):600–612, 2004.

[112] John Weier and David Herring. Measuring vegetation (ndvi & evi). *NASA Earth Observatory*, 20, 2000.

[113] Tomihisa Welsh, Michael Ashikhmin, and Klaus Mueller. Transferring color to greyscale images. In *ACM Transactions on Graphics (TOG)*, volume 21, pages 277–280. ACM, 2002.

[114] Marek Wójtowicz, Andrzej Wójtowicz, and Jan Piekarczyk. Application of remote sensing methods in agriculture. *Communications in Biometry and Crop Science*, 11(1):31–50, 2016.

[115] Guorong Wu, Brent C Munsell, Yiqiang Zhan, Wenjia Bai, Gerard Sanroma, and Pierrick Coupé. *Patch-Based Techniques in Medical Imaging: Third International Workshop, Patch-MI, Held in Conjunction with MICCAI, Quebec City, QC, Canada, Proceedings*, volume 10530. Springer, 2017.

[116] Lili Xu, Baolin Li, Yecheng Yuan, Xizhang Gao, and Tao Zhang. A temporal-spatial iteration method to reconstruct ndvi time series datasets. *Remote Sensing*, 7(7):8906–8924, 2015.

[117] Jiachen Yang, Bin Jiang, Zhihan Lv, and Na Jiang. A real-time image dehazing method considering dark channel and statistics features. *Journal of Real-Time Image Processing*, 13(3):479–490, 2017.

[118] Danny Ngo Lung Yao, Abdullah Bade, Norhaida Mohd Suaib, and Hamzah Asyrani Bin Sulaiman. Digital image enhancement using enhanced detail and dehaze technique (dde). *Advanced Science Letters*, 24(3):1559–1561, 2018.

[119] Haiping Yu, Fazhi He, and Yiteng Pan. A novel region-based active contour model via local patch similarity measure for image segmentation. *Multimedia Tools and Applications*, 77(18):24097–24119, 2018.

[120] Sergey Zagoruyko and Nikos Komodakis. Learning to compare image patches via convolutional neural networks. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 4353–4361, 2015.

[121] Jure Zbontar and Yann LeCun. Computing the stereo matching cost with a convolutional neural network. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 1592–1599, 2015.

[122] Jure Zbontar and Yann LeCun. Stereo matching by training a convolutional neural network to compare image patches. *Journal of Machine Learning Research*, 17(1-32):2, 2016.

[123] He Zhang, Vishwanath Sindagi, and Vishal M Patel. Joint transmission map estimation and dehazing using deep networks. *arXiv preprint arXiv:1708.00581*, 2017.

[124] Richard Zhang, Phillip Isola, and Alexei A Efros. Colorful image colorization. In *European Conference on Computer Vision*, pages 649–666. Springer, October 2016.

[125] Xiaopeng Zhang, Terence Sim, and Xiaoping Miao. Enhancing photographs with near infra-red images. In *Computer Vision and Pattern Recognition, IEEE Conference on*, pages 1–8. IEEE, 2008.

[126] Zhiqiang Zhao, Jiangbo Gao, Yanglin Wang, Jianguo Liu, and Shuangcheng Li. Exploring spatially variable relationships between ndvi and climatic factors in a transition zone using geographically weighted regression. *Theoretical and applied climatology*, 120(3-4):507–519, 2015.

[127] Tiancheng Zhi, Bernardo R Pires, Martial Hebert, and Srinivasa G Narasimhan. Deep material-aware cross-spectral stereo matching. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 1916–1925, 2018.

[128] Jun-Yan Zhu, Taesung Park, Phillip Isola, and Alexei A Efros. Unpaired image-to-image translation using cycle-consistent adversarial networks. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 2223–2232, 2017.

Publications

Published Journals

- Alonso Suárez, Patricia L. Suárez and Cristina Abad, “Applicability of Information Technology Cloud Computing in Ecuadorians Pymes". *ESPOL Technology Journal (RTE)*. Volume 28, no. 5 (2015).
- Patricia L. Suárez, Mónica Villavicencio, “Canny Edge Detection in Cross-Spectral Fused Images". *ENFOQUE UTE Journal (RTE)*. INCISCOS 2016 Special Edition, Volume 8, no. 1 (2016), Clarivate Analytics, Web of Science.

Journals in Review

- Patricia L. Suárez, Angel Sappa, Boris Vintimilla, “Deep Learning based Near Infrared Image Colorization". *IEEE Latin America Transactions:Computing and Processing*, (2019). (IEEE LATAM), IEEE Xplore.
- Patricia L. Suárez, Angel Sappa, Boris Vintimilla, “NDVI vegetation index estimation using deep learning". *IEEE Access The Multidisciplinary Open Access Journal*, (2019). (IEEE), IEEE Xplore.

Book Chapters

- Angel Fiallos, Patricia L. Suárez and Mónica Villavicencio, “Agile methodologies in software development: one based on Scrum and MDD approach". *book:Advances and Applications of Intelligent Systems and New Technologies*. Volume 28, no. 5 (2015).
- Marcos Espinosa, Patricia L. Suárez, “Study of use, privacy and dependence on social networks by students in the Ecuadorian Universities". *book:Communications in Computer and Information Science series:International Conference on Technologies and Innovation*. Volume 658, 114-128 (2016). Springer International Publishing.
- Patricia L. Suárez, Angel Sappa, Boris Vintimilla, “Learning to Colorize Infrared Images". *book:Trends in Cyber-Physical Multi-Agent Systems. The PAAMS Collection-15th International Conference, DOI 978-3-319-61578-3*, (2017). Springer International Publishing.

