Réseaux de neurones semi-supervisés pour la segmentation sémantique en télédétection

Javiera CASTILLO NAVARRO[1,2], Bertrand LE SAUX[1], Alexandre BOULCH[1], Sébastien LEFÈVRE[2]

[1]DTIS, ONERA, Université Paris-Saclay, FR-91123 Palaiseau, France

[2]Université Bretagne Sud, UMR 6074, IRISA, FR-56000 Vannes, France

{javiera.castillo_navarro, bertrand.le_saux, alexandre.boulch}@onera.fr, sebastien.lefevre@irisa.fr

Résumé – Cet article questionne différents aspects de la segmentation sémantique des données de télédétection par apprentissage profond. Les bonnes performances des réseaux de neurones actuels reposent sur la disponibilité de larges bases de données entièrement annotées. Or dans le domaine de la télédétection, bien que les images soient abondantes, les annotations sont rares ou difficiles à réaliser. Dans ce contexte, deux questions se posent : premièrement, quelle est la capacité de généralisation des réseaux supervisés actuels ? Et deuxièmement, est-il possible d'améliorer les performances des méthodes actuelles en utilisant des données non annotées ? Nos principales contributions sont : (i) une analyse approfondie de la robustesse des réseaux supervisés par rapport aux données d'observation de la Terre, et (ii) la présentation d'une architecture semi-supervisée, capable d'exploiter simultanément des images annotées et non-annotées.

Abstract – This work explores different aspects of semantic segmentation of remote sensing data using deep neural networks.The availability of large databases of fully annotated data is the basis for good performances of current neural networks. Although images on remote sensing are abundant, annotations are very rare or difficult to produce. In this context, two questions arise: first, how robust are existing supervised learning strategies with respect to data? Second, is it possible to improve performance of current methods by using non annotated data? Our main contributions are: (i) a strong robustness analysis of existing supervised learning strategies with respect to remote sensing data, (ii) the introduction of a semi-supervised architecture, capable of learning from annotated and non annotated images simultaneously.

1 Introduction et contexte

Le développement de techniques automatisées pour l'analyse des données d'observation de la Terre suscite beaucoup d'intérêt, du fait de son potentiel pour les domaines de l'écologie et la planification urbaine, par exemple. Aujourd'hui les méthodes par apprentissage profond constituent l'état de l'art en détection d'objet et segmentation d'images aériennes et satellites [2, 9] et permettent de réaliser rapidement et sans intervention humaine des cartographies sémantiques précises.

Néanmoins, ces méthodes sont supervisées et reposent largement sur la disponibilité de grandes banques d'images annotées pour atteindre des résultats satisfaisants. Par ailleurs, il existe à présent une grande quantité de données disponibles en libre accès, mais dont les annotations sont inexistantes ou alors très restreintes, à la fois en termes de sémantique comme en termes de couverture spatiale. Ainsi, nous nous intéressons à deux problèmes : quelles sont les caractéristiques souhaitées d'une base de données pour qu'elle soit de bonne qualité pour l'entraînement d'algorithmes de classification en télédétection ? Et par ailleurs, est-il possible d'exploiter les images non-annotées dans le but d'améliorer les performances de réseaux de neurones existants pour la segmentation sémantique ?

Les réseaux de neurones convolutifs (CNNs) et entièrement convolutifs (FCNs) constituent aujourd'hui l'état de l'art en analyse des données de télédétection : cartographie de l'occupation de sols dans des zones urbaines [6, 8], segmentation d'objets [1] ou encore segmentation sémantique pour des données multi-modales et multi-échelles [2].

Ces réseaux atteignent les meilleurs résultats sur la plupart des bases de données publiques de référence en télédétection [4, 7, 5, 10], avec des performances entre 80% ou 90% de précision globale. Cependant, ces méthodes sont supervisées et requièrent donc une grande quantité d'annotations denses durant l'apprentissage. De telles annotations sont rarement disponibles dans des situations réelles, et devront donc être réalisées au cas par cas. Dans ce contexte nous souhaitons répondre à la question : combien de données, et quelles données sont nécessaires pour entraîner un réseau de neurones supervisé pour la segmentation sémantique en télédétection ?

