Deep Learning of Radiometrical and Geometrical Sar Distorsions for Image Modality translations

Antoine Bralet, Abdourrahmane Atto, Jocelyn Chanussot, Emmanuel Trouve

DEEP LEARNING OF RADIOMETRICAL AND GEOMETRICAL SAR DISTORSIONS FOR IMAGE MODALITY TRANSLATIONS

Antoine BRALET † *Abdourrahmane M. ATTO* † *Jocelyn CHANUSSOT* ⋆ *Emmanuel TROUVÉ* †

†LISTIC, Université Savoie Mont Blanc, 74940 Annecy, France
⋆Univ. Grenoble Alpes, CNRS, Grenoble INP, GIPSA-Lab, 38000 Grenoble, France

ABSTRACT

Multimodal approaches for Earth Observations suffer from both the lack of interpretability of SAR images and the high sensitivity to meteorological conditions of optical images. Translation methods were implemented to solve them for specific tasks and areas. But these implementations lack of generalizability as they do not include samples with challenging characteristics. Firstly, this paper sums up the main problems that a general SAR to optical image translator should overcome. Then, a SAR Distorted Image to optical translator Network (SARDINet) alternating knowledgeable channel-wise spatial convolutions and cross-channel convolutions is implemented. It aims at solving a problem of major concern in remote sensing: translating layover disturbed SAR images into disturbance-free optical ones. SARDINet is trained through a classical and an adversarial framework and compared to cGAN and cycleGAN from the literature. Experimental results prove that adversarial approaches are more qualitative but worsen quantitative results.

Index Terms— Remote Sensing, Deep learning, Translator, Multimodal, SAR-Optical

1. INTRODUCTION

Synthetic Aperture Radar (SAR) and optical images have been widely used for Earth Observation, not only independently but especially jointly for their complementary representations. In particular, multimodal approaches were applied for building damage mapping [1], land cover mapping [2, 3, 4, 5], change detection [6, 7] and coarse segmentation [8]. But, these applications are limited as optical images are sensitive to meteorology and SAR images are hardly interpretable for non-experts. To solve it, SAR to optical translations - or reverse - were implemented and applied for cloud removal in optical images [5], for registration improvements [9, 10] or for the completion of time series with missing records in time dependent approaches [11, 12]. This paper aims at revealing the major issues that a global SAR to optical translator should overcome due to SAR particularities. In addition, a SAR Distorted Image to optical translator Network (SARDINet) trained to solve geometrical distortions is implemented. Based on images from SpaceNet6 dataset [13] it aims at demonstrating that neural networks can tackle SAR translation limitations.

The paper first introduces related works for SAR to optical translation in Section 2. SAR limitations are developed in Section 3 and Section 4 describes the dataset used to train SARDINet to correct geometrical effects. The latter is introduced in Section 5 and conclusions are drawn in Section 6.

2. RELATED WORKS

The introduction of Generative Adversarial Networks (GAN) [14] and conditional GAN (cGAN) [15] inspired methods for translating an image from a modality to another. In particular, pix2pix [16] removed cGAN's dependency to a random vector and added L1 norm constraint to improve image fidelity. On the other hand, [17] introduced cycleGAN designed in two parallel cGANs learning respectively the direct and reciprocal translation based on the cycle consistency loss calculated between the input and the back and forth translated image.

cGAN and cycleGAN were of particular interest in remote sensing domain for SAR to optical and reverse translations. For instance, cGAN architectures [16] were used by [5] for cloud removal in optical images using SAR. [9] and [10] learnt optical to SAR translation to find ground control points for high precision registration. [11] replaced the U-Net [18] generator by a CNN with Residual Blocks [19] for SAR to optical translation and showed performance improvements with an adversarial training. [20] introduced two discriminators to respectively improve image fidelity and global brightness, contrast and colors. [21] set in place an additional constraint based on Structural Similarity. [22] modified the architecture to enhance contour sharpness and improve texture generation. They also calculated a chromatic loss for color fidelity improvements. [23] evaluated image quality on Discrete Cosine Transforms and on features extracted by an aside network to improve spectral and feature consistencies.

