Multimodal SAR-optical fusion for cloud removal in Sentinel-2 imagery using conditional GANs

A multi-seasonal approach leveraging Pix2pix architecture and automated data curation.

LEDERMANN Quentin
January 2026
Master 2 Observation de la Terre et Géomatique (OTG) – 3 rue de l'Argonne, 67000 Strasbourg

Abstract : Cloud cover remains a primary obstacle for operational Sentinel-2 monitoring, particularly in multi-seasonal agricultural applications. This study proposes a multimodal cloud removal framework leveraging Sentinel-1 Synthetic Aperture Radar (SAR) data through a modified 5-channel Pix2Pix architecture. We introduce an automated data curation pipeline that adaptively filters triplets based on seasonal contrast thresholds, resulting in a high-quality dataset of 4,464 samples. Our results demonstrate that early fusion of SAR VV/VH and cloudy RGB channels allows the model to reconstruct fine-scale structural features, such as urban grids and agricultural field layouts, with a PSNR of 27.18 dB and an SSIM of 0.791. Qualitative analysis confirms the model's robustness in translating SAR geometries into realistic optical textures across both summer and winter landscapes. This research provides a scalable solution for continuous Earth observation, ensuring that data gaps caused by atmospheric interference no longer hinder operational monitoring.

Keywords : Conditional GANs, Pix2Pix, Sentinel-1/2 fusion, Cloud removal, Multi-seasonal analysis, SAR to optical

1. Introduction

Optical satellite imagery, particularly from the ESA Sentinel-2 mission, has revolutionized global Earth observation. However, its operational use is severely hindered by cloud cover, which affects approximately 60% of the Earth's surface at any given time. This atmospheric interference creates significant data gaps, making continuous time-series analysis challenging for applications such as precision agriculture or disaster management.

To overcome this limitation, Synthetic Aperture Radar (SAR) data from Sentinel-1 offers a viable alternative. SAR sensors operate in the microwave spectrum, allowing them to penetrate clouds and acquire data regardless of illumination or weather conditions. Despite this advantage, SAR imagery is difficult to interpret due to speckle noise and its fundamentally different physical interaction with the terrain compared to optical reflectance.

Recent advances in Conditional Generative Adversarial Networks (cGANs), specifically the Pix2Pix architecture (Isola et al., 2017), have shown great promise in "*translating*" images between domains. By treating cloud removal as an image-to-image translation problem, we can train a model to synthesize cloud-free optical patches from SAR inputs.

In this study, we implement an early fusion approach using 5-channel inputs (SAR VV/VH polarizations combined with cloudy RGB channels). Unlike single-season models, our work utilizes a multi-seasonal dataset from SEN12MS (Schmitt et al., 2019), including both summer and winter acquisitions. This integration aims to improve the model's robustness against seasonal phenological changes, such as vegetation cycles and snow cover.

The objective of this paper is to demonstrate that a rigorous automated data cleaning pipeline combined with a multimodal GAN can effectively reconstruct fine-scale ground features (urban and agricultural) even under dense cloud layers.

2. Materials and Methods

2.1. Dataset and seasonal variability

The dataset used in this study is derived from SEN12MS (Schmitt et al., 2019), a large-scale collection of synchronized Sentinel-1 (SAR) and Sentinel-2 (optical) patches. To enhance the model's ability to generalize across different environmental conditions, we curated a muti-seasonal subset comprising both summer and winter acquisitions. The inclusion of winter data is particularly significant as it introduces diverse phenological states, such as senescent vegetation, bare soils, and occasional snow cover, which differ drastically from the dense green canopies of summer. By training on this combined dataset, the model is forced to decouple structural information (from SAR) from seasonal spectral signatures (from optical).

2.2. Dataset and seasonal variability

A critical contribution of this work lies in the development of a custom automated quality control pipeline (clean_dataset.py). Raw satellite datasets often contain corrupted tiles, sensor artefacts, or non-informative samples (e.g., open ocean or cloud-saturated pixels) that can destabilize GAN training.