Dans ce but, dans la section 2, nous étudions le comportement de l'entraînement des réseaux par rapport aux données. Premièrement, nous souhaitons évaluer la quantité de données nécessaires pour obtenir de bonnes performances avec une configuration standard. Deuxièmement, nous renouvelons l'expérience avec un jeu de données à grande échelle afin de comprendre la relation entre l'entraînement et la capacité de généralisation des modèles résultants. Dans ces expériences, nous utilisons un FCN standard, efficace et polyvalent : SegNet [2, 3]. Ce réseau présente une architecture encodeur-décodeur, la par-

tie encodeur est similaire à VGG16 [11] et peut donc être initialisée avec des poids pré-entraînés. L'entraînement est réalisé avec une descente de gradient stochastique et comme fonction de coût l'entropie croisée standard.

Dans la section 3 nous nous posons une autre question : les données sans annotation peuvent-elles améliorer les performances des réseaux, avec une approche semi-supervisée ? Nous présentons une nouvelle architecture, nommée Berunda-Net, pour la segmentation sémantique d'images par apprentissage profond semi-supervisé.

2 Limites de l'apprentissage supervisé

Nous étudions ici le rapport entre les données d'entraînement disponibles et les performances des CNNs en classification. Deux cas de figures sont envisagés : d'abord un jeu de données classique, correspondant à un seul site géographique (quelques quartiers adjacents d'une petite ville allemande), puis une deuxième base de données à grande échelle qui regroupe plusieurs villes françaises.

2.1 Jeu de données mono-site

Le jeu des données *ISPRS Vaihingen 2D Semantic Labeling* [1] comprend 33 tuiles IRRV (infrarouge-rouge-vert) orthorectifiées à une résolution de 9 cm/px, acquises au-dessus de Vaihingen. 16 tuiles sont publiées avec des vérités-terrain (VT) denses, selon six classes d'intérêt : routes, bâtiments, végétation basse, arbres et véhicules, ainsi qu'une classe fourre-tout. Nous partitionnons ce jeu en 12 tuiles pour l'entraînement et 4 tuiles pour la validation.

La première expérience vise à analyser la sensibilité de l'apprentissage supervisé à la quantité de données lors de l'entraînement. Nous faisons varier la quantité d'images annotées de 1 à 12 tuiles. Deux configurations sont comparées : une initialisation de la partie encodeur avec les poids pré-entraînés sur ImageNet et une initialisation aléatoire des poids.

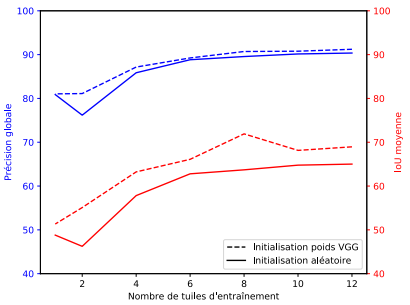

FIGURE 1 – Influence de la quantité d'annotations sur l'entraînement en termes de précision globale de classification et d'intersection-sur-union moyenne.

Les résultats sont présentés dans la figure 1. Les performances avec 12 tuiles reproduisent correctement l'état de l'art (près de 90 % de précision globale). Cependant, il est surprenant de constater que réduire la quantité de données ne dégrade que légèrement ces performances, et qu'un apprentissage avec une seule tuile (soit quelques rues seulement) conduit à 80 % de bonnes classifications. Cela s'explique par la grande similarité des images de ce jeu de référence, à cause de leur proximité géographique, qui ne posent donc pas de problème de généralisation. De plus le transfert de connaissances depuis ImageNet n'améliore que peu les performances.

