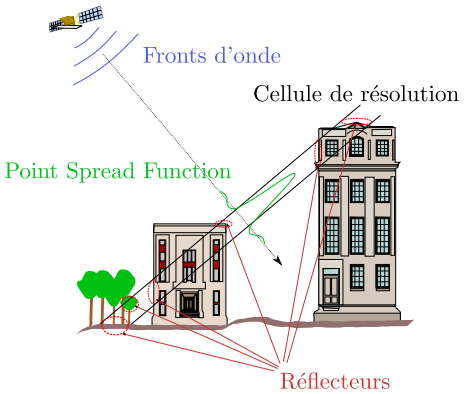
Apprentissage profond et espaces de représentation pour la reconstruction 3D en imagerie de télédétection multi-capteurs

1 Objet de la thèse

L'objectif de cette thèse est de développer un cadre générique et robuste pour la reconstruction 3D de scènes en télédétection pouvant être alimenté par diverses données (capteurs actif/passif, de différentes modalités, résolutions, dates d'acquisition). Ce cadre sera développé en s'appuyant sur les espaces de représentation appris par des réseaux de neurones profonds en intégrant des a priori physiques et structurels.

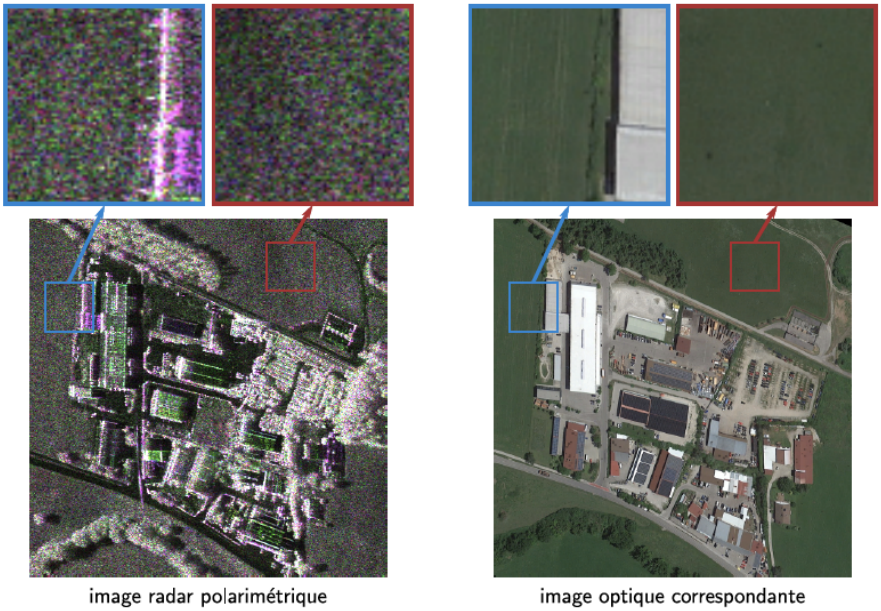
2 Descriptif de la thèse

L'intelligence artificielle (IA) et les approches par apprentissage ont révolutionné la vision par ordinateur et le traitement d'images ces dernières années. Le domaine de la télédétection, qui exploite des images satellitaires ou aériennes pour l'observation de la terre, profite lui aussi de l'efficacité de ces approches qui ont conduit à de nombreuses avancées. Néanmoins, une application directe des méthodes d'apprentissage se heurte à de multiples difficultés. Tout d'abord, le domaine de la télédétection a de fortes spécificités qui nécessitent des adaptations significatives des méthodes d'apprentissage. Par exemple en imagerie SAR (Synthetic Aperture Radar), la focalisation de l'image, les spécificités de l'imagerie cohérente (speckle), ou encore la nature complexe et éventuellement vectorielle (en polarimétrie, interférométrie ou tomographie) des données doivent être prises en compte. Par ailleurs, il existe actuellement peu de jeux de données de télédétection étiquetés, en particulier en imagerie SAR, et la création de tels jeux est particulièrement coûteuse en ressources. D'un point de vue applicatif, le domaine de la télédétection connaît actuellement une très forte évolution avec la mise à disposition de quantités énormes de données avec des fréquences temporelles élevées, comme les données Sentinel-1 et Sentinel-2 de l'agence spatiale européenne. Ainsi, les défis actuels de l'imagerie satellitaire portent sur l'exploitation conjointe de multiples sources de données acquises aussi bien avec des capteurs actifs (comme les SAR) que passifs (capteurs optiques), à différentes résolutions spatiales, spectrales et temporelles.



