Multi-Modal Image Synthesis with Attention Conditional GANs: SAR, Optical, and DEM

David Nhemaphuki[1], Ajay Kumar Thapa[2] & Umesh Bhurtyal[3]
davis10ge@gmail.com, ajay.thapa@ku.edu.np, umbhurtyal@wrc.edu.np
[1] Survey Department, [2] Kathmandu University, [3] Pashchimanchal Campus, Tribhuvan University

KEYWORDS

Image Synthesis, Cloud Removal, SAR, DEM, cGAN

ABSTRACT

This research addresses challenges in satellite-derived Earth observation data, such as cloud cover, atmospheric condition, seasonality and calibration inconsistencies, by leveraging Synthetic Aperture Radar for cloud penetration and Generative Adversarial Networks (GANs) for cloud removal and image synthesis. The main objective of this research is to develop conditional GAN (cGAN) models capable of synthesizing multispectral images from SAR (Sentinel-1), Optical (Sentinel-2) images and DEM (SRTM 30). The research dataset consists of above three products downloaded for 10 locations distributed across different geographical regions of Nepal from 2022-2023. Using these dataset the cGAN models are trained with different configuration. An initial cGAN model by Bermudez saw improved performance with DEM inclusion and further enhanced by integrating an attention block in the residual block. Thus, attention cGAN (A_ cGAN) model was selected for further analysis. Among, the A_cGAN models the A_cGAN_SOD model performed the best which uses the dataset (SAR, DEM and optical images) while comparing the MAE, RMSE, SSIM and PSNR values. The image generated from A_cGAN_SOD model for Hetauda was used for LULC classification using random forest algorithm obtaining the overall accuracy of 89.96% which shows that the model output is applicable

1. INTRODUCTION

The rapid increase in satellite sensors with lower revisit times and higher spatial resolutions has led to the acquisition of vast remotely sensed data products, essential for Earth observation tasks like urban planning, disaster management, and weather forecasting. However, these products are often hindered by factors like cloud cover, atmospheric conditions (haze, fog, and aerosols), seasonal variability, and instrument calibration issues. These challenges create data gaps in spatial and temporal domains, limiting their use in dynamic environmental processes such as crop monitoring, disaster response, and deforestation tracking. Cloud cover, a significant contributor to these gaps, affects around 67% of the Earth's surface, with 55% of land surfaces impacted, as demonstrated by King et al. (2013) in a study on MODIS data. Robust cloud removal and image synthesis methods are therefore crucial to ensuring consistent data availability for various applications.

Active sensors like Synthetic Aperture Radar (SAR), which can penetrate clouds, offer a typical solution to cloud-related issues but are limited by their less descriptive and complex data. To overcome these limitations, SAR data is often integrated with complementary imagery, such as near-date optical images and Digital Elevation Models (DEM), to enhance image recovery quality (Li et al., 2017). Additionally, Generative Adversarial Networks (GANs), introduced by Goodfellow et al., are a deep learning approach for generative modeling, often leveraging Convolutional Neural Network (CNN) architectures. Their objective is to learn the distribution of input data, enabling the generation of new, similar data samples. GANs consist of two neural networks: the generator, a CNN that creates realistic synthetic data, and the discriminator, a deconvolutional neural network that distinguishes between real and generated data. Conditional GANs (cGANs), as proposed by Mirza and Osindero (2014), extend traditional GANs by incorporating conditional information, making them effective in remote sensing for tasks like image synthesis and cloud removal. Studies such as Enomoto et al. (2017), Bermudez et al. (2018), and Christovam et al. (2021) have demonstrated the potential of cGANs in synthesizing high-quality, cloud-free images, enabling better applications of remote sensing data.

Attention mechanisms are pivotal in modern deep learning, enabling models to focus on critical input data while filtering out noise (Niu et al., 2021). Introduced by Bahdanau et al. (2014) in their seminal work on neural machine translation, the concept of "soft" attention allows models to dynamically assign weights to different parts of the input sequence, greatly improving alignment between source and target languages. This innovation has been widely adopted across various architectures, including Conditional Generative Adversarial Networks (cGANs), where attention mechanisms enhance the model's ability to selectively process relevant information, leading to the generation of high-quality and coherent outputs.

