CLASSIFICATION À TRÈS LARGE ÉCHELLE D'IMAGES SATELLITES À TRÈS HAUTE RÉSOLUTION SPATIALE PAR RÉSEAUX DE NEURONES CONVOLUTIFS

Tristan Postadjian[1], Arnaud Le Bris[1], Hichem Sahbi[2], Clément Mallet[1]

1: Univ. Paris Est, LASTIG MATIS, IGN, ENSG, F-94160 Saint-Mandé, France

2: CNRS, LIP6 UPMC Sorbonne Universités, Paris, France

Résumé

Les algorithmes de classification supervisée d'images satellites constituent un outil fondamental pour le calcul de cartes d'occupation des sols, à toutes les résolutions spatiales existantes. Ils ont permis d'établir la télédétection comme moyen le plu sfiable pour la génération de ces cartes. Les récents progrès en apprentissage automatique ont montré les très grandes performances des réseaux de neurones convolutifs pour de nombreuses applications, y compris l'interprétation d'images aériennes et satellites. Le travail présenté dans cet article établit une stratégie quant à l'utilisation d'un réseau de neurone convolutif pour la classification d'images satellites à très haute résolution spatiale (à savoir SPOT 6/7), couvrant de très larges régions géographiques, avec pour perspective future le calcul de cartes d'occupation des sols à l'échelle d'un pays.

Mots clés : Carte d'occupation des sols, Image satellite, Très Haute Résolution Spatiale, SPOT 6/7, large échelle, apprentissage, apprentissage profond, base de données géographiques.

Abstract

Supervised classification is the fundamental task for land-cover map generation, whatever the spatial resolution. Such techniques allowed to now consider remote sensing as the most reliable solution for generating such kinds of maps. Deep neural networks recently outperformed other state-of-the-art classifiers in many machine learning challenges, from semantic segmentation to speech recognition. Such strategies are now commonly employed in the literature for the purpose of land-cover mapping. This paper develops the strategy for the use of deep networks in order to label very high resolution satellite images (namely SPOT 6/7 images), with the perspective of land-cover mapping regions at the scale of a country.

Keywords : *Land-cover mapping, satellite images, Very High Spatial Resolution, SPOT 6/7, large-scale, learning, deep neural networks, geodatabases.*

1. Introduction

La classification d'une image acquise par télédétection géospatiale (aérienne ou satellite), dans un contexte de cartographie d'occupation des sols (OCS), consiste à attribuer à chaque pixel de cette image une classe : l'ensemble des classes retenues pour représenter la scène d'intérêt constitue une taxonomie, ou nomenclature, qui varie selon les besoins et/ou l'utilisateur final. L'attribution de cette classe s'appuie sur l'analyse visuelle propre au pixel mais peut aussi s'appuyer sur la description visuelle de son voisinage (Tupin et al., 2014). Les travaux existants utilisent tous des méthodes de classification supervisée (Khatami et al., 2016), reposant sur des algorithmes d'apprentissage automatique. Le terme *supervisé* provient de l'étape d'entraînement de l'algorithme qui consiste à modéliser les classes en jeu à partir d'un jeu de données de référence : on peut ensuite inférer les classes de données non étiquetées en leur appliquant le modèle ainsi entraîné. Le jeu de données de référence est un ensemble de pixels, décrits par des attributs (observations), et dont on connaît à l'avance les classes. Chaque classe est ainsi modélisée par rapport à ces attributs, qu'il s'agisse de méthodes génératives ou discriminatives. Il est important de noter que la qualité d'une classification dépend (i) du choix des attributs pour discriminer au mieux les classes entre elles et (ii) des données d'apprentissage. Jusqu'à très récemment, les méthodes discriminatives (à base de Séparateurs à Vaste Marge (Fauvel, 2007) ou de Forêts Aléatoires (Belgiu et Drăguţ, 2016)) ont été préférées aux méthodes génératives. Elles sont plus flexibles puisqu'elles lient directement les attributs avec les étiquettes fournies via l'ensemble d'apprentissage, résolvant ainsi directement le problème qui est évalué (Ng et Jordan, 2002).

Les techniques d'apprentissage profond, et plus particulièrement les **réseaux de neurones convolutifs** (CNNs pour *Convolutional Neural Networks*) répondent parfaitement aux critères énoncés. On trouvera une explication détaillé dans (Zhu et al., 2017). A l'inverse des méthodes nécessitant l'extraction d'attributs pertinents en amont du classifieur, les CNNs *apprennent* de bout en bout, au regard de la tâche de classification à effectuer (i) les attributs optimaux (filtres convolutifs), (ii) le classifieur, à partir du jeu d'apprentissage. Ainsi, on s'affranchit de

l'étape fastidieuse d'extraction d'attributs, requérant des connaissances expertes, subjectives, et qui varient selon les tâches à effectuer. Avec des images à très haute résolution spatiale et présentant un grand nombre de pixels et la couverture de grandes zones géographiques, les approches exhaustives les plus objectives possibles ne deviennent plus viables(Gressin et al., 2014; Tokarczyk et al., 2015). Les attributs *appris* par le CNN dépendant du jeu d'apprentissage, ils prennent implicitement en compte les relations entre les classes d'objets d'occupation des sols. La conséquence directe de l'apprentissage des attributs en même temps que du classifieur est le besoin massif de données d'apprentissage, puisqu'il s'agit de déterminer l'ensemble des filtres constituant le CNN.
C'est en partie grâce aux très grandes bases de données d'images annotées que les CNNs ont été récemment remis au devant de la scène dans la communauté de vision par ordinateur avec le succès de Krizhevsky et al. (2012) lors du challenge ILSVRC 2012, avec des performances bien au-dessus de méthodes autres que celles employant des CNNs. Les CNNs sont, aujourd'hui, la référence dans de nombreuses applications de la communauté vison par ordinateur, de la reconnaissance d'objets (Girshick et al., 2014; Redmon et Farhadi, 2017) à la classification d'images (He et al., 2016). Simonyan et Zisserman (2014) et Szegedy et al. (2015) ont démontré que les performances d'un CNN augmentent avec sa profondeur ; cependant, multiplier les couches de convolution d'un CNN multiplie également le nombre de paramètres à optimiser, pouvant mener à un cas de surapprentissage. Le modèle s'adapte exactement au jeu d'entraînement, menant à de faibles performances sur de nouvelles données non étiquetées.

Ce travail vise à l'analyse les images satellites des capteurs SPOT 6/7 fournies annuellement par le pôle de données THEIA [1]. Il s'agit d'évaluer la capacité à fournir une occupation du sol simple et générique sur l'intégralité de la France, à très haute résolution spatiale (environ celles des capteurs en entrée) à partir des méthodes CNN. L'objectif et l'évaluation sont ainsi opérationnels dans le sens où on cherche à proposer une solution avec un paramétrage et des temps de calculs limités, valide sur l'intégralité du territoire français. L'occupation des sols a une finalité statistique ou cartographique et nous nous situons en amont de ce choix : on ne se focalisera donc pas sur la délimitation précise des objets couvrant la surface terrestre.
Cet article met en œuvre une approche dite « au patch », qui sera décrite en section 2 dans le cadre d'une analyse des travaux existants pour l'occupation des sols à l'aide de CNNs. Ce choix est notamment motivé par sa relative simplicité d'implémentation dans le cadre d'une étude d'évaluation du potentiel des CNNs pour de la cartographie d'occupation des sols à l'échelle d'un pays. En effet, à notre connaissance, l'ensemble des investigations

1. http ://www.theia-land.fr/

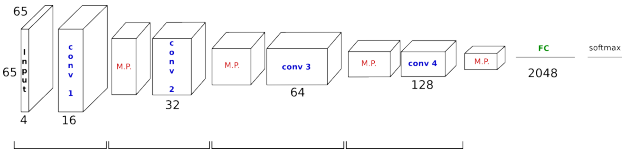


FIGURE 1: Architecture à quatre couches mis en place dans le cadre de notre étude.

sur le sujet de l'occupation des sols réalisée par apprentissage profond ré-utilisaient des réseaux pré-entraînés, qui ne permettent que l'analyse d'images composées de 3 bandes spectrales. Or, les images utilisées dans ce travail ont été acquises par les capteurs des satellites SPOT 6/7. Elles comportent 4 bandes spectrales (Rouge, Vert, Bleu, Infra-Rouge) à 1,5 m de résolution. L'intérêt dans l'utilisation de réseaux pré-entraînés est de profiter des entraînements massifs pré-existants. Si ceux-ci ont été effectués sur des images à trois bandes, les filtres calculés lors de cet entraînement sont valides pour de telles données : on ne peut donc pas simplement ajouter une bande, afin de gérer des images quatre bandes. L'approche proposée ne met pas en jeu de séries temporelles, car nous ne disposons que d'une seule image par année par région géographique. En outre, pour valoriser l'apport des images SPOT 6 et 7 d'une part et pour des raisons de fréquence de mise à jour d'autre part, nous n'utilisons pas de données d'altitude, sous la forme d'un Modèle Numérique de Surface par exemple, en complément de l'information radiométrique contenue dans les images. Enfin, afin d'analyser le comportement d'un CNN sur de nouvelles zones, on se restreint à faible nombre de classes génériques (*bâti, végétation, champs, routes, eau*). Toutefois, nous montrons le potentiel d'une nomenclature hiérarchique en partant d'une nomenclature générique pour affiner les classes existantes en sous-classes, en affinant le réseau lui-même pour qu'il puisse détecter ces nouvelles sous-classes.

2. État de l'art

Deux approches existent pour la cartographie de l'occupation des sols. Les approches « par patch », comme décrit dans Girshick et al. (2014), infèrent pour chaque patch centré sur un pixel le score d'appartenance pour chacune des classes. Les approches « denses » (initiées par Long et al. (2015)) permettent d'apprendre au sein de chaque patch des arrangement spatiaux entre classes au niveau du pixel. Ces dernières permettent notamment une géométrie plus précise des objets à classer. Ces approches denses reposent sur des architectures *encoder-decoder*. La première partie (*encoder*) permet de traduire l'information contenue dans l'image sous forme de vecteurs d'attributs de haut niveau, et la seconde (*decoder*) produit une carte de chaleur relative aux probabilités d'appartenance aux classes considérées, de résolution semblable à l'image d'origine, en utilisant l'information issue de l'encoder.

