SAR-OPTICAL IMAGE TRANSLATION USING SUPERVISED CYCLE GENERATIVE ADVERSARIAL NETWORK (S-(PRO)-CYCLEGAN)

Taresh Batra,
BITS Pilani, K.K. Birla Goa Campus
South Goa, Goa, India
f20190388@goa.bits-pilani.ac.in

ABSTRACT

Optical remote sensing data is highly susceptible to weather conditions, especially cloud formation, whereas synthetic aperture radar (SAR) can work under all weather conditions and overcome this disadvantage of optical RS data. However, SAR imagery poses different challenges like speckle noise caused by its imagery mechanism. It is also hard to recognise land cover types visually from SAR images. Thus, SAR-to-Optical image translation has vast utility in remote sensing. However, the geometric and radiometric differences between the two modalities make this task tricky. The model discussed here is primarily inspired by the great work done by Wang et al. [1] and Karras et al. [2]. The discussed supervised cycle consistency GAN has classic U-NET architecture and PatchGAN classifier (discriminator) [3]. The dataset was generated from Sentinel-1 (SEN-1) SAR images and Sentinel-2 (SEN-2) multi-spectral images. The model uses techniques like equalised learning rate and layer normalisation introduced in the ProGAN paper to smoothen the training process. The model has not been "progressively grown." However, doing so is a future objective.

***K*eywords** SAR-Optical Image Translation · Synthetic Aperture Radar (SAR) · Generative Adversarial Networks (GANs) · Sentinel

1 Introduction

Synthetic aperture radar remote sensing (RS) sensors have found wide applications in land use planning, disaster prevention, target detection [4] - [6], etcetera. The possibility of acquiring calibrated signals from a satellite platform and the absence of restrictions related to atmospheric conditions make SAR imagery a very fruitful prospect. SAR imagery can overcome the disadvantages of optical RS data and has the advantages of working all day and night under all weather conditions, large scope, and specific penetration capacity. However, it is hard to distinguish structural information in SAR images, which does not necessarily ease with an increase of spatial resolution, primarily for three reasons. Firstly, the colour information of land cover types in SAR images is very different from optical RS images. Secondly, SAR images are plagued by speckle noise, which never appears in optical RS images. Lastly, the imaging mechanism of SAR leads to geometric distortion and shadows. Thus, it is expected that users of SAR data may wish to use additional tools to facilitate the interpretation of SAR images. This paper attends to this idea and investigates the potential of image translation GANs in translating SAR images to optical images to ease the now cumbersome process of their interpretation. SAR images translated into near-optical representations motivated by the human eye's familiarity with such representations can be of great use. Besides providing a visual aid in interpreting SAR data, it also finds applications in follow-up techniques like fusing complementary SAR and optical data (image matching).

2 Background of Optical and SAR sensors

This paper section has been inspired mainly by the dissertation written by Dr Nina Merkle [7]. The objective behind this section is to provide readers with a more profound sense of the and theory and technicalities involved with the problem at hand.

2.1 Principles

Optical satellite sensors are passive systems that measure the sunlight reflected from objects on the ground with a strong dependence on local atmospheric and lighting conditions. More precisely, they detect the reflected or emitted electromagnetic radiation from objects in the visible and infrared (near-infrared, intermediate infrared, thermal infrared) range of the electromagnetic spectrum. Each object on the ground reflects and absorbs a specific part of the spectrum, showing a unique spectral reflectance signature in the generated images. Unlike optical satellites, radar satellites have an active sensor onboard, which emits electromagnetic signals and measures the strength and time delay of the returned signal backscattered from the ground objects. During image acquisition, the reflected signal's range, magnitude, and Doppler shift are collected by an antenna and later processed to a two-dimensional image of the surface (see Figures 1 and 2). Due to the active emitting of a signal and usage of a longer wavelength than optical sensors, images can be captured almost independently from local weather conditions and lighting conditions (i.e., day and night).

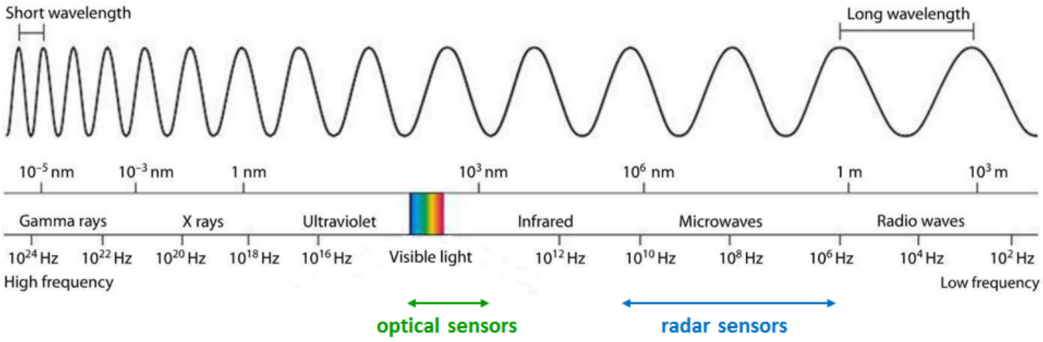

Figure 1: The electromagnetic spectrum and the operation ranges of optical and radar sensors (image source: [8])