For unsupervised trainings, most of networks are based on cycleGAN architecture [17]. Authors of [8] demonstrated the possibility to transfer SAR to optical and reverse for urban areas segmentation on SAR images. [24] improved results by including Residual Blocks [19] in the generator. Similarly,

authors from [25] introduced cascaded-residual connections between the input and each U-Net [18] deconvolution stage to reach a more qualitative high resolution translated image.

Several works intended to improve translation results using images from the past. In [11], a cGAN like network was designed to take as input a past SAR-optical pair and the SAR image to be translated. They showed significant improvements in comparison with their mono-temporal equivalent. [12] used an additional encoder network to extract features from past images. These features acted as masks used to condition the behaviour of the generator encoder to focus on areas of interest. They also shown that an adversarial training increased qualitative results but decreased quantitative ones.

Indeed, adversarial trainings are not necessary to get strong results. Given a pre- and a post-event image from different modalities, [6] used reciprocal translation to detect changes (*e.g.* floods, forest fires,...). They designed a cyclic but non adversarial network and forced latent spaces alignment with a specific loss. An improved version [7] described Ace-Net and included a discriminator at latent space level. Another representation learning framework has been detailed in [26] for a classification task. They learnt a common representation by extracting modality specific features with encoders and projecting one feature space to the other using an additional network.

3. CHALLENGING SAR IMAGE PROPERTIES

The lack of interpretability of SAR images and the high sensitivity to meteorological conditions of optical images encourage to focus on SAR to optical translation. This section discusses the main characteristics of SAR imagery that can lead to intricacy in a large-scale translation framework.

3.1. Sensor Characteristics

Whether SAR images are obtained by satellites or airplanes, both are dependent of the characteristics of the sensor on board. The wavelength used for the acquisition modulates the interpretation as a same area is different on X-band or C-band images (resp. *e.g.* TerraSAR-X and Sentinel-1). The pulse bandwidth and the height of the sensor affect image resolution. The polarization used for emission or reception (Horizontal or Vertical) also reveals different properties of the area of interest. Thus, depending on sensor characteristics, an area can be imaged differently but should be understood as unique by a general translator framework. Multi-scale [22], domain adaptation [27] and representation learning [6, 7, 26] could be further studied to overcome these issues.

3.2. Geometry of acquisition

Depending on the path followed by the sensor (*orbit* for satellites), its altitude and its direction (ascending/descending), a scene can be imaged with a different local incidence angle (*i.e.* between the radar line of sight and the vertical at the pixel location). The wavelength, the resolution and this incidence angle have an impact on the backscattered signal and especially on the *speckle* noise (see Section 3.3). In addition, range sampling of SAR images along the line of sight is responsible for geometrical distortions: depending on the slope and incidence angles, areas might be *dilated* or *compressed* along the range direction [28]. Furthermore, slopes oriented towards the sensor appear as brighter areas and the *fold-over* (or *layover*) disturbance occurs when the slope angle is greater than the incidence angle. On the opposite slopes (oriented backwards the sensor) radar shadows may occur in areas which are not reached by radar waves. Both phenomena happen when an object is big enough (e.g. buildings or mountains): the top of the object is imaged before its base resulting in a shift and superposition artefact (layover) whereas the non-illuminated distances result in dark areas (shadow), as visible in SAR images of Figure 2.

Previous works are rarely influenced by geometrical distortions since flat areas are not affected by these phenomena. But, in the perspective of a general translator, a dataset containing all these challenging perturbations should be acquired for a relevant training. Section 5 describes the proposed network designed to be robust to geometrical distortions.

3.3. Spatiotemporality

The environment itself can be limiting as the sensor catches not only the target backscattered signal but also echoes of the surrounding distributed targets creating the so-called *speckle* noise. Its statistical properties are affected by all sensor and geometry characteristics aforementioned as echoes are linked to object reflection ability with respect to the recording configuration used. Finally, for time series studies, the delay between two acquisitions is a major constraint. The shorter satellite revisit time for now is 6 days for Sentinel-1 (SAR) and 5 days for Sentinel-2 (optical) delaying the ground truth correspondence from hours to three days. Depending on the application and on the imaged scene the delay might not be short enough. However, while free releases of satellites carrying both optical and SAR sensors are unavailable, this temporal constraint cannot be overcome.

