CLOUDMASKGAN: A CONTENT-AWARE UNPAIRED IMAGE-TO-IMAGE TRANSLATION ALGORITHM FOR REMOTE SENSING IMAGERY

Sorour Mohajerani[1], Reza Asad[2], Kumar Abhishek[2], Neha Sharma[2], Alysha van Duynhoven[3], Parvaneh Saeedi[1]

[1]School of Engineering Science, Simon Fraser University, Canada
[2]School of Computing Science, Simon Fraser University, Canada,
[3]Department of Geography, Simon Fraser University, Canada

ABSTRACT

Cloud segmentation is a vital task in applications that utilize satellite imagery. A common obstacle in using deep learning-based methods for this task is the lack of a large number of images with their annotated ground truths. This work presents a content-aware unpaired image-to-image translation algorithm. It generates synthetic images with different land cover types from original images, while it preserves the locations and the intensity values of the cloud pixels. Therefore, no manual annotation of ground truth in these images is required. The visual and numerical evaluations of the generated images by the proposed method prove that their quality is better than that of competitive algorithms.

Index Terms— Cloud, Landsat 8, remote sensing imagery, unpaired image-to-image translation

1. INTRODUCTION

Obtaining an accurate measure of cloud cover is a crucial step in applications of satellite imagery. Presence of clouds and variant levels of cloud coverage could affect the value of an image. Indeed, transferring images with a high level of cloud coverage to a ground station is wasteful and unnecessary. On the other hand, cloud pixels in satellite images hold valuable information about the atmospheric parameters in weather studies [1]. The necessity for accurate identification of cloud-covered regions in satellite images therefore opens up an active research area. Automatic cloud detection and segmentation is a more challenging task when there is access to only a limited number of bands (e.g . Red, Green, Blue, and Near-infrared (NIR)). Such spectral band limitation exists in many satellites, such as HJ-1 and GF-2 [2].

In recent years, many cloud detection algorithms have been developed [3, 4, 5]. Among these, deep learning-based approaches have shown promising performance [6, 7, 8]. Access to vast satellite imagery datasets with corresponding ground truths becomes essential for successful deep learning-based segmentation algorithms. Unfortunately, remotely sensed data requires manual annotation of images, which is tedious and time-consuming due to the large size of images (an average of 9000×9000 pixels for Landsat 8 images). Unlike many other computer vision tasks, collecting additional data and relevant ground truths is unfeasible, making the expansion of existing training sets problematic.

In this work, we propose an approach for generating synthetic satellite imagery. Generated images should satisfy two main criteria: (1) they should exhibit realistic appearances with the original training images in texture, style, etc., and (2) the cloud pixel location and intensity values in the generated images should remain intact, so that the existing ground truths can be reused in further cloud detection algorithm.

Clouds share similar reflection characteristics with a nuber of ground objects such as snow and ice. Therefore, distinguishing these regions from clouds is a challenging task for most cloud segmentation algorithms [9]. To generate more diverse and challenging examples for the cloud detection frameworks, we narrow our focus to converting snowy landscapes to non-snowy ones, and vice versa. For this task, we propose an image-to-image translation model called CloudMaskGAN, which is a modification of the well-known CycleGAN [10] model. Our model is capable of capturing features from both snowy and non-snowy land covers, converting them to one another, while preserving cloud pixels.

2. RELATED WORKS

Generative Adversarial Networks (GANs), introduced by Goodfellow et al. [11], have attained success in realistic image generation tasks. Various versions of these generative models have been developed for a variety of applications, including text to image synthesis [12] and generating videos [13].

Presenting an approach to unpaired image-to-image translation, CycleGAN has gained attention for its ability to transfer characteristics from images in one domain to another one and vice versa. A cycle-consistensy loss makes the training process unsupervised, meaning that no example of paired images in both domains is required. Another well-known approach is called Multimodal Unsupervised Image-to-Image Translation (MUNIT) [14]. In this method, a system learns

the style of the objects in one domain and the shared content between two domains. Then it converts the style of one domain to the other one while preserving the content. Multiple translated images can be obtained using this approach.

