Segmentation Sémantique d'Images de Télédétection Combinant Modèles Graphiques Probabilistes Hiérarchiques et Réseaux de Neurones Convolutifs Profonds

Martina Pastorino, Gabriele Moser, Sebastiano B. Serpico, Josiane Zerubia

Segmentation Sémantique d'Images de Télédétection Combinant Modèles Graphiques Probabilistes Hiérarchiques et Réseaux de Neurones Convolutifs Profonds

M. Pastorino[1,2] G. Moser[1] S. B. Serpico[1] J. Zerubia[2]

[1] Università di Genova, DITEN, Italie
[2] Inria, Université Côte d'Azur, Sophia-Antipolis, France

Contact : martina.pastorino@inria.fr

Résumé

Dans cet article, une nouvelle méthode pour traiter la segmentation sémantique des données de télédétection à très haute résolution est présentée. Les progrès récents de l'apprentissage profond, en particulier les réseaux de neurones convolutifs et les réseaux entièrement convolutifs, ont montré des performances exceptionnelles dans cette tâche. Mais, comme pour les autres méthodes, la précision de la classification dépend de la quantité et de la qualité de la vérité de terrain utilisée pour les entraîner. Dans le même temps, les modèles de graphes probabilistes (PGMs) ont suscité beaucoup d'intérêt au cours des dernières années, en raison de la disponibilité toujours croissante des données à très haute résolution et, en conséquence, du besoin plus important de prévisions structurées. Les thèmes de recherche proposés dans cet article visent à relier et à combiner différents aspects de ces approches (modèles d'apprentissage profond et stochastiques) pour développer de nouvelles méthodes de classification d'images de télédétection. Afin de développer un pipeline mêlant apprentissage en profondeur et PGMs pour répondre au besoin croissant de cartographie sémantique précise dans les images de télédétection, deux architectures d'apprentissage bien connues telles que U-Net et SegNet ont été considerées. La validation expérimentale est menée avec l'ensemble de données "ISPRS 2D Semantic Labelling Challenge" sur la ville de Vaihingen, dans certains cas avec quelques modifications. Ceci afin de simuler les vérités de terrain courantes dans les applications réelles de télédétection, pour évaluer si la méthode proposée pouvait apporter des améliorations à la précision de la classification. Les résultats sont significatifs, car le pipeline étudié a un score de "rappel" plus élevé par rapport aux réseaux entièrement convolutifs standard considérés.

Mots Clefs

Télédétection, segmentation sémantique, modèles graphiques probabilistes, réseaux de neurones convolutifs, images multi-résolutions.

Abstract

In this paper, a novel method to tackle semantic segmentation of very high resolution remote sensing data is presented. Deep learning techniques, such as convolutional neural networks (CNNs) and fully convolutional networks (FCNs), have shown exceptional performances in this task. But the accuracy of their classification depends on the quantity and quality of the ground truth used to train them. On the other hand, probabilistic graphical models (PGMs) have sparked even more interest in the past few years, because of the ever-growing availability of very high resolution data and the correspondingly increasing need for structured predictions. The research themes proposed in this paper aim to link and combine different ideas of these approaches (deep learning and stochastic models) to develop new methods of classification of remote sensing images. In order to develop a pipeline combining deep learning and PGM to meet the growing need for precise semantic mapping in remote sensing images, two well-known deep learning architectures such as U-Net and SegNet were considered. The experimental validation was carried out with the "ISPRS 2D Semantic Labeling Challenge" data set on the city of Vaihingen, in some cases with some modifications, in order to approximate the ground truths common in real remote sensing applications, to assess whether the proposed method could improve the accuracy of classification in several cases. The results are significant, because the pipeline studied has a higher recall compared to the standard FCNs considered.

Keywords

Remote sensing, semantic segmentation, probabilistic graphical models, convolutional neural networks, multiresolution images.

L'Université de Gênes (UniGe) et l'Université Côte d'Azur (UCA) font partie de l'Alliance Ulysseus (Universités Européennes).

1 Introduction

Les modèles de données multi-modales, généralement fondés sur des méthodes multi-vues, multi-échelles et multi-résolutions, sont de plus en plus importants pour faire face aux exigences du traitement d'image de télédétection [1]. Des travaux récents ont montré que les techniques d'apprentissage profond peuvent atteindre des précisions par pixel très élevées et même reproduire les formes correctes des objets segmentés. Ce sont actuellement les méthodes dominantes pour la segmentation d'image [2, 3], qui ont également suscité un intérêt croissant pour les applications de télédétection [4].

Les architectures les plus pertinentes sont les réseaux entièrement convolutifs (en anglais "fully convolutional networks", FCNs) [5], par exemple U-Net [6] et SegNet [7], qui présentent des performances exceptionnelles [8]. Cependant, les architectures d'apprentissage profond nécessitent de grands ensembles de données avec des vérités de terrain densément étiquetées, qui représentent avec précision toutes les caractéristiques de l'objet, y compris leurs frontières. Ces vérités de terrain détaillées ne sont disponibles que sur des ensembles de données de référence.

Par ailleurs, les modèles graphiques probabilistes ("probabilistic graphical models", PGMs) [9], tels que les modèles de Markov sur des graphes planaires ou multi-couches, sont connus pour être des modèles stochastiques flexibles relativement à des informations spatiales et éventuellement multi-modales [10]. Deux sous-classes de modèles de Markov pour l'analyse d'image bidimensionnelle, pour lesquelles la causalité est formalisée, sont les champs aléatoires de maillage de Markov (MMRF) sur les treillis planaires [11] et les champs aléatoires hiérarchiques de Markov (HMRF) sur les arbres quaternaires [12, 13]. Pour les deux modèles, des algorithmes efficaces d'inférence sont disponibles. Ces deux familles présentent des propriétés complémentaires : un MMRF décrit les interactions spatiales entre les pixels, mais est un modèle à résolution unique ; un HMRF sur un arbre quaternaire appréhende les relations entre des sites situés à différentes échelles grâce à l'utilisation d'une chaîne de Markov, mais ne caractérise pas explicitement les dépendances spatiales au sein des couches à chaque résolution [12]. Dans une approche récente [14], les deux stratégies sont combinées et la markovianité est postulée à la fois à travers les échelles d'un arbre quaternaire et par rapport au système de voisinage associé à chaque couche de l'arbre.