We implemented a two-stage filtering process based on statistical properties :

1. SAR validation : Samples were rejected if the standard deviation (σ) was below 0.0001 or the maximum pixel value was below 0.001. This effectively removed "*dead*" pixels and areas with no backscatter variance.
2. Optical validation : To ensure the presence of reconstructible features, we enforced a threshold on the mean RGB standard deviation. While a strict threshold ($\sigma > 10$) was used for summer data, it was adjusted to $\sigma > 3$ for winter data to account for the naturally lower contrast of snow-covered or dormant landscapes without including corrupted "*flat*" images.

The final curated dataset consists of 4,069 summer triplets and 395 winter triplets, totaling 4,464 high-quality samples. The dual-stage filtering logic, illustrated in Figure 1, ensures that the GAN learns from geophysically meaningful textures rather than sensor noise or uninformative 'flat' imagery.

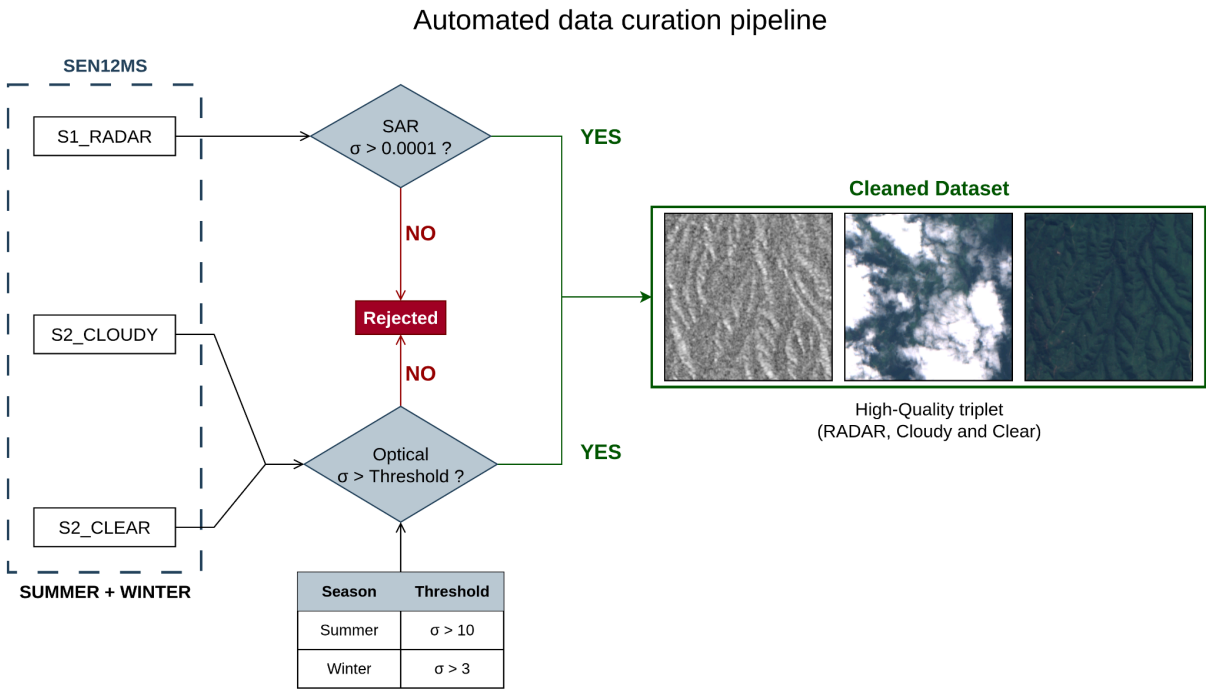

Figure 1 : Automated data curation and quality control pipeline. *The flowchart illustrates the statistical filtering logic used to refine the dataset. Triplets are evaluated based on their standard deviation (σ); samples failing the SAR threshold (σ < 0.0001) or the seasonally-adjusted optical contrast test (σ < 10 for summer, σ < 3 for winter) are discarded to ensure training stability.*

2.3. Model architecture : 5-channel multimodal Pix2Pix

The core of our translation framework is based on the Conditional Adversarial Network (cGAN) architecture, specifically the Pix2Pix model (Isola et al., 2017). However, we modified the standard 3-channel input to a 5-channel early fusion configuration.

