Débruitage multi-temporel d'images radar à synthèse d'ouverture par apprentissage profond auto-supervisé

Inès MERAOUMIA[1], Emanuele DALSASSO[1], Loïc DENIS[2], Florence TUPIN[1]

[1]LTCI, Télécom Paris, Institut Polytechnique de Paris, Palaiseau, France

[2]Univ Lyon, UJM-Saint-Etienne, CNRS, Institut d Optique Graduate School, Laboratoire Hubert Curien UMR 5516, F-42023, SAINT-ETIENNE, France

ines.meraoumia@telecom-paris.fr, emanuele.dalsasso@telecom-paris.fr
loic.denis@univ-st-etienne.fr, florence.tupin@telecom-paris.fr

Résumé – Les satellites imageurs radar (SAR) représentent une modalité très utilisée pour l'observation de la terre, fournissant à chaque revisite une nouvelle image de la zone d'intérêt. L'interprétation des images SAR est cependant difficile à cause du phénomène de chatoiement: des fluctuations apparaissent dans l'image, d'autant plus fortes là où la réflectivité radar est élevée. Un grand nombre de méthodes de réduction du chatoiement ont donc été développées. Nous proposons ici une approche d'apprentissage profond présentant deux originalités: 1) l'exploitation d'une *série temporelle* d'images afin d'améliorer la restauration d'une image d'intérêt et 2) l'entraînement *sans référence* du réseau de neurones.

Abstract – Synthetic aperture radar (SAR) satellites represent a widely used modality for Earth observation, providing a new image of the area of interest at each revisit. However, the interpretation of SAR images is difficult because of the speckle phenomenon: fluctuations appear in the image, stronger in areas where the radar reflectivity is high. A large number of speckle reduction methods have therefore been developed in the literature. We propose here a deep learning approach with two original features: 1) the exploitation of a *time series* of images in order to improve the restoration of an image of interest and 2) the *self-supervised* training of the neural network.

1 Introduction

Les images SAR (Synthetic Aperture Radar) sont acquises par des satellites munis de capteurs dits "actifs": une onde électromagnétique est émise par le capteur, puis la réponse du milieu observé est mesurée par ce même capteur. L'image SAR synthétisée retranscrit les propriétés physiques de la scène. Ces images sont particulières car ce sont des données à valeurs complexes comportant des informations d'amplitude, de phase et éventuellement de polarisation. Contrairement aux capteurs optiques, les images SAR peuvent être exploitées indépendamment de la météo, rendant l'observation en continu de la planète possible, tout particulièrement les zones aux climats humides et nuageux. Le survol d'un même point de la terre par le satellite se répète périodiquement. Ainsi, des piles multi-temporelles peuvent être générées : elles sont constituées d'images d'une même zone, acquises à différentes dates. Ces images contiennent des informations sur l'évolution de la zone au cours du temps. La détection ou l'identification des structures pour lesquels des changements sont survenus est cependant rendue difficile par les fortes fluctuations présentes dans ces images en raison du phénomène de "chatoiement". Ce phénomène est directement lié à la physique d'acquisition : il résulte d'interférences constructives ou destructives entre les échos produits par les réflecteurs présents au sein de chaque pixel de l'image SAR.

La synthèse des images SAR se termine par un post-traitement propre à chaque capteur. Cette étape a pour but d'améliorer l'image en réduisant les lobes secondaires des forts réflecteurs dus à la bande passante limitée du radar. Ce traitement implique généralement l'application d'une fonction d'apodisation sur le spectre de l'image. On notera $\mathbf{H}$ l'opérateur linéaire correspondant à ce post-traitement. L'application de ce filtrage a néanmoins pour effet de corréler spatialement le chatoiement, ce qui représente une difficulté supplémentaire pour la tâche de réduction du chatoiement.

L'objectif de cet article est de proposer une approche exploitant une piles d'images correspondant à une série temporelle d'observations d'une même zone afin de réduire le phénomène de chatoiement. Le cadre proposé s'appuie sur un réseau profond entraîné de façon auto-supervisée.

2 Réduction du chatoiement

2.1 Statistiques et propriétés du chatoiement

L'imagerie SAR donne accès en chaque pixel à une amplitude complexe z, correspondant à la résultante des échos radar produits par chacun des réflecteurs élémentaires. La distribution de z suit une loi Gaussienne complexe circulaire de la forme ci-dessous (modèle de Goodman [1]):

$$\mathrm{p}_Z(z) = \frac{1}{\pi v}\exp\left(\frac{-|z|^2}{v}\right) \quad (1)$$

où $v = \mathbb{E}[|Z|^2]$ correspond à la réflectivité radar. On peut réécrire l'amplitude complexe z en faisant apparaître explicite-

ment sa partie réelle a et sa partie imaginaire b : $z = a + ib$. Cette décomposition permet de montrer que a et b sont indépendants et suivent une même loi Gaussienne centrée en 0 et de variance $v/2$.