La communauté de télédétection s'intéresse depuis récemment aux CNNs pour les tâches de classification monotemporelles d'images aériennes et satellites. La plupart des travaux de télédétection s'appuient sur les architectures *fully convolutional* (FCN). Les FCN permettent une classification au pixel très précise, levant des ambiguïtés que l'approche par patch aurait, par exemple sur la distinction entre une voiture et le parking sur lequel elle se situe. L'utilisation jointe de l'information issue de l'image et d'un Modèle Numérique de Surface a été exploitée par Marmanis et al. (2018) en créant deux réseaux en parallèle, un pour chaque modalité, puis en les fusionnant à haut niveau d'abstraction. L'ajout de « skip-connections » permet de réinjecter de l'information haute fréquence sur la partie decoder pour retrouver la résolution initiale (réseaux Sharpmask ou RefineNet par exemple). Les mêmes auteurs se sont également efforcés d'améliorer la géométrie des objets détectés dans Marmanis et al. (2016).

En utilisant les mêmes données, Sherrah (2016) crée un FCN sans l'inconvénient du sous-échantillonnage dû aux couches de max-pooling, grâce à l'algorithme « a trous » (Chen et al., 2014) qui remplace ces couches de max-pooling, préservant ainsi la dimension initiale de l'image. Leur résultat est l'un des meilleurs sur les jeux de données de parangonnage de Potsdam et Vaihingen de l'ISPRS avec Volpi et Tuia (2017), travail dans lequel les auteurs utilisent un FCN avec une couche de déconvolution 3×3 en remplacement des couches fully-connected. Une comparaison entre approches « au patch » et *fully convolutional* montre que la seconde permet des temps de calculs au moment de la prédiction bien plus rapide qu'en n'utilisant une approche « au patch ».

Audebert et al. (2016) utilisent des données similaires dans le réseau SegNet(Badrinarayanan et al., 2015) pour effectuer une analyse multi-échelles dans la partie decoder. L'utilisation de réseaux de neurones convolutifs nécessitant un volume d'apprentissage très conséquent, l'extraction automatique de jeux d'apprentissage issus de bases de données géographiques est devenu indispensable : dans Kaiser et al. (2017), les auteurs utilisent OpenStreetMap pour générer leurs données d'entraînement. L'utilisation massive de telles données permet de s'affranchir de paramétrages spécifiques quant à la génération automatique de ces données comme il a pu être nécessaire dans d'autres travaux (Gressin et al., 2013; Inglada et al., 2015; Maas et al., 2016; Pelletier et al., 2016; Dechesne et al., 2017).

Plus récemment, Chen et al. (2018a) utilisent le « shuffling operator », introduit par Shi et al. (2016) afin d'améliorer la détection d'objets de taille réduite en accroissant la résolution des couches en sortie du réseau. Le réseau utilisé s'appuie sur la mécanique « atrous » (*dilated convolutions*)(Chen et al., 2018c) permettant d'accroître le champ réceptif des filtres sans augmenter le nombre de paramètres. Chen et al. (2018b) améliorent le processus en utilisant conjointement des attributs géométriques issus d'un MNS et radiométriques, pré-calculés en amont du réseau « atrous ». Enfin, les processus de fusion multi-capteurs est abordée dans (Audebert et al., 2018), où les différentes stratégies (fusion précoce ou tardive) apportent des avantages différents. L'ensemble de ces travaux mettant en jeu l'apprentissage profond pour le calcul de carte d'occupation des sols contraste avec notre propre approche pour plusieurs raisons : l'utilisation de strictement plus de trois bandes spectrales n'a pas été exploré à notre connaissance en raison des réseaux qu'utilisent les travaux existants. Ceux-ci utilisent en particulier des architectures pré-entraînés (Simonyan et Zisserman, 2014; Szegedy et al., 2015) à l'aide d'images 3 bandes. Par ailleurs, le contexte dans lequel nous nous plaçons est éloigné des études citées précédemment d'un point de vue de l'objectif voulu : s'ils visent à améliorer plus ou moins sensiblement les classifications existantes sur des données issues de parangonnage tels que les jeux de données de Potsdam et Vaihingen proposés par l'ISPRS, le but est ici de fournir des méthodes pour que le calcul de carte d'occupation des sols soit opérationnel et ainsi déployable à large échelle.

3. Méthodologie

Les composantes principales de la chaîne sont décrites dans cette partie : (i) les deux architectures testées pour mettre en évidence l'impact du nombre de couches, (ii) la création du jeu de données d'entraînement, (iii) les deux stratégies d'apprentissage du CNN, (iv) la phase de test sur de nouvelles données.

Toutes les expériences ont été menées sur un ordinateur standard de bureau avec une carte graphique Nvidia GeForce GTX 980 dotée de 4 Go de mémoire vidéo, et un processeur Intel Core i7-4790 (8 Go de RAM).

3.1. Architectures

L'approche au « patch » permet de construire des réseaux facilement, sans contrainte. En revanche, classifier des images à très large échelle peut être lourd en temps de calcul, surtout sur une machine dont les spécifications techniques sont mentionnées précédemment. Ceci explique pourquoi nous contraignons le réseau à être moins dense que des architectures déjà existantes.

Les patch en entrée du réseau doivent avoir une taille fixe : pour décider de cette taille, nous avons dû prendre en compte le fait que l'on cherche aussi bien à classer de petits objets (bâtiments) que des objets étendus (champs, forêts). Une fenêtre de $65\times65\times4$ (on utilise les quatre bandes disponibles) pixels est un compromis qui permet de prendre en compte les classes d'intérêt dont certains objets seraient petits (bâtiment isolé en milieu rural). Cela correspond à une emprise au sol d'environ 100 m de côté.

Deux architectures ont été créées pour constater l'impact du nombre de couches : 3 couches et 4 couches ont été testées. Celle à quatre couches est illustrée en Figure 1. Successivement, des convolutions opèrent sur les couches précédentes. On accroit alors l'abstraction et la complexité des représentations lorsqu'on s'éloigne de l'image d'entrée. Chaque couche est constituée de filtres

de convolution 3×3, auxquels succèdent des fonctions d'activation non linéaires, prenant ainsi en compte le caractère non linéaire du problème (Zhu et al., 2017). La fonction d'activation choisie est celle commune à toutes les architectures récentes, le « Rectified Linear Unit » (ReLU). Les couches de convolution sont séparées par des couches dites de « max-pooling 2×2 » (MP). Cela consiste à ne garder, sur un voisinage 2×2, que le maximum, préservant ainsi l'information locale la plus essentielle. Celles-ci permettent également (i) d'introduire une invariance locale en translation, (ii) d'accroître le champ réceptif des filtres au fur et à mesure que l'on progresse dans le réseau, (iii) de réduire les temps de calcul.
En fin de réseau, une couche « fully connected » combine l'ensemble des filtres de la couche précédente pour créer des attributs à haut niveau d'abstraction à partir du patch en entrée. Enfin, le critère de « cross-entropy » à minimiser est suivi d'un « softmax » pour rendre les scores homogènes à des probabilités.
L'ensemble des implémentations CNNs ont été menées en utilisant Torch [2].

3.2. Construction du jeu d'apprentissage

La constitution des données qui permettront l'apprentissage se fait en exploitant les bases de données géographiques nationales existantes (Figure 2). Une stratégie alternative consiste, en l'absence de données de référence suffisantes, à générer synthétiquement ces données (Kemker et al., 2018). Dans notre cas de figure, ces bases sont très massives et mettent à notre disposition de nombreux polygones appartenant aux classes d'occupation des sols que l'on cherche à discriminer. Elles sont géoréférencées, se superposant aux images SPOT parfaitement. La précision des polygones est de 1 m, ce qui est compatible avec la résolution spatiale de nos images. Une imprécision géométrique sur la base de données aura pour conséquence une erreur au plus inférieure au pixel sur la cartographie finale. En revanche, des erreurs sémantiques peuvent apparaître car ces bases de données ne sont pas nécessairement à jour sur l'ensemble du territoire. L'utilisation de CNNs pour apprendre massivement sur ces bases de données permet de réduire l'influence de ces erreurs lors de la phase d'entraînement.
Pour constituer nos patch d'apprentissage (65×65×4), on sélectionne régulièrement sur la zone d'apprentissage des pixels autour desquels sont extraits leur voisinage 65×65. Le nombre total d'échantillons s'élève à 10 000, dont 10% sont conservés pour valider le modèle. Dans un souci de généralisation, les échantillons subissent aléatoirement des transformations afin (i) de limiter le surapprentissage, (ii) d'obliger le réseau à modéliser certaines invariances. Notamment, les images étant acquises depuis des satellites, une invariance aux rotations par rapport au nadir est nécessaire dans la modélisation de nos classes. Nous présentons sur la Figure 3 des échantillons d'apprentissage construits selon le protocole décrit précédemment. Pour chaque classe, on constate une variabilité importante des objets représentant ces classes, motivant davantage l'utilisation des CNN pour le calcul d'attributs représentatifs de chaque classe.

2. http ://torch.ch/

3.3. Phase d'entraînement

Les réseaux décrits en Section 3.1 sont entraînés sur les données d'apprentissage construites selon la section 3.2 sur des zones géographiques spécifiques (voir Section 4). Deux stratégies (complémentaires) ont été considérées : (i) le réseau RWI pour « Random Weight Initialization », pour laquelle les paramètres des CNNs sont initialisés aléatoirement et (ii) le réseau FT pour « Fine-Tune », dont les paramètres sont copiés depuis un réseau existant. Le réseau prend, en pratique, des batchs de 200 échantillons d'apprentissage en entrée, pour accélérer la convergence et les temps de calcul. Les transformations aléatoires sont appliquées sur chaque batch avant son passage dans le réseau.