- Generator (U-Net) : The generator uses a U-Net based architecture consisting of an encoder-decoder structure with skip connections. These connections are vital in remote sensing as they shuttle low-level structural information (such as urban geometrics and field boundaries from the SAR signal) directly to the decoding layers, bypassing the bottleneck. The input tensor [256 x 256 x 5] concatenates Sentinel-1 VV/VH polarizations with the cloudy Sentinel-2 RGB channels.
- Discriminator (PatchGAN) : We employed a 70 x 70 PatchGAN discriminator. Instead of classifying the entire image as real or fake, the discriminator looks at local patches. This encourages the generator to produce high-frequency textures, effectively reducing the "*blurring*" effect common in L1-only optimization. The discriminator receives an 8-channel input : the 5-channel source (S1 + S2 cloudy) concatenated with the 3-channel output (either the real clear S2 or the synthesized fake S2).

Our architecture leverages an early fusion strategy by concatenating SAR and optical channels into a single 5-channel input tensor, as shown in Figure 2. This structure allows the model to utilize SAR structural information as a geometric guide for the optical reconstruction.

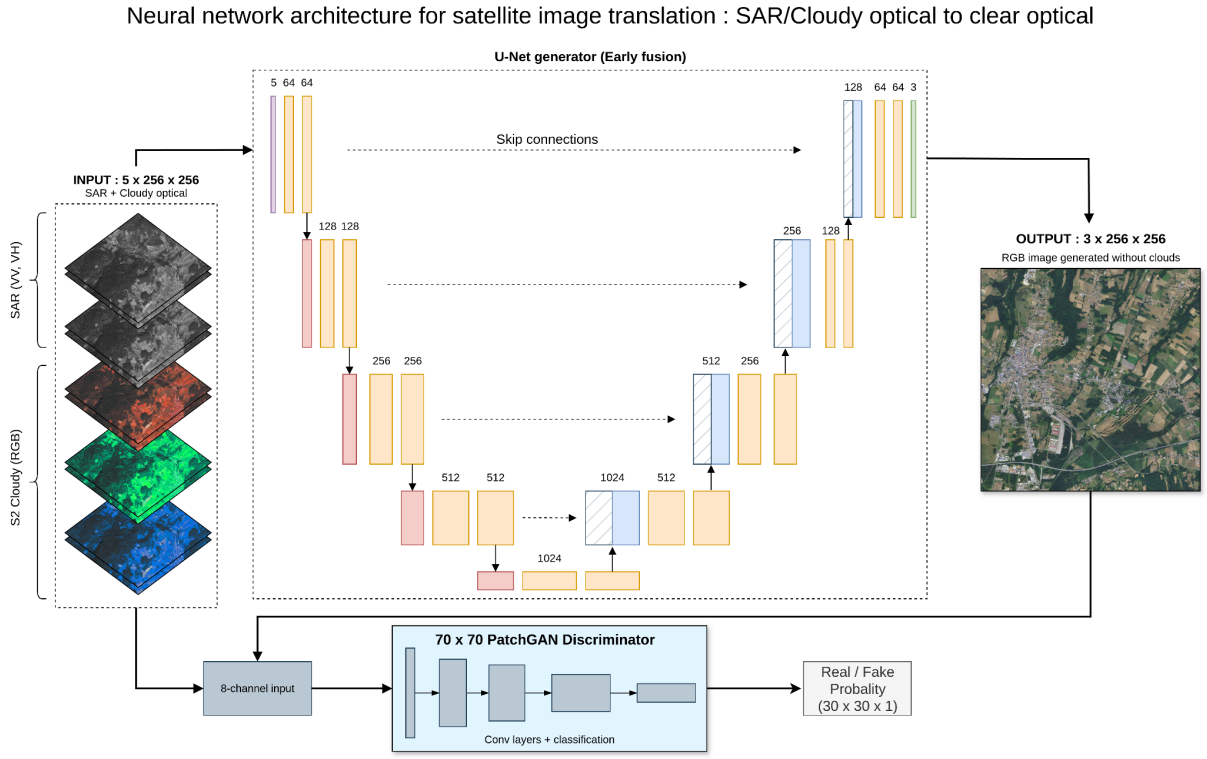

Figure 2 : Proposed 5-channel multimodal Pix2Pix architecture. *Overview of the modified GAN framework. The generator uses a 5-channel early fusion input (VV, VH, R, G, B) to preserve structural geometries via skip connections. The discriminator is a 70 x 70 PatchGAN that outputs a 30 x 30 prediction grid, enforcing local texture realism.*

