DEM-assisted neural network for SAR-to-optical image translation

Antoine Bralet, Trong Nghia Ngo, Emmanuel Trouvé, Jocelyn Chanussot, Abdourrahmane Atto

DEM-ASSISTED NEURAL NETWORK FOR SAR-TO-OPTICAL IMAGE TRANSLATION

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

†Université Savoie Mont Blanc, LISTIC, 74940 Annecy, France
⋆Univ. Grenoble Alpes, Inria, CNRS, Grenoble INP, LJK, Grenoble, 38000, France.

ABSTRACT

SAR-to-optical remote sensing modality translator neural networks are mainly trained on flat areas preventing their use to detect gravitational movements as landslides in steep sloped areas. In this paper, we first propose a new SAR-DEM-optical dataset in mountainous regions to improve performances of SAR-to-optical image translators under these extreme conditions. Then we upgrade SARDINet (SAR Distorted Image translator Network) model previously developped for urban areas, to take a Digital Elevation Model (DEM) together with the SAR image as input and perform translation in natural mountainous environment. Multiple fusion strategies are explored to merge efficiently SAR and DEM images: late fusion, early fusion but also an intermediate fusion based on balanced separable convolutions. These approaches are compared to the original SARDINet and two standard adversarial networks - Pix2pix and CycleGAN - showing improvements in distorted regions.

***Index Terms*—** Multimodal Images, SAR, DEM, Image Translation, Deep Learning

1. INTRODUCTION

Disaster detection [1] and early warning systems [2] require regular data acquisitions to make possible the identification of suspicious phenomena, monitor them and early react to threats. But due to clouds sensitivity, rainfall induced phenomena (*e.g.* floods or landslides) detection from optical images is often delayed or even impossible. Adversarial SAR-to-optical translators neural networks are increasingly considered to fill missing optical acquisitions. In particular, Pix2pix [3] and CycleGAN [4] are shown as the most efficient in remote sensing [5]. Supervised approaches mostly focus on modifying Pix2pix [3], for instance by training latent features to mimic an adjoint optical autoencoder in [6] or by translating the SAR input to the optical wavelet decomposition in [7]. Unsupervised networks leverage CycleGAN [4] architecture as in [8] where multiscale cascaded residual connections are introduced in the generators or in [1] where their latent spaces are forced to be aligned. Authors of [9] merge both architectures to implement a semi-supervised approach.

These methods are trained on flat landscapes, avoiding steep slopes which cause strong geometrical distortions in SAR images: foreshortening, layovers and shadows. Recently, we proposed a network called SARDINet (SAR Distorted Image translator Network) [10] to tackle distortions in urban areas. In this paper, the focus is set on mountainous regions affected by wider distortions and more variable shapes and textures to be reconstructed, making them more challenging to translate. We first introduce a new SAR-DEM-optical dataset containing 53.475 patches of SAR-optical couples with the corresponding DEM over the French Alps. Then we propose to include the DEM as input to deep learning models to strengthen the identification of distorted areas and therefore their translation. Taking SARDINet [10] as backbone, three fusion strategies are explored: early fusion, late fusion and intermediate fusion which leverages balanced separable convolutions [11]. To the best of our knowledge, this is the first SAR-to-optical translator leveraging the DEM to tackle strong geometrical distortions.

The remaining of the paper is organized as follows. Section 2 introduces our fusion strategies and Section 3 describes the mountainous dataset. Experiments and results are developed in Section 4 before concluding in Section 5.

2. METHOD

2.1. SARDINet architecture

SARDINet [10] is a deep encoder-decoder network targeting SAR-to-optical translation in areas with geometrical and radiometrical SAR distortions as foreshortening or shadows. This non-adversarial architecture is three-stepped. First, input conditioning layers are applied for noise removal and coarse feature extraction. Second, a three branches encoder computes finer feature extraction depending on the resolution. Branches are based on separable convolutions [11] for successive spatial and inter-channel feature extractions. Finally, latent features are input to the decoder which reconstructs each output channel in an independent branch.