3.3.1. Random Weight Initializationé (RWI)

Aucun réseau existant n'accepte, à notre connaissance, en entrée, quatre bandes RVB-IR. L'entraînement d'un réseau *from scratch* est donc obligatoire en passant par une initialisation aléatoire des paramètres de notre architecture. Nous avons paramétré le nombre d'époques à 5 000, n'ayant pas d'*a priori* sur le nombre nécessaire pour atteindre un point de convergence satisfaisant.

3.3.2. Fine-Tuning (FT)

De nombreuses architectures ont été largement entraînées sur des bases d'images 3 canaux RVB très volumineuses, offrant ainsi l'occasion d'utiliser des paramètres pré-calculés et très robustes. Cependant, des architectures telles que (Simonyan et Zisserman, 2014; Szegedy et al., 2015) possèdent de très nombreux hyperparamètres et y sont très sensibles. Elles peuvent rendre leur utilisation difficile lorsqu'elles sont utilisées dans des cas éloignés de leurs cas d'utilisation premiers (Papadomanolaki et al., 2016).
Affiner (ou « fine-tune ») un réseau existant repose sur le même principe itératif que pour le RWI, sauf au moment de l'initialisation, pour laquelle on récupère des paramètres pré-entraînés, que l'on affine sur le nouveau cas à traiter. En effet, de nouvelles données peuvent présenter des spécificités absentes du jeu d'entraînement sur lequel un réseau à été précédemment entraîné. L'idée est d'absorber ces nouvelles spécificités en ajoutant des données d'apprentissage du nouveau problème à un réseau qui convergeait déjà sur un autre jeu. Les couches les plus profondes sont souvent les seules à être *affinées* : le travail de Zeiler et Fergus (2014) a montré que les couches les plus proches de l'image extraient des attributs bas niveaux, tels que des contours, ou coins. Ils sont communs à l'ensemble des images que l'on peut trouver. En revanche, les couches les plus profondes modélisent des caractéristiques de plus en plus dépendantes

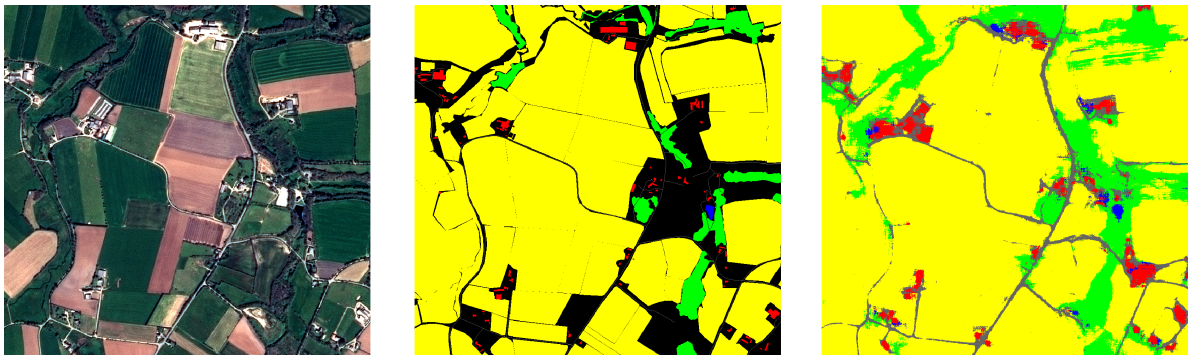


FIGURE 2: De gauche à droite : image SPOT, données de référence, classification. ● *pas de données de référence,* ● *bâti,* ● *route,* ● *culture,* ● *végétation,* ● *eau.*

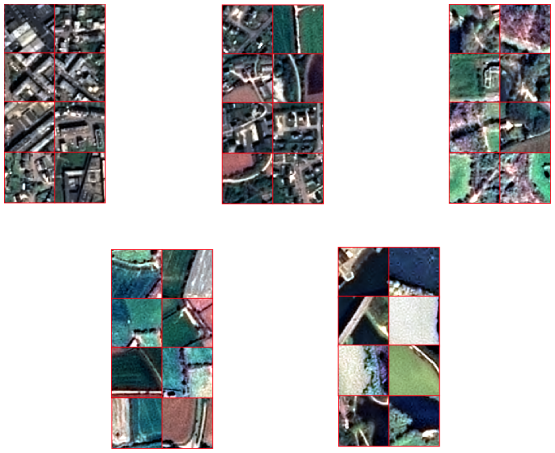


FIGURE 3: Aperçu des échantillons pour chaque classe. De gauche à droite et de haut en bas : *bâti*, *route*, *végétation*, *culture* et *eau*.

du cas d'utilisation. Ce constat permet ainsi, pour des réseaux très profonds, de ne ré-entraîner que les couches les plus profondes, accélérant les temps de calcul, et réduisant grandement le nombre de paramètres.
Le fine-tuning facilite grandement la convergence sur de nouvelles zones ou de nouvelles époques d'acquisition, tout en réduisant le nombre d'échantillons nécessaires à l'entraînement et le nombre d'itérations. Nous rappelons une nouvelle une fois, comme expliqué en Section 1, que l'utilisation de réseaux d'état de l'art dans notre cas est impossible du fait de l'utilisation d'images satellites quatre bandes.

3.4. Phase de classification

Avec une approche par patch, une fenêtre glissante parcourt l'ensemble de la zone à classifier. La fenêtre est centrée sur chaque pixel, et on applique le réseau sur un voisinage 65×65. La classe majoritaire à l'issue du réseau est affectée au pixel central. Les images SPOT étant très volumineuses (jusqu'à 20 Go), afin d'éviter toute perte d'information en cas d'erreur, celles-ci sont découpées en tuiles de 2 000×2 000 pixels.

4. Expérimentations

Toutes les applications reposent sur cinq classes d'occupation : *bâti, route, végétation, culture, eau*. Cependant, si nous verrons en Section 5 qu'une telle nomenclature entraîne des limitations importantes, nous étudions également le cas en Section 4.3 où l'on souhaite améliorer (i) la granularité sémantique et (ii) le résultat global en ajoutant une nouvelle classe de haies. Enfin, le processus de classification étant long en terme de temps de calcul, une stratégie mettant en jeu une segmentation préalable de l'image est présentée en Section 4.4, dans le but de réduire ces temps de calcul.

4.1. Données images et choix des régions d'intérêt

Les images ont été acquises par les capteurs des satellites SPOT 6/7. Elles et sont mises à disposition par le pôle de données THEIA. Nous disposons d'une image par année et par zone. Les quatre bandes, à 1,5m une fois le processus de pan-sharpening effectué, sont utilisées pour extraire le plus d'information des images (ne disposant pas d'information d'altitude, nous souhaitons tirer parti du maximum d'information radiométrique disponible dans les images satellites).
Les tests ont été effectués sur le département du Finistère pour sa grande diversité de paysage. Deux sous-régions sont considérées plus particulièrement pour entraîner les réseaux et étudier leur capacité à bien généraliser sur de nouvelles données et paysages : autour de la ville de Brest (notée ROI-1, milieu urbain) et de Le Faou (notée ROI-2, milieu rural). En plus de l'impact d'un changement géographique, différentes dates sont considérées : 2014 et 2016. Ce choix sur l'étude d'un changement temporel permet de jauger le potentiel du modèle proposé dans le cadre de production de cartes d'occupation des sols à diverses dates pour de futures analyses (détection de changement, évolution des forêts par exemple).

4.2. Stratégies d'entraînement

Il est essentiel d'analyser les différentes stratégies d'entraînement pour être capable de produire une classification à large échelle, pour laquelle il faudra classifier

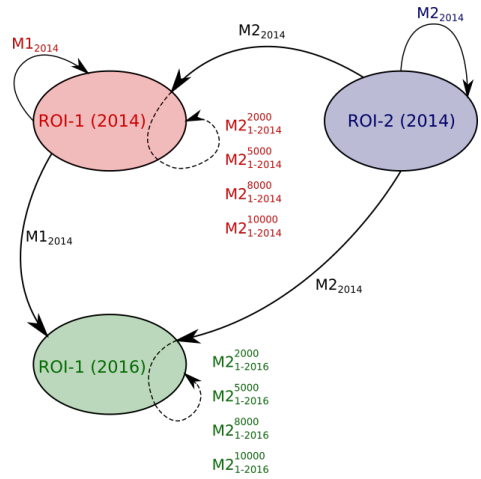


FIGURE 4: Stratégies d'entraînement (cf. Section 3) - pointillés : réseau FT sur de nouvelles zones ROI-1(2014), ROI-1(2016), ROI-2(2014) représentent respectivement les images SPOT acquises sur Brest en 2014 et 2016, et Le Faou en 2014.

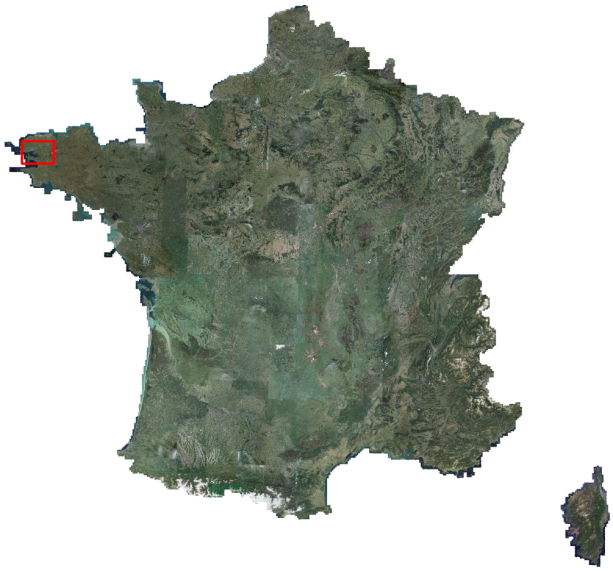


FIGURE 5: Couverture SPOT 6 et 7 en 2014 sur la France - ROI-3 correspond au cadre rouge.

des régions thématiques très différentes, à des dates différentes. La Figure 4 détaille l'ensemble de ces stratégies et cas d'application.