Il a été démontré que l'utilisation de données multi-bandes et multi-résolutions favorisait la précision spatiale des cartes de classification [1], grâce à l'exploitation de l'information à différentes résolutions : vue synoptique des plus grossières et détail spatial des plus fines. Les opérations de traitement exécutées par les réseaux de neurones convolutifs (en anglais CNNs pour "convolutional neural networks") [15] impliquent plusieurs étapes de traitement multi-échelles, à travers des convolutions et des opérations de "pooling", qui correspondent intrinsèquement à la structure des topologies de graphe multi-résolutions sur lesquelles les PGMs peuvent être formulés efficacement [9, 13].

Dans cet article, la segmentation sémantique des images multi-résolutions fondée sur les modèles hiérarchiques de Markov [10] et les FCNs est abordée. Les activations des FCNs à différents blocs (c'est-à-dire à plusieurs résolutions spatiales) et les canaux spectraux IRRV de l'image d'origine sont utilisés pour construire un arbre quaternaire d'apprentissage et les chaînes de Markov sont formulées à la fois à travers les échelles et par rapport à un balayage 1D du pixel en treillis de chaque couche (voir Fig. 1).

Le modèle est combiné avec des ensembles d'arbres de décision, tels que les forets aléatoires (en anglais "random forest" (RF)) [16], pour calculer les probabilités a posteriori nécessaires pour l'inférence sur le PGM en utilisant le mode marginal a posteriori (MPM en anglais), un critère particulièrement avantageux pour les tâches de classification sur des modèles multi-résolutions [12]. L'intégration de ces composantes méthodologiques permet d'exploiter les représentations extraites par le FCN à travers toutes les couches, en incorporant des informations sur le comportement spatial et la structure de la sortie de prédiction. Ceci vise à atténuer les limites du FCN dans l'apprentissage des relations spatiales de vérités de terrain rares, où les frontières spatiales de classe peuvent ne pas être présentes ou peuvent être mal représentées.

1.1 Travaux antérieurs

La grande disponibilité des données qu'offrent les récentes missions d'observation spatiale, ainsi que les progrès des capteurs pour l'acquisition d'images, offrent un grand potentiel d'applications dans le domaine de la télédétection. En particulier, les techniques de classification d'images dans ce secteur peuvent être utilisées pour des applications de cartographie de la couverture terrestre dans des domaines tels que l'urbanisme, l'agriculture de précision et le suivi des espèces forestières, en identifiant et en distinguant les catégories de pixels ou d'objets. Un rôle important dans ce contexte est joué par les méthodes de classification supervisée. En particulier, parmi les approches actuellement reconnues d'une grande efficacité, les méthodes fondées sur les MRFs jouent un rôle important.

Différentes techniques fondées sur les MRFs ont été appliquées au problème de la classification spatio-contextuelle de la couverture terrestre, en utilisant également des modèles hiérarchiques [9], multi-résolutions et multi-échelles [1]. Les approches introduites dans [14] pour résoudre le problème de la classification conjointe de plusieurs images acquises sur la même scène à différentes résolutions spatiales impliquent l'utilisation d'une approche graphique probabiliste avec un cadre de maillage de Markov hiérarchique, qui modélise la classification spatio-contextuelle multi-résolutions et éventuellement des images multi-capteurs.

Dans le même temps, de nombreuses tentatives ont été

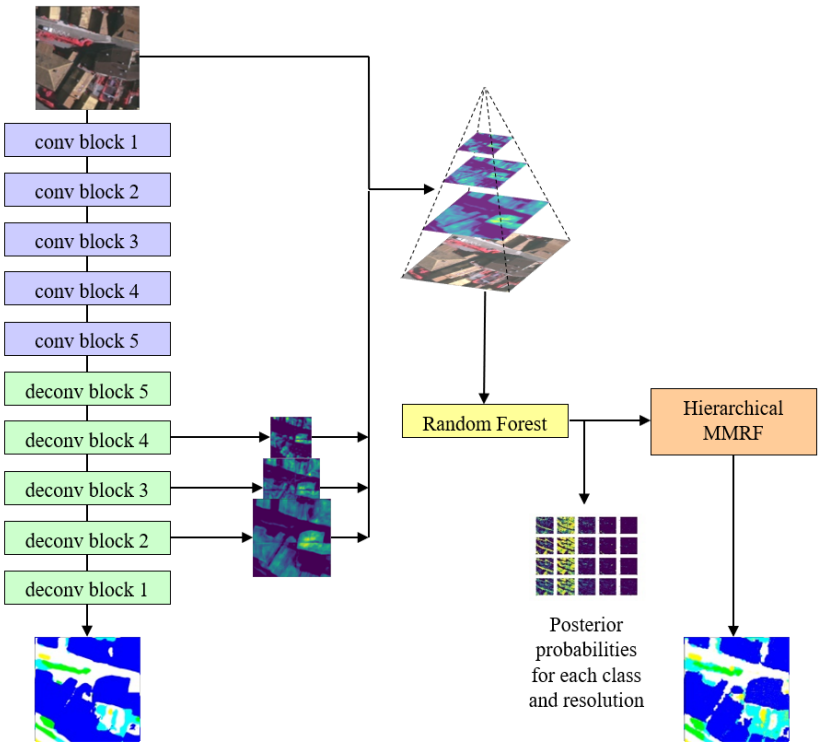

FIGURE 1 – Architecture de la methode proposée.

