

Universitat de Lleida

Economic Viability

Made by
Oriol Alàs Cercós

Delivery
25th of May, 2022

Universitat de Lleida
Escola Politècnica Superior
Màster en Enginyeria Informàtica
Technological Business Management and Entrepreneurship

Professorate:

Josep Escribà Garriga

Contents

1	Introduction	2
1.1	Sentinel2	2
1.2	Related work and state-of-the-art	2
2	Data	12
2.1	Academic Datasets	12
3	Model architecture	12
4	Experiments & results	12
5	Conclusions	12
6	Bibliography	12

List of Figures

1	Residual learning: a building block.	3
2	DSen2-CR model diagram.	3
3	Examples of the influence of SAR data. Being (a, f) cloudy images, (b, g), SAR images, (c, h), the target cloudless images, (d, i), the synthesized images without SAR input data and (e, j) the predictions using SAR data as input.	4
4	Prediction results by McGAN with the synthesized cloud images	6
5	Prediction results by McGAN with the real cloud images	6
6	Comparison between the cloudy images, the outputs of McGAN and Pix2Pix-based models and the ground-truth image.	8
7	Comparison between cloudy images, the outputs of McGAN versions and the ground-truth image.	8
8	Failure cases in CloudGAN attempting to removal overly clouded images over-smoothening or completely failing.	11
9	Failure cases in CloudGAN attempting to removal overly clouded images over-smoothening or completely failing.	11

Acronyms

cGAN Conditional Generative Adversarial Network.

CNN Convolutional Neural Network.

GAN Generative Adversarial Network.

McGAN Multi-spectral Conditional Generative Adversarial Network.

NIR Near Infra Red.

NLP Natural Language Processing.

PSNR Peak signal-to-noise ratio.

RS Remote Sensing.

SAR Synthetic Aperture Radar.

SSIM Structural Similarity Index Metric.

1 Introduction

Remote Sensing (RS) imagery is critical to perform challenges such climate change or natural resources management, including zone monitoring for reforestation, disaster mitigation and land surface change detection. Nevertheless, on average 55% of the Earth's land surfaces is covered by clouds, being then a significant impediment to carry out a broad range of applications. Satellite imagery plagued by films of clouds that obstructs the scene implies a great loss of information or causing effects such as blurring, which mitigates the power of RS. Hence, RS applications definitely needs a generic technique to detect and remove the cloudy region with an in-painting of the underlying scene.

1.1 Sentinel2

Sentinel2 images are provided by two satellites, Sentinel 2A and Sentinel 2B, which orbit each other with a 180° phase shift. Generally, the acquisition of the images is 10 days per satellite or more or less 5 days altogether. Therefore, a new updated image of a specific area is available in periods of time not exceeding five days. This makes Sentinel-2 data an excellent choice for studying environmental challenges. Sentinel-2 data is multi-spectral with 13 bands in the visible, near-infrared, and short-wave infrared spectrum. These bands come in a different spatial resolution ranging from 10m to 60m, so the images can be classified as medium-high resolution.

1.2 Related work and state-of-the-art

Deep learning have been a popular and efficient technique to solve challenges from satellite imagery. Specifically, Convolutional Neural Network (CNN) have been the main architecture of neural networks to provide a solution from image-based problems.

In [1], they create a deep learning approach to Sentinel-2 super-resolution. Their hypothesis was the existence of a complex mixture of correlations across many spectral bands over a large spatial context. Hence, the input of the model is a concatenation of the high-level resolution bands with the low-level resolution bandwidths upsampled to 10m by simple bi-linear interpolations. The model itself is a clear reference of residual networks [2]. Furthermore, as in ResNet architectures, it uses skip connections to reduce the average effective path length through the network, alleviate the vanishing gradient problem and

greatly accelerates the learning during the training.