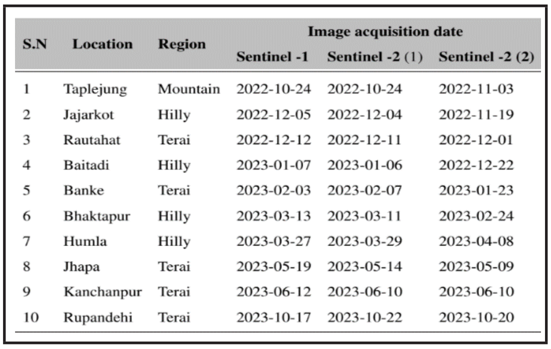
The main objective of this research is to develop attention-based cGAN models capable of synthesizing multispectral images from SAR (Sentinel-1), Optical (Sentinel-2) images and DEM (SRTM 30).

2. DATASETS AND PREPROCESSING

2.1 Datasets

The dataset for this research consists of Sentinel 1 (SAR) images, Sentinel 2 (optical) images and DEM of ten locations of Nepal as mentioned in section 2.1.

Table 1: Image acquisition dates with regions

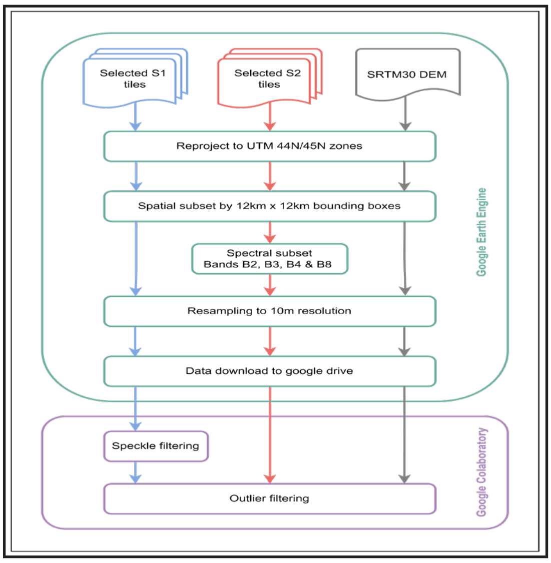


S.N	Location	Region	Image acquisition date		
			Sentinel -1	Sentinel -2 (1)	Sentinel -2 (2)
1	Taplejung	Mountain	2022-10-24	2022-10-24	2022-11-03
2	Jajarkot	Hilly	2022-12-05	2022-12-04	2022-11-19
3	Rautahat	Terai	2022-12-12	2022-12-11	2022-12-01
4	Baitadi	Hilly	2023-01-07	2023-01-06	2022-12-22
5	Banke	Terai	2023-02-03	2023-02-07	2023-01-23
6	Bhaktapur	Hilly	2023-03-13	2023-03-11	2023-02-24
7	Humla	Hilly	2023-03-27	2023-03-29	2023-04-08
8	Jhapa	Terai	2023-05-19	2023-05-14	2023-05-09
9	Kanchanpur	Terai	2023-06-12	2023-06-10	2023-06-10
10	Rupandehi	Terai	2023-10-17	2023-10-22	2023-10-20

For each location, one Sentinel-1 image and two Sentinel-2 images were acquired with acquisition dates as near as possible. The images span a time frame of one year from 2022 October to 2023 October covering the various seasons. The images have a cloud cover threshold of 5% and thus, there are no images from July, August and September as these are the heavy rainfall months. A detailed overview of the locations with the region and image acquisition dates is given in table 1 above.

2.2 Data download and pre-processing

Nepal lies in the UTM zones 44N and 45N. Among the 10 locations, 6 lie in 44N and 4 in 45N. They were spatially subset by bounding boxes of size 12km x 12km and spectral subset was applied to S2 images to select 4 bands with 10m resolution (B4-Red, B3-Green, B2-Blue and B8-NIR) out of 12 bands. Finally, the datasets were resampled to 10m resolution and downloaded to Google Drive.



Selected S1 tiles
Selected S2 tiles
SRTM30 DEM
Reproject to UTM 44N/45N zones
Spatial subset by 12km x 12km bounding boxes
Spectral subset Bands B2, B3, B4 & B8
Resampling to 10m resolution
Data download to google drive
Google Earth Engine
Speckle filtering
Outlier filtering
Google Colaboratory

Figure 1: Data pre-processing

SAR images contain granular noise known as speckle. A median filtering that replaces each value of the pixel with the median value within the local window was used to remove speckles. Furthermore, the outliers in the pixel values for SAR, optical and DEM images are filtered out using histogram based outlier filtering. This filtering helps in reduction of noise, enhance interpretation and analysis and improve visual quality.