Principe d'acquisition des images SAR :
après émission d'une onde électro-magnétique, le capteur enregistre l'onde rétrodiffusée qui peut intégrer plusieurs éléments situés à la même distance du capteur.

L'objectif de cette thèse est de construire un cadre de reconstruction 3D de scènes permettant l'interprétation de données satellitaires ou aériennes, voire acquises par drones à partir de données hétérogènes. Pour cela, nous proposons de construire une représentation de haut niveau de la scène qui pourra être utilisée pour différentes tâches, en nous appuyant sur une architecture d'apprentissage profond. Cette représentation pourra être alimentée par différents types de données (données optiques, SAR de différentes modalités -amplitude, interférométrie, tomographie-) pour donner une interprétation de la scène à différents niveaux (information de classes, hauteurs, évolutions temporelles, types de matériaux ou de rétro-diffuseurs, ...).

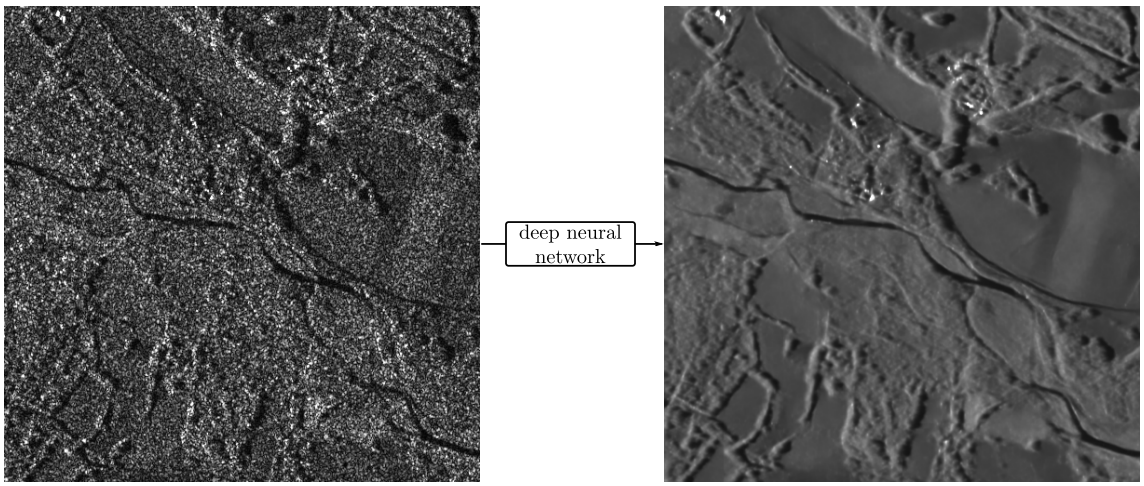


Apparence des objets de la scène pour une image SAR et une image optique

Les verrous méthodologiques à lever sont liés d'une part à l'introduction de contraintes physiques et structurelles dans les réseaux utilisés et d'autre part à la définition d'un modèle de représentation unifié permettant la fusion multi-capteurs. Cette approche est en rupture par rapport aux approches usuelles qui s'intéressent à une tâche particulière et entraînent un réseau spécifique pour la réaliser. Nous proposons plutôt de nous appuyer sur l'efficacité des représentations générées par les réseaux pour développer une représentation générique, invariante au capteur et donc exploitable pour un ensemble de tâches. Cette approche devrait également permettre de progresser dans l'explicabilité des approches d'apprentissage dans le domaine de la télédétection par la construction d'un espace latent interprétable.