- Patricia L. Suárez, Angel Sappa, Boris Vintimilla, “Infrared Image Colorization based on a Conditional Triplet DCGAN Architecture". *book:Image Processing, Computer Vision, Pattern Recognition, and Graphics:Image Analysis and Processing - ICIAP 2017, DOI 10.1007/978-3-319-68560-1*, (2017). Proceedings, Part I, Springer International Publishing.

- Patricia L. Suárez, Angel Sappa, Boris Vintimilla, “Vegetation Index Estimation from Monospectral Images". *book:Image Processing, Computer Vision, Pattern Recognition, and Graphics:15th International Conference, ICIAR 2018, DOI DOI 10.1007/978-3-319-93000-8*, (2018). Proceedings, Springer International Publishing.

Conference Papers

- Raúl Mira, Patricia L. Suárez, Rafael Rivadeneira and Angel Sappa, “PETRA: A Crowdsourcing-Based Platform for Rocks Data Collection and Characterization”, *4th Ecuador Technical Chapters Meeting, (ETCM)*, Ecuador, 2019.

- Patricia L. Suárez, and Angel D Sappa and Vintimilla, Boris X, "Image patch similarity through a meta-learning metric based approach". In *14th International Conference on Signal-Image Technology & Internet-Based Systems (SITIS)*, Italy, 2019.

- Rafael Rivadeneira, Patricia L, Suárez, Sappa Angel D and Boris X Vintimilla, "Thermal Image SuperResolution through Deep Convolutional Neural Network". In *International Conference Image Analysis and Recognition*, Springer; Canada 2019.

- Patricia L, Suárez, Angel D Sappa and Boris X Vintimilla, "Image Vegetation Index Through a Cycle Generative Adversarial Network", *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops (CVPR-PBVS)*, USA, 2019.

- Patricia L, Suárez, Angel D Sappa and Boris X. Vintimilla, "Cross-spectral image dehaze through a dense stacked conditional GAN based approach". In *14th International Conference on Signal-Image Technology & Internet-Based Systems (SITIS)*, Spain, 2018.

- Patricia L, Suárez, Angel D Sappa and Boris X. Vintimilla, & Hammoud, R. I, “Near InfraRed Imagery Colorization”, *25th IEEE International Conference on Image Processing, (ICIP)*, Greece, 2018.

- Patricia L, Suárez, Angel D Sappa and Boris X. Vintimilla, “Vegetation Index Estimation from Monospectral Images”, *15th International Conference on Image Analysis and Recognition, (ICIAR)*, Portugal, 2018.

- Patricia L, Suárez, Angel D Sappa and Boris X. Vintimilla, & Hammoud, R. I, “Infrared Image Colorization based on a Triplet DCGAN Architecture”, *Deep Learning based Single Image Dehazing. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops, (CVPR-PBVS)*, USA, 2018.

- Patricia L, Suárez, Angel D Sappa and Boris X. Vintimilla, “Adaptive Harris Corner Detector Evaluated with Cross-Spectral Images”, *The International Conference on Information Technology & Systems, (ICITS)*, Ecuador, 2018.
- Patricia L, Suárez, Angel D Sappa and Boris X. Vintimilla, “Learning Image Vegetation Index through a Conditional Generative Adversarial Network”, *2nd Ecuador Technical Chapters Meeting, (ETCM)*, Ecuador, 2017.
- Patricia L, Suárez, Angel D Sappa and Boris X. Vintimilla, “Infrared Image Colorization based on a Conditional Triplet DCGAN Architecture”, *19th International Conference on Image Analysis and Processing, (ICIAP)*, Italy, 2017.
- Patricia L, Suárez, Angel D Sappa and Boris X. Vintimilla, “Infrared Image Colorization based on a Triplet DCGAN Architecture”, *13th IEEE Workshop on Perception Beyond the Visible Spectrum, (CVPR-PBVS)*, Hawaii, 2017.
- Patricia L, Suárez, Angel D Sappa and Boris X. Vintimilla, “Learning to Colorize Infrared Images”, *15th International Conference on Practical Applications of Agents and Multi-Agent Systems, (PAAMS)*, Portugal, 2017.
- Patricia L, Suárez, Angel D Sappa and Boris X. Vintimilla, "Cross-spectral image patch similarity using convolutional neural network", *Electronics, Control, Measurement, Signals and their Application to Mechatronics, (ECMSM)*, Spain, 2017.
- Patricia L, Suárez, Mónica Villavicencio “Applying Canny Edge Detection with Cross-Spectral Fused Images using Morphological Filters”, *International Conference on Information Systems and Computer Science, (INCISCOS)*, Quito, Ecuador, 2016.
- Patricia L. Suárez, “The Importance of Parallelism in Algorithms Design”, *3rd International Scientific Congress, Technology, College and Society*, Samborondon, Ecuador, 2015.
- Alonso Suárez, Patricia L. Suárez, Cristina Abad, “Applicability of Information Technology Cloud Computing in Ecuadorians Pymes". In *Second Andean Congress of Computing, Information and Education*, (CACIED 2015), ESPOL. Guayaquil-Ecuador.