2.4. Training protocol and Loss function

The objective function of the network follows the cGAN formulation combined with an L1 regularization term to ensure spatial consistency :

$$G^* = \arg\min_G \max_D \mathcal{L}_{cGAN}(G, D) + \lambda \mathcal{L}_{L1}(G)$$

Training was conducted on an NVIDIA RTX 3080 GPU using Automatic Mixed Precision (AMP) to optimize VRAM usage and accelerate computation. We used the Adam optimizer with a learning rate of 2 x 10^{-4} and momentum parameters $\beta = 0.5$, $\beta_2 = 0.999$. The model was trained for 200 epochs with a batch size of 16. Data augmentation, including random horizontal and vertical flips, was applied to the training set to prevent overfitting. This was particularly critical given the smaller size of the winter subset (395 samples compared to 4,069 for summer), as it allowed the model to see a greater variety of orientations for the same dormant landscapes, thereby improving its generalization capabilities.

3. Results and discussion

3.1. Training convergence and quantitative evaluation

The proposed model was trained for 200 epochs, reaching a stable equilibrium between the generator and discriminator losses. The integration of the L1 regularization term ($\lambda = 100$) proved critical in maintaining spatial consistency throughout the translation process.

To assess the reconstruction quality, we utilized two standard computer vision metrics : Peak Signal-to-Noise Ratio (PSNR) and Structural Similarity Index (SSIM). The quantitative results are summarized below :

- PSNR : The model achieved an average score of 27.18 dB on the validation set. This relatively high value indicates that the synthesized images maintain a low pixel-wise error compared to the ground truth Sentinel-2 clear imagery.
- SSIM : A score of 0.791 was obtained, demonstrating that the structural information specifically urban boundaries and agricultural field layouts is successfully preserved. This confirms the effectiveness of the U-Net skip connections in transferring SAR-derived geometries to the optical output.

