CHAPTER 9

Deep learning-based vegetation index estimation

Patricia L. Suárez[a], **Angel D. Sappa**[a,b], **and Boris X. Vintimilla**[a]
[a]ESPOL Polytechnic University, CIDIS-FIEC, Guayaquil, Ecuador
[b]Computer Vision Center, Edifici O, Campus UAB, Bellaterra, Barcelona, Spain

9.1 Introduction

Computer vision applications can be found in almost every domain, including topics such as medical imaging, gaming, video surveillance, multimedia, industrial applications, and 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. There are some cases, in particular in medical imaging and remote sensing, where cross-spectral and multispectral images are considered. The appealing factor of using images from different spectral bands lies on the one hand on the possibility to obtain information that cannot be seen at the visible spectrum; on the other hand, on the combined use of information that can be considered to generate some kind of high-level reasoning; for instance, in remote sensing the combined use of images from different spectral bands is considered to generate vegetation indexes (VIs). These VIs are used to determine the health and strength of vegetation and their definitions involve several factors, such as soil reflectance, vegetation density, etc. All this information would help to increase the yield of crops [1, 2]. The obtained information is used for monitoring and evaluating the Earth's vegetative cover using several factors, such as soil reflectance, atmosphere, vegetation density, etc., with the aim to obtain those formulas that get more reliable information about vegetation based on remotely sensed values.

The usual form of a VI is a ratio of reflectance measured in two bands, or their algebraic combination. Spectral ranges (bands) to be used in VI calculation are selected depending on the spectral properties of plants. Lately, techniques based on sensors sensitive to multiple spectra have been implemented to perform remote sensing to evaluate the biophysical variables of vegetation in both forestry and agriculture [3, 4]. Furthermore, Panda et al. [5] proposed a method for processing high-end images in order to determine the importance of spectral VIs in the field of agricultural crop yield prediction using a neural network. In Ref. [6], the authors proposed to analyze the climatological phenomena that affect the local climate. According to their theory, these phenomena have a direct effect on crop yield.

Generative Adversarial Networks for Image-to-Image Translation
https://doi.org/10.1016/B978-0-12-823519-5.00013-0

The index could be computed using several spectral bands that are sensitive to plant biomass and health. 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, the information between the light reflected in the near-infrared and in the visible spectrum is generally used to detect areas that potentially have healthy vegetation.

Among the different indexes proposed in the literature, the Normalized Difference Vegetation Index (NDVI) is the most widely used [7]; NDVI is often used to monitor drought, forecast agricultural production, assist in forecasting fire zones and desert offensive maps [8]. 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:

$$NVDI = \frac{R_{\text{NIR}} - R_{\text{RED}}}{R_{\text{NIR}} + R_{\text{RED}}}, \tag{9.1}$$

where R_{NIR} is the reflectance of NIR radiation and R_{RED} is the reflectance of visible red 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 0 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 [9, 10].

Proposals that use images of several spectra, whether crossed or multispectral, depend on the use of multiple sensors. In the case of VIs such as NDVI, it is required to have images of the visible spectrum and near-infrared spectrum of the same scene, which are acquired by different cameras at the same time . These images are required to calculate the values of Eq. (9.1). It should be noted that before calculating Eq. (9.1) images must be accurately recorded, that is, the information must be referred to the same reference system. Since the images of different spectra can be displayed differently, the challenge is to find the same reference points in the images of both spectra [11]. Recently, techniques with convolutional networks have been proposed focusing on solving this problem and finding correspondences in crossed spectral domains [12, 13]. With the correlated information, the images can be recorded in a single reference system.

In Ref. [14], the authors proposed to use the NDVI to measure the changes in the ecosystem in a given interval of time. The changes in the index values allow us to infer how the climate impacts the health of the crops. With this method, the impact of climate

change can be determined and controlled planning can be managed, focusing efforts on the most affected areas. This is valuable in determining effective and smart reforestation plans.

In this chapter, a novel approach to perform an image-to-image translation is proposed, in which the NDVI is estimated using a synthetic NIR image. The proposed model is able to use unpaired data to estimate a synthetic NIR just from an grayscale image using a CycleGAN. Actually, a similar technique has been recently presented in Ref. [15] where an NDVI is generated from a near-infrared (NIR) image, and also in Ref. [16] where the VI is estimated just from the single image of the visible spectrum. Although interesting results have been obtained, the weak point of these approaches lies on the need of having NIR images, which are not that common such as visible spectrum images. In other words, the disadvantage of these approaches depends on paired samples for the training process. The solution proposed in the current chapter consists of a model where the index is estimated from an unpaired learning-based approach, where a cycled generative adversarial network (CycleGAN) [17] is trained with a large data set.

In the proposed approach, an unsupervised learning model with a set of unpaired images is used as an input, one from the visible spectrum and the other image corresponds to an NIR image; each one is fed into a CycleGAN to perform the image domain translation. Additionally, a multiple loss function is used to obtain a better optimization of the model; a residual network (ResNet) architecture is used to go deeper without degradation in accuracy and error rate. The chapter is organized as follows. Section 9.2 presents works related to the NDVI problem, as well as the basic concepts and notation of GAN and CycleGAN networks. The proposed approach is detailed in Section 9.3. The experimental results with a set of real images are presented in Section 9.4. Finally, the conclusions are given in Section 9.5.

9.2 Related work

Solutions based on computer vision to tackle problems related to precision agriculture have been widely used. This technology enables better identification, analysis, and management of this temporal and spatial in-field variability. Nowadays, with NIR sensors, all the captured crop information could be filed 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 these kinds of agricultural prediction problems. These methods came from mathematical and statistical to deep learning neural networks.

This section review works related to VI estimation, using classical approaches as well as convolutional neural network (CNN)-based approaches.

9.2.1 Vegetation index: Formulations and applications

In this section, agricultural approaches focused on the use of the NDVI to perform adequate control of crop production using the index information to monitor plant health at each stage of their growth are reviewed.

In Ref. [18], the authors proposed 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 NDVI. Another approach is presented in Ref. [19]; the authors propose a technique to predict the vegetation dynamics behavior using Moderate Resolution Imaging Spectroradiometer (MODIS) NDVI time series data sets and long short term memory model network, an advanced technique adapted from the artificial neural network.

Another approach, presented by Ulsig et al. [20], introduces an automated technique 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 normalized difference VI and another step used to detect individual palm trees using a combination of circular Hough transform (CHT) and the morphological operators. Damian et al. [21] propose to use the information obtained from the normalized difference VI using satellite images to increase the productivity improving the task of delimiting management zones for annual crops. For this research three crop productivity maps, from 2009 to 2015, were used for each area of analysis, developing a descriptive and geostatistical case study.

According to Ulsig et al. [22], 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 VIs such as the normalized difference VI 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. In addition, Li et al. present 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 NDVI data from satellite imagery, using the spatial lag-Tobit models [23], which predict tree canopy cover from NDVI. In another research [24], the authors focus on temporal NDVI and surface temperature, the methodology used altogether for the assessment of resolution dynamic Urban Heat Island (UHI) change on environmental condition with different environmental conditions, geographical locations, and demography. 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 Ref. [25], the authors present a

method to reconstruct NDVI 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 will be most applicable when large numbers of contaminated pixels exist.

Also, in Ref. [26], the authors present a method to analyze the use of the NDVI to evaluate crop yields, using a multispectral sensor mounted on a UAV with the objective of predicting biomass variations and grain production. In another work, presented by Taghizadeh et al. [27], the authors propose an approach to extract the phenological parameters based on time series of the NDVI, and these variables are used with crop rotation to predict the organic carbon content of the surface layer.

Also in Ref. [28], 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 captured 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 of the soil in a farmland area. Additionally, in Ref. [29], the authors illustrate how the normalized difference VI, leaf area index (LAI), and fractional vegetation cover are related to each other, using a simple radiative transfer model with vegetation, soil, and atmospheric components. Another approach [30] 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 represented significant improvements in model performance over the ordinary least-squares (OLS) models. The results revealed 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.

