INTERNATIONAL INSTITUTE OF INFORMATION TECHNOLOGY NAYA RAIPUR

MASTERS THESIS

Image to Image Translation of Sentinel-1 SAR images to Sentinel-2 Optical Images

---

Author:
Paritosh TIWARI

Supervisor:
Dr. Muneendra OJHA

A thesis submitted in fulfillment of the requirements for the degree of Master of Technology

in the

Deep Learning Lab
Department of Computer Science and Engineering

January 13, 2021

Declaration of Authorship

I, Paritosh TIWARI, declare that this thesis titled, "Image to Image Translation of Sentinel-1 SAR images to Sentinel-2 Optical Images" and the work presented in it are my own. I confirm that:

- This work was done wholly or mainly while in candidature for a research degree at this University.
- Where any part of this thesis has previously been submitted for a degree or any other qualification at this University or any other institution, this has been clearly stated.
- Where I have consulted the published work of others, this is always clearly attributed.
- Where I have quoted from the work of others, the source is always given. With the exception of such quotations, this thesis is entirely my own work.
- I have acknowledged all main sources of help.
- Where the thesis is based on work done by myself jointly with others, I have made clear exactly what was done by others and what I have contributed myself.

Signed: _____

Date: _____

"I'd rather die drunk, broke at 34 and have people at a dinner table talk about me, than live to be rich, and sober at 90, and nobody remember who I was."

Andrew (Whiplash 2014)

INTERNATIONAL INSTITUTE OF INFORMATION TECHNOLOGY NAYA RAIPUR

Abstract

Dr. Muneendra Ojha
Department of Computer Science and Engineering

Master of Technology

Image to Image Translation of Sentinel-1 SAR images to Sentinel-2 Optical Images

by Paritosh TIWARI

In this report we explore the difficulties associated with image-to-image translation when working with complex images and their possible solutions. Here, complex is equivalent to a neural network having to learn a greater number of features in order to set a certain standard of quality with its output. This is a walkthrough of the thought processes that ultimately culminate into approaches and ideas that reduce or remove certain roadblocks such as, colourising, assignment of colour channels, image sharpness and SAR noise and distortion. As our aim is to make this model robust enough to be deployed into production, our approach towards improving this model has to reflect the same. We finally end with suggestions of further research that could supplement this model such as, Cyclic GANs and data distillation and compression.

Keywords: Synthetic Aperture Radar (SAR), Generative Adversarial Networks (GAN), Sentinel-1, Sentinel-2, Noise, Distortion, Despeckling.

Contents

List of Figures

List of Abbreviations

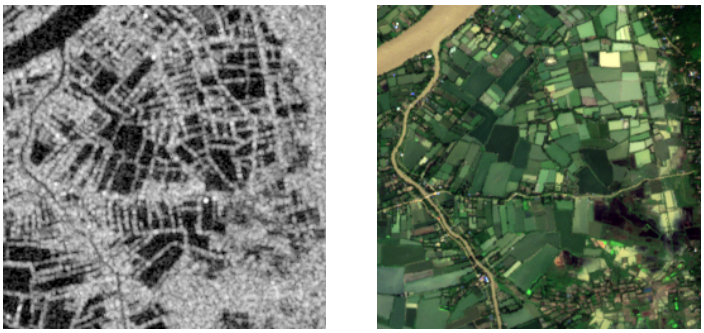
SAR	**S**ynthetic **A**perture **R**adar
RGB	**R**ed **G**reen **B**lue colour channels in the visual spectrum
S1	**S**entinel-**1** satellite
S2	**S**entinel-**2** satellite
GAN	**G**enerative **A**dversarial **N**etwork
PSNR	**P**eak **S**ignal to **N**oise **R**atio
SSIM	**S**tructural **Sim**ilarity Index
CNN	**C**onvolutional **N**eural **N**etwork

Chapter 1

Introduction

1.1 Objective

The goal here was to produce/construct RGB optical images (as captured by Sent-2) from SAR (Synthetic Aperture Radar) images (as captured by Sent-1) at production level.



(A) Sentinel-1 SAR (B) Sentinel-2 RGB (optical)

FIGURE 1.1: SAR and optical image pair

Ideally, we would want the final solution to take in S1 (Sentinel-1) images of a cloud covered region from a service which hosts such images, and output an image from the visual spectrum i.e., what S2 (Sentinel-2) would have captured had there been no cloud cover. Since, radar images are not affected by obstructions arising due to bad weather, it makes this project an ideal solution to reproduce data useful for human inspection that SAR images cannot capture, such as vegetation cover.

This image to image reproduction or translation can be performed by using a conditional GAN. For this report, we will treat the actual architecture of the project model as a black box. The generated image will be compared to the ground truth image (right) for accuracy and quality comparison.

The image dataset [**?**] is divided into 4 seasons: spring, summer, fall, winter and the images were captured across various geographic locations at different times. Their dimensions are $256 \times 256 \times 3$. The SAR image is actually single channel but the same pixel values are replicated across 3 channels so as to maintain symmetry in the model.

This report is an exploration and a survey of sorts into the domain of image-to-image translation. Our previous work had aimed to establish a theoretical and practical foundation; we now present our experiments conducted to establish a baseline and to test the limits of our model. This report will also be a critique of related work conducted in this domain, specifically that of satellite images.

Through our experiments we have come to realize that image translation models need to be improved as a whole. Separate components, that of a generator and

discriminator, cannot be improved separately as this often leads to one component overpowering the other and subsequent destabilization of the equilibrium and failure of convergence. We also came to know that, owing to their complexity, significant pre-processing of data is necessary when it comes to satellite images because a lot of times the data needs to be adjusted to the model instead of the other way round.

Having established our base of data distributions, GAN stabilization, diagnosing failure modes, usage of custom loss functions and varied architectures, we now begin work on establishing and improving our model. We also explore the feasibility of this model in production.

PSNR and SSIM scores have not been shown in this report as the quality of outputs does not warrant them. Human inspection is enough to draw out useful conclusions.

1.2 Related Work

In their paper, [**?**] have leveraged the pix2pix model [**?**] for SAR to RGB image translation. Their network is of a greater depth with several layers of 1024 feature maps, something which is beyond the scope of our GPU. We have thus scaled down our model to 512 feature maps. The data they have used is restricted to a single season suggesting loss of generalization. They have made no mention of solutions to mitigate this problem.

[**?**] have used residual networks in a fashion similar to Cycle-GANs to translate images. They have extended their model to include all 13 bands of Sentinel-2 and have incorporated a cloud adaptive loss, owing to the inclusion of a cloud contaminated image as an input along with the SAR image. However, their claim that their model will perform better because GANs are unstable seems to be baseless because all of the GANs in use for current research in this domain are relatively stable.

Chapter 2

Experiments

2.1 Experiment 1

We first ran the model over the entire dataset ($\approx 560,000$ images) for close to 50 epochs. We selected the image pairs randomly to maintain generalization. Due to the vastly varied nature of the images, the model was unable to learn anything significant. Terrain features (elevations and depressions) were captured to very little extent. The images shown 2.1 were generated in the last few training steps.

We then reduced the number of images trained the model over the fall season ($\approx 140,000$ images), but the results were similar.

As is evident from 2.1, our baseline model fails to capture the target image domain. Although major contours are recreated, the general colour scheme and reconstruction of landmarks seen in the optical images is non-existent.

This suggests that our network is not deep enough to capture all the features necessary to represent the images in the target domain or has not been trained for enough steps or both. Lack of data is out of question.

