Segmentation sémantique profonde par régression sur cartes de distances signées

Nicolas Audebert, Alexandre Boulch, Bertrand Le Saux, Sébastien Lefèvre

Segmentation sémantique profonde par régression sur cartes de distances signées

Nicolas Audebert[1,2] Alexandre Boulch[1] Bertrand Le Saux[1] Sébastien Lefèvre[2]

[1] ONERA, *The French Aerospace Lab*, F-91761 Palaiseau, France
[2] Univ. Bretagne-Sud, UMR 6074, IRISA, F-56000 Vannes, France
{nicolas.audebert,alexandre.boulch,bertrand.le_saux}@onera.fr, sebastien.lefevre@irisa.fr

Résumé

La compréhension de scène est une tâche visuelle reposant à l'heure actuelle sur une segmentation sémantique des images, obtenue par des réseaux profonds entièrement convolutifs. Toutefois, la nature convolutive de ces réseaux rend les frontières imprécises et les formes mal segmentées, alimentant un besoin croissant en régularisation a posteriori. Nous proposons ici de reformuler la tâche de segmentation sémantique en termes de régression de cartes de distance. Nous montrons qu'une telle formulation permet d'entraîner des réseaux convolutifs multi-tâches dont les segmentations générées sont plus régulières qu'avec les méthodes usuelles basées directement sur une classification dense.

Mots Clef

Segmentation sémantique, apprentissage profond, cartes de distance.

Abstract

Understanding of visual scenes relies more and more on dense pixel-wise classification obtained via deep fully convolutional neural networks. However, due to the nature of the networks, predictions often suffer from blurry boundaries and ill-segmented shapes, fueling the need for post-processing. This work formulates the standard semantic segmentation problem in terms of distance regression. We show that it is possible to train a multi-task fully convolutional neural network that builds more regular segmentations than those produced by existing methods based on direct dense classification.

Keywords

Semantic segmentation, deep learning, distance transform.

1 Introduction

La segmentation sémantique est une brique de base pour la compréhension de scène. En classifiant tous les pixels d'une image de façon dense, il est alors possible de construire des représentations abstraites s'intéressant aux objets et à leurs formes.

Les réseaux entièrement convolutifs (*Fully Convolutional Networks* ou FCN) sont un outil particulièrement efficace pour la segmentation sémantique pour de nombreux types d'images : multimédia [10], aériennes [31], médicales [37] ou de véhicule autonome [8].

Cependant, la littérature fait régulièrement face à des problèmes de frontières inter-classes imprécises ou de segmentations bruitées, nécessitant de faire appel à des régularisations a posteriori pour lisser les segmentations [42, 21]. La communauté s'est ainsi penchée sur différents post-traitements pour améliorer la netteté des contours et contraindre les segmentations à respecter la même topologie que la vérité terrain. Bien souvent, il s'agit de modèles graphiques ajoutés en fin de réseau [40] ou faisant appel à des connaissances a priori [22, 5].

Ce travail propose une approche directe consistant en une régularisation implicite intégrée dans la représentation de la vérité terrain. En effet, nous proposons d'utiliser les cartes de distance issues des masques de segmentation comme tâche auxiliaire. Les cartes de distance indiquent non seulement l'appartenance d'un pixel à une classe donnée, mais également sa proximité spatiale vis-à-vis des autres classes d'intérêt et contient donc une information plus riche concernant la structure spatiale des données. Cette approche s'inscrit dans la veine de travaux sur l'utilisation de primitives géométriques pour régulariser la segmentation sémantique, comme la prédiction de l'orientation des objets [36] ou de la position de leur centre de masse [12] De fait, en modifiant de façon minimale des réseaux de segmentation existants, nous parvenons à obtenir des segmentation plus régulières sans post-traitement ou connaissance a priori. Nous validons notre méthode sur plusieurs architectures de réseaux convolutifs profonds et sur plusieurs applications en compréhension de scène urbaine, en segmentation d'images 2,5D et en observation de la Terre.

2 Contexte

La segmentation sémantique est une tâche particulièrement importante en vision par ordinateur. Plusieurs jeux de données ont été introduits pour comparer différentes méthodes de segmentation sur des images multimédia [24, 10], de conduite autonome [8, 6], de télédétection [27, 31] et médicales [37]. De nombreuses applications en compréhension de scène dépendent d'une classification dense au ni-

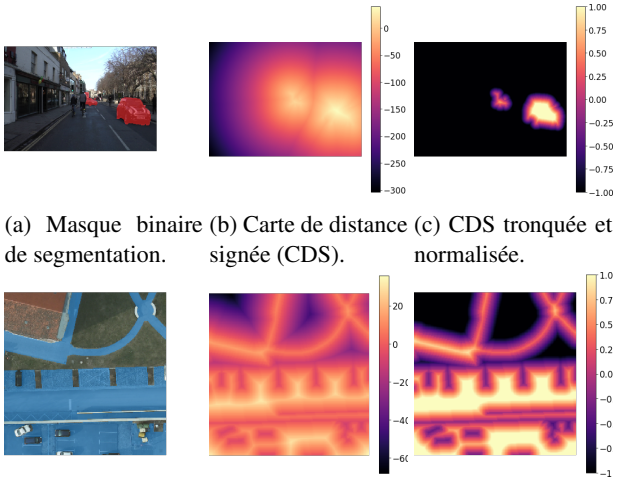
veau pixel, notamment en détection et segmentation d'objets [14, 2, 3, 34, 30, 7].