The flexibility of the above mentioned image-to-image translation algorithms to work with unpaired images from two domains is suitable for remote sensing data, where capturing a scene from a specific location with consistent cloud regions is not possible. Given our satellite image dataset, we would like to translate images with non-snowy/non-icy land cover to snowy/icy land cover while preserving the cloud pixels. Unfortunately, CycleGAN does not guarantee this. Our experiments in the non-snow to snow translation task show a tendency for CycleGAN to convert cloudy regions to snow (cloud elimination). We also observe that in the snow to non-snow translation, MUNIT converts some snow-covered areas to fake clouds. These drawbacks motivated us to modify CycleGAN to preserve clouds.

3. PROPOSED METHOD

3.1. CloudMaskGAN

To translate images from domain X to Y, CycleGAN model estimates two mappings: $G : X \rightarrow Y$ and $F : Y \rightarrow X$. In this work, domain X is non-snow land cover and domain Y is snow land cover. The mappings must satisfy a few requirements: (1) G should map an image $x \in X$ to a realistic image in domain Y (and similarly, F should create a realistic image in domain X given $y \in Y$), and (2) it should translate an image $x \in X$ into domain Y and translating the result back via F should give a good estimate of x. The same logic should also hold for the other direction of the translation. The first requirement is encoded using an adversarial loss and the second one is encoded as a cycle consistency loss. Starting from domain X, the cycle consistency loss is denoted as:

$$L_{cyc}(G, F, X) = E_{x\sim p_{data}(x)}\big[\|F(G(x)) - x\|_1\big]. \quad (1)$$

A similar equation describes the cycle consistency starting from Y domain. Also, an identity constraint helps to preserve the color information that may not have been preserved via the adversarial or the cycle consistency losses. This constraint requires the mapping G to act as identity when applied to an image $y \in Y$:

$$L_{identity}(G, Y) = E_{y\sim p_{data}(y)}\big[\|G(y) - y\|_1\big]. \quad (2)$$

Similarly, the mapping F applied to an image in domain X should roughly result in the same image.

To make sure that translated images are realistic enough and are generally similar to the original images of the target domain, a discriminator D_Y tries to distinguish between real images of y and the translated versions, $G(x)$. This can be summarized as an adversarial loss function by:

$$L_{GAN}(G, D_Y, X, Y) = E_{y\sim p_{data}(y)}[log(D_Y(y))] + E_{x\sim p_{data}(x)}[log(1 - D_Y(G(x)))]. \quad (3)$$

The same equation exists for the target domain X and D_x.

CloudMaskGAN forces the model to generate new synthetic information only outside of the cloud pixels while retaining the cloud pixels. There are two ingredients to achieve this goal: cloud ground truths, and a cloud consistency loss. Cloud ground truths are utilized in all mentioned loss function components to force the model to change only the non-cloud regions. Therefore, modified loss functions for translating images from domain X to Y are:

$$L_{cyc}(G, F, X, M_X) = E_{x\sim p_{data}(x)}[\|F(G(x)) * M_X - x * M_X\|_1],$$
$$L_{identity}(F, X, M_X) = E_{x\sim p_{data}(x)}[\|F(x) * M_X - x * M_X\|_1],$$
$$L_{GAN}(G, D_Y, X, Y, M_X, M_Y) = E_{y\sim p_{data}(y)}[log(D_Y(y * M_Y))] + E_{x\sim p_{data}(x)}[log(1 - D_Y(G(x) * M_X))]. \quad (4)$$

M_X and M_Y denote the logical inverse of the cloud ground truths of images in domains X and Y, respectively. $*$ denotes element-wise matrix multiplication. For example, M_X is a matrix of the same size as the real image x where all the cloud pixels are replaced with zeros, and all the non-cloud pixels are replaced with ones.