3. METHODOLOGY

The process begins with data collection from three primary sources: Sentinel-1, Sentinel-2, and DEM. Once collected, the data undergoes pre-processing to ensure consistency and quality for training. This stage involves noise filtering and normalization to standardize the data. Then, different datasets were prepared consisting a combination of different sources of data mentioned above and patches were extracted and divided into training, validation, and test sets.

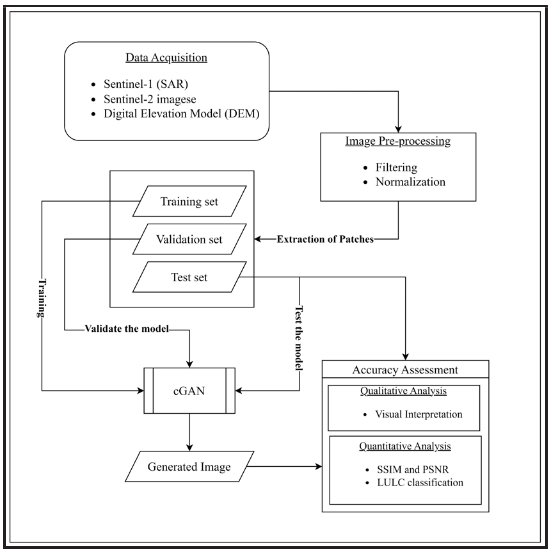


Figure 2: Research Methodology

Next step was to develop cGAN models with different settings and to train them to learn the complex relationships between SAR, optical, and DEM data. After training, the model's performance is evaluated using validation and test sets. Metrics such as MAE and RMSE were used for evaluation of model performance and the best performing model was selected for further analysis. While SSIM and PSNR were used to quantitatively evaluate the synthesized image, visual inspection was done for qualitative evaluation. SSIM is a widely used metric for measuring the similarity between two images while PSNR is a quantitative metric for assessing the quality of a reconstructed image relative to its original version. The overall flow of the proposed framework is shown in Figure 2.

3.1 Data Preparation

Data preparation involved three key steps: normalization, image concatenation, and patch extraction. Composite images were created by combining SAR, optical, and DEM data into four datasets:

i. *SAR_only (Sentinel-1),*
ii. *SAR_DEM (Sentinel-1 with DEM),*
iii. *SAR_OPT (Sentinel-1 with Sentinel-2) and*

iv. *SAR_OPT_DEM (Sentinel-1, Sentinel-2, and DEM)*

Each datasets are paired with target Sentinel-2 optical images. Patches of size 256x256 with a 25-pixel overlap were extracted from these datasets, covering approximately 6.55 km² per patch. A total of 640 patch pairs were generated, augmented, and divided into training (80%), validation (10%), and test (10%) sets.

3.2 cGAN Model Development

The architecture design of a Conditional Generative Adversarial Network (cGAN) consists of two main components: the generator and the discriminator.

Generator: A generator takes noise and additional conditional information as inputs to generate realistic images. The model proceeds with a sequence of the encoder which progressively downsamples the input increasing feature representation. Incorporating an attention module enhances the encoder's ability to focus on salient image regions. Following this, a series of residual blocks refine the feature maps, mitigating gradient vanishing issues. Subsequently, decoder blocks are employed to upsample the features, ultimately generating the output image. The function concludes with the compilation of the generator model, including an activation layer to finalize the output.

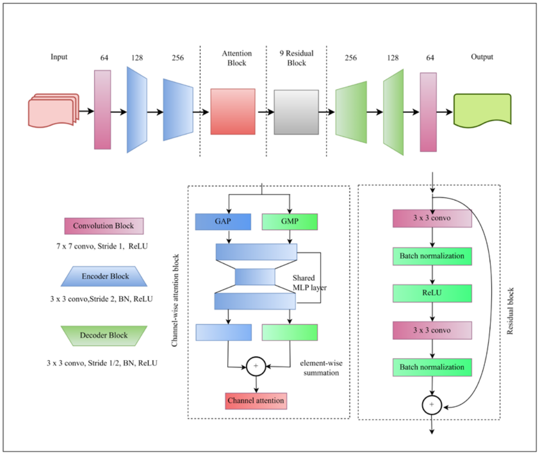


Figure 3: Generator

Discriminator: A Discriminator Model distinguishes real image and generated images by Generator. It starts with an input layer configured to match the dimension and channels of the input images. Then sequentially an encoder blocks is applied to the input image. These blocks, comprised of convolutional layers, progressively down sample the input while enhancing feature representation through activation functions such as Leaky ReLU. Following the encoder blocks, a final convolutional layer computes the logits, representing the discriminative decision for each input. An activation layer employing the sigmoid function subsequently generates the discriminator's output, producing a probability score indicating the likelihood of the input being real or fake.

