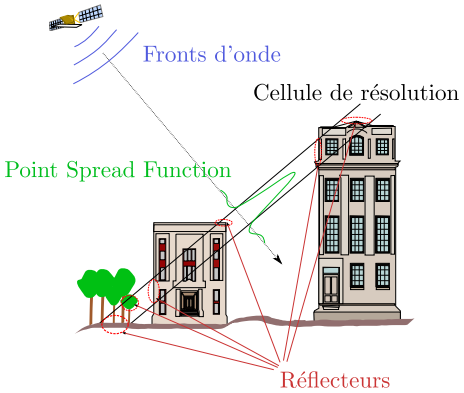
Deep learning and representation spaces for multi-sensor satellite image interpretation

1 Objective of the PhD

The objective of this thesis is to develop a generic and robust framework for the interpretation of remote sensing scenes that can be fed by various data (active/passive sensors, different modalities, resolutions, acquisition dates). This framework will be developed based on representation spaces learned by deep neural networks integrating physical and structural priors.

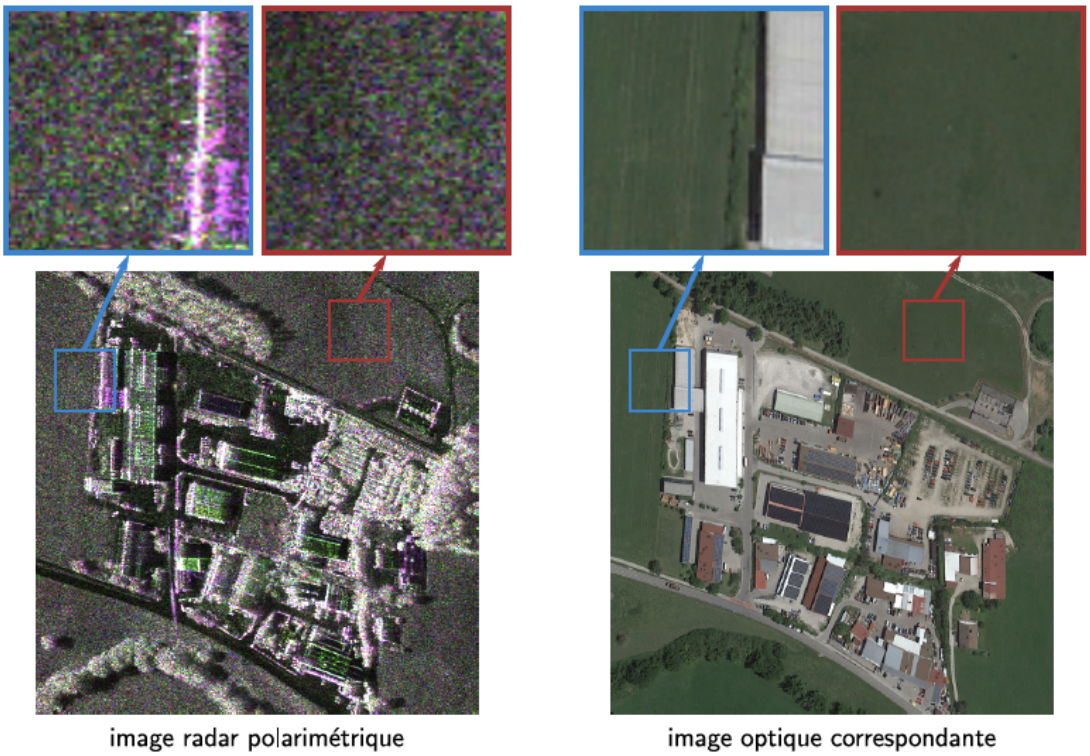
2 PhD description

Artificial intelligence (AI) and machine learning approaches have revolutionized computer vision and image processing in recent years. The field of remote sensing, which uses satellite or aerial images for Earth observation, also benefits from the efficiency of these approaches which have led to many advances. Nevertheless, a direct application of learning methods faces multiple difficulties. First of all, the field of remote sensing has strong specificities that require significant adaptations of learning methods. For example, in SAR (Synthetic Aperture Radar) imagery, the image focusing, the specificities of coherent imagery (speckle), or the complex and possibly vectorial nature (in polarimetry, interferometry or tomography) of the data must be taken into account. Moreover, there are currently few labeled remote sensing datasets, especially in SAR imagery, and the creation of such datasets is particularly resource-intensive. From an application point of view, the field of remote sensing is currently undergoing a very strong evolution with the availability of huge amounts of data with high temporal frequencies, such as the Sentinel-1 and Sentinel-2 data from the European Space Agency. Thus, the current challenges of satellite imaging involve the joint exploitation of multiple sources of data acquired both with active sensors (such as SAR) and passive sensors (optical sensors), at different spatial, spectral and temporal resolutions.