4. DATASET USED

The following study is based on the SpaceNet6 multi-sensor dataset described in [13] and composed by SAR and RGBNIR images available at Kaggle repository. SAR data (provided by Capella Space and Metasensing) were acquired in X-band by a North- and South-facing airborne sensor over Rotterdam on August 4^{th}, 23^{rd} and 24^{th}, 2019 with a relative off-nadir look angle ranging from 53.4° to 56.6°. Intensity images are available at a half-meter spatial resolution in full polarization

(HH, HV, VH, and VV). The optical data (provided by Maxar Worldview-2 satellite) are composed by four pan-sharpened bands corresponding to Blue, Green, Red and Near Infrared (NIR). These images were acquired on August 31^{st}, 2019 at 10:44 am with an off-nadir look angle of 18.4°.

The dataset is further preprocessed to coregister SAR and RGBNIR images, remove mono-modal views and crop images to fit 450x900px resulting in 3338 multimodal matching pairs. For this work, a sub-dataset available at IEEE Dataport is extracted. It contains 333 images recorded over vertical cylindrical tanks which create geometrical distortions (both layover and shadow phenomena). To take the strong dynamic of SAR images into account, each channel i is thresholded between $[\mu_i - 3\sigma_i, \mu_i + 3\sigma_i]$ based on its mean and standard deviation and normalized between 0 and 1.

5. SAR TO OPTICAL TRANSLATOR

This Section addresses a deep learning disintrication solution based on a SAR Disturbed Image to optical translator Network (SARDINet) trained on the challenging sub-dataset. Its architecture is first introduced before presenting the results for classical and adversarial trainings.

5.1. Architecture

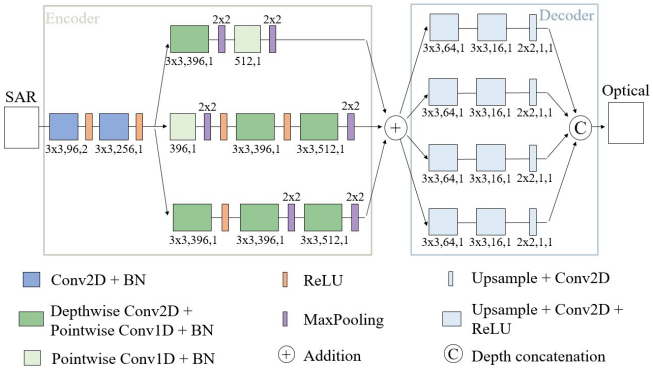

Fig. 1: Proposed SARDINet architecture. The encoder is divided in three branches for a multi-scale feature extraction and the decoder reconstructs optical channels with four independent branches.

The architecture of SARDINet visible in Figure 1 consists in three main steps : an input conditioning, a multi-scale feature extractor and a channel-independent reconstruction.

Input conditioning: This encoder first step is used to take full advantage of the early fusion of the polarizations. Successive 2D convolutions with Rectified Linear Unit (ReLU) activations are implemented to extract the most relevant cross-channel spatial features and to filter disturbing patterns.

Multi-scale feature extractor: The feature extractor is divided in three independent branches to extract global, intermediate and fine features. Each branch is composed of a succession of depthwise and pointwise convolutions (dark green blocks in Figure 1) following Xception architecture [29]. Depthwise convolutions extract channel specific spatial features and pointwise convolutions focus on cross-channel features. The scale of extracted features depends on the depth of the branch and the subsampling positions - ensured by 2x2 Max Pooling. Features are finally summed to obtain the latent features of the input image.

Channel-independent reconstructions: The decoder is composed of four independent branches for channel specific reconstruction. Each one is composed of three stages of 2x2 spatial upsampling and 2D convolutions demonstrating better results than transpose convolutions. The resulting images are concatenated to obtain the final optical image.

5.2. Experimental results

5.2.1. Training parameters

Results described in Section 5.2.2 are obtained with the following configuration. The dataset is divided in 80% training, 10% evaluation and 10% testing - *i.e.* respectively 266, 33 and 33 images. Mean squared error is calculated as loss function and an Adam optimizer with a learning rate of 1.10^{-3} is set for weights optimization. For computing time and memory issues and to increase dataset variability, five 200x200px images are randomly extracted from each training at each epoch which increases the training dataset to 1330 variable images. The training is performed with a batch size of 32 on a NVIDIA RTX A3000 GPU and computes 100 epochs within an hour and requires 3GB of memory.

