Apprentissage profond pour la segmentation et la détection automatique en imagerie multi-modale : application à l'oncologie hépatique

Vincent Couteaux

INSTITUT POLYTECHNIQUE DE PARIS

NNT : 2021IPPAT009

Thèse de doctorat

Apprentissage profond pour la segmentation et la détection automatique en imagerie multi-modale. Application à l'oncologie hépatique

Thèse de doctorat de l'Institut Polytechnique de Paris
préparée à Télécom Paris

École doctorale n° 626 Institut Polytechnique de Paris (ED IP Paris)
Spécialité de doctorat : Informatique, Données et Intelligence Artificielle

Thèse présentée et soutenue à Palaiseau, le 19 mai 2021, par

VINCENT COUTEAUX

Composition du Jury :

Chloé Clavel Professeure, Télécom Paris (LTCI)	Présidente
Jean-Philippe Thiran Professeur, EPFL (LTS5)	Rapporteur
Caroline Petitjean Maître de conférences, Université de Rouen Normandie (LITIS)	Rapporteuse
Pierre-Jean Valette Professeur, Hospices Civils de Lyon	Examinateur
Isabelle Bloch Professeure, Télécom Paris (LTCI)	Directrice de thèse
Olivier Nempont Philips Research Paris	Examinateur
Guillaume Pizaine Philips Research Paris	Invité

Apprentissage profond pour la segmentation et la détection automatique en imagerie multi-modale

Application à l'oncologie hépatique

Auteur
Vincent Couteaux

Directrice de thèse
Isabelle Bloch

Co-encadrants
Olivier Nempont
Guillaume Pizaine

Soutenue le 19 mai 2021 à Palaiseau

Table des matières

Table des figures

Chapitre 1

Introduction

La radiologie, qui est la spécialité médicale consistant à utiliser l'imagerie pour diagnostiquer et traiter des maladies, est une des disciplines de la médecine dont les progrès dépendent le plus directement de ceux de la technologie. Après la découverte des rayons X par Röntgen à la fin du XIX^{e} siècle, des examens de plus en plus informatifs et de moins en moins invasifs ont été rendus possibles notamment grâce au développement de l'échographie à partir des années 1950, de la tomodensitométrie (qu'on appellera « CT » pour *Computed Tomography* dans la suite du document), puis de l'imagerie par résonance magnétique (IRM) dans les années 1970. Les progrès récents de l'intelligence artificielle permis par l'essor de l'apprentissage profond suscitent aujourd'hui de grands espoirs en radiologie.

Un enjeu important de la recherche médicale actuelle est celui de la médecine de précision : comment caractériser le plus précisément possible la maladie d'un patient, pour lui proposer le traitement le plus adapté ? Le phénotypage à partir d'images médicales est une des pistes privilégiées pour répondre à cette problématique : peu invasives (par rapport à une biopsie notamment), déjà acquises en routine, les images contiennent beaucoup d'information sur l'état d'un patient et à plus forte raison si plusieurs modalités ou protocoles (différents temps d'injection, différentes séquences IRM...) sont combinés. L'étude du génome (génomique) et des protéines (protéomique) sont des exemples de disciplines pouvant fournir d'autres types de données utiles à la caractérisation précise de maladies, à la place ou en complément des images médicales.

Je vais brièvement présenter dans ce chapitre le domaine d'application de cette thèse, qui est l'oncologie hépatique en radiologie, en mettant l'accent sur l'importance du développement d'outils de traitement automatique d'images médicales, et plus particulièrement de segmentation et de détection automatique de structures dans des images multi-modales.

1.1 Les lésions hépatiques en radiologie : caractérisation en IRM de lésions courantes

Pour illustrer l'utilisation des images médicales dans la pratique clinique et se familiariser avec les images que je présenterai dans le reste de cette thèse, je propose de prendre l'exemple du diagnostic de quatre types de lésions du foie parmi les plus courants en utilisant les images IRM. Cette section est basée sur le livre *Liver MRI* de Hussain et Sorrell (2015), ainsi que l'encyclopédie en ligne *radiopaedia*[1].

Les images IRM peuvent être acquises selon différentes *séquences* qui mettent en valeur des propriétés différentes des tissus. Les quatre types de lésions qui nous intéressent dans cette section, que sont les kystes, les hémangiomes hépatiques, les carcinomes hépato-cellulaires, et les métastases de cancers d'autres organes, présentent des aspects distincts dans les séquences IRM dites pondérées en T1 et celles dites pondérées en T2.

L'utilisation d'un produit de contraste à base de Gadolinium est également utile pour différencier ces lésions. On acquiert en général quatre images pour chaque injection de produit de contraste : la première avant l'injection, dite *pré-contraste*, la seconde lorsque le produit de contraste est dans les artères du foie, qu'on appelle *temps artériel* (une dizaine de secondes après l'injection), la troisième lorsqu'il est présent dans les veines portales du foie et qu'on appelle *temps veineux* ou plus couramment *temps portal* (une minute après l'injection), et la dernière lorsque le produit s'est diffusé plus uniformément dans le foie et qu'on appelle *temps tardif* (quelques minutes après l'injection).

D'après Hussain et Sorrell (2015), l'image pondérée en T2, l'image pondérée en T1 sans contraste, ainsi que celle acquise au temps artériel et celle acquise au temps tardif suffisent à caractériser un certain nombre de lésions du foie, dont les quatre de notre exemple.

Dans la suite de la thèse on parlera - abusivement - de modalités différentes pour désigner des images de séquences IRM différentes ou de temps d'injection différents.

Les kystes hépatiques simples

Ces kystes sont des lésions bénignes, la plupart du temps asymptomatiques. Leur forme est ronde ou ovale, et leur taille varie de quelques millimètres à plusieurs centimètres

1. https://radiopaedia.org/articles/liver-lesions

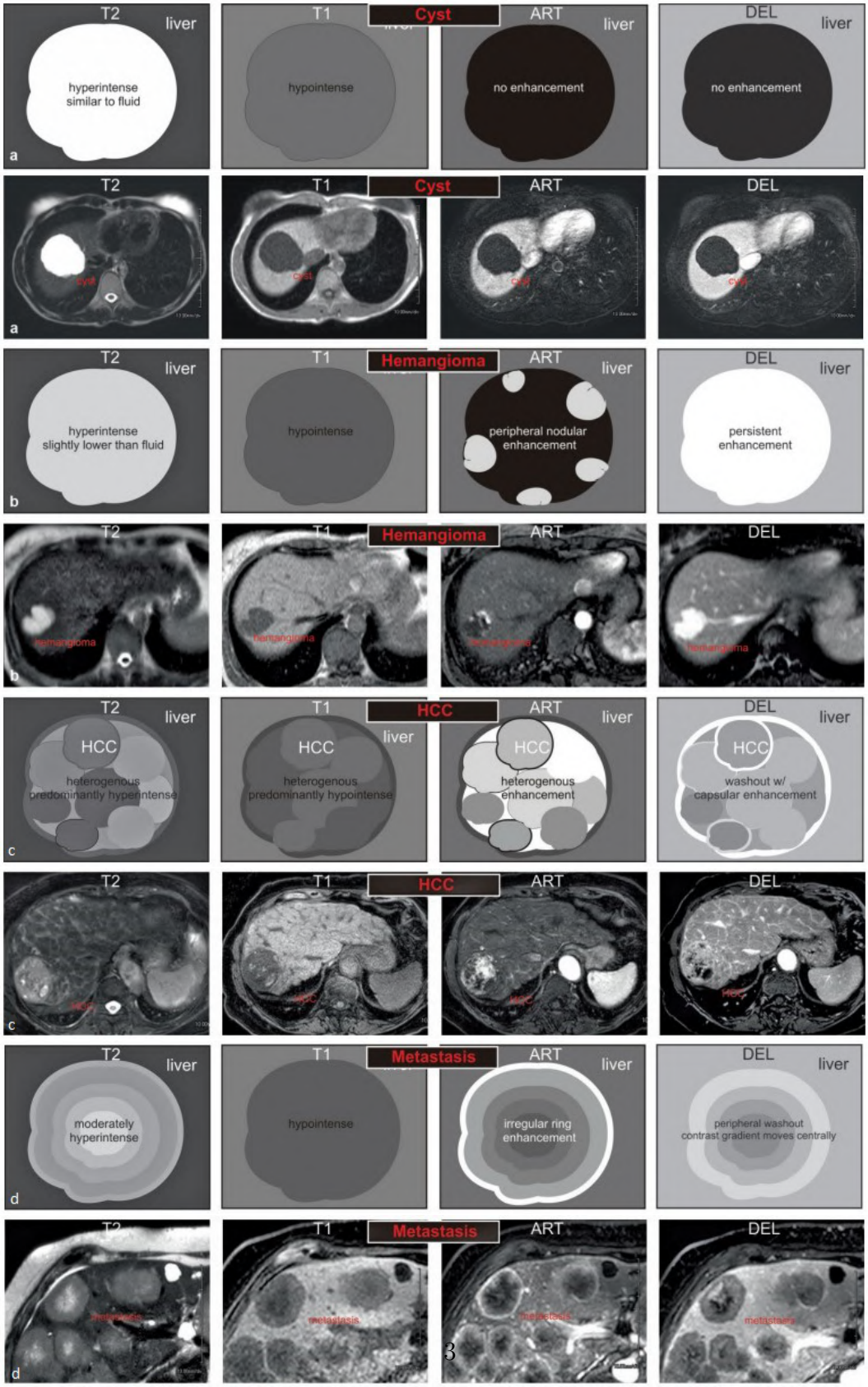


FIGURE 1.1 – Schémas et exemples de l'apparence de quatre lésions du foie parmi les plus fréquentes, dans des images IRM pondérées en T2 (à gauche), en T1 (milieu gauche), au temps artériel (milieu droite) et tardif (à droite). Tiré de HUSSAIN et SORRELL (2015).

de diamètre.

L'image pondérée en T2 montre les kystes du foie avec un signal fortement hyperintense, tandis qu'ils apparaissent hypointenses dans l'image en T1. Ces kystes n'étant pas vasculaires, leur particularité est de ne pas être mis en évidence par le produit de contraste, et par conséquent d'apparaître sombres quel que soit le temps d'injection (voir la figure 1.1a).

Les hémangiomes hépatiques

Les hémangiomes hépatiques sont les lésions vasculaires bénignes les plus fréquentes. La plupart des patients sont asymptomatiques, et ces lésions sont en général détectées incidemment lors d'examens d'imagerie. Il est important d'un point de vue radiologique de les distinguer des lésions malignes, et notamment des métastases.

Ils apparaissent modérément hyperintenses dans l'image pondérée en T2, et hypointense dans les images pondérées en T1. L'image au temps artériel met en évidence des nodules sur la périphérie de la lésion (qui peuvent apparaître comme un anneau interrompu au bord). Ce signal persiste au temps tardif, et s'étend dans toute la lésion pour montrer un signal hyperintense dans l'ensemble de la lésion. Cette dernière caractéristique permet de les différencier des métastases (voir la figure 1.1b).

Les carcinomes hépato-cellulaires

Il s'agit de la tumeur primaire du foie la plus fréquente. Elle est fortement associée aux cirrhoses, indifféremment d'origine alcoolique ou virale.

Les carcinomes hépato-cellulaires apparaissent principalement hyperintenses dans les images pondérées en T2, avec une certaine hétérogénéité due à la présence de sous-nodules montrant des signaux d'intensité variable. Cette hétérogénéité se retrouve dans l'image pondérée en T1, avec des signaux hypo, iso, et hyperintenses qui cohabitent dans la lésion. À la phase artérielle, l'intensité augmente de manière hétérogène, et s'atténue à la phase tardive, en conservant une intensité accrue sur le bord de la tumeur (voir la figure 1.1c).

Les métastases

Les métastases hépatiques sont 18 à 40 fois plus fréquentes que les tumeurs primaires du foie (Namasivayam, Martin et Saini 2007). Elles sont le plus souvent asymptomatiques tant que la charge tumorale reste faible. Elles proviennent le plus souvent de cancers primaires du tube digestif via la veine porte, des cancers du sein ou du poumon.

Elles apparaissent légèrement hyperintenses dans les images pondérées en T2, avec une intensité plus faible sur les bords de la tumeur pour les grosses lésions. Les images T1 les montrent légèrement hypointenses. Au temps artériel, l'intensité du bord de la tumeur augmente de manière irrégulière, et cette augmentation d'intensité s'atténue au temps tardif (voir la figure 1.1d).

1.2 Critères quantitatifs en radiologie pour l'oncologie hépatique

Si l'interprétation qualitative des images par les radiologues est souvent suffisante pour établir des diagnostics, comme on l'a illustré à la section précédente, certaines applications demandent en revanche d'estimer l'état des patients avec des critères quantitatifs, que ce soit en routine clinique (par exemple pour sélectionner les patients éligibles à des interventions chirurgicales ou de radiothérapie) ou pour des problématiques de recherche (par exemple pour comparer objectivement l'effet d'un traitement par rapport à un placebo).

Cette section vise à illustrer l'intérêt des différents outils de détection et de segmentation automatique pour évaluer quantitativement l'avancée de la maladie chez des patients atteints de lésions dans le foie, avec l'exemple de quatre critères couramment utilisés en radiologie.

1.2.1 Le critère RECIST

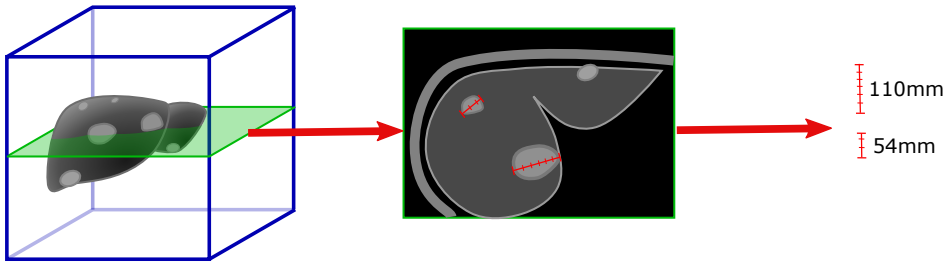


FIGURE 1.2 – Procédure RECIST : sélection de la coupe où la tumeur a le diamètre maximal, puis mesure du diamètre, pour chaque tumeur cible.

De la nécessité d'évaluer quantitativement et objectivement l'efficacité de traitements lors d'essais clinique sont nées, à l'initiative de l'organisation mondiale de la santé dans les années 1990, les recommandations RECIST (pour *Response Evaluation Criteria In Solid Tumors*) qui ont été publiées en 2000 (voir EISENHAUER et al. (2009) pour un historique plus complet). Celles-ci préconisent, en substance, de choisir un certain

nombre de tumeurs *cibles* (avec un maximum de 10, et 2 maximum par organe), et pour chacune des tumeurs, de sélectionner la coupe d'une image volumique où son diamètre est maximal, et de mesurer ce diamètre. Cette procédure est illustrée sur la figure 1.2. En prenant en compte l'évolution du diamètre des lésions cibles, ainsi que des critères subjectifs sur l'évolution des lésions non-cibles (comme « l'évolution indiscutable » de la taille de ces lésions), on attribue à l'évolution de la maladie l'une des quatre catégories suivantes : réponse complète, réponse partielle, progression ou stabilisation.

L'avantage principal de cette procédure est sa simplicité de mise en place, puisque mesurer le diamètre 2D des tumeurs peut être fait rapidement par un radiologue sans outil particulier, le tout sur un nombre restreint de tumeurs.

Son principal inconvénient est qu'elle ne prend en compte qu'une faible quantité d'information quantitative pour chaque image, et repose sur des critères subjectifs peu précis pour le reste des lésions. De plus, le diamètre maximal mesuré dans les coupes axiales ne donne qu'une information partielle sur chaque tumeur, en ignorant notamment l'étalement vertical. On peut noter également qu'une tumeur peut évoluer sans changer de taille ni même de forme, et qu'un critère uniquement basé sur la taille ne pourra rendre compte de cette évolution.

La segmentation automatique des lésions dans le foie permettrait de faire gagner du temps pour estimer ce critère, tout en le rendant plus reproductible, notamment en sélectionnant automatiquement la coupe montrant le diamètre maximal.

1.2.2 Estimation du volume tumoral

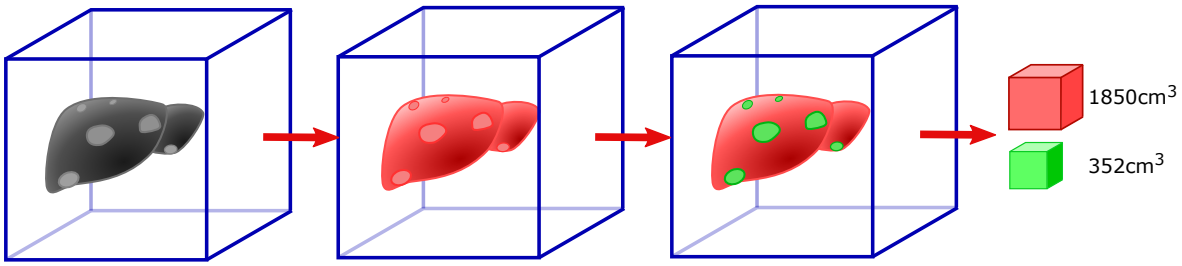


FIGURE 1.3 – Une chaîne de traitement possible pour l'estimation de la charge tumorale : segmentation du foie, segmentation du tissu tumoral, calcul des volumes.

Une autre manière de quantifier l'évolution de la maladie consiste à estimer le volume de tissu tumoral contenu dans le foie et de le comparer au volume du foie lui-même. Le rapport de ces deux volumes est une manière de mesurer la charge tumorale. Cette mesure est par ailleurs couramment utilisée comme critère pour sélectionner les patients éligibles à une hépatectomie, une transplantation, ou un traitement par radiothérapie (HSU et al. 2010 ; Y.-H. LEE et al. 2013 ; PALMER et al. 2020).

Pour la calculer, on a besoin de segmenter le foie et les tumeurs, c'est-à-dire classifier chaque voxel de l'image en fonction de son appartenance à du tissu tumoral, du parenchyme sain ou au reste de l'image. La procédure est illustrée sur la figure 1.3.

L'évolution de ce rapport de volumes renseigne sur la progression de la maladie, en prenant en compte toutes les tumeurs du foie, ainsi que celles qui apparaissent et disparaissent. Le volume des tumeurs donne également plus d'information que le diamètre 2D. L'étape de segmentation nécessaire au calcul de la charge tumorale est cependant fastidieuse, voire impossible à effectuer à la main de manière suffisamment rapide pour une utilisation clinique de routine. Elle nécessite donc des outils de segmentation automatique performants, et si possible capables de traiter des images provenant de modalités, séquences, ou temps d'injection différents. Certaines lésions ne sont en effet détectables que dans des images de certaines modalités.

1.2.3 LI-RADS

Depuis 2011, l'American College of Radiology (ACR) met à jour des recommandations pour catégoriser les lésions du foie par rapport à leur probabilité d'être des carcinomes hépato-cellulaires, dans le but de standardiser l'interprétation des images [2]. Ces recommandations sont appelées LI-RADS (pour *LIver Reporting And Data System*), et le principe est d'attribuer à chaque lésion une catégorie parmi cinq (de LR-1 à LR-5) en fonction de la probabilité de malignité. Une lésion probablement maligne qui n'est pas un carcinome hépato-cellulaire est classifiée à part (LR-M).

Pour faire cette caractérisation, l'ACR propose une procédure qui consiste, d'abord, à faire une première estimation de la catégorie à partir de critères principaux (la mise en évidence de la lésion au temps artériel, la taille, l'atténuation du contraste ou « *washout* »après le temps artériel, la mise en évidence d'une « capsule »...) qui doivent être reportés dans un tableau, puis de l'affiner dans un second temps avec des caractéristiques auxiliaires, dont notamment l'hyperintensité en IRM pondérée en T2. Ils préconisent comme dernière étape d'estimer si la catégorisation obtenue « semble raisonnable et appropriée ». Cette procédure doit être répétée pour chaque lésion.

Pour appliquer cette procédure, la segmentation automatique représente un gain de temps important, surtout dans le cas où le patient a beaucoup de lésions. Elle pourrait permettre notamment d'automatiser le calcul de la taille de la tumeur. La segmentation individuelle des tumeurs et leur identification dans les images de plusieurs modalités seraient dans ce cas souhaitable, de manière a pouvoir évaluer les critères auxiliaires qui nécessitent de s'appuyer sur plusieurs temps d'injection et séquences IRM. Une automatisation complète de la procédure est même envisageable, avec un calcul automatique des caractéristiques à prendre en compte pour la catégorisation. Cela permettrait, en

2. https ://www.acr.org/Clinical-Resources/Reporting-and-Data-Systems/LI-RADS

plus de faire gagner du temps, de diminuer l'importance des critères subjectifs et ainsi rendre plus reproductible l'estimation de la catégorie LI-RADS.

1.2.4 Vers la radiomique

La radiomique est un champ de recherche récent (on considère que les articles fondateurs sont Lambin et al. (2012) et H. J. W. L. Aerts et al. (2014)), qui part de l'hypothèse selon laquelle les images médicales contiennent de l'information sur l'état d'un patient qu'un radiologue ne pourrait pas voir à l'œil nu. À partir d'images et des segmentations des lésions, l'approche radiomique consiste à calculer, pour chaque lésion, un certain nombre de caractéristiques qui la décrivent le plus exhaustivement possible, et d'utiliser ces caractéristiques pour prédire une variable, qui peut être selon les cas la survie, la réponse à un traitement futur ou le diagnostic, par exemple. Ces caractéristiques sont des valeurs calculées à partir d'une image et d'un masque de segmentation, qui peuvent décrire la forme (diamètre 3D, volume, sphéricité...), les intensité (énergie, variance, entropie...), ou encore la texture en calculant des statistiques sur les intensités des voxels voisins. La prédiction de la variable d'intérêt à partir des caractéristiques extraites se fait à l'aide d'une méthode d'apprentissage statistique (régression linéaire, méthodes à vecteurs de support, arbres de décision, etc.).

L'attrait principal de cette approche vient de la possibilité, offerte par l'apprentissage statistique, de prendre en compte un grand nombre de caractéristiques pour chaque lésion, en ne se limitant plus à la taille ou à des critères binaires (mise en évidence au temps artériel, hyperintensité en T2, etc.). Cela a l'avantage, d'une part, de diminuer l'importance du facteur humain dans la caractérisation de lésions, mais surtout d'offrir la possibilité d'exploiter de l'information invisible à l'œil nu.

Cependant cette approche manque encore de maturité pour pouvoir être utilisée en routine, et du travail de recherche est encore nécessaire pour déterminer un ensemble de caractéristiques robustes, reproductibles, à utiliser avec des procédures standardisées dans la veine de RECIST et LI-RADS. Une manière de quantifier l'évolution d'une lésion maligne dans le foie, par exemple, serait de trouver une signature capable de prédire la survie, et de suivre l'évolution de cette prédiction au cours du traitement. La figure 1.4 représente la chaîne de traitement pour prédire la survie avec cette approche.

Dans ce cas, cette approche demanderait donc de segmenter individuellement les tumeurs dans des images de plusieurs modalités (qui contiennent potentiellement des informations complémentaires sur la maladie). Des outils automatiques adaptés sont par conséquent indispensables pour appliquer cette approche dans un contexte clinique.

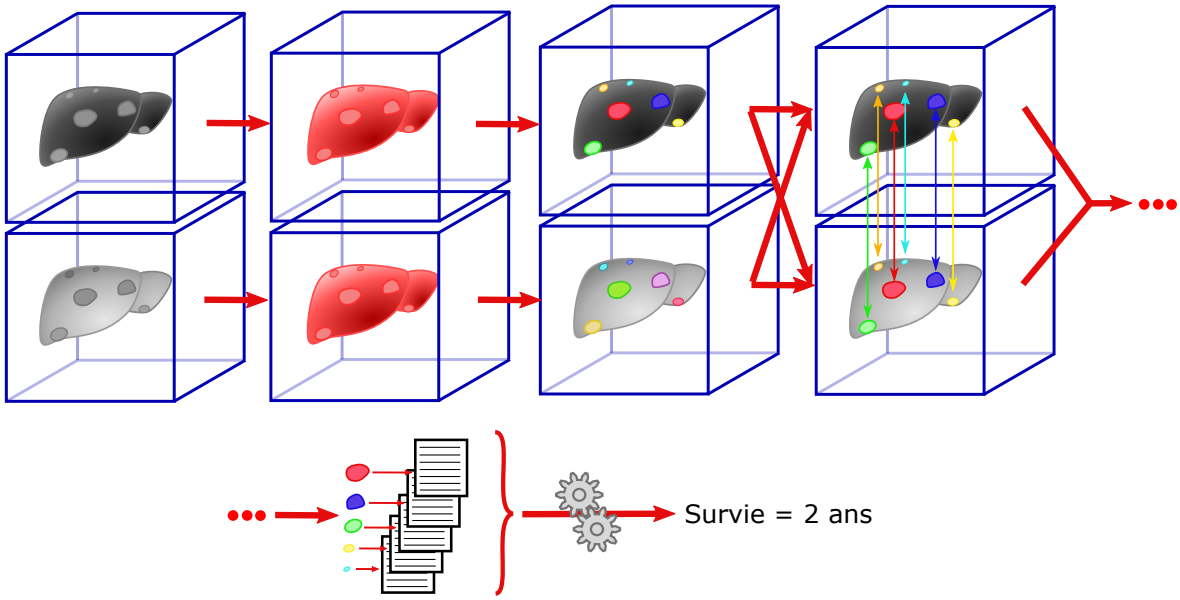


FIGURE 1.4 – Chaîne de traitement pour l'analyse radiomique en imagerie multi-modale pour l'oncologie hépatique. Les cinq étapes sont : segmentation du foie, segmentation individuelle des tumeurs, identification des tumeurs, extraction de la signature de chaque tumeur en utilisant les deux images, prédiction de la variable.

1.3 Problématique et contributions

1.3.1 Problématique

Pour automatiser l'extraction de caractéristiques de lésions hépatiques qui permettent d'estimer quantitativement l'état d'un patient (selon l'une des quatre approches présentées dans la section précédente par exemple), on considère la chaîne de traitement décrite ci-dessous. À partir d'une série d'images de différentes modalités acquises chez un patient ayant des tumeurs dans le foie, on applique les traitements suivants :

Segmentation du foie. Dans toutes les images, on classifie les voxels en fonction de leur appartenance au foie ou au reste de l'image. Le résultat de cette étape est un masque de segmentation par image.

Segmentation individuelle des lésions. Chaque lésion doit être segmentée, éventuellement individuellement dans chaque image afin de pouvoir calculer des caractéristiques sur chacune d'entre elles. Autrement dit, on a besoin d'un masque de segmentation par image et par lésion.

Identification des lésions. Pour les approches qui nécessitent d'extraire des caractéristiques de plusieurs images pour chaque lésion (comme LI-RADS, ou l'approche

radiomique), il faut faire correspondre les masques obtenus à l'étape précédente. Après cette étape, on a autant de masques de segmentation que d'images pour chaque lésion.

Extraction des caractéristiques. Les masques ainsi obtenus permettent d'obtenir la taille des lésions, ce qui est notamment utile pour automatiser l'estimation du RECIST. En utilisant les images en plus des masques, on peut obtenir des caractéristiques nécessitant les intensités (la mise en évidence de la capsule, ou l'hyperintensité en T2 pour LI-RADS). Pour l'approche radiomique, les images et les masques permettent de calculer un certain nombre de caractéristiques (qu'on appelle dans ce cas une signature) parmi celles standardisées par Zwanenburg et al. (2020). Cet ensemble de caractéristiques contient des caractéristiques de taille, de forme, d'intensité et de texture.

Estimation de la variable. On combine ensuite ces caractéristiques, soit en suivant des recommandations pour attribuer une catégorie au patient (comme RECIST et LI-RADS), soit, comme pour l'approche radiomique, en les mettant en entrée d'un modèle d'apprentissage statistique préalablement entraîné pour prédire la variable d'intérêt (la survie ou la réponse à un traitement par exemple).

Cette thèse vise à étudier les méthodes permettant d'automatiser les trois premiers maillons de cette chaîne de traitement, avec une emphase sur le traitement multi-modal de ces problèmes. Plus précisément, je m'intéresse aux questions suivantes : Quels sont les enjeux inhérents au traitement multi-modal des images ? Quels en sont les bénéfices potentiels ? Nous verrons que l'apprentissage profond (que nous appellerons « Deep Learning » dans toute cette thèse, pour suivre l'usage dominant) domine l'état-de-l'art pour tous les problèmes liés aux trois premières étapes de cette chaîne de traitement. Dès lors, comment cette classe de méthodes permet-elle de faire face à ces enjeux ? Comment, en particulier, les méthodes basées sur le Deep Learning s'accommodent-elles des décalages géométriques entre les images de différentes modalités ? Également, quelles caractéristiques apprennent les modèles obtenus par Deep Learning ? Comment gagner en compréhension sur leur fonctionnement ? Sur ce qui influence leurs prédictions ?

1.3.2 Contributions

La suite du document est subdivisée en 4 chapitres.

Le chapitre 2 étudie l'étape de segmentation, dans le contexte ou des images de plusieurs modalités sont appariées mais pas recalées. Je cherche, dans un premier temps, à répondre à la question suivante : quel apport pourraient avoir les informations multi-modales sur la tâche de segmentation ? Pour cela je compare différentes stratégies d'apprentissage sur des données d'IRM du foie de séquences T1 et T2. Je propose dans

un second temps une méthode basée sur l'optimisation jointe d'une tâche de segmentation et d'une tâche de recalage permettant d'intégrer une contrainte de similarité à l'apprentissage d'un réseau de segmentation.

Dans le chapitre 3, je m'écarte momentanément de la chaîne de traitement évoquée à la section précédente pour m'intéresser au problème de l'interprétabilité en Deep Learning. En partant du constat de l'intérêt de l'interprétabilité des réseaux de segmentation d'images médicales, d'une part, et de l'incapacité des méthodes d'interprétation proposées jusque là dans l'état de l'art à s'appliquer aux réseaux de segmentation d'autre part, je propose une méthode d'interprétabilité spécifique aux réseaux de segmentation.

Le chapitre 4 présente un travail préliminaire sur une méthode visant à effectuer en une seule passe les étapes de segmentation individuelle des lésions et d'identification des lésions détectées. Je montre son potentiel, d'abord avec une expérience sur des données synthétiques, puis sur les données d'IRM de foie du chapitre 2.

Enfin le chapitre 5 propose une conclusion de chacun des chapitres précédents, avec une réflexion sur l'état de l'art de chacun des problèmes traités, et comment mes contributions s'y insèrent. Je propose ensuite dans ce chapitre des perspectives de recherche dans l'objectif d'une automatisation plus fiable et plus complète de toute la chaîne de traitement.

Chapitre 2

Segmentation en imagerie multi-modale

Certaines modalités, comme le CT-scan, sont anatomiques, c'est-à-dire qu'elles montrent des contours précis entre les différents tissus du corps. D'autres sont fonctionnelles, comme la Tomographie par Emission de Positons (PET) ou certaines séquences d'IRM, c'est-à-dire qu'elles renseignent sur l'activité (fonction, métabolisme, etc.) des organes ou des tumeurs.

Lors d'un examen d'imagerie, les radiologues acquièrent souvent plusieurs modalités, à plusieurs intervalles de temps après l'injection d'un produit de contraste, de manière à avoir le plus d'information possible sur l'état du patient. Les contours d'une zone d'intérêt peuvent ainsi apparaître plus nets sur certaines images que sur d'autres. Lorsqu'un radiologue segmente manuellement une lésion ou un organe dans une image ou celui-ci n'apparaît pas avec des contours précis, il peut s'appuyer sur d'autres images pour améliorer la qualité de sa segmentation.

Le but de ce chapitre est d'étudier quelles méthodes de segmentation automatique peuvent être utilisées lorsque l'on dispose d'images de plusieurs modalités pour chaque patient. Le cas d'application étudié est celui de la segmentation du foie dans les images IRM en pondération T1 et en pondération T2. La séquence IRM pondérée en T1 donne une image anatomique, où le foie apparaît bien contrasté et est facile à segmenter manuellement. À l'inverse, le foie en IRM pondérée en T2 est difficile à distinguer des tissus qui l'entourent. De plus, cette séquence est plus longue à acquérir, ce qui résulte en une plus faible résolution dans la direction orthogonale au plan de coupe et la rend plus sujette à des artefacts d'acquisition. La segmentation manuelle du foie y est par conséquent plus difficile.

Je propose dans la section 2.1 une revue des différents problèmes liés à la segmentation multi-modale abordés dans la littérature, et des différentes techniques utilisées pour les

résoudre.

La section 2.2 présente mes expériences qui comparent différentes stratégies proposées dans la littérature pour segmenter le foie en IRM pondérée en T1 et T2.

Enfin, dans la section 2.3 je propose une méthode de segmentation jointe de deux images de modalités différentes, qui intègre une information a priori de similarité entre les deux masques de segmentation.

2.1 Les différents problèmes de segmentation d'images médicales

Deux revues de littérature sur les méthodes de Deep Learning pour la segmentation d'images médicales sont récemment parues : celle présentée par Tajbakhsh et al. (2020) se concentre sur le problème de la rareté des annotations en imagerie médicale et décrit quelles méthodes ont été proposées dans la littérature pour y répondre, tandis que Taghanaki et al. (2020) font une revue plus générale des méthodes de segmentation, à la fois pour les images naturelles (problème généralement désigné par l'appellation « segmentation *sémantique* ») et les images médicales. D'autres revues abordant le problème selon des angles différents existent, et sont répertoriées dans les introductions de ces deux articles.

Dans cette section, je propose une revue des méthodes de segmentation d'images médicales en distinguant les problèmes traités selon les trois critères suivants :

— Le caractère *multi-modal*. La base de données considérée contient-elle des images de différentes modalités ? Pour simplifier, on considérera également que le problème est multi-modale si la base de données contient des images de différentes séquences IRM.

— L'*appariement* des données. On dit que les données sont appariées si, pour chaque patient, on a plusieurs images différentes de la même zone du corps. Cela implique la plupart du temps une segmentation jointe de plusieurs images.

— Si les données sont appariées, le *recalage* des images entre elles. Étant donné une série d'images, les voxels de mêmes coordonnées correspondent-ils à un même point du corps du patient sur toutes les images ? Certaines images sont acquises en même temps, et sont donc parfaitement recalées. D'autres sont acquises les unes après les autres, ou sur des machines différentes, et ne le sont par conséquent pas (à cause des mouvements ou de la respiration du patient entre les acquisitions).

Cette classification a pour but de replacer les différentes innovations proposées dans le contexte du problème auxquelles elles visent à répondre. Elle permet également de pré-

ciser l'appellation « segmentation multi-modale », qui est employée dans la littérature pour désigner des problèmes qui nécessitent des méthodologies différentes.

Cinq classes de problèmes découlent de cette classification : la segmentation mono-modale, la segmentation non-appariée, la segmentation appariée recalée, la segmentation appariée non-recalée mono-modale, la segmentation appariée non-recalée multi-modale. C'est cette dernière classe de problème qui est l'objet principal de ce chapitre

L'objectif de cette section est de présenter l'état de l'art des méthodes utilisées pour résoudre les problèmes des différentes classes, et de discuter du bénéfice que pourraient avoir ces innovations pour d'autres problèmes.

Cette revue se concentre sur les méthodes par apprentissage profond (qu'on désignera par le terme anglais « Deep Learning » dans la suite, pour suivre l'usage dominant), celles-ci ayant concentré l'immense majorité de l'innovation en segmentation médicale de ces dernières années.

2.1.1 Segmentation mono-modale

Lorsque la base de données est mono-modale, et que l'on n'a pas de données appariées, on se trouve dans le contexte de segmentation le plus simple, où chaque image de la base provient de la même distribution.

Architecture

Un axe de recherche favorisé par la littérature est celui de l'architecture des réseaux de neurones. On peut toutefois affirmer que l'architecture U-net, proposée par Ronneberger, Fischer et Brox (2015), est devenue standard en segmentation d'images médicales. Si quelques améliorations à cette architecture ont été étudiées (mécanismes d'attention, blocs de convolution résiduels ou denses par exemple, voir l'introduction de Tajbakhsh et al. (2020)), certains travaux (Hofmanninger et al. 2020 ; Isensee, Petersen, Klein et al. 2018) suggèrent que des apports méthodologiques à cette base n'auraient qu'une influence mineure sur les performances par rapport à la qualité des données. Isensee, Petersen, Klein et al. (2018) proposent une procédure systématique pour adapter les métas-paramètres (taille des *batchs*, résolution des images d'entrée...) d'un U-net à un nouveau jeu données, qu'ils baptisent « nnUnet », et parviennent avec elle à remporter des compétitions de segmentation (Kavur et al. 2021 ; Simpson et al. 2019), battant des équipes aux méthodes novatrices.

D'autres architectures, pourtant populaires pour la segmentation sémantique en vision par ordinateur, comme Deeplab (L.-C. Chen et al. 2017) et PSPNet (H. Zhao et al.

2017), peinent à s'imposer dans le domaine médical.

Supervision partielle et régularisation

Un autre axe de recherche privilégié par la littérature est celui de la supervision partielle, où l'on cherche à tirer parti de données sans annotations. C'est le fil conducteur de la revue de littérature de Tajbakhsh et al. (2020). On peut distinguer deux sous-problèmes : d'une part, l'apprentissage faiblement supervisé, où les annotations de chaque image sont partielles (par exemple, des boîtes englobantes (Dai, K. He et Sun 2015), des points (Bearman et al. 2016) ou des gribouillages (D. Lin et al. 2016). et d'autre part, l'apprentissage semi-supervisé, qui consiste à tirer parti pendant la phase d'apprentissage des images non-annotées. Ces deux problèmes visent à répondre à la rareté des annotations, particulièrement importante en imagerie médicale puisque celles-ci sont souvent coûteuses à obtenir.

Pour tirer parti des données sans annotations, ces méthodes reposent pour la plupart sur l'optimisation, en plus de l'erreur de segmentation des données annotées, d'une fonction de coût non-supervisée. Cette dernière peut être vue comme un terme de régularisation, et peut prendre différentes formes selon l'information que l'on a *a priori* sur le problème.

Ben Ayed, Desrosiers et Dolz (2019) proposent de discerner trois types d'informations a priori :

— Les informations sur la structure : un pixel est plus probablement de la classe de ses voisins que d'une autre classe, et d'autant plus si son intensité est similaire. Les champs aléatoires conditionnels (*CRF*) sont une solution populaire pour intégrer cette information, aussi bien en post-traitement (L.-C. Chen et al. 2017) que comme régularisation de la fonction de coût (Tang et al. 2018).

— Les informations basées sur la connaissance du problème. Cela peut être une probabilité a priori de localisation des objets à segmenter dans l'image, comme dans l'article de Y. Zou et al. (2018) ; ou des contraintes sur la taille des objets : des contraintes d'égalité, comme par exemple Y. Zhou et al. (2019) (l'objet segmenté doit faire une certaine taille), ou des contraintes d'inégalités, comme pour la méthode de Kervadec et al. (2019) (l'objet segmenté doit avoir une taille comprise entre deux valeurs) ;

— L'information directement apprise des données, grâce à l'apprentissage adversaire. Un réseau discriminateur est entraîné à distinguer les segmentations issues des annotations manuelles des segmentations prédites par un réseau segmenteur, tandis que le réseau segmenteur apprend à tromper le discriminateur en plus de segmenter les images. Le terme non-supervisé de la fonction de coût du segmenteur dépend donc du réseau discriminateur, et peut prendre plusieurs formes. Des

méthodes se basant sur ce principe ont notamment été proposées par Samson et al. (2019), Luc et al. (2016), Ghafoorian et al. (2018), Hung et al. (2018), Yizhe Zhang et al. (2017).

L'intégration d'informations a priori est un axe de recherche intéressant, non seulement parce qu'il permet de répondre à l'une des problématiques les plus importantes de l'apprentissage automatique en imagerie médicale, qui est la rareté des annotations, mais aussi parce que l'on peut imaginer toute sortes de contraintes, notamment anatomiques (taille et localisation relative des organes, forme (comme Zeng et al. (2019)), aspect...). Cet axe offre par conséquent des perspectives prometteuses. La méthode que je détaille à la section 2.3 se base sur l'information a priori que les masques des foies dans les images T1 et T2 ne doivent différer que d'une transformation élastique.

Si ces axes de recherche (architecture, supervision partielle et régularisation) sont surtout étudiés dans un contexte de segmentation mono-modale, il demeure qu'ils peuvent être pertinents dans d'autres contextes. Dans la suite, je me concentre sur les méthodes visant à répondre à un problème spécifique au contexte pour lesquelles elles ont été proposées.

2.1.2 Segmentation non-appariée

Cette classe correspond aux contextes pour lesquels les images à segmenter sont issues de plusieurs modalités différentes, mais on n'a pas d'information sur un quelconque appariement entre elles. On cherche dans ce cas à obtenir un modèle capable de segmenter ces images indifféremment, quelle que soit leur modalité. L'enjeu est d'étudier le compromis entre variété de données et mélange de distributions : alors qu'on sait qu'un réseau généralise mal a des images venant de distributions qu'il n'a pas vues à l'entraînement, vaut-il mieux entraîner des réseaux différents sur chaque modalité, quitte à réduire la taille des bases d'entraînement, ou tout mélanger et espérer que les caractéristiques pertinentes pour une distribution le soit aussi pour l'autre ?