9.2.2 Deep learning-based approaches

Deep learning models have obtained state-of-the-art results on some computer vision complex problems. Nevertheless, there are many challenging problems in agricultural pending to be solved and deep learning approaches are the most likely to be used, obviating the need for a pipeline of specialized and hand-crafted methods used before. Some researchers have proposed deep learning-based approaches for remote sensing and agricultural applications. In Ref. [31], 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 VI.

Huang et al. [32] 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 CNN, operating directly on indexes input due to its remarkable classification performance, exploiting the spatial contextual information and spectral features for effective topographic extraction. In addition, in Ref. [33], a decision-level fusion approach is proposed with a simpler architecture for the task of dense semantic labeling. This method first obtains two initial probabilistic labelings resulting from a fully CNN and a simple classifier, for example, logistic regression exploiting spectral channels and LiDAR data, respectively. The conditional random field (CRF) inference will estimate the final dense semantic labeling results. In Ref. [34], 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.

All the approaches presented earlier are just a selection of recent publications where the usefulness of VIs, in particular the NDVI, can be appreciated. Unfortunately, to compute the NDVI registered images from different spectra (i.e., visible and NIR) are needed, which sometimes is a challenging task since they may look different. So, the problem is how to find the same set of features in both spectra (e.g., points [11]) to be used as a reference for the registration process. Recently, some deep learning-based approaches have been proposed to overcome this problem and to obtain correspondences in cross-spectral domains (e.g., [13, 35]). Once correspondences are obtained, the image registration can proceed by mapping both images to a single reference system; then VIs can be easily computed. As mentioned in the previous section, recently some approaches for estimating NDVI have been proposed (e.g., [15, 16]) implementing GAN's networks using NIR or RGB images; both approaches depend on the existence of accurately registered images.

Having in mind the registration drawback needed to estimate the NDVI and to overcome this problem, in the current work an unsupervised learning model is proposed (a CycleGAN architecture). The model is trained with a set of unpaired images (grayscale and NDVI image) under an unsupervised scheme. To understand generative adversarial networks (GANs), a summary is given here.

GANs are powerful and flexible tools quite useful in several computer vision problems; one of their most common applications is image generation. Fig. 9.1 depicts this architecture. In the GAN framework [36], generative models are estimated via an adversarial process, in which simultaneously two models are trained: (i) a generative model G that captures the data distribution, and (ii) a discriminative model D that estimates the probability that a sample came from the training data rather than G. The training procedure for G is to maximize the probability of D making a mistake. In this architecture, it

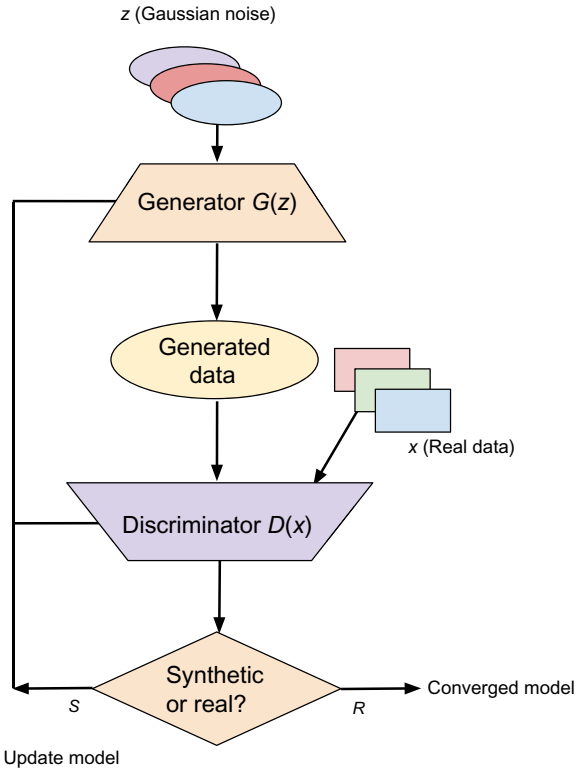

Fig. 9.1 Illustration of a generative adversarial network.

is possible to apply certain conditions to improve the learning process. According to Ref. [37], 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 D are both trained simultaneously, the parameters for G are adjusted to minimize $\log(1-D(G(z)))$ and for D to minimize $\log D(x)$ with a value function $V(D, G)$:

$$\min_G \max_D V(D,G) = \mathbb{E}_{x\sim p\ \mathrm{data}(x)}[\log D(x)] + \mathbb{E}_{z\sim p\ \mathrm{data}(z)}[\log(1-D(G(z)))]. \quad (9.2)$$

GANs can be extended to a conditional model if both the generator and discriminator are conditioned on some extra information y (see Fig. 9.2). This information could be any kind of auxiliary information, such as class labels or data from other modalities. We can perform the conditioning by feeding y into both discriminator and generator as additional input layer. The objective function of a two-player minimax game would be

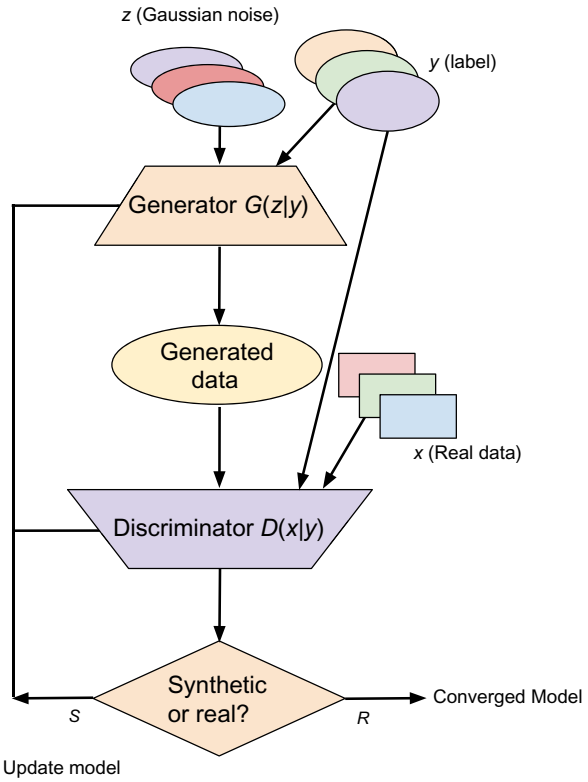

Fig. 9.2 Illustration of a conditional generative adversarial network.

$$\min_G \max_D V(D,G) = \mathbb{E}_{x\sim p\ \mathrm{data}(x)}[\log D(x|y)] + \mathbb{E}_{z\sim p\ z(z)}[\log(1 - D(G(z|y)))]. \quad (9.3)$$

The discriminator performs a binary classification including the extra information fed to the network, as a result, the discriminator and generator will gain more accurate gradients. Conditional GANs enhance the stability of the model, but it affects the learning of the semantic characteristics of the image samples.

9.3 Proposed approach

This work proposes to estimate the NDVI VI using a synthetic NIR generated from just from a single image of the visible spectrum using a CycleGAN. The architecture used in this approach is based on the one presented in Ref. [17], a previous work that presents an unpaired image-to-image translation, through a CycleGAN. This type of network permits domain style transfer, which is a convenient method for image-to-image translation problems because it is not necessary to have a set of input images that capture the scene at

the same time and place from different spectra. Obtaining this type of set of images could be time consuming and quite difficult based on what type of domain style the image data set we are trying to translate between. In Ref. [38], 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 but also learn a loss function to train the corresponding mapping. Before presenting the proposed approach, a brief description of CycleGAN is presented.

