Impact de la stratégie de décodage sur la traduction de modalité radar-optique d'images de télédétection

Antoine BRALET[1] Abdourrahmane ATTO[1] Jocelyn CHANUSSOT[2] Emmanuel TROUVÉ[1]

[1]LISTIC, Université Savoie Mont Blanc, 74940 Annecy, France

[2]GIPSA-Lab, Univ. Grenoble Alpes, CNRS, Grenoble INP, 38000 Grenoble, France

Résumé – Pour de nombreuses applications en vision par ordinateur, des approches d'apprentissage profond basées sur une architecture encodeur-décodeur ont été proposées. En imagerie de télédétection, cette architecture est notamment employée pour la traduction de modalité radar vers optique afin de disposer d'images optiques indépendamment des conditions météorologiques. L'architecture neuronale doit donc prendre en compte les distorsions des images radar sans introduire d'artefacts optiques. Cet article étudie l'impact de la stratégie de décodage sur la reconstruction résultant de notre traducteur radar-optique : SARDINet. Trois stratégies sont comparées ici : la convolution post-sur-échantillonnage, la convolution transposée et la convolution sub-pixellique.

Abstract – Numerous computer vision applications are tackled by deep learning approaches with an encoder-decoder architecture. In remote sensing imagery, this architecture is leveraged for radar to optical image translation to ensure all-weather availability of optical images. The network should not only correct radar distortions but also generate artifact-free optical images. This paper focuses on the impact of the decoding strategy on the reconstruction resulting from our radar-to-optical translator : SARDINet. Three strategies are compared here : post-upsampling convolutions, transposed convolutions and sub-pixel convolutions.

1 Introduction

La traduction de modalité est une des nombreuses applications pour lesquelles les architectures neuronales encodeur-décodeur améliorent significativement les résultats. L'encodeur permet d'extraire les caractéristiques pertinentes des images d'entrée puis le décodeur les traduit, projette ou encore transfère dans l'espace des images de sortie. En particulier, en télédétection, la traduction radar-optique permet de rendre les images radar interprétables et d'obtenir une image optique indépendamment des conditions météorologiques. L'architecture doit donc identifier et corriger les distorsions liées aux acquisitions radar et reconstruire une image optique sans artefacts. De fait, le décodeur se base sur des caractéristiques "basse résolution" extraites afin de générer une image haute résolution nettoyée, corrigée et crédible. Dans ces travaux, nous comparons trois stratégies de décodage permettant de réaliser ce processus pour étudier leur impact sur la reconstruction de notre traducteur radar-optique : SARDINet [2]. Les trois stratégies en question sont : la convolution post-sur-échantillonnage, la convolution transposée et la convolution sub-pixellique [8, 1].

L'article est structuré de la façon suivante : la Section 2 présente l'état de l'art sur la traduction de modalité radar-optique et sur les stratégies de décodage. La Section 3 expose les modifications appliquées à SARDINet et la Section 4 présente les données utilisées pour obtenir les résultats de la Section 5. La Section 6 conclut et donne les perspectives de ces travaux.

2 État-de-l'art

2.1 Traduction de modalité radar-optique

Les approches de traduction de modalité en télédétection sont généralement basées sur des approches adversaires génératives. En témoigne l'étude menée dans [10] qui améliore les performances d'une architecture cGAN [4] à l'aide d'un générateur à deux branches, un discriminateur multi-échelle et une fonction de coût mesurant l'aberration chromatique de la reconstruction. Le réseau mis en place dans [7] est basé sur une approche adversaire générative non supervisée : CycleGAN [11]. La contribution majeure réside dans l'alignement des caractéristiques latentes des deux traducteurs radar-optique et optique-radar. L'approche proposée dans [5] vise à faciliter le travail du générateur en reconstruisant la décomposition en ondelettes de l'image optique. Un second réseau permet ensuite de coloriser l'image. Une approche temporelle est mise en place dans [6] afin de prendre en compte des connaissances passées pour améliorer la reconstruction optique. Cette étude a également conclu à une dégradation des métriques quantitatives lors d'un entraînement adversaire. Nous avons aussi pu le constater dans [2] en implémentant SARDINet (*SAR Distorted Image translator Network*), un traducteur radar-optique surpassant les distorsions géométriques dans les images radar. Celui-ci est détaillé en Section 3.1.