3 Programme de la thèse

L'objectif de cette thèse est la construction d'une représentation générique et robuste d'une scène à partir de données multi-capteurs. Nous proposons de nous appuyer sur les espaces de représentations appris par les réseaux d'apprentissage profond, en particulier les auto-encodeurs. Nous décrivons ci-dessous les trois axes méthodologiques qui seront développés dans ces travaux et les applications associées.



L'exploitation d'un modèle physique des images SAR permet d'entraîner, sans vérité terrain (entraînement auto-supervisé), des réseaux à supprimer le speckle [5].

3.1 Exploitation d'algorithmes déroulés pour la reconstruction tomographique en imagerie SAR

La reconstruction tomographique permet de combiner plusieurs acquisitions SAR pour reconstruire des profils d'élévation. Au lieu de retrouver la hauteur d'un seul rétrodiffuseur dans un pixel comme en interférométrie, on peut extraire les contributions de plusieurs rétrodiffuseurs (en milieu urbain) ou le profil complet de rétrodiffusion en hauteur (pour de la végétation par exemple). Plusieurs algorithmes existent, dont des algorithmes d'optimisation d'une fonctionnelle avec contrainte de parcimonie. Nous proposons d'étudier l'intérêt des algorithmes "déroulés" ("unrolling") pour cette tâche et en particulier les méthodes d'équilibre profond "deep equilibrium" [1] pour le TomoSAR. L'approche de DEQ part de l'observation que les architectures de réseaux profonds tendent généralement à converger au fur et à mesure des couches du réseau. Les méthodes de "deep equilibrium" cherchent donc à trouver directement un tel point d'équilibre, et permettent notamment un gain important en termes de mémoire requise pour le modèle. Nous proposons donc ici de coupler les approches de déploiement d'algorithmes existantes en TomoSAR [7] avec celles de "deep equilibrium", ce qui permettrait d'améliorer ces résultats.

3.2 Construction d'espaces de représentation invariants

Un deuxième axe de recherche concerne l'exploitation des architectures des auto-encodeurs, et en particulier les auto-encodeurs variationnels pour construire des espaces de représentations possédant des propriétés d'invariance. Nous proposons dans un premier temps d'étudier comment entraîner un réseau à apprendre des paramètres géométriques simples (comme par exemple la position, la taille et la hauteur de bâtiments isolés) en imagerie optique ou SAR à haute résolution [11]. L'objectif est de rendre ces apprentissages invariants à différents paramètres d'acquisition : la nature (active ou passive) du capteur, la position du capteur par rapport à la scène, son angle d'incidence... Une attention particulière sera portée à la géométrie de l'espace latent ainsi construit et à ses capacités génératives. On pourra s'inspirer des travaux sur le démélange (*disentanglement*) des variables qui influencent la représentation (comme par exemple en édition de visages où les variables de genre, d'âge, de pose du visage sont démélangées dans l'espace latent) [16] ou de fonctions de coût favorisant la structuration de l'espace latent sous la forme d'une variété lisse [12]. Nous pourrons aussi considérer la construction de réseaux de neurones fortement contraints, comme par exemple le PCA-autoencodeur [11], pour mieux démélanger les paramètres d'intérêt de la scène.