4.2.1. Réseau entraîné par RWI

Deux jeux d'entraînement ont été constitués pour les régions ROI-1(2014) et ROI-2(2014). Un entraînement « from scratch » est obligatoire dans les deux cas, comme mentionné précédemment, puisqu'on classe 4 bandes et que les réseaux existants n'en acceptent que 3 au plus en entrée. Les deux expérimentations sont notées $M1_{2014}$ et $M2_{2014}$ sur la Figure 4. $M1_{2014}$ correspond à l'entraînement RWI sur ROI-2, en 2014, du réseau à 4 couches. En revanche, $M1_{2014}$ correspond à l'entraînement RWI sur ROI-1, en 2014, des réseaux à 4 couches et à 3 couches. La comparaison entre 3 et 4 couches n'est étudiée qu'ici. Tous les autres réseaux seront des réseaux à 4 couches.

4.2.2. Fine-tuning : changement géographique et temporel

L'intérêt du fine-tuning est notamment d'affiner un réseau sur une nouvelle donnée en apprenant ses nouvelles spécificités. Le processus est très intéressant ici : le réseau entraîné par RWI sur ROI-2, qui décrit un paysage rural, extrait des attributs qui caractérisent difficilement ROI-1 qui est un milieu urbain dense. La capacité de généralisation du réseau, via le fine-tuning, va permettre de capturer le caractère "urbain dense" de la nouvelle scène qu'est ROI-1 en sélectionnant des patch d'apprentissage sur cette zone. Sur la Figure 4, les réseaux FT sont notés MN^{xxxx}_{P-year}, ou N et P désignant, respectivement, la région où le réseau a été pré-entraîné (par RWI), et la région où le réseau est affiné à la date $year$, en prenant un nombre $xxxx$ sur la région P pour l'affiner.

Si le fine-tuning d'une région géographique vers une autre a un intérêt évident, c'est aussi le cas quant à l'utilisation d'un réseau existant entraîné sur une zone à un instant *t* sur la même zone à un instant *t'*. La détection de changement est, en effet, un cas d'utilisation direct des cartes d'occupation des sols. Des différences d'apparences de végétation, dues à des saisons différentes, ou des rotations de culture, peuvent induire une variabilité intra-classe très forte d'une époque à une autre. Pour étudier cela, nous avons tout d'abord appliqué le réseau 4 couches $M1_{2014}$ sur ROI-1(2016), ce qui correspond au test du réseau appris en 2014 sur la **même zone** en 2016. Puis, pour tester un changement géographique et temporel, nous avons appliqué le réseau $M2_{2014}$ sur ROI-1(2016).

4.2.3. Classification sur de larges régions

Nous avons mené une classification sur une partie du Finistère (26%), représentant 1 755 km^2 (notée ROI-3), soit environ 780 millions de pixels. La classification s'est faite par application du réseau 4 couches $M1_{2014}$. Aucun ajustement n'a été effectué au moment du passage à l'échelle, ce qui pourrait nous porter préjudice car la région est principalement composée de paysages urbains, tandis que $M1_{2014}$ caractérise un paysage urbain dense (car entraîné sur Brest). De plus, les classes d'occupation varient beaucoup spectralement par rapport aux patch qui ont servi d'apprentissage pour $M1_{2014}$. En effet, les estuaires sont des zones humides et les landes sont des objets présents sur ROI-3 mais tous deux absents absents du jeu de données d'apprentissage. L'emprise de la zone ainsi classifiée est visible sur la Figure 5.

4.3. Affinage sémantique

L'étude du potentiel des CNN pour le calcul de carte d'occupation des sols à large échelle a été menée en

considérant cinq classes génériques, afin de jauger l'impact d'autres facteurs. Toutefois, la question de la gestion de classes supplémentaires est pertinente. Nous avons traité le cas des haies sur une zone extraite de la région Finistère dans cette étude (Figure 7). En effet, nous avons constaté que les classes de *routes* et de *végétation* étaient fortement confondues. La cause majeure de ceci provient de la nature de certains objets de la classe *végétation* qui se trouve être des haies, objets très fins et similaires morphologiquement aux routes. Par ailleurs, on peut voir sur la figure que les haies bordent souvent les routes dans les paysages aussi bien urbains que ruraux. Afin d'obtenir de meilleurs résultats sur la classe *route*, nous avons donc subdivisé la classe *végétation* initiale en classes *haies* et *végétation* (qui n'inclue donc plus les haies).
Nous mettons en place le procédé de fine-tuning présenté en Section 3.3.2, déjà mis en œuvre pour classer (i) de nouvelles zones et (ii) des zones imagées à des dates différentes (Section 4.2.2). A l'instar de ces deux cas d'étude, il a fallu constituer au préalable un jeu d'apprentissage sur lequel affiner le réseau, en incluant la classe *haies* et en retravaillant la classe *végétation* pour que le jeu associée à celle-ci ne comporte plus de haies. Ce jeu, de 5 000 échantillons a été extrait de la zone ROI-3 dans sa totalité (Figure 5).

4.4. Sur-segmentation de l'image à classifier

Nous nous intéressons au calcul de carte d'occupation des sols dans un contexte de large échelle (régions, pays, continents). Cela implique certaines contraintes : en particulier, une classification au pixel est un processus qui peut devenir très vite coûteux en temps de calcul et en mémoire machine. De plus, l'approche adoptée pour la phase de prédiction est celle de la fenêtre glissante détaillée en Section 3.4. Dans le but de réduire l'impact de cette stratégie sur les temps de traitement, nous proposons de segmenter l'image à classer en amont de cette classification. Pour cela, deux options sont possibles : (i) adopter une approche objet, (ii) produire une sur-segmentation de l'image en segments plus petits que des objets mais homogènes en radiométrie. Le choix s'est porté sur une sur-segmentation, ou approche superpixels. En effet, cela produit un résultat pour lequel les objets sémantiquement significatifs sont divisés en segments, au lieu d'avoir l'objet en entier pour une approche purement OBIA (Object Based Image Analysis). Toutefois, nous privilégions la pureté de chaque superpixel : chacun d'entre eux doit rassembler tout ou une partie d'un objet. Une approche OBIA est très susceptible de produire des segments incluant des pixels n'appartenant pas à celui-ci. Un grand nombre d'algorithmes de sur-segmentation existent, présentant tous des avantages et des inconévients différents. Par exemple, l'algorithme SLIC (Simple Linear Iterative Clustering), développé par Achanta et al. (2012) permet de produire une grille de superpixels à l'aspect très régulier (mosaïque), en calculant un K-means intégrant les coordonnées dans l'image de chaque pixel en plus de son information radiométrique. Les algorithmes de segmentations produisant de tels résultats (Stutz et al., 2018) ne peuvent être envisagés dans notre car les objets que nous souhaitons classés sont (i) différents en superficie et (ii) présentent des morphologies alternant entre fine (*route*) et large (*culture*). L'algorithme utilisé (Felzenszwalb et Huttenlocher, 2004) permet de produire des segments de formes variées et paramétré pour que ceux-ci soient restreints spatialement. Cette méthode repose sur l'analyse d'un graphe et la comparaison entre deux scores. La première représente le score entre deux superpixels adjacents et correspond à la variation minimale de radiométrie entre deux pixels (chacun appartenant à l'un des superpixels). Le second représente l'homogénéité au sein de chaque superpixel. Si celui-ci est supérieur au premier, alors le poids entre ces deux superpixels est important (chaque superpixel définit une région homogène en son sein et différente de l'autre). En plus d'un paramètre de lissage gaussien en amont de la segmentation, *m* permet de contrôler la taille minimale des superpixels est possible (enpixels) et *k* définit une échelle d'analyse (*k* élevé produit de plus larges superpixels). Nous avons ainsi opté pour cette méthode. Idéalement, le partitionnement devrait être appris au même titre que la classification, de manière supervisé. C'est un champs d'investigation naissant (Tu et al., 2018).
La stratégie de classification est la suivante :

1. Segmentation de l'image en superpixels ;
2. Pour chaque superpixel :
   (1) sélectionner régulièrement au sein du superpixel un pourcentage de pixels ;
   (2) classifier chaque pixel ;
   (3) attribuer au superpixel la classe majoritaire.

On peut voir un exemple de segmentation sur la Figure 6, produite avec le paramétrage suivant : $(k, m) = (30, 20)$. On peut noter que des valeurs voisines de celles choisies produisent des cartes de segmentation similaires. Le résultat de segmentation montre des segments cohérents avec la morphologie semi-urbaine de la scène, et ne présente pas de segment mixtes contenant des pixels de classes d'occupation différentes.