The cloud consistency loss to preserve the cloud pixels between real and generated images is described as:

$$L_{cloud}(G, X, M_X) = E_{x\sim p_{data}(x)}[\|G(x) * (1 - M_X) - x * (1 - M_X)\|_2^2]. \quad (5)$$

The CloudMaskGAN loss function is a linear combination of the components in equations 4 and 5:

$$L(G, F, D_Y, X, Y, M_X, M_Y) = L_{GAN}(G, D_Y, X, Y, M_X, M_Y) + \lambda_1 L_{identity}(F, X, M_X) + \lambda_2 L_{cyc}(G, F, X, M_X) + \lambda_3 L_{cloud}(G, X, M_X). \quad (6)$$

Here, λ_1, λ_2 and λ_3 control the degree to which we enforce the identity, cycle consistency, and cloud consistency. By setting $\lambda_1 = 25$, $\lambda_2 = 10$ and $\lambda_3 = 20$, found through trial and error, we achieved the best results. The complete loss will have four additional loss terms for translation from domain Y to X. Fig. 1 displays CloudMaskGAN model in detail. The utilized network for F and G is the one introduced in [15] and for D is PatchGAN [10]. Both of these networks led to the best results of CycleGAN model and we have reused them in this work. The images are resized to $192 \times 192 \times 4$ before being fed to the model. CloudMaskGAN is trained for 200 epochs with a starting learning rate of 0.0002. After the 100th epoch, the learning rate is linearly reduces to zero.

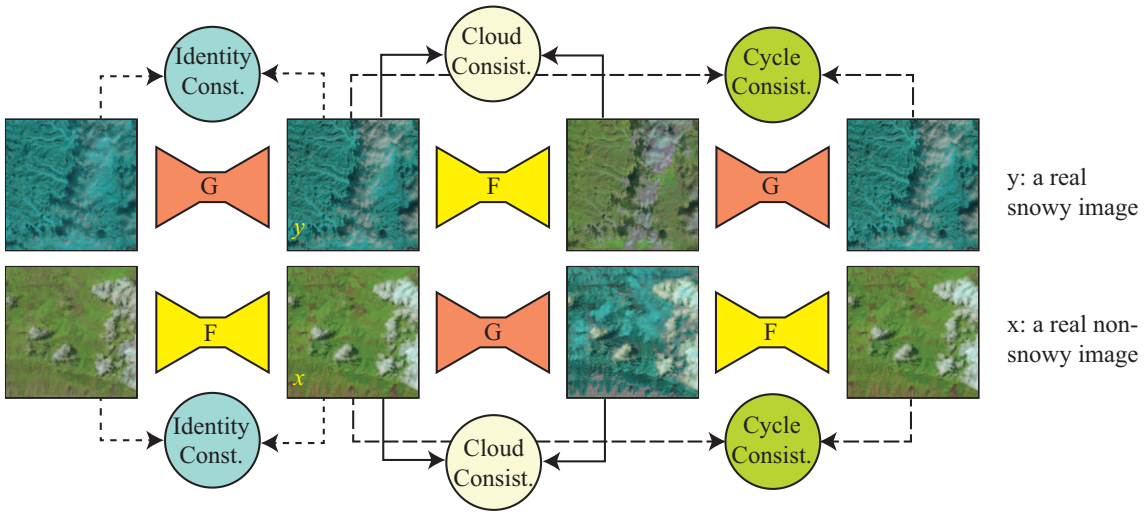


Fig. 1: CloudMaskGAN model. This model learns F and G mappings such that G translates a non-snowy image (x) to a snowy one ($G(x)$) and F translates a snowy image (y) to a non-snowy one ($F(y)$) while preserving clouds. For simplicity, GAN loss is not displayed in this figure. False color images are used here for further clarification. Regions in cyan represent snow.

4. EXPERIMENTS AND RESULTS

4.1. Dataset

38-Cloud, the dataset introduced in [16], is utilized for image-to-image translation in this work. This dataset is a modified version the dataset introduced in [2]. The images collected cover various locations in North America. 38-Cloud contains 8400 patches that have been extracted from 18 Landsat 8 images for training. The ground truths of these images have been manually annotated. Patches are of 384×384 pixels with a spatial resolution of 30 meters. Each patch consists of four channels of Blue, Green, Red, and NIR, which are band 2 to band 5 of the Landsat 8 products.

