Classification multi-modale de l'occupation du sol à partir de réseaux de neurones convolutifs et de l'auto-distillation

Yawogan Jean Eudes Gbodjo, Dino Ienco

Classification multi-modale de l'occupation du sol à partir de réseaux de neurones convolutifs et de l'auto-distillation

Yawogan Jean Eudes Gbodjo[1] Dino Ienco[1]

[1] INRAE, UMR TETIS, Université de Montpellier

500 Rue Jean Francois Breton, 34090 Montpellier
jean-eudes.gbodjo@inrae.fr

Résumé

Plusieurs études ont été menées ces dernières années afin d'exploiter la complémentarité entre les données multi-modales de télédétection pour des applications notables telles que la cartographie de l'occupation des sols. Afin de franchir une étape supplémentaire dans l'exploitation de données multi-modales, par rapport aux études précédentes qui ont étudié la combinaison de données multi-temporelles radar et optique ou de données optiques multi-temporelles et multi-échelles, nous proposons dans ce travail une méthode combinant simultanément toutes ces sources d'entrées, en particulier des séries temporelles d'images radar et optique à échelle moyenne ainsi qu'une image optique à échelle fine. La méthode proposée est une architecture multi-branche de réseaux de neurones convolutifs (CNN) qui exploite un encodeur CNN par source. De plus, elle est équipée d'une stratégie d'auto-distillation afin de renforcer les encodeurs par source et permettre au réseau « d'apprendre de lui-même ». Les expériences sont menées sur un site d'étude, à savoir l'île de la Réunion, un territoire français d'outre-mer où les données de référence ont été collectées sous des contraintes opérationnelles et sont annotées de manière éparse. Les résultats obtenus, notamment une précision globale de classification d'environ 94%, témoignent de l'efficacité de la méthode proposée. Notre étude confirme à nouveau l'avantage de l'analyse multi-modale pour des tâches de cartographie de l'occupation du sol.

Mots Clef

Images satellitaires multi-modales, réseaux de neurones convolutifs, auto-distillation, cartographie de l'occupation des sols, données spatialement éparses.

Abstract

Several studies have been conducted in recent years to exploit the complementarity among multi-modal remote sensing data for notable applications such as land cover mapping. In order to make a step further, with respect to previous studies that investigated multi-temporal SAR and optical data or multi-temporal/multi-scale optical combination scenarios, here we propose a deep learning framework that simultaneously combine all these input sources, specifically multi-temporal SAR/optical data and fine scale optical information. Our proposal relies on a patch-based multi-branch convolutional neural network (CNN) that exploits different per source CNN encoders to deal with the specificity of the input signals. Furthermore, our framework is equipped with a self-distillation component to boost the per source encoders supporting the network to learn from itself. Experiments were carried out on a real world benchmark, namely the Reunion island, a french overseas department, where the annotated reference data collected under operational constraints are sparse. Obtained results, providing an overall accuracy of about 94%, highlight the effectiveness of our proposal based on CNNs and self-distillation to combine heterogeneous multi-sensor remote sensing data and confirm the benefit of multi-modal analysis for downstream tasks such as land cover mapping.

Keywords

Multi-modal satellite images, convolutional neural networks, self-distillation, land use and land cover mapping, sparsely annotated data.

1 Introduction

De nos jours, il existe une pléiade de missions satellitaires fournissant continuellement des images de la surface terrestre dans diverses modalités (ex. radar ou optique) et à des échelles spatiales et temporelles variées. Par conséquent, une même zone d'étude peut être efficacement couverte par des informations multi-sources et complémentaires provenant de différents capteurs. En particulier, l'avènement des missions Sentinel de l'Agence spatiale européenne [1] permet d'accéder en quasi synchronisation à un ensemble de données radar et optique à hautes résolutions spatiale (jusqu'à 10-m) et temporelle (jusqu'à une acquisition tous les cinq/six jours) sur la plupart des surfaces continentales. La communauté concentre depuis un certain temps ses efforts pour démontrer l'intérêt de combiner les informations multi-modales fournies par différents

capteurs [2]. En mettant particulièrement l'accent sur la cartographie de l'occupation des sols, plusieurs études ont récemment évalué le potentiel des techniques d'apprentissage profond (AP) pour exploiter autant que possible la complémentarité entre données de différents capteurs sur la même zone d'étude [3]. Contrairement aux approches traditionnelles où, dans un premier temps, des caractéristiques par source étaient extraites et ensuite, une méthode standard d'apprentissage automatique était déployée sur l'agrégation de ces caractéristiques [4], les méthodes d'apprentissage profond permettent de combiner directement les données multi-sources sans étapes intermédiaires. Dans les travaux présentés par Liu et al. [5] ou Gaetano et al. [6], les bandes panchromatiques et multispectrales de diverses résolutions spatiales sont directement combinées pour fournir une cartographie de l'occupation du sol à la résolution la plus fine. Récemment, Hong et al. [7] ont proposé de fusionner des données Lidar multispectrales avec des données optiques hyperspectrales pour la classification de l'occupation des sols en milieu urbain. En ce qui concerne la classification de données multi-modales de télédétection où au moins une des sources est une série temporelle d'images satellitaires (STIS), Kussul et al. [8] ou encore Ienco et al. [9] ont combiné ensemble des STIS radar et optique dans le but de tirer parti de la complémentarité entre les capteurs actifs et passifs. Plus encore, Benedetti et al. [10] ainsi que Gadiraju et al. [11] ont proposé de combiner des STIS optiques avec une information optique mono-date à très haute résolution spatiale (THRS) dans le but d'exploiter conjointement ces informations optiques multi-temporelles et multi-échelles.