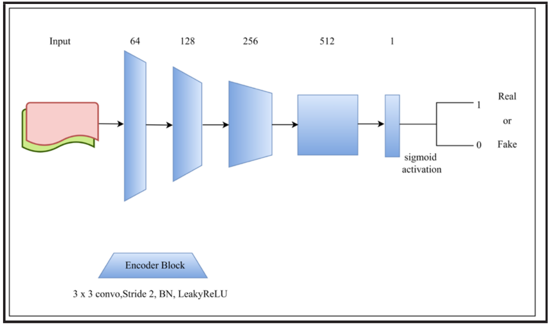


Figure 4: Discriminator

Detailed network architecture of generator and discriminator model used in the study are shown in Figure 3 and 4.

3.2 Training of the cGAN Models

A base cGAN model based on Bermudez, was trained and modified by incorporating various components as mentioned in the sub-sections below. Same hyperparameters from Table 2 were used for training in order to make uniform comparison for all the models.

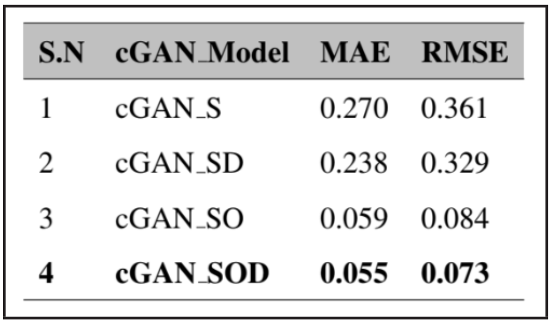
Table 2: Hyperparameter used in the models

Patch Size	Training Size	Validation Size	Test Size	Epoch	Patience	Batch Size	Learning Rate
256	0.8	0.1	0.1	50	20	2	0.0002

3.2.1 Experiment I: DEM Integration

The models cGAN_S, cGAN_SD, cGAN_SO and cGAN_SOD were trained from datasets SAR only, SAR DEM, SAR OPT and SAR OPT DEM respectively. The primary goal of this experiment is to see whether adding the DEM data as additional data to the SAR and optical improves the performance of the model or not.

Table 3: Performance of cGAN models with DEM integration



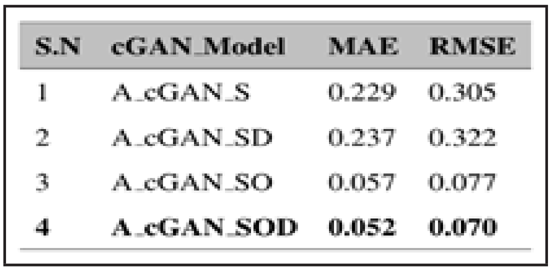
S.N	cGAN_Model	MAE	RMSE
1	cGAN_S	0.270	0.361
2	cGAN_SD	0.238	0.329
3	cGAN_SO	0.059	0.084
4	**cGAN_SOD**	**0.055**	**0.073**

The addition of DEM data contributes in reducing errors, as evidenced by the notable improvement from cGAN S to cGAN SD. Furthermore, incorporating optical data further enhances model accuracy, leading to the lowest MAE and RMSE values in the cGAN SOD variant, which integrates SAR, optical, and DEM data. This underscores the importance of leveraging multiple types of data to improve the predictive capability of the cGAN model.

3.2.2 Experiment II: Adding Attention Block

An attention block was added to the base cGAN model architecture which helps to improve the model's ability to focus on relevant information enhancing its capacity to generate high-quality and coherent outputs. Thus, A_cGAN_S, A_cGAN SD, A_cGAN_SO and A_cGAN_SOD models were obtained.

Table 4: Attention cGAN models



S.N	cGAN_Model	MAE	RMSE
1	A_cGAN_S	0.229	0.305
2	A_cGAN_SD	0.237	0.322
3	A_cGAN_SO	0.057	0.077
4	**A_cGAN_SOD**	**0.052**	**0.070**

These models performed slightly better than the corresponding models in the previous experiment as it was able to achieve lower MAE and RMSE values (Table 4). A total of 8 models were trained and among them the least values of MAE 0.052 and RMSE 0.070 was observed for the model A_cGAN_SOD and was selected for the hyperparameter tuning where the parameters learning rate, batch size and patch size were tuned for the better performance of the model.

