Complétion par apprentissage profond de séries temporelles d'images multi-spectrales à partir d'images hyper-spectrales

Cheick T. CISSÉ[1,2], Ahed ALBOODY[1], Matthieu PUIGT[1], Gilles ROUSSEL[1],
Vincent VANTREPOTTE[3], Cédric JAMET[3], Trung-Kien TRAN[3]

[1]Univ. Littoral Côte d'Opale, LISIC – UR 4491, F-62219 Longuenesse France*

[2]Institut FEMTO-ST, Univ. Bourgogne Franche-Comté, CNRS, Belfort, France & Orange Lab, Belfort, France

[3]Univ. Littoral Côte d'Opale, CNRS, LOG – UMR 8187, F-62930 Wimereux, France

cheick.cisse@utbm.fr, Prenom.NOM@univ-littoral.fr

Résumé – Dans cet article, nous nous intéressons à la complétion d'une série temporelle d'images satellitaires multi-spectrales à partir d'une série temporelle d'images satellitaires hyper-spectrales. Nous proposons pour cela une nouvelle approche d'apprentissage profond. Notre contribution principale réside dans la tâche de complétion de l'erreur qui permet d'améliorer la qualité de complétion. Nous montrons que l'approche proposée est capable de fournir des prédictions haute fidélité avec de meilleurs indices de qualités que les approches de la littérature, sur des images réelles issues des bases de données CIA / LGC et Sentinel-2 / Sentinel-3.

Abstract – In this paper, we are interested in the completion of a time series of multi-spectral images from a time series of hyper-spectral ones. To that end, we propose a new deep learning approach to that end. Our main contribution lies in the error completion task which allows to improve the completion performance. We show that our proposed method is able to produce high fidelity predictions with better quality indices than state-of-the-art methods on true images taken from the CIA / LGC database and Sentinel-2 / Sentinel-3 data.

1 Introduction

L'observation satellitaire de notre planète a connu des avancées instrumentales significatives depuis plusieurs décennies. Cependant, afin de maintenir un rapport signal-à-bruit (RSB) constant, l'augmentation du nombre de bandes spectrales dans les imageurs hyperspectraux se fait au détriment de la résolution spatiale. Ainsi, notre planète est aujourd'hui observée par des imageurs multi-spectraux et hyper-spectraux fournissant respectivement une bonne résolution spatiale et spectrale. De plus, la période d'échantillonnage de ces imageurs peut aussi fortement varier. Par exemple, cette période est d'environ 5 jours pour les satellites Sentinel-2 (S-2) contre 1,4 jours pour Sentinel-3 (S-3). Or, pour certaines applications comme l'observation marine côtière, il est nécessaire d'observer le milieu avec une bonne résolution à la fois spatiale, spectrale et temporelle. Malheureusement, aucun satellite aujourd'hui n'est capable de combiner ces trois propriétés. Par exemple, S-2 fournit une résolution spatiale variant de 10 à 60 m pour 13 bandes spectrales alors que S-3 fournit une résolution spatiale de 300 m pour 21 bandes. Alors que la fusion d'images multi- et hyper-spectrales acquises en même temps a été très étudiée [1, 2], la complétion de séries temporelles d'images multi-spectrales à partir d'images hyper-spectrales (aussi connue sous le nom de fusion spatio-temporelle) l'a été beaucoup moins [3]. Nous nous intéressons ici à ce dernier problème.

La littérature est principalement axée sur des méthodes pondérées de filtrage [4], de krigeage [5] et de régression [6]. D'autres techniques sont basées sur le démélange [7] ou sur l'apprentissage de dictionnaires [8]. Enfin plus récemment, des approches fondées sur l'apprentissage profond [9] et/ou sur les réseaux antagonistes génératifs [10] ont été proposées. Dans cet article, nous proposons une nouvelle méthode de complétion de série temporelle d'images multi-spectrales basée sur l'apprentissage profond. La structure de l'article est organisée comme suit. Dans la section 2, nous introduisons le problème considéré et notre méthode. Une validation expérimentale est fournie dans la section 3 alors que nous concluons dans la section 4.