SARDINet is trained for 200 epochs for fair comparison with pix2pix [16] and cycleGAN [17] available on Github. The comparisons are based on four metrics : RMSE, MAE and Structural SIMilarity (SSIM) with a window size of 3 and 11. Those metrics are only calculated on RGB channels as state of the art methods are limited to three input/output channels. To measure the impact of an adversarial framework, an Adversarial (Adv-) SARDINet is trained from scratch with a discriminator designed as three stages of 2D convolutions ending by a fully connected and a Softmax layer.

5.2.2. Results

Table 1 compares SARDINet variants with state of the art on images of size 200x200. Among experienced methods, Table 1 demonstrates that the non-adversarial SARDINet reaches the best metric results. In contrast, Figure 2 shows that resulting images are blurrier than adversarial approaches but is better structured which explains the stronger SSIM-3 results by about 0.06. Similar conclusions can be extracted from the comparison with Adv-SARDINet as SARDINet quantitative results are 0.01 and 0.016 better in MAE and RMSE and 0.05 better in terms of SSIM. But when it comes to qualitative results, adversarial ones are less blurry despite bright artefacts

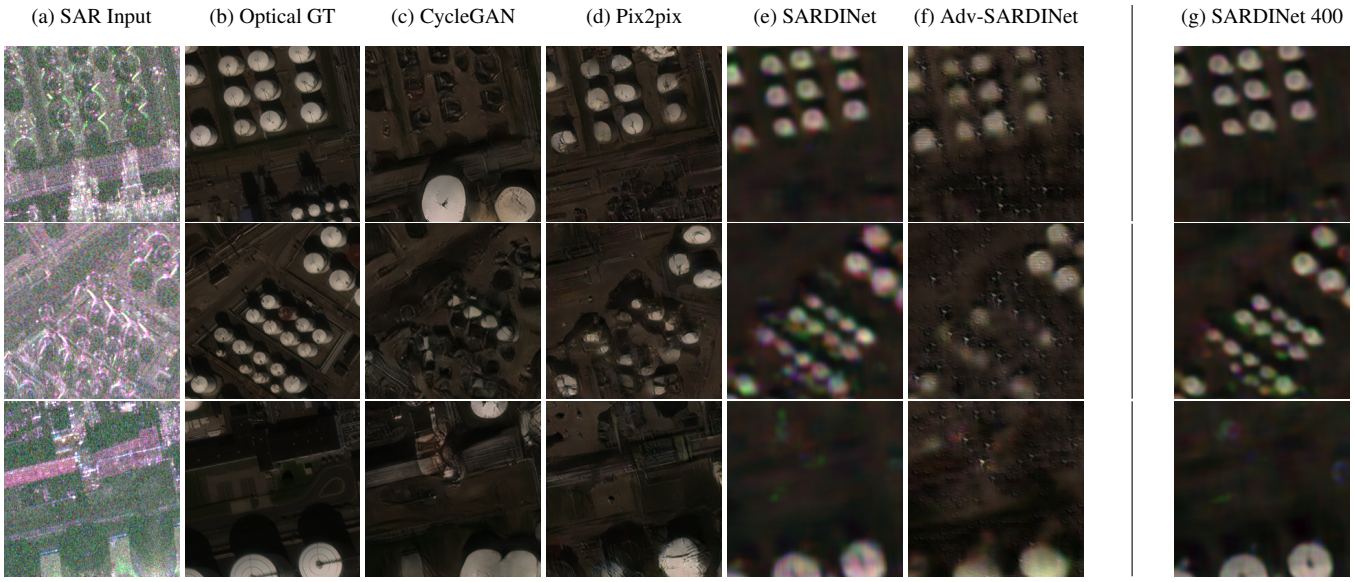

Fig. 2: Qualitative results comparisons of our networks with the state of the art for three testing images. SAR images channels are displayed as (VV, $\frac{1}{2}$(VH+HV), HH) and RGB for optical images.

which appear on the ground. Depending on the application, one may choose classical or adversarial training. But for most of remote sensing applications, accurate results are privileged against good-looking results denoting SARDINet relevancy.

