



HAL
open science

Modèles d'Apprentissage Machine pour la Compréhension de Scènes

Bertrand Le Saux

► **To cite this version:**

Bertrand Le Saux. Modèles d'Apprentissage Machine pour la Compréhension de Scènes. Vision par ordinateur et reconnaissance de formes [cs.CV]. UNIVERSITE PARIS-SACLAY, 2019. tel-02918956

HAL Id: tel-02918956

<https://hal.science/tel-02918956v1>

Submitted on 21 Aug 2020

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.

Modèles d'Apprentissage Machine pour la Compréhension de Scènes

Habilitation à diriger des recherches de l'Université Paris-Saclay
préparée à l'Université Paris-Sud

Ecole doctorale n°580 Sciences et Technologies de l'Information et de la
Communication (STIC)
Spécialité : Automatique, Traitement du Signal, Traitement des Images, Robotique

HDR présentée et soutenue à Palaiseau, le 18 Décembre 2019, par

BERTRAND LE SAUX

Composition du Jury :

Renaud Marlet Professeur, École des Ponts ParisTech (Imagine) Chercheur, Valeo.ai	Président
Jocelyn Chanussot Professeur, Institut National Polytechnique de Grenoble (Gipsa-Lab)	Rapporteur
Florence Tupin Professeur, Télécom Paris (LTCl)	Rapporteuse
Josiane Zerubia Directrice de Recherche, Institut National de Recherches en Informatique et Automatique (Ayana)	Rapporteuse
François Fleuret Chercheur, IDIAP Professeur ordinaire, École Polytechnique Fédérale de Lausanne	Examineur
Sylvie Le Hégarat-Masclé Professeur, Université de Paris Sud (Satie)	Examinatrice

Résumé

La compréhension de scènes vise à répondre à la question : comment construire un modèle d'une région du monde réel afin d'y agir et d'y interagir ? Il s'agit donc d'extraire la sémantique et la géométrie des données disponibles : images, nuages de points 3D, etc. Dans ce but, plusieurs approches d'apprentissage machine sont présentées : elles diffèrent par la proportion d'a priori de conception et d'apprentissage introduits tout au long des algorithmes.

Trois aspects du problème sont envisagés. Les premiers travaux visent à la compréhension du contenu sémantique des images, c'est à dire la classification, la détection d'objets et la segmentation sémantique. Puis, plusieurs approches d'apprentissage sont proposées pour l'observation de la Terre et la télédétection, notamment pour l'apprentissage interactif, la classification sémantique multimodale et la détection de changements sémantiques. Enfin, l'accent est mis sur la vision 3D, avec l'estimation de la profondeur à partir d'une seule image et la classification de nuages de points 3D par des réseaux de neurones.

Ces approches variées reposent sur des mécanismes sous-jacents communs qui prennent une importance croissante. Elles réalisent une analyse multimodale pour bénéficier des données complémentaires disponibles, issues de capteurs différents mais aussi de sources et méta-données hétérogènes. Symétriquement, l'optimisation jointe d'objectifs multiples permet de régulariser l'apprentissage de modèles performants. Surtout, elles ont de plus en plus recours à une multiplicité des points de vue sur la scène pour relier, tant en apprentissage qu'en inférence, des invariances spatiales qui servent une analyse locale et une reconstruction sémantique globale. Cela est rendu possible par une intégration croissante de l'apparence et de la structure 3D, et conduit à une meilleure compréhension sémantique de la scène.

Remerciements

Je remercie en premier lieu Jocelyn Chanussot, Florence Tupin et Josiane Zerubia d'avoir accepté de relire ce mémoire et d'en être rapporteurs. Leurs remarques et leurs encouragements m'ont accompagné dans l'amélioration de ce document et la préparation de la présentation, et continueront à nourrir ma réflexion à l'avenir. Je remercie également les membres du jury : François Fleuret, Sylvie Le Hégarat Mascle, et Renaud Marlet, ce dernier m'ayant de plus fait l'honneur d'en être le président. J'espère que les discussions nées des multiples questions posées lors de la soutenance seront le point de départ de recherches passionnantes.

De nombreux travaux présentés dans ce manuscrit sont le fruit de collaborations avec les étudiants que j'ai eu la chance d'encadrer au cours de ces dernières années. Elles ont eu lieu dans le cadre de stages de fin d'études, de thèses mais aussi de projets universitaires, forme d'enseignement que j'affectionne particulièrement. Qu'ils en soient tous remerciés, et pour ne citer que ceux avec qui la relation s'est échelonnée sur plusieurs années : Nicolas Chauffert (stagiaire et moniteur lors des Modals Image et Vision de l'École Polytechnique), Adrien Lagrange (stagiaire que j'ai retrouvé en tant que membre de son jury de doctorat), Hicham Randriarivo (doctorant), Joris Guerry (doctorant et moniteur lors des Modal Image et Vision de l'École Polytechnique), Nicolas Audebert (doctorant), Marcela Pinheiro de Carvalho (doctorante), Rodrigo Caye Daudt (doctorant et moniteur à l'ENSTA pour les cours d'apprentissage machine), Javiera Castillo-Navarro (stagiaire puis doctorante), Antoine Manier (doctorant), Gaston Lenczner (doctorant), Clément Rambour (post-doctorant) et enfin Rémy Leroy (stagiaire puis doctorant).

Outre les étudiants, j'ai eu de précieux collaborateurs dans les différents laboratoires où j'ai travaillé. Pour ne citer que les co-auteurs ou co-encadrants les plus fréquents : Giuseppe Amato (ISTI / CNR di Pisa), Horst Bunke (Université de Berne), Bernard Chalmond, Jiaping Wang et Alain Trouvé (ENS de Cachan) Martial Sanfourche, Pauline Trouvé-Peloux, Frédéric Champagnat, Aurélien Plyer, Guy Le Besnerais, Élise Colin-Koeniguer, Stéphane Herbin, Adrien Chan Hon Tong et Alexandre Boulch (ONERA). Merci à eux tous et aux nombreux collègues (scientifiques, cadres, ou personnels de soutien) avec lesquels j'ai eu plaisir à travailler au quotidien.

Hors mes équipes d'accueil, de nombreuses personnes m'ont permis d'effectuer des activités diverses mais toujours passionnantes dans le monde académique. Il serait difficile de lister toutes celles et tous ceux que j'ai croisés, mais les projets communs et les rencontres ont d'une manière ou d'une autre contribué à mon cheminement, tant sur le plan des outils que des domaines d'études, et je les en remercie. Cependant, le co-encadrement de thèses a été l'occasion de projets de recherche avec des enseignants-chercheurs que je remercie tout particulièrement : Marin Ferecatu et Michel Crucianu (CNAM Paris), David Filliat (ENSTA Paris), Sébastien Lefèvre (Université de Bretagne Sud), Andrés Almansa (CNRS), Yann Gousseau (Telecom

Paris). je suis également très reconnaissant aux nombreux responsables d'enseignements et collègues avec qui j'ai donné des cours devant des auditoires variés : Fabrice Rossi (alors à l'Université Paris Dauphine), Yvan Bonnassieux, Razvigor Ossikovski et Jean-Charles Vanel (École Polytechnique), François Goudail (Institut d'Optique Graduate School), Thierry Bernard (ENSTA Paris), Loïc Landrieu et à nouveau Sébastien Lefèvre (dans le cadre d'EuroSDR), etc. Enfin, l'[IEEE GRSS](#) m'a permis de développer des activités particulièrement ambitieuses au cours de ces dernières années. Merci tout spécialement à Devis Tuia de m'avoir mis le pied à l'étrier en m'invitant de le rejoindre à l'[IADF TC](#), ainsi qu'à Gabriele Moser, Ronny Hänsch et Naoto Yokoya (co-responsables du comité technique avec qui j'ai travaillé quotidiennement pendant quatre ans en jonglant avec les fuseaux horaires), Irena Hanjsek, Paolo Gamba et les membres de l'[IEEE GRSS](#) adcom, et enfin Avik Bhattacharya éditeur en chef des *Geoscience and Remote Sensing Letters* pour m'avoir invité à faire partie du comité éditorial de la revue.

Table des matières

Résumé	iii
Remerciements	v
Table des matières	vi
Liste des figures	viii
Liste des tableaux	x
Avant-propos	xiii
1 Compréhension du contenu sémantique des images	1
1.1 Classification d'images et de scènes	2
1.1.1 Navigation dans les bases d'images	2
1.1.2 Reconnaissance de scènes	5
1.2 Observation de scène et reconnaissance d'objets pour drones	8
1.3 Segmentation sémantique image et 3D mono-vue	11
1.4 Résumé et conclusions	16
2 Télédétection	19
2.1 Apprentissage automatique pour l'observation de la Terre	20
2.2 Apprentissage profond pour l'observation de la Terre	25
2.3 Jeux de données pour l'apprentissage machine en télédétection	33
2.4 Résumé et conclusions	40
3 Vision et compréhension 3D	41
3.1 Reconstruction 3D en microscopie confocale	41
3.2 Estimation 3D mono-image	44
3.3 Segmentation sémantique de scènes 3D	49
3.4 Résumé et conclusions	53
4 Perspectives	55
4.1 Conclusions et vue d'ensemble	55
4.2 Perspectives en compréhension de scènes à large-échelle	56
4.2.1 À court terme : apprendre avec peu de données	57
4.2.2 À moyen terme : apprendre avec des données géospatiales multiples	59

4.2.3	À plus long terme : vers une analyse globale de la Terre en 3D	60
4.3	Perspectives en compréhension de l'environnement local	61
4.3.1	À court terme : prédire et sémantiser des nuages de points	63
4.3.2	À moyen terme : vers des modèles plus abstraits	65
4.3.3	À plus long terme : analyser la scène 3D multitemporelle	66
5	Notice individuelle et rapports d'activité	69
5.1	Curriculum Vitae	69
5.1.1	Activités professionnelles	69
5.1.2	Parcours académique	70
5.1.3	Enseignements / Encadrements	70
5.2	Rapport d'activité de recherche	71
5.2.1	Production scientifique et brevets	71
5.2.2	Prix scientifiques	72
5.2.3	Activités d'intérêt collectif	73
5.2.4	Activités d'expertise et jury	74
5.3	Rapport d'activité lors d'études et recherches contractuelles	75
5.3.1	Activités de recherche collaborative	75
5.4	Rapport d'activité d'enseignement et d'encadrement	76
5.4.1	Enseignement et formation	76
5.4.2	Encadrement de stages et de thèses	77
	Glossaire	85
	Glossaire	85
	Bibliographie	89
	Articles de revue internationale avec comité de lecture	89
	Articles de revue nationale avec comité de lecture	91
	Communications en congrès internationaux avec actes et comité de lecture	91
	Communications en congrès nationaux avec actes et comité de lecture	95
	Thèse et manuscrits	96
	Références	97

Table des figures

1.1	Interface de navigation dans une base d'images : vue d'ensemble	2
1.2	Résumés de séquences vidéo	4
1.3	Analyse de résumé de séquence vidéo	4
1.4	Classification non-supervisée de régions d'images	6
1.5	Construction des vecteurs d'occurrence de régions	6
1.6	Classification de scènes génériques et régions pertinentes pour la décision	7
1.7	Mots visuels pertinents pour reconnaître des scènes d'actualités	7
1.8	Approche d'apprentissage interactif pour l'analyse de scène depuis un drone	9
1.9	Analyse de scène par drone : carte de classification d'environnement	9
1.10	Analyse de scène par drone : classification dans la géométrie de la vidéo de la caméra embarquée	10
1.11	Augmentation de données cohérente avec la 3D pour données RGB-D	12
1.12	Algorithme SnapNet-R pour la segmentation sémantique de données RGB-D	13
1.13	Résultats de SnapNet-R pour la segmentation sémantique	15
1.14	Prédictions de détecteurs de personnes RGB-D	16
2.1	Modèle discriminants en télédétection	20
2.2	Classification interactive d'objets en télédétection	22
2.3	Détection interactive de changements en télédétection.	23
2.4	Modèles à parties déformables en télédétection	23
2.5	Détection d'objet par <i>Discriminatively-trained Model Mixture</i> (DtMM)	24
2.6	Cartographie automatique par réseaux de neurones entièrement convolutifs	26
2.7	Cartes de densité de trafic urbain	27
2.8	Réseau de neurones multimodal avec correction résiduelle.	28
2.9	Segmentation sémantique multimodale avec correction résiduelle	28
2.10	Réseau de neurones pour l'apprentissage joint à partir de données fortement hétérogènes.	29
2.11	Résultats de cartographie automatique par apprentissage conjoint avec OpenStreetMap	30
2.12	Segmentation sémantique de bâtiments	31
2.13	Réseau de neurones 3D pour la classification de données hyperspectrales	32
2.14	Jeu de données OSCD et détection de changements par réseaux convolutifs	32
2.15	Réseau de neurones multitâche pour la détection de changement sémantique	33
2.16	Détection de changement sémantique	34
2.17	Données diffusées dans le cadre du DFC2016	37
2.18	Distribution mondiale globale des données du DFC2017	38

2.19	Données multimodales 3D du DFC2018	39
3.1	Plans de coupe par micro-rotations du microscope confocal.	42
3.2	Reconstructions 3D de cellules.	43
3.3	Coupes xz de cellules.	43
3.4	Estimation de profondeur mono-image par flou de défocus et apprentissage profond	44
3.5	<i>Deep Belief Network</i> (DBN) pour l'estimation de profondeur mono-image par flou de défocus	45
3.6	Reconstructions de cartes de profondeur par DBNs	46
3.7	Architecture D3-Net pour l'estimation de carte de profondeur	46
3.8	D3-Net avec pénalité LS-GAN	47
3.9	Comparaison de la qualité de prédiction de D3-Net sur des images avec et sans flou de défocalisation	48
3.10	Estimation de profondeur sur données floues synthétiques et réelles, en intérieur et extérieur	48
3.11	Algorithme SnapNet pour la segmentation sémantique 3D.	49
3.12	Vues virtuelles générées par SnapNet	50
3.13	Résultats de segmentation sémantique 3D par SnapNet	51
3.14	Résultats de segmentation sémantique 3D par SnapNet pour les opérations de Recherche et Sauvetage	52
4.1	Évolution de la cartographie	56
4.2	Réseaux multitâches pour la compréhension de scènes à large échelle	57
4.3	Apprentissage faiblement supervisé pour la détection de changement	58
4.4	Cartographie sémantique et prédiction de la 3D locale par apprentissage multitâche	59
4.5	Architecture de réseau de neurones pour la prédiction de nuages de points mono- image	63
4.6	Prédictions de nuages de points 3D avec une seule image	64
5.1	Nombre de citations de mes articles et communications par année	72
5.2	Nombre de communications et articles publiés par chaque doctorant	82
5.3	Statistiques annuelles de publication des doctorants encadrés	82
5.4	Évolution annuelle du nombre de citations par doctorant	83

Liste des tableaux

1.1	Performances de SnapNet-R en segmentation sémantique sur SUNRGBD	13
1.2	Performances de SnapNet-R en segmentation sémantique sur NYUDv2	14
2.1	Co-évolution des données et des méthodes de classification en télédétection	20
2.2	Résultats de segmentation sémantique sur le jeu de test aérien d'ISPRS Potsdam	29
2.3	Principaux jeux de données en télédétection.	35
2.4	Principaux jeux de données publics en imagerie hyperspectrale	36
3.1	Liste des fonctions de coût pour la régression	47
5.1	Résumé de la production scientifique	71
5.2	Indicateurs bibliométriques globaux de la production scientifique	72
5.3	Collaborations nationales ou internationales	75
5.4	Activités d'enseignement	76
5.5	Liste des supports de cours produits	77
5.6	Stages encadrés	78
5.7	Retombées des stages encadrés sur le plan scientifique et contractuel	79
5.8	Thèses encadrées	80
5.9	Résumé de la production scientifique des thèses encadrées	81
5.10	Indicateurs bibliométriques de l'impact des thèses encadrées	83
5.11	Devenir des doctorants encadrés après leur soutenance.	83

Avant-propos

Mes travaux visent à développer des outils d'*apprentissage automatique pour la compréhension de scènes*. Ils sont donc à la croisée de l'intelligence artificielle, de la vision par ordinateur et de l'analyse d'image. Les domaines applicatifs concernent l'image multimédia, l'imagerie biologique, la télédétection et la robotique.

De manière très générale, ces recherches visent à comprendre une scène, c'est à dire répondre à la question : comment construire un modèle d'un lieu du monde réel afin d'y agir et d'y interagir ? Pour cela, un tel modèle doit accéder à une connaissance du contenu de la scène, c'est à dire y mettre un sens, l'expliquer selon des concepts intelligibles, ou plus prosaïquement identifier les éléments qui s'y trouvent. Il doit aussi intégrer la structure de la scène, pour pouvoir s'y déplacer, connaître la disposition des éléments, prendre des décisions et planifier des actions.

Comment donc trouver ce sens et cette structure ? Par appétence pour une certaine tradition empiriste, je me suis intéressé à l'apprentissage automatique qui vise à construire de tels modèles (ou dans l'état actuel des connaissances, des modèles très réduits qui effectuent une action, une tâche) selon des approches statistiques à partir d'échantillons du monde réel. La question sous-jacente est alors de déterminer l'algorithme qui créera le modèle permettant de prendre les meilleures décisions et de définir les meilleures actions dans de nouvelles scènes.

Enfin, ces échantillons ne sont que des représentations fortement dégradées de la réalité factuelle : des mesures, des signaux issus pour les machines de capteurs mêlant électronique, optique et mécanique. Comment parvenir à un modèle convenable à partir de cette perception partielle et sommaire ? La compréhension de scènes doit surmonter cette difficulté, par exemple par l'inclusion de connaissances a priori sur la nature des signaux reçus et un modèle général de la réalité.

Les recherches menées pour répondre à ces questions s'organisent selon plusieurs thèmes. L'axe principal concerne la *compréhension du contenu sémantique des images*, c'est à dire principalement la classification d'image, la détection d'objets et la segmentation sémantique. Ce thème est abordé dans le chapitre 1. Au cours de mon évolution de carrière, deux axes connexes s'y sont adjoints. Tout d'abord en *télédétection*, où mes travaux recouvrent les mêmes problématiques de compréhension du contenu des scènes observées, mais en s'adaptant au contexte particulier : capteurs spécifiques (SAR, hyperspectral), fusion d'information, détection de changements. Mes travaux dans ce domaine sont relatés dans le chapitre 2. Ensuite dans le domaine de la *vision 3D*, mes travaux visent à étudier le lien entre image et 3D sous l'angle de l'apprentissage statistique, avec des applications pour la reconstruction 3D, la classification 3D, ou l'estimation de profondeur. Ils sont décrits dans le chapitre 3. Enfin, les perspectives de ces travaux et mon projet de recherche actuel sont décrits dans le chapitre 4.

Ce mémoire est un résumé des mes travaux scientifiques. il vise à montrer un parcours

et à dégager une vue d'ensemble organisée. En revanche, ce n'est pas un traité scientifique, notamment car il ne vise pas à l'exhaustivité. Ce n'est pas non plus une thèse résolvant un problème selon un agenda précis, car mon activité a en partie été guidée par les chances de travailler sur des questions nouvelles, soulevées par les équipes d'accueil et les collaborations spontanées. Ce mémoire n'est pas non plus une simple compilation d'articles, car un de ses objectifs est aussi de raconter un cheminement.

Selon une classification plus universitaire que celle utilisée initialement, et pour reprendre les sensibilités des laboratoires où j'ai travaillé, mes travaux touchent au génie informatique, à l'automatique et au traitement du signal (section 61 du Conseil National des Universités), informatique (section 27) et mathématiques appliquées (section 26). Ils s'inscrivent dans le large domaine des Sciences et Technologies de l'Information et de la Communication (STIC). En tout état de cause, ils relèvent des sciences expérimentales, car les hypothèses et les modèles proposés sont évalués à l'aune de données et des expériences.

Chapitre 1

Compréhension du contenu sémantique des images

L'axe structurant de mes recherches est la compréhension du contenu sémantique de scène [1, 2, 3]. Une scène est une représentation d'un environnement réel, et cette représentation peut être plus ou moins complexe : simple ou multivue, selon une ou plusieurs modalités de perception. L'objectif est donc de décrire par des mots le contenu de la scène représentée : le type de scène, l'agencement et la composition de la scène, les objets présents, leur position, etc...

Ce chapitre explore plusieurs facettes de ce problème sous l'angle de la compréhension d'images, déclinée selon plusieurs tâches et applications. Tout d'abord, la recherche d'images par le contenu (*Content-based Image Retrieval (CBIR)*) dans la partie 1.1. L'objectif est de classer les images, soit seulement visuellement en regroupant les images semblables entre elles, soit avec de la sémantique en attribuant des images à une classe connue, identifiée par un seul mot ou concept. Ensuite, en partie 1.2 nous aborderons des scènes vues par un drone, qui n'ont donc à chaque instant qu'une vue très partielle de leur environnement, et chercherons à montrer comment il est possible de percevoir et comprendre les régions de cet environnement. Enfin, toujours avec l'objectif d'aider un robot à comprendre son environnement, nous nous intéresserons en partie 1.3 à la segmentation fine en régions et objets de ce qui est perçu, et à la reconnaissance de chacun de ces éléments.

Ainsi, tout au long de ce chapitre, la représentation sera de plus en plus complexe : une simple image qui ne donne qu'une vue statique de la scène, une représentation multivue qui permet d'estimer une reconstruction 3D sommaire, et enfin une représentation mono-vue mais multimode (image et 3D). En même temps, la complexité du contenu et la finesse avec laquelle il peut être décrit seront également croissantes : d'un simple *tag* à une cartographie de l'environnement exploré par un robot, et finalement à une description fine de tout ce qui apparaît dans le champ de vision.

Une idée sous-jacente dans l'évolution de ces travaux est que la complexité de représentation peut être utilisée pour construire une description riche de la scène. Dans ces approches de compréhension de scènes, de nombreux blocs sont appris de manière automatique, et pour cela des modèles de plus en plus complexes sont utilisés. Néanmoins, l'algorithme complet de compréhension de la scène peut garder une certaine simplicité en tirant parti d'a priori de vision, et notamment en combinant sémantique et 3D [4, 5].

1.1 Classification d'images et de scènes

1.1.1 Navigation dans les bases d'images

Approche de classification non-supervisée

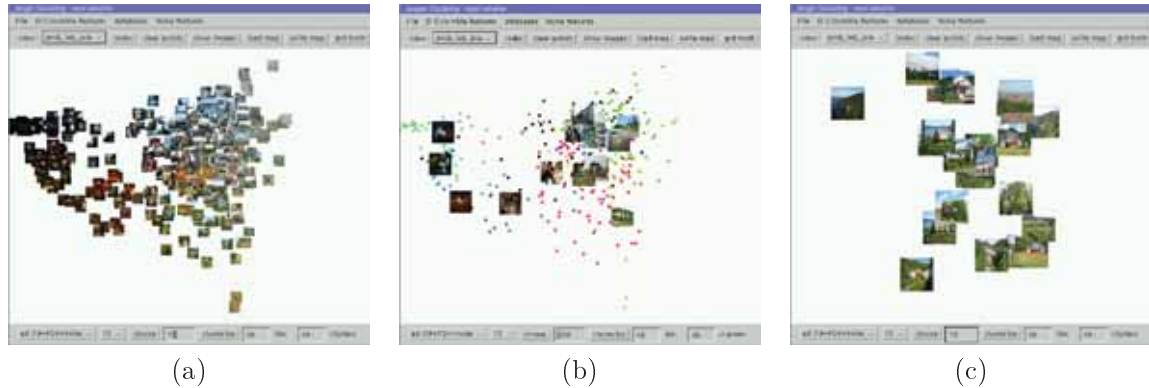


FIGURE 1.1 – Interface de navigation dans une base d'images. (a) Les imageries générées à partir des images de la base sont disposées dans l'espace de visualisation 2D. Cette vue globale de la base permet de se faire une idée grossière des teintes dominantes dans la base et de leurs répartitions dans le plan. (b) Puis seules les représentants des catégories obtenues par classification sont affichés, pour permettre à l'utilisateur d'accéder à chaque classe. (c) Lorsque l'utilisateur sélectionne et zoome sur une catégorie, toutes les images y appartenant sont affichées. La forme et la densité de la catégorie informent sur la proximité des images. Le gradient de la signature (ici le gradient de couleur) permet d'arranger les images suivant les directions du plan. Source : [PhD-2003].

Aperçu Mes premiers travaux (thèse [PhD-2003] et post-doctorat) s'inscrivaient dans le domaine des bases de données multimédia, et visaient à la **classification d'images par le contenu** [CI-1]. Tout d'abord de manière non-supervisée, avec Nozha Boujemaa à l'INRIA, pour répondre au problème de la *page zéro*, ou comment résumer une base d'images en quelques images pertinentes par **catégorisation visuelle**. Nous avons alors développé *Adaptive Robust Clustering* (ARC), une approche de *clustering* capable de trouver les catégories majoritaires dans un ensemble d'images [CI-3, CI-2, CI-4] (voir partie 1.1.1). Puis, de manière supervisée avec des méthodes à noyaux, soit avec les mêmes descripteurs standard pour le raffinement interactif des catégories [CI-7], soit avec des **vecteurs d'occurrences de régions visuelles** lors de mes travaux avec Giuseppe Amato (ISTI/CNR de Pise, Italie) puis Horst Bunke (Université de Berne, Suisse). Ces descripteurs sont construits par indexation par rapport à un dictionnaire des mots visuels appris sur la base [CI-5, CI-6], et ils sont complétés ensuite par une classification jointe avec un graphe d'organisation des régions pour apporter la structure spatiale de l'image [CI-8, CI-9] (voir partie 1.1.2). Ces travaux ont été appliqués pour la **classification d'images**, photos et vidéos d'actualités pour des utilisateurs tels que TF1 ou l'agence de presse italienne ANSA.

Contexte Ces travaux s'inscrivaient dans le contexte plus général de l'accès intelligent aux documents par le contenu visuel. Les recherches en ce domaine comprennent l'indexation et

la recherche d'images par le contenu (CBIR) [6, CI-1, 7], la gestion et l'interrogation de bases d'images et de bases multimédia. Une difficulté notable de ce domaine est de rapprocher les techniques développées des usages des utilisateurs potentiels. Plusieurs systèmes et plate-formes ont été proposés pour démontrer la pertinence des approches proposées, à commencer par QBIC [8], mais aussi Netra [9], RETIN [10].

Sans méta-données. Face à une base d'images inconnue, c'est à dire une collection de fichiers non visualisés, l'utilisateur commence par s'interroger sur son contenu et les thèmes qui y sont présents. Elle est en général trop grande pour être parcourue fichier par fichier. La recherche par mot-clé ou image exemple est une approche possible, mais pré-suppose d'avoir une idée de ce que contient la base. Une grande part de la base risque alors de rester inexploitée. Il y a donc besoin d'une vue d'ensemble qui renseigne sur les principales caractéristiques de la base d'images considérée et permette de l'appréhender dans sa globalité. Nous avons proposé de rechercher comment les images pouvaient être regroupées au sein de la base, et d'utiliser les groupes ainsi formés pour générer automatiquement son résumé visuel. Ce résumé, nommé *page zéro* [11], constitue alors un point de départ pour l'exploration et la navigation dans la base d'images. Notre solution est présentée en partie 1.1.1.

Avec méta-données. Une approche alternative, utilisée dans les moteurs de recherche traditionnels, est la recherche par mot-clé. La recherche s'effectue alors en comparant la requête aux méta-données, soit générées simultanément à la production du document, soit a posteriori par annotation. Hélas, les méta-données associées aux images sont rares. En effet, soit les méta-données de production sont perdues par manque d'automatisation des processus, soit les annotations sont coûteuses à produire. Une solution est alors la classification automatique des images. Cette tâche de vision par ordinateur n'était pas encore résolue par les algorithmes qui se sont développés à partir de l'apparition des challenges *Pascal* [12] (en 2005) et *ImageNet Large Scale Visual Recognition Challenge (ILSVRC)* [13] associé à [14] (en 2009). Cependant, les corpus existants d'images et d'annotations associées permettaient déjà d'envisager des approches supervisées pour résoudre ce problème, comme par exemple avec des *Machine à Vecteurs de Support – Support-Vector Machine (SVM)* [15]. En parallèle, des représentations de l'image fondées sur ses régions étaient apparues, notamment dans le domaine du CBIR, comme par exemple BLOBWorld [16]. Cette représentation par régions suivie d'une quantification par rapport à un dictionnaire était notamment l'idée mise en oeuvre pour la recherche d'images ou de scènes similaires [17, 18], l'annotation d'image (traduction d'images en mots) [19], et sera plus tard formalisée pour la classification de scènes sous nom de sacs de mots visuels (*Sac de Mots – Bag of Words (BoW)*) [20, 21]. Nous présentons notre propre approche par dictionnaire de régions et méthode à noyau en partie 1.1.2.

Classer les images d'une collection en groupes visuellement similaires pose plusieurs problèmes : comment définir la notion de similarité visuelle ? quel est le nombre de classes présentes ? que faire d'images isolées ?

Les images sont d'abord projetées dans un espace de représentation composé de descripteurs caractéristiques de l'apparence (histogrammes de couleur, descripteurs de Fourier, etc.), puis projetées dans un espace de dimension réduite (typiquement par *Analyse en Composantes Principales (ACP)*). La visualisation des images dans cet espace fait apparaître des groupes de densités et formes variées, qui correspondent cependant à des images de même classe et d'apparence visuelle similaire.

Nous avons développé un algorithme de classification non-supervisée (ou *clustering*) adapté à ces données, et qui de manière plus générale répond à plusieurs limitations connues des



FIGURE 1.2 – (a) Visualisation des images extraites de la vidéo dans l'espace des signatures. Les images sont disposées selon les deux premières composantes de la signature histogramme de couleur pondéré réduite par ACP. Les images groupées dans une même catégorie par notre méthode sont indiquées avec la même couleur. (b) Résumé de la séquence vidéo obtenus avec le nouvel algorithme ARC.



FIGURE 1.3 – Les catégories retrouvées correspondent à des scènes similaires selon un critère visuel, ce qui permet de regrouper des images d'une même scène indépendamment des changements de focales (a) et d'une grande variabilité dans la séquence (b).

méthodes standard de partitionnement. En particulier, la méthode ARC permet de déterminer automatiquement le nombre de classes, s'adapte à des groupes de données de densités diverses, et ne cherche pas à classer les données aberrantes et ambiguës [CI-2, CI-3]. En pratique, Le nombre de catégories existant dans les données est estimé au cours de la classification en minimisant une fonction d'Agglomération Compétitive [22] qui rend compte de la répartition des données et comporte un terme de régularisation fondé sur la validité de la partition obtenue.

Soit $X = \{x_i | i \in \{1 \dots N\}\}$ l'ensemble des représentations à classer. Soit $J : M_{fc} \times (\mathbb{R}^p)^c \mapsto \mathbb{R}^+$

$$(U, B) \rightarrow J(U, B) = \sum_{j=1}^C \sum_{i=1}^N (u_{ji})^2 d^2(x_i, \beta_j) - \sum_{j=1}^C \alpha_j \left[\sum_{i=1}^N (u_{ji}) \right]^2 \quad (1.1)$$

où u_{ji} est l'appartenance de x_i au groupe X_j , M_{fc} est l'ensemble des partitions floues de X , et ;

$$B = \{\beta_j | j \in \{1 \dots C\}\} \subset (\mathbb{R}^p)^c \text{ avec } \beta_j \in \mathbb{R}^p \quad (1.2)$$

est l'ensemble des prototypes des groupes X_j , tandis que α_j est le facteur d'adaptation du groupe X_j dans le terme de régularisation. Pour $1 \leq j \leq C$;

$$d^2(x_i, \beta_j) = \|x_i - \beta_j\|^2 \quad (1.3)$$

où $\|\cdot\|$ est n'importe quelle norme de \mathbb{R}^p induite par un produit scalaire. En pratique la distance de Gustafson-Kessel [23] est utilisée pour toutes les classes, sauf pour une classe virtuelle de données aberrantes [24] telles que $d^2(x_i, \beta_1) \leq \delta^2$ (les données dont la distance des classes réelles est supérieure à cette constante sont attribuées à cette classe virtuelle.).

Diverses applications ont été proposées, et tout d'abord pour la gestion d'albums de photos numériques personnels. En effet quantités de photos sont stockées sur les disques durs des ordinateurs [25], et les utilisateurs ne prennent plus la peine de les trier ou les classer. Des outils d'organisation automatique de leur collection d'images leur permettent donc de la parcourir et de retrouver un cliché particulier. La figure 1.1 présente le résultat d'une catégorisation de collection particulière par l'algorithme ARC et la navigation dans la base. Des procédures de raffinement interactif par SVM permettent aussi à l'utilisateur de corriger les catégories obtenues avec seulement quelques exemples [CI-7].

Notre approche a également été appliquée à des séquences vidéo de journal télévisé [CI-4]. L'objectif est alors de regrouper les images selon leur contenu et de présenter un résumé en image. En particulier, les services d'archives de la chaîne de télévision qui fournissait le corpus de vidéos sont intéressés par retrouver les scènes de plateau qui permettent de découper le journal en différents sujets. En pratique, des images sont simplement extraites de la vidéo, projetées dans l'espace des descripteurs et classés au moyen de l'algorithme ARC. La figure 1.2 montre l'ensemble des images-échantillons extraites de la vidéo et le résumé visuel des prototypes identifiés pour chaque classe. En haut à droite du résumé se trouve la figure connue d'un célèbre présentateur de TF1, entrée vers la catégorie des images de plateau. La figure 1.3 montre deux classes particulières qui illustrent la capacité de la méthode à capturer une similarité visuelle pertinente malgré des changements d'échelle ou le mouvement des protagonistes.

1.1.2 Reconnaissance de scènes

Approche de classification supervisée

Pour les cas où il existe déjà un corpus d'images annotées, il est possible de tirer parti de ces informations pour entraîner des classifieurs supervisés aptes à annoter des images inconnues. Notre objectif est ici la classification d'image par mot-clé significatif de la scène représentée. Afin de décrire la scène imagée, nous utilisons les différentes régions de l'image pour constituer un index de mots visuels présents ou non dans l'image. Le premier classifieur utilisé est une méthode à noyau, un Kernel-adatron qui implémente une version rapide de SVM. Le deuxième classifieur proposé rajoute l'information de structure de l'image en considérant le graphe des régions. Dans les deux approches, une procédure de sélection de caractéristiques pertinentes fondée sur l'information mutuelle permet d'accélérer l'apprentissage et d'améliorer grandement les performances.

En pratique, l'étape initiale consiste à segmenter chaque image en régions homogènes. L'algorithme Mean-Shift [26] est choisi pour des raisons d'efficacité algorithmique et de qualité des résultats. Les régions obtenues sont rassemblées (Fig. 1.4-a) et l'algorithme ARC (voir partie 1.1.1) de classification non-supervisée est appliqué à la base de régions ainsi constituée. Les prototypes des classes obtenues sont présentés sur la figure 1.4-b. Cet ensemble de classes de

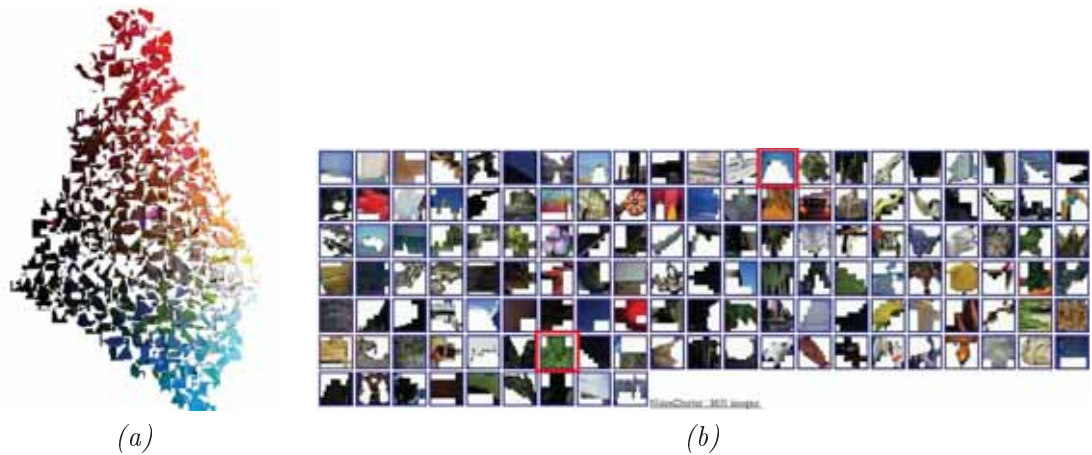


FIGURE 1.4 – (a) Images des régions représentées selon les coordonnées de leur couleur moyenne dans l'espace LUV . Les imageries sont obtenues en coupant la zone correspondant à la région dans l'image originale et en la plaçant sur un fond blanc. Pour des raisons de visibilité, seulement 10% de la base est représentée. (b) Prototypes résultant de la classification non-supervisée des régions par l'algorithme ARC : divers groupes sont constitués, contenant des ciels, zones de verdure, terrain, fleurs, visage, vêtement, etc. Source : [PhD-2003].

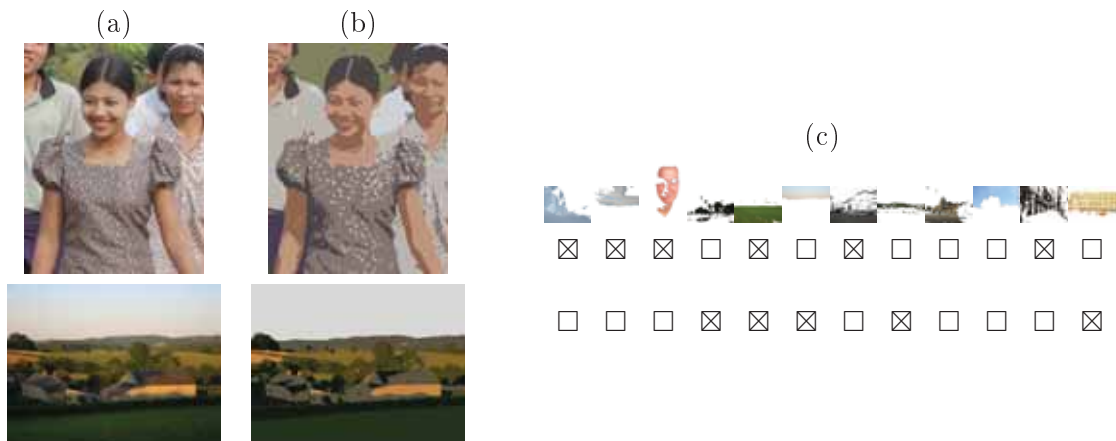


FIGURE 1.5 – Construction des vecteurs d'occurrence de régions : l'image originale (a) est tout d'abord segmentée (b) puis les régions sont mises en correspondance avec les classes du dictionnaire (voir figure 1.4) pour construire le vecteur booléen d'occurrences de régions. Source : [CI-5, CI-6].

régions similaires constitue un dictionnaire de caractéristiques locales ou *mots visuels*. Toute image (en apprentissage ou en test) peut alors être décrite par rapport à ce dictionnaire. L'image est segmentée, et chacune de ses régions est mise en correspondance avec sa classe dans le dictionnaire. Puis, un vecteur d'occurrences de mots visuel (une classe de caractéristiques) présents ou non dans l'image est construit. Ce vecteur est de taille fixe (la dimension du dictionnaire) et différent selon le type de scène de l'image, comme le montre la figure 1.5.

La première approche de classification proposée exploite directement les vecteurs d'occurrence. Elle est entraînée en deux étapes sur une collection de scènes annotées. D'abord, une



FIGURE 1.6 – Classification de scènes génériques et régions pertinentes pour la décision. 1ère rangée, les images originales à classer ; 2e rangée, les régions pertinentes pour la classification, c'est à dire celles dont l'entrée dans le dictionnaire a une forte information mutuelle avec la classe à prédire ; 3e rangée, le type de scène prédit. Source : [CI-6].

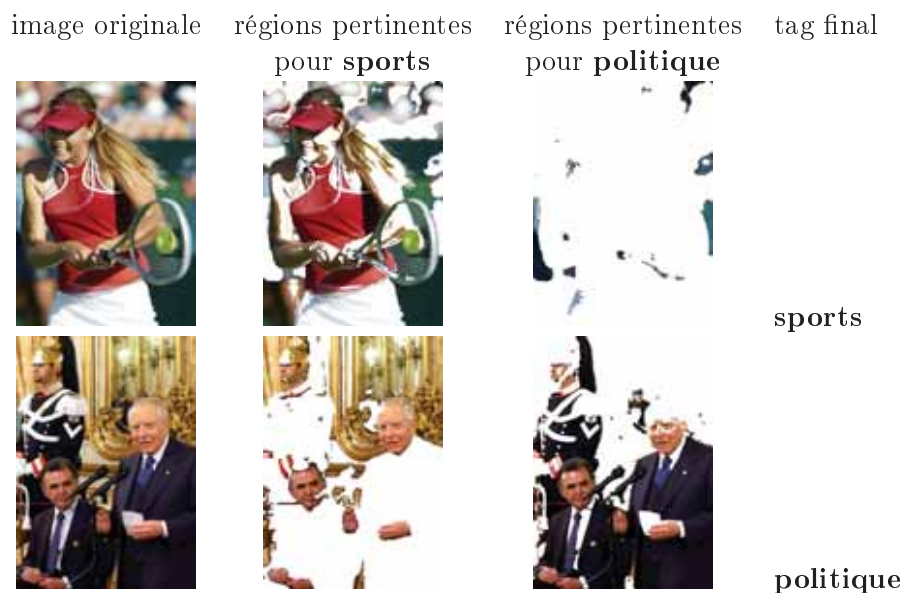


FIGURE 1.7 – Mots visuels pertinents pour reconnaître des scènes d'actualités : les images originales sont présentées dans la 1ère colonne, uniquement les mots visuels pertinents pour reconnaître les images de *sport* dans la 2e colonne, les mots visuels pertinents pour reconnaître les images de *politique* dans la 3e colonne, et la classe finalement attribuée dans la dernière colonne. Source : [CI-5]

première étape de **sélection de caractéristiques** vise à sélectionner les mots visuels pertinents pour prédire un type de scène donné. À cet effet, l'information mutuelle [27, 28] entre la classe à prédire et chaque classe de mot visuel est calculée. Les vecteurs d'occurrences sont alors réduits pour ne garder que les mots visuels les plus importants, c'est à dire les mots visuels très fréquents dans les scènes recherchées, ou inversement ceux qui sont rarement présents. Ensuite, la **classification** est assurée par une **SVM** à faible coût calculatoire, le **Kernel-adatron**. L'algo-

rithme Adatron a tout d’abord été introduit dans [29] comme une approche de type perceptron pour la classification linéaire. Une version à noyau a ensuite été proposée [30]. Concrètement, il procède par descente de gradient pour résoudre le problème d’optimisation de la marge (et donc de réduction du risque empirique) entre les deux classes de l’ensemble d’apprentissage, et est donc par une implémentation simple des SVMs [31]

Néanmoins, l’approche précédente considère les régions comme un simple BoW et ne prend pas en compte l’organisation des régions les unes par rapport aux autres. Nous avons donc proposé une deuxième approche qui inclue la structure spatiale des régions sous la forme d’un graphe de régions. L’étape de sélection de caractéristiques est à nouveau utilisée pour d’une part choisir les régions pertinentes et d’autre part élaguer le graphe afin de réduire la complexité des traitements. La classification repose ici sur un classifieur de type k -plus-proches-voisins avec une distance d’édition de graphe (*Graph-edit distance*) définie à dessein pour comparer les graphes de régions d’image [CI-8]. Finalement, les deux approches ont été combinées dans un système à classifieurs multiples (*Multiple Classifier System (MCS)*) qui parvient à tirer parti des complémentarités des deux représentations pour obtenir d’excellents taux de classification pour toutes les classes [CI-9].

Ces algorithmes ont notamment été utilisés pour la classification et la reconnaissance de scènes génériques [CI-6] comme illustré dans la figure 1.6. Notamment, la procédure de sélection de mots visuels pertinents permet de comprendre quelles régions de l’image ont contribué à la prise de décision et fournissent donc des éléments d’explicabilité de l’algorithme d’apprentissage machine. Ces classifieurs ont également été appliqués dans le domaine des bibliothèques numériques (*digital libraries*) [32] pour l’annotation automatique d’images [CI-5]. Pour la gestion d’archives d’agence de presse (l’ANSA italienne), ils servent notamment à taguer les images selon des catégories de sujets : sports, politique, etc. Des résultats avec les régions de l’image considérées pour prédire le type de scène sont présentés dans la figure 1.7.

1.2 Observation de scène et reconnaissance d’objets pour drones

Aperçu Plus récemment, mes travaux pour l’analyse de scène et la reconnaissance d’objets ont eu pour cadre la robotique. La problématique est alors de reconnaître des objets d’intérêt pour la mission du robot : base mobile d’atterrissage pour un drone, personnes, objets de l’environnement pour l’aide à la navigation. La caméra est embarquée sur un drone ou un robot terrestre, avec des visées variables allant de la vue d’oiseau à la vue frontale. En collaboration avec Martial Sanfourche (ONERA), j’ai proposé plusieurs approches pour la détection et la reconnaissance de véhicules [CI-11] ou de personnes [RN-1, CN-2, CI-19], ainsi que des approches génériques pour la classification interactive et générique de l’environnement [CI-16, CI-18].

Contexte Le contexte est ici défini par des scénarios d’intervention rapide en environnement inconnu. C’est le cas par exemple pour les missions de sauvetage et recherche (*Search-and-Rescue (SaR)*) qui font suite à une catastrophe, qu’elle soit d’origine naturelle ou humaine. Il est alors nécessaire de disposer d’une connaissance précise de l’environnement pour aider et guider les équipes d’intervention. Or, il n’existe pas toujours de carte préalable. De plus, dans tous les cas l’environnement a subi des modifications et les objets mobiles tels que les véhicules ou les personnes doivent être repérés dans l’instant. Par ailleurs, à l’époque de ces travaux, la



FIGURE 1.8 – Notre approche d'apprentissage interactif pour l'analyse de scène depuis un drone. (1) le drone collecte vidéos, 3D issue d'un LiDAR et positions GPS. (2) Une carte 3D de l'environnement est construite sur la station sol par ajustement de faisceaux global. (3) L'interprète construit des classificateurs *ad hoc* pour la vue d'ensemble de la scène. (4) Les classificateurs sont adaptés géométriquement au format de la caméra embarquée pour utilisation en vol, connaissant la position courante. Source : [CI-16]

reconstruction géométrique de l'environnement est rendue possible par de nombreuses approches de cartographie et localisation simultanées (*Self-Localization And Mapping (SLAM)*) [33, 34] y compris pour les drones [35]. Le défi est alors de combiner ces cartes d'occupation de l'espace avec des informations sémantiques de plus haut niveau [36, 37].