Afin de se placer dans un cadre proche du traitement d'images traditionnel, on considère l'image intensité $w = |z|^2$. On peut relier l'image d'intensité w, l'image des réflectivités radar v (non corrompue par le phénomène de chatoiement) et un terme de pur chatoiement u via le modèle multiplicatif:

$$w = v \times u \tag{2}$$

Si les réflectivités v sont considérées comme déterministes, alors les intensités w sont statistiquement indépendantes d'un pixel à l'autre avant l'application de l'apodisation spectrale.

2.2 Restauration des images SAR

2.2.1 D'une approche mono-image vers un débruitage multi-temporel

La tâche de débruitage des images SAR (c'est à dire, de réduction du chatoiement) est un véritable défi, du fait des propriétés statistiques du chatoiement qui diffèrent des modèles classiques employés dans la littérature de débruitage des images naturelles, mais également à cause de la corrélation spatiale de ce chatoiement qui représente une difficulté supplémentaire. Les méthodes traditionnelles mono-image nécessitent généralement une étape préliminaire de blanchiment ou a minima de sous-échantillonnage afin de réduire cette corrélation spatiale. Les approches très récentes par apprentissage profond en mono-image SAR2SAR [2] et MERLIN [3] sont, elles, robustes aux corrélations spatiales du chatoiement et sont donc applicables sans sous-échantillonnage préalable.

Lorsque de multiples images d'une même zone sont disponibles, il peut être bénéfique de les combiner lors de l'étape de réduction du chatoiement. Cela nécessite néanmoins un traitement robuste aux changements temporels survenus d'une image à l'autre. Le débruitage multi-temporel a ainsi pour but d'exploiter la redondance d'information présente dans la pile d'images. La littérature actuelle contient principalement des méthodes ne faisant pas appel à un apprentissage profond pour résoudre ce problème.

Le filtre de Quegan [4] se base sur un moyennage après la compensation des changements. Ses performances dépendent directement du nombre d'images disponibles. L'approche [5] propose une méthode de filtrage par bloc de patches similaires pouvant être identifiés au sein de la pile temporelle, elle repose principalement sur la méthode de filtrage mono-image SAR-BM3D [6]. La technique développée dans [7] réalise une moyenne pondérée par la similarité entre les patches sélectionnés dans un premier temps selon la dimension temporelle, puis dans un second temps dans les dimensions spatiales. RABASAR [8] est algorithme basé sur le débruitage du quotient entre l'image à débruiter et une "super-image" obtenue par débruitage de la moyenne temporelle de toute la pile (cette "super-image" présente une très bon rapport signal-sur-bruit car le moyennage temporel réduit fortement les fluctuations du chatoiement). Un sous-échantillonage (ou un blanchiment) est nécessaire en pré-traitement pour l'ensemble des méthodes présentées dans ce paragraphe sans quoi des artefacts importants apparaissent dans les images filtrées.

2.2.2 Importance d'un entraînement auto-supervisé

Les approches par apprentissage profond ont récemment permis de repousser les performances des méthodes de débruitage d'images. Principalement basées sur un apprentissage supervisé, elles sont difficilement transposables aux images SAR pour lesquelles il est très difficile d'accéder à l'image des réflectivités radar v (la "vérité terrain"). En effet, le chatoiement observé est lié à la physique de la scène, ainsi, il est quasiment impossible de mesurer directement une image de la réflectivité radar de la scène sans chatoiement. Les méthodes d'apprentissage supervisé sont naturellement limitées par l'utilisation d'une vérité terrain approximative. Celle-ci est souvent construite à partir de la moyenne temporelle d'une grande pile d'images. Cette "super-image" dans laquelle les différents changements ainsi que le chatoiement ont été moyennés présente un très bon rapport signal sur bruit. Un chatoiement synthétique peut alors être introduit en utilisant le modèle statistique de Goodman (équations (1) et (2)) pour produire les paires d'images corrompues et de référence nécessaires à un entraînement supervisé. L'absence de modélisation de l'effet de la pondération spectrale correspondant au post-traitement $\mathbf{H}$ implique que le réseau obtenu ne peut être appliqué que des images radar sous-échantillonnées ou blanchies.