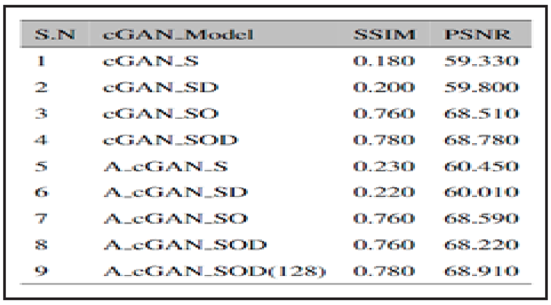
3.3 Hyperparmater Tuning

Firstly, the A_cGAN_SOD model was trained with the batch size of 1, 2, 4 and 8 while other parameters remains the same (learning rate = 0.0002 and patch size = 256). Secondly, the learning rate of 0.02, 0.002, 0.0002 and 0.0001 were taken into consideration while setting the batch size of 2 and patch size 256. Lastly, the model was then trained with the patch size of 128, 256 and 512 where batch size was set to 2 and learning rate 0.0002. The *batch size of 2, learning rate of 0.0002* and *patch size of 128* demonstrates the most favorable results, exhibiting the least MAE of 0.0451 and an RMSE of 0.0632.

4. RESULT AND DISSCUSSION

Table 5 summarizes the performance of various cGAN models based on SSIM and PSNR metrics, emphasizing the advantages of integrating SAR, Optical imagery, and DEM data. Models combining all three data sources outperform those using only SAR or SAR+DEM. Attention-based architectures further boost performance by emphasizing relevant features.

Table 5: Performance of cGAN models



S.N	cGAN_Model	SSIM	PSNR
1	cGAN_S	0.180	59.330
2	cGAN_SD	0.200	59.800
3	cGAN_SO	0.760	68.510
4	cGAN_SOD	0.780	68.780
5	A_cGAN_S	0.230	60.450
6	A_cGAN_SD	0.220	60.010
7	A_cGAN_SO	0.760	68.590
8	A_cGAN_SOD	0.760	68.220
9	A_cGAN_SOD(128)	0.780	68.910

The "A_cGAN_SOD (128)" model achieves the highest SSIM (0.780) and PSNR (68.910), along with low RMSE and MAE during training, demonstrating excellent learning and generalization, making it a strong candidate for practical applications. A_cGAN_SOD(128) model was used to generate the optical image for the date 2023-10-22. Then, the real Sentinel-2 image for the date 2023-10-22 and synthesized image was used for LULC classification whose results will be referred to as LULC_real and LULC_synth with an overall accuracies of 93.21% and 89.96% respectively.

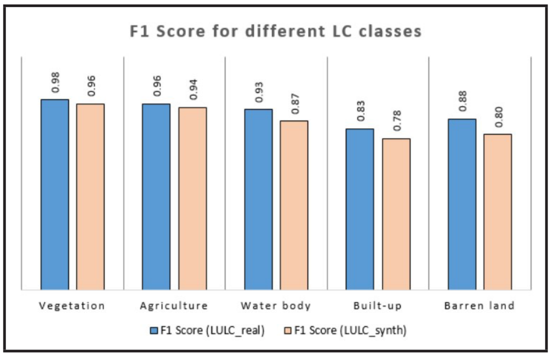


Figure 5: F1 Score for different LC classes

Figure 5 compares the F1 scores for different land cover (LC) classes between LULC_real and LULC_synth. Across all land cover classes, the F1 scores for LULC_synth are slightly lower than those for LULC_real, indicating a marginally reduced performance for the synthetic dataset. However, the scores remain within an acceptable range demonstrating the synthetic dataset's reliability for land cover classification tasks.