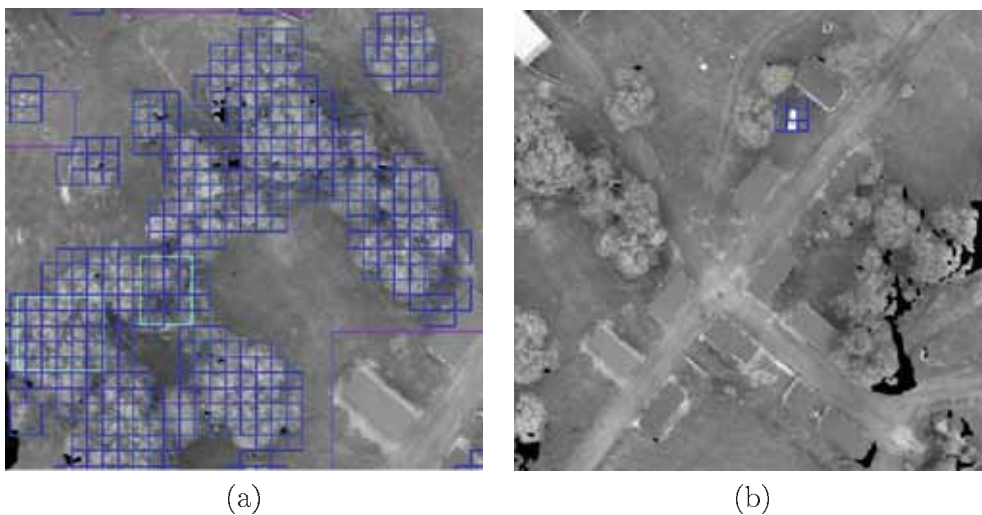


FIGURE 1.9 – Analyse de scène par drone : carte de classification d'environnement superposée à la vue d'ensemble ortho-rectifiée générée à partir des prises de vues du drone. (a) : classification d'arbres pour évitement d'obstacle à l'atterrissage ; (b) : classification de véhicules pour localisation de cibles. Source : [CI-16].

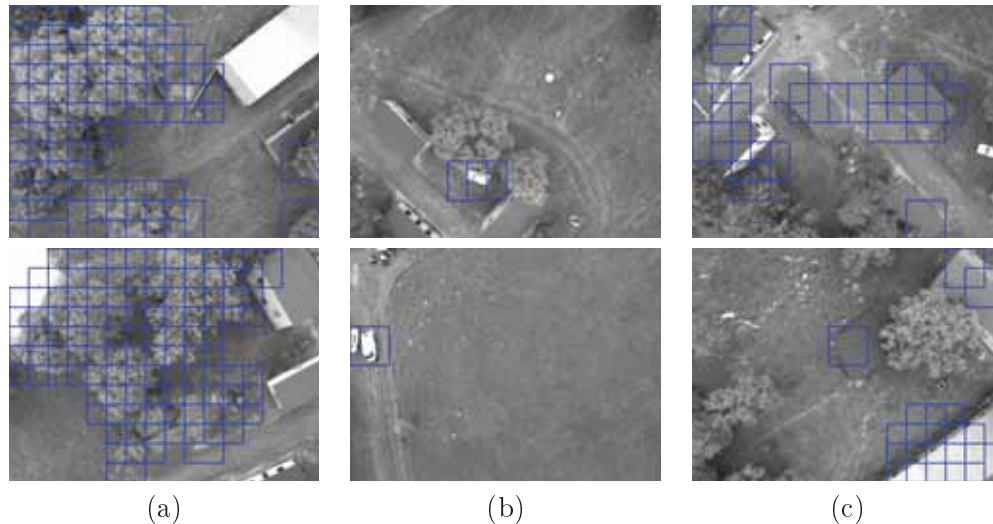


FIGURE 1.10 – Analyse de scène par drone : classification dans la vidéo. Les résultats de détection (carrés bleus) sont obtenus directement dans la vidéo de la caméra embarquée après transformation géométrique. (a) : classification d’arbres ; (b) : classification de véhicules ; (c) : classification de bâtiments. Source : [CI-16].

Approche d’analyse de scènes par drone La solution que nous avons proposée consistait à insérer l’homme dans la procédure en réalisant un **apprentissage interactif** de classes sémantiques d’intérêt. L’objectif est alors de produire des cartes sémantiques de la scène observée à partir des images acquises lors du vol du drone. Au contraire de [38, 39] qui utilisent des classifieurs pré-entraînés, notre approche permet de construire des classifieurs sémantiques adaptés à la situation à traiter [CI-16, CI-18].

Le processus complet est résumé dans la figure 1.8. Une carte 3D géométrique de l’environnement est construite sur la station de contrôle au sol à partir des données acquises en vol depuis le drone (vidéo, mesures laser et GPS). Une estimation fine de la trajectoire est réalisée par ajustement de faisceaux épars [40] afin d’agrèger les données LiDAR pour constituer un modèle 3D de l’environnement (Fig. 1.8-2). À partir de ce modèle, une vue d’ensemble de la scène est générée sous forme d’une image ortho-rectifiée. L’expert sélectionne ensuite des zones de l’image pour définir l’objet d’intérêt, et génère ainsi des patches-exemples. Ces patches sont indexés typiquement par des *Histogram of Oriented Gradients* (HOGs) [41] et des *Local Binary Patterns* (LBPs) [42] (reproduisant l’expérience de combinaison de ces descripteurs déjà prouvée efficace sur *Pascal VOC Challenge* (Pascal) [43]). Ils servent ensuite à entraîner en quelques itérations un classifieur par une procédure d’*Online Gradient-Boost* [44]. Gradient-Boost [45, 46] est choisi pour sa capacité à minimiser différentes fonctions objectifs, et notamment des fonctions qui accordent moins d’importance aux exemples mal annotés [47] [CI-13, CI-18]. En effet, de par l’aspect interactif et grossier de la création du jeu d’apprentissage, de nombreux exemples ne sont pas fiables (Fig. 1.8-3). Enfin le classifieur ainsi construit peut alors être adapté à la géométrie des images de la caméra. En effet, connaissant la position courante du drone grâce au GPS, l’homographie entre le plan image de la caméra et celui de la vue d’ensemble est calculable. Les classifieurs adaptés sont alors transférables au drone afin de le doter de la capacité à reconnaître cibles et obstacles apparaissant dans son propre champ visuel (Fig. 1.8-3).

Cette approche de cartographie interactive de l'environnement¹ a été appliquée à l'aide à la navigation des drones, afin de leur donner des capacités qui permettent de gagner en autonomie. En effet, l'analyse interactive de scène permet de doter un drone opérant sur une zone nouvelle de fonctionnalités utiles. Par exemple, la classification de zones d'arbres ou de bâtiments permet de faire de l'évitement d'obstacle, notamment lors de la recherche d'un lieu d'atterrissage. La détection de véhicules permet de trouver et localiser des cibles d'intérêt. Dans tous ces cas, l'approche proposée permet la classification de terrain selon le point de vue global, celui de l'opérateur et de l'équipe qui planifie la mission (voir figure 1.9), et selon le point de vue du drone, pour la détection en temps réel (voir figure 1.10). Ces travaux ont par la suite été complétés par des méthodes pour la cartographie et la localisation de personnes en détresse² pour l'aide aux équipes d'intervention dans un cadre SaR [RN-1, CN-2, CI-19]. Pour la localisation et reconnaissance de véhicules, et notamment afin de permettre le retour du drone à sa plate-forme mobile, nous avons également proposé une approche de détection et d'estimation de la pose d'objet 3D par *template matching* [CI-11]. Cette approche a été implémentée et utilisée pendant plusieurs années dans le drone ReSSAC de l'ONERA.

1.3 Segmentation sémantique image et 3D mono-vue

Aperçu Dans le cadre de la thèse de Joris Guerry [48] (dirigée par David Filliat - ENSTA ParisTech - et soutenue en 2017), l'objectif était la reconnaissance d'objet pour la robotique, et visait d'une part à tirer parti des données Rouge-Vert-Bleu-Profondeur (*Red - Green - Blue - Depth* (RGB-D)) souvent disponibles sur ce type de plates-formes, et d'autre part à permettre l'utilisation dans tous types de conditions environnementales. Pour le premier axe, nous avons d'abord proposé des algorithmes de fusion multimodale de réseaux convolutifs orientés objet, plus précisément des Réseau de Neurones Convolutif basé Régions – *Region Convolutional Neural Networks* (R-CNNs) [CI-33, CN-6]. Puis nous avons proposé des algorithmes de classifications de gestes [CI-29] et des réseaux de segmentation sémantique pour des robots terrestres [CI-34]. Pour le deuxième axe, nous avons proposé des réseaux de neurones capables de sélectionner des classificateurs pour l'adaptation de domaine [CN-5] et rendu publics un nouveau jeu de données avec des conditions (illuminations, artefacts, etc) très variables : ONERA.ROOM [CI-33].

Contexte Ces travaux sont à nouveau définis par des scénarios d'exploration robotique. Mais à l'inverse de la partie 1.2, il s'agit plutôt de robots terrestres avec une visée frontale, qui produisent donc des images standard similaires à celles qui étaient analysées en partie 1.1. Cependant les approches ont évolué entre temps : l'état de l'art en classification d'image est dorénavant obtenu avec des réseaux de neurones convolutifs (Réseau Neuronal Convolutif – *Convolutional Neural Networks* (CNNs)) [50], la détection d'objet avec des R-CNNs [51, 52, 53], et la segmentation sémantique (c'est à dire la classification au niveau pixel) avec des réseaux de neurones entièrement convolutifs (Réseau Entièrement Convolutif – *Fully Convolutional Networks* (FCNs)) [54, 55, 56]. Par ailleurs, le contexte robotique apporte également des spécificités. Avec l'apparition des capteurs à lumière structurée à bon marché, telle la Microsoft

1. Cartographie interactive de l'environnement : <https://www.youtube.com/watch?v=0TxaLcou0HE>

2. Cartographie et localisation de personnes en détresse (projet FP7 Darius) : <https://www.youtube.com/watch?v=JyHaeBkvKTQ>

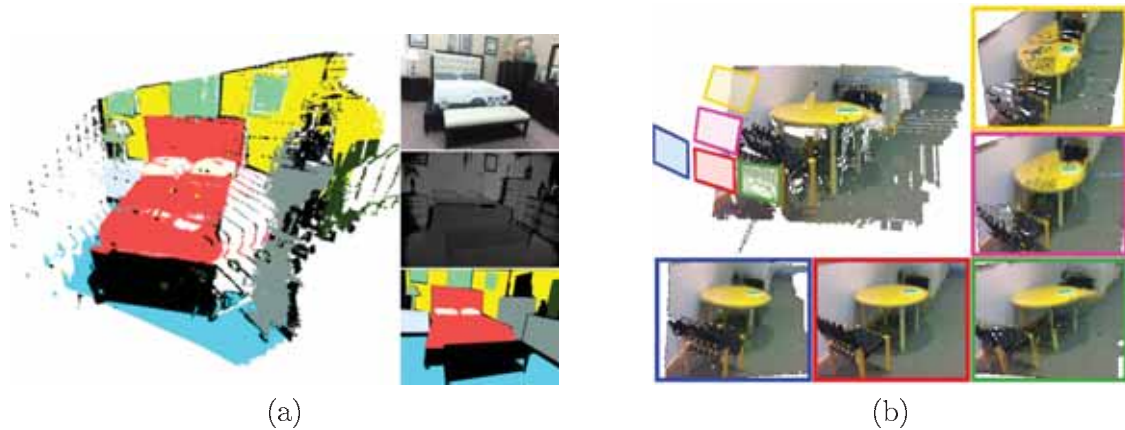


FIGURE 1.11 – [Augmentation de données cohérente avec la 3D pour données RGB-D issues de la base SUNRGBD [49]. (a) Données RGBD monovue comparées aux données 3D équivalentes. (b) Échantillonnage des points de vue pour l'augmentation de données cohérente avec la 3D. Source : [CI-34].

KinectTM, les robots sont souvent équipés de capteurs RGB-D [57] qui fournissent une information cruciale pour les tâches précédentes. Le type d'application est toujours guidé par des besoins en sauvetage et recherche, mais cette fois ci à l'intérieur des bâtiments. Il s'agit alors de reconnaître l'environnement pour détecter des personnes restées dans un bâtiment évacué par exemple.

Approche pour la segmentation sémantique RGB-D La plupart des approches de l'état de l'art pour le traitement des données RGB-D et notamment la **segmentation sémantique** utilisent des évolutions des FCNs. Notamment, Gupta *et al.* [58] proposèrent une méthode de détection d'objet fondée sur les R-CNN [52] avec un encodage de la profondeur et une sortie de classification dense. L'architecture FuseNet [59] réalise une intégration progressive et multiéchelle de l'information de profondeur dans un FCN encodeur-décodeur tel que SegNet [55] ou U-Net [56]. Pour tirer parti du contexte robotique et du fait que le robot évolue dans son environnement, MVCNet (*MultiView Consistent network*) [60] estime la trajectoire courante et utilise les projections des vues acquises avant et après l'image à classifier pour rendre plus robuste la prédiction.

Le changement de point de vue pour gagner de l'information est l'idée de base de notre approche. Cependant, à la différence des méthodes précédentes, nous l'avons utilisée pour effectuer de l'**augmentation de données cohérente en 3D** à l'apprentissage pour améliorer les prédictions avec une seule image RGB-D en conditions de test. En effet, l'information contenue dans un couple image et profondeur est plus facilement interprétable en 3D, comme le montre la figure 1.11-a, même avec une seule vue. Connaissant l'information de profondeur, en utilisant les paramètres caméra et des hypothèses simples de projection, il est alors possible de générer des vues virtuelles qui représentent la scène sous différents points de vue légèrement différents et donc de réaliser l'augmentation de données [61, 62]. La figure 1.11-b montre les différents positions de caméra virtuelles choisies, avec plusieurs décalages de 10° par rapport à la position réelle.

L'algorithme *Snapshot Network for Robotics* (SnapNet-R) [CI-34] est détaillé sur le gra-

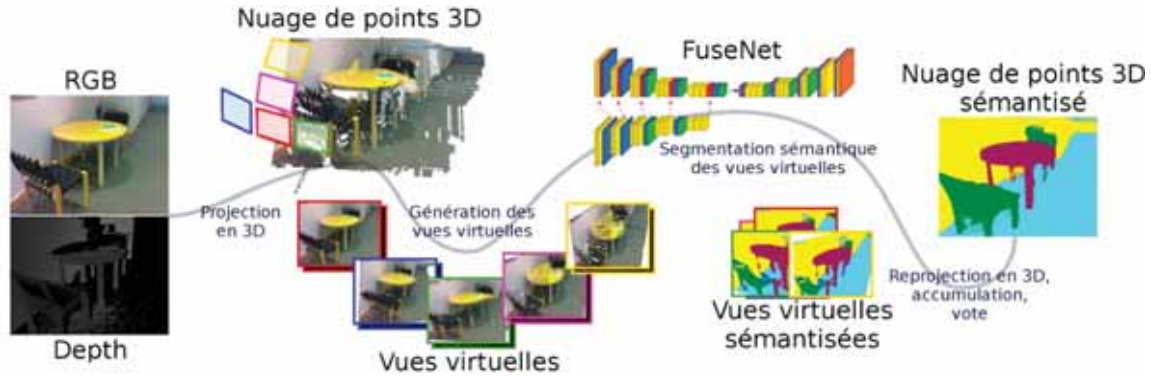


FIGURE 1.12 – Résumé graphique de l’algorithme `SnapNet-R(obot)` pour la segmentation sémantique de données `RGB-D` par augmentation de données cohérentes en 3D. Source : [48].

phique 1.12. Le processus est le même en apprentissage et en test. À partir de la donnée `RGB-D`, des vues virtuelles (image et profondeur) sont synthétisées selon les points de vue explicites en figure 1.11-b. Un réseau de type FuseNet [59] est entraîné ou utilisé en inférence lors du test, pour produire des vues sémantisées de la scène. Enfin, les cartes de segmentations sont reprojétées dans le référentiel de la vue initiale pour obtenir la classification finale par accumulation et vote majoritaire.

Tableau 1.1 – Performances de `SnapNet-R` en segmentation sémantique sur `SUNRGBD` [49]. Nous avons reproduit les résultats du FuseNet SF5 dans l’expérience (*rep*) que nous pouvons comparer à notre stratégie multivue. P = taux de bonne classification global ; PM = taux de bonne classification par classe moyen ; mIoU = Intersection-sur-Union par classe moyenne. Pour chaque métrique, la meilleure valeur est en gras et la deuxième meilleure valeur est en italique. Source : [CI-34].

Approche	P	PM	mIoU
LSTM-CF [63] (RGB)	–	48.1	–
FCN 8s [54] (RGB)	68.2	38.4	27.4
Bayesian SegNet [64] (RGB)	71.2	45.9	30.7
Context-CRF [65] (RGB-D)	78.4	<i>53.4</i>	42.3
*FuseNet SF5[59] (RGB-D)	76.3	48.3	37.3
DFCN-DCRF [66] (RGB-D)	76.6	50.6	39.3
3D GNN [67] (RGB-D)	–	52.5	<i>40.2</i>
FuseNet SF5 (<i>rep.</i>)	77.21	54.81	39.11
<code>SnapNet-R</code> [CI-34]	<i>78.0</i>	58.1	39.6

* Calculés à basse résolution (224×224) comme dans [59] au contraire des autres méthodes dont les résultats sont calculés à la résolution d’origine.

Notre approche `SnapNet-R` a été évaluée sur les jeux de données de référence *Scene Understanding RGB-D* (`SUNRGBD`) [49] et *New-York University dataset version 2* (`NYUv2`) [68]. La segmentation sémantique de `SUNRGBD` est une tâche difficile. En effet, comme le montre

Tableau 1.2 – Performances de **SnapNet-R** en segmentation sémantique sur **NYUv2** [68]. Pour chaque métrique, la meilleure valeur est en gras, la deuxième meilleure valeur est en italiques. Source : [CI-34].

Approche	P	PM	mIoU
NYUv2 40 classes			
RCNN [51] (RGB-HHA)	60.3	35.1	28.6
FCN 16s [54] (RGB-HHA)	65.4	46.1	34.0
Eigen et al.[69](RGB-D-N)	65.6	45.1	34.1
Context-CRF [65] (RGB-D)	67.6	49.6	37.1
*FuseNet SF3[60] (RGB-D)	66.4	44.2	34.0
*MVCNet-MP [60](RGB-D)	70.66	51.78	40.07
3D GNN [67] (RGB-D)	-	<i>54.0</i>	<i>39.9</i>
FuseNet SF5 (RGB-D)	62.19	48.28	31.01
SnapNet-R (RGB-D)	<i>69.20</i>	60.55	38.33
NYUv2 13 classes			
Coupric et al.[70] (RGB-D)	52.4	36.2	–
Hermans et al.[71] (RGB-D)	54.2	48.0	–
SceneNet (DHA)[72] (DHA)	67.2	52.5	–
Eigen et al.[69] (RGB-D-N)	75.4	66.9	52.6
*FuseNet SF3 [60] (RGB-D)	75.8	66.2	54.2
*MVCNet-MP [60] (RGB-D)	<i>79.13</i>	70.59	<i>59.07</i>
FuseNet SF5 (RGB-D)	78.41	<i>72.07</i>	56.33
SnapNet-R (RGB-D)	81.95	77.51	61.78

* Calculés à basse résolution (320x240) au contraire des autres méthodes dont les résultats sont calculés à la résolution d’origine.

le tableau 1.1, la meilleure méthode précédente était Context-CRF [65] qui n’atteignait que les 42,3% de *mean Intersection-over-Union* (mIoU) et 53.4% de précision moyenne. Pour cela ils mettaient en oeuvre une architecture très profonde couplée avec un CRF Dense [73]. Les réseaux de neurones pour graphes 3D (3D GNN) [67] publiés simultanément à notre approche affichaient 40.2% de mIoU et 52.5% de précision moyenne, tandis que **SnapNet-R** obtenaient une valeur équivalente de 39.6 de mIoU et une performance de 58.1% en précision moyenne, supérieure de 4 points à Context-CRF. Les résultats sur le jeu de données **NYUv2** sont compilés dans le tableau 1.2. Là encore, **SnapNet-R** était dans le trio de tête des approches de l’état de l’art, obtenant notamment la meilleure valeur en taux de classification moyen par classe, et les meilleures performances quelque soient les métriques sur la taxonomie à seulement 13 classes, c’est à dire sans classification à grain fin. Pour comparaison, l’approche concurrente MVCNet-MP [60] utilise plusieurs images successives du flux vidéo et estime leur pose respective pour rendre son estimation plus robuste, au contraire de notre approche qui n’utilise qu’une seule image. Enfin, la figure 1.13 permet de comparer quelques résultats de segmentation sur **NYUv2**. De manière générale, les formes des objets sont plus précises et complètes, notamment pour les objets fins qui sont mieux discriminés sous différents points de vue.

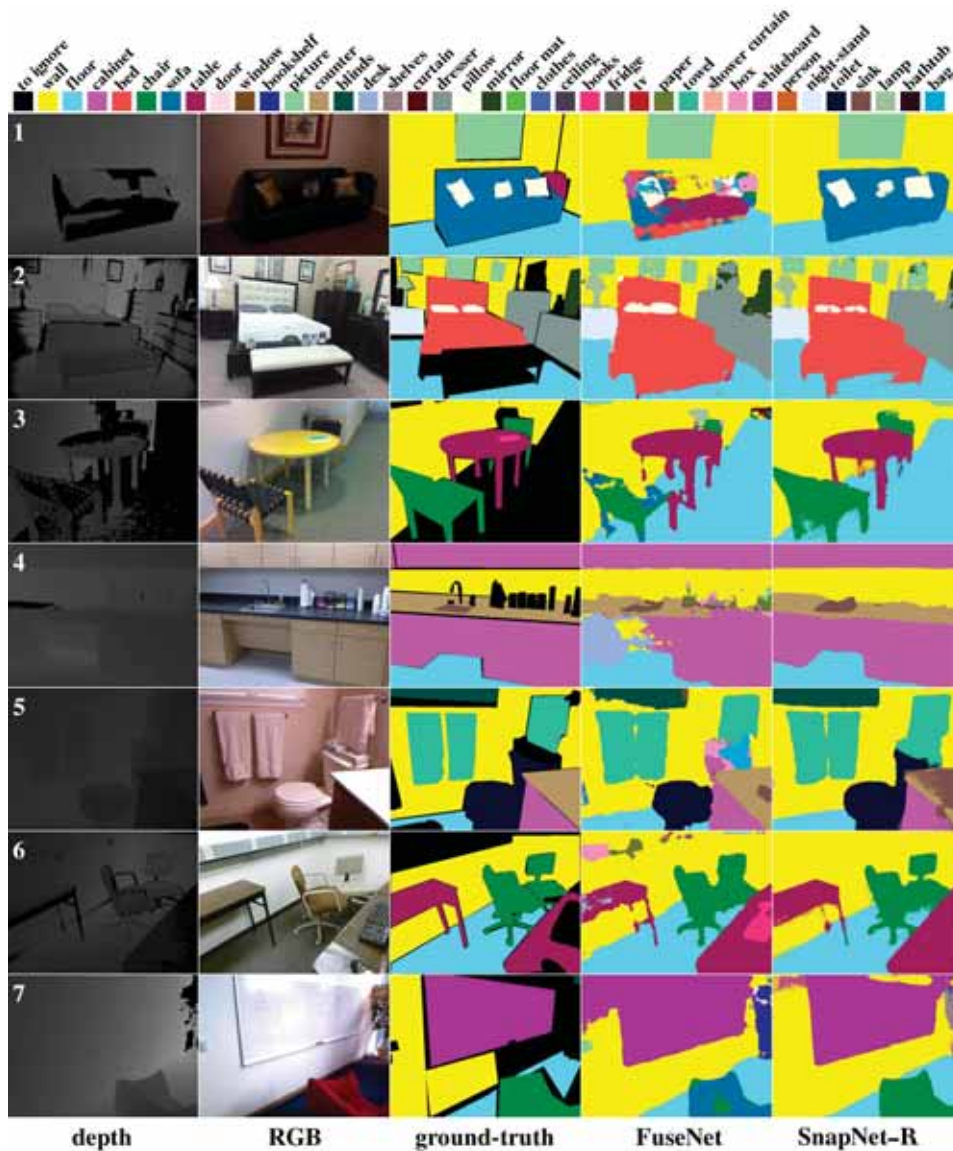


FIGURE 1.13 – Résultats de *SnapNet-R* pour la segmentation sémantique sur des images de la base *SUNRGBD*. Source : [CI-34].

Approche pour la détection d'objet en RGB-D Par ailleurs nous avons proposé plusieurs approches de détection d'objet sur données RGB-D [CI-33, CN-6]. L'architecture reprend la structure d'un R-CNN [52] en multicanal, c'est à dire une branche pour l'image *Red - Green - Blue* (RGB) et une branche pour la profondeur (D). Cependant, les régions détectées par l'une ou l'autre branche sont toutes proposées au classifieur d'objets, ce qui permet de lever des ambiguïtés. Ces approches ont établi un nouvel état de l'art sur les jeux de données du domaine, et nous avons par ailleurs proposé à la communauté un nouveau jeu de données plus grand, en intérieur et plus varié en termes de variation d'illumination : *ONERA.ROOM*³ Des résultats sur des scènes difficiles (avec flou de bougé, contre-jour, occlusion, etc) sont présentés

3. *ONERA.ROOM* est disponible en téléchargement : <http://jorisguerry.fr/onera-room/>.



FIGURE 1.14 – Exemples de prédictions de nos méthodes multimodales pour la détection d’objets RGB-D. Les modèles sont entraînés sur *InOutdoor RGBD People* [74] et testés sur *InOutdoor RGBD People* et *ONERA.ROOM*.

sur la figure 1.14⁴.

1.4 Résumé et conclusions

Les travaux présentés dans ce chapitre visent à la classification et la détection d’objets dans des images de scènes. En partie 1.1 sont présentées des approches pour la classification d’images supervisée (par SVMs) et non-supervisée (par *clustering ARC*). Respectivement, elles permettent la reconnaissance automatique de scènes et la navigation dans des bases d’images. Puis, nous avons présenté en partie 1.2 des approches de classification de scènes vues depuis un drone par apprentissage interactif fondé sur des techniques de *boosting* incrémental. Enfin, des approches de segmentation sémantique de scène et de détection d’objets par réseaux convo-

4. Également en vidéo : <https://www.youtube.com/watch?v=jEHyG2BSnGc>

lutifs ont été introduites en partie 1.3. Notamment, ces dernières proposent des modèles pour l'apprentissage à partir de données RGB-D dans le contexte de la robotique.

L'évolution majeure dans la manière de traiter le problème, au-delà des techniques employées, est dans la progression vers une prise en compte implicite de l'aspect 3D de la scène observée pour mieux la comprendre et lui associer des éléments sémantiques. Cette tendance sera poursuivie au chapitre 3 et ouvre les perspectives sur la compréhension de scènes 3D du chapitre 4

Chapitre 2

Télédétection

Je me suis intéressé aux problématiques de télédétection et d'analyse des données d'observation de la Terre lorsque j'ai rejoint l'ONERA. L'objectif de la compréhension de scènes est alors de *sémantiser* ces données, c'est à dire reconnaître des objets (véhicules, bâtiments, etc), classifier les différentes zones (végétation, routes, etc), ou en mesurer l'évolution.

Le fil conducteur de ces travaux est le développement de modèles d'apprentissage automatique pour effectuer cette sémantisation. Les différentes approches présentées en parties 2.1 et 2.2 montrent l'évolution vers des méthodes de plus en plus performantes pour effectuer cette tâche. Étant donné la **Très Haute Résolution (THR)** des images satellitaires ($\sim 50\text{cm}/\text{pixel}$) ou aériennes ($\sim 10\text{cm}/\text{pixel}$) disponibles depuis une dizaine d'années, les modèles développés ici sont très liés à ceux présentés précédemment (chapitre 1). Néanmoins, le changement drastique de point de vue introduit de nombreuses spécificités : apparence inhabituelle, absence de perspective, invariance en rotation, etc. En termes de vision par ordinateur, des a priori spécifiques guident alors la conception des algorithmes. Les approches présentées dans ce chapitre offrent également des réponses à des problèmes spécifiques de la télédétection, tels que l'analyse d'images issues de capteurs particuliers (multispectraux, hyperspectraux, *Synthetic Aperture Radar (SAR)*, etc.), la fusion de données multimodales, et l'analyse de données multitemporelles.

Par ailleurs, les données de télédétection sont particulièrement volumineuses (issues par exemple de satellites avec des délais de revisite de quelques jours), et constituées d'images de très grande taille. Cela impose de prendre en compte l'aspect calculatoire des méthodes de classification mises en oeuvre. En revanche, si les données images sont nombreuses, il y a une disponibilité bien moindre de données de référence qui permettraient d'entraîner des modèles statistiques. Mes travaux dans ce domaine ont donc été structurés selon deux axes principaux. (1) Concevoir des modèles et approches qui permettent de remédier au manque d'exemples (notamment par apprentissage interactif) et d'entraîner puis classifier rapidement (données massives). (2) Construire et diffuser des jeux de données permettant d'entraîner des modèles efficacement, et ainsi contribuer au développement de nouvelles recherches dans le domaine de la télédétection et l'interprétation d'images d'observation de la Terre.

Le reste de ce chapitre est organisé comme suit. La partie 2.1 présente différentes approches d'apprentissage automatique pour construire des modèles efficaces de classification et détection d'objet. La partie 2.2 introduit plusieurs méthodes de réseaux de neurones profonds pour l'observation de la Terre, avec des applications en cartographie automatique et détection de changement notamment. Enfin, la partie 2.3 liste les efforts faits pour constituer et diffuser

des jeux de données à même de structurer la communauté de la télédétection autour de tâches spécifiques et de compétitions d'évaluation de méthodes.

2.1 Apprentissage automatique pour l'observation de la Terre

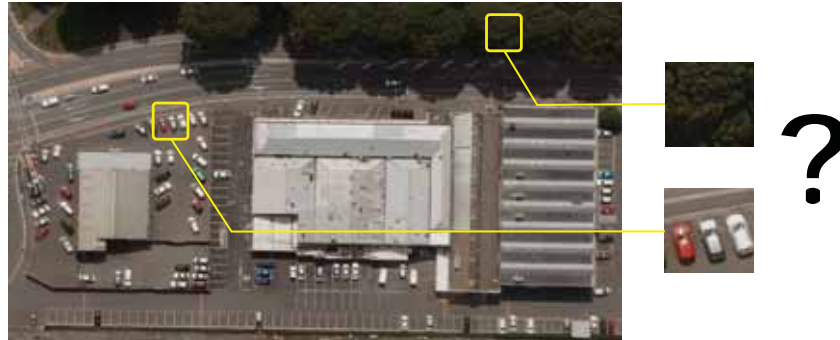


FIGURE 2.1 – Objectif des modèles discriminants en télédétection : apprendre $p(\text{classe}|\text{position}) = f(\text{patch})$.

Tableau 2.1 – Co-évolution des données et des méthodes de classification en télédétection, ou comment les algorithmes tirent parti de la résolution de plus en plus fine et de la quantité croissante de données annotées.

<i>Évolution données</i>	<i>Description</i>	<i>Classification</i>
resol. –	Filtrage niveau pixel	Modélisation expert, seuillage
resol. +	Textures	
data +	Descripteurs locaux	Estimateurs de distribution (<i>GMM</i> , etc.)
resol. ++	Descripteurs complexes ; Modélisation objet	Classifieurs par apprentissage (<i>SVMs</i> , méthodes ensemblistes : forêts aléatoires, boosting)
data ++	Patches de pixels ; Filtres	<i>active learning</i> , <i>latentSVM</i>
		Réseaux de neurones profonds (<i>RBM</i> , <i>CNNs</i>)

Aperçu En télédétection, l'objectif est d'être capable d'associer de la sémantique à une position (géolocalisation) ou une région pour laquelle on dispose de données image. Aujourd'hui, il est raisonnable de supposer un continuum dans les images de la Terre, chaque image acquise étant complétée par tuilage avec une autre acquisition. La question de la classification d'image soulevée au chapitre 1 se reformule en classification de *patch* local extrait autour d'une position (voir figure 2.1). Nous avons cherché à construire des **modèles discriminants** pour classer ces patches, c'est à dire des modèles capables d'estimer $p(\text{classe}|\text{position}) = f(\text{patch})$.

J'ai notamment proposé des approches à base de boosting ou de SVMs pour l'**apprentissage interactif** pour la télédétection, où l'utilisateur vient spécifier sa cible d'intérêt : activité humaine dans des images optiques à moyenne et Très Haute Résolution [CI-13, CI-18] ou détection de changements dans des images SAR [CI-14]. Dans le cadre de la thèse d'Hicham Randrianarivo (dirigée par Michel Curcianu, co-encadrée par Marin Ferecatu, tous deux CNAM ParisTech, et soutenue en 2016), nous nous sommes intéressés à l'**apprentissage contextuel** pour la détection d'objets par des modèles à parties déformables (*Deformable Part-Models* (DPMs)) en imagerie RGB ou multispectrale [CI-15, CI-17]. Ces travaux ont conduit à proposer une méthode par mélange de modèles et entraînement par *hard negative mining* pour la détection de véhicules et d'objets [CN-3, CI-21].

Contexte de la classification en télédétection En télédétection, la compréhension de scène revient souvent à une classification de l'occupation et l'utilisation du sol. Le tableau 2.1 vise à résumer l'évolution des approches de classification en télédétection ces vingt dernières années, en relation avec la résolution de plus en plus fine des images et la disponibilité de plus en plus grande de données de référence associées à ces images. Nous distinguons plusieurs familles de méthodes selon leur niveau de complexité tant en caractérisation de l'information qu'en classification. (i) Au niveau pixel, peu d'information est disponible, et les approches modélisent des statistiques de luminance pour la détection de véhicules [75]. Le spectre d'une image hyperspectrale est plus signifiant et les distributions de différentes classes peuvent être modélisées, par exemple avec des *Gaussian Mixture Models* (GMMs) [76] ou des SVMs [77]. Le voisinage local est plus riche en information et peut également être caractérisé par divers descripteurs, par exemple par des graphes [78] ou des textures extraites via des *Markov Random Fields* (MRFs) [79, 80, 81]. (ii) La complexité des descripteurs s'est ensuite accrue, en confiant au pouvoir de discrimination de méthodes avancées d'apprentissage le soin de séparer les différentes classes. Par exemple, des combinaisons complexes de caractéristiques (moments géométriques, transformées de Fourier, détecteurs de lignes, etc.) sont utilisées avec des SVMs dans [82]. Les HOGs [41], parfois combinés avec les LBPs [42], se sont imposés comme le descripteur dense générique en imagerie optique, utilisés pour la détection de véhicules [83] ou la classification [84]. Des descripteurs de bords et de textures tels que les filtres de Haar ou de Gabor [85] sont également très utilisés en imagerie optique [86, 87], et connaissent depuis un grand succès en SAR également [88, 89]. En hyperspectral, les approches spatiales-spectrales se sont développées pour inclure l'information de voisinage [90, 91, 92, 93]. (iii) Avec la THR spatiale (inférieure à $\sim 1\text{m}/\text{pixel}$), plusieurs approches de modélisation objet sont apparues : *Scale-Invariant Feature Transforms* (SIFTs) [94] et appariement de graphe [95], modèle hiérarchique [96], etc. Des revues complètes et exhaustives de la littérature sont disponibles tant en optique [97, 98] que pour les données hyperspectrales [99, 100] ou pour le SAR [101, 102, 103, 104]. Une introduction à la classification multimodale peut être trouvée dans [105, 106, 107]

Une première limite de l'évolution des modèles présentés est la prise en compte des spécificités du domaine de la la télédétection. De fait, les volumes d'images de télédétection, notamment en THR, sont particulièrement importants, et dans de nombreux cas d'usage nécessitent un traitement rapide : analyse de site inconnu ou chamboulé suite à une catastrophe, recueil d'informations afin de préparer des interventions humaines. Cela nécessite donc des traitements rapides. De plus, il y a peu d'annotations disponibles pour des applications génériques, et encore moins ou pas d'annotations dans des applications qui visent à l'analyse d'un site nouveau ou la recherche d'une nouvelle classe. Le domaine requiert donc dans la plupart des cas des

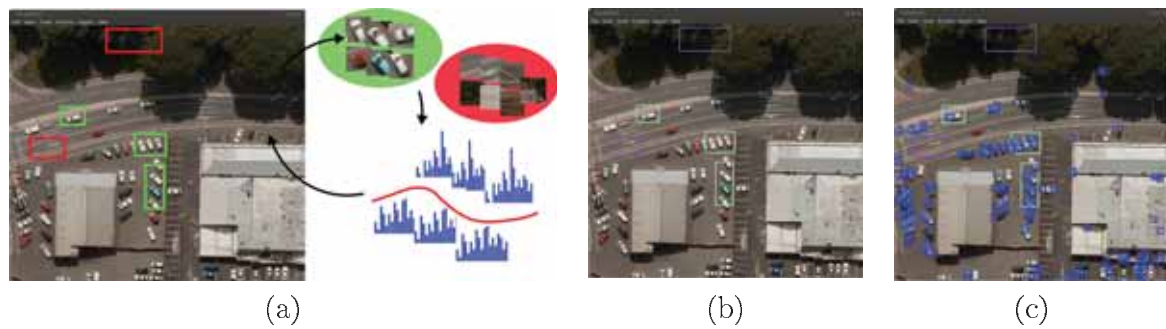


FIGURE 2.2 – Classification interactive d’objets en télédétection. (a) Boucle interactive : sélection d’exemples dans l’image via une interface ergonomique ; extraction de patches et indexation (par HOG-LBP) ; classification des patches de la ou des images entières. Exemple : (b) sélection de zones contenant l’objet recherché (ici, véhicules) ; (c) classification (carrés bleus) des patches de l’image contenant des véhicules. Source : [CI-18].

modèles qui peuvent être entraînés avec peu d’exemples en temps contraint, afin de traiter rapidement de grands volumes de données, tout en étant performants. Or, ce n’est pas le cas de toutes les approches présentées précédemment. Mes travaux ont donc visé à construire des modèles performants pour la THR actuelle, en tirant avantage d’a priori forts sur les objets et classes d’intérêt pour compenser le peu d’exemples. Ces a priori sont fournis par apprentissage interactif ou en imposant une structure forte sur le modèle.

Approches pour l’apprentissage rapide et interactif Nous avons proposé une approche pour construire des classifieurs de manière interactive, et donc coller tant au contexte géographique de la zone à classer qu’à la nature de la cible d’intérêt. La figure 2.2-a présente le processus. Dans une interface de *Système d’Information Géographique (SIG)*, l’utilisateur sélectionne des zones d’exemples, dont sont extraites des patches (plusieurs patches par zone) qui sont indexés par des descripteurs visuels. Dans [CI-13], nous avons identifié la combinaison HOG et LBP comme le meilleur index, généralisant ainsi au domaine de la télédétection le résultat de [43] pour les images multimédias. Nous avons proposé plusieurs approches d’apprentissage [CI-18], soit par SVM, sur le même principe que [CI-7], soit par *online boosting*. Cette dernière approche présente l’intérêt d’être incrémentale et de construire le classifieur au fur et à mesure des itérations sans ré-entraînement complet. De plus nous avons proposé des fonctions de coût peu sensibles les erreurs d’étiquetage : en effet, les entrées de l’utilisateur peuvent être approximatives [CI-18].

Ces approches ont été appliquées à la détection de bâtiments et structures artificielles dans des images satellitaires optiques ($\sim 70\text{cm/pixel}$) [CI-13], à la détection de véhicules dans des images aériennes optiques ($\sim 10\text{cm/pixel}$) [CI-18] (voir figure 2.2-b et c) et à la détection de changements dans des images SAR ($\sim 50\text{cm/pixel}$) [CI-14]. Dans ce dernier cas, la sélection est effectuée en comparant les images originales, mais l’extraction de caractéristiques est effectuée sur la carte de changement statistique calculée à partir de ces images (voir figure 2.3). Cet apprentissage interactif est également utilisé pour construire à la volée les classifieurs pour la vision des drones ($\sim 20\text{cm/pixel}$) en partie 1.2 au chapitre 1. Enfin, toujours pour remédier au manque d’exemples, nous avons également proposé des implémentations sur *Graphics Processing Units (GPUs)* de SVMs et notamment de *one-class SVMs* pour la détection d’anomalies.

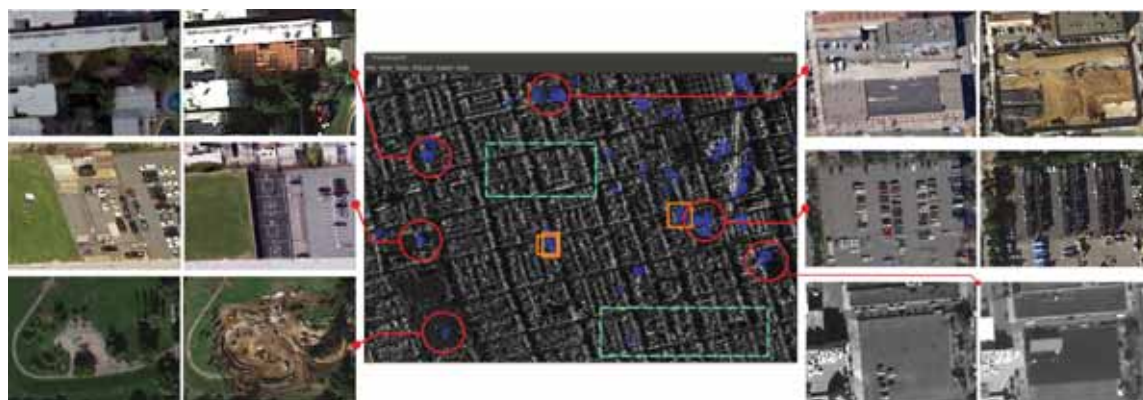


FIGURE 2.3 – Détection interactive de changements en télédétection. Les zones exemples de changements (rectangles en orange) et de non-changements (rectangles en vert) permettent d’entraîner l’algorithme d’*online gradient-boosting* qui retourne les changements détectés (carrés en bleu). Une inspection des sites identifiés en imagerie optique THR multodate confirme que les changements : chantiers, nouvelles constructions, installation de terrains de sports ou de panneaux solaires. Source : [CI-18, CI-14].

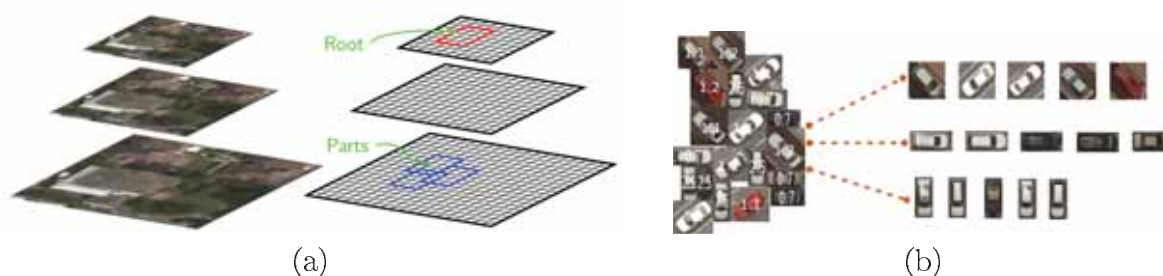


FIGURE 2.4 – Modèles à parties déformables (DPM) en télédétection : (a) modèle multiéchelle avec représentation globale (*racine*) et locale par parties ; (b) catégorisation des exemples selon l’apparence présente dans les DPM et reprise dans les DtMM afin de créer un mélange de modèles. Sources : [CI-17, CI-21].

Approches pour l’apprentissage contextuel Dans le cadre de la thèse d’Hicham Randriarivo [108] nous avons tout d’abord cherché à exploiter la **Très Haute Résolution** des images aériennes (5 à 20cm/pixel) pour proposer une modélisation fine des objets d’intérêt et de leur voisinage. Nous avons proposé d’utiliser les Modèles Déformables à Parties (DPMs) [109] en télédétection (voir figure 2.4). Les modèles déformables avaient ceci dit déjà été utilisés en aérien [110]. Ces modèles représentent les objets comme des graphes multiéchelles de parties locales décrites par des HOGs avec des coefficients de positionnement relatif. Ils s’inspirent des pyramides spatiales [111], des *pictorial structures* [112]. Par rapport à nos propres travaux, ils remettaient au goût du jour les graphes de descripteurs de régions que nous avons proposé dans [CI-8, CI-9]. Pour valider cette approche, nous avons d’ailleurs proposé un jeu d’annotations objet sur des images de l’Agence de Cartographie de Nouvelle-Zélande - *New-Zealand Aerial Mapping* (NZAM) sur Christchurch [113]¹. Les DPMs en observation de la

1. Le jeu de données est redistribué sous license libre à l’URL : <https://blesaux.github.io/data/>

Terre ont montré leur capacité à modéliser des objets urbains (bâtiments, arbres, véhicules) dans des milieux très riches en information, et ce malgré de grandes variations de taille et d'apparence [CI-15]. Nous avons également étendu les DPMs pour les données multimodales [CI-17] (hyperspectral et RGB THR) et montré que la fusion de sources permet de mieux caractériser les objets et notamment de détecter les véhicules.



FIGURE 2.5 – Détection d’objet en milieu urbain par DtMM sur les jeux de données (a) NZAM de Christchurch et (b) DFC2015 sur Zeebrugge. Les véhicules selon différentes orientations sont retrouvés par des composantes différentes du mélange de modèles estimé par un GMM. Source : [CI-21].

Le concept même d’objet est problématique en imagerie aérienne, car les classes les plus visibles n’en sont pas toujours : les routes ne sont jamais déconnectées, les bâtiments sont de formes très variables et contigus dans les centres-villes, les arbres sont groupés, etc. Nous verrons dans la partie 2.2 suivante que leur reconnaissance est plus adéquatement réalisée par segmentation sémantique. Nous nous sommes donc concentrés sur la détection de véhicules. La suite de ce travail a consisté à déconstruire les DPM pour concevoir des détecteurs d’objets rapides et des stratégies d’apprentissage à même de les entraîner avec peu d’exemples. La perspective réduite des images vues du ciel (tout au moins pour les images en visée vers le nadir) permet par exemple de s’affranchir de l’espace de recherche multiéchelle des DPMs. La taille des véhicules ne requiert plus de modèles à parties. En revanche, nous avons mis en évidence que plusieurs des mécanismes moins connus des DPMs étaient pertinents pour cette application. Nos mélanges de modèles entraînés de manière discriminative (DtMMs) [CI-21, CN-3] combinent une catégorisation des exemples par GMM qui conduit en pratique à séparer les objets avec des orientations différentes (voir figure 2.4-b), puis un entraînement de SVMs linéaires sur des HOGs selon une procédure itérative de recherche d’exemples négatifs (*hard-negative mining*). En détection, l’image entière est indexée et le problème est vu comme un filtrage (*template matching*) dans l’espace des HOGs. Il en résulte des cartes d’activation obtenues par corrélation. Les activations des différents modèles sont ensuite fusionnées par *Non-Maximum Suppression* (NMS). Des résultats en environnement anglo-saxon (plan en damier) et continental européen sont présentés en figure 2.5. Dans [CI-22], nous utilisons à nouveau un graphe des régions voisines pour modéliser la structure du contexte et régulariser les détections. Il est notable que ces méthodes orientées objet obtiennent de meilleures performances que les approches standards de l’état de l’art (voir supra) dans le comparatif compréhensif de [CI-20,

RI-3], et ne seront dépassées que par les approches par réseaux de neurones décrites dans la partie 2.2 suivante.