Les architectures de réseaux de neurones entièrement convolutives (FCN) [26] ont permis d'étendre le modèle de réseau convolutif à la segmentation sémantique, en transformant la classification par image en classification dense (par pixel). De nombreux modèles ont par la suite été proposés afin de tirer parti d'un contexte multi-échelles [21, 41] ou d'architectures symétriques encodeur-décodeur [4, 30]. Toutefois, les FCN tendent à produire des segmentations invariantes par translation locale, produisant un effet de flou au niveau des frontières. En outre, les segmentations peuvent être sujettes à des erreurs grossières comme des topologies d'objet non respectées (connexité, convexité, a priori polygonal…). La communauté s'est donc penchée sur des régularisations permettant de diminuer ces erreurs [11].

Ainsi, des modèles graphiques comme les champs aléatoires conditionnels ont été utilisés pour régulariser les frontières des segmentations [23], d'abord comme prétraitement séparé puis intégré au réseau de façon différentiable [42, 40]. Dans le même esprit, [22] reformule les méthodes variationnelles d'ensemble de niveau de façon à pouvoir les résoudre avec un FCN. Ces méthodes revisitent des algorithmes classiques de traitement d'image afin de les rendre différentiables, ce qui les rend compatibles avec la descente de gradient stochastique et la rétro-propagation utilisée pour l'entraînement des réseaux profonds. Cependant, elles nécessitent des ajustements considérables pour s'intégrer au cadre traditionnel de l'apprentissage profond et sont par ailleurs coûteuses en temps de calcul.

Une approche alternative consiste à réaliser une régularisation guidée par les données en modifiant la fonction objectif du modèle. Cette approche a notamment été étudiée dans le cadre de la détection de contours sémantisée [38]. Par exemple, [5, 20] introduisent une fonction objectif spécialement étudiée pour la détection de contours. CASENet [39] utilise des a priori sémantiques concernant les relations d'adjacence de différentes classes pour raffiner les contours, tandis que la stratégie COB [19] intègre des a priori géométriques dans sa fonction objectif. Des approches multi-échelles ont également démontré leur efficacité à améliorer la structure spatiale des segmentations par FCN, notamment en fusionnant des cartes d'activation issues de différentes couches [25]. La prédiction explicite des contours aide par ailleurs à régulariser la segmentation au niveau des bordures des objets, comme démontré dans [32, 28]. Ces deux approches peuvent être combinées pour en cumuler les effets [7, 9].

L'ensemble de ces méthodes s'intéresse à la réduction du bruit de classification et à l'amélioration de la structure spatiale de la classification en intégrant une notion de voisinage spatial et d'appartenance à une frontière. Toutefois, si ces informations sont difficiles à capturer dans des annotations par masque d'appartenance binaire à une classe, elles peuvent apparaître plus explicitement dans des cartes



(a) Masque binaire de segmentation. (b) Carte de distance signée (CDS). (c) CDS tronquée et normalisée.

(d) Masque binaire de segmentation. (e) Carte de distance signée (CDS). (f) CDS tronquée et normalisée.

FIGURE 1 – Différentes représentations de segments annotés.

de distances. Ainsi, [12] propose d'utiliser la prédiction de cartes de distances afin de segmenter des objets contenus dans des boîtes englobantes même si l'objet sort des limites de la boîte.

Nous proposons donc dans ces travaux d'utiliser la régression des cartes de distances comme tâche intermédiaire dans un réseau de segmentation sémantique. En particulier, avec une modification minime sur la structure du réseau, nous utilisons cette tâche auxiliaire comme régularisateur afin de raffiner les cartes de segmentation issues d'un FCN.

3 Régression des cartes de distances

Ce travail s'intéresse à l'utilisation des cartes de distances signées (CDS) pour régulariser des réseaux de segmentation sémantique. Le passage en carte de distances transforme un masque binaire en une représentation équivalente mais à valeurs continues. En l'occurence, nous travaillons avec des cartes de distances signées tronquées puis renormalisées dans $[-1, 1]$. Ces représentations des annotations sont illustrées dans la Figure 1. Nous émettons l'hypothèse que cette représentation permet toutefois d'accéder plus directement à la structure spatiale des données, notamment car elle contient pour un pixel sa distance spatiale relative à toutes les classes d'intérêt. Cette représentation est donc plus riche en quantité d'information que les masques binaires utilisés pour la classification. Nous montrons que l'utilisation de la régression des CDS en tâche auxiliaire d'un réseau de segmentation sémantique a des résultats bénéfiques sur la segmentation finale.

La régression directe des CDS ne permet pas d'obtenir de meilleurs résultats de segmentation que la classification dense pixel à pixel usuelle. De fait, nous proposons donc d'utiliser une stratégie d'apprentissage multi-tâches dans laquelle le réseau est optimisé à la fois sur la classification

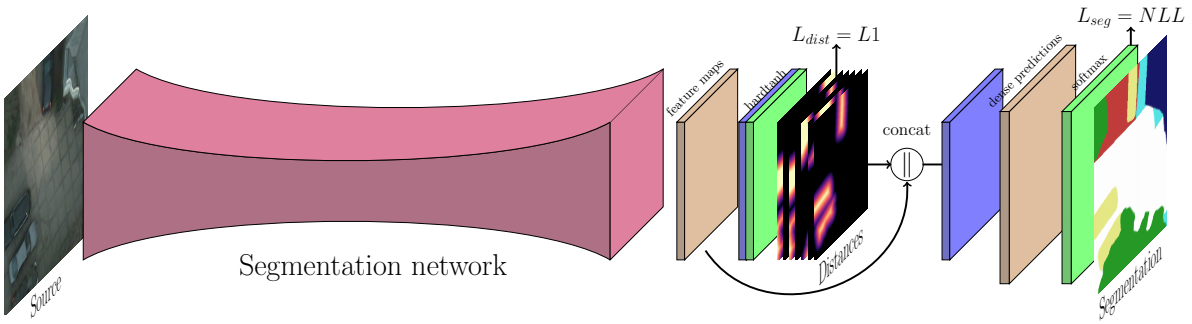


FIGURE 2 – Apprentissage multi-tâches (classification pixel à pixel et régression des cartes de distances). Les couches convolutives sont en bleu, les activations non-linéaires en vert et les cartes d'activation en marron.