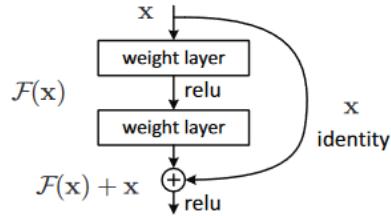


Figure 1: Residual learning: a building block.

Similarly, in [3], a residual network is used with the same skip connections mechanism to bring a solution to cloud removal challenge. Regarding the design of the neural network, DSen2-CR is a fully convolutional network, so it can accept input images of any spatial dimensions (m), as it can be seen in 2. The output of DSen2-CR is a 13-channel layer, representing the thirteen bands from Sentinel2.

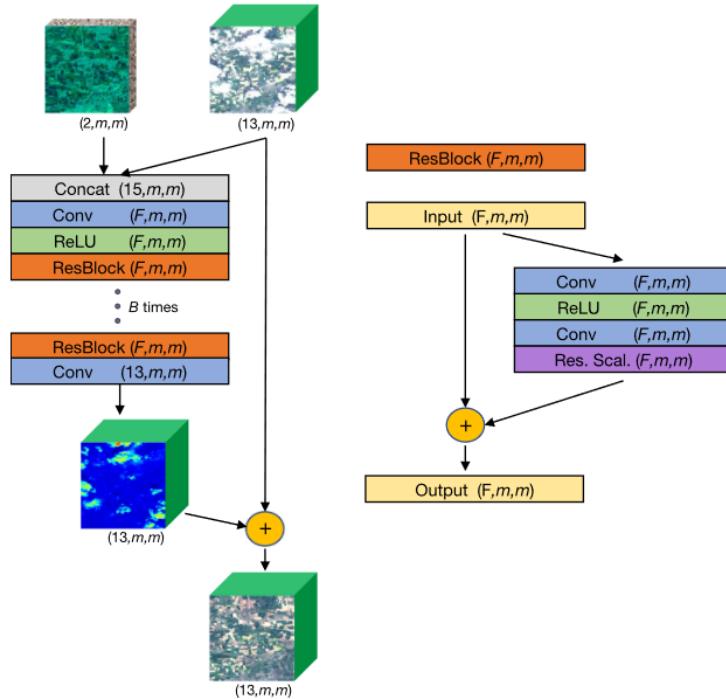


Figure 2: DSen2-CR model diagram.

It also uses Synthetic Aperture Radar (SAR) optical data, which represents an important complementary source to help the model to make greater results. However, SAR images are affected by a particular type of noise called speckle, which can difficult network's learning. In addition to that, the model depends on one more source to remove the clouds, as SAR

images cannot be downloaded in Sentinel-2. It can be seen in 3 that the lack of SAR data make the network less powerful.

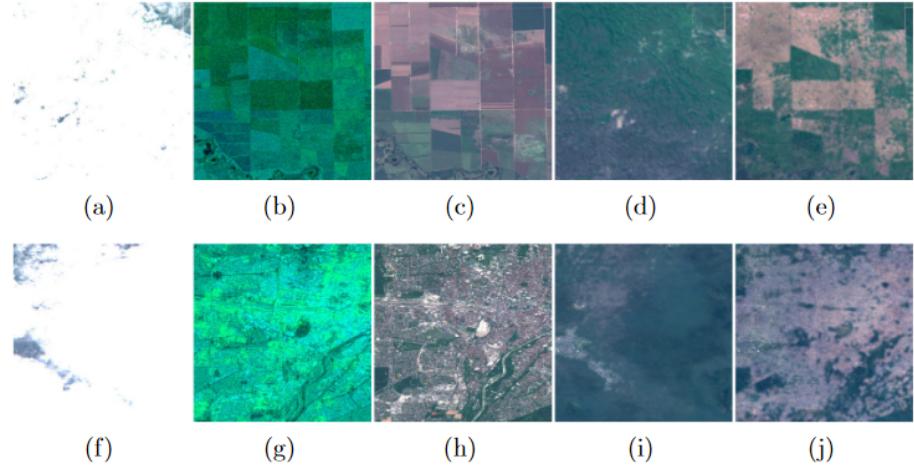


Figure 3: Examples of the influence of SAR data. Being (a, f) cloudy images, (b, g), SAR images, (c, h), the target cloudless images, (d, i), the synthesized images without SAR input data and (e, j) the predictions using SAR data as input.

