Proceedings of 13th IOE Graduate Conference
Peer Reviewed
ISSN: 2350-8914 (Online), 2350-8906 (Print)
Year: 2023 Month: April Volume: 13

Satellite Image to Map Transformation using Modified Generative Adversarial Network

Biplove Pokhrel [a], Badri Raj Lamichhane [b], Biplav Sharma Regmi [c]

[a, b] *Department of Electronics and Computer Engineering, Paschimanchal Campus, IOE, Tribhuvan University, Nepal*
[c] *Department of Information and Communications Technologies (ICT), Asian Institute of Technology, Thailand*
[a] biplovepokhrel101@gmail.com, [b] badri@wrc.edu.np, [c] biplav01regmi@gmail.com

Abstract
Satellite image processing, which is a crucial field of study and development, encompasses the analysis of photographs of the world taken by artificial satellites. Map Synthesis is a key area of satellite image processing. Deep learning techniques such as Generative Adversarial Networks (GANs) are increasingly often employed in satellite imaging research. They can be used for everything from navigation to urban planning to disaster response to medical imaging to super resolution. The produced maps can be used in addition to or in substitute of conventional map-making techniques because they are extremely accurate and detailed. In locations with incomplete or old data, GANs can also be helpful in creating maps. Moreover, data augmentation, machine learning, simulation, and visualization can be done using the synthesized maps. A deep learning model known as a "generative adversarial network" (GAN) includes major two parts: a generator and a discriminator. The discriminator assesses the veracity of the generated examples while the generator generates fresh samples of data. The suggested model is built using the Pix2Pix Conditional Generative Adversarial Network (cGAN) architecture to produce maps from satellite images. A Unet Network serves as the network's generator, and its output is passed into the patch discriminator. Blocks with fixed and variable kernel sizes are added to the U-Net architecture to broaden the receptive field. The discriminator attempts to determine whether each patch in a picture is real or fake using paired samples of datasets containing satellite images and their matching map images. The final output is produced by convolutionally running this discriminator across the image and averaging all responses.

Keywords
Dilation rate, Generative Adversarial Networks, Kernel, Patch Discriminators Receptive Field Block, U-net

1. Introduction

Satellite mapping is the process of using satellite imagery to create maps of the earth's surface. This technology has been in use since the 1960s and has revolutionized the way we understand and use information about the earth. The satellite is equipped with cameras and other sensors that are used to collect imagery of the earth's surface. This imagery is then transmitted back to earth and processed to create maps for different applications.The task of "image-to-image translation" is a specific type of creative problem where the aim is to convert an initial image into a manufactured one, or to associate an original image with a particular target image through automated means.[1]Deep neural networks termed Generative Adversarial Networks (GANs) are used to produce realistic data by combining two different neural networks. The GAN is trained on a dataset, with the generator learning to produce maps that are similar to the real maps in the dataset. The discriminator is learned to differentiate between fake maps and actual maps. The generator and discriminator are two networks that train one another through repeated cycles of generation and discrimination while also seeking to outsmart one another.

2. Problem Statement

Synthesizing high-resolution images using a Generative Adversarial Network (GAN) can be a challenging task, particularly when working with satellite images. One of the main difficulties in obtaining a strong pixel-level mask, which is necessary to ensure that the generated images are realistic and preserve the underlying features of the area being mapped. Capturing finer details from the satellite image and then transfer these details to neural network to produce map image is another challenging task. Traditional Convolution network suffers from Over fitting, Lack of translation invariance which has also limited to small field of view.

3. Related Works

Conditional adversarial networks are advised as a good solution for general image image translation issues[2] .GAN is explored in the conditional setting which learns a generative model suitable for image to image translation task where condition on an input image is created for generation of corresponding output image. This work differed from other research in architecture that used "U-Net" base generators and "Patch discriminator" classifiers based on convolutional neural networks that only penalized feature at a range of patched images. This research paved the idea for the use of skip connection which bypass one or more layers and connects input to output.