Dans la littérature existante, la majorité des approches multi-modales basées sur l'AP exploitent principalement deux sources de données en entrée. C'est particulièrement le cas lorsque des STIS sont exploitées dans l'analyse (ex. radar et optique ou optique multi-temporelle et multi-échelle). Dans le but d'aller plus loin par rapport aux études précédentes et d'exploiter encore plus le potentiel des techniques d'AP pour la combinaison d'images satellitaires multi-modales en cartographie de l'occupation des sols, nous proposons dans ce travail une méthode basée sur un ensemble de réseaux de neurones convolutifs (CNN) pour exploiter à la fois des STIS radar et optiques ainsi que d'une image optique THRS. Nous adoptons des CNNs comme encodeurs pour chacune des sources de données car ces approches sont prometteuses pour traiter les images THRS et récemment, leurs aptitudes ont également été mis en avant pour les données multi-temporelles telles que les STIS [8, 9, 12]. De plus, afin de tirer le meilleur parti des informations multi-modales, la méthode est équipée d'une stratégie d'auto-distillation [13] dans laquelle les encodeurs par source sont optimisés en tenant compte du résultat de la classification. Ceci permet en quelques sortes au modèle d'apprendre de lui-même. Plus précisément, les connaissances des couches les plus profondes (la sortie du modèle) sont distillées vers les couches moins profondes (les encodeurs par source) dans le but de guider le processus d'apprentissage. Bien que le processus d'auto-distillation ait récemment fait l'objet d'une attention particulière dans le domaine de la vision par ordinateur afin d'améliorer les performances des réseaux de neurones convolutifs [14], il reste encore inexploré dans le contexte de la classification multi-modale. La méthode proposée a été évaluée sur un site d'étude réel, à savoir l'île de la Réunion, un territoire français d'outre-mer sur lequel les données de référence ont été collectées sous des contraintes opérationnelles. Ces contraintes liées aux efforts de collecte et aux coûts des campagnes d'acquisition sur le terrain empêchent une annotation exhaustive impliquant de ce fait des données de référence spatialement éparses. Ceci fait par conséquent obstacle à l'utilisation d'approches standards de segmentation sémantique en vision par ordinateur comme U-Net qui requièrent des annotations denses (chaque pixel doit être étiqueté) [15]. Pour cette raison, dans le cas d'annotations spatialement éparses comme celles qui caractérisent la cartographie opérationnelle de l'occupation du sol, les approches basées sur des patchs [9, 16] sont généralement préférées, l'objectif étant de produire la carte d'une zone d'étude entière à partir de quelques échantillons collectés [17, 16].
La suite de l'article est organisée comme suit : les deux sites d'étude et leurs jeu de données sont présentés dans la section 2 ; la section 3 décrit la méthode proposée pour la classification multi-modale de l'occupation du sol, tandis que les expériences et les évaluations sont présentées et discutées dans la section 4. Enfin, la section 5 conclut le travail.

2 Données utilisées

L'étude a été réalisée sur l'île de la Réunion, un territoire français d'outre-mer situé dans l'océan Indien. Les données satellitaires consistent en une série temporelle radar Sentinel-1 (S1) et optique Sentinel-2 (S2) ainsi qu'une image optique THRS SPOT-6/7. Au total, 26 images S1 et 21 images S2 ont été collectées sur l'année 2017 correspondant à la vérité terrain. L'image THRS SPOT-6/7 a été obtenue à partir d'un mosaïquage, par une technique d'harmonisation colorimétrique [18], entre 4 images acquises respectivement le 26 Décembre 2016 et les 10 Mai, 11 Juin et 20 Novembre 2017, afin d'assurer une couverture sans nuages de l'ensemble du site d'étude.
Les données S1 utilisées ont été acquises en bande-C avec une double polarisation (VH et VV) dans le mode IW (Interferometric Wide Swath) et en orbite ascendante. Elles ont été téléchargées au niveau 1C GRD (Ground Range Detected) sur la plateforme PEPS [1]. Les images S1 ont d'abord été calibrées radiométriquement en valeurs de rétrodiffusion (décibels), puis orthorectifiées à 10-m de résolution spatiale et enfin un filtrage multi-temporel a été effectué sur les séries temporelles afin de réduire le chatoiement. Les images S2 ont été téléchargées à partir de la

1. https://peps.cnes.fr/

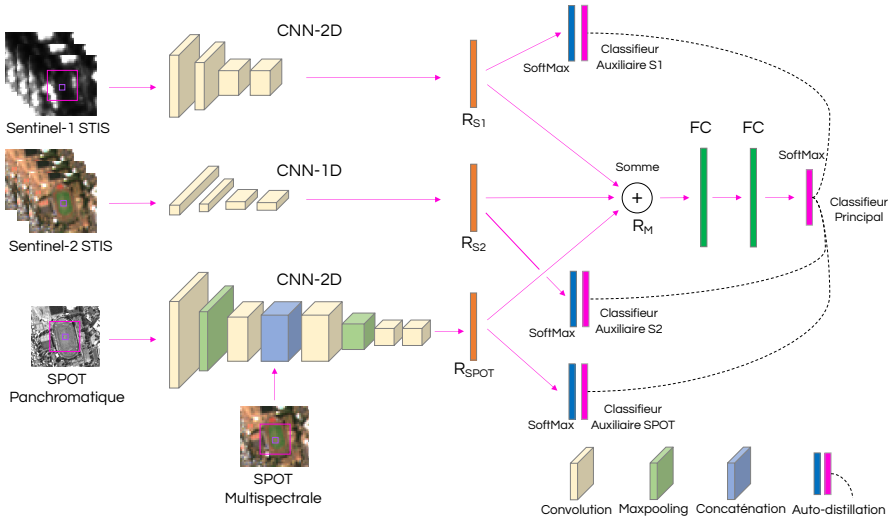

FIGURE 1 – Aperçu global de la méthode *MMCNN*$_{SD}$.