Wang et al. (2019) ont montré qu'il était possible d'utiliser un même réseau pour segmenter le foie en CT et en IRM. Valindria et al. (2018) ont proposé une comparaison d'approches où une partie des couches du réseau est entraîné spécifiquement sur une modalité. J'étudie ce compromis dans la partie 2.2 dans le cas particulier où les images du foie en IRM pondérées en T1 et T2 sont appariées.

Certains articles proposent des méthodes dans le cas où l'on ne dispose pas des segmentations annotées dans une des modalités (Y. Huo et al. 2018 ; Yuan et al. 2019 ; Z. Zhang, L. Yang et Zheng 2018). Cette classe de problèmes devient alors un cas particulier de l'adaptation de domaine non-supervisée, qui est un problème voisin de l'apprentissage semi-supervisé, où les images non-annotées proviennent d'une distribution différente de celle des images annotées. Le problème d'adaptation de domaine

non-supervisée suscite surtout beaucoup d'intérêt pour des applications de segmentation pour la conduite autonome, où il est très fastidieux d'obtenir des annotations (il faudrait assigner une classe à chaque pixel d'une vidéo prise en caméra embarquée) alors que des images issues de jeux vidéos (comme GTA5, voir Y. Zou et al. 2018) sont disponibles en grande quantité.

Le problème d'adaptation de domaine non-supervisée a gagné de l'intérêt grâce à l'essor de l'apprentissage adversaire, et notamment la technique du Cycle-GAN (proposée par Zhu et al. 2017) qui consiste à entraîner un réseau pour transformer une image d'une modalité A vers une autre modalité B, et un autre de B vers A, en contraignant les images à être les mêmes après un cycle. Cette technique permet donc de passer d'une modalité à l'autre sans avoir de données appariées, et donc d'entraîner un segmenteur sur plusieurs modalités à la fois en n'ayant que les annotations d'une seule. Elle est maintenant largement adoptée par d'autres méthodes qui ne se limitent pas aux problèmes de supervision partielle, comme la synthèse d'images CT à partir d'images IRM (avec l'objectif de pouvoir se passer d'acquisitions CT pour la planification de radiothérapie notamment, voir Wolterink et al. (2017), H. Yang et al. (2018)), ou le recalage multi-modal (Guo et al. 2020).

2.1.3 Segmentation appariée recalée

Lorsque l'on a des images appariées de différentes modalités, qui sont recalées entre elles, c'est-à-dire qu'un voxel de même coordonnées dans les toutes les images correspond au même point du corps, dans la plupart des cas un seul masque de segmentation suffit pour segmenter toutes les images d'un même appariement.

Toutes les méthodes proposées pour résoudre ce problème entraînent un réseau prédisant ce masque en prenant en entrée chacune des images dans un canal. L'axe de recherche privilégié est celui de l'architecture, qui a donné lieu un certain nombre de propositions assez élaborées, comme *HyperDenseNet* (Dolz et al. 2019), ou *OctopusNet* (Y. Chen et al. 2019). T. Zhou, Ruan et Canu (2019) proposent une revue des méthodes et des jeux de données publics de la littérature pour ce problème.

Néanmoins je n'ai pas trouvé de consensus se dégager sur une architecture réellement avantageuse pour la segmentation multi-modale. J'ai donc privilégié l'architecture U-net pour mes expériences de la suite de ce chapitre.

2.1.4 Segmentation appariée non-recalée mono-modale

Certaines applications cliniques requièrent la segmentation d'un organe dans deux images de même modalité d'un même patient, acquises à des moments différents. C'est le cas par exemple en radiothérapie, où l'organe est segmenté dans une image ac-

quise pendant la phase de planification, et une image acquise le jour de l'intervention (Elmahdy et al. 2019).

Pour résoudre ce problème, des méthodes dont le principe est un apprentissage joint de la tâche de segmentation et la tâche de recalage (c'est-à-dire la prédiction d'une transformation géométrique qui aligne une image sur l'autre) ont été proposées (Beljaards et al. 2020 ; Elmahdy et al. 2019). Ces méthodes ont été proposées dans les cas où la tâche principale est le recalage, et elles se servent de la segmentation comme une tâche auxiliaire. Si l'on s'intéresse plutôt à la tâche de segmentation comme tâche principale, comme dans ce chapitre, l'avantage d'optimiser conjointement les deux n'est pas clair.

2.1.5 Segmentation appariée non-recalée multi-modale

C'est cette classe de problèmes qui nous intéresse principalement dans ce chapitre. La différence avec la classe précédente est que les images sont ici issues de modalités différentes. Les méthodes de segmentation et recalage appris conjointement décrites à la section 2.1.4, ne peuvent pas être appliquées telles quelles dans ce cas là. En effet, l'apprentissage de la tâche de recalage se fait typiquement en minimisant la différence des intensités des voxels entre l'image fixe et l'image transformée, ce qui n'est possible que lorsque ces intensités sont comparables, c'est-à-dire si les images sont de même modalité. La tâche de recalage devient donc plus difficile.

Pour faire du recalage d'images multi-modales avec du Deep Learning, deux stratégies sont en concurrence : d'une part en utilisant des descripteurs invariants selon la modalité, comme Z. Xu et al. (2020) qui utilisent MIND (Heinrich et al. 2012), et d'autre part en « traduisant » l'image d'une modalité vers une autre en utilisant les Cycle-GAN (par exemple Guo et al. 2020), comme évoqué à la section 2.1.2.

C'est cette dernière stratégie sur laquelle se basent F. Liu et al. (2020), en ajoutant aux tâches de « traduction » et de recalage une tâche de segmentation. Chartsias et al. (2019) apprennent une représentation commune aux deux modalités, dans le but d'apprendre conjointement les tâches de segmentation et de recalage. Comme dans la section précédente, les articles s'intéressant à la segmentation d'images appariées sans recalage le font sous l'angle de l'apprentissage joint avec le recalage, en utilisant la segmentation comme tâche auxiliaire.

Contrairement à ces articles, c'est la tâche de segmentation qui nous intéresse principalement dans ce chapitre. C'est pour cette raison que je m'intéresse d'abord, dans la section 2.2, aux stratégies ne faisant pas intervenir de recalage. Dans la section 2.3, je propose une méthode qui se sert du recalage comme tâche auxiliaire dans le but d'intégrer une contrainte de similarité entre les segmentations d'une paire d'images de modalités différentes.

2.2 Quelle stratégie pour la segmentation d'images appariées mais pas recalées ?

Le but de cette section est d'estimer dans quelle mesure certaines techniques et stratégies d'apprentissage utilisées pour traiter les problèmes décrits dans la section 2.1 sont pertinentes dans le cadre de la segmentation du foie dans des paires d'images IRM pondérées en T1 et T2.

Pour cela, je propose une comparaison de différentes stratégies s'articulant autour de deux axes. Le premier axe est celui de la stratégie multi-modale : est-il plus efficace, dans notre cas où les données sont appariées mais pas recalées entre elles, d'entraîner un réseau qui segmente indifféremment chacune des modalités mais une image à la fois, comme le font Wang et al. (2019) en segmentation non-appariée ? Ou d'entraîner un réseau prenant en entrée les deux images de chaque paire, comme c'est l'usage en segmentation d'images appariées et recalées (voir la section 2.1.3) ? Si le réseau prend en entrée les deux images de la paire, sera-t-il plus performant en prédisant les deux à la fois, et en étant ainsi *multitâche*, ou en en prédisant qu'une seule ? Vaut-il mieux entraîner un réseau spécifique pour chaque modalité, ou bien un réseau généraliste sera-t-il plus performant ? De même, pré-recaler les deux images permet-il d'améliorer les performances ?

Mes expériences sur des données synthétiques, détaillées dans la section 2.2.1, tendent à montrer qu'un réseau de segmentation est capable de localiser l'information pertinente dans toutes les images mises en entrée, même si celles-ci ne sont pas recalées, au moins lorsque le problème est simple et les décalages pas trop importants. On peut alors émettre l'hypothèse que sur des données réelles, ajouter de l'information anatomique en entrée du réseau par le biais de l'image pondérée en T1, peut permettre d'améliorer les performances de segmentation de l'image en T2, même si ces deux images ne sont pas recalées.

Le second axe est celui de la fonction de coût optimisée durant l'entraînement. Alors que Jadon (2020), Sudre et al. (2017) ont montré l'importance de ce paramètre en segmentation d'images médicales, ces dernières années ont vu l'essor de fonctions de coût apprises par apprentissage adversaire (Ghafoorian et al. 2018 ; Luc et al. 2016 ; Samson et al. 2019). Je compare trois fonctions classiques de coût voxel à voxel et trois fonctions de coût adversaires. L'hypothèse pour cet axe est que les fonctions de coût adversairesdevraient conduire à des modèles plus robustes, et à des segmentations plus précises en général. L'intuition est qu'en apprenant la distribution des masques de segmentation, on évite les prédictions impossibles (avec des trous par exemple) ou aux formes trop inhabituelles.

Alors que les articles de segmentation d'images appariées sans recalage abordent le problème sous l'angle de l'optimisation jointe des tâches de segmentation et de recalage (Chartsias et al. 2019 ; F. Liu et al. 2020), pour cette section on se restreint à des stratégies plus simples selon les deux axes décrits ci-dessus. Nous étudierons une méthode de segmentation et recalage appris conjointement dans la section 2.3.

2.2.1 Expérience préliminaire avec des données synthétiques

Lors de la segmentation jointe de deux images, un réseau de segmentation prend en entrée les deux images, chacune dans un canal différent. Cette expérience vise à déterminer l'importance de l'alignement des canaux sur la capacité d'un tel réseau à utiliser l'information pertinente dans les deux images.

Pour cela, l'idée est de créer un ensemble de données où l'information utile est séparée entre les deux canaux. Ainsi, on pourra mesurer la performance d'un réseau en fonction de l'alignement de ces deux canaux.

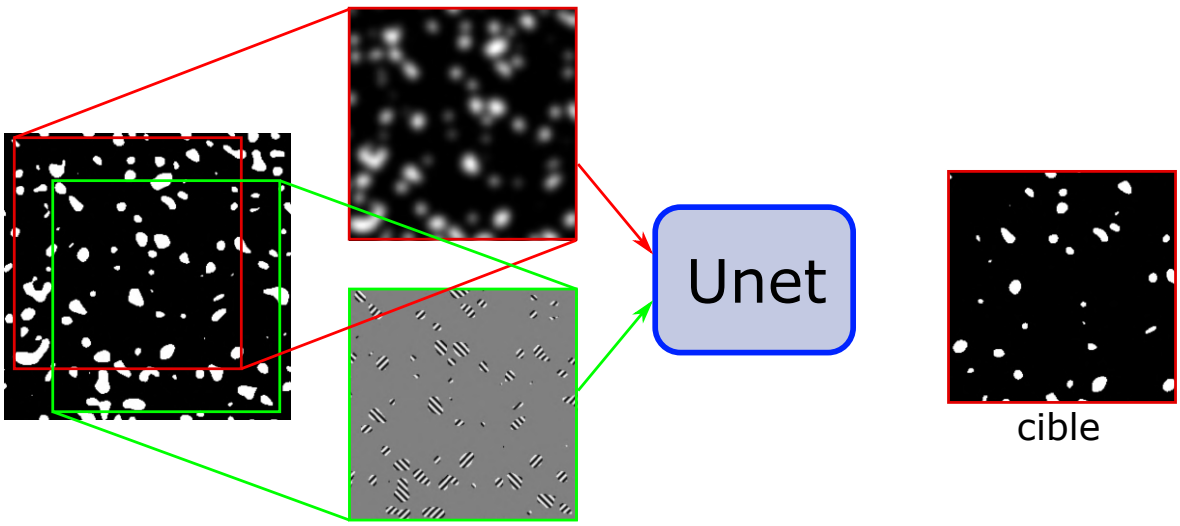


Figure 2.1 – Principe de l'expérience avec des données synthétiques. Le canal vert contient l'information qui permet de discriminer les patatoïdes à segmenter (striures orientées selon un angle inférieur à 90 degrés), mais la cible est alignée avec le canal rouge. De plus, les contours sont flous dans le canal rouge.

Partons d'un générateur de champ de patatoïdes striés, comme représenté figure 2.1. Un des canaux (représenté en vert sur la figure 2.1) comporte des striures sur chaque patatoïde, qui sont soit orientées à 45 degrés, soit à -45 degrés. L'autre canal contient les mêmes patatoïdes mais sans striures. On veut apprendre à un réseau à segmenter uniquement les patatoïdes striées à 45 degrés dans le canal vert, mais le masque de

segmentation doit être aligné avec le canal rouge. Ainsi, l'information de position (où sont les patatoïdes à segmenter) est uniquement dans le canal rouge, tandis que l'information discriminante (quels patatoïdes doivent être segmentés) est uniquement dans le canal vert. Pour rendre la tâche plus difficile, le canal rouge est flouté, de manière à ce que l'information de contour ne soit présente que dans le canal vert. Le décalage est une translation tirée aléatoirement d'une distribution uniforme. L'amplitude de cette distribution est un paramètre de l'expérience.

De cette manière, on s'assure que le réseau doive faire la correspondance entre les patatoïdes des deux canaux s'il veut les segmenter correctement. Pour éviter le sur-apprentissage, les images d'entraînement sont générées aléatoirement à la volée, de sorte que la même image n'est jamais utilisée deux fois pendant l'entraînement.

On remarque qu'il suffit de quelques époques à un U-net pour segmenter correctement les patatoïdes striés dans le bon sens. A-t-il pour autant appris à faire correspondre les patatoïdes des deux canaux ? Vérifions d'abord que le réseau n'ait pas simplement appris à utiliser le patatoïde strié le plus proche. La figure 2.2 montre des résultats de segmentation où le décalage est suffisamment important pour que les patatoïdes correspondants dans les deux canaux ne se touchent pas, de manière à ce que de nombreuses formes soient segmentées correctement alors que des formes plus proches dans le canal de discrimination ont des orientations différentes. On peut donc rejeter cette hypothèse. Comme mentionné précédemment, on peut également rejeter l'hypothèse du sur-apprentissage. On remarque également que lorsqu'il manque l'information de discrimination pour segmenter une patate, comme ça peut être le cas sur les bords (voir figure 2.2), le réseau prédit une probabilité de 0.5 que le patatoïde soit à segmenter. C'est un argument supplémentaire pour s'assurer que le réseau apprenne bien une correspondance entre les deux canaux.

Toutefois, cette capacité à extraire l'information non-alignée a des limites : la limite d'amplitude de la distribution des décalages au-dessus de laquelle l'apprentissage ne parvient plus à converger est de 35 pixels (pour des images 256×256). De plus, un réseau ne parvient pas facilement à généraliser aux décalages d'amplitude différentes de ceux qu'il a vus pendant l'entraînement, même pour des décalages d'amplitude plus faible (figure 2.3).

Néanmoins ces expériences montrent qu'un réseau de segmentation est capable, au moins pour des problèmes simples, d'apprendre à trouver la correspondance entre les deux canaux lorsqu'ils ne sont pas alignés. Cela justifie les expériences décrites dans la section 2.2.3 qui consistent à entraîner un réseau prenant en entrée les deux modalités.

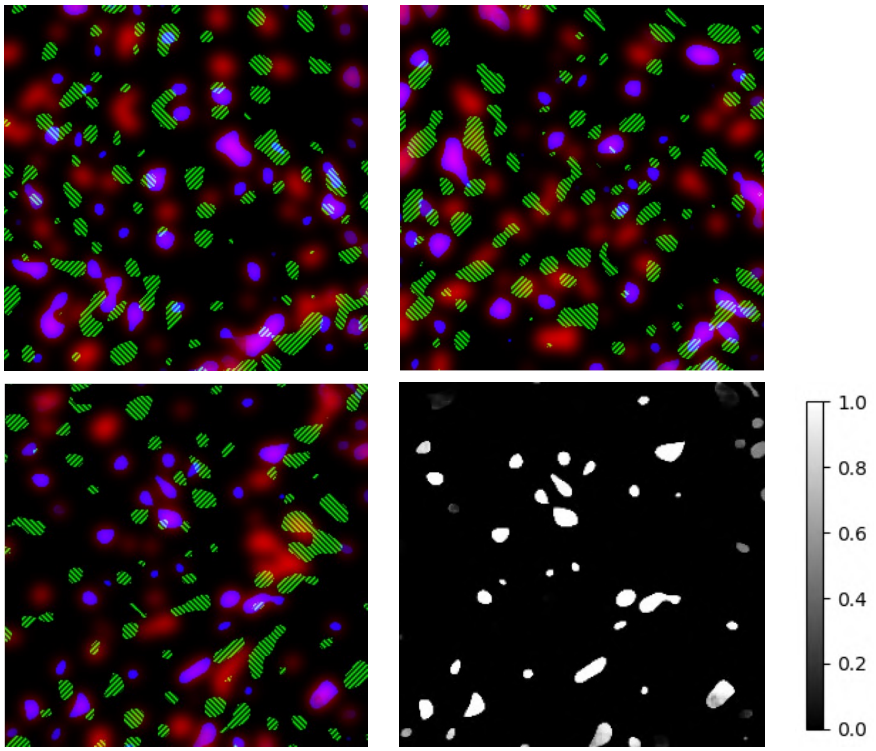


FIGURE 2.2 – Résultat de segmentation pour des différentes orientations de décalage. Pour les trois premières images : canal rouge = canal 1 (localité) de l'image d'entrée ; canal vert = canal 2 (discrimination) de l'image d'entrée ; canal bleu = segmentation prédite par le réseau. On remarque qu'elle est alignée sur le canal rouge, comme prévu, et qu'elle ne contient que les patatoïdes striés avec un angle inférieur à 90 degrés sur le canal vert. Dernière image : sortie du réseau seule. Sur les bords où l'information de discrimination a été coupée, on remarque que le réseau prédit une probabilité de 0.5.

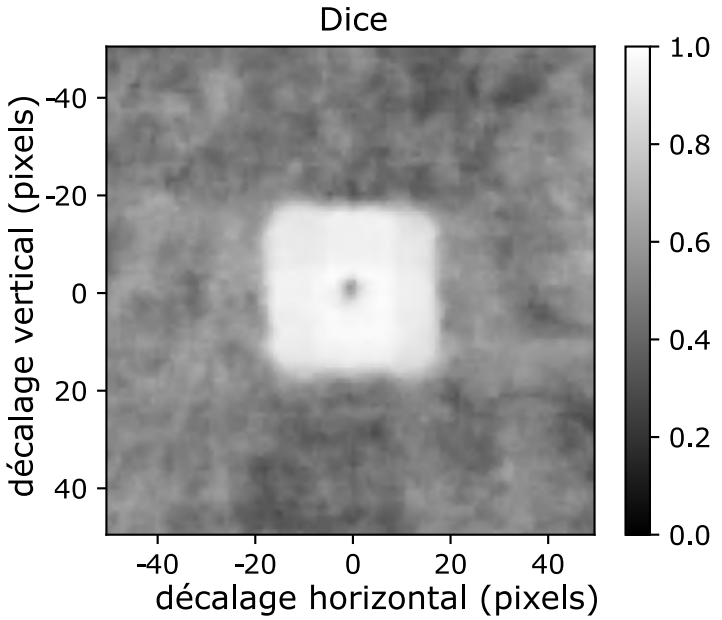


FIGURE 2.3 – Performance d'un réseau entraîné sur les données synthétiques en fonction du décalage. En abscisses, le décalage horizontal ; en ordonnée, le décalage vertical (en pixels). Plus un pixel de la carte est blanc, plus le réseau obtient de bonnes performances (mesurées par le coefficient de Dice). Le réseau à été entraîné avec des désalignements de 10 à 30 pixels. La tâche grise au milieu suggère que le réseau n'arrive pas à segmenter des images parfaitement alignées.

2.2.2 Données

La base de données utilisée dans ce chapitre contient 88 paires d'images IRM pondérées en T1 et T2, centrées sur le foie. Ces 88 paires d'images proviennent de 51 patients ayant tous des lésions hépatiques (majoritairement des métastases, avec quelques kystes, tumeurs primaires et angiomes). Toutes les images sont pré-recalées grossièrement en utilisant les méta-données de position enregistrées au moment de l'acquisition. Toutefois, le foie peut apparaître à des positions différentes dans les deux images, notamment à cause de la respiration (l'écart pouvant aller jusqu'à 15 centimètres environ). Chaque image est également ré-échantillonnée de manière à ce que qu'un voxel corresponde à 3mm verticalement, et 1,5mm dans les deux dimensions horizontales.

Les segmentations de référence sont obtenues par annotation manuelle, effectuée par une interne en radiologie en utilisant des outils interactifs 3D. Il est à noter que la segmentation manuelle du foie est une tâche fastidieuse et qu'en conséquence la précision des annotations n'est jamais parfaite. L'image en T2, à cause de sa faible résolution verticale, du faible contraste du foie et de sa plus grande propension à avoir des artefacts d'acquisition (voir la figure 2.5 est encore plus difficile à segmenter, ce qui rend les annotations de l'image en T2 moins précises encore.

La base de donnée est divisée en trois, en gardant 12 paires d'images pour l'ensemble de test, 6 paires pour la validation et 70 paires pour l'entraînement.

2.2.3 Méthodes

Paramètres communs

Pour toutes les stratégies et fonctions de coût comparées, on fixe les paramètres suivants. L'architecture du réseau est U-net 3D (Milletari, Navab et Ahmadi 2016 ; Ronneberger, Fischer et Brox 2015), initialisé avec les poids fournis par Z. Zhou

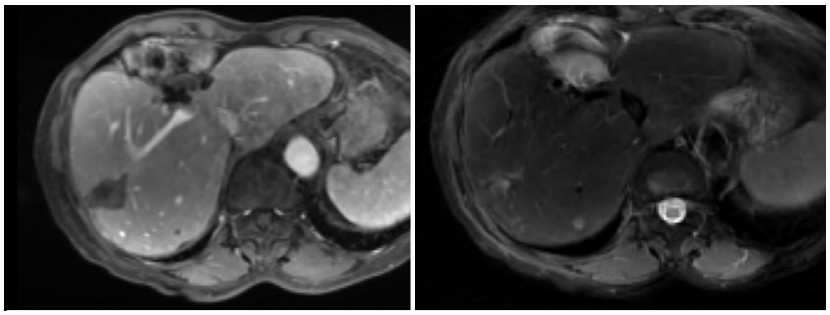


Figure 2.4 – Coupes axiales d'une paire d'images de la base de données. À gauche, l'image pondérée en T1, à droite, l'image pondérée en T2.

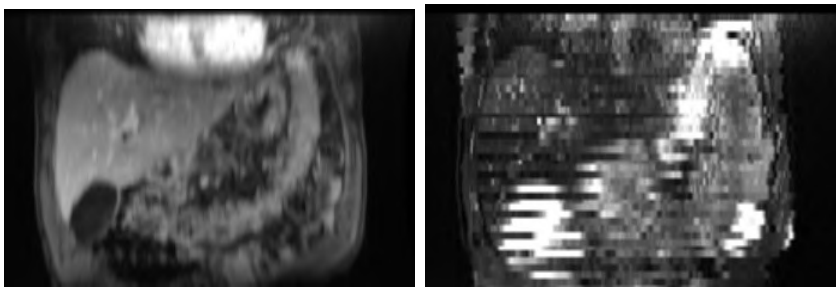


FIGURE 2.5 – Coupes coronales d'une paire d'images de la base pour illustrer la plus faible qualité des images en T2. À gauche : l'image en T1, à droite : l'image en T2.

et al. (2019). Ce choix d'architecture est maintenant standard en segmentation d'images médicales (ISENSEE, PETERSEN, KOHL et al. 2019). Pour l'entraînement j'utilise l'optimiseur Adam avec un pas de 1.10^{-4}, pendant 900 époques de 150 étapes, en sauvegardant le réseau à la fin de l'époque où l'on enregistre le meilleur score sur l'ensemble de validation.

Pour des raisons de mémoire, chaque étape d'optimisation n'utilise qu'une image (taille de *batchs* de 1), et chaque image est rognée aléatoirement.

Augmentation de données

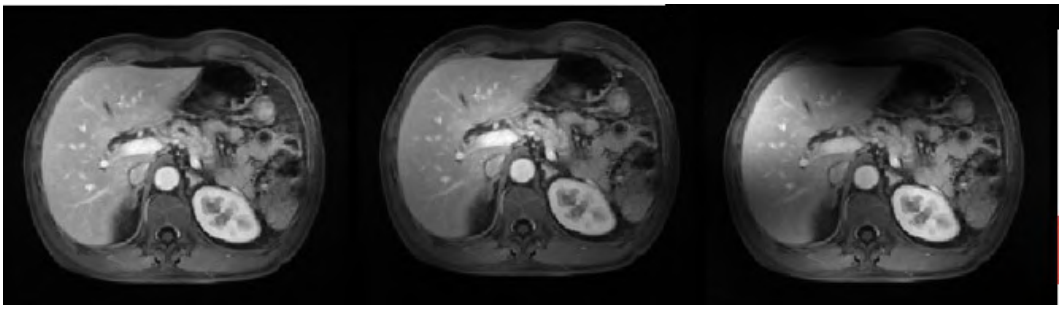


FIGURE 2.6 – Augmentation artificielle des données en générant et appliquant un champ de biais multiplicatif. À gauche, l'image originale, au milieu et à droite, après application du biais synthétique.

Pour toutes les stratégies entrée/sortie et fonctions de coût testées, on utilise la même stratégie d'augmentation artificielle des données, qui consiste à appliquer un champ multiplicatif généré aléatoirement à la volée et qui imite le champ de biais dû à l'hétérogénéité du champ magnétique, un signal de basse fréquence très lisse, qui altère les images et qu'on trouve souvent dans les images IRM. La figure 2.6 montre un exemple de l'application d'un tel champ. Mes expériences montrent qu'une telle augmentation,

au minimum, ne dégrade pas les performances.

Stratégies entrée/sortie

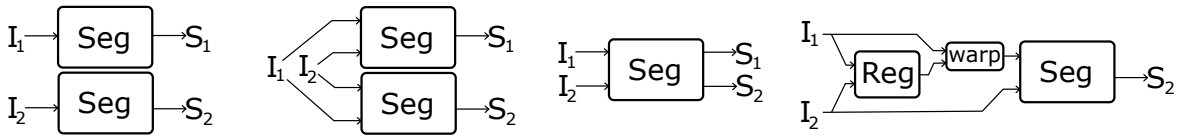


FIGURE 2.7 – Inférence d'une paire d'images pour les différentes stratégies d'apprentissage testées. De gauche à droite : simple entrée ; double entrée, sortie simple ; double entrée, sortie double ; double entrée pré-recalée, sortie simple.

Pour le premier axe de comparaison, on compare plusieurs contextes d'entrée/sorties qui sont illustrés sur la figure 2.7. Chacune de ces stratégies a deux versions, qu'on appelle *spécialisée* et *non-spécialisée*, que l'on précise dans la suite.

Simple entrée : le réseau n'a qu'un seul canal d'entrée et qu'un seul canal de sortie, de sorte qu'il ne peut prédire le masque que d'une image à la fois. Dans sa version *spécialisée*, on entraîne deux réseaux distincts, chacun des deux étant spécialisé sur une modalité. Dans sa version *non-spécialisée*, on entraîne un seul réseau, qui est entraîné indifféremment sur l'une ou l'autre des modalités.

Double entrée, sortie simple : le réseau prend en entrée les deux images de la paire, et prédit la segmentation d'une seule d'entre elles (celle qui est dans le premier canal d'entrée). Dans sa version *spécialisée*, on entraîne deux réseaux, l'un prenant dans son premier canal d'entrée l'image T1 et inversement pour l'autre. Dans sa version *non-spécialisée*, on entraîne un seul réseau, que l'on entraîne en choisissant aléatoirement quelle image on place dans le canal d'entrée.

Double entrée, sortie double : Le réseau prend en entrée les deux images de la paire, et sort les deux masques. Lorsqu'il est *spécialisé*, le premier canal d'entrée du réseau reçoit toujours l'image T1 tandis que le second reçoit l'image T2. Lorsqu'il est *non-spécialisé*, on l'entraîne en échangeant aléatoirement l'ordre de la paire.

Double entrée pré-recalée, sortie simple : C'est une variante de la stratégie « double entrée, sortie simple », qui vise à tester si déformer l'image T1 pour la recaler sur l'image T2 (qui est la modalité la plus difficile) permet d'aider le réseau à utiliser l'information contenue dans l'image T1 pour faire des prédictions plus précises sur l'image T2. Pour cela, on applique un algorithme de recalage non-linéaire à toute la base de données avant l'apprentissage.

Fonction de coût

Le deuxième axe de comparaison porte sur la fonction de coût utilisée pour l'apprentissage. Dans la suite, on considère une image $x \in X$, où $X = \mathbf{R}^{H \times L \times P}$ est l'espace des images de taille H (nombre de coupes), L (largeur en voxels), P (profondeur en voxels), et son masque annoté $y \in Y$ où $Y = [0,1]^{H \times L \times P}$ est l'espace des masques, ainsi qu'un réseau de segmentation $S : X \to Y$.

On compare trois fonctions voxel à voxel simples, c'est-à-dire qu'elles n'utilisent pas de paramètre appris :

Entropie croisée binaire : c'est la fonction de coût classiquement utilisée pour les tâches de classifications binaires où les modèles prédisent une probabilité, et notamment par Ronneberger, Fischer et Brox (2015).

$$L_{ecb}(x, y) = -\sum_{i,j,k} y_{i,j,k} \log(S(x)_{i,j,k})$$

où $S(x)_{i,j,k}$ désigne la sortie du réseau pour l'entrée x au point de coordonnées (i, j, k), et $y_{i,j,k}$ désigne la valeur du masque y au même point.

Dice : Sudre et al. (2017) ont proposé d'utiliser le coefficient de Dice, souvent utilisé comme métrique de performance en segmentation, comme fonction de coût à optimiser. La normalisation permet d'obtenir de meilleures performances en cas de fort déséquilibre entre les classes. En segmentation, cela correspond aux cas où les objets à segmenter sont petits par rapport à l'image.

$$L_{Dice}(x, y) = \frac{2 \sum_{i,j,k} y_{i,j,k} S(x)_{i,j,k}}{\sum_{i,j,k} y_{i,j,k} \sum_{i,j,k} S(x)_{i,j,k}}$$

Entropie croisée binaire + Dice : pour avoir un compromis entre les deux approches.

$$L_{\Sigma} = \lambda_{ecb} L_{ecb} + \lambda_{Dice} L_{Dice}$$

Pour cette expérience je fixe $\lambda_{ecb} = \lambda_{Dice} = 1$ pour garder un compromis équitable.

On compare également trois fonctions de coût *adversaires*, c'est-à-dire qui nécessitent l'entraînement d'un autre réseau $D : Y \to [0, 1]$, appelé discriminateur, dont la tâche est d'apprendre à reconnaître les masques issus des annotations. L'idée sous-jacente est qu'en apprenant la distribution qui régit les masques annotés, on peut obtenir des modèles qui sont robustes à des prédictions improbables (par exemple des prédictions avec des trous, ou des formes étranges), et par conséquent plus précis en général.

La technologie de l'apprentissage adversaire a récemment connu un gain de popularité important en segmentation (voir par exemple Hung et al. 2018 ; Yizhe Zhang et al. 2017). Pour une revue plus complète de comment cette technologie (au départ proposée pour la génération d'images naturelles) est utilisée en segmentation, voir la section de revue de littérature proposée par Samson et al. (2019).

Fonction de coût GAN de base, utilisée par exemple par Luc et al. (2016) :

$$L_{GAN}(x, y) = L_{ecb}(x, y) - \log(D(S(x)))$$

Fonction de coût Embedded, proposée par Ghafoorian et al. (2018) pour stabiliser l'entraînement :

$$L_{EL}(x, y) = L_{ecb}(x, y) - ||D_k(S(x)) - D_k(y)||_2^2$$

où D_k représente la k-ième couche du réseau D (k est un paramètre).

Fonction de coût parieuse, proposée par Samson et al. (2019). A la place du discriminateur D, on entraîne un réseau *parieur* $G : X \times Y \to Y$ (pour *gambler*) qui prend en entrée une image et une segmentation, et qui sort une carte de pari. Un voxel à la position i, j, k de la carte de pari $G(x, y)$ correspond à une estimation de la probabilité que $y_{i,j,k}$ soit faux. La dernière couche de sortie de G est normalisée, de manière à ce que le budget alloué au parieur soit limité, et qu'il doive placer judicieusement ses paris en apprenant à détecter les erreurs de segmentation. Le réseau parieur est entraîné en minimisant L_G :

$$L_G(x, y) = \sum_{i,j,k} G(x, S(x))_{i,j,k} y_{i,j,k} \log(S(x)_{i,j,k})$$

La fonction de coût parieuse pour le segmenteur est

$$L_S = L_{ecb} - L_G$$

2.2.4 Résultats

Stratégie multimodale

La performance de chacune des stratégies est évaluée en calculant le score de Dice moyen sur la base de test. D'autres métriques, comme la distance d'Hausdorff, donnent une comparaison des approches équivalente.

Le caractère stochastique de l'algorithme d'optimisation, ainsi que l'aléa de l'initialisation des poids du réseaux, impliquent que l'on puisse mesurer des différences de performances entre plusieurs exécutions d'une même expérience. Pour connaître l'ordre

	Non-spécialisée			Spécialisée		
Stratégie	T1	T2	p	T1	T2	p
Simple entrée	0,961	0,938	-	0,959	0,929	0,4
Double entrée, sortie simple	0,955	0,932	0,004	0,956	0,930	0,012
Double entrée, sortie double	0,938	0,907	0,0009	0,942	0,897	0,0008
Entrée pré-recalée, sortie simple	-	-	-	-	0,925	0,007

TABLE 2.1 – Score de Dice moyen en fonction de la stratégie d'entrée/sortie.

	Non-spécialisée.		Spécialisée	
Stratégie	T1	T2	T1	T2
Double entrée, sortie simple	<0,001	0,016	0,001	0,030
Double entrée, sortie double	0,042	0,083	0,023	0,103
Entrée pré-recalée, sortie simple	-	-	-	0,054

TABLE 2.2 – Perte de Dice lorsque la modalité auxiliaire ne correspond pas, en fonction de la stratégie d'entrée/sortie.

de grandeur de ces différences, et ainsi avoir une meilleure idée de la différence de performance qui suggère un effet réel de la stratégie, on lance trois exécutions d'une validation croisée à quatre blocs sur toute la base avec la stratégie « simple entrée ». En moyenne, les scores de Dice mesurés sur les différentes exécutions ne diffèrent pas de plus de 0.005.

Les résultats moyens de chaque stratégie sont rapportés dans le tableau 2.1. La stratégie en simple entrée non-spécialisée obtient les meilleurs scores. Pour chacune des stratégies, on effectue un test statistique des rangs signés de Wilcoxon (WILCOXON 1945), qui est indiqué dans le cas de données appariées sur lesquelles on ne peut pas faire d'hypothèse de normalité, ce qui est le cas des scores de Dice. L'hypothèse nulle que l'on teste est que la médiane des scores des 24 images de la base de test obtenue par une stratégie est identique à celle obtenue par la stratégie en simple entrée non-spécialisée. Plus cette hypothèse apparaît improbable (p faible dans le tableau 2.1), plus la différence de performance avec la stratégie la plus performante apparaîtra réelle.

Ainsi, la stratégie simple entrée spécialisée ne montre pas de différence significative de performance par rapport à son pendant spécialisé (la différence de score en T2 étant majoritairement expliqué par une mauvaise performance sur le cas difficile de la base de test, *cf infra*). Les stratégies en double entrée et sortie simple montrent une perte de performance probablement réelle ($p < 0.05$) mais faible. On ne note en tout cas pas

d'amélioration de la performance lorsqu'on ajoute une modalité auxiliaire en entrée, et ce même avec un pré-recalage. La perte de performance des stratégies en sorties doubles est en revanche non seulement significative, mais également forte (jusqu'à 4 points de Dice en T2). Ces stratégies sont un cas particuliers d'apprentissage multi-tâches. Or, comme montré par Wu, H. R. Zhang et Ré (2020), il n'est pas trivial de mettre en place des stratégies reposant sur l'apprentissage multi-tâches, et l'approche naïve que j'ai essayée ici a rapidement montré ses limites. C'est potentiellement particulièrement difficile à utiliser en segmentation 3D, où augmenter la capacité des réseaux est rapidement coûteux en termes de mémoire. La méthode que je propose dans la section 2.3 se base sur une stratégie multi-tâches, à laquelle on rajoute une régularisation pour aider l'apprentissage.

L'idée des stratégies à double-entrée est d'étudier si un réseau peut bénéficier de l'information contenue dans l'image auxiliaire, c'est-à-dire l'image de la paire qu'on ne cherche pas à segmenter. Par exemple si un réseau doit segmenter le foie dans l'image T2, peut-il utiliser l'image T1 pour augmenter sa précision ? Un moyen simple d'estimer à quel point un réseau utilise l'information de l'image auxiliaire et de mesurer la chute de performance que l'on observe lorsqu'on remplace l'image auxiliaire de chaque paire par une image de même modalité mais provenant d'une paire différente. Ainsi, plus un réseau aura appris à utiliser la modalité auxiliaire pour prendre sa décision, plus on observera une chute de performance lorsque l'image auxiliaire ne correspond pas. Inversement si l'on n'observe aucune baisse du score de Dice, on pourra conclure que le réseau n'utilise pas la modalité auxiliaire. Le tableau 2.2 montre les résultats de cette expérience. On remarque que l'utilisation de la modalité auxiliaire est négligeable dans le cas des stratégies « sortie simple » pour l'image T1, et plus forte pour l'image T2 (ce qui est cohérent avec l'idée selon laquelle l'image T2 est plus difficile à segmenter, et qu'ainsi le réseau apprenne à s'aider de l'image T1), surtout après un pré-recalage, ce qui est également attendu puisque l'information paraît plus facile à utiliser lorsque les images sont recalées. Les stratégies avec « double sortie » montrent la plus grande dépendance à la sortie auxiliaire.

La question est maintenant de savoir si l'utilisation d'information de l'image auxiliaire peut permettre de meilleures performances, ce qui était notre hypothèse de départ. La comparaison des tableaux 2.2 et 2.1 suggère précisément le contraire : une utilisation accrue de l'image auxiliaire semble liée à une perte de performance. Une explication peut être que donner les deux images en entrée offre au réseau plus de chance de surapprendre : cette utilisation accrue de l'image auxiliaire pourrait suggérer qu'il a appris par cœur certains exemples de la base d'entraînement en utilisant l'image auxiliaire, ce qui peut lui faire perdre de la précision sur les données de test.

Fonction de coût

Le tableau 2.3 montre les performances des réseaux entraînés avec différentes fonctions de coût. On ne trouve pas d'effet clair de ce paramètre sur la performance, et les fonctions de coût adversaires ne semblent pas apporter de bénéfices.

La figure 2.8 montre la carte de paris prédite par un réseau parieur. On remarque que ce réseau parie sur des erreurs à proximité de la bordure du foie, ce qui est cohérent avec l'intuition selon laquelle, comme les annotations ne sont pas parfaitement précises, il est probable de trouver des incohérences entre la segmentation prédite et la segmentation annotée au bord du foie. On peut donc interpréter ces cartes comme une estimation de l'incertitude sur le résultat.

FIGURE 2.8 – Carte de paris prédite par le réseau parieur. Chacune des images correspond à la coupe centrale selon les trois dimensions. En blanc, les voxels pour lesquels la segmentation est le plus probablement fausse, et noire où elle est vraie. En vert la segmentation de référence, en rouge le masque prédit par le segmenteur.

Le résultat de cette expérience corrobore l'idée, exprimée notammant par ISENSEE, PETERSEN, KOHL et al. (2019) et HOFMANNINGER et al. (2020), que l'amélioration des performances d'un réseau passe avant tout par l'attention portée à la variété des données et la qualité de l'annotation, en tout cas avant la recherche d'innovations méthodologiques, dont les avantages sont souvent difficiles à répliquer sur d'autres jeux de données. Notons également que la division de la base de donnée choisie pour comparer les stratégies était particulièrement difficile, comme l'a montré l'expérience de validation croisée puisque l'on a obtenu un score de Dice de 0.983 sur une des divisions.

Inspection visuelle

Les figures 2.9 et 2.10 montrent quelques exemples de prédictions des réseaux « simple entrée non-spécialisé » et « double sortie spécialisé » respectivement. On peut voir que les prédictions des deux réseaux restent précises, même en présence d'une charge tumorale

Fonction de coût	T1	T2	p
Entropie croisée binaire	0,961	0,938	-
Dice	0,959	0,931	0,42
$L_{Dice} + L_{ecb}$	0,959	0,932	0,74
GAN standard	0,950	0,930	0,00011
Embedding	0,960	0,935	0,71
Parieur	0,959	0930	0,134

TABLE 2.3 – Score de Dice moyen en fonction de la fonction de coût utilisée.

importante (comme en bas à gauche), ou lorsque de grosses lésions sont présentes près du bord (colonne du milieu, en bas). Toutefois, on remarque que le réseau « simple entrée » est bien plus précis sur l'image en T2 en général (notamment sur les cas de la colonne de gauche).