2.2 Apprentissage profond pour l'observation de la Terre

Aperçu Je travaille depuis plusieurs années maintenant au développement de **réseaux de neurones pour la télédétection**. Nous avons établi un état de l'art des approches de classification d'image au niveau pixel qui a montré d'une part la supériorité des réseaux convolutifs et d'autre part la possibilité de les entraîner avec des jeux de données de taille raisonnable pour le domaine [CI-20, RI-3]. Ces résultats, qui coïncident avec ceux de la vision par ordinateur, nécessitent néanmoins un apprentissage beaucoup plus long que les travaux présentés dans la partie 2.1. Puis dans la thèse de Nicolas Audebert [114] (dirigée par Sébastien Lefèvre, Université de Bretagne-Sud, et soutenue en 2018), nous avons proposé des architectures de réseaux multimodales (imagerie optique et modèle d'élévation de terrain par exemple) et notamment la correction résiduelle [CI-25, RI-11]. Sur le plan applicatif, ces réseaux se sont montrés particulièrement performants pour la segmentation sémantique (classification au niveau pixel) et même la détection d'objets : approche *Segment-before-detect* primée par l'*Int. Society of Photogrammetry and Remote Sensing* (ISPRS) [CI-24, RI-7]. Ces travaux évoluent maintenant vers l'utilisation de sources ouvertes (telles qu'*OpenStreetMap* (OSM)) pour inclure un a priori qui guide la classification [CI-32] et l'extension aux capteurs hyperspectraux [RI-16]. Dans la thèse de Rodrigo Daudt (en cours, dirigée par Yann Gousseau à Télécoms Paris, et co-encadrée par Alexandre Boulch, ONERA), nous nous intéressons à la détection de changements et à l'analyse d'activité dans des séries d'images multitemporelles [CN-9, CI-38, CI-42]. Les premiers résultats montrent la capacité de développer des réseaux convolutifs à détecter des changements ciblés (urbains par exemple) et ainsi effectuer une réelle détection de changements sémantiques.

Contexte de l'apprentissage profond en télédétection Avant l'attrait suscité par l'apprentissage profond en vision par ordinateur [115], les réseaux de neurones convolutifs (CNNs) avaient déjà montré leur efficacité en télédétection dans un article séminal de Mnih et Hinton de 2010 [116] pour la détection de routes à large-échelle. Une extension pour le cas des données incomplètes et mal-étiquetées fut proposée deux ans plus tard [117]. Bien que les CNNs soient également utilisés en classification en imagerie hyperspectrale [118], l'apprentissage profond dans ce domaine utilise tout d'abord plutôt des auto-encodeurs (*Stacked Auto-Encoders* (SAEs) [119]) [120] et des machines de Boltzmann (*Restricted Boltzmann Machines* (RBMs)) [121, 122] entraînés de manière non-supervisée dans un but d'apprentissage de représentation, comme une alternative à l'ACP par exemple.

Les premières approches utilisant les CNNs pour la classification d'images optiques (RGB) apparaissent en 2014 et 2015. Tout d'abord en utilisant les CNNs comme extracteur de caractéristique suivant des travaux sur la capacité de transfert [123] des réseaux pré-entraînés sur ImageNet [14]. Ils sont appliqués sur des patches [CI-20, 124] ou des superpixels [125]. Les FCNs de Long, Shelhamer et Darrell [54] permettent la segmentation sémantique, c'est à dire classification dense de chaque pixel du patch. Ils sont appliqués directement avec succès à l'imagerie aérienne [126, 127]. Des revues de la littérature en apprentissage profond sont disponibles pour la télédétection en général [128] et pour l'imagerie hyperspectrale en particulier [129, RI-16].

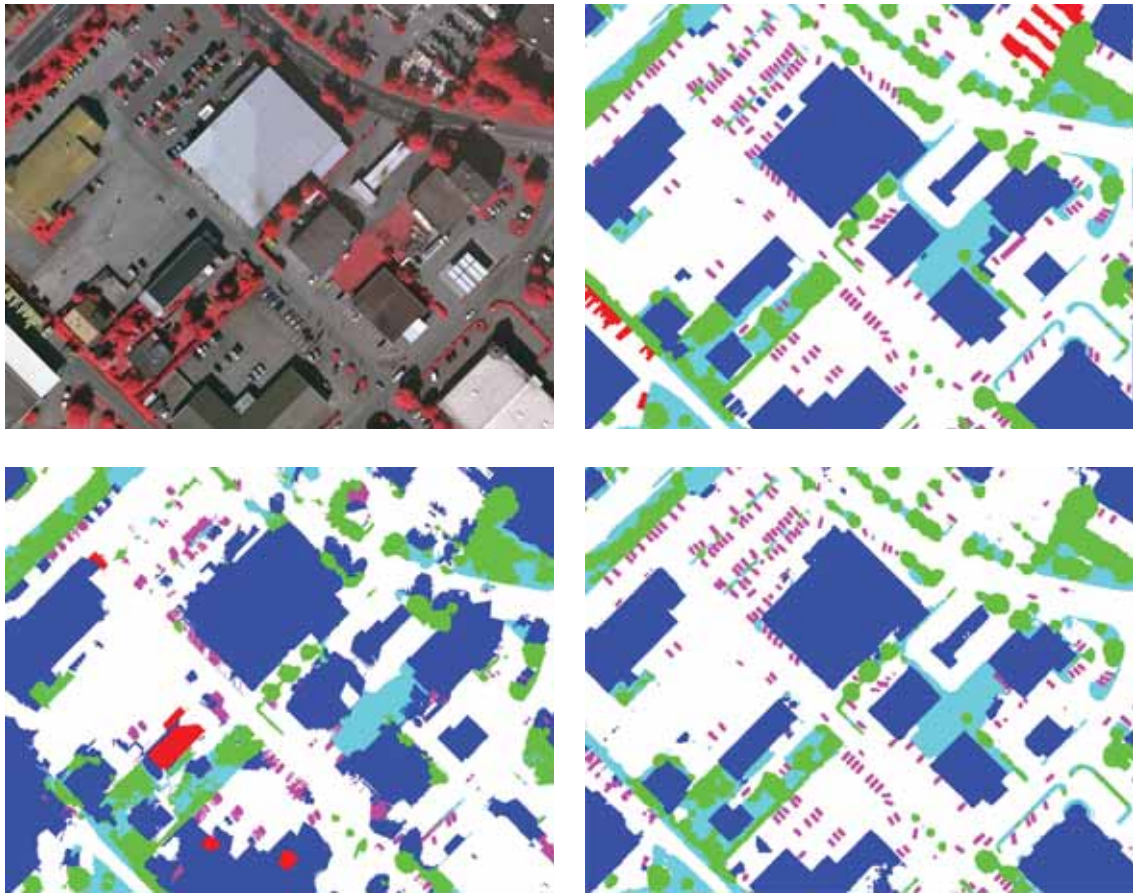


FIGURE 2.6 – Cartographie automatique par réseaux de neurones entièrement convolutifs ((FCNs)) encodeurs-décodeurs SegNet. En haut, image IR/R/G et données de référence associées issues de la base ISPRS Vaihingen. En bas, résultats de l’approche de base (superpixels et CNN [CI-23]) puis segmentation sémantique par SegNet [RI-7]. Source : [RI-7].

Approches de cartographie automatique Le problème est de proposer une approche capable d’interpréter le contenu des images pour effectuer la segmentation sémantique (classification au niveau pixel) des données. Or, pour des images vues du ciel, les cartes de segmentation sont très similaires à des cartes géographiques précises. En effet, les catégories visuelles les plus reconnaissables dans ces images sont les routes, le bâti, ou la végétation, telles que visibles dans les plans routiers ou cadastraux. La segmentation sémantique est donc un moyen d’effectuer la cartographie automatique à partir d’images aériennes ou satellitaires.

Nos premiers travaux ont donc consisté à proposer des moyens de tirer parti des CNNs pour cette segmentation sémantique et les comparer aux approches de l’état de l’art (niveau pixel, SVMs, approches orientées objet, etc). Les expériences réalisées sur les jeux de données DFC2015² (pour lequel nous avons réalisé les données de référence - voir partie 2.3) et ISPRS Vaihingen [130]³ ont montré les meilleures performances des CNNs et surtout leur capacité à

2. Données DFC2015 Zeebruges disponibles sur : <http://dase.grss-ieee.org/>

3. Données ISPRS Vaihingen disponibles sur : <http://www2.isprs.org/commissions/comm2/wg4.html>

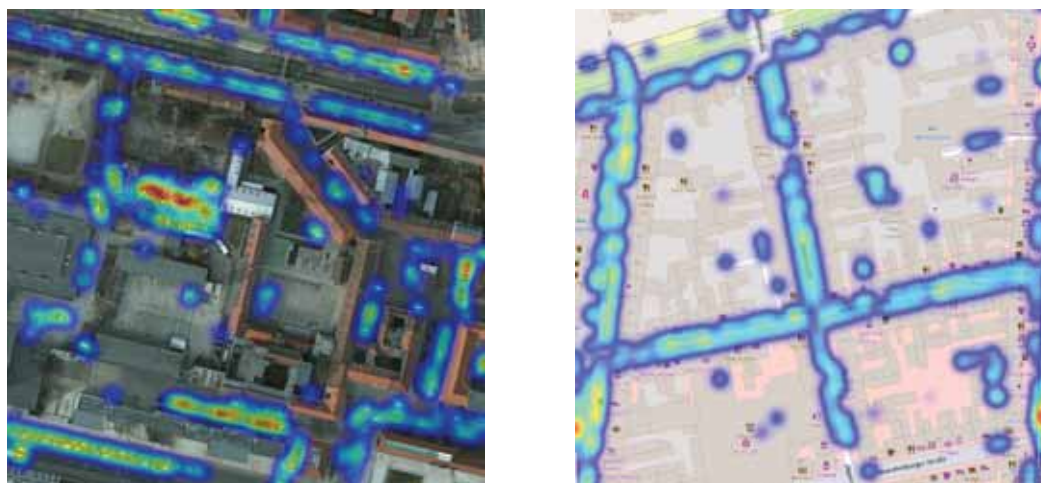


FIGURE 2.7 – Cartes de densité de trafic de véhicules en milieu urbain sur des données ISPRS Potsdam et sur une carte OpenStreetMap : les points d’accumulation permettent d’évaluer l’occupation des parkings et des routes.

apprendre des classifieurs génériques performants, quelque soit la classe [CI-20, RI-3]. Nous avons évalué l’influence de la segmentation en superpixels dans les approches fondées région (segmentation en superpixels puis classification par CNN du patch englobant) et montré la complémentarité de *Simple Linear Iterative Clustering* (SLIC) [131] avec les réseaux convolutifs [CI-23].

Néanmoins, les approches de classification par région sont fortement tributaires du grain de segmentation choisi. En parallèle de [126, 127] et de [132, 133, 134], nous avons donc proposé dans le cadre de la thèse de Nicolas Audebert [114] des architectures entièrement convolutives (FCNs) qui permettent la **segmentation sémantique**. L’approche proposée est fondée sur une architecture encodeur-décodeur de type SegNet [55] et minimise une fonction de pénalité d’entropie croisée standard pour la classification multiclasse [RI-7, CI-24]. En particulier, la qualité de la classification dense (voir figure 2.6) permet de détecter individuellement les véhicules (approche *Segment-before-detect*) et ensuite de les classer par type (berlines, camionnettes, etc.) à l’aide d’un simple CNN [RI-7]). Ces travaux ont notamment été appliqués à l’analyse d’image orientée objet et à l’analyse du milieu urbain : la figure 2.7 montre par exemple des cartes de densité de véhicules en milieu urbain, superposées soit à l’image analysée (issue de ISPRS Potsdam) soit au plan OSM⁴. Cela permet de visualiser la distribution des véhicules dans les images et notamment de repérer les lieux d’intérêt comme les routes à forte circulation ou engorgement et les parkings plus ou moins occupés [CI-24, CI-26].

En télédétection, les données sont rarement limitées à l’image seule : différents capteurs peuvent être combinés lors des acquisitions aériennes et satellitaires. La géolocalisation des données spatiales permet également de mettre en relation différentes sources de données. Nous en avons tiré avantage et proposé des réseaux pour l’**analyse de données multimodales**. Par exemple, quand l’imagerie est combiné à un *Modèle Numérique d’Élévation* (MNE), l’information 3D est une source d’information cruciale pour distinguer des objets (comme précédemment montré en vision en partie 1.3). Nous avons donc proposé des réseaux à flux multiples, toujours

4. Données OpenStreetMap : © les contributeurs OpenStreetMap, disponibles selon la licence Open Database - <http://www.openstreetmap.org/copyright>.

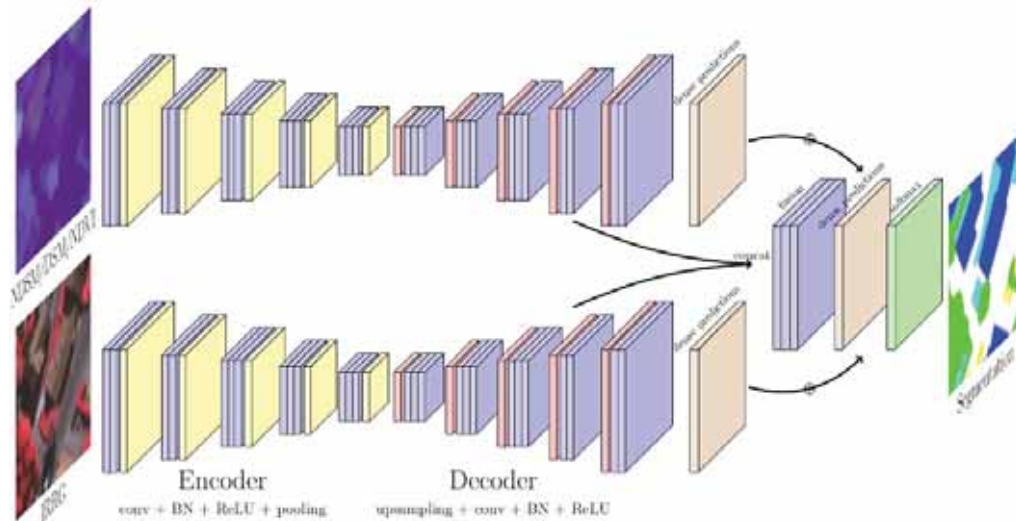


FIGURE 2.8 – Réseau de neurones multimodal avec correction résiduelle. Source : [CI-25].

fondés sur des FCNs encodeurs-décodeurs, ainsi qu'une procédure de fusion entraînable : voir figure 2.8. En élaborant sur l'idée de l'apprentissage résiduel [135], la *correction résiduelle* vise à apprendre la correction du deuxième ordre nécessaire pour améliorer la simple moyenne des sorties de chaque flux [CI-25, CI-27]. Dans un deuxième temps, cette idée a été étendue à une fusion précoce sous le nom VFuseNet [RI-11]. La figure 2.9 et le tableau 2.2 montrent l'apport de la 3D (détection des véhicules sur le parking sur le toit d'un bâtiment) et la segmentation qui peut être atteinte par notre approche.



FIGURE 2.9 – Segmentation sémantique multimodale avec correction résiduelle. De gauche à droite, extrait d'une image IR/R/G de ISPRS Vaihingen, référence, segmentation sans MNE par SegNet [55], segmentation avec MNE par FuseNet [59] et par notre approche avec correction résiduelle [RI-11]. Source : [RI-11].

Grâce au géoréférencement des données spatiales, des données plus hétérogènes que celles issues de capteurs peuvent être mises en correspondance avec les données d'observation de la Terre. C'est le cas des cartes géographiques, et par exemple celles créées collaborativement telles qu'OSM. Elles contiennent des couches d'annotations sur les routes, l'emprise des bâtiments ou la végétation, et peuvent de fait être utilisées comme cible de l'apprentissage. Cependant, elles peuvent également être erronées, imprécises ou incomplètes de part leur mode même de production. Il est donc plus sûr de ne les considérer que comme une donnée d'entrée ancillaire, complémentaire aux données issues de capteurs. Nous avons donc utilisé cette idée pour proposer des architectures de réseaux de neurones qui permettent d'intégrer l'information car-

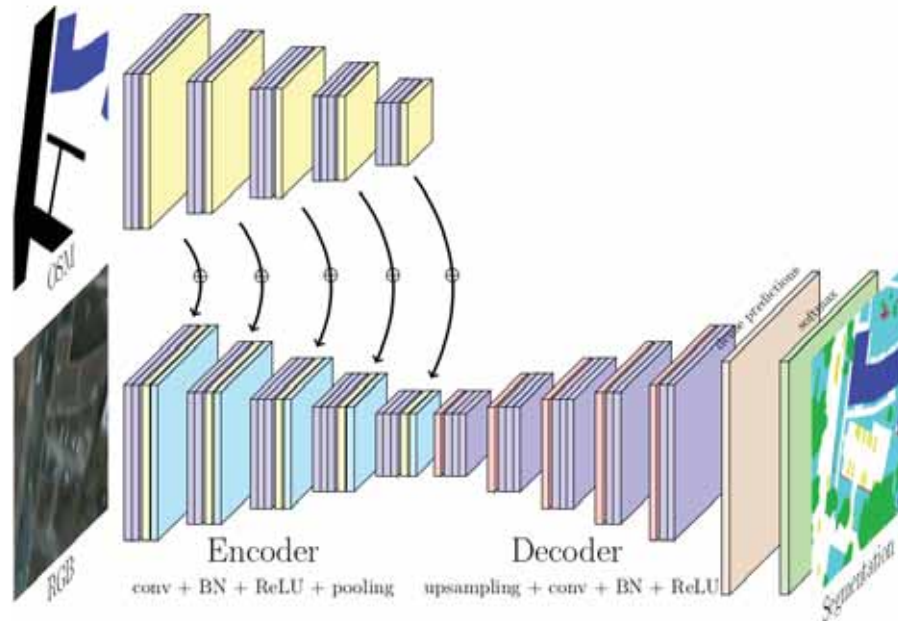


FIGURE 2.10 – Réseau de neurones pour l'apprentissage joint à partir de données fortement hétérogènes. Source : [CI-32].

Tableau 2.2 – Résultats sur le jeu de test d'ISPRS Potsdam (PM = précision moyenne (*overall accuracy* et scores F1 par classe). La meilleure valeur est indiquée en **gras**. Les méthodes du bloc supérieur sont des approches de l'état de l'art extraites du classement officiel (DST5 [127] est la première approche par FCNs, CASIA2 [136] propose une classification multi-échelle en cascade, SWJ2 [137] introduit des couches d'attention dans les réseaux). Les cinq méthodes du bloc inférieur sont nos approches présentées partie 2.2.

Méthode	surfaces imp	bâti	veg.	arbres	voit.	PM
DST5 [127]	92.5	96.4	86.7	88.0	94.7	90.3
CASIA2 [136]	93.3	97.0	87.7	88.4	96.2	91.1
SWJ2 [137]	94.4	97.4	87.8	87.6	95.6	91.7
SegNet [CI-26] RGB	93.0	92.9	85.0	85.1	95.1	89.7
VFuseNet [RI-11] RGB+MNE	92.7	96.3	87.3	88.5	95.4	90.6
FuseNet [CI-32] RGB+OSM	95.3	95.9	86.3	85.1	96.8	92.3
SegNet-SDT [RI-19] RGB	94.3	96.5	88.5	86.5	96.8	92.2
FuseNet (SDT) [RI-19] RGB+OSM	95.2	95.9	86.4	85.0	96.5	92.3

tographique partielle (les classes de la carte diffèrent des classes cibles) comme un moyen de guider l'apprentissage [CI-32, CN-8]. Ces réseaux utilisent l'idée de la correction résiduelle ou bien une intégration progressive de l'information ancillaire par une architecture FuseNet [59], comme illustré sur la figure 2.10. Ces approches permettent d'obtenir des cartes encore plus précises que précédemment, avec un gain de près de 2.5 points de pourcentage (voir tableau 2.2,

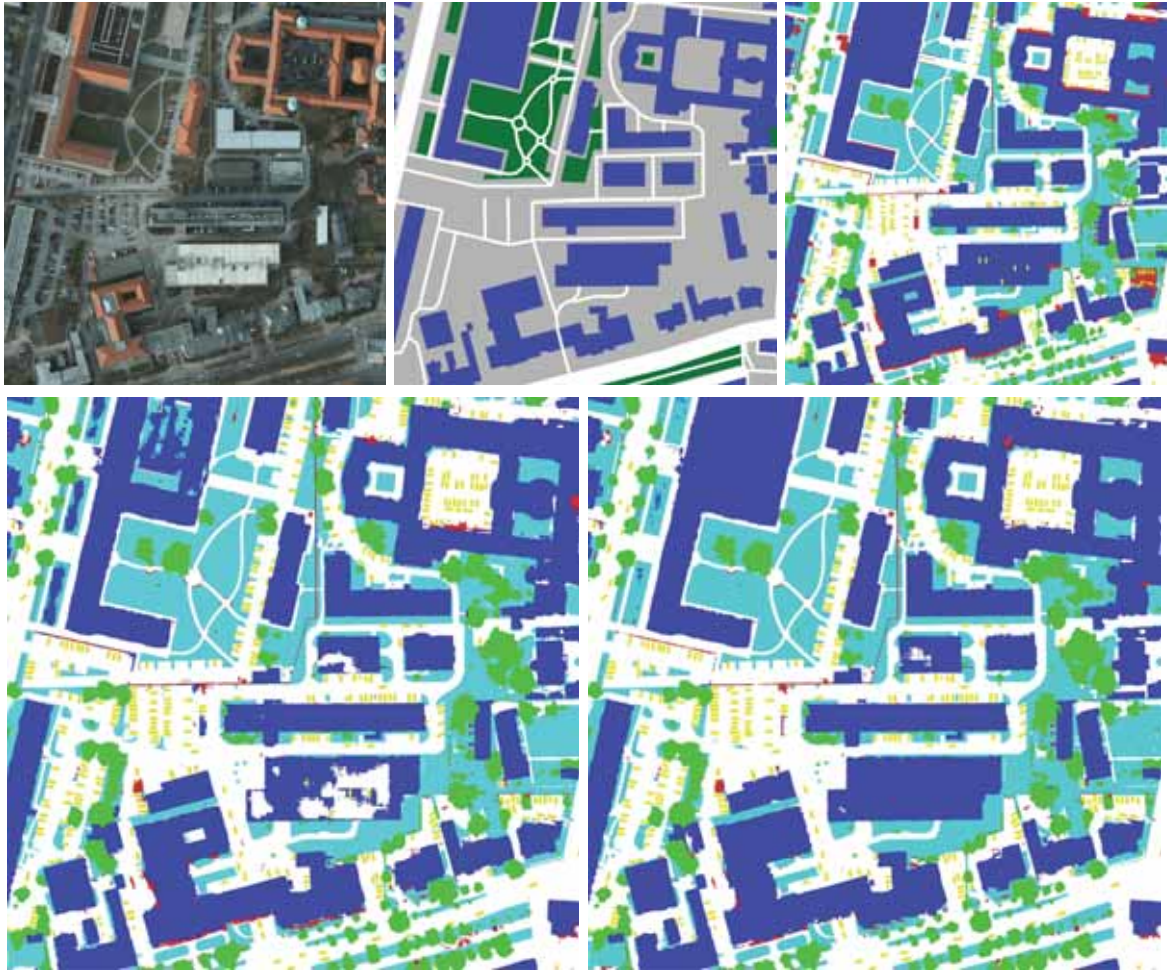


FIGURE 2.11 – Résultats de cartographie automatique par apprentissage conjoint avec données OSM sur ISPRS Potsdam. Source : [CI-32]. En haut : image RGB,

avec comparaison aux méthodes de l'état de l'art⁵). En particulier, l'apport d'OSM permet d'établir les meilleures performances globales sur *ISPRS Potsdam*, grâce notamment à un gain sur les routes et les véhicules dont la localisation est guidée par l'information ancillaire. De plus la cartographie automatique par régression sur des cartes de transformées de distances proposée en [RI-19] permet également de renforcer les algorithmes de segmentation sémantique, y compris les plus simples comme SegNet.

La figure 2.11 montre des bâtiments aux contours mieux définis. Par ailleurs, l'usage de cartes ancillaires accélère la vitesse de convergence de l'apprentissage : de près de 25% sur *ISPRS Potsdam* par exemple [CN-8]. Enfin, dans ce travail est introduit l'idée de compléter l'apprentissage de la classification par une régression sur la carte de distance signée générée à partir des données de référence : cet apprentissage multitâche permet d'inclure de l'information spatiale utile pour prédire des formes d'objets plus régulières [CN-10, RI-19] dans une approche duale aux approches fondées contours [132, 138]. Cela a permis d'être parmi les meilleures approches sur le jeu de données *INRIA Aerial* [139] dédié à la segmentation sémantique de bâtiments et à l'étude de la généralisation de modèles à large-échelle [CI-35] (voir figure 2.12).



FIGURE 2.12 – Segmentation sémantique de bâtiments sur *INRIA Aerial*. De gauche à droite : image RGB de Chicago, référence, segmentation par SegNet et segmentation avec SegNet-SDT (apprentissage multitâche avec cartes de distance signées). Les pixels correctement classés sont en vert, les faux positifs en rose et les faux négatifs en bleu. L'apprentissage multitâche permet au réseau de mieux capturer la structure spatiale des bâtiments. Source : [RI-19].

Enfin, nous nous sommes également intéressés aux données issues de capteurs hyperspectraux. Elles ont la particularité d'être des cubes de données plus que de simples images, puisque l'information spectrale est une dimension à part entière. La rareté des données de référence associées est un problème récurrent dans ce domaine. Nous avons proposé des modèles génératifs par *Generative Adversarial Network (GAN)* pour la synthèse de spectres hyperspectraux correspondants à des classes données [CI-39]. Par ailleurs, à destination de la communauté, souvent composée de thématiciens, nous avons proposé la boîte à outils logicielle *DEEP Learning for HYPERspectral toolboX (DeepHyperX)* associée à la revue comparative des approches de l'état de l'art [RI-16]. *DeepHyperX* comporte différents modèles permettant d'analyser les cubes hyperspectraux : par spectre avec des convolutions 1D, selon une approche spatiale-spectrale avec des convolutions 2D et 1D, ou selon des motifs locaux spatiaux-spectraux avec des convolutions 3D (voir figure 2.13). Sur les jeux de données les plus volumineux (*Pavia* et *DFC2018* [RI-17]), nous avons montré que les réseaux de neurones 3D obtenaient les meilleures performances de classification, en raison probablement de leurs capacité à apprendre des motifs

5. Classement des approches sur *ISPRS Potsdam* : <http://www2.isprs.org/commissions/comm2/wg4/potsdam-2d-semantic-labeling.html>

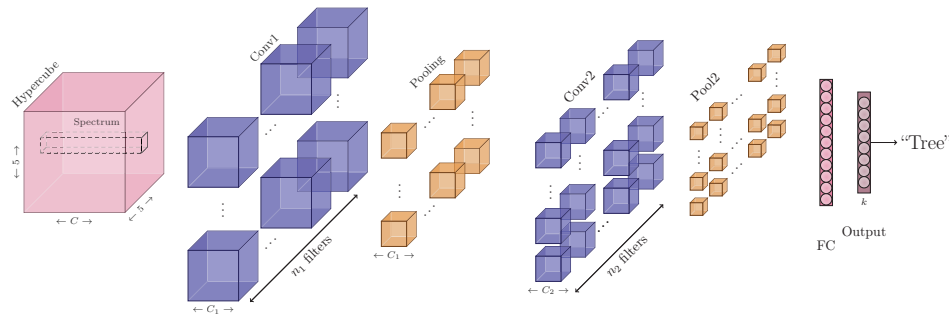


FIGURE 2.13 – Réseau de neurones 3D pour la classification de données hyperspectrales disponible dans [DeepHyperX](#). Il reproduit l’architecture proposée dans [140] et alterne convolutions 3D et couches de *max-pooling* 3D. Source : [RI-16].

spatiaux-spectraux précis et caractéristiques à certaines longueurs d’onde.

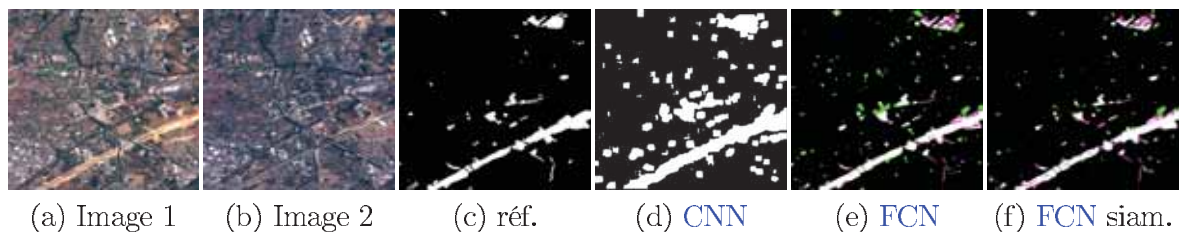


FIGURE 2.14 – Jeu de données [OSCD](#) (extrait sur Montpellier) et détection de changements par réseaux convolutifs, en utilisant les 13 canaux des images Sentinelle 2. (d) [CNN](#) ; (e) [FCN](#) avec concaténation des images en entrée ; (f) [FCN](#) siamois. Dans les images (e) et (f), blanc indique les vrais positifs, noir les vrais négatifs, vert les faux positifs, et magenta les faux négatifs. Source : [CI-38, CI-42].

Approches de détection de changement La détection de changements est de longue date une finalité de la télédétection [141, 142]. Elle consiste à comparer deux images ou plus à des dates différentes pour identifier les zones qui ont changé, et ainsi analyser l’activité du site concerné [143, 144]. Les approches cherchent à construire des fonction de différence entre les deux images [145]. La plupart des approches historiques sont non-supervisées et cherchent à automatiser la découverte de motifs de changement [146, 147, 148]. En mode supervisé, les [SVM](#) ont été utilisées [149], et notamment en mode interactif [CI-14]. Par manque de jeu d’apprentissage, les rares méthodes récentes à base de [CNN](#) procèdent par transfert d’apprentissage, en réutilisant des poids pré-entraînés pour une autre tâche [150, 151]. Or il y a un avantage certain à entraîner de bout en bout un réseau pour une tâche dédiée, et d’autant plus si les images correspondent au cadre d’usage.

Dans le cadre de la thèse de Rodrigo Daudt (en cours), nous avons donc proposé un jeu de données pour la détection de changement. la base [ONERA Sentinel Change Detection \(OSCD\)](#) [CI-38] comporte des couples d’images multispectrales Sentinelle-2 sur plusieurs villes du monde (24 paires d’images $600 * 600$ pixels), associées à des cartes de changement pixeliques

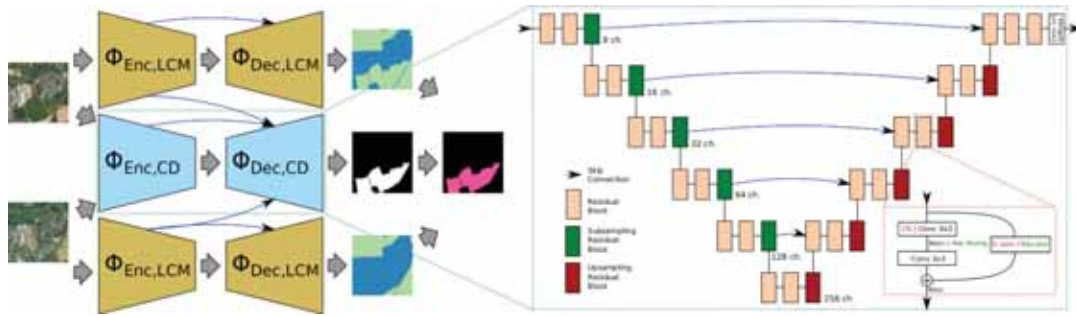


FIGURE 2.15 – Réseau de neurones multitâche pour la détection de changement sémantique. Source : [RI-18].

réalisées par analyse visuelle⁶. Nous avons alors proposé des premières approches par **CNN** : soit par réseau à simple-flux, soit par réseau siamois à deux flux [CI-38] pour gérer les deux images d'entrée. L'approche a montré la meilleure réussite pour cette tâche. En particulier, les expériences ont montré de meilleures performances quand les 13 canaux des images multispectrales Sentinelle-2 étaient utilisées. Dans la foulée, nous avons proposé des approches par **FCN**, à nouveau selon les modes simple-flux et siamois, qui ont conduit à une amélioration des performances [CN-9, CI-42] (voir figure 2.14).

Cependant, pour d'une part attaquer les problèmes environnementaux et sociétaux à grande échelle, et d'autre part bénéficier pleinement de la capacité d'apprentissage des modèles actuels, il faut avoir recours à des données massives représentant de larges zones géographiques. C'est pourquoi nous avons constitué le jeu de données *High-Resolution Semantic Change Dataset (HRSCD)*, constitué d'imagerie aériennes à 50cm de résolution (Base de Données Orthophotographique (BD ORTHO) de l'IGN) et de cartes d'occupation du sol et de changement (Urban Atlas de l'European Environment Agency (EEA)). HRSCD contient les deux villes de Rennes et Caen et leurs environs, en 2005 et 2012. La taille de HRSCD (291 paires d'images $10k * 10k$ pixels) permet non seulement d'entraîner des détecteurs de changements, ou des réseaux de segmentation sémantique, mais de cartographier les changements sémantiques (passage d'une classe donnée à une autre). Nous avons également présenté une approche d'**apprentissage multitâche** [152] pour la **détection de changement sémantique**. Elle repose sur une architecture de réseau de neurones qui exécute la détection de changement et la cartographie automatique simultanément. L'information est propagée depuis les images à différentes dates vers les différents objectifs par des branches distinctes qui partagent leur paramètres via des connexions croisées. Nous avons montré que cet apprentissage joint donne de meilleures performances que des approches séquentielles ou plus naïves où l'information est traitée en bloc [RI-18].

2.3 Jeux de données pour l'apprentissage machine en télédétection

Aperçu En parallèle des travaux des parties 2.1 et 2.2, au sein du comité technique de l'*Institute of Electrical and Electronics Engineers (IEEE) Geosciences and Remote Sensing Society (GRSS)* consacré à l'analyse d'image et la fusion de données (*Image Analysis and*

6. Données sur : <https://iee-dataport.org/open-access/oscd-onera-satellite-change-detection>; et serveur d'évaluation : <http://dase.grss-ieee.org/>

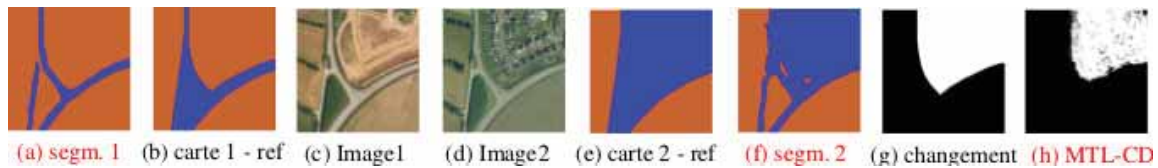


FIGURE 2.16 – Détection de changement sémantique sur des données HRSCD. Les résultats de l'apprentissage multitâche présenté figure 2.15 sont indiqués en rouge. Source : [RI-18].

Data Fusion Technical Committee (IADF TC)), nous travaillons à rendre publics des jeux de données pour permettre le développement de nouvelles méthodes d'apprentissage et leur comparaison de manière rigoureuse et impartiale. Les deux axes mis en avant sont les capteurs innovants (vidéo depuis l'espace, multiples capteurs, LiDAR multispectral) comme dans le cas du DFC2016 [RI-8] et du DFC2018 [RI-12], ainsi que les données à large échelle, si possible ouvertes et collaboratives. Cet axe comprend notamment le DFC2017 [RI-13] qui proposait d'utiliser des données Landsat et Sentinelle-2 avec des cartes OSM pour classifier la typologie urbaine, tout en vérifiant la capacité d'adaptation de domaine des méthodes évaluées (9 villes autour du globe). Plus récemment, le DFC2019 [RI-15] proposait des données satellitaires THR et des données 3D LiDAR à des fins de reconstruction 3D sémantique sur de larges territoires urbains.

Contexte Si les parties 2.1 et 2.2 ont montré que les méthodes d'apprentissage sont essentielles et performantes pour automatiser la compréhension de données d'observation de la Terre, il n'en reste pas moins qu'elles nécessitent des données pour apprendre. Or, ces données, et surtout les données annotées, ne sont pas si communes. Cela provient notamment de la multiplicité des capteurs, du commerce de ces données, et d'une prise de conscience plus tardive de l'intérêt du développement collaboratif et partagé. L'initiative des *Data Fusion Contests (DFCs)* commence en 2006 avec l'action de Paolo Gamba et Jocelyn Chanussot [153] et un jeu de données consacré à l'affinage panchromatique (*pan-sharpening*) d'images THR à partir d'images à haute résolution (simulations Pléiades et images QuickBird). Sept équipes de par le monde participèrent à la compétition qui établit la supériorité des résultats des approches par analyse multirésolution. Depuis lors, la série des DFC poursuit deux objectifs. Le premier consiste à rendre publiques des **données multimodales** dont certaines **inédites** et issues de **capteurs innovants** : imagerie optique et SAR multitemporelle [160], imagerie hyperspectrale [161, 162, 163], données multiangulaires [164]. Le deuxième but est de diffuser des **données multimodales** (si possible **avec des données de référence**) à **large-échelle** (plus grande que l'état de l'art courant). Cela vise à faire émerger de nouvelles familles d'algorithmes. Cet axe inclue la détection de changement [165], la fusion large-échelle de données optiques, SAR et LiDAR [166], et la classification [162, 163]. En parallèle, l'ISPRS et notamment le comité technique II/4 sur l'analyse et la reconstruction de scènes a diffusé plusieurs jeux de données pour la classification urbaine et la reconstruction 3D, la segmentation sémantique 3D et 2D : notamment les données ISPRS *Vaihingen* et ISPRS *Potsdam* précédemment utilisées [156, 155]. En maintenant ouvert un serveur d'évaluation de résultats de 2013 à 2018, ces jeux de données sont devenus des références pour la segmentation sémantique aérienne. Enfin, une autre base de référence pour la détection de bâtiments à échelle globale est l'*INRIA Aerial Image Labeling Dataset* [139] : un

Tableau 2.3 – Statistiques des principaux jeux de données en télédétection. Les tirets – indiquent que l’information est sans objet, les points d’interrogation ? indiquent que l’information n’est pas disponible.

Nom	Source	Tâche	Taille (Go)	Emprise (km ²)	Résol. (m/pixel)	Classes
DFC06 [153]	QuickBird/ Pléïades	<i>Pansharpening</i>	4	105	2.8/0.7	-
DFC2015 [RI-3]	Aérien	Segmentation sémantique	x	35	0.05	8
DFC2016 [RI-8]	Deimos-2	-	25.2	-	1/4	-
DFC2017 [RI-13]	Landsat/S2	Classification	x	x	10/20	17
DFC2018 [RI-17]	Aérien	Classification	10.1	4.9	0.05/1	20
DFC2019 [RI-15]	WorldView3	3D / Classif.	320	100/ 20)	(~0.3	6
URBAN1-2 [154]	?	Routes & bâtiments	?	578	1.2	2
ISPRS Vaihingen [155]	Aérien [156]	Classification	2.2	7.02	0.09	6
ISPRS Potsdam [155]	Aérien [156]	Classification	16.9	68.4	0.05	6
xView [157]	WorldView3	Détection d’objets	33	1415	0.3	60
DeepGlobe18 [158]	Vivid / WorldView3	Routes & bâtiments / Classification	1632/983 /1717	0.5/0.3/ 0.5	1/1/7	
OSCD [CI-38]	S2	Changement	0.49	86.4	10	2
xView2/xBD [159]	WorldView3 ?	Changement / Bâtiment	17.2	19804	0.3	5×4
HRSCD [RI-18]	BDORTHO	Changement	19.4	29100	0.5	5 ² × 2

serveur d’évaluation⁷ y est associé et les meilleures approches après un an de compétition ont été résumées dans [CI-35] (voir également partie 2.2). Dans le domaine de l’imagerie hyperspectrale, plusieurs jeux de données sont couramment utilisés : *Indian Pines*, *Pavia (University & City)*, etc (voir tableau 2.4). Les références (vérités-terrain) complètes de ces jeux de données ont souvent été publiées, ce qui conduit hélas à des évaluations différentes d’un article à l’autre et réduit la portée de ces données en tant que *benchmark* d’évaluation, comme nous l’avons souligné dans notre revue du domaine [RI-16]. Pour y remédier, le serveur *Data and Algorithms Standard Evaluation (DASE)*⁸ de l’IEEE GRSS a instauré des partitions d’entraînement et de test de référence qui permettent une comparaison loyale.

En parallèle, les communautés de vision par ordinateur et d’apprentissage machine promeuvent également des jeux de données pour l’imagerie aérienne, mais avec des résultats de diffusion variés. De fait, les jeux de données URBAN1-2 pour la détection de routes et bâtiments à large échelle de [116, 117] ont été rendus publics [154] plusieurs années après les

7. <https://project.inria.fr/aerialimagelabeling/>

8. <http://dase.grss-ieee.org/>

Tableau 2.4 – Principaux jeux de données publics en imagerie hyperspectrale. Source : [RI-16].

Nom	Pixels	Canaux	Gamme de fréquences	Résol.	Labels	Classes	Mode
Pavia (U & C)	991,040	103	0.43-0.85 μm	1.3 m	50,232	9	Aérien
Indian Pines	21,025	224	0.4-2.5 μm	20 m	10,249	16	Aérien
Salinas	111,104	227	0.4-2.5 μm	3.7 m	54,129	16	Aérien
KSC	314,368	176	0.4-2.5 μm	18 m	5,211	13	Aérien
Botswana	377,856	145	0.4-2.5 μm	30 m	3,248	14	Satellite
DFC 2018	5,014,744	48	0.38-1.05 μm	1 m	547,807	20	Aérien

articles⁹ tandis que TorontoCity [167] n'est toujours pas public deux ans après l'annonce de sa création. Aujourd'hui, des acteurs industriels ou institutionnels américains proposent également des défis accompagnés de jeux de données. C'est le cas des *challenges* SpaceNet [168] qui ont été consacrés depuis 2017 à la détection de bâtiments, de réseaux routiers ou encore l'estimation de temps de trajets. Également, les *challenges* XView ont été organisés avec comme but la détection d'objets (véhicules)¹⁰ [157] ou l'évaluation de dégâts aux bâtiments post-catastrophe (xView2¹¹) en utilisant le jeu de données à échelle globale xBD [159]. Enfin, les *challenges* multidisciplinaires et réunissant des acteurs académiques et industriels sont peut-être les plus réussis, à l'image de DeepGlobe [158]. DeepGlobe 2018¹² proposait trois compétitions parallèles : détection de routes, de bâtiments, et classification de l'usage du sol. Les principaux jeux de données et leurs statistiques sont recensés dans le tableau 2.4 pour l'imagerie hyperspectrale et dans le tableau 2.3 pour l'imagerie optique classique.

Pour remédier au manque de données pour l'apprentissage, j'ai donc contribué à rassembler et diffuser des jeux de données multisourceq et de données de référence (vérité-terrain) pour permettre l'entraînement d'algorithmes supervisés. L'objectif est de permettre le développement de nouvelles méthodes en télédétection et analyse d'image pour l'observation de la Terre, et aussi évaluer et comparer les techniques concurrentes sur des tâches spécifiques (fonction de *benchmark*).

IEEE GRSS Data Fusion Contest 2015¹³ Cette compétition, organisée par l'IADF TC (Gabriele Moser et Devis Tuia) en collaboration avec l'Académie Militaire Royale de Belgique (Michal Shimoni), distribuait deux types de données acquises sur la ville de Zeebruges : des images aériennes ortho-rectifiées à 5cm/pixel et des données LiDAR à 10cm/pixel. L'objectif était la réalisation d'un travail de recherche documenté dans un article. Il y avait deux *challenges* parallèles : 2D [RI-3] et 3D [169]. Dans le cadre du stage d'Adrien Lagrange, nous avons constitué une référence de segmentation sémantique pour le jeu d'images selon plusieurs classes visuelles urbaines : eau, routes, bâtiments, végétation haute et basse, voitures et bateaux. Cela a par ailleurs servi au comparatif d'approches sur la segmentation d'ortho-photos et des MNEs associés présenté en partie 2.2 [CI-20]. Après le DFC, nos cartes de référence ont servi à établir

9. Disponibles sur l'ancienne page personnelle de V. Mnih : <https://www.cs.toronto.edu/~vmnih/data/>

10. <http://xviewdataset.org/>

11. <https://xview2.org/>

12. <http://deepglobe.org>

13. <http://www.grss-ieee.org/community/technical-committees/data-fusion/2015-ieee-grss-data-fusion-contest/>

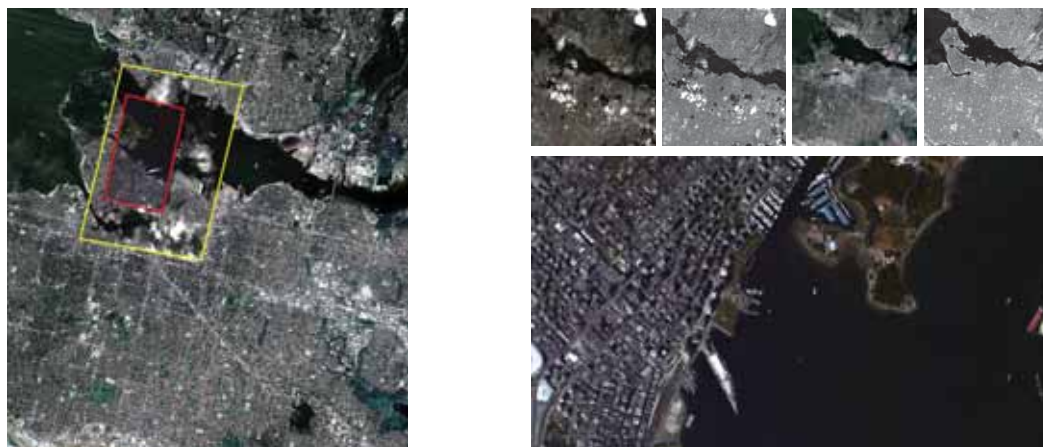


FIGURE 2.17 – Données diffusées dans le cadre du DFC2016. À gauche, emprises superposées de l'image multispectrale (cadre complet), image panchromatique THR (cadre jaune), et vidéo haute définition (cadre rouge). À droite, images multispectrales et panchromatiques pour deux dates, et image de la vidéo haute définition. Source : [RI-8]

le *benchmark* de classification toujours disponible sur le serveur DASE.

IEEE GRSS Data Fusion Contest 2016 ¹⁴ Cette compétition fut organisée par l'IADF TC (Devis Tuia, Gabriele Moser et moi-même) en collaboration avec Deimos Imaging (Roberto Fabrizi) et UrtheCast (Sven Cowan), opérateurs de satellites et fournisseurs de données. Diverses données satellitaires multisources et multitemporelles étaient rendues publiques : imagerie optique à haute résolution 1m/pixel ; imagerie multispectrale à 4m/pixel ; et surtout vidéo-depuis-l'espace à haute définition (1m/pixel). Ce DFC appartient donc à l'axe *données innovantes*. Les acquisitions étaient faite au-dessus de Vancouver, Canada, à deux dates différentes. Il s'agissait d'un défi ouvert évalué sur la base d'un article, visant à promouvoir les meilleurs travaux de recherche possibles avec ces données. Comparativement avec l'année précédente, les travaux primés proposaient des approches d'apprentissage profond plus avancées, appliquées à l'interprétation de scène et au pistage vidéo depuis l'espace d'une part, et au recalage, à la segmentation sémantique et la détection de changement simultanés d'autre part [RI-8].