Le réseau SAR2SAR [2] de débruitage mono-image est obtenu au terme d'un apprentissage semi-supervisé: en sélectionnant des paires de véritables images radar (à la différence des images dont le chatoiement est synthétique), dont les changements temporels ont été compensés, l'entraînement est réalisé directement sur des images corrélées spatialement. L'approche MERLIN [3] quant à elle est une méthode d'apprentissage profond auto-supervisée fournissant une estimation des réflectivités à partir de la décomposition en parties réelle et imaginaire de l'image SAR.

Cet article propose une extension multi-temporelle de l'approche MERLIN, permettant d'exploiter également la redondance temporelle contenue dans les piles d'images SAR afin d'améliorer la performance du débruitage, tout en étant robuste aux corrélations spatiales du chatoiement (grâce à l'entraînement auto-supervisé sur des véritables images radar).

3 Approche proposée

3.1 Extension multi-date du cadre MERLIN

L'approche MERLIN [3] découle de la même idée que Noise2Noise [9]: la possibilité d'entraîner un réseau de débruitage uniquement à partir d'une collection de paires d'images bruitées, chaque paire correspondant à deux réalisations

indépendantes et identiquement distribuées, c'est à dire deux versions bruitées d'une même scène.

Les parties réelle et imaginaire de l'image des amplitudes complexes z sont indépendantes sous certaines conditions discutées ci-dessous. L'approche MERLIN (coMplex sElf-supeRvised despeckLINg) consiste à entraîner le réseau à produire une estimation des réflectivités v seulement à partir de a^2 et à utiliser la composante b^2 afin de définir la fonction de coût (le rôle de a et b peut être échangé lors de l'entraînement).

L'approche proposée dans cet article est une extension multi-temporelle de cette stratégie d'entraînement auto-supervisé: le débruitage mono-image est étendu en ajoutant des canaux supplémentaires en entrée du réseau afin d'injecter également les images observées aux autres dates. Cette approche *Multi-Input Single-Output* (MISO) est naturellement robuste à la corrélation spatiale du chatoiement car l'entraînement est réalisé sur des images présentant ce type de corrélations.

Soit $z_1, ..., z_T$ une pile de T images. Soit N un entier tel que $N \geq 2$. On appellera "réseau MISO$_N$", le réseau ayant N images en entrée et une unique sortie.

Lors de la phase d'entraînement, les patches fournis en entrée du réseau MISO$_N$ sont sélectionnés de la manière suivante: une date de référence $1 \leq t_{\text{ref}} \leq T$ est fixée (cela correspond à la date pour laquelle le réseau doit prédire la réflectivité en sortie) et $N-1$ dates additionnelles distinctes les unes des autres et différentes de t_{ref} sont tirées aléatoirement (le tirage aléatoire permet d'obtenir un réseau quasi-invariant aux permutations des entrées).

La fonction de coût utilisée pour l'entraînement auto-supervisée est la co-log-vraisemblance [3]:

$$\mathcal{L}(\hat{u}_{\text{ref}}, b_{\text{ref}}) = \sum_k \frac{1}{2}\log(\hat{u}_{\text{ref},k}) + \frac{b^2_{\text{ref},k}}{\hat{u}_{\text{ref},k}} \qquad (3)$$

avec k un pixel, $\hat{u}_{\text{ref}}$ l'image des réflectivités estimées pour la date t_{ref} à partir de la composante "partie réelle" a_{ref} et des images correspondant aux dates additionnelles, b_{ref} la composante "partie image" à la date t_{ref}. Ainsi, le réseau est récompensé lorsque sa sortie est vraisemblable vis à vis de la composante "partie imaginaire" dont le chatoiement est indépendant du chatoiement présent dans les différents canaux d'entrée.

Puisque le réseau est entraîné à n'exploiter que l'information présente dans la composante "partie réelle" (ou "partie imaginaire", le rôle de a_{ref} et b_{ref} pouvant être interverti dans la fonction de coût $\mathcal{L}$), en phase de test deux estimations sont réalisées, correspondant aux deux configurations possibles: une première image débruitée $\hat{u}_{\text{ref}}$ est obtenue à partir de a^2_{ref} puis une seconde estimation est produite à partir de b^2_{ref}. Ces deux estimations sont combinées pour produire l'estimation finale (Fig.1).