2.2 Jeu de données large-échelle

Dans cette section nous introduisons une nouvelle base de données, nommée *MiniFrance*, dans le but d'ajouter de la variabilité aux tests de segmentation sémantique. MiniFrance est constituée de 16 conurbations de différentes régions en France : Brest, Nantes, Lille, Clermont-Ferrand, Nice, etc. Nous collectons les données à partir de deux sources en licence libre [2] :

— Les images aériennes de la BD ORTHO de l'IGN (de 2012 à 2014), fournies sous la forme de tuiles RVB de dimensions 10.000 × 10.000px à 50cm/px de résolution.

— Les annotations du projet Copernicus *Urban Atlas 2012* [3], réparties en 14 classes d'utilisation du sol (urbain dense, forêts, champs, etc.). Pour faciliter l'apprentissage, des images de référence à résolution de 50cm/px sont générées à partir des données vectorielles originales.

8 villes sont utilisées pour l'entraînement et 8 autres pour l'évaluation, avec une diversité semblable en termes d'architectures et de design urbain dans les deux ensembles. MiniFrance est constituée de 2121 images (sa taille est donc 2719 fois plus grande que celle de ISPRS Vaihingen en nombre de pixels).

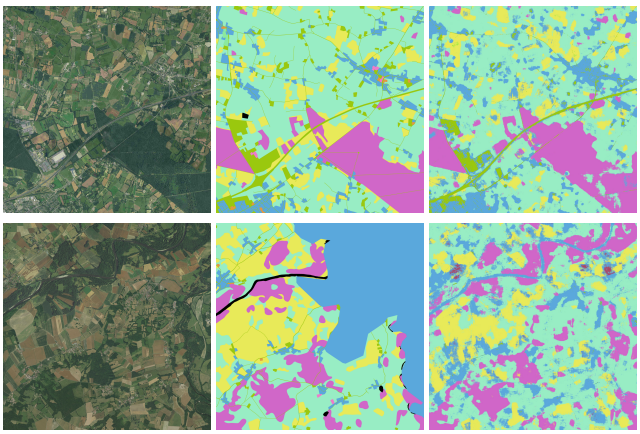

FIGURE 2 – Résultats du modèle supervisé sur deux images de MiniFrance. De gauche à droite, l'image d'entrée, la référence associée et la prédiction du réseau.

La figure 2 compare les cartes d'occupation du sol produites

1. http://www2.isprs.org/commissions/comm2/wg4/vaihingen-2d-semantic-labeling-contest.html

2. Pour la BD ORTHO, nous ne considérons que les régions disponibles en licence libre.

3. https://land.copernicus.eu/local/urban-atlas/urban-atlas-2012/view

par un SegNet avec les données de référence. Les résultats correspondent visuellement aux images, à deux nuances près. La carte semble avoir un grain plus fin que la référence (voir ligne 1, la zone de forêts et de champs en rose). Par ailleurs (ligne 2, zone bleue annotée comme ville), la référence peut ne pas être fiable et la prédiction obtenue par notre méthode semble mieux correspondre à la réalité.

Comme en Sec. 2.1, nous étudions l'influence de la quantité de données disponibles durant l'entraînement sur la classification. Cependant, en raison des temps de calcul (avec un GPU Titan X, l'entraînement sur MiniFrance prend 40h et l'évaluation 25h), nous menons des expériences plus ciblées : entraînement sur les 8 villes, 10% équirépartis, et une seule ville (Caen, soit 12,5%). Les résultats sont présentés dans le tableau 1.

TABLE 1 – Performances de la classification par rapport à la quantité de données.