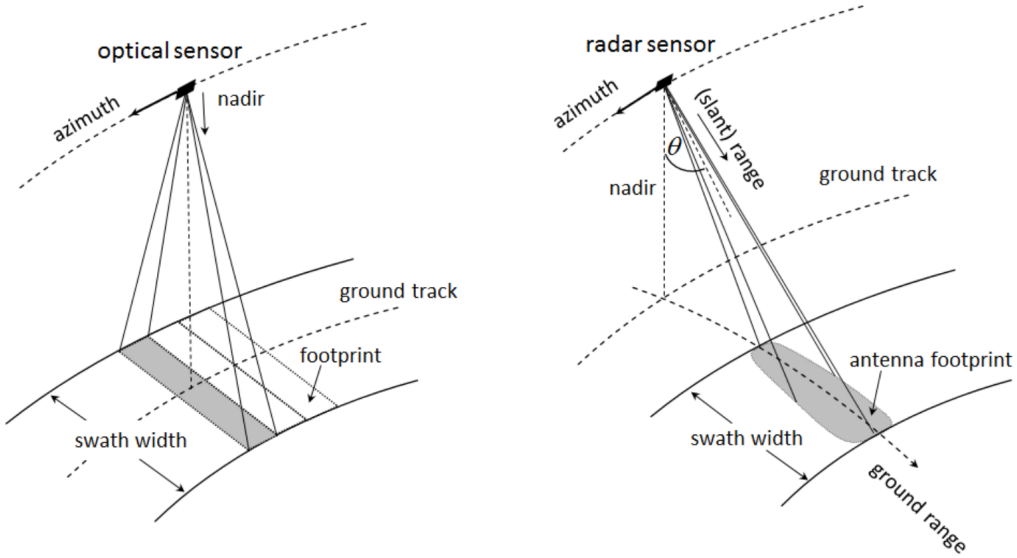

Figure 2: Comparison of the different acquisition geometries between optical and SAR sensors (sources of the images: [7],[9])

2.2 Radiometric and Geometric Characteristics

In the case of optical and SAR satellites, the images acquired by both sensors exhibit pretty different properties that characterise the images. In particular, the specific acquisition principle of a radar sensor and the resulting image effects make the visual interpretation and usage of SAR images a challenging task [10].

2.2.1 Radiometric Properties

The different wavelengths measured and utilised by optical and SAR sensors lead to different radiometric properties in the images. These differences occur because the response of an object depends on the signal properties (wavelength, polarisation), the surface properties (roughness, randomness of local reflectors and reflectance properties) and sensor perspective. The visible and near-infrared part of the spectrum utilised in optical satellite imagery enables the collection of information about the chemical structure of an object. Therefore, the pixel intensity values of optical images contain information about the chemical characteristics of an observed area. While SAR sensors, on the other hand, utilise electromagnetic signals with a much lower frequency and energy (i.e., higher wavelength). Therefore, the obtained images mainly capture the physical and geometrical properties of the objects, where the pixel intensity values contain information about the roughness, the electrical conductivity and the orientation of an object to the sensor [11]. Thus, the same object in an optical and SAR image may appear with high intensity for one sensor and low intensity for the other. Another effect in SAR images is called speckle, which further complicates the human and automatic interpretation of the images. Speckle is due to the coherent interference of waves reflected from many scatterers in each resolution cell. Thus, the neighbouring pixels may show a high variation in their pixel intensity values (see Figure 3).

Figure 3: Illustration of an optical (top) and SAR image (bottom) covering the same area. Both images have a ground sampling distance of 1.25 m. (image source: [7])

2.2.2 Geometric Properties

Due to the different imaging techniques used in optical and SAR satellites, the corresponding images exhibit different geometric properties for all objects above the ground. For optical systems, objects located perpendicular to the direction of flight get projected away from the sensor in the image plane. The opposite is the case for radar systems, where the image geometry gets derived through the travelling time of the backscattered signal (see Figures 4 and 5). SAR images

show three typical distortion effects: foreshortening, layover, and shadowing. These effects strongly influence the visual appearance within an image of the above-ground objects.

Foreshortening denotes the shortening of a distance between two points during the projection in the image plane. This effect occurs if a slope facing the sensor has an angle smaller than the incident angle θ of the SAR sensor or a slope facing away from the sensor with an angle smaller than $90° \theta$.