The CycleGAN framework introduced by [3] Zhu et al.is one of the most consistently used unpaired approaches for image-to-image translation. This model included two mappings function between different domain. Two adversarial discriminators were introduced to get adversarial loss and cycle consistency losses to prevent the learned mappings contradicting from each other. Zhang et al.[4]developed an updated GAN

model that uses external geographic data as implicit guidance to produce better quality map images. To enable smooth satellite to picture conversion, the text-based geographic data is transformed into an image.

A series of convolutions is used to translate the conditional generator described by Ganguli et al. [5] to the standard layer of a map, and the real/generated map and the satellite image are concatenated as the discriminator input. In addition to the GAN losses, reconstruction loss and style transfer loss were also introduced. Inagale et al. [5]series of encoder and decoder blocks using U-net architecture. This model take satellite image as a input down sample it to its bottleneck representation and then up sample from the bottle neck(vector space representation) into the size of output image which is same as input size. Discriminator model is deep convolution neural network model which simply perform the image classification. The training time for the results was very short. Authors implied on improving of results by increasing the training time for the model.

The multi-scale Receptive Fields Blocks (RFB)[6] in the generator network are used by OPPO-proposed Research's RFB-ESRGAN, which is based on ESRGAN and restores finer features and texture. From input LR photos, RFB can extract both coarse and fine features. RFB reduces the model complexity and processing time by employing multiple small kernels rather than a single large kernel. With the use of adversarial learning, Song et al. introduced the MAPGen-GAN[7] unsupervised domain mapping model, which can swiftly translate remote sensing images into general maps. Circulatory consistency and geometrical consistency constraints were added to the loss function of the suggested model to increase the fidelity and geometry accuracy of the output maps.

Majority of the Map Synthesis work in Conditional GAN have used U-net generator and patch gan discriminator. Zhou et al [8] has stated in his paper regarding the major limitations of Unet regarding finding the Optimal depth of the network and aggregate limitations of the skip connections in both scale in encoder and decoder. Inagale et al[5] found out resnet-9 was generating image with better quality then the Unet generator however the margin was not that much higher. Shang et al[9] used Receptive field blocks [6] in the image resolution which enhanced the lower resolution images to higher resolution image with greater accuracy. This research focus on the generator network with Receptive field blocks that enchances the learning of very vital detail of the image when passing through the networks.Different U-Net Architecture are tested using RFB blocks in generator.The discriminator used is patch discriminator which penalizes in every patches of the image taken as a input.

4. Methodology

The GAN, which consists of a generator network and a discriminator network, serves as the basic foundation for this study. The research's approach is shown in Figure 1.

4.1 Implemented Model

The generator produces a counterfeit image from a genuine photo, and both the real and fake images are given to the discriminator. In this case, unlike previous GANs, both the real image and its corresponding map image are provided for comparison. The discriminator aims to determine which image is authentic by merging both images. During the training of the discriminator, both discriminator loss and generator loss are experienced. The discriminator loss is calculated using the L1 loss, which is the Mean Absolute Error between the actual target image, which is the corresponding map of the original satellite image, and the fake image created by the generator. The discriminator is then updated with this loss.

In Pix2Pix GAN, the generator's loss function often combines the adversarial loss and the L1 or L2 loss. The adversarial loss trains the generator to create images that can deceive the discriminator, similar to a standard GAN.

$$L_{cGAN}(G,D) = E_{x,y}[logD(x,y)] + E_x[log(1-D(x,G(x))] \quad (1)$$

Our objective for the min max game:

$$(G^*,D^* = arg\frac{min}{G}\frac{max}{D}(L_cGAN(G,D) + \lambda L_{L1}(G)) \quad (2)$$