La colonne de droite montre deux cas particulièrement difficiles : en haut, il s'agit du seul cas de la base où le patient a subit une hépatectomie, ce qui rend l'anatomie du foie atypique. Malgré cette difficulté les deux réseaux parviennent à faire une prédiction correcte, surtout pour l'image en T1. En bas, le foie montre une charge tumorale très importante. Le réseau « simple entrée » fait une estimation précise des contours du foie malgré cette difficulté, contrairement à celle du réseau « double sortie ».

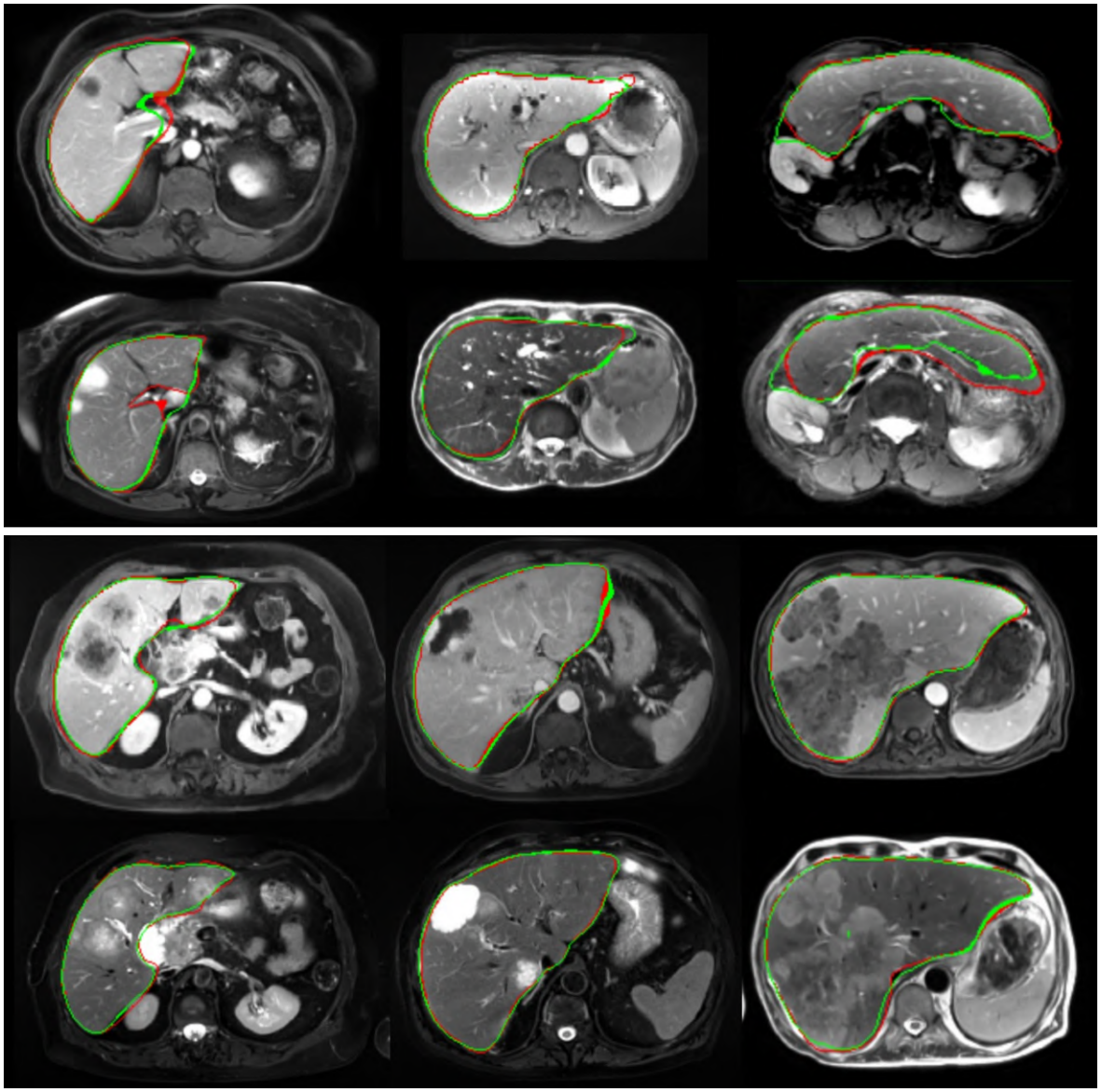


FIGURE 2.9 – Résultat du modèle « simple entrée, non spécialisé » sur 6 paires d'images de la base de test. Pour chaque paire, l'image en T1 est positionnée au dessus de l'image en T2. Le contour vert correspond à la prédiction du réseau, et le contour rouge à la segmentation de référence.

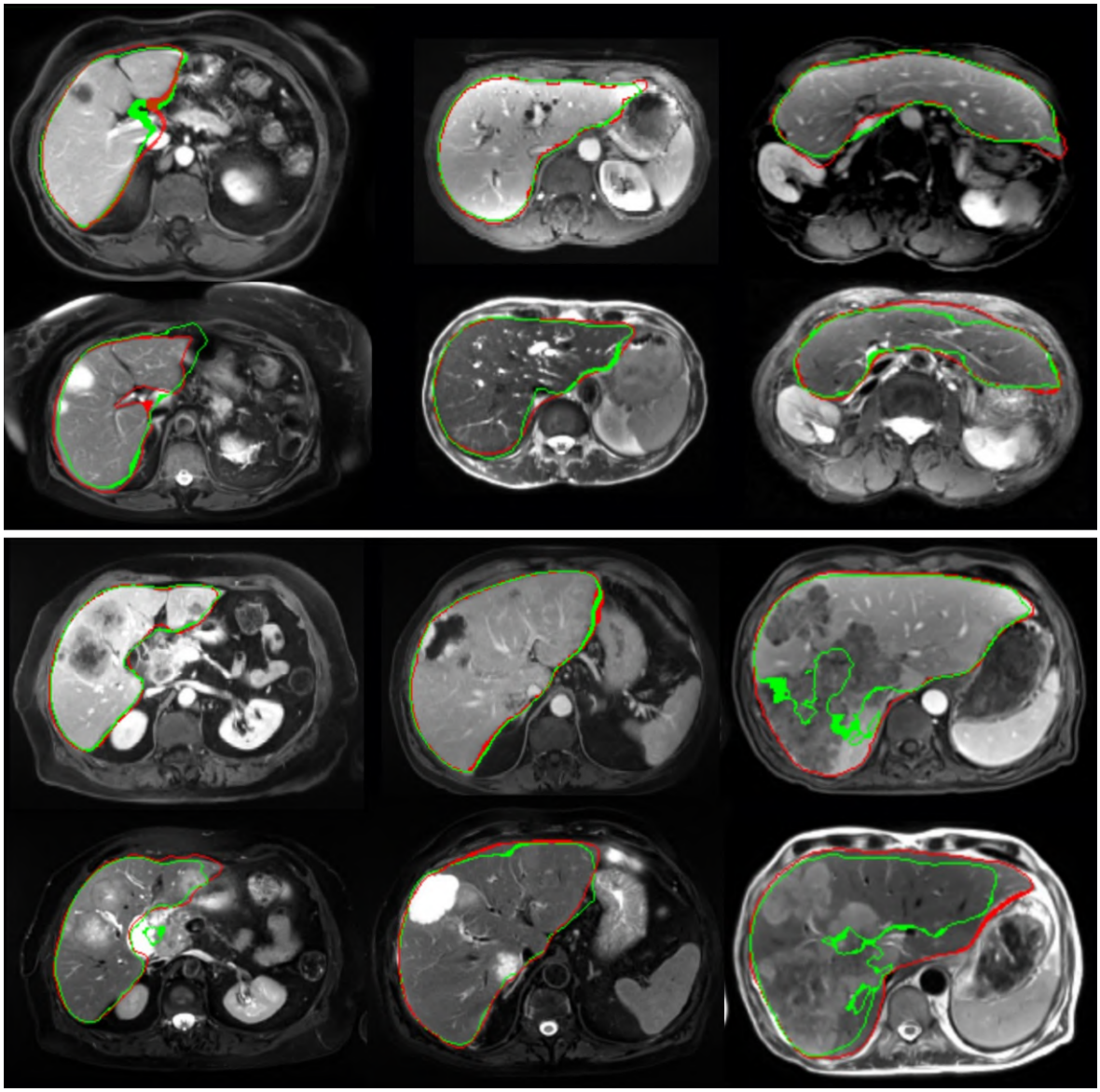


FIGURE 2.10 – Résultat du modèle « double sortie, spécialisé » sur 6 paires d'images de la base de test. Pour chaque paire, l'image en T1 est positionnée au dessus de l'image en T2. Le contour vert correspond à la prédiction du réseau, et le contour rouge à la segmentation de référence.

2.3 Intégration d'une contrainte de similarité

L'imparfaite précision des annotations, particulièrement pour les images T2 et dont nous avons discuté à la section précédente, impose un plafond sur la performance des réseaux si on la définit comme l'écart entre les masques prédits et les masques annotés.

Un axe d'amélioration auquel on peut alors penser dans le cas de la segmentation de paires d'images est celui de la *similarité*. En effet, les deux images d'une même paire étant acquises à quelques minutes d'intervalle seulement, on s'attend à ce que les masques de segmentation du foie dans le deux images soient identiques à une déformation lisse près, induite par les mouvements du patient ou sa respiration. Dans la suite, on dit que deux masques de segmentation sont *similaires à une classe de déformations près* s'il existe une déformation de cette classe qui transforme un masque en l'autre. Le but est alors, étant donnée une classe de déformations qu'on jugera acceptable, de contraindre l'apprentissage pour favoriser les prédictions similaires entre elles, à cette classe de déformations près. Cette classe dépend de notre connaissance *a priori* sur l'objet à segmenter : si l'on veut segmenter un os par exemple, on s'attend à ce que les masques des deux images ne diffèrent que par une transformation rigide ; pour le foie, constitué de tissus mous, la déformation peut être élastique.

Cette approche a plusieurs buts. Premièrement, des masques plus similaires pourront donner des mesures quantitatives plus cohérentes entre les modalités. Ensuite, elle peut permettre de limiter l'effet des biais dans les annotations spécifiques à une modalité (par exemple, la faible résolution en z des images pondérées en T2 peut causer des imprécisions dans l'annotation). Enfin, dans le cas où l'organe à segmenter n'est pas également facile à segmenter dans les deux modalités (comme c'est le cas lorsque qu'une modalité est anatomique et l'autre fonctionnelle, et comme c'est le cas pour l'application qui nous intéresse), on veut apprendre au réseau à chercher l'information utile dans la modalité facile pour segmenter la plus difficile, ce qu'il ne fait pas spontanément comme on l'a vu dans la section précédente.

Cette idée s'inscrit dans l'approche, évoquée dans la section 2.1.1, qui consiste à intégrer de l'information anatomique à l'apprentissage des réseaux de segmentation par le biais d'un terme de régularisation à la fonction de coût. La méthode que je propose dans cette section consiste à simultanément entraîner un réseau *segmenteur*, qui prend en entrée les deux images et prédit les deux masques, et un réseau *recaleur*, qui prend en entrée les deux masques prédits et estime les paramètres de la transformation qui les sépare. Les deux réseaux coopèrent pour minimiser l'erreur de segmentation et maximiser la similarité entre les deux masques prédits (voir figure 2.11). L'objectif

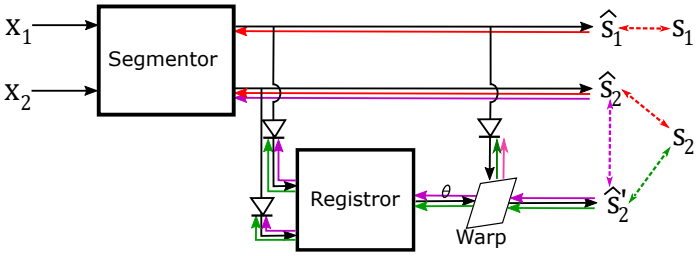


FIGURE 2.11 – Schéma de la méthode pour appliquer une contrainte de similarité. Les flèches pointillées représentent les fonctions de coût minimisées, les flèches pleines de couleur représentent les gradients correspondant à la fonction de coût de la même couleur. Les diodes représentent l'opération d'arrêt des gradients.

est de créer une boucle de rétroaction positive entre le segmenteur et le recaleur pour améliorer la similarité : au fur et à mesure que le segmenteur prédit des paires de masques plus similaires, le recaleur devient alors capable d'améliorer la qualité de ses prédictions, ce qui en retour permet d'affiner les prédictions du recaleur, en apprenant à chercher l'information pertinente dans les deux images.

2.3.1 Methode proposée

On considère une base de données de paires d'images que l'on note (x_1, x_2), avec leurs segmentations annotées correspondantes (s_1, s_2). On entraîne un segmenteur S et on note les masques estimés $(\hat{s_1}, \hat{s_2}) = S(x_1, x_2)$ On entraîne simultanément un recaleur R tel que $\theta = R(\hat{s_1}, \hat{s_2}) \in \Omega$, où le paramètre Ω est l'ensemble des transformations acceptables pour l'objet à segmenter. Par exemple, si l'on veut segmenter des objets rigides (comme un os), on peut choisir Ω comme étant l'ensemble des transformations rigides. On note $\hat{s_2}' = \theta(\hat{s_1})$ le masque prédit de la première modalité auquel on a appliqué la transformation prédite par R. On veut que $\hat{s_2}$ et $\hat{s_2}'$ soient proches de s_2, et que $\hat{s_1}$ soit proche de s_1.

Comme les images de la paire proviennent de différentes modalités, on ne peut pas directement comparer les intensités des voxels, ce qui implique qu'on ne peut pas directement optimiser une fonction de coût basée sur les images pour entraîner le recaleur, comme d'autres méthodes d'apprentissage joint d'une tâche de recalage et de segmentation le proposent (BELJAARDS et al. 2020 ; ELMAHDY et al. 2019).

Différentes stratégies ont été proposées pour outrepasser cette difficulté inhérente au recalage multi-modal : Z. XU et al. (2020) utilisent des descripteurs invariants par modalité, F. LIU et al. (2020) apprennent en même temps une tâche de traduction d'image pour passer d'une modalité à l'autre, et CHARTSIAS et al. (2019) apprennent une représentation commune aux deux images. La méthode que je présente ici s'appuie

uniquement sur le recalage des masques prédits et non des images, ce qui simplifie nettement la tâche de recalage.

Fonctions de coût

On définit trois fonctions de coût, représentées par la flèches en couleur pointillées sur la figure 2.11.

Erreur de segmentation : (en rouge sur la figure Figure 2.11)

$$L_r(s_1, s_2, \hat{s_1}, \hat{s_2}) = L_{ecb}(s_1, \hat{s_1}) + L_{ecb}(s_2, \hat{s_2})$$

où L_{ecb} représente l'entropie croisée binaire ($L_{ecb}(x, y) = -\sum_i y_i \log(x_i)$).

Erreur de recalage : (en vert sur la figure 2.11)

$$L_g = L_{mse}(\hat{s_2}' * f, s_2 * f)$$

où L_{mse} est la fonction d'erreur quadratique moyenne, et f est un filtre passe-bas. On floute les masques avant d'appliquer L_{mse} dans le but d'adoucir les contours des masques, évitant ainsi les discontinuités, et d'obtenir des gradients plus fiables par rapport à cette fonction de coût. Cette fonction de coût sert à entraîner le recaleur.

Erreur de similarité : (en rose sur la figure 2.11)

$$L_p = L_{mse}(\hat{s_2}', \hat{s_2})$$

Les deux réseaux sont entraînés pour minimiser

$$L = \lambda_r L_r + \lambda_g L_g + \lambda_p L_p$$

Les opérateurs d'arrêt de gradients (représentés par le symbole d'une diode sur la figure 2.11) sont placés de manière à ce que L_g n'influence pas le segmenteur. Minimiser L pour le segmenteur revient donc à minimiser $L_s = L_r + \lambda_p L_p$, en fixant $\lambda_r = 1$. Le terme $\lambda_p L_p$ peut ainsi être considéré comme un terme de régularisation, et L_r comme le terme d'attache aux données.

Il est important de noter que la fonction de coût L_p ne conditionne que le deuxième canal (voir figure 2.11), en le contraignant à être similaire au premier. Cette asymétrie entre les deux images d'entrée est justifiée si l'on considère, comme pour l'application qui nous intéresse, qu'une des deux modalité (que l'on met dans le deuxième canal) est plus difficile que l'autre, et par conséquent que ce soit elle qui bénéficie de la régularisation.

Entraînement

L'entraînement se fait en trois étapes. La première étape consiste à pré-entraîner le segmenteur, en ne minimisant que L_r, jusqu'à convergence. La deuxième étape est un pré-entraînement du recaleur, en minimisant L_g, en donnant en entrée du réseau les masques annotés. Enfin on entraîne l'ensemble en minimisant L.

Test

Pour évaluer l'effet de notre méthode sur la similarité des masques prédits, on définit la métrique de similarité des masques s_1 et s_2, à une déformation τ de Ω près.

$$D_S(s_1, s_2, \Omega) = \max_{\tau \in \Omega} Dice(\tau(s_1), s_2)$$

où $Dice(y_1, y_2)$ est le score Dice de deux masques binaires y_1 et y_2.

Pour calculer cette similarité relative lorsque Ω est l'ensemble des déformations denses paramétrées par un champ de déformation, on en calcule une approximation en calculant le champ de déformation optimal par une descente de gradient, selon le schéma itératif suivant :

$$\tau_0 = I_d$$

$$\tau_{k+1} = \tau_k - \alpha \frac{\partial}{\partial \tau_k} ||\tau_k(s_1 * f) - (s_2 * f)||^2$$

où f est un filtre passe-bas (comme pour la définition de L_g) qui sert à convexifier le problème (voir section 2.3.2) et α est le pas de la descente de gradient.

2.3.2 Choix du filtre : expérience sur les gradients

Le filtre f présent dans la définition de L_g et utilisé pour l'approximation de la similarité relative a pour but de fournir des gradients utiles par rapport au champ de déformation.

Pour préciser, considérons deux masques binaires s_1 et s_2. Le problème de recalage, qui consiste à trouver la transformation $\tau \in \Omega$ qui minimise $L(\tau(s_1), s_2)$, où la classe de transformation Ω est un paramètre du problème et où L est une certaine fonction de distance, n'a aucune raison d'être convexe. Cela implique que les gradients $\frac{\partial}{\partial \tau} L(\tau(s_1), s_2)$ n'ont aucune raison de pointer vers le minimum global du problème, ni même dans une direction un tant soit peu intéressante. Or, la minimisation de la fonction de coût L_g ainsi que la procédure d'optimisation pour approximer la similarité relative se basent sur ces gradients.

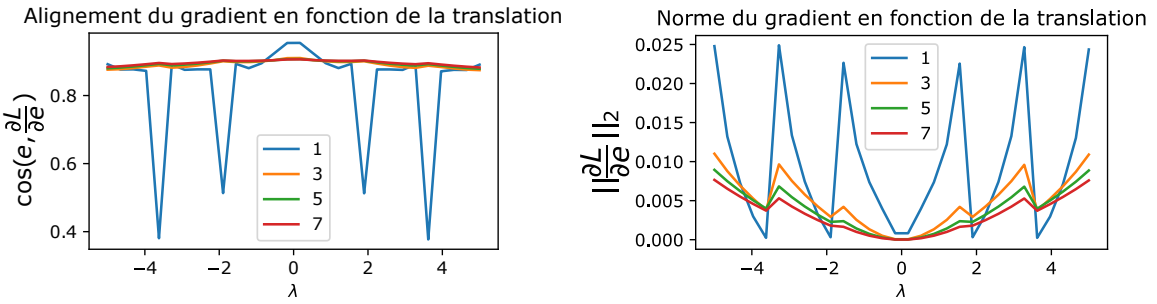


FIGURE 2.12 – Gradients de l'erreur quadratique moyenne de s_1 et de sa translatée d'amplitude λ, par rapport au vecteur de translation. À gauche, son alignement par rapport au vecteur de translation en fonction de λ (une valeur proche de 1 montre un alignement). À droite, sa norme en fonction de λ. Les différentes couleurs correspondent à plusieurs tailles de filtres (un filtre de taille 1 équivaut à une absence de filtre)

Le but de l'expérience décrite dans cette section est d'étudier, dans un contexte où l'on connaît le minimum global, si les gradients sont orientés en direction de ce minimum global, ainsi que l'influence sur l'orientation des gradients d'un filtrage passe-bas des masques.

Ce contexte est le suivant : s_1 est un masque issu des annotations de la base de données, et s_2 est la translatée de s_1 par un vecteur λe, où e est un vecteur unitaire et de direction arbitraire, et λ est un réel qui correspond à l'amplitude de la translation. Dans ces conditions on sait donc que le recalage optimal est la translation de vecteur $-\lambda e$. On peut donc calculer les gradients de la distance (pour cette expérience on choisit l'erreur quadratique moyenne) entre s_1 et s_2 par rapport au vecteur e.

Les courbes bleues de la figure 2.12 montrent l'alignement du gradient par rapport à e et son amplitude, en fonction de λ. On remarque que pour certaines valeurs de lambda, les gradients sont à la fois mal alignés et de forte amplitude. On remarque d'ailleurs qu'un recalage par descente de gradients ne converge pas vers le minimum global.

Un moyen simple d'atténuer ce phénomène consiste à flouter les deux filtres à l'aide d'un filtre passe pas. Pour cette expériences je teste 3 filtres binomiaux séparables de longueur 3, 5 et 7. L'alignement et l'amplitude des gradients obtenus après filtrage des masques est représenté par les courbes orange, vertes et rouges sur la figure 2.12. On constate que le filtrage permet de se débarrasser des gradients mal alignés, quelle que soit la taille du filtre. Je choisis le filtre binomial séparable de longeur 3 dans la suite, pour des raisons de rapidité de calcul.

2.3.3 Expérience sur des données synthétiques

Cette expérience a pour but de s'assurer que la méthode fonctionne correctement dans un contexte simplifié en 2D. Comme pour l'expérience décrite dans la section 2.2.1, on génère un masque binaire avec des formes aléatoires. Pour cette expérience un génère un autre masque légèrement différent qui sera utilisé pour l'autre image. Les deux masques sont translatés l'un par rapport à l'autre (les deux masques sont visibles sur la figure 2.14a). Pour simuler les deux modalités, l'une des deux images reçoit des sinusoïdes de fréquence fixe avec les parties positive et négative différenciées par l'angle, et l'inverse pour l'autre image (voir figure 2.13).

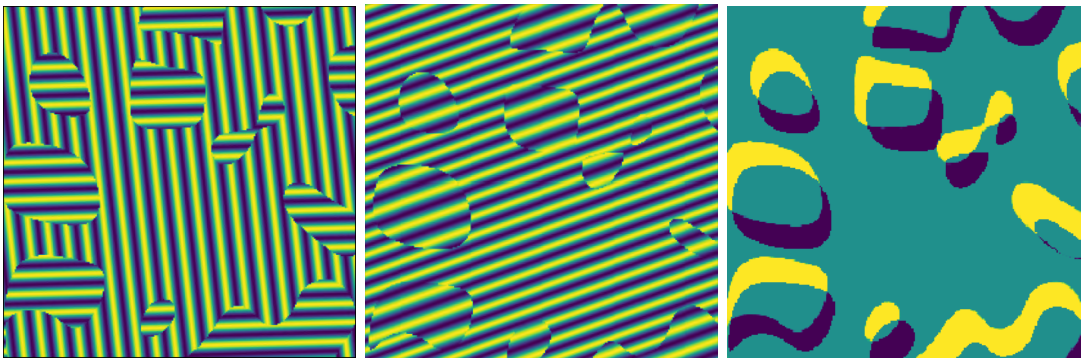
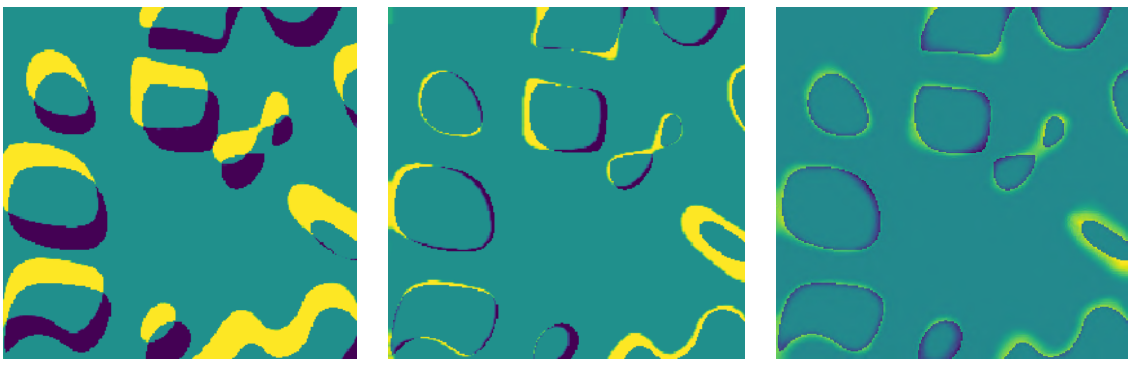


FIGURE 2.13 – Une paire d'images de l'expérience sur les données synthétiques. L'image de gauche est codée en angle, celle du milieu en fréquence. Les masques correspondant sont montrés sur l'image de droite (en jaune : masque de l'image 1, en bleu foncé : masque de l'image 2, en cyan là où ils coïncident).



(a) (b) (c)

FIGURE 2.14 – Différence entre les sorties du réseaux et s_2. Les masques coïncident sur les parties en cyan, les parties jaunes correspondent au premier masque seulement et les parties bleu marine au deuxième. (a) $\hat{s_1} - s_2$; (b) $\hat{s_2}' - s_2 = \theta(\hat{s_1}) - s_2$; (c) $\hat{s_2} - s_2$.

On entraîne les réseaux segmenteur et recaleur en générant à volée de telles paires d'images, en choisissant Ω comme étant l'ensemble des translations (R régresse donc deux paramètres). La figure 2.14 montre la différence entre les trois sorties du réseau ($\hat{s_1}$, $\hat{s_2}'$ et $\hat{s_2}$) et s_2, en lui donnant en entrée les images de la figure 2.13. La figure 2.14b permet de mettre en évidence la différence entre les masques, après recalage. 2.14c montre la sortie du deuxième canal du réseau segmenteur. On constate que celle-ci est proche de la translatée de s_1, et que le segmenteur a donc bien appris à chercher l'information du premier canal en le recalant sur le second, ce qui est le comportement attendu.

Pour mettre en évidence l'effet de la régularisation sur la similarité relative, on entraîne 30 réseaux avec différentes valeurs de λ_p. Pour chaque réseau, on choisit $\lambda_r = 1 - \lambda_p$, et $\lambda_g = 1$. Pour chaque réseau entraîné, et chaque élément d'un ensemble de paires d'images, on calcule la similarité des prédictions relativement aux translations, et on la compare à la similarité relative des masques de la segmentation de référence s_2.

La figure 2.15 montre les résultats de cette expérience. On constate un net gain de similarité (à une translation près) lorsque les réseaux sont entraînés avec $\lambda_p (= 1 - \lambda_r) < 0,5$

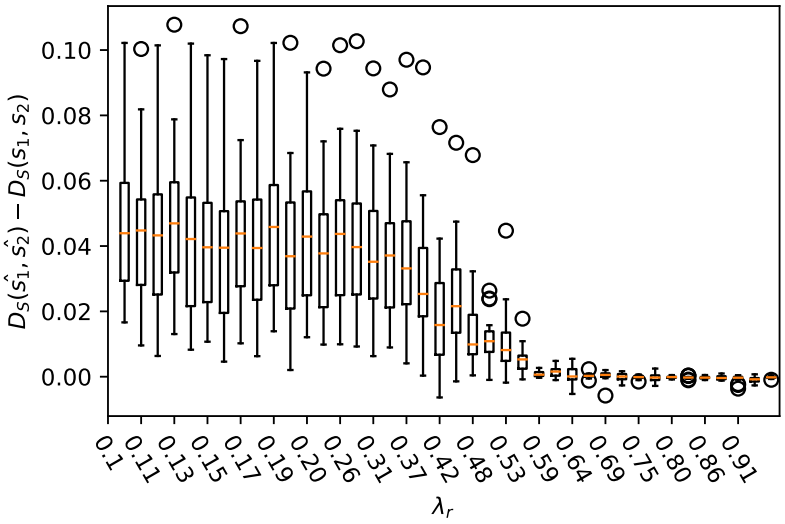


FIGURE 2.15 – Distribution de la différence de similarité relative entre les masques prédits et la vérité terrain, pour chaque réseau entraîné avec un λ_r différent. Des valeurs positives indiquent un gain de similarité par rapport aux masques de la vérité terrain.

2.3.4 Expérience sur les données réelles

Paramètres

Pour le segmenteur, on utilise comme à la section 2.2 un U-net 3D, pré-entraîné avec les poids fournis par Z. Zhou et al. (2019). On choisit Ω comme étant un ensemble de déformations lisses, paramétrées par un champ de vecteurs de déplacements définis sur une grille de basse résolution, qu'on ramène à la résolution de l'image par une interpolation tri-linéaire. C'est cette basse résolution qui impose la régularité du champ de déformation.

Pour estimer le champ de déplacement, on choisit pour le recaleur une architecture totalement convolutionnelle, qui sous-échantillonne l'entrée par un facteur 16 : 4 blocs de deux couches convolutionnelles de noyau $3 \times 3 \times 3$ d'activation linéaire rectifiée, suivies d'une couche de *max-pooling* de taille $2 \times 2 \times 2$. Les couches convolutionnelles de chaque bloc ont 16, 32, 64 et 128 filtres respectivement. La couche de sortie est une couche convolutionnelle de taille $1 \times 1 \times 1$ avec une activation linéaire.

Ce champ de déformation est ensuite rééchantillonné avant de passer dans la couche de déformation (désignée par *Warp* sur la figure 2.11 à la page 37). Ce rééchantillonnage ainsi que la couche de déformation utilisent une interpolation tri-linéaire. On fixe $\lambda_r = 0,1$, $\lambda_p = 1$ et $\lambda_g = 1$

On utilise la même division de la base de données pour l'entraînement qu'à la section 2.2, avec la même stratégie d'augmentation artificielle de données et de coupes aléatoires, et on entraîne pendant 1500 époques de 100 étapes d'optimisation.

Résultats

Sans recalage, le coefficient de Dice des paires de masques issus des annotations est en moyenne de 0,751. En calculant la similarité à la classe de déformations Ω (telle que définie au paragraphe précédent) près, on trouve une valeur de $0,955$.

Le réseau « double entrée, double sortie » sans recaleur produit des masques de similarité à Ω près de 0,954, tandis que le réseau simple entrée montre une similarité de $0,959$. Notre méthode produit des masques ayant une similarité à Ω près de $0,966$.

On compare les similarités obtenues avec notre méthode et les similarités des masques annotés avec un test des rangs signés de Wilcoxon, qui donne $p = 0,0028$. Cela tend à indiquer que cette différence de similarité ne peut probablement pas être expliquée uniquement par le bruit statistique. En comparaison, le test donne $p = 0,08$ pour la différence de similarité entre le réseau en simple entrée sans régularisation et les annotations.

Quant aux performances, on mesure un Dice des prédictions par rapport aux annotations de $0,946$ pour les images T1, et $0,918$ pour les images T2. Si ces performances n'égalent pas la stratégie « simple entrée non spécialisée » (voir section 2.2.4), on constate cependant que l'ajout de la coopération avec un réseau recaleur permet d'améliorer sensiblement les performances d'un réseau « double entrée, double sortie ». En guidant l'apprentissage multi-tâche, le réseau apprend donc plus efficacement à chercher l'information pertinente de la modalité auxiliaire.

La figure 2.16 montre des résultats de segmentation pour 6 paires d'images de la base de test. Pour comparer avec les stratégies étudiées à la section 2.2, ce sont les mêmes paires que sur les figures 2.9 et 2.10 aux pages 34 et 35.

On constate que les prédictions se rapprochent en qualité de celles du réseau « simple entrée » sur les cas moins difficiles (colonne de gauche et colonne du milieu).

Sur le cas avec une forte charge tumorale (en bas à droite), le contour estimé reste d'assez mauvaise qualité. Sur le cas avec hépatectomie (en haut à droite), on remarque que le réseau confond une partie du rein avec le foie. Toutefois, cette erreur est intéressante puisqu'elle est identique dans les deux modalités, ce qui souligne la similarité entre les prédictions du réseau.

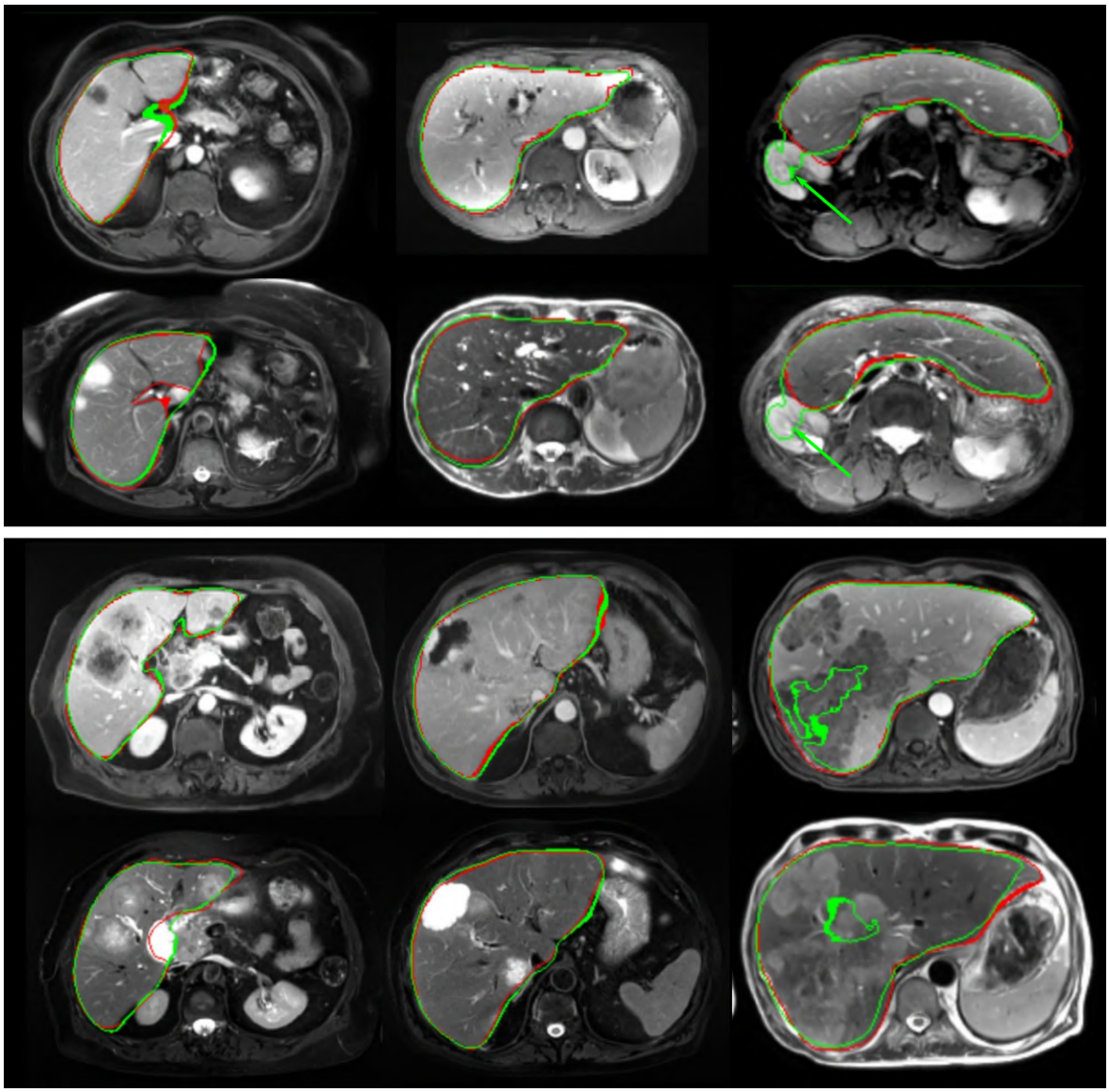


FIGURE 2.16 – Résultats de segmentation avec notre méthode, sur 6 paires d'images de la base de test. Pour chaque paire l'image en T1 est au-dessus de l'image en T2. Le contour prédit par le réseau est représenté en vert, et le contour de la segmentation de référence est représenté en rouge.

2.4 Conclusion

Ce chapitre vise principalement à étudier l'intérêt d'utiliser des informations multi-modales pour la segmentation. Pour cela j'ai commencé par étudier un contexte très simple, avec des données synthétiques en 2D, pour montrer qu'un réseau était capable d'utiliser l'information dans deux images non recalées, jusqu'à une certaine amplitude de décalages. Mes expériences sur des paires d'images IRM pondérées en T1 et T2 ont cependant eu tendance à montrer que l'ajout de l'image auxiliaire avait un effet délétère sur les performances de segmentation du foie. Contrairement à mes intuitions initiales, l'image en T2 contient bien toute l'information nécessaire pour y segmenter précisément le foie. Je discute à la section 5.1.1 d'applications qui seraient plus susceptibles de bénéficier d'une telle combinaison d'informations, par exemple si l'une des modalités contient très peu d'information anatomique.

Une autre conclusion de ces expériences est que segmenter conjointement les deux images de la paire, avec un réseau *multi-tâches*, peut largement détériorer les performances par rapport à un réseau qui n'en segmente qu'une à la fois. De même, mes expériences de comparaison des fonctions de coût n'ont pas permis de mettre en évidence un quelconque avantage apporté par des fonctions de coût basées sur l'apprentissage adversaire, proposées récemment dans la littérature pour la segmentation.

La deuxième partie du chapitre vise à étudier si l'ajout d'informations multi-modales, à défaut de permettre l'amélioration des performances de segmentation, peut aider à obtenir des masques de segmentation plus similaires entre eux. Pour cela, je propose une méthode pour intégrer à l'apprentissage la connaissance *a priori* que l'on a sur le problème, selon laquelle les deux masques prédits ne doivent différer que par une déformation lisse. Cette méthode se base sur une optimisation conjointe de la segmentation et du recalage des deux images, en utilisant le recalage comme tâche auxiliaire pour aider celle de segmentation. Mes expériences sur des données synthétiques et sur les données réelles montrent que la méthode permet effectivement d'augmenter la similarité des deux masques prédits, sans trop compromettre la qualité des segmentations. Elles montrent également que l'ajout de la tâche auxiliaire de recalage permet d'améliorer sensiblement les performances d'un réseau *multi-tâches* entraîné à segmenter les deux images de la paire simultanément. Il serait alors intéressant d'appliquer la méthode à un problème où l'une des modalités contient peu d'information anatomique, avec l'espoir de surpasser les performances d'un réseau segmentant les images une par une.

Chapitre 3

Interprétabilité en segmentation

Aborder un problème en utilisant une technique d'apprentissage automatique suppose l'obtention d'un modèle prédictif dont les paramètres sont optimisés sur un jeu de données. Comme Yu Zhang et al. (2020), on considérera dans ce chapitre qu'un tel modèle est interprétable s'il on peut *expliquer* ses prédictions – c'est-à-dire, fournir des règles logiques ou probabilistes menant à cette prédiction – dans des *termes compréhensibles* par un humain – c'est-à-dire, qui se rapportent à des notions connues du domaine d'application (superpixels ou objets en vision par ordinateur, concepts anatomiques en imagerie médicale par exemple).

Certaines techniques d'apprentissage automatique produisent des modèles plus interprétables que d'autres. Par exemple, une régression linéaire fournit directement un poids pour chacune des dimensions du vecteur d'entrée. Si les dimensions correspondent à des caractéristiques bien identifiées, on peut donc directement connaître l'importance relative de ces caractéristiques pour influencer les prédictions du modèle. De même, un arbre de décision, qui seuille une caractéristique à chaque nœud, est immédiatement interprétable (si l'arbre n'est pas trop grand). En revanche, un réseau de neurones multi-couches ne l'est pas : il n'est pas aisé de déterminer l'utilité de chacun de ses nombreux paramètres (on parle même de couches et d'états *cachés*). C'est vrai à plus forte raison encore pour les réseaux profonds, qui peuvent facilement avoir des dizaines de millions de paramètres à optimiser, répartis sur un grand nombre de couches, et à qui l'on apprend à résoudre des tâches de plus en plus abstraites (détection d'émotions, compréhension du sens d'un texte...).