IEEE GRSS Data Fusion Contest 2017 ¹⁵ L'IADF TC (Devis Tuia, Gabriele Moser et moi-même) organisa cette édition avec WUDAPT ¹⁶ (Benjamin Bechtel) et Geo-Wiki ¹⁷. Ce *challenge* visait des buts environnementaux globaux : il s'agissait d'estimer les Zones de Climat Local - *Local Climate Zones* (LCZ) [170] sur plusieurs grandes villes du monde à partir de données *open source*. Les LCZ sont notamment utiles pour déterminer les îlots de chaleurs urbains. À cet effet, des données multitemporelles, multisources et multimodales (couches images et sémantiques) furent diffusées. Précisément, pour chaque ville furent distribuées des images

14. <http://www.grss-ieee.org/community/technical-committees/data-fusion/2016-ieee-grss-data-fusion-contest/>

15. <http://www.grss-ieee.org/community/technical-committees/data-fusion/2017-ieee-grss-data-fusion-contest-2/>

16. <http://www.wudapt.org/>

17. <https://www.geo-wiki.org/>

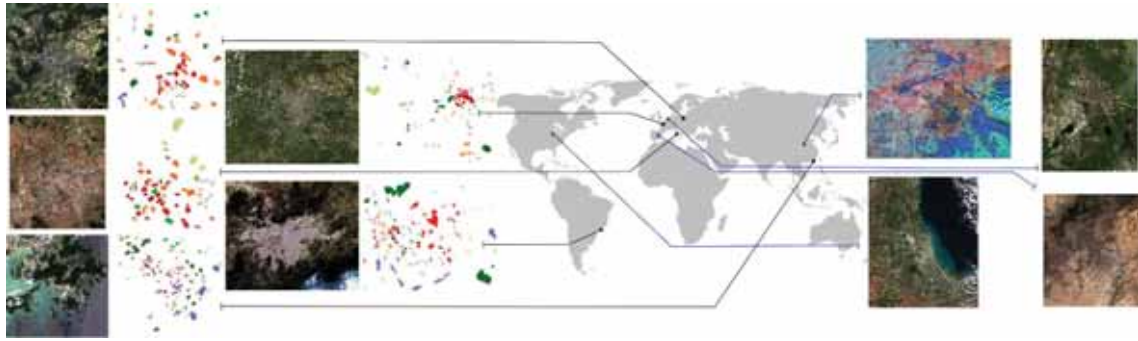


FIGURE 2.18 – Distribution mondiale globale des données du DFC2017. Des images satellitaires Landsat 8 et Sentinelle 2 ainsi que des couches OSM et des cartes de climat locales (LCZ) sont distribuées pour les villes servant à l'apprentissage (à gauche : Berlin, Hong Kong, Paris, Rome, and Sao Paulo). Seules les images satellitaires et les couches OSM sont rendues publiques pour les villes de test (à droite : Amsterdam, Chicago, Madrid, and Xi'An). Source : [RI-13].

multispectrales Landsat et Sentinelle-2 à différentes dates, ainsi que des données vectorielles et rasterisées OSM. La compétition consistait en un *benchmark* de classification, avec des données de références publiées uniquement pour la phase d'apprentissage, et pas pour la phase de test. L'évaluation avait lieu à nouveau sur le serveur DASE. Bien que les méthodes d'apprentissage profond furent les premières à donner des performances élevées, les approches qui obtinrent les meilleurs scores étaient fondées sur le *boosting* et les forêts aléatoires et furent celles qui intégraient des a priori sur les données (correction atmosphérique par exemple) et qui tiraient parti de données libres additionnelles, prouvant que la masse de données est un levier efficace pour une analyse précise [RI-13].

IEEE GRSS Data Fusion Contest 2018 ¹⁸ En 2018, l'accent était à nouveau mis sur des capteurs innovants, avec des données issues d'un capteur aérien LiDAR *multispectral*. Cependant, un objectif majeur de ce DFC était l'analyse d'un milieu urbain en mêlant diverses sources d'information en 3D (voir figure 2.19). Les données comportaient : LiDAR multispectral (sous forme de nuages de points et de MNE à 0.5m/pixel, images THR à 0.05m/pixel et images hyperspectrales à 1m/pixel. Ces données étaient distribuées par l'Université de Houston (Saurabh Prasad) avec qui l'IADF TC (moi-même, Naoto Yokoya et Ronny Hänsch) co-organisait le *challenge*. La compétition était à nouveau sous la forme d'un *benchmark* de classification au niveau pixel, c'est à dire de segmentation sémantique. Elle comportait une phase d'entraînement où un tiers des données était rendu public avec une carte sémantique associée de 20 classes (allant de catégories d'usage du sol à des objets composés de matériaux spécifiques) suivie d'une phase de test, évaluée sur DASE. En raison du volume de données disponible, les réseaux de neurones profonds obtinrent les meilleures performances, se révélant commodes pour fusionner diverses sources grâce à des architectures ad hoc. Néanmoins, des post-traitements et des classifieurs spécifiques pour certaines classes étaient utilisés dans les méthodes qui établirent l'état de l'art, avec des gains d'environ 15 points de pourcentage. Au niveau des capteurs, le LiDAR multispectral se révéla particulièrement informatif, puisqu'il obtint le meilleur taux de

18. <http://www.grss-ieee.org/community/technical-committees/data-fusion/2018-ieee-grss-data-fusion-contest/>



FIGURE 2.19 – Données multimodales 3D du DFC2018. Le nuage de points acquis sur Houston, Texas, USA, est coloré (de gauche à droite) par les fausses couleurs de l’image hyperspectrale, les fausses couleurs du LiDAR multispectral et les couleurs RGB de l’imagerie THR à 5cm/pixel.

classification, devant même la fusion de données avec l’imagerie THR [RI-17].

IEEE GRSS Data Fusion Contest 2019 ¹⁹ La thématique explorée concernait la reconstruction 3D sémantique à large échelle [RI-15]. L’université Johns Hopkins (Myron Brown) et la IARPA (Hakjae Kim et Gary O’Brien) contribuèrent à organiser cette édition avec l’IADFTC (moi-même, Naoto Yokoya et Ronny Hänsch). Jusqu’ici, la 3D avaient souvent été une donnée d’entrée, comme en 2015 et 2018 avec du LiDAR aérien. Ce DFC accompagnait l’avancée technologique vers l’estimation 3D à haute résolution depuis l’espace, possible avec l’imagerie satellitaire THR, qui présente l’avantage d’être disponible partout dans le monde avec des mises à jour bien plus aisées et fréquentes. De plus, l’aspect sémantique était également promu de manière à inciter les participants à proposer des approches qui dépassaient la classification de MNE 2.5D, soit dans des cadres multitâches, soit dans un cadre nuage de points 3D. De fait, quatre compétitions étaient proposées. Estimation ou reconstruction 3D, dans les trois cadres mono-image, stéréo avec une paire d’images, et stéréo multivue. Ces trois compétitions étaient multiobjectifs et s’accompagnaient de segmentation sémantique. Enfin, un quatrième et dernier *challenge* visait à la segmentation sémantique 3D de nuages de points. À ces fins, des jeux de données d’images WorldView3 de résolution $\sim 30\text{cm}/\text{pixel}$ à divers angles d’incidence de la base de données US3D [171] étaient distribués, ainsi que des données issues de LiDAR aérien (densité des points $\sim 80\text{cm}$) comme référence pour la reconstruction 3D ou comme entrée pour la classification 3D. Les sites étudiés étaient Jacksonville, en Floride and Omaha, au Nebraska, USA, pour une emprise totale de 100km^2 , et 20km^2 par *challenge*. L’évaluation était effectuée sur le serveur CodaLab²⁰. Ce DFC montra que les approches mono-vue pour l’estimation 3D étaient hautement compétitives grâce aux réseaux de neurones, rivalisant avec les approches plus géométriques. En segmentation sémantique 3D, à nouveau des approches neuronales fondées sur des encodages locaux des points telles que PointCNN [172] et PointSIFT [173] obtinrent les meilleures performances.

Analyse et impact Au niveau de la participation, ces *challenges* montrent une globalisation croissante et un accroissement de la multidisciplinarité. En effet, si des traitements en imagerie hyperspectrale ou en reconstruction 3D multivue nécessitent de l’expérience et une connaissance spécifique du domaine, des approches issues des sciences des données ou de la vision par ordinateur sont aussi, voire plus performantes sur des données standard, pour des problèmes posés simplement. Par ailleurs, les DFC poursuivent un objectif de formation avec de nom-

19. <http://www.grss-ieee.org/community/technical-committees/data-fusion/2019-ieee-grss-data-fusion-contest/>

20. <https://competitions.codalab.org/competitions/?q=2019+IEEE+GRSS+Data+Fusion+Contest>

breux étudiants qui y participent, encouragés en cela par leurs professeurs dans de nombreuses universités de par le monde.

Sur le plan du contenu, l'évolution que j'ai pu contribuer à mettre en oeuvre a été d'aller vers des jeux de données de plus en plus grands, que ce soit à l'échelle locale (centres urbains) ou bien à large échelle : soit de manière globale avec des sites répartis à la surface du globe ou bien avec la couverture de très grandes régions (voir tableau 2.3). Nous avons également systématisé la diffusion de données de référence à grande échelle pour offrir l'opportunité d'entraîner des modèles d'une complexité nouvelle et valider les différents algorithmes. Un défi restant concerne la capacité au transfert d'apprentissage et à l'adaptation de domaine. En effet les modèles actuels sont souvent surappris sur la zone géographique initiale, comme l'a montré le DFC2017 qui testait différents villes de par le monde. Un objectif ambitieux serait de fournir un jeu de données suffisamment grand et générique pour donner la possibilité de pré-entraîner des modèles pour d'autres applications. Cela reviendrait à fournir un équivalent d'ImageNet [14] pour la télédétection, ce but sera exploré plus avant dans le chapitre 4, partie 4.2.

2.4 Résumé et conclusions

Ce chapitre présentait des travaux dans le domaine de la télédétection et de l'observation de la Terre, c'est à dire la compréhension de scènes vues du ciel. Tout d'abord la partie 2.1 présentait plusieurs approches d'apprentissage automatique pour la classification et la détection d'objets dans des images aériennes ou satellitaires. Notamment, elle comporte l'introduction et la validation des modèles à parties déformables (DPM) dans ce contexte et des approches par *gradient-boosting* pour l'apprentissage en ligne. Puis la partie 2.2 présentait des approches d'apprentissage profond par réseaux de neurones pour ces mêmes problématiques, montrant un gain de performances considérable. En particulier, les problématiques d'apprentissage multi-modal (multiples capteurs et sources de données hétérogènes) et de détection de changement sémantique ont été explorées. Au delà de leur qualité, l'intérêt des résultats actuels est de révéler des questions plus subtiles et de dégager des perspectives ambitieuses qui seront exposés au chapitre 4. Enfin, la partie 2.3 récapitule nos efforts pour créer l'environnement propice au développement de l'apprentissage automatique dans ce domaine, en créant et diffusant des bases de données pour entraîner les algorithmes.

Chapitre 3

Vision et compréhension 3D

Mes activités autour de la 3D visent à explorer les liens entre image et 3D sous l'angle de l'apprentissage et de l'estimation statistique, et non selon des approches fondées sur la géométrie ou la photogrammétrie. Elles ont des applications en imagerie biologique, reconstruction 3D et interprétation de scènes 3D.

Précisément, la question récurrente dans tous mes travaux sur ce thème est de pouvoir estimer la position (l'état) ou une valeur associée à des points 3D de la scène, alors que cette scène n'est que partiellement échantillonnée. L'échantillonnage est ici très structuré : il s'agit surtout d'images 2D (vues générées par coupe ou projection) de la scène 3D. L'estimation est possible par l'utilisation d'a priori apportés par le contexte local, sous la forme d'information de flou ou de cohérence sémantique locale.

Deux idées / mécanismes sont ici à l'oeuvre :

- D'abord un échantillonnage de l'espace 3D, typiquement fondé sur des vues 2D (images ou vues en coupe), qui permet un accès direct aux données ;
- Ensuite un moyen d'accéder à l'information manquante, par exemple via le flou d'acquisition, qui diffuse l'information des zones non imagées dans échantillons.

Les trois travaux suivants illustrent ce principe avec différentes combinaisons : échantillonnage massif et flou de diffusion pour la reconstruction 3D en microscopie confocale 3.1, une seule vue et flou d'acquisition pour l'estimation de profondeur mono-image 3.2, et échantillonnage sous un grand nombre de points de vue pour la sémantisation 3D (voir partie 3.3).

3.1 Reconstruction 3D en microscopie confocale

Aperçu Mes premiers travaux sur la 3D avaient pour cadre la tomographie et la microscopie confocale et étaient menés avec Jiaping Wang (Université Paris-13) et Bernard Chalmond (ENSEA) à l'ENS Cachan. Ils avaient pour objectif la **reconstruction 3D** de cellules vivantes de lymphocytes à partir d'images 2D (plans de coupe tournant autour d'un axe de rotation). J'ai pu montrer l'apport du flou de diffraction du microscope pour reconstruire le volume 3D complet par estimation alternée de la position des coupes et du volume 3D de la cellule [RI-1, CI-10].

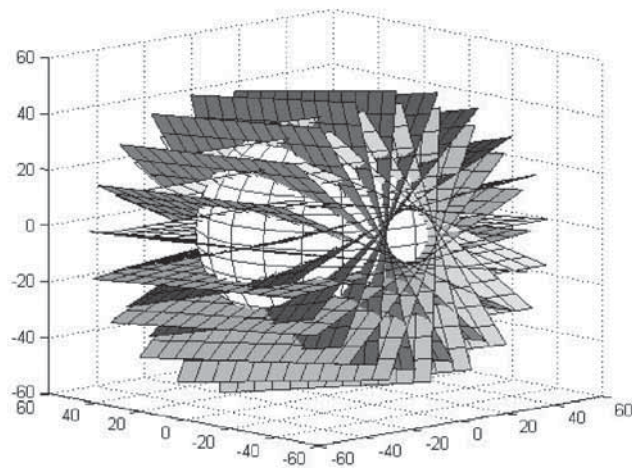


FIGURE 3.1 – Plans de coupe par micro-rotations du microscope confocal dans le repère de la cellule. Des artefacts tels que le "trou noir" autour de l'axe de rotation peuvent apparaître en raison de la dérive de la cellule. Source : [RI-1].

Contexte de la microscopie confocale La microscopie confocale est un procédé optique qui réalise des images de très faible profondeur de champ et permet de produire des coupes de l'échantillon observé. Pour permettre cette précision, un marquage par fluorescence et une source laser sont couramment utilisés. Traditionnellement, ces microscopes produisent des piles de coupes en positionnant le plan focal à diverses positions.

Le principe du microscope confocal mis en oeuvre était de garder l'objet observé (typiquement des cellules) en suspension dans un fluide grâce à un champ électromagnétique. Le principal avantage est de préserver l'intégrité de la cellule vivante et donc de fournir une meilleure interprétation de ses mécanismes fonctionnels aux biologistes. Ainsi, c'est l'échantillon qui tourne dans le plan focal grâce à un champ électro-magnétique. De plus, le grand avantage de cette technique est d'obtenir une très bonne résolution en 3D sur les trois axes.

Le champ permet d'imposer des micro-rotations à la cellule, et ainsi de faire varier la vue de coupe imagée au niveau du plan focal. Les images acquises sont donc globalement positionnées en rotation autour d'un axe passant par le centre de la cellule, comme représenté dans la figure 3.1. Deux problèmes sont cependant à résoudre : le positionnement précis des coupes, et l'estimation du volume 3D. Ces images de coupe sont plus précisément des transformations par convolution d'une réalité 3D. Ainsi, plus qu'une simple interpolation, ce deuxième problème est une reconstruction 3D à partir de projections. Par ailleurs, les images peuvent être légèrement dégradées par un flou résiduel [174].

Approche de reconstruction 3D L'approche proposée consiste à estimer alternativement position des coupes et reconstruction 3D, de manière itérative via une procédure *Expectation-Maximization*. Précisément, les étapes suivantes sont effectuées tout à tour :

- **[Espérance]** estimation de la position des plans de coupes connaissant la reconstruction 3D courante ;

- **[Maximisation]** estimation de la luminance de tous les voxels du volume 3D (et donc y compris les données latentes non imagées directement), connaissant les vues en coupe et leurs positions, grâce à la fonction d'étalement de flou (*Point-Spread Function* (PSF)) du microscope mesurée au préalable. En effet, pour les points non imagés directement, soit dans le "trou noir" soit entre deux plans de coupe, la reconstruction intègre l'information des voxels dans un voisinage défini par le support de la PSF. Cela assure la cohérence globale et le transfert de l'information effectivement mesurée quand celle-ci est disponible.

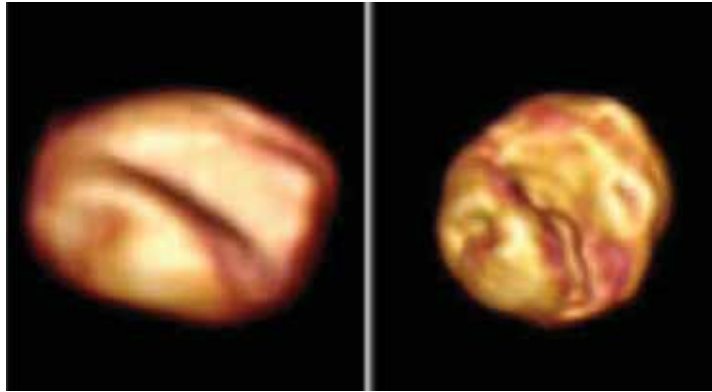


FIGURE 3.2 – Reconstructions 3D de cellule (l'axe optique est horizontal) : à gauche par microscope avec coupes selon l'axe vertical en Z, à droite par microscope avec microrotations et reconstruction avec PSF. Source : [RI-1].

La reconstruction avec imagerie standard à partir d'une pile d'image déforme la cellule et produit un résultat peu précis, même avec défloutage. Au contraire, la prise en compte itérative de la PSF permet une reconstruction nette en tout point du volume, y compris les points non imagés directement, comme le montre la figure 3.2. De nombreux détails sont révélés en utilisant cette technique d'imagerie et reconstruction. La figure 3.3 met en évidence le gain de précision dans la visualisation de la cellule : la lamina (réseau protéique fibreux) sur la face interne de l'enveloppe, l'invagination de l'enveloppe cellulaire dans le noyau et la présence d'éléments caractéristiques de la dynamique interne de la cellule (marqués par des flèches).

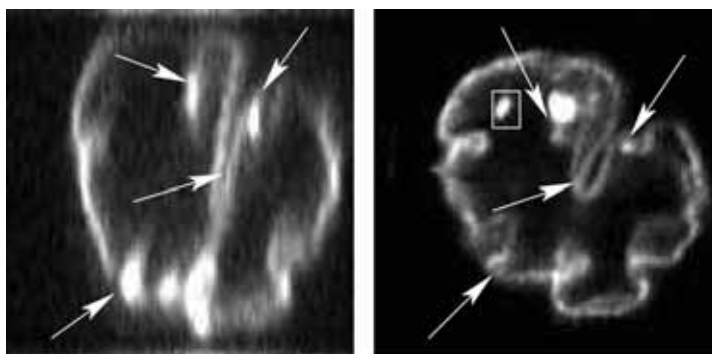


FIGURE 3.3 – Coupes selon le plan xz (l'axe optique est vertical) du volume estimé 3D : à gauche par microscope avec coupes selon l'axe vertical en Z, à droite par microscope avec micro-rotations et reconstruction avec PSF. Source : [RI-1].

3.2 Estimation 3D mono-image

Aperçu À bien plus grande échelle qu'en partie 3.1, et de manière similaire à la reconnaissance visuelle du contenu d'une image (partie 1.3), je cherche également à concevoir des méthodes capables de prédire la 3D d'une scène à partir d'une seule vue par apprentissage statistique, tel qu'illustré par la figure 3.4. Les applications concernent ici plutôt la robotique et touchent à la photographie computationnelle et à la co-conception de capteurs innovants. En utilisant des optiques à faible profondeur de champ, Pauline Trouvé-Peloux et moi avons proposé des approches par auto-encodeur profond [CN-4] ou plus récemment dans le cadre de la thèse de Marcela Carvalho (soutenue en 2019, dirigée par Frédéric Champagnat, ONERA, et André Almansa, CNRS) par réseaux convolutifs, et notamment par apprentissage adversaire [CI-40, CI-41, CN-7, CN-11].



FIGURE 3.4 – *Deep Depth from Defocus* : estimation de profondeur mono-image par flou de défocalisation et apprentissage profond, avec mesure de l'incertitude de prédiction du modèle. Source : [CI-40].

Contexte L'estimation de la 3D locale d'une scène a de nombreuses applications pour les interactions homme-machine, la réalité virtuelle ou la robotique. De nombreuses approches standards existent, telles que la stéréo-vision [175], la lumière structurée (utilisée pour la *kinect*) ou la *structure-from-motion* [176], mais elles requièrent soit des capteurs encombrants ou coûteux, soit plusieurs vues de la scène.

Des approches mono-vues statistiques [177, 178] sont apparues dans les années 2000 : à l'instar de Make3D [179], elles exploitent diverses propriétés de l'image (planarité, colinéarité, etc.) combinées avec une modélisation de la scène par modèles de Markov. Plus récemment, grâce à l'introduction de données RGB-D massives telles que le jeu de données NYUv2 [68], des approches par apprentissage profond fondées sur des réseaux convolutifs [180, 69] ont permis un gain significatif de performances. Néanmoins toutes ces méthodes sont purement statistiques et n'incluent pas une mesure physique de l'environnement, au contraire de l'imagerie active ou même de la stéréo passive. Elles sont donc sensibles à de nombreux paramètres difficilement contrôlables, tels que la constitution du jeu d'entraînement, et il est risqué de s'y fier pour des applications critiques, telles que la perception pour la conduite autonome.

Par ailleurs, dans le domaine de la photographie computationnelle, plusieurs travaux ont étudié l'utilisation du flou de défocalisation - ou *Depth from Defocus* (DFD) - pour déduire la profondeur, en commençant par [181] et plus récemment [182, 183]. En effet, le flou de l'image est directement lié à la distance au plan focal. En utilisant une optique à faible profondeur de champ qui amplifie le flou sur une grande part de la zone imagée, il est alors possible d'estimer la profondeur en utilisant un indice physique. Néanmoins, cette approche est pénalisée par une ambiguïté par rapport au plan focal (en deçà ou au-delà) et une zone aveugle à l'estimation de

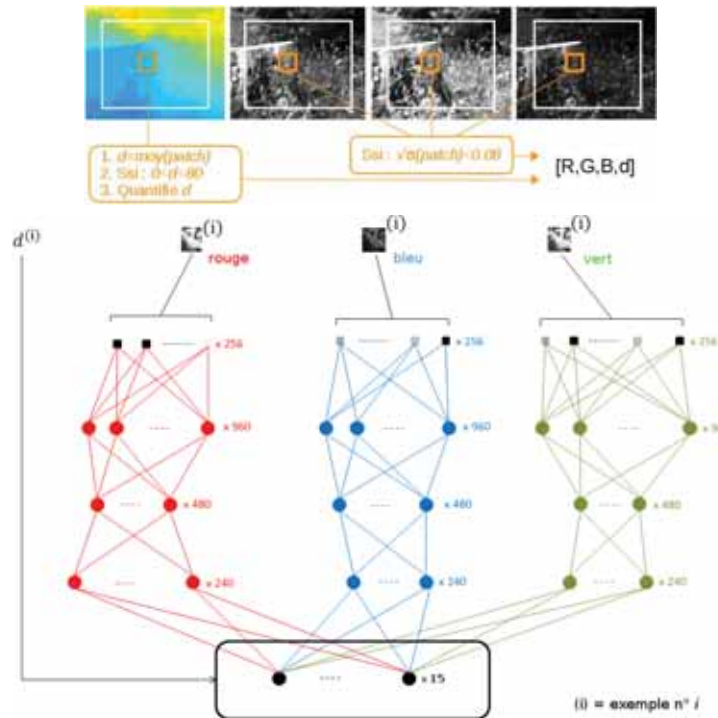


FIGURE 3.5 – Schéma de sélection de fenêtres (*patch* $[R, G, B, d]$ dans un triplet d’images RGB chromatique et la carte de profondeur obtenue par stéréo (en haut) et architecture globale du DBN à 3 canaux avec contrainte de parcimonie pour encoder l’information locale (en bas). Source : [CN-4].

profondeur dans les zones nettes de l’image. De plus, la DFD nécessite un modèle de scène et un calibrage de flou pour la relier à une valeur de profondeur, ce qui réduit son emploi pour des environnement inconnus.

Estimation 3D mono-vue par machines de Boltzmann restreintes Nos premiers travaux sur ce thème visaient à remplacer l’estimation analytique de la profondeur de DFD par une prédiction rapide par réseau de neurones. Nous avons alors proposé un apprentissage par machines de Boltzmann restreintes - ou RBMs - des informations de défocalisation issues d’un capteur DFD [CN-4]. Plus précisément, le réseau employé était un réseau de croyances profond - ou DBN- qui consiste en trois auto-encodeurs de type RBM suivis par une étape de classification de la profondeur discrétisée. Chaque couche de RBM est entraînée en minimisant l’erreur de reconstruction avec une contrainte de parcimonie pour l’encodage latent, et la couche de classification optimise les probabilités de classe issues d’un softmax.

Sur le plan optique, le prototype de caméra choisi permet la DFD chromatique [184, 185], c’est à dire que chaque canal RGB a un plan focal différent, ce qui permet de minimiser la zone aveugle pour la glsdfd et de lever l’ambiguïté de position par rapport au plan focal, en combinant les information des trois canaux. Pour en tirer parti, notre réseau traite donc chacun des canaux séparément avant d’en combiner l’information extraite avec l’information de profondeur au niveau du softmax, tel qu’indiqué sur le graphique de la Figure 3.5

Cette approche de co-design de caméra DFD et réseau d’estimation de la carte de profondeur

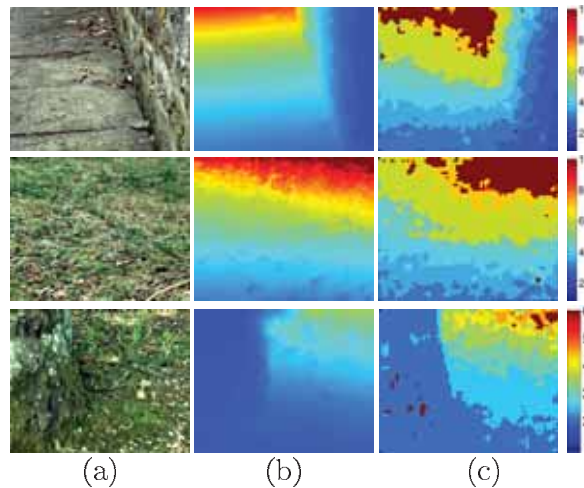


FIGURE 3.6 – Reconstructions de cartes de profondeur par DBNs : (a) image couleur, (b) carte de profondeur stéréo (référence), (c) DBNs parcimonieux. Source [CN-4]

a été testée sur des scènes en champ proche ($d \leq 10\text{m}$) en validant la mesure par une acquisition stéréo. Quelques résultats sont présentés sur la figure 3.6 et démontrent la preuve de concept.

Estimation 3D mono-vue par réseaux convolutifs Nous avons poursuivi ces travaux lors de la thèse de Marcela Carvalho. L’objectif est toujours l’estimation de profondeur mono-image par apprentissage profond et flou de défocalisation ou *Deep Depth from Defocus (Deep-DFD)*. Cependant l’approche est ici fondée sur les réseaux convolutifs et vise à bénéficier du passage à l’échelle de l’apprentissage sur de grandes bases d’images.

Les contributions sont multiples : (1) la proposition d’un modèle convolutif adversaire performant pour l’estimation de profondeur, et une étude approfondie de l’influence de la fonction de pénalité; (2) apport de la DFD; (3) apport de l’informations sémantique.

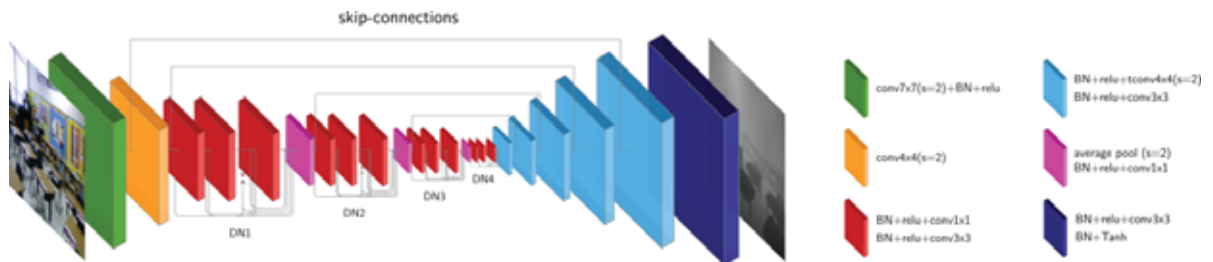


FIGURE 3.7 – Architecture *Deep Depth-from-Defocus Network* (D3-Net) pour l’estimation de carte de profondeur. Source : [CI-41].

(1) **Modèle convolutif pour l’estimation de profondeur.** Nous avons proposé d’une part une nouvelle architecture de réseau de neurones dédiée à l’estimation de profondeur à partir d’images génériques : *D3-Net*. Nous avons montré comment des choix particuliers de fonctions de coût et d’apprentissage affectent les performances de la prédiction de profondeur. Les fonctions de régression standard, comme l’erreur absolue \mathcal{L}_1 ou quadratique \mathcal{L}_2 ou leurs composées, ont été comparées à des fonctions de coût spécifiques telles que des pénalités dédiées

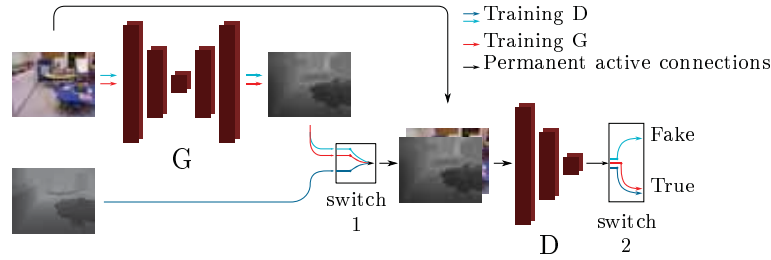


FIGURE 3.8 – D3-Net avec pénalité LS-GAN, où le générateur G est le réseau de la figure 3.7. Source : [CI-41].

à l'estimation de profondeur \mathcal{L}_{eigen} [180, 69] (définies en détails dans le tableau 3.1). Surtout, nous avons, à notre connaissance, été les premiers [CN-7] avec [186] à proposer des pénalités adversaires pour améliorer le réalisme des cartes de profondeur générées. Spécifiquement, nous avons proposé des réseaux génératifs adversaires conditionnels (C-GAN) [187] qui évaluent le réalisme du couple image-profondeur associée. Nos études ont permis de mettre en évidence l'apport des modèles adversaires, qui permettent d'obtenir un gain de performances notamment avec un très grand nombre d'exemples. Cette analyse a par ailleurs permis d'atteindre en 2018 le trio de tête de l'état de l'art en estimation de profondeur monoculaire [CI-41], au niveau de [186] et [188] sur le jeu de données NYUv2 [68]. La figure 3.7 présente l'architecture D3-Net tandis que la figure 3.8 illustre la stratégie d'apprentissage conditionnel adversaire, où la fonction de pénalité est elle-même apprise par un réseau cherchant à discriminer les cartes de profondeurs prédites de cartes réelles.

Fonction de coût		Équation
Absolue moyenne	\mathcal{L}_1	$\frac{1}{N} \sum_i^N l_i $
Quadratique moyenne	\mathcal{L}_2	$\frac{1}{N} \sum_i^N (l_i)^2$
Invariante à l'échelle [180]	\mathcal{L}_{eigen}	$\frac{1}{N} \sum_i^N d_i^2 - \frac{\lambda}{2N^2} (\sum_i^N d_i)^2$
Invariante à l'échelle avec gradients [69]	gra- $\mathcal{L}_{eigengrad}$	$\frac{1}{N} \sum_i^N d_i^2 - \frac{\lambda}{2N^2} (\sum_i^N d_i)^2 + \frac{1}{N} \sum_i^N [(\nabla_x d_i)^2 + (\nabla_y d_i)^2]$
BerHu [189]	\mathcal{L}_{berhu}	$\mathcal{L}_1(l_i)$ if $\mathcal{L}_1(l_i) \leq c$, else $\frac{\mathcal{L}_2(l_i)+c^2}{2c}$
Huber [189]	\mathcal{L}_{huber}	$\mathcal{L}_1(l_i)$ if $\mathcal{L}_1(l_i) \geq c$, else $\frac{\mathcal{L}_2(l_i)+c^2}{2c}$
Least Squared Adversarial [190, 187]	\mathcal{L}_{gan}	$\frac{1}{2} \mathbb{E}_{x,y \sim p_{data}(x,y)} [(D(x,y) - 1)^2] + \frac{1}{2} \mathbb{E}_{x \sim p_{data}(x)} [(D(x, G(x)) - C)^2] + \lambda \mathcal{L}_{L1}(G(x))$

Tableau 3.1 – Liste des fonctions de coût pour la régression. Soient y_i et \hat{y}_i la vérité terrain et la distance estimée en mètres, $l_i = y_i - \hat{y}_i$, $d_i = \log(y_i) - \log(\hat{y}_i)$, G , le réseau du générateur, D , le réseau discriminatoire et x , l'entrée RGB.

Par ailleurs, nous avons étudié comment remédier à la difficulté du problème en tirant parti des indices de profondeur disponibles dans le cas mono-image.

(2) **Apport de la DFD** La question sous-jacente qui se pose ici est d'étudier et de quantifier comment l'optique et le design du capteur pouvaient aider à l'estimation 3D. Nous avons pour cela créé des bases d'images avec flou de défocalisation afin d'entraîner D3-Net. Nous montrons que l'association d'images défocalisées avec un réseau de neurones permet à la fois de dépasser

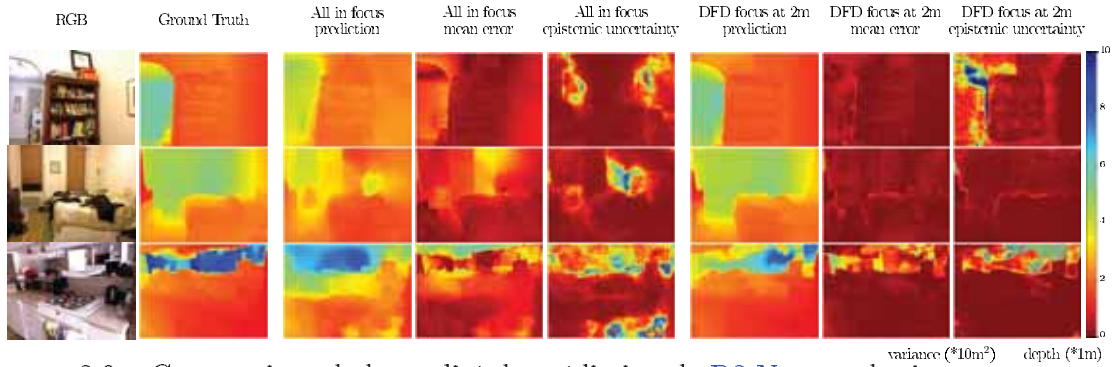


FIGURE 3.9 – Comparaison de la qualité de prédiction de **D3-Net** sur des images avec et sans flou de défocalisation (plan focal à 2m) : cartes de profondeur, erreur moyenne et incertitude épistémique [188] sur des données NYUv2 [68]. Source [CI-40]

les performances obtenues avec des images nettes mais également d'éviter les limitations classiques de la **DFD** (zone nette, ambiguïté proche / lointain). Nous avons également caractérisé l'incertitude du modèle en utilisant des réseaux de neurones Bayésiens [188]. Nous avons montré que la **DFD** permettait de réduire cette incertitude (le réseau génère une carte de confiance en sa prédiction) tout en obtenant des estimations plus précises (l'erreur de profondeur est réduite) [CN-11, CI-40] (voir Fig. 3.9).

Enfin, nous avons développé plusieurs prototypes de caméras avec une optique pour la **DFD**. L'estimation de profondeur est alors possible pour des scènes et des environnements jamais vus jusque là, selon une procédure de transfert d'apprentissage et *fine-tuning* illustrée Fig. 3.10

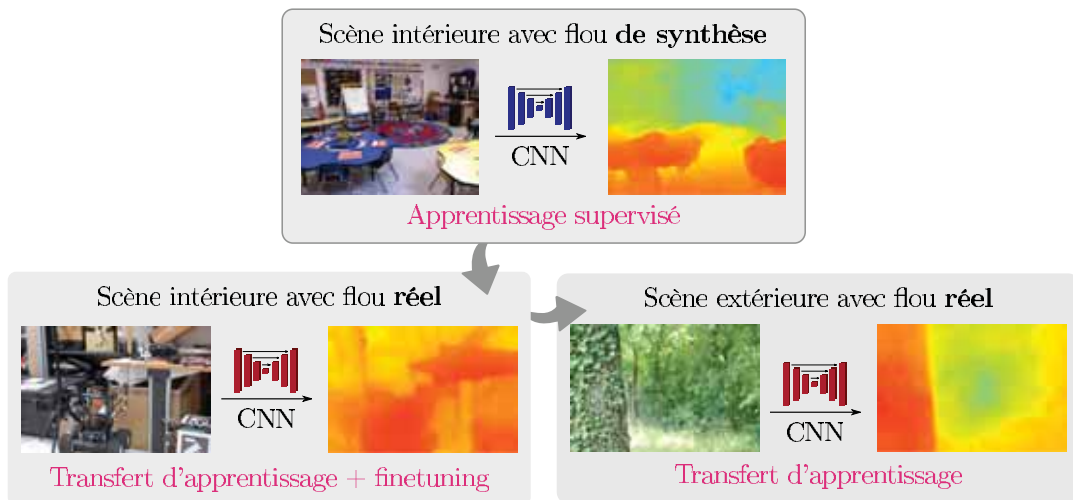


FIGURE 3.10 – Estimation de profondeur sur données floues synthétiques et réelles, en intérieur et extérieur. Les résultats montrent la flexibilité d'adaptation des modèles de **Deep-DFD** entraînés avec un jeu de données en intérieur avec flou synthétique, adapté par *fine-tuning* sur des données avec flou réel en intérieur, et finalement testés sur des scènes extérieures réelles sans ré-entraînement. Source [CI-40].

(3) **Apport de l'informations sémantique.** Enfin, nous avons montré que l'estimation de la 3D est également améliorée par la compréhension sémantique de la scène observée. Pour

cela, nous avons recours à des modèles d'apprentissage multitâche (*Multi-Task Learning (MTL)*) qui apprennent simultanément à estimer la 3D et prédire des classes sémantiques pour chaque pixel. Dans un cadre robotique, nous avons proposé une approche pour la compétition *3D Reconstruction Meets Semantics (3DRMS 2018)* [191]. Elle vise à la reconstruction 3D et à la classification sémantique de l'environnement d'un robot. L'approche que nous avons proposé combine estimation de profondeur mono-image avec *D3-Net*, raffinement de profondeur et segmentation sémantique 2D à partir de l'image et de la profondeur estimée, reconstruction 3D de la scène entière et enfin segmentation sémantique du nuage de points 3D par projection des classes 2D. En particulier, une de ses difficultés réside dans le protocole *Simulation to Reality (Sim2Real)* choisi : entraînement et validation sur données synthétiques, test sur données réelles. Notre approche a réalisé les meilleures performances en termes d'annotation sémantique 2D sur données synthétiques (grâce à l'estimation jointe de profondeur et sémantique) et annotation sémantique 3D sur données réelles après *fine-tuning* [192]. Dans la partie 4.2, nous verrons également qu'il est possible de combiner deux tâches d'estimation 3D et sémantique dans un cadre télédétection. La 3D est alors le modèle numérique d'élévation et la sémantique consiste en une classification du sursol. Nous montrerons que le modèle multitâche améliore simultanément les performances des deux tâches.

3.3 Segmentation sémantique de scènes 3D

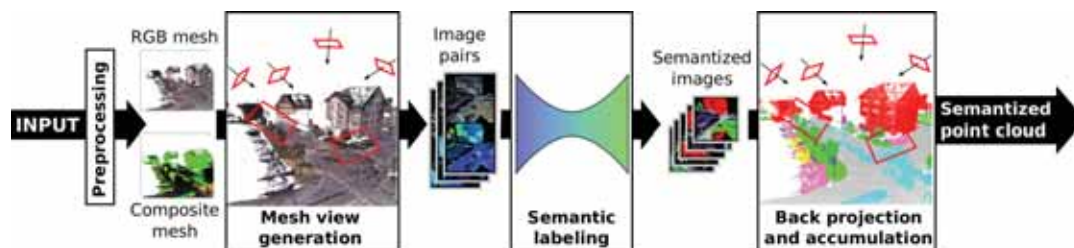


FIGURE 3.11 – Algorithme SnapNet : génération de vues virtuelles à partir du nuage de points, segmentation sémantique en 3D, re-projection en 3D pour classification du nuage de points. Source : [RI-10].

Aperçu Enfin, je cherche également à concevoir des approches pour la **compréhension sémantique de scènes en 3D**. En termes applicatifs, c'est l'extension naturelle des travaux sur la compréhension du contenu des images 2D du chapitre 1. Au sein de ce chapitre 3 sur la vision 3D, ces travaux viennent compléter ceux sur la compréhension géométrique de scènes des parties 3.1 et 3.2. En particulier, j'ai proposé avec Alexandre Boulch et nos collaborateurs des approches pour la sémantisation (ou classification) de nuages de points 3D, issus de capteur **LiDAR** ou de photogrammétrie. L'algorithme SnapNet est un réseau convolutif multivue pour les points 3D, qui procède par échantillonnage d'images 2D du nuage de points, classification dense en 2D, reprojection des prédictions en 3D, et vote [CI-30, RI-10]. Ces approches ont été appliquées avec succès à la cartographie urbaine (*benchmark semantic3D* [193]¹) et à la classification de bâtiments détruits pour aider les secours lors d'opérations de recherche et sauvetage (projet FP7 Inachus).

1. Voir <http://www.semantic3d.net/>

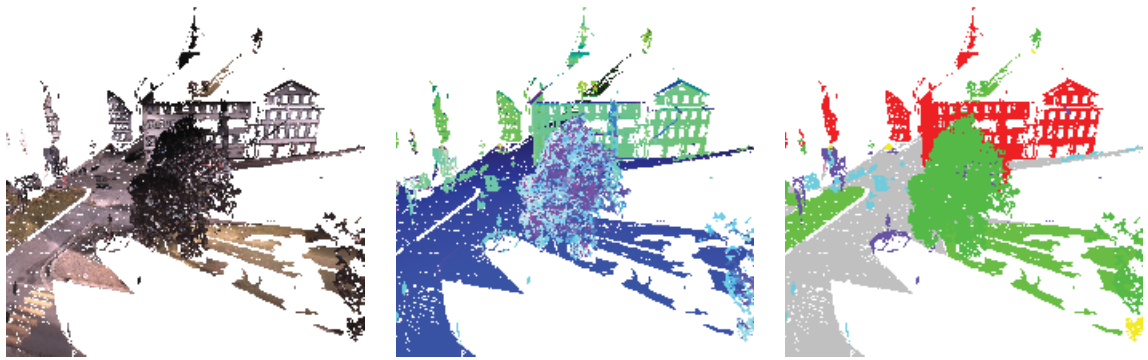


FIGURE 3.12 – Vues virtuelles générées par SnapNet : image couleur, carte d’information géométrique, vérité-terrain. Source : [RI-10].

Contexte Les nuages de points 3D s’imposent peu à peu comme un mode de représentation standard de scènes réelles 3D, bien que des alternatives existent : maillages par exemple [194]. Cela est dû notamment à des moyens d’acquisitions (capteurs lasers de type LiDAR ou photogrammétrie) pour lesquels c’est le format natif, à l’essor de domaines tels que les véhicules autonomes, et à la compacité inhérente au modèle, qui ne stocke que l’information utile des objets de la scène. Pour comprendre la scène, il est donc nécessaire de traiter ces points 3D et identifier à quoi ils correspondent [195]. On retrouve alors les tâches usuelles de la vision par ordinateur : segmentation sémantique (attribuer chaque point du nuage à des classes prédéfinies), détection et localisation d’objets 3D, etc. Au contraire des images, les nuages de points sont des ensembles non-structurés, invariants par permutation des éléments, et particulièrement épars.

Nous nous sommes intéressés en particulier à la segmentation sémantique de nuages de points. Pour cette tâche, on dispose généralement de nuages de points où chaque point est associé à une étiquette pour l’apprentissage, et l’objectif est d’étiqueter de nouveaux nuages de points. La première question qui s’est posée est de savoir comment traiter un nuage de points 3D ? En effet, le domaine est beaucoup plus ouvert qu’en 2D et plusieurs approches co-existent. Un échantillonnage avec des éléments en 3D conduit aux approches par voxels et convolution 3D, telles que VoxNet [196] ou OctNet [197]. Elles sont une transposition directe des modèles convolutifs 2D pour le traitement. Elles héritent également des représentations utilisées en robotique, les cartes d’occupation de l’espace de type Octomap [198]. Cependant utiliser des représentations 3D denses pour des données éparses n’est pas optimal et aboutit à un compromis entre la mémoire utilisée et la finesse de résolution atteinte. Un échantillonnage 1D conduit à traiter les points directement, avec des modèles sans convolution, comme les perceptrons multicouches de PointNet [199]. L’inconvénient est la perte d’information de contexte, que même les variantes comme PointNet++ [200] ne parviennent pas à compenser pour les grandes scènes. Une tendance récente vise à réintroduire l’idée de représentation locale et l’équivariance en translation en proposant diverses convolutions éparses adaptées au points de l’espace, comme par exemple [172].