faites pour concevoir des architectures d'apprentissage profond, telles que les CNNs [17, 18], qui permettent d'effectuer une segmentation sémantique. Ces réseaux ont atteint des performances de classification excellentes sur divers ensembles de données [17, 18], mais nécessitent de grands ensembles de données d'apprentissage qui peuvent ne pas être disponibles pour de nombreuses applications de télédétection. Les traitements exécutés par un CNN [15] impliquent plusieurs étapes de traitements multi-échelles, à la fois par des convolutions avec des tailles de fenêtres données et par des opérations de regroupement. Ces processus correspondent intrinsèquement à la structure des topologies de graphes multi-résolutions - régulières ou irrégulières - sur lesquelles des modèles probabilistes peuvent être formulés de manière efficace et efficiente [9, 13]. Pour ce qui concerne la combinaison de techniques d'apprentissage profond (tels que les CNNs) et les PGMs (tels que les MRFs ou les CRFs), les PGMs sont principalement utilisés comme étapes de post-traitement [19, 20, 21, 22, 23].

2 Méthodologie

2.1 Modèle de Markov hiérarchique

Le PGM considéré consiste en un HMRF qui modélise les informations multi-résolutions à travers les différentes couches de l'arbre quaternaire via une chaîne de Markov, combiné avec un modèle de Markov planaire fondé sur une chaîne de Markov par rapport à un scan 1 D du réseau de pixels, qui modélise les informations spatiales contextuelles.

Soit $\{S^0, S^1, \ldots, S^L\}$ un ensemble de grilles de pixels arrangées comme un arbre quaternaire (0 étant l'échelle la plus grossière et L la plus fine) : chaque site $s \in S^l$ a un site parent $s^- \in S^{l-1}$ et quatre sites enfants $s^+ \subset S^{l+1}$ ($l = 1, 2, \ldots, L-1$). Une hiérarchie sur l'arbre $S = \bigcup_{l=0}^{L} S^l$ de la racine aux feuilles est déterminée. Si une étiquette de classe discrète x_s dans un ensemble fini Ω de M classes ($x_s \in \Omega$, $s \in S$) est associée à chaque $s \in S$, alors $\mathcal{X} = \{x_s\}_{s \in S}$ est un MRF hiérarchique si [10, 12] :

$$P(\mathcal{X}^l | \mathcal{X}^{l-1}, \mathcal{X}^{l-2}, \ldots, \mathcal{X}^0) = P(\mathcal{X}^l | \mathcal{X}^{l-1}), \quad (1)$$

où $\mathcal{X}^l = \{x_s\}_{s \in S^l}$ ($l = 0, 1, \ldots, L$), c'est-à-dire si la markovianité est valable à travers les échelles. Dans ce modèle hiérarchique, les probabilités de transition se factorisent [12], d'où :

$$P(\mathcal{X}^\ell | \mathcal{X}^{\ell-1}) = \prod_{s \in S^\ell} P(x_s | x_{s^-}). \quad (2)$$

Le modèle est étendu pour incorporer des informations spatiales tout en maintenant la causalité. Soit R un treillis rectangulaire et $\prec$ une relation d'ordre dans le treillis de pixels, représentant les pixels qui précèdent chaque site $s \in R$ (c'est-à-dire les sites $r \in R$ tels que $r \prec s$). Une relation de voisinage est supposée dans R de manière cohérente avec cette relation d'ordre, et $r \precsim s$ indique que r

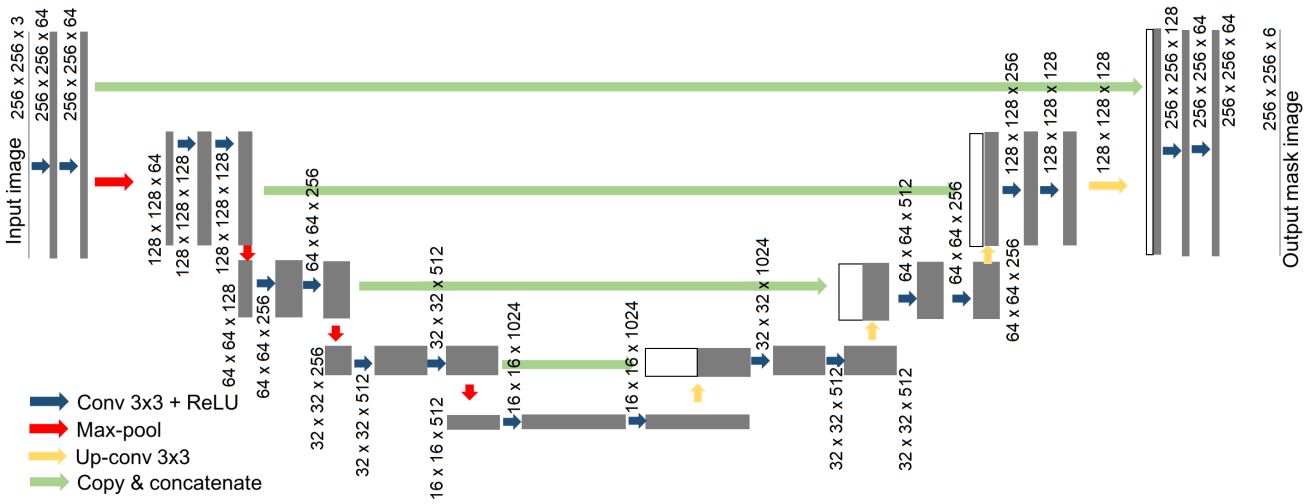

FIGURE 2 – Architecture d'un réseau U-Net [6].