To be fair, the target domain is vast and the images we are using here are extremely complex as is expected from satellite images of terrain. Making the network deeper would be a brute-force approach and does not seem to be an elegant solution.

2.2 Experiment 2

Next, we followed the experimental guidelines of [**?**] and selected 24,000 images for training, and 1,070 images for testing, from the fall season. These images were evenly distributed (as close as possible, given our manual selection) over different types of terrain. The model was trained for 10 epochs. The results improved to certain degree 2.2. Our model was able to learn major geographical features, but image colourization remained inaccurate. Also, it was unable to learn the finer features of images, such as narrow roads, boundaries, rooftops etc.

Training the model on a relatively smaller dataset definitely shows some improvement. The domain has been reduced to a single season although the types of terrain still varies greatly. The restriction of seasons gives us many advantages such as: our model will no longer be confused with the colour scheme when the same terrain type is covered with snow/ice during the winter season.

The contours are captured is much greater detail but the colourisation still remains less that adequate. The reduction in the size of the target domain does not reflect on the colours generated.

This shortcoming now points to the quality of SAR images. The model will not learn what it cannot distinguish. The noise and distortion in captured SAR images hinders the ability of our model to distinguish among building tops, slightly grassy

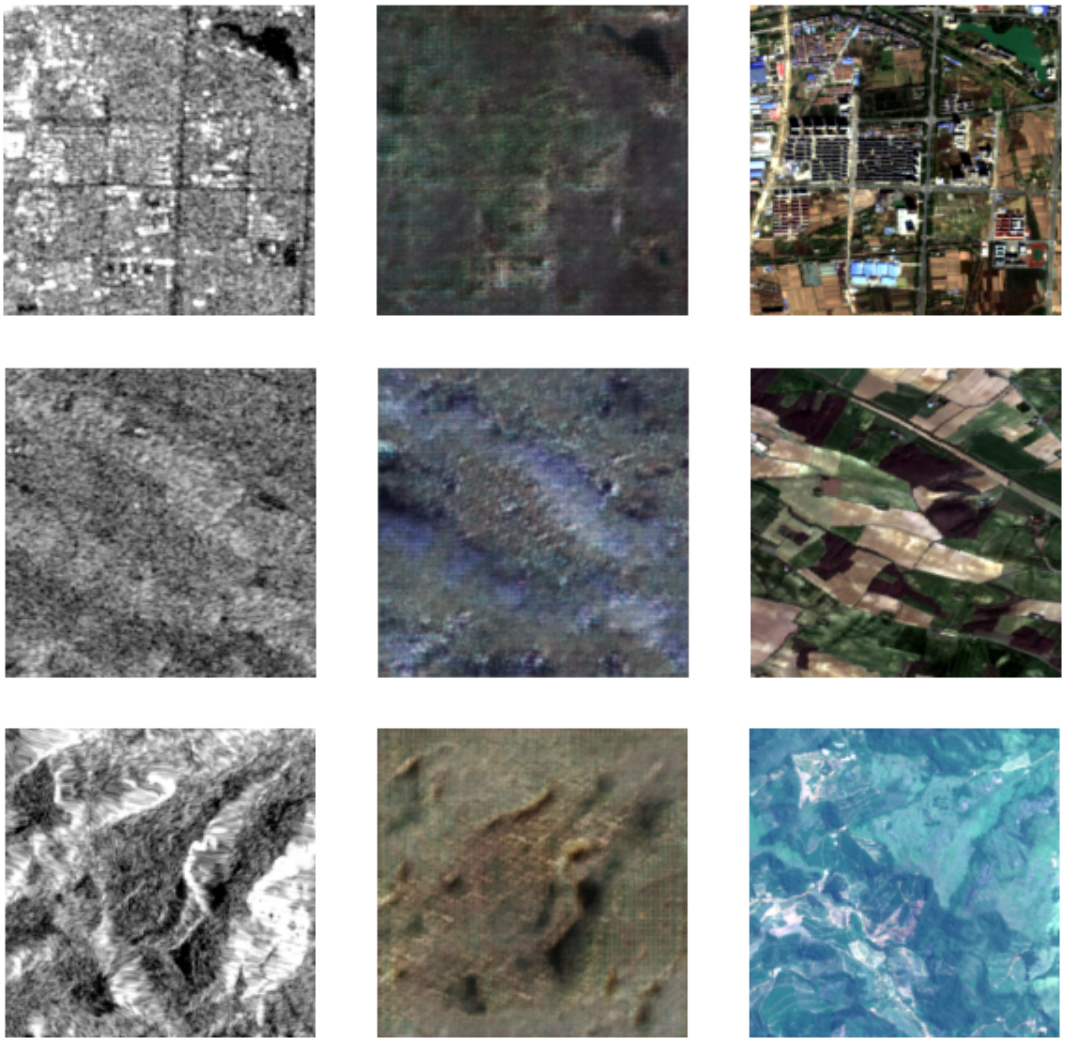


FIGURE 2.1: Exp 1: SAR (left), generated image (centre), ground truth (right)

and barren terrain as the SAR images for all three are more or less similar. Restricting the target domain further based on terrain type would make us lose generalization.

Still, on the bright side, our model has extracted features from the SAR images and translated then to optical images.

2.3 Experiment 3

To test the model on simpler images i.e., images with relatively less features to be learned by the neural net, the same model, with unchanged parameters was trained on 1,200 RGB images of dogs, with dimensions of 256×256, from the Linnaeus 5 dataset **[?]**. Here, the source was the 3-channel grayscale image generated by the program and the target was the original dog image 2.3.

The grayscale images were generated by a matrix dot product between the RGB image and the vector $[0.2989, 0.5870, 0.1140]$ which is the matplotlib standard. The single channel result was then repeatedly stacked thrice to generate the 3-channel grayscale image.