5. Résultats

5.1. Métriques d'évaluation

Les résultats obtenus ont été qualifiés par validation croisée avec les bases de données géographiques existantes (Figure 2). Il s'agit de la vérité terrain la plus exhaustive à disposition même s'il faut bien indiquer qu'elles comportent un certain taux de fausses étiquettes (erreurs ou changements depuis la génération de ces bases). Les polygones de la base de données ont subi une érosion d'un pixel pour s'affranchir des incertitudes liées aux frontières des objets. L'évaluation porte donc davantage sur les pixels situés à l'intérieur des objets, ce qui rejoint notre approche « au patch » décrite en Section 1 : plutôt que de se concentrer sur les frontières des objets,

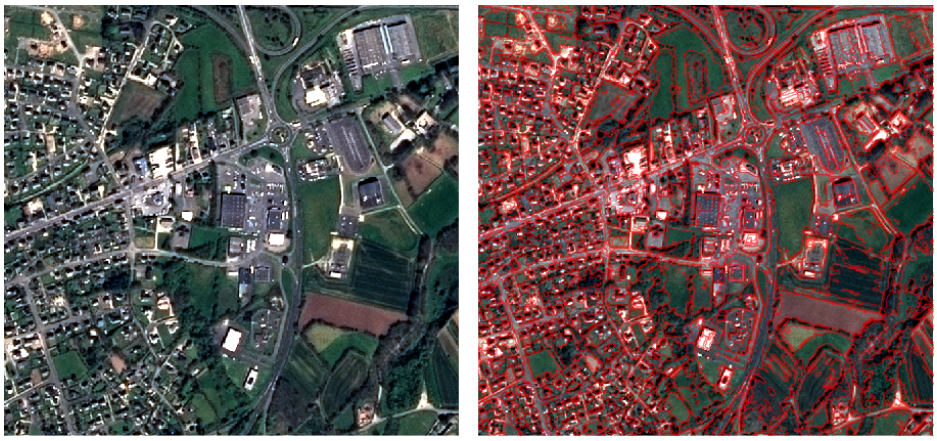
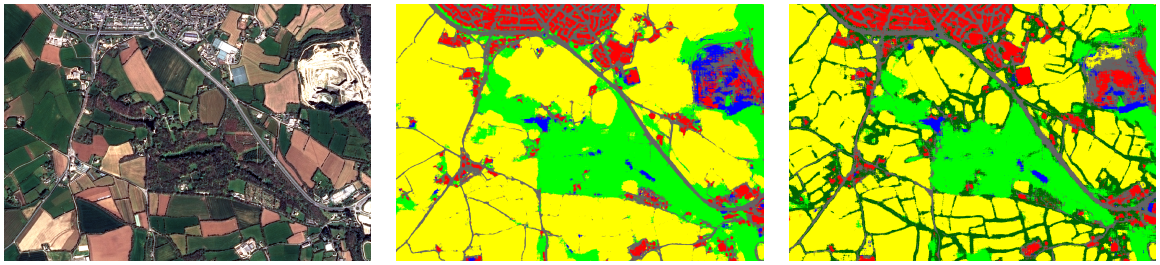


FIGURE 6: A gauche : Image SPOT - A droite : segmentation utilisant la méthode (Felzenszwalb et Huttenlocher, 2004) avec $k = 30$, $m = 20$.



(1) Image SPOT

(2) 5 classes : ● *Bâti,* ● *Route,* ● *Culture,* ● *Végétation,* ● *Eau.*

(3) 6 classes : ● *Bâti,* ● *Route,* ● *Culture,* ● *Végétation,* ● *Haie,* ● *Eau.*

FIGURE 7: Passage d'une occupation des sols à 5 classes vers 6 classes, en affinant un réseau pré-entraîné. On présente le résultat sur une zone de 3,5 km$\times$1,2 km, extraite de ROI-1 (Figure 5).

on cherche à quantifier le nombre d'objets bien retrouvés dans l'image. Les matrices de confusion calculées par comparaison pixel à pixel permettent de dériver un certains nombres d'indicateurs. Deux d'entre eux sont spécifiques à chaque classe : le premier est l'« Average Accuracy » (AA), représentant la part de pixels correctement classifiés, le second est le F1-score, mettant en jeu les mesures de précision (exactitude) et rappel (exhaustivité), toutes deux souvent en opposition. En plus de ces indicateurs "par classe", on utilise l'« Overall Accuracy » (OA) et l'indice Kappa qui sont des mesures de qualité globale, le premier étant le taux global de pixels bien classés, le second comparant la classification avec une attribution aléatoire des étiquettes.

La classe *eau* est souvent dure à évaluer car (i) très diverse dans ses objets (étangs, mares dans les carrières, océan, estuaires), (ii) les ombres sont souvent classées en eau, diminuant la précision sur cette classe.

Qualifier le résultat sur ROI-3 nous permet d'évaluer la capacité de généralisation du réseau en nous éloignant de la zone d'apprentissage.

5.2. Entraînement par RWI et application directe du réseau

Sur GPU, les réseaux à 3 et 4 couches ont requis respectivement 52,5 et 61 heures d'entraînement. Comme mentionné auparavant, les réseaux ont convergé avant la complétion des 5 000 itérations. 300 itérations ont suffi dans les deux cas. La comparaison entre les deux réseaux est visible sur la Figure 9. On note les meilleures performances du réseau 4 couches qui conduisent à ne considérer que celui-ci dans la suite.

Ici, les trois premières lignes du Tableau 3 nous intéressent. La première ligne retranscrit l'application du réseau $M1_{2014}$ sur ROI-1(2014). On peut constater que le réseau routier et le bâti ressortent très bien. Ce résultat est déjà satisfaisant, les méthodes telles que les Forêts Aléatoires peinant à détecter les routes sans l'utilisation d'une information d'altitude (Gressin et al., 2013). La classe de *culture* obtient un très bon F-score du fait du nombre très important à l'intérieur des polygones de cultures étant correctement détectées. Cela masque ainsi les problème aux frontières. En analysant la seconde ligne, qui correspond à l'application du modèle rural sur la zone urbain dense, on peut constater que toutes les classes subissent une détérioration. Les classes de *bâti*

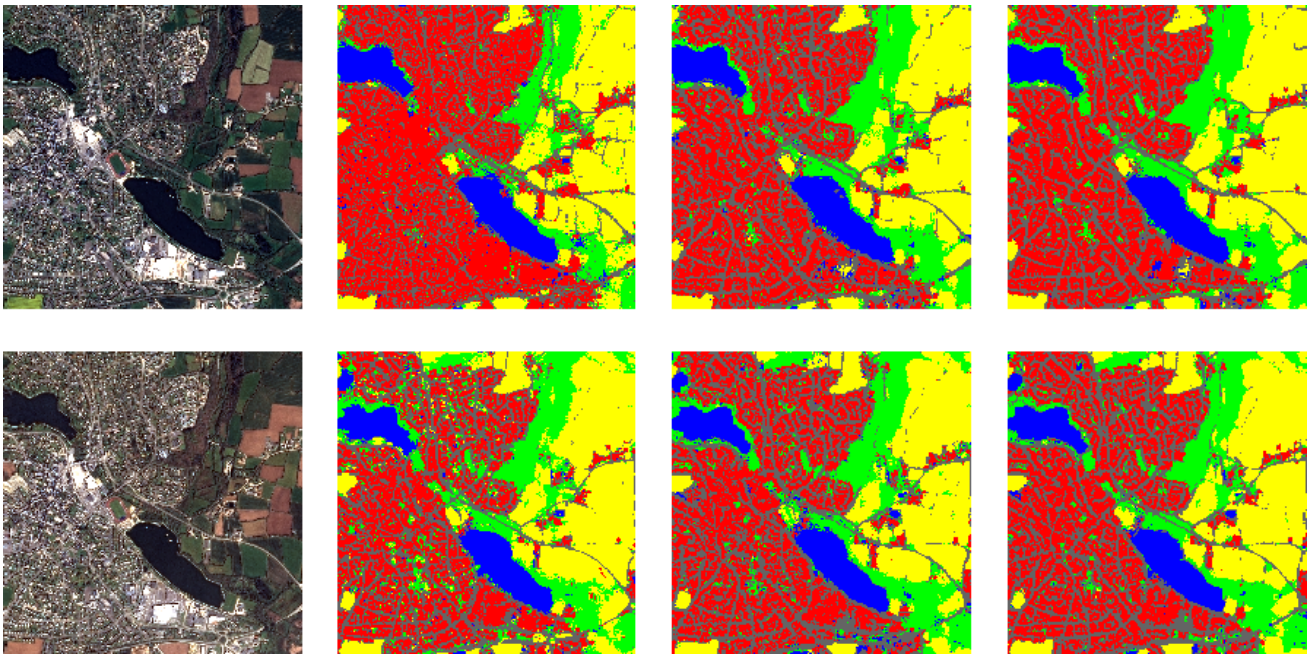


FIGURE 8: Impact des différentes stratégies d'entraînement sur une partie de ROI-1.
1ère ligne - de gauche à droite : 2014 image, $M2_{2014}$, $M2^{2000}_{1-2014}$, $M2^{8000}_{1-2014}$.
2ème ligne - de gauche à droite : 2016 image, $M1_{2014}$, $M2^{2000}_{1-2016}$, $M2^{8000}_{1-2014}$.
● *Bâti,* ● *Route,* ● *Culture,* ● *Végétation,* ● *Eau.*

et de *route* sont, comme prévu, les plus impactées car mal caractérisées par le réseau "rural". Enfin, les derniers résultats proviennent du passage du modèle "urbain" vers la même scène urbaine, mais deux ans plus tard. Les scores sont donc très similaires, indiquant une très bonne robustesse du réseau à travers le temps. Les résultats peuvent être appréciés visuellement sur la Figure 8.
Malgré une architecture légère, une application naïve du réseau permet de détecter des classes habituellement difficiles à trouver. Le réseau routier est spécifiquement très compliqué à récupérer sans MNS. Or, nous arrivons ici à le détecter sans cette donnée d'altitude, par ailleurs quasi systématiquement intégrée dans d'autres méthodes de classification, afin de limiter les confusions avec le bâti.

5.3. Fine-tuning de réseau initialisé par RWI

Pour cette phase de tests, nous avons concentré nos efforts sur la région ROI-1 étant donné qu'elle affichait les résultats les moins bons par application directe de réseau pré-entraîné. L'étude comporte aussi l'impact du nombre d'échantillons utilisés pour effectuer ce fine-tuning. En général, même si l'entraînement était programmé pour 300 itérations, environ 100 étaient suffisantes. Dans cette partie, le fine-tuning agit toujours sur le réseau pré-entraîné sur ROI-2(2014), et est appliqué sur ROI-1, aux deux dates.
1. Pour lever les ambiguïtés en région urbaine, $M2_{2014}$ a été affiné en utilisant quelques échantillons d'entraînement de ROI-1(2014). Ces tests correspondent aux quatre lignes du milieu du Tableau 3. Des améliorations par rapport à l'application directe sont à noter, notamment pour les deux classes qui avaient été le plus affectées (*bâti* et *route*). Un résultat attendu est également l'augmentation des précisions avec le nombre d'échantillons. On améliore de 23% et 14% les classes de *route* et *bâti* en utilisant 8 000 patchs d'apprentissage de ROI-1. Cette amélioration montre le caractère polyvalent et extrêmement adaptatif des CNNs. On peut constater également visuellement l'amélioration sémantique grâce au fine-tuning (Figure 8).
2. Les trois dernières rangées du tableau concernent le fine-tuning d'un réseau que l'on souhaite appliquer sur une région géographiquement éloignée, et acquise à deux ans d'intervalle. En l'occurrence, $M2_{2014}$ est appliqué sur ROI-1(2016). Les métriques nous donnent encore satisfaction.
L'utilisation de 8 000 échantillons apporte la plus grande amélioration, mais 2 000 échantillons suffisent à gagner 10% au moins sur chaque classe. Il est également intéressant de noter que les durées d'entraînement par ce biais sont biens moindres qu'en Section 5.2. Elles varient selon le nombre d'échantillons, mais environ 3h suffisent pour obtenir un modèle adéquat.