plateforme THEIA [2] au niveau 2A en réflectance de surface et sont fournies avec des masques de nuages. Seules les bandes à 10-m de résolution spatiale ont été prises en compte dans cette analyse, c'est-à-dire les longueurs d'onde du Bleu, du Vert, du Rouge et du proche infrarouge. Étant donné la nébulosité affectant les images dans l'optique, la technique de gap-filling multi-temporel [17] a été appliquée en tant que pré-traitement sur chaque bande, afin de remplacer les valeurs des pixels de nuages détectés. Ces valeurs sont linéairement interpolées en utilisant les valeurs non nuageuses précédentes et suivantes dans la séquence temporelle. De plus, deux indices spectraux communément utilisés pour caractériser l'activité de la végétation ont été extraits : le NDVI [19] et le NDWI [20]. Ces derniers sont ensuite incorporés dans l'analyse ce qui donne un total de six bandes décrivant chaque image S2. Les images SPOT sont constituées d'une bande panchromatique et de quatre bandes multispectrales (spectre bleu, vert, rouge et proche infrarouge) à 1.5-m et 6-m de résolution spatiale respectivement. Elles sont orthorectifiées et prétraitées en réflectance au dessus de l'atmosphère.

La vérité terrain sur l'île de la Réunion a été construite sur base de plusieurs sources : le Registre Parcellaire Graphique (RPG), des prises de points GPS ainsi que l'interprétation visuelle d'images THRS SPOT conduit par un expert connaissant le terrain. Les données de référence sont publiquement accessibles [21]. Au total, le jeu de données est constitué de 880 723 pixels étiquetés, répartis sur 11 classes d'occupation du sol (Tableau 1). Rappelons qu'en raison des contraintes opérationnelles, la vérité terrain collectée sur les deux sites est spatialement éparse. Ainsi, nous concentrons nos efforts, non pas sur l'emploi de techniques de segmentation sémantique qui nécessitent des données densement étiquetés, mais plutôt sur des stratégies de classification à base de patchs d'images.

2. http://theia.cnes.fr

TABLE 1 – Nombre de pixels par classe sur la Réunion

Classe	Pixels
1 – CANNE À SUCRE	88 962
2 – PÂTURAGE ET FOURRAGE	68 098
3 – MARAÎCHAGE	17 488
4 – CULTURE SOUS SERRE OU OMBRAGÉE	1 908
5 – ARBORICULTURE	33 721
6 – ESPACE BOISE	205 023
7 – LANDE ET SAVANE	155 231
8 – ROCHER ET SOL NU NATUREL	154 343
9 – OMBRE DUE AU RELIEF	54 301
10 – EAU	82 592
11 – ESPACE ARTIFICIALISÉ	19 056
TOTAL	880 723

3 Méthode

La Figure 1 donne un aperçu global du fonctionnement de la méthode *MMCNN*$_{SD}$ (Multi-Modal CNN with per source Self-Distillation). Elle a 3 branches, une dédiée à chaque source d'information complémentaire qu'elle prend en entrée : les séries temporelles S1 et S2 ainsi que l'image THRS SPOT. Chaque branche est composée d'un encodeur CNN qui extrait une représentation spécifique par source. Les représentations par source, respectivement R_{S1}, R_{S2} et R_{SPOT}, sont par la suite agrégées selon un schéma de fusion tardive [22] par le biais de la somme. La représentation multi-source R_M, résultant de l'agrégation, est finalement transmise à deux couches entièrement connectées suivies par une couche de sortie avec une activation $SoftMax$ (classifieur principal), afin d'effectuer la classification finale. De plus, la méthode *MMCNN*$_{SD}$ est dotée d'une stratégie d'auto-distillation [13] lui permettant « d'apprendre de lui-même ». En effet, chaque encodeur CNN par source est également muni d'une couche de sortie (classifieur auxiliaire) avec une activation $SoftMax$, dans le but de forcer le modèle à extraire des représentations plus discriminantes et complémentaires entre elles. Les classi-

fieurs auxiliaires par source sont entraînés pour mimer la sortie du classifieur principal, ceci dans le but de distiller ou transférer les connaissances des couches les plus profondes, notamment la sortie du modèle, vers les couches les moins profondes c'est-à-dire celles des encodeurs par source. Tandis que le processus classique de distillation de connaissances [13] est basé sur une architecture « enseignant – étudiant » où le transfert de connaissances se fait de l'enseignant à l'étudiant, celui de l'auto-distillation [14] ne requiert pas de paire de modèles distincts puisque la distillation de connaissances se fait à partir du modèle lui même de manière autonome. Pour faire le lien avec l'architecture « enseignant – étudiant », dans notre cas, la sortie du classifieur principal du modèle *MMCNN*$_{SD}$ peut être assimilée aux connaissances du modèle enseignant tandis que les encodeurs par source représentent les modèles étudiants ayant pour objectif de mimer le comportement de l'enseignant. Nous employons de cette façon la stratégie d'auto-distillation dans un contexte d'analyse multi-modale. À cette fin, la perte (L) à minimiser dans l'entraînement du modèle est définie comme suit :

$$L = CE(Y, CL(R_M)) + \lambda \sum_{s \in \{S1, S2, SPOT\}} CE(CL(R_M), OUT(R_s))$$

où Y est l'information de référence c'est-à-dire la classe d'occupation du sol à prédire ; CE est l'entropie croisée dans un contexte multi-classe ; CL correspond au classifieur principal c'est-à-dire un réseau de neurones constitué de deux couches entièrement connectées avec une activation ReLU et la normalisation par lot, suivies d'une couche de sortie avec une activation $SoftMax$; OUT correspond à un classifieur auxiliaire constitué d'une couche de sortie avec une activation $SoftMax$. Enfin, λ est un hyperparamètre contrôlant l'importance relative des coûts liés aux classifieurs auxiliaires et donc à la stratégie d'auto-distillation par rapport à celle du classifieur principal associé à la représentation multi-modale. Bien que le modèle implique, pendant la phase d'entraînement, l'utilisation du classifieur principal et des classifieurs par source associés à la stratégie d'auto-distillation, seule la prédiction fournie par le classifieur principal, c'est-à-dire $CL(R_M)$, est considérée en temps d'inférence. L'ensemble des paramètres associés au modèle *MMCNN*$_{SD}$ est appris de bout en bout.