The data for the training of CloudMaskGAN was selected from the training set of the 38-Cloud dataset. We gathered 1854 patches for training and testing/evaluation of the proposed CloudMaskGAN. The cloud coverage of these patches is lower than 85%. 1069 of these patches have non-snowy and 785 have snowy land cover. Snow or non-snow image annotations were manually extracted for these patches. We have kept the ratio of images in these two domains close to that of Yosemite Winter and Summer Flickr dataset described in [10]. Since the number of images available was not large enough, we trained CloudMaskGAN on all 1854 images and then used the trained model to generate synthetic images for those 1854 images.

4.2. Evaluation Metrics and Experimental Design

By evolution of synthetic image generation algorithms, many metrics for measuring the quality and diversity of the generated images have been developed, such as Inception Score (IS) [17] and Fréchet Inception Distance (FID) [18]. Both of these metrics are dependent on the response of generated images to a pre-trained CNN on ImageNet dataset [19]. For computing FID, the feature maps of real and generated images are obtained using an Inception-v3 [20] network pre-trained on ImageNet. Those features are then compared to each other to obtain the disturbance level between them.

Unfortunately, these metrics are not a proper choice for different types of images such as biomedical and remote sensing [21]. Indeed, weights of a network pre-trained on ImageNet are not able to capture the representative features of remote sensing data since not only are this dataset's images completely different from remote sensing data, but also they are colored. On the contrary, even RGB remote sensed images are generally dark because of the specific settings of satellites' image acquisition equipment. As a result, no natural color is visible in these images. Observing the limitations of some numerical metrics for measuring the quality of generated images, some researchers utilize the objective evaluation methods such as Amazon Mechanical Turk, which involves human participants [22] to label fake and real images. Unfortunately, these evaluations can be biased and far from accurate.

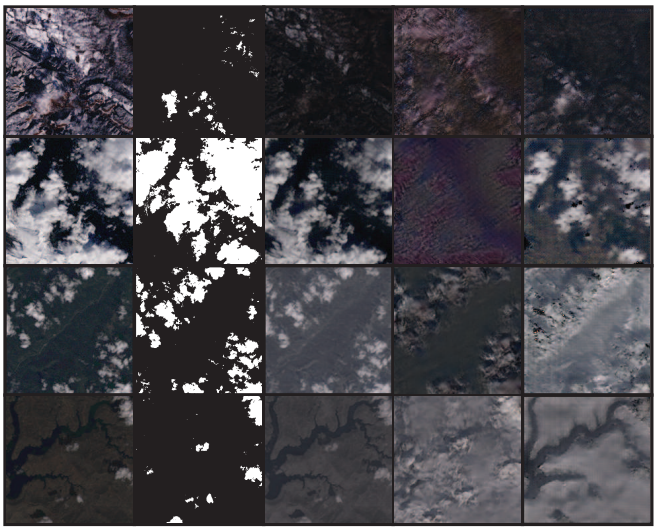
For evaluation of our image translation framework, we have utilized the method introduced in [21]. This method consists of two experiments: GAN-test and GAN-train. GAN-test is to use a third party pre-trained network on real images to test the generated data and obtain the performance. GAN-train is to train a third party network on only the generated data and test it on real test images and obtain the numerical performance. The higher the performance measures of these experiments, the better the quality of the generated data. In our case, we should measure the performance of cloud seg-

mentation in GAN-test and GAN-train. Four pixel-wise metrics of precision, recall, Jaccard Index, and overall accuracy are utilized to evaluate the performance with the following formulations:

$$\begin{aligned} \text{Jaccard Index} &= \frac{TP}{TP+FN+FP}, \\ \text{Precision} &= \frac{TP}{TP+FP}, \\ \text{Recall} &= \frac{TP}{TP+FN}, \\ \text{Overall Accuracy} &= \frac{TP+TN}{TP+TN+FP+FN}. \end{aligned} \tag{7}$$

Here TP, TN, FP, and FN are the total number of true positive, true negative, false positive, and false negative pixels, respectively.