Improvements have been found using Generative Adversarial Network (GAN) [4], which is a model architecture that belongs to the set of generative models. GAN is an unsupervised model made up of two neural networks: the generator and the discriminator. The idea is based on a game theoretic scenario in which the generator network must compete against an adversary. While the generator network produces samples, the aim of the discriminator is to distinguish between the real samples and the drawn by the generator. The discriminator is a binary classifier trying not to be fooled. The generator loss is calculated using the discriminator as a reference of how much far is from real images while the discriminator loss is calculated by how much accurate is discerning between the synthesized data and the real one. The standard function can be known as the min-max loss:

$$\begin{aligned} Loss_D(D) &= E_x[\log(D(x))] \\ Loss_G(G) &= E_y[\log(1 - D(G(y)))] \\ Loss_{GAN}(G, D) &= Loss_D(D) + Loss_G(G) \end{aligned}$$

During training, both networks constantly try to outsmart each other, in a zero-sum game. At some point of the training, the game may end up in a state that game theorists call a Nash equilibrium, when no player would be better off changing their own strategy, assuming

the other players do not change theirs. GANs can only reach one possible Nash equilibrium: when the generator produces so realistic images that the discriminator is forced to guess by 50 % probability that the image is real or not. Nevertheless, the training process not always can converge to that equilibrium. There are several factors that make the training hard to reach the desired state. For instance, there is a possibility that the discriminator always outsmarts the generator so that it can clearly distinguish between fake and real images. As it never fails, the generator is stuck trying to produce better images as it cannot learn from the errors of the discriminator. Possible solutions can be carried out such as making the discriminator less powerful, decreasing the learning rate or adding noise to the discriminator target. Another big obstacle is when the generator becomes less diverse, and it learns only to perfectly generate realistic images of a single class, so it forgets about the others. This is called mode collapse. At some point, the discriminator can learn how to beat the generator, but then, the latter is forced to do the same but in another class, cycling between classes never becoming good at any of them. A popular technique to avoid is experience replay, which consists in storing synthetic images at each iteration in a replay buffer. There is a lot of literature of obstacles and solutions to improve GAN training and it is still very active, as it is in its applications too. The tuning of hyper-parameters and the design of the model will be a key to pursue the Nash equilibrium. For instance, there is a variant called Conditional Generative Adversarial Network (cGAN). Traditionally, the generative network only produces the image from a random vector as an input, which is also called latent vector since it cannot be manipulated or with prior convictions of how will be. Unfortunately, this only allows to generate a random image from the domain of the latent space, which is hard to map to the generated images. However, cGAN can be trained so that both generator and discriminator models can be conditioned to some class labels or multi-dimensional vectors and produce synthetic images from a specific domain.

Actually, it is hard to find another GAN that can fit in cloud removal problem without being cGAN, since the model needs to be conditioned to the input image. In [5], they proposed a McGAN trained using RGB and Near Infra Red (NIR) cloud-free bands as input and synthetic RBG cloudy images as target. It is true that short wave bands are unaffected by cloud cover and using NIR images to guide to uncover the clouds of satellite imagery is great since NIR bands posses have higher penetration through fog than visible light bands. Nevertheless, synthesizing the target might not be realistic enough to feasibly

deploy the model in real-state. Regarding the experimental results, they have only showed qualitative metrics by comparing the ground-truth with cloud-free images, ones synthesized and ones real, accompanied both by the cloud obscure image and the cloud mask. As the NIR channel is not altered by the perlin noise, there is a huge difference about the effectiveness comparing the synthesized images with the real ones, as it can be seen in 4 and 5