5.4. Performance du fine-tuning en ajoutant la classe *Haie*

Afin d'enrichir la nomenclature par l'ajout d'une nouvelle classe de haies, le réseau quatre couches entraîné pour détecter cinq classes sur le Finistère (Section 5.2) a été affiné, et non pas ré-entraîné « from scratch », afin de limiter les durées d'apprentissage. En ce qui concerne ceux-ci, ils sont similaires à ceux indiqués en Section 5.3

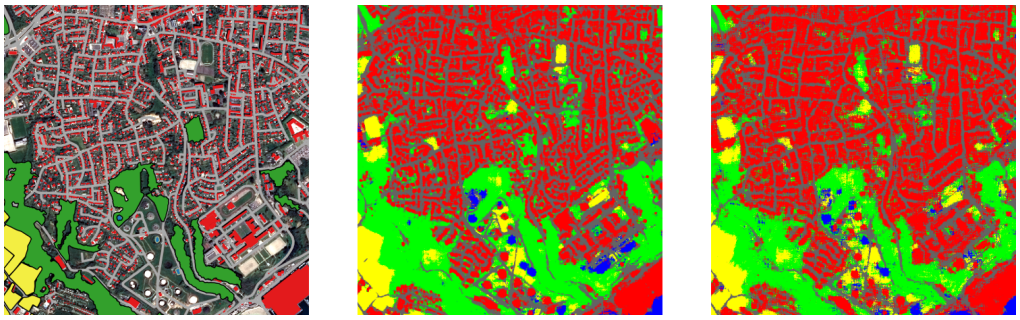


FIGURE 9: Comparaison entre un réseau 3 couches (milieu) et 4 couches (droite) : la donnée de référence est à gauche. ● *Bâti*, ● *Route*, ● *Culture*, ● *Végétation*, ● *Eau*.

(environ 3h pour atteindre une solution pertinente).
Qualitativement, la classification peut être appréciée sur la Figure 7, qui représente une partie de la zone totale classifiée. Une inspection visuelle permet de confirmer l'efficacité du processus de fine-tuning dans le cas d'ajout de classe supplémentaire étant donné que les haies séparant les parcelles agricoles ainsi que celles longeant le réseau routier sont bien détectées, tout en préservant les classes initiales.
Une validation quantitative a également été effectuée, dont les résultats sont fournis dans le tableau 1. La première ligne correspond aux performances, calculées sous forme de F-Score pour chaque classe, lors de la classification à cinq classes tandis que la seconde indique les F-Scores après ajout de la classe *haie*. L'objectif qui était de séparer les classes *végétation* et *route* est en partie résolue. En effet, la première obtient un meilleur résultat qu'en ne considérant pas les haies à part entière. Toutefois, on peut constater une décroissance sur les autres classes en terme de qualité, malgré un visuel très satisfaisant. En particulier, la classe *culture* subit la décroissance la plus importante : cela provient des données de référence pour la classe *haie* utilisées pour effectuer notre validation croisée. La donnée est fortement impactée par des erreurs de mise à jour, avec des objets *haies* réellement existants, mais absents de cette vérité terrain. Ces haies, à juste titre classifiées comme telles par le réseau, sont donc considérés à tort comme des parties de parcelles agricoles par la donnée de référence (la seule à notre disposition pour mener cette étude). Par ailleurs, les confusions *route* / *végétation* sont transposées vers des confusions entre *route* et *haie*. On note beaucoup de haies qui longent les routes, avec des frontières floues entre les deux objets. En ajoutant le fait que les routes sont des objets fins, une erreur de classification à leurs frontières impacte fortement les performances.
En disposant de données de référence à jour pour la classe *haie*, il est certain que les scores augmenteraient significativement pour les classes de *culture* et de *végétation*. En outre, le F-score de la classe *route* n'a que peu diminué pour un objectif cartographique, tout en offrant une granularité sémantique plus riche.

Bâti	*Route*	*Végétation*	*Culture*	*Eau*	*Haie*
73,30	**80,85**	83,66	**95,84**	67,76	x
72,15	79,94	**84,28**	92,36	**68,98**	43,54

TABLE 1: Évolution des performances en ajoutant la classe *Haie*. La métrique correspond au F-Score par classe.

5.5. Réduction des temps de calcul : approche superpixels

Afin d'apprécier la pertinence de l'utilisation de superpixels pour réduire les temps de traitement à la phase de prédiction, nous avons comparé une classification effectuée selon le processus évoqué en Section 3.4, où chaque pixel est classifié par le réseau de neurones convolutifs, au procédé impliquant les superpixels décrit en Section 4.4. Visuellement, le résultat est très satisfaisant comme en atteste la Figure 10. La scène choisie montre que l'approche reste valide que l'on soit en milieu urbain dense, semi-urbain ou encore rural. Les résultats de validation croisée, visible sur le tableau 2 indiquent une nette amélioration de la classification en présegmentant l'image (ligne 1 du tableau), en comparaison de l'approche utilisée jusqu'ici (ligne 2). Ceci se fait en ne classifiant que 20% des pixels au sein de chaque superpixel. L'amélioration globale sur toutes les classes provient de la nature même de l'approche superpixels qui limite fortement le bruit, au contraire présent sur la classification « pixels purs ». La classe *eau* obtient de faible performance car une forte sur-détection survient du fait de l'inondation de friches après des précipitations qui ont précédé l'acquisition de cette image.
Par ailleurs, un peu moins de 5 minutes suffisent à établir une classification, de qualité supérieure. Cette durée, qui inclut le calcul de la segmentation, est bien plus satisfaisante face aux 30 minutes environ nécessaire pour l'approche précédente. Dans le but de compléter cette étude, nous avons mené ce traitement en utilisant jusqu'à 80% des pixels au sein de chaque superpixel. Toutefois, cela n'améliore les performances au mieux que de 0,6%, pour un temps de calcul considérablement accru (augmentation linéaire avec le nombre de pixels classifiés).

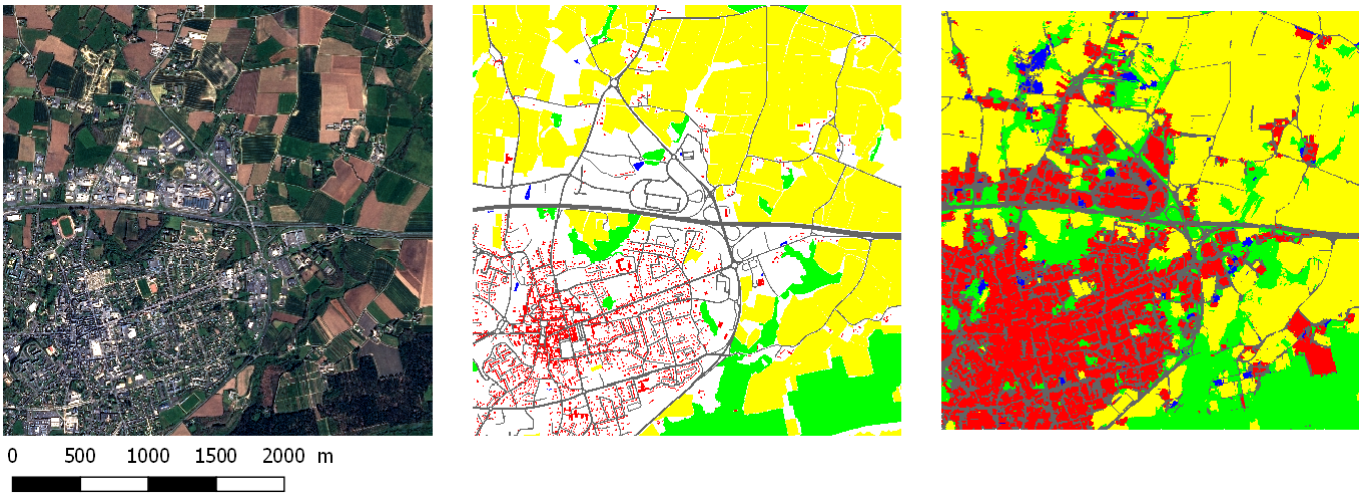


FIGURE 10: - De gauche à droite : Image SPOT, donnée de référence, classification sur les superpixels : ● *Bâti,* ● *Route,* ● *Culture,* ● *Végétation,* ● *Eau*.

	Kappa	F-score					Temps (min.)
		Bâti	*Route*	*Végétation*	*Culture*	*Eau*	
Superpixels (20%)	**75,76**	**68,19**	**76,77**	**81,90**	**92,27**	**14,44**	**4,5**
« Pixels purs »	73,82	66,61	75,29	79,79	91,64	10,99	28

TABLE 2: Comparaison des performances entre une approche « pixels purs » et l'utilisation de superpixels.