des pixels et sur la régression des CDS.
En particulier, nous modifions l'architecture du réseau pour, dans un premier temps, effectuer la régression des CDS ; puis nous ajoutons une couche convolutive additionnelle pour fusionner les activations de la dernière couche avec les CDS prédites afin de réaliser la classification finale. Le réseau est ainsi entraîné en multi-tâches, la régression des CDS étant utilisée comme tâche intermédiaire avant la classification.
L'altération du réseau se résume comme suit. La dernière couche, habituellement suivie d'un *softmax* est ici utilisée comme couche de régression des CDS. Les distances étant normalisées entre -1 et 1, la fonction $hardtanh$ est utilisée comme activation non-linéaire. Puis nous concaténons les activations de la couche précédente aux CDS ainsi prédites pour alimenter une couche convolutive additionnelle suivie d'un *softmax*. L'architecture complète est illustrée dans la Figure 2.
Par souci d'équité dans nos expériences, les modèles de référence présentés dans la Section 4 utilisent également une couche de convolution additionnelle afin que tous les modèles équivalents possèdent le même nombre de paramètres optimisables.
Les fonctions de coût utilisées dans ces travaux sont la log-vraisemblance négative (NLL) pour la classification et la distance L1 pour la régression.
En notant respectivement $Z_{seg}, Z_{dist}, Y_{seg}, Y_{dist}$ la classification après *softmax*, la carte de distances prédite, les annotations de vérité terrain et la carte de distances réelle, la fonction de coût à minimiser est :

$$L = NLL(Z_{seg}, Y_{seg}) + \lambda L1(Z_{dist}, Y_{dist}) \qquad (1)$$

où λ est un hyperparamètre contrôlant l'amplitude de la régularisation.

4 Expériences

4.1 Références

Afin de pouvoir mesurer l'effet de la régression des cartes de distance, nous entraînons des réseaux avec l'architecture SegNet [4] ou PSPNet [41] de référence, soit en régression pure, soit en classification pure.

SegNet [4] est un modèle encodeur-décodeur conçu pour la conduite autonome. Sa conception dérive du modèle VGG-16 [33]. L'encodeur produit des cartes d'activation à résolution 1:32. Ces cartes sont ensuite suréchantillonnées et projetées dans l'espace sémantique par le décodeur.
PSPNet [41] est une architecture de segmentation sémantique récente ayant établi un nouvel état de l'art sur plusieurs jeux de données [8, 10]. Elle dérive du modèle ResNet [16] et utilise un module de concaténation en pyramide d'activations pour prendre en compte plusieurs niveaux de contexte spatial. Dans notre cas, nous utilisons une version réduite de PSPNet conçue sur l'architecture ResNet-101. ResNet-101 produit des cartes d'activation à résolution 1:32 qui sont suréchantillonnées par déconvolution.

4.2 Jeux de données

Nous validons notre approche sur plusieurs jeux de données afin de démontrer sa capacité à généraliser dans des contextes de segmentation mono et multi-classes sur plusieurs types d'images.

ISPRS 2D Semantic Labeling Le jeu de données ISPRS 2D Semantic Labeling [31] est constitué de deux ensembles d'images aériennes. Le premier a été acquis sur la ville de Vaihingen et comporte 33 images infrarouge-rouge-vert (IRRV) à une résolution de 9cm/px d'une taille moyenne d'environ 2000×1500px. Des annotations denses sur 16 images sont disponibles sur les six classes suivantes : routes, bâtiments, végétation basse, arbres, véhicules ainsi qu'une classe "autre". Le second ensemble a été acquis sur la ville de Potsdam et comporte 38 images infrarouge-rouge-vert-bleu (IRRVB) à une résolution de 5cm/px, pour une taille moyenne de 6000×6000px. Des annotations denses sont disponibles pour les mêmes classes que précédemment sur 24 images. L'évaluation est réalisée par validation croisée en divisant les jeux de données en trois subdivisions.

INRIA Aerial Image Labeling Benchmark Le jeu de données INRIA Aerial Image Labeling [27] contient 360 images RVB de taille 5000×5000px à une résolution de 30cm/px, couvrant 10 agglomérations de divers points du

Méthode	Ville	% classification	Routes	Bâtiments	Vég. basse	Arbres	Véhicules
SegNet* (régression)	Vaihingen	89.49	91.03	95.60	81.23	88.31	0.00
SegNet* (classification)		90.00	91.98	95.53	80.91	88.07	87.94
SegNet* (multi-tâches)		**90.43**	**92.46**	**95.99**	**81.30**	**88.34**	**88.16**
SegNet* (classification)	Potsdam	91.85	94.12	96.09	88.48	85.44	96.62
SegNet* (multi-tâches)		**92.22**	**94.33**	**96.52**	**88.55**	**86.55**	**96.79**

TABLE 1 – Résultats de validation croisée sur les jeux de données ISPRS. Les valeurs indiquées représentent le taux global de bonne classification et le score F1 pour chaque classe.

globe. La moitié des villes sont utilisées pour l'apprentissage et associées à des annotations publiques d'empreintes de bâtiments. Le reste du jeu de données est utilisé pour l'évaluation.