SnapNet Notre approche repose sur un échantillonnage 2D de l’espace 3D du nuage de points, simplement en générant diverses vues de la scène. Cela permet notamment de transformer notre problème en un problème de segmentation sémantique 2D pour lequel des outils performants existent, tels que les réseaux entièrement convolutifs [54, 55]. En détails, la méthode se décom-

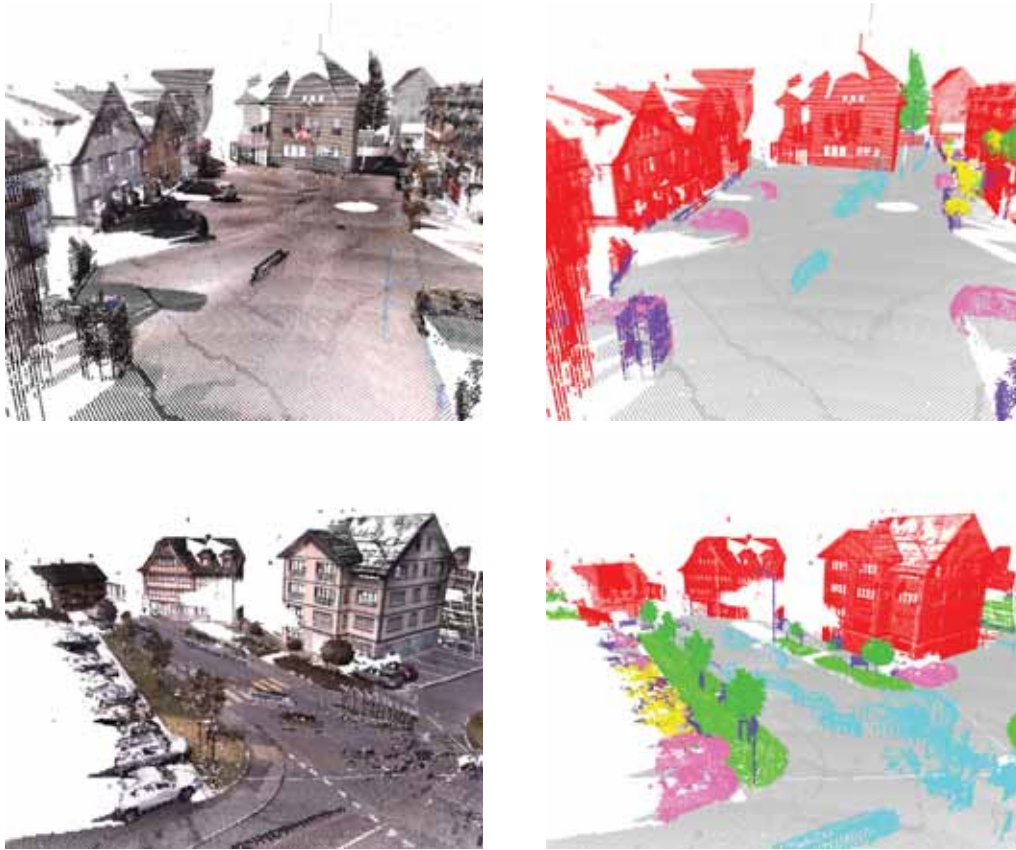


FIGURE 3.13 – Résultats de segmentation sémantique 3D par SnapNet : (gauche) nuage de points 3D colorés (scènes de Semantic3D [193]) et (droite) nuage segmenté selon les classes bâtiment, route, terrain naturel, véhicules, mobilier urbain, végétation haute et basse, artefacts de scan.

pose comme suit (voir aussi figure 3.11) :

- Une étape de *pré-traitement* consiste à calculer des descripteurs locaux de points (normales ou bruit local) et générer un maillage du nuage de points ;
- *Génération des vues* : à partir du maillage, des vues sont générées en prenant diverses positions de caméra. Deux types de vues sont considérés : des images RGB pour l'apparence, notamment quand le nuage de point est coloré, ainsi que des cartes locales de descripteurs géométriques ;
- La *segmentation sémantique 2D* produit des cartes de segmentation 2D à partir des deux vues, par exemple avec un réseau SegNet [55] ;
- Enfin, on re-projète sur le maillage 3D les étiquettes obtenues. Les étiquettes de points 3D sont attribuées par un vote majoritaire qui permet de régulariser les incohérences de prédiction.

Un aspect fondamental de notre approche par rapport à d'autres approches multivues issues de l'informatique graphique comme MV-CNN [201] (12 vues) ou Panorama [202] (1 vue sur

le cylindre englobant) est d'utiliser un très grand nombre de vues (de l'ordre de 10^2 ou 10^3). Cela permet donc de traiter des scènes complètes et non des objets isolés, et surtout d'avoir un échantillonnage fin et complet de la scène observée. Par ailleurs, cette randomisation évite un biais de sélection lié à l'algorithme, et le grand nombre d'échantillons aléatoires ainsi que la procédure de vote permettent d'être robuste aux échantillons non-significatifs et aux erreurs de classification. Cependant, des a priori liés au type de scène à traiter peuvent être incorporés dans l'échantillonnage pour induire des invariances aptes à aider la classification : par exemple pour les scènes urbaines de Semantic3D, des vues proches et lointaines étaient générées pour chaque position de caméra, de manière à inclure plusieurs niveaux de contexte (voir figure 3.12).

Des résultats de segmentation sémantique 3D avec sont présentés sur la figure 3.13 : les principales classes urbaines sont retrouvées et les artefacts d'acquisition sont identifiées. SnapNet a obtenu en 2017 les meilleures performances de l'état de l'art sur le *benchmark* large-échelle Semantic3D en 2017, avec des taux de classification globale de 91% et de plus de 60% de *mIoU*. Deux ans après, l'algorithme est toujours dans le top 5², face à des approches de conception très différente : en effet, les méthodes les plus prometteuses aujourd'hui cherchent à généraliser la convolution directement à des voisinages 3D, et non plus plus utiliser la 2D comme *proxy* au contexte géométrique et d'apparence local. Enfin, cette approche a été utilisée dans le cadre du projet européen FP7 Inachus³ pour fournir des des cartes d'évaluation des dégâts⁴ aux équipes de sauvetage et recherche intervenant suite à une catastrophe (voir figure 3.14) [RI-10, CI-31].

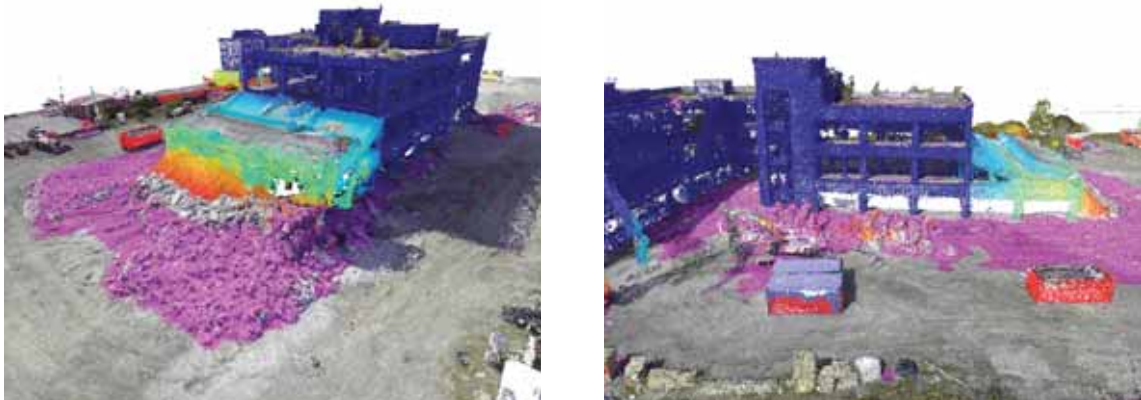


FIGURE 3.14 – Résultats de segmentation sémantique 3D par SnapNet pour les opérations de Recherche et Sauvetage : des cartes d'évaluation des dégâts aux bâtiments suite à une catastrophe (tremblement de terre, explosion) permettent de guider les équipes de secours vers les zones où ils peuvent retrouver et aider des survivants. Chaleur de la couleur proportionnelle au niveau de dégâts, depuis le bleu (intact) au violet (débris).

2. Voir http://semantic3d.net/view_results.php?chl=1, consultée le 16.09.2019.

3. Voir <https://www.inachus.eu/>

4. Détection de débris (projet FP7 Inachus) : <https://www.youtube.com/watch?v=xT4VrtCu8Po>

3.4 Résumé et conclusions

Les travaux présentés dans ce chapitre abordaient le problème de la modélisation 3D à travers trois problèmes de reconstruction, estimation et sémantisation 3D. Dans un cadre applicatif différent de la plupart des travaux présentés ici, la partie 3.1 visait à reconstruire des modèles 3D de cellules en microscopie confocale. Une optimisation alternée du modèle 3D et de la position des vues échantillonnées permettait la reconstruction précise. Les deux parties suivantes couvraient les deux aspects de géométrie et de sémantique de la compréhension de scènes. La partie 3.2 vise à estimer la structure 3D d'une scène dans le cadre difficile de la vision monoculaire. Nous avons proposé une approche par réseaux de neurones convolutifs qui permet de générer des cartes de profondeur précises à partir d'une seule image et des systèmes avec flou de défocalisation (DFD) pour réduire l'incertitude de prédiction locale par l'intégration d'a priori optiques. La partie 3.3 part de la 3D pour classer chaque point 3D et parvenir à la segmentation sémantique de la scène. l'algorithme SnapNet repose sur une multiplication des points de vue sur la scène observée, échantillonnée par des images, afin de la comprendre dans son ensemble. Dans sa construction, il résume et préfigure plusieurs éléments d'une approche de compréhension de scènes comme la représentation duale image et nuage de points 3D (défi *(viii)*) des perspectives du chapitre 4 ou l'analyse à échelles locale et globale.

Chapitre 4

Perspectives

Ce chapitre résume les conclusions des parties précédentes en partie 4.1, puis propose des perspectives de recherche en partie 4.2 et 4.3.

4.1 Conclusions et vue d'ensemble

Les travaux décrits dans les chapitres précédents s'ordonnaient selon trois axes. L'axe initial décrit au chapitre 1 concerne la compréhension du contenu sémantique des images, c'est à dire principalement la classification d'image, la détection d'objets et la segmentation sémantique. Trois cadres applicatifs ont été envisagés : le multimédia, la robotique volante et la compréhension de scènes de taille moyenne, la robotique terrestre et la compréhension de l'environnement immédiat du robot. À travers ces trois exemples est illustrée l'évolution des algorithmes qui ont permis une compréhension de plus en plus fine du contenu des images, jusqu'à une classification pixelique des objets et régions. En particulier, ces travaux intègrent au fur et à mesure l'a priori de structure 3D de la scène qui est à l'origine de l'image pour une meilleure compréhension de celle-ci. De fait, la connaissance du point de vue dans la géométrie de la scène (en partie 1.2) ou le décalage de point de vue comme moyen d'augmenter la prise de connaissance sur la scène (en partie 1.3) permettent de gagner en précision de sémantisation.

Le chapitre 2 rassemblait les contributions en matière de compréhension de la Terre vue du ciel. Ayant eu lieu à une période de forts changements dans ce domaine, ces travaux forment un corpus cohérent de résultats dans l'utilisation des réseaux de neurones profonds pour la télédétection, en tenant compte de ses spécificités : des capteurs variés (RGB mais aussi multispectraux et hyperspectraux) et des modes variés (imagerie mais aussi modèles d'élévation et cartographie). Les tâches visées s'appliquent à des domaines d'emploi particuliers qui prennent plus de sens immédiat qu'en vision par ordinateur : la segmentation sémantique est un moyen d'accéder à la cartographie géographique automatique (voir figure 4.1), la détection de changements entre images permet de suivre l'évolution urbaine ou environnementale. Deux tendances se dessinent. D'une part la multiplication des sources d'information et représentations d'une même scène, reliées par la géolocalisation. D'autre part, à nouveau la prise en compte de la 3D pour une meilleure compréhension des images (partie 2.2), et même l'évolution vers la 3D en tant que représentation de la scène (partie 2.3).

Les travaux précédents convergent donc vers la 3D, à laquelle était consacrée le chapitre 3 sur la vision 3D. Elle y est envisagée tout d'abord sous l'angle de la compréhension de la structure 3D. L'objectif est alors la reconstruction 3D, et plus précisément une estimation 3D de l'objet



FIGURE 4.1 – Évolution de la cartographie de l’Europe : (a) *Tabula Rogeriana* d’Al Idrissi (1154) ; (b) Carte de l’Europe de Beylet (1700) ; (c) Carte du réseau routier par traces GPS incluses dans OSM. (2018). Sources : © Gallica/Bibliothèque Nationale de France, les contributeurs Wikipedia, OpenStreetMap.

de l’observation car les approches proposées ont de moins en moins recours à des modèles géométriques exacts (parties 3.1 puis 3.2). Puis sous l’angle de la compréhension sémantique de cette structure 3D. Ce dernier travail sur la sémantisation de scènes représentées en 3D par un nuage de points colorés (partie 3.3) concentre plusieurs intuitions et techniques développées au cours de mes recherches : l’analyse de la scène en termes de géométrie et de colorimétrie ; la multiplication des points de vues amorcée au chapitre 1 et ici poussée à l’extrême avec l’introduction de l’aléatoire dans leur échantillonnage ; l’apprentissage de représentations locales et la régularisation par vote des prédictions au niveau global.

Dans la suite sont présentées les perspectives de mes travaux et mon projet de recherche. Il vise au développement de modèles d’apprentissage pour la compréhension de scènes perçues selon plusieurs modalités complémentaires, notamment image et 3D. Il s’articule selon deux niveaux de granularité : la compréhension de scènes à large-échelle (en partie 4.2) et la compréhension de l’environnement local (en partie 4.3).

4.2 Perspectives en compréhension de scènes à large-échelle

La *large échelle* désigne ici d’une part l’observation de la Terre, et en particulier l’étude globale d’un pays, d’un continent, voire de la planète entière. D’autre part, cela recouvre également la télédétection sur des scènes de taille moyenne (une ville, un quartier, etc.) quand elles sont représentées à une résolution très fine, générant alors des volumes de données considérables.

Cet axe regroupe des problématiques nouvelles qui émergent avec les évolutions de la télédétection vers une couverture globale, fréquemment renouvelée, et une résolution de plus en plus fine (de 30cm/pixel pour les satellites à 1cm/pixel pour un drone survolant un site d’étude). Le défi actuel est de répondre à la disponibilité toujours accrue de données d’observation de la Terre. Se pose alors la question de comment bénéficier de ces données pour mettre au point et entraîner les algorithmes d’apprentissage massifs tels que les réseaux de neurones profonds. De fait, si les données images sont nombreuses aujourd’hui, les annotations correspondantes sont toujours rares ou imprécises. Concrètement, les défis identifiés sont les suivants :

- (i) La couverture globale pose de manière très concrète le problème de la généralisation des modèles d’apprentissage, et du déploiement en tout site de modèles entraînés sur des zones géographiques limitées.

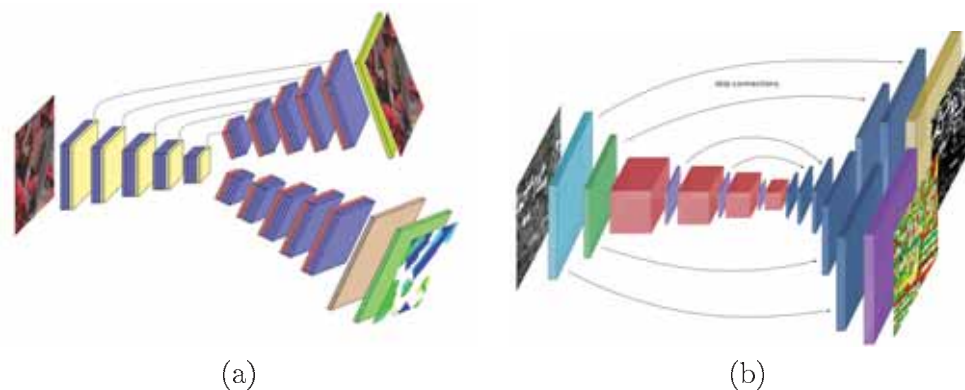


FIGURE 4.2 – Réseaux multitâches pour la compréhension de scènes à large échelle. (a) Architecture semi-supervisée pour la segmentation sémantique large échelle, capable d’apprendre sur des données annotées ou pas. (b) Architecture de modèle d’apprentissage multitâche pour la régression de hauteur et la classification sémantique. Sources : [CN-12, RI-20].

- (ii) Le géo-référencement permet de mettre en relation des données très variées. En tirer parti requiert des modèles pour l’analyse multimodale capables de gérer des données très hétérogènes : images, 3D, données vectorielles, texte, etc.
- (iii) Un corollaire des points précédents est le recours à des sources d’annotation automatique, des données ouvertes collaboratives ou encore des données synthétiques qui introduisent des erreurs (bruit d’étiquetage, de mise en correspondance erronée, ou de modélisation). Comment alors apprendre de manière robuste ?
- (iv) Enfin, au delà des images vues du ciel et des modèles numériques d’élévation 2.5D (de mieux en mieux résolus), le modèle de représentation est en passe d’évoluer vers la 3D globale à très haute résolution, ce qui amènera à définir et utiliser des modèles différents de ceux mis en oeuvre jusqu’ici.
- (v) Par ailleurs, en plus des domaines applicatifs déjà traités dans les chapitres précédents, l’aspect large-échelle permet également de contribuer à des études globales de surveillance de la planète pouvant servir aux objectifs de développement durable tels que définis par les Nations Unies¹ : faire en sorte que les villes et communautés soient durables, sans impact environnemental négatif ; gérer les ressources en eau salubre ; ou préserver les écosystèmes terrestres et surveiller la déforestation ou la désertification. Ces moyens d’observation peuvent également être mis au service d’objectifs de sécurité des populations et de gestion des crises internationales. Que mettre en oeuvre pour y parvenir ? Quelles applications responsables et utiles sont envisageables ?

Plusieurs pistes sont possibles pour apporter des solutions à ces questions.

4.2.1 À court terme : apprendre avec peu de données

Afin de tirer parti d’un faible nombre de données annotées et de données brutes abondantes, et par là même apporter des éléments de solution aux défis (i) et (iii), l’**apprentissage semi-supervisé** [203, 204] apparaît comme une stratégie possible pour entraîner des réseaux de

1. <https://www.un.org/sustainabledevelopment/fr/objectifs-de-developpement-durable/>

neurones. Nous explorons cette approche dans le cadre de la thèse de Javiera Castillo Navarro débutée en 2019, et qui fait suite à la thèse de Nicolas Audebert décrite au chapitre 2. À l’ONERA, nous avons commencé par constituer un jeu de données composé d’images aériennes THR associées à des cartes d’occupation des sols issues d’Urban Atlas, regroupant 16 cités françaises : MiniFrance. Nous l’avons utilisé pour mettre en évidence le problème de la généralisation des modèles de l’état de l’art en télédétection. En effet, si un modèle de classification entraîné sur ISPRS Vaihingen est très robuste à une baisse drastique des exemples (un modèle entraîné sur une seule des 16 images obtient le déjà très bon taux de classification de 80%), cela est dû à la forte homogénéité de ce *benchmark*. Au contraire, un modèle entraîné sur une ou même plusieurs villes de MiniFrance va avoir une grande variance selon le site de test de près de 20 points de pourcentage, en raison de la variabilité des apparences [CI-43]. Nous développons maintenant des architectures de réseaux de neurones pour l’apprentissage semi-supervisé, en combinant des tâches supervisées et non-supervisées, telles que de l’auto-encodage (par reconstruction de l’image d’entrée : voir figure 4.2-a) ou de la catégorisation non-supervisée. Les premiers résultats montrent que la stratégie semi-supervisée permet d’améliorer la classification [CN-12]. Pour parvenir au large-échelle, ces travaux se poursuivront par le développement d’architectures multimodales adaptées aux modalités et aux résolutions des images satellitaires à couverture globale (Pléiades, Sentinelle). Ils aborderont également les problèmes de la prédiction structurée, pour prédire des classes à différents niveaux de détails en fonction des ressources images disponibles, et de l’estimation du niveau de prédiction possible en fonction d’un lieu donné, pour savoir quelles cartes peuvent être produites et combien de données sont nécessaires pour cela.

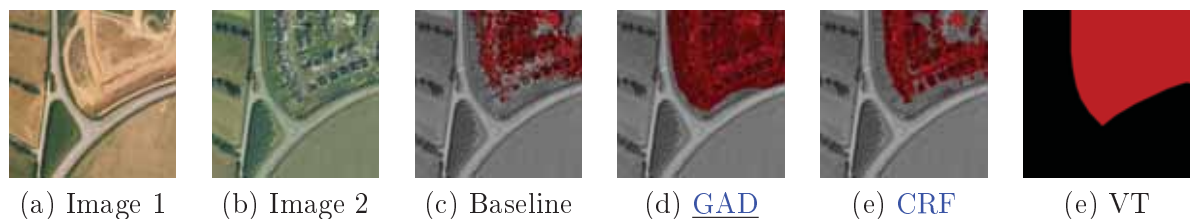


FIGURE 4.3 – Apprentissage faiblement supervisé pour la détection de changement. Cartes de changement obtenues par différentes approches : (c) supervisé ; (d) faiblement supervisé avec GAD ; et (e) régularisé par un CRF dense. (f) Vérité-terrain. Les changements sont marqués en rouge. Source : [CI-44]. .

Une autre approche pour répondre au défi (iii) des données peu fiables est l’**apprentissage faiblement supervisé** (ou *weak-learning* [205, 206, 207]). Dans le cadre de la thèse de Rodrigo Daudt, nous avons rencontré ce problème en traitant les données HRSCD dont les annotations viennent justement de sources créées automatiquement. Nous avons alors proposé une approche d’optimisation alternée du modèle de classification et des données de références [CI-44]. La mise à jour de la référence (les annotations) utilise la diffusion anisotropique [208] avec comme base l’image, de manière à adapter peu à peu la référence aux gradients et aux régions de l’image. La figure 4.3 montre un résultat de notre approche *Guided Anisotropic Diffusion* (GAD) comparée à d’autres approches supervisée ou de régularisation par *Conditional Random Field* (CRF) : elle parvient à prédire un changement qui correspond mieux à la réalité, malgré un apprentissage sur des annotations imparfaites.

Enfin, toujours pour le défi (iii) mais pour les problèmes de classification dus à une mauvaise

généralisation (défi (i)), nous reprenons dans la thèse de Gaston Lenczner débutée en 2019 l'idée de l'**apprentissage interactif** et itératif décrite dans le chapitre 1, partie 1.2, mais en l'adaptant aux réseaux de neurones profonds. Plusieurs questions se posent. D'abord, si par leur construction même les algorithmes de *boosting* utilisés dans [CI-13, CI-16, CI-18] se prêtent assez facilement à des variantes incrémentales, les réseaux de neurones sont des modèles plus complexes et donc plus lents à entraîner. Comment alors interagir avec des réseaux de neurones pour l'apprentissage? Une approche consiste à ajouter des dimensions virtuelles pour l'interaction lors de l'apprentissage et de tirer certaines annotations d'interaction au hasard depuis la référence [209, 210]. En déploiement, la prédiction prendra en compte les interactions réelles. Comment ensuite interagir avec l'utilisateur? Une piste est l'apprentissage actif, qui consiste à lui proposer des exemples qui seront informatifs pour l'algorithme, c'est à dire qui conduiront à un meilleur modèle. Si de nombreuses stratégies d'apprentissage actif existent pour les SVMs par exemple [211, 212], c'est encore un problème ouvert avec les CNNs.

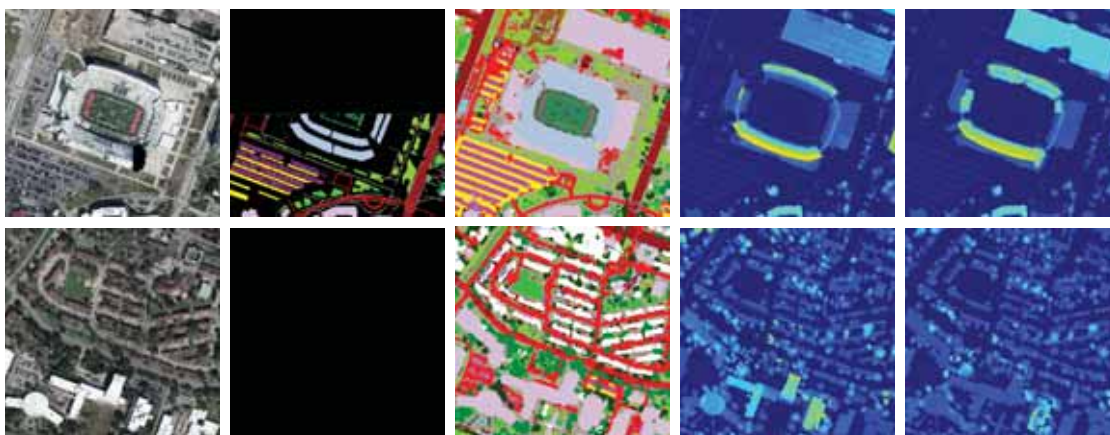


FIGURE 4.4 – Cartographie sémantique et prédiction de la 3D locale par apprentissage multitâche. De gauche à droite, donnée image RGB issue du DFC2018 sur Houston, référence sémantique et prédiction (noir signifie une absence d'information), référence d'élévation et prédiction. Source : [RI-20].

4.2.2 À moyen terme : apprendre avec des données géospatiales multiples

La problématique (ii) renvoie à la mise en commun de données multimodales et de données hétérogènes : images aériennes ou au sol, 3D, cartes, mais aussi bases de données géoréférencées, traces téléphoniques, données de l'"internet des objets". Le couplage de ces données peut se faire par référencement spatial : la géolocalisation commune est alors ce qui les met en relation. Cette mise en correspondance fait l'objet de plusieurs initiatives de standardisation, telle INSPIRE en Europe, dont il est possible de bénéficier [213, 214]. Cela amène à développer des algorithmes d'apprentissage fortement multimodaux (image et 3D, image et texte, etc), tant pour les entrées que pour les tâches à prédire. Les réseaux de neurones, par la plasticité de leurs architectures, sont une solution de choix pour cet apprentissage multitâche.

Plusieurs formes d'**apprentissage multimodal** avec des réseaux de neurones ont été proposées au chapitre 2, partie 2.2. L'**apprentissage multitâche** est une autre solution pour tirer parti des données multiples : deux exemples sont le réseau de détection de changement

sémantique présenté en partie 2.2 et le réseau d'apprentissage semi-supervisé présenté dans cette partie (4.2, voir figure 4.2-a). Dans le cadre de la thèse de Marcela Carvalho, nous avons également utilisé ces techniques pour prédire simultanément l'élévation locale et la cartographie du sol à partir d'images aériennes [RI-20]. Cela permet donc d'aborder également le défi (iv) sur la 3D. Nous avons proposé une architecture qui minimise simultanément une régression sur la hauteur donnée par le MNE et une fonction de pénalité de classification (entropie croisée), qui est représentée sur la figure 4.2-b. La figure 4.4 montre des cartes d'occupation du sol et de hauteur locale générées à partir des images RGB THR du DFC2018. Les réseaux produisent des hauteurs précises pour le sol, les zones résidentielles et la végétation, mais ont des prédictions plus variables sur les bâtiments plus élevés. Les cartes sémantiques sont détaillées. Surtout, nous avons montré que l'apprentissage multitâche permettait d'améliorer les performances pour les deux tâches par rapport aux modèles simples.

Les exemples précédents confirment que la piste multimodal / multitâche est prometteuse. Cependant, le nombre d'entrées ou d'objectifs est encore limité. Nous avons donc en projet de tester l'hypothèse d'un apprentissage multiple massif et d'en mesurer l'apport.

Tout d'abord dans le cadre du post-doctorat de Clément Rambour, débuté en 2019 entre l'ONERA (Élise Koeniguer et moi-même) et le CNAM Paris (Michel Crucianu, Nicolas Audebert et Mihai Datcu). L'objectif est l'analyse multitemporelle multimodale de séries SAR et optiques (Sentinelles 1 et 2), à des fins de suivi de l'activité à la surface du globe, et notamment des catastrophes naturelles telles que les crues et inondations, les feux de forêts, etc. Cela s'inscrit donc également dans le défi (v). La première question est de savoir comment utiliser les deux sources d'information de manière optimale. De fait, l'imagerie SAR traverse les nuages et est souvent plus adaptée pour détecter des changements et variations entre images. En revanche, ces changements sont plus aisément caractérisables dans des images optiques. Une possibilité est d'utiliser des approches de détection de changement en SAR [215] pour la constitution de bases de données multimodales. Alors, des réseaux de détection de changement sémantique peuvent être entraînés pour caractériser l'activité. Une deuxième question consiste à être capable de prévoir l'activité et les risques. Les réseaux récurrents sont un moyen d'apprendre des motifs spatio-temporels et ainsi comprendre l'évolution des espaces urbains et naturels.

4.2.3 À plus long terme : vers une analyse globale de la Terre en 3D

Porté à grande échelle, les outils développés dans le cadre de ces travaux permettront une véritable analyse géospatiale d'un pays ou d'un continent. Un tel *jumeau numérique* de la planète permettrait de modéliser son état et son évolution, et prendre les bonnes décisions quant à l'aménagement du territoire, l'environnement ou son développement économique. Deux projets allant en ce sens sont maintenant détaillés.

Tout d'abord, un projet d'apprentissage automatique pour l'**analyse géospatiale multimodale** (aérienne, streetview et texte) qui aborde le défi (ii). De fait, de plus en plus de données géolocalisées sont produites à chaque instant : imagerie vue du ciel bien sûr, mais aussi photos depuis nos téléphones, commentaires sur des lieux et des commerces, textes relatifs à un lieu ou un événement, trajectoires *Global Positioning System* (GPS), traces de connexion de téléphone, etc. De nouvelles applications peuvent alors être imaginées. Par exemple, où est-ce que cette photo de rue a été prise [216] ? Cela peut être utile pour l'égo-localisation pour la conduite autonome, mais aussi pour contrôler et désambiguïser des *fake news* si la photo est incohérente avec sa légende... Ou par ailleurs, qu'est-ce que l'on voit depuis le ciel, étant donné une image aérienne et du texte géolocalisé décrivant cet endroit (par exemple un extrait de

Wikipedia) [217] ? Cela permet une classification de l'occupation du sol plus précise que la télédétection classique. Enfin, comment décrire un site vu du ciel ou avec quelques photos ? Similaire au légendage d'images ou de vidéos, le légendage d'endroit (*place captioning*) est un moyen de rendre la classification du sol intelligible à tout à chacun.

Pour y parvenir, plusieurs problèmes doivent être résolus. Comment concevoir des algorithmes pour des données hautement hétérogènes ? Les réseaux de neurones sont une piste vraisemblable, étant donné leur capacité prouvée pour le multimodal / multitâche. Cependant, ils doivent être étendus pour intégrer des données éparées [218], et de nouveaux modes comme le texte [219, 220]. Comment construire un espace de représentation adapté à des tâches très variées ? Projeter des données multiples dans une unique espace de représentation commun crée un encodage riche, qui doit être cependant élaboré soigneusement afin d'être optimal et générique [221].

Par ailleurs, la **3D à large échelle** (défi (*iv*)) est en soi un projet complexe et ambitieux : quelles méthodes développer pour avoir les mêmes fonctionnalités que dans les plateformes d'aujourd'hui en 2D (par exemple *GoogleEarthEngine* ou *Copernicus Data Hub*) et en développer de nouvelles ? À partir de 2022, les satellites de la **Constellation Optique en 3D (CO3D)** fourniront une imagerie 3D à 50cm/pixel de résolution spatiale et une précision en altitude de 1m. Le projet AI4GEO porté par un consortium incluant des instituts de recherche tels que le **CNES**, l'**IGN** et l'**ONERA** visera à développer les outils pour la production automatique d'information géospatiale 2D et 3D à l'échelle de la planète. Il y a donc besoin de reconstruire des surfaces 3D, et d'extraire et classer les objets 3D. D'ailleurs, des maintenant, la constellation TanDEM-X² fournit des modèles d'élévation mis à jour régulièrement. Cela soulève le problème de l'analyse de données 3D multitemporelles.

Comment en effet avoir une segmentation sémantique de données 3D performante pour ces sources qui sont différentes des **LiDAR** décrits en partie 3.3 ? Les mêmes approches sont-elles directement transposables ? Les résultats du **DFC2019** [RI-21] donnés en partie 2.3 indiquent que la réponse est sûrement affirmative, au moins pour certaines d'entre elles, mais cela doit être confirmé et élargi aux différents types d'approches de segmentation sémantique 3D. Ensuite, comment identifier et caractériser les changements pour des acquisitions à différentes dates ? Le problème soulevé ici est la variabilité des données, dans la scène ou due au mode d'acquisition : les points échantillonnés changent d'une acquisition à une autre, y compris lorsqu'ils échantillonnent la même surface. Différentes stratégies seront à étudier, directement sur les données, ou bien selon des méthodes orientées objet après une première phase de sémantisation. Dans le premier cas, et si les approches par réseaux de neurones sont privilégiées, la structure globale des réseaux présentés en partie 2.2 est vraisemblablement toujours valide, mais les réseaux en eux-mêmes devraient être adaptés au traitement de données 3D, par exemple avec des modules PointNet [199] ou des convolutions 3D [172, 222].

4.3 Perspectives en compréhension de l'environnement local

La définition d'une scène à l'échelle locale est restée à peu près la même sur la période couverte par ces travaux : une pièce ou l'étage d'un bâtiment en intérieur, une portion de terrain ou une route en extérieur. En revanche, les données enregistrées de cette scène ont considérablement changé : d'une seule image à plusieurs et à la vidéo, puis à la 3D indirectement (par reconstruction) ou directement par **LiDAR**.

2. <https://www.dlr.de/dlr/en/desktopdefault.aspx/tabid-10378/>

Dans le domaine de **la compréhension de scènes locales**, mon objectif est de développer des approches pour répondre aux problématiques soulevées par une modélisation en 3D de la scène d'une richesse et d'une précision inégalées. Longtemps, le paradigme de Marr [223] a constitué l'agenda de la recherche en vision par ordinateur. En bref, ce modèle computationnel [224, 225] pour le traitement et la représentation de l'information visuelle comprend trois niveaux : extraction de caractéristiques 2D (*primal sketch*), calcul d'une représentation 2.5D égo-centrée, représentation 3D avec décomposition en formes organisées dans l'espace, préalable à la reconnaissance. De fait, la scène en 3D, qui était un objectif de nombreux travaux de vision par ordinateur, est aujourd'hui grâce aux nouveaux capteurs, **LiDARs** ou caméras optiques monoculaires de la partie 3.2, le point de départ de l'analyse et des traitements possibles pour la sémantisation. Cela crée un contexte ouvert et passionnant de par la convergence d'approches issues de la vision par ordinateur, de la robotique et de l'informatique graphique, cette dernière étant habituée depuis longtemps à manipuler ce type d'objet informatique.

Cet axe regroupe donc les problématiques qui apparaissent avec l'arrivée de la vision 3D dans le domaine de la compréhension de scènes :

- (vi) Quelle représentation pour la 3D ? Cette représentation peut dériver des outils de production 3D, de choix de stockage (besoin de compacité), mais également être choisie de manière à être optimale en fonction des tâches à accomplir (sémantisation) ou des applications (compréhension de l'environnement, réalité virtuelle, etc.). Cela pourrait être les nuages de points 3D (standard d'acquisition par **LiDARs** ou photogrammétrie dans le cas du **SLAM**), des modèles de surface ou des maillages issus de l'informatique graphique [226, 194], des cartes de rendu et de normales [227] ou des modèles voxeliques (composés de voxels, c'est-à-dire des volumes 3D élémentaires) hérités des cartes d'occupation de l'espace de la robotique...
- (vii) Se pose également le problème du traitement de cette représentation 3D. Au contraire du traitement des images pour lequel les réseaux convolutifs recueillent un consensus actuellement, le traitement de la 3D est beaucoup plus ouvert en raison notamment de son cadre multidisciplinaire.
- (viii) Comment traiter des données multimodales dans le cadre de la vision 3D ? Un véhicule autonome (par exemple une voiture telle qu'utilisée pour le *benchmark* KITTI [228]) est bardé de capteurs différents : caméras sur banc stéréo, scanner laser, etc. Pour bénéficier au mieux de cette prise d'informations, il est nécessaire d'intégrer les entrées 3D directes, les images, ou toute autre entrée d'une source différente dans le modèle de représentation 3D choisi en (vi).
- (ix) La dimension temporelle dépasse le cadre de l'odométrie et du **SLAM**, qui visent à reconstituer une scène statique, pour devenir une composante du modèle 3D. Comment alors gérer et traiter un modèle 3D multitemporel ? Il peut par exemple servir pour la vision active (par exemple en robotique, avec le robot évoluant dans la scène) et l'analyse spatio-temporelle (renforcement de la probabilité de reconnaissance et fermeture de boucle, ou détection de changement et d'objets mobiles).

Sont détaillés dans la suite plusieurs programmes de travaux pour apporter des réponses à ces questions.

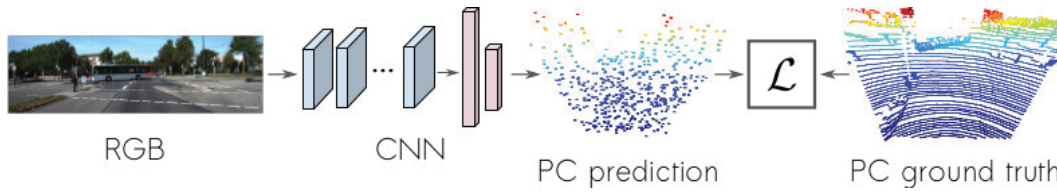


FIGURE 4.5 – Architecture de réseau de neurones pour la prédiction de nuages de points mono-image. Source : [229].

4.3.1 À court terme : prédire et sémantiser des nuages de points

Dans la suite de nos travaux des parties 3.2 et 3.3, nous avons deux objectifs à court terme qui répondent au défi (vii).

Prédiction de nuages de points 3D mono-image Dans le cadre de la thèse de Marcela Carvalho, nous avons développé des approches pour la prédiction de cartes de profondeur 2.5D (voir partie 3.2) qui réalisent d’un seul coup les deux premières étapes du programme de Marr. Sur la fin de cette thèse, et dans le cadre de la thèse de Rémy Leroy débutée en 2019, nous développons des approches pour la prédiction de nuages de points 3D à partir d’une seule image, c’est à dire les trois étapes de Marr en une seule fois.

Ces approches sont basées sur un CNN pour l’encodage de l’information image et une fonction de coût de transport optimal pour la prédiction de nuages de points correspondant aux données de références (voir figure 4.5). En détails, la prédiction se fait en deux temps pour des raisons de capacité mémoire. Tout d’abord la prédiction d’un nuage-squelette, c’est à dire un nuage de point épars, en s’inspirant des travaux de [230] pour la prédiction des points d’un modèle d’objet 3D. Nous avons comparé différentes fonctions de pénalités adaptées aux nuages de points, la distance de chanfrein (*Chamfer distance*) et la distance de Wasserstein, dans sa version approchée par l’algorithme d’optimisation de Sinkhorn [231, 232]. Dans un deuxième temps, nous densifions le nuage squelette par une variante de l’algorithme DensePCR [233], qui utilise des modules PointNet [199] pour calculer des caractéristiques locales et globales afin de prédire des points à rajouter. Nous avons testé notre approche sur le jeu de données KITTI [228] qui contient des images de scènes pour la conduite autonome de voitures associées à des nuages de points LiDAR : des résultats de prédiction de scène sont montrés figure 4.6. La structure et la distribution globales du nuage de points sont retrouvées (sans encoder des artefacts d’acquisition tels que les lignes du LiDAR) et permet de visualiser la route et les obstacles. Nous avons montré qu’il est donc possible de prédire directement le nuage de points d’une scène à partir d’une seule vue avec des réseaux de neurones. Nous avons montré que la distance de Sinkhorn issue du transport optimal était la fonction de pénalité qui donne la distribution des points la plus réaliste et la mieux répartie. Nous avons également introduit un réseau de densification avec une approche résiduelle [135] qui permet d’apprendre de petites variations de position par rapport à la distribution du nuage-squelette. Il s’agit, à notre connaissance, de la première méthode mono-vue pour prédire le nuage du point correspondant à une scène. La prédiction directe de nuages de points 3D modélise de plus implicitement les paramètres de la caméra nécessaires pour le passage de la profondeur à la 3D. Les limites actuelles sont liés à la densité des nuages de points et à la capacité mémoire.

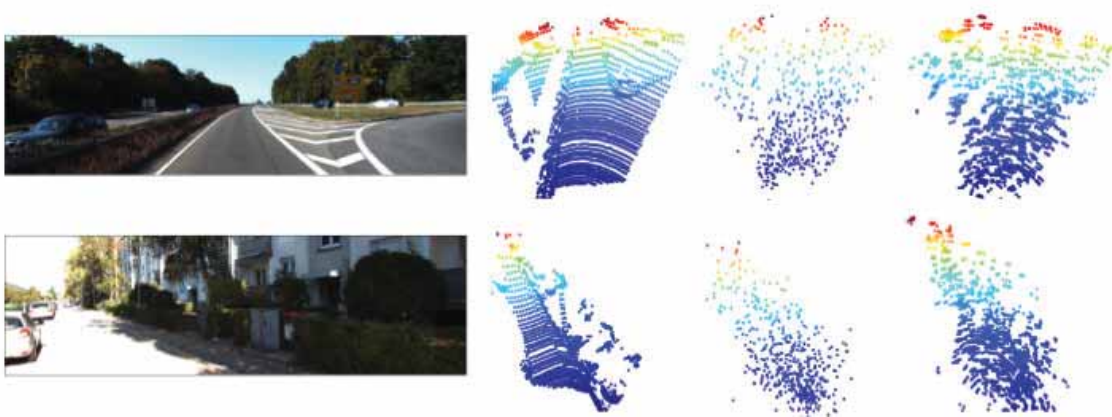


FIGURE 4.6 – Prédiction de nuages de points 3D avec une seule image : image originale RGB du jeu KITTI [228] pour la conduite autonome, nuage de points LiDAR correspondant, prédiction du nuage squelette par réseaux de neurones et transport optimal, densification par réseaux de neurones. Source : [229].

Segmentation sémantique de nuages de points 3D Dans le cadre de sa thèse débutée en 2019, Antoine Manier, avec qui je collabore, cherche à développer des algorithmes plus performants pour sémantiser des nuages de points 3D, dans la lignée des travaux de la partie 3.3. Le contexte est fourni par le partenaire SNCF Réseaux, et concerne la maintenance de sites et d'équipements en milieu ferroviaire.

Depuis les travaux de SnapNet [RI-10], de nombreuses approches ont été proposées visant à modéliser le voisinage local des points 3D pour une meilleure classification. D'abord, SuperPointGraph [234] produit une sur-segmentation du nuage de points sous forme de SuperPoints (inspirés des sperpixels en 2D [131] et classe les graphes locaux par des réseaux récurrents. Ensuite, toute une famille d'approches vise à émuler des convolutions en 3D, afin de transposer le principe des réseaux convolutifs : les χ -transforms de PointCNN [172], TangentConv [235], ConvPoint [222] ou encore KPConv [236]. La difficulté est alors de construire des convolutions sur des ensembles non-réguliers de points 3D, au contraire de la grille de pixels 2D.

Notre premier problème est de faire le tri entre ces méthodes : quelles sont les plus performantes, mais aussi quelles approches sont les plus prometteuses même si elles n'obtiennent pas encore les meilleurs résultats de l'état de l'art. La convergence vers une solution combinant leurs avantages respectifs permettrait de gagner en performances. Le problème sous-jacent est également de trouver des moyens d'inclure de l'invariance au point de vue dans des méthodes d'apprentissage automatique. Ce problème bien connu de l'analyse d'image prend une complexité supplémentaire en 3D, mais en même temps bénéficie d'une géométrie plus naturelle (pas de projection de la scène). La seconde question qui se pose est de savoir comment introduire des a priori géométriques qui permettraient d'améliorer les performances des réseaux actuels. Les fils caténaux peuvent ainsi être modélisés par des lignes, les murs et les routes par des surfaces planes, ce qui permet de régulariser la segmentation et d'alléger la représentation de la scène 3D.

4.3.2 À moyen terme : vers des modèles plus abstraits

Les défis (*viii*) (sur les entrées multiples) et (*vi*) (sur la représentation cible) nécessitent de créer un formalisme commun pour le traitement simultané des images et des nuages de points (soit issus d'une seule acquisition, soit accumulés). Les deux projets suivants suivent des programmes qui convergent en ce sens. Pour le défi (*vii*), le gain en qualité de sémantisation passe aussi par la détection d'objets et la segmentation fine.

Détection d'objets 3D Plusieurs travaux récents en analyse d'image utilisent la prédiction de profondeur pour la détection et localisation en 3D d'objets 3D. Par exemple, PseudoLiDAR [237] et PseudoLiDAR++ [238] prédisent une la carte de profondeur et l'utilisent pour obtenir une estimation 3D de la scène sous forme de nuage de points dans laquelle il est plus facile de repérer les objets. L'*Orthographic Feature Transform* [239] cherche à prédire directement une représentation voxelique de l'espace 3D. Ces exemples, tout comme que SnapNet-R [CI-34], montrent que pour détecter en 3D, un espace de représentation 3D est mieux adapté que la simple carte de profondeur.

Les travaux sur la prédiction de nuages de points de la thèse de Rémy Leroy se poursuivront par le développement d'approches pour la détection d'objets 3D et leur localisation, notamment dans le cadre des véhicules autonomes avec des jeux de données tels que KITTI [228] ou Waymo [240]. Mieux que les cartes 2.5D de pseudo-LiDAR [237], les nuages de points 3D non-structurés prédits en partie 4.3.1 sont l'espace de recherche idéal pour la détection 3D. Enfin, une optimisation conjointe du nuage et de la sémantique permettrait sûrement d'obtenir des estimations plus fines.

Enfin, plusieurs de résultats précédents montrent qu'il est possible d'estimer la 3D d'une scène avec des capteurs peu onéreux et compacts (simple caméra ou caméra avec DFD). Cependant, nul ne confierait la conduite d'un véhicule automatique à un tel système de perception. Le problème qui se pose est de comprendre les mécanismes à l'oeuvre dans les réseaux pour s'assurer de la qualité de la prédiction de 3D. Le calcul de l'incertitude de prédiction comme dans [CI-40] ou [188], qui permet aux réseaux de s'auto-évaluer à travers une estimation de la confiance que l'on peut accorder à la prédiction, mais aussi la compréhension fine des indices visuels de l'image (bords, statistiques de la scène) permettra de rendre les prédictions plus fiables. De plus, certains mécanismes optiques issus de la photographie computationnelle permettraient également de rendre les systèmes plus sûrs. La DFD fournit déjà une perception physique de la scène qui correspond donc à une mesure d'une réalité factuelle. Les aberrations chromatiques ou astigmatiques, ou bien encore les pupilles codées [241] permettraient de renforcer son effet.

Vers le jumeau numérique de la scène Dans la suite de la thèse d'Antoine Manier (voir partie 4.3.1) sur la sémantisation de nuages de points 3D, l'objectif est de fournir les moyens aux gestionnaires d'infrastructures de veiller à l'évolution des sites par des approches de maintenance prédictive comme le jumeau numérique. Le but est de modéliser les infrastructures et leur état pour prédire la durée de vie des sites et des matériels et les risques potentiels d'exploitation.

À cette fin, la sémantisation de la scène 3D doit être plus fine que la simple segmentation sémantique et inclure la recherche des objets 3D d'intérêt dans le nuage de la scène. Outre les approches par apprentissage déjà évoquées, les objets étant particulièrement normalisés dans ce cadre particulier, des approches par appariement telles que [CI-11] ou [242] sont envisageables. En parallèle de la recherche d'éléments géométriques, cela permet de tendre vers la vectorisation

du nuage de points. La mise en relation avec des méta-données d’usage des matériels permettra alors d’identifier chaque objet pour le suivi de son évolution.

4.3.3 À plus long terme : analyser la scène 3D multitemporelle

Représentation de la scène 3D Les travaux décrits précédemment apportent probablement une réponse possible au défi (vi) de la représentation optimale de la scène 3D.

En effet, le nuage de points 3D est une représentation couramment utilisée en vision et en robotique. Elle présente les avantages d’être compacte, de constituer le format natif de systèmes d’acquisition tels que les scanner lasers ou la photogrammétrie, et plusieurs méthodes d’analyse 3D existent pour son traitement, à l’instar de PointNet et de ses variantes [199, 172, 173].