2.2 Stratégies de décodage

Les approches précédentes se basent sur une architecture de type encodeur-décodeur. Les décodeurs en question sont construits sur une succession de couches convolutionnelles et de mise à l'échelle visant à générer une image haute résolution à partir de caractéristiques basse résolution. Pour ce faire, deux stratégies sont généralement envisagées : appliquer des convolutions post-sur-échantillonnage ou des convolutions transposées. La première (Figure 1a) consiste à sur-échantillonner les caractéristiques basse résolution par interpolation - plus proche voisin, bilinéaire, bicubique, etc. - puis à appliquer des convolutions 2D. La seconde (Figure 1b) sur-échantillonne les caractéristiques basse résolution en intercalant entre chaque

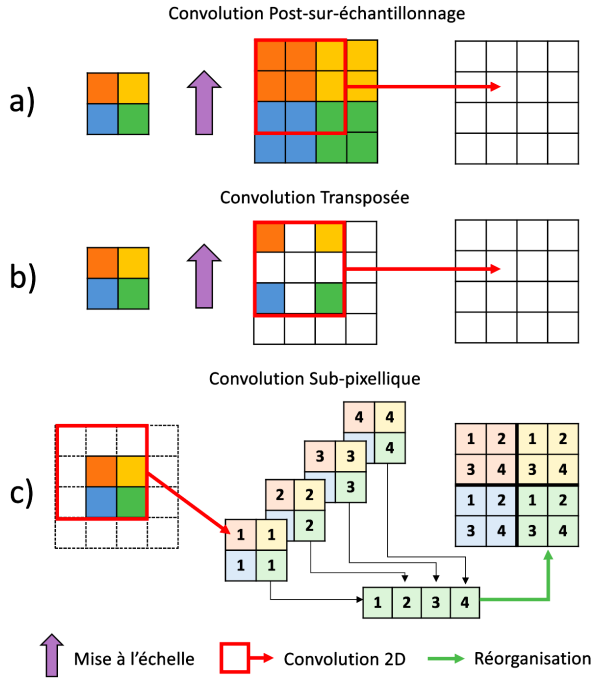

FIGURE 1 : Stratégies comparées afin de décoder l'espace latent et reconstruire une image optique. L'exemple est donné ici pour une mise à l'échelle d'un facteur deux.

pixel existant un pixel nul. Ces derniers sont ensuite déterminés numériquement par l'application de convolution 2D. Une troisième approche issue de la super-résolution peut aussi être exploitée en télédétection : la convolution sub-pixellique [8] (Figure 1c). Pour obtenir une image de résolution k fois plus grande que les caractéristiques basse résolution, il s'agit de convoluer ces dernières avec k^2 noyaux 2D. Les pixels des caractéristiques résultantes sont alors réorganisés pour reconstruire leurs propres voisinages sur l'image haute résolution finale. Ainsi, la valeur du pixel $p^m_{i;j}$ à la position $i;j$ de la caractéristique m est associée à celle du pixel $p_{ki+q;kj+r}$ de l'image haute résolution - avec q et r respectivement le quotient et le reste de la division euclidienne de m par k. Afin d'éviter les effets de damiers, les auteurs de [1] proposent d'initialiser les noyaux de convolution de façon identique.

3 Méthode

3.1 Traducteur radar-optique : SARDINet

Le traducteur de modalité sélectionné comme référence est un réseau que nous avons précédemment implémenté dans [2] : SARDINet. Il a été développé afin de pouvoir contourner les problématiques de distorsions géométriques suite à l'acquisition radar. Celles-ci peuvent apparaître si des objets 3D volumineux sont imagés (*e.g.* des silos dans [2]) mais également si les pentes de la zone d'étude sont trop fortes comme ce peut être le cas en milieu montagneux. Cette étude permet donc aussi de tester sa robustesse dans des milieux naturels végétalisés induisant des mécanismes de rétrodiffusion supplémentaires dans les images radar.

SARDINet est une architecture encodeur-décodeur. L'encodeur se divise en trois branches. Si la première extrait des caractéristiques globales, les deux autres extraient des caractéristiques plus fines à un niveau de résolution différent. L'implémentation de convolutions séparables [3] permet une extraction successive de caractéristiques spatiales et inter-canal. L'espace latent est alors formé par la somme des caractéristiques extraites par chaque branche puis fourni à un décodeur à trois branches indépendantes. Chaque branche est dédiée à la reconstruction d'un seul canal et est composée de convolutions post-sur-échantillonnage sur trois étages successifs. C'est cette stratégie de décodage que nous modifions en Section 3.2 pour étudier son impact sur la reconstruction finale.

Dans ces travaux, nous nous concentrons sur une entrée bi-canal : VV et VH pour reconstruire une image RGB comme décrit en Section 4. Nous avons également choisi de symétriser le nombre de noyaux entre l'encodeur et le décodeur pour permettre une plus grande variété de motifs reconstructibles.