SUN RGB-D Le jeu de données SUN RGB-D [35] contient 10 335 images RVB accompagnées d'une carte de profondeur. Ces images ont été annotées sur 37 classes d'intérêt comportant le mobilier, les murs, le sol…

CamVid Le jeu de données CamVid [6] comporte 701 images extraites de plusieurs vidéos filmées par une caméra embarquée dans une voiture, avec une résolution de 360×480px. Nous utilisons la même division du jeu de données que [4], c'est-à-dire 367 images d'apprentissage, 101 images de validation et 233 images de test. Les annotations recouvrent 11 classes d'intérêt telles que "bâtiment", "piéton", "voiture" ou encore "trottoir".

4.3 Protocole expérimental

Nous expérimentons avec les modèles SegNet et PSPNet-101.
SegNet est entraîné pendant 50 000 itérations sur des *mini-batches* de 10 images. L'optimisation se fait par descente de gradient stochastique avec un taux d'apprentissage de 0,01, divisé par 10 après 25 000 et 45 000 itérations. Les poids de l'encodeur sont initialisés avec ceux de VGG-16 [33] pré-entraîné sur ImageNet. Les poids du décodeur sont initialisés aléatoirement en utilisant la stratégie proposée dans [15]. Pour le jeu de données multi-modal SUN RGB-D, nous utilisons le modèle FuseNet [13], qui consiste en un SegNet à double entrée. Sur les images aériennes, nous augmentons le nombre d'échantillons d'apprentissage en extrayant des images de 256×256 (384×384 pour le jeu de données INRIA Aerial Image) et en procédant aléatoirement à des symétries horizontales ou verticales. L'inférence est réalisée avec une fenêtre glissante de même dimension et un recouvrement de 75%.
PSPNet est entraîné sur CamVid pendant 750 000 itérations sur 10 images en parallèle par descente de gradient stochastique avec un taux d'apprentissage de 0,01, divisé par 10 après 500 000 itérations. Nous extrayons aléatoirement des imagettes de 224×224 et nous appliquons aléatoirement une symétrie horizontale. Suivant le protocole de [18], nous raffinons l'apprentissage en entraînant pendant 200 000 itérations sur les images à pleine résolution. Notre implémentation de PSPNet utilise les poids de ResNet-101 [16] pré-entraînés sur ImageNet pour l'initialisation, et n'utilise pas la fonction de coût auxiliaire présentée dans [41].
Finalement, nous compensons le déséquilibre des classes dans les jeux de données SUN RGB-D et CamVid en utilisant une pondération relativement à la fréquence médiane. Toutes les expériences sont réalisées à l'aide de la bibliothèque PyTorch [1]. Les CDS sont calculées sur CPU à l'aide de la bibliothèque Scipy [17] et conservées en mémoire pour éviter les calculs inutiles.

Méthode	I/U	% classification
SegNet* (classification)	65.04	94.74
SegNet* (multi-tâches)	**71.02**	**95.63**

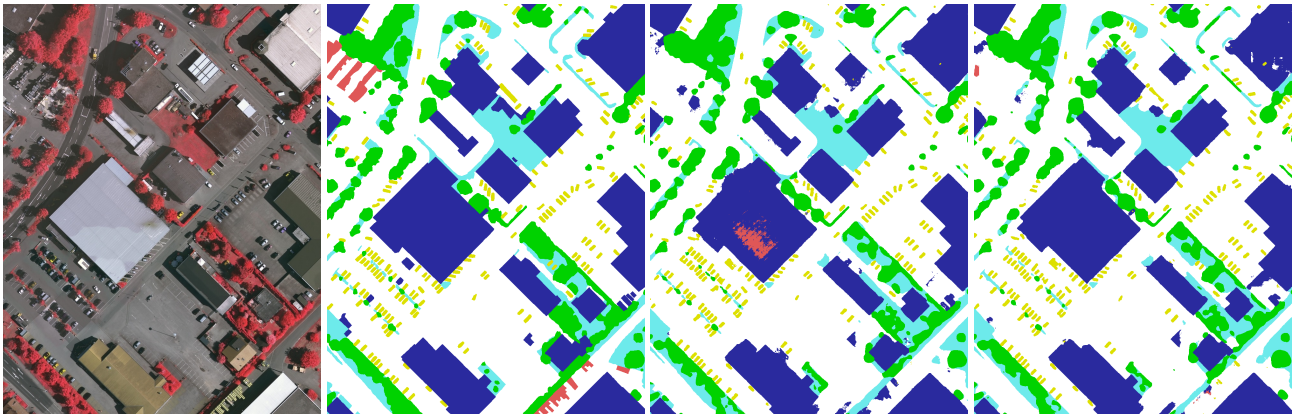
TABLE 2 – Résultats sur le jeu de données INRIA Aerial Image Labeling. Nous indiquons le taux global de bonne classification ainsi que le ratio intersection sur union (I/U).

4.4 Résultats

Dans les Tableaux 1 à 4, les modèles suffixés par "*" sont ceux proposés dans le cadre cette étude.