Principle of SAR image acquisition :
after emitting an electro-magnetic wave, the sensor records the backscattered signal which can mix different elements of the scene if located at the same distance from the sensor.

The objective of this thesis is to build a generic and robust representation framework allowing the interpretation of satellite scenes from heterogeneous data. To do so, we propose to build a high-level representation of the scene that can be used for different tasks, based on a deep learning architecture. This representation can be fed by different types of data (optical data, SAR of different modalities -amplitude, interferometry, tomography-) to give an interpretation of the scene at different levels (class information, heights, temporal evolution, types of materials or backscatterers, ...).

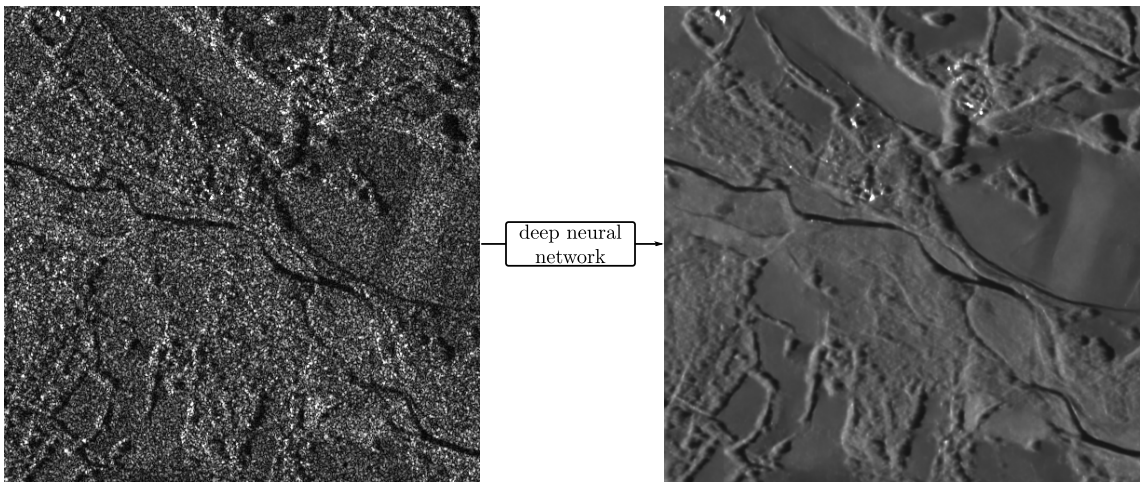


Appearance of the objects in a scene for a SAR image (on the left, polarimetric modality) and an optical image (on the right)

The methodological obstacles to overcome are related to the introduction of physical and structural constraints in the networks used and to the definition of a unified representation model allowing multi-sensor fusion. This approach is a break from the usual approaches that focus on a particular task and train a specific network to perform it. Instead, we propose to build on the efficiency of the representations generated by the networks to develop a generic representation, invariant to the sensor and thus exploitable for several tasks. This approach should also allow progress in the explicability of learning approaches in the field of remote sensing through the construction of an interpretable latent space.

3 Program of the PhD

The objective of this thesis is to construct a generic and robust representation of a scene from multi-sensor data. We propose to rely on representation spaces learned by deep learning networks, in particular auto-encoders. We describe below the two methodological axes that will be developed in this work and the associated applications.



Exploiting a physical model of SAR images allows to train without ground-truth (auto-supervised mode) deep neural networks to do speckle reduction in SAR images [4].

3.1 Building invariant representation spaces

We propose to exploit the architectures of auto-encoders, and in particular variational auto-encoders, to build representation spaces with invariance properties. We first propose to study how to train a network to learn simple geometric parameters (such as the position, size and height of isolated buildings) in high-resolution optical or SAR imagery. The objective is to make this learning invariant to different acquisition parameters : the nature (active or passive) of the sensor, the position of the sensor with respect to the scene, its angle of incidence, ... Particular attention will be paid to the geometry of the constructed latent space and its generative capacities. One can take inspiration from the works on the unmixing (*disentanglement*) of the variables which influence the representation (as for example in face edition where the variables of gender, age, pose of the face are unmixed in the latent space) [14] or of cost functions favoring the structuring of the latent space in the form of a smooth variety [10].