Pourquoi au juste voudrait-on des modèles interprétables ? De prime abord, dans certains domaines au moins, la perte d'interprétabilité peut sembler un coût à payer bien faible au vu des progrès spectaculaires qu'a permis l'introduction du Deep Learning. De plus, les attentes des concepteurs des algorithmes et de leurs utilisateurs (médecins, patients en traitement de l'image médicale) peuvent différer vis-à-vis de l'interprétabilité des modèles, et ainsi soulever des questions différentes.

Cependant, on peut relever plusieurs enjeux liés à ce problème. D'abord, celui de la confiance que l'on peut accorder aux prédictions de tels modèles. La mesure de bonnes performances d'un modèle suffit-elle à s'assurer que celui-ci fonctionne comme attendu, sur toutes sortes de données ? Pour mesurer la performance d'un modèle en s'affranchissant du problème de sur-apprentissage, il est d'usage de mettre de côté une partie des données pendant la phase d'apprentissage, et de le tester uniquement sur ces données. Or il peut arriver que l'ensemble de test présente les mêmes biais que celui d'entraînement. Par exemple, toutes les images de chevaux de l'ensemble de données public PASCAL VOC contenaient un court texte, ce qui permettait à certains classifieurs d'avoir de bonnes performances sans réellement apprendre à reconnaître un cheval (Lapuschkin et al. 2019). Ainsi, il est toujours intéressant de mieux comprendre le fonctionnement d'un modèle, même s'il montre d'excellentes performances. C'est un enjeu particulièrement important en imagerie médicale, puisque ces modèles y sont parfois destinés à aider les médecins à prendre des décisions.

Un autre enjeu et celui de l'amélioration des méthodes. Par exemple, après avoir établi que les réseaux de classification entraînés sur le jeu de données ImageNet comptaient plus que les humains sur les textures des objets pour les classifier, Geirhos et al. (2018) ont montré qu'ils pouvaient améliorer les performances en rajoutant des images aux textures modifiées à la base d'entraînement.

On peut aussi citer - même s'il est moins important en imagerie médicale qu'en vision par ordinateur - l'enjeu de l'équité (problématique qu'on trouve sous le nom de *fairness* dans la littérature). Kim et al. (2018) ont notamment montré que les réseaux de classification d'images naturelles apprennent les biais racistes et sexistes contenus dans les données d'apprentissage (ils associent par exemple la classification de tablier à la présence d'une femme dans l'image, ou bien les raquettes de ping-pong à la présence de personnes asiatiques).

Toutefois, le problème de l'interprétabilité est mal défini. Comme relevé par Yu Zhang et al. (2020), les définitions et les motivations des différents articles l'abordant sont souvent différentes. La définition sur laquelle on s'est accordé au début de ce paragraphe reste large, et peut englober des problématiques variées. Les critères d'explicabilité et de compréhension demeurent subjectifs, et l'ont peut imaginer toutes sortes de manières de gagner en compréhension sur le fonctionnement d'un modèle.

L'une des problématiques principales de cette thèse est la question de l'apport du Deep Learning pour la segmentation automatique d'images médicales. La question de l'interprétabilité étant en enjeu majeur inhérent à l'application de cette technologie, je propose dans ce chapitre d'étudier quelles réponses ce champ de recherche peut apporter à l'interprétabilité des réseaux de segmentation d'images médicales, avec pour application la segmentation du tissu tumoral du foie dans des coupes axiales d'images

CT.

Dans la section 3.1, je présente une revue de l'état de l'art en Deep Learning interprétable, au moins tel qu'il était au moment de mon travail sur cette problématique. Je montre comment la littérature se focalise principalement sur les réseaux de classification, et discute de pourquoi les méthodes proposées s'appliquent difficilement aux réseaux de segmentation. Dans la section 3.2, je discute des objectifs qu'une méthode d'interprétabilité des réseaux de segmentation pourrait remplir. Autrement dit, je propose une *manière* d'interpréter les réseaux de segmentation. Enfin dans la section 3.3, je propose une méthode et détaille les expériences que j'ai menées pour montrer dans quelle mesure elle permet de remplir les objectifs décrits dans la section précédente. Cette méthode a été présentée au *workshop* iMIMIC associé à la conférence MICCAI 2019 [1] (Couteaux, Nempont et al. 2019).

3.1 Interprétabilité en Deep Learning

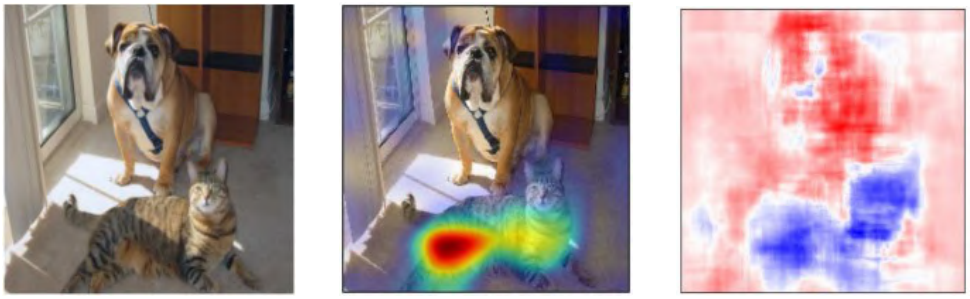
Dans cette section je propose une revue des méthodes d'interprétabilité les plus populaires, au moins au moment où j'ai commencé à travailler sur le sujet mi-2019. Deux revues de la littérature ont récemment paru sur le Deep Learning interprétable (champ de recherche que l'on trouve parfois désigné par l'acronyme *XAI* pour *eXplainable Artificial Intelligence*) : Yu Zhang et al. (2020) proposent une revue générale sur les méthodes d'interprétabilité, tandis que Reyes et al. (2020) se concentrent sur leurs applications à la radiologie, et proposent une réflexion sur les opportunités et les défis spécifiques à cette application.

Pour cette section, je me concentre sur la pertinence des méthodes de la littérature pour interpréter les réseaux de segmentation. J'essaie de rendre compte de la diversité de ces méthodes, en les divisant en trois catégories : les méthodes à base de saillance, les méthodes de visualisation et les méthodes basés concepts.

3.1.1 Cartes de saillance

Cette classe de méthodes, qui compte les méthodes d'interprétabilité les plus populaires, vise à répondre à la question suivante : étant donné un réseau entraîné, un vecteur d'entrée et la réponse du réseau à ce vecteur d'entrée, qu'est-ce qui, dans ce vecteur d'entrée, contribue à la réponse du réseau ? Pour un réseau de classification d'images, qui est le contexte le plus répandu dans la recherche en interprétabilité, cela peut se reformuler en : étant donné une image, un classifieur, et la prédiction du classifieur, quelles parties de l'image ont le plus été utilisées par le réseau ? Ce problème est

1. https://imimic-workshop.com/previous_editions/2019/index.html



(a) Image originale (b) Grad-CAM (Chat) (c) Occlusion (Chat)

(d)Image originale (e) Grad-CAM (Chien) (f) Occlusion (Chien)

FIGURE 3.1 – Exemple de cartes de saillance obtenues avec Grad-CAM (c et d), superposées avec l'image originale, et obtenues par une méthode d'occlusion (c et g). Pour (c) et (g), les pixels bleus correspondent à ceux dont l'occlusion fait baisser le score de la classe chat (c) ou chien (g), tandis que l'occlusion des pixels rouges le fait augmenter. Tiré de SELVARAJU et al. (2017), qui introduisent Grad-CAM.

parfois appelé *explicabilité* : étant donné une prédiction en particulier, on cherche à l'expliquer.

Cela reste toutefois une question vague, à laquelle des approches très différentes ont tenté de répondre. Leur point commun est de fournir des cartes de saillance, c'est-à-dire une image de la taille de celle que l'on cherche à expliquer, où les parties importantes sont mises en valeur.

Comme discuté dans l'article de REYES et al. (2020), l'intérêt de ces méthodes pour l'imagerie médicale est substantiel : si l'on prend l'exemple d'un diagnostic fourni par un classifieur à partir d'une radio, le simple résultat de classification n'a pas grand intérêt. Au contraire, déterminer les zones de l'image qui l'ont amené à faire cette classification peut être très utile pour un radiologue. On peut citer ZECH et al. (2018) qui utilisent une telle méthode pour interpréter un réseau profond entraîné à détecter des pneumonies dans des radios de la poitrine.

Une grande partie de ces méthodes repose sur les gradients d'un objectif (activation

d'un neurone, fonction de coût...) par rapport à l'image, que toutes les bibliothèques de Deep Learning permette facilement de calculer par rétro-propagation. Si l'on calcule les gradients de l'activation d'un neurone de la couche de sortie qui code une certaine classe par rapport à l'image d'entrée, on peut visualiser les pixels dont une petite modification de valeur modifierait le plus la probabilité d'appartenir à cette classe. Simonyan, Vedaldi et Zisserman (2013) ont proposé cette méthode en premier. Plusieurs autres méthodes ont par la suite été proposées pour améliorer la qualité visuelle des cartes de saillance par gradients, comme *SmoothGrad* (Smilkov et al. 2017) ou *Integrated Gradients* (Sundararajan, Taly et Yan 2017). Grad-CAM (Selvaraju et al. 2017) est une autre méthode populaire basée sur les gradients, proposée plus récemment. Une autre idée consiste à re-projeter les activations d'un réseau dans l'espace image, et ainsi visualiser ce qui excite un neurone dans l'image (Zeiler et Fergus 2014, Springenberg et al. 2014).

Ensuite viennent les méthodes d'« attribution », dont le but est d'attribuer à chaque pixel un score, positif ou négatif, sur sa contribution à l'excitation d'un neurone. *Layer-wise Relevance Propagation* (Bach et al. 2015), qui est un cas particulier de la *Deep Taylor Decomposition* (Montavon et al. 2017), est la plus diffusée. La figure 3.2 montre un exemple de cartes de saillance qu'il est possible d'obtenir avec cette méthode, pour un classifieur de chiffres manuscrits.

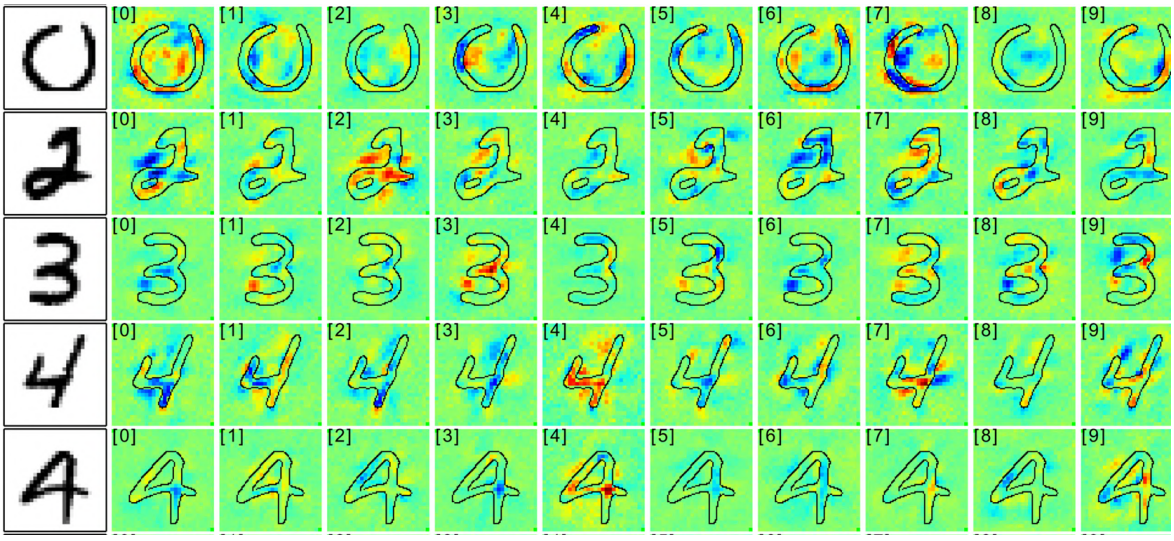


Figure 3.2 – Cartes de saillance obtenues par *Layer-wise Relevance propagation* (LRP) sur un classifieur de chiffres manuscrits. Les pixels rouges correspondent à ceux qui ont une influence positive sur le score de la classe indiquée dans le coin en haut à gauche, tandis que les pixels bleus ont une influence négative. Tiré de Bach et al. (2015)

La fiabilité de toutes les méthodes citées ci-desssus est toutefois remise en question

(Ghorbani, Abid et J. Zou 2019 ; Yeh et al. 2019). Kindermans, Hooker et al. 2017 montrent notamment qu'un simple décalage des intensité d'une image peut donner de explications très différentes pour la plupart des méthodes, et proposent PatternNet et PatternAttribution (Kindermans, Schütt et al. 2017) pour remédier à ces manques de fiabilité. Hooker et al. 2018 montrent que beaucoup de ces méthodes basées sur les gradients ne parviennent pas à mettre en évidences les parties de l'image réellement importantes pour la classification.

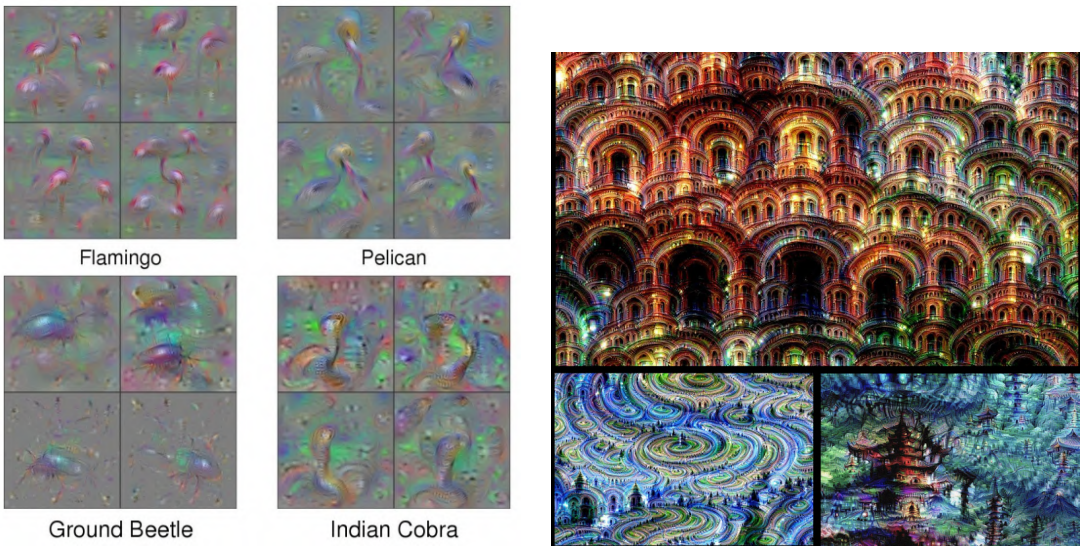
Une approche différente de toutes ces méthodes à base de rétro-propagation consiste simplement à altérer l'image d'entrée pour mesurer comment cette altération modifie la sortie du réseau. On dit alors que la méthode est *model-agnostic*, car elle ne nécessite pas d'avoir accès aux paramètres du modèle (contrairement aux méthodes à base de gradients par exemple). La manière la plus simple suivant ce principe est celle d'« occlusion », qui consiste à cacher un petit morceau de l'image, et mesurer comment la sortie en est affectée. En déplaçant la partie cachée en chaque pixel de l'image, on peut obtenir une carte de saillance, comme sur la figure 3.1 (c) et (f). Les méthodes LIME, proposées par Ribeiro, Singh et Guestrin (2016, 2018) se basent également sur ce principe, en perturbant des ensemble de pixels conjoints et homogènes appelés super-pixels. Une autre méthode *model-agnostic* populaire basée sur la perturbation de caractéristiques est SHAP, proposée par Lundberg et S.-I. Lee (2017).

Toutefois la question à laquelle tentent de répondre toutes les méthodes de cette classe, qui se base sur la localisation dans l'image des parties importantes, est peu pertinente pour les réseaux de segmentation. En effet, la segmentation est par essence une tâche de localisation, et le masque prédit contient déjà cette information.

3.1.2 Visualisation

Une autre manière de gagner en compréhension sur un réseau de neurone est de générer une image qui maximise l'activation d'un neurone du réseau. Le principe est d'optimiser cette activation par montée de gradient, en utilisant les gradients calculés par rétro-propagation, comme dans la section précédente. Plusieurs articles (Simonyan, Vedaldi et Zisserman 2013 ; Yosinski et al. 2015) ont proposé des méthodes qui reposent sur ce principe, qui ne varient que par la régularisation utilisée : L_2, flou Gaussien, translations aléatoires, zooms. Sans régularisation, en effet, les images générées sont plus difficiles à interpréter. La figure 3.3a montre le genre d'images qu'on peut obtenir avec cette méthode.

Cette méthode est connue sous le nom de *DeepDream*, d'après un article de blog de chercheurs de Google (Mordvintsev, Olah et Tyka 2015), dans lequel ils montrent les images très stylisées qu'ils ont réussi à obtenir avec cette méthode (voir figure 3.3b).



(a) (b)

FIGURE 3.3 – Exemples d'images obtenues par maximisation d'activation. (a) est tirée de SIMONYAN, VEDALDI et ZISSERMAN (2013), (b) de MORDVINTSEV, OLAH et TYKA (2015)

Une variante de cette méthode, proposée par MAHENDRAN et VEDALDI (2015, 2016), consiste à trouver l'image qui minimise la distance de sa représentation à une couche donnée du réseau (c'est-à-dire les activations de cette couche) à une représentation cible, et ainsi inverser cette représentation. Cela permet de visualiser quelle information sur l'image a été gardée, à une couche donnée du réseau.

NGUYEN et al. (2016) ont proposé de visualiser une image qui maximise l'activation d'un neurone non pas en optimisant directement sur les pixels de l'image, mais sur le vecteur d'entrée d'un réseau générateur, préalablement entraîné par apprentissage adversaire. Ainsi, les images générées ressemblent davantage à des images réelles, ce qui rend la visualisation plus aisée.

Si ces méthodes de visualisation par maximisation d'activation peuvent sans problème être appliquées aux réseaux de segmentation, à ma connaissance aucun article ne s'y est intéressé. Il faut aussi noter que la visualisation d'images générées est surtout adaptée aux réseaux entraînés sur des images naturelles. En effet, un humain peut facilement reconnaître les objets qui apparaissent à l'intérieur (par exemple les animaux de la figure 3.3a). Cependant, on peut soutenir que cette interprétation est moins facile pour les images médicales, ce qui rend l'interprétation par visualisation moins utile dans ce cas.

3.1.3 Interprétabilité par concepts

Une autre question à laquelle on peut chercher à répondre est celle de la topologie de l'espace des représentations, dans les différentes couches du réseau. En apprentissage automatique, on cherche toujours à représenter les données dans un espace où elles sont linéairement séparables (autrement dit, en considérant une tâche de classification binaire, un espace où il existe un hyperplan tel que les échantillons d'une même classe soient tous du même côté de celui-ci). C'est le principe des méthodes à noyau par exemple.

Alain et Bengio (2016) montrent que les réseaux de neurones convolutionnels apprennent une représentation dans laquelle les données deviennent progressivement linéairement séparable, au fur et à mesure des couches du réseau. Ils obtiennent ce résultat à l'aide de « sondes », qui sont des modèles de classification linéaires qu'ils entraînent sur les sorties de chaque couche. Une idée d'interprétabilité est alors de visualiser l'ensemble des échantillons de la base de données à différentes couches du réseau, comme mentionné par Chakraborty et al. (2017), à l'aide de méthodes de visualisation par réduction de dimensionnalité comme t-SNE (Van der Maaten et G. Hinton 2008) ou UMAP (McInnes, Healy et Melville 2018). K. Xu et al. (2018) proposent une méthode de réduction de dimensionnalité utilisant les probabilités d'appartenance à une classe prédites par un réseau de classification.

L'idée qui suit, exploitée par Kim et al. (2018), est de se demander si l'espace des représentations appris par le réseau sépare linéairement les données, non seulement en fonction de leur classe, mais aussi en fonction de concepts compréhensibles par les humains. L'intuition est que pour classifier correctement un tigre, un réseau doit apprendre une représentation dans laquelle les objets rayés sont séparés des autres. On peut comme cela tester différents concepts pour lesquels on a une base annotée.

Yeche, Harrison et Berthier (2019) ont proposé une méthode pour interpréter les réseaux de classification d'images médicales, qui se base sur ce principe, en utilisant des caractéristiques continues à la place de concepts discrets. Mes essais d'appliquer cette stratégie avec un réseau de segmentation n'ont cependant pas étés concluants, et ont donné des résultats moins intéressants qu'avec la méthode, plus simple, que je propose dans la suite.

Cette classe d'approches m'intéresse particulièrement parce qu'elle est applicable aux images médicales. Il est en effet possible de décrire de manière compréhensible un objet à segmenter dans une image médicale avec ses caractéristiques (grand, petit, clair, sombre, texturé, homogène...), qu'on peut voir comme des concepts interprétables. La méthode que je décris dans ce chapitre fait le pont entre les méthodes de visualisation, adaptées aux réseaux de segmentation, et les méthodes par concepts, adaptées aux images médicales.

3.2 Comment interpréter un réseau de segmentation ?

On a vu dans la section précédente que de nombreuses manières d'interpréter un réseau avait été proposées dans la littérature, et que l'interprétabilité par concepts était l'approche la plus prometteuse pour interpréter les réseaux de segmentation parmi les trois décrites. L'idée est de choisir un ensemble de concepts compréhensibles par un humain et de déterminer lesquels de ces concepts un réseau utilise pour prendre la décision de ranger une entrée dans une certaine classe. On aimerait être capable de répondre à des questions du genre « ce réseau détecte-t-il des rayures pour reconnaître un tigre ? Est-il sensible au rouge pour détecter un camion de pompiers ? »

Dans le cas d'un réseau de segmentation d'images médicales, on peut se poser le même genre de questions à l'échelle des groupe de pixels : l'intensité d'un tel groupe joue-t-elle un rôle dans la décision de le considérer comme une tumeur ? Sa taille, sa forme ont-elles une importance ? La figure 3.4 montre un exemple d'une coupe dans laquelle une tumeur du foie est visible. On peut y voir différents groupes homogènes de pixels dans le foie et aux alentours, dont quelques-uns sont mis en évidence par une flèche. Un oeil humain (qui a l'habitude) peut aisément reconnaître que la tache pointée par la flèche jaune, au vu de son intensité plus importante que les alentours, de sa taille et de sa forme, est un vaisseau sanguin, alors que celles pointées par la flèche rose, qui sont plus sombres, est une lésion. De la même façon qu'un humain peut souvent facilement expliquer à quoi il reconnaît ce qu'il voit, il serait intéressant de savoir si un réseau prend en compte des critères similaires à ceux des humains pour prendre ses décisions.

Ainsi, une manière d'interpréter le fonctionnement d'un réseau de segmentation serait de quantifier quelles caractéristiques compréhensibles par un humain (comme la taille, la forme ou la texture) doivent avoir certains groupes de pixels pour être plus susceptibles d'être segmentés par le réseau.

Pour être plus précis, introduisons les notions de *sensibilité* et de *robustesse* d'un réseau à une caractéristique. Dans la suite, on dit qu'un réseau est *sensible* à une caractéristique si un groupe de pixels a plus de chance d'être classifié positivement par ce réseau lorsque, toutes choses égales par ailleurs, la valeur de cette caractéristique calculée sur ce groupe de pixels augmente. À l'inverse, on dit qu'un réseau est *robuste* (ou *indifférent* selon ce à quoi on s'attend) à une caractéristique si la sortie du réseau ne varie pas lorsqu'on modifie, toutes choses égales par ailleurs, la valeur d'une caractéristique d'un groupe de pixels. Interpréter un réseau reviendrait à estimer, pour un ensemble choisi de caractéristiques compréhensibles, les sensibilités du réseau à chacune d'entre elles.

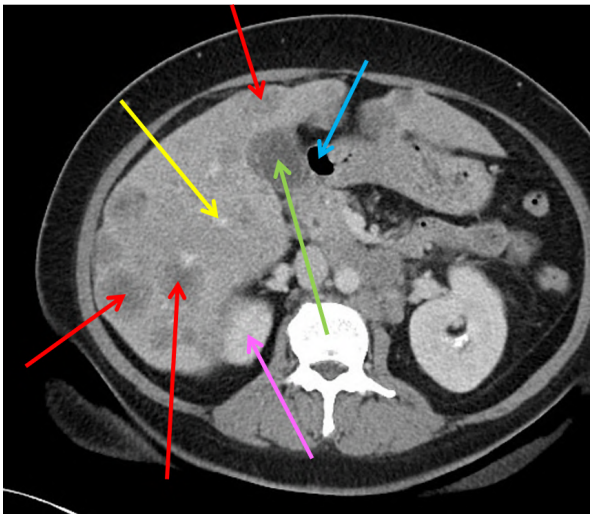


FIGURE 3.4 – Plusieurs taches sont visibles dans le foie et à proximité. Comment le réseau parvient-il à faire la différence entre les métastases (flèches rouges), un vaisseau sanguin (flèche jaune), la vésicule biliaire (flèche verte), un morceau de rein (flèche rose) ou d'intestin (flèche bleue) ?

Par exemple, on sait qu'en imagerie CT au temps d'injection portal, les métastases apparaissent comme des taches hypointenses (voir les flèches rouges sur la figure 3.4), à peu près sphériques et que leur taille peut varier de quelques millimètres à plusieurs centimètres. On s'attend donc à ce qu'un réseau performant pour segmenter les métastases dans ce type d'images soit sensible aux faibles intensités, et relativement robuste à la taille. Admettons au contraire qu'on se rende compte que le réseau obtenu est très sensible au diamètre des tumeurs, et qu'il soit d'autant plus disposé à segmenter une tache sombre qu'elle est grosse. On pourrait alors en déduire que le réseau risquerait de passer à côté de petites tumeurs, ce que l'on aurait pas pu savoir en se contentant de tester le réseau avec le score Dice moyen sur une base de test. Pour y remédier, on pourrait par exemple rajouter à la base d'entraînement des images montrant des petites lésions, ou pénaliser plus fortement les erreurs sur les petites lésions pendant l'apprentissage, afin de faire diminuer cette sensibilité et ainsi augmenter la robustesse générale du réseau.

Cependant, la notion de « toutes choses égales par ailleurs » implique qu'il est très difficile de mesurer directement ces sensibilités pour un ensemble arbitraire de caractéristiques : on ne peut pas artificiellement faire varier la taille d'une tumeur, par exemple, sans toucher à d'autres caractéristiques d'intensité ou de texture et on ne pourrait pas déterminer à quoi seraient dues les variations de réponse du réseau. Dans la section suivante, je propose une méthode qui permet de se faire une idée des sensibilités et robustesses de n'importe quelle caractéristique calculable.

Parmi elles, les caractéristiques radiomiques, qui sont un ensemble de descripteurs conçus pour extraire l'information utile des organes ou lésions segmentées dans les images, sont de bons candidats. Une centaine de caractéristiques, plus ou moins interprétables, ont été standardisées par Zwanenburg et al. (2020), comprenant des descripteurs de formes (diamètre, volume, sphéricité, élongation...), des statistiques sur les intensités (énergie, moyenne, écart type, entropie...), et des descripteurs de texture basés sur des statistiques de voisinages (par exemple, le nombre de pixels voisins ayant la même intensité). Elles serviront de base et de motivation pour utiliser des caractéristiques continues plutôt que des concepts discrets (comme la présence ou non de rayure, par exemple).

3.3 L'analyse de *Deep Dreams*

3.3.1 Principe

Pour illustrer le principe de la méthode, prenons l'exemple d'un jeu de données bi-dimensionnelles, tel que représenté sur la figure 3.5, à gauche. Les échantillons de ce jeu de données sont alignés sur une ligne, représenté par une flèche verte sur la figure, de sorte qu'un simple seuil sur cette ligne suffise à distinguer les classes positives et négatives. Admettons ensuite l'existence d'une fonction de classification continue qui sépare le plan en deux selon la ligne représentée en gris sur la figure, de telle manière que les points du plan en-dessous de la ligne soient classifiés négativement, et les points au-dessus positivement. On voit que ce classifieur permet d'avoir une précision parfaite sur ces données.

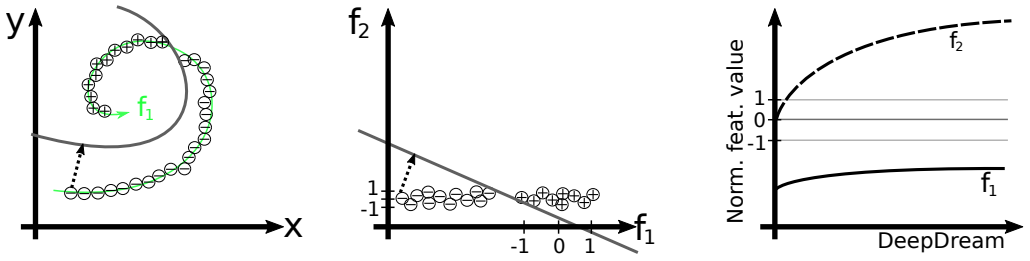


Figure 3.5 – Illustration de la méthode avec un classifieur bi-dimensionnel. À gauche : espace des entrées, $\oplus$ and $\ominus$ représentent respectivement les examples positifs et négatifs. Le classifieur est représenté par la ligne grise, et les données peuvent être décrites par les caractéristiques f_1 (flèche verte) et f_2 (orthogonale à f1). Au milieu : les données représentées dans l'espace des caractéristiques. À gauche et au milieu, le chemin de plus forte pente est représenté par une flèche pointillée. Á droite, la projection de ce chemin sur f_1 et f_2

Admettons ensuite qu'on puisse calculer deux caractéristiques f_1 et f_2 telles que pour chacun de ces échantillons, f_1 corresponde à la position sur la ligne verte, et f_2 l'écart à la ligne. On peut ainsi représenter le jeu de données dans l'espace de ces deux caractéristiques (figure 3.5, milieu). Les classes positives et négatives étant complètement déterminées par la la position sur la ligne f_1, on peut s'attendre, intuitivement, à ce qu'un classifieur idéal sur ce jeu de données s'appuie uniquement sur f_1, et que la ligne qui sépare l'espace apparaisse donc verticale dans l'espace des caractéristiques. Or, on remarque que ce n'est pas le cas lorsqu'on représente cette ligne pour le classifieur considéré (la ligne grise sur la figure du milieu), et qu'il n'est donc pas idéal (on peut imaginer par exemple un échantillon qui soit mal classifié parce qu'un peu trop écarté de la ligne).

Autrement dit, on aimerait que le classifieur soit *sensible* à f_1 et *robuste* à f_2. Comment, dès lors, déterminer la trop faible sensibilité de notre classifieur à f_1, et son manque de robustesse à f_2 ?

L'idée de la méthode que je détaille dans la suite, qui fonctionne avec tous les classifieurs dérivables, consiste à suivre le chemin de plus forte pente depuis un échantillon négatif dans l'espace d'entrée, et de suivre l'évolution des caractéristiques qui nous intéressent le long de ce chemin. Ce chemin de plus forte pente, qu'on appellera chemin *Deep dream* pour des raisons qu'on précisera dans la suite, est représenté par une flèche pointillée sur la figure 3.5 à gauche et au milieu. À droite, on peut voir l'évolution des caractéristiques f_1 et f_2 le long de ce chemin.

L'hypothèse principale de la méthode est que le chemin de plus forte pente dans l'espace d'entrée favorise les caractéristiques auxquelles le réseau est le plus sensible. Ainsi, si la valeur d'une caractéristique varie beaucoup le long de ce chemin, on peut en déduire que le réseau est sensible à celle-ci ; inversement, si la valeur d'une caractéristique reste constante, on peut en déduire que le réseau y est robuste.

Normaliser les valeurs des différentes caractéristiques est crucial pour savoir si les variations de valeurs sont significatives. Pour cela, une solution simple est de normaliser selon la moyenne et l'écart-type calculés sur les exemples positifs de la base d'entraînement. De cette manière, la valeur d'une caractéristique s'éloignant fortement des valeurs normales (comme c'est le cas pour f_2 dans l'exemple de la figure 3.5) peut indiquer une trop forte sensibilité à cette caractéristique. De la même manière, une caractéristique évoluant vers des valeurs normales peut indiquer une sensibilité attendue, de même qu'une stagnation hors des valeurs normales peut indiquer un manque de sensibilité.

3.3.2 L'algorithme

Deep Dream est le nom donné par les chercheurs de Google (Mordvintsev, Olah et

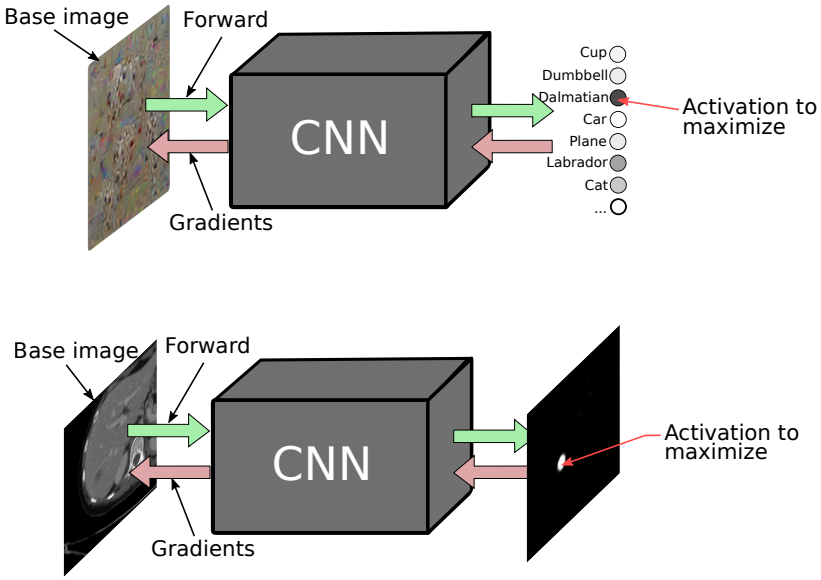


FIGURE 3.6 – En haut : principe de *Deep Dream* pour un réseau de classification d'images naturelles. En bas : même principe appliqué à un réseau de segmentation d'images médicales.

TYKA 2015) à l'algorithme de maximisation d'activation, qui permet de générer une image maximisant l'activation d'un neurone du réseau. Si l'on choisit un neurone de sortie d'un réseau de classification, on peut visualiser le genre d'image qui maximise la probabilité d'appartenir à une certaine classe (figure 3.6, haut). On peut aussi appliquer le même algorithme à un réseau de segmentation (figure 3.6, bas). Dans ce cas, les neurones de la dernière couche ne représentent plus les probabilités d'appartenance à des classes, mais la probabilité que le pixel correspondant fasse partie d'un objet à segmenter (un organe ou une lésion dans le cas d'images médicales).

Étant donné un réseau de segmentation S et un pas α, on calcule le *DeepDream* grâce à un algorithme itératif de montée de gradient. On part d'une image X_0 qui ne contient pas d'objet à segmenter (une image de foie sans lésion dans le cas d'un réseau de segmentation de tumeurs du foie par exemple), et d'un neurone i de la couche de sortie choisi arbitrairement.

À chaque itération j et jusqu'à convergence :

— On fait passer l'image X_j dans le réseau et on récupère ainsi le masque $M_j = S(X_j)$, de même que le gradient de l'activation du neurone par rapport à l'image : $\frac{\partial i}{\partial X}$

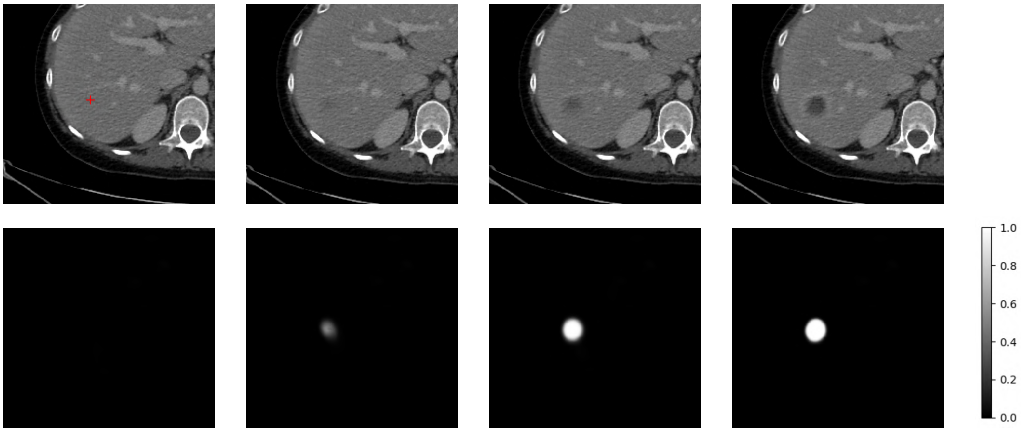


FIGURE 3.7 – Différentes étapes d'une montée de gradient appliqué à une coupe CT montrant un foie sain, avec un réseau entraîné à segmenter des tumeurs du foie dans des coupes CT. La ligne du haut montre les images X_j, et la ligne du bas montre les masques M_j (le blanc indique une probabilité proche de 1). La croix rouge sur l'image de gauche montre l'emplacement du neurone maximisé par montée de gradient. On observe l'apparition d'une zone répondant positivement.

— On met à jour l'image pour l'itération suivante $X_{j+1} = X_j + \alpha \frac{\partial i}{\partial X}$.
— On calcule les caractéristiques de l'objet généré, qui sont des fonctions de l'image et du masque $f_k(X_j, M_j)$.

À chaque étape, on maximise donc la sortie du réseau à un endroit précis de l'image. Une groupe de pixels répondant positivement va donc apparaître au fur et à mesure des itérations (voir figure 3.7). La trajectoire des X_j va donc suivre le chemin de plus forte pente dans l'espace image.

L'avantage d'appliquer cet algorithme à un réseau de segmentation est qu'on récupère gratuitement à chaque itération le masque de l'objet généré, ce qui permet de calculer des caractéristiques f_k qui dépendent à la fois des intensités de l'objet généré, et de son masque. Le produit final de l'*analyse de Deep Dream* est donc formé des courbes $j \to f_k(X_j, M_j)$, dont on pourra interpréter l'évolution en termes de sensibilité et robustesse, en fonction de ce que l'on attend d'un réseau selon l'application.

3.3.3 Expériences

La pertinence de cette méthode repose sur l'hypothèse que le chemin de plus forte pente favorise l'évolution des caractéristiques auxquelles le réseau est le plus sensible. Si celle-ci peut paraître raisonnable, elle n'est néanmoins pas soutenue par des bases

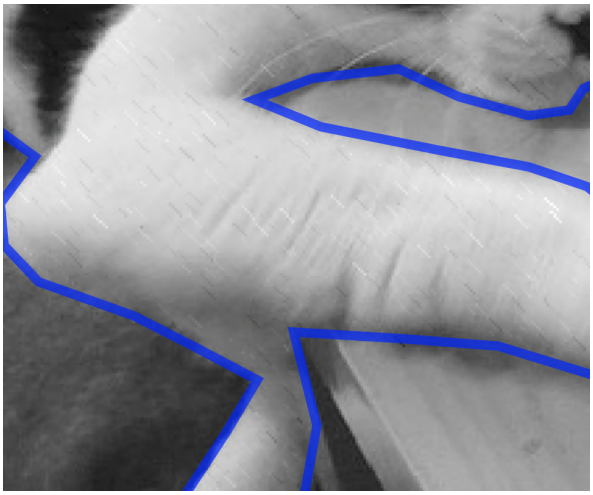


FIGURE 3.8 – Image marquée pour l'expérience contrôlée. Les lignes bleues représentent le contour de la segmentation de référence.

théoriques. Le but de cette section est de montrer, dans des contextes contrôlés, quel genre de résultats la méthode peut effectivement fournir.

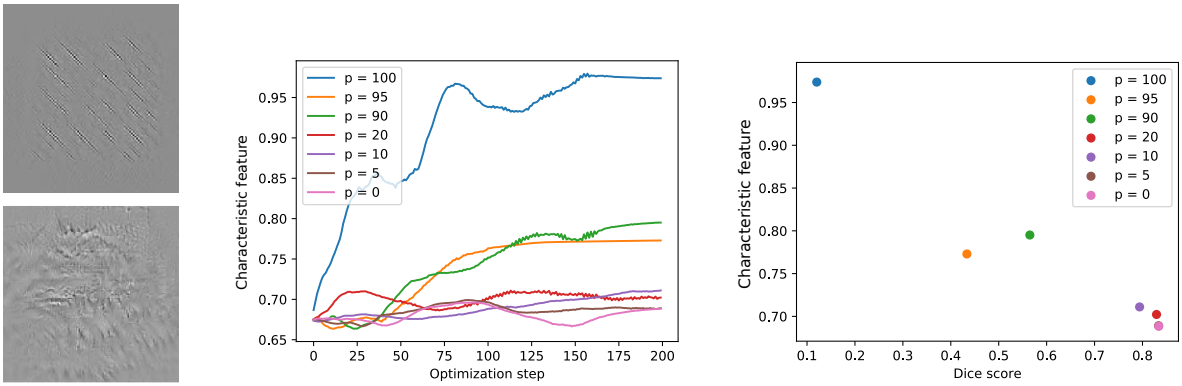
Expérience contrôlée

Cette expérience vise à montrer que l'analyse de *Deep Dream* peut permettre, dans un contexte où l'on connaît par ailleurs la sensibilité réelle d'un réseau à une caractéristique, d'estimer cette sensibilité. Pour cela, on part d'une tâche de segmentation de chats et de chiens en utilisant le jeu de données public COCO (T.-Y. LIN, MAIRE et al. 2014).