Furthermore, training SARDINet for 200 more epochs with a learning rate of 1.10^{-4} demonstrates a decrease of about 0.06 in terms of MAE and RMSE and a SSIM increase of 0.01 and reach more qualitative results (last column of Figure 2). It means that the network does not suffer from overfitting and can reach better results with a longer training.

The striking point with the current approach is its structural similarity fidelity. Indeed, even if adversarial approaches look better in terms of color transitions, the shapes are deformed due to the geometrical distortions in the input SAR images - a common SAR difficulty which SARDINet is able to overcome. Further experiments will be conducted to confront these conclusions to larger datasets.

6. CONCLUSIONS AND PERSPECTIVES

The paper explored image modality translation for SAR layover decompression issue. The network implemented successfully de-distorted SAR characteristics that are so far considered as unrecoverable with standard SAR processing techniques. Results were quantitatively better than the adversarial frameworks and could still significantly be enhanced. Improvements can be reached not only with a longer training or a further hyperparameter optimization but also by training on the initial 450x900 images - at the expense of an increased processing cost. This study is promising for the synthesis of a general SAR to optical translator. The generalization requires wider and more challenging training datasets to be relevant.

Network	RMSE	MAE	SSIM-3	SSIM-11
CycleGAN [17]	0.163	0.102	0.780	0.800
Pix2pix [16]	0.155	0.098	0.79	**0.814**
SARDINet	**0.080**	**0.053**	**0.850**	0.813
Adv-SARDINet	0.096	0.063	0.802	0.762
SARDINet 400	0.074	0.046	0.860	0.825

Table 1: Quantitative comparison of our methods with the state of the art. Best 200 epochs results are displayed in bold.

7. ACKNOWLEDGMENT

This work is supported by the region Auvergne-Rhône-Alpes (AURA, France) through the project IATOAURA.

8. REFERENCES

[1] B. Adriano, N. Yokoya, J. Xia, et al., "Learning from multimodal and multitemporal earth observation data for building damage mapping," *ISPRS J. Photogramm. Remote Sens.*, vol. 175, pp. 132–143, 2020.

[2] D. Ienco, R. Interdonato, R. Gaetano, et al., "Combining Sentinel-1 and Sentinel-2 satellite image time series for land cover mapping via a multi-source deep learning architecture," *ISPRS J. Photogramm. Remote Sens.*, vol. 158, pp. 11–22, 2019.

[3] N. Kussul, M. Lavreniuk, S. Skakun, and A. Shelestov, "Deep learning classification of land cover and crop types using remote sensing data," *IEEE Geosci. Remote Sens. Lett.*, vol. 14, no. 5, pp. 778–782, 2017.

[4] N. Yokoya, P. Ghamisi, and J. Xia, "Multimodal, multitemporal, and multisource global data fusion for local climate zones classification based on ensemble learning," in *2017 IEEE IGARSS*, 2017, pp. 1197–1200.

[5] J. D. Bermudez, P. N. Happ, D. a. B. Oliveira, et al., "SAR to optical image synthesis for cloud removal with generative adversarial networks," in *ISPRS Ann. Photogramm. Remote Sens. Spat. Inf. Sci.* 2018, vol. IV-1, pp. 5–11, Copernicus GmbH.

[6] L. T. Luppino, M. A. Hansen, M. Kampffmeyer, et al., "Code-aligned autoencoders for unsupervised change detection in multimodal remote sensing images," *arXiv preprint arXiv:2004.07011*, 2020.

[7] L. T. Luppino, M. Kampffmeyer, F. M. Bianchi, et al., "Deep image translation with an affinity-based change prior for unsupervised multimodal change detection," *IEEE Trans. Geosci. Remote Sens.*, pp. 1–22, 2021.

[8] L. Liu and B. Lei, "Can SAR images and optical images transfer with each other?," in *IGARSS 2018 - 2018 IEEE Int. Geosci. Remote Sens.*, 2018, pp. 7019–7022, ISSN: 2153-7003.