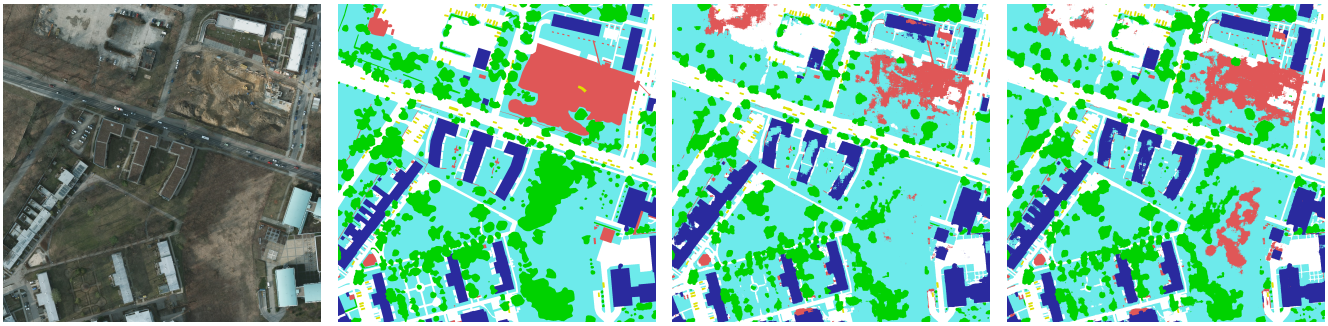
ISPRS dataset Les résultats de validation croisée sur les jeux de données ISPRS Vaihingen et Potsdam sont détaillés dans le Tableau 1. Toutes les classes semblent bénéficier de la régression des cartes de distances. En particulier, les arbres sur Potsdam sont significativement mieux segmentés, la régression de la CDS contraignant le réseau à prendre en compte la convexité naturelle de l'objet en dépit de l'absence de feuilles. Deux exemples de segmentation sont présentés en Figure 3 et Figure 4, dans lesquelles on peut voir que les bâtiments bénéficient grandement de l'approche multi-tâches (apparence plus lisse et moins de bruit de classification). Nous avons également testé l'approche par régression seule sur le jeu de données ISPRS Vaihingen avec des résultats mitigés. En effet, la plupart des classes bénéficient de ce traitement mais le réseau devient alors incapable de segmenter les véhicules, provoquant dans l'ensemble une baisse du taux de bonne classification.

INRIA Aerial Image Labeling Benchmark Les résultats sur le jeu de données INRIA Aerial Image Labeling



(a) Image IRRV (b) Vérité terrain (c) SegNet (classification) (d) SegNet (multi-tâches)

FIGURE 3 – Extrait des résultats de segmentation sur le jeu de données ISPRS Vaihingen. Légende : blanc : routes, bleu : bâtiments, cyan : végétation basse, vert : arbres, jaune : véhicules, rouge : autre, noir : indéfini.



(a) Image RVB (b) Vérité terrain (c) SegNet (classification) (d) SegNet (multi-tâches)

FIGURE 4 – Extrait des résultats de segmentation sur le jeu de données ISPRS Potsdam. Légende : blanc : routes, bleu : bâtiments, cyan : végétation basse, vert : arbres, jaune : véhicules, rouge : autre, noir : indéfini.

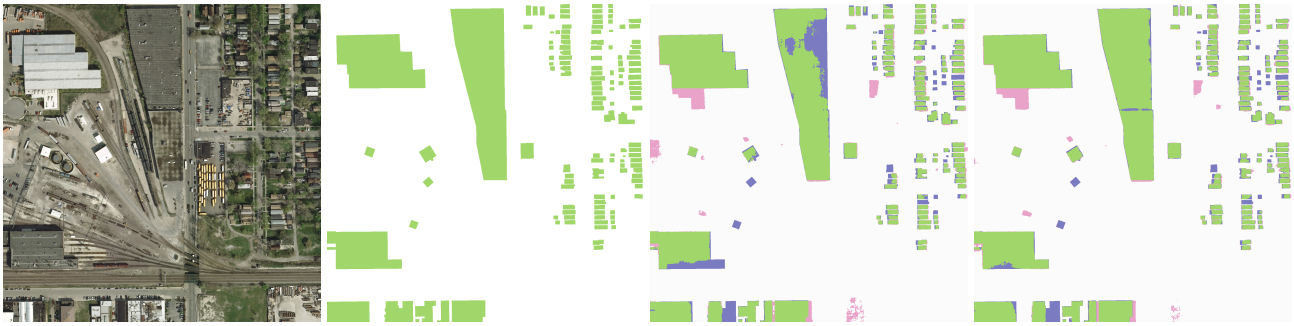
Méthode	% classification	I/U	Précision
3D Graph CNN [29]	-	42.0	55.2
3D Graph CNN [29] (multi-échelles)	-	**43.1**	55.7
FuseNet* [13]	76.8	39.0	55.3
FuseNet* (multi-tâches)	**77.0**	38.9	**56.5**

TABLE 3 – Résultats sur le jeu de données SUN RGB-D dataset (images de 224 × 224px). Les métriques utilisées sont le taux de bonne classification, la moyenne du rapport intersection sur union (I/U) et la précision moyenne.

sont détaillés dans le Tableau 2. L'utilisation de la régression sur les cartes de distances améliore significativement le ratio d'intersection sur union. Comme illustré dans la Figure 5, les formes des bâtiments respectent mieux l'a priori polygonal et la connexité des objets. Les bâtiments qui étaient déjà détectés sont segmentés avec plus de régularité.

SUN RGB-D Nous indiquons dans le Tableau 3 les résultats détaillés de segmentation sur le jeu de données SUN RGB-D. Le passage à un modèle multi-tâche améliore légèrement la précision moyenne et le taux moyen de bonne classification, contre une très faible diminution du rapport I/U. Ces résultats montrent que l'utilisation de la régression des cartes de distances s'étend également à des architectures multi-modales à double entrée. En outre, nos résultats sont comparables à ceux obtenus par [29] utilisant un réseau de neurones convolutif sur le graphe 3D de la scène, qui utilise donc une information plus riche.

CamVid Les résultats sur le jeu de données CamVid sont détaillés dans le Tableau 4 avec notamment une comparaison à la méthode de [18]. Plusieurs exemples qualitatifs sont illustrées dans la Figure 6. Le passage de PSPNet au mode de fonctionnement multi-tâches permet d'améliorer le rapport I/U global de presque 2 points et améliore la majorité des classes, à l'exception du ciel et des routes. Ceci est notamment dû à la présence de pixels non annotés, nombreux aux frontières de ces classes, provoquant la génération de cartes de distances inexactes. Dans l'ensemble, nos résultats sont compétitifs avec les méthodes de l'état



(a) RGB image (b) Ground truth (c) SegNet (classification) (d) SegNet (multi-task)

FIGURE 5 – Extrait des résultats de segmentation sur le jeu de données INRIA Aerial Image Labeling. Les pixels corrects sont en vert, les faux positifs en rose et les faux négatifs en bleu. L'approche multi-tâches capture mieux la structure spatiale des objets.