3.2 Mise en œuvre

On suppose les images de la pile multi-temporelle préalablement recalées. L'apprentissage auto-supervisé n'est possible que si le chatoiement de l'observation utilisée dans la fonction de coût $\mathcal{L}$ est indépendant du chatoiement présent dans les différentes images d'entrée, on supposera cette condition vérifiée

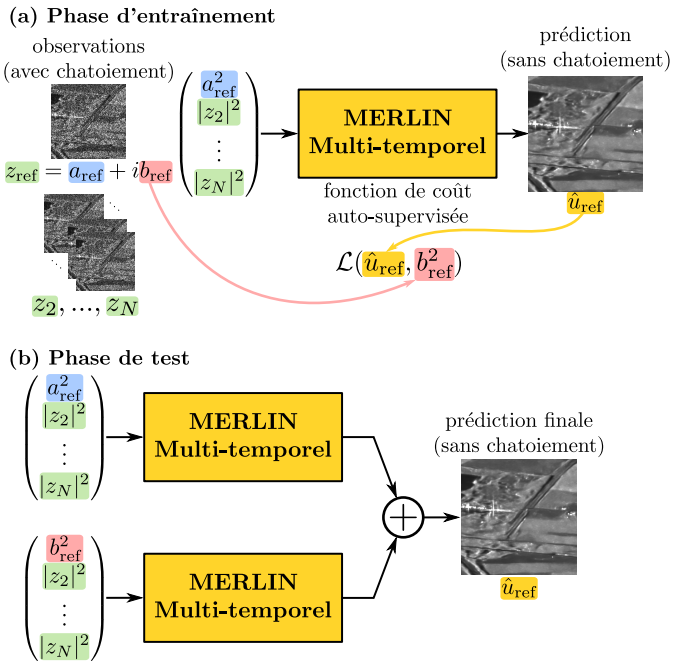

Figure 1: Principe de l'entraînement auto-supervisé

en s'assurant que les dates sont suffisamment espacées pour que le chatoiement soit temporellement décorrélé (c'est à dire incohérent). Pour garantir l'indépendance entre la partie réelle et imaginaire, une condition suffisante est que le prétraitement par l'opérateur **H** (l'apodisation spectrale) soit à valeurs réelles, c'est à dire que la fonction de transfert correspondante soit à symétrie hermitienne. On effectuera donc un recentrage, en range et en azimuth, du spectre de l'image autour de zéro avant l'ingestion des données par les réseaux MISO (on place l'image au zéro Doppler).

Une transformée logarithmique et une normalisation est appliquée aux images en entrée du réseau. La transformation logarithmique stabilise la variance du chatoiement (elle devient indépendente de la réflectivité) et compresse la dynamique des images.

4 Résultats expérimentaux

4.1 Bruit simulé

L'intérêt du filtrage multi-temporel est mis en évidence sur des images pour lesquelles le chatoiement est simulé (ce qui permet de calculer une distance à une image de référence sans chatoiement). Les images vérités terrain ont été construites en utilisant la méthode présentée dans [10]. Le jeu de données d'entraînement est composé de 5 piles Sentinel-1 débruitées. Le bruit est simulé selon le modèle de Goodman eq. (2). On s'intéresse aux valeurs de $1 \leq N \leq 5$ (le cas $N = 1$ correspondant à un filtrage mono-date).

La figure 2 montre que le PSNR moyen augmente lorsque le nombre de canaux additionnels augmente. Un gain de 3 dB est

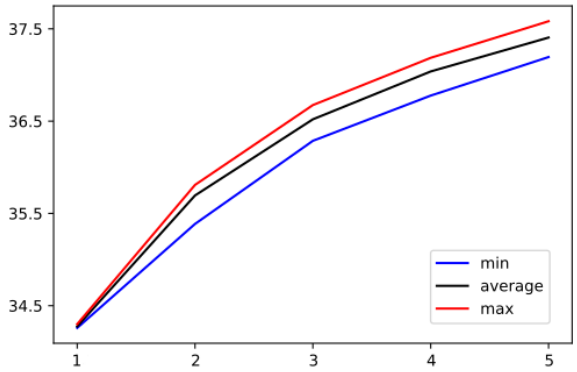

Figure 2: Evolution du PSNR en fonction du nombre N d'entrées du réseau (test sur l'ensemble des piles d'entraînement).

observé entre les résultats obtenus avec l'approche mono-date et le réseau MISO_5.

4.2 Résultats sur des images TSX

La validation de notre méthode est réalisée en ré-entraînant les réseaux sur 2 piles d'images de la ville de Saint-Gervais acquises par le satellite allemand TerraSAR-X en mode Stripmap. Chaque pile est constituée de 26 images de taille 1024x1024 pixels.

Les résultats de la figure 3 montrent une meilleure reconstruction des structures fines qui sont stables dans le temps. Une meilleure qualité d'image semble être obtenue par comparaison aux méthodes de filtrage multi-temporel de l'état de l'art [5, 7].