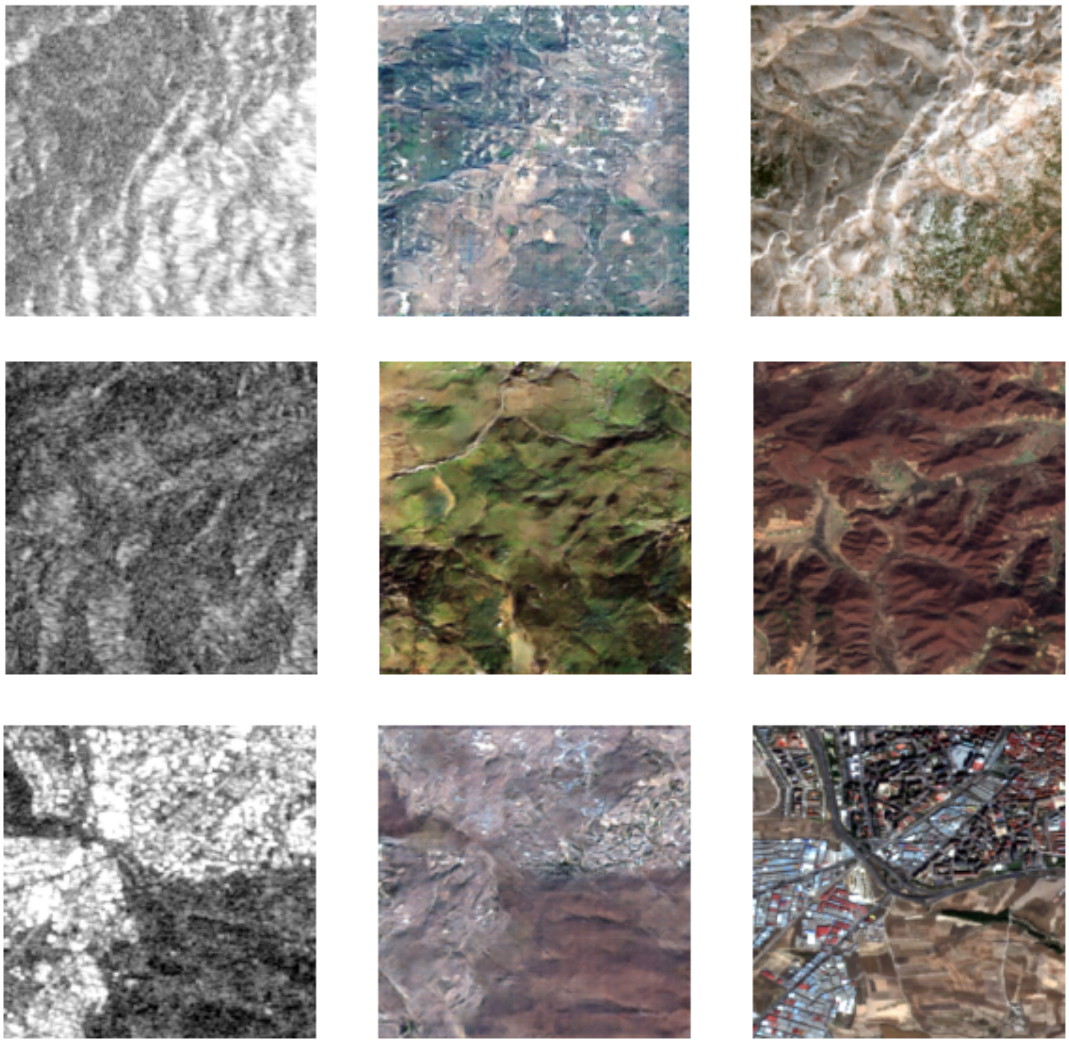


FIGURE 2.2: Exp 2: SAR (left), generated image (centre), ground truth (right)

Training our model on a reduced target domain of a smaller number of images gives us positive results. The dog images in the data set are simpler i.e., they have a smaller number of features to be learnt. Our model quickly learns to colourise major features in the grayscaled images. Further training for more epochs and a slightly larger number of images will yield better results.

As we can see from some of the generated images, there are undesirable artifacts with the colours assigned to the same feature. This suggests that our model is still learning to assign proper colour channels to the grayscale image features. This could also be a problem plaguing our original satellite model.

2.4 Experiment 4

To emulate the noise/distortion in SAR images, we trained the model to generate coloured versions of noisy grayscale images of the same dataset of dog images. The additive noise was introduced via Gaussian noise with a mean of zero and a standard deviation of 0.4.

FIGURE 2.3: Exp 3: Grayscale (left), generated image (centre), ground truth (right)

Since, the images were re-scaled to pixels values of $[-1, 1]$ to account for the tanh activation function in the generator network, the introduction of Gaussian noise pushed some pixel values beyond the tanh threshold. As a result, the generated plots were clipped to accommodate for the value overflow. Therefore, the noisy images appear relatively less noisy (when compared to SAR images), but the actual distortion is much greater. This plot clipping had no effect on the model as the generator was trained before the pixel values were adjusted 2.4.

Our model colourises our generated images relatively well, despite the noise and distortion. The generated images, however, are not as sharp as the images in our previous experiment. This leads us to the conclusion that reducing the noise and distortion in SAR images will only serve to make the generated images sharper but might leave the colour assignment problem unsolved.

2.5 Experiment 5

In order to reduce the noise/distortion in SAR images, we trained a CNN model to reduce the noise in our dog image dataset. The objective was to first build the model around a simple dataset and then scale it up to handle more complex images. The

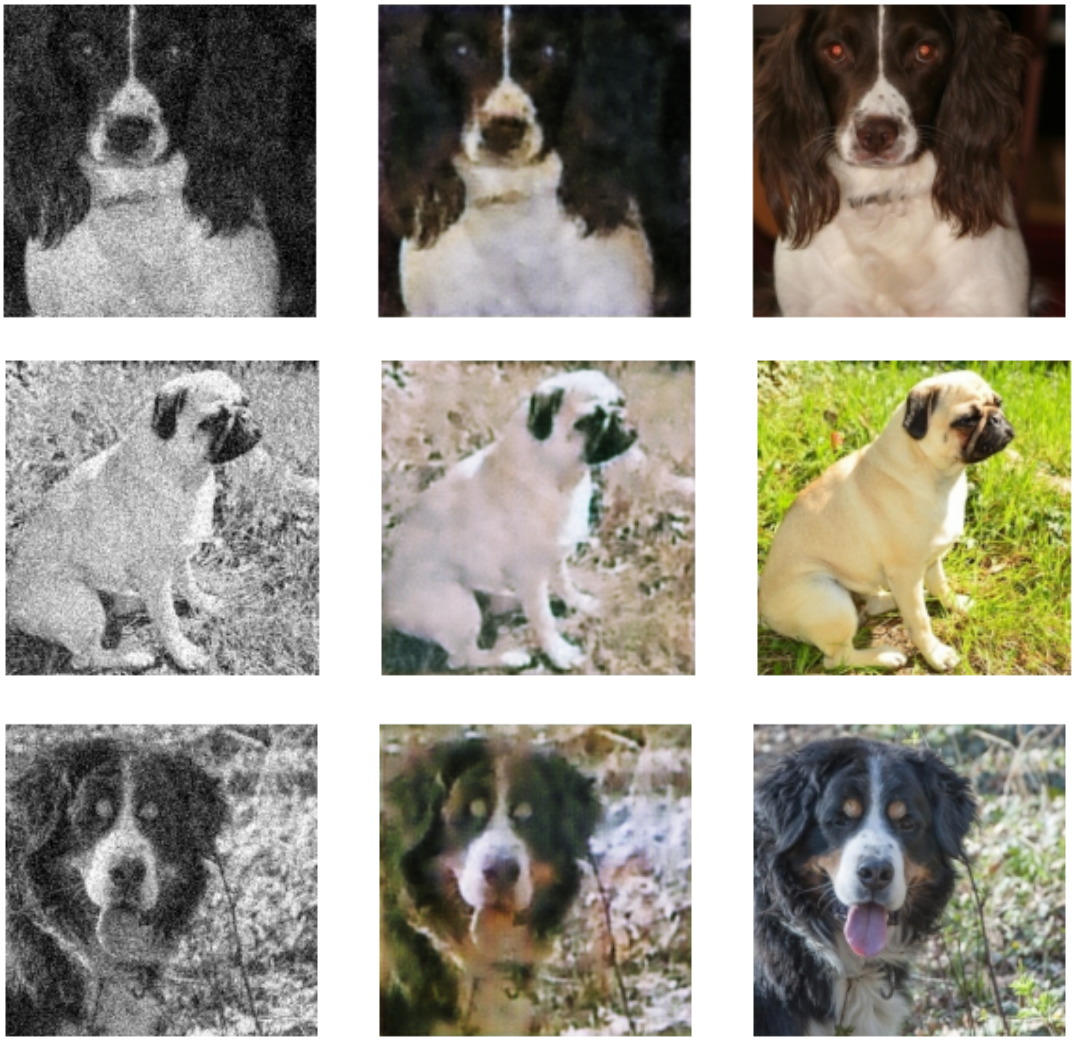


FIGURE 2.4: Exp 4: Noisy grayscale (left), generated RGB image (centre), ground truth RGB image(right)

CNN model was referenced from a paper by [**?**]. The subsequent process is called despeckling of images, where the noise is referred to as speckle.