Des alternatives existent cependant. Par exemple, des travaux sur la prédiction de surfaces et de maillages apparaissent [226, 243]. Il s’agit d’un format usuel en informatique graphique, et de plus en plus en robotique car l’usage de fonctions de distance signée permet une représentation compacte de l’occupation de l’espace utile pour la planification de la trajectoire du robot. Par ailleurs, les approches voxeliques héritées de la robotique sont toujours couramment utilisées, tels que les Oct-trees (arbres de partition binaire en 3D, établis de manière à allouer des voxels mieux résolus aux zones occupées de l’espace) d’Octomap [198]. Enfin, des modélisation plus haut niveau, orientées objet sont également possibles, par exemple pour la modélisation urbaine [244]. Une comparaison des modèles possibles sera nécessaires.

Tous ces modèles de représentation sont explicites et humainement compréhensibles. Cependant, une solution est peut-être tout simplement de considérer le choix de cette représentation optimale comme un problème d’apprentissage à optimiser. Ici encore, l’apprentissage multitâche tenant compte des différents objectifs possibles (robotique, synthèse de scène, compréhension) et des différentes sources (LiDAR, caméra RGB ou bien par photogrammétrie) offre un moyen possible pour créer cette représentation.

Analyse multitemporelle de scènes 3D Le dernier défi (ix) est l’analyse de scènes 3D multitemporelles, c’est à dire de scènes 4D. En effet, les nouveaux moyens physiques et algorithmiques pour représenter la scène en 3D sont suffisamment rapides et maniables pour permettre la revisite de scènes et de multiplier les acquisitions sur un même site. Les espoirs suscités par la voiture autonome aboutissent aussi à la diffusion de nombreuses données permettant d’entraîner des modèles statistiques.

Comment exploiter la redondance temporelle pour reconstruire à la fois la sémantique et accumuler l’information géométrique ? L’objectif est ici de développer des outils pour la reconstruction sémantique 3D multitemporelle, par exemple pour les scènes traversées par un véhicule de SemanticKITTI [245]. Lorsque le temps entre deux acquisitions est relativement faible, les changements d’une acquisition à l’autre sont principalement caractérisés par le changement de point de vue. Ainsi l’exploitation des données antérieures est une opportunité pour le raffinement sémantique et géométrique de la scène reconstruite, mais aussi le traitement des occlusions. Cela ouvre également la possibilité de raisonner sur les objets dynamiques de la scène, notamment de pister leurs déplacements et d’anticiper leurs trajectoires.

Le cadre multitemporel permet également d’aborder la vision active [246], c’est à dire une approche où le *point de vue* sur la scène peut être choisi de manière à améliorer la perception. Pour un robot ou un véhicule se déplaçant dans son environnement, des vues images ou même des vues 3D peuvent être acquises sur les zones de la scène qui permettent de compléter la prédiction et de lever les ambiguïtés, c’est à dire de générer une *conscience de la scène*. Par

rapport à SnapNet-R ([CI-34], partie 1.3), cela apporte une information réelle au lieu des vues virtuelles générées en mono-acquisition et les points de vues ne sont pas répartis systématiquement mais choisis et orientés à propos. De plus, dans le cas où le robot explore vraiment son environnement, sa trajectoire peut être optimisée de manière à améliorer sa compréhension de la scène.

Chapitre 5

Notice individuelle et rapports d'activité

Ce chapitre fait la synthèse de mon parcours académique et professionnel. La partie 5.1 présente en bref mon curriculum vitae. Le bilan scientifique des mes activités en matière de recherche est dressé en partie 5.2. Mes activités en termes de projets sont détaillées en partie 5.3. Enfin, mes activités d'enseignement et d'encadrement sont résumées en partie 5.4.

5.1 Curriculum Vitae

5.1.1 Activités professionnelles

1999-2003 Doctorant *INRIA/Imédia, Rocquencourt, France.*

- **Machine learning, multimédia** : Indexation et recherche par le contenu d'images et de vidéos, développement de techniques de classification supervisée et non-supervisée pour catégoriser et gérer des collections d'images par similarité visuelle ou résumer des vidéos [PRIAMM MediaWorks].

2003-2005 Chercheur post-doctoral *ERCIM CNR de Pise / Univ. de Berne, Italie / Suisse.*

- **Machine learning, multimédia** : Classification de scènes par méthodes à noyaux et appariement de graphes, pour l'annotation automatique de documents multimédia et la recherche dans les bibliothèques numériques [FP6 NoE Delos].

2005-2007 Chercheur post-doctoral *ENS Cachan/CMLA, Cachan, France.*

- **Imagerie 3D, microscopie confocale** : Tomographie et reconstruction de volumes 3D de cellules vivantes par des techniques d'inférence bayésienne et de déconvolution [FP6 Automation].

2015-2019 Vice-responsable puis responsable *IEEE GRSS Image Analysis and Data Fusion Technical Committee.*

2008-2020 Chercheur *ONERA/DTIS, Palaiseau, France.*

- **Machine learning, télédétection** : Apprentissage statistique fondé sur des méthodes de classification non-paramétriques (réseaux convolutifs, boosting, machines à vecteurs de support, etc.) pour l'imagerie satellitaire : cartographie automatique, détection de changements, fouille interactive d'images [ONERA DELTA, Partenariats ONERA-SNCF, ONERA-TOTAL].
- **Vision, robotique** : Détection et reconnaissance d'objet pour la vision des drones et les capteurs aéroportés (optique, **LiDAR**). Applications en recherche et sauvetage en milieu urbain [ONERA AZUR, FP7 SEC Darius, FP7 SEC Inachus].
- **Vision 3D, fusion de données** : Apprentissage profond pour l'interprétation de données 3D (nuages de points, données **RGB-D**), l'estimation de profondeur mono-image et *Depth from Defocus* [ONERA DELTA].

5.1.2 Parcours académique

1996-1999 Ingénieur ENSERG (aujourd'hui Phelma) *INPG, Grenoble, France*.

1998-1999 DEA Signal Image Parole, *INPG, Grenoble, France*.

Mémoire sur l'estimation de mouvement dans des séquences vidéos (sous la direction de Jürgen Stauder). *IRISA/Temics, Rennes, France*.

1999-2003 Thèse en Informatique, *Univ. Versailles-Saint-Quentin-en-Yvelines & INRIA/Imédia, Rocquencourt, France*. Classification non-exclusive et personnalisation par apprentissage : application à la navigation dans les bases d'images (sous la direction de Nozha Boujema).

Jury :

- Michel Scholl (Président) ;
- Bernadette Bouchon-Meunier (Rapporteuse) ;
- Françoise Prêteux (Rapporteuse) ;
- Nozha Boujema (Directrice de thèse) ;
- Carl Frélicot (Examineur) ;
- Claude Timsit (Examineur).

5.1.3 Enseignements / Encadrements

2000-2002 Algorithmique et programmation orientée objet *Univ. Paris IX Dauphine, Paris, France*¹. TD et TP Java en DEUG MASS, 80h ;

2004 Qualification par les sections CNU 27 et 61 ;

2010 Détection et suivi d'objet pour véhicules autonomes *IPSA Air et Espace, Paris, France*². Cours magistral et TP, 10h ;

1. <https://www.dauphine.psl.eu/>

2. <https://www.ipsa.fr/>

- 2010-2017** Traitement d'image et vision par ordinateur *École Polytechnique, Palaiseau, France*³. Modal Dpt. Physique, projet 60h ;
- 2017-2019** Apprentissage automatique *ENSTA ParisTech, Palaiseau, France*⁴. Cours et TD, 2e année, module 21h ;
- 2016-...** Reconnaissance des formes et apprentissage machine *Institut d'Optique Graduate School, Palaiseau, France*⁵. Cours et TD, 3e année, module 18h ;
- 2019-...** *Deep learning for remote sensing EuroSDR, Europe*⁶. Formation continue (doctorants et professionnels) : séminaire suivi de 15 jours de cours et projets en ligne (30h).

Encadrements en bref : 4 PhD soutenus, 5 PhD en cours, 1 post-doc en cours, 10 stages Master ou Projets de Fin d'Étude (X, CPE Lyon, ENSEA, ENSTA, IMT Atlantique...).

Voir partie 5.4 pour un descriptif détaillé des activités en matière d'enseignement et d'encadrement.

5.2 Rapport d'activité de recherche

5.2.1 Production scientifique et brevets

	Total (int. / nat.)	4 dernières années (int. / nat.)
Nombre de publications dans des revues avec comité de lecture	22 (21 / 1)	20 (19 / 1)
Nombre de publications dans des actes de congrès avec comité de lecture	58 (46 / 12)	32 (24 / 8)
Nombre de livres ou de chapitres de livres	0	0
Nombre de conférences invitées dans des congrès internationaux	6	4
Nombre de brevets	0	0
Nombre de rapports techniques ONERA	25	12

Tableau 5.1 – Résumé de la production scientifique

La production scientifique est résumée dans le tableau 5.1. Pour les revues, elle comprend 14 articles avec comité de relecture en simple aveugle et 8 articles en relecture ouverte. Caractéristique du domaine des Sciences et Techniques de l'Information et de la Communication (STIC, correspondant aux sections 27 et 61 de la Commission Nationale Universitaire - CNU), une majeure partie des travaux ont été publiés dans des actes de conférence, et ont eux aussi fait l'objet d'une relecture en simple ou double aveugle, pour un total de 58 communications, dont 6 conférences invitées.

L'impact de cette production est en partie mesurable par le nombre de citations de ces travaux : 1212 citations cumulées au total, et 828 depuis 4 ans (2016). 28 publications ont reçu plus de 10 citations. Pour une analyse plus fine, l'histogramme annuel du nombre de citations

3. <https://www.polytechnique.edu/>

4. <https://www.ensta-paristech.fr/>

5. <https://www.institutoptique.fr/>

6. <http://www.eurosdnet.net/>

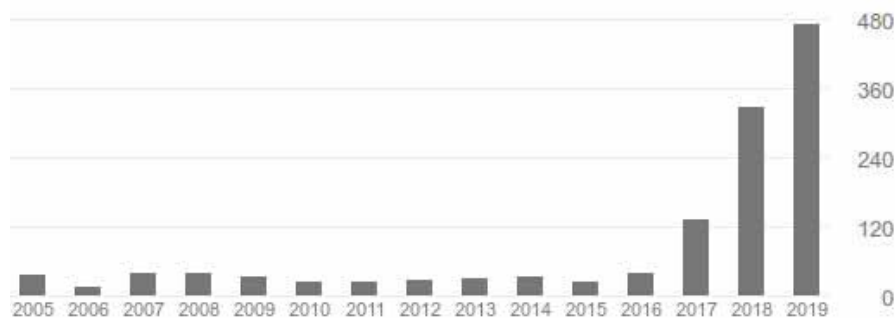


FIGURE 5.1 – Nombre de citations de mes articles et communications par année. Source <https://scholar.google.fr/> consultée le 10/12/2019.

	h-index ^a	i10-index ^b	HIC ^c
toutes publications	21	28	65
publications post-2014	17	21	-

Tableau 5.2 – Indicateurs bibliométriques globaux de la production scientifique.

a. nombre h de publications citées au moins h fois, source <https://scholar.google.fr/> consultée le 10/12/2019

b. nombre de publications avec au moins 10 citations, source <https://scholar.google.fr/> consultée le 10/12/2019

c. *High influence citations*, source <https://www.semanticscholar.org/> consultée le 14/10/2019

est présenté en figure 5.1⁷. Il montre un impact régulier tout au long de ma carrière, et en forte croissance sur les dernières années. Une analyse d'impact selon divers indices bibliométriques calculés à l'aide des sources publiques disponibles est également donnée dans le tableau 5.2.

5.2.2 Prix scientifiques

Plusieurs de mes projets (travaux, collaborations ou articles) ont été récompensés par des prix scientifiques :

[2003] Récipiendaire d'une bourse de recherche *European Research Consortium for Informatics and Mathematics (ERCIM)*⁸ ;

[2005] *Prix CNISF - L'Usine Nouvelle des ingénieurs de l'année - PIA 2005*. Prix pour la science ayant contribué au développement et à la réalisation inédite d'un instrument, d'un matériel ou d'une technique indispensable à la recherche scientifique. : B. Chalmond et S. Shorte et leurs équipes ;

[2015] *2e Prix Data Fusion Contest 2015* au congrès IGARSS 2015, A. Lagrange, B. Le Saux et al. ;

[2016] *Prix du meilleur poster* pour Joris Guerry aux rencontres JJCR 2016 (*Étudiant encadré*) ;

7. Et consultable en temps réel : <https://scholar.google.fr/citations?user=SiGd2-YAAAAJ&hl=fr>

8. <https://fellowship.ercim.eu/>

[2016] *Award for Best Contribution to the ISPRS 2D Semantic Labeling Contest* au congrès GeoBIA 2016, N. Audebert, B. Le Saux et S. Lefèvre ;

[2017] *2e Prix du meilleur article étudiant* du congrès JURSE 2017 pour Nicolas Audebert (*Étudiant encadré*) ;

[2018] *Prix du meilleur article* du congrès RFIAP 2018, M. Carvalho, B. Le Saux et al. ;

[2018] *Prix de thèse 2018* pour Joris Guerry, décerné par l'ED Interfaces, pôle "Ingénierie des Systèmes Complexes" (*Étudiant encadré*) ;

[2019] *Prix du meilleur article étudiant* de l'atelier CVPR / Earth Vision 2019 pour Rodrigo Daudt (*Étudiant encadré*).

5.2.3 Activités d'intérêt collectif

Je contribue à l'animation de la communauté scientifique par diverses actions :

[2015-2017] Vice-responsable nommé du comité technique sur l'analyse d'image et la fusion de données (IADF TC) de l'IEEE GRSS.

- Organisation du Data Fusion Contest 2016 "*Very High Temporal Resolution from Space*" avec Deimos Imaging et UrTheCast.
- Organisation du Data Fusion Contest 2017 "*Open data for global multimodal land use classification*" avec l'Université de Hambourg et WUDAPT.

[2017] *Keynote talk* à l'université d'été en télédétection (RSSS'2017) co-organisée par l'Université Technique de Munich (TUM) et le Centre Aérospatial Allemand (DLR) : <https://www.lmf.bgu.tum.de/en/tumdlrssi17/>

[2017-2019] Responsable élu du comité technique sur l'analyse d'image et la fusion de données (IADF TC) de l'IEEE GRSS⁹ en charge de l'organisation des compétitions Data Fusion Contest¹⁰

- Organisation du Data Fusion Contest 2018 "*Advanced multi-sensor optical remote sensing for urban land use and land cover classification*" avec l'Université de Houston.
- Organisation du Data Fusion Contest 2019 "*Large-Scale Semantic 3D Reconstruction*" avec l'Université Johns Hopkins et la IARPA.

[2018] Co-organisateur atelier "TerraData" en marge du congrès RFIAP / CFPT'2018 : <https://sites.google.com/view/terradata2018>

[2018] *Guest editor* d'un numéro spécial de IEEE Geoscience and Remote Sensing Letters (GRSL) consacré à "*Multimodal Optical Remote Sensing Imagery*"

[2019] Tutoriel "Deep Learning for Remote Sensing" avec Loïc Landrieu (IGN) lors du congrès JURSE 2019 ;

9. <http://www.grss-ieee.org/community/technical-committees/data-fusion/>

10. <http://www.grss-ieee.org/community/technical-committees/data-fusion/data-fusion-contest/>

[2019] Co-organisateur de l'atelier "EarthVision" en marge du congrès CVPR 2019 : <https://www.grss-ieee.org/earthvision2019/>

[2019] Tutoriel "Machine Learning for Remote Sensing" avec Ronny Hänsch (TU Berlin), Yuliya Tarabalka (LuxCarta) et Devis Tuia (Wangeningen Univ. & Research) lors du congrès IGARSS 2019 ;

[2020] *Guest editor* d'un numéro spécial de IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing (JSTARS) consacré à " Computer Vision-based Approaches for Earth Observation".

[2020] Co-organisateur de l'atelier "EarthVision" en marge du congrès CVPR 2020 : <https://www.grss-ieee.org/earthvision2020/>

Je suis relecteur récurrent pour les revues IEEE TGRS, IEEE JSTARS, IEEE GRSL, ISPRS JPRS et MDPI Remote Sensing dans le domaine de la télédétection, ainsi que Patt. Rec. Letters, CVIU, IEEE TIP dans le champ de la vision par ordinateur. J'ai été régulièrement membre du comité de programme ou relecteur de IGARSS, JURSE ou CFPT dans le domaine de la télédétection, et de CVPR, IROS, ICRA, RFIA(P), GRETSI, CBMI dans celui de la vision par ordinateur.

5.2.4 Activités d'expertise et jury

J'ai apporté mon expertise dans l'évaluation de projets pour :

- Agence Nationale de la Recherche (ANR) ;
- Institut National des Sciences de l'Univers (CNRS) / Programme national de télédétection spatiale ;
- Consortium de recherche et d'innovation en aérospatiale au Québec (CRIAQ) ;
- *Indo-French Centre for the Promotion of Advanced Research* (CEFIPRA).

J'ai été examinateur dans le cadre de deux jurys de thèse de doctorat :

- Vinicius Ferraris, encadrée par Nicolas Dobigeon et Marie Chabert, INP Toulouse, 2018 ;
- Adrien Lagrange, encadrée par Mathieu Fauvel et Nicolas Dobigeon, INP Toulouse, 2019.

Depuis 2015, j'ai par ailleurs participé à 4 comités de suivi de thèse :

- Benjamin Kellenberger (encadré par Devis Tuia) à l'Université de Zurich ;
- Mathieu Laroze (encadré par Sébastien Lefèvre, Romain Dambreville, Chloé Friguet et Ewa Kijak) à l'Université de Rennes-I ;
- Onur Tasar (encadré par Yuliya Tarabalka et Pierre Alliez) à l'Université de Nice ;
- Rim Zaiani (encadrée par Mathieu Fauvel et Jordi Inglada) à l'INP Toulouse.

5.3 Rapport d'activité lors d'études et recherches contractuelles

Une grande part de ma carrière s'est déroulée au [Département Traitement de l'Information et Systèmes \(DTIS\)](#), à l'ONERA, Établissement Public à vocation Industrielle et Commerciale (EPIC) sous tutelle du ministère de la Défense. Cette partie recense les études et projets menés dans le cadre des activités contractuelles et commerciales de l'ONERA avec ses partenaires institutionnels et industriels.

5.3.1 Activités de recherche collaborative

Période	Projet	Participants
2002-2003	PRIAMM MediaWorks	<i>INRIA</i> , Univ. Paris 11 / LIMSI, TF1, Aegis, etc.
2003-2004	FP6 DELOS Network of Excellence for Digital Libraries	ERCIM, <i>CNR</i> , ETHZ, Imperial College, <i>INRIA</i> , etc.
2005-2008	Projet Européen FP6 Automation	Institut Pasteur, <i>ENS Cachan</i> , etc..
2008-2009	Contrat de Recherche DGA AvalH2	<i>ONERA</i> , IGN, CNES, CEA, ArMines, etc.
2008-2009	Projet "Pôle de compétitivité" IdF IMVN InfoMagic	Univ. Paris 11, Univ. Paris 8, Telecom ParisTech, <i>INRIA</i> , Thalès, <i>ONERA</i> , CEA, etc.
2010-2012	Contrat de Recherche DGA EFUSION	<i>ONERA</i> , IGN, CNES, Telecom ParisTech, CEA, ArMines
2010-2013	Projet Européen FEDER LIMA	<i>ONERA</i> , Magellium, Noveltis, Oktal-SE, etc.
2012-2015	Projet Européen FP7 Darius	NTUA, Cork Institute of Tech., <i>ONERA</i> , SINTEF, Airbus, BAE systems, etc.
2015-2018	Projet Européen FP7 Inachus	ICCS, NTUA, Univ. Twente, FOI, <i>ONERA</i> , Fraunhofer EMI, EPLFM, etc.
2014-2018	PRI Naomi	<i>ONERA</i> , Total
2015-2019	PRI Drosfiles	<i>ONERA</i> , SNCF Réseau, Altamétris
2017-2019	Contrat de Recherche DGA Optimom	<i>ONERA</i> , IGN, CNES, CEA
2019-2023	Projet BPI AI4GEO	CS, CNES, <i>ONERA</i> , IGN, Airbus, Qwant, etc.

Tableau 5.3 – Collaborations nationales ou internationales (ayant donné lieu à une production scientifique en commun) et/ou valorisation de la recherche dans un contexte extra-académique. Organisme employeur en italiques.

Les collaborations nationales ou internationales auxquelles j'ai participé au cours de ma carrière sont récapitulées dans le tableau [5.3](#).

Date	Établ. ^{nt}	LMD	Vol.	Intitulé	Type
2000 - 2002	Univ. Paris Dauphine	L1	80h	Algorithmique	TD
2010 - 2011	IPSA	M2	10h	Vision	Cours / Lab
2011 - 2017	École Polytechnique	L3	60h	Vision et Traitement d'Image	Cours / Lab
2017	ONERA	Prof.	16h	Formation continue "Apprentissage machine"	Cours
2016 - 2019	IOGS	M2	18h	Apprentissage et Reconnaissance des Formes	Cours / TD
2017 - 2019	ENSTA ParisTech	M1	21h	Apprentissage Automatique	Cours / TD
2019 -	EuroSDR	Prof.	30h	<i>Deep Learning for Remote Sensing</i>	Cours et projet

Tableau 5.4 – Activités d'enseignement

5.4 Rapport d'activité d'enseignement et d'encadrement

5.4.1 Enseignement et formation

Bien que mon poste actuel ne comporte pas de charge d'enseignement, je me suis investi régulièrement dans plusieurs actions d'enseignement et de formation en école d'ingénieur (récapitulatif dans le tableau 5.4).

Actuellement, j'interviens dans les cours d'*apprentissage automatique* de l'Institut d'Optique Graduate School (niveau M2) et de l'ENSTA ParisTech (niveau M1). Les cours suivent le schéma classique cours magistral et travaux dirigés. À l'ENSTA, ils comportent également la réalisation d'un mini-projet sur la compréhension d'un corpus d'articles de recherche ou la réalisation d'une solution pour un problème bien défini d'apprentissage automatique (par exemple : challenge Kaggle). Les cours comportent une présentation des principes généraux de l'apprentissage statistique, l'étude de plusieurs algorithmes standard pour la classification supervisée (méthodes ensemblistes, méthodes à noyaux) et non-supervisée, et l'étude des réseaux de neurones et des approches d'apprentissage profond.

De 2011 à 2017, je me suis occupé du Modal *traitement d'image numérique et vision* du département de physique de l'École Polytechnique. Cet enseignement de niveau L1 (2e année d'école) consiste en un projet mené sur un trimestre avec des journées de travail hebdomadaires (pour un total de 50 à 60h). Les élèves y développent une application qui comporte acquisition des données et traitement informatique, tout en étant familiarisé avec des notions de projet au sens entreprise ou recherche. Les réalisations passées incluent par exemple de la super-résolution, du SLAM et de l'odométrie pour la robotique, de la détection de panneaux pour la conduite autonome ou encore la détection et la reconnaissance de visage par réseaux de neurones.

Par ailleurs, j'ai également participé à des enseignements d'algorithmie et programmation à l'Université Paris-Dauphine et de vision par ordinateur à l'IPSA - École d'ingénieurs aéronautique et spatiale.

Enfin, j'interviens dans des cursus de formation continue. D'une part à l'ONERA en coordination avec le service Formation de la Direction des Ressources Humaines. Ces cours sous forme d'exposés invités et de cours magistraux s'adressent à un public d'ingénieurs de recherche de tous âges et de domaines variés : optique, électromagnétisme, matériaux, dynamique des fluides,

Établ. ^{nt}	Intitulé	Documents
Univ. Paris Dauphine	Algorithmique	TD
IPSA	Vision	Cours "Détection d'objets" / Lab
École Polytechnique	Vision et Traitement d'Image	Cours "Vision par ordinateur" ^a Cours "Reconnaissance d'objets"
ONERA	Apprentissage machine	Cours "Apprentissage" et "Deep Learning" ^b
ENSTA ParisTech	Apprentissage Automatique	Cours "App. non-supervisé" / TD ^c Cours "App. Random Forests / boosting" / TD
I0GS	Apprentissage et Reconnaissance des Formes	Cours "Apprentissage non-supervisé" / TD ^d Cours "Approches ensemblistes" / TD Cours "Auto-encodeurs et GANs" / TD
EuroSDR	Deep Learning for Remote Sensing	Cours "Deep learning for automatic mapping" ^e

Tableau 5.5 – Liste des supports de cours produits et liens si disponibles

a. <https://blesaux.github.io/teaching/X-modal>

b. <https://delta-onera.github.io/education/>

c. <https://blesaux.github.io/teaching/ENSTA>

d. <https://blesaux.github.io/teaching/I0GS-machine-learning>

e. <http://www.eurocdr.net/education/course/eduserv17-2019>

etc. D'autre part je participe aux formations de EuroSDR, organisme pan-européen dédié à la recherche sur les données spatiales et réunissant organismes de recherche et instituts cartographiques nationaux. La formation vise des professionnels du domaine (interprètes d'image aérienne, utilisateurs de système d'information géographique) et a lieu sous la forme d'un séminaire d'une demi-journée, suivi de cours en ligne et de la réalisation d'un projet.

Les documents conçus pour ces cours sont pour la plupart (et pour les plus récents) en ligne sur le site <http://blesaux.github.io/teaching/> (voir Tableau 5.5).

5.4.2 Encadrement de stages et de thèses

Au fil des années, j'ai eu la chance de travailler avec de nombreux et talentueux étudiants lors leur stage ou leur thèse.

5.4.2.1 Encadrement de stages

J'ai encadré ou encadre 10 stagiaires (voir tableau récapitulatif 5.6). Il s'agit majoritairement de stages de master ou de fin d'étude d'école d'ingénieur (niveau M2), à l'exception de deux stages de dernière année de l'École Polytechnique (niveau M1 en raison du cursus particulier de la formation). Ces stages ont concerné les thèmes de la vision par ordinateur (compréhension d'images, de vidéos), la robotique et notamment les drones, la télédétection et l'imagerie aérienne, et l'apprentissage machine. Il est notable que ces stages ont été majoritairement proposés en co-encadrement avec mes collègues (7/10) afin de créer des dynamiques collaboratives au sein des équipes.

n°	An	Étudiant	Sujet	Formation	Co-encadr.
1	2002	Nizar Grira	Clustering for video-summary	M.Eng ENIT	Nozha Boujemaa
2	2009	Fabien Giannesini	GPU-based anomaly detection for large image browsing	M.Eng ENSEA	-
3	2011	Caroline Henry	Vehicle detection for UAV vision systems	M.Eng ENS2M	M. Sanfourche
4	2011	Nicolas Chauffert	Active learning of regions-of-interest in satellite images	M.Eng École Polytechnique	Jonathan Israël
5	2012	Roman Garcia	Tracking and recognition in videos from camera networks	M.Eng CPE Lyon	Valérie Leung
6	2013	Morgane Rivière	Domain adaptation for object recognition in aerial imagery	M.Eng École Polytechnique	-
7	2014	Thierry Dumas	Depth from defocus and learning	MSc. / M.Eng Centrale Marseille	Pauline Trouvé
8	2015	Adrien Lagrange	Classification for Big Remote Sensing Data	MSc. / M.Eng ENSTA ParisTech	-
9	2018	Javiera Castillo Navarro	Large-scale semi-supervised semantic segmentation	MSc. CentraleSupélec / Master Data Science	A. Boulch, N. Audibert et S. Lefèvre
10	2019	Rémy Leroy	Neural networks for 3D prediction	IMT Atlantique	M. Carvalho et P. Trouvé

Tableau 5.6 – Stages encadrés

Le tableau 5.7 résume les retombées de ces stages tant sur le plan scientifique que sur les contributions apportées aux études menées à l'ONERA. 6 stages sur 10 ont donné lieu à une communication en colloque, et l'un d'entre eux a une publication en revue. Deux stages ont obtenu les félicitations du jury du prix du stage de recherche de l'École Polytechnique dans l'option maths applis. Le stage d'Adrien Lagrange a remporté la deuxième place de la compétition annuelle du Data Fusion Contest en 2015, ce qui a donné lieu à une conférence invitée au congrès IGARSS 2015 à Milan. Par ailleurs, ces stages ont toujours servi les problématiques des études menées à l'ONERA, tant au niveau de la recherche amont (Projets de Recherche et Projets de Recherche Fédérateur) que pour les études contractuelles menées avec l'Europe, l'industrie ou les autorités de tutelle.

5.4.2.2 Encadrement de thèses

J'ai encadré ou encadre 9 doctorants : 4 doctorants ayant soutenu leur thèse et 5 doctorats en cours. Le tableau 5.8 récapitule les doctorants, périodes de thèse et co-encadrants / directeurs. 6 thèses sur 9 ont été menées en collaboration avec des chercheurs et professeurs issus de laboratoires extérieurs afin de créer des liens avec le monde universitaire. 5 thèses sur 9 ont de plus été initiées ou menées en collaboration avec des chercheurs ONERA afin de contribuer au dynamisme interne. Dans tous les cas, j'ai initié les projets de thèse, (co-)défini les sujets et encadré au quotidien les doctorants. Quatre thèses sur neuf ont fait l'objet d'un (co-)financement extérieur, par Total dans le cadre du PRI Naomi, le CNES dans le cadre de leur appel à sujets de thèses doctorales, SNCF Réseaux et Delair dans le dispositif CIFRE, soit 44% en moyenne. Ces thèses s'inscrivent dans les axes thématiques "Perception et Traitement de l'Information" et "Intelligence Artificielle et Décision" du DTIS.

n°	Année	Nom	Sujet	Publi.	Prix	Contract.
1	2002	N. Grira	Clustering for video-summary	FUZZ-IEEE [CI-4]	-	-
2	2009	F. Giannesini	GPU-based anomaly detection for large image browsing	IGARSS [CI-12]	-	-
3	2011	C. Henry	Vehicle detection for UAV vision systems	-	-	PR AZUR
4	2011	N. Chauffert	Active learning of regions-of-interest in satellite images	IGARSS [CI-13]	Prix stage de recherche X option maths applis	FP7 Darius
5	2012	R. Garcia	Tracking and recognition in videos from camera networks	-	-	PR Copernic ; FP7 Subito
6	2013	M. Rivière	Domain adaptation for object recognition in aerial imagery	-	Prix du stage recherche option maths applis	PR Azur ; FP7 Darius
7	2014	T. Dumas	Depth from defocus and learning	GRETSE [CN-4]	-	PR Copernic
8	2015	A. Lagrange	Classification for Big Remote Sensing Data	IGARSS [CI-20], JS-TARS [RI-3]	Data Fusion Contest 2nd rank Award	PRI Naomi
9	2018	J. Castillo	Large-scale semi-supervised semantic segmentation	JURSE [CI-43]	-	AI4GEO
10	2019	R. Leroy	Neural networks for 3D prediction	-	-	-

Tableau 5.7 – Retombées des stages encadrés sur le plan scientifique et contractuel (réutilisation des méthodes et résultats).

Le tableau 5.9 compile différents indicateurs de réussite sur les thèses encadrées. Notamment, les travaux menés dans les quatre premières thèses ont été récompensés par de nombreux prix : prix étudiant ou meilleur papier en congrès, prix scientifique, prix de thèse. Les thèses soutenues ont duré 40 mois en moyenne (médiane à 38 mois). Après la première thèse qui a duré trop longtemps, les suivantes respectent un calendrier plus vertueux de quelques semaines au-delà de l'objectif des 3 ans. Chaque thèse soutenue a donné lieu à au moins un article publié en revue, et 1.75 article en moyenne. Pour chaque thèse soutenue, les doctorants ont contribué à au moins 4 communications en congrès (7 en moyenne, médiane à 5.5).

Les thèses en cours de plus d'un an ont déjà donné lieu à des communications en congrès. La figure 5.2 permet de visualiser les publications par doctorant (par ordre chronologique), tandis que les diagrammes de la figure 5.3 montrent l'évolution temporelle des publications. On observe une régularité de la production scientifique au fil des thèses encadrées, mis à part un étudiant particulièrement prolifique. La thèse ayant duré plus que la normale a notamment eu une production standard au final.

La figure 5.4 montrent l'évolution au fil des ans des citations des travaux des doctorants (source <https://scholar.google.fr/>, consultée le 15/10/2019). Chaque courbe indique, pour chaque doctorant, le nombre de citations dans l'année cumulées sur l'ensemble des publications. Ces courbes montrent, notamment pour les thèses soutenues, l'impact croissant des travaux menés auprès de la communauté. Le tableau 5.10 compile divers indices bibliométriques. Les travaux des thèses sont largement cités. Pour les trois premiers doctorants, avec un recul d'au

n°	Début	Soutenance	Nom	Sujet	Co-encadr.	Fin. ext.
1	01/10/2012	13/12/2016	Hicham Randrianarivo	Statistical learning of semantic classes for aerial image interpretation	Marin Ferecatu et Michel Crucianu (CNAM ParisTech)	
2	01/10/2014	18/11/2017	Joris Guerry	Robust visual recognition by neural networks in robotic exploration scenarios. Detect me if you can!	David Filliat (ENSTA ParisTech)	
3	01/10/2015	17/10/2018	Nicolas Audebert	Classification of Big Remote Sensing Data	Sébastien Lefèvre (Univ. Bretagne Sud)	100% Total
4	01/10/2016	25/11/2019	Marcela Pinheiro de Carvalho	3D Camera by Depth from Defocus and Deep Learning	Pauline Trouvé-Peloux, Frédéric Champagnat (ONERA) et Andrés Almansa (Univ. Paris Descartes)	
5	01/10/2017		Rodrigo Caye Daudt	Deep networks for multi-temporal activity analysis of Earth-observation data	Alexandre Boulch (ONERA) et Yann Gousseau (Telecoms ParisTech)	
6	02/01/2019		Javiera Castillo Navarro	Large-scale semi-supervised semantic segmentation	Alexandre Boulch (ONERA) et Sébastien Lefèvre (Univ. Bretagne Sud)	50% CNES
7	20/03/2019		Antoine Manier	Compréhension sémantique de scènes tridimensionnelles pour l'accompagnement de la maintenance en milieu ferroviaire	Alexandre Boulch (ONERA)	100% SNCF
8	10/09/2019		Gaston Lenczner	Réseaux de neurones interactifs pour l'analyse de scènes acquises par drone	Guy Le Besnerais (ONERA)	100% Delair
9	02/11/2019		Rémy Leroy	Deep neural networks for 3D prediction in the wild	Pauline Trouvé-Peloux, Frédéric Champagnat (ONERA)	

Tableau 5.8 – Thèses encadrées. La double ligne sépare les thèses soutenues des thèses en cours.

n°	Nom	Sujet	Durée	Nb. comm.	Nb. art.	Prix
1	H. Randrianarivo	Aerial image interpretation	49m	6	1	Data Fusion Contest Award 2015 Award (2nd rank)
2	J. Guerry	Robust visual recognition	38m	5	1	Prix du meilleur poster JJCR 2015; Prix de thèse option "Ingénierie des systèmes complexes" ED Interfaces 2018
3	N. Audebert	Classification of Big Remote Sensing Data	37m	13	4	Best student award JURSE 2017 (2nd rank); ISPRS Award for best contribution to the semantic labeling benchmark 2016
4	M. Carvalho	3D Camera by DFD and Deep Learning	38m	4	1	Prix du meilleur article RFIAP 2018
5	R. Daudt	Multi-temporal activity analysis		4	1	Best student award CVPR 2019 / Earth Vision workshop
6	J. Castillo	Semi-supervised semantic segmentation		2		
7	A. Manier	Compréhension de scènes 3D				
8	G. Lenczner	Analyse interactive de scènes vues par drone				
9	Rémy Leroy	3D prediction in the wild				

Tableau 5.9 – Résumé de la production scientifique des thèses encadrées : durée de la thèse, nombre de communications en congrès, nombre d'articles publiés en revue, prix scientifiques.

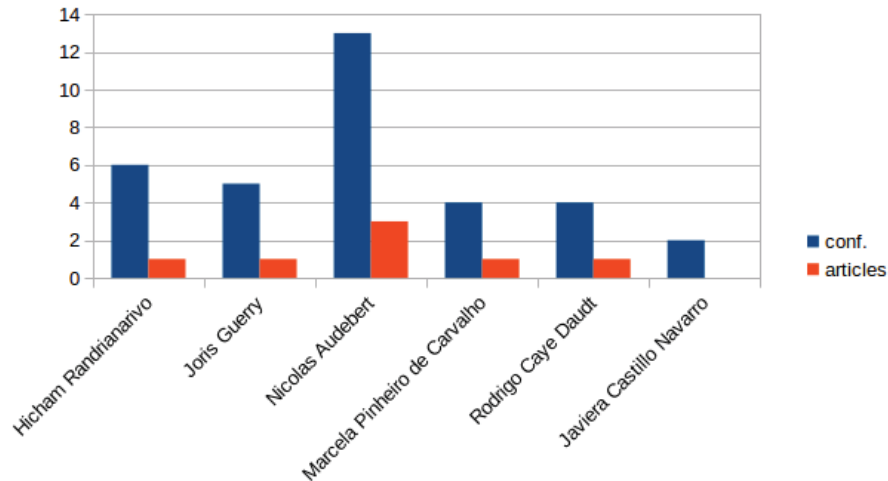


FIGURE 5.2 – Nombre de communications et articles publiés par chaque doctorant.

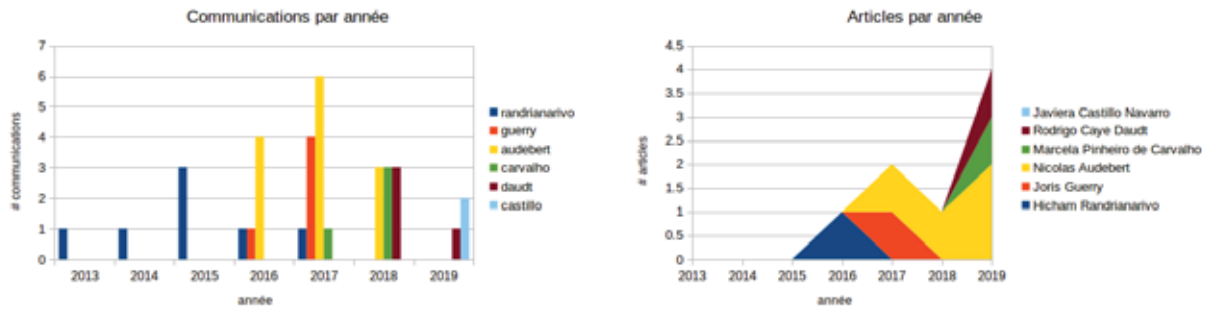


FIGURE 5.3 – Statistiques annuelles de publication des doctorants encadrés : nombre de communications par année et nombre d'articles publiés par année.

moins un an après la soutenance, au moins 2 publications sont citées plus de 10 fois (*i10-index*), le h-index est au moins égal à 5, et les nombres de citations totales vont de 84 à 503. Cette influence conséquente sur la communauté découle de la qualité des travaux bien évidemment, mais aussi d'une volonté marquée de contribution à cette communauté, notamment en contribuant à diffuser des moyens de reproductibilité de la science, comme des bases de données pour l'évaluation des méthodes et du code informatique.

Enfin, le tableau 5.11 indique ce que sont devenus les doctorants après leur soutenance. Sur quatre docteurs, un a rejoint le secteur académique (Maître de Conférences - MCF), un le département Recherche et Développement (R&D) d'une grande entreprise, et deux travaillent dans des Petites ou Moyennes Entreprises (PME) de haute technologie. Tous étaient en emploi moins de trois mois après leur soutenance.

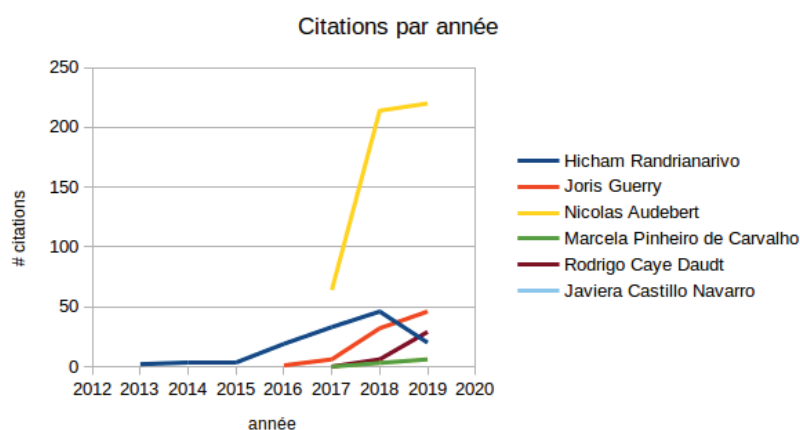


FIGURE 5.4 – Évolution annuelle du nombre de citations cumulées pour chaque doctorant. Source : <https://scholar.google.fr/>, consultée le 15/10/2019.

n°	Nom	Nb. comm.	Nb. art.	Nb. cit. total	Nb. top. cit.	h-index ^a	i10-index ^b
1	H. Randrianarivo	6	1	126	57	5	2
2	J. Guerry	5	1	84	34	5	4
3	N. Audebert	13	4	503	133	9	9
4	M. Carvalho	4	1	9	6	2	0
5	R. Daudt	4	1	35	16	3	2
6	J. Castillo	2					

Tableau 5.10 – Indicateurs bibliométriques de l'impact des thèses encadrées : nombre de communications en congrès, nombre d'articles publiés en revue, nombre total de citations sur l'ensemble de la production scientifique, nombre de citations de la publication la plus citée, h-index, i10-index. Source : <https://scholar.google.fr/>, consultée le 15/10/2019.

- a. nombre h de publications citées au moins h fois
b. nombre de publications avec au moins 10 citations

n°	Nom	Fin thèse	Poste	Employeur
1	H. Randrianarivo	2016	Ingénieur R&D	Qwant Research ^a
2	J. Guerry	2017	Ingénieur de Recherche Fondateur, <i>Chief Technical Officer</i>	EDF R&D Viz2Code ^b
3	N. Audebert	2018	MCF	CNAM ParisTech
4	M. Carvalho	2019	Ingénieure R&D	UpCiti ^c

Tableau 5.11 – Devenir des doctorants encadrés après leur soutenance.