In summary for generator loss and discriminator loss we get,

$$L_{gen}(G,D) = BCE[D(x,G(g),1] + \lambda L_{l1}(G) \quad (3)$$

$$L_{dis}(G,D) = BCE[D(x,G(x)),0] + BCE(D(x,y),1) \quad (4)$$

Equation 1, represents a Generative Adversarial Network (GAN) where G is the generator and D is the discriminator. The equation measures the success of G and D in a game-like setting where G tries to create realistic outputs and D tries to distinguish between real and fake inputs. Equation 3, is the generator's loss function, which consists of two terms: binary cross-entropy (BCE) loss between D's output on the generated sample (G(g)) and the label "1", indicating that the generator wants to fool the discriminator into thinking that the generated sample is real, and a regularization term L1 loss that encourages diversity in the generated samples. lamda is a hyperparameter that controls the trade-off between the two terms. Equation 4, is the discriminator's loss function, which also consists of two terms: BCE loss between D's output on the generated sample and the label "0", indicating that the discriminator correctly identifies the sample as fake, and BCE loss between D's output on the real sample (x) and its label "1", indicating that the discriminator correctly identifies the sample as real.

Generator The U-Net generator is a type of generator used in the Pix2Pix GAN model, which is used for converting one image to another. Originally designed for medical image segmentation, the U-Net architecture has been adapted for various other computer vision applications. The U-Net design consists of two networks: an encoder network and a decoder network, with skip links between matching layers in both networks. The encoder network, which consists of a number of convolutional layers with batch normalization and ReLU activation functions, uses a max-pooling layer for downsampling. The decoder network performs upsampling on a number of layers with batch normalization and ReLU activation functions using a transposed convolutional layer. The skip connections between the encoder and decoder layers allow the model to maintain picture information while delivering output visuals that are vivid and lifelike. In this research U-Net generator is used along with the receptive field block with variation dilation rate.The block takes an input applies a series of convolutional layers with different kernel sizes and dilation

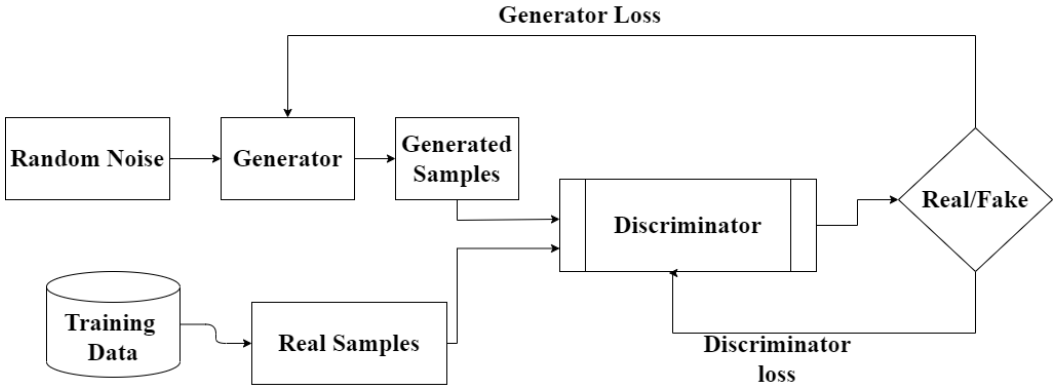

Figure 1: GAN Model

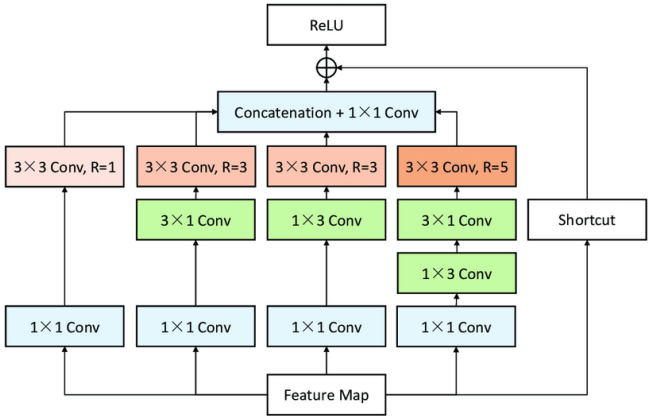

Figure 2: Receptive Field Block

rates to capture features at different scales. The output of each branch is concatenated, and then multiplied by a scaling factor of 0.2 before being added to a shortcut connection that bypasses the block. Finally, the output is passed through another convolutional layer before being returned.