On ajoute à une proportion p de chats et chiens un marquage, qui est une texture synthétique constituée de courts segments orientés à 135°, d'intensités et positions variables (voir figure 3.8) (Les chats et chiens marqués sont tirés aléatoirement avec une probabilité p). On entraîne plusieurs réseaux G_p, d'architecture U-net, avec différentes valeurs de p.

Pour estimer à quel point un réseau compte sur le marquage pour segmenter les images, on calcule simplement le score Dice obtenu sur la base de test sans le marquage. Ainsi, on considère qu'un réseau qui obtient un bon score sur la base de test sans le marquage ne l'utilise pas pour prendre ses décisions, tandis qu'un réseau qui obtient un plus mauvais score en a plus besoin.

Pour l'analyse de DeepDream, on a besoin de suivre une caractéristique f_m qui répond à la présence du marquage. Pour cela, j'utilise le maximum de la convolution de l'image avec un segment orienté à 135° de la même taille que celui utilisé pour le marquage,



(a) (b) (c)

FIGURE 3.9 – Expérience contrôlée. 7 réseaux ont été entraînés avec différentes probabilités p de marquer les zones annotées positivement. (a) *Deep Dream* du réseau avec $p = 100\%$ en haut, $p = 0\%$ en bas. (b) Evolution de la caractéristique du marquage pendant la procédure de montée de gradient. (c) Score Dice sur l'ensemble de test sans marquage, et caractéristique du marquage à la fin du *Deep Dream* pour chacun des 7 réseaux. On remarque que plus un réseau obtient de mauvaises performances sur l'ensemble de test, et par conséquent compte sur le marquage pour faire ses segmentations, plus son *Deep Dream* montre une caractéristique du marquage élevée.

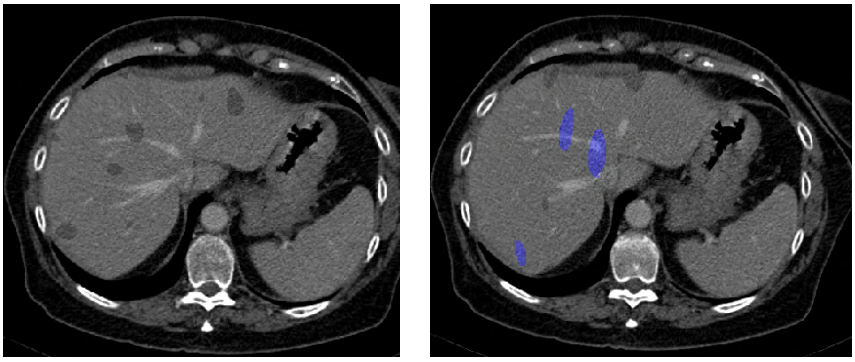
que j'appelle « caractéristique du marquage » dans la suite :

$$f_m(X, M) = \max_{i,j \in \{1,\dots,h\}\times\{1,\dots,l\}} (X \star s_{135^\circ})_{i,j} \times M_{i,j}$$

où X est une image de taille $h \times l$, $M \in \{0,1\}^{h\times l}$ est le masque, s_{135° est le segment orienté à 135°. i et j représentent les coordonnées des pixels de l'image.

La figure 3.9a montre les *DeepDreams* obtenus pour les réseaux avec $p = 100\%$ et $p = 0\%$. On peut clairement voir une texture semblable à la texture synthétique rajoutée pour le réseau avec $p = 100\%$. La figure 3.9b montre l'évolution de la caractéristique f_m au cours de l'optimisation, pour différentes valeurs de p. Ces courbes montrent une augmentation rapide de f_m pour le réseau entraîné avec $p = 100\%$, ainsi que pour les réseaux entraînés avec $p = 95\%$ et $p = 90\%$, ce qui suggère une certaine sensibilité de ces réseaux à cette caractéristique. En revanche, f_m reste constante pour les réseaux entraînés avec $p \leq 20\%$, suggérant cette fois que ces réseaux ne sont pas sensibles au marquage.

La figure 3.9c met en relation le score Dice obtenu sur la base de test sans marquage, qui indique à quel point un réseau peut se passer du marquage, à la sensibilité à la caractéristique f_m évaluée par notre méthode. On constate que les réseaux entraînés



(a) (b)

FIGURE 3.10 – Tumeurs synthétiques ajoutées (zones sombres) dans le foie sur une image montrant un foie sain. (b) Image avec des tumeurs synthétiques allongées. En bleu le masque de la vérité terrain pour l'expérience sur l'élongation.

avec $p \leq 20\%$ n'ont pas de perte de performance lorsqu'on enlève le marquage (le réseau $p = 0\%$ servant d'étalon), ce qui cohérent avec le fait que l'analyse de *DeepDream* ne détecte de pas de sensibilité à la caractéristique du marquage. De manière générale, on constate que plus un réseau compte sur le marquage pour prendre ses décisions, plus on détecte une sensibilité élevée à f_m.

Cette expérience tend donc à montrer que, dans certains contextes au moins, la méthode permet de se faire une bonne idée de la sensibilité (ou robustesse) d'un réseau de segmentation à certaines caractéristiques.

Tumeurs synthétiques

Pour cette expérience, on considère un réseau entraîné à segmenter des tumeurs du foie dans des coupes CT, ainsi qu'un réseau entraîné à segmenter des « fausses tumeurs ».

Pour générer la base de fausses tumeurs, on récupère les coupes CT de la base de données qui contiennent du foie mais pas de tissu tumoral. Pour chacune de ces coupes, on génère aléatoirement un masque, à la manière de l'expérience décrite à la section 2.2.1, dont on ne garde que l'intersection avec le masque du foie (qui a été segmenté manuellement au préalable). Ensuite, on abaisse simplement l'intensité des pixels de l'image où le masque est positif. Le résultat de la procédure est visible sur la figure 3.10a.

Le but de cette expérience est d'étudier ce que l'analyse de DeepDream peut nous apprendre sur la différence de comportement de deux réseaux entraînés sur des tâches qui peuvent *a priori* sembler similaires, si l'on considère qu'un humain va chercher des

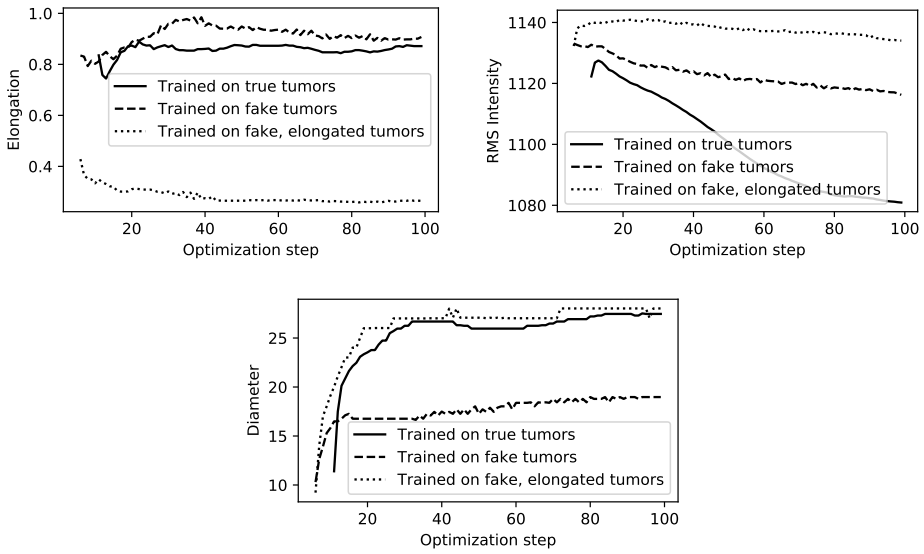


FIGURE 3.11 – Analyze de DeepDream de 3 caractéristiques, comparant les 3 réseaux considérés pour l'expérience des fausses tumeurs. L'élongation, suivant le standard des caractéristiques radiomiques, est proche de 0 pour les objets allongés, et proche de 1 pour les objets ronds.

taches sombres dans le foie pour segmenter des tumeurs en CT.

Parallèlement, on entraîne un autre réseau sur une autre base de tumeurs synthétiques, dans laquelle les pseudo-tumeurs sont générées avec des formes plus allongées. De plus, on entraîne ce réseau pour ne segmenter que les pseudo-tumeurs dont la hauteur dépasse de deux fois la largeur (voir figure 3.10b). De cette manière, on force le réseau à être sensible à l'élongation. L'idée ici est de vérifier que l'analyse de DeepDream est capable également de mettre en valeur les sensibilités à des caractéristiques de forme, et pas seulement des caractéristiques d'intensité et de texture.

On effectue l'analyse de ces trois réseaux sur trois caractéristiques : l'élongation, la moyenne quadratique des intensités, et le diamètre. Les résultats sont visibles sur la figure 3.11. Le premier résultat à discuter est celui des différences de sensibilités à l'intensité entre le réseau entraîné sur les vraies tumeurs, et celui entraîné sur les fausses. En effet, on remarque que le réseau entraîné sur les vraies tumeurs montre une sensibilité aux basses intensités plus importante. L'interprétation qu'on peut en tirer est que le réseau entraîné sur les fausses tumeurs n'a pas eu besoin de comprendre que celles-ci avaient une intensité plus basse pour être performant, et qu'il a trouvé d'autres caractéristiques sur lesquelles prendre ses décisions, comme peut-être les bords saillants, qui sont peu réalistes. À l'inverse, le réseau entraîné sur les vraies tumeurs montre une

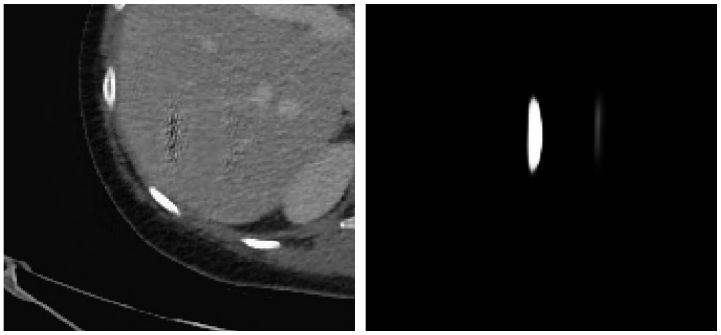


FIGURE 3.12 – DeepDream et le masque correspondant générés avec le réseau entraîné sur les fausses tumeurs allongées.

meilleure sensibilité aux basses intensités, ce qui tend à montrer que, pour cette tâche plus difficile, le réseau doit apprendre plus en détail les caractéristiques réelles des tumeurs. De manière générale, ce résultat est conforme à l'intuition que moins la base d'entraînement est variée, plus le réseau peut se concentrer sur des détails ou sur une partie seulement des caractéristiques des objets à segmenter. La sensibilité plus faible au diamètre montré par le réseau entraîné sur les pseudo-tumeurs peut aussi aller dans le sens de cette intuition, dans le sens où le critère de taille aurait moins d'importance pour cette tâche plus facile, en admettant que les pseudo-tumeurs générées soient de même taille en moyenne que les vraies.

On remarque ensuite que les DeepDreams obtenus avec le réseau entraîné sur les pseudo-tumeurs allongées reflètent bien la sensibilité à l'élongation, comme on peut le voir sur la figure 3.12, et sur les courbes en haut à gauche de la figure 3.11. Cela tend à montrer que la méthode peut également mettre en valeur les sensibilités à des caractéristiques de forme, comme l'élongation.

Exemple d'interprétation en utilisant les caractéristiques radiomiques

Pour cet exemple on effecture l'analyse de DeepDream avec 6 caractéristiques issues des caractéristiques radiomiques standardisées décrites par ZWANENBURG et al. (2020). Deux d'entre elles sont des caractéristiques de forme (périmètre et sphéricité), deux sont des statistiques sur les intensités (entropie et intensité), et deux sont des caractéristiques de texture (Contraste GLCM, qui mesure la tendance des pixels voisins de l'objet à avoir de grande variations d'intensité, et GLDM *large dependance emphasis*, qui mesure la propension de l'objet à contenir de larges zones d'intensités proches). On s'attend à ce que le réseau ait appris à reconnaître des tâches rondes de faible intensité, et donc qu'il soit sensible aux faibles intensité et à la sphéricité. En revanche, comme les lésions sont en général peu texturées, et que le parenchyme du foie duquel il doit les discriminer non plus, on a peu d'attentes sur les caractéristiques de texture.

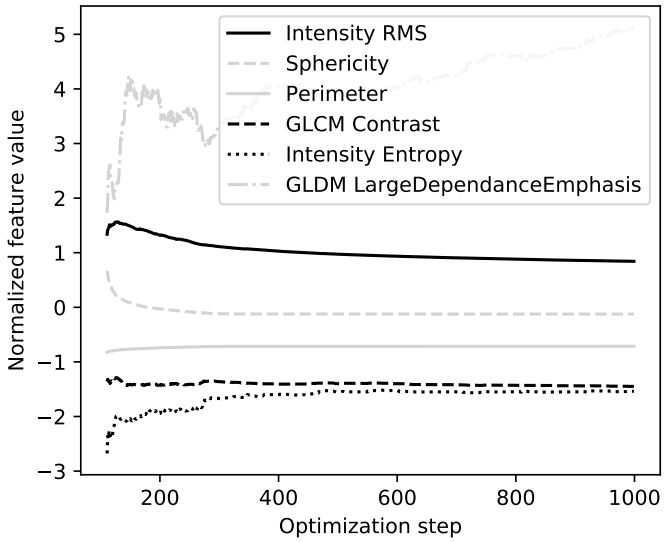


FIGURE 3.13 – Analyse de DeepDream d'un réseau de segmentation de tumeurs de foie dans des coupes CT

Chacune d'entre elles est normalisée sur les valeurs calculées sur la base d'entraînement, de manière à ce que 0 corresponde à la valeur moyenne, et que $[-1, 1]$ corresponde à l'intervalle normal. Les résultats sont montrés sur la figure 3.13.

On peut alors discerner plusieurs tendances, qu'ont peut interpréter comme suit :

- La valeur d'une caractéristique évolue en se rapprochant des valeurs normales, comme c'est le cas ici pour l'intensité et la sphéricité. Cela montre une bonne sensibilité du réseau à ces caractéristiques.
- La valeur d'une caractéristique évolue peu, tout en restant à l'intérieur de l'intervalle normal, comme c'est le cas ici pour la périmètre. On peut interpréter cela comme une bonne robustesse à cette caractéristique.
- La valeur d'une caractéristique évolue peu, ou reste en dehors des valeurs normales comme c'est le cas ici pour le contraste GLCM, ou l'entropie. Cela peut montrer une trop faible sensibilité à ces caractéristiques.
- La valeur d'une caractéristique évolue fortement et rapidement hors des valeurs normales, comme la *Large Dependance Emphasis GLDM* ici. On peut interpréter cela comme une trop forte sensibilité (ou un manque de robustesse) à cette caractéristique.

Il est important de noter que ces caractéristiques n'ont pas la même utilité pour se faire une idée du fonctionnement du réseau. En fonction des informations a priori qu'on a sur ce fonctionnement (par exemple qu'un réseau de segmentation de tumeurs

doit être sensible à des taches de faible intensité) où de l'interprétabilité même des caractéristiques (ici il est difficile d'avoir une idée précise de ce que les caractéristiques GLCM et GLDM signifient), les courbes apportent plus ou moins d'information.

Au total, l'analyse de DeepDream nous apprend que ce réseau est sensible aux taches sombres, avec une préférence pour les formes rondes, et qu'il semble peu tenir compte de la texture des tumeurs. Ceci est cohérent avec nos ce que l'on connaissait a priori du problème, et suggère donc que le réseau a bien appris à reconnaître une lésion.

3.4 Conclusion

Le travail présenté dans ce chapitre se base sur le constat que la littérature en Deep Learning interprétable s'intéresse peu aux réseaux de segmentation (section 3.1). Or, le constat selon lequel l'interprétabilité des modèles de traitement d'image est particulièrement importante quand il s'agit de traiter des images médicales, et selon lequel la segmentation automatique est à la fois l'un des problèmes les plus importants en imagerie médicale et l'un de ceux qui ont le plus bénéficié de l'essor du Deep Learning, fournit une motivation importante pour s'attaquer à ce problème.

Je propose ainsi, dans la section 3.2 une manière d'interpréter les réseaux de segmentation, différente des approches basées sur l'explication d'exemples qui sont maintenant standard pour les réseaux de classification mais peu pertinentes pour les réseaux de segmentation. Celle-ci consiste à estimer les *sensibilités* et *robustesses* du réseau à certaines caractéristiques des objets qu'il a appris à segmenter (la forme, l'intensité des tumeurs par exemple).

Je propose également, dans la section 3.3, une méthode pour estimer ces sensibilités localement, c'est-à-dire au voisinage d'un point négatif, où il n'y a pas d'objet à segmenter. Je montre avec des expériences que cette méthode peut effectivement estimer correctement les sensibilités d'un réseau à certaines caractéristiques, notamment de texture et de forme, dans des contextes où l'on connaît ces sensibilités par ailleurs.

Celle-ci pourrait être utilisée pour s'assurer du fonctionnement correct d'un réseau, en complément du calcul du score sur une base de test, qui peut souffrir des mêmes biais que la base d'entraînement. Ce calcul pourrait ainsi ne pas mettre en valeur certains comportements indésirables, que l'on pourrait par ailleurs exprimer en termes de sensibilité et de robustesse à des caractéristiques. Elle pourrait être également utile pour détecter et corriger certains de ces biais, et ainsi obtenir un modèle plus robuste en général.

Chapitre 4

Détection multi-modale de tumeurs

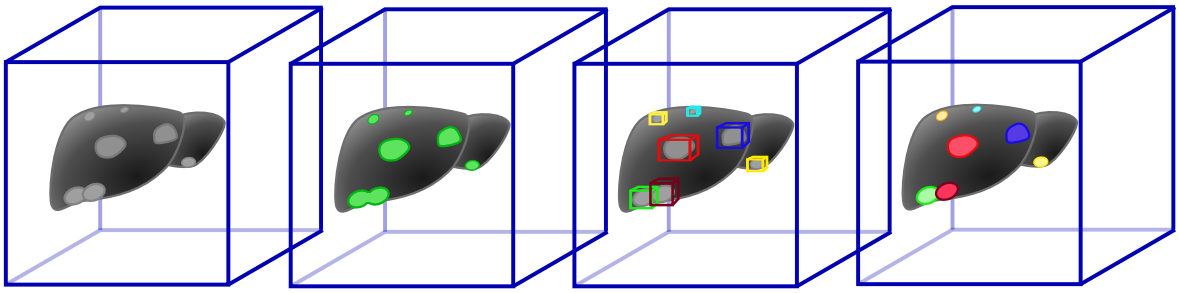


FIGURE 4.1 – Illustration des différents problèmes de segmentation et détection de lésions. De gauche à droite : une image volumétrique de foie avec des lésions ; segmentation (dite « sémantique ») ; détection de boîtes englobantes ; segmentation d'instance

La détection d'objets en traitement d'images est un problème qui consiste à prédire l'emplacement et la nature des objets présents dans une image. En vision par ordinateur, ce problème trouve par exemple des applications pour la détection d'obstacles (piétons, véhicules, poteaux...) dans des systèmes embarqués, et a ainsi été intensivement étudié dans la littérature (voir L. LIU et al. 2020 ; XIAO et al. 2020). Il est en général abordé sous l'angle de la régression de boîtes englobantes, qui sont également classifiées pour déterminer la nature de l'objet.

Ce problème est voisin de deux autres, la segmentation standard (appelée « sémantique » en vision par ordinateur), qui est une simple classification des voxels de l'image, et la segmentation d'instances, qui consiste à prédire autant de masques de segmentation qu'il y a d'objets dans l'image, ce qui permet notamment de séparer deux objets qui se chevauchent ou se touchent (voir la figure 4.1). Notons que la segmentation d'instance est un sous-problème de la détection d'objets.

En imagerie médicale, il peut être particulièrement utile d'aborder le problème de détection (et éventuellement caractérisation) automatique de lésions sous l'angle de la

segmentation d'instances plus que de la segmentation standard. En effet, les tumeurs doivent souvent être considérées individuellement, que ce soit pour suivre leur évolution au cours du temps, ou pour calculer des caractéristiques quantitatives (notamment dans le cadre de la chaîne de traitements détaillée à la page 9, qui est le fil conducteur de cette thèse). Une simple segmentation du tissu malade ne permet pas de différencier des tumeurs qui se touchent par exemple. Dans ce chapitre, on s'intéresse à la détection de lésions sous forme de boîtes englobantes, qui est une base pour la segmentation d'instances.

Les radiologues ayant besoin de plusieurs modalités pour caractériser une lésion (voir chapitre 1), il est naturel de vouloir les détecter automatiquement dans plusieurs modalités à la fois, à la manière de ce que l'on a étudié pour la segmentation du foie dans le chapitre 2. Comme pour mon travail sur la segmentation, un des objectifs est d'obtenir des détections plus précises en fournissant l'information des deux modalités au réseau. En effet, certaines lésions n'étant pas visibles (ou seulement partiellement) dans des images de certaines modalités, combiner l'information de plusieurs images peut être dans certains cas indispensable. Mais l'enjeu le plus important est d'avoir, pour chaque tumeur, son emplacement dans les deux images (comme expliqué dans la section 1.2.4 à la page 8). En effet cet emplacement peut être très différent si, comme dans le cas d'application qui nous intéresse, les différentes images ne sont pas acquises simultanément, et d'autant plus si les objets à détecter sont situés dans du tissu mou qui se déforme pendant la respiration. Je me limite dans ce chapitre au cas où les images sont acquises dans la même journée, et que la maladie n'a pas eu le temps d'évoluer significativement [1].

Cette contrainte demande de prédire, en plus des boîtes englobantes de toutes les tumeurs, une correspondance entre les boîtes prédites dans les deux modalités. Certains patients, notamment ceux atteints de cancers métastatiques (voir la figure 1.1 page 3), ont plusieurs dizaines de lésions dans le foie, de sorte qu'il est ardu de faire la correspondance entre les tumeurs des deux images. La détection jointe des lésions dans plusieurs images pourrait donc permettre aux radiologues de gagner un temps important.

Le but de ce chapitre est de proposer et d'étudier une méthode de détection de lésions hépatiques dans les paires d'images IRM pondérées en T1 et T2. Premièrement, je décris l'état de l'art en détection dans la section 4.1, avant de présenter la méthode que je propose dans la section 4.2. Ce chapitre est le résultat d'un travail préliminaire, qui vise à montrer la pertinence d'une telle méthode plutôt que ses performances, comme illustré dans les sections 4.3 et 4.4. Je discute dans la section 5.1.3 des expériences qu'il resterait à mener pour avoir une vue plus globale des atouts et limites de cette

1. Pour une discussion sur les enjeux liés à la détection jointe dans des images acquises à plusieurs mois d'intervalle, voir la section 5.2.3

méthode.

4.1 Etat de l'art en détection

4.1.1 Detection en Deep Learning

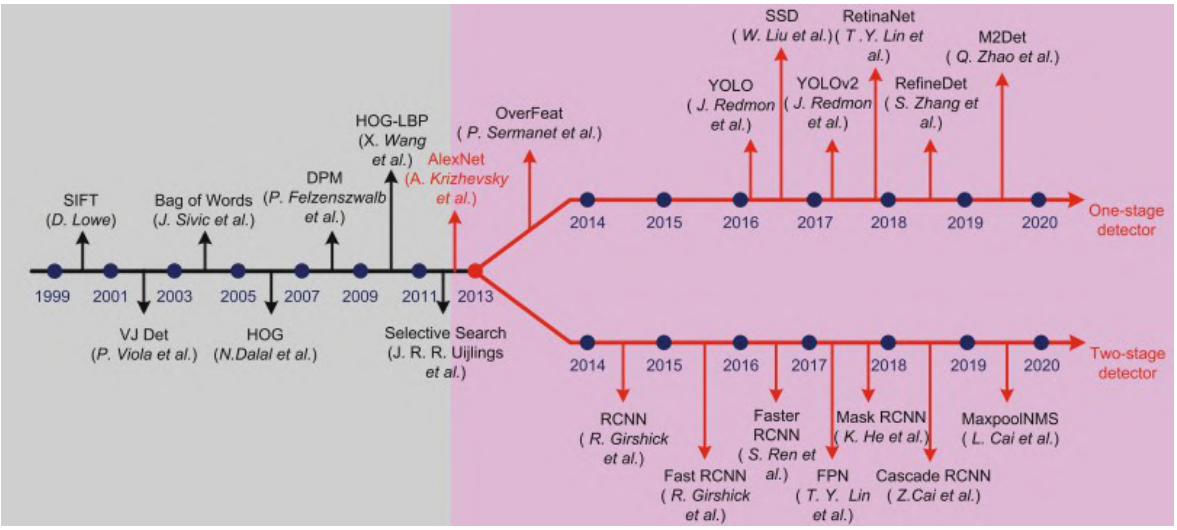


FIGURE 4.2 – Chronologie des innovations importantes pour la détection d'objets. Tirée de Z.-Q. ZHAO et al. 2019.

A plus forte raison encore que pour la segmentation ou la classification d'images, la majeure partie de l'innovation pour la détection automatique d'objets dans les images est portée par la recherche en vision par ordinateur. Cela peut s'expliquer notamment, comme on l'a dit, par les applications directes de ce problème à la conduite autonome, qui bénéficie d'incitations économiques importantes.

Toutes les revues de la littérature sur la détection d'objets (L. LIU et al. 2020 ; XIAO et al. 2020 ; Z.-Q. ZHAO et al. 2019) s'accordent à dire que les approches utilisant le Deep Learning ont supplanté les autres dans l'état de l'art et que deux branches sont maintenant en concurrence : l'une regroupant les méthodes à deux étapes, et l'autre celles à une seule étape (voir la figure 4.2 tirée de Z.-Q. ZHAO et al. (2019)).

Le principe des méthodes à deux étapes est dans un premier temps de déterminer des propositions d'objets, c'est à dire une liste de boîtes qui pourraient potentiellement englober des objets, et dans un second temps de classifier ces propositions tout en leur assignant un score de confiance, et en ajustant les bornes des boîtes proposées à la première étapes. Cette branche comprend principalement le RCNN et ses évolutions (pour

Region-based Convolutional Neural Network et proposé par Girshick et al. 2014), et notamment Fast-RCNN (Girshick 2015) et Faster-RCNN (Ren et al. 2015). Alors que le RCNN et le Fast-RCNN se concentraient sur la deuxième étape de classification, qui consiste à faire la classification et l'ajustement d'une liste de propositions, Faster-RCNN propose d'aborder la première également avec du Deep Learning, en introduisant le RPN, pour *Region Proposal Network*. Il introduit pour cela le concept d'*ancres*, que je détaille dans la section 4.2. Le Mask-RCNN de K. He, Gkioxari et al. 2017 est une variante très populaire du Faster-RCNN qui ajoute une branche de segmentation à la deuxième étape pour devenir un modèle de segmentation d'instance.

Le principe des méthodes à une seule étape est de déterminer la position, la classe, et les coordonnées de la boîte englobante des objets en une seule passe d'un réseau de neurones. Deux architectures dominent cette branche : YOLO (pour *You Only Look Once*) et ses évolutions (Redmon, Divvala et al. 2016 ; Redmon et Farhadi 2017), et SSD (pour *Single-Shot Detector*) et ses évolutions (C.-Y. Fu et al. 2017 ; Z. Li et F. Zhou 2017 ; W. Liu et al. 2016). Le principe de YOLO est de partitionner l'image et de prédire un nombre de fixe de boîtes englobantes par subdivision. SSD se base sur le principe du RPN, mais en utilisant plusieurs cartes de caractéristiques de différentes résolutions d'un réseau « colonne vertébrale » en même temps.

Cette astuce a pour but de répondre à une difficulté importante et inhérente aux objets des images naturelles, qui est leur grande variabilité de taille. Le FPN (pour *Feature Pyramid Network*), introduit par T.-Y. Lin, Dollár et al. (2017), exploite le même principe, en ajoutant des connexions remontantes et en l'adaptant à une architecture de type Faster-RCNN.

Le consensus qui semble maintenant établi (par exemple par T.-Y. Lin, Goyal et al. 2017 ; S. Zhang et al. 2018 ; Z.-Q. Zhao et al. 2019) est que l'approche en deux étapes donne des résultats plus précis, tandis que l'approche en une seule étape est plus rapide (ce qui peut être crucial pour certaines applications où l'on a besoin de détection en temps réel dans un flux vidéo, notamment pour la conduite autonome). Récemment, des méthodes ont tenté de réunir le meilleur des deux mondes, soit par des innovations dans l'architecture du réseau lui-même, comme S. Zhang et al. 2018 en imitant les deux étapes en un seul réseau, soit en concevant une fonction de coût plus adaptée au fort déséquilibre de classes inhérent aux méthodes de détection par classification d'ancres (comme le RPN et le SSD). Cette piste est celle de T.-Y. Lin, Goyal et al. 2017, qui introduit la fonction de coût *focale* (que je détaille dans la section 4.2.3), et soutiennent qu'un simple RPN utilisant la méthode FPN peut être aussi performant qu'un Faster-RCNN en utilisant cette fonction de coût. Ils appellent cette combinaison « Retina-net ».

On voit donc que le choix d'architecture est varié lorsqu'on s'attaque à un problème

de détection d'objets, contrairement par exemple à la segmentation où le choix du U-net semble maintenant assez standard, en tout cas pour les images médicales (voir section 2.1). Je choisis pour ce chapitre la méthode Retina-net comme base, majoritairement pour la simplicité d'implémentation d'une méthode à une seule étape, et parce que le problème de déséquilibre de classes peut s'avérer encore plus important en 3D qu'en 2D.

4.1.2 Détection en imagerie médicale

Les images IRM et CT ont la spécificité, par rapport aux images naturelles notamment, d'être en 3D, ce qui demande une adaptation méthodologique importante pour les traiter. Kern et Mastmeyer 2020 ont récemment publié une revue de littérature qui présente les différentes méthodes basées sur le Deep Learning utilisées pour la détection automatique dans les images volumiques (3D). Ils relèvent une majorité de méthodes reposant sur des réseaux de convolution 2D, en utilisant diverses stratégies pour les utiliser avec des images 3D (en général soit en traitant indépendemment les coupes selon les 3 axes, et en les recombinant dans un second temps, soit avec une approche dite « 2.5D », qui consiste à traiter trois (ou plus) coupes successives que l'on décale petit à petit. Voir la section IV.B de Kern et Mastmeyer 2020). Ils ne relèvent que deux articles présentant des méthodes se basant sur une implémentation de Faster-RCNN avec des convolutions 3D (Kaluva et al. 2020 ; X. Xu et al. 2019, respectivement pour la localisation d'organes et la détection de nodules dans le poumon), et un autre proposant sa propre architecture de détection (Wei et al. 2019, pour les lésions du foie).

Cette dernière approche basée sur des réseaux de convolution 3D m'intéresse cependant davantage, car elle me semble plus appropriée pour le problème que j'aborde dans ce chapitre (la détection jointe de tumeurs dans une paire d'images non-recalées). En effet, une même tumeur pouvant ne pas apparaître dans la même coupe dans les deux images, l'approche coupe par coupe semble moins pertinente.

Étonnamment, et malgré une approche qu'ils décrivent comme systématique, Kern et Mastmeyer (2020) semblent oublier une partie de l'état de l'art qui a émergé à la suite de l'introduction du Retina-Unet de Jaeger et al. (2020). Les auteurs proposent dans cet article d'utiliser les annotations de segmentation en faisant remonter la pyramide de caractéristiques de l'architecture Retina-Net jusqu'à la résolution de l'image (et la faisant ainsi ressembler à un U-net) pour apprendre en même temps à segmenter les objets à détecter. Le but est d'obtenir des détections plus précises en ajoutant de la supervision. Ils appliquent leur méthode à la détection de nodules du poumon avec une implémentation 3D du Retina-net, qu'ils comparent au Faster-RCNN et au Mask-RCNN. Depuis, plusieurs articles ont proposé des méthodes de segmentation d'instance se basant sur un Mask-RCNN 3D : segmentation du foie et du rein par C.-Y. Chen et al.

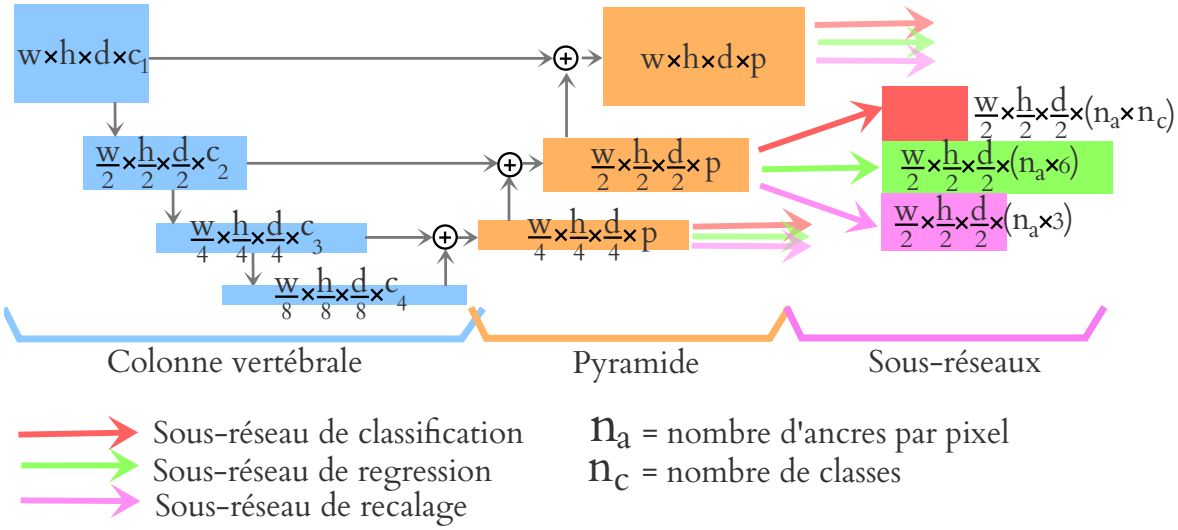


FIGURE 4.3 – Architecture Retina-net avec le sous-réseau de recalage. Les rectangles colorés représentent les cartes de caractéristiques. Les sorties des trois sous-réseaux ne sont représentées qu'au 2e niveau de pyramide pour des raisons de lisibilité.

(2019), détection de métastases dans le cerveau par LEI, TIAN et al. (2020), détection de nodules du poumon par KOPELOWITZ et ENGELHARD (2019) ou de tumeurs de le sein par LEI, X. HE et al. (2020).

De manière générale, on constate que l'utilisation de réseaux de convolution 3D pour la détection d'objets dans des images médicales volumiques est une tendance très récente (quelques articles datent de 2019, et la plupart de 2020). Pour comparer, en segmentation le U-net de RONNEBERGER, FISCHER et BROX 2015 n'a eu qu'un an à attendre pour se voir adapté en 3D (par MILLETARI, NAVAB et AHMADI 2016 notamment), alors que le Faster-RCNN date de 2015 comme le U-net (REN et al. 2015). Si ce retard est assez étonnant au vu de l'importance de la tâche de détection d'objets en imagerie médicale, l'adoption de réseaux 3D semble maintenant acquise, ce qui justifie mon choix d'utiliser cette technologie pour mon expérience décrite à la section 4.4.

4.2 Méthode de détection multimodale

4.2.1 Principe général de la méthode proposée

La méthode que je propose dans cette section se base sur l'architecture Retina-net, proposée par T.-Y. LIN, GOYAL et al. 2017. Le principe de cette architecture (illustrée figure 4.3) est de construire une « pyramide de caractéristiques » à partir d'un réseau

« colonne vertébrale », et d'associer à chaque voxel des cartes de caractéristiques de cette pyramide un nombre fixe d'« ancres », qui correspondent à des boîtes de différentes tailles et rapports de forme.À chaque niveau de la pyramide, un sous-réseau de classification et un sous-réseau de régression des boîtes est attaché.

Le sous-réseau de classification doit prédire, pour chaque ancre, si la boîte englobante d'un objet de l'image chevauche l'ancre au-dessus d'un certain seuil sur le taux de recouvrement, et éventuellement la classe de cet objet. Le sous-réseau de régression est quant à lui entraîné pour prédire, pour chaque ancre, l'écart des coordonnées de la boîte de l'objet le plus proche avec les coordonnées de l'ancre.

Dans toute la suite, on suppose que le réseau prend en entrée une paire d'images de modalité différente. L'idée de la méthode de détection multimodale que je propose dans cette section est d'ajouter à ces deux sous-réseaux un sous-réseau de recalage. Ce troisième sous-réseau est entraîné pour prédire le vecteur de déplacement entre l'emplacement de la lésion dans une image vers son emplacement dans l'autre image. Cet entraînement se fait de manière supervisée : les annotations sont effectuées en même temps dans les deux images, en prenant soin de faire correspondre les lésions dans les deux modalités. Ainsi, le vecteur de déplacement à prédire est directement disponible pendant l'entraînement. C'est la différence principale avec les méthodes de recalage par Deep Learning évoquées dans la section 2.1.5, qui cherchent à estimer le champ de déformation dans toute l'image : celles-ci ne peuvent pas compter sur des annotations manuelles de tout le champ de déformation, et minimisent donc en général une fonction de coût basée sur les intensités de l'image pour entraîner leurs réseaux.

Lors de l'entraînement, on attribue aux lésions de chaque modalité une classe différente. Pour la prédiction, on considère que deux boîtes détectées dans des modalités différentes correspondent à la même tumeur si la boîte de l'une translatée par le vecteur prédit par le sous-réseau de recalage chevauche l'autre au-dessus d'un certain seuil (voir figure 4.5).

4.2.2 Architecture

Étant donné un réseau colonne vertébrale, on note $B_k(x)$ les cartes de caractéristiques de ce réseau en sortie du k-ième bloc de convolution, si x est l'image d'entrée. On admet que l'architecture de la colonne vertébrale est telle qu'au niveau k, la résolution est diminuée d'un facteur 2^{k-1} par rapport à l'image originale (par exemple avec un *max-pooling* ou une convolution *stridée* dans chaque bloc). Pour la détection d'objets dans les images naturelles, ce réseau colonne vertébrale est en général un réseau de classification pré-entraîné (d'architecture VGG ou ResNet par exemple).

On construit le k-ième étage de la pyramide par l'opération $P_k(x) = p_k(B_k(x) +$

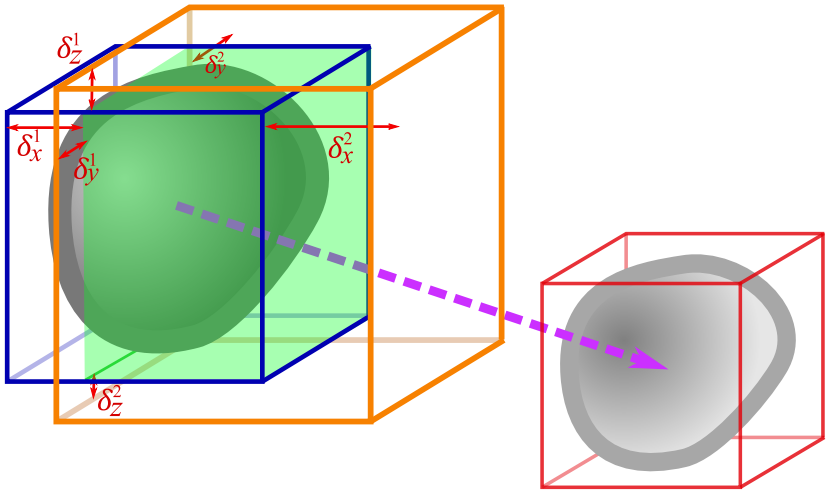


FIGURE 4.4 – Illustration du calcul des vérités terrain, pour une ancre représentée en orange. La boîte des annotations englobant l'objet le plus proche de l'ancre est représentée en bleu. Celle englobant le même objet dans l'autre image est représentée en rouge. La zone verte correspond à l'intersection entre la boîte bleue et l'ancre. Si le rapport entre le volume de l'intersection et le volume de l'union de l'ancre et de la boîte dépasse s_p, alors l'ancre sera considérée positive. Les double flèches rouges indiquent les valeurs à régresser par le sous-réseau R. La flèche pointillée mauve représente le vecteur de déplacement à estimer par le sous-réseau D.

$u_s(P_{k+1}(x)))$, où p_k est un bloc de convolution à résolution constante et u_s l'opération de sur-échantillonnage 2×2 (ou $2 \times 2 \times 2$ dans le cas 3D). Pour l'étage le plus profond de la pyramide K, $P_K = p_K(B_K(x) + u_s(B_{K+1}(x))$. En général, on ne remonte pas la pyramide jusqu'au niveau $k = 1$ de la résolution initiale. Le plus haut et le plus profond (K) étage de la pyramide sont des méthodes basées sur le principe du FPN.