5 Conclusion

L'approche proposée permet une réduction du chatoiement en exploitant des séries temporelles, et ce sans qu'une vérité terrain ne soit nécessaire lors de la phase d'entraînement.

Les différents réseaux présentés sont adaptés à la corrélation spatiale du chatoiement. En augmentant le nombre de canaux secondaires, les valeurs de PSNR augmentent, les structures fines sont beaucoup mieux préservées et les images débruitées contiennent plus de détails. Le nombre maximum d'images en entrée du réseau a été fixé à 20 dans cette étude pour des raisons de consommation mémoire sur la carte graphique, mais les résultats présentés semblent montrer qu'une asymptote en terme de qualité de restauration est rapidement atteinte lorsqu'on ajoute des dates supplémentaires en entrée du réseau.

References

[1] J. W. Goodman, *Speckle phenomena in optics: theory and applications*. Roberts and Company Publishers, 2007.

[2] E. Dalsasso, L. Denis, and F. Tupin, "SAR2SAR: a semi-supervised despeckling algorithm for SAR images," *IEEE JSTARS*, 2021.

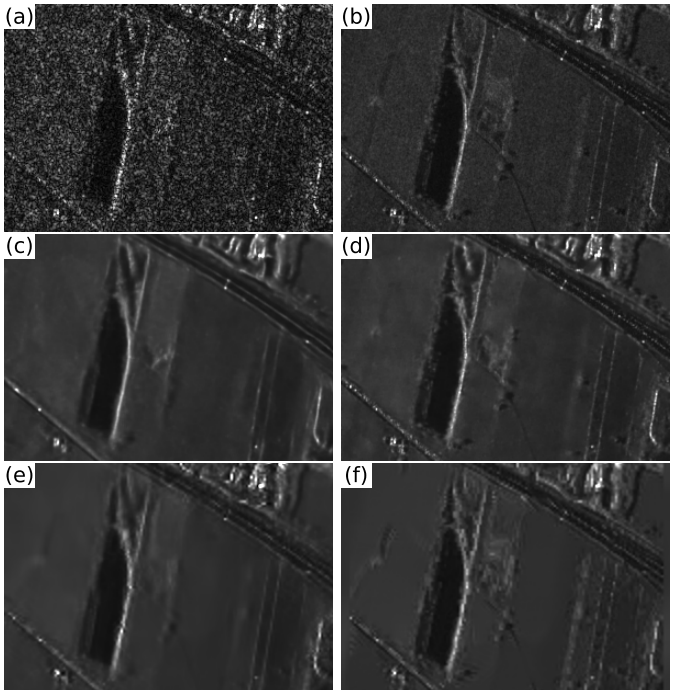

Figure 3: (a) Image bruitée; (b) Moyenne temporelle de la pile d'images; (c) MERLIN [3]; (d) MISO_{16}; (e) MSAR-BM3D [5]; (f) Multi-temporal NL-Means [7].

[3] ——, "As if by magic: self-supervised training of deep despeckling networks with merlin," *IEEE TGRS*, 2021.

[4] S. Quegan and J. J. Yu, "Filtering of multichannel sar images," *IEEE TGRS*, 2001.

[5] G. Chierchia, M. El Gheche, G. Scarpa, and L. Verdoliva, "Multitemporal sar image despeckling based on block-matching and collaborative filtering," *IEEE TGRS*, 2017.

[6] S. Parrilli, M. Poderico, C. V. Angelino, and L. Verdoliva, "A nonlocal SAR image denoising algorithm based on LLMMSE wavelet shrinkage," *IEEE TGRS*, 2011.

[7] X. Su, C.-A. Deledalle, F. Tupin, and H. Sun, "Two-step multitemporal nonlocal means for synthetic aperture radar images," *IEEE TGRS*, 2014.

[8] W. Zhao, C.-A. Deledalle, L. Denis, H. Maître, J.-M. Nicolas, and F. Tupin, "Ratio-Based Multitemporal SAR Images Denoising: RABASAR," *IEEE TGRS*, 2019.

[9] J. Lehtinen, J. Munkberg, J. Hasselgren, S. Laine, T. Karras, M. Aittala, and T. Aila, "Noise2Noise: Learning Image Restoration without Clean Data," in *ICML*, 2018.

[10] E. Dalsasso, I. Meraoumia, L. Denis, and F. Tupin, "Exploiting multi-temporal information for improved speckle reduction of Sentinel-1 SAR images by deep learning," in *IGARSS* , 2021.