Comme mentionné précédemment, cet axe pourra tirer bénéfice des travaux menés actuellement dans l'équipe sur les espaces latents [16, 11] ainsi que sur le débruitage de

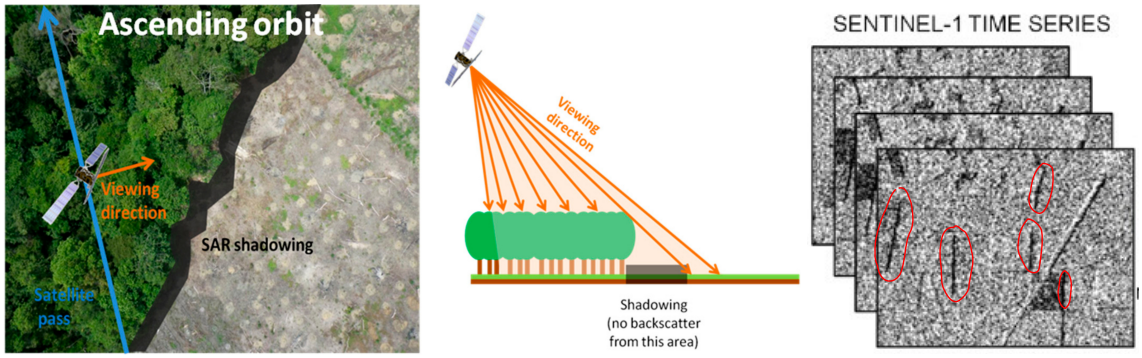
données SAR dans un cadre non supervisé [5, 4]. Par ailleurs, l'expérience de l'équipe en interférométrie [8] et en tomographie radar [13] lui permet d'envisager plusieurs types de contraintes physiques spécifiques. Pour finir, la disponibilité de couples de données optiques et radar à différentes résolutions et dans différentes configurations, ansi que l'accès à des données simulées permettra de travailler sur l'invariance de la représentation au capteur considéré.

L'objectif de cette étape est la définition d'une représentation générique (invariante à la source d'acquisition et à la tâche à mener) qui permettrait de combiner ensuite différentes acquisitions hétérogènes (avec des résolutions spatiales, temporelles et spectrales différentes). Elle pourra être exploitée pour décliner différentes tâches (segmentation sémantique, estimation de l'élévation, débruitage, suivi temporel, etc.) en fonction des sources disponibles.

3.3 Relations spatiales

Un troisième axe de recherche concernera la prise en compte des relations spatiales entre les objets ou les parties des objets de la scène. Ces relations spatiales sont très exploitées dans le domaine de l'imagerie médicale pour laquelle on a une forte connaissance a priori des positions relatives des "objets" les uns par rapport aux autres [2]. Cette information est également structurante en imagerie de télédétection, par exemple avec les positions relatives des zones de repliement et d'ombres en imagerie SAR ou des zones occultées et ombres en imagerie optique. Des travaux récents explorent l'introduction de ces relations selon différentes stratégies : à travers des représentations à base de graphes ou en les intégrant dans des cartes de relations spatiales et en modifiant les fonctions de coût optimisées par le réseau [14]. Si cette information pouvait être exploitée dans le cas d'objets isolés et dans la construction de leur représentation abordée dans l'axe précédent, la prise en compte de relations spatiales structurant la scène urbaine et donc les relations entre objets est également un élément important (comme par exemple l'alignement de bâtiments). On pourrait également exploiter cette information sous forme hiérarchique notamment dans un contexte multi-échelles comme proposé dans [9] pour prendre en compte les aspects multi-résolutions.

3.4 Exemples d'applications



La présence de zones d'ombres est un indicateur fort de la déforestation d'une parcelle dans une image SAR. Illustration adaptée de [3].

Comme mentionné précédemment, l'originalité de l'approche proposée repose sur un travail de fond sur la représentation des données, indépendamment d'applications spécifiques. Néanmoins, les méthodes développées pourront être appliquées pour différents objectifs. En particulier la définition d'une représentation générique, indépendante du capteur pourrait permettre de lever les verrous de la détection de changement multi-capteurs ou même mono-capteur mais avec angles d'incidence différents par exemple en imagerie SAR. Les travaux existants considèrent des angles d'acquisition identiques ou proches [15, 10], et cette situation reste particulièrement difficile à traiter car elle nécessite une

représentation de haut niveau. De nombreux jeux de données sont actuellement disponibles : les données Sentinel-1 et Sentinel-2, librement accessibles avec de haute fréquence de répétition temporelle mais une résolution modérée. L'équipe possède également de nombreux jeux de données SAR à plus haute résolution acquis avec TerraSAR-X, CosmoSky-Med ou RadarSat-2 dans différentes modalités (Strip Map, Spotlight, Staring Spotlight, polarimétrie) avec des configurations interférométriques et tomographiques.