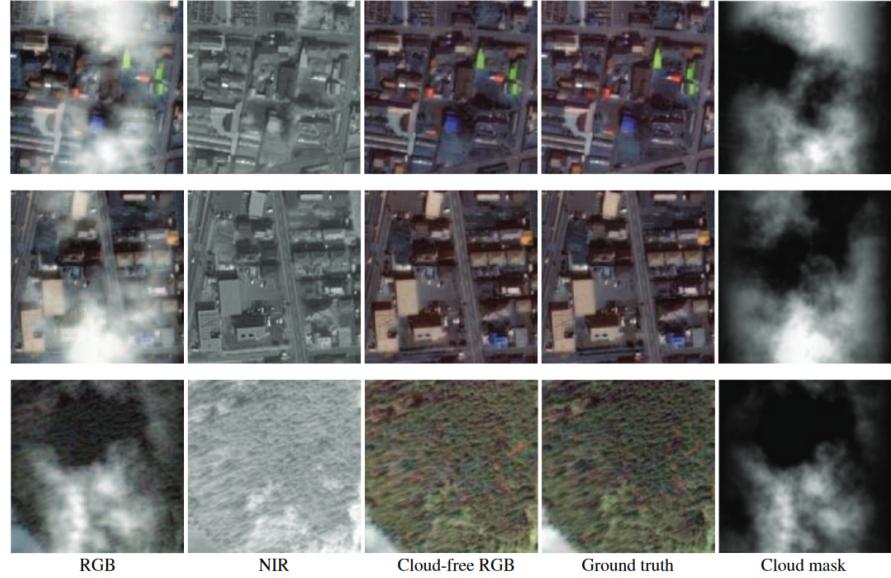


Figure 4: Prediction results by McGAN with the synthesized cloud images

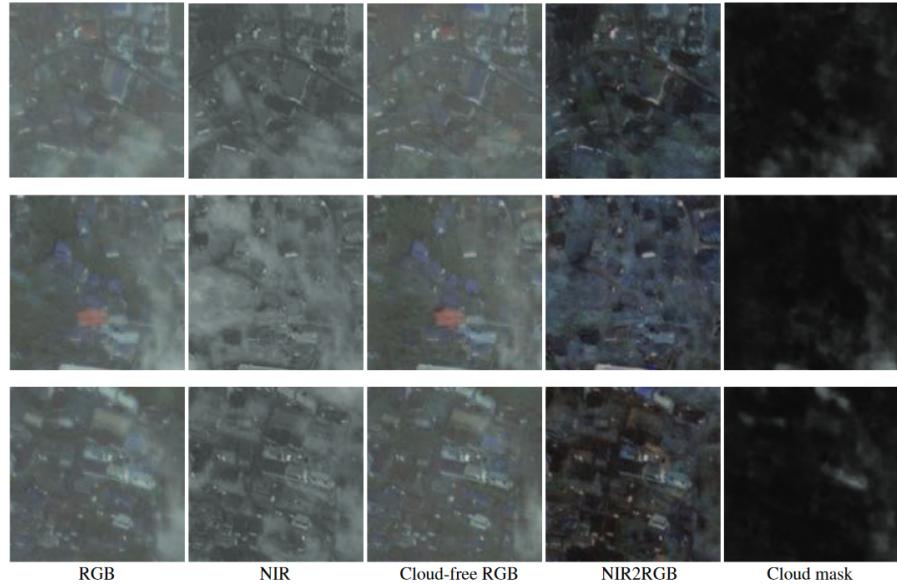


Figure 5: Prediction results by McGAN with the real cloud images

To overcome the need of synthesized data, [6] proposed two datasets and two models

respective to each of the datasets. Both neural networks have versions depending on including or excluding the NIR channel as an input. The first is based on Pix2Pix [7] model, a cGAN which performs very well in image-to-image translation tasks. The generator network of Pix2Pix uses a encoder-decoder architecture but with skip-connections to avoid loss of information of the bottleneck, similar to U-Net architecture [8]. The model is trained by a paired dataset of single-images. That means the dataset have a cloudy image as input and a real cloud-free image of the same zone. The generator tries to produce the image as real as the target ones, so it is compulsory to have both images. The second model proposed is a spatio-temporal cGAN trained by pairing clear images with three cloudy corresponding from diverse points in time. Instead of only trying the U-Net architecture as in the first model, they have proposed two designs for the generator:

- A branched ResNet generator. It has three encoder-decoder residual networks to learn feature maps from each image, then concatenating their output to learn from these features again and finally, concatenating the result through a final encoder-decoder to generate the new image.
- A branched U-Net generator. A slightly modification of the U-Net to allow multiple input images to encode them separately but decoding together and skip connections.

Moreover, they have also created two paired datasets to train the single-image models and the multi-temporal models respectively to overcome the need of synthesizing the images to make them cloud obscured.

Regarding experimental results, they have compared the single-image model with McGAN. On the other hand, they have compared the multi-temporal model with their different versions since it was the first of its kind. The results can be found in 6 and 7. All models from the same comparison were trained using the same datasets created by them. In both cases they have used Peak signal-to-noise ratio (PSNR) and Structural Similarity Index Metric (SSIM) as quantitative metrics and random figure comparison as qualitative metrics. While PSNR is a pixel-based metric capturing the difference between the corresponding pixels from two images, SSIM tracks similarity between visible structures and large-scale features in the images.

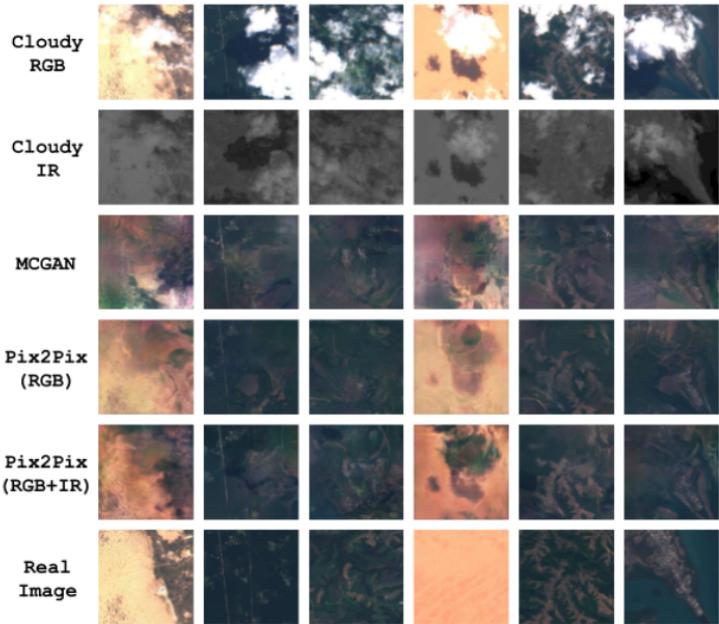


Figure 6: Comparison between the cloudy images, the outputs of McGAN and Pix2Pix-based models and the ground-truth image.