As mentioned previously, this axis will benefit from the work currently being carried out in the team on latent spaces [14, 9] and on the denoising of SAR data in an unsupervised framework [4, 3]. In addition, the team's experience in interferometry [6] and radar tomography [11] allows to consider several types of specific physical constraints. Finally, the availability of optical and radar data pairs at different resolutions and in different configurations, as well as the availability of simulated data, will allow to work on the invariance of the representation to the considered sensor.

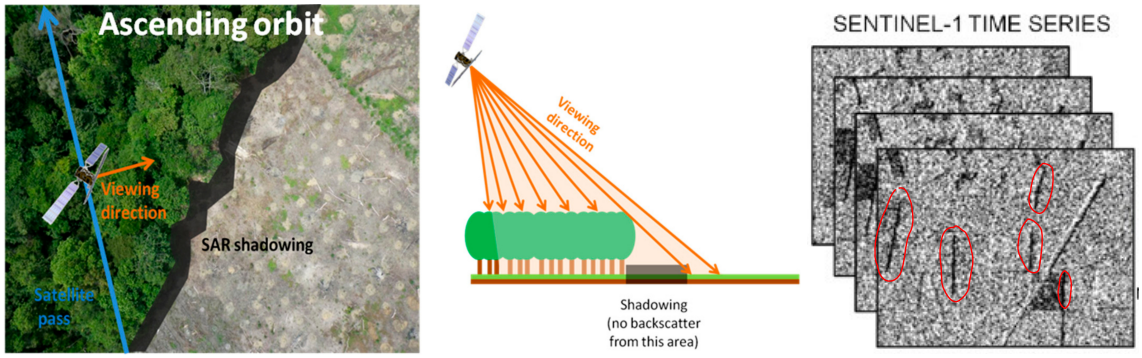
The objective of this step is the definition of a generic representation (invariant to the acquisition source and to the task to be carried out) which will allow to combine different heterogeneous acquisitions (with different spatial, temporal and spectral resolutions). It will be used for different tasks (semantic segmentation, elevation estimation, denoising, temporal tracking, etc.) depending on the available sources.

3.2 Spatial relationships

A second research axis will focus on the inclusion of knowledge on the spatial relations between objects from the scene or parts an object. These spatial relations are often exploited in the field of medical imagery for which we have a strong a priori knowledge of the relative positions of organs or structures within organs [1]. This information is also meaningful in remote sensing imaging, for example with the relative positions of the overlay and shadow areas in SAR imagery or of the occluded and shadow zones in optical imagery. Recent works explore the introduction of these relationships using different strategies : through graph-based representations or by integrating them into spatial relation maps and modifying the cost functions optimized by the network [12]. This information

can be exploited in the case of isolated objects and in the construction of their representation as discussed in the previous axis. But taking into account the spatial relations structuring the urban scene and thus the relations between objects is also an important element (as for example the alignment of buildings). This information can also be exploited in a hierarchical form, in particular in a multi-scale context as proposed in [7] to take into account the multi-resolution aspects.

3.3 Examples of applications



The presence of shadows is a strong indication of the deforestation of a plot in a SAR image. Illustration adapted from [2].

As previously mentioned, the originality of the proposed approach is based on an in depth work on data representation, independently of specific applications. Nevertheless, the developed methods can be applied for different purposes. In particular, the definition of a generic representation, independent of the sensor, makes it possible to remove the obstacles to the detection of multi-sensor or even single-sensor changes but with different angles of incidence, for example in SAR imaging. Existing works consider identical or close acquisition angles [13, 8], while larger changes of incidence angles is much more challenging because it requires a high-level representation. Many data sets are currently available : Sentinel-1 and Sentinel-2 data, freely available with high temporal repetition rate but moderate resolution. The team also has many higher resolution SAR data sets acquired with TerraSAR-X, CosmoSkyMed or RadarSat-2 in different modalities (Strip Map, Spotlight, Staring Spotlight, polarimetry) with interferometric and tomographic configurations.

Another application is optical and SAR matching and joint scene interpretation. One can imagine an approach of incremental updating of the interpretation in a broad sense of the scene each time a new data is acquired, either to improve the current interpretation, or to update it if an evolution has taken place.

3.4 Planning of the PhD

The first year will be devoted to the understanding of optical and SAR imaging and to the mathematical and physical modelings associated with them (sensor modeling, SAR synthesis, statistical models, ...). A research axis on auto-encoders and domain adaptation approaches will be conducted, in connection with the introduction of physical and structural knowledge on image acquisition.