SARDINet [10] is originally applied in urban areas, *i.e.* with a structured context (circular tanks, rectangular buildings, straight streets). In this work we first propose to evaluate

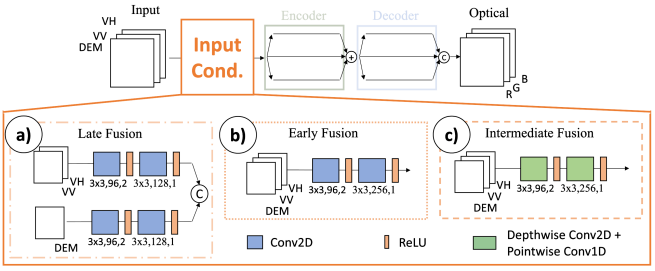

Fig. 1: SAR-DEM fusion strategies within the input conditioning. Encoder and decoder remain as in [10]. Networks are trained whether with a) late, b) early or c) intermediate fusion strategy.

its performances on a more challenging mountainous dataset. Indeed, the environmental context is less structured than urban areas and SAR geometrical distortions are stronger due to steep mountainous slopes which affects the range sampling of radar images as depicted by red and purple circles in Figure 4. To account for these distorted areas during the translation process and improve the relevancy of their optical reconstructions by the network, we modify SARDINet input conditioning to leverage the DEM as an additional input to the network.

2.2. Fusion strategies

As depicted in Figure 1 three DEM-SAR fusion strategies are implemented. **Late Fusion SARDINet**$_L$ in Figure 1a processes both SAR and DEM images through independent branches. Features are then concatenated and input to the encoder. **Early Fusion SARDINet**$_E$ in Figure 1b first combines both modalities and processes them using the standard input conditioning of SARDINet [10]. **Intermediate Fusion SARDINet**$_I$ in Figure 1c leverages separable convolutions [11]. The purpose is to temper feature merging to avoid an input to rule over the other with channel-wise convolutions while allowing feature exchanges in point-wise convolutions.

Contrary to the original paper [11], we balance separable convolutions as illustrated in Figure 2. Depth-wise convolution first extracts K_S spatial features from the same channel which are then concatenated along the channel dimension and input to the same point-wise convolutional layer to extract K_C inter-channel features. All resulting features are finally concatenated to output $K_S \times K_C$ feature maps.

3. MOUNTAINOUS SAR-DEM-OPTICAL DATASET

To study distortion robustness of deep translators, we create a new mountainous SAR-optical dataset released on IEEE Dataport (https://ieee-dataport.org/documents/sar-dem-optical-mountainous-dataset-distortion-management). It is composed by 53.475 patches of size 256×256 and resolution 10×10 meters from 36 SAR-DEM-optical triplets of images acquired over the northern French Alps between 2018 and 2021. Selected SAR images are standard Ground

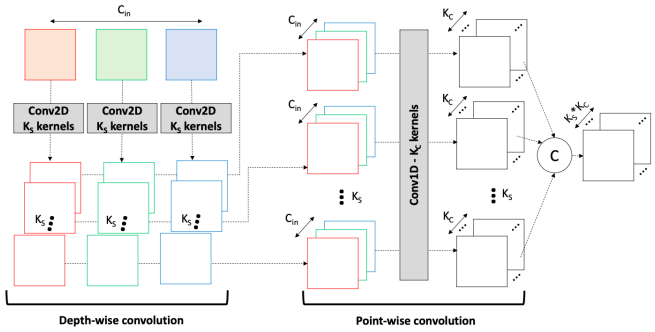

Fig. 2: Balanced separable convolutions. Depth-wise convolutions extract multiple features with K_S channel-wise independent convolutional kernels. Point-wise convolutions apply the same K_C kernels to the K_S groups of features maps.

Range Detected ESA products from Sentinel 1 constellation acquired during descending passes, in interferometric wide swath mode both in VV and VH polarizations. Images are orthorectified with the SRTM 1sec HGT DEM and calibrated. The long tail distribution of values is compressed with a logarithmic transformation. We select the aforementioned DEM to be input to the network after being normalized based on the maximum elevation. Optical images are RGB channels of level 2A products from Sentinel 2 constellation with a cloud coverage below 25%. These images are masked using provided snow and cloud maps. This step results in regions where unavailable values are set to 0 for each optical channel. Intensities are linearly transformed channel-wise to exploit the whole $[0, 255]$ range of values.

4. EXPERIMENTS

4.1. Training configuration

Based on results from [5], Pix2pix and CycleGAN are selected as state-of-the-art translators. They are trained using the Pytorch implementation available on Github with the SAR-DEM couple as input. To assess the impact of the DEM on the reconstruction, the original SARDINet [10] is also trained only with the SAR input. Each network is trained for 100 epochs with a learning rate of 1.10^{-5} and a batch size of 24. Optimization of SARDINet-derived networks is based on the MAE loss with an Adam optimizer. Pix2Pix and CycleGAN optimizations are left unchanged. The dataset is splitted in 80% training, 10% evaluation and 10% test.