Un autre exemple d'application est la mise en correspondance optique / radar et l'interprétation conjointe de scènes avec prise en compte de la géométrie 3D. On peut imaginer une approche de mise à jour incrémentale de l'interprétation au sens large de la scène chaque fois qu'une nouvelle donnée est acquise, soit pour améliorer l'interprétation courante, soit pour la mettre à jour si une évolution a eu lieu.

3.5 Planning prévisionnel de la thèse

La première année sera consacrée à la compréhension de l'imagerie optique et SAR et aux modélisations mathématiques et physiques qui leur sont associées (modélisation des capteurs, synthèse SAR, modèles statistiques, ...). Les travaux en tomographie SAR avec des aproches de type deep equilibrium seront menés. Un axe de recherche sur les auto-encodeurs et les approches de type adaptation de domaine [8] sera mené, en lien avec l'introduction de connaissances physiques et structurelles sur l'acquisition.

La deuxième année sera consacrée aux aspects multi-capteurs et la construction de caractéristiques invariantes. L'utilisation d'un cadre multi-tâches (segmentation sémantique, carte d'élévations, rétro-diffuseurs / matériaux) servira de point de départ à cet axe. Des jeux de données hétérogènes (SAR, optique, multi-résolutions, et angles) disponibles dans l'équipe seront exploités pour développer et valider la représentation construite.

La troisième année sera consacrée au développement d'applications à partir de cette représentation. Deux grands domaines sont envisagés : la caractérisation et la reconstruction tri-dimensionnelle en milieu urbain multi-capteurs et le suivi temporel de milieux urbains ou naturels (forêts, réseau hydrologique).

4 Encadrement

L'équipe IMAGES de Télécom Paris a une longue expérience en imagerie de télédétection [5, 6, 4, 10, 11] (`https://perso.telecom-paristech.fr/tupin/radarteam/staffEN.php`) qui est internationalement reconnue. La thèse sera réalisée sous la direction de Florence Tupin, professeure dans cette équipe, en co-encadrement avec Christophe Kervazo (maître de conférences dans l'équipe IMAGES) et Loïc Denis de Télécom Saint-Etienne (Laboratoire Hubert Curien, Univ. de Saint-Etienne / CNRS / Institut d'Optique Graduate School) qui est un professeur invité de Télécom Paris, collaborateur de longue date de l'équipe et dont les travaux en imagerie non conventionnelle (astronomie, holographie, imagerie SAR) sont internationalement reconnus.

Par ailleurs, cette thèse s'inscrit dans le cadre du projet ALIA du CIEDS (centre inter-disciplinaire pour la défense et la sécurité de l'Institut Polytechnique de Paris) dont l'objectif est de développer des méthodes d'apprentissage en imagerie de télédétection. Elle donnera potentiellement lieu à des collaborations avec Yann Gousseau et Saïd Ladjal dans le cadre de ce projet, ainsi qu'avec les deux autres doctorants/doctorantes et stagiaires qui seront recrutés.

Références

[1] Shaojie Bai, J. Zico Kolter, and Vladlen Koltun. Deep equilibrium models, 2019.

[2] 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.

[3] 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.

[4] 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.

[5] 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.

[6] 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.

[7] K. Qian et al. γ-Net : Superresolving SAR Tomographic Inversion via Deep Learning, 2021. arXiv 2112.04211v1.

[8] 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.

[9] 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.

[10] 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.

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

[12] 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.

[13] 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.

[14] 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.

[15] 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.

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