est un voisin causal de s. Par conséquent, la markovianité spatiale est exprimée comme suit :

$$P(x_s|x_r, r \prec s) = P(x_s|x_r, r \precsim s) \tag{3}$$

La relation d'ordre total est définie par un schéma de visite de pixels impliquant la combinaison de quatre balayages en zig-zag et de deux balayages de courbes de Hilbert. Plus de détails sur la trajectoire de balayage des pixels peuvent être trouvés dans [13, 14]. Les équations (1)-(3) sont supposées vraies, définissant ainsi un cadre de modélisation des dépendances entre les pixels inter-couches et intra-couche. L'indépendance conditionnelle est un autre concept important car elle peut être utilisée pour décomposer des distributions de probabilités complexes en un produit de facteurs, chacun étant constitué du sous-ensemble de variables aléatoires correspondantes. Soit $\mathcal{Y} = \{y_s\}_{s\in S}$ le champ aléatoire des observations associé à tous les pixels de l'arbre quaternaire, le modèle d'observation $P(\mathcal{Y}|\mathcal{X})$ est défini par une factorisation :

$$P(\mathcal{Y}|\mathcal{X}) = \prod_{s\in S} P(y_s|x_s) = \prod_{\ell=0}^{L} \prod_{s\in S^\ell} P(y_s|x_s) \tag{4}$$

2.2 Apprentissage profond

L'apprentissage profond a prouvé être efficace dans la tâche de segmentation sémantique, notamment via les FCNs [5], comme U-Net [6] (voir Fig. 2) et SegNet [7]. Les FCNs sont des réseaux de neurones qui ne contiennent aucune couche dense, ayant ainsi la possibilité d'obtenir une sortie ayant la même taille que l'entrée. Les deux réseaux utilisent une architecture encodeur-décodeur, avec des couches de "pooling" et de "un-pooling" effectuant respectivement des processus de sous-échantillonnage et de sur-échantillonnage, réalisant ainsi des prédictions par pixel à la résolution d'origine. L'encodeur est basé sur VGG16 [24].

Ici, cette approche a été étendue pour exploiter le comportement intrinsèquement multi-échelles des CNNs avec un HMRF pour le traitement d'images multi-résolutions. La formulation multi-échelles d'un MRF hiérarchique sur un arbre quaternaire avec des treillis de pixels, ayant une relation de puissance de 2 entre les couches, correspond directement aux résolutions obtenues dans les couches intermédiaires d'un réseau convolutif par des couches de "pooling" de taille 2×2.

Les architectures considérées ont 5 blocs convolutifs, chacun contenant des couches convolutives, avec un filtre de taille 3×3 et un "zero padding" de dimension 1, suivis par l'activation ReLU et une "batch normalization". Chaque bloc convolutif est suivi d'une couche de "max pooling". A chaque étape de sous-échantillonnage, le nombre de filtres est doublé. Le décodeur, symétrique à l'encodeur, effectue le sur-échantillonnage et la classification, apprenant à restaurer la résolution spatiale complète tout en transformant les cartes d'entités codées en étiquettes finales. La dimension des patchs utilisés pour entraîner le réseau est de 256×256 pixels. La fonction de perte est calculée par un softmax pixel par pixel [15] sur la carte des caractéristiques finales, combinée avec la fonction de perte d'entropie croisée. Les activations du réseau à trois résolutions différentes sont insérées dans l'arbre quaternaire à travers trois connexions de saut, pour connecter le FCN au PGM hiérarchique, afin d'exploiter les informations multi-échelles (voir Fig. 1).

2.3 Algorithme d'inférence et critère MPM

Comme indiqué dans [13], l'estimation du maximum a posteriori (MAP) n'est pas satisfaisante pour la classification d'images multi-échelles. Le critère MPM [9, 12] est

particulièrement approprié pour les MRF hiérarchiques car il pénalise les erreurs selon l'échelle, évitant l'accumulation d'erreurs le long des couches [12]. On peut prouver que le MPM sur le modèle de Markov proposé est obtenu via les étapes récursives suivantes [14] :

$$P(x_s) = \sum_{x_{s^-}} P(x_s|x_{s^-})P(x_{s^-}), \tag{5}$$

$$P(x_s|y_s^d) \propto P(x_s|y_s) \prod_{t\in s^+} \sum_{x_t} \frac{P(x_t|y_t^d)P(x_t|x_s)}{P(x_t)}, \tag{6}$$

$$P(x_s^c|x_s, y_s^d) \propto \frac{P(x_s|y_s^d)P(x_s|x_{s^-})P(x_{s^-})}{P(x_s)^{n_s}} \cdot \prod_{r\lesssim s} P(x_s|x_r)P(x_r), \tag{7}$$

$$P(x_s|\mathcal{Y}) = \sum_{x_s^c} P(x_s^c|x_s, y_s^d)P(x_{s^-}|\mathcal{Y}) \prod_{r\lesssim s} P(x_r|\mathcal{Y}), \tag{8}$$

où y_s^d représente les observations de tous les descendants de s dans l'arbre (y compris s lui-même), x_s^c recueille les étiquettes de tous les sites connectés à s (ie, x_{s^-} et $\{x_r\}_{r\lesssim s}$), et n_s est le nombre de ces sites. Par conséquent, l'inférence MPM dans le cadre considéré est réalisée par trois étapes récursives. Premièrement, (5) est utilisé pour calculer $P(x_s)$ sur tous les sites via un passage descendant de la racine aux feuilles. Ensuite, (6) et (7) sont utilisés pour calculer $P(x_s|x_s^c, y_s^d)$ par un passage ascendant des feuilles vers la racine. Enfin, (8) est utilisé pour obtenir $P(x_s|\mathcal{Y})$ par une deuxième passe descendante.
Plus précisément, (5) est une application directe du théorème de probabilité totale, et la preuve de (6) dans le cas d'un modèle de Markov hiérarchique est identique à celle rapportée dans [12]. (7) et (8) sont valables sous les hypothèses d'indépendance conditionnelle suivantes :

$$\text{A1}: \quad P(x_s|x_s^c, \mathcal{Y}) = P(x_s|x_s^c, y_s^d) \tag{9}$$

$$\text{A2}: \quad P(x_s^c|\mathcal{Y}) = P(x_{s^-}|\mathcal{Y}) \prod_{r\lesssim s} P(x_r|\mathcal{Y})$$

$$\text{A3}: \quad P(x_s^c|x_s, y_s^d) = P(x_s^c|x_s) = P(x_{s^-}|x_s) \prod_{r\lesssim s} P(x_r|x_s)$$