2 Problématique et méthode proposée

Dans cette section, nous introduisons le problème considéré et la méthode développée. Plus précisément, nous considérons deux séries temporelles d'images multi- et hyper-spectrales. Nous supposons par ailleurs que (i) la période d'échantillonnage des images hyper-spectrales est plus petite que celle des images multi-spectrales et que (ii) certaines images multi- et hyper-spectrales sont acquises les mêmes jours. De plus, nous supposons que pour une image multi-spectrale à l'instant t_2,

*Ce travail est financé par la SFR « Campus de la Mer » et par le CNES, dans le cadre du projet TOSCA « OSYNICO ». Les expériences présentées dans cet article ont été en partie réalisées sur la plate-forme de calcul scientifique CALCULCO, gérée par le SCoSI de l'ULCO.

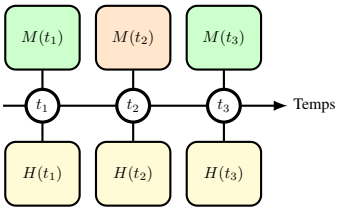

FIGURE 1 – Périodes d'échantillonnage considérées.

il existe deux paires d'images multi- et hyper-spectrales respectivement acquises aux instants t_1 et t_3, et une image hyper-spectrale acquise à l'instant t_2. Plus précisément, nous notons $M(t_i)$ et $H(t_i)$ les images respectivement multi- et hyper-spectrales acquises à l'instant t_i. Ainsi, comme dans [11] et comme on peut le voir sur la figure 1, nous cherchons à estimer $M(t_2)$ à partir de $M(t_1)$, $M(t_3)$, $H(t_1)$, $H(t_2)$ et $H(t_3)$.

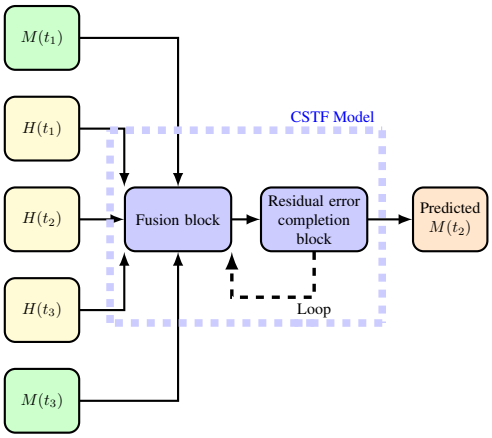

FIGURE 2 – Structure de la méthode proposée.

Pour résoudre ce problème, nous proposons une approche d'apprentissage profond nommée CSTF (pour *Completion Spatio-Temporal Fusion* en anglais). Cette approche est itérative (voir figure 2) et est composée de deux blocs – un bloc de fusion et un bloc de complétion d'erreur résiduelle – qui sont alternativement lancés plusieurs fois. Un bloc de fusion est exécuté une dernière fois pour fournir l'estimation de $M(t_2)$.

Le bloc de fusion est présenté sur la figure 3 et peut être vu comme une extension à 5 images de la stratégie proposée dans [9] qui utilise 3 images, c.-à-d. 2 images hyper-spectrales et 1 image multi-spectrale. Dans notre extension, nous estimons la différence entre des paires d'images hyper-spectrales prises à des instants adjacents, que nous concaténons ensuite avec les images multi-spectrales disponibles pour prédire $M(t_2)$. Pour cela, la fonction de perte considérée est l'erreur quadratique moyenne entre l'image prédite et la cible.