Méthode	I/U	% classif.	Bâtiments	Arbres	Ciel	Voiture	Panneau	Route	Piéton	Barrière	Poteau	Trottoir	Cycliste
SegNet [4]	46.4	62.5	68.7	52.0	87.0	58.5	13.4	86.2	25.3	17.9	16.0	60.5	24.8
DeepLab-LFOV [21]	61.6	–	81.5	74.6	89.0	**82.2**	42.3	92.2	48.4	27.2	14.3	75.4	50.1
DenseNet56 [18]	58.9	88.9	77.6	72.0	92.4	73.2	31.8	92.8	37.9	26.2	32.6	79.9	31.1
DenseNet103 [18]	**66.9**	**91.5**	**83.0**	**77.3**	**93.0**	77.3	**43.9**	**94.5**	**59.6**	37.1	**37.8**	**82.2**	50.5
PSPNet* (classification)	60.3	89.3	74.7	64.1	89.0	71.8	36.6	90.8	44.5	38.5	25.4	77.4	50.3
PSPNet* (multi-tâches)	62.2	90.0	76.2	66.4	88.8	78.0	37.6	90.7	47.2	**40.1**	28.6	78.9	**51.2**

TABLE 4 – Résultats sur le jeu de données CamVid incluant le rapport d'intersection sur union (I/U) global et pour chaque classe, ainsi que le taux de bonne classification.

de l'art [18, 21].

4.5 Discussion

L'intégration de la régression des cartes de distances dans un cadre d'optimisation multi-tâches permet d'améliorer et de lisser la structure spatiale des segmentations prédites par le réseau. En particulier, le modèle est contraint d'apprendre la notion de proximité spatiale d'un pixel par rapport à des classes voisines. En particulier, dans le cas des images aériennes, des arbres dont le feuillage tombe en hiver peuvent révéler le sol. La réponse spectrale des filtres correspond alors à un mélange de texture, bien que l'annotation recherchée corresponde à l'enveloppe convexe de l'arbre. La régression des cartes de distances permet d'orienter le réseau vers des recherches de structure géométriques, moins dépendantes de la radiométrie locale. En outre, cela permet de limiter la présence du bruit de classification poivre et sel qui est habituellement corrigé par des modèles graphiques a posteriori.

5 Conclusion

Dans ces travaux, nous avons étudié la tâche de segmentation sémantique à l'aide de réseaux de neurones entièrement convolutifs. La segmentation sémantique est souvent utilisée comme première étape pour de nombreuses méthodes de compréhension de scènes et est donc une tâche de grande importance. En dépit des excellents résultats qu'ils permettent d'obtenir, les FCN nécessitent souvent une régularisation explicite afin de pouvoir obtenir une segmentation visuellement régulière, soit en altérant la fonction objectif, soit par post-traitement à base de modèle graphique. Nous avons étudié dans ces travaux une représentation alternative des annotations pour la segmentation sémantique. Plus précisément, nous avons proposé d'utiliser les cartes de distances signées issues de la transformation des masques binaires. Nous avons montré que l'intégration de la régression de ces cartes de distances comme tâche auxiliaire dans un FCN permettait de régulariser les segmentations. Notamment, cela contraint le modèle à apprendre une notion de proximité spatiale d'un pixel à une classe d'intérêt. Cette régularisation s'obtient de façon implicite, uniquement à partir des données et sans connaissance a priori ou hypothèse faite sur les données. En outre, elle nécessite une altération minimale du réseau et de fait n'allonge quasiment pas le temps de calcul. Nous avons obtenu de façon systématique une amélioration de la segmentation obtenue par des FCN sur des applications en conduite autonome, télédétection et imagerie robotique avec carte de profondeur.