5.6. Vers une classification à large échelle

La carte d'occupation des sols que nous obtenons en appliquant $M1_{2014}$ sur ROI-3 est en Figure 11. La classification de cette zone a pris 3,5 jours, pour 780 millions pixels. La zone qui a servi à l'apprentissage est détourée en noire, et correspond à environ 8% de la zone totale. Il est intéressant de regarder la Figure 12 qui montre l'indice Kappa calculé sur chaque tuile 2 000×2 000. Sur cette figure, la zone détourée en noire correspond à la zone d'apprentissage. On constate que le résultat est très satisfaisant globalement (l'indice kappa étant global), même en s'éloignant de la zone d'apprentissage. Les pixels noirs sur la figure indiquant le kappa correspondent à des zones où il n'y pas de données de référence (zones exclusivement maritimes si on regarde la Figure 11). Les estuaires sont les éléments que l'on souhaiterait être classés en *eau* mais aucun estuaire n'étant présent sur la zone d'apprentissage, il est normal que des confusions surviennent sur ces régions.

6. Conclusion

Une architecture légère d'apprentissage profond par Convolutional Neural Networks par patchs a été mise en place afin de pouvoir tester efficacement différent schémas d'entraînement, liés aux possibilités réelles d'avoir ou non des données de référence en plus ou moins grandes quantités. Une approche par patch a été choisie pour la classification large échelle de l'occupation des sols d'une image à très haure résolution spatiale SPOT 6/7. Elle s'avère rapide à mettre en place et se montre efficace et polyvalente comme l'attestent les différentes expérimentations. Ces travaux montrent également la pertinence de créer sans effort des jeux d'apprentissage à partir de bases de données géographiques existantes et libres d'accès. On peut ainsi s'en servir pour entraîner des algorithmes d'apprentissage profond, pour des résultats en milieu urbain inégalé par des méthodes différentes des CNNs. Le schéma d'apprentissage RWI offre déjà une bonne délimitation des cinq classes d'intérêt, mais aussi une description urbaine très satisfaisante. On obtient notamment de bons résultats sur les classes de *bâti* et de *route* lors de la classification de scènes urbaines en utilisant un modèle appris sur de telles scènes. Cela est d'autant plus intéressant qu'aucun *a priori* n'a été conjecturé sur ces classes. De plus, nous n'utilisons pas de données autres que la radiométrie. Une baisse des performances apparaît toutefois quand on souhaite classifier une zone différente d'un point de vue thématique.

Si le RWI peine lorsque l'on change de contexte thématique, le fine-tuning pallie très bien ce problème, les classes de *bâti* et de *route* bénéficiant grandement de ce processus. Accroître le nombre d'échantillons améliore les résultats même si une limite semble être atteinte autour de 8 000 échantillons par classe. La sémantisation à grande échelle vient alors à moindres coûts pour des résultats très satisfaisants.

Nous avons également pu appliquer le processus de fin-tuning dans le cas où l'on souhaiterait accroître la granularité sémantique de la description de l'occupation des sols, et ce, en ré-utilisant un réseau entraîné à l'origine

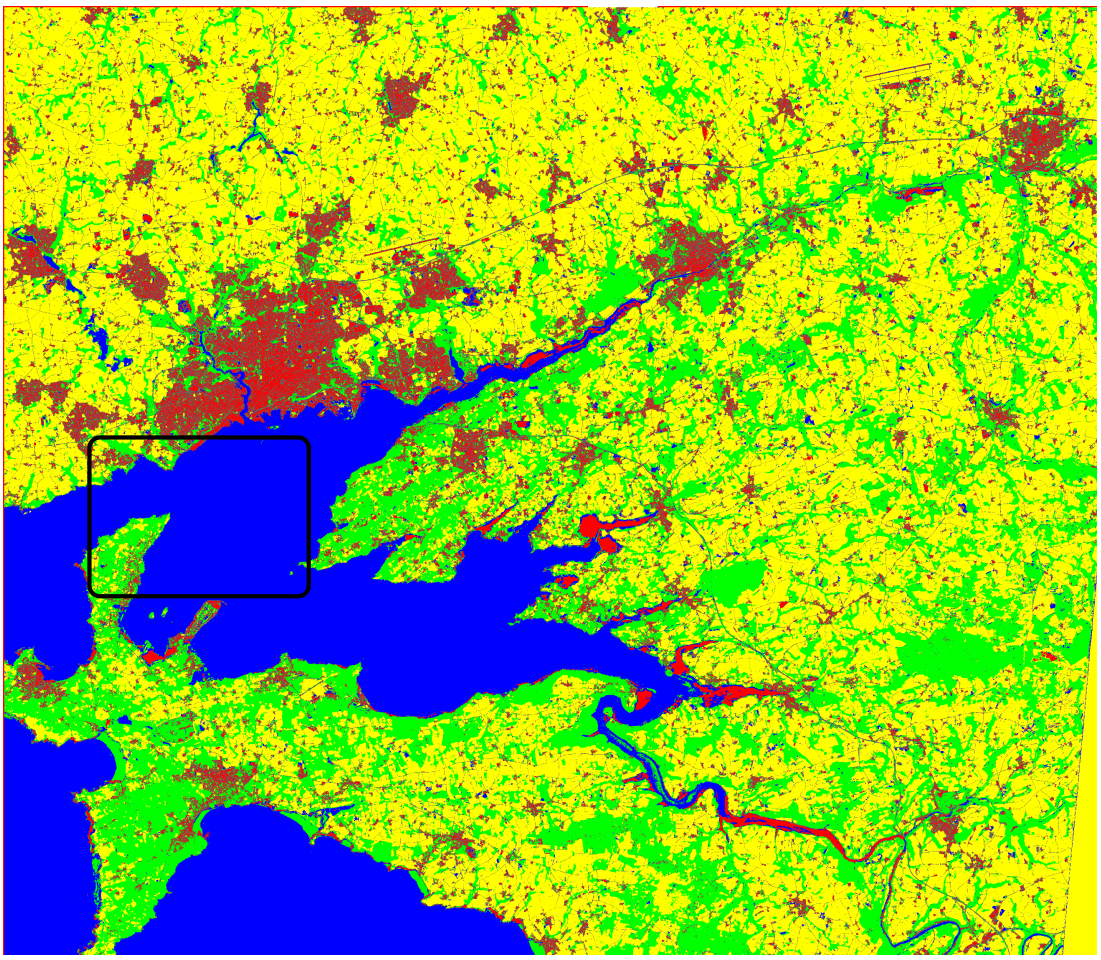


FIGURE 11: Occupation des sols sur ROI-3 avec le réseau 4 couches entraîné sur ROI-1(2014). La superficie de ROI-3 est de 39×45 km. La région en noire est la région d'apprentissage. ● *Bâti,* ● *Route,* ● *Culture,* ● *Végétation,* ● *Eau.*

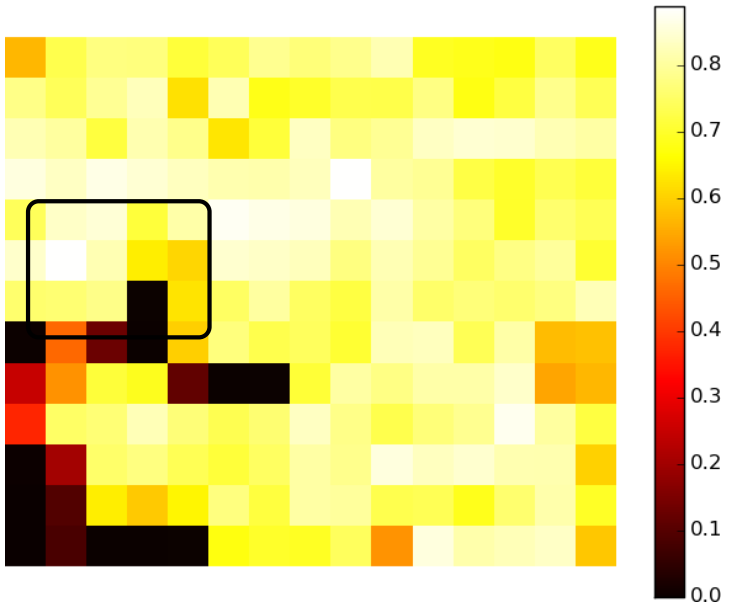


FIGURE 12: Indice Kappa sur ROI-3 : la qualité ne décroît pas en s'éloignant de la zone d'apprentissage (en noire). Les pixels noirs correspondent aux régions sans données de référence pour la validation croisée.

Expérimentation	K	OA	AA	F_{moy}	F_{route}	$F_{végétation}$	$F_{bâti}$	$F_{culture}$	F_{eau}
$M1_{2014} \rightarrow$ ROI-1 (2014)	0,84	0,92	0,88	0,85	0,75	0,85	0,81	0,97	0,86
$M2_{2014} \rightarrow$ ROI-1 (2014)	0,71	0,84	0,80	0,73	0,54	0,73	0,68	0,94	0,77
$M1_{2014} \rightarrow$ ROI-1 (2016)	0,83	0,91	0,86	0,83	0,75	0,82	0,79	0,97	0,84
$M2^{2000}_{1\text{-}2014}$	0,81	0,90	0,86	0,82	0,71	0,84	0,76	0,96	0,82
$M2^{5000}_{1\text{-}2014}$	0,83	0,91	0,88	0,84	0,74	0,85	0,80	0,96	0,86
$M2^{8000}_{1\text{-}2014}$	0,85	0,92	0,89	0,86	**0,77**	0,87	**0,82**	0,97	0,86
$M2^{10000}_{1\text{-}2014}$	0,86	0,92	0,90	0,86	0,78	0,87	0,82	0,97	0,84
$M2^{2000}_{1\text{-}2016}$	0,80	0,89	0,86	0,81	0,71	0,83	0,76	0,96	0,78
$M2^{8000}_{1\text{-}2016}$	0,84	0,92	0,89	0,85	**0,76**	0,86	**0,81**	0,97	0,85
$M2^{10000}_{1\text{-}2016}$	0,84	0,92	0,89	0,85	0,77	0,86	0,81	0,97	0,84

TABLE 3: Évaluation des stratégies d'entraînement envisagées.

pour un problème comportant moins de classes.
Enfin, malgré des temps de calcul qui peuvent paraître très longs en phase de test, une pré-segmentation de la scène permet de produire des cartes d'occupation des sols dans des temps cette fois-ci raisonnables.