The model has 8 stacked convolutional layers with the last layer being a component wise division of the noisy input and the generated speckle layer, followed by a tanh activation 2.5. It is trained over Euclidian loss between the generated image (despeckled) and the ground truth (grayscale).

An important point to note here is that noise is *additive* but the removal of it is *multiplicative* (treated as having been added multiplicatively, and thus divided).

The CNN model successfully reduces the noise and distortion to a great degree but also reduces the sharpness of the image 2.6. With further training and a larger dataset, the model could perform really well. The sharpness of generated images could be regained by other techniques like super-resolution.

2.6 Experiment 6

The same CNN model was trained to despeckle SAR images, with grayscaled RGB images as ground truth. Although the distortion in SAR images is not speckling but

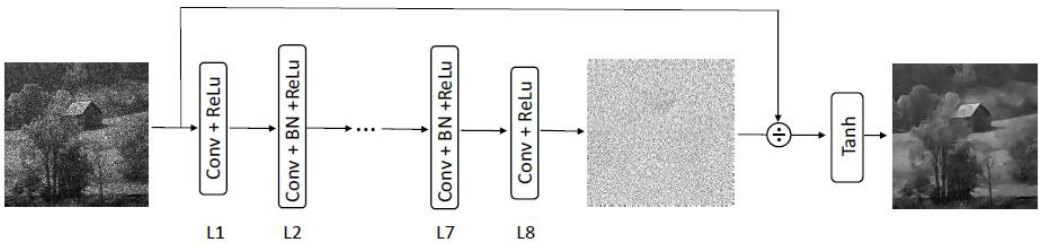


FIGURE 2.5: Exp 5: CNN for despeckling of images
Source: [?]

it appears similar. The results however, were undesirable.

The CNN model, when applied to SAR images leads to a severe loss of information. The resultant images are extremely undesirable. This leads us to believe that perhaps the model should first be trained on grayscaled Sen-2 RGB images with additive Gaussian noise, or that the distortion and noise in SAR images cannot be treated the same as Gaussian noise. Further profiling and analysis is needed.

2.7 Experiment 7

Next, we trained our model on extremely restricted image domain of satellite images. We selected satellite images that were captured over a single swath (single run of image capturing by a satellite over a particular region of interest). This set had $\approx 1,200$ images. Number of epochs were 5. The model seemed to be progressing towards what would eventually be an optical representation of radar images 2.8. This confirmed that our model architecture was valid.

2.8 Experiment 8

Finally, we selected 12,000 images, evenly distributed over 4 different types of major terrain/regions. Barren land, further divided into 2 classes, pure barren land and mixture of barren and urban land; grasslands, further divided into 2 classes, pure grasslands and a mixture of grasslands and urban areas; grid, which is agricultural land; and urban land, which are images majorly showing rooftops, roadways, narrow paths, boundaries etc. This gave us $\approx 2,000$ images per class.

As can be seen from the images 2.9, the model learns and generates images with a smaller number of features (barren land, grassland) easily, but highly complex images i.e., images with a greater number of features or information per unit area, such as urban areas and agricultural land, have not been learnt accurately 2.10. This guides us towards further restricting the image domain.

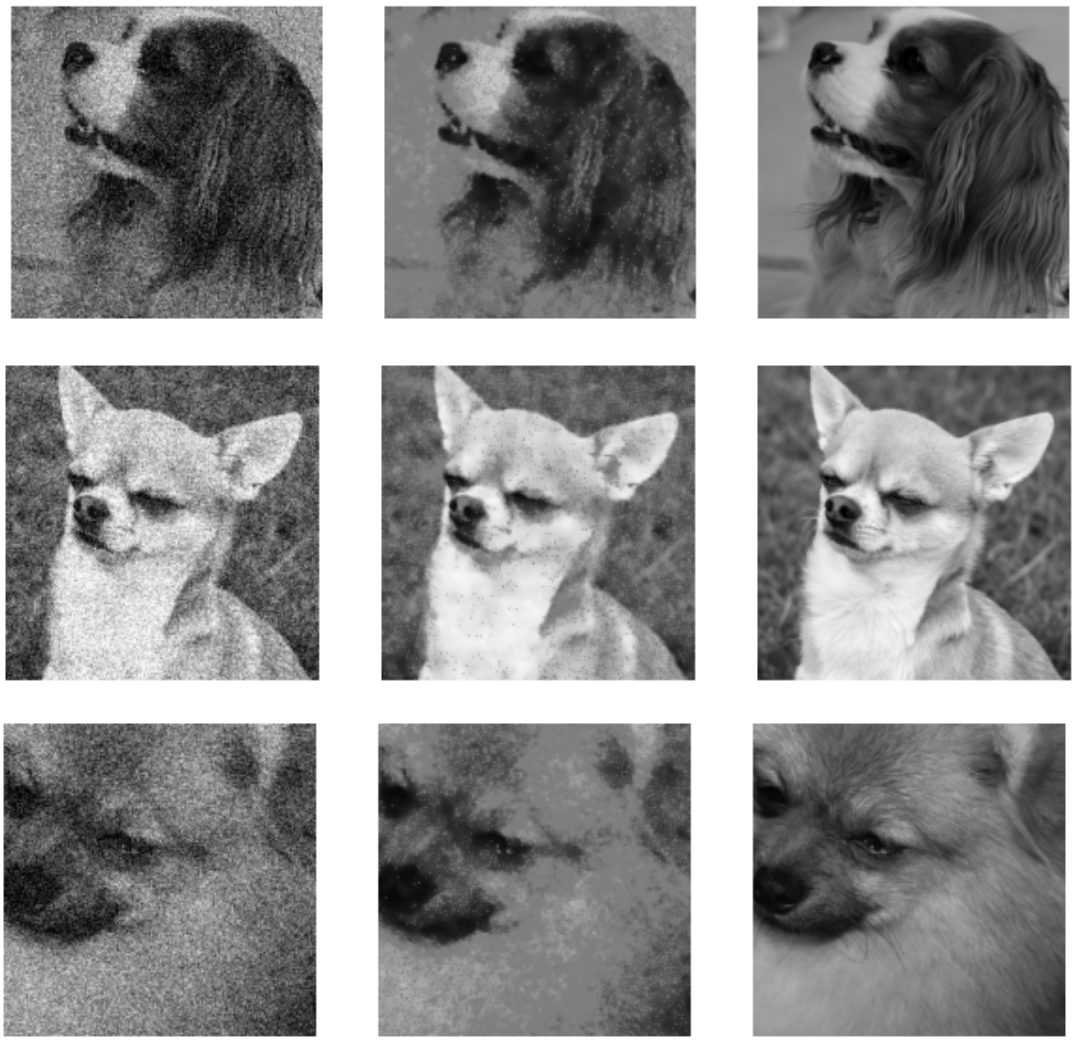


FIGURE 2.6: Exp 5: Noisy grayscale (left), generated noise-free grayscale image (centre), ground truth grayscale image (right)