9.3.1 Cycle 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. This task has been generally performed by using a training set of aligned image pairs. However, for many tasks, paired training data will not be available, and to prepare them often takes a lot of work from specialized personnel to obtain thousands of paired image datasets, especially with complex image translations. CycleGAN 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 (see Fig. 9.3) to observe the corresponding scheme. CycleGAN will learn to perform style transfer from the two sets despite every image having vastly different compositions. According to Zhu et al. [17], the CycleGAN presents an approach for learning to translate an image from a source domain X to a target domain Y in the absence of paired examples (see Fig. 9.4) to observe a description of the domain translation with paired samples (*left of the graph*); and unpaired samples (*right of the graph*); in our case, we use a translation of unpaired images. Thus, the goal is to learn a mapping G: $X \rightarrow Y$ such that the distribution of images from $G(X)$ is indistinguishable from the distribution Y using an adversarial loss. Because this mapping

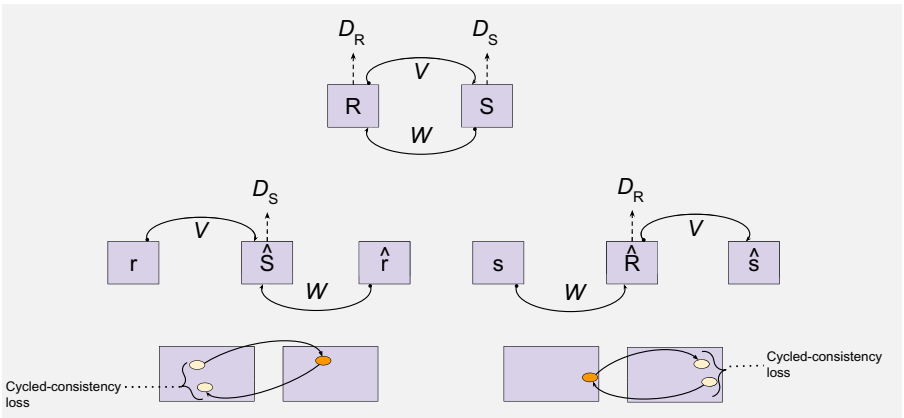

Fig. 9.3 Cycle generative adversarial network, original scheme proposed in Ref. [17].

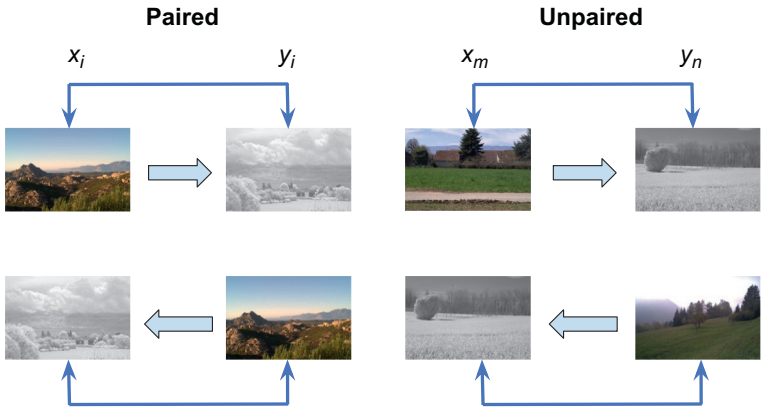

Fig. 9.4 (*Left*) Supervised training (paired data). (*Right*) Unsupervised training (unpaired data).

is highly under-constrained, it is necessary an inverse mapping F: $Y \rightarrow X$ and introduce a cycle consistency loss to enforce $F(G(X)) \approx X$ (and vice versa).

The model includes two mappings functions G: $X \rightarrow Y$ and F: $Y \rightarrow X$. In addition, 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 [36] for matching the distribution of generated 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.

9.3.2 Residual learning model (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. One of the first implementation of deep convolutional networks is presented in Ref. [39], where the authors present an approach to classify 1000 different classes from the ImageNet dataset. The model have been designed to support very deep CNN training to classify the 1.2 million high-resolution images into the 1000 different classes. The model has 60 million parameters and 650,000 neurons. The architecture consists of five convolutional layers followed by max-pooling layers in some cases, and three fully connected layers with a last softmax layer of 1000 elements. The authors also have implemented a very efficient convolutional operation with multiple GPU to reduce training time and overfitting. Additionally in the fully connected layers they have employed a dropout operation to perform regularization, which proved to be very effective. Another technique that continues the work in very deep learning networks, is the one presented in Ref. [40]; according to the authors, 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 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 Ref. [40], 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. Previously, the desired underlying mapping was denoted by $H(x)$. 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 feed-forward neural networks with "shortcut connections" to perform identity mapping, and their outputs are added to the outputs of the stacked layers (see Fig. 9.5), 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.

It avoids the vanishing gradient problem, as the gradient is backpropagated 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, we implement our generator and discriminator 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. ResNet is a model designed to be applied in a deep neural network layer architecture, which consists of convolution layers known as building blocks, where a residue of input is added to the output.

9.3.3 Proposed architecture

This section presents the approach proposed for NDVI vegetation estimation just with a single image from the visible spectrum. As mentioned earlier, it uses a similar architecture like the one proposed in Ref. [17], a recent work for unpaired image-to-image translation, where the use of a CycleGAN has been proposed. CycleGANs 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 one is trying to translate between. Another approach presented

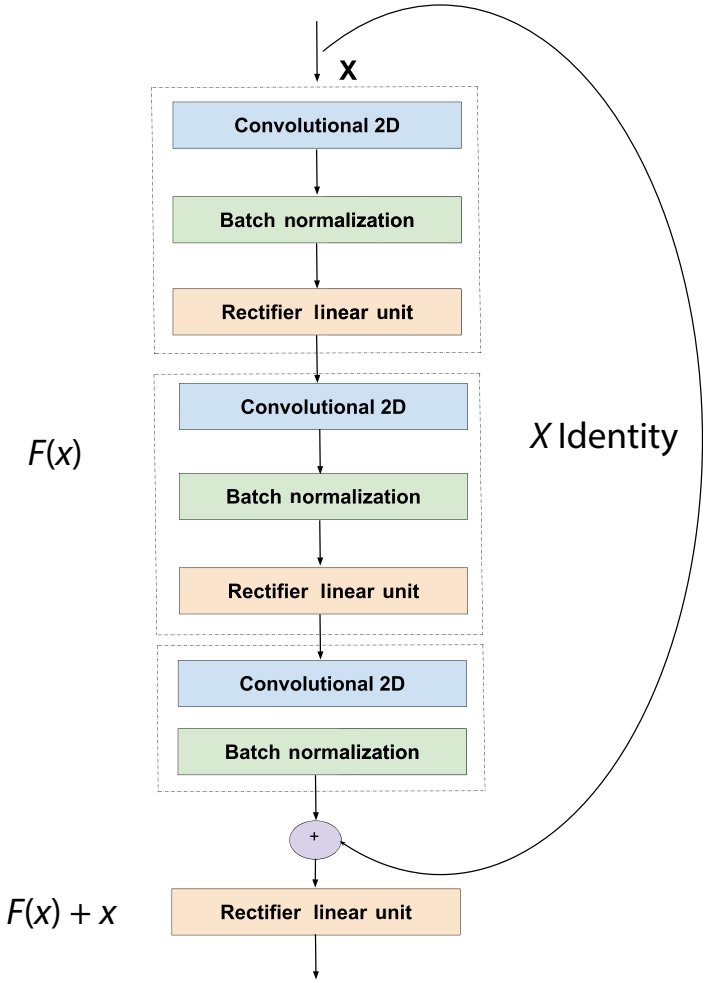

Fig. 9.5 Residual block used on the generator network.

in Ref. [38] has shown results synthesizing photos from label maps, reconstructing objects from edge maps, but still dependent on some kind of correlated labeling.

Our architecture is based on the approach presented in Ref. [17] in relation to cycle consistent learning and loss functions; in our work, it is used to estimate the synthetic NIR images. The proposed model can learn to translate the images between the visible spectrum to the 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 VI 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, our architecture uses the ResNet [40] to perform the image transformation from one spectrum to another.

