A CONDITIONAL GENERATIVE ADVERSARIAL NETWORK TO FUSE SAR AND MULTISPECTRAL OPTICAL DATA FOR CLOUD REMOVAL FROM SENTINEL-2 IMAGES

Claas Grohnfeldt[1], *Michael Schmitt*[1], *Xiaoxiang Zhu*[1,2]

[1]Signal Processing in Earth Observation, Technical University of Munich (TUM), Munich, Germany
[2]Remote Sensing Technology Institute (IMF), German Aerospace Center (DLR), Wessling, Germany

ABSTRACT

In this paper, we present the first conditional generative adversarial network (cGAN) architecture that is specifically designed to fuse synthetic aperture radar (SAR) and optical multi-spectral (MS) image data to generate cloud- and haze-free MS optical data from a cloud-corrupted MS input and an auxiliary SAR image. Experiments on Sentinel-2 MS and Sentinel-1 SAR data confirm that our extended SAR-Opt-cGAN model utilizes the auxiliary SAR information to better reconstruct MS images than an equivalent model which uses the same architecture but only single-sensor MS data as input.

Index Terms— SAR, optical remote sensing, data fusion, deep learning, generative adversarial network (GAN), cloud-removal

1. INTRODUCTION

The Sentinel-1 and Sentinel-2 satellite missions have been providing global *synthetic aperture radar (SAR)* data and optical *multi-spectral (MS)* imagery, respectively, with high temporal and medium-to-high spatial resolution for some years now. In comparison to SAR observations, MS images contain rich spectral information and are readily interpretable by the human eye. However, they suffer from inevitable problems of spaceborne sensors operating in the optical wavelength range: Their measurements are strongly affected by the atmosphere. In particular, optical signals, as measured by MS sensors, cannot penetrate clouds whereas microwaves, as transmitted and received by radar instruments, can. Considering that the MODIS cloud mask showed that about 67% of the Earth's surface is covered by clouds on average [1], significant information gaps occur in MS acquisitions every day.

Due to that, techniques such as dehazing and cloud-removal of spaceborne remote sensing data acquired by optical sensors have long been important topics in the community. The most recent – and thus far most powerful – approaches for dehazing and cloud-removal are based on either SAR-optical image fusion (e.g. [2]) or data-driven machine learning procedures (e.g. [3, 4]).

Recently, many important remote sensing problems, including hyperspectral image classification, SAR and optical image interpretation, multi-modal data fusion, and 3D reconstruction, have been addressed successfully via deep learning [5]. A particularly dynamic sub-field within deep learning concerns the concept of *generative adversarial networks (GANs)*, for they allow to generate artificial data from seed information [6]. Among different GAN rationales, *conditional GANs (cGANs)* have attracted considerable interest in the remote sensing community, as they allow to generate desired artificial data based on a specified target output [7]. As an example related to the content of this paper, an approach proposed in [4] uses the cGAN concept to generate cloud-free RGB images from combined cloud-affected RGB and cloud-free *near infrared (NIR)* measurements. The problem with that approach is that the majority of types of clouds are impenetrable not only to visible but also to infrared light [8].

In this paper we overcome the above-outlined problem by advancing the idea of [4] in that we use co-registered SAR instead of NIR data as auxiliary input information about the ground covered by clouds. Furthermore, our implementation allows for MS input images with more than 3 bands, it incorporates deeper adversarial networks, and supports a large array of remote sensing data formats and bit-depths. By combining the generative deep learning and multi-modal data fusion rationales [9], we propose the first cGAN architecture that is designed to fuse SAR and optical MS image data in order to predict cloud- and haze-free MS images from cloud-corrupted MS measurements. Based on representative experiments, we demonstrate that additional long-wavelength SAR information and pixel-level data fusion yield superior dehazing results relative to an equivalent cGAN model that is designed and trained in an analogous manner, but merely with optical input data.