Ensemble d'entraînement	*Précision*	*mIoU*
100 %	52,40 %	15,79 %
10 %	50,14 %	15,25 %
Caen ($\sim$ 12,5%)	42,09 %	10,05 %

Les valeurs des performances sont moindres que sur ISPRS Vaihingen (52% de précision globale), car les classes sont sémantiquement plus complexes. À quantité de données égale, un entraînement mono-site dégrade les résultats de 10 points, au contraire d'un échantillonage régulier de l'ensemble d'entraînement. Ainsi, l'entraînement sur une seule ville n'offre pas suffisamment de diversité pour appréhender toutes les images d'évaluation. Cela montre l'importance d'avoir des localisations variées pour l'apprentissage, et le soin à accorder à la généralisation dans l'objectif de concevoir des modèles applicables n'importe où. En complément, nous appliquons le modèle global sur chaque ville d'évaluation séparément (voir tableau 2). Les résultats varient de 20 points de Marseille à Cherbourg (de 46 à 67% de précision), ce qui montre les limites d'une architecture classique même avec un entraînement varié.

3 Apprentissage semi-supervisé

Les expériences menées dans la section 2 ont mis en évidence certaines faiblesses des réseaux de neurones complètement supervisés actuels. Dans cette section, nous proposons une approche semi-supervisée afin de profiter aussi des données disponibles sans annotation pour apprendre la structure et les caractéristiques intrinsèques aux images, et améliorer ainsi la capacité de généralisation des modèles.

3.1 BerundaNet

L'architecture que nous proposons s'inspire à la fois des *Stacked What-Where Autoencoders* [12] et des réseaux entièrement convolutifs pour la segmentation sémantique tel SegNet.

BerundaNet possède ainsi un encodeur et deux décodeurs (basés sur SegNet pour des raisons de compacité du modèle) :

TABLE 2 – Résultats pour chaque ville d'évaluation du modèle entraîné avec 100% des données.

Score	Marseille	Rennes	Angers	Quimper	Vannes	Clerm.-F.	Lille	Cherbourg
OA	46,13	51,56	44,85	50,82	49,51	46,51	61,35	67,54
mIoU	12,77	15,05	13,15	13,93	12,66	11,40	16,93	15,82

le premier décodeur (en haut sur la figure 3) a pour but la reconstruction de l'image d'entrée (à l'instar d'un auto-encodeur), alors que le second décodeur est chargé de la segmentation sémantique. En pratique, le réseau avec les deux décodeurs est entraîné lorsque les données sont annotées, tandis que seule la partie auto-encodeur est apprise sur les images seules.

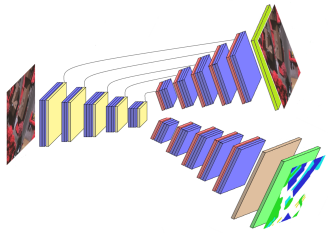

FIGURE 3 – Architecture semi-supervisée BerundaNet.

La fonction de coût doit être adaptée à cette nouvelle architecture. A la manière de [12], elle est composée de trois termes :

$$\ell = \ell_{\text{seg}} + \lambda_{\text{rec}}\ell_{\text{rec}} + \lambda_{\text{m}}\ell_{\text{m}}, \quad (1)$$

où ℓ_{seg} est la pénalité de segmentation, ℓ_{rec} la pénalité de reconstruction, et ℓ_{m} une pénalité qui contrôle la qualité des reconstructions intermédiaires (correspondant aux lignes grises dans la figure 3). λ_{rec} et λ_{m} permettent de pondérer ces différentes pénalités. En pratique, ℓ_{seg} est la fonction d'entropie croisée standard. Pour ℓ_{rec} et ℓ_{m}, termes qui visent à contrôler les différences entre l'image originale ou les couches intermédiaires et leur reconstruction, la norme L_2 est utilisée.

3.2 Expériences et résultats

Afin de valider notre architecture, nous présentons une première série de résultats sur le jeu de données *ISPRS Vaihingen*. Notre objectif est de quantifier l'apport de la semi-supervision avec peu de données annotées. Par conséquent, l'ensemble d'entraînement contient 12 tuiles dont une seule est annotée. L'ensemble de validation est le même que dans la section 2.1. Nous comparons BerundaNet en mode semi-supervisé avec deux modèles complètement supervisés sur la tuile annotée : SegNet et BerundaNet sans données auxiliaires. Les hyperparamètres de la fonction de coût ont été fixés à $\lambda_{\text{rec}} = 1$ et $\lambda_{\text{m}} = 0{,}1$ et nous nous assurons que le nombre d'itérations avec annotations soit le même pour tous les modèles. Les résultats globaux sont présentés dans le tableau 3 et la figure 4.