Plus de détails peuvent être trouvés dans [13, 14]. L'hypothèse A1 signifie que l'étiquette de s, étant donné les étiquettes parent et frère/sœur, ne dépend que des observations des descendants de s et non de celles des autres sites. A2 implique que, étant donné le champ d'observation, les étiquettes parent et frère/sœur de s sont conditionnellement indépendantes. A3 signifie que les étiquettes parent et frère/sœur de s, lorsqu'elles sont conditionnées à l'étiquette s, sont indépendantes des observations des descendants de s et mutuellement indépendantes. Ces hypothèses sont similaires aux conditions d'indépendance conditionnelle qui sont généralement acceptées pour des raisons de commodité analytique dans le cas des MRF hiérarchiques [12] ou planaires [10].

Un schéma de visite symétrique sur la grille de pixels est utilisé pour éviter les artefacts anisotropes, par exemple une combinaison symétrisée du balayage en zig-zag et des courbes de Hilbert, comme expliqué dans [14]. La probabilité de transition $P(x_s|x_{s^-})$ à travers les échelles est définie par le modèle de Bouman [25], $P\{x_s = \omega|x_{s^-} = \omega\} = \vartheta$ pour tout $\omega \in \Omega$, où ϑ est un paramètre de la méthode, et $P\{x_s = \omega|x_{s^-} = \omega'\}$ constante pour tout $\omega \neq \omega'(\omega, \omega' \in \Omega)$ [14]. La probabilité de transition spatiale $P(x_s|x_r)$ ($r \precsim s$) est modélisée de manière analogue avec un paramètre ψ [14].

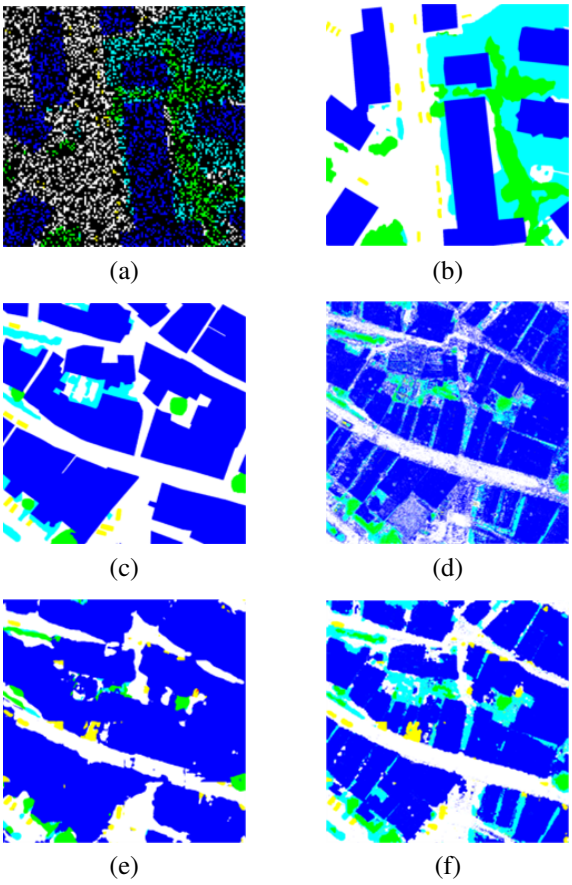

FIGURE 3 – **Cartes de vérité au sol et de classification :** (a) ensemble d'apprentissage détérioré (70 % de pixels non étiquetés), (b) ensemble d'apprentissage d'origine, (c) ensemble de test ; cartes de classification obtenues par : (d) RF, (e) U-Net, et (f) la méthode proposée ("Ver. 4"). Les classes : bâtiments, routes, végétation basse, arbres, véhicules.

Dans la méthode proposée dans cet article, les treillis S^l correspondent aux différentes résolutions impliquées dans le FCN, le vecteur d'observation y_s de chaque site $s \in S^l$ est obtenu en empilant toutes les activations de réseau associées à ce pixel positionné dans les couches à la résolution S^l, et la technique RF [16] est utilisée pour estimer les