The core idea of ResNet is to introduce a so-called "identity shortcut connection" that skips one or more layers. These skip connections 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 input; otherwise, the characteristics of original images will not be retained in the output and results will be very unreal. The formulation of $F(x) + x$ can be realized by feed-forward neural networks with "shortcut connections" (see Section 9.3.2). Shortcut connections are those skipping one or more layers. In our case, the shortcut connections simply perform identity mapping, and their outputs are added to the outputs of the stacked layers. Identity shortcut connections add neither extra parameter nor computational complexity. The entire network can still be trained end-to-end by SGD with backpropagation. These skip connections 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 RGB input (grayscale); otherwise, the characteristics of original images will not be retained in the output and results will be very unreal. Fig. 9.6 depicts the CycleGAN model proposed in the current work. As shown in

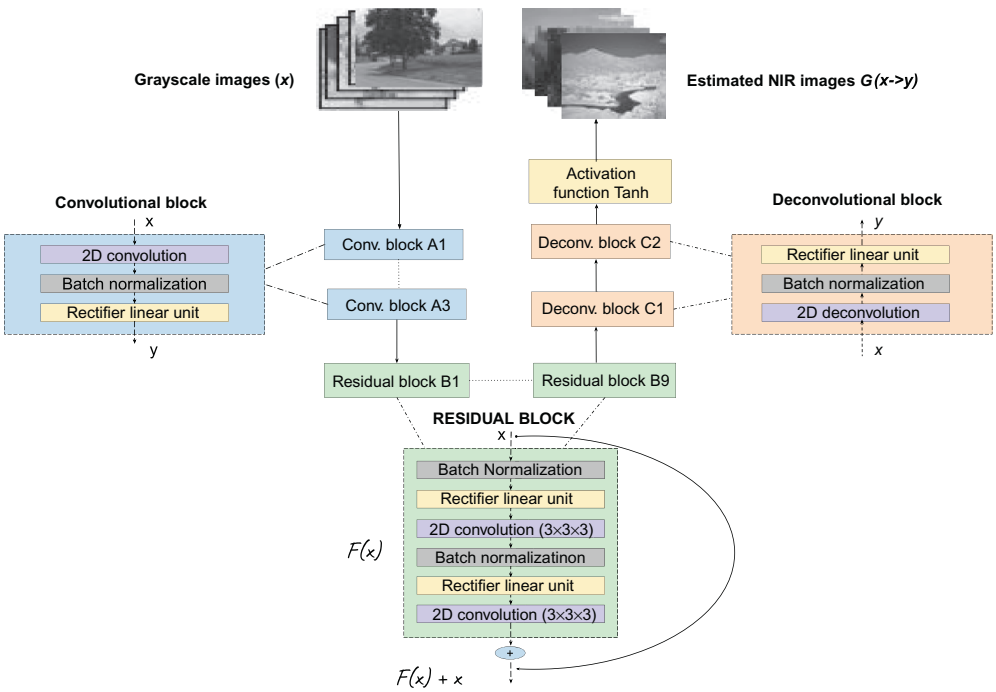

Fig. 9.6 Cycle generative adversarial generator network detailed architecture.

Fig. 9.6, CycleGAN architecture to generate NIR synthetic images is composed of two generators G, F and two discriminators Dx, Dy. In order to generate a synthetic image, the architecture takes the advantage from the joint of cycle consistency and least-square losses [41] 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 and NIR images and generate uniform synthetic outputs. According to Zhu et al. [17], the objective of a CycleGAN is to learn mapping functions between two domains X and Y given training samples $x_i{}_{i=1}^{N} \in X$ and $x_i{}_{i=1}^{N} \in Y$.

The generator network architecture designed to estimate NIR synthetic VI is described in Fig. 9.6. Also, Figs. 9.9 and 9.10 depict the CycleGAN scheme proposed in the current work. The model includes two mapping functions $G: X \rightarrow Y$ and $F: Y \rightarrow X$. In addition, 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 [36] 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.

9.3.4 Loss functions

The adversarial losses, according to Goodfellow et al. [36], are applied to both mapping functions. For the mapping function $G: X \rightarrow Y$ and its discriminator D_y, the objective is defined as

$$\mathcal{L}_{GAN}(G, D_y, X, Y) = \mathbb{E}_{y \sim p\,\mathrm{data}(y)}[\log D_Y(y)] + \mathbb{E}_{x \sim p\,\mathrm{data}(x)}[\log(1 - D_Y(G(x)))], \quad (9.4)$$

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 $F: Y \rightarrow X$ and its discriminator D_x, the objective is defined as

$$\mathcal{L}_{GAN}(F, D_x, Y, X) = \mathbb{E}_{x \sim p\,\mathrm{data}(x)}[\log D_X(x)] + \mathbb{E}_{y \sim p\,\mathrm{data}(y)}[\log(1 - D_X(F(y)))], \quad (9.5)$$

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.

Also, according to Zhu et al. [17], 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, that is, $x \rightarrow G(x) \rightarrow F(G(x)) \approx x$, 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

$$\mathcal{L}_{cycle}(G,F) = \mathbb{E}_{x \sim p\ \mathrm{data}(x)}[\|F(G(x)) - x\|_1] + \mathbb{E}_{y \sim p\ \mathrm{data}(y)}[\|G(F(y)) - y\|_1]. \quad (9.6)$$

9.3.5 Least-square GAN's loss

In the current work, a least-square loss has been implemented [41] 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 our case the synthetic NIR image. The experiments performed with this loss instead of negative log likelihood shown better results. Eqs. (9.4), (9.5) are replaced with the least-square losses, which are defined as

$$\mathcal{L}_{LSGAN}(G, D_y, X, Y) = \mathbb{E}_{y \sim p\ \mathrm{data}(y)}[(D_Y(y) - 1)^2] + \mathbb{E}_{x \sim p\ \mathrm{data}(x)}[D_Y(G(x))^2] \quad (9.7)$$

and

$$\mathcal{L}_{LSGAN}(F, D_x, Y, X) = \mathbb{E}_{x \sim p\ \mathrm{data}(x)}[(D_X(x) - 1)^2] + \mathbb{E}_{y \sim p\ \mathrm{data}(y)}[(D_X(F(y))^2)]. \quad (9.8)$$

For this unsupervised approach, the standard CycleGAN $\mathcal{L}_{CYCLE}$: (cycle-consistent loss), and $\mathcal{L}_{LSGAN}$: (least-square loss), have been implemented, both with their corresponding weights distributions for the multiple 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

$$\mathcal{L}_{\mathrm{FINAL-SYN-NIR_{CYCLE-GAN}}} = 0.38\mathcal{L}_{\mathrm{GAN}} + 0.62\mathcal{L}_{\mathrm{CYCLE}}. \quad (9.9)$$

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

$$\mathcal{L}_{\mathrm{FINAL-SYN-NIR_{CYCLE-LSGAN}}} = 0.65\mathcal{L}_{\mathrm{LSGAN}} + 0.35\mathcal{L}_{\mathrm{CYCLE}}. \quad (9.10)$$

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 VI. An inappropriate weight balance increases the risk that the model generates synthetic indexes with too many artifacts and that it cannot generalize properly.

Once the synthetic NIR image is estimated, the NDVI is computed by using Eq. (9.1) together with the information from the red channel of the given image.

9.4 Results and discussions

9.4.1 Datasets for training and testing

The proposed approach has been evaluated using a grayscale and with an unpaired NVDI VI; the architecture of the U-Net generator implemented is presented in Fig. 9.6; the model receives as an input the a single image of the visible spectrum representation from Brown and Süsstrunk [42]. From the aforementioned data set, the *country*, *mountain*, and *field* categories have been considered for evaluating the performance of the proposed approach; examples of this dataset are presented in Fig. 9.7. This 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 network to generate the VI from each of these categories, a data lengthening process has been applied to avoid overfitting or underfitting the model, so that it can converge and generalize; this process is carried out automatically by a specialized algorithm. It should be noted that during the training process paired images do not belong to the same scene, because there is no need to have correspondences as input for the CycleGAN proposed model.