Layover appears if a slope facing the sensor has an angle higher than the incident angle of the SAR sensor. This effect is quite common in urban and mountainous areas. Consequently, areas affected by overlay appear relatively bright (due to the overlay of the signal response), and buildings and steep mountains can appear upside down in SAR images.

Shadowing appears if a terrain slope is oriented away from the sensor at an angle higher than the incident angles of the sensor. An important thing to note is that no information is gained from these shadowed regions. These appear as dark areas in the images. All introduced radar effects depend on the viewing direction of the sensor and the geometry of the targeted object on the ground.

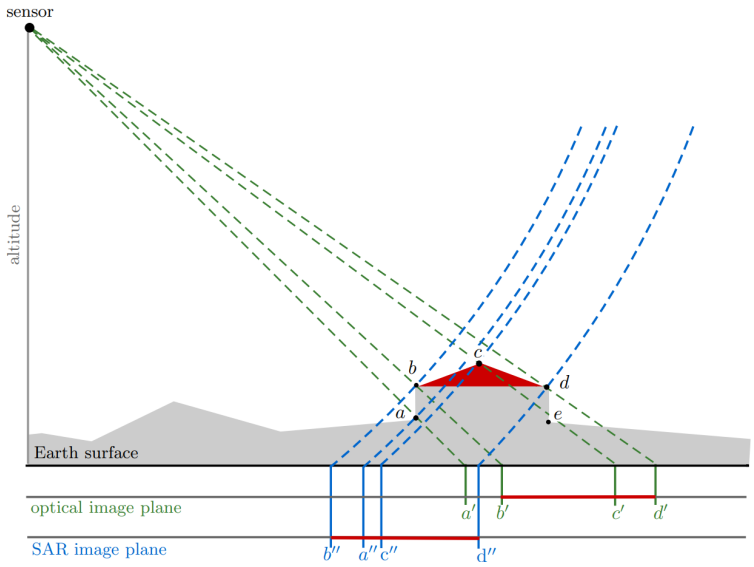

Figure 4: Comparison of optical and SAR imaging. The green (blue) marked lines illustrate the projection of the four points a to d on the Earth surface into the optical (SAR) image plane. Elevated points such as point c are shifted away from the sensor in the optical image plane and towards the sensor in the SAR image plane. The point e is neither seen by the optical nor SAR sensor, and hence not present in the acquired images. (image and caption source: [7])

3 Related Work

Having looked at the various challenges imaging faces in both modalities (Optical and SAR), it is evident that a viable solution is required. This problem has received the attention of many researchers and thus has had some exquisite solutions proposed. Most researchers focused on the problem of image matching between the two modalities. Optical and synthetic aperture radar (SAR) images contain complementary information about objects on the Earth's surface, which enriches the information conveyed by an object for several applications. Most successful solutions proposed using a deep network to translate SAR images to optical images to perform image matching. This paper only looks at the first half of this problem. Dr Nina Merkle was one of the first researchers to focus on the image translation aspect of the aforementioned SAR-Optical image matching problem. She trained a conditional GAN (cGAN) to translate SAR images to optical images, which aided the matching process [7]. This clever method of dealing with this problem has since been a hot research topic.

One of the most revolutionary papers in image-to-image translation is pix2pix [3]. Isola et al. popularised using U-NET [12] architecture for image translation and introduced the PatchGAN classifier as the discriminator of the model. It achieved compelling results. The PatchGAN classifier helped stabilise the GAN training process, which is a painstaking job. Instead of classifying an entire image as real or fake, PatchGAN classifies patches in an image as real or fake. Translation of unpaired data is known to be a notoriously tricky task. One very effective method to address this problem was introduced by making GANs consistent with their results [13]. Zhu et al. introduced CycleGANs which used the

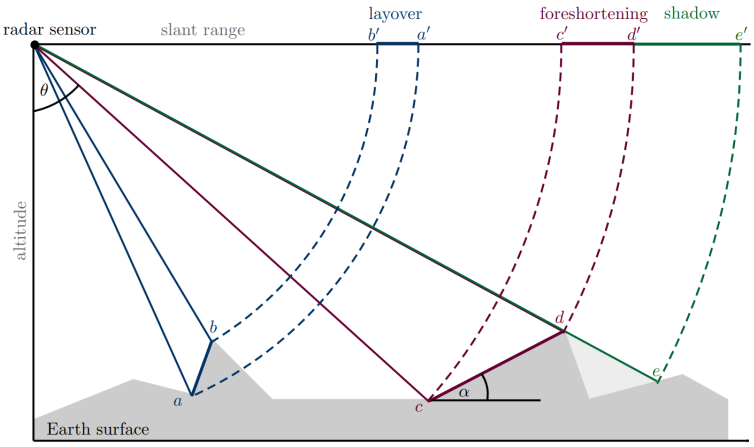

Figure 5: Illustration of the geometric distortion effects layover (marked blue), foreshortening (marked dark red) and shadowing (marked green) for SAR images. Layover: an observed object appears upside down in the image plan; Foreshortening: an observed object or ground segment appears shortened in the image plan; Shadow: non-visible regions appear as dark areas in the image.(image and caption source: [7])

