IGARSS 2021 - 2021 IEEE International Geoscience and Remote Sensing Symposium | 978-1-6654-0369-6/21/$31.00 ©2021 IEEE | DOI: 10.1109/IGARSS47720.2021.9554268

INTERNAL LEARNING FOR SEQUENCE-TO-SEQUENCE CLOUD REMOVAL VIA SYNTHETIC APERTURE RADAR PRIOR INFORMATION

Patrick Ebel[1], *Michael Schmitt*[2,3], *Xiao Xiang Zhu*[1,2]

[1]Data Science in Earth Observation(SiPEO), Technical University of Munich (TUM), Munich, Germany
[2]Remote Sensing Technology Institute (IMF), German Aerospace Center (DLR), Wessling, Germany
[3] Department of Geoinformatics, Munich University of Applied Sciences, Munich, Germany

ABSTRACT

Many observations acquired via optical satellites are polluted by cloud coverage, impeding a continuous and on-demand monitoring of the Earth. Recent advances in the field of cloud removal consider multi-temporal data to reconstruct pixels covered by clouds at a time point of interest. Yet, the limitation of preceding work is that information gets integrated over time, removing any temporal resolution from the declouded end products. In this work we consider a sequence-to-sequence approach, translating cloudy time series to a series of cloud-free multi-spectral images without the need of any external cloud-free data set. Our network is guided by synthetic aperture radar (SAR) information providing a strong prior for the reconstruction of cloud-covered information. We analyze the proposed method by visual inspection of predictions and in terms of error metrics to highlight its benefits. Finally, an ablation study is conducted in which the our network is compared against a baseline model and the effectiveness of the proposed SAR prior is demonstrated.

***Index Terms*—** synthetic aperture radar, optical imagery, cloud removal, time series, data fusion, deep learning

1. INTRODUCTION

On average over 60 % of the earth's total surface is covered by clouds [1], obstructing a continuous and seamless monitoring of our planet. With an increasing demand for a continuous observation of the environment, there is a need for methods that can reconstruct information of the Earth's surface from cloud-covered satellite image pixels. Previous work on cloud-removal from optical Sentinel-2 (S2) data utilized synthetic aperture radar data (SAR), as acquired via Sentinel-1 (S1) satellites, to process radar information not affected by clouds and inpaint cloud-covered information in a given optical image [2]. Complementary to the multi-sensor data fusion idea, recent models follow a multi-temporal approach that uses optical information recorded at a different acquisition time to inpaint cloudy pixels [3, 4]. The approach proposed in this work integrates multi-temporal and multi-modal information, reconstructing cloud-covered pixels in a time series of S2 imagery with the aid of a paired and co-registered time series of S1 observations. Furthermore, while preceding multi-temporal approaches have often integrated cloud-free information over the time dimension at the cost of loosing all temporal resolution our model allows for a sequence-to-sequence translation approach from cloudy to cloud-free multi-spectral observations, preserving temporal information.

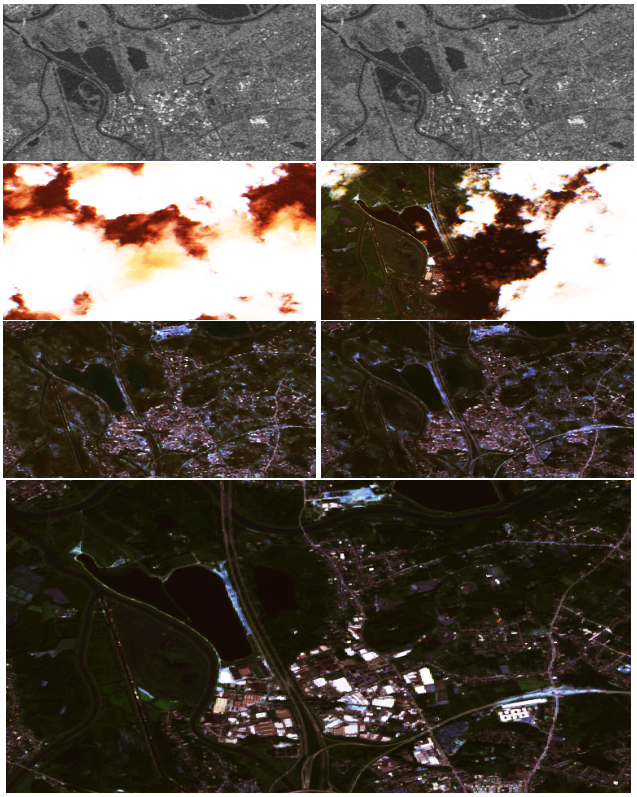

Fig. 1: Example observations and cloud-free predictions. Columns: Samples at two different time points. Rows: S1 data (in grayscale), cloudy S2 data (in RGB), predicted cloud-free $\hat{S}2$ data, reference cloud-free S2 data of a later point in time. The results highlight that our network is able to integrate multi-modal and multi-temporal information to predict a clear-view sequence of multi-spectral observations, even in the presence of heavy cloud coverage.