Notre principale contribution réside dans le second bloc, présenté en figure 4. En sortie du bloc de fusion, nous estimons $M(t_2)$ et nous avons alors 3 images multi-spectrales qui peuvent être comparées aux 3 images hyper-spectrales. En particulier, en dégradant spatialement ces dernières et en considérant les longueurs d'ondes communes avec les premières, nous obte-

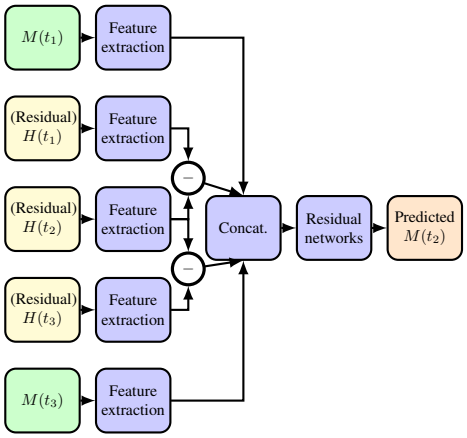

FIGURE 3 – Structure du bloc de fusion.

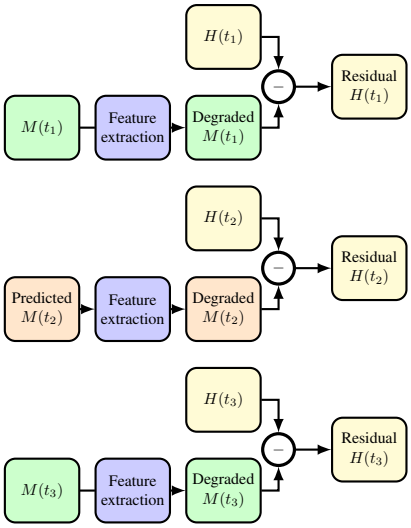

FIGURE 4 – Structure du bloc de complétion d'erreur résiduelle.

nons des images comparables dont les différences sont utilisées pour améliorer l'estimation de $M(t_2)$, à la fois en terme de contenu spatial et spectral, dans un autre bloc de fusion.

D'un point de vue mathématique, l'image $M(t_2)$ prédite dans le bloc de fusion est liée aux données disponibles selon

$$\begin{aligned} M(t_2) =& h\Big(f_1\big(M(t_1)\big), f_2\big(M(t_3)\big), f_3(H(t_3)) - f_4(H(t_2)), \\ & f_5\big(H(t_1) - f_4(H(t_2))\big)\Big), \end{aligned} \tag{1}$$

où $\forall i = 1, \ldots, 5$, $f_i(\cdot)$ est une fonction d'extraction de caractéristique (composée d'une couche de convolution et d'une fonction d'activation ReLu) et $h(\cdot)$ est une fonction qui concatène les caractéristiques des différentes (différnces des) images et des réseaux résiduels. Il est à noter que durant le premier passage dans le bloc de fusion, nous traitons les images hyper-spectrales pour prédire $M(t_2)$. Cependant, cette image prédite est utilisée comme entrée du bloc de complétion d'erreur résiduelle dont les sorties sont utilisées comme les images hyper-spectrales d'entrée du prochain passage du bloc de fusion.

Méthode	PSNR	SAM	SSIM	SCC	UQI	PSNR	SAM	SSIM	SCC	UQI
Tests avec 25 m de résolution spatiale (CIA / LGC)						Tests avec 60 m de résolution spatiale (S-2 / S-3)				
STARFM	24.2	0.6897	0.3974	0.1712	0.5357	17.7	0.9165	0.0531	**0.0316**	0.0692
DCSTFN	23.6	0.2187	0.7309	0.1211	0.7434	25.6	**0.4090**	0.1816	0.0214	**0.0791**
DMnet	24.0	0.2147	0.7421	0.2012	0.7529	19.2	0.4096	0.0555	0.0252	0.07382
DL-SDFM	21.4	0.2591	0.3146	0.0359	0.1720	30.1	0.4878	0.5037	-0.0023	0.0008
CSTF	**36.4**	**0.1277**	**0.9215**	**0.5225**	**0.9701**	**32.9**	0.7891	**0.6372**	0.0311	0.0748

TABLEAU 1 – Performances des méthodes testées pour les bases de données considérées.