Remerciements

Ce travail a bénéficié de fonds publics reçus dans le cadre de GEOSUD, un projet (ANR-10-EQPX-20) du programme "Investissements d'Avenir" de l'Agence Nationale pour la Recherche (ANR). Il a été conduit sous l'égide du projet TOSCA "Occupation des SOls" du CNES et du Centre d'Expertise Scientifique du même nom du pôle de données THEIA.

Références

Achanta, R., Shaji, A., Smith, K., Lucchi, A., Fua, P., Susstrunk, S., 2012. Slic superpixels compared to state-of-the-art superpixel methods. IEEE Transactions on Pattern Analysis and Machine Intelligence 34 (11), 2274–2282.

Audebert, N., Le Saux, B., Lefèvre, S., 2016. Semantic segmentation of earth observation data using multimodal and multi-scale deep networks. Dans : Asian Conferenceon Computer Vision, 20-24 November, Taipei, Taiwan.

Audebert, N., Saux, B. L., Lefèvre, S., 2018. Beyond rgb : Very high resolution urban remote sensing with multimodal deep networks. ISPRS Journal of Photogrammetry and Remote Sensing 140, 20 – 32.

Badrinarayanan, V., Kendall, A., Cipolla, R., 2015. Segnet : A deep convolutional encoder-decoder architecture for image segmentation. arXiv :1511.00561.

Belgiu, M., Draguţ, L., 2016. Random forest in remote sensing : A review of applications and future directions. ISPRS Journal of Photogrammetry and Remote Sensing 114, 24–31.

Chen, K., Fu, K., Yan, M., Gao, X., Sun, X., Wei, X., Feb 2018a. Semantic segmentation of aerial images with shuffling convolutional neural networks. IEEE Geoscience and Remote Sensing Letters 15 (2), 173–177.

Chen, K., Weinmann, M., Gao, X., Yan, M., Hinz, S., Jutzi, B., Weinmann, M., 2018b. Residual shuffling convolutional neural networks for deep semantic image segmentation using multi-modal data. ISPRS Annals of Photogrammetry, Remote Sensing and Spatial Information Sciences IV-2, 65–72.

Chen, L.-C., Papandreou, G., Kokkinos, I., Murphy, K., Yuille, A. L., 2014. Semantic image segmentation with deep convolutional nets and fully connected CRFs. arXiv :1412.7062.

Chen, L.-C., Papandreou, G., Kokkinos, I., Murphy, K., Yuille, A. L., 2018c. Deeplab : Semantic image segmentation with deep convolutional nets, atrous convolution, and fully connected crfs. IEEE transactions on pattern analysis and machine intelligence 40 (4), 834–848.

Dechesne, C., Mallet, C., Bris, A. L., Gouet-Brunet, V., 2017. Semantic segmentation of forest stands of pure species combining airborne lidar data and very high resolution multispectral imagery. ISPRS Journal of Photogrammetry and Remote Sensing 126, 129 – 145.

Fauvel, M., 2007. Spectral and spatial methods for the classification of urban remote sensing data. Thèse de doctorat, Institut National Polytechnique de Grenoble - INPG, France ; University of Iceland, Iceland.

Felzenszwalb, P. F., Huttenlocher, D. P., 2004. Efficient graph-based image segmentation. International Journal of Computer Vision 59 (2), 167–181.

Girshick, R., Donahue, J., Darrell, T., Malik, J., 2014. Rich feature hierarchies for accurate object detection and semantic segmentation. Dans : IEEE Conference on Computer Vision and Pattern Recognition, IEEE/CVF, 24-27 June, Colombus, USA.

Gressin, A., Mallet, C., Vincent, N., Paparoditis, N., 2013. Updating land cover databases using a single very high resolution satellite image. ISPRS Annals of Photogrammetry, Remote Sensing and Spatial Information Sciences II-3/W2, 13–18.

Gressin, A., Vincent, N., Mallet, C., Paparoditis, N., 2014. A unified framework for land-cover database update and enrichment using satellite imagery. Dans : IEEE International Conference on Image Processing, IEEE, 27-30 October, Paris, France. pp. 5057–5061.

He, K., Zhang, X., Ren, S., Sun, J., 2016. Deep residual learning for image recognition. Dans : IEEE Conference on Computer Vision and Pattern Recognition, IEEE/CVF, 26 June-1 July, Las Vegas, USA.

Inglada, J., Arias, M., Tardy, B., Hagolle, O., Valero, S., Morin, D., Dedieu, G., Sepulcre, G., Bontemps, S., Defourny, P., Koetz, B., 2015. Assessment of an operational system for crop type map production using high temporal and spatial resolution satellite optical imagery. Remote Sensing 7 (9), 12356–12379.

Kaiser, P., Wegner, J., Lucchi, A., Jaggi, M., Hofmann, T., Schindler, K., 2017. Learning aerial image segmentation from online maps. IEEE Transactions on Geoscience and Remote Sensing.

Kemker, R., Salvaggio, C., Kanan, C., 2018. Algorithms for semantic segmentation of multispectral remote sensing imagery using deep learning. ISPRS Journal of Photogrammetry and Remote Sensing.

Khatami, R., Mountrakis, G., Stehman, S., 2016. A meta-analysis of remote sensing research on supervised pixel-

based land-cover image classification processes : General guidelines for practitioners and future research. Remote Sensing of Environment 177, 89–100.

Krizhevsky, A., Sutskever, I., Hinton, G. E., 2012. Imagenet classification with deep convolutional neural networks. Dans : Neural Information Processing Systems, 3-8 December, Lake Tahoe, USA. pp. 1097–1105.

Long, J., Shelhamer, E., Darrell, T., 2015. Fully convolutional networks for semantic segmentation. Dans : IEEE Conference on Computer Vision and Pattern Recognition, IEEE/CVF, 7-12 June, Boston, USA.

Maas, A., Rottensteiner, F., Heipke, C., 2016. Using label noise robust logistic regression for automated updating of topographic geospatial databases. ISPRS Annals of Photogrammetry, Remote Sensing and Spatial Information Sciences III-7, 133–140.

Marmanis, D., Schindler, K., Wegner, J., Galliani, S., Datcu, M., Stilla, U., 2016. Classification with an edge : Improving semantic image segmentation with boundary detection. arXiv :1612.01337.

Marmanis, D., Schindler, K., Wegner, J., Galliani, S., Datcu, M., Stilla, U., 2018. Classification with an edge : Improving semantic image segmentation with boundary detection. ISPRS Journal of Photogrammetry and Remote Sensing 135, 158 – 172.

Ng, A., Jordan, M., 2002. On discriminative vs. generative classifiers : A comparison of logistic regression and naive Bayes. Dans : Neural Information Processing Systems, 9-14 December, Vancouver, Canada. pp. 841–848.

Papadomanolaki, M., Vakalopoulou, M., Zagoruyko, S., Karantzalos, K., 2016. Benchmarking deep learning frameworks for the classification of very high resolution satellite multispectral data. ISPRS Annals of Photogrammetry, Remote Sensing and Spatial Information Sciences III-7, 83–88.

Pelletier, C., Valero, S., Inglada, J., Champion, N., Dedieu, G., 2016. Assessing the robustness of random forests to map land cover with high resolution satellite image time series over large areas. Remote Sensing of Environment 187, 156–168.

Redmon, J., Farhadi, A., 2017. Yolo9000 : Better, faster, stronger. Dans : IEEE Conference on Computer Vision and Pattern Recognition, IEEE/CVF, 21-26 June, Honolulu, USA.

Sherrah, J., 2016. Fully convolutional networks for dense semantic labelling of high-resolution aerial imagery. arxiv :1606.02585.

Shi, W., Caballero, J., Huszár, F., Totz, J., Aitken, A. P., Bishop, R., Rueckert, D., Wang, Z., 2016. Real-time single image and video super-resolution using an efficient sub-pixel convolutional neural network. Dans : Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 1874–1883.

Simonyan, K., Zisserman, A., 2014. Very deep convolutional networks for large-scale image recognition. arXiv :1409.1556.

Stutz, D., Hermans, A., Leibe, B., 2018. Superpixels : An evaluation of the state-of-the-art. Computer Vision and Image Understanding 166, 1 – 27.

Szegedy, C., Liu, W., Jia, Y., Sermanet, P., Reed, S., Anguelov, D., Erhan, D., Vanhoucke, V., Rabinovich, A., 2015. Going deeper with convolutions. Dans : IEEE Conference on Computer Vision and Pattern Recognition, IEEE/CVF, 7-12 June, Boston, USA.

Tokarczyk, P., Wegner, J. D., Walk, S., Schindler, K., Jan 2015. Features, color spaces, and boosting : New insights on semantic classification of remote sensing images. IEEE Transactions on Geoscience and Remote Sensing 53 (1), 280–295.

Tu, W.-C., Liu, M.-Y., Jampani, V., Sun, D., Chien, S.-Y., Yang, M.-H., Kautz, J., 2018. Learning superpixels with segmentation-aware affinity loss. Dans : IEEE Conference on Computer Vision and Pattern Recognition, 18-22 JUne, Salt Lake City, USA.

Tupin, F., Inglada, J., Nicolas, J.-M. (Eds.), 2014. Remote Sensing Imagery. ISTE - Wiley.

Volpi, M., Tuia, D., 2017. Dense semantic labeling of subdecimeter resolution images with convolutional neural networks. IEEE Transactions on Geoscience and Remote Sensing 55 (2), 881–893.

Zeiler, M. D., Fergus, R., 2014. Visualizing and understanding convolutional networks. Dans : European Conference on Computer Vision, 6-12 September, Zurich, Switzerland. pp. 818–833.

Zhu, X. X., Tuia, D., Mou, L., Xia, G., Zhang, L., Xu, F., Fraundorfer, F., 2017. Deep learning in remote sensing : A comprehensive review and list of resources. IEEE Geoscience and Remote Sensing Magazine 5 (4), 8–36.