2. RELATED WORK

Most recent approaches to cloud removal for remote sensing are based on deep learning solutions. More specifically, contributions can be categorized based on either mono- or multi-temporal observations. Mono-temporal approaches process an observations just at the target time point and often utilize a complementary SAR observation [7]. Multi-temporal approaches [3, 4] are designed to process a time series of observations and integrate cloud-free information over time to impaint cloud-covered pixels at a given target date. While the one recovered image is de-clouded at a high fidelity the limitation of preceding multi-temporal approaches is in the loss of any temporal resolution.

3. METHODOLOGY

The network considered in this work is based on a 3D Encoder-Decoder architecture [5]. That is, the model consists of an encoder and a decoder part, arranged symmetrically in a U-Net like manner with skip connections between paired layers. The input is a set of multi-temporal $S1$ observations. The output is a set of cloud-free multi-temporal multi-spectral samples predicted via a model $G_{S1\rightarrow S2}$ and given by

$$\hat{S2} = G_{S1\rightarrow S2}(S1)$$

where the sequence S1 is of dimensions $[N \times 2 \times H \times W]$ and S2 is of dimensions $[N \times 13 \times H \times W]$, with N denoting the number of observations in a given time series, H and W are the spatial dimensions while 2 and 13 are the numbers of channels of SAR and multi-spectral optical observations, respectively. Furthermore, binary cloud and cloud shadow masks m of dimensions $[N \times 1 \times H \times W]$ are predicted on the $S2$ data via the method of [8]. To model relations across samples in time, convolution operators are 3-dimensional across both spatial and the temporal dimensions. The network is trained to minimize the loss $\mathcal{L}_{all}$ with

$$\begin{aligned}\mathcal{L}_2 &= ||S2 \cdot (1-m), \hat{S2} \cdot (1-m)||_2 \\ \mathcal{L}_{perc} &= ||VGG16(S2) \cdot (1-m), VGG16(\hat{S2}) \cdot (1-m)||_2 \\ \mathcal{L}_{all} &= \lambda_2 \mathcal{L}_2 + \lambda_{perc} \mathcal{L}_{perc},\end{aligned}$$

where $\lambda_2 = 1$ and $\lambda_{perc} = 0.01$ are hyper-parameters to linearly combine the individual losses within $\mathcal{L}_{all}$. $\mathcal{L}_2$ denotes a pixel-wise reconstruction loss evaluated over the cloud-free area, $\mathcal{L}_{perc}$ is a perceptual loss evaluated via an auxiliary pre-trained VGG16 network results in sharper predictions [9]. Importantly, the network is trained directly on the target sequence of images as in [6, 5], incorporating their information via the loss $\mathcal{L}_{all}$. That is, no external training data set is needed for the network to learn removing clouds on the target sequence $S2$. The network is trained to predict the given cloud-free pixels and effectively learns inpainting the cloud-obscurred information. This is thanks to the 3-dimensional convolution kernels modeling the domain's spatio-temporal regularities (i.e. cloud-free land surface) before overfitting to the noise (i.e. irregular cloud coverage). Pseudo-code outlining the internal learning approach followed in this study is provided in Algorithm 1 and further details are given in the seminal work of [6].

Algorithm 1 Internal Learning to Remove Clouds

procedure SEQ2SEQDECLOUDING($S1, S2, iterMax$)
  $G_{S1\rightarrow S2}$ = init. new NeuralNetwork()
  iterCount = 0
  while iterCount $<$ iterMax **do**
    $\hat{S2} = G_{S1\rightarrow S2}(S1)$
    $G_{S1\rightarrow S2}.backpropagate(\mathcal{L}_{all}(S2, \hat{S2}))$
    iterCount = iterCount + 1
  Return $\hat{S2}$