Le bloc de complétion d'erreur résiduelle consiste en 3 fonctions similaires $g_i(\cdot)$ qui sont respectivement appliquées à $M(t_i)$ et $H(t_i)$ pour $i = 1, \ldots, 3$, selon

$$H_r(t_i) = g_i(H(t_i), M(t_i)) = H(t_i) - f'_i\left(M(t_i)\right), \qquad (2)$$

où $f'_i(\cdot)$ est la fonction d'extraction de caractéristique de la i-ème image multi-spectrale, et $H_r(t_i)$ est l'image hyper-spectrale résiduelle qui est utilisée comme entrée du bloc de fusion suivant, permettant d'y améliorer la prédiction de $M(t_2)$. Il est à noter que chaque fonction $f'_i(\cdot)$ réalise une dégradation des images multi-spectrales – pour les rendre comparables à leurs images hyper-spectrales respectives – mais elles fournissent aussi des traitements spécifiques, c.-à-d. 3 couches convolutionnelles, chacune étant suivie d'une fonction d'activation ReLu.

3 Validation expérimentale

Nous étudions ici les performances de notre méthode proposée CSTF. Pour cela, nous considérons des bases de données d'images réelles où l'image $M(t_2)$ est connue mais pas utilisée dans l'étape d'apprentissage. En particulier, nous considérons les bases de données [1] CIA / LGC [12] – aux résolutions spatiales de 25 m (pour Landsat) et 500 m (pour MODIS) pour les images respectivement multi-spectrales et hyper-spectrales, et 6 bandes spectrales dans tous les cas – et des données S-2 à 60 m de résolution spatiale et S-3 pré-traitées [13] [2].

Pour ces jeux de données, nous comparons les performances obtenues par notre méthode CSTF avec celles obtenues par de nombreuses approches, c.-à-d. STARFM [4], DCSTFN [9], DMnet [3] et DL-SDFM [11]. Mis à part DL-SDFM, les méthodes de la littérature que nous testons utilisent 3 images acquises aux instants t_1 et t_2 pour réaliser la complétion. L'apprentissage est réalisé par extraction de patchs extraits d'une série temporelle de 24 paires d'images. Afin de pouvoir utiliser des indices quantitatifs de performance, nous considérons pour chaque base de données 3 images multi-spectrales à prédire.

Pour des raisons de calcul, toutes les méthodes d'apprentissage profond utilisent des patchs de taille 150×150 durant la phase d'apprentissage. Lors de l'étape de prédiction, il devient alors nécessaire de reconstruire l'image $M(t_2)$ complète à partir des patchs estimés. Nous avons remarqué dans des tests préliminaires que le *zero padding* – qui est la stratégie par défaut dans TensorFlow – engendrait des artefacts visibles. Aussi, nous avons remplacé cette stratégie par un padding symmétrique qui supprime ces effets. Dans ces tests, le bloc de fusion est appelé 6 fois alors que le bloc de complétion d'erreur résiduelle est lancé 5 fois.

Afin de quantifier les performances des méthodes testées, nous utilisons des critères de performance classiques [3], à savoir (i) le PSNR [14] (qui est le rapport entre la plus grande valeur possible de puissance du signal et la puissance du bruit), (ii) le SAM moyen [15] (le SAM mesure l'angle entre le spectre de référence et celui estimé, pour chaque pixel), (iii) le SSIM [14] qui est une mesure de similarité entre deux images en terme de perception visuelle, (iv) le SCC [16] (le coefficient de corrélation spatiale) et (v) l'UIQI [17] qui modélise la distorsion de l'image prédite selon la perte de corrélation, la distorsion de luminance et la distorsion de contraste.