The second year will be devoted to multi-sensor aspects and the construction of invariant features. The use of a multi-task framework (semantic segmentation, elevation map, backscatterers/materials) will be the starting point for this axis. Heterogeneous datasets (SAR, optical, multi-resolution, and multi-angles) available in the team will be exploited to develop and validate the representation built.

The third year will be devoted to the development of applications based on this representation. Two main domains are envisaged : characterization and three-dimensional reconstruction in multi-sensor urban environments and temporal monitoring of urban or natural environments (forests, hydrological network).

4 Supervision

The IMAGES team of Télécom Paris has a long experience in SAR imaging [4, 5, 3, 8, 9] (`https://perso.telecom-paristech.fr/tupin/radarteam/staffEN.php`). The PhD will be co-supervised by Loïc Denis at Télécom Saint-Etienne (Laboratoire Hubert Curien, Univ. de Saint-Etienne / CNRS / Institut d'Optique Graduate School) who is an invited researcher of Télécom Paris, a long-time collaborator of the team whose work in non-conventional imaging (astronomy, holography, SAR imaging) is internationally recognized.

Moreover, this thesis is part of the ANR ASTRAL project whose objective is to develop learning methods that take into account the physics of SAR imaging acquisition and which includes the CNAM, the Hubert Curien laboratory, ONERA and Télécom Paris.

Références

[1] Isabelle Bloch. Fuzzy spatial relationships for image processing and interpretation : a review. *Image and Vision Computing*, 23(2) :89–110, 2005. Discrete Geometry for Computer Imagery.

[2] Alexandre Bouvet, Stéphane Mermoz, Marie Ballère, Thierry Koleck, and Thuy Le Toan. Use of the sar shadowing effect for deforestation detection with sentinel-1 time series. *Remote Sensing*, 10(8) :1250, 2018.

[3] Emanuele Dalsasso, Loïc Denis, and Florence Tupin. SAR2SAR : A Semi-Supervised Despeckling Algorithm for SAR Images. *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, 14 :4321–4329, 2021.

[4] Emanuele Dalsasso, Loïc Denis, and Florence Tupin. As if by magic : self-supervised training of deep despeckling networks with MERLIN. *IEEE Transactions on Geoscience and Remote Sensing*, pages 1–1, 2022.

[5] Loïc Denis, Emanuele Dalsasso, and Florence Tupin. A Review of Deep-Learning Techniques for SAR Image Restoration. In *2021 IEEE International Geoscience and Remote Sensing Symposium IGARSS*, pages 411–414, 2021.

[6] Giampaolo Ferraioli, Charles-Alban Deledalle, Loïc Denis, and Florence Tupin. PA-RISAR : Patch-based estimation and regularized inversion for multi-baseline SAR interferometry. *IEEE Transactions on Geoscience and Remote Sensing*, 56(3) :1626–1636, March 2018.

[7] Joy Hsu, Jeffrey Gu, Gong-Her Wu, Wah Chiu, and Serena Yeung. Capturing implicit hierarchical structure in 3D biomedical images with self-supervised hyperbolic representations, 2021. arXiv 2012.01644.

[8] Gang Liu, Yann Gousseau, and Florence Tupin. A contrario comparison of local descriptors for change detection in very high spatial resolution satellite images of urban areas. *IEEE Transactions on Geoscience and Remote Sensing*, 57(6) :3904–3918, 2019.

[9] A. Newson, A. Almansa, Y. Gousseau, and S. Ladjal. Taking apart auto-encoders, how do they encode geometric shapes ?, 2018. HAL preprint hal-01676326.

[10] Alon Oring, Zohar Yakhini, and Yacov Hel-Or. Autoencoder image interpolation by shaping the latent space. In *International Conference on Machine Learning*, pages 8281–8290. PMLR, 2021.

[11] Clement Rambour, Loic Denis, Florence Tupin, Helene Oriot, Yue Huang, and Laurent Ferro-Famil. Urban surface reconstruction in SAR tomography by graph-cuts. *Computer Vision and Image Understanding*, 188 :102791, 2019.

[12] Mateus Riva, Pietro Gori, Florian Yger, Roberto Cesar, and Isabelle Bloch. Approximation of dilation-based spatial relations to add structural constraints in neural networks, 2021. arXiv 2102.10923.

[13] X. Su, C. Deledalle, F. Tupin, and H. Sun. NORCAMA : Change analysis in SAR time series by likelihood ratio change matrix clustering. *ISPRS Journal of Photogrammetry and Remote Sensing*, 2015.

[14] X. Yao, A. Newson, Y. Gousseau, and P. Hellier. A Latent Transformer for Disentangled Face Editing in Images and Videos. *ICCV*, 2021.