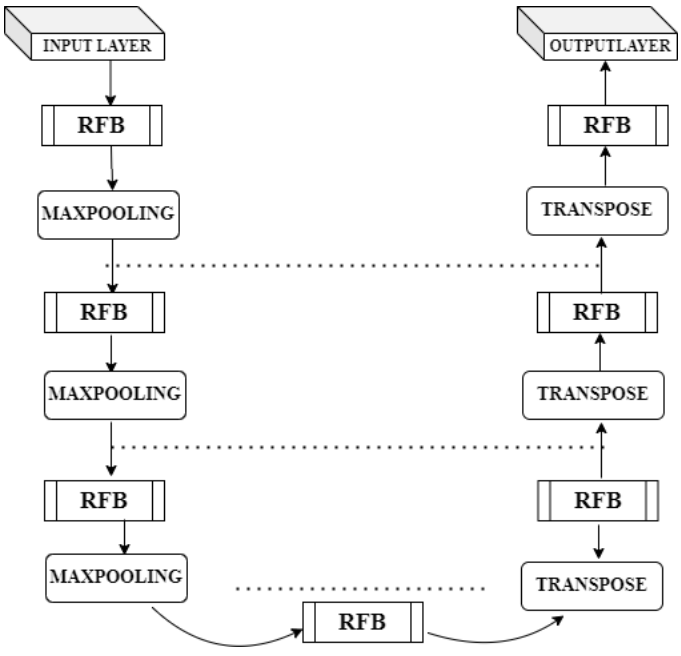

Figure 3: Modified Generator Architecture

The block has four branches, each consisting of several convolutional layers with different configurations. The first branch applies a 1x1 convolution followed by a 3x3 convolution with a dilation rate of 1. The second and third branches apply a 1x1 convolution followed by two separate 3x3 convolutions with dilation rates of 1 and 3, respectively, along the two dimensions. The fourth branch applies a 1x1 convolution followed by three separate convolutions with kernel sizes of (1, 3), (1, 3), and (3,1) respectively, with increasing dilation rates of 1, 3, and 5.

Discriminator Convolutional Neural Networks (CNNs) have been extensively tested and shown to be effective for tasks related to image classification and generation. Consequently, we employ CNNs as the basis for both the generator and discriminator networks. The role of the discriminator network is to utilize convolutional layers to decrease the dimensions of input images, ultimately producing a binary output that categorizes input images as either genuine or fake (generated by the generator network). In this case, a PatchGAN[2] is used. As stated in the pix2pix paper, the discriminator in the pix2pix cGAN is a convolutional PatchGAN classifier; it attempts to categorize whether each picture patch is real or not real. The discriminator consists of three blocks: convolution, batch normalizing, and leaky reLU. After the final layer, the output is shaped as follows: (batchsize, 30, 30, 1). A (70 x 70) area of the input image is classified for each 30 x 30 image patch of the output.

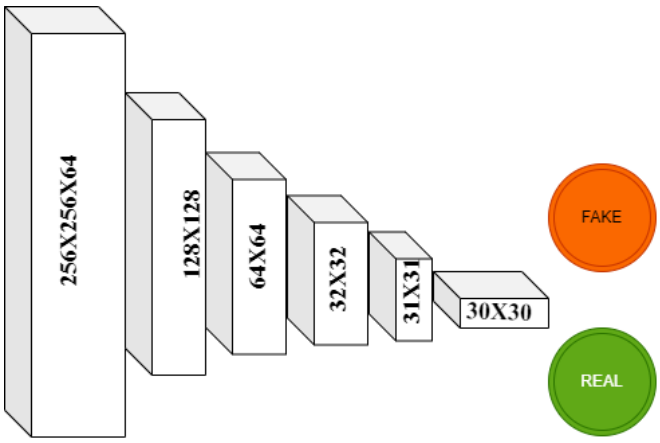

Figure 4: Patch Gan Discriminator

5. Experimental Results and Discussion

5.1 Dataset

This research utilized maps dataset that was also used in the initial pix2pix release. Almost 85 percent of earlier GAN research in Map Synthesis has used this dataset. Following the archive's decompression, two folders—"training" (1096 photos) and "validation"—were created (1098 images). Each image has a resolution of 1200 x 600 pixels and compares satellite and map modes side by side. As part of the pre-processing, the image is divided in half, creating two 600x600 images. Each image is then resized to 256 x 256 pixels to meet the conditional GAN model's dimensions.