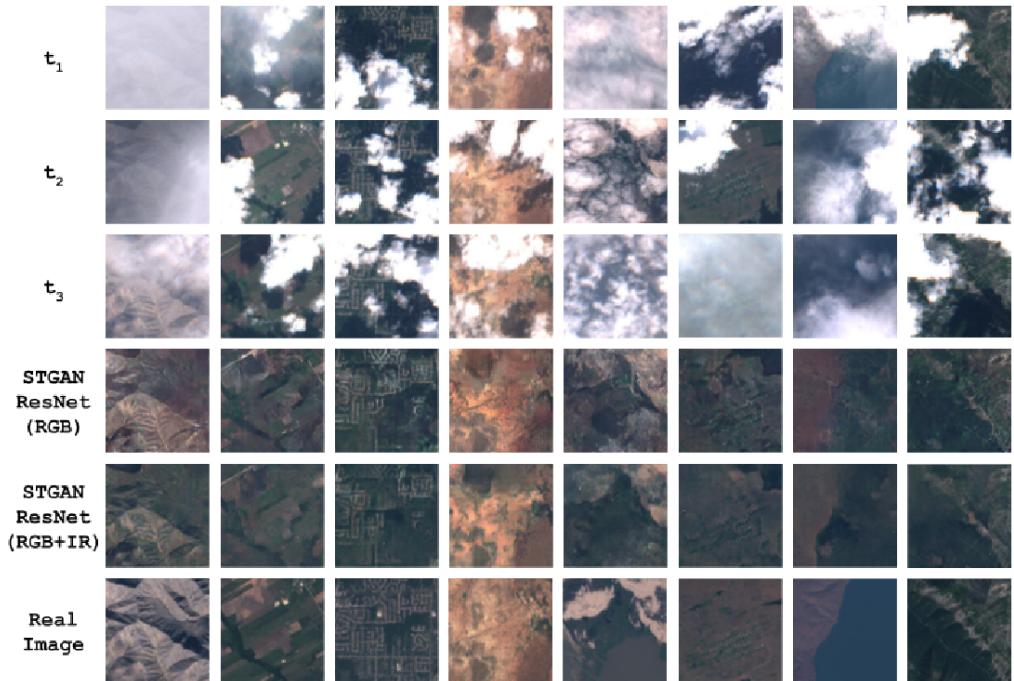


Figure 7: Comparison between cloudy images, the outputs of McGAN versions and the ground-truth image.

Table 1: PSNR and SSIM of the single-image and multi-temporal model versions .

	Validation Set		Test Set	
Model	PSNR	SSIM	PSNR	SSIM
Pix2Pix (RGB)	23.130	0.442	22.894	0.437
Pix2Pix (RGB + IR)	21.352	0.485	21.146	0.481
MCGAN (RGB + IR)	20.871	0.424	21.013	0.381
STGAN U-Net (RGB)	25.484	0.534	25.822	0.564
STGAN ResNet (RGB)	25.519	0.550	26.000	0.573
STGAN U-Net (RGB + IR)	25.142	0.651	25.388	0.661
STGAN ResNet (RGB + IR)	25.628	0.724	26.186	0.734
Raw Cloudy Images	7.926	0.389	8.289	0.422

To overcome the paired-data, Cloud-GAN [9] is a CycleGAN [10] which uses two generators (G_A and G_B) and two discriminators (D_A and D_B). CycleGANs can create new samples of output data, but also transforming the desired data to samples of input data. In essence, they learn to transform data from the two sources by the two generators respectively being also these sources the target. In other words, CycleGAN is a model that learns two data transformation functions between two domains. In our case, the generator G_A is generating cloud-free images from cloudy images while the generator G_B is turning the cloud-free images into cloudy. Hence, there is no need to train the model by paired cloudy-cloud-free imagery. The datasets for G_A and G_B are respectively:

$$X_A : \{x \mid x \text{ is a real cloud obscured image}\}$$

$$X_B : \{x \mid x \text{ is a real cloud-free image}\}$$

$$X = X_A \cup X_B$$

The synthesized data Y can be split into:

$$Y_A : \{y \mid y \text{ is a synthesized cloud obscured image}\}$$

$$Y_B : \{y \mid y \text{ is a synthesized cloud-free image}\}$$

$$Y = Y_A \cup Y_B$$

$$X_A \cap Y_A = \emptyset, \quad X_B \cap Y_B = \emptyset$$

The transformation functions from each generator can be expressed as:

$$G_A : \text{Generates } y \in Y_B \text{ from } z \in Z_A, \quad Z_A = X_A \cup Y_A$$

$$G_A : Z_A \longrightarrow Y_B$$

$$G_B : \text{Generates } y \in Y_A \text{ from } z \in Z_B, \quad Z_B = X_B \cup Y_B$$

$$G_B : Z_B \longrightarrow Y_A$$

It is well-known that each generator will learn its corresponding transformation function by minimizing the loss, which is calculated by how much the discriminator can discern between their produced data and the real data. On the other hand, the discriminator loss is how good it is to distinguish between the real and the synthesized data.