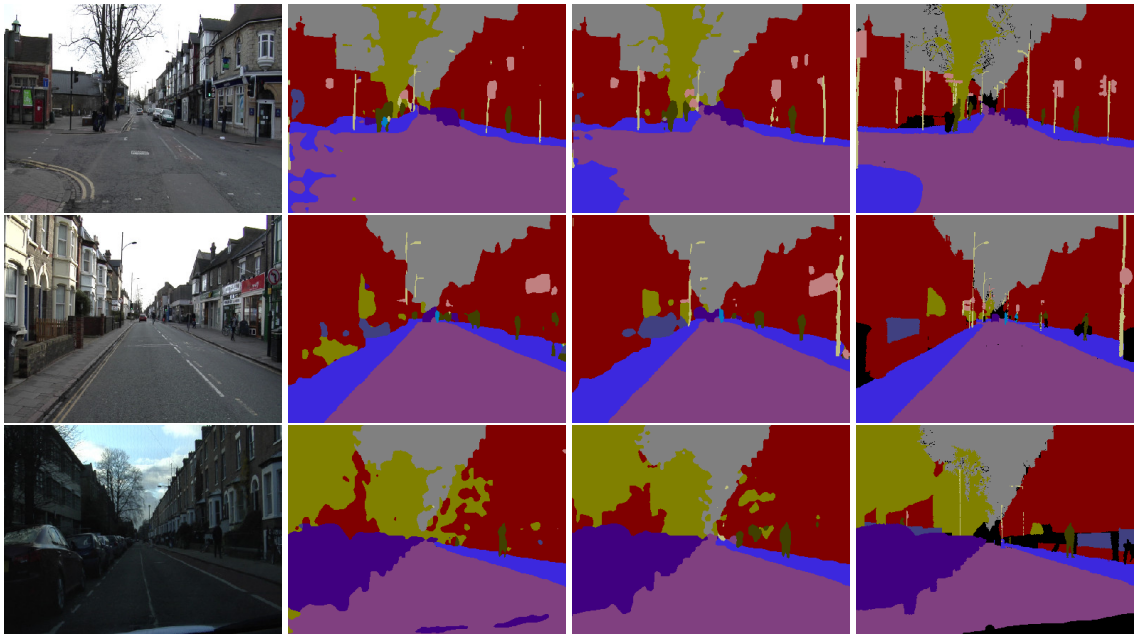


FIGURE 6 – Exemple de résultats de segmentation sémantique sur le jeu de données CamVid. De gauche à droite : image RVB, PSPNet (classification), PSPNet (multi-tâches), annotations.

Références

[1] PyTorch : Tensors and Dynamic neural networks in Python with strong GPU acceleration, 2016. http://pytorch.org/. 4

[2] A. Arnab and P. H. S. Torr. Pixelwise Instance Segmentation With a Dynamically Instantiated Network. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 441–450, 2017. 2

[3] 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) :368, Apr. 2017. 2

[4] V. Badrinarayanan, A. Kendall, and R. Cipolla. SegNet : A Deep Convolutional Encoder-Decoder Architecture for Scene Segmentation. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 39(12) :2481–2495, 2017. 2, 3, 4, 6

[5] G. Bertasius, J. Shi, and L. Torresani. Semantic Segmentation With Boundary Neural Fields. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 3602–3610, 2016. 1, 2

[6] G. J. Brostow, J. Fauqueur, and R. Cipolla. Semantic object classes in video : A high-definition ground truth database. *Pattern Recognition Letters*, 30(2) :88–97, Jan. 2009. 1, 4

[7] H. Chen, X. Qi, L. Yu, and P.-A. Heng. DCAN : Deep Contour-Aware Networks for Accurate Gland Segmentation. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 2487–2496, 2016. 2

[8] M. Cordts, M. Omran, S. Ramos, T. Rehfeld, M. Enzweiler, R. Benenson, U. Franke, S. Roth, and B. Schiele. The Cityscapes Dataset for Semantic Urban Scene Understanding. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 3213–3223, June 2016. 1, 3

[9] D. Cheng, G. Meng, S. Xiang, and C. Pan. FusionNet : Edge Aware Deep Convolutional Networks for Semantic Segmentation of Remote Sensing Harbor Images. *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, 10(12) :5769–5783, 2017. 2

[10] M. Everingham, S. M. A. Eslami, L. V. Gool, C. K. I. Williams, J. Winn, and A. Zisserman. The Pascal Visual Object Classes Challenge : A Retrospective. *International Journal of Computer Vision*, 111(1) :98–136, June 2014. 1, 3

[11] A. Garcia-Garcia, S. Orts-Escolano, S. Oprea, V. Villena-Martinez, and J. Garcia-Rodriguez. A Review on Deep Learning Techniques Applied to Semantic Segmentation. *arXiv :1704.06857 [cs]*, Apr. 2017. arXiv : 1704.06857. 2

[12] Z. Hayder, X. He, and M. Salzmann. Boundary-aware Instance Segmentation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2017. 1, 2

[13] C. Hazirbas, L. Ma, C. Domokos, and D. Cremers. FuseNet : Incorporating Depth into Semantic Segmentation via Fusion-Based CNN Architecture. In *Computer Vision ACCV 2016*, pages 213–228. Springer, Cham, Nov. 2016. 4, 5

[14] K. He, G. Gkioxari, P. Dollár, and R. Girshick. Mask R-CNN. In *International Conference on Computer Vision (ICCV)*, Mar. 2017. arXiv : 1703.06870. 2

[15] K. He, X. Zhang, S. Ren, and J. Sun. Delving Deep into Rectifiers : Surpassing Human-Level Performance on ImageNet Classification. In *IEEE International Conference on Computer Vision (ICCV)*, pages 1026–1034, Dec. 2015. 4

[16] K. He, X. Zhang, S. Ren, and J. Sun. Deep Residual Learning for Image Recognition. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 770–778, June 2016. 3, 4

[17] E. Jones, T. Oliphant, P. Peterson, and others. SciPy : Open source scientific tools for Python, 2001. https://www.scipy.org/. 4

[18] S. Jégou, M. Drozdzal, D. Vazquez, A. Romero, and Y. Bengio. The One Hundred Layers Tiramisu : Fully Convolutional DenseNets for Semantic Segmentation. In *IEEE Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, pages 1175–1183, July 2017. 4, 5, 6

[19] K. K. Maninis, J. Pont-Tuset, P. Arbelaez, and L. Van Gool. Convolutional Oriented Boundaries : From Image Segmentation to High-Level Tasks. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, PP(99) :1–1, 2017. 2

[20] I. Kokkinos. Pushing the Boundaries of Boundary Detection using Deep Learning. *arXiv :1511.07386 [cs]*, Nov. 2015. arXiv : 1511.07386. 2

[21] L. C. Chen, G. Papandreou, I. Kokkinos, K. Murphy, and A. L. Yuille. DeepLab : Semantic Image Segmentation with Deep Convolutional Nets, Atrous Convolution, and Fully Connected CRFs. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, PP(99) :1–1, 2018. 1, 2, 6

[22] N. Le, K. Gia Quach, K. Luu, M. Savvides, and C. Zhu. Reformulating Level Sets as Deep Recurrent Neural Network Approach to Semantic Segmentation. *ArXiv e-prints*, 1704 :arXiv :1704.03593, Apr. 2017. 1, 2