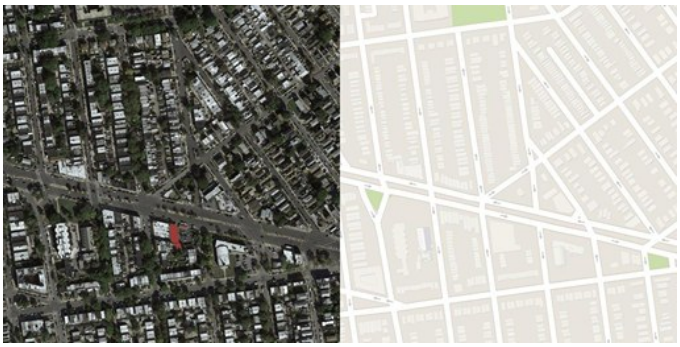

Figure 5: Paired Data Set

5.2 Model Training Details

We trained the model using Google Colab pro with GPU backend.Python version 3.8.10 was used for implementing the model. The optimizer we used for training is Adam. For each steps during training,we use binary cross entropy loss(BCE) for the adversarial loss and non saturated version of the discriminator loss. The discriminator trains faster than the modified generator because classfication is simple in task than synthesizing the image. In this study we slow down the learning of Discriminator by a factor of 2 as suggested by Isola et al. We use value for beta as 0.5 and 0.9 with learning rate alpha as 0.0005.Lambda value is taken as 100 as mentioned in Pix2Pix original paper. The dropout rate for each decoder block is 0.5 while no dropout is employed for encoder block. Batch size for the image is kept to 1 and the model is trained for 80 thousand steps. In Pix2Pix, the generator's objective is to create images that are so realistic that they are impossible to differentiate from real images. This is achieved through two components of the total generator loss: the adversarial loss and the L1 loss. The adversarial loss measures the generator's ability to deceive the discriminator, a separate neural network that learns to distinguish between real and generated images. This loss encourages the generator to create images that are very similar to the real ones, making it difficult for the discriminator to tell the difference. Meanwhile, the L1 loss measures how closely the pixel values of the generated images match those of the real images.

Discriminator loss at the very beginning is low since the noise of the image is easily classified fake. As the training increases the capacity of generator to synthesize more real images increases resulting in balancing the discriminator loss. The model was trained upto 80k iteration steps with batch size one. The loss at around 77k to 79k training was nominal.

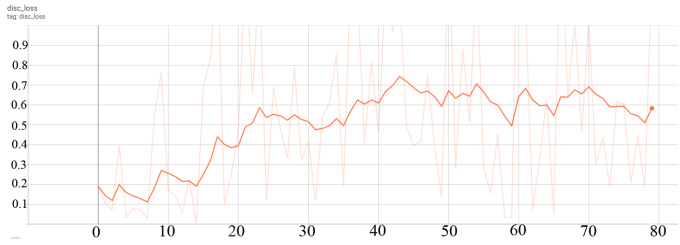

Figure 6: Discriminator Loss

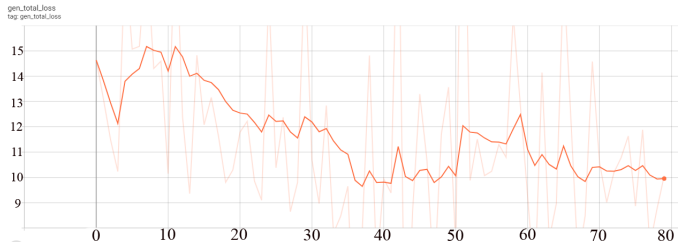

Figure 7: Total GAN Loss

The adversarial loss and the L1 loss are added together to create the overall generator loss. As mentioned above, the hyperparameter Lamda regulates the trade-off between the adversarial loss and the L1 loss. The total generator loss motivates the generator to create realistic and structurally similar images.