On considère ensuite trois sous-réseaux, celui de classification C, celui de régression des boîtes R et celui d'estimation des déplacements D. Ils ne modifient pas la résolution d'entrée, de sorte que si l'entrée d'un sous-réseau $P_k(x)$ est un tenseur de dimension (w_k, h_k, d_k, p), ou w_k, h_k et d_k sont les dimensions d'espace et p le nombre de canaux, la sortie d'un sous-réseau sera un tenseur de dimension $(w_k, h_k, d_k, n_a \times n_r)$, où n_a correspond aux nombre d'ancres de base, et n_r dépend du sous-réseau.

Les ancres de base sont un ensemble de pavés (ou de rectangles dans le cas 2D) dont le nombre et les dimensions sont des méta-paramètres. À chaque coordonnée spatiale de chaque niveau de pyramide k, on associe n_a ancres, qui sont les ancres de base centrées sur cette coordonnée, et agrandies d'un facteur 2^{k-1}. Au total pour une image de taille w_1, h_1, d_1, on a donc $\sum_{k \in L} n_a w_k h_k d_k = \sum_{k \in L} n_a w_1 h_1 d_1 / 2^{3(k-1)}$ ancres, ou L (un autre

méta-paramètre) est l'ensemble des niveaux de la pyramide que l'on considère.

La figure 4.4 illustre comment sont calculées les valeurs à prédire par les trois sous-réseaux. Pour le sous-réseau de classification C, $n_r = 2n_c$, où n_c est le nombre de classes : pour chaque classe, l'objet détecté peut être soit dans la première image, soit dans la seconde. Pour mes expériences j'utilise $n_c = 1$, ce qui correspond au cas où l'on ne souhaite pas différencier les objets à détecter. Ainsi la sortie du réseau $C(P_k(x))_{i_x,i_y,i_z,(a\times c\times m)}$ est une estimation de la probabilité que l'ancre a à la position spatiale i_x, i_y, i_z englobe un objet de la classe c dans l'image m, avec $0 \leq a < n_a$, $0 \leq c < n_c$ et $m \in \{0, 1\}$.

Pour le sous-réseau de regression des boîtes englobantes R, $n_r = 6$ (ou $n_r = 4$ dans le cas 2D). Pour une ancre $a \in \{0, \ldots, n_a - 1\}$ à une position spatiale i_x, i_y, i_z et à un niveau $k \in L$, on note la sortie du sous-réseau de régression

$$(\delta_x^1, \delta_x^2, \delta_y^1, \delta_y^2, \delta_z^1, \delta_z^2) = R(P_k(x))_{i_x,i_y,i_z,a}$$

Pour chaque dimension $d \in \{x, y, z\}$, δ_d^1 et δ_d^2 correspondent à l'écart des bornes de la boîte englobante détectée avec celles de l'ancre, selon la dimension d.

Pour le sous-réseau d'estimation de déplacements D, $n_r = 3$ (ou $n_r = 2$ dans le cas 2D). Pour la même ancre, on note la sortie du sous-réseau de régression

$$(d_x, d_y, d_z) = 2^{k-1} D(P_k(x))_{i_x,i_y,i_z,a}$$

Ce sont les coordonnées du vecteur de déplacement entre l'objet détecté à l'ancre a et le même objet dans l'autre image. Le facteur 2^{k-1} permet de remettre le vecteur prédit à l'échelle de l'image.

4.2.3 Optimisation

Pour entraîner le modèle, on a besoin des valeurs cibles de classes, coordonnées de boîtes englobantes et vecteurs de déplacement pour chacune des ancres. Ces valeurs sont calculées à partir des boîtes englobantes annotées à la main sur l'ensemble d'entraînement. Pour chaque ancre, on détermine la boîte englobante issue de l'annotation la plus proche, c'est-à-dire celle qui a le plus grand rapport intersection/union avec elle. On considère qu'une ancre est *positive* si ce rapport dépasse un seuil s_p, et *négative* si ce rapport est en-dessous d'un autre seuil s_n. Entre s_n et s_p, on ignore cette ancre pour l'apprentissage. Dans la suite, on note IoU le rapport intersection/union d'une ancre avec la boîte annotée la plus proche.

Fonction de coût *focale* pour le classifieur

Le postulat de T.-Y. Lin, Goyal et al. (2017) est que tout problème de détection d'objet dans des images doit faire face à un fort déséquilibre de classes : le nombre d'ancres par image est très important (plusieurs par pixel de l'image), tandis qu'il n'y a que quelques dizaines d'objets au maximum par image, si bien que le nombre d'exemples négatifs dépasse largement le nombre d'exemples positifs.

Suivant T.-Y. Lin, Goyal et al. (2017), on utilise une fonction de coût *focale* pour entraîner le sous-réseau de classification. Cette fonction de coût a pour but, par rapport à une entropie croisée binaire classiquement utilisée pour les tâches de classification, de défavoriser les ancres faciles, c'est-à-dire les ancres que le réseau classifie correctement avec une forte confiance. L'hypothèse est que le réseau arrivera facilement à classifier correctement les exemples négatifs avec une forte confiance, et qu'une simple entropie croisée ne pénaliserait pas assez les faux négatifs par rapport aux vrais négatifs.

Si p est la probabilité qu'une ancre a contienne un objet de classe c estimée par le réseau, et $y \in \{0, 1\}$ la probabilité cible, la fonction de coût focale à la forme suivante :

$$\mathcal{L}_f(p, y, IoU) = \begin{cases} -(1-p)^\gamma \log(p) & \text{si } y = 1 \text{ et } IoU \geq s_p \\ 0 & \text{si } s_n < IoU < s_p \\ -p^\gamma \log(1-p) & \text{sinon} \end{cases}$$

où γ est un méta-paramètre qui contrôle le poids des exemples faciles dans la fonction de coût. Notons que si $IoU < s_n$ alors $y = 0$. Si $\gamma = 0$ on retrouve une entropie croisée binaire classique, et plus γ est grand, moins les exemples faciles vont avoir de poids dans le coût.

L_1 lissée pour le régresseur de boîtes englobantes

Pour entraîner le réseau régresseur de boîtes englobantes, on optimise une fonction de coût L_1 lissée. Ce choix est maintenant standard pour toutes les méthodes de détection à base d'ancres (Girshick 2015 ; T.-Y. Lin, Goyal et al. 2017).

Si $\hat{\delta}$ est l'écart prédit par le sous-réseau R entre une borne de l'ancre et celle d'une boîte, et δ l'écart réel (calculé par rapport à une boîte annotée), on entraîne le sous-réseau R en minimisant :

$$\mathcal{L}_r(\hat{\delta}, \delta, IoU) = \begin{cases} \frac{1}{2}(\delta - \hat{\delta})^2 & \text{si } |\delta - \hat{\delta}| < \alpha \text{ et } IoU \geq s_p \\ \alpha(|\delta - \hat{\delta}| - \frac{\alpha}{2}) & \text{si } |\delta - \hat{\delta}| \geq \alpha \text{ et } IoU \geq s_p \\ 0 & \text{si } IoU < s_p \end{cases}$$

où α est un méta-paramètre qui contrôle le seuil de transition entre le comportement L_2 et L_1.

Cette fonction de coût n'est sensible qu'aux ancres positives, et par conséquent n'est pas sensible au déséquilibre de classes.

Erreur quadratique moyenne pour l'estimation de vecteurs de déplacement

Si $\hat{d} \in \mathbb{R}^3$ est le vecteur de déplacement estimé pour l'ancre a, et d celui de référence, on optimise

$$\mathcal{L}_{eqm}(\hat{d}, d, IoU) = \begin{cases} ||\hat{d} - d||_2^2 & \text{si } IoU \geq s_p \\ 0 & \text{si } IoU < s_p \end{cases}$$

4.2.4 Inférence et post-traitement

Une boîte englobante issue des annotations va correspondre à plusieurs ancres positives (toutes celles dont le rapport intersection/union avec la boîte dépasse s_p). Ainsi, un réseau entraîné prédira un certain nombre de boîtes pour chaque objet détecté. Un post-traitement est donc nécessaire, et le plus populaire (depuis Girshick 2015) est l'algorithme de *non-maximum suppression* (NMS), que je décris rapidement ci-dessous :

À l'initialisation de l'algorithme, toutes les boîtes prédites par le réseau sont candidates. À chaque étape, on retient la boîte avec le plus grand score, et on la retire de la liste des candidates, avec toutes celles donc le rapport intersection/union dépasse un seuil s_{NMS} (qui est un paramètre de l'algorithme). À la fin on considère que chaque boîte retenue correspond à un objet détecté. Cet algorithme est effectué indépendamment pour toutes les classes.

Pour la correspondance entre les boîtes, on utilise le critère illustré sur la figure 4.5 : si la translatée par le vecteur détecté par D d'une boîte détectée dans une modalité et une boîte d'une autre modalité ont un rapport intersection/union supérieur à s_{NMS}, on considère que ces deux boîtes correspondent au même objet. Si aucune boîte de l'autre modalité ne correspond, alors on considère que la translatée de la boîte correspond à l'objet dans l'autre modalité. De cette manière tous les objets sont détectés en paires, et cela réduit la probabilité de faux négatifs.

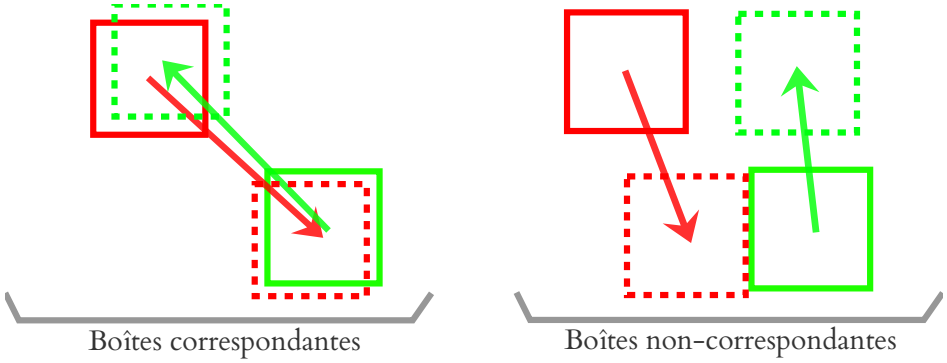


FIGURE 4.5 – Illustration du critère utilisé pour la correspondance des boîtes détectées. Les boîtes rouges et vertes en trait plein représentent des détections dans les images différentes, les flèches représentent la prédiction du sous-réseau de recalage, et les boîtes en pointillé représentent la translatée de la boîte pleine par le vecteur prédit.

4.3 Expérience avec des données synthétiques

Afin de tester la méthode dans un contexte où le réseau est plus simple à entraîner qu'avec des images 3D, j'ai mené une expérience préliminaire avec des données synthétiques.

Comme pour l'expérience décrite dans la section 2.3.3, les deux modalités sont simulées avec des images de sinusoïdes, l'une codée en orientation et l'autre en fréquence. On cherche à détecter des formes dans les deux modalités. Un exemple est illustré figure 4.6.

Génération des données

On part d'un champ de formes patatoïdales généré aléatoirement. Chaque composante connexe est numérotée, puis on déforme le champ numéroté qui servira de base à la deuxième modalité. Ainsi, à chaque forme d'une image correspond exactement une forme dans l'autre image (voir figure 4.7).

Ces champs sont ensuite utilisés pour générer les deux images de sinusoïdes : une modalité est codée en angle (les formes à détecter sont striées selon un angle compris entre O et $\pi/2$, et l'arrière plan entre $-\pi/2$ et 0 ; l'arrière plan et les formes ont la même fréquence tirée aléatoirement). L'autre modalité est codée en fréquence (les formes à détecter sont striées avec une période supérieure à 12 pixels, et l'arrière-plan une période inférieure à 12 pixels ; l'angle est tiré aléatoirement). La figure 4.6 montre

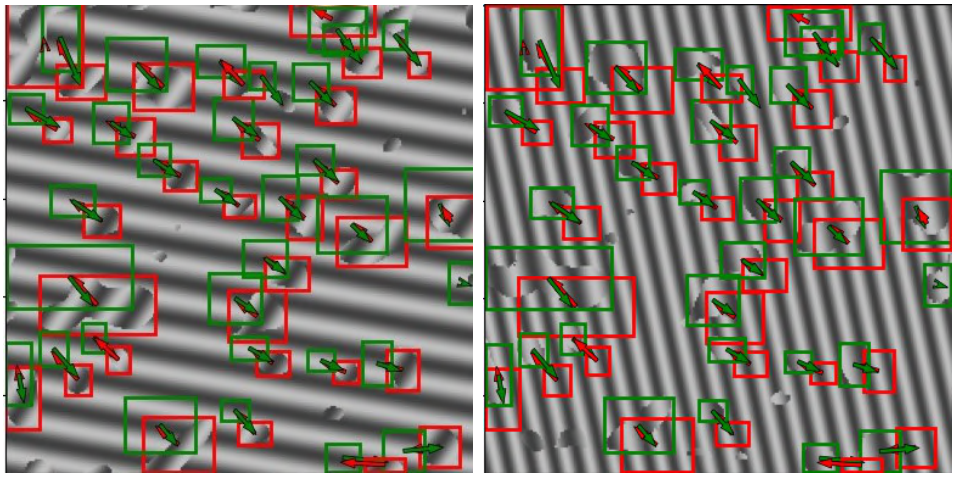


FIGURE 4.6 – Une paire d'images synthétiques pour tester la méthode de détection jointe. Les boîtes rouges correspondent aux détections dans la première modalité (image de gauche), les boîtes vertes dans la seconde (image de droite). Les flèches correspondent au déplacement prédit par le réseau.

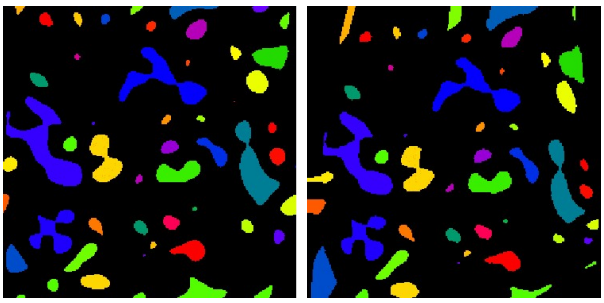


FIGURE 4.7 – Premières étapes de la génération de paires d'images synthétiques. Chaque couleur correspond à une composante connexe, et donc à un objet à détecter. A gauche, avant la déformation (modalité 1). A droite, après la déformation (modalité 2).

un exemple d'une paire d'images.

Les vérités terrain sont obtenues en calculant les bornes de chaque composante connexe, et les vecteurs de déplacements en soustrayant le centre de la boîte ainsi obtenue avec le centre de la boîte correspondante de l'autre image.

Paramètres

Pour cette expérience, j'utilise un ResNet-50 (K. HE, X. ZHANG et al. 2016) comme réseau colonne vertébrale. J'utilise les niveaux de pyramides 2, 3, 4 et 5. Les ancres

de base sont des rectangles de rapport 0.5 (deux fois plus larges que haut), 1 (carrés) et 2 (deux fois plus hauts que larges), mis à l'échelle d'un facteur 1, $2^{1/3}$ et $2^{2/3}$, soit $n_a = 9$ ancres par pixel. Au niveau 2, le coté de l'ancre de base carrée d'échelle 1 fait 16 pixels, jusqu'au niveau 5 où il en fait 128.

Le seuil s_n en-dessous duquel on considère qu'une ancre ne contient pas d'objet est fixé à 0,4, et le seuil s_p au dessus du quel on considère qu'une ancre en contient un est fixé à $0,6$.

Pour l'optimisation, j'utilise une fonction de coût focale avec $\gamma = 2$, et l'entraînement dure 1000 époques de 100 étapes. À chaque étape, une paire d'images est générée à la volée.

Ce sont des paramètres classiques pour la détection d'objets dans les images naturelles (proches de ceux utilisé par T.-Y. Lin, Goyal et al. (2017) notamment), en rajoutant le niveau de pyramide 2, pour les petites formes.

Résultats

La figure 4.6 montre un exemple de résultat de détection. On peut voir que la plupart des formes sont bien détectées, à l'exception des plus petites d'entre elles (qui ne passent pas le seuil s_p même pour les plus petites ancres), et que les déplacements prédits pointent bien sur la boîte correspondante de l'autre modalité.

Sur 100 paires d'images générées, le réseau obtient une sensibilité de 72% (en admettant qu'une boîte de la vérité terrain est détectée si le réseau en prédit une avec un rapport intersection/union supérieur à 0,6), et une spécificité de 92%. Cette sensibilité relativement basse est majoritairement due aux petites formes, que le réseau a tendance à manquer. Pour l'augmenter on pourrait soit rajouter le niveau 1 de la pyramide, soit ajouter une échelle plus petite que 1 à toutes les ancres de base.

En moyenne, le rapport intersection/union des boîtes détectées et des boîtes prédites est 0,89. Les boîtes prédites déplacées par le vecteur prédit coïncident avec les boîtes correspondantes de la vérité terrain avec un rapport intersection/union de 0,73. Ceci montre que le réseau recaleur a bien le comportement souhaité, car il difficile de s'attendre a un très haut score : La déformation entre les deux masques générés étant élastique, les formes des deux images ne se superposent pas exactement. De plus, certaines formes au bord sortent de l'image.

De manière générale, cette expérience tend à montrer que la méthode fonctionne correctement, à la fois pour détecter et faire la correspondance entre les formes, dans un cas simple en 2D.

4.4 Résultats préliminaires sur les données réelles

Les expériences que je présente dans la suite de cette section visent à étudier si la méthode peut fonctionner pour l'application qui nous intéresse, à savoir la détection de lésions hépatiques dans des paires d'images IRM pondérées en T1 et T2. Par rapport à l'expérience de la section précédente, la difficulté vient surtout d'une part du passage d'images 2D à 3D, et du passage de données que l'on peut générer à l'infini à un ensemble de données restreint.

Je présente des résultats préliminaires illustrant la faisabilité d'une telle approche. Des travaux complémentaires seraient nécessaires pour évaluer le potentiel de la méthode et feront l'objet d'une étude ultérieure.

4.4.1 Données et annotation

La base de données est la même que pour les expériences du chapitre 2, et est constituée de paires d'images IRM pondérées en T1 (temps d'injection portal) et T2, centrées sur le foie. Ces images sont acquises à quelques minutes d'intervalle et ne sont donc pas parfaitement recalées, à cause notamment de la respiration du patient.

J'ai moi-même annoté 48 paires d'images issues de 41 patients, à l'aide d'un outil permettant de visualiser simultanément les deux images (voir figure 4.8). Chaque lésion est ainsi localisée dans les deux images, et les dimensions des boîtes peuvent être ajustées. Chaque cas a en moyenne 6 lésions environ, le nombre de lésions variant de 1 à 54.

Je garde 5 paires pour tester l'algorithme.

De manière générale, à cause du manque de temps et de ma faible expérience pour reconnaître des tumeurs, ces annotations sont relativement peu précises et en faible quantité. On ne peut par conséquent pas s'attendre à d'excellentes performances de la méthode.

Comme pour mes expériences du chapitre 2, les images sont redimensionnées pour que les voxels fassent 3mm verticalement, et 1,5mm pour les deux dimensions horizontales. Pour faciliter la tâche de détection, j'applique un masque prédit par un réseau de segmentation du foie aux images, de manière à ce que seuls les voxels du foie n'apparaissent pas noirs (voir la figure 4.9 à la page 86). On utilise les masques prédits par un réseau pour que la méthode reste complètement automatique.

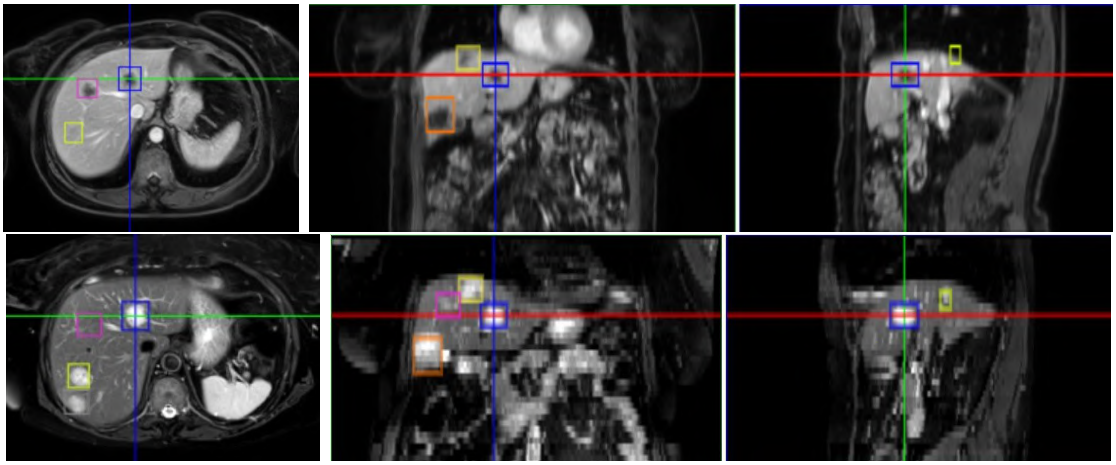


FIGURE 4.8 – Paire d'images de la base (la ligne du haut est l'image en T1, celle du bas l'image en T2) dans l'outil d'annotation. Les boîtes d'une même couleur correspondent à la même lésion. les lignes de couleurs rouges, vertes et bleues correspondent respectivement aux plans de coupes axiaux, coronaux et sagittaux

4.4.2 Choix des paramètres

Comme noté par JAEGER et al. (2020), la construction de la pyramide à partir d'un réseau colonne vertébrale ressemble à la partie montante d'un réseau d'architecture U-net. Je choisis donc comme colonne vertébrale la même architecture U-net que pour mes expériences du chapitre 2. J'utilise les cartes de caractéristiques des niveaux de pyramide 2, 3 et 4 comme entrée des trois sous-réseaux.

La taille, le nombre et les dimensions des ancres de base sont des paramètres aussi importants que le nombre de possibilités est grand. Je choisis seulement trois ancres de base qui ne diffèrent que par l'échelle. Le but, en prenant un faible nombre d'ancres par pixel, est de limiter le déséquilibre de classes entre ancres positives et ancres négatives. De plus, comme la plupart des tumeurs de la base de données sont à peu près sphériques, on peut choisir des ancres ayant le même aspect. Je choisis des ancres de tailles $(3; 4; 4)$, $(4; 6; 6)$ et $(5; 7.5; 7.5)$ voxels respectivement au niveau 2 de la pyramide. Je choisis une plus petite taille pour la première dimension (verticale), parce que cette dimension a une plus faible résolution et que par conséquent la plupart des boîtes de la vérité terrain ont une hauteur plus petite que leurs largeurs et profondeurs en nombre de voxels.

Les seuils s_n et s_p (qui déterminent si les ancres sont négatives ou positives en fonction du chevauchement de la boîte de la vérité terrain la plus proche) sont également potentiellement critiques pour le succès de la méthode. Il ne serait pas judicieux d'utiliser telles quelles les valeurs couramment utilisées pour la détection d'objets dans les images

naturelles (en général respectivement 0,4 et 0,6). En effet en rajoutant une dimension, les rapports intersection/union deviennent en moyenne bien plus faibles. En laissant ces seuils trop haut, on risque de n'avoir aucune ancre positive pour certaines boîtes de la vérité terrain. Si on les fixe à une valeur trop élevée au contraire, des ancres positives peuvent correspondre à plusieurs boîtes de l'annotation et rendre les tâches de régression et recalage ambiguës. J'ai choisi $s_n = s_p = 0,3$.

Quant à la fonction de coût focale, je choisis de garder $\gamma = 2$, comme préconisé par T.-Y. Lin, Goyal et al. (2017). Il serait toutefois intéressant d'essayer d'autres valeurs plus élevées pour ce paramètre, étant donné que le passage à la 3D augmente encore le déséquilibre de classes.

Pour augmenter artificiellement les données, j'applique des décalages d'intensité aléatoires aux images, et je translate les deux images de chaque paire aléatoirement, de manière à éviter que le réseau d'estimation des déplacements ne sur-apprenne.

J'effectue trois expériences : pour la première, on ne cherche à détecter que les boîtes dans l'image T1. Les boîtes de l'image T2 sont retrouvées grâce à l'estimation du déplacement des boîtes détectées dans l'image T1. La seconde est la même à modalité inversée, tandis que pour la troisième on estime les boîtes des deux images de la paire à la fois, comme pour l'expérience préliminaire décrite à la section 4.3.

Pour cette dernière, j'ai trouvé qu'utiliser un sous-réseau différent par classe plutôt qu'un seul pour les deux classes donnait de meilleurs résultats : on entraîne ainsi deux sous-réseaux de classification, deux sous-réseaux de régression des boîtes et deux sous-réseaux d'estimation des déplacements.

Il est à noter que la plupart de ces choix sont davantage basés sur l'intuition que sur l'expérience, et qu'un grand nombre de combinaisons de ces paramètres sont possibles.

4.4.3 Résultats et discussion

La figure 4.9 montre le résultat de détection d'un cas de la base de test par le réseau entraîné à détecter uniquement les tumeurs de l'image en T2. Pour les deux tumeurs sur lesquelles sont centrées les coupes présentées dans cette figure, on constate que les détections (les boîtes rouges) sont correctes, mais surtout que les déplacements prédits sont suffisamment précis pour que les boîtes translatées englobent bien la tumeur correspondante de l'image en T1. On remarque cependant la présence de quelques faux positifs et faux négatifs.

La figure 4.10 montre le résultat sur le même cas du réseau entraîné à détecter les deux tumeurs dans les deux modalités. On constate que les boîtes translatées se superposent

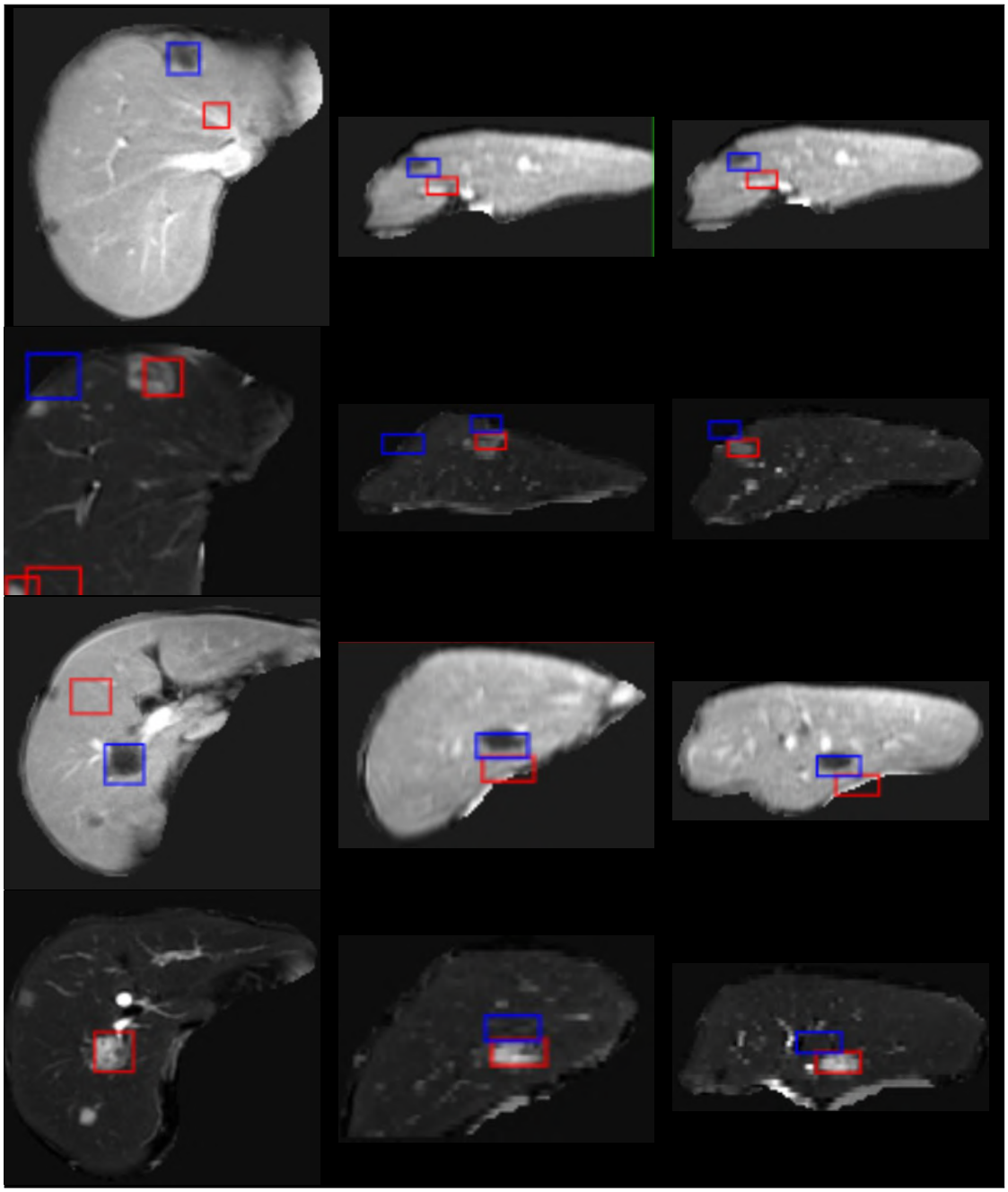


FIGURE 4.9 – Deux tumeurs détectées par le réseau entraîné sur les tumeurs en T2. Lignes 1 et 3 : images en T1. Lignes 2 et 4 : images en T2. Les colonnes de gauche, milieu et droite représentent respectivement les coupes axiales, coronales et sagittales. Les boîtes rouges représentent les tumeurs détectées dans l'image en T2, et les boîtes bleues les même boîtes translatées par le vecteur de déplacement prédit par le sous-réseau de recalage.

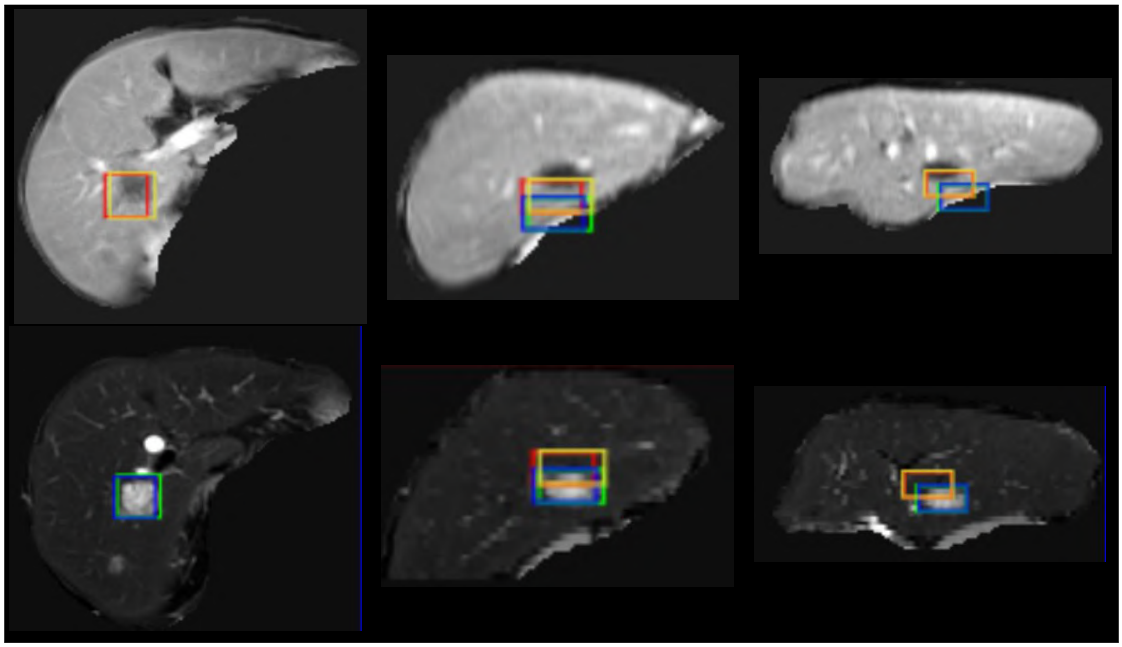


FIGURE 4.10 – Une tumeur détectée par le réseau entraîné sur les deux modalités. La boîte rouge correspond à la tumeur détectée dans l'image en T1, et la boîte bleue la même boîte translatée par le vecteur de déplacement prédit. La boîte verte correspond à la tumeur détectée dans l'image en T2, et la boîte jaune à la même boîte translatée.

assez bien sur les boîtes détectées de l'autre image (malgré une précision assez faible de la localisation de la tumeur dans l'image en T1).

La visualisation de ces résultats tend à montrer que la tâche d'estimation du déplacement, dont l'ajout est le cœur de l'innovation que je propose dans ce chapitre, semble fonctionner correctement. C'est la tâche de détection en elle-même, qui revient pour cette méthode à de la classification d'ancres, qui souffre encore de mauvaises performances.

Cependant, on a toutes les raisons de penser que la tâche de détection fonctionnerait mieux avec des annotations plus nombreuses et de meilleure qualité, ainsi qu'un peu de temps pour expérimenter les différentes combinaisons de paramètres, puisqu'un consensus semble maintenant s'être établi sur l'efficacité des méthodes de détection d'objets par les méthodes à base d'ancres (Faster R-CNN, et Mask R-CNN en tête).

Au total, les expériences décrites dans les sections 4.3 et 4.4 suggèrent que la détection jointe d'objets dans des paires d'images au moyen de l'ajout d'un sous-réseau de recalage fonctionne correctement. Du travail est encore nécessaire pour parvenir à une méthode réellement utilisable pour la détection de tumeurs dans le foie, mais la perspective de recevoir bientôt des annotations faites par un expert permet d'être

optimiste.

Chapitre 5

Conclusion

5.1 Contributions et discussion

5.1.1 Segmentation du foie en imagerie multi-modale

La segmentation multi-modale en Deep Learning peut s'appréhender de nombreuses manières, en répondant à des questions et problèmes très différents : la fusion de modalités (quelle architecture pour un réseau pouvant prendre en entrée des images issues de différentes distributions ?), l'adaptation de domaine non-supervisée (comment obtenir un réseau capable de segmenter plusieurs modalités, en ayant les annotations que pour une seule d'entre elles ?), et l'optimisation jointe avec le recalage principalement. Ces problèmes demandent des méthodologies très différentes selon que les données sont appariées ou non.

Dans le contexte de la chaîne de traitements pour la radiomique, le problème qui nous intéresse est celui de la segmentation avec des données appariées mais pas recalées, et qui a peu été traité dans la littérature jusqu'à récemment. Mes travaux reposent sur l'intuition selon laquelle l'information contenue dans les deux modalités peut être utile pour segmenter chacune des images, et qui vient de l'observation que regarder toutes les images peut parfois aider lorsqu'on en segmente une à la main. L'objectif initial était d'imaginer une méthode pour fusionner l'information contenue dans les deux images non recalées, en s'inspirant éventuellement des réseaux de type *spatial transformer* proposés par Jaderberg et al. (2015) et qui consistent à insérer des transformations géométriques dans les couches du réseau.

Toutefois cette idée n'a pas passé l'étape de l'expérience sur les données synthétiques, avec lesquelles les réseaux parviennent toujours à de bonnes performances, même sans de telles transformations géométriques. Ce résultat étonnant m'a amené à tester si un simple U-net pouvait s'aider d'une image en T1 pour segmenter plus précisément le

foie dans l'image en T2. Mais mes expériences tendent à montrer que le T2 contient bien toute l'information nécessaire pour y segmenter précisément le foie. Une autre application, où l'information serait plus diluée entre les modalités, aurait pu donner des résultats intéressants sur le comportement des réseaux dans ce cas. Il en va de même pour mon travail sur la similarité, où j'ai l'intuition que la méthode pourrait améliorer les performances pour une telle application, en guidant l'apprentissage à chercher l'information au bon endroit. On peut penser à des applications utilisant d'autres modalités fonctionnelles montrant très peu d'information anatomique (comme la tomographie par émission de positons (TEP)), associée à une modalité anatomique comme le CT (acquises avec des machines combinées TEP-CT, ou éventuellement après un pré-recalage manuel si les images sont acquises sur des machines différentes). Il aurait été intéressant de tester ces méthodes sur les lésions en IRM pondérée en T1 et T2, puisqu'elles peuvent avoir un aspect très différent dans ces deux séquences. Malheureusement je n'avais pas suffisamment d'annotations de lésions au moment de mon travail sur la segmentation.

Ces travaux m'ont poussé à m'interroger sur le futur de la recherche méthodologique en segmentation d'images. Est-ce que, depuis 2015 et l'introduction du U-net par Ronneberger, Fischer et Brox (2015), le problème serait résolu, et les progrès dans ce domaine ne seraient à attendre que du côté de l'effort d'annotation ? Ou plutôt, pour nuancer un peu, est-ce que tout gain de performance permis par une avancée méthodologique pourrait être simplement égalé par l'ajout de données ? C'est ce que suggèrent Hofmanninger et al. (2020), ainsi que les résultats de beaucoup de compétitions de segmentation (Bakas et al. 2018 ; Kavur et al. 2021 ; Simpson et al. 2019), puisque l'on constate que les équipes bien placées à ces compétitions utilisent en général un réseau proche du U-net original, en agrégeant éventuellement les prédictions d'un ensemble de ces réseaux (ce qui était aussi notre stratégie pour remporter la compétition de segmentation de cortex rénal présentée dans Couteaux, Si-Mohamed, Renard-Penna et al. (2019), voir l'annexe B). Notamment, Isensee, Petersen, Klein et al., avec leur nnUnet, qui est une procédure pour entraîner un simple U-net sur n'importe quel jeu de données (Isensee, Petersen, Klein et al. 2018), parviennent régulièrement à remporter ces compétitions de segmentation (Kavur et al. 2021 ; Simpson et al. 2019) ou se placer dans les premières places (Bakas et al. 2018), ce qui fournit un argument fort en faveur de cette affirmation. C'est ce que suggèrent aussi - plus humblement - les résultats de mes expériences sur les fonctions de coût adversaires (voir la section 2.2.4), puisque je n'ai pas réussi à obtenir un quelconque avantage en utilisant un apprentissage adversaire, technique pourtant très populaire en segmentation.

Toutefois, la collecte des données et surtout l'effort d'annotation peuvent être très coûteux, d'autant plus en imagerie médicale où les images 3D sont fastidieuses à seg-

menter à la main, et que ces segmentations doivent être faites par des experts dont le temps est précieux. Est-ce qu'alors, au contraire, les progrès méthodologiques, notamment dans des contextes de faible supervision (où les segmentations manuelles sont partielles et donc plus rapides à faire) ou semi-supervision (en tirant parti des images non annotées à l'apprentissage) seraient cruciaux ? Il est à noter que beaucoup d'innovations proposées ces dernières années en segmentation reposent sur l'ajout d'un terme de régularisation à l'apprentissage, souvent par apprentissage adversaire (voir section 2.2.3), par l'ajout, dans le cadre d'une optimisation conjointe, d'une tâche ne nécessitant pas d'annotations comme le recalage (voir section 2.1.4), ou en ajoutant des contraintes anatomiques à l'apprentissage (voir section 2.1.1). Cette régularisation peut dans la plupart des cas être utilisée pour faire de l'apprentissage semi ou faiblement supervisé, simplement en ne minimisant que le terme de régularisation sur les images qui ne sont pas annotées (ou les voxels qui ne sont pas annotés, dans le cas de l'apprentissage faiblement supervisé). Beaucoup d'idées, et notamment toutes sortes de contraintes anatomiques (emplacements relatifs, taille ou topologie des organes...), sont alors possibles, et surtout, potentiellement utiles.

Quoi qu'il en soit, la communauté pousse fortement vers l'innovation méthodologique : les conférences d'imagerie médicale font la part belle aux applications de méthodes proposées dans les dernières conférences de vision par ordinateur, souvent justifiées par un petit gain de performance sur un ensemble de données précis. Si cette tendance permet bien, à terme, de mettre en valeur les innovations qui résistent à l'épreuve du temps et qui ont par conséquent un impact réel sur un champ de recherche, on peut regretter que trop peu d'articles poussent ces innovations dans leur retranchements, en cherchant par exemple des données pour lesquelles elles seraient inutiles ou délétères. Par exemple, mes expériences sur le recalage par réseaux de neurones, que j'ai menées en parallèles de mes travaux sur la similarité, m'ont fait prendre conscience de certaines des limites de ces méthodes, notamment lorsque que les décalages sont importants. On trouve peu de mentions de ces limites dans les articles présentant des méthodes de recalage par Deep Learning (maintenant incontournables dans l'état de l'art), qui se contentent souvent d'évoquer leur performances sur une base de données précise. De plus, la course aux performances sur certains jeux de données publics (l'exemple du jeu de données de lésions du cerveau BRATS (Bakas et al. 2018) est sans doute le plus emblématique) aboutit à une sur-spécialisation des méthodes, les rendant difficilement généralisables à des jeux de données quelconques.