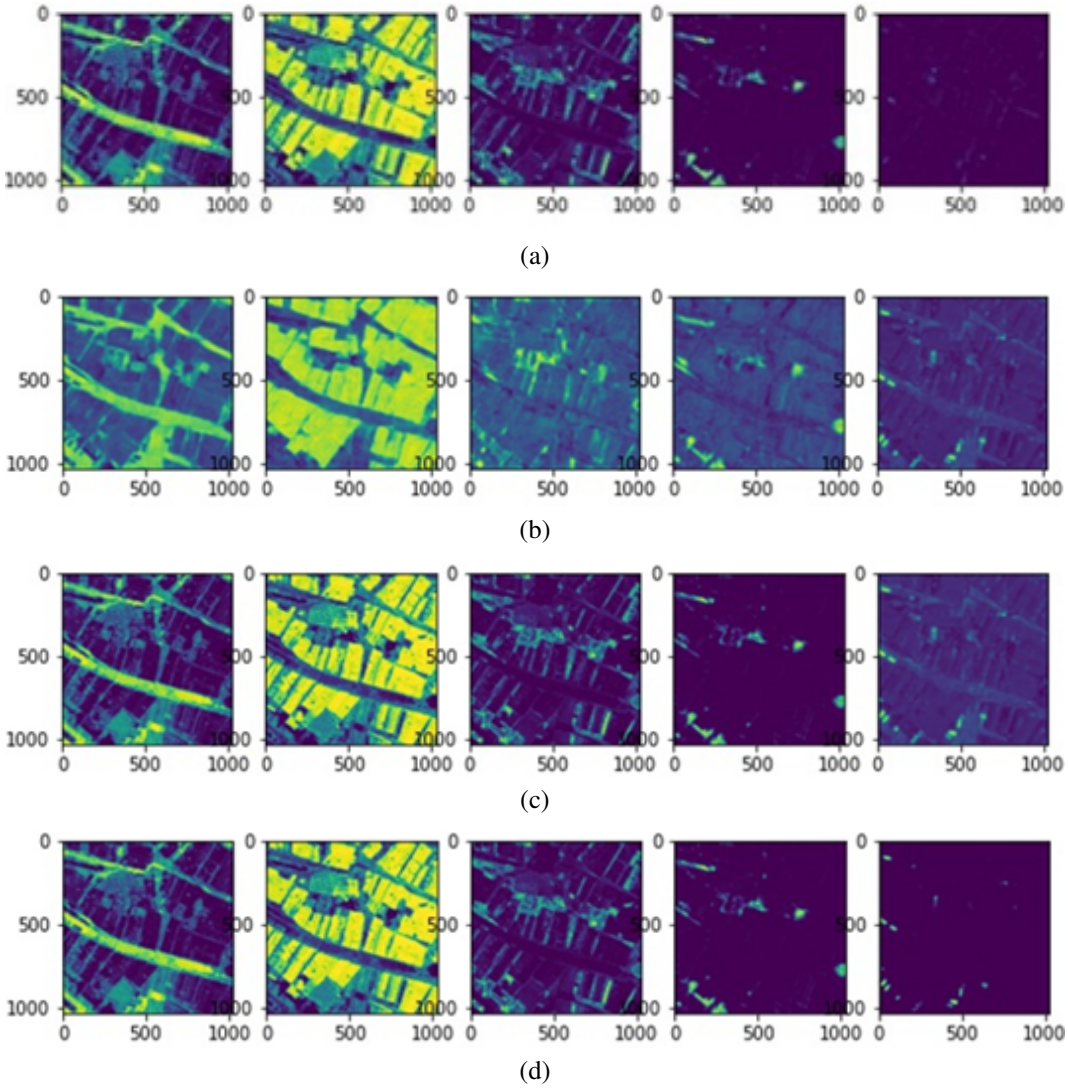

FIGURE 4 – **Probabilités a posteriori, différentes configurations :** (a) “Ver. 1”, (b) “Ver. 2”, (c) “Ver. 3” et (d) “Ver. 4”.

probabilités a posteriori $P(x_s|y_s)$ à partir des échantillons d’apprentissage des classes.

3 Validation expérimentale

La méthode proposée a été validée expérimentalement avec l’ensemble de données “ISPRS 2D Semantic Labeling Challenge Vaihingen” [1]. Il se compose d’images aériennes à très haute résolution (9 cm/pixel) avec six classes : routes, bâtiments, végétation basse, arbres, véhicules ainsi qu’une classe “autre”. Cependant, la classe “autre” est très mélangée et d’un intérêt relativement limité en tant que classe de couverture terrestre car elle comprend toutes les surfaces non incluses par les autres classes. Il y a 33 images IRRV et données DSM extraites d’un nuage de points LiDAR. Parmi les 16 images avec vérités de terrain (VT) “publiques”, 12 ont été choisies pour entraîner le réseau (images 1, 3, 23, 26, 7, 11, 13, 28, 17, 32, 34 et 37) et 4 pour le tester (images 5, 15, 21 et 30). Des tests ont été réalisés sur Google Colab. Une sous-partie des 1024 × 1024 pixels de l’image d’apprentissage 1 et une autre de l’image de test 5 ont été sélectionnées pour entraîner le RF et pour appliquer le PGM hiérarchique, respectivement, car il était impossible d’effectuer l’analyse sur des patchs plus grands en raison des limitations de RAM sur Google Colab. Ces images manquent d’instances de la classe “autre”, que nous avons exclue de l’expérimentation.

Les résultats de classification présentés dans cet article ont été obtenus avec : $L = 4$, soit quatre niveaux dans chaque arbre quaternaire, avec une relation de puissance de 2 entre les couches ; ϑ et ψ fixés à 0,82 (des valeurs plus élevées ou plus basses donnaient de moins bons résultats). Plusieurs conditions d’apprentissage ont été envisagées, car l’ensemble de données Vaihingen est “idéal”, avec des VT densément étiquetées, rarement disponibles dans des scénarios réels. Par conséquent, l’U-Net a été entraîné avec un

1. https ://www2.isprs.org/commissions/comm2/wg4/benchmark/

TABLE 1 – Résultats de la méthode proposée, appliquée sur l'ensemble de test avec quatre configurations différentes, et comparée au FCN standard.