2. THE SAR-OPT-CGAN FRAMEWORK FOR OPTICAL IMAGE ENHANCEMENT

The SAR-Optical-cGAN (SAR-Opt-cGAN) architecture proposed in this paper builds upon the well-established cGAN architecture known as `pix2pix` [10]. We adapt and extend one of its established TensorFlow-based implementations [11] to

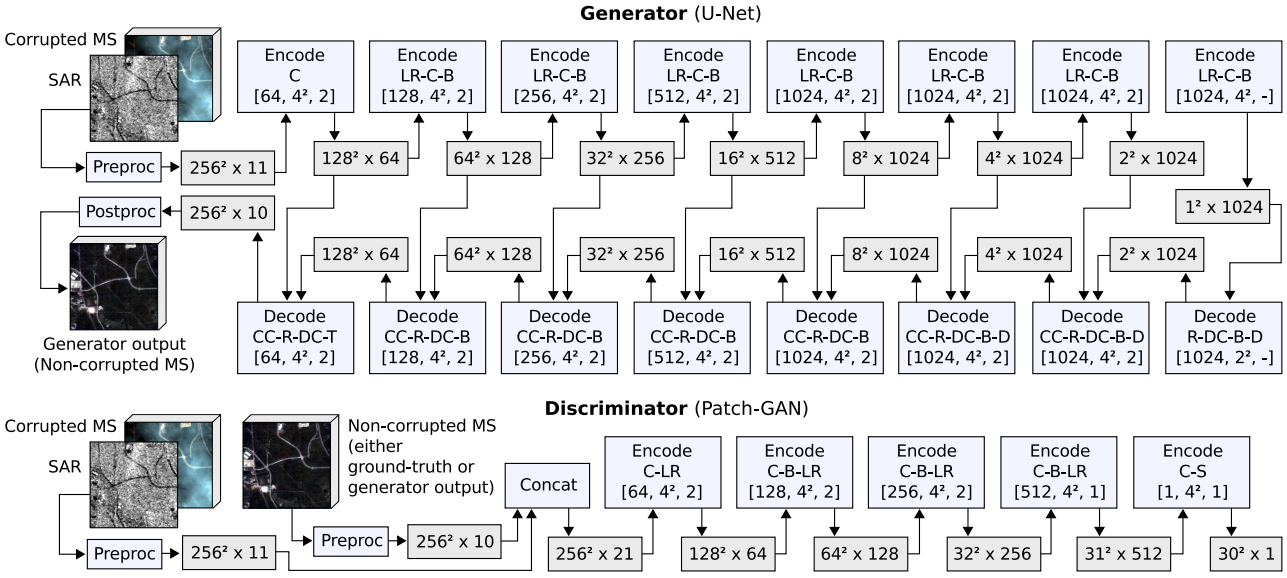

Fig. 1: Architectures of the generator (top) and discriminator (bottom) networks used in the SAR-Opt-cGAN. Acronyms in the encoder and decoder units are as follows: C=Convolution, R=ReLU, LR=Leaky ReLU, B=Batch Normalization, D=Dropout, DC=Deconvolution, CC=Concatenation, T=Tanh, S=Sigmoid. The three numbers in parentheses shown in all encoding and decoding layers indicate the number of filters, filter size and stride, respectively.

accommodate peculiarities of remote sensing data such as arbitrary numbers of spectral channels, multi-modal data, radiometric depths other than 8-bit, and unconventional data formats and intensity ranges. In order to avoid falsification of data-inherent information, we discard colorization and histogram adaption steps, which are conventionally applied in computer vision for visual enhancement purposes. Our adaption steps are described in the following.

2.1. Underlying cGAN Framework

The cGAN architecture used in `pix2pix` was designed for color and grayscale image-to-image translation purposes originally [10]. Based on the *generative adversarial network (GAN)* concept [6], its objective is to find a generator function $G : \mathcal{X} \mapsto \mathcal{Y}$ capable of producing an artificial image $\boldsymbol{Y} \in \mathcal{Y} \subseteq \mathbb{R}^{m \times n}$ that is indistinguishable from real data by an adversarially trained discriminator $D : \mathcal{X} \mapsto \{0, 1\}$. Early conditional GANs incorporate a random noise vector $\boldsymbol{z} \in \mathcal{Z} \subseteq \mathbb{R}^d$ in addition to an observed source image $\boldsymbol{X} \in \mathcal{X}$ as input for G. However, since z was found to be largely ignored by trained generators [10], it is discarded and randomness limited to dropout in this work. As opposed to G, D is trained to detect the generator's counterfeits. Concretely, the adversaries are computed via