Si le SegNet adapté à ce cas de figure obtient toujours les meilleures performances, l'architecture BerundaNet permet de mettre en évidence certains aspects de l'entraînement semi-supervisé qui fournit des régions mieux définies (mIoU en augmentation) et ce d'autant plus sur les petits objets (7 points

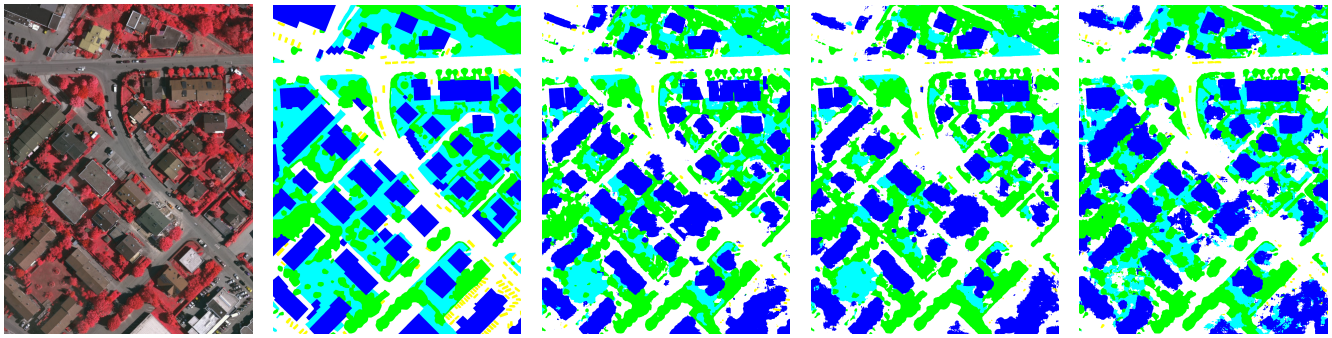

image | référence | SegNet (sup) | BerundaNet (sup) | BerundaNet

FIGURE 4 – Résultats de l'entraînement semi-supervisé ou supervisé avec 1 tuile annotée pour différentes méthodes.

TABLE 3 – Comparaison de résultats par apprentissage supervisé et semi-supervisé.

Modèle	*Précision*	*mIoU*	*F1-voiture*
SegNet (sup)	81,03 %	51,33 %	62,10%
BerundaNet (sup)	79,58 %	46,56 %	42,05%
BerundaNet	78,16%	47,28 %	49,44 %

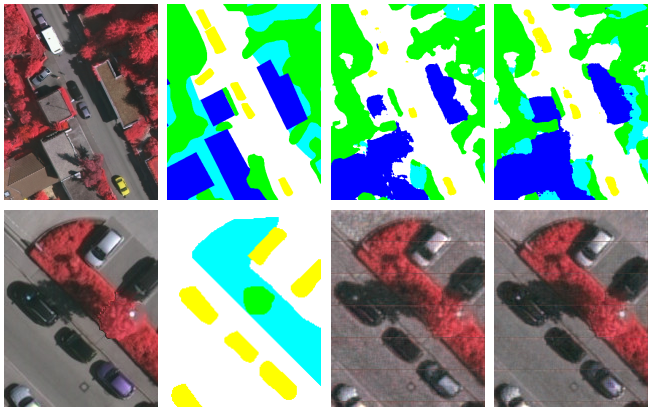

image orig. | référence | sup. | semi-sup.