9.4.2 Data augmentation

The proposed architecture uses as an input an unpaired dataset from Brown and Süsstrunk [42], the RGB converted to grayscale and the NIR images. In order to enlarge the size of the training dataset, we have implemented an automatic data augmentation process to create a modified version of images in the dataset of grayscale and NIR 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 this approach (see Fig. 9.8) has provided us with a total of 70 different variations with a size of 256 × 256 for each image per category existent in the data set; 3500 pairs of images of (256 × 256 pixels) have been generated, both in a grayscale of the RGB images as well as in the corresponding NIR 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, which can be used to feed the learning network to synthesize VIs to increase the performance and accelerate the generalization of the model.

Fig. 9.7 Some examples of cross-spectral images, where the (*first row*) RGB images; (*second row*) unpaired NIR images; (*third row*) ground-truth NDVI images. *(Images from M. Brown, S. Süsstrunk, Multi-spectral SIFT for scene category recognition, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, IEEE, 2011, pp. 177–184.)*

9.4.3 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

Algorithm 1 Augmenting dataset per category, *country*, *field*, and *mountain*; *variations*, is the parameter for the number of iterations per image (in the current experiment, the variables have been set as follows: $variations = 70$, to improve the performance and ability to generalize the model and $sizecrop = 256$ to resize the crop randomly obtained).

```
for Number of images available per category do
   for variations steps do
      • x, y = pixels coordinate on image, randomly obtained
      • h, w = height and width to extract from image, randomly obtained
      • croplist(variations) = crop(x,y,w,h) crop image using
        the coordinates previously obtained
      • croplist(variations) = resize(croplist(variations), sizecrop,
            sizecrop) resize the crop previously obtained
    end for
  savetodisk(croplist)
  croplist = []
end for
```

Fig. 9.8 Algorithm proposed for data augmentation.

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 mean square error (MSE), one of the most used image quality metrics. The MSE metric measures the average of the squares of the errors or deviations. This is to say that large differences between actual and predicted are punished more with MSE. This error (MSE) does not match with human visual perception. In contrast to MSE recently, a perceptual metric that measures image quality level, Structural Similarity (SSIM) index , has been developed with a view to comparing the structural and feature similarity measures between restored and original objects on the basis of perception. For our approach, we have used Root Means Squared Error and SSIM index as metrics, with which we were able to compute the results of the experiments and obtain consistent results. However, RMSE does not measure representation of the textures of the images, instead of SSIM index, which is an absolute value of the representation perspective presented in the images. Additionally, from a semantic perspective, SSIM index gives better results to measure over RMSE error. Also, the SSIM index performs well to obtain perception and saliency-based errors. According to Wang et al. [43], SSIM index evaluates images accounting for the fact that the human visual perception system is sensitive to changes in the 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. The *structural loss* for a pixel p is defined as

$$\mathcal{L}_{SSIM} = \frac{1}{NM}\sum_{p=1}^{P} 1 - SSIM(p), \tag{9.11}$$

where $SSIM(p)$ is the structural similarity index (see Ref. [43] for more details) centered in pixel p of the patch P.

9.4.4 Experimental results

The proposed approach (see Figs. 9.9 and 9.10) has been evaluated using NIR and RGB images together with the corresponding NDVI obtained from Eq. (9.1), in which the

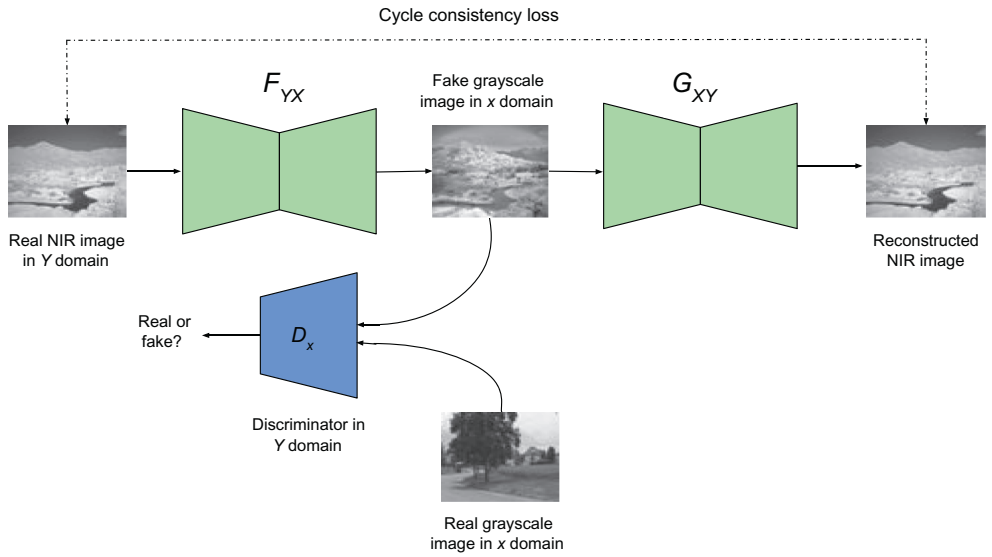

Fig. 9.9 Cycle generative adversarial model F: Y (*NIR*) $\rightarrow$ X(*grayscale*) and its discriminator D_x.

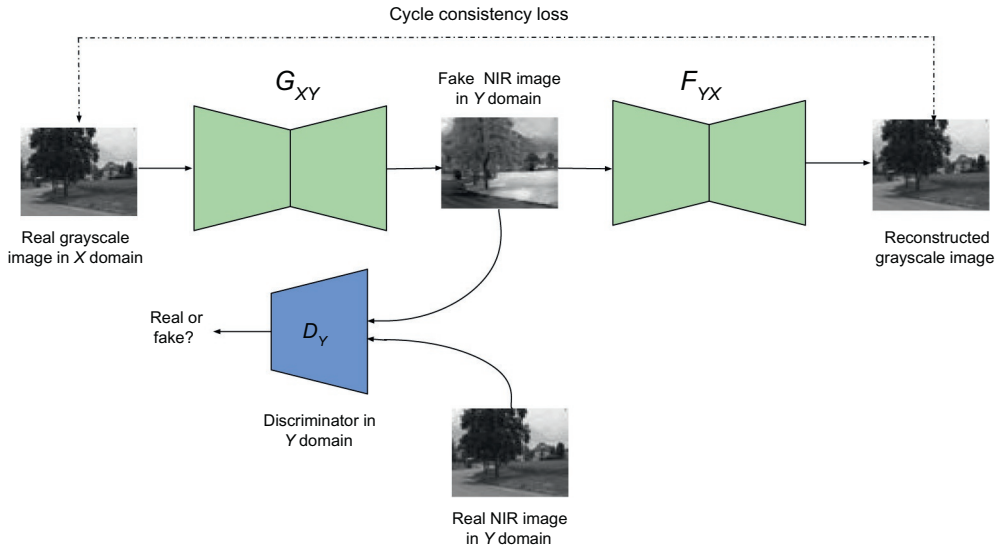

Fig. 9.10 Cycle generative adversarial model G: X(*grayscale*) $\rightarrow$ Y (*NIR*) and its discriminator D_y.

RGB red channel was used; the cross-spectral data set used in our implementation came from Brown and Süsstrunk [42]. This dataset consists of 477 registered images categorized into 9 groups captured in RGB (visible) and NIR (near-infrared) spectral bands. The *country*, *mountain*, and *field* categories have been considered for evaluating the performance of the proposed approach. The *country* category contains 52 pairs of images of (1024 × 680 pixels), *mountain* category contains 55 pairs of images of (1024 × 680 pixels), while the *field* contains 51 pairs of images of (1024 × 680 pixels). In order to increase the training dataset, a data augmentation process was performed, to improve the accuracy of our network to generate synthetic NIR images. The data augmentation consists of applying flipping, rotating, and transposing over the original images. After the data augmentation process, for each category 600 pairs of images from visible and NIR spectrum have been generated. Additionally, for each category 40 pairs of images for testing and 20 pairs of images for validation from visible and NIR spectrum have been used. It is important to emphasize that despite the dataset images are registered, for the CycleGAN training process, we use unpaired images.