concept of "consistency". Consistency, here, refers to a GAN's ability to be consistent with its output. Consistency was achieved using 2 GANs, say G_1, which converts images of type A to type B and its counterpart G_2, which converts images from B to A (see Figure 6). The model minimises the difference between $G_2(G_1(A))$ and the original image A. Thus, making the model reason about its consistency.

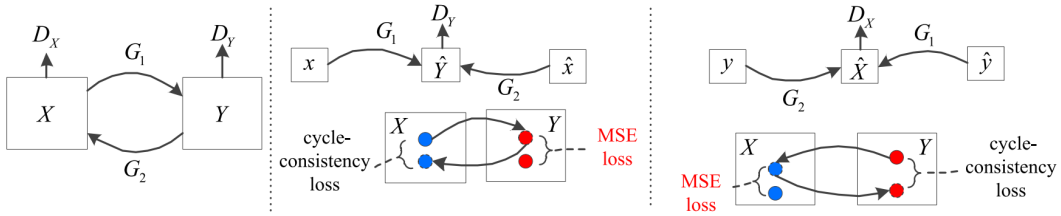

Figure 6: Illustration of cycle consistency loss (source of image: [1])

Although the Sentinel dataset used to address the SAR-Optical image translation contains paired images. Combining the virtues of pix2pix and CycleGAN yields superior results in paired image translation. Wang et al. did so for SAR-Optical image translation using an S-CycleGAN (Supervised Cycle GAN) [1]. The authors of this paper took the U-NET architecture and the PatchGAN classifier from the pix2pix network and added the consistency aspect of the cycle GAN. The resulting image translation network outperformed both its parent networks individually. The S-CycleGAN approach is the current state-of-the-art method in image translation nets. The model presented in this paper is a slight modification of the S-CycleGAN as presented by Wang et al. [1].

4 Key Ideas Used

4.1 Generative Adversarial Network (GAN)

Proposed by Goodfellow et al. in 2014, GANs have since been one of the hottest topics of research in generative modelling using deep learning. The concept of GANs earned plenty of attention in machine learning and offered new possibilities for several research problems through the generation of high-quality samples. GANs belong to the class of generative models and pursue the goal of learning the data distribution of a given dataset, commonly images, to generate new data from the learned distribution. It is well known that generative models are challenging to learn and come with high computational costs. Despite these problems, generative models and especially GANs made decisive progress over the last years. The success originates from the idea of reformulating the image generation problem by integrating it into a newly defined task, which can be learned through supervised learning techniques. Through this strategy, the original task will be learned as a by-product and difficulties of unsupervised learning can be avoided. Traditionally,

GANs consist of two networks, a generator G and a discriminator D. The generator learns the input data distribution to produce realistic data to fool the discriminator. On the other hand, the discriminator tries to distinguish whether its input is from the training data or the generator. This framework corresponds to a minimax two-player game (see Figure 7). GANs are unconditioned generative models that map input random noisy z to output y.

$G : z \rightarrow y$

The loss function of a traditional GAN is composed of two parts,

First, the predicted log probability of D that y is real,

$E_{y \sim p_{data}(y)}[logD(y)]$

Second, the predicted log probability of D that $G(z)$ is real,

$E_{z \sim p_z(z)}[log(1 - D(G(z)))]$

Combining the two,

$$\min_G \max_D L_{GAN}(G, D) = \min_G \max_D \quad E_{y \sim p_{data}(y)}[logD(y)] \quad + \quad E_{z \sim p_z(z)}[log(1 - D(G(z)))]$$

Conditional GANs (cGANs), a slight modification of GANs, are more useful in the context of our problem. In cGANs, a conditional setting is applied, meaning that both the generator and discriminator are conditioned on some auxiliary information (such as class labels or data) from other modalities. As a result, the ideal model can learn a multi-modal mapping from inputs to outputs by being fed with different contextual information (see Figure 8).