Architecture des encodeurs CNN par source. Afin de tirer partie de la complémentarité des diverses sources d'information dont nous disposons, nous avons conçu des encodeurs CNN qui leur sont spécifiques. Pour traiter la série temporelle S1, nous utilisons un réseau convolutif à deux dimensions (CNN-2D). Ainsi, les données S1 sont organisés en un empilement d'images successives dont le nombre total de bandes est équivalent au nombre de dates d'acquisitions multiplié par un facteur de 2 puisque les données S1 sont analysés en double polarisation (VV et VH). Des patchs d'images de mêmes dimensions spatiales centrées sur le pixel à classifier sont dès lors extraites de cet empilement et constituent l'information en entrée du CNN-2D. Pour traiter la série temporelle S2, nous avons suivi la littérature récente en cartographie de l'occupation du sol [12] et adopté un réseau convolutif unidimensionnel (CNN-1D). Ici, c'est l'information séquentielle du pixel, provenant de la série temporelle, qui constitue l'entrée du modèle CNN-1D. Enfin pour traiter l'image THRS SPOT, nous adoptons à nouveau un CNN-2D, a fortiori dans le but d'exploiter, autant que possible, cette information spatiale à échelle fine. Les images SPOT disposent d'une bande panchromatique et de bandes multispectrales à des résolutions spatiales différentes (1.5-m et 6-m respectivement). Afin de traiter les deux types d'informations avec leur résolution native en évitant dans le mesure du possible toute étape intermédiaire pouvant engendrer des coûts de calcul supplémentaires comme le rééchantillonnage ou le pan-sharpening [6], le modèle CNN-2D adopté traite en premier l'information panchromatique et une fois que les cartes d'activation produites ont les mêmes dimensions spatiales que les bandes multispectrales, ces dernières sont intégrées au processus par concaténation. Ici également, des patchs d'images (panchromatique et multispectrales), prises autour de la position géographique correspondante au centre du pixel à classifier sur l'image Sentinel, sont extraites de l'image SPOT.

TABLE 2 – Détails sur l'architecture du modèle *MMCNN*$_{SD}$. (Par souci de lisibilité, les classifieurs auxiliaires ont été omis.)

Sentinel-1	**Sentinel-2**	**SPOT**
		7×7 Conv2D (128) avec PAN
		MaxPooling2D 3×3
3×3 Conv2D (128)	5×1 Conv1D (128)	3×3 Conv2D (256)
3×3 Conv2D (128)	3×1 Conv1D (128)	Concaténation avec MS
3×3 Conv2D (256)	3×1 Conv1D (256)	3×3 Conv2D (256)
1×1 Conv2D (256)	1×1 Conv1D (256)	MaxPooling2D 3×3
GlobAvgPooling2D	GlobAvgPooling1D	3×3 Conv2D (256)
		1×1 Conv2D (256)
		GlobAvgPooling2D
Agrégation par somme		
Couche entièrement connectée (512) + ReLU + Normalisation par lot		
Couche entièrement connectée (512) + ReLU + Normalisation par lot		
Couche de sortie entièrement connectée avec activation $SoftMax$		

Nous reportons dans le tableau 2, les détails concernant l'architecture du modèle *MMCNN*$_{SD}$. La partie initiale du tableau, incluant les couches de Pooling, décrit les encodeurs CNN par source présentés précédemment. Conv1D et Conv2D désignent respectivement des convolutions 1D et 2D. Les valeurs associées (128, 256 et 512) correspondent au nombre de filtres de convolution. Chaque couche convolutive est activée par une fonction ReLU, suivie de la normalisation par lot et d'une couche de Dropout.

4 Expériences

4.1 Protocole expérimental

Dans le processus d'évaluation du modèle proposé, nous avons tout d'abord réalisé une étude spécifique sur les en-