On average, every training process took about 80 hours using a 3.2 GHz 8 core processor with 32 GB of memory with a NVIDIA TITAN XP GPU. Some illustrations with the corresponding NIR results obtained with the proposed CycleGAN approach are depicted in Fig. 9.11 for qualitative evaluation.

These synthetic NIR images obtained with the CycleGAN are then used for estimating the NDVIs. Figs. 9.12–9.14 present some illustrations of NDVIs estimated per category *country*, *field*, and *mountain* using the generated synthetic NIR images. Also, Figs. 9.15–9.17 present illustrations of the NVDI generated with the proposed approach compared with results from Ref. [17], showing better qualitative results. Quantitative evaluations are presented in Table 9.1. In this table, average root mean square error (RMSE) and SSIM index computed over the validation set are depicted, when different combinations of the proposed loss functions were considered. Our experiments used the standard loss function for GANs, which are based on negative log likelihood and also used the least-square loss to obtain better quantitative 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. We implement least-square loss to accelerate and maintain stable the training process. Additionally, in this table, results from Refs. [16, 17] are presented. It can be appreciated that in all the cases the results obtained with the least-square loss in the proposed CycleGAN are better than those obtained with the approach presented in Refs. [16, 17]. It should be mentioned that the least-square losses permits to accelerate the network convergence, allowing a better optimization of the network.

To increase the cycle loss effect over the network we used $L1$ (λ). The CycleGAN network proposed has been trained using Stochastic AdamOptimazer since it is well

Fig. 9.11 Illustration of NIR images obtained by the proposed CycleGAN, which are later on used to estimate the corresponding NDVIs. (*First row*) RGB images. (*Second row*) Grayscale image used as input into the CycleGAN. (*Third row*) Estimated NIR images. (*Fourth row*) Ground-truth NIR images. *(Images from M. Brown, S. Süsstrunk, Multi-spectral SIFT for scene category recognition, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, IEEE, 2011, pp. 177–184,* country, field*, and* mountain *categories.)*

suited for problems with deep network, large datasets, and avoid overfitting. The image dataset was normalized in a $(-1, 1)$ range and rescaled to 256×256 to avoid memory problems during the training process. The following hyperparameters were used during the training process: learning rate 0.0003, epsilon $= 1e-08$, exponential decay rate for the first moment momentum 0.6, $L1$ (λ) 10.5, weight decay $1e-2$, leak ReLU 0.20.

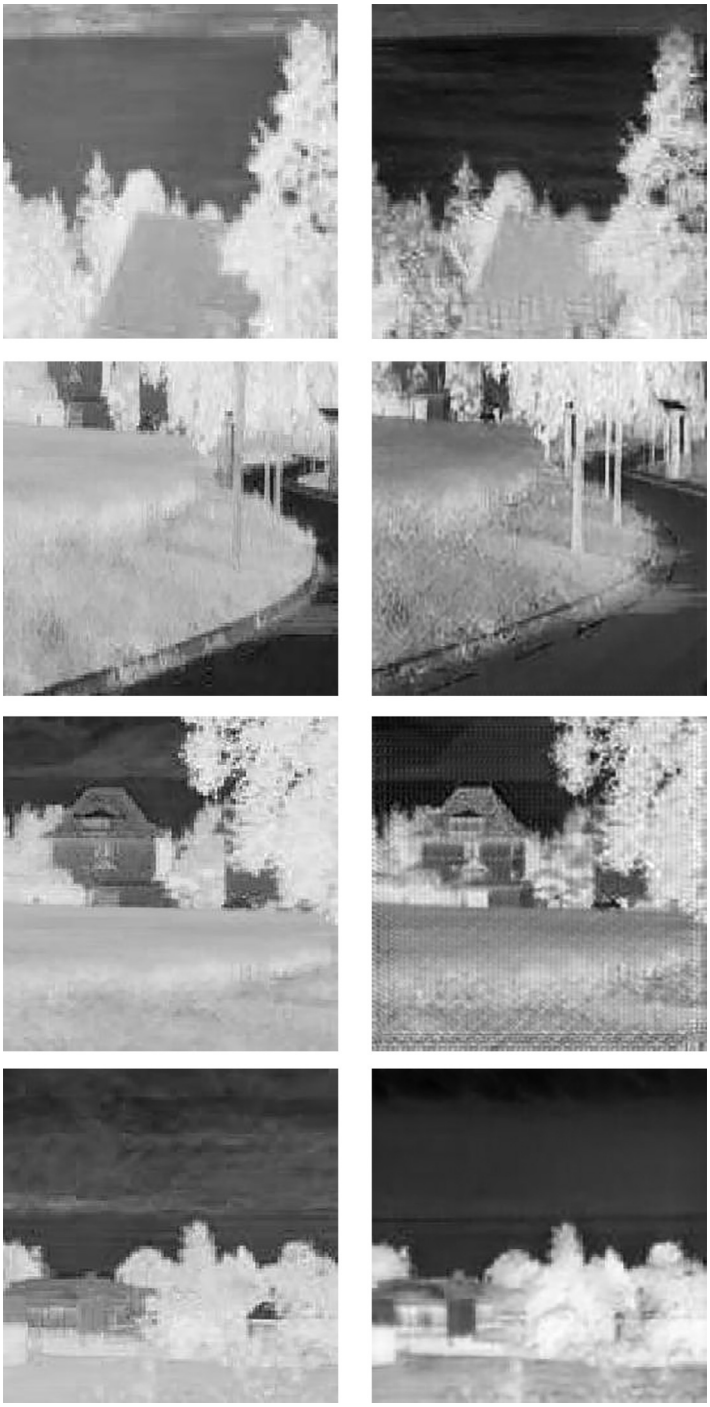

Fig. 9.12 Images of NDVI VIs from *Country* category obtained with the synthetic NIR generated by the proposed CycleGAN. (*Left*) Ground-truth NDVI VI images. (*Right*) Estimated NDVI VIs. *(Images from M. Brown, S. Süsstrunk, Multi-spectral SIFT for scene category recognition, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, IEEE, 2011, pp. 177–184.)*