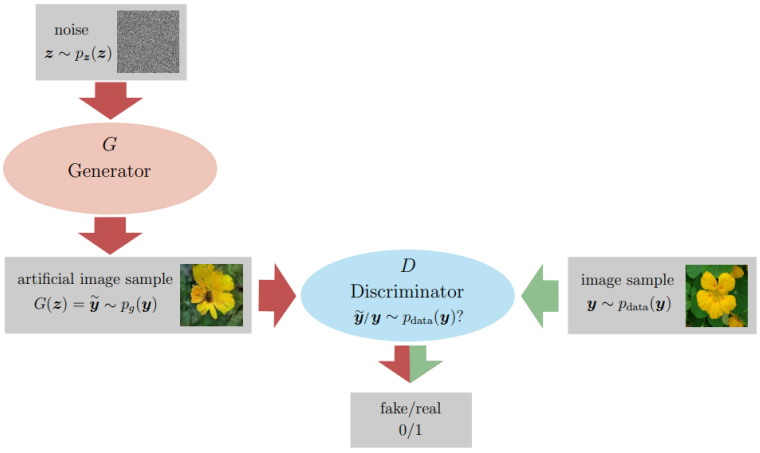

Figure 7: Illustration of the general GAN concept(source of image: [7])

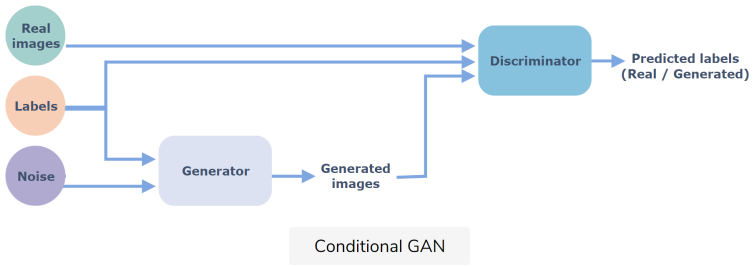

Figure 8: Illustration of the concept of conditional GANs (cGANs) (source of image: [15])

4.2 U-Net

U-Net is called so because of its structure (see Figure 9). Almost all state-of-the-art paired image-to-image translation solutions make use of a U-Net architecture. This autoencoder network steadily decreases the size of the input image while increasing its channels to encode it, and then using the encoding, it steadily increases the size while reducing the

number of channels to translate the image into its desired counterpart. This method is highly effective and follows a relatively stable training process due to the skip connections.

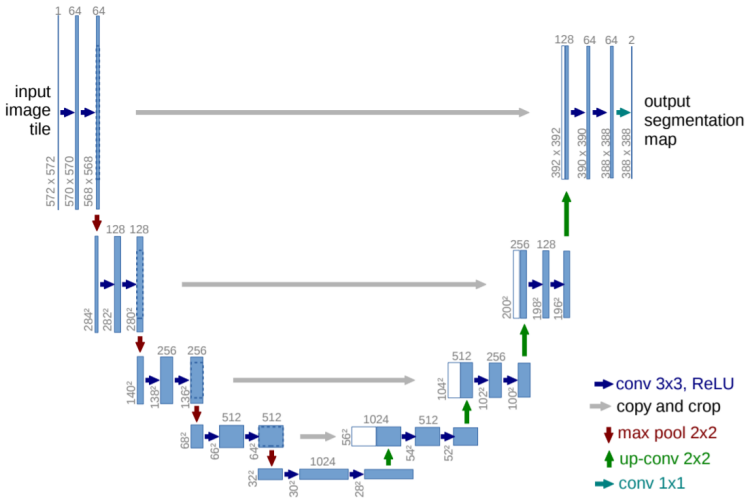

Figure 9: The original U-Net architecture as proposed by Ronneberger et al. [12] (source of image: [12])

4.3 Cycle Consistency Loss

This loss is pretty intuitive to understand. In order to nudge GANs towards being consistent with their results, a consistency loss was introduced by Zhu et al. [13]. The idea here is to train 2 GANs instead of just 1, such that both GANs translate the image of one kind into the other in an inverted way. G_1 maps the optical image to SAR image, and G_2 maps SAR image to optical image. The loss function in the learning process comprises the adversarial loss functions of both the GANs and a "cycle consistency" loss. The consistency loss has been discussed at length later in the paper.

4.4 Pixel Normalisation

Instead of using batch normalisation, as is commonly done, the authors of ProGANs [2] used pixel normalisation. This "pixelnorm" layer has no trainable weights. It normalises the feature vector in each pixel to unit length and is applied after the convolutional layers in the generator. In simpler terms, it normalises pixel values across channels. Doing so prevents signal magnitudes from spiralling out of control during training.

$$b_{x,y} = \frac{a_{x,y}}{\sqrt{\frac{1}{N}\sum_{j=0}^{N-1}(a_{x,y}^{j})^2+\epsilon}}$$

$\epsilon = 10^{-8}$, N is the number of feature maps, and $a_{x,y}$ and $b_{x,y}$ are the original and normalised feature vector in pixel (x, y), respectively (pixelnorm as presented by Karras et al. [2].)