- a. PME développant moteur de recherche et outils de cartographie : <https://www.qwant.com/>
b. Start-up spécialisée en apprentissage profond pour la reconnaissance de documents
c. PME spécialisée dans les *smart cities* : <https://www.upciti.com>

Glossaire

- 3DRMS 2018** *3D Reconstruction Meets Semantics*. 48
- ACP** Analyse en Composantes Principales. 3, 25
- ANSA** *Agenzia Nazionale Stampa Associata*. 2, 8
- ARC** *Adaptative Robust Clustering*. 2, 3, 5, 16
- BD ORTHO** Base de Données Orthophotographique. 33
- BoW** Sac de Mots – *Bag of Words*. 3, 8
- C-GAN** *Conditional Generative Adversarial Network*. 46
- CBIR** *Content-based Image Retrieval*. 1, 2, 3
- CNES** Centre National d'Études Spatiales. 61
- CNN** Réseau Neuronal Convolutif – *Convolutional Neural Network*. 11, 20, 25, 26, 27, 32, 58, 63
- CO3D** Constellation Optique en 3D. 61
- CRF** *Conditional Random Field*. 58
- D3-Net** *Deep Depth-from-Defocus Network*. x, 46, 47, 48
- DASE** *Data and Algorithms Standard Evaluation*. 34, 36, 37, 38
- DBN** *Deep Belief Network*. x, 44, 45
- Deep-DFD** *Deep Depth from Defocus*. 44, 46, 48
- DeepHyperX** *DEEP Learning for HYPERspectral toolboX*. 31
- DFC** *Data Fusion Contest*. 34, 36, 37, 38, 39
- DFC2015** *IEEE GRSS Data Fusion Contest 2015*. 24, 26, 34, 36
- DFC2016** *IEEE GRSS Data Fusion Contest 2016*. ix, 33, 34, 36, 37
- DFC2017** *IEEE GRSS Data Fusion Contest 2017*. ix, 33, 34, 37, 40

- DFC2018** *IEEE GRSS Data Fusion Contest 2018*. ix, 31, 33, 34, 38, 59
- DFC2019** *IEEE GRSS Data Fusion Contest 2019*. 33, 34, 39, 61
- DFD** *Depth from Defocus*. 44, 45, 46, 47, 48, 52, 65, 70
- DPM** *Deformable Part-Model*. 20, 22, 23, 24, 40
- DTIS** Département Traitement de l'Information et Systèmes. 74, 78
- DtMM** *Discriminatively-trained Model Mixture*. ix, 22, 24
- EEA** *European Environment Agency*. 33
- ERCIM** *European Research Consortium for Informatics and Mathematics*. 69, 72
- FCN** Réseau Entièrement Convolutif – *Fully Convolutional Network*. 11, 12, 25, 27, 28, 32
- GAD** *Guided Anisotropic Diffusion*. 58
- GAN** *Generative Adversarial Network*. 31
- Geo-Wiki** *Geo-Wiki: Earth Observation and Citizen Science*. 37
- GMM** *Gaussian Mixture Model*. 20, 21, 24
- GPS** *Global Positioning System*. 8, 10, 55, 60
- GPU** *Graphics Processing Unit*. 22
- Graph-edit distance** Distance d'appariement de graphe, calculable par l'algorithme A^* , et définie par un ensemble d'opérations d'édition des sommets et arêtes des graphe, chacune étant liée à un coût. 8
- GRSS** *Geosciences and Remote Sensing Society*. v, 33, 34, 69
- HOG** *Histogram of Oriented Gradients*. 10, 21, 22, 23, 24
- HRSCD** *High-Resolution Semantic Change Dataset*. 33, 34, 58
- IADF TC** *Image Analysis and Data Fusion Technical Committee*. v, 33, 36, 37, 38, 39
- IEEE** *Institute of Electrical and Electronics Engineers*. v, 33, 34, 69
- IGN** Institut Géographique National. 33, 61
- ILSVRC** *ImageNet Large Scale Visual Recognition Challenge*. 3
- ImageNet** Jeu de données images *ImageNet* à large-échelle. 3
- INRIA Aerial** *INRIA Aerial Image Labeling Dataset*. 31, 34
- INSPIRE** *INfrastructure for SPatial InfoRmation in Europe*. 59

- IR/R/G** *Infra-Red - Red - Green*. 25, 28
- ISPRS** *Int. Society of Photogrammetry and Remote Sensing*. 25, 34
- ISPRS Vaihingen** *ISPRS Semantic Labeling dataset - Vaihingen*. 25, 26, 28, 34, 57
- ISPRS Potsdam** *ISPRS Semantic Labeling dataset - Potsdam*. xi, 26, 27, 28, 31, 28, 34
- Kernel-adatron** Algorithme de perceptron adaptatif. 5, 6
- LBP** *Local Binary Pattern*. 10, 21, 22
- LCZ** *Zones de Climat Local - Local Climate Zones*. 37
- LiDAR** *Light Detection And Ranging*. 8, 10, 33, 34, 36, 38, 39, 49, 61, 62, 63, 65, 66, 70
- LS-GAN** *Least-Square Generative Adversarial Network*. x, 46
- LUV** *Lightness-Color (u,v)*. 5
- MCS** *Multiple Classifier System*. 8
- Mean-Shift** Algorithme de classification non-supervisée non-paramétrique qui pratique des estimation locales de densité pour trouver les classes pertinentes. 5
- mIoU** *mean Intersection-over-Union*. 13, 52
- MNE** *Modèle Numérique d'Élévation*. 27, 28, 36, 38, 39, 59
- MRF** *Markov Random Field*. 21
- MTL** *Multi-Task Learning*. 48
- NMS** *Non-Maximum Suppression*. 24
- NYUv2** *New-York University dataset version 2*. 13
- NZAM** *Agence de Cartographie de Nouvelle-Zélande - New-Zealand Aerial Mapping*. 23, 24
- Online Gradient-Boost** Algorithme de boosting incrémental. 10
- ONERA** *Office National d'Études et de Recherches Aérospatiales*. 10, 19, 57, 60, 61
- ONERA.ROOM** Jeu de données **RGB-D** pour la détection de personnes en robotique, acquis en conditions difficiles : illuminations variables, flou de bougé, etc.. 11, 15, 14
- OSCD** *ONERA Sentinel Change Detection*. ix, 32, 34
- OSM** *OpenStreetMap*. 25, 26, 27, 28, 33, 37, 55
- Pascal** *Pascal VOC Challenge*. 3, 10
- PSF** *Point-Spread Function*. 42

- RBM** *Restricted Boltzmann Machine*. 20, 25, 45
- R-CNN** Réseau de Neurones Convolutif basé Régions – *Region Convolutional Neural Network*. 11, 12, 15
- ReSSAC** Recherche et Expérimentations en vol sur les systèmes drones et Systèmes embarqués Sûrs Autonomes Coopérants. 10
- RGB** *Red - Green - Blue*. 13, 15, 23, 25, 28, 31, 38, 44, 45, 47, 51, 55, 59, 63, 66
- RGB-D** *Red - Green - Blue - Depth*. ix, 11, 12, 13, 15, 14, 16, 44, 70, 87, 88
- SAE** *Stacked Auto-Encoder*. 25
- SaR** *Search-and-Rescue*. 8, 10
- SAR** *Synthetic Aperture Radar*. 19, 20, 21, 22, 34, 60
- SIFT** *Scale-Invariant Feature Transform*. 21
- SIG** Système d'Information Géographique. 22
- Sim2Real** *Simulation to Reality*. 48
- SLAM** *Self-Localization And Mapping*. 8, 62
- SLIC** *Simple Linear Iterative Clustering*. 26
- SnapNet-R** *Snapshot Network for Robotics*. ix, xi, 12, 13, 14
- SUNRGBD** *Scene UNderstanding RGB-D*. 11, 13, 14
- SVM** Machine à Vecteurs de Support – *Support-Vector Machine*. 3, 5, 6, 16, 20, 21, 22, 24, 26, 32, 58
- TF1** Télévision française 1. 2, 5
- THR** Très Haute Résolution. 19, 20, 21, 22, 23, 33, 34, 36, 38, 39, 57, 59
- Urban Atlas** *Copernicus Land Monitoring Service - Urban Atlas*. 33, 57
- WUDAPT** *World Urban Database and Access Portal Tools*. 37

Bibliographie

Articles de revue internationale avec comité de lecture

- [RI-1] B. LE SAUX, B. CHALMOND, Y. YU, A. TROUVÉ, O. RENAUD et S. L. SHORTE. « Isotropic high resolution 3D confocal micro-rotation imaging for non-adherent living cells ». In : *Journal of Microscopy* 233.3 (mar. 2009), p. 404–416. URL : <https://www.ncbi.nlm.nih.gov/pubmed/19250461>.
- [RI-2] S. HERBIN, F. CHAMPAGNAT, J. ISRAEL, F. JANEZ, B. LE SAUX, V. LEUNG et A. MICHEL. « Scene Understanding from Aerospace Sensors : What can be Expected ? » In : *AerospaceLab* 4 (mai 2012), p. 1–15. URL : <https://hal.archives-ouvertes.fr/hal-01183709>.
- [RI-3] M. CAMPOS-TABERNER, A. ROMERO-SORIANO, C. GATTA, G. CAMPS-VALLS, A. LAGRANGE, B. LE SAUX, A. BEAUPÈRE, A. BOULCH, A. CHAN-HON-TONG, S. HERBIN, H. RANDRIANARIVO, M. FERECATU, M. SHIMONI, G. MOSER et D. TUIA. « Processing of Extremely High-Resolution LiDAR and RGB Data : Outcome of the 2015 IEEE GRSS Data Fusion Contest–Part A : 2-D Contest ». In : *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing* 9.12 (déc. 2016), p. 5547–5559.
- [RI-4] D. TUIA, G. MOSER et B. LE SAUX. « 2016 IEEE GRSS Data Fusion Contest : Very high temporal resolution from space Technical Committees ». In : *IEEE Geoscience and Remote Sensing Magazine* 4.1 (mar. 2016), p. 46–48.
- [RI-5] D. TUIA, G. MOSER et B. LE SAUX. « 2016 IEEE GRSS Data Fusion Contest : Multitemporal Very High Resolution from Space [Technical Committees] ». In : *IEEE Geoscience and Remote Sensing Magazine* 5.1 (mar. 2017), p. 66–70.
- [RI-6] D. TUIA, G. MOSER, B. LE SAUX, B. BECHTEL et L. SEE. « 2017 IEEE GRSS Data Fusion Contest : Open Data for Global Multimodal Land Use Classification [Technical Committees] ». In : *IEEE Geoscience and Remote Sensing Magazine* 5.1 (mar. 2017), p. 70–73.
- [RI-7] N. AUDEBERT, B. LE SAUX et S. LEFÈVRE. « Segment-before-Detect : Vehicle Detection and Classification through Semantic Segmentation of Aerial Images ». In : *Remote Sensing* 9.4 (avr. 2017), p. 1–18. URL : <https://hal.archives-ouvertes.fr/hal-01529624>.

- [RI-8] L. MOU, X. ZHU, M. VAKALOPOULOU, K. KARANTZALOS, N. PARAGIOS, B. LE SAUX, G. MOSER et D. TUIA. « Multitemporal Very High Resolution From Space : Outcome of the 2016 IEEE GRSS Data Fusion Contest ». In : *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing* 10.8 (août 2017), p. 3435–3447.
- [RI-9] D. TUIA, G. MOSER, B. LE SAUX, B. BECHTEL et L. SEE. « The 2017 IEEE Geoscience and Remote Sensing Society Data Fusion Contest : Open Data for Global Multimodal Land Use Classification [Technical Committees] ». In : *IEEE Geoscience and Remote Sensing Magazine* 5.4 (déc. 2017), p. 110–114.
- [RI-10] A. BOULCH, J. GUERRY, B. LE SAUX et N. AUDEBERT. « SnapNet : 3D point cloud semantic labeling with 2D deep segmentation networks ». In : *Computers & Graphics* 71 (2017), p. 189–198.
- [RI-11] N. AUDEBERT, B. LE SAUX et S. LEFEVRE. « Beyond RGB : Very High Resolution Urban Remote Sensing With Multimodal Deep Networks ». In : *ISPRS Journal of Photogrammetry and Remote Sensing* 140 (2018). Geospatial Computer Vision, p. 20–32. URL : <http://www.sciencedirect.com/science/article/pii/S0924271617301818>.
- [RI-12] B. LE SAUX, N. YOKOYA, R. HANSCH et S. PRASAD. « 2018 IEEE GRSS Data Fusion Contest : Multimodal Land Use Classification [Technical Committees] ». In : *IEEE Geoscience and Remote Sensing Magazine* 6.1 (mar. 2018), p. 52–54.
- [RI-13] N. YOKOYA, P. GHAMISI, J. XIA, S. SUKHANOV, R. HEREMANS, C. DEBES, B. BECHTEL, B. LE SAUX, G. MOSER et D. TUIA. « Open data for global multimodal land use classification : Outcome of the 2017 IEEE GRSS Data Fusion Contest ». In : *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing* 11.5 (mai 2018), p. 1363–1377.
- [RI-14] B. LE SAUX, N. YOKOYA, R. HANSCH et S. PRASAD. « Advanced Multisource Optical Remote Sensing for Urban Land Use and Land Cover Classification [Technical Committees] ». In : *IEEE Geoscience and Remote Sensing Magazine* 6.4 (déc. 2018), p. 85–89.
- [RI-15] B. LE SAUX, N. YOKOYA, R. HANSCH, M. BROWN et G. HAGER. « 2019 Data Fusion Contest [Technical Committees] ». In : *IEEE Geoscience and Remote Sensing Magazine* 7.1 (mar. 2019), p. 103–105.
- [RI-16] N. AUDEBERT, B. LE SAUX et S. LEFÈVRE. « Deep Learning for Classification of Hyperspectral Data : A Comparative Review ». In : *IEEE Geoscience Remote Sensing Magazine* 7.1 (juin 2019).
- [RI-17] Y. XU, B. DU, L. ZHANG, D. CERRA, M. PATO, E. CARMONA, S. PRASAD, N. YOKOYA, R. HÄNSCH et B. LE SAUX. « Advanced multi-sensor optical remote sensing for urban land use and land cover classification : Outcome of the 2018 IEEE GRSS Data Fusion Contest ». In : *IEEE Journal of Selected Topics in Applied Earth Obs. and Remote Sensing* 12.6 (juin 2019), p. 1709–1724.
- [RI-18] R. DAUDT, B. LE SAUX, A. BOULCH et Y. GOUSSEAU. « Multitask learning for large-scale semantic change detection ». In : *Computer Vision and Image Understanding* 187 (juil. 2019), p. 102783.

- [RI-19] N. AUDEBERT, A. BOULCH, B. LE SAUX et S. LEFÈVRE. « Distance transform regression for spatially-aware deep semantic segmentation ». In : *Computer Vision and Image Understanding* 189 (sept. 2019), p. 102809.
- [RI-20] M. CARVALHO, B. LE SAUX, P. TROUVÉ-PELOUX, A. ALMANSA et F. CHAMPAGNAT. « Multi-Task Learning of Height and Semantics from Aerial Images ». In : *IEEE Geosci. and Remote Sensing Letters* (nov. 2019).
- [RI-21] B. LE SAUX, N. YOKOYA, R. HANSCH et M. BROWN. « Report on the 2019 IEEE GRSS Data Fusion Contest : Large-Scale Semantic 3D Reconstruction [Technical Committees] ». In : *IEEE Geoscience and Remote Sensing Magazine* 7.4 (déc. 2019).

Articles de revue nationale avec comité de lecture

- [RN-1] M. SANFOURCHE, B. LE SAUX, A. PLYER et G. LE BESNERAIS. « Cartographie et interprétation de l'environnement par drone ». In : *Revue Française de Photogrammétrie et Télédétection* 213-214 (jan. 2017), p. 55–62.

Communications en congrès internationaux avec actes et comité de lecture

- [CI-1] N. BOUJEMAA, F. FAUQUEUR, M. FERECATU, F. FLEURET, V. GOUET, B. LE SAUX et H. SAHBI. « Interactive Specific and Generic Image Retrieval ». In : *Proc. of MultiMedia and Content-Based Information Retrieval*. 2001.
- [CI-2] B. LE SAUX et N. BOUJEMAA. « Unsupervised Categorization for Image Database Overview ». In : *Int. Conf. on Visual Information System (VISUAL), LNCS 2314*. Hsin-chu, Taiwan, mar. 2002.
- [CI-3] B. LE SAUX et N. BOUJEMAA. « Unsupervised Robust Clustering for Image Database Categorization ». In : *IEEE-IAPR Int. Conf. on Pattern Recognition (ICPR)*. Quebec, Canada, août 2002.
- [CI-4] B. LE SAUX, N. GRIRA et N. BOUJEMAA. « Adaptive Robust Clustering with Proximity-Based Merging for Video-Summary ». In : *IEEE Int. Conf. on Fuzzy Systems (FUZZ-IEEE)*. Saint-Louis, USA, mai 2003.
- [CI-5] B. LE SAUX et G. AMATO. « Image recognition for digital libraries ». In : *ACM MultiMedia/Int. Workshop on Multimedia Information Retrieval*. New-York, USA, oct. 2004, p. 91–98.
- [CI-6] B. LE SAUX et G. AMATO. « Image classifiers for scene analysis ». In : *Int. Conf. on Computer Vision and Graphics (ICCVG)*. Warsaw, Poland, sept. 2004.
- [CI-7] B. LE SAUX et N. BOUJEMAA. « Image database clustering with SVM-based class personalization ». In : *IS&T SPIE Conf. on Storage and Retrieval Methods and Applications for Multimedia*. San José, USA, jan. 2004.

- [CI-8] B. LE SAUX et H. BUNKE. « Feature selection for graph-based image classifiers ». In : *IAPR Iberian Conf. on Pattern Recognition and Image Analysis (IbPRIA)*. Estoril, Portugal, juin 2005.
- [CI-9] B. LE SAUX et H. BUNKE. « Combining SVM and Graph Matching in a Multiple Classifier System for Image Content Recognition ». In : *IEEE-IAPR Int. Conf. on Pattern Recognition / Statistical Techniques in Pattern Recognition Workshop (ICPR/S+SSPR)*. Hong Kong, China, août 2006.
- [CI-10] B. LE SAUX, B. CHALMOND, Y. YU, A. TROUVÉ, O. RENAUD et S. L. SHORTE. « Micro-rotation Imaging Deconvolution ». In : *IEEE Int. Symposium on Biomedical Imaging (ISBI) : From Nano to Macro*. Paris, France, mai 2008, p. 1367–1370.
- [CI-11] B. LE SAUX et M. SANFOURCHE. « Robust vehicle categorization from aerial images by 3D-template matching and multiple classifier system ». In : *IEEE Int. Symposium on Image and Signal Processing and Analysis (ISPA)*. Dubrovnik, Croatia, sept. 2011.
- [CI-12] F. GIANNESINI et B. LE SAUX. « GPU-accelerated One-Class SVM for exploration of remote sensing data ». In : *IEEE Int. Geoscience and Remote Sensing Symposium (IGARSS)*. Munich, Germany, juil. 2012.
- [CI-13] N. CHAUFFERT, J. ISRAËL et B. LE SAUX. « Boosting for interactive man-made structure classification ». In : *IEEE Int. Geoscience and Remote Sensing Symposium (IGARSS)*. Munich, Germany, juil. 2012.
- [CI-14] B. LE SAUX et H. RANDRIANARIVO. « Urban change detection in SAR images by interactive learning ». In : *IEEE Int. Geoscience and Remote Sensing Symposium (IGARSS)*. Melbourne, Australia, juil. 2013.
- [CI-15] H. RANDRIANARIVO, B. LE SAUX et M. FERECATU. « Man-made structure detection with deformable part-based models ». In : *IEEE Int. Geoscience and Remote Sensing Symposium (IGARSS)*. Melbourne, Australia, juil. 2013.
- [CI-16] B. LE SAUX et M. SANFOURCHE. « Rapid Semantic Mapping : Learn Environment Classifiers On the Fly ». In : *Int. Conf. on Intelligent Robots and Systems (IROS)*. Tokyo, Japan, 2013.
- [CI-17] H. RANDRIANARIVO, B. LE SAUX et M. FERECATU. « Multimodal Classification with Deformable Part Models for Urban Cartography ». In : *IEEE Int. Geoscience and Remote Sensing Symposium (IGARSS)*. Quebec City, Canada, juil. 2014.
- [CI-18] B. LE SAUX. « Interactive Design of Object Classifiers in Remote Sensing ». In : *IEEE-IAPR Int. Conf. on Pattern Recognition (ICPR)*. Stockholm, Sweden, août 2014.
- [CI-19] M. SANFOURCHE, B. LE SAUX, A. PLYER et G. LE BESNERAIS. « Environment Mapping and Interpretation by Drone ». In : *Joint Urban Remote Sensing Event (JURSE)*. Lausanne, Switzerland, mai 2015.
- [CI-20] A. LAGRANGE, B. LE SAUX, A. BEAUPÈRE, A. BOULCH, A. CHAN-HON-TONG, S. HERBIN, H. RANDRIANARIVO et M. FERECATU. « Benchmarking classification of Earth-observation data : from learning explicit features to convolutional networks ». In : *IEEE Int. Geoscience and Remote Sensing Symposium (IGARSS)*. Milan, Italy, juil. 2015.

- [CI-21] H. RANDRIANARIVO, B. LE SAUX, M. CRUCIANU et M. FERECATU. « Discriminatively-trained model mixture for object detection in aerial images ». In : *ESA Image Info. Mining (IIM)*. Bucarest, Romania, 2015.
- [CI-22] H. RANDRIANARIVO, B. LE SAUX, N. AUDEBERT, M. CRUCIANU et M. FERECATU. « Structural classifiers for contextual semantic labeling of aerial images ». In : *ESA Big Data in Space (BiDS)*. Tenerife, Spain, 2016.
- [CI-23] N. AUDEBERT, B. LE SAUX et S. LEFÈVRE. « How useful is region-based classification of remote sensing images in a deep learning framework ? » In : *IEEE Int. Geoscience and Remote Sensing Symposium (IGARSS)*. Beijing, China, 2016.
- [CI-24] N. AUDEBERT, B. LE SAUX et S. LEFÈVRE. « On the usability of deep networks for object-based image analysis ». In : *Conf. on Geo Object-Based Image Analysis (GEOBIA)*. Enschede, Netherlands, 2016.
- [CI-25] N. AUDEBERT, B. LE SAUX et S. LEFÈVRE. « Semantic Segmentation of Earth Observation Data Using Multimodal and Multi-scale Deep Networks ». In : *Asian Conf. on Computer Vision (ACCV)*. Taipei, Taiwan, nov. 2016. URL : <https://hal.archives-ouvertes.fr/hal-01360166>.
- [CI-26] N. AUDEBERT, A. BOULCH, H. RANDRIANARIVO, B. LE SAUX, S. LEFÈVRE et R. MARLET. « Deep learning for Urban Remote Sensing ». In : *Joint Urban Remote Sensing Event (JURSE)*. Dubai, UAE, mar. 2017.
- [CI-27] N. AUDEBERT, B. LE SAUX et S. LEFÈVRE. « Fusion of heterogeneous data in convolutional networks for urban semantic labeling ». In : *Joint Urban Remote Sensing Event (JURSE)*. Dubai, UAE, mar. 2017.
- [CI-28] F. LIMBERGER, R. WILSON, M. AONO, A. BOULCH, B. BUSTOS, A. GIACHETTI, A. GODIL, B. LE SAUX, B. LI, Y. LU, H.-D. NGUYEN, V.-T. NGUYEN, V.-K. PHAM, I. SIPIRAN, A. TATSUMA, M.-T. TRAN et S. VELASCO-FORERO. « SHREC : Point-Cloud Shape Retrieval of Non-Rigid Toys ». In : *Eurographics Workshop on 3D Object Retrieval (EG/3DOR)*. Sous la dir. d'I. PRATIKAKIS, F. DUPONT et M. OVSJANIKOV. Lyon, France : The Eurographics Association, 2017.
- [CI-29] Q. DE SMEDT, H. WANNOUS, J.-P. VANDEBORRE, J. GUERRY, B. LE SAUX et D. FILLIAT. « SHREC : 3D Hand Gesture Recognition Using a Depth and Skeletal Dataset ». In : *Eurographics Workshop on 3D Object Retrieval (EG/3DOR)*. Sous la dir. d'I. PRATIKAKIS, F. DUPONT et M. OVSJANIKOV. Lyon, France : The Eurographics Association, 2017.
- [CI-30] A. BOULCH, B. LE SAUX et N. AUDEBERT. « Unstructured Point Cloud Semantic Labeling Using Deep Segmentation Networks ». In : *Eurographics Workshop on 3D Object Retrieval (EG/3DOR)*. Sous la dir. d'I. PRATIKAKIS, F. DUPONT et M. OVSJANIKOV. Lyon, France : The Eurographics Association, 2017.
- [CI-31] N. RIVIERE, A. AMDITIS, A. AMIEZ, G. ATHANASIOU, J. BERGGREN, A. BOULCH, N. BOZABALIAN, D. DUARTE, P.-E. DUPOUY, P. ESCALAS, M. GERKE, F. GIROUD, C. GRAND, N. KERLE, Y. LAMBERT, F. NEX, B. LE SAUX, A. SCHILLING et G. TOLD. « 3D laser imaging techniques to improve USaR operations for wide-area surveillance and monitoring of collapsed buildings ». In : *Int. Conf. on Information Systems for Crisis Response And Management (ISCRAM)*. Albi, France, 2017.

- [CI-32] N. AUDEBERT, B. LE SAUX et S. LEFÈVRE. « Joint Learning from Earth Observation and OpenStreetMap Data to Get Faster Better Semantic Maps ». In : *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR) Workshops / IEEE/ISPRS EarthVision : Large Scale Computer Vision for Remote Sensing Imagery*. Honolulu, United States, juil. 2017. URL : <https://hal.archives-ouvertes.fr/hal-01523573>.
- [CI-33] J. GUERRY, B. LE SAUX et D. FILLIAT. « “Look At This One”, Detection sharing between modality-independent classifiers for robotic discovery of people ». In : *Eur. Conf. on Mobile Robotics (ECMR)*. Paris, France, 2017.
- [CI-34] J. GUERRY, A. BOULCH, B. LE SAUX, A. PLYER, J. MORAS et D. FILLIAT. « SnapNet-R : Consistent 3D Multi-View Semantic Labeling for Robotics ». In : *Int. Conf. of Computer Vision / Workshop on 3D Reconstruction meets Semantics (ICCV/3DRMS)*. Venice, Italy, oct. 2017.
- [CI-35] B. HUANG, K. LU, N. AUDEBERT, A. KHALEL, Y. TARABALKA, J. MALOF, A. BOULCH, B. LE SAUX, L. COLLINS, K. BRADBURY, S. LEFÈVRE et M. EL-SABAN. « Large-scale semantic classification : outcome of the first year of INRIA Aerial Image Labeling Benchmark ». In : *IEEE Int. Geoscience and Remote Sensing Symposium (IGARSS)*. Valencia, Spain, juil. 2018.
- [CI-36] A. BOULCH, P. TROUVÉ-PELOUX, É. KOENIGUER, F. JANEZ et B. LE SAUX. « Learning speckle suppression in SAR images without ground truth : application to Sentinel-1 time-series ». In : *IEEE Int. Geoscience and Remote Sensing Symposium (IGARSS)*. Valencia, Spain, juil. 2018.
- [CI-37] B. LE SAUX, A. BEAUPÈRE, A. BOULCH, J. BROSSARD, A. MANIER et G. VILLEMEN. « Railway detection : from filtering to segmentation networks ». In : *IEEE Int. Geoscience and Remote Sensing Symposium (IGARSS)*. Valencia, Spain, juil. 2018.
- [CI-38] R. CAYE DAUDT, B. LE SAUX, A. BOULCH et Y. GOUSSEAU. « Urban Change Detection for Multispectral Earth Observation Using Convolutional Neural Networks ». In : *IEEE Int. Geoscience and Remote Sensing Symposium (IGARSS)*. Valencia, Spain, juil. 2018.
- [CI-39] N. AUDEBERT, B. LE SAUX et S. LEFÈVRE. « Generative adversarial networks for realistic synthesis of hyperspectral samples ». In : *IEEE Int. Geoscience and Remote Sensing Symposium (IGARSS)*. Valencia, Spain, juil. 2018.
- [CI-40] M. CARVALHO, B. LE SAUX, P. TROUVÉ-PELOUX, F. CHAMPAGNAT et A. ALMANSA. « Deep Depth from Defocus : how can defocus blur improve 3D estimation using dense neural networks? » In : *IEEE Eur. Conf. on Computer Vision / Workshop on 3D Reconstruction in the Wild (ECCV / 3DRIW)*. Munich, Germany, sept. 2018.
- [CI-41] M. CARVALHO, B. LE SAUX, P. TROUVÉ-PELOUX, F. CHAMPAGNAT et A. ALMANSA. « On Regression Losses for Deep Depth Estimation ». In : *IEEE Int. Conf. on Image Processing (ICIP)*. Athens, Greece, oct. 2018.
- [CI-42] R. CAYE DAUDT, B. LE SAUX, A. BOULCH et Y. GOUSSEAU. « Fully Convolutional Siamese Networks for Change Detection ». In : *IEEE Int. Conf. on Image Processing (ICIP)*. Athens, Greece, oct. 2018.

- [CI-43] J. CASTILLO-NAVARRO, N. AUDEBERT, A. BOULCH, B. LE SAUX et S. LEFÈVRE. « What Data are needed for Deep Learning in Earth Observation ? » In : *Joint Urban Remote Sensing Event (JURSE)*. Vannes, France, mai 2019.
- [CI-44] R. CAYE DAUDT, B. LE SAUX, A. BOULCH et Y. GOUSSEAU. « Guided Anisotropic Diffusion and Iterative Learning for Weakly Supervised Change Detection ». In : *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR) Workshops / IEEE/ISPRS EarthVision : Large Scale Computer Vision for Remote Sensing Imagery*. Long Beach, United States, juin 2019.
- [CI-45] R. CAYE DAUDT, A. CHAN-HON-TONG, B. LE SAUX et A. BOULCH. « Learning to understand Earth-observation images with weak and unreliable ground-truth ». In : *IEEE Int. Geoscience and Remote Sensing Symposium (IGARSS)*. Yokohama, Japan, juil. 2019.
- [CI-46] G. MOSER, F. DELL'ACQUA, R. HÄNSCH, J. KERÉKES, B. LE SAUX, L. PIERCE et N. YOKOYA. « The IEEE GRSS Data and Algorithms Standard Evaluation (DASE) platform and the IEEE GRSS Data Fusion Contest initiative ». In : *Int. Symposium on Digital Earth*. Florence, Italy, sept. 2019.

Communications en congrès nationaux avec actes et comité de lecture

- [CN-1] B. LE SAUX. « Apprentissage interactif par Online Gradient Boost en télédétection ». In : *Colloque GRETSI*. Brest, France, sept. 2013.
- [CN-2] M. SANFOURCHE, B. LE SAUX, A. PLYER et G. LE BESNERAIS. « Cartographie et interprétation de l'environnement par drone ». In : *Congrès de la Société Française de Photogrammétrie et Télédétection - Colloque Drones*. Montpellier, France, 2014.
- [CN-3] H. RANDRIANARIVO, B. LE SAUX et M. FERECATU. « Détection de véhicules en imagerie aérienne par mélange de modèles discriminatifs ». In : *Colloque GRETSI*. Lyon, France, 2015.
- [CN-4] T. DUMAS, B. LE SAUX et P. TROUVÉ-PELOUX. « Réseaux de neurones profonds pour estimer la profondeur grâce au flou de défocalisation ». In : *Colloque GRETSI*. Lyon, France, 2015.
- [CN-5] J. GUERRY, B. LE SAUX et D. FILLIAT. « Sélection d'algorithmes de classification par réseau de neurones ». In : *Conf. de Reconnaissance des Formes et Intelligence Artificielle (RFIA)*. Clermont-Ferrand, France, 2016.
- [CN-6] J. GUERRY, B. LE SAUX et D. FILLIAT. « RCNN RGBD pour la détection de personnes en conditions difficiles ». In : *Colloque GRETSI*. Juan-les-Pins, France, sept. 2017.
- [CN-7] M. CARVALHO, B. LE SAUX, P. TROUVÉ-PELOUX, A. ALMANSA et F. CHAMPAGNAT. « Estimation de profondeur à partir d'une seule image avec un réseau adversaire ». In : *Colloque GRETSI*. Juan-les-Pins, France, sept. 2017.

- [CN-8] N. AUDEBERT, B. LE SAUX et S. LEFÈVRE. « Couplage de données géographiques participatives et d'images aériennes par apprentissage profond ». In : *Colloque GRETSI*. Juan-les-Pins, France, sept. 2017.
- [CN-9] R. CAYE DAUDT, B. LE SAUX, A. BOULCH et Y. GOUSSEAU. « Détection dense de changements par réseaux de neurones siamois ». In : *Congrès Reconnaissance des Formes, Image, Apprentissage et Perception (RFIAP)*. Marne-la-Vallée, France, mai 2018.
- [CN-10] N. AUDEBERT, A. BOULCH, B. LE SAUX et S. LEFÈVRE. « Segmentation sémantique profonde par régression sur cartes de distances signées ». In : *Congrès Reconnaissance des Formes, Image, Apprentissage et Perception (RFIAP)*. Marne-la-Vallée, France, mai 2018.
- [CN-11] M. CARVALHO, B. LE SAUX, P. TROUVÉ-PELOUX, F. CHAMPAGNAT et A. ALMANSA. « Estimation de profondeur monoculaire par réseau de neurones et l'apport du flou de défocalisation ». In : *Congrès Reconnaissance des Formes, Image, Apprentissage et Perception (RFIAP)*. Marne-la-Vallée, France, mai 2018.
- [CN-12] J. CASTILLO-NAVARRO, B. LE SAUX, A. BOULCH et S. LEFÈVRE. « Réseaux de neurones semi-supervisés pour la segmentation sémantique en télédétection ». In : *Colloque GRETSI*. Lille, France, sept. 2019.

Thèse et manuscrits

- [PhD-2003] B. LE SAUX. « Classification non exclusive et personnalisation par apprentissage : Application à la navigation dans les bases d'images ». 2003.

Références

- [1] A. OLIVA et A. TORRALBA. « Modeling the Shape of the Scene : A Holistic Representation of the Spatial Envelope ». In : *International Journal of Computer Vision (IJCV)* 42.3 (mai 2001), p. 145–175. URL : <https://doi.org/10.1023/A:1011139631724>.
- [2] F. BREMOND. « Scene Understanding : perception, multi-sensor fusion, spatio-temporal reasoning and activity recognition ». Habilitation à diriger des recherches. Université Nice Sophia Antipolis, juil. 2007. URL : <https://tel.archives-ouvertes.fr/tel-00275889>.
- [3] J. VOGEL et B. SCHIELE. « Semantic Modeling of Natural Scenes for Content-Based Image Retrieval ». In : *International Journal of Computer Vision (IJCV)* 72.2 (avr. 2007), p. 133–157. URL : <https://doi.org/10.1007/s11263-006-8614-1>.
- [4] A. TORRALBA et A. OLIVA. « Depth Estimation from Image Structure ». In : *IEEE Trans. Pattern Analysis and Machine Intelligence (TPAMI)* 24.9 (sept. 2002), p. 1226–1238. URL : <https://doi.org/10.1109/TPAMI.2002.1033214>.
- [5] Z. ZIA, M. STARK et K. SCHINDLER. « Towards Scene Understanding with Detailed 3D Object Representations ». In : *International Journal of Computer Vision (IJCV)* 112.2 (2015), p. 188–203.
- [6] A. SMEULDERS, M. WORRING, S. SANTINI, A. GUPTA et R. JAIN. « Content based image retrieval at the end of the early years ». In : *IEEE Trans. on Pattern Analysis and Machine Intelligence (TPAMI)* 22.12 (2000), p. 1349–1380.
- [7] M. CORD, J. FOURNIER, P.-H. GOSSELIN et S. PHILIPP-FOLIGUET. « Interactive Exploration for Image Retrieval ». In : *EURASIP Journal on Applied Signal Processing* 14 (2006), p. 2173–2186.
- [8] M. FLICKHER, H. SAWHNEY, W. NIBLACK, J. ASHLEY, Q. HUANG, B. DOM, M. GORKANI, J. HAFNER, D. LEE, D. PETKOVIC, D. STEELE et P. YANKER. « Query by Image and Video Content : The QBIC System ». In : *IEEE Computer* 28.9 (1995), p. 23–32.
- [9] W.-Y. MA et B. MANJUNATH. « NeTra : A toolbox for navigating large image databases ». In : *Multimedia Systems* 7.3 (1999), p. 184–198.
- [10] J. FOURNIER, M. CORD et S. PHILIPP-FOLIGUET. « RETIN : A Content-Based Image Indexing and Retrieval System ». In : *Pattern Analysis and Applications* 4.2-3 (2001), p. 153–173.

- [11] S. SCLAROFF, M. LA CASCIA, S. SETHI et L. TAYCHER. « Unifying Textual and Visual Cues for Content-Based Image Retrieval on the World Wide Web ». In : *Computer Vision and Image Understanding (CVIU)* 75.1/2 (1999), p. 86–98.
- [12] M. EVERINGHAM, L. VAN-GOOL, C. K. I. WILLIAMS, J. WINN et A. ZISSERMAN. « The Pascal Visual Object Classes (VOC) Challenge ». In : *International Journal of Computer Vision (IJCV)* 88.2 (juin 2010), p. 303–338.
- [13] O. RUSSAKOVSKY, J. DENG, H. SU, J. KRAUSE, S. SATHEESH, S. MA, Z. HUANG, A. KARPATY, A. KHOSLA, M. BERNSTEIN, A. C. BERG et L. FEI-FEI. « ImageNet Large Scale Visual Recognition Challenge ». In : *International Journal of Computer Vision (IJCV)* 115.3 (2015), p. 211–252.
- [14] J. DENG, W. DONG, R. SOCHER, L.-J. LI, K. LI et L. FEI-FEI. « ImageNet : A large-scale hierarchical image database ». In : *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. IEEE. 2009, p. 248–255. URL : <https://ieeexplore.ieee.org/abstract/document/5206848/>.
- [15] O. CHAPELLE, P. HAFFNER et V. VAPNIK. « SVMs for Histogram-Based Image Classification ». In : *IEEE Trans. on Neural Networks* 10 (1999), p. 1055–1065.
- [16] C. CARSON, M. THOMAS, S. BELONGIE, J. HELLERSTEIN et J. MALIK. « Blobworld : A System for Region-Based Image Indexing and Retrieval ». In : *VISUAL 1999. LNCS*. T. 1614. Déc. 1998, p. 509–517.
- [17] G. QIU. « Indexing chromatic and achromatic patterns for content-based colour image retrieval ». In : *Pattern Recognition* 35.8 (2002), p. 1675–1686.
- [18] J. SIVIC et A. ZISSERMAN. « Video Google : A Text Retrieval Approach to Object Matching in Videos ». In : *The IEEE International Conference on Computer Vision (ICCV)*. T. 2. Nice, France, 2003, p. 1470–1477.
- [19] P. DUYGULU, K. BARNARD, J. DE FREITAS et D. FORSYTH. « Object Recognition as Machine Translation : Learning a Lexicon for a Fixed Image Vocabulary ». In : *The European Conference on Computer Vision (ECCV)*. T. 4. Copenhagen, Denmark, mai 2002, p. 97–112.
- [20] G. CSURKA, C. R. DANCE, L. FAN, J. WILLAMOWSKI et C. BRAY. « Visual categorization with bags of keypoints ». In : *The European Conference on Computer Vision (ECCV) / Workshop on Statistical Learning in Computer Vision*. 2004, p. 1–22.
- [21] L. FEI-FEI et P. PERONA. « A Bayesian hierarchical model for learning natural scene categories ». In : *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. Juin 2005.
- [22] H. FRIGUI et R. KRISHNAPURAM. « Clustering by competitive agglomeration ». In : *Pattern Recognition* 30.7 (1997), p. 1109–1119.
- [23] E. E. GUSTAFSON et W. C. KESSEL. « Fuzzy clustering with a fuzzy covariance matrix ». In : *IEEE CDC*. San Diego, California, 1979, p. 761–766.
- [24] R. N. DAVÉ. « Characterization and detection of noise in clustering ». In : *Pattern Recognition Letters* 12.11 (1991), p. 657–664.
- [25] P. ANANDAN. « Personal Digital Media : It’s about sharing experiences ». In : *Proceedings of MMCBIR 2001*. Rocquencourt, France, 2001.

- [26] D. COMANICIU et P. MEER. « Robust Analysis of Feature Spaces : Color Image Segmentation ». In : *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. San Juan, Porto Rico, juin 1997, p. 750–755.
- [27] R. BATTITI. « Using mutual information for selecting features in supervised neural network learning ». In : *IEEE Trans. on Neural Networks* 5.4 (1994), p. 537–550.
- [28] F. FLEURET. « Fast Binary Feature Selection with Conditional Mutual Information ». In : *Journal of Machine Learning Research (JMLR)* 5 (nov. 2004), p. 1531–1555.
- [29] J. ANLAUF et M. BIEHL. « The Adatron : an Adaptive Perceptron Algorithm ». In : *Neurophysics Letters* 10 (1989), p. 687–692.
- [30] T.-T. FRIESS, N. CHRISTIANINI et C. CAMPBELL. « The Kernel-Adatron Algorithm : a Fast and Simple Learning Procedure for Support Vector Machines ». In : *International Conference on Machine Learning (ICML)*. Madison, Wisconsin, juil. 1998.
- [31] V. VAPNIK. *The Nature of Statistical Learning Theory*. New-York, N.Y. : Springer Verlag, 1995.
- [32] G. AMATO, C. GENNARO, P. SAVINO et F. RABITTI. « MILOS : a Multimedia Content Management System for Digital Library Applications ». In : *European Conference on Digital Libraries*. Bath, U.K., sept. 2004.
- [33] H. DURRANT-WHYTE et T. BAILEY. « Simultaneous localization and mapping : part I & II ». In : *IEEE Robotics Automation Magazine* 13.2 (juin 2006), p. 99–110.
- [34] M. CUMMINS et P. NEWMAN. « FAB-MAP : Probabilistic Localization and Mapping in the Space of Appearance ». In : *The International Journal of Robotics Research* 27.6 (2008), p. 647–665.
- [35] F. FRAUNDORFER, L. HENG, D. HONEGGER, G.-H. LEE, L. MEIER, P. TANSKANEN et M. POLLEFEYS. « Vision-based autonomous mapping and exploration using a quadrotor MAV ». In : *IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*. Vilamoura, Portugal, 2012.
- [36] A. NUECHTER et J. HERTZBERG. « Towards semantic maps for mobile robots ». In : *Robotics and Autonomous Systems* 56.11 (2008), p. 915–926.
- [37] O. MOZOS, R. TRIEBEL, P. JENSFELT, A. ROTTMAN et W. BURGARD. « Supervised semantic labeling of places using information extracted from sensor data ». In : *Robotics and Autonomous Systems* 55.5 (2007), p. 391–402.
- [38] P.-E. FORSSSEN, D. MEGER, K. LAI, S. HELMER, J. LITTLE et D. LOWE. « Informed visual search : combining attention and object recognition ». In : *IEEE International Conference on Robotics and Automation (ICRA)*. Pasadena, California, USA, 2008.
- [39] M. ANDRILUKA, P. SCHNITZSPAN, J. MEYER, S. KOHLBRECHER, K. PETERSEN, O. VON STRYK, S. ROTH et B. SCHIELE. « Vision Based Victim Detection from Unmanned Aerial Vehicles ». In : *IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*. Taipei, Taiwan, 2010.
- [40] M. LOURAKIS et A. ARGYROS. « SBA : A Software Package for Generic Sparse Bundle Adjustment ». In : *ACM Trans. Math. Software* 36.1 (2009), p. 1–30.

- [41] N. DALAL et B. TRIGGS. « Histograms of Oriented Gradients for Human Detection ». In : *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. Washington DC, USA, 2005, p. 886–893.
- [42] T. OJALA, M. PIETIKAINEN et D. HARWOOD. « A Comparative Study of Texture Measures with Classification Based on Feature Distributions ». In : *Pattern Recognition* 29 (1996), p. 51–59.
- [43] J. ZHANG, K. HUANG, Y. YU et T. TAN. « Boosted Local Structured HOG-LBP for Object Localization ». In : *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. Colorado Springs, USA, 2011.
- [44] C. LEISTNER, A. SAFFARI, P. ROTH et H. BISCHOF. « On Robustness of On-line Boosting : A Competitive Study ». In : *The IEEE International Conference on Computer Vision (ICCV) / Workshop on On-line Learning for Computer Vision*. Kyoto, Japan, 2009.
- [45] Y. FREUND et R. E. SCHAPIRE. « A decision-theoretic generalization of on-line learning and an application to boosting ». In : *Journal of Computer and System Sciences* 55.1 (1997).
- [46] L. MASON, J. BAXTER, P. BARTLETT et M. FREAN. « Boosting Algorithms as Gradient Descent ». In : *Advances in Neural Information Processing Systems 13 (NIPS)*. T. 13. 2000, p. 512–518.
- [47] P. M. LONG et R. A. SERVEDIO. « Random Classification Noise Defeats All Convex Potential Boosters ». In : *Machine Learning* 78.3 (2010), p. 287–304.
- [48] J. GUERRY. « Reconnaissance visuelle robuste par réseaux de neurones dans des scénarios d’exploration robotique. Détecte-moi si tu peux! » 2017. URL : <https://hal.archives-ouvertes.fr/tel-01680372/>.
- [49] S. SONG, S. P. LICHTENBERG et J. XIAO. « SUN RGB-D : A RGB-D scene understanding benchmark suite ». In : *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. Boston, USA, 2015, p. 567–576.
- [50] A. KRIZHEVSKY, I. SUTSKEVER et G. E. HINTON. « ImageNet Classification with Deep Convolutional Neural Networks ». In : *Advances in Neural Information Processing Systems 25 (NIPS)*. 2012, p. 1097–1105.
- [51] R. GIRSHICK, J. DONAHUE, T. DARRELL et J. MALIK. « Rich feature hierarchies for accurate object detection and semantic segmentation ». In : *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. 2014.
- [52] R. GIRSHICK. « Fast R-CNN ». In : *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. 2015.
- [53] S. REN, K. HE, R. GIRSHICK et J. SUN. « Faster R-CNN : Towards real-time object detection with region proposal networks ». In : *Advances in Neural Information Processing Systems 28 (NIPS)*. 2015.
- [54] J. LONG, E. SHELHAMER et T. DARRELL. « Fully Convolutional Networks for Semantic Segmentation ». In : *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. Juin 2015, p. 3431–3440.

- [55] V. BADRINARAYANAN, A. KENDALL et R. CIPOLLA. « SegNet : A Deep Convolutional Encoder-Decoder Architecture for Scene Segmentation ». In : *IEEE Trans. on Pattern Analysis and Machine Intelligence (TPAMI)* 39.12 (déc. 2017), p. 2481–2495.
- [56] O. RONNEBERGER, P. FISCHER et T. BROX. « U-Net : Convolutional Networks for Biomedical Image Segmentation ». In : *MICCAI*. Munich, 2015, p. 234–241. URL : http://dx.doi.org/10.1007/978-3-319-24574-4_28.
- [57] K. LAI, L. BO, X. REN et D. FOX. « Sparse distance learning for object recognition combining rgb and depth information ». In : *IEEE International Conference on Robotics and Automation (ICRA)*. IEEE. 2011, p. 4007–4013.
- [58] S. GUPTA, R. GIRSHICK, P. ARBELÁEZ et J. MALIK. « Learning rich features from RGB-D images for object detection and segmentation ». In : *The European Conference on Computer Vision (ECCV)*. 2014.
- [59] C. HAZIRBAS, L. MA, C. DOMOKOS et D. CREMERS. « FuseNet : Incorporating Depth into Semantic Segmentation via Fusion-based CNN Architecture ». In : *Proc. ACCV*. T. 2. 2016.
- [60] L. MA, J. STUECKLER, C. KERL et D. CREMERS. « Multi-View Deep Learning for Consistent Semantic Mapping with RGB-D Cameras ». In : *IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*. Vancouver, Canada, sept. 2017.
- [61] D. C. CIREŞAN, U. MEIER, L. M. GAMBARDELLA et J. SCHMIDHUBER. « Deep, Big, Simple Neural Nets for Handwritten Digit Recognition ». In : *Neural Computation* 22.12 (2010), p. 3207–3220.
- [62] N. CHAWLA, K. BOWYER, L. HALL et W. KEGELMEYER. « SMOTE : Synthetic Minority Over-sampling Technique ». In : *J. Artif. Intell. Res. (JAIR)* 16 (jan. 2002), p. 321–357.
- [63] Z. LI, Y. GAN, X. LIANG, Y. YU, H. CHENG et L. LIN. « LSTM-CF : Unifying context modeling and fusion with LSTMs for RGB-D scene labeling ». In : *The European Conference on Computer Vision (ECCV)*. Springer. 2016, p. 541–557.
- [64] A. KENDALL, V. BADRINARAYANAN et R. CIPOLLA. « Bayesian SegNet : Model Uncertainty in Deep Convolutional Encoder-Decoder Architectures for Scene Understanding ». In : *British Machine Vision Conference (BMVC)*. Sept. 2017.
- [65] G. LIN, C. SHEN, A. VAN DEN HENGEL et I. REID. « Exploring context with deep structured models for semantic segmentation ». In : *IEEE Trans. on Pattern Analysis and Machine Intelligence (TPAMI)* (2017).
- [66] J. JIANG, Z. ZHANG, Y. HUANG et L. ZHENG. « Incorporating depth into both CNN and CRF for indoor semantic segmentation ». In : 2017, p. 525–530.
- [67] X. QI, R. LIAO, J. JIA, S. FIDLER et R. URTASUN. « 3D Graph Neural Networks for RGBD Semantic Segmentation ». In : *The IEEE International Conference on Computer Vision (ICCV)*. 2017.
- [68] N. SILBERMAN, D. HOIEM, P. KOHLI et R. FERGUS. « Indoor Segmentation and Support Inference from RGBD Images ». In : *The European Conference on Computer Vision (ECCV)*. 2012.

- [69] D. EIGEN et R. FERGUS. « Predicting depth, surface normals and semantic labels with a common multi-scale convolutional architecture ». In : *The IEEE International Conference on Computer Vision (ICCV)*. 2015, p. 2650–2658.
- [70] C. COUPRIE, C. FARABET, L. NAJMAN et Y. LECUN. « Indoor semantic segmentation using depth information ». In : *Int. Conf. on Learning Representations (ICLR)*. Scottsdale, AZ, United States, mai 2013.
- [71] A. HERMANS, G. FLOROS et B. LEIBE. « Dense 3d semantic mapping of indoor scenes from rgb-d images ». In : *IEEE International Conference on Robotics and Automation (ICRA)*. IEEE. 2014, p. 2631–2638.
- [72] A. HANDA, V. PATRAUCEAN, V. BADRINARAYANAN, S. STENT et R. CIPOLLA. « Understanding real world indoor scenes with synthetic data ». In : *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. 2016, p. 4077–4085.
- [73] P. KRÄHENBÜHL et V. KOLTUN. « Efficient inference in fully connected CRFs with Gaussian edge potentials ». In : *Advances in Neural Information Processing Systems 24 (NIPS)*. 2011, p. 109–117.
- [74] O. MEES, A. EITEL et W. BURGARD. « Choosing smartly : Adaptive multi-modal fusion for object detection in changing environments ». In : *IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*. IEEE. 2016, p. 151–156.
- [75] G. SHARMA, C. MERRY, P. GOEL et M. MCCORD. « Vehicle detection in 1-m resolution satellite and airborne imagery ». In : *International Journal of Remote Sensing* 27.4 (2006).
- [76] F. MELGANI et L. BRUZZONE. « Covariance estimation with limited training samples ». In : *IEEE Trans. Geoscience and Remote Sensing (TGRS)* 37 (2001), p. 2113–2118.
- [77] F. MELGANI et L. BRUZZONE. « Classification of Hyperspectral Remote Sensing Images with Support Vector Machines ». In : *IEEE Trans. Geoscience and Remote Sensing (TGRS)* 42.8 (2004).
- [78] F. TUPIN, B. HOUSHMAND et M. DATCU. « Road detection in dense urban areas using SAR imagery and the usefulness of multiple views ». In : *IEEE Trans. on Geoscience and Remote Sensing (TGRS)* 40.11 (2002), p. 2405–2414.
- [79] A. LORETTE, X. DESCOMBES et J. ZERUBIA. « Texture Analysis through a Markovian Modelling and Fuzzy Classification : Application to Urban Area Extraction from Satellite Images ». In : *International Journal of Computer Vision (IJCV)* 36.3 (2000).
- [80] G. RELIER, X. DESCOMBES, F. FALZON et J. ZERUBIA. « Texture feature analysis using a Gauss-Markov model in hyperspectral image classification ». In : *IEEE Trans. on Geoscience and Remote Sensing (TGRS)* 42.7 (juil. 2004), p. 1543–1551.
- [81] M. WALESSA et M. DATCU. « Model-based Despeckling and Information Extraction from SAR Images ». In : *IEEE Trans. on Geoscience and Remote Sensing (TGRS)* 38.5 (2000).
- [82] J. INGLADA. « Automatic recognition of man-made objects in high resolution optical remote sensing images by SVM classification of geometric image features ». In : *ISPRS Journal of Photogrammetry and Remote Sensing* 62 (2007).