Cependant, d'autres tendances portées par la communauté vont dans le bon sens. Premièrement, l'essor des compétitions de segmentation, concours souvent proposés en marge des conférences d'imagerie médicale (comme les Journées Francophones de Radiologie, voir l'annexe), qui permettent à plusieurs équipes de proposer la méthode montrant les meilleures performances sur un problème de segmentation donné, en four-

nissant ainsi un cadre équitable pour réellement comparer les méthodes entre elles. Même si, en contraignant les participants à utiliser un même jeu de données, ces compétitions favorisent l'optimisation de méthodes sur un jeu de données précis et rendent ainsi ces résultats difficilement généralisables, elles permettent de mettre en évidence les tendances sur les innovations réellement bénéfiques. Deuxièmement, le partage de code source permet directement de reproduire les résultats mentionnés dans les articles, et de tester les méthodes proposées sur ses propres jeux de données. Mais on peut regretter que le partage des poids des réseaux entraînés reste une pratique encore assez rare, contrairement par exemple à la classification des images naturelles, tâche pour laquelle il est très facile de récupérer des réseaux pré-entraînés.

5.1.2 Interprétabilité des réseaux de segmentation

Mon travail sur l'interprétabilité des réseaux de segmentation m'a amené à me questionner sur les enjeux du Deep Learning interprétable. Jusqu'alors le jeune champ de recherche qu'est le Deep Learning interprétable s'était concentré sur l'interprétation de réseaux de classification, à l'exception de Bau et al. (2018) qui abordent les réseaux génératifs d'images naturelles.

La raison est en premier lieu historique, car c'est la classification d'images qui a amené les chercheurs en vision par ordinateur à développer le Deep Learning (LeCun, Boser et al. 1989), et cette technologie a connu ses premiers succès en s'attaquant à cette tâche (Krizhevsky, Sutskever et G. E. Hinton 2012). Elle restait encore jusqu'à récemment le premier moteur des innovations sur les architectures des réseaux de neurones (Inception par Szegedy et al. (2017), ResNet par K. He et al. (2016), DenseNet par G. Huang et al. (2017)). Une autre raison est le succès rencontré par les méthodes par cartes de saillance (voir la section 3.1), qui permettent de récupérer de l'information de localité sur des prédictions globales, ouvrant la voie à des méthodes de segmentation sémantique faiblement supervisées utilisant des réseaux entraînés pour classifier des images (Selvaraju et al. 2017 ; Simonyan, Vedaldi et Zisserman 2013).

Ce succès a entraîné une réduction du débat autour de l'interprétabilité à la question de l'explicabilité des prédictions : étant donné un réseau de classification et une de ses prédictions, quelle partie de l'image d'entrée a joué un rôle dans cette prédiction ?

En traitement de l'image et en science en général, faire des prédictions implique de modéliser un problème à l'aide de données d'une part, et de connaissances *a priori* que l'on a sur la régularité du problème d'autre part. Par exemple en mécanique newtonienne, prédire l'attraction de deux objets physiques suppose l'existence d'une force d'intensité proportionnelle au produit des masses et inversement proportionnelle au carré de la distance (connaissance *a priori*), et ce coefficient de proportionnalité est fixé par des observations (les données). Ces dernières années, notamment grâce aux progrès de

l'informatique, on est passé de modèles réalistes avec peu de paramètres fixés en fonction des données, comme la théorie de la gravité newtonienne mais aussi comme les algorithmes à base de minimisation d'énergie en traitement d'images « classique », à des modèles d'apprentissage automatique reposant sur des caractéristiques façonnées à la main en utilisant les connaissances du problème *a priori*, pour arriver à des modèles profonds aux millions de paramètres, où cette connaissance est limitée à guider le choix de l'architecture du réseau. En ce sens, le Deep Learning est l'aboutissement d'une évolution épistémologique qui vise à réduire l'importance des connaissances sur celle des données pour modéliser un problème [1]. Il est important de noter que cette évolution n'est pas limitée au traitement de l'image, ce qu'illustrent parfaitement les résultats spectaculaires obtenus avec du Deep Learning par Senior et al. (2020) sur le problème crucial de repli des protéines en biochimie, réputé très complexe. À mon sens, cette évolution pose des questions plus larges que la simple explicabilité des prédictions, en particulier à l'heure où l'on s'apprête à laisser ces algorithmes faire des diagnostics, ou leur laisser conduire nos voitures.

D'une part, la confiance en ces modèles est une question clef. Est-ce que de bonnes performances sur un ensemble de test suffisent pour avoir confiance en les prédictions que le réseau fera par la suite ? Qu'est-ce que le réseau a appris des données d'entraînement ? A-t-il réellement appris à reconnaître une tumeur ou un piéton ou a-t-il focalisé sur un biais des données ? Des travaux comme ceux de Kim et al. (2018) vont dans cette direction. Les auteurs ont proposé une méthode pour interpréter les réseaux de neurones de classification d'images naturelles en termes de concepts compréhensibles par des humains. Ils ont montré que dans certains cas, les réseaux retenaient les biais sexistes ou racistes présents dans les données.

D'autre part, une meilleure compréhension des réseaux est utile pour améliorer les méthodes existantes. Geirhos et al. (2018) ont montré que leurs réseaux privilégiaient l'information de texture à l'information de forme, et en ont tenu compte pour modifier la phase d'apprentissage et obtenir de meilleures performances.

C'est pour ces raisons que la recherche sur le Deep Learning interprétable doit s'élargir aux tâches pour lesquelles cette technologie a permis des progrès conséquents : en vision par ordinateur la segmentation sémantique, la détection d'objets et la synthèse d'images ; en traitement du son la reconnaissance et synthèse vocale ; en traitement du langage la traduction automatique et la génération de texte ; en imagerie médicale, la segmentation, le recalage et la détection, pour en citer quelques-unes. Cependant, si l'on peut appliquer les mêmes techniques de Deep Learning pour résoudre ces différentes tâches à quelques modifications près, la manière d'interpréter les modèles obtenus est

1. Sur les questions épistémologiques liées au Deep Learning, je conseille les leçons de Stéphane Mallat, disponibles sur la chaîne Youtube du Collège de France (https://www.youtube.com/watch?v=u8zKhpWoJPw)

beaucoup plus dépendante de la tâche. Par exemple, dans la section 3.1.1, j'ai évoqué pourquoi les méthodes de saillance, qui sont très populaires et utiles pour les réseaux de classification, sont peu pertinentes pour les réseaux de segmentation.

En proposant une méthode d'interprétabilité pour ces réseaux, j'ai donc également dû proposer une *manière* de les interpréter, en termes de *sensibilité* et *robustesse* à des caractéristiques. Cette méthode se base sur l'hypothèse qu'en trouvant l'image qui maximise l'activation des neurones de sortie, on peut en déduire quelles caractéristiques de l'image favorisent une réponse positive du réseau. Bien que les expériences que je décris à la section 3.3.3 corroborent cette hypothèse, celle-ci bénéficierait d'un meilleur fondement théorique.

On peut en outre trouver des limites à cette méthode. D'abord, cette estimation de sensibilité et robustesse est locale, c'est-à-dire qu'elle n'est valable qu'en un voisinage d'un point négatif choisi arbitrairement (voire figure 3.5). Il est tout à fait possible que le chemin de plus forte pente partant d'un autre point négatif ait une trajectoire très différente dans l'espace des caractéristiques. Une manière simple d'améliorer la méthode serait donc de calculer plusieurs trajectoires, en partant de points négatifs différents (c'est-à-dire, dans le cas de la segmentation de tumeurs, de choisir plusieurs coupes différentes, et plusieurs emplacements pour chaque coupe). On pourrait ainsi étudier la distribution de ces différentes trajectoires, et de mettre ainsi en évidence des tendances globales, plutôt que de se contenter d'une seule. Cela permettrait en outre de rendre la comparaison avec les valeurs normales calculées sur la base d'entraînement plus pertinente. J'ai choisi, en présentant ma méthode, d'en rester à l'analyse de trajectoires simples pour des raisons de simplicité et de clarté, d'autant que les expériences que je montre me semblaient suffisamment pertinentes avec une seule trajectoire. Une autre limite de la méthode est que l'idée de calculer l'évolution des caractéristiques pendant l'optimisation reporte la nécessité d'interpréter visuellement des images générées (comme pour les méthodes de visualisation classiques) à la nécessité d'interpréter des courbes, sans fournir de réponse claire et quantitative à la question de la sensibilité. Je pense toutefois que visualiser ces courbes facilite la tâche d'interprétation, surtout dans le cas d'images médicales, sur lesquelles le cerveau humain est moins entraîné que sur les images naturelles.

De plus, il demeure encore certaines idées autour de la maximisation d'activation à étudier, par exemple l'influence de l'emplacement du neurone à optimiser : le réseau se comporte-t-il différemment à proximité d'un bord du foie, ou juste à l'extérieur ? Il est également possible, et probablement intéressant, de faire l'étude des sensibilités et robustesses au voisinage d'un point positif réel (une tumeur annotée de la base pour continuer avec le même exemple), cette fois-ci en minimisant l'activation du neurone, par descente de gradient.

Si toutes ces pistes sont intéressantes à étudier pour répondre à la question de la sensibilité d'un réseau à des caractéristiques, il demeure que l'on peut imaginer beaucoup d'autres questions auxquelles tenter de répondre pour mieux comprendre les réseaux de neurones, et ainsi imaginer d'autres manières d'interpréter un tel réseau, tout en restant dans la définition de l'interprétabilité telle qu'on l'a donnée au début du chapitre 3. Par exemple : Que détecte un neurone en particulier ? À partir de quelle couche un U-net a-t-il discriminé le tissu sain du tissu malade ? Peut-on expliquer en termes compréhensibles l'influence des différents méta-paramètres ? Y a-t-il des filtres inutiles ?

Depuis la présentation de mon travail au *workshop* iMIMIC (Couteaux, Nempont et al. 2019), deux articles ont abordé l'interprétabilité des réseaux de segmentation. Natekar, Kori et Krishnamurthi (2020) comparent trois réseaux de segmentation de tumeurs du cerveau ayant des architectures différentes au moyen de plusieurs méthodes d'interprétabilité. L'une d'elles est similaire à celle que je propose (par maximisation d'activation), sans toutefois faire le suivi de l'évolution des caractéristiques de l'objet généré. Les auteurs concluent que les motifs obtenus sont difficiles à interpréter dans le cas de l'imagerie médicale, ce qui était aussi mon hypothèse (voir la section 3.1.2). Une autre est Grad-CAM (Selvaraju et al. 2017), qui est une méthode de saillance. Natekar, Kori et Krishnamurthi (2020) soutiennent que celle-ci peut tout de même être utile à l'interprétation des réseaux de segmentation car elle permet de décrire comment évoluent les régions d'intérêt au fur et à mesure des couches du réseau.

Santamaria-Pang et al. (2020) proposent une méthode de segmentation qui intègre un module de « langage émergent », dont le but est d'apprendre une représentation que l'on peut facilement corréler à des caractéristiques compréhensibles par les humains (comme la taille et l'excentricité).

5.1.3 Détection de tumeurs dans des images multi-modales

Comme évoqué dans la section 4.4, les méthodes de détection d'objets en Deep Learning à base d'ancres nécessitent l'ajustement d'un nombre important de méta-paramètres, en comparaison des méthodes de segmentation ou de classification d'images par exemple. De plus, le passage à trois dimensions demande de réestimer toutes les valeurs de ces paramètres, notamment parce que les rapports intersection/union entre deux boîtes ne sont pas comparables selon que l'on est en 2D ou en 3D. Ainsi, ce passage demande un travail assez long, d'autant plus qu'il faut 4 à 5 jours d'entraînement pour avoir les résultats d'une expérience, et j'ai manqué de temps pour explorer suffisamment de combinaisons. Les résultats que j'ai présentés dans le chapitre 4 me rendent cependant optimiste pour obtenir un modèle performant en détection de lésions hépatiques dans des paires d'images de modalités différentes.

La méthode Retina-net, sur laquelle je base mon travail présenté dans le chapitre 4 et proposée par T.-Y. Lin, Goyal et al. (2017), est en fait une simplification du très populaire Faster-RCNN de Ren et al. (2015). L'approche de ce dernier est constituée de deux étapes, où un réseau de type Fast-RCNN prend en entrée les boîtes proposées par un réseau de proposition de régions (RPN, pour *Region Proposal Network*). T.-Y. Lin, Goyal et al. (2017) soutiennent qu'utiliser une fonction de coût *focale* permet de se passer de la deuxième étape, en utilisant directement les sorties du RPN pour obtenir les résultats de détection. J'ai principalement choisi d'utiliser Retina-net comme base de travail pour sa simplicité d'implémentation, mais il est tout à fait possible d'ajouter un réseau Fast-RCNN au bout du modèle que je décris à la section 4.2 pour obtenir une architecture de type Faster-RCNN, plus standard en détection d'objets, et notamment en imagerie médicale (voir Kern et Mastmeyer 2020).

Pour l'application clinique qui nous intéresse, plus que la détection de tumeurs par boîtes englobantes, c'est le problème voisin de segmentation d'instances qui est pertinent. La méthode Mask-RCNN (K. He, Gkioxari et al. 2017), qui se base sur Faster-RCNN et qui consiste à rajouter une branche de segmentation au modèle Fast-RCNN de la deuxième étape, semble tout indiquée et compatible avec l'ajout du sous-réseau de recalage que je propose pour la détection jointe dans des paires d'images. C'est celle que j'essaierai dès que j'aurais accès aux masques de segmentation des tumeurs annotés par un radiologue. C'est aussi celle que mon équipe avait utilisée pour gagner la compétition de diagnostic de fissures du ménisque du genou à partir d'images IRM aux Journées Francophones de Radiologie (voir Couteaux, Si-Mohamed, Nempont et al. (2019) en annexe).

Ces masques de segmentation annotés pourront aussi me servir à essayer l'idée proposée par Jaeger et al. (2020), qui semble prometteuse et consiste à ajouter une couche de segmentation à un modèle de détection, dans le but de rajouter de la supervision au niveau des voxels, et ainsi guider la tâche de détection. Plus précisément, les auteurs proposent de faire remonter la pyramide de caractéristiques jusqu'au premier niveau (à la résolution d'entrée) pour faire une classification des voxels en tissu tumoral/tissu sain.

Une autre idée intéressante et que je n'ai pas eu le temps d'explorer consiste à ajouter une fonction de coût qui ne nécessite pas de supervision pour l'apprentissage du modèle. L'hypothèse est que si $B_1 \in \mathbb{R}^6$ est le vecteur des coordonnées d'une boîte de la modalité 1 prédit par le réseau, et $B'_1 = \tau(B_1, d_1)$ la même boîte translatée par le vecteur de déplacement prédit d_1, on veut que $IoU(B'_1, B_2)$ soit proche de 1, où B_2 est la boîte correspondante détectée dans l'autre modalité, et IoU est la fonction qui mesure le rapport intersection/union de deux boîtes. Une idée pour appliquer cette contrainte est la suivante : en notant $(d_x, d_y, d_z) = \lfloor C(P_{k_0}(X))_{i,j,k,a} \rfloor$ l'arrondi du déplacement prédit pour l'ancre a à la position spatiale de coordonnées (i, j, k) au niveau de pyramide k_0,

si cette ancre est classifiée comme positive par le sous-réseau C on optimise les trois sous-réseaux à la position $(i + d_x, i + d_y, i + d_z)$ et à l'ancre a, comme si la boîte B_1' faisait partie des annotations. En plus d'encourager le réseau à faire des détections plus cohérentes entre les modalités, cette fonction de coût non supervisée permettrait de tirer parti des paires d'images de la base qui ne sont pas annotées.

Quand j'aurais obtenu des performances de détection satisfaisantes, le réseau obtenu sera tout indiqué pour tester l'analyse de DeepDream (la méthode d'interprétabilité que je décris au chapitre 3). Il sera intéressant de l'utiliser pour estimer si les ancres de différentes tailles sont bien sensibles aux bonne tailles, et également pour tester si les différentes classes correspondant aux deux images sont bien sensibles à des caractéristiques spécifiques aux modalités qui correspondent. J'essaierai également de l'utiliser avec le sous-réseau de recalage, en générant une lésion dans les deux images avec un décalage précis.

De manière générale, beaucoup de travail est encore à faire et beaucoup d'idées sont encore à tester pour avoir un aperçu de tout le potentiel de cette méthode. Les masques de segmentation des tumeurs annotés par un radiologue que je vais bientôt avoir à ma disposition offrent d'excitantes perspectives d'amélioration.

5.2 Perspectives

5.2.1 Fin de la chaîne de traitement : extraction des descripteurs, et prédiction de la variable d'intérêt

Les travaux que j'ai présentés portent sur l'automatisation de trois premiers maillons de la chaîne de traitement pour la radiomique. La continuité naturelle de cette thèse consiste donc à étudier l'apport potentiel du Deep Learning sur les deux maillons suivants : l'extraction de descripteurs et la prédiction de la variable d'intérêt (survie, évolution de la maladie par exemple).

Dans l'approche radiomique classique, un ensemble de descripteurs fixe est utilisé. Comme je l'ai évoqué à la section 3.3.3, un ensemble de caractéristiques a été standardisé (Zwanenburg et al. 2020), qui comprend des descripteurs de forme, des statistiques sur les intensités et sur les textures. Ce travail de standardisation a pour principal but la reproductibilité, en fournissant des caractéristiques clairement définies. La dernière étape de la chaîne, la prédiction d'une variable, est ensuite faite à l'aide de techniques d'apprentissage automatique telles que des méthodes à vecteurs de support, des arbres de décision, des régressions linéaires ou logistiques, qui prennent en entrée tout ou une partie (qu'on appelle dans ce cas une « signature ») des caractéristiques calculées.

Cette approche en deux temps (calcul de descripteurs façonnés à la main, qu'on met en entrée d'un classifieur *shallow*) a longtemps été l'approche prépondérante en classification d'image naturelles, avec des caractéristiques comme les histogrammes de gradients (HOG) ou les SIFT. L'approche *deep*, popularisée par Krizhevsky, Sutskever et G. E. Hinton 2012 et incontournable depuis, consiste à s'affranchir de l'étape de façonnage des caractéristiques en utilisant des modèles prenant directement l'image en entrée. Ceux-ci (les réseaux de neurones convolutionnels) sont capables d'apprendre une représentation des images qu'ils prennent en entrée adaptée à la tâche pour laquelle ils sont entraînés.

Devant un tel succès pour les images naturelles, l'approche *deep* en une étape est maintenant largement utilisée pour l'aide au diagnostic (voir la revue de bibliographie de Fujita 2020), ou pour faire des pronostics à partir des images (Ravichandran et al. 2018 ; Ypsilantis et al. 2015 par exemple). Le terme de *discovery radiomics* est parfois employé pour désigner cette méthode (Kumar et al. 2017).

Cette approche a néanmoins des inconvénients par rapport à la méthode classique. Outre les problèmes d'interprétabilité (qui était l'objet du chapitre 3) et de reproductibilité, qu'on a déjà évoqués, la quantité de données annotées nécessaire pour entraîner ces modèles peut être rédhibitoire pour certaines applications.

Pour tirer parti du pouvoir de représentation des réseaux convolutionnels avec une quantité de données réduite, une idée est alors d'utiliser les caractéristiques apprises par un réseau sur une autre tâche, comme la classification d'images naturelles. Par exemple, Huynh, H. Li et Giger 2016 utilisent les cartes de caractéristiques du réseau AlexNet (celui de Krizhevsky, Sutskever et G. E. Hinton 2012) auxquelles ils ajoutent des caractéristiques radiomiques standard, pour classifier des tumeurs dans des images mammographiques. Cependant, on sait que si ces modèles sont si performants, c'est en partie parce qu'ils sont capables d'apprendre des représentations qui s'affranchissent de l'information inutile pour la tâche pour laquelle ils sont entraînés. À l'inverse des caractéristiques radiomiques, qui sont conçues pour décrire les objets d'une image de la manière la plus complète possible, les caractéristiques apprises par des réseaux de classification ne fourniront donc qu'une représentation partielle de l'image.

Pour trouver des caractéristiques à la fois apprises (et donc potentiellement meilleures que les caractéristiques standards façonnées à la main) et générales (et donc utilisables avec un modèle d'apprentissage automatique *shallow* pour différentes tâches), l'apprentissage de représentations a peut-être des réponses à apporter. Ce champ de recherche a pour objet d'étude les méthodes permettant d'apprendre une représentation à partir des données, c'est-à-dire un ensemble de caractéristiques qui permettent de décrire chaque image en conservant le plus d'informations possible sur elle. Celui-ci a beaucoup progressé depuis l'essor du Deep Learning notamment grâce aux modèles de type

auto-encodeurs.

Un auto-encodeur est un réseau dont la tâche est de prédire exactement la même image que celle qui lui est fournie. Plus précisément, un modèle de ce type est constitué d'un réseau encodeur, qui prédit une représentation dite *latente* de basse dimension à partir d'une image, et d'un réseau décodeur qui génère une image à partir d'une représentation. L'encodeur et le décodeur sont entraînés simultanément pour minimiser l'erreur de reconstruction des images. Ainsi, le réseau apprend une représentation de basse dimension qui conserve le plus possible d'information sur l'image d'entrée, le tout sans supervision.

Une variante aujourd'hui incontournable est l'auto-encodeur *variationnel* (proposé par Kingma et Welling 2013) dont le principe est de contraindre la représentation latente à suivre une distribution fixe (souvent une loi normale centrée de variance égale à 1). L'intérêt est qu'avec une telle contrainte, tout échantillon tiré selon cette distribution pourra être décodé en une image plausible. De cette manière, chaque dimension de l'espace latent représente une caractéristique qui varie continuellement selon les échantillons de la base de données. On peut citer d'autres méthodes qui reposent sur ce principe, qui diffèrent de l'auto-encodeur variationnel par la manière de contraindre l'espace latent à suivre une distribution : alors que l'autoencodeur variationnel minimise une distance de Kullback-Leibler sur chacun des échantillons, les auto-encodeurs de Wasserstein (Kolouri et al. 2018 ; Tolstikhin et al. 2017) s'inspirent de la théorie du transport optimal et les auto-encodeurs adversaires (Makhzani et al. 2015) utilisent une fonction de coût adversaire. La figure 5.1 montre l'espace latent à 3 dimensions que j'ai obtenu avec le jeu de données MNIST (LeCun, Bottou et al. 1998), qui sont des chiffres manuscrits.

Un enjeu important de ce domaine de recherche est celui du démêlage (*disentanglement* en anglais). L'hypothèse est qu'un jeu d'images vit sur une variété de basse dimension, et que chacune des dimensions de cette variété hypothétique représente une « vraie » caractéristique des images. Par exemple si l'on a des images de visages, une dimension pourrait représenter la longueur, des cheveux, une autre la teinte, une autre la barbe, l'âge ou les lunettes par exemple. Si l'on entraîne un auto-encodeur avec une des méthodes mentionnées plus haut, ces caractéristiques se retrouvent « emmêlées » dans la représentation latente. Le problème de démêlage consiste alors à trouver une représentation dont chacune des dimensions représente une caractéristique interprétable.

C'est la piste que j'ai commencé à explorer au début de ma thèse. Mon idée était d'entraîner un tel modèle pour trouver une représentation de basse dimension pour les tumeurs du foie en CT, avec certaines dimensions réservées pour des caractéristiques connues comme la taille ou la forme. Ainsi, on pourrait obtenir un ensemble réduit de caractéristiques qui décrit la tumeur avec une perte d'information limitée, tout en

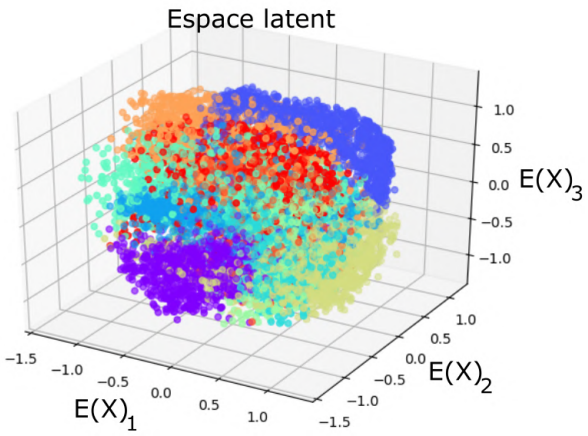


FIGURE 5.1 – Visualisation de l'espace latent à trois dimensions d'un auto-encodeur de Wasserstein entraîné sur le jeu de données MNIST. Chaque point représente un échantillon de la base de test, chaque couleur représentant un chiffre.

gardant une certaine interprétabilité.

J'ai réalisé quelques expériences qui montrent le potentiel de ces méthodes. Avec un jeu de données jouet, qui sont des images de cercle de taille et d'intensité variables, les résultats sont montrés sur la figure 5.2. Pour cette expérience j'utilise un auto-encodeur de Wasserstein avec un espace latent de dimension 2, comme on sait que les images vivent réellement sur une variété de dimension 2 (l'intensité et le diamètre étant indépendants). Pour démêler, on veut rajouter un peu de supervision pour qu'une des dimension de l'espace latent coïncide avec une grandeur que l'on connaît (ici le diamètre). Plutôt que d'ajouter une fonction de coût pour contraindre cette dimension (à la manière de MAKHZANI et al. 2015), j'ai trouvé une autre manière de faire qui donne de bons résultats (voir les figures 5.2d et 5.2e) : si $X \in \mathbb{R}^{N \times N}$ est une image de la base, $g(X) \in \mathbb{R}$ la grandeur que l'on veut encoder et $E(X)_0, ..., E(X)_{d-1}$ les d caractéristiques latentes, pour une certaine quantité d'images de la base (j'avais choisi 1/5 des échantillons) on minimise

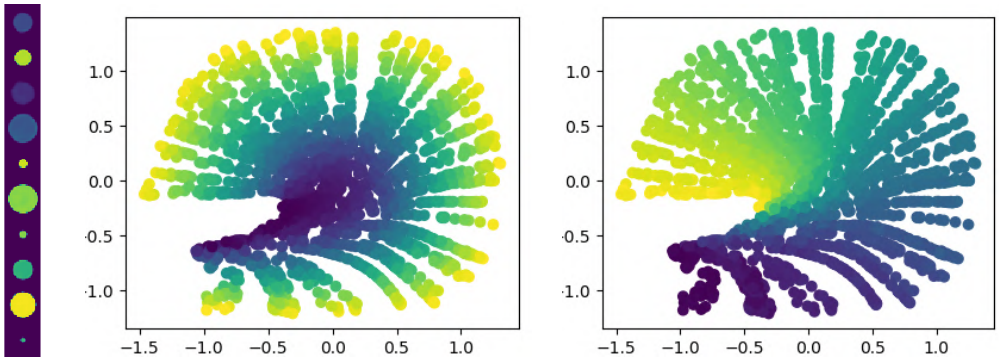
$$||X - D(g(X), E(X)_1, ..., E(X)_{d-1})||$$

au lieu de

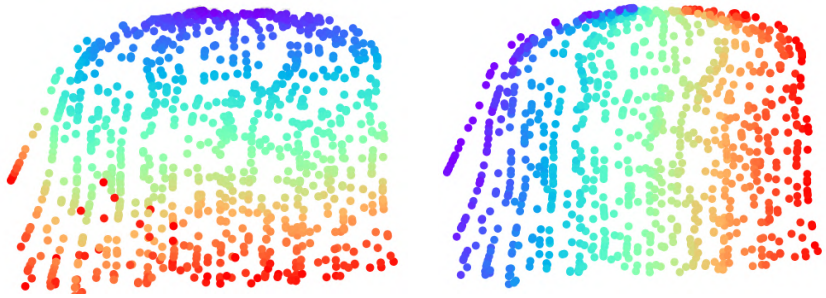
$$||X - D(E(X)_0, E(X)_1, ..., E(X)_{d-1})||$$

pour le reste des échantillons, où $D : \mathbb{R}^d \to \mathbb{R}^{N \times N}$ est le réseau décodeur qui génère une image à partir d'une représentation latente.

J'ai également appliqué le même principe sur une base de données de coupes CT centrées sur la tumeur, en utilisant un espace latent à 15 dimensions et en faisant



(a) (b) intensité sans démêlage (c) diamètre sans démêlage



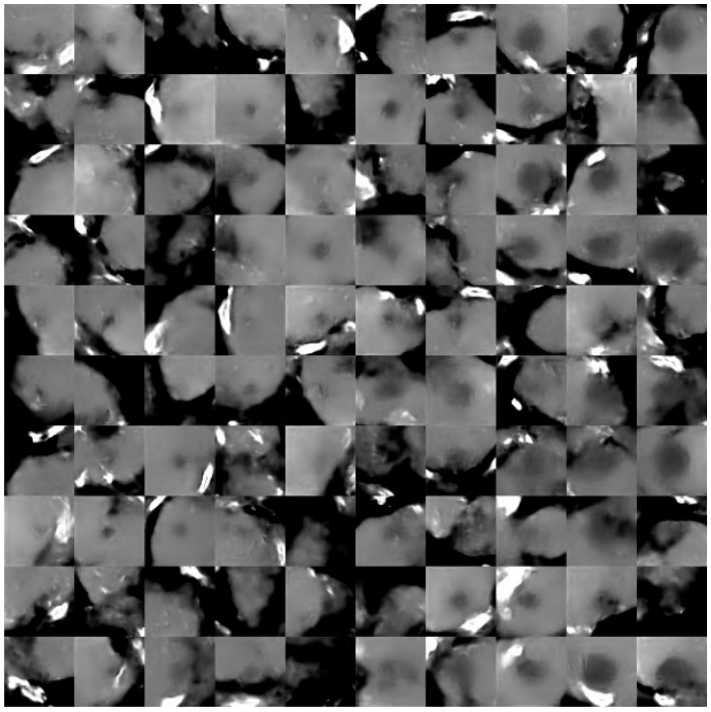
(d) intensité avec démêlage (e) diamètre avec démêlage

FIGURE 5.2 – Expérience de démêlage avec un jeu de données jouet. (a) : 10 exemples de données d'entraînement. (b), (c) , (d) et (e) : espace latent d'un auto-encodeur. Chaque point représente un échantillon. (b) et (c) : sans démêlage. (d) et (e) : avec démêlage. (b) et (d) : la couleur représente l'intensité. (c) et (e) : la couleur représente le diamètre.

coïncider la première avec le diamètre de la tumeur. Des échantillons générés avec ce réseau sont visibles sur la figure 5.3a.

Pour savoir si cette méthode permettait d'encoder des caractéristiques très abstraites, j'ai aussi fait l'expérience avec une base de données de visages en essayant d'encoder l'âge sur une des dimensions. Pour cette expérience j'ai utilisé un auto-encodeur variationnel *introspectif*, tel que proposé par H. HUANG et al. (2018). Ce modèle utilise le principe de l'apprentissage adversaire pour obtenir des échantillons générés plus réalistes. La figure 5.4 montre les résultats de cette expérience. On voit clairement que la variable d'âge a été encodée sur la première dimension de l'espace latent, et que les autres dimensions retiennent suffisamment d'information pour correctement recomposer le visage initial. On peut ainsi modifier l'âge d'un visage.

Je me suis aussi demandé si ces modèles pouvaient être utiles pour apprendre directement à régresser la variable d'intérêt (la survie par exemple) lorsque la quantité



(a) Échantillons générés aléatoirement par un auto-encodeur de Wasserstein entraîné sur des coupes CT centrées sur des tumeurs. De gauche à droite, la première dimension, qui encode le diamètre de la tumeur, augmente linéairement tandis que les autres sont tirées aléatoirement.

de données annotée est faible. La rareté des annotations est, comme on l'a déjà dit, un obstacle majeur pour l'application d'algorithmes de Deep Learning aux problèmes de diagnostic ou pronostic assistés par ordinateur. Autrement dit, les auto-encodeurs variationnels permettent-ils de faire de l'apprentissage semi-supervisé ? En prenant la tâche de régression de l'âge à partir de photos de visages comme point de départ, je suis parvenu à des résultats corrects avec seulement 3% d'annotations sur une base de 20000 images (une corrélation de 0,76 entre les âges prédits et les vérités terrain), alors qu'un réseau entraîné de manière totalement supervisée ne convergeait pas du tout avec si peu de données annotées.

Si ces quelques résultats suggèrent que la piste des auto-encodeurs génératifs (variationnels, de Wasserstein ou introspectifs) est prometteuse, beaucoup de chemin reste à parcourir avant qu'elle fournisse un ensemble de caractéristiques à la fois complet (qui rende compte de toute l'information contenue dans l'image d'une tumeur), de basse dimension, et qui ait un pouvoir de prédiction important pour plusieurs tâches différentes.

La suite naturelle à ce travail consisterait donc à appliquer l'une de ces méthodes à un ensemble d'images avec une variable à prédire disponible, et de comparer le pouvoir

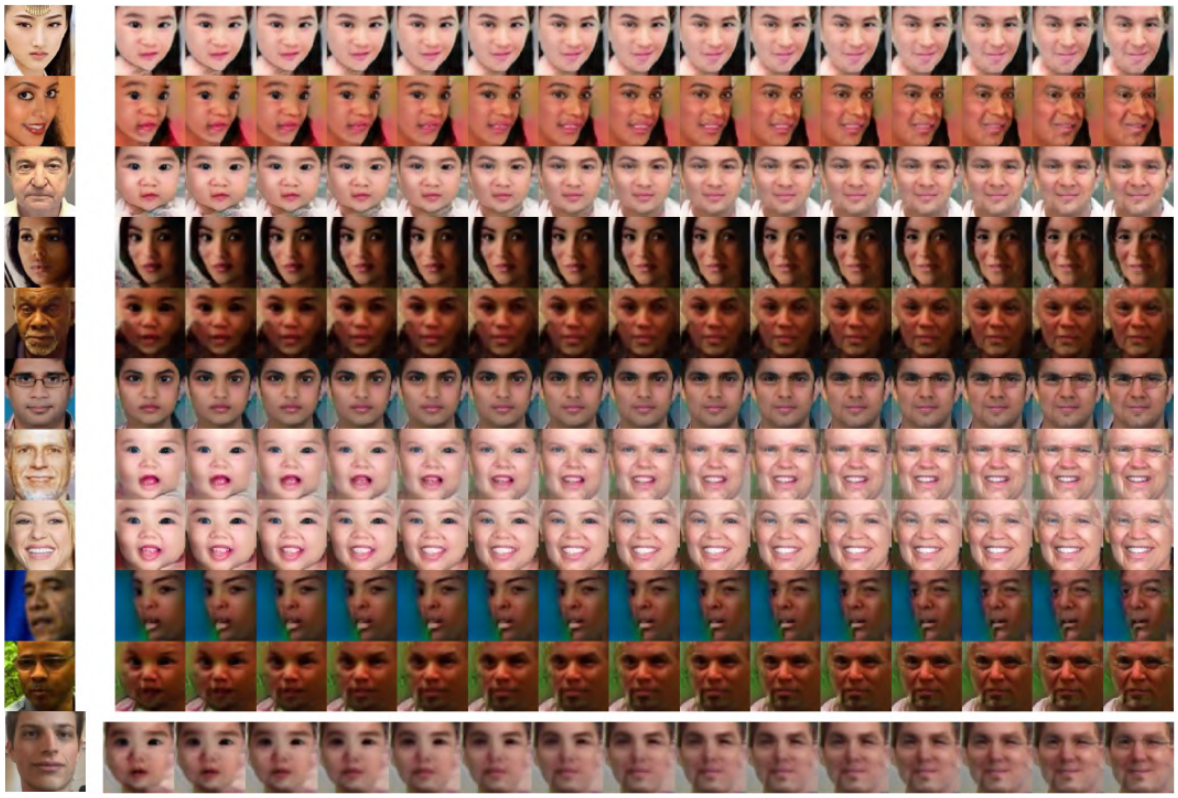


FIGURE 5.4 – Reconstruction de visages avec un auto-encodeur entraîné pour encoder l'âge sur une caractéristique latente. À gauche, les visages originaux de la base de test. À droite, les visages reconstruits, en changeant la caractéristique liée à l'âge.

prédictif des caractéristiques ainsi obtenues (avec ou sans guidage semi-supervisé) avec les caractéristiques radiomiques standard.

On peut également penser à d'autres pistes pour bénéficier de l'efficacité du Deep Learning avec peu de données annotées, et notamment aux méthodes d'apprentissage semi-supervisé à base d'apprentissage adversaire, comme ODENA 2016.

5.2.2 Interprétabilité des réseaux de recalage

Alors que la recherche sur le recalage en Deep Learning propose une littérature de plus en plus fournie (voir Y. FU et al. 2020), on manque toujours, à ma connaissance, de la moindre méthode qui permette d'interpréter les réseaux de recalage, tout comme il n'y avait pas de méthodes pour interpréter les réseaux de segmentation au moment où j'ai commencé à travailler sur l'interprétabilité.

La compréhension des réseaux de recalage me paraît pourtant particulièrement inté-

ressante, puisque l'idée communément admise du fonctionnement des réseaux de classification (les premières couches détectent des caractéristiques de bas niveau, comme les gradients ou les contours, et les couches suivantes détectent des caractéristiques de plus en plus abstraites, comme des textures et même des formes, jusqu'à prédire la classe de l'image) semble difficilement applicable à ce type de réseau. S'il est maintenant admis que ces réseaux permettent d'obtenir des recalages précis, aucun travail à ma connaissance ne s'intéresse à *comment* ils fonctionnent.

La première étape pour concevoir une méthode d'interprétabilité est de chercher une question à laquelle la méthode essaiera d'apporter une réponse, en se basant sur une hypothèse de fonctionnement. Dans le cas des réseaux de segmentation, j'avais proposé au chapitre 3 d'essayer de répondre à la question « à quelle caractéristiques compréhensible par un humain un réseau est-il sensible ? ». L'hypothèse sous-jacente est qu'un réseau prend ses décisions de classification en fonction de certaines caractéristiques qu'ont les objets présents dans l'image.

Pour les réseaux de recalage, une idée intuitive de leur fonctionnement est qu'ils repèrent des points d'intérêt dans les deux images à recaler, puis en font la correspondance pour *in fine* produire un champ de déplacement. Si tel était le cas, déterminer ou au moins localiser ces paires de points d'intérêts se correspondant fournirait un éclaircissement intéressant vers le fonctionnement d'un tel réseau. Peut-être qu'utiliser des méthodes d'attribution à base de gradient sur un des vecteurs du champ de déplacement prédit pourrait fournir cette information, même si rien ne garantit qu'une seule paire de points d'intérêt suffirait à expliquer ce vecteur. Toutefois, rien à ce jour ne semble corroborer cette hypothèse, et il est tout à fait possible que les réseaux de recalage ne s'appuient pas du tout sur une telle correspondance de points d'intérêt.

Une autre question potentiellement intéressante serait : quels types de déformations une architecture est-elle capable de régresser ? En particulier, mes expériences de segmentation non appariée suggèrent que l'amplitude des décalages est un facteur critique pour qu'un réseau de neurones apprenne à faire correspondre l'information de deux images décalées. Il serait intéressant de creuser cette piste, en étudiant quels paramètres pourraient avoir une influence sur l'amplitude maximale qu'un réseau serait capable de prédire, et notamment la profondeur de l'architecture puisque le champ réceptif des neurones augmente avec la profondeur.

Au total, cette piste n'est pour l'instant faite que de questions ouvertes. Cependant, ce travail me semble important pour que le champ de recherche du recalage par Deep Learning progresse, et pour que la recherche en Deep Learning interprétable ne se limite plus aux réseaux de classification.

5.2.3 Identification de lésions pour le suivi longitudinal

Le problème d'identification de lésions dans des images de modalités différentes acquises à quelques minutes d'intervalle, qui était l'objet du chapitre 4, n'est qu'une étape vers le problème d'identification de lésions dans des images acquises à des dates différentes. Comme évoqué en introduction, c'est l'évolution des tumeurs qui intéresse le plus les radiologues pour l'établissement du pronostic, plus que leur aspect à un instant donné. Ce problème est peut-être par conséquent plus important encore pour l'application clinique, mais également plus difficile parce qu'il demande de relever de nombreux défis, décrits ci-dessous.

D'abord, le changement potentiel de taille et d'aspect des lésions implique qu'il ne suffit plus d'estimer un déplacement pour identifier une même lésion dans deux images, mais qu'il faudrait également réestimer la taille et la forme des boîtes englobantes.

Ensuite, il serait nécessaire de prendre en compte la possibilité que des lésions apparaissent et disparaissent d'une date à l'autre, et par conséquent que seules certaines des détections doivent être faites par paire.

Enfin, l'annotation elle-même devient problématique puisqu'il n'est pas aisé, même pour une personne entraînée, de déterminer si des lésions apparaissant dans deux images différentes correspondent, surtout en présence de plusieurs dizaines de lésions, comme sur le cas présenté dans la figure 5.5.