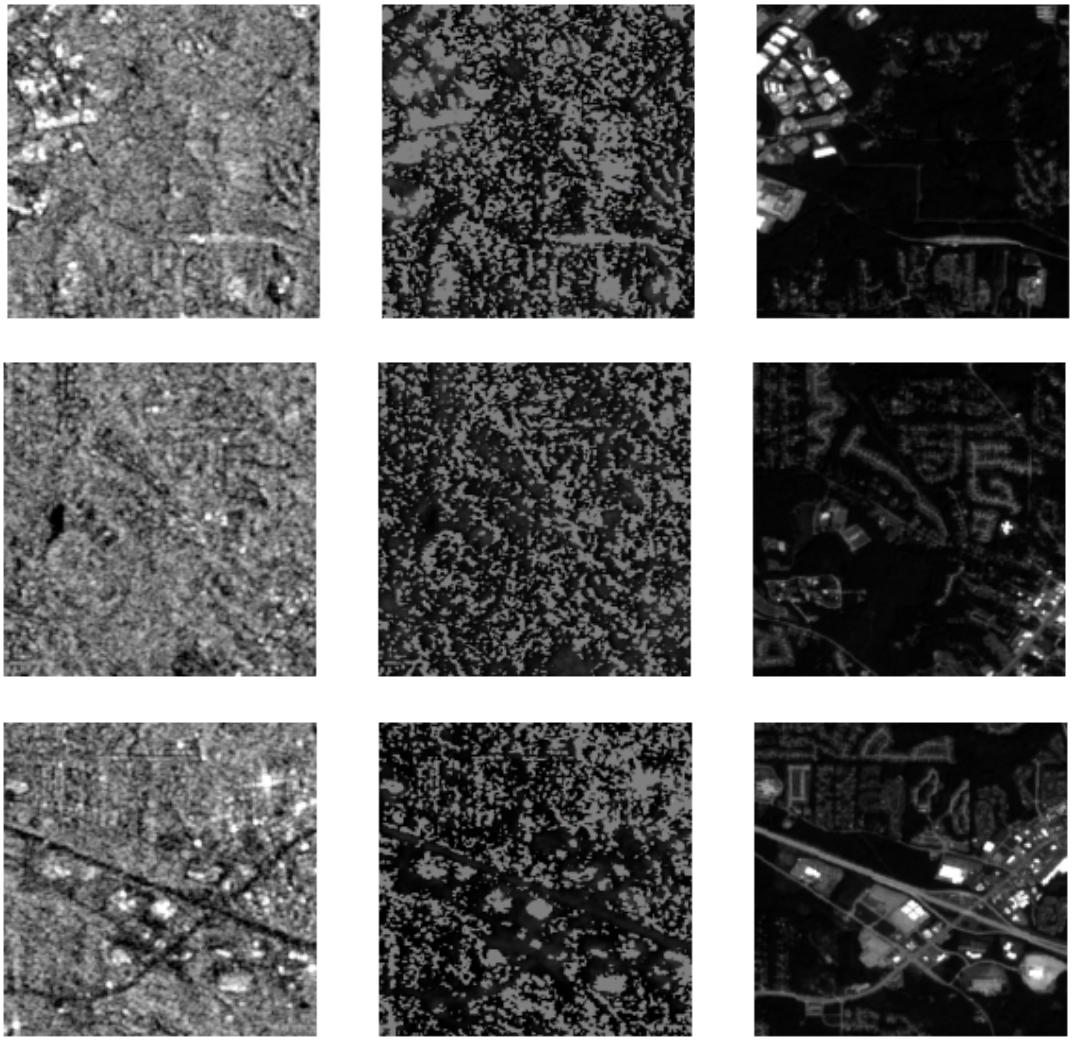


FIGURE 2.7: Exp 6: SAR (left), generated despeckled grayscale SAR (centre), ground truth grayscale optical image (right)

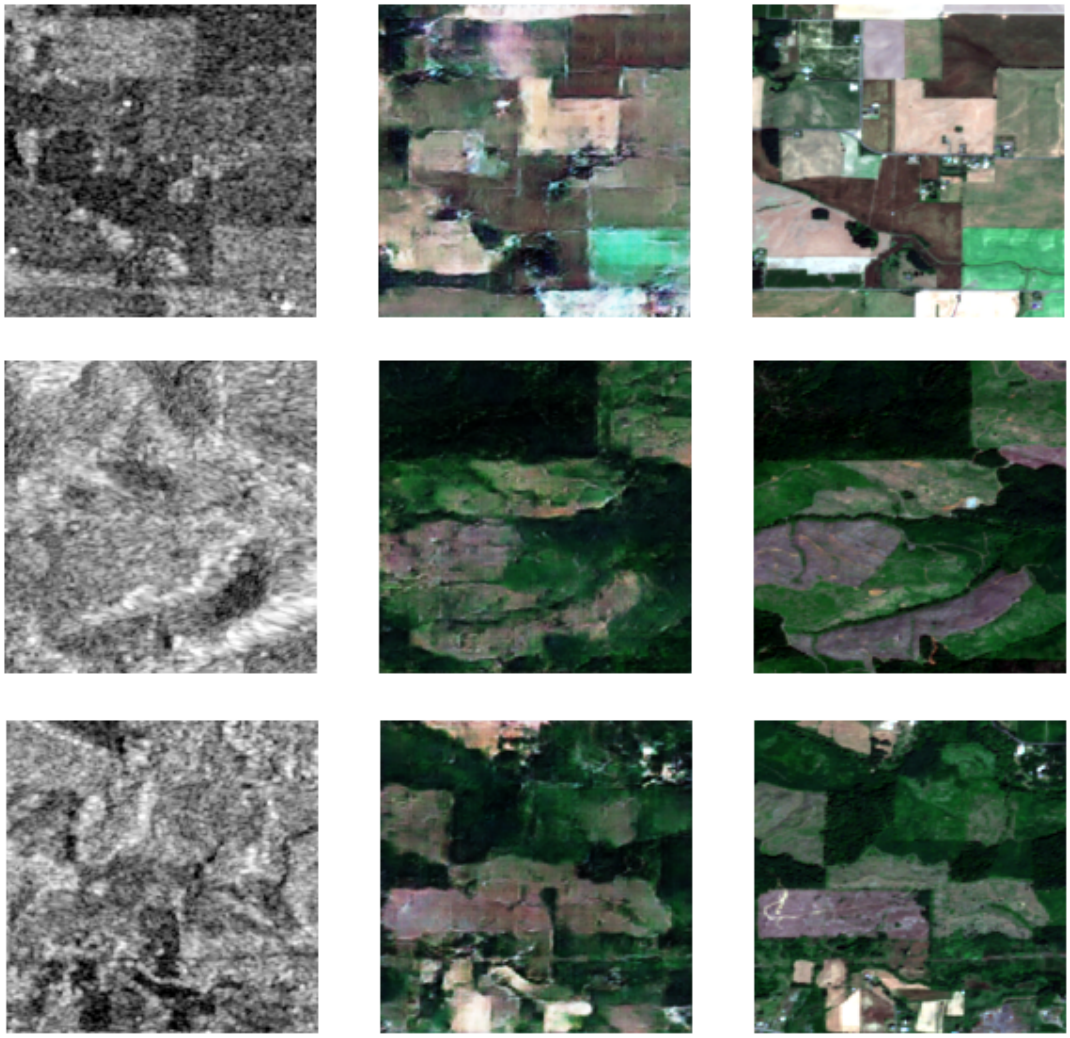


FIGURE 2.8: Exp 7: SAR (left), generated image (centre), ground truth (right)