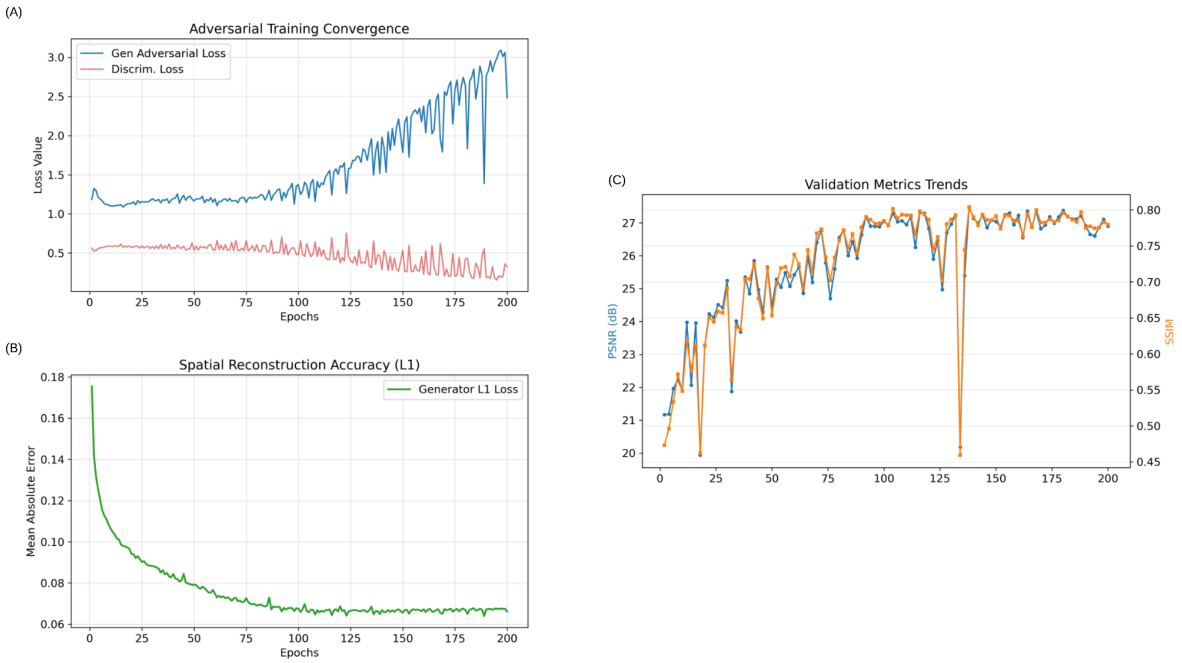

Figure 3 : Quantitative training and validation performances. *(A) Adversarial Loss evolution showing the competitive equilibrium between Generator and Discriminator. (B) Generator L1 Loss (MAE) highlighting the convergence of pixel-wise geographical reconstruction. (C) Validation metrics (PSNR and SSIM) computed every 2 epochs, illustrating the steady improvement in image fidelity and structural consistency.*

The quantitative evolution of the training process is detailed in Figure 3. As observed in the loss curves (Fig. 3A and 3B), the model exhibits a rapid initial convergence of the L1 term, followed by a stabilization of the adversarial loss, confirming a healthy GAN training dynamic without mode collapse.

The validation metrics (Fig. 3C) corroborate this progress: the PSNR gradually increases to reach 27.18 dB, while the SSIM stabilizes around 0.791. The synchronization between the decrease in L1

loss and the increase in SSIM demonstrates that the model effectively learns to use SAR geometries to reconstruct valid optical structures, rather than merely producing 'hallucinated' textures.

3.2. Training convergence and quantitative evaluation

The visual inspection of the synthesized patches (Figure 4) reveals the model's ability to "see through" dense cloud cover. While the original Sentinel-2 cloudy inputs are completely opaque in the visible spectrum, the 5-channel early fusion approach allows the network to use SAR backscatter as a structural guide. In urban areas, the 70 x 70 PatchGAN discriminator effectively enforces local realism, preventing the "blurring" typically associated with pure L1 optimization. Even in winter samples, where the spectral contrast is naturally lower, the model successfully reconstructs the ground features by decoupling the seasonal signatures from the permanent topographic structures.

The qualitative performance of the model is illustrated in Figure 4. By comparing the synthesized outputs with the ground truth clear imagery, it is evident that the early fusion of SAR data provides a reliable geometric backbone, allowing the generator to reconstruct missing optical information with high spatial precision.