codeurs CNN par source afin de valider nos choix architecturaux. Pour ce faire, les données S1 et S2 sont analysées en considérant successivement des réseaux CNN-1D, CNN-2D et CNN-3D. Les CNN-1D/2D sont les mêmes que ceux adoptés pour le modèle (voir Tableau 2). En ce qui concerne le CNN-3D, nous avons adopté un modèle similaire aux deux premiers en gardant presque la même architecture notamment le même nombre de couches convolutives et de filtres associés ainsi qu'une couche de Pooling pour l'extraction des caractéristiques. Seuls la taille des filtres de convolution et le pas dans le domaine temporel ont été modifiés pour les couches convolutives. Ainsi, une taille de $3 \times 3 \times 3$ a été adoptée pour les trois premières couches convolutives suivant les recommandations de [23] tandis que nous avons gardé une taille de 1 dans les 3 dimensions pour la dernière couche convolutive, similairement aux CNN-1D/2D. Par ailleurs, le pas des convolutions dans le domaine temporel est fixé à 2 pour les deuxième et troisième couches convolutives afin de tirer d'avantage parti du signal temporel. Notons que le module de classification des encodeurs CNN-1D/2D/3D est équivalent à celui du modèle $MMCNN_{SD}$ c'est-à-dire son classifieur principal. Par la suite, nous avons intégré dans l'analyse la combinaison des données multi-modales par le biais de la méthode $MMCNN_{SD}$. À cette fin, nous avons réalisé également une étude d'ablation sur la combinaison des sources afin de démêler leurs interactions. Deux variantes du modèle proposé avec S1 et S2 uniquement puis S2 et SPOT ont ainsi été testées. Ces variantes sont dénommées : $MMCNN_{SD}^{S1+S2}$ et $MMCNN_{SD}^{S2+SPOT}$. De plus, nous évaluons la contribution des classifieurs auxiliaires par source associés à la stratégie d'auto-distillation proposée afin de comprendre leur bénéfice par rapport à notre modèle. Cette variante, dénommée $MMCNN_{noSD}$, peut être assimilée à un processus de fusion tardive classique tel que présentée par Hong et al. [3]. Par ailleurs, nous considérons un autre compétiteur qui s'appuie sur les travaux de Benedetti et al [10] sur la cartographie multi-source de l'occupation du sol dans lesquels les classifieurs auxiliaires par source sont supervisés à partir des étiquettes originales. Ce compétiteur est dénommé $MMCNN_{HardLabels}$.

En ce qui concerne les données en entrée des modèles, comme évoqué auparavant, nous avons extrait des patchs d'images pour décrire un emplacement géographique spécifique. Ainsi, la taille des patchs S1/S2 a été fixée à 9×9 soit 4 pixels dans chaque direction autour du pixel cible à classifier tandis que pour les patchs SPOT, des tailles de 8×8 et 32×32 ont été prises respectivement pour les bandes multispectrales et panchromatique, pareillement à [10]. Rappelons que les patchs SPOT sont extraits autour du même emplacement géographique correspondant au centre du pixel à classifier sur les images S1/S2. Rappelons également que l'encodeur CNN-1D ne prend uniquement en compte que l'information séquentielle provenant du pixel central des patchs d'images extraites.

Pour le reste, les valeurs des patchs sont normalisées par bande, considérant les série temporelles S1/S2 ou l'image SPOT, dans l'intervalle [0,1]. Les données ont été divisées en jeu d'entraînement, de validation et test avec des proportions respectives de 50%, 20% et 30%. Pour ce faire, nous nous assurons aussi que tous les pixels appartenant à une même unité d'occupation du sol sur la vérité terrain se retrouvent exclusivement dans l'une des 3 partitions (entraînement, validation ou test) afin d'éviter au mieux de possibles biais d'auto-corrélation spatiale dans la procédure d'évaluation. Par ailleurs, les modèles sont optimisés à travers une procédure d'entraînement et validation, consistant à sauvegarder les paramètres ou poids permettant à un modèle au moment de l'entraînement de généraliser le mieux sur le jeu de validation. La phase d'entraînement a été conduite sur 300 époques en utilisant l'optimiseur Adam [24] avec un taux d'apprentissage de 10^{-4}. Le taux de Dropout a été fixé à 0.4 et la taille des lots à 256 échantillons. L'hyper-paramètre λ est quant à lui fixé à 0.3 pour toutes les approches multi-sources incluant des classifieurs auxiliares par source. Les métriques précision globale, F1 score et cœfficient Kappa sont considérées pour évaluer les performances des modèles sur le jeu test. Du fait que les performances sont susceptibles de varier en fonction de la division des données, en raison d'échantillons plus simples ou plus complexes impliqués dans les différentes partitions, ces métriques sont moyennées sur 5 divisions aléatoires du jeu de données suivant la stratégie précédemment décrite.

4.2 Évaluation des encodeurs par source

TABLE 3 – Performances moyennes des encodeurs CNN par source

Sources		F1 Score	Kappa	Précision globale
S1	CNN-1D	64.82 ± 1.32	0.587 ± 0.018	65.63 ± 1.64
	CNN-2D	**73.09** ± 2.62	**0.684** ± 0.030	**73.39** ± 2.66
	CNN-3D	72.35 ± 2.94	0.673 ± 0.036	72.63 ± 3.16
S2	CNN-1D	87.98 ± 1.12	0.859 ± 0.017	88.09 ± 1.06
	CNN-2D	87.41 ± 1.61	0.851 ± 0.021	87.41 ± 1.66
	CNN-3D	**88.62** ± 1.45	**0.866** ± 0.017	**88.66** ± 1.36
SPOT		88.35 ± 1.33	0.862 ± 0.017	88.35 ± 1.39

Nous reportons respectivement dans le tableau 3, les performances moyennes des encodeurs CNN par source sur le site de la Réunion. En premier, nous constatons que tirer parti des dépendances spatiales ou temporelles des données S1/S2 conduit à des résultats différents. En effet, les convolutions 2D (spatiales) sont largement plus efficaces que les convolutions 1D (temporelles) pour les données S1 tandis que les performances sont comparables pour les données S2. Ce gain en performance par rapport à l'emploi des convolutions 2D pour les données S1 peut se traduire par le fait que les filtres convolutifs 2D aident à atténuer encore plus le bruit spatial lié au chatoiement des images radar [25], ce qui permet sans doute d'obtenir des représentations plus discriminantes. En ce qui concerne l'encodeur CNN-3D, ses performances sont légèrement en dessous du

CNN-2D avec S1 et légèrement au dessus des autres modèles, notamment le CNN-1D avec S2. Toutefois, le bénéfice qu'apporte l'emploi de convolutions à la fois dans les dimensions spatiale et temporelle (convolutions 3D) reste minimal, surtout si l'on prend en compte le nombre de paramètres largement plus important du CNN-3D par rapport aux autres encodeurs ainsi que ses temps en calcul plus longs. Notons par ailleurs, l'importance de l'information spatiale fine sur l'île de la Réunion qui donne des performances compétitives par rapport à celles de toute la série temporelle S2. Pour résumer, cette étude spécifique sur les encodeurs CNN par source suggère que le CNN-2D est plus pertinent pour la représentation des données S1 tandis que, pour des raisons de parcimonie (nombre de paramètres plus léger et temps de calcul réduit), il est plus approprié d'encoder la série temporelle S2 avec un CNN-1D. Ci-après, les résultats présentés pour S1 et S2 sont respectivement ceux obtenus avec les encodeurs CNN-2D et CNN-1D.