For our GAN-test, we have utilized the U-Net network in [2] pre-trained on real Landsat 8 images of the introduced dataset in [2]. For our GAN-train, we simplified this U-Net to be trainable with 1854 images. Indeed, we have removed two encoder blocks (Encode 5 and Encode 6) and two decoder blocks (Decode 1 and Decode 2). The training and the test settings of our GAN-train and Gan-test are the same as reported in [2]. We have conducted GAN-test and GAN-train experiments for the images generated by CycleGAN, MUNIT, and CloudMaskGAN.



(a) (b) (c) (d) (e)

Fig. 2: Image-to-image translation visual results: (a) RGB image, (b) ground truth, (c) CycleGAN, (d) MUNIT, (e) CloudMaskGAN. First two rows are samples of snow to non-snow translation, while the last two rows show samples of non-snow to snow translation. Pixel values are adjusted for better visualization.

4.3. Results and Discussion

Table 1 shows the quantitative results associated with the generated images. As it is clear, the GAN-train Jaccard Index and overall accuracy of the proposed CloudMaskGAN outperforms that of CycleGAN by 76.7% and 22.1%. Please note that the images obtained by MUNIT led to overfitting of the GAN-train training phase. Therefore, we report only the overall accuracy of this experiment. Also, the GAN-test Jaccard Index and overall accuracy of the CloudMaskGAN are higher than that of CycleGAN by 156.2% and 91.1% and that of MUNIT by 81.7% and 20.9%. This shows that images generated by CloudMaskGAN are closer to the real remote sensing data than those generated by CycleGAN and MUNIT. Therefore, a cloud detection network can perform better over these images. Also, training a cloud segmentation network on the generated images leads to better weights. Thus, those weights yield superior cloud masks on unseen real remote sensing images. Some visual results of the proposed augmentation method are shown in Fig. 2. In the first row of this figure, one can see the problem of cloud elimination made by CycleGAN model and fake clouds generated by MUNIT. Also, some fake clouds appeared in the MUNIT result of the last row in Fig. 2. These problems do not exist in the generated image by the proposed method.

Table 1: GAN-train and GAN-test of generated images (in %).

Method	Jaccard	Precision	Recall	Overall Accuracy
CycleGAN GAN-train	26.17	54.14	38.56	69.09
MUNIT GAN-train	-	-	-	31.21
CloudMaskGAN GAN-train	**46.24**	**66.50**	**68.57**	**84.36**
CycleGAN GAN-test	10.76	21.49	37.60	39.14
MUNIT GAN-test	15.17	33.25	42.35	61.86
CloudMaskGAN GAN-test	**27.57**	**47.05**	**49.96**	**74.80**

5. CONCLUSION

The proposed CloudMaskGAN is capable of generating realistic synthetic remote sensing images. By incorporating ground truths, CloudMaskGAN provides a content-aware image-to-image translation approach that can be extended to other computer vision tasks. Since cloud regions remain intact in the generated images, no expensive annotation of these images is required. Given imbalanced or limited datasets (and ground truths), CloudMaskGAN has the potential to generate high quality and diverse synthetic data that can retain pixel values in specific regions. In future works, translating other types of land covers in satellite images (such as vegetation, water, bare soil, etc.) will be experimented.

6. REFERENCES

[1] R. S. Reddy, D. Lu, F. Tuluri, and M. Fadavi, "Simulation and prediction of hurricane lili during landfall over the central gulf states using mm5 modeling system

and satellite data," in *IEEE Int. Geo. and Remote Sens. Symp. (IGARSS)*, 2017, pp. 36–39.

[2] S. Mohajerani, T. A. Krammer, and P. Saeedi, "A cloud detection algorithm for remote sensing images using fully convolutional neural networks," in *IEEE Int. Workshop on Multimedia Sign. Proc. (MMSP)*, 2018, pp. 1–5.

[3] Y. Zhang, B. Guindon, and J. Cihlar, "An image transform to characterize and compensate for spatial variations in thin cloud contamination of Landsat images," *Remote Sens. of Env.*, vol. 82, pp. 173 – 187, 2002.

[4] Z. Zhu and C. Woodcock, "Object-based cloud and cloud shadow detection in landsat imagery," *Remote Sens. of Env.*, vol. 118, pp. 8394, 2012.