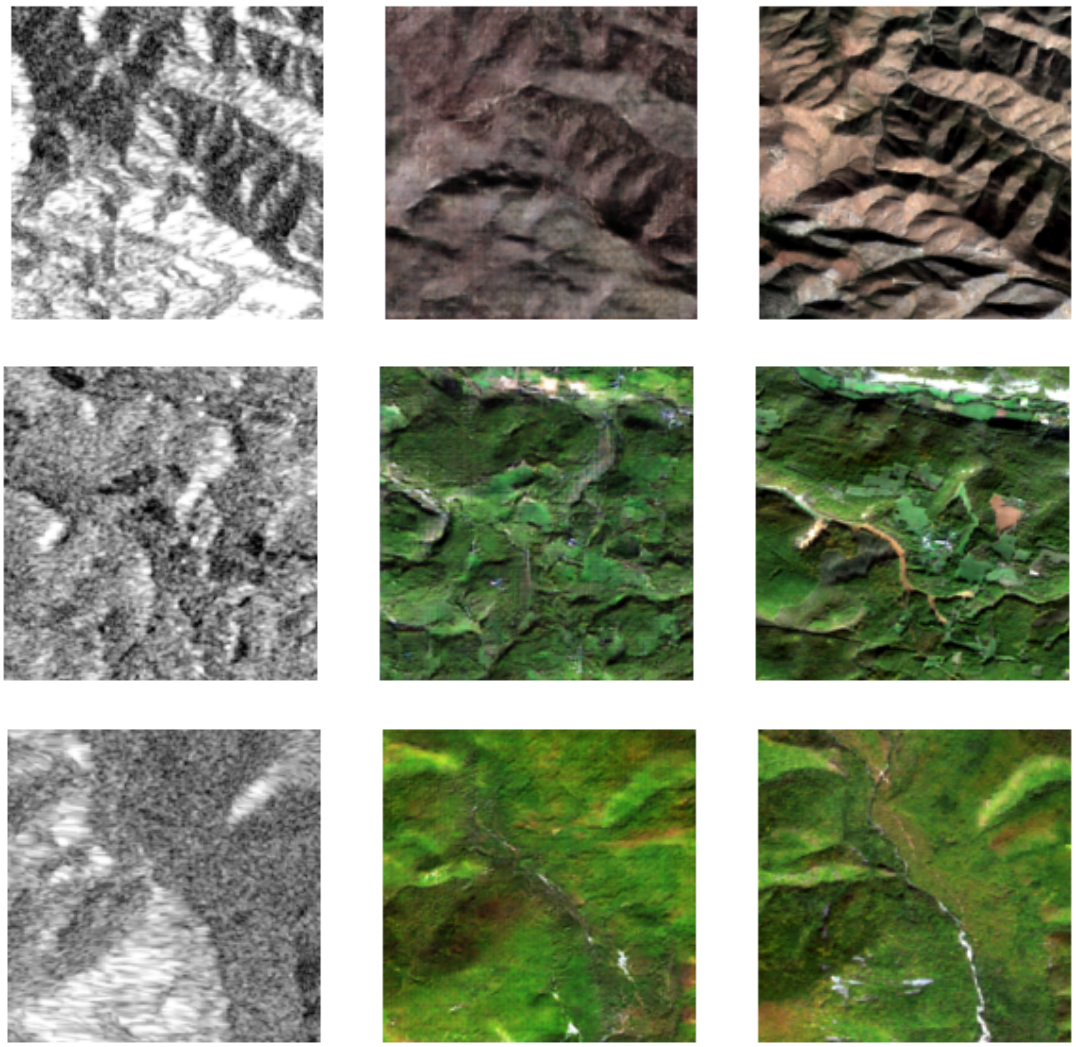


FIGURE 2.9: Exp 8-1: SAR (left), generated image (centre), ground truth (right)

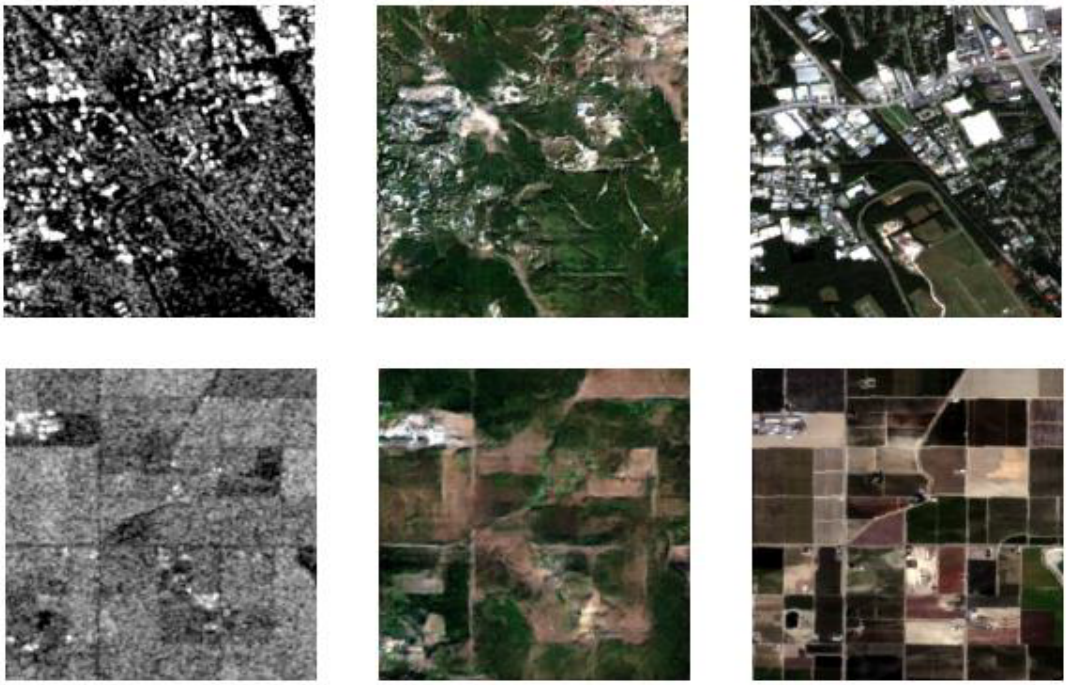


FIGURE 2.10: Exp 8-2: SAR (left), generated image (centre), ground truth (right)

Chapter 3

Proposed Architecture

3.1 SAR Classifier

The most accurate and sharpest results were produced in experiment 7 2.8. However, restricting the image domain to such a small number does not make our model quite practical. Thus we propose to classify SAR images into a certain class of terrain before being fed into a generator model trained specifically on that type of terrain.