4.5 Equalised Learning Rate

Here the weights are initialised normally (He's initialiser) and scaled according to the following equation.

$$W_f = W_i * \sqrt{\frac{2}{k*k*c}}$$

Learning rates can be equalised across layers by scaling the weights before every forward pass (as presented by Karras et al. [2]). The benefit of separate initialisation and scaling (done at every forward propagation), i.e., doing this dynamically instead of both during initialisation, relates to the scale-invariance in commonly used optimisation methods such as ADAM [16] and RMSProp [17]. These methods normalise a gradient update by its estimated standard deviation, thus making the update independent of the scale of the parameter. As a result, if some parameters have a larger dynamic range than others, they will take longer to adjust. This approach ensures that the dynamic range, and thus the learning speed, is the same for all weights.

4.6 PatchGAN Discriminator

Before pix2pix [3], most networks used a single Binary Cross-Entropy(BCE) loss function for their discriminator, only labelling the image as either 'fake' or 'real'. This method was not viable for deep GANs, especially the generator, as it suffered from the vanishing gradient problem. A PatchGAN discriminator, on the other hand, classifies patches of images as 'fake' or 'real', providing more implementable feedback for the generator to improve.

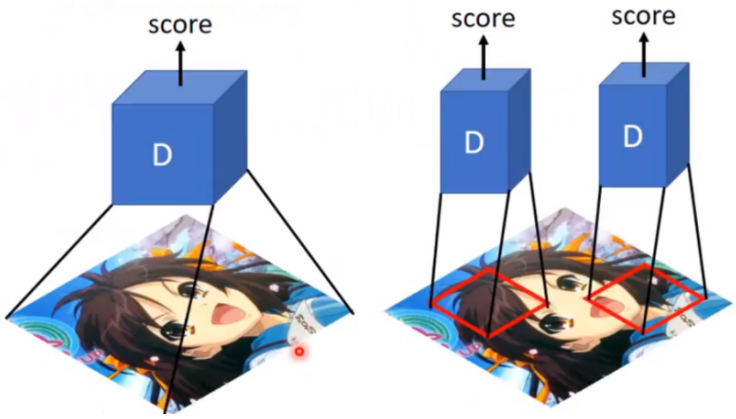

Figure 10: Illustration explaining the idea of a PatchGAN classifier. The usual GAN discriminator on the left scores the entire image as real (1) or fake (0). The PatchGAN discriminator, on the other hand, scores patches of the image. [12] (source of image: [20])

5 Proposed Model

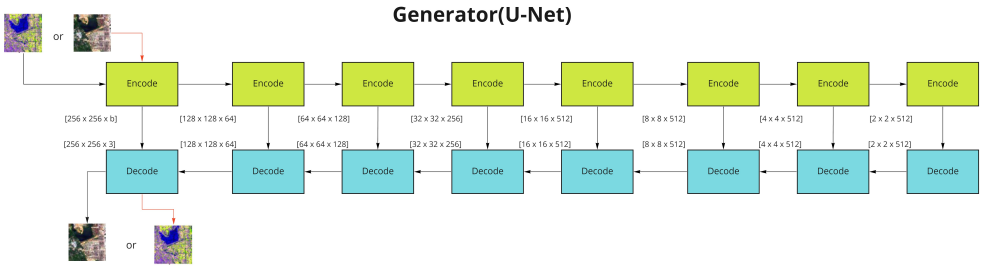

Figure 11: Generator architecture used

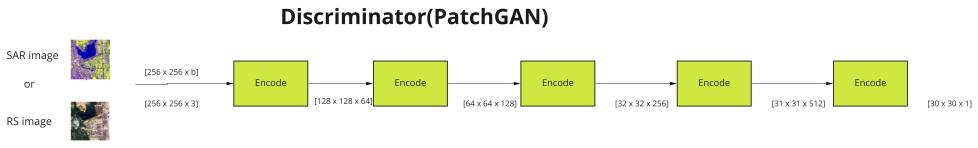

Figure 12: Discriminator architecture used

5.1 Generator

The generator used follows an autoencoder U-Net architecture with eight encoder blocks and eight decoder blocks since the model proposed is trained on the Sentinel 1-2 dataset (256 x 256, i.e., 2^8). The equalised learning rate algorithm is

used throughout the generator to stabilise training. Each encoder block reduces the image size and increases the number of channels by a factor of two (after the fourth block, the number of channels is frozen). In contrast, each decoder block increases the image size and reduces the number of channels by a factor of two (until the fourth block, the number of channels is frozen) (see Figure 13). The ProGAN [2] generator architecture inspires the structure of these modules. Upscaling in the decoder module is done using the nearest neighbour (NN) interpolation algorithm. This method is chosen over the transpose convolution operation to avoid the classic chessboard pattern caused by them. Using NN followed by convolution yields better results and is also more scalable [18]. The encoder uses average pooling for downscaling. Most translation models have one convolutional/transpose-convolutional layer per encoder/decoder block; however, the ProGAN architecture had two. So, taking the middle road, an extra skip connection was added to ease any potential difficulties caused by increasing the number of parameters. Two such generators are used in the consistency loss scheme (G_1 and G_2).