[5] Z. Zhu, S. Wang, and C. E. Woodcock, "Improvement and expansion of the fmask algorithm: cloud, cloud shadow, and snow detection for Landsats 47, 8, and sentinel 2 images," *Remote Sens. of Env.*, vol. 159, pp. 269 – 277, 2015.

[6] F. Xie, M. Shi, Z. Shi, J. Yin, and D. Zhao, "Multilevel cloud detection in remote sensing images based on deep learning," *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, vol. 10, no. 8, pp. 3631–3640, 2017.

[7] D. Tuia, B. Kellenberger, A. Prez-Suey, and G. Camps-Valls, "A deep network approach to multitemporal cloud detection," in *IGARSS 2018 - 2018 IEEE International Geoscience and Remote Sensing Symposium*, 2018, pp. 4351–4354.

[8] K. Yuan, G. Meng, D. Cheng, J. Bai, S. Xiang, and C. Pan, "Efficient cloud detection in remote sensing images using edge-aware segmentation network and easy-to-hard training strategy," in *2017 IEEE International Conference on Image Processing (ICIP)*, Sep. 2017, pp. 61–65.

[9] Y. Zi, F. Xie, and Z. Jiang, "A cloud detection method for Landsat 8 images based on pcanet," *Remote Sensing*, vol. 10, no. 6, 2018.

[10] J. Zhu, T. Park, P. Isola, and A. A. Efros, "Unpaired image-to-image translation using cycle-consistent adversarial networks," in *2017 IEEE International Conference on Computer Vision (ICCV)*, 2017, pp. 2242–2251.

[11] I. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville, and Y. Bengio, "Generative adversarial nets," in *Advances in Neural Information Processing Systems (NeurIPS)*, 2014, pp. 2672–2680.

[12] S. E. Reed, Z. Akata, X. Yan, L. Logeswaran, B. Schiele, and H. Lee, "Generative adversarial text to image synthesis," *CoRR*, vol. abs/1605.05396, 2016.

[13] C. Vondrick, H. Pirsiavash, and A. Torralba, "Generating videos with scene dynamics," *CoRR*, vol. abs/1609.02612, 2016.

[14] X. Huang, M. Liu, S. Belongie, and J. Kautz, "Multimodal unsupervised image-to-image translation," *European Conference on Computer Vision (ECCV)*, 2018.

[15] J. Johnson, A. Alahi, and F. Li, "Perceptual losses for real-time style transfer and super-resolution," in *European Conference on Computer Vision (ECCV)*, 2016.

[16] S. Mohajerani and P. Saeedi, "Cloud-Net: An end-to-end cloud detection algorithm for Landsat 8 imagery," *CoRR*, vol. abs/1901.10077, 2019.

[17] T. Salimans, I. Goodfellow, W. Zaremba, V. Cheung, A. Radford, and X. Chen, "Improved techniques for training gans," in *Advances in Neural Information Processing Systems (NeurIPS)*, 2016, pp. 2234–2242.

[18] M. Heusel, H. Ramsauer, T. Unterthiner, B. Nessler, and Sepp Hochreiter, "Gans trained by a two time-scale update rule converge to a local nash equilibrium," in *Advances in Neural Information Processing Systems (NeurIPS)*, 2017, pp. 6626–6637.

[19] J. Deng, W. Dong, R. Socher, L. Li, K. Li, and F. Li, "Imagenet: A large-scale hierarchical image database," in *Computer Vision and Pattern Recognition, 2009. CVPR 2009. IEEE Conference on*. Ieee, 2009, pp. 248–255.

[20] C. Szegedy, V. Vanhoucke, S. Ioffe, J. Shlens, and Z. Wojna, "Rethinking the inception architecture for computer vision," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016, pp. 2818–2826.

[21] K. Shmelkov, C. Schmid, and K. Alahari, "How good is my gan?," in *European Conference on Computer Vision (ECCV)*, 2018.

[22] X. Wu, J. Shao, L. Gao, and H. T. Shen, "Unpaired image-to-image translation from shared deep space," in *2018 25th IEEE International Conference on Image Processing (ICIP)*, 2018, pp. 2127–2131.