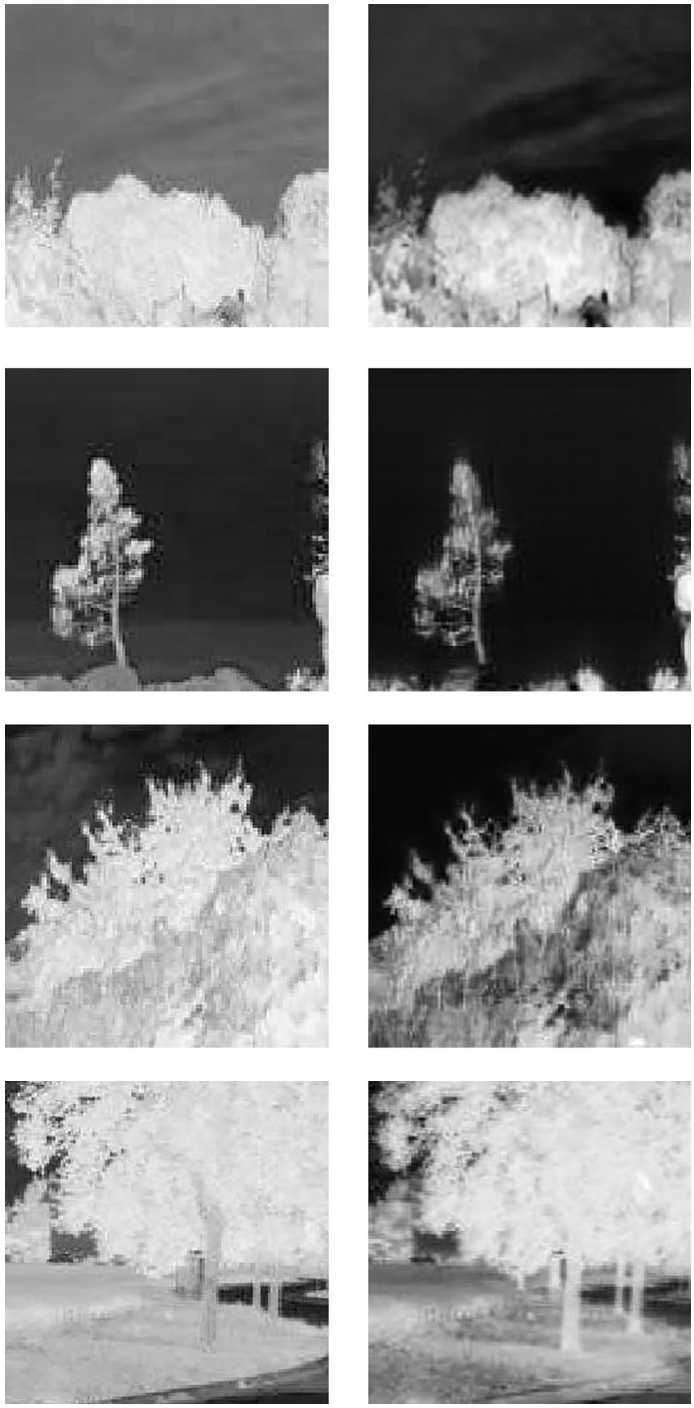

Fig. 9.13 Images of NDVI VIs from *Field* category obtained with the synthetic NIR generated by the proposed CycleGAN. (*Left*) Ground-truth NDVI VI images. (*Right*) Estimated NDVI VIs. *(Images from M. Brown, S. Süsstrunk, Multi-spectral SIFT for scene category recognition, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, IEEE, 2011, pp. 177–184.)*

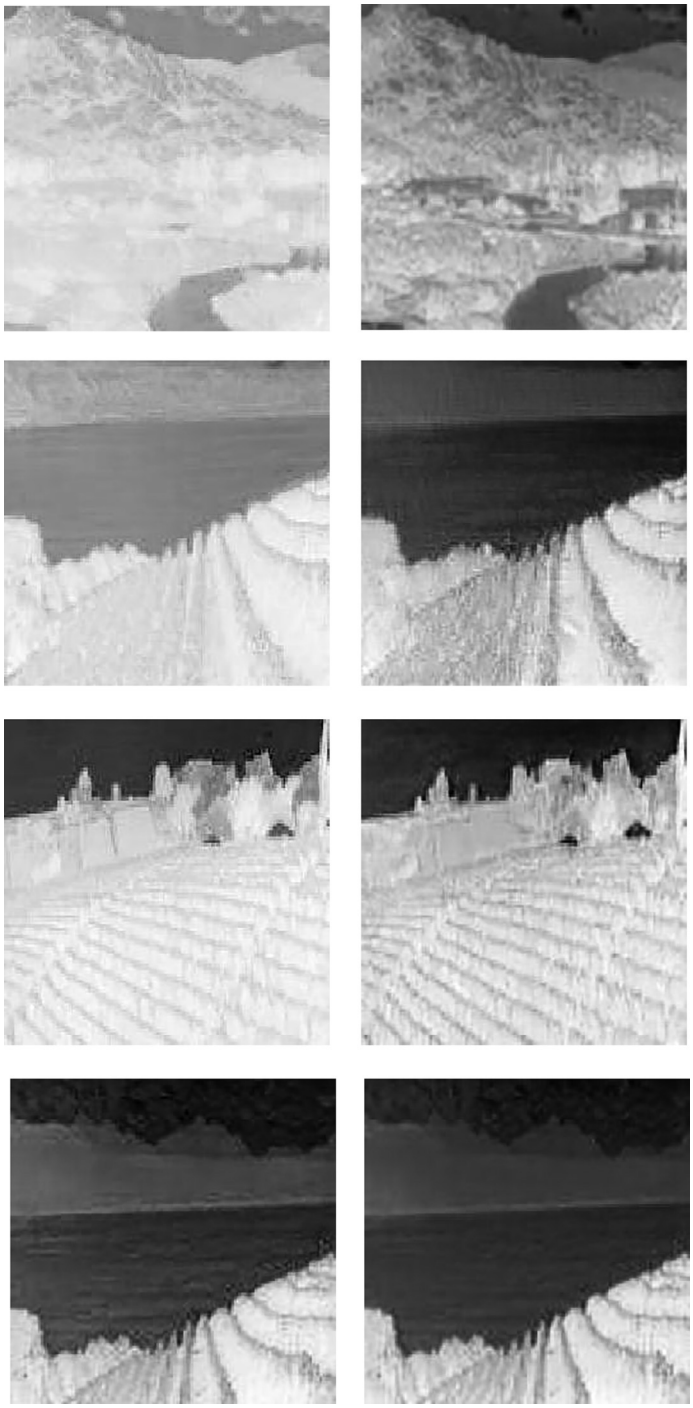

Fig. 9.14 Images of NDVI VIs from *mountain* category obtained with the synthetic NIR generated by the proposed CycleGAN. (*Left*) Ground-truth NDVI VI images. (*Right*) Estimated NDVI VIs. *(Images from M. Brown, S. Süsstrunk, Multi-spectral SIFT for scene category recognition, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, IEEE, 2011, pp. 177–184.)*

Fig. 9.15 Images of NDVI VIs obtained from *country* with the proposed CycleGAN implemented in this paper: (*first col*) NDVI estimated with Ref. [17]; (*second col*) NDVI estimated by the proposed CyclicGAN; (*third col*) Ground-truth NDVI VI. *(Images from M. Brown, S. Süsstrunk, Multi-spectral SIFT for scene category recognition, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, IEEE, 2011, pp. 177–184.)*