5.2 Discriminator (PatchGAN)

This discriminator tries to classify if each $N \times N$ patch in an image is real or fake. This discriminator is run across the image convolutionally, averaging all responses to provide the ultimate output. N can be much smaller than the full image size and still produce high-quality results. This is advantageous because a smaller PatchGAN has fewer parameters, runs faster, and can be applied to arbitrarily large images. Such a discriminator effectively models the image as a Markov random field, assuming independence between pixels separated by more than a patch diameter. PatchGAN can be understood as a form of texture/style loss, as suggested by the authors pix2pix [3]. PatchGAN, as presented in pix2pix, was used as is in the proposed model. Wasserstein Loss (Earth mover's loss) with gradient penalty was also considered; however, the findings of Makow[19] suggested that Wasserstein GANs (WGANs) do not aid the process of image-to-image translation and take a longer time to train. Thus, PatchGAN seemed to be the optimal choice.

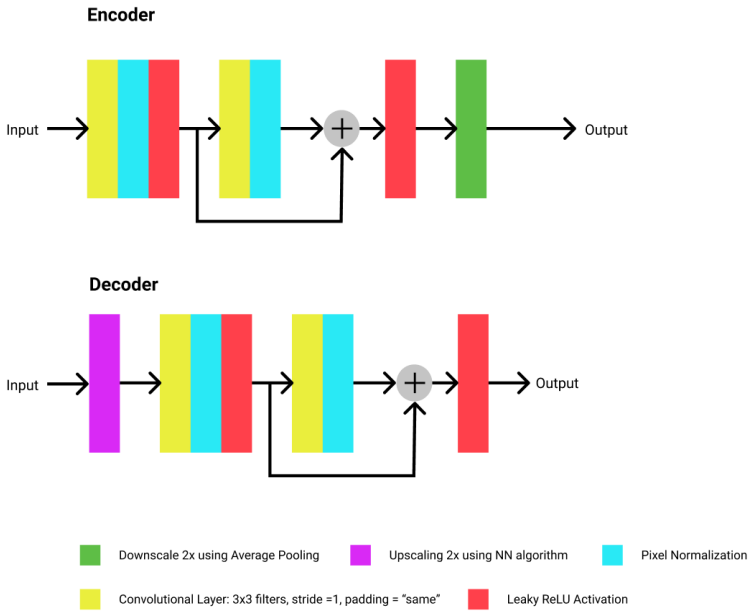

Figure 13: Encoder and decoder architectures used

5.3 Loss Function

5.3.1 Consistency Loss

For each input image x, the image translation cycle should be able to bring x back to the original image, i.e., $x \rightarrow G_1(x) \rightarrow G_2(G_1(x)) \approx x$ Similarly, for each image y from domain Y, G_1 and G_2 should also satisfy backward cycle consistency. $y \rightarrow G_2(y) \rightarrow G_1(G_2(y)) \approx y$. Therefore, the cycle consistency loss is formulated as follows.

$$L_{cons}(G_1, G_2) = E_{x \sim p_{data}(x)}[||G_2(G_1(x)) - x||_1] + E_{y \sim p_{data}(y)}[||G_1(G_2(y)) - y||_1]$$

5.3.2 GAN Loss

The generator G_1 tries to generate images $G_1(x)$ that look similar to images from domain Y and the discriminator D_Y aims to discriminate between y and $G_1(x)$. In the same way, G_2 tries to generate images $G_2(y)$ that look similar to images from domain X and the discriminator D_X aims to discriminate between x and $G_2(y)$. The adversarial losses for the two mappings are as follows.

$L_{GAN}(G_1, D_Y, X, Y) = E_{y \sim p_{data}(y)}[logD_Y(y)] + E_{x \sim p_{data}(x)}[log(1 - D_Y(G_1(x))]$

$L_{GAN}(G_2, D_X, Y, X) = E_{x \sim p_{data}(x)}[logD_X(x)] + E_{y \sim p_{data}(y)}[log(1 - D_X(G_2(y))]$

5.3.3 Mean Squared Error (MSE) Loss (Supervised)

The pixel-level MSE loss named L_p is added to the two mappings of CycleGAN. It is calculated as follows.

$L_p(G_1, G_2) = E_{x,y \sim p_{data}(x,y)}[||G_1(x) - y||_2] + E_{y,x \sim p_{data}(y,x)}[||G_2(y) - x||_2]$

5.3.4 Combining the Losses

$L_{CycleGAN}(G_1, G_2, D_X, D_Y) = L_{GAN}(G_1, D_y, X, Y) + L_{GAN}(G_2, D_X, Y, X) + \lambda L_{cons}(G_1, G_2)$

finally, the proposed model's loss function, L_{model} is as follows.