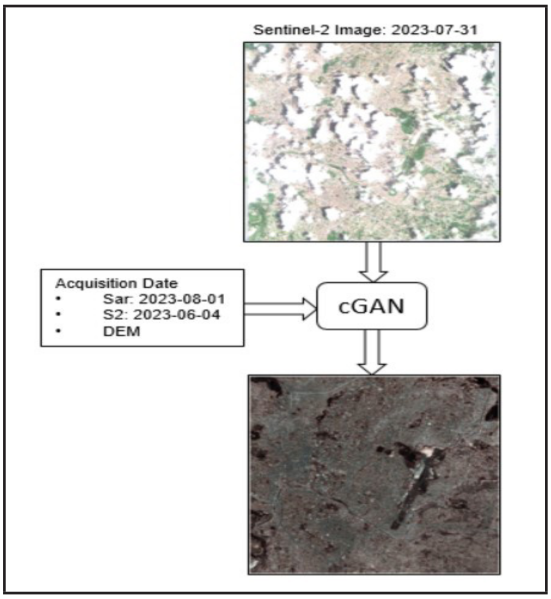


Figure 6: A_cGAN_SOD outcome

Figure 6 illustrate the outcome produced by the proposed model in the research. Initially, a Sentinel-2 image from July 31, 2023, was observed to contain significant cloud cover. To address this, the model utilized a combination of inputs: a SAR image from August 1, 2023, a cloud-free Sentinel-2 image from June 4, 2023, and a DEM. By combining these inputs, the model reconstructs a clear and usable image corresponding to July 31, 2023, overcoming the limitations caused by cloud cover in the original Sentinel-2 data.

5. CONCLUSION

Various models underwent training and testing, and models incorporating DEM data alongside SAR and optical data resulted in a marginal improvement than without it, indicating that supplementary conditioning data can enhance overall model accuracy. Also, A_cGAN_SOD, cGAN model with attention mechanism and using SAR, optical and DEM dataset, demonstrated superiority over other models, boasting the lowest RMSE and MAE values. It also achieved the highest SSIM and PSNR values showing that adding attention mechanism can generate more realistic and detailed outputs that capture important characteristics of the input data.

Moreover, the images produced by the A_cGAN_SOD model demonstrate substantial promise for geospatial analysis, effectively serving as substitutes for optical data compromised by cloud cover or other distortions. The model achieves a commendable Overall Accuracy (OA) score of 89.96% in Land Use Land Cover (LULC) classification tasks, underscoring its effectiveness in generating reliable and accurate imagery for analytical purposes.

REFERENCES

Bahdanau, D., Cho, K., & Bengio, Y. (2014). Neural machine translation by jointly learning to align and translate. arXiv preprint arXiv:1409.0473.

Bermudez, J., Happ, P., Oliveira, D., & Feitosa, R. (2018). Sar to optical image synthesis for cloud removal with generative adversarial networks. ISPRS Annals of the Photogrammetry, Remote Sensing and Spatial Information Sciences, 4, 5–11.

Bermudez, J. D., Happ, P. N., Feitosa, R. Q., & Oliveira, D. A. (2019). Synthesis of multispectral optical images from sar/optical multitemporal data using conditional generative adversarial networks. *IEEE Geoscience and Remote Sensing Letters*, 16(8), 1220–1224.

Christovam, L., Shimabukuro, M., Galo, M. T., & Honkavaara, E. (2021). Evaluation of SAR to optical image translation using conditional generative adversarial network for cloud removal in a crop dataset. *The International Archives of the Photogrammetry, Remote Sensing and Spatial Information Sciences*, 43, 823–828.

Enomoto, K., Sakurada, K., Wang, W., Fukui, H., Matsuoka, M., Nakamura, R., & Kawaguchi, N. (2017). Filmy cloud removal on satellite imagery with multispectral conditional generative adversarial nets. *Proceedings of the IEEE conference on computer vision and pattern recognition workshops*, 48–56.

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

King, M. D., Platnick, S., Menzel, W. P., Ackerman, S. A., & Hubanks, P. A. (2013). Spatial and temporal distribution of clouds observed by modis onboard the terra and aqua satellites. *IEEE transactions on geoscience and remote sensing*, 51(7), 3826–3852.

Li, Y., Fu, R., Meng, X., Jin, W., & Shao, F. (2020). A sar-to-optical image translation method based on conditional generation adversarial network (cgan). IEEE Access, 8, 60338–60343

Mirza, M., & Osindero, S. (2014). Conditional generative adversarial nets. arXiv preprint arXiv:1411.1784.

Niu, Z., Zhong, G., & Yu, H. (2021). A review on the attention mechanism of deep learning. Neurocomputing, 452, 48–62.

Author's Information

Name	:	David Nhemaphuki
Academic Qualification	:	MS in Geoinformatics
Organization	:	Survey Department
Current Designation	:	Survey Officer
Work Experience	:	9 yrs.
No. Published paper/article	:	2