4. EXPERIMENTS AND ANALYSIS

Experiments are conducted on a novel data set of co-registered and paired Sentinel-1 and Sentinel-2 time series curated for this study. Each time series consists of $N = 30$ images sampled over a given ROI at subsequent points in time. The data set consists of 3 geospatially separate ROIs, with each ROI being of height $H = 384$ and width $W = 768$. Analogous to the preprocessing pipeline for a previous multi-modal cloud removal data set [7], all images in this study are value-clipped and then rescaled for every pixel to take values within $[0, 1]$. The modalities $S1$ and $S2$ are value-clipped within $[-25; 0]$ and $[0; 10000]$, respectively. For a given ROI the model is trained for 20 passes over batches of size $n = 5$ containing temporally adjacent observations, for 100 iterations each. Parameters are optimized via ADAM at a learning rate of 0.01 and hyperparameters including parameter $iterMax$ of Algorithm 1 as set in [5]. To quantitatively evaluate the performance of the model we implement an approach similar to the the data simulation method of [3]. Specifically, we synthesize 1 cloudy observation $S2_t$ from an originally cloudless observation $S2_t$ in the time series $S2$ at time point t by blending the formerly cloud-free image with clouds from most cloud-covered image in the time series. The cloud-removed prediction $\hat{S2}_t$ is then compared against $S2_t$ in order to get a measure of goodness of cloud removal. The cloud-coverage of the whole data set following the aforementioned target image generation is on average at 61.10 (± 38.32) $\%$. That is, a large portion of all optical observations, roughly corresponding to what has been observed empirically [1], is shrouded by clouds. Qualitative example outcomes for the proposed model and the described optimization procedure are illustrated in Fig. 1. The results underline that the considered model is capable of integrating multi-modal and multi-temporal information to predict a clear-view sequence of multi-spectral observations, even in the presence of intense cloud coverage. Complementarily, a quantitative

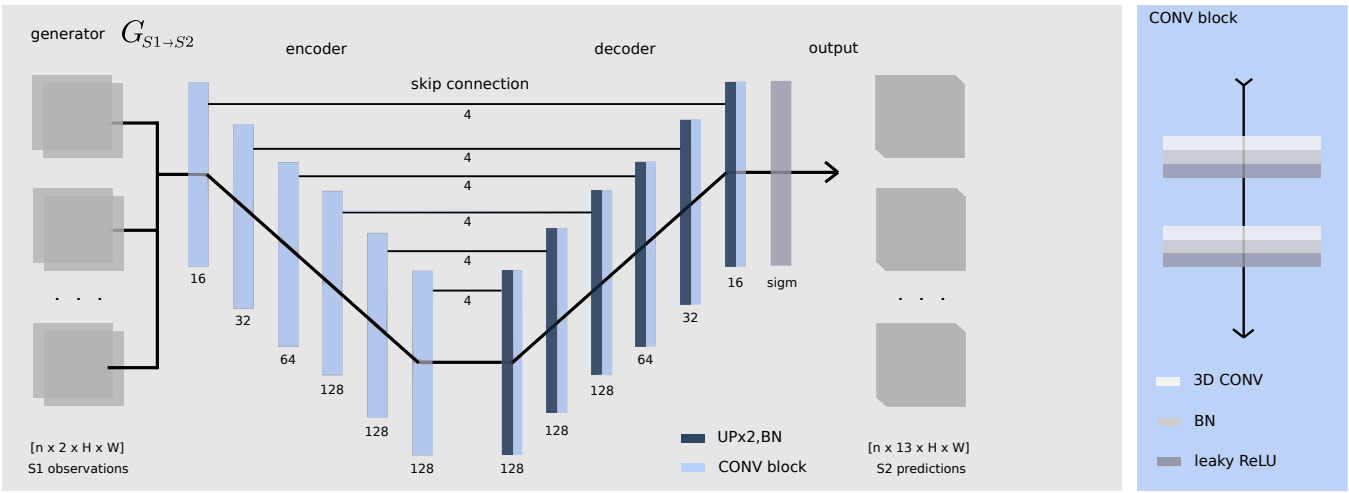

Fig. 2: An overview of the 3D Encoder-Decoder architecture $G_{S1\rightarrow S2}$ employed in our work. The network is based on the architecture of [5] and consists of encoder and decoder parts arranged symmetrically in a U-Net like manner with skip connections between paired layers. Input to the network is a batch of multi-temporal $S1$ observations. The output is a predicted batch of multi-temporal multi-spectral $S2$ observations. For the ablation study, Gaussian noise is used as an input as in [6, 5].

	NRMSE	PSNR	SSIM	SAM
full model	**0.27**	**11.59**	**0.51**	**27.73**
ablation model	0.30	11.43	0.49	28.13

Table 1: Quantitative evaluation of our proposed model in terms of root mean squared error (RSME), peak signal-to-noise ratio (PSNR), structural similarity (SSIM) and the Spectral Angle Mapper (SAM) [10] metric. Our multi-temporal network with SAR guidance outperforms the multi-temporal ablation model without prior SAR information.

analysis is conducted by evaluating the network in terms of root mean squared error (RSME), peak signal-to-noise ratio (PSNR), structural similarity (SSIM) and the Spectral Angle Mapper (SAM) [10] metric. The results are provided in Table 1. Finally, a comparative evaluation of the proposed model is conducted. For this purpose, the network is re-trained but without the S1 input sequence. Instead, the model receives white noise as input driving its predictions, as is standard in the literature on internal learning [6, 5]. This variant serves as an ablation model and its performance is reported in Table 1. In comparison, the full model including the SAR prior outperforms the ablation network on all considered metrics. Finally, exemplary outcomes comparing both models are provided in Fig. 3. The illustrations show that the model incorporating prior SAR information produces sharper predictions compared to the model without biasing, which is also prone to minor discoloring artifacts at times. The structural information provided by the SAR input provides a strong prior to the model, guiding it towards learning to reliably remove clouds in the cloudy input time series.