Ensemble de données complet	**bâtiments**	**routes**	**végétation basse**	**arbres**	**véhicules**	**OA**	**rappel**	**précision**	**Cohen's κ**	**score F1**
U-Net standard	**0.85**	**0.97**	0.42	0.84	0.80	**0.90**	0.78	**0.72**	**0.81**	**0.75**
Méthode proposée - "Ver. 1"	0.81	0.92	**0.60**	0.80	0.25	0.86	0.68	0.64	0.74	0.66
Méthode proposée - "Ver. 2"	0.80	0.90	**0.55**	**0.92**	0.92	0.86	**0.82**	0.60	0.74	0.68
Méthode proposée - "Ver. 3"	0.78	0.91	0.59	0.80	**0.95**	0.86	0.81	0.58	0.74	0.68
Méthode proposée - "Ver. 4"	0.80	0.92	**0.60**	0.80	0.88	0.87	0.80	0.63	0.75	0.70
70% de pixels non étiquetés	**bâtiments**	**routes**	**végétation basse**	**arbres**	**véhicules**	**OA**	**rappel**	**précision**	**Cohen's κ**	**score F1**
U-Net standard	0.65	**0.97**	0.15	0.72	0.71	0.83	0.64	**0.67**	0.64	**0.65**
Méthode proposée - "Ver. 1"	**0.76**	0.90	**0.65**	0.65	0.24	**0.84**	0.64	**0.67**	**0.70**	**0.65**
Méthode proposée - "Ver. 2"	0.74	0.70	0.59	**0.83**	**0.86**	0.71	0.74	0.46	0.53	0.57
Méthode proposée - "Ver. 3"	0.70	0.85	0.64	0.66	0.68	0.79	**0.75**	0.54	0.63	0.63
Méthode proposée - "Ver. 4"	0.75	0.89	**0.65**	0.65	0.80	0.83	**0.75**	0.57	0.69	**0.65**
Érosion morphologique	**bâtiments**	**routes**	**végétation basse**	**arbres**	**véhicules**	**OA**	**rappel**	**précision**	**Cohen's κ**	**score F1**
U-Net standard	0.91	**0.92**	0.15	0.62	0.27	**0.87**	0.57	**0.77**	**0.75**	0.66
Méthode proposée, "Ver. 1"	0.76	0.90	**0.65**	0.66	0.24	0.86	0.58	0.55	0.73	0.56
Méthode proposée, "Ver. 2"	0.89	0.73	0.49	**0.67**	0.62	0.76	0.68	0.55	0.60	0.61
Méthode proposée, "Ver. 3"	0.88	0.86	0.50	0.60	**0.70**	0.85	**0.71**	0.57	0.72	0.63
Méthode proposée, "Ver. 4"	**0.93**	0.88	0.51	0.60	0.43	0.86	0.66	0.68	0.74	**0.67**

ensemble d'apprentissage "détérioré", avec des VT ayant un pourcentage de pixels non étiquetés (illustré à la Fig. 3(a)) ou par des opérateurs morphologiques (voir le tableau 1). Cette deuxième approche a été conçue comme une approximation des VT avec des patchs isolés de pixels étiquetés associés à différentes classes, généralement trouvées dans des applications réelles. Dans ce cas, cependant, la sélection des pixels étiquetés était bien équilibrée, préservant la probabilité a priori.

Les probabilités a posteriori par pixel sont insérées dans le PGM hiérarchique, comme mentionné précédemment. Cependant, celles obtenues par RF pour la classe 5, "véhicules", sur le treillis le plus fin (1024 × 1024 pixels) n'apparaissent pas suffisamment détaillées pour avoir une estimation appropriée dans les images résultantes (voir Fig. 3(d)). Cette configuration a été appelée "Ver. 1". Une façon de surmonter ce problème est de substituer ces probabilités a posteriori par celles obtenues dans la couche de sortie du réseau ("Ver. 2"), ou, en se concentrant uniquement sur les probabilités a posteriori de classe 5, soit en substituant l'estimation RF par la sortie du réseau ("Ver. 3"), ou avec un redimensionnement aux plus proches voisins de la même classe en treillis 512 × 512 ("Ver. 4", voir Fig. 3(f)). La justification de ces deux formulations est de permettre à l'approche proposée de se concentrer sur la discrimination des classes minoritaires (comme les arbres et les voitures). Les différentes configurations des probabilités a posteriori sont représentées sur la Fig. 4.

Les résultats quantitatifs présentés dans le tableau 1 confirment que l'approche proposée obtient des résultats plus élevés pour les classes minoritaires susmentionnées, et ces améliorations sont d'autant plus remarquables que les données d'entraînement se rapprochent des ensembles de données peu annotés disponibles pour les applications de cartographie réelles. Par exemple, avec 70 % de pixels non étiquetés, il y a une amélioration globale de la précision de la classification, particulièrement notable pour la classe "végétation basse", avec une augmentation de 50 %. Avec l'érosion morphologique, par exemple, alors que l'U-Net standard a obtenu une précision de 27 % dans la classe des véhicules, la méthode proposée a pu atteindre 70 %. Le tableau 1 montre que, dans ce cas, la methode proposée atteint des résultats plus précis pour la classification de chaque classe de l'ensemble de données, à l'exception de la classe "routes". En particulier, pour la classe "végétation basse", la methode proposée a obtenu une précision de 65 %, tandis que l'U-Net avait une précision de 15 %. De plus, dans toutes les situations considérées, le rappel atteint par l'approche proposée est supérieur à ceux des FCNs standards. En particulier, dans le cas de l'érosion morphologique, le gain en termes de rappel est de 14 %.

4 Conclusion

Une nouvelle approche pour la segmentation sémantique des images de télédétection fondée sur les CNNs, les PGMs hiérarchiques et les ensembles d'arbres de décision a été proposée dans cet article. Les résultats présentés indiquent que cette approche dépasse la précision du FCN standard pour le rappel. Ils montrent la possibilité de l'approche proposée d'exploiter la capacité de modélisation spatiale des modèles hiérarchiques de Markov pour atténuer les limites des approches FCNs en termes d'exigence de données d'apprentissage. Cette nouvelle méthode surpasse l'état de l'art dans la classification des classes minoritaires, tout en maintenant des résultats de classification adéquats pour les autres classes, comme le montre l'analyse quantitative des résultats. Une analyse visuelle des cartes de segmentation sémantique résultantes confirme ces conclusions et suggère que la méthode proposée est capable de récupérer les détails spatiaux qui ont été perdus dans les prédictions obtenues par le FCN à partir de cartes d'entraînement non denses sous-optimales.