3.2 Décodeurs mis en place

En Section 5.2, nous comparons les résultats obtenus en modifiant la stratégie employée dans le décodeur. Cinq variantes sont implémentées avec différents types de convolutions :

- SARDINet$_{up}$ correspond à la convolution post-sur-échantillonnage - donc l'architecture originale de SARDINet,
- SARDINet$_{tr}$ correspond aux convolutions transposées,
- SARDINet$^{pix}_{pix}$ correspond aux convolutions sub-pixelliques.
- SARDINet$^{pix}_{up}$ correspond à la convolution post-sur-échantillonnage avec une dernière couche de convolution sub-pixellique,
- SARDINet$^{pix}_{tr}$ correspond aux convolutions transposées avec une dernière couche de convolution sub-pixellique.

Le nombre de noyaux utilisés ainsi que le nombre d'étages de mise à l'échelle est le même dans chacune des architectures. Une seule différence est à noter : les architectures SARDINet$^{pix}_{xx}$ résultent directement sur l'image de sortie tandis que les deux autres ont une couche de convolution supplémentaire pour équilibrer la démultiplication du nombre de noyaux requise par la convolution sub-pixellique. Pour assurer des comparaisons équitables, nous avons initialisé les convolutions sub-pixelliques aléatoirement - différemment de [1].

4 Jeu de données

Les réseaux sont entraînés sur un sous-groupe du jeu de données *BigEarthNet-MM* [9]. Ce dernier contient une collection de plus de 590.000 couples d'images radar-optique de taille 120x120 pixels acquises par les satellites Sentinel-1 et Sentinel-2 en Europe de juin 2017 à Mai 2018. Les images de chaque couple ont été choisies aussi proche que possible temporellement parlant pour assurer correspondance et cohérence entre l'entrée du réseau et sa vérité de terrain. Pour ces travaux nous utilisons les polarisations Vertical-Vertical (VV) et Vertical-Horizontal (VH) des images radar afin de reconstruire les canaux Rouge, Vert et Bleu (RGB) des images optiques parmi les neuf disponibles. Pour des raisons de capacité mémoire, nous avons sélectionné aléatoirement un sous-groupe de

195∶109 paires d'images sans présence de nuages ni de neige qui engendrerait de nouvelles problématiques non étudiées ici.

5 Résultats expérimentaux

5.1 Paramètres expérimentaux

Les résultats en Section 5.2 sont obtenus en divisant le jeu de données aléatoirement en deux sous-jeux de données : un jeu d'entraînement avec 80% des couples et un jeu de validation avec les 20% restants - soit respectivement 156∶087 et 39∶022 images. Le réseau a été entraîné avec des lots de 8 images et un taux d'apprentissage de $5{\cdot}10^{-5}$ sur 100 époques. Contrairement à notre approche initiale [2], la fonction de coût est calculée à partir de l'erreur absolue moyenne (MAE) ainsi que la similarité structurelle (SSIM), chaque terme contribuant équitablement et évaluant respectivement la radiométrie générale et la cohérence de la structure des motifs reconstruits. Nous utilisons un optimisateur Adam de paramètres $(\beta_1; \beta_2)$ initialisés respectivement à 0.9 et 0.999. L'évaluation des performances est menée sur le jeu de validation par trois métriques : l'erreur quadratique moyenne (MSE), le rapport signal sur bruit (PSNR) ainsi que la distance de Fréchet (FID). Elles permettent de mesurer respectivement la qualité générale de l'image (couleur, contraste, etc), sa fiabilité et la similarité de sa distribution avec les images de référence.

5.2 Résultats

Les résultats obtenus lors de nos expériences sont visibles quantitativement dans le Tableau 1 et visuellement sur la Figure 2. On peut alors constater que malgré le changement de fonction de coût qui favorise la récupération de détails grâce au terme de similarité structurelle, les résultats visuels de SARDINet$_{up}$ sont toujours flous. Ce flou est d'ailleurs répercuté dans SARDINet$^{pix}_{up}$ malgré l'utilisation finale de convolution sub-pixellique. Ainsi donc le manque de contraste des images observé dans [2] peut certes être expliqué par la fonction de coût basée sur l'erreur quadratique, mais aussi par l'utilisation de convolution post-sur-échantillonnage. Il convient donc de laisser le réseau déterminer sa propre interpolation pour la reconstruction de caractéristiques et pour le rendu final : lui en imposer une tend à perturber sa reconstruction.