$L_{model} = L_{CycleGAN}(G_1, G_2, D_X, D_Y) + \beta L_p(G_1, G_2)$

where, λ and β are hyperparameters.

5.4 Dataset

A paired SAR-optical image dataset was obtained from the Sentinel-1 (SEN-1) and Sentinel-2 (SEN-2) images. SEN-1 was conducted by the European Space Agency (ESA) and contained two satellites, Sentinel-1A and Sentinel-1B. Both satellites carry a C-band SAR sensor and provide dual-polarization SAR images in all-weather, day or night. SEN-2, which contains two satellites (Sentinel-2A and Sentinel-2B), was developed by ESA and can provide multi-spectral RS images.

References

[1] L. Wang et al., "SAR-to-Optical Image Translation Using Supervised Cycle-Consistent Adversarial Networks," in IEEE Access, vol. 7, pp. 129136-129149, 2019, doi: 10.1109/ACCESS.2019.2939649.

[2] Karras, Tero, et al. "Progressive growing of gans for improved quality, stability, and variation." arXiv preprint arXiv:1710.10196 (2017).

[3] Isola, Phillip, et al. "Image-to-image translation with conditional adversarial networks." Proceedings of the IEEE conference on computer vision and pattern recognition. 2017.

[4] J.-S. Lee and E. Pottier, Polarimetric Radar Imaging: From Basics to Applications. Boca Raton, FL, USA: CRC Press, 2009.

[5] G. Cheng, P. Zhou, and J. Han, "Learning rotation-invariant convolutional neural networks for object detection in VHR optical remote sensing images," IEEE Trans. Geosci. Remote Sens., vol. 54, no. 12, pp. 7405–7415, Dec. 2016.

[6] C. O. Dumitru, S. Cui, D. Faur, and M. Datcu, "Data analytics for rapid mapping: Case study of a flooding event in Germany and the tsunami in Japan using very high resolution SAR images," IEEE J. Sel. Topics Appl. Earth Observ. Remote Sens., vol. 8, no. 1, pp. 114–129, Jan. 2015.

[7] Merkle, Nina Marie, "Geo-localization Refinement of Optical Satellite Images by Embedding Synthetic Aperture Radar Data in Novel Deep Learning Frameworks," Dissertation, Universität Osnabrück, 2018.

[8] The Electromagnetic Spectrum. https://www.miniphysics.com/electromagnetic-spectrum$_2$5.$html$.

[9] S. Auer, "3D Synthetic Aperture Radar Simulation for Interpreting Complex Urban Reflection Scenarios," Dissertation, Technische Universität München, München, 2011.

[10] S. Auer and S. Gernhardt, "Linear Signatures in Urban SAR Images - Partly Misinterpreted?," IEEE Geoscience and Remote Sensing Letters, 11(10):1762–1766, 2014.

[11] C. Heipke. Photogrammetrie und Fernerkundung: Handbuch der Geodäsie, herausgegeben von Willi Freeden und Reiner Rummel. Springer Reference Naturwissenschaften. Springer Berlin Heidelberg, 2017.

[12] Ronneberger, O., Fischer, P. and Brox, T., 2015, October. U-net: Convolutional networks for biomedical image segmentation. In International Conference on Medical image computing and computer-assisted intervention (pp. 234-241). Springer, Cham.

[13] Zhu, J.Y., Park, T., Isola, P. and Efros, A.A., 2017. Unpaired image-to-image translation using cycle-consistent adversarial networks. In Proceedings of the IEEE international conference on computer vision (pp. 2223-2232).

[14] Goodfellow, I., Pouget-Abadie, J., Mirza, M., Xu, B., Warde-Farley, D., Ozair, S., Courville, A. and Bengio, Y., 2014. Generative adversarial nets. Advances in neural information processing systems, 27.

[15] What is a Conditional GAN (cGAN)?. https://www.educative.io/edpresso/what-is-a-conditional-gan-cgan

[16] Kingma, D.P. and Ba, J., 2014. Adam: A method for stochastic optimization. arXiv preprint arXiv:1412.6980.

[17] RMSProp. https://paperswithcode.com/method/rmsprop

[18] Zhang, X., Karaman, S. and Chang, S.F., 2019, December. Detecting and simulating artifacts in gan fake images. In 2019 IEEE International Workshop on Information Forensics and Security (WIFS) (pp. 1-6). IEEE.

[19] Wasserstein GANs for Image-to-Image Translation. https://cs230.stanford.edu/projects$_spring_2018/reports/8289943.pdf$

[20] Unpaired Image to Image Translation with CycleGAN. https://blog.paperspace.com/unpaired-image-to-image-translation-with-cyclegan/