FIGURE 5 – Exemples de segmentation (en haut) et reconstruction (en bas) obtenus par BerundaNet avec apprentissage supervisé et semi-supervisé.

gagnés pour la classification de voitures). Les zooms de la figure 5 montrent des contours de voitures et de bâtiments mieux délimités. Pour la reconstruction, il est intéressant de noter que le modèle semi-supervisé obtient des textures plus distinctes et encode mieux les couleurs rares (cf. voiture mauve).

4 Conclusion

Nous avons montré que la taille des données annotées nécessaires pour un apprentissage supervisé varie grandement en fonction du besoin de généralisation entre les données d'entraînement et celles de test. Peu de données sont nécessaires pour la classification sémantique sur une seule ville beaucoup plus sont requises pour généraliser à plusieurs villes d'aspects et de plans variés. Nous proposons donc MiniFrance, un nouveau jeu de données très varié et multi-site apte à contribuer au développement d'approches plus génériques pour des applications à large-échelle. Nous avons également proposé BerundaNet, une architecture de FCN multi-tâche pour l'entraînement semi-supervisé avec un auto-encodage qui agit comme une régularisation de l'apprentissage. Nous avons mis en évidence la capacité de ce modèle à bénéficier des images non-annotées, notamment pour les classes ou caractéristiques rares. L'étape suivante consistera à explorer les possibilités de l'apprentissage semi-supervisé dans le contexte ardu de MiniFrance.

Références

[1] N. Audebert, B. Le Saux, and S. Lefèvre. Segment-before-Detect : Vehicle Detection and Classification through Semantic Segmentation of Aerial Images. *Remote Sensing*, 9(4) :1–18, Apr. 2017.

[2] N. Audebert, B. Le Saux, and S. Lefevre. Beyond RGB : Very High Resolution Urban Remote Sensing With Multimodal Deep Networks. *ISPRS J. of Photogrammetry and Remote Sensing*, 140 :20–32, 2018.

[3] V. Badrinarayanan, A. Kendall, and R. Cipolla. SegNet : A Deep Convolutional Encoder-Decoder Architecture for Image Segmentation. *IEEE Trans. on Patt. Anal. and Mach. Intel.*, 39(12) :2481–2495, 2017.

[4] M. Campos-Taberner et al. Processing of extremely high-resolution LiDAR and RGB data : Outcome of the 2015 IEEE GRSS Data Fusion Contest-Part A : 2-D Contest. *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, 9(12) :5547–5559, 2016.

[5] E. Maggiori, Y. Tarabalka, G. Charpiat, and P. Alliez. Can semantic labeling methods generalize to any city ? the INRIA aerial image labeling benchmark. In *Proc. of IGARSS*, 2017.

[6] D. Marmanis et al. Semantic segmentation of aerial images with an ensemble of CNNs. *ISPRS Annals*, 3 :473, 2016.

[7] L. Mou et al. Multitemporal Very High Resolution from Space : Outcome of the 2016 IEEE GRSS Data Fusion Contest. *IEEE J. of Sel. Topics in Applied Earth Obs. and Remote Sensing*, 10(8) :3435–3447, 2017.

[8] S. Paisitkriangkrai et al. Effective semantic pixel labelling with convolutional networks and conditional random fields. In *Proc. of the IEEE CVPR Workshops*, pages 36–43, 2015.

[9] N. Rey, M. Volpi, S. Joost, and D. Tuia. Detecting animals in African Savanna with UAVs and the crowds. *Remote Sensing of Environment*, 200 :341–351, 2017.

[10] F. Rottensteiner et al. The ISPRS benchmark on urban object classification and 3D building reconstruction. In *ISPRS Annals*, 2012.

[11] K. Simonyan and A. Zisserman. Very deep convolutional networks for large-scale image recognition. In *Proc. of ICLR*, 2015.

[12] J. J. Zhao, M. Mathieu, R. Goroshin, and Y. LeCun. Stacked what-where auto-encoders. In *Proc. of ICLR*, 2015.