Il est à noter que les images S-2 et S-3 présentent des nuages – qui ne sont pas nécessairement observés en même temps dans les images acquises le même jour – qui peuvent avoir des positions différentes et faire diminuer les valeurs des critères de performance. De plus, le pré-traitement remplace les valeurs aux niveaux des nuages par des NaN qui ne sont pas utilisés pour les calculs de performance. Le tableau 1 montre les performances atteintes pour chacune des bases de données, pour une unique bande spectrale à 60 m de résolution spatiale. L'approche CSTF que nous proposons surpasse toutes les méthodes de la littérature qui ont été testées, pour tous les indices de performance, sur les données CIA / LGC. Lorsque nous considérons les tests sur les données S-2 / S-3, notre approche donne les meilleures performances pour deux indices, c.-à-d. le PSNR et le SSIM. Mis à part pour le SAM [4], les performances atteintes pour tous les autres indices sont proches des meilleurs valeurs obtenues. Il est notable que notre approche fournit toujours de meilleures performances que l'approche DL-SDFM, qui utilise elle-aussi 5 images pour prédire $M(t_2)$ mais qui n'utilise pas de bloc de complétion d'erreur résiduelle, qui est la principale innovation de notre méthode.

Pour illustrer ces résultats, nous fournissons sur la figure 5 une image $M(t_2)$ prédite par chaque méthode, pour une longueur d'onde, et l'image de référence de la base CIA / LGC. La plupart des détails sont perdus avec STARFM et DL-SDFM.

1. Les bases de données CIA et LGC peuvent être respectivement obtenues à `http://dx.doi.org/10.4225/08/5111AC0BF1229` et à `http://dx.doi.org/10.4225/08/5111AD2B7FEE6`.

2. Ce pré-traitement permet de corriger les effets atmosphériques dans les données, afin de pouvoir comparer les images satellitaires à des acquisitions in situ. Cependant, une telle comparaison n'est pas réalisée dans cet article.

3. Nous avons utilisé l'implémentation fournie à `https://github.com/andrewekhalel/sewar`.

4. Le SAM peut être amélioré en exécutant plus de blocs dans la boucle.

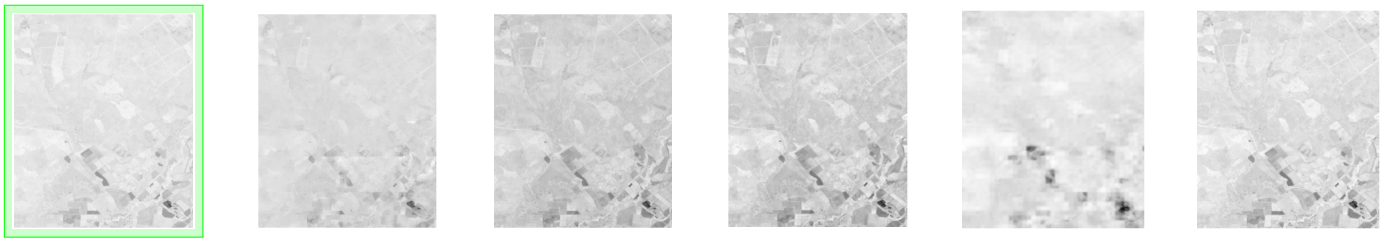

FIGURE 5 – De g. à d. : image de référence et prédictions obtenues avec STARFM, DCSTFN, DMnet, DL-SDFM et CSTF.

Ces détails sont préservés avec les autres méthodes mais semblent légèrement plus précis avec CSTF qu'avec DCSTFN et DMnet.

4 Conclusion

Nous avons proposé une nouvelle méthode d'apprentissage profond pour compléter une série temporelle d'images multi-spectrales à partir d'images hyper-spectrales. Elle consiste en une boucle où sont exécutés de manière alternée un bloc de fusion et un bloc de complétion d'erreur résiduelle. La principale innovation de notre méthode réside dans ce dernier et les résultats expérimentaux montrent l'intérêt de notre approche. Dans de futurs travaux, nous aimerions mieux caractériser les performances de notre méthode en fonction de ses paramètres et appliquer notre stratégie à d'autres architectures et modèles.

Références

[1] N. Yokoya, C. Grohnfeldt, and J. Chanussot, "Hyperspectral and multispectral data fusion : A comparative review of the recent literature," *IEEE Geosci. Remote Sens. Mag.*, vol. 5, no. 2, pp. 29–56, 2017.