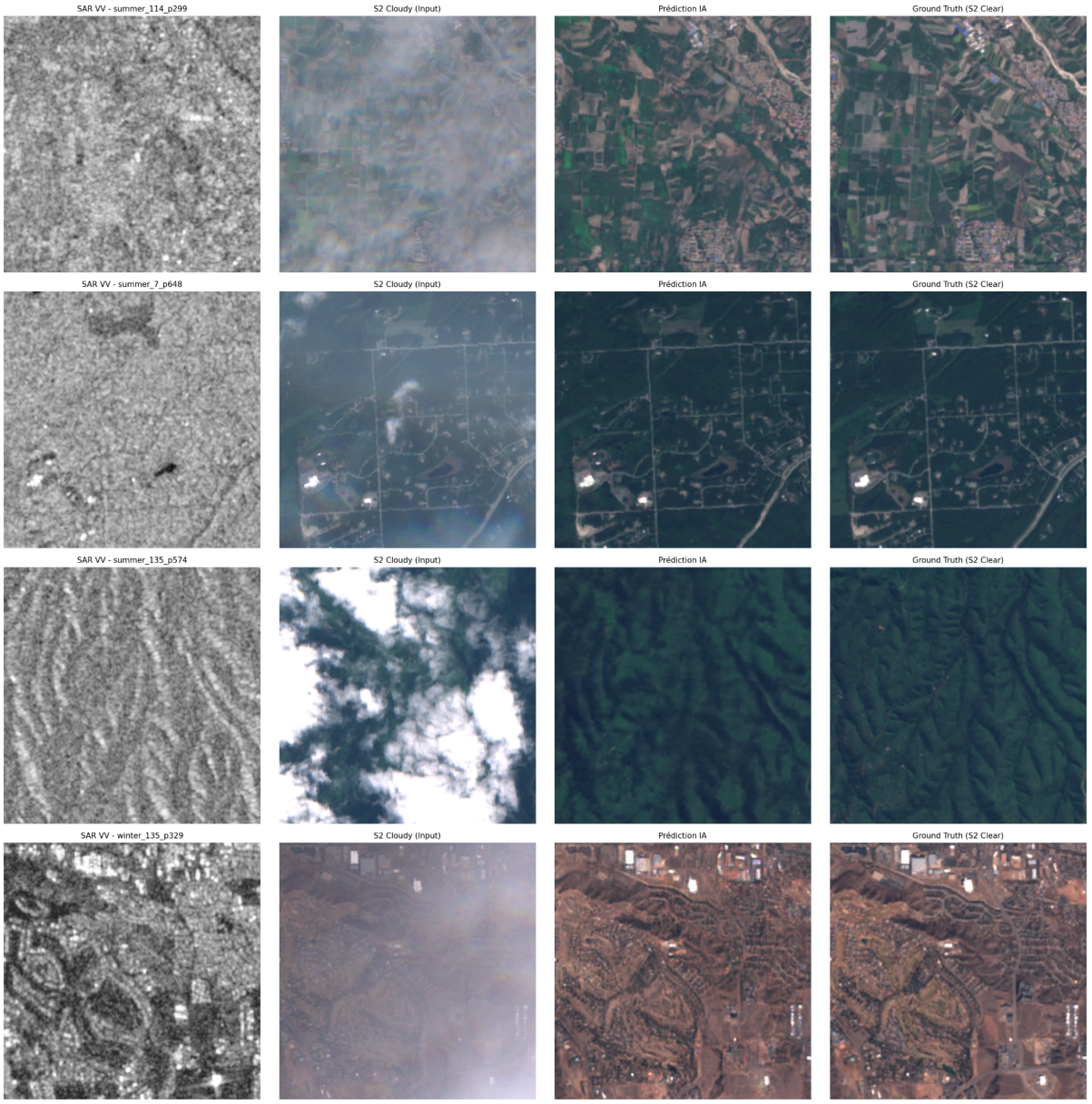

Figure 4 : Qualitative cloud removal results on the validation set (Epoch 200). *From left to right: (a) Sentinel-1 SAR input (VV), (b) Sentinel-2 cloudy input, (c) Synthesized cloud-free output (AI prediction), and (d) Sentinel-2 ground truth. The model effectively reconstructs urban geometries (line 2) and complex topography (line 3) by using SAR as a structural guide. The bottom row (winter sample) demonstrates the model's robustness to seasonal spectral variations.*

As shown in Figure 4, the model demonstrates high fidelity in reconstructing various landscapes. The second row highlights the preservation of linear features (roads and urban grids) even when the optical input is obscured by haze. Most importantly, the fourth row (winter sample) confirms that the model successfully decouples the permanent land structure from the seasonal phenology. Although extreme snow cover is not present in this specific set, the accurate reconstruction of the dormant vegetation's brown tones proves that the seasonally-adjusted thresholds in our curation pipeline (Section 2.2) effectively preserved the necessary spectral diversity for robust multi-seasonal learning.

4. Conclusion

This study demonstrates the effectiveness of multimodal SAR-optical fusion for cloud removal in Sentinel-2 imagery using a conditional GAN framework. By implementing a 5-channel early fusion architecture and an adaptive multi-seasonal data curation pipeline, we successfully reconstructed cloud-covered regions with high structural fidelity (SSIM: 0.791) and spectral accuracy (PSNR: 27.18 dB).

Our findings highlight that while SAR backscatter provides a robust geometric backbone, the integration of seasonal contrast thresholds is essential for training stability in varied environments. This research provides a scalable solution for continuous Earth observation, ensuring that data gaps caused by atmospheric interference no longer hinder operational monitoring in critical fields such as precision agriculture and environmental management.

Bibliography

J. D. Bermudez, P. N. Happ, D. A. Oliveira, and R. Q. Feitosa, "SAR to optical image synthesis for cloud removal with conditional generative adversarial networks," ISPRS Annals of Photogrammetry, Remote Sensing & Spatial Information Sciences, vol. 4, no. 2, 2018.