Ces expériences nous permettent également de noter que l'utilisation de convolutions sub-pixelliques en dernière couche est particulièrement bénéfique puisque l'on observe pour SARDINet$^{pix}_{up}$ (resp. SARDINet$^{pix}_{tr}$) un gain de 0.82 (resp. 1.44) points de MSE et de 0.09 (resp. 0.03) points de PSNR sur SARDINet$_{up}$ (resp. SARDINet$_{tr}$). Seul le FID est légèrement augmenté de 0.04 points avec SARDINet$^{pix}_{up}$ alors qu'il décroit fortement de 0.82 points pour SARDINet$^{pix}_{tr}$. Ce bénéfice s'observe aussi qualitativement : on peut noter un bien meilleur contraste voire l'apparition de certaines textures de forêt à l'aide de la convolution sub-pixellique - en particulier sur la dernière ligne de la Figure 2.

De plus, au regard des résultats quantitatifs, malgré que SARDINet$^{pix}_{pix}$ ait obtenu les meilleurs résultats en termes de PSNR et de FID, on observe une forte augmentation de la MSE. Celle-ci s'explique qualitativement par la présence d'un effet de damier sur la reconstruction identifiée par des zones cerclées de rouge sur la Figure 2. Néanmoins, il est notable que cet effet n'est pas ou très peu visible sur les images reconstruites par SARDINet$^{pix}_{up}$ et SARDINet$^{pix}_{tr}$. On en conclut que cet effet est accentué par la cascade de convolutions sub-pixelliques.

TABLE 1 : Étude comparative des résultats quantitatifs. Les deux meilleurs résultats sont ordonnés en police **grasse** et soulignée.

	MSE ↓	PSNR ↑	FID ↓
SARDINet$_{up}$	10.69e-3	29.19	<u>6.06</u>
SARDINet$_{tr}$	11.99e-3	29.24	6.21
SARDINet$^{pix}_{pix}$	10.78e-3	**29.34**	**5.39**
SARDINet$^{pix}_{up}$	**9.87e-3**	<u>29.28</u>	6.10
SARDINet$^{pix}_{tr}$	<u>10.55e-3</u>	29.27	**5.39**

Fort de ces résultats, il semble que le meilleur compromis réside en SARDINet$^{pix}_{tr}$. En effet, cette stratégie permet d'obtenir le meilleur FID et second meilleur MSE et le PSNR n'est qu'à 0.01 point du second meilleur PSNR d'après le Tableau 1. De plus, les effets de flou et de damier sont particulièrement estompés sur la Figure 2. On peut alors noter que la convolution transposée permet de reconstruire des caractéristiques pertinentes et bien contrastées, mais que lors de la reconstruction finale elle perd un peu de son efficacité au sens des métriques quantitatives mesurées ; en témoignent les résultats de SARDINet$_{tr}$. De fait, l'ajout de la convolution sub-pixellique en dernière couche permet de réorganiser les caractéristiques extraites et d'en tirer le meilleur parti.

Enfin, en se focalisant sur la dernière ligne de la Figure 2, on peut noter que l'image radar d'entrée contient des variations radiométriques probablement liées à la topographie et à leur dilatation/compression lors de l'orthorectification des images SAR. Notre approche originelle [2] avait démontré les capacités de SARDINet à surpasser les distorsions géométriques en milieu urbain. Cette étude permet d'illustrer son aptitude à tenir compte des artefacts d'orthorectification. L'architecture de SARDINet, à l'aide de son encodeur à trois branches permet effectivement de récupérer des caractéristiques pertinentes sans pour autant être affectée par ces changements radiométriques. Un modèle numérique de terrain serait nécessaire pour confirmer ces hypothèses. Néanmoins, la reconstruction reste fidèle à la vérité de terrain démontrant la robustesse de SARDINet dans les milieux naturels.

6 Conclusion et perspectives

Cette étude a permis d'identifier que le manque de contraste dans nos précédents travaux [2] n'était pas seulement dû à la fonction de coût, mais également à l'utilisation de convolutions post-sur-échantillonnage. Ceci nous pousse donc à explorer l'utilisation d'autres stratégies de décodage consistant à laisser le réseau décider de l'interpolation la plus adéquate. De fait, suite à ces travaux, il est possible de statuer quant à la véritable efficacité des convolutions sub-pixelliques. En particulier lorsqu'elles sont associées à une autre stratégie de décodage et utilisées seulement pour la dernière mise à l'échelle, les résultats sont significativement améliorés et l'effet de damier tout comme le manque de contraste sont réduits. Une étude pa-