- [83] T. T. NGUYEN, H. GRABNER, B. GRUBER et H. BISCHOF. « On-line Boosting for Car Detection from Aerial Images ». In : *IEEE International Conference on Research, Innovation and Vision for the Future (RIVF'07)*. 2007, p. 87–95.
- [84] J. MICHEL, M. GRIZONNET, J. INGLADA, J. MALIK, A. BRICIER et O. LAHLOU. « Local feature based supervised object detection : Sampling, learning and detection strategies ». In : *IEEE Int. Geoscience and Remote Sensing Symposium (IGARSS)*. Vancouver, Canada, 2011.
- [85] A. BARALDI et F. PARMIGIANI. « Segmentation of SAR images by means of Gabor filters working at different spatial resolutions ». In : *IEEE Int. Geoscience and Remote Sensing Symposium (IGARSS)*. Lincoln, Nebraska, 1996.
- [86] A. LORETTE, X. DESCOMBES et J. ZERUBIA. « Texture Analysis through a Markovian Modelling and Fuzzy Classification : Application to Urban Area Extraction from Satellite Images ». In : *International Journal of Computer Vision* 36 (fév. 2000), p. 221–236.
- [87] M.-T. PHAM, G. MERCIER, O. REGNIERS et J. MICHEL. « Texture Retrieval from VHR Optical Remote Sensed Images Using the Local Extrema Descriptor with Application to Vineyard Parcel Detection ». In : *Remote Sensing* 8.5 (2016). URL : <https://www.mdpi.com/2072-4292/8/5/368>.
- [88] X. PERROTTON, M. STURZEL et M. ROUX. « Automatic Object Detection on Aerial Images Using Local Descriptors and Image Synthesis ». In : *Proceedings of International Conference on Vision Systems*. Santorini, Greece, 2008.
- [89] J. LEITLOFF, S. HINZ et U. STILLA. « Vehicle detection in very high resolution satellite images of city areas ». In : *IEEE Trans. On Geoscience and Remote Sensing (TGRS)* 48.7 (2011).
- [90] F. DELL'ACQUA, P. GAMBA, A. FERRARI, J. A. PALMASON, J. A. BENEDIKTSSON et K. ARNASON. « Exploiting spectral and spatial information in hyperspectral urban data with high resolution ». In : *IEEE Geoscience and Remote Sensing Letters* 1.4 (oct. 2004), p. 322–326.
- [91] Y. TARABALKA, M. FAUVEL, J. CHANUSSOT et J. A. BENEDIKTSSON. « SVM and MRF-based method for accurate classification of hyperspectral images ». In : *IEEE Geoscience and Remote Sensing Letters* 7.4 (2010), p. 736–740.
- [92] M. FAUVEL, Y. TARABALKA, J. A. BENEDIKTSSON, J. CHANUSSOT et J. C. TILTON. « Advances in Spectral-Spatial Classification of Hyperspectral Images ». In : *Proceedings of the IEEE* 101.3 (2013), p. 652–675.
- [93] G. CAMPS-VALLS, L. GOMEZ-CHOVA, J. MUNOZ-MUNARI, J. VILA-FRANCÉS et J. CALPE-MARAVILLA. « Composite Kernels for Hyperspectral Image Classification ». In : *IEEE Geosci. and Remote Sens. Letters* 3.1 (2006).
- [94] D. LOWE. « Distinctive Image Features from Scale-Invariant Keypoints ». In : *Int. Journal of Computer Vision* 60.2 (2004), p. 91–110.
- [95] B. SIRMACEK et C. UNSALAN. « Urban-area and building detection using SIFT keypoints and graph theory ». In : *IEEE Trans. on Geoscience and Remote Sensing (TGRS)* 47.4 (avr. 2009), p. 1156–1167.

- [96] C. VADUVA, I. GAVAT et M. DATCU. « Deep learning in very high resolution remote sensing image information mining communication concept ». In : *European Signal Processing Conference (EUSIPCO)*. Bucharest, Romania, 2012, p. 2506–2510.
- [97] D. LU et Q. WENG. « A Survey of Image Classification Methods and Techniques for Improving Classification Performance ». In : *Int. J. Remote Sens.* 28.5 (2007), p. 823–870.
- [98] G. MOUNTRAKIS, J. IM et C. OGOLE. « Support Vector Machines in remote sensing : A review ». In : *ISPRS Journal of Photogrammetry and Remote Sensing* 66 (2011), p. 247–259.
- [99] A. PLAZA, J. A. BENEDIKTSSON, J. W. BOARDMAN, J. BRAZILE, L. BRUZZONE, G. CAMPS-VALLS, J. CHANUSSOT, M. FAUVEL, P. GAMBA, A. GUALTIERI, M. MARCONCINI, J. C. TILTON et G. TRIANNI. « Recent advances in techniques for hyperspectral image processing ». In : *Remote Sensing of Environment* 113 (2009). Imaging Spectroscopy Special Issue, S110–S122.
- [100] G. CAMPS-VALLS, D. TUIA, L. BRUZZONE et J. A. BENEDIKTSSON. « Advances in Hyperspectral Image Classification : Earth monitoring with statistical learning methods ». In : *IEEE Signal Processing Magazine* 31.1 (2014), p. 45–54.
- [101] A. MOREIRA, P. PRATS-IRAOLA, M. YOUNIS, G. KRIEGER, I. HAJNSEK et K. P. PAPATHANASSIOU. « A tutorial on synthetic aperture radar ». In : *IEEE Geoscience and Remote Sensing Magazine* 1.1 (mar. 2013), p. 6–43.
- [102] E. E. KURUOGLU et J. ZERUBIA. « Modeling SAR images with a generalization of the Rayleigh distribution ». In : *IEEE Transactions on Image Processing* 13.4 (avr. 2004), p. 527–533.
- [103] G. MOSER, J. ZERUBIA et S. B. SERPICO. « SAR Amplitude Probability Density Function Estimation Based on a Generalized Gaussian Model ». In : *IEEE Transactions on Image Processing* (juin 2006). URL : <https://hal.inria.fr/inria-00561372>.
- [104] K. KAYABOL et J. ZERUBIA. « Unsupervised amplitude and texture classification of SAR images with multinomial latent model ». In : *IEEE Transactions on Image Processing* 22.2 (fév. 2013), p. 561–572. URL : <https://hal.inria.fr/hal-00745387>.
- [105] A. VOISIN, V. A. KRYLOV, G. MOSER, S. B. SERPICO et J. ZERUBIA. « Supervised Classification of Multisensor and Multiresolution Remote Sensing Images With a Hierarchical Copula-Based Approach ». In : *IEEE Transactions on Geoscience and Remote Sensing* 52.6 (juin 2014), p. 3346–3358.
- [106] M. DALLA MURA, S. PRASAD, F. PACIFICI, P. GAMBA, J. CHANUSSOT et J. A. BENEDIKTSSON. « Challenges and Opportunities of Multimodality and Data Fusion in Remote Sensing ». In : *Proceedings of the IEEE* 103.9 (sept. 2015), p. 1585–1601.
- [107] P. GHAMISI, B. RASTI, N. YOKOYA, Q. WANG, B. HOFLE, L. BRUZZONE, F. BOVOLO, M. CHI, K. ANDERS, R. GLOAGUEN, P. M. ATKINSON et J. A. BENEDIKTSSON. « Multisource and Multitemporal Data Fusion in Remote Sensing : A Comprehensive Review of the State of the Art ». In : *IEEE Geoscience and Remote Sensing Magazine* 7.1 (mar. 2019), p. 6–39.

- [108] H. RANDRIANARIVO. « Apprentissage statistique de classes sémantiques pour l'interprétation d'images aériennes ». 2016. URL : <https://tel.archives-ouvertes.fr/tel-01482119v1>.
- [109] P. FELZENSZWALB, R. GIRSHICK, D. MCALLESTER et D. RAMANAN. « Object detection with discriminatively trained part-based models ». In : *IEEE Trans. Pattern Analysis and Machine Intelligence (TPAMI)* 32.9 (2010), p. 1627–1645.
- [110] A. RENO, D. GILLIES et D. BOOTH. « Deformable models for object recognition in aerial images ». In : *Proceedings of the SPIE Conference Automatic Target Recognition VIII*. T. 3371. 1998, p. 323–333.
- [111] S. LAZEBNIK, C. SCHMID et J. PONCE. « Beyond bags of features : Spatial pyramid matching for recognizing natural scene categories ». In : *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. 2006.
- [112] P. F. FELZENSZWALB et D. P. HUTTENLOCHER. « Pictorial Structures for Object Recognition ». In : *Int. J. of Computer Vision* 61 (2003), p. 2005.
- [113] NZAM. *New Zealand Aerial Mapping Limited, Aerial Christchurch after Earthquake on Feb, 22, 2011*. <http://nzam.com/>. 2011.
- [114] N. AUDEBERT. « Classification de données massives de télédétection ». 2018. URL : <https://tel.archives-ouvertes.fr/tel-02073908>.
- [115] A. KRIZHEVSKY, I. SUTSKEVER et G. HINTON. « ImageNet Classification with Deep Convolutional Neural Networks ». In : *Advances in Neural Information Processing Systems 25 (NIPS)*. Lake Tahoe, Nevada, USA, 2012.
- [116] V. MNIH et G. HINTON. « Learning to Detect Roads in High-Resolution Aerial Images ». In : *The European Conference on Computer Vision (ECCV)*. Crete, Greece, 2010.
- [117] V. MNIH et G. HINTON. « Learning to Label Aerial Images from Noisy Data ». In : *International Conf. on Machine Learning (ICML)*. Edinburgh, Scotland, 2012.
- [118] A. ROMERO, C. GATTA et G. CAMPS-VALLS. « Unsupervised Deep Feature Extraction Of Hyperspectral Images ». In : *Proc. of WHISPERS*. Lausanne, Switzerland, 2014.
- [119] Y. BENGIO, D. LAMBLIN, P. POPOVICI et H. LAROCHELLE. « Greedy Layer-Wise Training of Deep Networks ». In : *Advances in Neural Information Processing Systems 20 (NIPS)*. Vancouver, B.C., Canada, 2006, p. 153–160.
- [120] Y. CHEN, Z. LIN, X. ZHAO, G. WANG et Y. GU. « Deep Learning-Based Classification of Hyperspectral Data ». In : *Selected Topics in Applied Earth Observations and Remote Sensing, IEEE Journal of* 7.6 (2014), p. 2094–2107.
- [121] M. E. MIDHUN, S. R. NAIR, V. T. PRABHAKAR et S. S. KUMAR. « Deep Model for Classification of Hyperspectral image using Restricted Boltzmann Machine ». In : *Proc. of ICONIAAC*. New-York, USA, 2014.
- [122] P. TOKARCZYK, J. MONTOYA et K. SCHINDLER. « An Evaluation of Feature Learning Methods for High Resolution Image Classification ». In : *ISPRS Annals of Photogrammetry, Remote Sensing and Spatial Information Sciences*. T. 1. 3. Melbourne, Australia, 2012, p. 389–394.

- [123] J. YOSINSKI, J. CLUNE, G. HINTON et H. LIPSON. « How transferable are features in deep neural networks? » In : *Advances in Neural Information Processing Systems 27 (NIPS)*. T. 2. Montreal, Canada, 2014, p. 3320–3328.
- [124] M. CASTELLUCCIO, G. POGGI, C. SANSONE et L. VERDOLIVA. « Land use classification in remote sensing images by convolutional neural networks ». In : *arXiv :1508.00092 [cs]* (août 2015). URL : <http://arxiv.org/abs/1508.00092>.
- [125] J. E. VARGAS, P. T. M. SAITO, A. X. FALCÃO, P. J. DOS REZENDE et J. A. DOS SANTOS. « Superpixel-Based Interactive Classification of Very High Resolution Images ». In : *Proc. 27th SIBGRAPI Conference on Graphics, Patterns and Images*. Août 2014, p. 173–179.
- [126] S. PAISITKRIANGKRAI, J. SHERRAH, P. JANNEY et A. VAN DEN HENGEL. « Effective semantic pixel labelling with convolutional networks and conditional random fields ». In : *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops*. 2015, p. 36–43.
- [127] J. SHERRAH. « Fully Convolutional Networks for Dense Semantic Labelling of High-Resolution Aerial Imagery ». In : *arXiv :1606.02585 [cs]* (juin 2016). arXiv : 1606.02585.
- [128] X. X. ZHU, D. TUIA, L. MOU, G. XIA, L. ZHANG, F. XU et F. FRAUNDORFER. « Deep Learning in Remote Sensing : A Comprehensive Review and List of Resources ». In : *IEEE Geoscience and Remote Sensing Magazine* 5.4 (déc. 2017), p. 8–36.
- [129] P. GHAMISI, N. YOKOYA, J. LI, W. LIAO, S. LIU, J. PLAZA, B. RASTI et A. PLAZA. « Advances in Hyperspectral Image and Signal Processing : A Comprehensive Overview of the State of the Art ». In : *IEEE Geoscience and Remote Sensing Magazine* 5.4 (déc. 2017), p. 37–78.
- [130] F. ROTTENSTEINER, G. SOHN, J. JUNG, M. GERKE, C. BAILLARD, S. BENITEZ et U. BREITKOPF. « The ISPRS benchmark on urban object classification and 3D building reconstruction ». In : *ISPRS Ann. Photogramm. Remote Sens. Spat. Inf. Sci* 1 (2012), p. 3.
- [131] R. ACHANTA, A. SHAJI, K. SMITH, A. LUCCHI, P. FUA et S. SÜSTRUNK. « Slic superpixels ». In : (2010).
- [132] D. MARMANIS, J. D. WEGNER, S. GALLIANI, K. SCHINDLER, M. DATCU et U. STILLA. « Semantic Segmentation of Aerial Images with an Ensemble of CNNs ». In : *ISPRS Annals of Photogrammetry, Remote Sensing and Spatial Information Sciences* 3 (2016), p. 473–480.
- [133] E. MAGGIORI, Y. TARABALKA, G. CHARPIAT et P. ALLIEZ. « Fully convolutional neural networks for remote sensing image classification ». In : *IEEE Int. Geoscience and Remote Sensing Symposium (IGARSS)*. Juil. 2016, p. 5071–5074.
- [134] M. VOLPI et D. TUIA. « Dense Semantic Labeling of Subdecimeter Resolution Images With Convolutional Neural Networks ». In : *IEEE Trans. on Geoscience and Remote Sensing (TGRS)* 55.2 (2017), p. 881–893.
- [135] K. HE, X. ZHANG, S. REN et J. SUN. « Deep residual learning for image recognition ». In : *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. 2016, p. 770–778.

- [136] Y. LIU, B. FAN, L. WANG, J. BAI, S. XIANG et C. PAN. « Semantic labeling in very high resolution images via a self-cascaded convolutional neural network ». In : *ISPRS Journal of Photogrammetry and Remote Sensing* 145 (2018), p. 78–95.
- [137] J. WANG, L. SHEN, W. QIAO, Y. DAI et Z. LI. « Deep Feature Fusion with Integration of Residual Connection and Attention Model for Classification of VHR Remote Sensing Images ». In : *Remote Sensing* 11 (juil. 2019), p. 1617.
- [138] D. MARCOS, D. TUIA, B. KELLENBERGER, L. ZHANG, M. BAI, R. LIAO et R. URTASUN. « Learning Deep Structured Active Contours End-to-End ». In : *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. Juin 2018.
- [139] E. MAGGIORI, Y. TARABALKA, G. CHARPIAT et P. ALLIEZ. « Can Semantic Labeling Methods Generalize to Any City? The Inria Aerial Image Labeling Benchmark ». en. In : *IEEE Int. Geoscience and Remote Sensing Symposium (IGARSS)*. Juil. 2017.
- [140] Y. CHEN, H. JIANG, C. LI, X. JIA et P. GHAMISI. « Deep Feature Extraction and Classification of Hyperspectral Images Based on Convolutional Neural Networks ». In : *IEEE Trans. on Geoscience and Remote Sensing (Trans)* 54.10 (oct. 2016), p. 6232–6251.
- [141] M. HUSSAIN, D. CHEN, A. CHENG, H. WEI et D. STANLEY. « Change Detection from Remotely Sensed Images : From Pixel-based to Object-based Approaches ». In : *ISPRS Journal of Photogrammetry and Remote Sensing* 80 (2013), p. 91–106.
- [142] A. SINGH. « Review Article Digital Change Detection Techniques Using Remotely-sensed Data ». In : *International Journal of Remote Sensing* 10.6 (1989), p. 989–1003.
- [143] P. BLANC. « Development of methods for detection of change ». Theses. École Nationale Supérieure des Mines de Paris, déc. 1999. URL : <https://pastel.archives-ouvertes.fr/tel-00477115>.
- [144] C. BENEDEZ, X. DESCOMBES et J. ZERUBIA. « Building Development Monitoring in Multitemporal Remotely Sensed Image Pairs with Stochastic Birth-Death Dynamics ». In : *IEEE Trans. on Pattern Analysis and Machine Intelligence (TPAMI)* 34.1 (2012), p. 33–50.
- [145] F. BOVOLO et L. BRUZZONE. « A wavelet-based change-detection technique for multitemporal SAR images ». In : *International Workshop on the Analysis of Multi-Temporal Remote Sensing Images*. IEEE. 2005, p. 85–89.
- [146] L. BRUZZONE et F. BOVOLO. « A novel framework for the design of change-detection systems for very-high-resolution remote sensing images ». In : *Proceedings of the IEEE* 101.3 (2013), p. 609–630.
- [147] G. LIU, J. DELON, Y. GOUSSEAU et F. TUPIN. « Unsupervised change detection between multi-sensor high resolution satellite images ». In : *European Signal Processing Conference (EUSIPCO)*. IEEE. 2016, p. 2435–2439.
- [148] M. VAKALOPOULOU, K. KARANTZALOS, N. KOMODAKIS et N. PARAGIOS. « Simultaneous registration and change detection in multitemporal, very high resolution remote sensing data ». In : *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR) Workshops / IEEE/ISPRS EarthVision : Large Scale Computer Vision for Remote Sensing Imagery*. 2015, p. 61–69.

- [149] M. VOLPI, D. TUIA, F. BOVOLO, M. KANEVSKI et L. BRUZZONE. « Supervised change detection in VHR images using contextual information and support vector machines ». In : *International Journal of Applied Earth Observation and Geoinformation* 20 (2013), p. 77–85.
- [150] A. M. EL AMIN, Q. LIU et Y. WANG. « Convolutional neural network features based change detection in satellite images ». In : *First International Workshop on Pattern Recognition*. International Society for Optics et Photonics. 2016, 100110W.
- [151] A. M. EL AMIN, Q. LIU et Y. WANG. « Zoom out CNNs features for optical remote sensing change detection ». In : *Int. Conference on Image, Vision and Computing*. 2017, p. 812–817.
- [152] R. CARUANA. « Multitask Learning ». In : *Machine Learning* (1997).
- [153] L. ALPARONE, L. WALD, J. CHANUSSOT, C. THOMAS, P. GAMBA et L. M. BRUCE. « Comparison of pansharpening algorithms : Outcome of the 2006 GRS-S Data Fusion Contest ». In : *IEEE Trans. on Geoscience and Remote Sensing (TGRS)* 45.10 (2007), p. 3012–3021.
- [154] V. MNIH. « Machine Learning for Aerial Image Labeling ». Thèse de doct. University of Toronto, 2013.
- [155] F. ROTTENSTEINER, G. SOHN, M. GERKE et J. D. WEGNER. *Journal of Photogrammetry and Remote Sensing : Special issue on Urban object detection and 3D building reconstruction*. T. 93. Elsevier, juil. 2014.
- [156] N. HAALA, M. CRAMER et K. H. JACOBSEN. « The german camera evaluation project - results from the geometry group ». In : *Canadian Geomatics Conference And Symposium Of Commission I - Geometry*. 2010.
- [157] D. LAM, R. KUZMA, K. MCGEE, S. DOOLEY, M. LAIELLI, M. KLARIC, Y. BULATOV et B. MCCORD. *xView : Objects in Context in Overhead Imagery*. Fév. 2018.
- [158] I. DEMIR, K. KOPERSKI, D. LINDENBAUM, G. PANG, J. HUANG, S. BASU, F. HUGHES, D. TUIA et R. RASKAR. « DeepGlobe 2018 : A Challenge to Parse the Earth Through Satellite Images ». In : *The IEEE Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*. Juin 2018.
- [159] R. GUPTA, B. GOODMAN, N. PATEL, R. HOSFELT, S. SAJEEV, E. HEIM, J. DOSHI, K. LUCAS, H. CHOSET et M. GASTON. « Creating xBD : A Dataset for Assessing Building Damage from Satellite Imagery ». In : *The IEEE Conference on Computer Vision and Pattern Recognition / Workshop on Computer Vision for Global Challenges (CVPRW)*. Juin 2019.
- [160] F. PACIFICI, F. DEL FRATE, W. J. EMERY, P. GAMBA et J. CHANUSSOT. « Urban Mapping Using Coarse SAR and Optical Data : Outcome of the 2007 GRSS Data Fusion Contest ». In : *IEEE Geoscience and Remote Sensing Letters* 5.3 (juil. 2008), p. 331–335.
- [161] G. LICCIARDI, F. PACIFICI, D. TUIA, S. PRASAD, T. WEST, F. GIACCO, J. INGLADA, E. CHRISTOPHE, J. CHANUSSOT et P. GAMBA. « Decision fusion for the classification of hyperspectral data : Outcome of the 2008 GRS-S Data Fusion Contest ». In : *IEEE Trans. on Geoscience and Remote Sensing (TGRS)* 47.11 (2009), p. 3857–3865.

- [162] C. DEBES, A. MERENTITIS, R. HEREMANS, J. HAHN, N. FRANGIADAKIS, T. van KASTEREN, W. LIAO, R. BELLENS, A. PIZURICA, S. GAUTAMA, W. PHILIPS, S. PRASAD, Q. DU et F. PACIFICI. « Hyperspectral and LiDAR Data Fusion : Outcome of the 2013 GRSS Data Fusion Contest ». In : *IEEE J. Sel. Topics Appl. Earth Observ. Remote Sens.* 7.6 (2014), p. 2405–2418.
- [163] W. LIAO, X. HUANG, F. V. COILLIE, S. GAUTAMA, A. PIZURICA, W. PHILIPS, H. LIU, T. ZHU, M. SHIMONI, G. MOSER et D. TUIA. « Processing of Multiresolution Thermal Hyperspectral and Digital Color Data : Outcome of the 2014 IEEE GRSS Data Fusion Contest ». In : *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing* 8.6 (juin 2015), p. 2984–2996.
- [164] F. PACIFICI et Q. DU. « Foreword to the special issue on optical multiangular data exploitation and outcome of the 2011 GRSS Data Fusion Contest ». In : *IEEE J. Sel. Topics Appl. Earth Observ. Remote Sens.* 5.1 (2012), p. 3–7.
- [165] N. LONGBOTHAM, F. PACIFICI, T. GLENN, A. ZARE, M. VOLPI, D. TUIA, E. CHRISTOPHE, J. MICHEL, J. INGLADA, J. CHANUSSOT et Q. DU. « Multi-modal change detection, application to the detection of flooded areas : outcome of the 2009-2010 Data Fusion Contest ». In : *IEEE J. Sel. Topics Appl. Earth Observ. Remote Sens.* 5.1 (2012), p. 331–342.
- [166] C. BERGER, M. VOLTERSEN, R. ECKARDT, J. EBERLE, T. HEYER, N. SALEPCI, S. HESE, C. SCHMULLIUS, J. TAO, S. AUER, R. BAMLER, K. EWALD, M. GARTLEY, J. JACOBSON, A. BUSWELL, Q. DU et F. PACIFICI. « Multi-modal and multi-temporal data fusion : Outcome of the 2012 GRSS Data Fusion Contest ». In : *IEEE J. Sel. Topics Appl. Earth Observ. Remote Sens.* 6.3 (2013), p. 1324–1340.
- [167] S. WANG, M. BAI, G. MATTYUS, H. CHU, W. LUO, B. YANG, J. LIANG, J. CHEVERIE, S. FIDLER et R. URTASUN. « TorontoCity : Seeing the World with a Million Eyes ». In : *The IEEE International Conference on Computer Vision (ICCV)*. Oct. 2017, p. 3028–3036.
- [168] A. V. ETTEN, D. LINDENBAUM et T. M. BACASTOW. *SpaceNet : A Remote Sensing Dataset and Challenge Series*. 2018. arXiv : [1807.01232](https://arxiv.org/abs/1807.01232) [cs.CV].
- [169] A.-V. VO, L. TRUONG-HONG, D. LAEFER, D. TIEDE, S. D’OLEIRE-OLTMANN, A. BARALDI, M. SHIMONI, G. MOSER et D. TUIA. « Processing of Extremely high resolution LiDAR and RGB data : Outcome of the 2015 IEEE GRSS Data Fusion Contest. Part B : 3D contest ». In : *IEEE J. Sel. Topics Appl. Earth Observ. Remote Sens.* 9.12 (2016), p. 5560–5575.
- [170] I. D. STEWART et T. R. OKE. « Local climate zones for urban temperature studies ». In : *Bull. Amer. Meteor. Soc.* 93 (2012), p. 1879–1900.
- [171] M. BOSCH, G. FOSTER, G. CHRISTIE, S. WANG, G. HAGER et M. BROWN. « Semantic Stereo for Incidental Satellite Images ». In : *Proc. Winter Conf. on Applications of Computer Vision*. Waikoloa Village, Hawaii, USA, 2019.
- [172] Y. LI, R. BU, M. SUN, W. WU, X. DI et B. CHEN. « PointCNN : Convolution On X-Transformed Points ». In : *Advances in Neural Information Processing Systems 31 (NIPS)*. 2018, p. 820–830.

- [173] M. JIANG, Y. WU, T. ZHAO, Z. ZHAO et C. LU. *PointSIFT : A SIFT-like Network Module for 3D Point Cloud Semantic Segmentation*. 2018.
- [174] N. DEY, L. BLANC-FÉRAUD, C. ZIMMER, Z. KAM, P. ROUX, J. OLIVO-MARIN et J. ZERUBIA. « Richardson-Lucy Algorithm with Total Variation Regularization for 3D Confocal Microscope Deconvolution ». In : *Microscopy Research Technique* 69 (2006), p. 260–266.
- [175] O. FAUGERAS. *Three-dimensional Computer Vision : A Geometric Viewpoint*. Cambridge, MA, USA : MIT Press, 1993.
- [176] R. HARTLEY et A. ZISSERMAN. *Multiple View Geometry in Computer Vision*. 2^e éd. New York, NY, USA : Cambridge University Press, 2003.
- [177] A. SAXENA, S. H. CHUNG et A. Y. NG. « Learning Depth from Single Monocular Images ». In : *Advances in Neural Information Processing Systems 19 (NIPS)*. 2006.
- [178] « A dynamic bayesian network model for autonomous 3D reconstruction from a single indoor image ». In : *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. 2006.
- [179] A. SAXENA, M. SUN et A. Y. NG. « Make3D : Learning 3D scene structure from a single still image ». In : *IEEE Trans. on Pattern Analysis and Machine Intelligence (TPAMI)* 31.5 (2009), p. 824–840.
- [180] D. EIGEN, C. PUHRSCHE et R. FERGUS. « Depth map prediction from a single image using a multi-scale deep network ». In : *Advances in Neural Information Processing Systems 27 (NIPS)* (2014).
- [181] A. P. PENTLAND. « A new sense for depth of field ». In : *IEEE Trans. Pattern Analysis and Machine Intelligence (TPAMI)* 9 (1987).
- [182] P. TROUVÉ, F. CHAMPAGNAT, G. LE BESNERAIS et J. IDIER. « Single image local blur identification ». In : *IEEE ICIP* (2011).
- [183] M. MARTINELLO et P. FAVARO. « Single Image Blind Deconvolution with Higher-Order Texture Statistics ». In : *Video Processing and Comp. Video* (2011).
- [184] P. TROUVÉ-PELOUX. « Conception conjointe optique / traitement pour un imageur compact à capacité 3D ». Thèse de doct. École centrale de Nantes, 2012.
- [185] P. TROUVÉ, F. CHAMPAGNAT, G. LE BESNERAIS, J. SABATER, T. AVIGNON et J. IDIER. « Passive depth estimation using chromatic aberration and a depth from defocus approach ». In : *Applied Optics* 52.29 (2013).
- [186] H. JUNG, Y. KIM¹, D. MIN, C. OH et K. SOHN. « Depth prediction from a single image with conditional adversarial networks ». In : *ICIP*. 2017.
- [187] P. ISOLA, J.-Y. ZHU, T. ZHOU et A. A. EFROS. « Image-To-Image Translation With Conditional Adversarial Networks ». In : *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. Juil. 2017.
- [188] A. KENDALL et Y. GAL. « What Uncertainties Do We Need in Bayesian Deep Learning for Computer Vision ? » In : *Advances in Neural Information Processing Systems 31 (NIPS)*. 2017, p. 5580–5590.

- [189] I. LAINA, C. RUPPRECHT, V. BELAGIANNIS, F. TOMBARI et N. NAVAB. « Deeper depth prediction with fully convolutional residual networks ». In : *3D Vision (3DV), 2016 Fourth International Conference on*. IEEE. 2016, p. 239–248.
- [190] X. MAO, Q. LI, H. XIE, R. Y. LAU, Z. WANG et S. P. SMOLLEY. « Least squares generative adversarial networks ». In : *The IEEE International Conference on Computer Vision (ICCV)*. 2017, p. 2813–2821.
- [191] R. TYLECEK, T. SATTLER, H.-A. LE, T. BROX, M. POLLEFEYS, R. B. FISHER et T. GEVERS. « The Second Workshop on 3D Reconstruction Meets Semantics : Challenge Results Discussion ». In : *The European Conference on Computer Vision (ECCV) / Workshop on 3D Reconstruction meets Semantics*. Sous la dir. de L. LEAL-TAIXÉ et S. ROTH. Cham : Springer International Publishing, 2019, p. 631–644.
- [192] *3D Reconstruction meets Semantics 2018 – Challenge*. <http://trimbot2020.webhosting.rug.nl/events/3drms/challenge/>. Accessed : 2019-09-5.
- [193] T. HACKEL, N. SAVINOV, L. LADICKY, J.-D. WEGNER, K. SCHINDLER et M. POLLEFEYS. « Large-Scale Point Cloud Classification Benchmark ». In : *The IEEE Conference on Computer Vision and Pattern Recognition / Workshop on Large-Scale 3D Data (CVPRW)*. 2016. URL : <http://www.semantic3d.net/>.
- [194] M. ROUHANI, F. LAFARGE et P. ALLIEZ. « Semantic segmentation of 3D textured meshes for urban scene analysis ». In : *ISPRS Journal of Photogrammetry and Remote Sensing* 123 (2017).
- [195] A. BOULCH, S. HOULLIER, R. MARLET et O. TOURNAIRE. « Semantizing Complex 3D Scenes using Constrained Attribute Grammars ». In : *Computer Graphics Forum* 32.5 (août 2013), p. 33–42.
- [196] D. MATURANA et S. SCHERER. « VoxNet : A 3D Convolutional Neural Network for real-time object recognition ». In : *IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*. Hamburg, Germany, 2015, p. 922–928.
- [197] G. RIEGLER, A. OSMAN ULUSOY et A. GEIGER. « Octnet : Learning deep 3D representations at high resolutions ». In : *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. Honolulu, Hawaii, 2017.
- [198] A. HORNUNG, K. M. WURM, M. BENNEWITZ, C. STACHNISS et W. BURGARD. « OctoMap : An Efficient Probabilistic 3D Mapping Framework Based on Octrees ». In : *Autonomous Robots* (2013). Software available at <http://octomap.github.com>. URL : <http://octomap.github.com>.
- [199] C. R. QI, H. SU, K. MO et L. J. GUIBAS. « PointNet : Deep learning on point sets for 3D classification and segmentation ». In : *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. Honolulu, Hawaii, USA, juil. 2017, p. 77–85.
- [200] C. R. QI, L. YI, H. SU et L. J. GUIBAS. « Pointnet++ : Deep hierarchical feature learning on point sets in a metric space ». In : *Advances in Neural Information Processing Systems 30 (NIPS)*. 2017, p. 5099–5108.
- [201] H. SU, S. MAJI, E. KALOGERAKIS et E. LEARNED-MILLER. « Multi-view convolutional neural networks for 3D shape recognition ». In : *The IEEE International Conference on Computer Vision (ICCV)*. 2015, p. 945–953.

- [202] K. SFIKAS, T. THEOHARIS et I. PRATIKAKIS. « Exploiting the PANORAMA Representation for Convolutional Neural Network Classification and Retrieval ». In : *Eurographics Workshop on 3D Object Retrieval*. Lyon, France, 2017.
- [203] O. CHAPELLE, B. SCHÖLKOPF et A. ZIEN. *Semi-Supervised Learning*. The MIT Press, 2006.
- [204] T. DURAND, T. MORDAN, N. THOME et M. CORD. « WILDCAT : Weakly Supervised Learning of Deep ConvNets for Image Classification, Pointwise Localization and Segmentation ». In : *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. Juil. 2017.
- [205] J. DAI, K. HE et J. SUN. « Boxsup : Exploiting bounding boxes to supervise convolutional networks for semantic segmentation ». In : *Proceedings of the IEEE International Conference on Computer Vision*. 2015, p. 1635–1643.
- [206] Z. LU, Z. FU, T. XIANG, P. HAN, L. WANG et X. GAO. « Learning from weak and noisy labels for semantic segmentation ». In : *IEEE transactions on pattern analysis and machine intelligence* 39.3 (2017), p. 486–500.
- [207] A. KHOREVA, R. BENENSON, J. HOSANG, M. HEIN et B. SCHIELE. « Simple does it : Weakly supervised instance and semantic segmentation ». In : *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. 2017, p. 876–885.
- [208] P. PERONA et J. MALIK. « Scale-space and edge detection using anisotropic diffusion ». In : *IEEE Trans. on Pattern Analysis and Machine Intelligence (TPAMI)* 12.7 (1990), p. 629–639.
- [209] N. XU, B. PRICE, S. COHEN, J. YANG et T. S. HUANG. « Deep interactive object selection ». In : *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. 2016, p. 373–381.
- [210] K. RAKELLY, E. SHELHAMER, T. DARRELL, A. A. EFROS et S. LEVINE. « Few-shot segmentation propagation with guided networks ». In : *arXiv preprint arXiv :1806.07373* (2018).
- [211] D. TUIA, M. VOLPI, L. COPA, M. KANEVSKI et J. MUNOZ-MARI. « A Survey of Active Learning Algorithms for Supervised Remote Sensing Image Classification ». In : *IEEE Journal of Selected Topics in Signal Processing* 5.3 (juin 2011), p. 606–617.
- [212] E. PASOLLI, F. MELGANI, D. TUIA, F. PACIFICI et W. J. EMERY. « SVM Active Learning Approach for Image Classification Using Spatial Information ». In : *IEEE Trans. on Geoscience and Remote Sensing (TGRS)* 52.4 (avr. 2014), p. 2217–2233.
- [213] L. DING, D. DIFRANZO, A. GRAVES, J. R. MICHAELIS, X. LI, D. L. MCGUINNESS et J. HENDLER. « Data-gov wiki : Towards linking government data ». In : *2010 AAAI Spring Symposium Series*. 2010.
- [214] G. BARTHA et S. KOCSIS. « Standardization of geographic data : The European INSPIRE directive ». In : *European Journal of Geography* 2.2 (2011), p. 79–89.
- [215] É. COLIN-KOENIGUER, A. BOULCH, P. TROUVÉ-PELOUX et F. JANEZ. « Colored visualization of multitemporal SAR data for change detection : issues and methods ». In : *Proc. Eur. Conf. on Synthetic Aperture Radar*. Aachen, Germany, 2018.

- [216] N. VO, N. JACOBS et J. HAYS. « Revisiting IM2GPS in the Deep Learning Era ». In : *The IEEE International Conference on Computer Vision (ICCV)*. Venice, Italy, 2017.
- [217] E. SHEEHAN, B. UZKENT, C. MENG, Z. TANG, M. BURKE, D. LOBELL et S. ERMON. « Learning to interpret satellite images using wikipedia ». In : *Proc. 28th International Joint Conference on Artificial Intelligence (IJCAI)*. Macao, China, 2019.
- [218] J. MAIRAL, F. BACH et J. PONCE. « Sparse Modeling for Image and Vision Processing ». In : *Found. Trends. Comput. Graph. Vis.* 8.2-3 (déc. 2014), p. 85–283. URL : <http://dx.doi.org/10.1561/06000000058>.
- [219] X. ZHANG, J. ZHAO et Y. LECUN. « Character-level convolutional networks for text classification ». In : *Advances in Neural Information Processing Systems 28 (NIPS)*. 2015, p. 649–657.
- [220] S. LAI, L. XU, K. LIU et J. ZHAO. « Recurrent convolutional neural networks for text classification ». In : *AAAI conference on artificial intelligence*. 2015, p. 2267–2273.
- [221] A. R. ZAMIR, A. SAX, W. B. SHEN, L. J. GUIBAS, J. MALIK et S. SAVARESE. « Taskonomy : Disentangling Task Transfer Learning ». In : *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. Salt Lake City, Utah, USA, 2018.
- [222] A. BOULCH. « Generalizing Discrete Convolutions for Unstructured Point Clouds ». In : *Eurographics Workshop on 3D Object Retrieval*. 2019.
- [223] D. MARR. *Vision : A Computational Investigation into the Human Representation and Processing of Visual Information*. New-York : W. H. Freeman et Company, 1982.
- [224] R. P. HORAUD et O. MONGA. *Vision par ordinateur : outils fondamentaux*. Traité des nouvelles technologies. Série informatique. Hermès, 1995.
- [225] D. HOIEM et S. SAVARESE. *Representations and Techniques for 3D Object Recognition and Scene Interpretation*. Morgan & Claypool, 2011.
- [226] M. BLOESCH, T. LAIDLAW, R. CLARK, S. LEUTENEGGER et A. J. DAVISON. « Learning Meshes for Dense Visual SLAM ». In : *The IEEE International Conference on Computer Vision (ICCV)*. Seoul, Korea, 2019.
- [227] V. DESCHAINTRE, M. AITTALA, F. DURAND, G. DRETTAKIS et A. BOUSSEAU. « Single-image SVBRDF Capture with a Rendering-aware Deep Network ». In : *ACM Trans. on Graphics (TOG)* 37.4 (juil. 2018), 128 :1–128 :15.
- [228] A. GEIGER, P. LENZ, C. STILLER et R. URTASUN. « Vision meets Robotics : The KITTI Dataset ». In : *International Journal of Robotics Research (IJRR)* (2013).
- [229] M. PINHEIRO DE CARVALHO. « Deep Depth from Defocus : Neural Networks for Monocular Depth Estimation ». Theses. Université Paris Saclay, à paraître, nov. 2019.
- [230] H. FAN, H. SU et L. J. GUIBAS. « A point set generation network for 3d object reconstruction from a single image ». In : *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. 2017, p. 605–613.
- [231] G. PEYRÉ et M. CUTURI. « Computational optimal transport ». In : *Foundations and Trends® in Machine Learning* 11.5-6 (2019), p. 355–607.

- [232] C. VILLANI. *Optimal transport : old and new*. T. 338. Springer Science & Business Media, 2008.
- [233] P. MANDIKAL et R. V. BABU. « Dense 3D Point Cloud Reconstruction Using a Deep Pyramid Network ». In : *Winter Conference on Applications of Computer Vision (WACV)*. 2019, p. 1052–1060.
- [234] L. LANDRIEU et M. SIMONOVSKY. « Large-Scale Point Cloud Semantic Segmentation with Superpoint Graphs ». en. In : *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. Salt Lake City, UT : IEEE, juin 2018, p. 4558–4567. URL : <https://ieeexplore.ieee.org/document/8578577/> (visité le 27/03/2019).
- [235] M. TATARCHENKO, J. PARK, V. KOLTUN et Q.-Y. ZHOU. « Tangent Convolutions for Dense Prediction in 3D ». en. In : *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. Salt Lake City, UT, USA : IEEE, juin 2018, p. 3887–3896. URL : <https://ieeexplore.ieee.org/document/8578507/> (visité le 29/03/2019).
- [236] H. THOMAS, C. R. QI, J.-E. DESCHAUD, B. MARCOTEGUI, F. GOULETTE et L. J. GUIBAS. « KPConv : Flexible and Deformable Convolution for Point Clouds ». In : *The IEEE International Conference on Computer Vision (ICCV)*. Oct. 2019. URL : openaccess.thecvf.com/content_ICCV_2019/papers/Thomas_KPConv_Flexible_and_Deformable_Convolution_for_Point_Clouds_ICCV_2019_paper.pdf.
- [237] Y. WANG, W. CHAO, D. GARG, B. HARIHARAN, M. CAMPBELL et K. Q. WEINBERGER. « Pseudo-LiDAR from Visual Depth Estimation : Bridging the Gap in 3D Object Detection for Autonomous Driving ». In : *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. Juin 2019. URL : <http://arxiv.org/abs/1812.07179>.
- [238] Y. YOU, Y. WANG, W. CHAO, D. GARG, G. PLEISS, B. HARIHARAN, M. CAMPBELL et K. Q. WEINBERGER. *Pseudo-LiDAR++ : Accurate Depth for 3D Object Detection in Autonomous Driving*. 2019. URL : <http://arxiv.org/abs/1906.06310>.
- [239] T. RODDICK, A. KENDALL et R. CIPOLLA. « Orthographic Feature Transform for Monocular 3D Object Detection ». In : *Proc. of the British Machine Vision Conference (BMVC)*. 2019.
- [240] *Waymo Open Dataset : An autonomous driving dataset*. 2019.
- [241] P. TROUVÉ-PELOUX, J. SABATER, A. BERNARD-BRUNEL, F. CHAMPAGNAT, G. LE BESNERAIS et T. AVIGNON. « Turning a conventional camera into a 3D camera with an add-on ». In : *Applied Optics* 57.10 (2018), p. 2553–2563.
- [242] D. HOLZ, A.-E. ICHIM, F. TOMBARI, R. B. RUSU et S. BEHNKE. « Registration with the Point Cloud Library A Modular Framework for Aligning in 3-D ». In : *IEEE Robotics & Automation Magazine* 22.4 (2015), p. 110–124.
- [243] P.-A. LANGLOIS, A. BOULCH et R. MARLET. « Surface Reconstruction from 3D Line Segments ». In : *Int. Conf. on 3D Vision (3DV)*. Quebec City, Canada, sept. 2019.
- [244] F. LAFARGE et C. MALLET. « Creating Large-Scale City Models from 3D-Point Clouds : A Robust Approach with Hybrid Representation ». In : *International Journal of Computer Vision (IJCV)* 99.1 (août 2012), p. 69–85.

- [245] J. BEHLEY, M. GARBADE, A. MILIOTO, J. QUENZEL, S. BEHNKE, C. STACHNISS et J. GALL. « SemanticKITTI : A Dataset for Semantic Scene Understanding of LiDAR Sequences ». In : *the IEEE/CVF International Conference on Computer Vision (ICCV)*. 2019.
- [246] J. ALOIMONOS, I. WEISS et A. BANDYOPADHYAY. « Active vision ». In : *International Journal of Computer Vision (IJCV)* 1.4 (jan. 1988), p. 333–356. URL : <https://doi.org/10.1007/BF00133571>.

Titre : Modèles d'Apprentissage Machine pour la Compréhension de Scènes

Mots clés : Apprentissage machine; Réseaux de neurones; Compréhension de scènes; Vision par ordinateur; Observation de la Terre; Vision 3D; Analyse d'images

Résumé :

La compréhension de scènes vise à répondre à la question : comment construire un modèle d'une région du monde réel afin d'y agir et d'y interagir ? Il s'agit donc d'extraire la sémantique et la géométrie des données disponibles : images, nuages de points 3D, etc. Dans ce but, plusieurs approches d'apprentissage machine sont présentées : elles diffèrent par la proportion d'a priori de conception et d'apprentissage introduits tout au long des algorithmes. Trois aspects du problème sont envisagés. Les premiers travaux visent à la compréhension du contenu sémantique des images, c'est à dire la classification, la détection d'objets et la segmentation sémantique. Puis, plusieurs approches d'apprentissage sont proposés pour l'observation de la Terre et la télédétection, notamment pour l'apprentissage interactif, la classification sémantique multimodale et la détection de changements sémantiques. Enfin, l'accent est mis sur la

vision 3D, avec l'estimation de la profondeur à partir d'une seule image et la classification de nuages de points 3D par des réseaux de neurones. Ces approches variées reposent sur des mécanismes sous-jacents communs qui prennent une importance croissante. Elles réalisent une analyse multimodale pour bénéficier des données complémentaires disponibles, issues de capteurs différents mais aussi de sources et méta-données hétérogènes. Symétriquement, l'optimisation jointe d'objectifs multiples permet de régulariser l'apprentissage de modèles performants. Sur-tout, elles ont de plus en plus recours à une multiplicité des points de vue sur la scène pour relier, tant en apprentissage qu'en inférence, des invariances spatiales qui servent une analyse locale et une reconstruction sémantique globale. Cela est rendu possible par une intégration croissante de l'apparence et de la structure 3D, et conduit à une meilleure compréhension sémantique de la scène.

Title : Machine Learning Models for Scene Understanding

Keywords : Machine learning; Neural networks; Scene understanding; Computer vision; Earth observation; 3D Vision; Image analysis

Abstract : Scene understanding aims to answer the question : how to build a model of a real-world region in order to act and interact with it? It is therefore necessary to extract the semantics and geometry of the available data : images, 3D point-clouds, etc. For this purpose, several machine learning approaches are presented: they differ in the proportion of prior assumptions and learning introduced throughout the algorithms. Three aspects of the problem are envisaged. The first works aim at understanding the semantic content of images, through classification, object detection and semantic segmentation. Then, several learning approaches are proposed for Earth observation and remote sensing, notably for interactive learning, multimodal semantic classification and semantic change detection. Finally, the focus is on 3D vision, with depth estimation from a single image and

classification of 3D point-clouds by neural networks. These various approaches are based on common underlying mechanisms that are becoming increasingly important. They perform a multimodal analysis in order to benefit from the available, complementary data, obtained from different sensors but also from heterogeneous sources and meta-data. Symmetrically, joint optimization of multiple objectives helps to regularize the learning of efficient models. Moreover, they increasingly rely on a multiplicity of points of view on the scene to relate, in both learning and inference, spatial invariances that serve a local analysis and a global semantic reconstruction. This is made possible by a growing integration of the appearance and 3D structure, and leads to a better semantic understanding of the scene.