Translation performances are evaluated based on Mean Square Error (MSE), Structural SIMilarity (SSIM) and Peak Signal To Noise Ratio (PSNR). Metric calculation are performed on the test set by discriminating for each image the areas impacted by foreshortening and shadows based on maps obtained through the algorithm detailed in [12]. The Fréchet Inception Distance (FID) is also measured only on the whole images to assess the credibility of optical patterns. Networks stability is studied by averaging the performances resulting from 5 random initializations.

	↓ MSE $.10^{-2}$	↑ SSIM.10^{-2}	↑ PSNR	↓ FID	↓ MSE.10^{-2}	↑ SSIM.10^{-2}	↑ PSNR
	Global				Foreshortening areas		
CycleGAN [4]	8.67±2.10	61.11±4.49	27.99±0.06	**10.67**±13.14	14.22±2.89	57.95±2.18	26.06±0.03
Pix2Pix [3]	8.19±1.27	59.14±1.46	28.03±0.04	21.71±8.67	8.05±1.12	59.94±1.91	26.25±0.07
SARDINet [10]	2.91±0.06	74.44±0.07	28.58±0.04	38.73±0.54	3.41±0.12	74.78±0.11	**26.99**±0.09
SARDINet$_L$	2.82±0.06	**74.48**±0.07	**28.59**±0.05	38.94±0.51	**3.20**±0.08	**74.89**±0.15	26.97±0.07
SARDINet$_E$	**2.81**±0.04	74.47±0.02	**28.59**±0.03	39.10±0.70	**3.20**±0.04	74.84±0.08	26.98±0.10
SARDINet$_I$	2.99±0.04	74.26±0.06	28.55±0.03	34.01±0.54	3.40±0.05	74.71±0.04	26.97±0.07
	Non distorted areas				Shadows areas		
CycleGAN [4]	8.42±2.19	61.27±4.62	27.99±0.07	*NA*	15.33±0.35	55.34±3.12	25.68±0.01
Pix2Pix [3]	8.19±1.29	59.12±1.45	28.03±0.04	*NA*	12.27±1.60	55.24±1.18	25.79±0.01
SARDINet [10]	2.84±0.06	74.53±0.07	28.58±0.04	*NA*	6.16±0.16	66.92±0.06	26.04±0.02
SARDINet$_L$	2.76±0.06	74.55±0.07	**28.60**±0.05	*NA*	5.85±0.14	67.00±0.05	**26.05**±0.02
SARDINet$_E$	**2.75**±0.05	**74.56**±0.02	28.59±0.03	*NA*	**5.80**±0.09	**67.02**±0.05	26.04±0.02
SARDINet$_I$	2.92±0.04	74.34±0.06	28.55±0.03	*NA*	6.18±0.14	66.65±0.09	26.01±0.02

Table 1: Quantitative comparisons of the fusion methods with the literature depending on SAR distortions. Mean and standard deviation over 5 experiments with different random initializations are displayed. Best results are displayed in bold.

4.2. Results

4.2.1. Fusion strategy comparison

Based on quantitative results displayed in Table 1, adding the DEM as input to SARDINet shows improvements in distorted areas. SARDINet$_E$ and SARDINet$_L$ both decrease strongly the MSE score compared to SARDINet. They also show qualitative improvements as visible in red circled shadow areas of Figure 4 where SARDINet confuses shadows with water body. The confusion is strongly reduced with SARDINet$_E$ and disappears with SARDINet$_L$ proving that separating the sources enforces the network to leverage appropriately the DEM despite quantitative similarities. Indeed, SARDINet$_E$ reconstructions look mostly ruled by the SAR image as water body is still reconstructed in shadow areas and inaccurate building patterns are reconstructed in urban zones (yellow circles). Finally, SARDINet$_I$ reaches competitive performances while reducing the number of learning weights from 223.776 to 9.852 in the input conditioning, which is encouraging towards lightweight networks implementations.