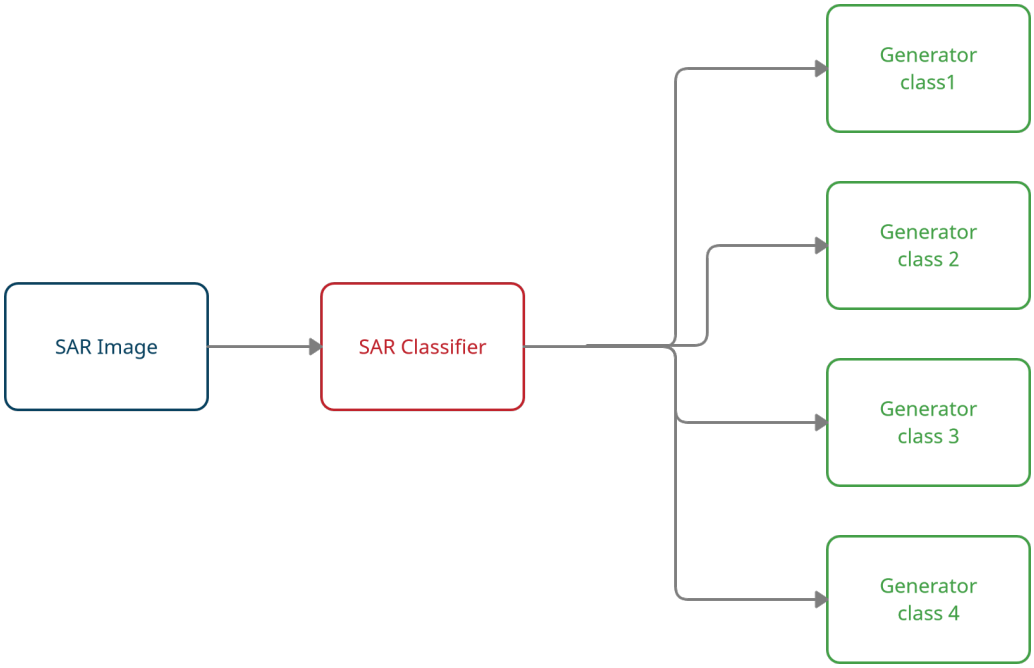


FIGURE 3.1: SAR Classifier

Image dataset was the same as the last experiment (8), only this time the images were separated and annotated into 4 major classes of barren land, grassland, grid and urban, with each class containing $\approx 2,000$ images.

3.2 Classifier Experiments

We first constructed a CNN accepting a single channel of an SAR images (since the other two channels had the same data). Starting at 32 cells in the first layer and then doubling the number at each layer until the image was downsampled from 256×256 to 4×4. All the layers had a 4×4 kernel and 2 units of stride with zero padding,

and a batch normalization layer and rectified linear unit layer was added at the end. Adam optimizer was used. Batch size was 1.

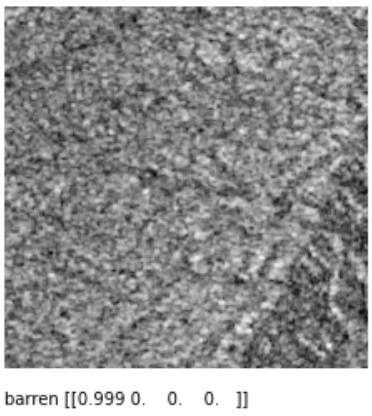
The model classification remained stagnant at 0.25 for each class after 10 epochs. For some reason, our classifier was not learning to differentiate among the types of land.

Batch size was changed to 3. Our model started giving a higher score to the barren land class.

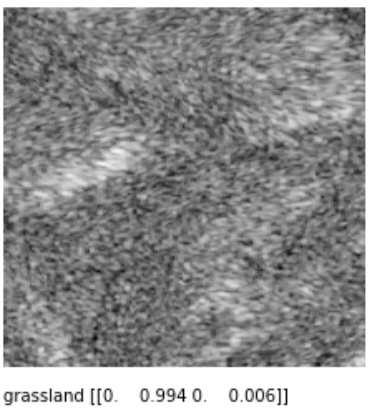
Several other configurations were checked. We increased the network depth, downsampled the image to 2×2, increased/decreased the number of epochs etc. Following observations were made. A model with greater depth and trained for longer was eventually memorizing the entire dataset, as the classification accuracy jumped to $\approx 100\%$ during training after a certain number of epochs.

As expected, the testing accuracy was zero. The testing We needed to find the right balance between downsampling the image and restricting the model memory.

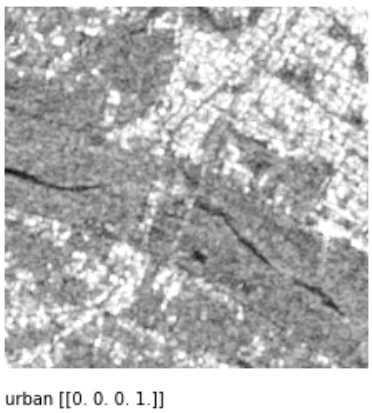
(A) Agricultural land class



(B) Barren land class



(C) Grassland class

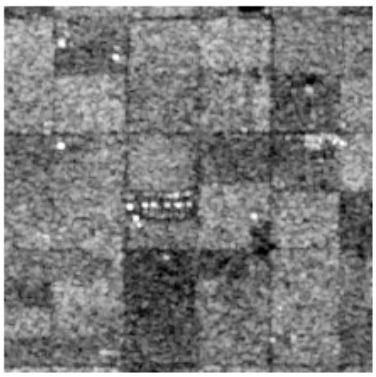


(D) Urban land class

FIGURE 3.2: SAR Classifier over-fitting on data
(true class name) [barren land score, grassland score, grid score, urban score]

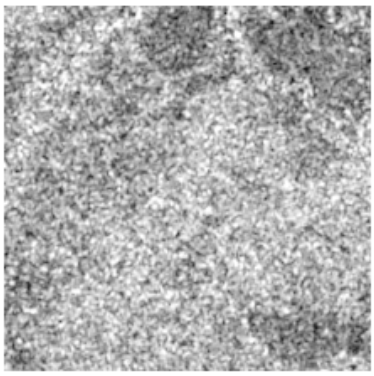
Final configuration changes include reduced depth, 4 units of strides and reducing training steps to 15 epochs with batch size of 3. This gives us a well regularized

and generalized model with an accuracy of $\approx$ 85%. The accuracy can be improved further.



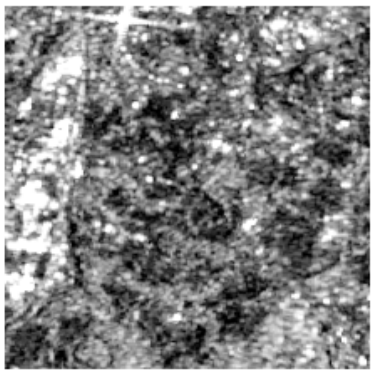
agri_grid [[0.069 0.111 0.771 0.05]]

(A) Agricultural land class



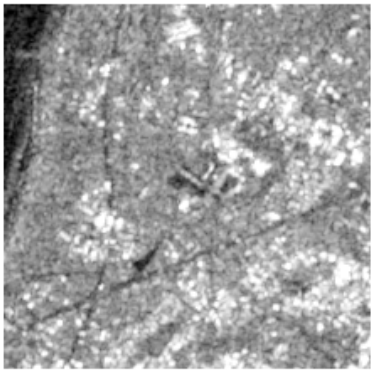
barren [[0.645 0.002 0.007 0.346]]

(B) Barren land class



grassland [[0.163 0.739 0.089 0.008]]

(C) Grassland class



urban [[0.157 0.04 0.176 0.628]]

(D) Urban land class

FIGURE 3.3: SAR Classifier final model results
(true class name) [barren land score, grassland score, grid score, urban score]

The results of this model architecture will be close to experiment 7 2.8 of our image to image translation model.

Chapter 4

Final Analysis, Conclusion and Future Work

Even with the classifier, we could still face difficulty in image mapping due to the amount of colour variation in images of different geographical features. This would require an increased number of generators which could mitigate the advantage we had over the brute force method.