M. Drusch, U. Del Bello, S. Carlier, O. Colin, V. Fernandez, F. Gascon, et al., "Sentinel-2: ESA's optical high-resolution mission for operational services," Remote Sensing of Environment, vol. 120, pp. 25-36, 2012.

P. Ebel, A. Meraner, M. Schmitt, and X. X. Zhu, "Multi-sensor data fusion for cloud removal in global and multi-temporal Sentinel-2 imagery," IEEE Transactions on Geoscience and Remote Sensing, vol. 58, no. 11, pp. 7582-7595, 2020.

K. Enomoto, K. Sakurada, W. Wang, H. Fukui, M. Matsuoka, R. Nakamura, and H. Nobuhara, "Filmy cloud removal on satellite imagery with combined use of SAR gradient information," in IEEE Conference on Computer Vision and Pattern Recognition Workshops, 2017, pp. 1533-1542.

J. Gao, Q. Yuan, J. Li, H. Zhang, and X. Sun, "Cloud removal with fusion of high resolution optical and SAR images using generative adversarial networks," Remote Sensing, vol. 12, no. 12, p. 1917, 2020.

I. Goodfellow, J. Pouget-Abadie, J. Mirza, B. Xu, D. Warde-Farley, S. Ozair, et al., "Generative adversarial nets," Advances in Neural Information Processing Systems, vol. 27, 2014.

C. Grohnfeldt, M. Schmitt, and X. X. Zhu, "A conditional generative adversarial network to fuse SAR and optical satellite imagery for cloud removal," in IEEE International Geoscience and Remote Sensing Symposium (IGARSS), 2018, pp. 5255-5258.

K. He, X. Zhang, R. Ren, and J. Sun, "Deep residual learning for image recognition," in Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2016, pp. 770-778.

P. Isola, J. Y. Zhu, T. Zhou, and A. A. Efros, "Image-to-image translation with conditional adversarial networks," in Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2017, pp. 1125-1134.

J. Johnson, A. Alahi, and L. Fei-Fei, "Perceptual losses for real-time style transfer and super-resolution," in European Conference on Computer Vision, 2016, pp. 694-711.

D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," arXiv preprint arXiv:1412.6980, 2014.

Y. Luo, X. Li, Y. Wang, and G. Chen, "Deep learning based cloud removal from satellite imagery: A review," Earth Science Informatics, pp. 1-17, 2020.

A. Meraner, P. Ebel, M. Schmitt, and X. X. Zhu, "Cloud removal in Sentinel-2 imagery using a deep residual neural network and SAR-optical data fusion," ISPRS Journal of Photogrammetry and Remote Sensing, vol. 166, pp. 333-346, 2020.

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

O. Ronneberger, P. Fischer, and T. Brox, "U-net: Convolutional networks for biomedical image segmentation," in International Conference on Medical Image Computing and Computer-Assisted Intervention, 2015, pp. 234-241.

M. Schmitt, L. H. Hughes, C. Qiu, and X. X. Zhu, "SEN12MS – A Curated Dataset of Sentinel-1 and Sentinel-2 Imagery for Deep Learning," arXiv preprint arXiv:1906.07789, 2019.

M. Schmitt and X. X. Zhu, "Data fusion and remote sensing: An overview," IEEE Geoscience and Remote Sensing Magazine, vol. 4, no. 3, pp. 6-23, 2016.

R. Torres, P. Snoeij, D. Geudtner, J. Bibby, M. Davidson, E. Attema, et al., "GMES Sentinel-1 mission," Remote Sensing of Environment, vol. 120, pp. 9-24, 2012.

Z. Wang, A. C. Bovik, H. R. Sheikh, and E. P. Simoncelli, "Image quality assessment: from error visibility to structural similarity," IEEE Transactions on Image Processing, vol. 13, no. 4, pp. 600-612, 2004.

Q. Zhang, Q. Yuan, J. Li, Z. Yang, and X. Ma, "Missing data reconstruction in remote sensing image with a unified spatial–temporal–spectral deep convolutional neural network," IEEE Transactions on Geoscience and Remote Sensing, vol. 56, no. 12, pp. 7274-7288, 2018.