4.2.2. State-of-the-art comparison

As expected, in Table 1, Pix2pix [3] and CycleGAN [4] reach the two best FID scores - the latter being an indicator of the credibility of reconstructed patterns but not on their relevancy with respect to a ground truth. Indeed, all three other metrics show significant decrease and less stability in performances compared to SARDINet-like networks. As visible in Figure 4, Pix2pix creates cloud-like patterns and reconstructs rocks instead of a lake in the third row - confusing shadows and water bodies. On the other hand, CycleGAN is affected by the land-cover distribution in the training dataset. On the first and second row, it mostly generates forest-like patterns independently from the input image while on the third, it is not able to generate urban patterns in the yellow circle. In addition, in distorted areas, none of them is able to distinguish the forests from the fields in purple circles and CycleGAN

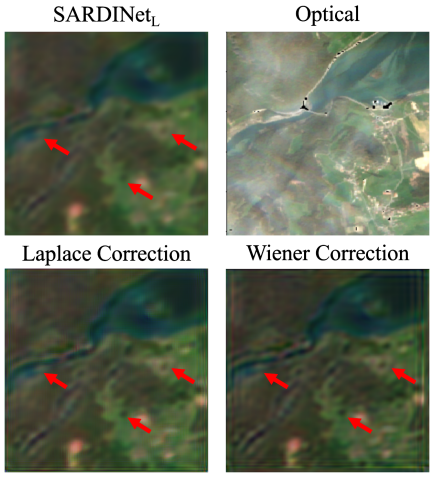

Fig. 3: Contrast enhancement applied on SARDINet$_L$ translations.

translates radar shadows as optical shadows which explains the strong quantitative drops in performances in distorted areas. Thus, despite their optical-like aspects, those adversarial reconstructions lack of stability in the training, relevancy in the reconstructed patterns and robustness to distortions compared to SARDINet derivatives.

4.2.3. Contrast enhancement

Our reconstructions show high quantitative performance but suffer from a low contrast compared to the ground truth images. We demonstrate here that this blurriness can be tackled with straightforward post-processings whereas missing informations in adversarial reconstructions are unrecoverable. Two standard image processing deblurring approaches are exploited: convolving the blurry image with the Laplacian filter or applying the Wiener filter on the image - which consists in dividing the Fourier Transform of the input image by the Fourier Transform of a gaussian kernel of size 9×9 and of standard deviation 29. For illustration purposes, these strategies are applied on SARDINet$_L$ translations which shown the

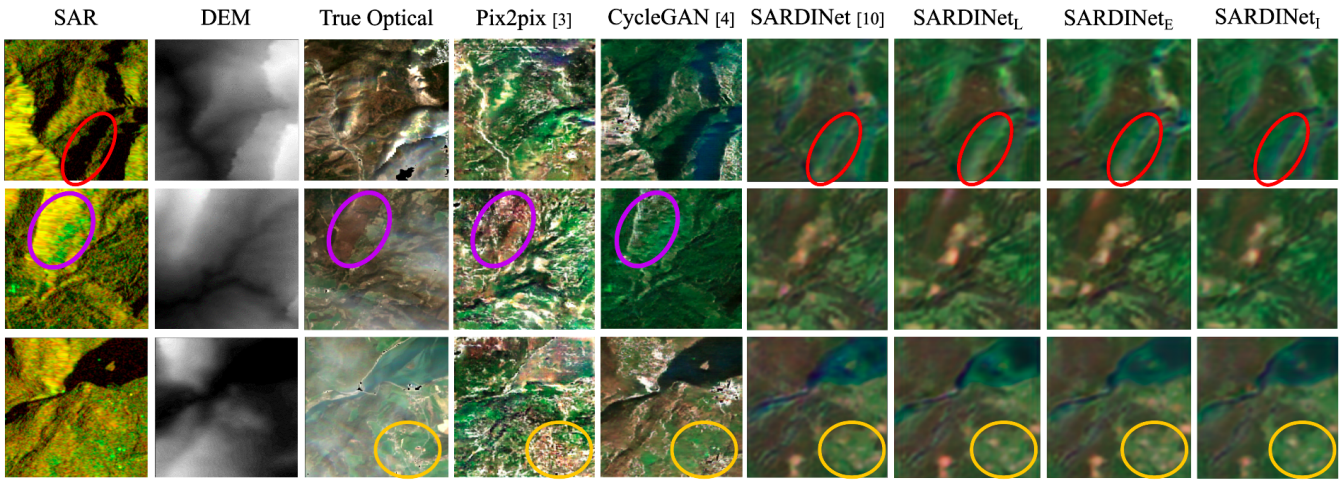

Fig. 4: Qualitative comparison of our fusion strategies with the literature. SAR image is displayed as [VV,VH,0]. Red and purple circles identify shadow and foreshortening areas respectively. Yellow circles focus on an urban high frequency area.