$$G^* = \arg\min_G \max_D \mathcal{L}_{\text{cGAN}}(G, D) + \lambda \mathcal{L}_{L_1}(G), \quad (1)$$

where the traditional GAN loss function (noise excluded)

$$\mathcal{L}_{\text{cGAN}}(G, D) = \mathbb{E}_{\boldsymbol{X},\boldsymbol{Y}}[\log D(\boldsymbol{X}, \boldsymbol{Y})] + \mathbb{E}_{\boldsymbol{X}}[\log(1 - D(\boldsymbol{X}, G(\boldsymbol{X})))] \quad (2)$$

is augmented by an L_1-distance term

$$\mathcal{L}_{L_1}(G) = \mathbb{E}_{\boldsymbol{X},\boldsymbol{Y}}[\|\boldsymbol{Y} - G(\boldsymbol{X})\|_1]. \quad (3)$$

While the former term promotes solutions of G that produce images undetectable by D, the latter term ensures the output of G to be close to the ground truth in an L_1-sense during training. L_1- is favorable over L_2-based regularization in this context for it preserves details without entailing blurring effects.

2.2. Adaptations for SAR and Multi-spectral Data Fusion

Figure 1 displays the architectures of the generator and discriminator networks composing the SAR-Opt-cGAN. For the generator, we employ a U-Net architecture with skip connections similar to the one designed in [10]. In order to improve its spectral mapping capacity, we add filters to three encoder and decoder layers resulting in a bottleneck layer of 1024 feature maps. Our network is not restricted to 3-band RGB images, but capable of reading and writing an arbitrary number of MS channels and one or more auxiliary SAR images. Multiple SAR channels may be considered to account for different polarizations.

The SAR data should be co-registered and concatenated with the MS input. In addition to preprocessing procedures

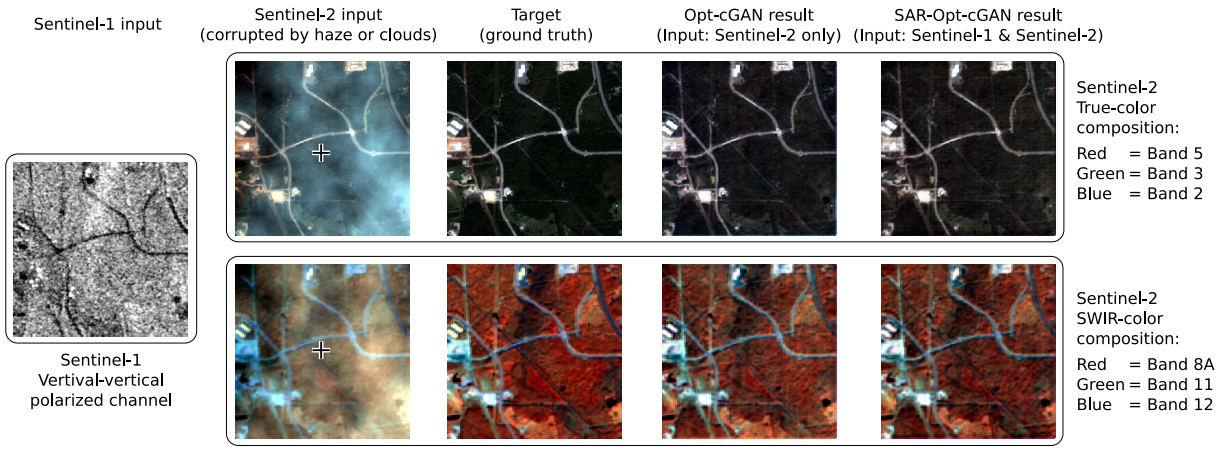

Fig. 2: Exemplary visual results for Sentinel-2 MS image dehazing based on the proposed SAR-Opt-cGAN architecture and fusion with Sentinel-1 SAR imagery.

typically performed on Earth observation data such as atmospheric and radiometric correction, filtering and geocoding, the pre- and postprocessing modules indicated in Fig. 1 involve a new data I/O interface, which utilizes the *geospatial data abstraction library (GDAL)* for versatility. Moreover, they perform data type conversions, domain scaling, and transformations from SAR to MS and back to the SAR data domain. Altogether, SAR-Opt-Net possesses close to 170 million trainable weights. In order to be able to measure the benefit of additional SAR information, we built and trained an analogous network, named Opt-cGAN, which ignores the SAR input.