4.3 Combinaison multi-modale

TABLE 4 – Performances moyennes des modèles CNN multi-sources

Sources	F1 Score	Kappa	Précision globale
$MMCNN_{SD}^{S1+S2}$	91.99 ± 0.42	0.906 ± 0.004	92.05 ± 0.30
$MMCNN_{SD}^{S2+SPOT}$	93.07 ± 1.18	0.918 ± 0.014	93.12 ± 1.16
$MMCNN_{SD}$	**94.34** ± 0.49	**0.934** ± 0.006	**94.38** ± 0.49
$MMCNN_{noSD}$	93.21 ± 0.79	0.920 ± 0.009	93.25 ± 0.77
$MMCNN_{HardLabels}$	93.74 ± 0.94	0.926 ± 0.011	93.77 ± 0.96

Les performances moyennes des modèles multi-sources sont présentées respectivement dans le tableau 4. Nous remarquons tout d'abord que la combinaison de données multi-modales améliore systématiquement les performances de classification par rapport aux résultats précédents obtenus avec les sources individuelles. La combinaison simultanée de toutes les sources d'information disponibles par le biais du modèle proposé se révèle de loin la plus efficace. Ceci démontre à nouveau le potentiel pour la classification de l'occupation du sol de combiner des données multi-modales de télédétection.

Par rapport aux autres aspects étudiés du modèle proposé, notamment l'apport des classifieurs auxiliaires associés à la stratégie d'auto-distillation, nous remarquons dans l'étude d'ablation menée ($MMCNN_{noSD}$ vs $MMCNN_{HardLabels}$ vs $MMCNN_{SD}$) que ces composants architecturaux contribuent aux performances obtenues sur les un site d'étude. Tout d'abord, les modèles équipés de classifieurs auxiliaires c'est-à-dire $MMCNN_{HardLabels}$ et $MMCNN_{SD}$ sont plus performants que la variante assimilé à la fusion tardive classique c'est-à-dire $MMCNN_{noSD}$. Pour investiguer plus en détail cette observation en nous focalisant sur l'approche proposée, nous avons illustré sur la figure 2, les comportements des modèles $MMCNN_{noSD}$ et $MMCNN_{SD}$ pendant la phase d'apprentissage en considérant leurs performances sur les jeux d'entraînement et de validation.

FIGURE 2 – Comportement du modèle avec et sans stratégie d'auto-distillation pendant la phase d'apprentissage

S'il est à noter que les deux modèles arrivent très bien à s'adapter au jeu d'entraînement de façon quasi identique, c'est bien le modèle proposé et entraîné avec la stratégie d'auto-distillation qui performe le mieux sur le jeu de validation. Ceci montre que la stratégie d'auto-distillation contribue à une meilleure généralisation du modèle. En second lieu, la comparaison en termes de performances numériques entre le modèle proposé et celui qui adopte la stratégie de supervision des classifieurs auxiliaires à partir des étiquettes originales ($MMCNN_{HardLabels}$), montre également que la stratégie d'auto-distillation améliore systématiquement l'exploitation conjointe des données multi-modales. Ces résultats sont conformes aux études récentes sur la distillation des connaissances [13] où il est montré que les étiquettes fournis par le modèle enseignant (dans notre cas le classifieur principal) transmettent plus d'informations utiles et sont plus convenables aux étudiants (dans notre cas les classifieurs auxiliaires) que les étiquettes originales, leur permettant ainsi de mimer plus facilement le comportement du modèle enseignant.

4.4 Analyse qualitative

Dans cette évaluation qualitative, nous nous intéressons aux cartes d'occupation du sol produites à partir des modèles. Rappelons que les cartes d'occupation du sol sont générés à la résolution des images Sentinel (10-m). Par souci de simplicité, nous ne présentons que les cartes produites à partir des combinaisons multi-modales c'est-à-dire $MMCNN_{SD}^{S1+S2}$, $MMCNN_{SD}^{S2+SPOT}$ et $MMCNN_{SD}$. Par ailleurs, étant donné que nous utilisons des imagettes pour décrire des emplacements géographiques spécifiques, les pixels en bordure (4 pixels dans chaque direction puisque la taille des imagettes Sentinel est 9×9) restent non étiquetés. Nous présentons successivement deux séries d'extraits des cartes dans la figure 3. La première série d'extraits (a-d) représente une partie de Saint-Pierre, un quartier urbain côtier avec des plantations de canne à sucre et d'arboriculture. Des erreurs de classification peuvent être mises en évidence entre espaces artificialisés et culture sous serre sur l'extrait de $MMCNN_{SD}^{S1+S2}$ tandis que l'introduction d'information spatiale à échelle fine (extraits