5.3 Output

All outputs obtained from the model are after training the model for 80 thousand iteration. Figure 9 shows the output of the GAN model with modified Unet architecture. Input image along with predicted and ground truth is shown in the figure below.The output shape of the image is also 256 x 256 x3. The 1st row is the input image, the image in the middle is Ground truth and last image in the row is predicted image by the model. With each increase in iteration steps the image from the model becomes clearer. The series of images are shown in the figure below. The similarity between predicted and ground truth image in modified generator is very high. Resnet on the other hand outperformed unet in generating image with more clearance. The image was tested after 80k iteration.

Evaluation Metric The SSIM metric assesses the structural likeness of two images by evaluating their brightness, contrast, and composition. It generates a score from -1 to 1, where a score of 1 means the images are indistinguishable, and a score of -1 indicates they are entirely different.The SSIM loss function is used to measure how similar the generated image is to the target image, and the generator is taught to reduce this loss. The Multiscale Structural Similarity Index (MS-SSIM) is a widely-used image quality assessment metric that evaluates the similarity between two images. It is intended to capture variations in picture structure at various scales and is an variant of the Structural Similarity (SSIM) index. MS-SSIM computes a set of structural similarity values at different scales and combines them to produce a final similarity score. This approach allows MS-SSIM to account for the perceptual