However, training a cyclic GAN using only the two network losses does not guarantee that cycle consistency is held. Cycle consistency means that given an input, it is desired the back-and-forth transformation $G_B(G_A(z_A)) = z'_A$ to output the the original input z_A . In other words, we want the composition of the transformation functions to be the identity function.

$$G_A \circ G_B = Id \iff G_B \circ G_A = Id$$

Thus, an additional cycle consistency loss is used to enforce this property. This loss is defined as the uniform norm between an input value z_A from the dataset Z_A and it's forward-cycle prediction $G_B(G_A(z_A))$ and the same for the data $z_B \in Z_B$, $G_A(G_B(z_B))$. The higher the difference, the more distant the predictions are from the original inputs. Ideally, our network would also minimize this loss by a weighting factor λ which is used to control the relevancy of this loss compared to the others.

$$Loss_{Cycle} = E[G_B(G_A(z_A)) - z_A] + E[G_A(G_B(z_B)) - z_B]$$

$$Loss_{CycleGAN} = Loss_{GAN}(G_A, D_A) + Loss_{GAN}(G_B, D_B) + \lambda \cdot Loss_{Cycle}$$

As mentioned before, in [9] they did novel deep learning model for cloud removal, since CloudGAN did not need paired-data and they had promising results removing thin clouds, small cloud patches and synthetic clouds. They have great PSNR in 8 although they could only do it for synthetic clouds as the data is not paired, so that there is no quantitative

results for real data, neither a comparison with other models proposed. In addition to that, the images to train and test seem to be not diverse enough since they only download satellite images over Paris region.

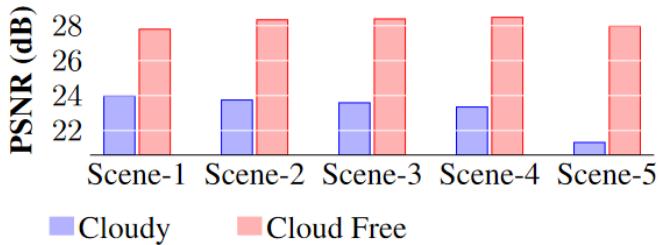


Figure 8: Failure cases in CloudGAN attempting to remove overly clouded images over-smoothening or completely failing.

Even though cyclic GANs work excellently to transforming images to different styles, they seemed to be not enough to create new shapes or providing geometric changes as well as they are not general enough providing unexpected results when the input data fed is quite different from the trained. This issue can be also found in Cloud-GAN as its removal cannot be performed well in overly clouded images 9.

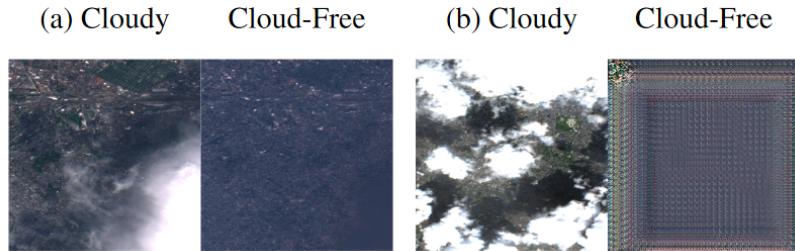


Figure 9: Failure cases in CloudGAN attempting to remove overly clouded images over-smoothening or completely failing.