3. EXPERIMENTS AND RESULTS

3.1. Test Dataset

We carry out experiments on a subset of the *SEN1-2* dataset [12], which consists of co-registered Sentinel-2 MS and Sentinel-1 SAR image patches of size 256×256 px with spatial sampling distance of 10 m for all input data. In contrast to the original *SEN1-2* dataset, we extracted patches from only 23 globally distributed scenes all of which were acquired in fall 2017. In contrast to [12], we use the full 16-bit MS information and 10 of 13 Sentinel-2 bands (discarding only the 60 m resolution bands) instead of just 8-bit RGB images. In total, we use 24720 Sentinel-1/Sentinel-2 patch-pairs for training, and 1117 patch-pairs for testing. In extension to what [4] proposed for synthetic corruption of RGB channels, we synthesized cloud- and haze-corrupted input data by adding Perlin noise to all Sentinel-2 channels in an adaptive manner.

3.2. Computational Setup

We used the following set of hyper-parameters throughout the experiments: $\lambda = 100$, learning rate $\alpha = 2^{-4}$, dropout rate $= 0.4$. We used the Adam optimizer with an exponential moving average decay of 0.99. We trained both SAR-Opt-cGAN and Opt-GAN on the training dataset for 5 epochs while continuously assessing the reconstruction results relative to the ground truth in terms *root mean square error (RMSE)* and *spectral angle mapper (SAM)*. We did the processing for both networks on NVIDIA TitanX GPUs.

3.3. Results

The assessment results of this study are summarized in Figs. 2-5. Figure 2 displays input, ground truth and output of SAR-Opt-cGAN and Opt-cGAN for one sample validation patch. Visually, both networks seem to succeed dehazing the corrupted Sentinel-2 input data. More distinct differences between both performances are revealed by the pixel-wise-measured RMSE and SAM error maps shown in Fig. 3, band-wise errors shown in Fig. 4 (left) and sample spectral profiles displayed in Fig. 4 (right). In this patch example, the additional SAR information brings a clear benefit to the dehazing, i.e., information reconstruction performance of our proposed SAR-Opt-cGAN. Moreover, the overall (average) RMSE and SAM values shown in Fig. 5, which are measured based on all training and validation patches and models that have been trained for 1, 2, 3, 4 and 5 epochs, respectively, indicate that this observation holds true in general.

4. SUMMARY AND CONCLUSION

In this paper, we presented a novel cGAN based approach to fusing SAR and optical imagery for declouding, dehazing and, more generally, enhancement purposes of optical remote sensing data. We adapted and further developed the well-known `pix2pix` architecture to what we call SAR-Opt-cGAN. This new network is capable of reading, writing and fusing multi-modal remote sensing data. Experiments on

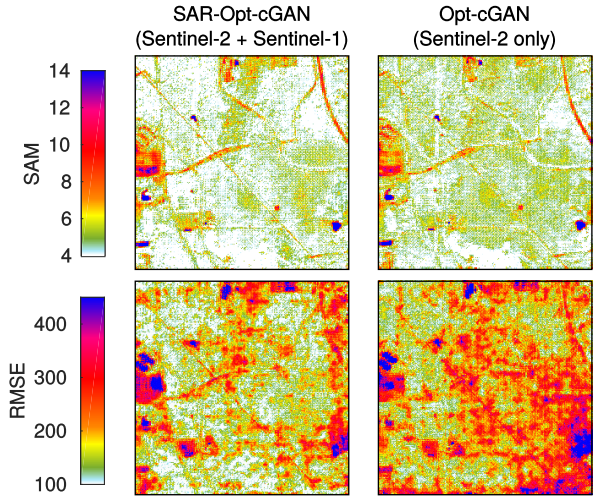

Fig. 3: Pixel-wise comparative quality assessment of the reconstructed sample patches shown in Fig. 2.