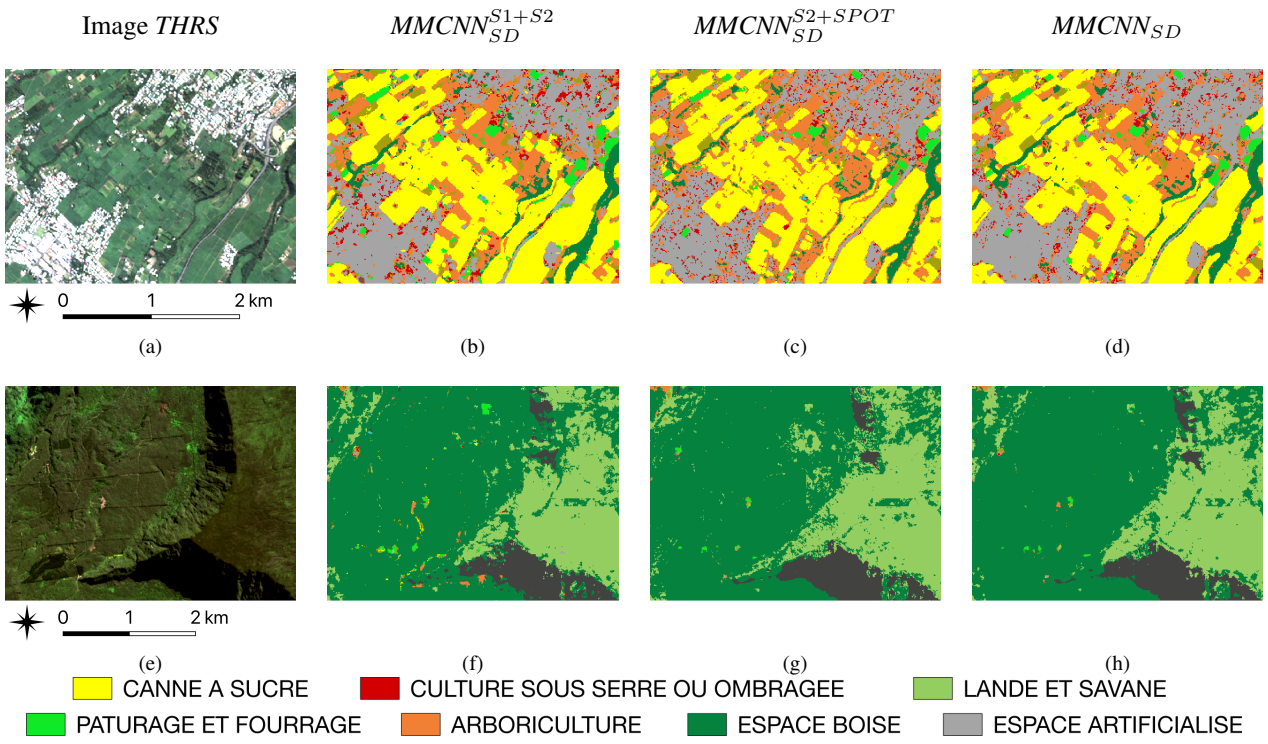

FIGURE 3 – Extraits des cartes d'occupation du sol sur l'île de la Réunion. L'image THRS est affichée en tant que référence. Sont présentés en haut Saint-Pierre et en bas la forêt de Belouve.

de $MMCNN_{SD}^{S2+SPOT}$ et $MMCNN_{SD}$) a significativement réduit ces imprécisions. La seconde série d'extraits (e-h) quant à elle est localisée dans la forêt de Belouve. Le paysage est constitué de forêts primaires et de plantations forestières. Nous pouvons constater quelques légères imprécisions dans la détection de la forêt qui est mal classée avec l'arboriculture ainsi que les lande et savane. Ces imprécisions sont pour la plupart supprimées avec le modèle $MMCNN_{SD}$. En somme, cette évaluation qualitative sous-tend également le bénéfice de combiner des données multi-modales de télédétection pour la cartographie de l'occupation du sol. Dans l'ensemble, les extraits analysées suggèrent que les cartes d'occupation du sol produites par $MMCNN_{SD}^{S2+SPOT}$ et $MMCNN_{SD}$ sont d'une qualité satisfaisante, tandis que celle générée par $MMCNN_{SD}^{S1+S2}$ présente encore des erreurs extensives. Ceci est probablement dû au bruit subsistant dans les données radar et à la fine information spatiale fournie par l'image SPOT qui est particulièrement pertinente sur l'île de la Réunion.

5 Conclusion

Dans ce travail, nous avons présenté la méthode $MMCNN_{SD}$ traitant de la cartographie de l'occupation du sol à partir de données de télédétection multi-sources, multi-temporelles et multi-échelles, à savoir des séries temporelles d'images radar (S1) et optique (S2) ainsi qu'une image optique THRS (SPOT). La méthode repose sur une architecture multi-branche composée d'encodeurs CNN par source et est également équipée d'une composante d'auto-distillation lui permettant de transférer les connaissances des couches les plus profondes vers les moins profondes en apprenant en quelques sortes lui-même. Les résultats expérimentaux, obtenus sur le site d'étude de l'île de la Réunion où les données de référence sont spatialement éparses, sous-tendent clairement l'intérêt de recourir à des informations complémentaires provenant de divers capteurs pour la tâche de cartographie de l'occupation du sol. Une extension directe de l'approche proposée pourra permettre l'intégration de données S1 en orbite descendante, des bandes spectrales restantes S2 ou d'autres sources comme un modèle numérique de surface afin de tenir compte d'effets de relief importants sur la Réunion.

Remerciements

Ce travail a bénéficié d'une aide de l'État gérée par l'Agence Nationale de la Recherche au titre du programme d'Investissements d'Avenir portant la référence ANR-16-CONV-0004.

Références

[1] M. Berger, J. Moreno, J. A. Johannessen, P. F. Levelt, and R. F. Hanssen. Esa's sentinel missions in support of earth system science. *Remote Sensing of Environment*, 120 :84 – 90, 2012.