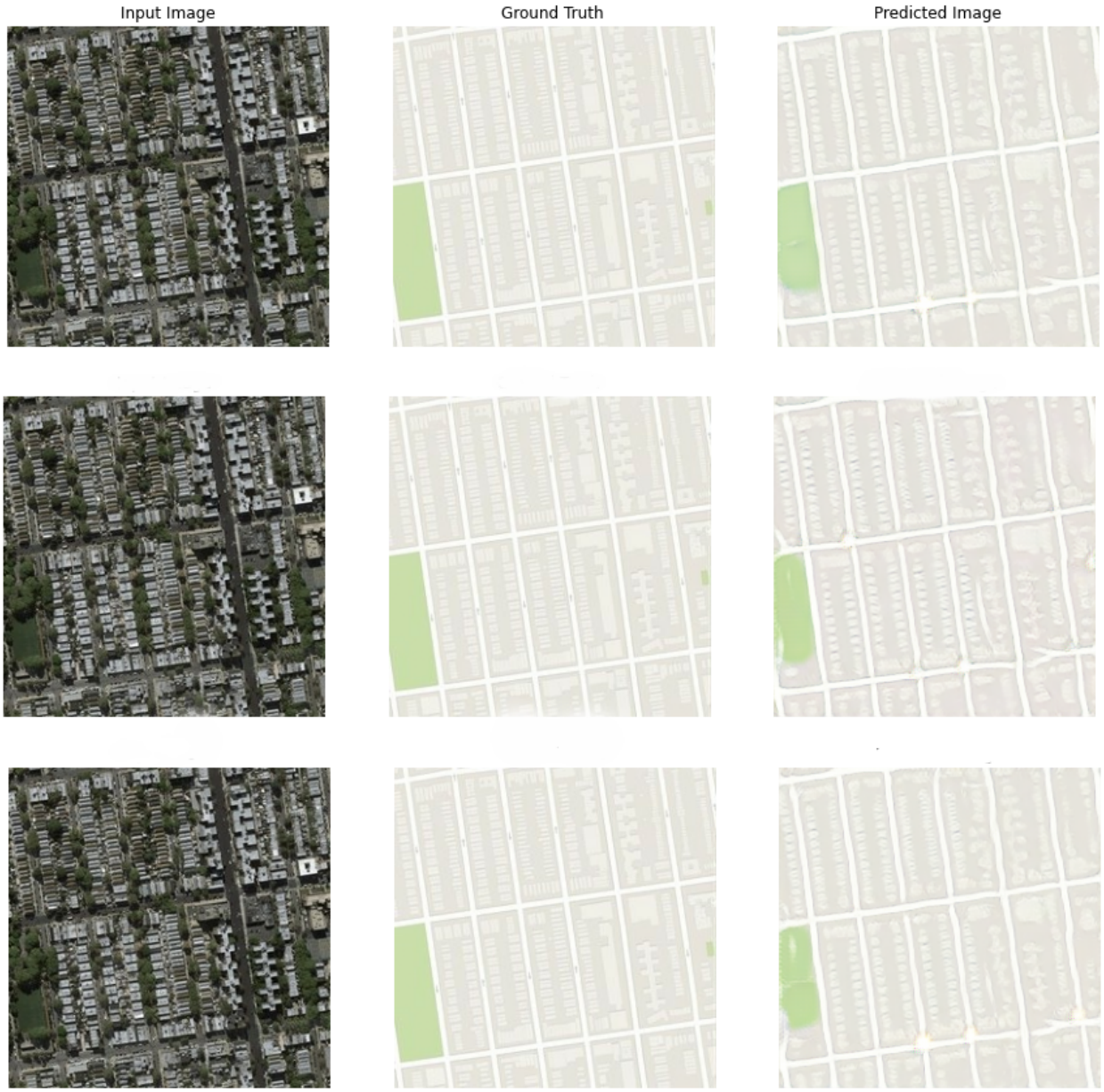

Figure 8: Output from modified archtiecture,Resnet 9 and Unet architecture from top .

significance of image structures that are visible only at specific scales. Table 1 shows the SSIM and MS-SSIM scores for the differenent models used.

Table 1: MS-SSIM and SSIM score for different model

Model	MS-SSIM	SSIM
Unet with RFB	0.82713	0.75
Resnet Base Model	0.76859	0.70
U-net Base Model	0.74758	0.68

For evaluation purposes, we generated 10 different images as test sample. In Table 1 we can see that the model with RFB blocks has higher MS-SSIM and SSIM than the Base model U-net and Resnet GAN model.MS-SSIM is considered better in image generating task.

Figure 9: Same test Output from modified archtiecture, Resnet 9 and Unet architecture in 30k training steps