most reliable results. As visible on Figure 3, applying standard post-processing not only strengthen underlying objects borders and make structures more visible - as buildings and river borders pointed by red arrows - but textures in forested areas look also partially reconstructed. These improvements have a direct quantitative impact as the FID score decreased to 24.26 (resp. 25.87) with the Laplace (resp. Wiener) filter which lies within the FID performances of Pix2pix and CycleGAN (see Table 1) without affecting other metrics.

5. CONCLUSION

In this paper, a new SAR-DEM-optical mountainous dataset is created to assess the impact of the DEM on SAR-to-optical translations in SAR distorted regions. Three DEM-SAR fusion strategies are explored and compared to standard translators: Pix2pix [3], CycleGAN [4] and SARDINet [10]. The DEM is shown beneficial to avoid confusions in distorted areas - especially with the late fusion configuration - improving network robustness. Finally, apparent blurriness can be straightforwardly tackled to reach more credible optical-like patterns, while saving reliable land-cover patterns the latter cannot restore. Further works aim at using the translation of post event SAR images for change detection from a multimodal SAR-optical series of images.

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

6. REFERENCES

[1] L. T. Luppino, M. A. Hansen, M. Kampffmeyer, F. M. Bianchi, et al., "Code-Aligned Autoencoders for Unsupervised Change Detection in Multimodal Remote Sensing Images," *IEEE Trans. Neural Netw. Learn. Syst.*, 2022.

[2] M. Desrues, P. Lacroix, and O. Brenguier, "Satellite Pre-Failure Detection and In Situ Monitoring of the Landslide of the Tunnel du Chambon, French Alps," *Geosci.*, vol. 9, no. 7, pp. 313, July 2019.

[3] P. Isola, J.-Y. Zhu, T. Zhou, and A. A. Efros, "Image-to-Image Translation with Conditional Adversarial Networks," *Proc. IEEE Comput. Soc. Conf.*, pp. 1125–1134, Nov. 2016.

[4] J.-Y. Zhu, T. Park, P. Isola, and A. A. Efros, "Unpaired Image-To-Image Translation Using Cycle-Consistent Adversarial Networks," in *Proc. IEEE Int. Conf. Comput. Vis.*, 2017, pp. 2223–2232.

[5] Y. Zhao, T. Celik, N. Liu, and H.-C. Li, "A Comparative Analysis of GAN-Based Methods for SAR-to-Optical Image Translation," *IEEE Geosci. Remote. Sens. Letters*, vol. 19, pp. 1–5, 2022.

[6] H. Wang, Z. Zhang, Z. Hu, and Q. Dong, "SAR-to-Optical Image Translation With Hierarchical Latent Features," *IEEE Trans. Geosci. Remote Sens.*, vol. 60, pp. 1–12, 2022.

[7] H. Li, C. Gu, D. Wu, G. Cheng, et al., "Multiscale Generative Adversarial Network Based on Wavelet Feature Learning for SAR-to-Optical Image Translation," *IEEE Trans. Geosci. Remote Sens.*, vol. 60, pp. 1–15, 2022.

[8] 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, Jan. 2021.

[9] W.-L. Du, Y. Zhou, H. Zhu, J. Zhao, et al., "A Semi-Supervised Image-to-Image Translation Framework for SAR–Optical Image Matching," *IEEE Geosci. Remote. Sens. Letters*, vol. 19, pp. 1–5, 2022.

[10] A. Bralet, A. M. Atto, J. Chanussot, and E. Trouvé, "Deep Learning of Radiometrical and Geometrical Sar Distorsions for Image Modality translations," in *IEEE Int. Conf. Image Processing (ICIP)*, Oct. 2022, pp. 1766–1770.

[11] F. Chollet, "Xception: Deep Learning With Depthwise Separable Convolutions," in *Proc. IEEE Int. Conf. Comput. Vis.*, 2017, pp. 1251–1258.

[12] E. Meier, U. Frei, and D. Nüesch, "Precise Terrain Corrected Geocoded Images," *SAR Geocoding: Data and Systems*, pp. 173–185, 1993.