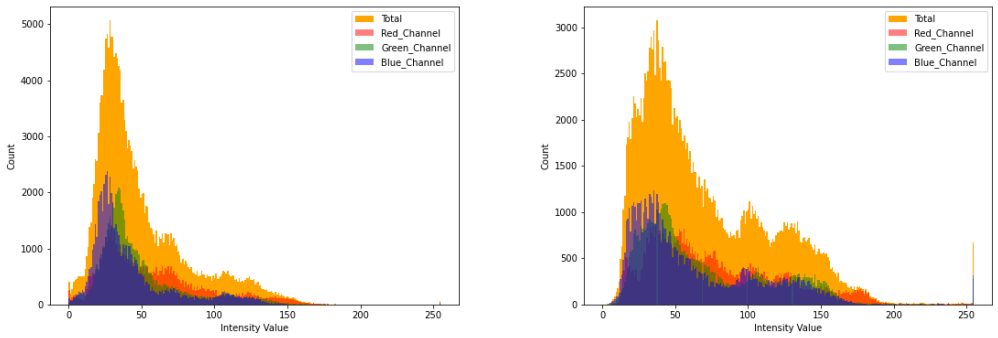


FIGURE 4.1: Agricultural land - colour histograms

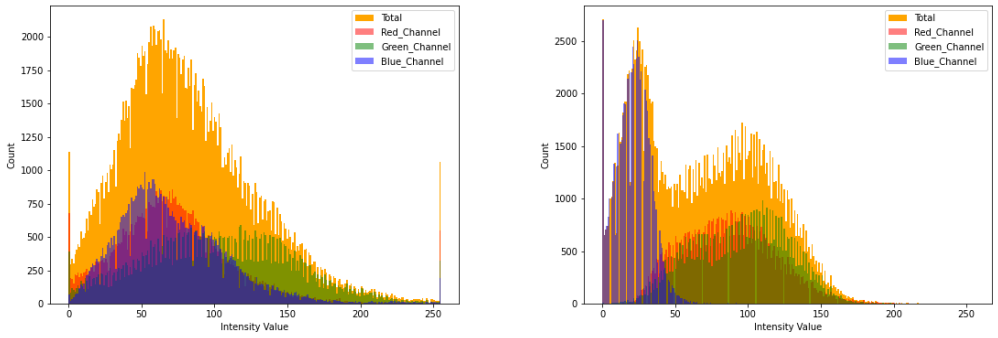


FIGURE 4.2: Grassland - colour histograms

The colour profile histograms 4.1, 4.2, 4.3, 4.4 show the amount of variation in barren land optical images, as compared to images of other classes.

Our comparatively lighter network will benefit if it is fed images from a certain domain/class only. This will reduce the number of colour profiles it has to learn (and fine features). The argument here is that, if it works for a shallow network like

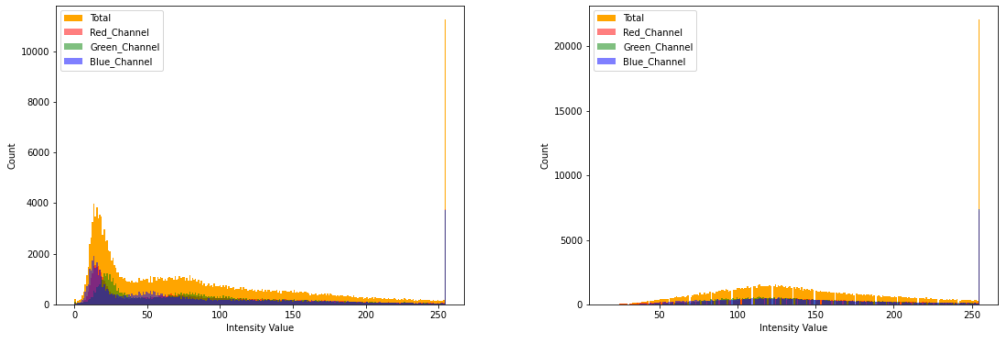


FIGURE 4.3: Urban land - colour histograms

ours, it will work very well for a deeper network.

The model might not produce exact sharp copies of optical images, but they will be accurate enough for *land classification*. If our goal is to monitor agricultural land, we will be better off training our model for a domain of land which primarily covers vegetation. Any other features contaminating the images will not matter as the agricultural land will be reproduced very well.

While this project shows promise, deploying it at production level in a real-world scenario does not seem viable. There is no way to guarantee the correctness of the generated image during $> 90\%$ cloud cover. We could solve this problem by comparing our generated image with optical images captured at an earlier point in time and stored in archives. Implementing this solution requires additional research.

All the satellites deployed into the earth's orbit have a lifespan and the life of the Sentinel program is nearly over. Even though the images might be useful for years to come, the data will not be useful for production level projects.

The major issue that we have with the generated images arise to the amount of noise and distortion in Sentinel-1 SAR images. In recent years however, other satellites have captured radar images with a greater amount of clarity 4.5, with some being practically grayscale versions of their optical counterparts. This makes our dataset obsolete in comparison.

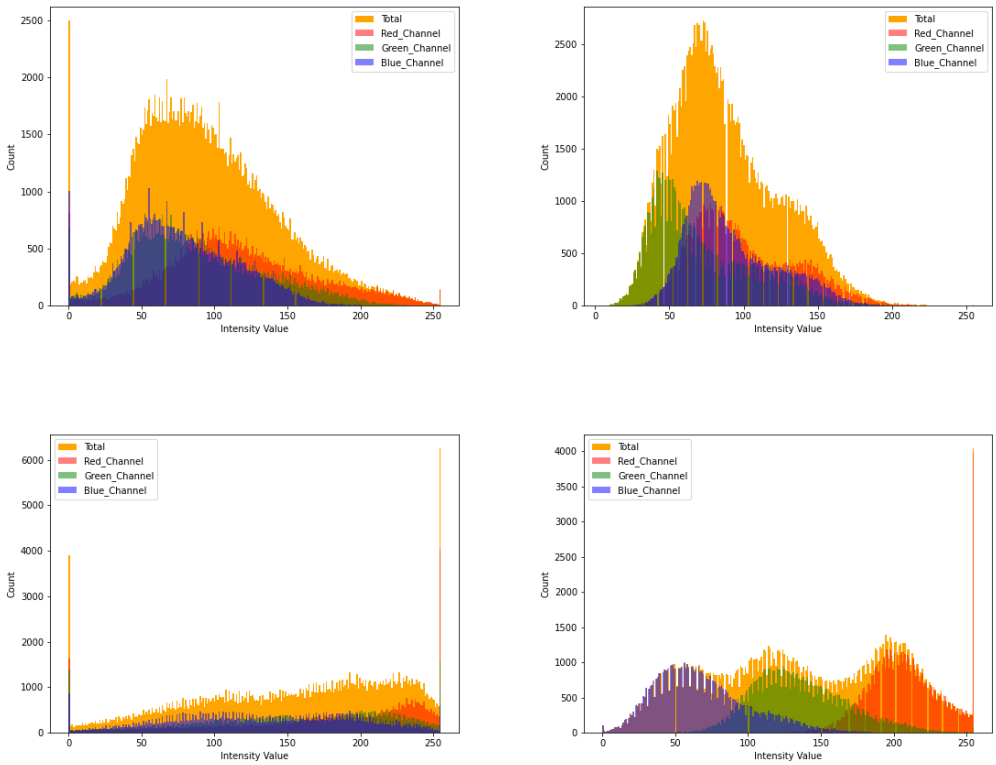


FIGURE 4.4: Barren land - colour histograms

FIGURE 4.5: High resolution SAR images
Source: Capella Space