[9] N. Merkle, P. Fischer, S. Auer, and R. Müller, "On the possibility of conditional adversarial networks for multi-sensor image matching," in *2017 IEEE IGARSS*, 2017, pp. 2633–2636.

[10] N. Merkle, S. Auer, R. Muller, et al., "Exploring the potential of conditional adversarial networks for optical and SAR image matching," *IEEE J. Sel. Top. Appl. Earth Obs. Remote Sens.*, vol. 11, no. 6, pp. 1811–1820, 2018.

[11] W. He and N. Yokoya, "Multi-temporal Sentinel-1 and -2 data fusion for optical image simulation," *ISPRS Int. J. Geo-Inf.*, vol. 7, no. 10, pp. 389, 2018.

[12] X. Liu, D. Hong, J. Chanussot, et al., "Modality translation in remote sensing time series," *IEEE Trans. Geosci. Remote Sens.*, pp. 1–14, 2021.

[13] J. Shermeyer, D. Hogan, J. Brown, et al., "SpaceNet 6: Multi-sensor all weather mapping dataset," in *2020 IEEE/CVF Conference on CVPRW*, 2020, pp. 768–777.

[14] I. Goodfellow, J. Pouget-Abadie, M. Mirza, et al., "Generative adversarial nets," *Adv. Neural Inf. Process. Syst.*, p. 9, 2014.

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

[16] P. Isola, J.-Y. Zhu, T. Zhou, and A. A. Efros, "Image-to-image translation with conditional adversarial networks," in *Proceedings of the IEEE CVPR*, 2016, pp. 1125–1134.

[17] J.-Y. Zhu, T. Park, P. Isola, and A. A. Efros, "Unpaired image-to-image translation using cycle-consistent adversarial networks," in *Proceedings of the IEEE ICCV*, 2017, pp. 2223–2232.

[18] O. Ronneberger, P. Fischer, and T. Brox, "U-Net: Convolutional networks for biomedical image segmentation," in *MICCAI 2015*. 2015, pp. 234–241, Springer International Publishing.

[19] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proceedings of the IEEE CVPR*, 2016, pp. 770–778.

[20] Y. Mourchid, M. Donias, and Y. Berthoumieu, "Dual color-image discriminators adversarial networks for generating artificial-SAR colorized images from Sentinel-1 images," in *MACLEAN (ECML/PKDD 2020)*, 2020.

[21] Y. Li, R. Fu, X. Meng, et al., "A SAR-to-optical image translation method based on conditional generation adversarial network (cGAN)," *IEEE Access*, vol. 8, pp. 60338–60343, 2020.

[22] X. Yang, J. Zhao, Z. Wei, et al., "SAR-to-optical image translation based on improved cGAN," *Pattern Recognition*, vol. 121, pp. 108208, 2022.

[23] J. Zhang, J. Zhou, and X. Lu, "Feature-guided SAR-to-optical image translation," *IEEE Access*, vol. 8, pp. 70925–70937, 2020.

[24] M. Fuentes Reyes, S. Auer, N. Merkle, C. Henry, and M. Schmitt, "SAR-to-optical image translation based on conditional generative adversarial networks—optimization, opportunities and limits," *Remote Sens.*, vol. 11, pp. 2067, 2019.

[25] S. Fu, F. Xu, and Y.-Q. Jin, "Reciprocal translation between SAR and optical remote sensing images with cascaded-residual adversarial networks," *Sci. China Inf. Sci.*, vol. 64, no. 2, pp. 1–15, 2021.

[26] P. Jain, B. Phelan, and R. Ross, "Multi-modal self-supervised representation learning for earth observation," *IEEE IGARSS*, pp. 3241–3244, 2021.

[27] Y. Ganin and V. Lempitsky, "Unsupervised domain adaptation by backpropagation," in *ICML*, 2015, pp. 1180–1189.

[28] S. Kaushik, Y. Yan, L. Ravanel, et al., "Visibility analysis of glaciers on steep slopes in the european alps using terrasar-x/paz data," in *2021 IEEE IGARSS*. IEEE, 2021, pp. 5505–5508.

[29] F. Chollet, "Xception: Deep learning with depthwise separable convolutions," in *Proceedings of the IEEE CVPR*, 2017, pp. 1251–1258.