[2] A. Alboody, M. Puigt, G. Roussel, V. Vantrepotte, C. Jamet, and T. K. Tran, "Experimental comparison of multi-sharpening methods applied to Sentinel-2 MSI and Sentinel-3 OLCI images," in *Proc. WHISPERS'21*, 2021.

[3] J. Li, Y. Li, L. He, J. Chen, and A. Plaza, "Spatio-temporal fusion for remote sensing data : An overview and new benchmark," *Science China Information Sciences*, vol. 63, no. 4, pp. 140301, 2020.

[4] F. Gao, J. Masek, M. Schwaller, and F. Hall, "On the blending of the Landsat and Modis surface reflectance : predicting daily Landsat surface reflectance," *IEEE Trans. Geosci. Remote Sens.*, vol. 44, no. 8, pp. 2207–2218, 2006.

[5] J. Wang and B. Huang, "A rigorously-weighted spatio-temporal fusion model with uncertainty analysis," *Remote Sensing*, vol. 9, no. 10, pp. 990, 2017.

[6] Q. Wang and P. M. Atkinson, "Spatio-temporal fusion for daily Sentinel-2 images," *Remote Sensing of Environment*, vol. 204, pp. 31–42, 2018.

[7] B. Zhukov, D. Oertel, F. Lanzl, and G. Reinhackel, "Unmixing-based multisensor multiresolution image fusion," *IEEE Trans. Geosci. Remote Sens.*, vol. 37, no. 3, pp. 1212–1226, 1999.

[8] B. Huang and H. Song, "Spatiotemporal reflectance fusion via sparse representation," *IEEE Trans. Geosci. Remote Sens.*, vol. 50, no. 10, pp. 3707–3716, 2012.

[9] Z. Tan, P. Yue, L. Di, and J. Tang, "Deriving high spatiotemporal remote sensing images using deep convolutional network," *Remote Sensing*, vol. 10, no. 7, pp. 1066, 2018.

[10] H. Zhang, Y. Song, C. Han, and L. Zhang, "Remote sensing image spatiotemporal fusion using a generative adversarial network," *IEEE Trans. Geosci. Remote Sens.*, vol. 59, no. 5, pp. 4273–4286, 2021.

[11] D. Jia, C. Song, C. Cheng, S. Shen, L. Ning, and C. Hui, "A novel deep learning-based spatiotemporal fusion method for combining satellite images with different resolutions using a two-stream convolutional neural network," *Remote Sensing*, vol. 12, no. 4, pp. 698, 2020.

[12] I. V. Emelyanova, T. R. McVicar, T. G. Van Niel, L. T. Li, and A. I. Van Dijk, "Assessing the accuracy of blending landsat–modis surface reflectances in two landscapes with contrasting spatial and temporal dynamics : A framework for algorithm selection," *Remote Sensing of Environment*, vol. 133, pp. 193–209, 2013.

[13] F. Steinmetz and D. Ramon, "Sentinel-2 MSI and sentinel-3 OLCI consistent ocean colour products using POLYMER," in *Proc. SPIE "Remote Sensing of the Open and Coastal Ocean and Inland Waters"*, 2018, vol. 10778.

[14] Z. Wang, A. C. Bovik, H. R. Sheikh, and E. P. Simoncelli, "Image quality assessment : from error visibility to structural similarity," *IEEE Trans. Image Process.*, vol. 13, no. 4, pp. 600–612, 2004.

[15] R. H. Yuhas, A. F. Goetz, and J. W. Boardman, "Discrimination among semi-arid landscape endmembers using the spectral angle mapper (sam) algorithm," in *Proc. Summaries 3rd Annu. JPL Airborne Geosci. Workshop*, 1992, vol. 1, pp. 147–149.

[16] J. Zhou, D. Civco, and J. Silander, "A wavelet transform method to merge landsat tm and spot panchromatic data," *International journal of remote sensing*, vol. 19, no. 4, pp. 743–757, 1998.

[17] Z. Wang and A. Bovik, "A universal image quality index," *IEEE Signal Process. Lett.*, vol. 9, no. 3, pp. 81–84, 2002.