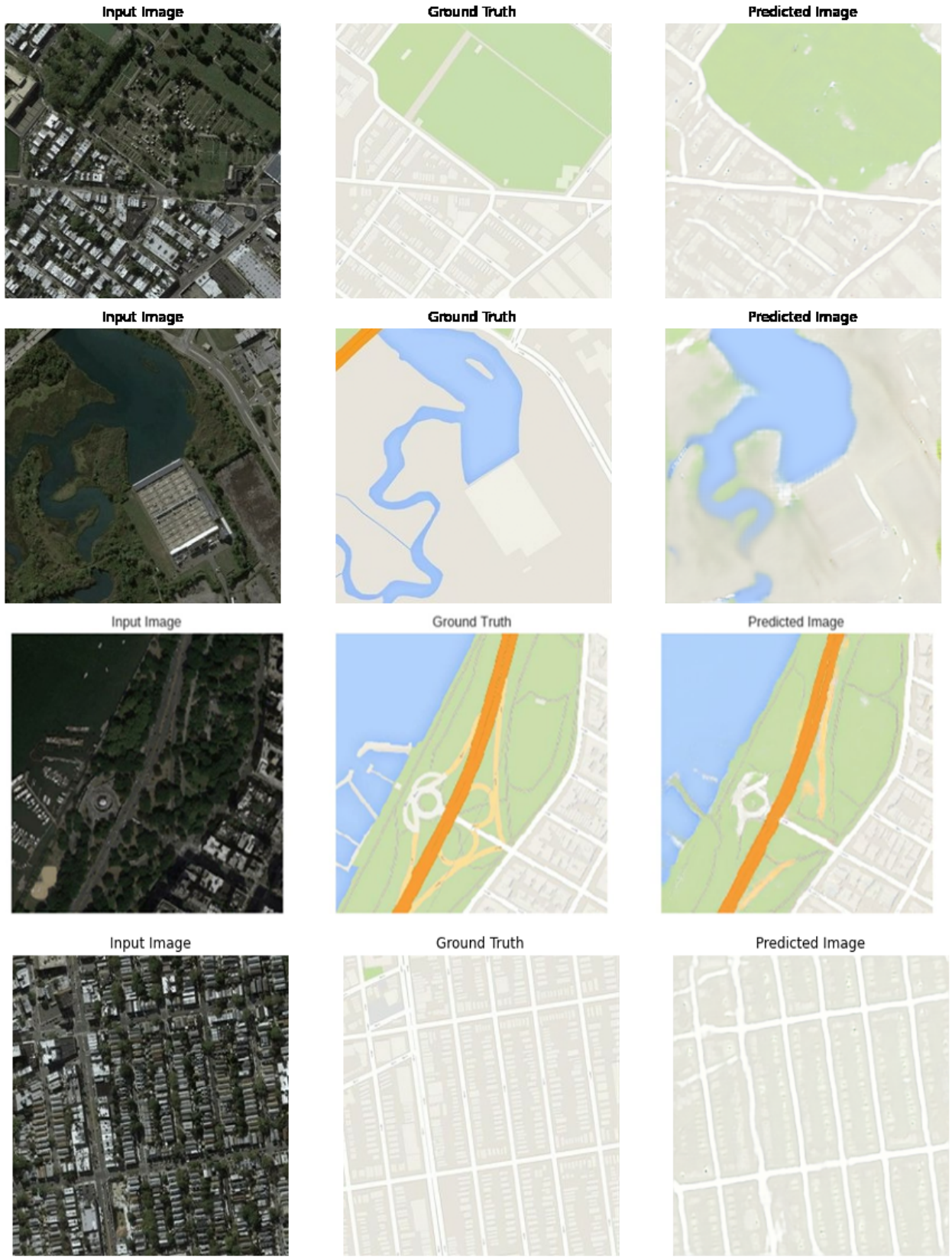

Figure 10: Test Output from modified archtiecture in different terrain of Fields, Shores, Water Bodies and Street

6. Conclusion

In this work, we tried to use Receptive Field Block for synthesizing images in base pix2pix GAN. Images generated had fine details than the original base model. The generator had less number of parameters than the base model U-Net and RESNET GAN. From the above output we can see that model using RFB can generate better images than U-Net and RESNET generator GAN.

References

[1] Yingxue Pang, Jianxin Lin, Tao Qin, and Zhibo Chen. Image-to-image translation: Methods and applications. *IEEE Transactions on Multimedia*, 24:3859–3881, 2021.

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

[3] Jun-Yan Zhu, Taesung Park, Phillip Isola, and Alexei A Efros. Unpaired image-to-image translation using cycle-consistent adversarial networks. In *Proceedings of the IEEE international conference on computer vision*, pages 2223–2232, 2017.

[4] Ying Zhang, Yifang Yin, Roger Zimmermann, Guanfeng Wang, Jagannadan Varadarajan, and See-Kiong Ng. An enhanced gan model for automatic satellite-to-map image conversion. *IEEE Access*, 8:176704–176716, 2020.

[5] Vaishali Ingale, Rishabh Singh, and Pragati Patwal. Image to image translation: Generating maps from satellite images. *arXiv preprint arXiv:2105.09253*, 2021.

[6] Kai Zhang, Shuhang Gu, and Radu Timofte. Ntire 2020 challenge on perceptual extreme super-resolution: Methods and results. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition workshops*, pages 492–493, 2020.

[7] Jieqiong Song, Jun Li, Hao Chen, and Jiangjiang Wu. Mapgen-gan: a fast translator for remote sensing image to map via unsupervised adversarial learning. *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, 14:2341–2357, 2021.

[8] Zhou Z Rahman Siddiquee MM Tajbakhsh. N liang j et al. stoyanov d et al. *UNet++: a nested U-net architecture for medical image segmentation Deep Learning in Medical Image Analysis and Multimodal Learning for Clinical Decision Support*, 2018.

[9] Taizhang Shang, Qiuju Dai, Shengchen Zhu, Tong Yang, and Yandong Guo. Perceptual extreme super-resolution network with receptive field block. In *computer vision and pattern recognition workshops*, pages 440–441, 2020.