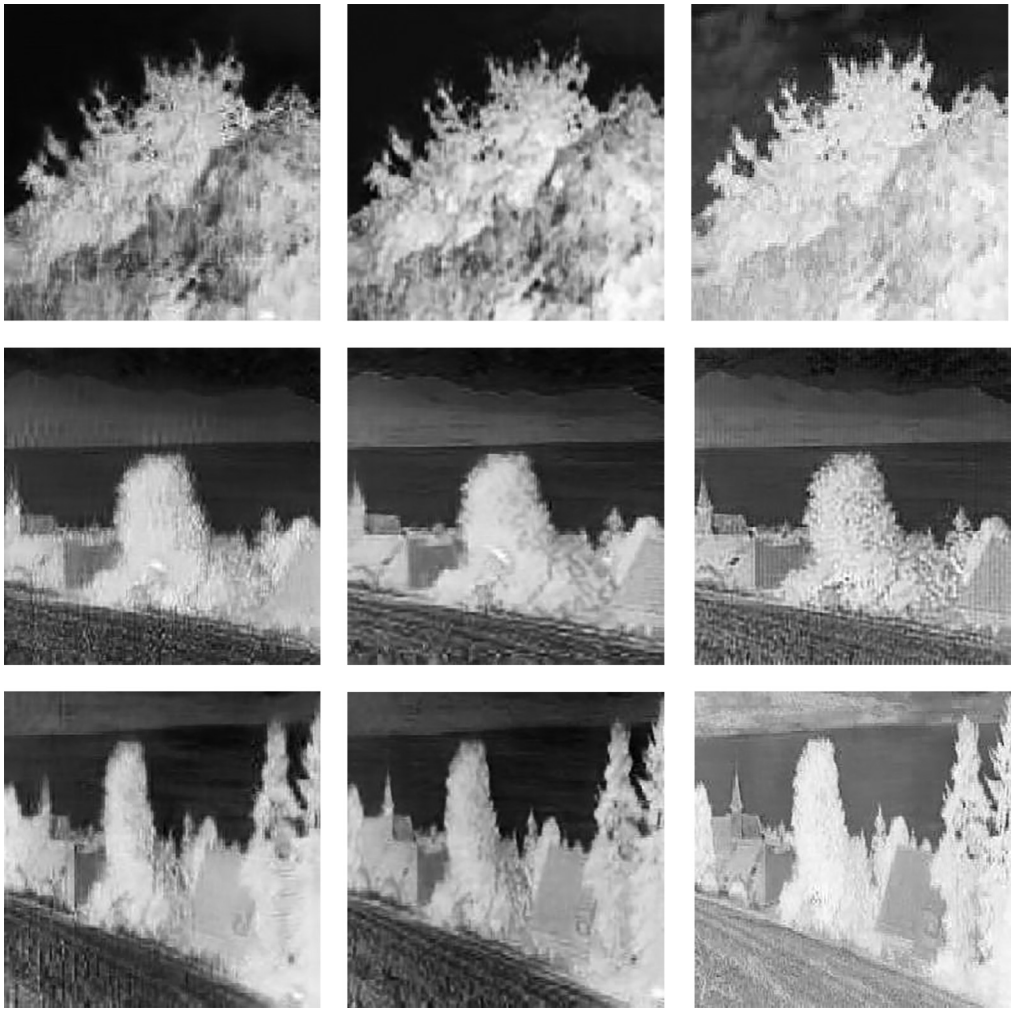

Fig. 9.16 Images of NDVI VIs from *field* obtained with the proposed CycleGAN implemented in this chapter: (*first col*) NDVI estimated with Ref. [17]; (*second col*) NDVI estimated by the proposed CyclicGAN; (*third col*) ground-truth NDVI VI. *(Images from M. Brown, S. Süsstrunk, Multi-spectral SIFT for scene category recognition, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, IEEE, 2011, pp. 177–184.)*

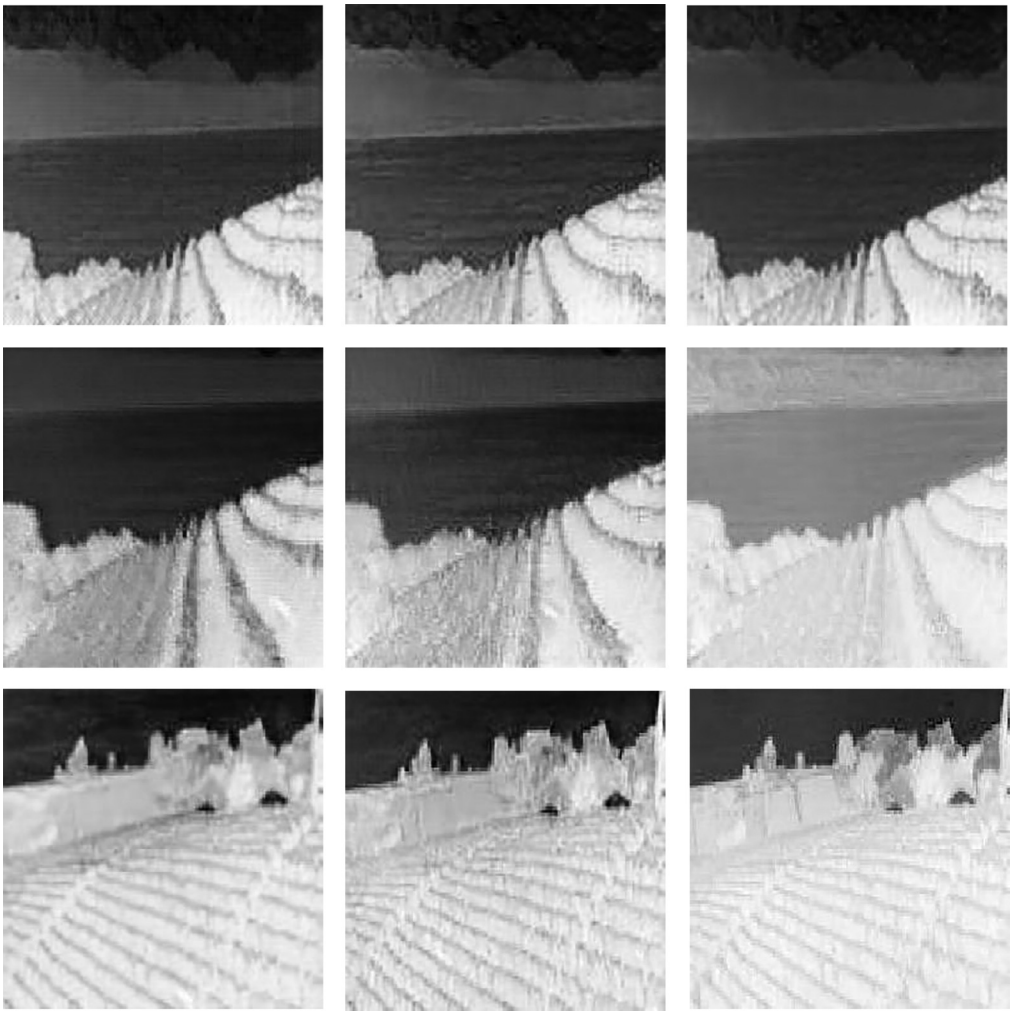

Fig. 9.17 Images of NDVI VIs from *mountain* obtained with the proposed CycleGAN implemented in this chapter: (*first col*) NDVI estimated with Ref. [17]; (*second col*) NDVI estimated by the proposed CyclicGAN; (*third col*) ground-truth NDVI VI. *(Images from M. Brown, S. Süsstrunk, Multi-spectral SIFT for scene category recognition, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, IEEE, 2011, pp. 177–184.)*

Table 9.1 Average root means squared errors (RMSE) and structural similarities (SSIM) obtained from the estimated NDVI and the real one computed from Eq. (9.1) (SSIM the bigger the better).

	RMSE			SSIM		
Training	**Country**	**Field**	**Mountain**	**Country**	**Field**	**Mountain**
Supervised approach: results from Ref. [16]	3.53	3.70	–	0.94	0.91	–
Unsupervised approach: results from Ref. [17]	3.46	3.53	3.82	0.93	0.90	0.88
Proposed NDVI estimation with $\mathcal{L}_{\text{FINAL}-\text{SYN}-\text{NIR}_{\text{CYCLE}-\text{LSGAN}}}$	3.39	3.56	3.81	0.94	0.92	0.89

Notes: NDVI values are scaled up to a range of [0–255] since they are depicted as images as shown in Figs. 9.15–9.17.

9.5 Conclusions

This chapter tackles the challenging problem of generating NDVI VI using an NIR synthetic image and its corresponding RGB representation. NIR images are estimated by using a CycleGAN network. Results have shown that in most of the cases the network is able to obtain reliable synthetic NIR representations that can be used to obtain VIs. As mentioned in Section 9.4, this approach has not the limitation of needing paired NIR-RGB images for training. As a future work, actually, as work in progress we are considering the use of a CycleGAN architecture with continual learning with deep generative display, but feed it with RGB and their corresponding NIR image in the generator to speed up the generalization. Future work will also consider other loss functions to improve the training process.

Acknowledgments

This work has been partially supported by the ESPOL project PRAIM (FIEC-09-2015); the Spanish Government under Project TIN2017-89723-P; and the "CERCA Programme/Generalitat de Catalunya." The authors also thank NVIDIA for GPU donations and the CYTED Network: "Ibero-American Thematic Network on ICT Applications for Smart Cities" (REF-518RT0559).

References

[1] S.F. Di Gennaro, F. Rizza, F.W. Badeck, A. Berton, S. Delbono, B. Gioli, P. Toscano, A. Zaldei, A. Matese, UAV-based high-throughput phenotyping to discriminate barley vigour with visible and near-infrared vegetation indices, Int. J. Remote Sens. 39 (15–16) (2018) 5330–5344.

[2] M.F. Dreccer, G. Molero, C. Rivera-Amado, C. John-Bejai, Z. Wilson, Yielding to the image: how phenotyping reproductive growth can assist crop improvement and production, Plant Sci. 282 (2019) 73–82.

[3] M. Wójtowicz, A. Wójtowicz, J. Piekarczyk, Application of remote sensing methods in agriculture, Commun. Biometry Crop Sci. 11 (1) (2016) 31–50.