Nos travaux futurs pourraient impliquer l'introduction de couches denses pour calculer les probabilités a posteriori

au lieu du classificateur RF. En outre, la nouvelle méthode, qui a été testée sur des ensembles de données de référence, sera étendue pour traiter des données associées à des applications réelles, par exemple pour la surveillance des catastrophes naturelles.

Références

[1] L. Gómez-Chova, D. Tuia, G. Moser et G. Camps-Valls, "Multimodal classification of remote sensing images : a review and future directions," *Proc. of the IEEE*, 103(9) : 1560–1584, 2015.

[2] A. Garcia-Garcia, S. Orts-Escolano, S. Oprea, V. Villena-Martinez et J. Garcia-Rodriguez, "A review on deep learning techniques applied to semantic segmentation," *arXiv preprint ArXiv :1704.06857*, 2017.

[3] X. Yuan, J. Shi et L. Gu, "A review of deep learning methods for semantic segmentation of remote sensing imagery," *Expert Systems with Applications*, 169, 2021.

[4] K. Nogueira, O. Penatti et J. dos Santos, "Towards better exploiting convolutional neural networks for remote sensing scene classification," *Pattern Recognit.*, 61 : 539–556, 2017.

[5] J. Long, E. Shelhamer et T. Darrell, "Fully convolutional networks for semantic segmentation," *IEEE Conf. Comput. Vis. Pattern Recognit.*, 3431–3440, 2015.

[6] O. Ronneberger, P. Fischer et T. Brox, "U-Net : Convolutional networks for biomedical image segmentation," *Med. Image Comput. Comput. Ass. Interv.*, 9351 : 234–241, 2015.

[7] V. Badrinarayanan, A. Kendall et R. Cipolla, "SegNet : a deep convolutional encoder-decoder architecture for image segmentation," *IEEE Trans. Pattern Anal. Mach. Intell.*, 39(12) : 2481–2495, 2017.

[8] N. Audebert, B. Saux et S. Lefèvre, "Semantic segmentation of earth observation data using multimodal and multi-scale deep networks," *Proc. Asian Conf. Comput. Vis.*, 180–196, 2016.

[9] Z. Kato et J. Zerubia, "Markov random fields in image segmentation," *Found. Trends Signal Process.*, 5(1-2) : 1–155, 2012.

[10] S.Z. Li, *Markov random field modeling in image analysis.* Springer, 2009.

[11] P. A. Devijver, "Hidden Markov mesh random field models in image analysis," *J. Appl. Stat.*, 20(5-6) : 187–227, 1993.

[12] J. Laferté, P. Pérez et F. Heitz, "Discrete Markov image modeling and inference on the quadtree," *IEEE Trans. Image Process.*, 9(3) : 390–404, 2000.

[13] I. Hedhli, G. Moser, S. B. Serpico et J. Zerubia, "Classification of multisensor and multiresolution remote sensing images through hierarchical Markov random fields," *IEEE Geosci. Remote Sens. Lett.*, 14(12) : 2448–2452, 2017.

[14] A. Montaldo, L. Fronda, I. Hedhli, G. Moser, S. B. Serpico et J. Zerubia, "A causal hierarchical Markov framework for the classification of multiresolution and multisensor remote sensing images," *ISPRS Ann. of Photogramm., Remote Sens., Spat. Inf. Sci.*, 3 : 269–277, 2020.

[15] I. Goodfellow, Y. Bengio et A. Courville, *Deep learning*. Boston, Massachussetts, USA : MIT Press, 2016.

[16] L. Breiman, "Random forests," *Mach. Learn.*, 45(1) : 5–32, 2001.

[17] X. X. Zhu, D. Tuia, L. Mou, G. Xia, L. Zhang, F. Xu et F. Fraundorfer, "Deep learning in remote sensing : A comprehensive review and list of resources," *IEEE Trans. Geosci. Remote Sens.*, 5(4) : 8–36, 2017.

[18] S. Nowozin et C. H. Lampert, "Structured learning and prediction in computer vision," *Found. Trends. Comput. Graph. Vis.*, 6(3–4) : 185–365, 2011.

[19] W. Zhao, S. Du, Q. Wang et W. Emery, "Contextually guided very-high-resolution imagery classification with semantic segments," *ISPRS J. Photogramm. Remote Sens.*, 132 : 48–60, 2017.

[20] W. Zhao, W. J. Emery, Y. Bo et J. Chen, "Land cover mapping with higher order graph-based co-occurrence model," *Remote Sens.*, 10, 1713(11), 2018.

[21] D. Buscombe et A. C. Ritchie, "Landscape classification with deep neural networks," *Geosci.*, 8(7), 2018.

[22] Y. Liu, S. Piramanayagam, S. Monteiro et E. Saber, "Semantic segmentation of multisensor remote sensing imagery with deep ConvNets and higher-order conditional random fields," *J. Appl. Remote Sens.*, 13 : 1–23, 2019.

[23] Z. Li, R. Wang, W. Zhang, F. Hu et L. Meng, "Multiscale features supported DeepLabV3+ optimization scheme for accurate water semantic segmentation," *IEEE Access*, 7 : 155787–155804, 2019.

[24] K. Simonyan et A. Zisserman, "Very deep convolutional networks for large-scale image recognition," *International Conference on Learning Representations*, 2015.

[25] C. A. Bouman et M. Shapiro, "A multiscale random field model for Bayesian image segmentation," *IEEE Trans. Image Process.*, 3(2) : 162–177, 1994.