5. CONCLUSION

We proposed a novel multi-temporal approach to cloud removal in optical satellite data that allows for sequence-to-sequence mapping, preserving the temporal resolution of the original cloudy input time series. The network and the reconstruction of cloud-obscurred pixels are guided by prior information given in the form of SAR observations. Experimental analysis highlighted that the proposed model produces cloud-removed observations of high quality, even in the presence of many and heavy clouds. We evaluated our approach in an ablation study against the same architecture not receiving guidance via any SAR information and thereby demonstrated that the benefits of radar data persist, even when optical observations at multiple time points are available. As demonstrated, providing prior SAR information allows to improve on the goodness of predictions, highlighting the benefits that complementary SAR images can provide even in the presence of repeated optical measures. To summarize, the contribution of this work is: First, an advancement of preceding cloud removal methods by allowing sequence-to-sequence translation and preserving temporal resolution. Second, a facilitation to the remote sensing practitioner by the internal learning approach no longer requiring external (cloud-free) training data. Finally, a improvement of the existing methodology by achieving better predictions thanks including the SAR prior guiding the model. In future work we aim to further investigate the incorporation of multi-temporal and multi-modal information in the context of cloud removal.

Fig. 3: Illustrations on the effect of prior guidance via SAR information. Columns: SAR input to the SAR-conditioned model, cloud-free prediction of the model conditioned on Gaussian noise, cloud-free prediction of the model conditioned on SAR information, cloud-free observation as a reference image. The structural information provided by the SAR input provides a strong prior to the model, guiding it towards learning to remove clouds in the cloudy input time series.

6. REFERENCES

[1] Michael D. King, Steven Platnick, W. Paul Menzel, Steven A. Ackerman, and Paul A. Hubanks, "Spatial and temporal distribution of clouds observed by MODIS onboard the terra and aqua satellites," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 51, no. 7, pp. 3826–3852, Jul 2013.

[2] Patrick Ebel, Michael Schmitt, and Xiao Xiang Zhu, "Cloud removal in unpaired Sentinel-2 imagery using cycle-consistent GAN and SAR-optical data fusion," *IGARSS 2020 IEEE International Geoscience and Remote Sensing Symposium, 2020*, 2020.

[3] Muhammad Usman Rafique, Hunter Blanton, and Nathan Jacobs, "Weakly supervised fusion of multiple overhead images," in *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*. IEEE, 2019, pp. 1479–1486.

[4] Vishnu Sarukkai, Anirudh Jain, Burak Uzkent, and Stefano Ermon, "Cloud removal from satellite images using spatiotemporal generator networks," in *The IEEE Winter Conference on Applications of Computer Vision*, 2020, pp. 1796–1805.

[5] Haotian Zhang, Long Mai, Ning Xu, Zhaowen Wang, John Collomosse, and Hailin Jin, "An internal learning approach to video inpainting," in *Proceedings of the IEEE International Conference on Computer Vision*, 2019, pp. 2720–2729.

[6] Dmitry Ulyanov, Andrea Vedaldi, and Victor Lempitsky, "Deep image prior," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2018, pp. 9446–9454.

[7] Patrick Ebel, Andrea Meraner, Michael Schmitt, and Xiao Xiang Zhu, "Multisensor data fusion for cloud removal in global and all-season Sentinel-2 imagery," *IEEE Transactions on Geoscience and Remote Sensing*, 2020.

[8] Michael Schmitt, Lloyd H Hughes, Chunping Qiu, and Xiao Xiang Zhu, "Aggregating cloud-free Sentinel-2 images with Google Earth Engine.," *ISPRS Annals of Photogrammetry, Remote Sensing & Spatial Information Sciences*, vol. 4, 2019.

[9] Justin Johnson, Alexandre Alahi, and Li Fei-Fei, "Perceptual losses for real-time style transfer and super-resolution," in *European conference on computer vision*. Springer, 2016, pp. 694–711.

[10] Fred A Kruse, AB Lefkoff, JW Boardman, KB Heidebrecht, AT Shapiro, PJ Barloon, and AFH Goetz, "The spectral image processing system (sips)-interactive visualization and analysis of imaging spectrometer data," in *AIP Conference Proceedings*. American Institute of Physics, 1993, vol. 283, pp. 192–201.