[23] G. Lin, C. Shen, A. Van Den Hengel, and I. Reid. Efficient piecewise training of deep structured models for semantic segmentation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, Apr. 2015. 2

[24] T.-Y. Lin, M. Maire, S. Belongie, J. Hays, P. Perona, D. Ramanan, P. Dollár, and C. L. Zitnick. Microsoft COCO : Common Objects in Context. In D. Fleet, T. Pajdla, B. Schiele, and T. Tuytelaars, editors, *Computer Vision ECCV 2014*, number 8693 in Lecture Notes in Computer Science, pages 740–755. Springer International Publishing, Sept. 2014. DOI : 10.1007/978-3-319-10602-1_48. 1

[25] Y. Liu, M.-M. Cheng, X. Hu, K. Wang, and X. Bai. Richer Convolutional Features for Edge Detection. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 3000–3009, 2017. 2

[26] J. Long, E. Shelhamer, and T. Darrell. Fully convolutional networks for semantic segmentation. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 3431–3440, June 2015. 2

[27] E. Maggiori, Y. Tarabalka, G. Charpiat, and P. Alliez. Can Semantic Labeling Methods Generalize to Any City ? The Inria Aerial Image Labeling Benchmark. In *IEEE International Symposium on Geoscience and Remote Sensing (IGARSS)*, July 2017. 1, 3

[28] D. Marmanis, K. Schindler, J. D. Wegner, S. Galliani, M. Datcu, and U. Stilla. Classification With an Edge : Improving Semantic Image Segmentation with Boundary Detection. *ISPRS Journal of Photogrammetry and Remote Sensing*, 2017. arXiv : 1612.01337. 2

[29] X. Qi, R. Liao, J. Jia, S. Fidler, and R. Urtasun. 3d Graph Neural Networks for RGBD Semantic Segmentation. In *IEEE International Conference on Computer Vision (ICCV)*, pages 5199–5208, 2017. 5

[30] O. Ronneberger, P. Fischer, and T. Brox. U-Net : Convolutional Networks for Biomedical Image Segmentation. In *Medical Image Computing and Computer-Assisted Intervention (MICCAI)*, pages 234–241. Springer International Publishing, Cham, 2015. 2

[31] F. Rottensteiner, G. Sohn, J. Jung, M. Gerke, C. Baillard, S. Benitez, and U. Breitkopf. The ISPRS benchmark on urban object classification and 3d building reconstruction. *ISPRS Annals of the Photogrammetry, Remote Sensing and Spatial Information Sciences*, 1 :3, 2012. 1, 3

[32] W. Shen, X. Wang, Y. Wang, X. Bai, and Z. Zhang. DeepContour : A deep convolutional feature learned by positive-sharing loss for contour detection. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 3982–3991, June 2015. 2

[33] K. Simonyan and A. Zisserman. Very Deep Convolutional Networks for Large-Scale Image Recognition. *arXiv :1409.1556 [cs]*, Sept. 2014. 3, 4

[34] L. Sommer, K. Nie, A. Schumann, T. Schuchert, and J. Beyerer. Semantic labeling for improved vehicle detection in aerial imagery. In *IEEE International Conference on Advanced Video and Signal Based Surveillance (AVSS)*, pages 1–6, Aug. 2017. 2

[35] S. Song, S. P. Lichtenberg, and J. Xiao. SUN RGB-D : A RGB-D scene understanding benchmark suite. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 567–576, June 2015. 4

[36] J. Uhrig, M. Cordts, U. Franke, and T. Brox. Pixel-Level Encoding and Depth Layering for Instance-Level Semantic Labeling. In *Pattern Recognition*, Lecture Notes in Computer Science, pages 14–25. Springer, Cham, Sept. 2016. 1

[37] V. Ulman, M. Maska, K. E. G. Magnusson, O. Ronneberger, C. Haubold, N. Harder, P. Matula, P. Matula, D. Svoboda, M. Radojevic, I. Smal, K. Rohr, J. Jaldén, H. M. Blau, O. Dzyubachyk, B. Lelieveldt, P. Xiao, Y. Li, S.-Y. Cho, A. C. Dufour, J.-C. Olivo-Marin, C. C. Reyes-Aldasoro, J. A. Solis-Lemus, R. Bensch, T. Brox, J. Stegmaier, R. Mikut, S. Wolf, F. A. Hamprecht, T. Esteves, P. Quelhas, . Demirel, L. Malmström, F. Jug, P. Tomancak, E. Meijering, A. Muñoz-Barrutia, M. Kozubek, and C. Ortiz-de Solorzano. An objective comparison of cell-tracking algorithms. *Nature Methods*, advance online publication, Oct. 2017. 1

[38] J. Yang, B. Price, S. Cohen, H. Lee, and M.-H. Yang. Object Contour Detection With a Fully Convolutional Encoder-Decoder Network. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 193–202, 2016. 2

[39] Z. Yu, C. Feng, M.-Y. Liu, and S. Ramalingam. CASENet : Deep Category-Aware Semantic Edge Detection. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 5964–5973, 2017. 2

[40] Z. Liu, X. Li, P. Luo, C. Change Loy, and X. Tang. Deep Learning Markov Random Field for Semantic Segmentation. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, PP(99) :1–1, 2017. 1, 2

[41] H. Zhao, J. Shi, X. Qi, X. Wang, and J. Jia. Pyramid Scene Parsing Network. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 2881–2890, 2017. 2, 3, 4

[42] S. Zheng, S. Jayasumana, B. Romera-Paredes, V. Vineet, Z. Su, D. Du, C. Huang, and P. H. S. Torr. Conditional Random Fields as Recurrent Neural Networks. In *IEEE International Conference on Computer Vision (ICCV)*, pages 1529–1537, Dec. 2015. 1, 2