Although CNN have worked very well for cloud removal, latest and disruptive state-of-the-art deep learning attention-based architectures [11] uncover new paths to achieve remarkable improvements and results. It has been demonstrated that transformers can excellently overcome challenges such Natural Language Processing (NLP) [12], Text-To-Image Generation [13] or Image Completion [14] with large datasets, great model size and enough compute. Recent contributions have been demonstrated that

2 Data

2.1 Academic Datasets

3 Model architecture

4 Experiments & results

5 Conclusions

6 Bibliography

References

- [1] Charis Lanaras, José Bioucas-Dias, Silvano Galliani, Emmanuel Baltsavias, and Konrad Schindler. Super-resolution of sentinel-2 images: Learning a globally applicable deep neural network. *ISPRS Journal of Photogrammetry and Remote Sensing*, 146:305–319, 2018.
- [2] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778, 2016.
- [3] Andrea Meraner, Patrick Ebel, Xiao Xiang Zhu, and Michael Schmitt. Cloud removal in sentinel-2 imagery using a deep residual neural network and sar-optical data fusion. *ISPRS Journal of Photogrammetry and Remote Sensing*, 166:333 – 346, 2020.
- [4] Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. Generative adversarial nets. In *Advances in neural information processing systems*, pages 2672–2680, 2014.
- [5] Kenji Enomoto, Ken Sakurada, Weimin Wang, Hiroshi Fukui, Masashi Matsuoka, Ryosuke Nakamura, and Nobuo Kawaguchi. Filmy cloud removal on satellite imagery with multispectral conditional generative adversarial nets. In *2017 IEEE Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, pages 1533–1541, 2017.

- [6] Vishnu Sarukkai, Anirudh Jain, Burak Uzkent, and Stefano Ermon. Cloud removal in satellite images using spatiotemporal generative networks. *arXiv preprint arXiv:1912.06838*, 2019.
- [7] Phillip Isola, Jun-Yan Zhu, Tinghui Zhou, and Alexei A Efros. Image-to-image translation with conditional adversarial networks. *CVPR*, 2017.
- [8] Olaf Ronneberger, Philipp Fischer, and Thomas Brox. U-net: Convolutional networks for biomedical image segmentation, 2015. cite arxiv:1505.04597Comment: conditionally accepted at MICCAI 2015.
- [9] Praveer Singh and Nikos Komodakis. Cloud-gan: Cloud removal for sentinel-2 imagery using a cyclic consistent generative adversarial networks. In *IGARSS 2018 - 2018 IEEE International Geoscience and Remote Sensing Symposium*, pages 1772–1775, 2018.
- [10] Jun-Yan Zhu, Taesung Park, Phillip Isola, and Alexei A Efros. Unpaired image-to-image translation using cycle-consistent adversarial networks. In *Computer Vision (ICCV), 2017 IEEE International Conference on*, 2017.
- [11] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. In Isabelle Guyon, Ulrike von Luxburg, Samy Bengio, Hanna M. Wallach, Rob Fergus, S. V. N. Vishwanathan, and Roman Garnett, editors, *Advances in Neural Information Processing Systems 30: Annual Conference on Neural Information Processing Systems 2017, 4-9 December 2017, Long Beach, CA, USA*, pages 6000–6010, 2017.
- [12] Tom B. Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel M. Ziegler, Jeffrey Wu, Clemens Winter, Christopher Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever, and Dario Amodei. Language models are few-shot learners, 2020. cite arxiv:2005.14165Comment: 40+32 pages.

- [13] Aditya Ramesh, Mikhail Pavlov, Gabriel Goh, Scott Gray, Chelsea Voss, Alec Radford, Mark Chen, and Ilya Sutskever. Zero-shot text-to-image generation. In Marina Meila and Tong Zhang, editors, *Proceedings of the 38th International Conference on Machine Learning*, volume 139 of *Proceedings of Machine Learning Research*, pages 8821–8831. PMLR, 18–24 Jul 2021.
- [14] Mark Chen, Alec Radford, Rewon Child, Jeffrey Wu, Heewoo Jun, David Luan, and Ilya Sutskever. Generative pretraining from pixels. In Hal Daumé III and Aarti Singh, editors, *Proceedings of the 37th International Conference on Machine Learning*, volume 119 of *Proceedings of Machine Learning Research*, pages 1691–1703. PMLR, 13–18 Jul 2020.