[2] M. Schmitt and X. X. Zhu. Data fusion and remote sensing : An ever-growing relationship. *IEEE Geosc. and Rem. Sens. Mag.*, 4(4) :6–23, 2016.

[3] D. Hong, L. Gao, N. Yokoya, J. Yao, J. Chanussot, Q. Du, and B. Zhang. More diverse means better : Multimodal deep learning meets remote-sensing imagery classification. *IEEE Transactions on Geoscience and Remote Sensing*, pages 1–15, 2020.

[4] J.J. Erinjery, M. Singh, and R. Kent. Mapping and assessment of vegetation types in the tropical rainforests of the western ghats using multispectral sentinel-2 and sar sentinel-1 satellite imagery. *Remote Sensing of Environment*, 216 :345–354, 2018.

[5] X. Liu, L. Jiao, J. Zhao, J. Zhao, D. Zhang, F. Liu, S. Yang, and X. Tang. Deep multiple instance learning-based spatial-spectral classification for PAN and MS imagery. *IEEE Trans. Geoscience and Remote Sensing*, 56(1) :461–473, 2018.

[6] R. Gaetano, D. Ienco, K. Ose, and Rémi Cresson. A two-branch CNN architecture for land cover classification of PAN and MS imagery. *Remote. Sens.*, 10(11) :1746, 2018.

[7] D. Hong, J. Chanussot, N. Yokoya, J. Kang, and X. Xiang Zhu. Learning-shared cross-modality representation using multispectral-lidar and hyperspectral data. *IEEE Geosci. Remote. Sens. Lett.*, 17(8) :1470–1474, 2020.

[8] N. Kussul, M. Lavreniuk, S. Skakun, and A. Shelestov. Deep learning classification of land cover and crop types using remote sensing data. *IEEE Geosci. Remote Sensing Lett.*, 14(5) :778–782, 2017.

[9] D. Ienco, R. Interdonato, R. Gaetano, and D. H. T. Minh. Combining sentinel-1 and sentinel-2 satellite image time series for land cover mapping via a multi-source deep learning architecture. *ISPRS Journal of Photogrammetry and Remote Sensing*, 158 :11 – 22, 2019.

[10] P. Benedetti, D. Ienco, R. Gaetano, K. Ose, R. G. Pensa, and S. Dupuy. M3 fusion : A deep learning architecture for multiscale multimodal multitemporal satellite data fusion. *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, 11(12) :4939–4949, 2018.

[11] K. K. Gadiraju, B. Ramachandra, Z. Chen, and R. R. Vatsavai. Multimodal deep learning based crop classification using multispectral and multitemporal satellite imagery. In *ACM SIGKDD Conference on Knowledge Discovery and Data Mining*, pages 3234–3242, 2020.

[12] C. Pelletier, G. Webb, and F. Petitjean. Temporal convolutional neural network for the classification of satellite image time series. *Remote Sensing*, 11(5) :523, 2019.

[13] L. Wang and K. J. Yoon. Knowledge distillation and student-teacher learning for visual intelligence : A review and new outlooks. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, pages 1–1, 2021.

[14] L. Zhang, J. Song, A. Gao, J. Chen, C. Bao, and K. Ma. Be your own teacher : Improve the performance of convolutional neural networks via self distillation. In *2019 IEEE/CVF International Conference on Computer Vision, ICCV 2019, Seoul, Korea (South), October 27 - November 2, 2019*, pages 3712–3721. IEEE, 2019.

[15] N. Audebert, B. Le Saux, and S. Lefèvre. Semantic segmentation of earth observation data using multimodal and multi-scale deep networks. In *Computer Vision – ACCV 2016*, pages 180–196. Springer International Publishing, 2017.

[16] R. Interdonato, D. Ienco, R. Gaetano, and K. Ose. Duplo : A dual view point deep learning architecture for time series classification. *ISPRS Journal of Photogrammetry and Remote Sensing*, 149 :91 – 104, 2019.

[17] J. Inglada, A. Vincent, M. Arias, B. Tardy, D. Morin, and I. Rodes. Operational high resolution land cover map production at the country scale using satellite image time series. *Remote Sensing*, 9(1) :95, 2017.

[18] R. Cresson and N. Saint-Geours. Natural color satellite image mosaicking using quadratic programming in decorrelated color space. *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, 8(8) :4151–4162, 2015.

[19] J. W. Rouse, R. H. Hass, J.A. Schell, and D.W. Deering. Monitoring vegetation systems in the great plains with ERTS. *Third Earth Resources Technology Satellite (ERTS) symposium*, 1 :309–317, 1973.

[20] B. Gao. Ndwi—a normalized difference water index for remote sensing of vegetation liquid water from space. *Remote Sensing of Environment*, 58(3) :257 – 266, 1996.

[21] S. Dupuy, R. Gaetano, and L. Le Mézo. Mapping land cover on reunion island in 2017 using satellite imagery and geospatial ground data. *Data in Brief*, 28 :104934, 2020.

[22] Y. Hu, A. Soltoggio, R. Lock, and S. Carter. A fully convolutional two-stream fusion network for interactive image segmentation. *Neural Networks*, 109 :31–42, 2019.

[23] S. Ji, C. Zhang, A. Xu, Y. Shi, and Y. Duan. 3d convolutional neural networks for crop classification with multi-temporal remote sensing images. *Remote Sensing*, 10(2) :75, Jan 2018.

[24] D. P. Kingma and J. Ba. Adam : A method for stochastic optimization. In *3rd International Conference on Learning Representations, ICLR 2015, San Diego, CA, USA, May 7-9, 2015, Conference Track Proceedings*, 2015.

[25] P. Wang, H. Zhang, and V. M. Patel. Sar image despeckling using a convolutional neural network. *IEEE Signal Processing Letters*, 24(12) :1763–1767, 2017.