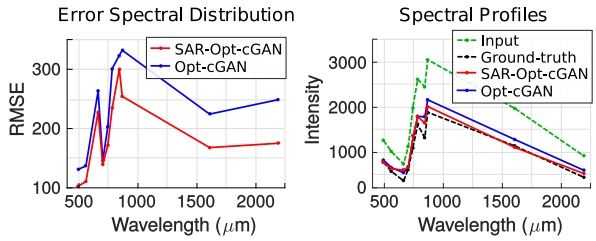

Fig. 4: Band-wise comparative quality assessment (left) and spectral profiles measured at the center pixel (right) of the sample patches shown in Fig. 2.

a large dataset of co-registered Sentinel-1 SAR and Sentinel-2 multi-spectral images demonstrate not only the cloud-removal capabilities of the proposed network but also the additional benefit in performance brought by auxiliary SAR data.

References

[1] M. D. King, S. Platnick, W. P. Menzel, S. A. Ackerman, and P. A. Hubanks, "Spatial and temporal distribution of clouds observed by MODIS onboard the Terra and Aqua satellites," *IEEE Trans. Geosci. Remote Sens.*, vol. 51, no. 7, pp. 3826–3852, 2013.

[2] B. Huang, Y. Li, X. Han, Y. Cui, W. Li, and R. Li, "Cloud removal from optical satellite imagery with SAR imagery using sparse representation," *IEEE Geosci. Remote Sens. Lett.*, vol. 12, no. 5, pp. 1046–1050, 2015.

[3] M. Xu, X. Jia, M. Pickering, and A. J. Plaza, "Cloud removal based on sparse representation via multitemporal dictionary learning," *IEEE Trans. Geosci. Remote Sens.*, vol. 54, no. 5, pp. 2998–3006, 2016.

[4] K. Enomoto, K. Sakurada, W. Wang, H. Fukui, M. Matsuoka, R. Nakamura, and N. Kawaguchi, "Filmy cloud removal on satellite imagery with multispectral conditional generative adversarial nets," in *Proc. CVPR Workshops*, Honolulu, HI, USA, 2017, pp. 1533–1541.

[5] X. Zhu, D. Tuia, L. Mou, G.-S. Xia, L. Zhang, F. Xu, and F. Fraundorfer, "Deep learning in remote sensing: A comprehensive review and list of resources," *IEEE Geosci. Remote Sens. Mag.*, vol. 5, no. 4, pp. 8–36,

[6] I. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville, and Y. Bengio, "Generative adversarial nets," in *Proc. NIPS*, Montreal, Canada, 2014, pp. 2672–2680.

[7] M. Mirza and S. Osindero, "Conditional generative adversarial nets," *ArXiv preprint arXiv:1411.1784*, 2014.

[8] G. Hunt, "Radiative properties of terrestrial clouds at visible and infra-red thermal window wavelengths," *Quart. J. R. Met. Soc.*, vol. 99, pp. 346–369, 1973.

[9] M. Schmitt and X. Zhu, "Data fusion and remote sensing – an ever-growing relationship," *IEEE Geosci. Remote Sens. Mag.*, vol. 4, no. 4, pp. 6–23, 2016.

[10] P. Isola, J. Y. Zhu, T. Zhou, and A. A. Efros, "Image-to-image translation with conditional adversarial networks," in *Proc. CVPR*, Honolulu, HI, USA, 2017, pp. 5967–5976.

[11] C. Hesse, *Tensorflow port of image-to-image translation with conditional adversarial nets*, `https://github.com/affinelayer/pix2pix-tensorflow`.

[12] M. Schmitt, L. H. Hughes, and X. X. Zhu, "The SEN1-2 dataset for deep learning in SAR-optical data fusion," *Submitted to ISPRS TCI Symposium*,

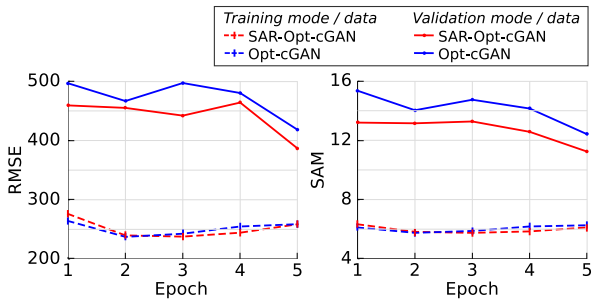

Fig. 5: Overall (average) assessment results measured for SAR-Opt-cGAN and Opt-cGAN on the full data set during training and testing over 5 epochs.