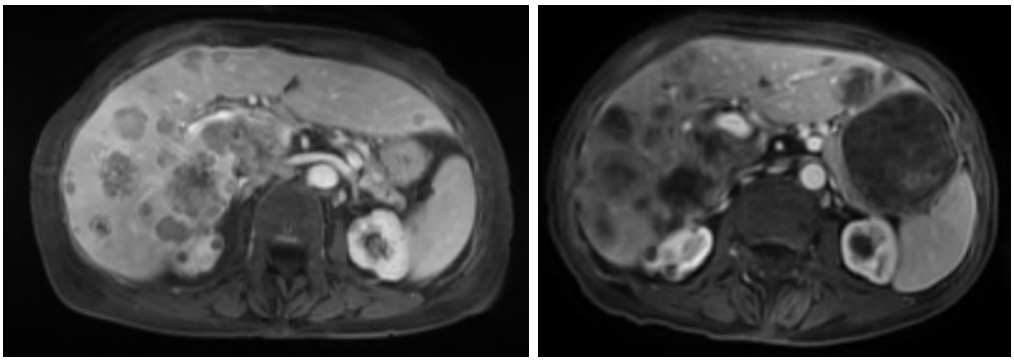


FIGURE 5.5 – Deux images IRM pondérées en T1 acquises à 9 mois d'intervalle. Il n'est pas aisé d'établir la correspondre entre les lésions dans les deux images chez ce patient qui en a beaucoup.

On aurait alors besoin, pour ce problème, d'un recalage précis du foie en entier, qui ne serait ni basé exclusivement sur une comparaison des intensités des voxels, puisque l'évolution des tumeurs entraînerait une modification de ces intensités et de l'aspect général du foie, ni basé sur l'annotation des lésions, à cause du changement de taille ainsi que l'apparition et la disparition des tumeurs d'une date à l'autre. On peut alors penser

à un recalage basé sur les vaisseaux sanguins, comme Vijayan et al. (2014), même si ceux-ci peuvent également être déformés par l'apparition de tumeurs. Plusieurs critères, basés sur des repères à l'intérieur et à l'extérieur du foie devraient alors probablement être combinés pour obtenir un recalage suffisamment précis.

Bibliographie

Aerts, Hugo J W L, Emmanuel Rios Velazquez, Ralph T. H. Leijenaar, Chintan A. Parmar, Patrick Grossmann, Sara Cavalho, Johan Bussink, René Monshouwer, Benjamin Haibe-Kains, Derek Rietveld, Frank J. P. Hoebers, M M Rietbergen, C. Rene Leemans, Andre Dekker, John Quackenbush, Robert J. Gillies et Philippe Lambin (2014). « Decoding tumour phenotype by noninvasive imaging using a quantitative radiomics approach ». In : *Nature Communications* (page 8).

Alain, Guillaume et Yoshua Bengio (2016). « Understanding intermediate layers using linear classifier probes ». In : *arXiv preprint arXiv :1610.01644* (page 54).

Bach, Sebastian, Alexander Binder, Grégoire Montavon, Frederick Klauschen, Klaus-Robert Müller et Wojciech Samek (2015). « On pixel-wise explanations for non-linear classifier decisions by layer-wise relevance propagation ». In : *PloS one* 10.7, e0130140 (page 51).

Bakas, Spyridon, Mauricio Reyes, Andras Jakab, Stefan Bauer, Markus Rempfler, Alessandro Crimi, Russell Takeshi Shinohara, Christoph Berger, Sung Min Ha, Martin Rozycki et al. (2018). « Identifying the best machine learning algorithms for brain tumor segmentation, progression assessment, and overall survival prediction in the BRATS challenge ». In : *arXiv preprint arXiv :1811.02629* (pages 90, 91).

Bau, David, Jun-Yan Zhu, Hendrik Strobelt, Bolei Zhou, Joshua B Tenenbaum, William T Freeman et Antonio Torralba (2018). « Gan dissection : Visualizing and understanding generative adversarial networks ». In : *arXiv preprint arXiv :1811.10597* (page 92).

Bearman, Amy, Olga Russakovsky, Vittorio Ferrari et Li Fei-Fei (2016). « What's the point : Semantic segmentation with point supervision ». In : *European conference on computer vision*. Springer, p. 549-565 (page 16).

Beljaards, Laurens, Mohamed S. Elmahdy, Fons Verbeek et Marius Staring (2020). « A Cross-Stitch Architecture for Joint Registration and Segmentation in Adaptive Radiotherapy ». In : *ArXiv :2004.08122* (pages 19, 37).

Ben Ayed, Ismail, Christian Desrosiers et Jose Dolz (2019). « Weakly supervized CNN segmentation : models and optimization ». In : url : https://github.

com/LIVIAETS/miccai_weakly_supervised_tutorial/blob/master/Documents/MICCAI-2019-Tutorial_On_WeakSemiSup.pdf (page 16).

Chakraborty, Supriyo, Richard Tomsett, Ramya Raghavendra, Daniel Harborne, Moustafa Alzantot, Federico Cerutti, Mani Srivastava, Alun Preece, Simon Julier, Raghuveer M Rao et al. (2017). « Interpretability of deep learning models : a survey of results ». In : *2017 IEEE smartworld, ubiquitous intelligence & computing, advanced & trusted computed, scalable computing & communications, cloud & big data computing, Internet of people and smart city innovation (smartworld/SCALCOM/UIC/ATC/CBDcom/IOP/SCI)*. IEEE, p. 1-6 (page 54).

Chartsias, Agisilaos, Giorgos Papanastasiou, Chengjia Wang, Scott Semple, David E. Newby, Rohan Dharmakumar et Sotirios A. Tsaftaris (2019). « Disentangle, align and fuse for multimodal and zero-shot image segmentation ». In : *ArXiv :1911.04417* (pages 19, 21, 37).

Chen, Chien-Ying, L. Ma, Y. Jia et Panli Zuo (2019). « Kidney and Tumor Segmentation Using Modified 3D Mask RCNN ». In : (pages 73, 74).

Chen, Liang-Chieh, George Papandreou, Iasonas Kokkinos, Kevin Murphy et Alan L Yuille (2017). « Deeplab : Semantic image segmentation with deep convolutional nets, atrous convolution, and fully connected crfs ». In : *IEEE transactions on pattern analysis and machine intelligence* 40.4, p. 834-848 (pages 15, 16).

Chen, Yu, Yuexiang Li, Jiawei Chen et Yefeng Zheng (2019). « OctopusNet : A Deep Learning Segmentation Network for Multi-modal Medical Images ». In : *ArXiv :1906.02031* (page 18).

Couteaux, Vincent, Salim Si-Mohamed, Olivier Nempont, Thierry Lefevre, Alexandre Popoff, Guillaume Pizaine, Nicolas Villain, Isabelle Bloch, Anne Cotten et Loic Boussel (2019). « Automatic knee meniscus tear detection and orientation classification with Mask-RCNN ». In : *Diagnostic and interventional imaging* 100.4, p. 235-242 (pages 96, 123).

Couteaux, Vincent, Salim Si-Mohamed, Raphaele Renard-Penna, Olivier Nempont, Thierry Lefevre, Alexandre Popoff, Guillaume Pizaine, Nicolas Villain, Isabelle Bloch, Julien Behr, Marie-France Bellin, Catherine Roy, Olivier Rouviere, Sarah Montagne, Nathalie Lassau et Loic Boussel (2019). « Kidney cortex segmentation in 2D CT with U-Nets ensemble aggregation ». In : *Diagnostic and Interventional Imaging* 100, p. 211-217 (pages 90, 123).

Couteaux, Vincent, Olivier Nempont, Guillaume Pizaine et Isabelle Bloch (2019). « Towards Interpretability of Segmentation Networks by Analyzing DeepDreams ». In : *Interpretability of Machine Intelligence in Medical Image Computing and Multimodal Learning for Clinical Decision Support*. Springer, p. 56-63 (pages 49, 95).

Dai, Jifeng, Kaiming He et Jian Sun (2015). « Boxsup : Exploiting bounding boxes to supervise convolutional networks for semantic segmentation ». In : *Proceedings of the IEEE international conference on computer vision*, p. 1635-1643 (page 16).

Dolz, Jose, Karthik Gopinath, Jing Yuan, Herve Lombaert, Christian Desrosiers et Ismail Ben Ayed (2019). « HyperDense-Net : A Hyper-Densely Connected CNN for Multi-Modal Image Segmentation ». In : *IEEE Transactions on Medical Imaging* 38, p. 1116-1126 (page 18).

Eisenhauer, Elizabeth A, Patrick Therasse, Jan Bogaerts, Lawrence H Schwartz, D Sargent, Robert Ford, Janet Dancey, S Arbuck, Steve Gwyther, Margaret Mooney et al. (2009). « New response evaluation criteria in solid tumours : revised RECIST guideline (version 1.1) ». In : *European journal of cancer* 45.2, p. 228-247 (page 5).

Elmahdy, Mohamed S., Jelmer M. Wolterink, Hessam Sokooti, Ivana Igum et Marius Staring (2019). « Adversarial optimization for joint registration and segmentation in prostate CT radiotherapy ». In : *ArXiv :1906.12223* (pages 19, 37).

Fu, Cheng-Yang, Wei Liu, Ananth Ranga, Ambrish Tyagi et Alexander C Berg (2017). « Dssd : Deconvolutional single shot detector ». In : *arXiv preprint arXiv :1701.06659* (page 72).

Fu, Yabo, Yang Lei, Tonghe Wang, Walter J Curran, Tian Liu et Xiaofeng Yang (2020). « Deep learning in medical image registration : a review ». In : *Physics in Medicine & Biology* 65.20, 20TR01 (page 103).

Fujita, Hiroshi (2020). « AI-based computer-aided diagnosis (AI-CAD) : the latest review to read first ». In : *Radiological physics and technology* 13.1, p. 6-19 (page 98).

Geirhos, Robert, Patricia Rubisch, Claudio Michaelis, Matthias Bethge, Felix A Wichmann et Wieland Brendel (2018). « ImageNet-trained CNNs are biased towards texture ; increasing shape bias improves accuracy and robustness ». In : *arXiv preprint arXiv :1811.12231* (pages 48, 93).

Ghafoorian, Mohsen, Cedric Nugteren, Nóra Baka, Olaf Booij et Michael Hofmann (2018). « El-gan : Embedding loss driven generative adversarial networks for lane detection ». In : *European Conference on Computer Vision (ECCV)*. Springer, p. 256-272 (pages 17, 20, 29).

Ghorbani, Amirata, Abubakar Abid et James Zou (2019). « Interpretation of neural networks is fragile ». In : *Proceedings of the AAAI Conference on Artificial Intelligence*. T. 33, p. 3681-3688 (page 52).

Girshick, Ross (2015). « Fast r-cnn ». In : *Proceedings of the IEEE international conference on computer vision*, p. 1440-1448 (pages 72, 78, 79).

Girshick, Ross, Jeff Donahue, Trevor Darrell et Jitendra Malik (2014). « Rich feature hierarchies for accurate object detection and semantic segmentation ». In : *Proceedings of the IEEE conference on computer vision and pattern recognition*, p. 580-587 (page 72).

Guo, Yi, Xiangyi Wu, Zhi Wang, Xi Pei et X George Xu (2020). « End-to-end unsupervised cycle-consistent fully convolutional network for 3D pelvic CT-MR defor-

mable registration ». In : *Journal of Applied Clinical Medical Physics* 21.9, p. 193-200 (pages 18, 19).

He, Kaiming, Georgia Gkioxari, Piotr Dollár et Ross Girshick (2017). « Mask r-cnn ». In : *Proceedings of the IEEE international conference on computer vision*, p. 2961-2969 (pages 72, 96).

He, Kaiming, Xiangyu Zhang, Shaoqing Ren et Jian Sun (2016). « Deep residual learning for image recognition ». In : *Proceedings of the IEEE conference on computer vision and pattern recognition*, p. 770-778 (pages 81, 92).

Heinrich, Mattias P, Mark Jenkinson, Manav Bhushan, Tahreema Matin, Fergus V Gleeson, Michael Brady et Julia A Schnabel (2012). « MIND : Modality independent neighbourhood descriptor for multi-modal deformable registration ». In : *Medical image analysis* 16.7, p. 1423-1435 (page 19).

Hofmanninger, Johannes, Forian Prayer, Jeanny Pan, Sebastian Röhrich, Helmut Prosch et Georg Langs (2020). « Automatic lung segmentation in routine imaging is primarily a data diversity problem, not a methodology problem ». In : *European Radiology Experimental* 4.1, p. 1-13 (pages 15, 32, 90).

Hooker, Sara, Dumitru Erhan, Pieter-Jan Kindermans et Been Kim (2018). « A benchmark for interpretability methods in deep neural networks ». In : *arXiv preprint arXiv :1806.10758* (page 52).

Hsu, Chia-Yang, Yi-Hsiang Huang, Cheng-Yuan Hsia, Chien-Wei Su, Han-Chieh Lin, Che-Chuan Loong, Yi-You Chiou, Jen-Huey Chiang, Pui-Ching Lee, Teh-Ia Huo et al. (2010). « A new prognostic model for hepatocellular carcinoma based on total tumor volume : the Taipei Integrated Scoring System ». In : *Journal of hepatology* 53.1, p. 108-117 (page 6).

Huang, Gao, Zhuang Liu, Laurens Van Der Maaten et Kilian Q Weinberger (2017). « Densely connected convolutional networks ». In : *Proceedings of the IEEE conference on computer vision and pattern recognition*, p. 4700-4708 (page 92).

Huang, Huaibo, Zhihang Li, Ran He, Zhenan Sun et Tieniu Tan (2018). « Introvae : Introspective variational autoencoders for photographic image synthesis ». In : *arXiv preprint arXiv :1807.06358* (page 101).

Hung, Wei-Chih, Yi-Hsuan Tsai, Yan-Ting Liou, Yen-Yu Lin et Ming-Hsuan Yang (2018). « Adversarial Learning for Semi-supervised Semantic Segmentation ». In : *BMVC* (pages 17, 29).

Huo, Yuankai, Zhoubing Xu, Shunxing Bao, Albert Assad, Richard G. Abramson et Bennett A. Landman (2018). « Adversarial synthesis learning enables segmentation without target modality ground truth ». In : *IEEE 15th International Symposium on Biomedical Imaging (ISBI)*, p. 1217-1220 (page 17).

Hussain, Shadid M et Michael M Sorrell (2015). *Liver MRI. Correlation with Other Imaging Modalities and Histopathology*. Springer (pages 2, 3).

Huynh, Benjamin Q, Hui Li et Maryellen L Giger (2016). « Digital mammographic tumor classification using transfer learning from deep convolutional neural networks ». In : *Journal of Medical Imaging* 3.3, p. 034501 (page 98).

Isensee, Fabian, Jens Petersen, André Klein, David Zimmerer, Paul F. Jaeger, onnon Kohl, Jakob Wasserthal, Gregor Koehler, Tobias Norajitra, Sebastian J. Wirkert et Klaus Maier-Hein (2018). « nnU-Net : Self-adapting Framework for U-Net-Based Medical Image Segmentation ». In : *ArXiv :1809.10486* (pages 15, 90).

Isensee, Fabian, Jens Petersen, Simon A. A. Kohl, Paul F. Jäger et Klaus Maier-Hein (2019). « nnU-Net : Breaking the Spell on Successful Medical Image Segmentation ». In : *ArXiv :1904.08128* (pages 26, 32).

Jaderberg, Max, Karen Simonyan, Andrew Zisserman et Koray Kavukcuoglu (2015). « Spatial transformer networks ». In : *arXiv preprint arXiv :1506.02025* (page 89).

Jadon, Shruti (2020). « A survey of loss functions for semantic segmentation ». In : *ArXiv :2006.14822* (page 20).

Jaeger, Paul F, Simon AA Kohl, Sebastian Bickelhaupt, Fabian Isensee, Tristan Anselm Kuder, Heinz-Peter Schlemmer et Klaus H Maier-Hein (2020). « Retina U-Net : Embarrassingly simple exploitation of segmentation supervision for medical object detection ». In : *Machine Learning for Health Workshop*. PMLR, p. 171-183 (pages 73, 84, 96).

Kaluva, Krishna Chaitanya, Kiran Vaidhya, Abhijith Chunduru, Sambit Tarai, Sai Prasad Pranav Nadimpalli et S. Vaidya (2020). « An Automated Workflow for Lung Nodule Follow-Up Recommendation Using Deep Learning ». In : *ICIAR* (page 73).

Kavur, A Emre, N Sinem Gezer, Mustafa Barış, Sinem Aslan, Pierre-Henri Conze, Vladimir Groza, Duc Duy Pham, Soumick Chatterjee, Philipp Ernst, Savaş Özkan et al. (2021). « CHAOS challenge-combined (CT-MR) healthy abdominal organ segmentation ». In : *Medical Image Analysis* 69, p. 101950 (pages 15, 90).

Kern, Daria et Andre Mastmeyer (2020). « 3D Bounding Box Detection in Volumetric Medical Image Data : A Systematic Literature Review ». In : *arXiv preprint arXiv :2012.05745* (pages 73, 96).

Kervadec, Hoel, Jose Dolz, Meng Tang, Eric Granger, Yuri Boykov et Ismail Ben Ayed (2019). « Constrained-CNN losses for weakly supervised segmentation ». In : *Medical image analysis* 54, p. 88-99 (page 16).

Kim, Been, Martin Wattenberg, Justin Gilmer, Carrie Cai, James Wexler, Fernanda B. Viégas et Rory Sayres (2018). « Interpretability Beyond Feature Attribution : Quantitative Testing with Concept Activation Vectors (TCAV) ». In : *ICML* (pages 48, 54, 93).

Kindermans, Pieter-Jan, Sara Hooker, Julius Adebayo, Maximilian Alber, Kristof T Schütt, Sven Dähne, Dumitru Erhan et Been Kim (2017). « The (un) reliability of saliency methods ». In : *arXiv preprint arXiv :1711.00867* (page 52).

Kindermans, Pieter-Jan, Kristof T Schütt, Maximilian Alber, Klaus-Robert Müller, Dumitru Erhan, Been Kim et Sven Dähne (2017). « Learning how to explain neural networks : Patternnet and patternattribution ». In : *arXiv preprint arXiv :1705.05598* (page 52).

Kingma, Diederik P et Max Welling (2013). « Auto-encoding variational bayes ». In : *arXiv preprint arXiv :1312.6114* (page 99).

Kolouri, Soheil, Phillip E Pope, Charles E Martin et Gustavo K Rohde (2018). « Sliced Wasserstein auto-encoders ». In : *International Conference on Learning Representations* (page 99).

Kopelowitz, Evi et Guy Engelhard (2019). « Lung Nodules Detection and Segmentation Using 3D Mask-RCNN ». In : *ArXiv* abs/1907.07676 (page 74).

Krizhevsky, Alex, Ilya Sutskever et Geoffrey E Hinton (2012). « Imagenet classification with deep convolutional neural networks ». In : *Advances in neural information processing systems* 25, p. 1097-1105 (pages 92, 98).

Kumar, Devinder, Audrey G Chung, Mohammad J Shaifee, Farzad Khalvati, Masoom A Haider et Alexander Wong (2017). « Discovery radiomics for pathologically-proven computed tomography lung cancer prediction ». In : *International Conference Image Analysis and Recognition*. Springer, p. 54-62 (page 98).

Lambin, Philippe, Emmanuel Rios-Velazquez, Ralph Leijenaar, Sara Carvalho, Ruud GPM Van Stiphout, Patrick Granton, Catharina ML Zegers, Robert Gillies, Ronald Boellard, André Dekker et al. (2012). « Radiomics : extracting more information from medical images using advanced feature analysis ». In : *European journal of cancer* 48.4, p. 441-446 (page 8).

Lapuschkin, Sebastian, Stephan Wäldchen, Alexander Binder, Grégoire Montavon, Wojciech Samek et Klaus-Robert Müller (2019). « Unmasking clever hans predictors and assessing what machines really learn ». In : *Nature communications* 10.1, p. 1-8 (page 48).

LeCun, Yann, Bernhard Boser, John S Denker, Donnie Henderson, Richard E Howard, Wayne Hubbard et Lawrence D Jackel (1989). « Backpropagation applied to handwritten zip code recognition ». In : *Neural computation* 1.4, p. 541-551 (page 92).

LeCun, Yann, Léon Bottou, Yoshua Bengio et Patrick Haffner (1998). « Gradient-based learning applied to document recognition ». In : *Proceedings of the IEEE* 86.11, p. 2278-2324 (page 99).

Lee, Yun-Hsuan, Cheng-Yuan Hsia, Chia-Yang Hsu, Yi-Hsiang Huang, Han-Chieh Lin et Teh-Ia Huo (2013). « Total tumor volume is a better marker of tumor burden

in hepatocellular carcinoma defined by the Milan criteria ». In : *World journal of surgery* 37.6, p. 1348-1355 (page 6).

Lei, Yang, X. He, Jincao Yao, Tonghe Wang, Lijing Wang, W. Li, W. Curran, T. Liu, D. Xu et X. Yang (2020). « Breast Tumor Segmentation in 3D Automatic Breast Ultrasound Using Mask Scoring R-CNN. » In : *Medical physics* (page 74).

Lei, Yang, Z. Tian, S. Kahn, W. Curran, T. Liu et X. Yang (2020). « Automatic detection of brain metastases using 3D mask R-CNN for stereotactic radiosurgery ». In : *Medical Imaging* (page 74).

Li, Zuoxin et Fuqiang Zhou (2017). « FSSD : feature fusion single shot multibox detector ». In : *arXiv preprint arXiv :1712.00960* (page 72).

Lin, Di, Jifeng Dai, Jiaya Jia, Kaiming He et Jian Sun (2016). « Scribblesup : Scribble-supervised convolutional networks for semantic segmentation ». In : *Proceedings of the IEEE conference on computer vision and pattern recognition*, p. 3159-3167 (page 16).

Lin, Tsung-Yi, Piotr Dollár, Ross Girshick, Kaiming He, Bharath Hariharan et Serge Belongie (2017). « Feature pyramid networks for object detection ». In : *Proceedings of the IEEE conference on computer vision and pattern recognition*, p. 2117-2125 (page 72).

Lin, Tsung-Yi, Priya Goyal, Ross Girshick, Kaiming He et Piotr Dollár (2017). « Focal loss for dense object detection ». In : *Proceedings of the IEEE international conference on computer vision*, p. 2980-2988 (pages 72, 74, 78, 82, 85, 96).

Lin, Tsung-Yi, Michael Maire, Serge J. Belongie, Lubomir D. Bourdev, Ross B. Girshick, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár et C. Lawrence Zitnick (2014). « Microsoft COCO : Common Objects in Context ». In : *ECCV* (page 61).

Liu, Fengze, Jinzheng Cai, Yuankai Huo, Chi-Tung Cheng, Ashwin Raju, Dakai Jin, Jing Xiao, Alan L. Yuille, Le Lu, Chien-Hung Liao et Adam P. Harrison (2020). « JSSR : A Joint Synthesis, Segmentation, and Registration System for 3D Multi-Modal Image Alignment of Large-scale Pathological CT Scans ». In : *ArXiv :2005.12209* (pages 19, 21, 37).

Liu, Li, Wanli Ouyang, Xiaogang Wang, Paul Fieguth, Jie Chen, Xinwang Liu et Matti Pietikäinen (2020). « Deep learning for generic object detection : A survey ». In : *International journal of computer vision* 128.2, p. 261-318 (pages 69, 71).

Liu, Wei, Dragomir Anguelov, Dumitru Erhan, Christian Szegedy, Scott Reed, Cheng-Yang Fu et Alexander C Berg (2016). « Ssd : Single shot multibox detector ». In : *European conference on computer vision*. Springer, p. 21-37 (page 72).

Luc, Pauline, Camille Couprie, Soumith Chintala et Jakob Verbeek (2016). « Semantic Segmentation using Adversarial Networks ». In : *ArXiv :1611.08408* (pages 17, 20, 29).

Lundberg, Scott et Su-In Lee (2017). « A unified approach to interpreting model predictions ». In : *arXiv preprint arXiv :1705.07874* (page 52).
Mahendran, Aravindh et Andrea Vedaldi (2015). « Understanding deep image representations by inverting them ». In : *Proceedings of the IEEE conference on computer vision and pattern recognition*, p. 5188-5196 (page 53).
– (2016). « Visualizing deep convolutional neural networks using natural pre-images ». In : *International Journal of Computer Vision* 120.3, p. 233-255 (page 53).
Makhzani, Alireza, Jonathon Shlens, Navdeep Jaitly, Ian Goodfellow et Brendan Frey (2015). « Adversarial autoencoders ». In : *arXiv preprint arXiv :1511.05644* (pages 99, 100).
McInnes, Leland, John Healy et James Melville (2018). « Umap : Uniform manifold approximation and projection for dimension reduction ». In : *arXiv preprint arXiv :1802.03426* (page 54).
Milletari, Fausto, Nassir Navab et Seyed-Ahmad Ahmadi (2016). « V-net : Fully convolutional neural networks for volumetric medical image segmentation ». In : *Fourth International Conference on 3D Vision (3DV)*. IEEE, p. 565-571 (pages 25, 74).
Montavon, Grégoire, Sebastian Lapuschkin, Alexander Binder, Wojciech Samek et Klaus-Robert Müller (2017). « Explaining nonlinear classification decisions with deep Taylor decomposition ». In : *Pattern Recognition* 65, p. 211-222 (page 51).
Mordvintsev, Alexander, Christopher Olah et Mike Tyka (2015). « Inceptionism : Going deeper into neural networks ». In : *Google Research Blog* (pages 52, 53, 58).
Namasivayam, Saravanan, Diego R Martin et Sanjay Saini (2007). « Imaging of liver metastases : MRI ». In : *Cancer Imaging* 7.1, p. 2 (page 4).
Natekar, Parth, Avinash Kori et Ganapathy Krishnamurthi (2020). « Demystifying Brain Tumor Segmentation Networks : Interpretability and Uncertainty Analysis ». In : *Frontiers in Computational Neuroscience* 14, p. 6 (page 95).
Nguyen, Anh, Alexey Dosovitskiy, Jason Yosinski, Thomas Brox et Jeff Clune (2016). « Synthesizing the preferred inputs for neurons in neural networks via deep generator networks ». In : *Advances in neural information processing systems* 29, p. 3387-3395 (page 53).
Odena, Augustus (2016). « Semi-supervised learning with generative adversarial networks ». In : *arXiv preprint arXiv :1606.01583* (page 103).
Palmer, Daniel H, Neil S Hawkins, Valérie Vilgrain, Helena Pereira, Gilles Chatellier et Paul J Ross (2020). « Tumor burden and liver function in HCC patient selection for selective internal radiation therapy : SARAH post-hoc study ». In : *Future Oncology* 16.01, p. 4315-4325 (page 6).
Ravichandran, Kavya, Nathaniel Braman, Andrew Janowczyk et Anant Madabhushi (2018). « A deep learning classifier for prediction of pathological complete response to neoadjuvant chemotherapy from baseline breast DCE-MRI ». In : *Medical Ima-*

ging 2018 : Computer-Aided Diagnosis. T. 10575. International Society for Optics et Photonics, p. 105750C (page 98).

Redmon, Joseph, Santosh Divvala, Ross Girshick et Ali Farhadi (2016). « You only look once : Unified, real-time object detection ». In : *Proceedings of the IEEE conference on computer vision and pattern recognition*, p. 779-788 (page 72).

Redmon, Joseph et Ali Farhadi (2017). « YOLO9000 : better, faster, stronger ». In : *Proceedings of the IEEE conference on computer vision and pattern recognition*, p. 7263-7271 (page 72).

Ren, Shaoqing, Kaiming He, Ross Girshick et Jian Sun (2015). « Faster r-cnn : Towards real-time object detection with region proposal networks ». In : *arXiv preprint arXiv :1506.01497* (pages 72, 74, 96).

Reyes, Mauricio, Raphael Meier, Sérgio Pereira, Carlos A Silva, Fried-Michael Dahlweid, Hendrik von Tengg-Kobligk, Ronald M Summers et Roland Wiest (2020). « On the interpretability of artificial intelligence in radiology : challenges and opportunities ». In : *Radiology : artificial intelligence* 2.3, e190043 (pages 49, 50).

Ribeiro, Marco Túlio, Sameer Singh et Carlos Guestrin (2016). « "Why Should I Trust You ?" : Explaining the Predictions of Any Classifier ». In : *HLT-NAACL Demos* (page 52).

– (2018). « Anchors : High-Precision Model-Agnostic Explanations ». In : *AAAI* (page 52).

Ronneberger, Olaf, Philipp Fischer et Thomas Brox (2015). « U-net : Convolutional networks for biomedical image segmentation ». In : *International Conference on Medical image Computing and Computer-Assisted Intervention*. Springer, p. 234-241 (pages 15, 25, 28, 74, 90).

Samson, Laurens, Nanne van Noord, Olaf Booij, Michael Hofmann, Efstratios Gavves et Mohsen Ghafoorian (2019). « I Bet You Are Wrong : Gambling Adversarial Networks for Structured Semantic Segmentation ». In : *IEEE International Conference on Computer Vision Workshops* (pages 17, 20, 29).

Santamaria-Pang, Alberto, James Kubricht, Aritra Chowdhury, Chitresh Bhushan et Peter Tu (2020). « Towards Emergent Language Symbolic Semantic Segmentation and Model Interpretability ». In : *International Conference on Medical Image Computing and Computer-Assisted Intervention*. Springer, p. 326-334 (page 95).

Selvaraju, R. R., M. Cogswell, A. Das, R. Vedantam, D. Parikh et D. Batra (2017). « Grad-CAM : Visual Explanations from Deep Networks via Gradient-Based Localization ». In : *2017 IEEE International Conference on Computer Vision (ICCV)*, p. 618-626. doi : 10.1109/ICCV.2017.74 (pages 50, 51, 92, 95).

Senior, Andrew W, Richard Evans, John Jumper, James Kirkpatrick, Laurent Sifre, Tim Green, Chongli Qin, Augustin Žídek, Alexander WR Nelson, Alex

Bridgland et al. (2020). « Improved protein structure prediction using potentials from deep learning ». In : *Nature* 577.7792, p. 706-710 (page 93).

Simonyan, Karen, Andrea Vedaldi et Andrew Zisserman (2013). « Deep inside convolutional networks : Visualising image classification models and saliency maps ». In : *arXiv preprint arXiv :1312.6034* (pages 51-53, 92).

Simpson, Amber L, Michela Antonelli, Spyridon Bakas, Michel Bilello, Keyvan Farahani, Bram Van Ginneken, Annette Kopp-Schneider, Bennett A Landman, Geert Litjens, Bjoern Menze et al. (2019). « A large annotated medical image dataset for the development and evaluation of segmentation algorithms ». In : *arXiv preprint arXiv :1902.09063* (pages 15, 90).

Smilkov, Daniel, Nikhil Thorat, Been Kim, Fernanda B. Viégas et Martin Wattenberg (2017). « SmoothGrad : removing noise by adding noise ». In : *CoRR* abs/1706.03825 (page 51).

Springenberg, Jost Tobias, Alexey Dosovitskiy, Thomas Brox et Martin A. Riedmiller (2014). « Striving for Simplicity : The All Convolutional Net ». In : *CoRR* abs/1412.6806 (page 51).

Sudre, Carole H, Wenqi Li, Tom Vercauteren, Sebastien Ourselin et M Jorge Cardoso (2017). « Generalised dice overlap as a deep learning loss function for highly unbalanced segmentations ». In : *Deep learning in medical image analysis and multimodal learning for clinical decision support*. Springer, p. 240-248 (pages 20, 28).

Sundararajan, Mukund, Ankur Taly et Qiqi Yan (2017). « Axiomatic Attribution for Deep Networks ». In : *ICML* (page 51).

Szegedy, Christian, Sergey Ioffe, Vincent Vanhoucke et Alexander Alemi (2017). « Inception-v4, inception-resnet and the impact of residual connections on learning ». In : *Proceedings of the AAAI Conference on Artificial Intelligence*. T. 31. 1 (page 92).

Taghanaki, Saeid Asgari, Kumar Abhishek, Joseph Paul Cohen, Julien Cohen-Adad et Ghassan Hamarneh (2020). « Deep semantic segmentation of natural and medical images : a review ». In : *Artificial Intelligence Review*, p. 1-42 (page 14).

Tajbakhsh, Nima, Laura Jeyaseelan, Qian Li, Jeffrey N Chiang, Zhihao Wu et Xiaowei Ding (2020). « Embracing imperfect datasets : A review of deep learning solutions for medical image segmentation ». In : *Medical Image Analysis* 63, p. 101693 (pages 14-16).

Tang, Meng, Federico Perazzi, Abdelaziz Djelouah, Ismail Ben Ayed, Christopher Schroers et Yuri Boykov (2018). « On regularized losses for weakly-supervised cnn segmentation ». In : *Proceedings of the European Conference on Computer Vision (ECCV)*, p. 507-522 (page 16).

Tolstikhin, Ilya, Olivier Bousquet, Sylvain Gelly et Bernhard Schoelkopf (2017). « Wasserstein auto-encoders ». In : *arXiv preprint arXiv :1711.01558* (page 99).

Valindria, Vanya V., Nick Pawlowski, Martin Rajchl, Ioannis Lavdas, Eric O. Aboagye, Andrea G. Rockall, Daniel Rueckert et Ben Glocker (2018). « Multi-modal Learning from Unpaired Images : Application to Multi-organ Segmentation in CT and MRI ». In : *IEEE Winter Conference on Applications of Computer Vision (WACV)*, p. 547-556 (page 17).

Van der Maaten, Laurens et Geoffrey Hinton (2008). « Visualizing data using t-SNE. » In : *Journal of machine learning research* 9.11 (page 54).

Vijayan, Sinara, Ingerid Reinertsen, Erlend Fagertun Hofstad, Anna Rethy, Toril A Nagelhus Hernes et Thomas Langø (2014). « Liver deformation in an animal model due to pneumoperitoneum assessed by a vessel-based deformable registration ». In : *Minimally Invasive Therapy & Allied Technologies* 23.5, p. 279-286 (page 106).

Wang, Kang, Adrija Mamidipalli, Tara Retson, Naeim Bahrami, Kyle Hasenstab, Kevin Blansit, Emily Bass, Timoteo Delgado, Guilherme Cunha, Michael S Middleton et al. (2019). « Automated CT and MRI liver segmentation and biometry using a generalized convolutional neural network ». In : *Radiology : Artificial Intelligence* 1.2, p. 180022 (pages 17, 20).

Wei, Yanan, X. Jiang, K. Liu, Cheng Zhong, Z. Shi, J. Leng et F. Xu (2019). « A Hybrid Multi-atrous and Multi-scale Network for Liver Lesion Detection ». In : *MLMI@MICCAI* (page 73).

Wilcoxon, Frank (1945). « Individual Comparisons by Ranking Methods ». In : *Biometrics Bulletin* 1.6, p. 80-83. issn : 00994987. url : http://www.jstor.org/stable/3001968 (page 30).

Wolterink, Jelmer M, Anna M Dinkla, Mark HF Savenije, Peter R Seevinck, Cornelis AT van den Berg et Ivana Išgum (2017). « Deep MR to CT synthesis using unpaired data ». In : *International workshop on simulation and synthesis in medical imaging*. Springer, p. 14-23 (page 18).

Wu, Sen, Hongyang R. Zhang et Christopher Ré (2020). « Understanding and Improving Information Transfer in Multi-Task Learning ». In : *International Conference on Learning Representations*. url : https://openreview.net/forum?id=SylzhkBtDB (page 31).

Xiao, Youzi, Zhiqiang Tian, Jiachen Yu, Yinshu Zhang, Shuai Liu, Shaoyi Du et Xuguang Lan (2020). « A review of object detection based on deep learning ». In : *Multimedia Tools and Applications* 79.33, p. 23729-23791 (pages 69, 71).

Xu, Kai, Dae Hoon Park, Chang Yi et Charles Sutton (2018). « Interpreting deep classifier by visual distillation of dark knowledge ». In : *arXiv preprint arXiv :1803.04042* (page 54).

Xu, Xuanang, F. Zhou, Bo Liu, D. Fu et X. Bai (2019). « Efficient Multiple Organ Localization in CT Image Using 3D Region Proposal Network ». In : *IEEE Transactions on Medical Imaging* 38, p. 1885-1898 (page 73).

Xu, Zhe, Jie Luo, Jiangpeng Yan, Xiu Li et Jagadeesan Jayender (2020). *F3RNet : Full-Resolution Residual Registration Network for Multimodal Image Registration*. arXiv : 2009.07151 [eess.IV] (pages 19, 37).

Yang, Heran, Jian Sun, Aaron Carass, Can Zhao, Junghoon Lee, Zongben Xu et Jerry Prince (2018). « Unpaired brain MR-to-CT synthesis using a structure-constrained CycleGAN ». In : *Deep Learning in Medical Image Analysis and Multimodal Learning for Clinical Decision Support*. Springer, p. 174-182 (page 18).

Yeche, Hugo, Justin Harrison et Tess Berthier (2019). « UBS : A Dimension-Agnostic Metric for Concept Vector Interpretability Applied to Radiomics ». In : *Interpretability of Machine Intelligence in Medical Image Computing and Multimodal Learning for Clinical Decision Support*. Springer, p. 12-20 (page 54).

Yeh, Chih-Kuan, Cheng-Yu Hsieh, Arun Sai Suggala, David Inouye et Pradeep Ravikumar (2019). « How Sensitive are Sensitivity-Based Explanations ? » In : *arXiv preprint arXiv :1901.09392* (page 52).

Yosinski, Jason, Jeff Clune, Anh Nguyen, Thomas Fuchs et Hod Lipson (2015). « Understanding neural networks through deep visualization ». In : *arXiv preprint arXiv :1506.06579* (page 52).

Ypsilantis, Petros-Pavlos, Musib Siddique, Hyon-Mok Sohn, Andrew Davies, Gary Cook, Vicky Goh et Giovanni Montana (2015). « Predicting response to neoadjuvant chemotherapy with PET imaging using convolutional neural networks ». In : *PloS one* 10.9, e0137036 (page 98).

Yuan, Wenguang, Jia Wei, Jiabing Wang, Qianli Ma et Tolga Tasdizen (2019). « Unified Attentional Generative Adversarial Network for Brain Tumor Segmentation From Multimodal Unpaired Images ». In : *ArXiv :1907.03548* (page 17).

Zech, John R, Marcus A Badgeley, Manway Liu, Anthony B Costa, Joseph J Titano et Eric Karl Oermann (2018). « Variable generalization performance of a deep learning model to detect pneumonia in chest radiographs : a cross-sectional study ». In : *PLoS medicine* 15.11, e1002683 (page 50).

Zeiler, Matthew D et Rob Fergus (2014). « Visualizing and understanding convolutional networks ». In : *European Conference on Computer Vision*. Springer, p. 818-833 (page 51).

Zeng, Qi, Davood Karimi, Emily HT Pang, Shahed Mohammed, Caitlin Schneider, Mohammad Honarvar et Septimiu E Salcudean (2019). « Liver Segmentation in Magnetic Resonance Imaging via Mean Shape Fitting with Fully Convolutional Neural Networks ». In : *International Conference on Medical Image Computing and Computer-Assisted Intervention*. Springer, p. 246-254 (page 17).

Zhang, Shifeng, Longyin Wen, Xiao Bian, Zhen Lei et Stan Z Li (2018). « Single-shot refinement neural network for object detection ». In : *Proceedings of the IEEE conference on computer vision and pattern recognition*, p. 4203-4212 (page 72).

Zhang, Yizhe, Lin Yang, Jianxu Chen, Maridel Fredericksen, David P Hughes et Danny Z Chen (2017). « Deep adversarial networks for biomedical image segmentation utilizing unannotated images ». In : *International Conference on Medical Image Computing and Computer-Assisted Intervention*. Springer, p. 408-416 (pages 17, 29).

Zhang, Yu, Peter Tiňo, Aleš Leonardis et Ke Tang (2020). « A Survey on Neural Network Interpretability ». In : *arXiv preprint arXiv :2012.14261* (pages 47-49).

Zhang, Zizhao, Lin Yang et Yefeng Zheng (2018). « Translating and Segmenting Multimodal Medical Volumes with Cycle- and Shape-Consistency Generative Adversarial Network ». In : *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, p. 9242-9251 (page 17).

Zhao, Hengshuang, Jianping Shi, Xiaojuan Qi, Xiaogang Wang et Jiaya Jia (2017). « Pyramid scene parsing network ». In : *Proceedings of the IEEE conference on computer vision and pattern recognition*, p. 2881-2890 (page 15).

Zhao, Zhong-Qiu, Peng Zheng, Shou-tao Xu et Xindong Wu (2019). « Object detection with deep learning : A review ». In : *IEEE transactions on neural networks and learning systems* 30.11, p. 3212-3232 (pages 71, 72).

Zhou, Tongxue, Su Ruan et Stéphane Canu (2019). « A review : Deep learning for medical image segmentation using multi-modality fusion ». In : *Array* 3-4, p. 100004 (page 18).

Zhou, Yuyin, Zhe Li, Song Bai, Chong Wang, Xinlei Chen, Mei Han, Elliot Fishman et Alan L Yuille (2019). « Prior-aware neural network for partially-supervised multi-organ segmentation ». In : *Proceedings of the IEEE/CVF International Conference on Computer Vision*, p. 10672-10681 (page 16).

Zhou, Zongwei, Vatsal Sodha, Md Mahfuzur Rahman Siddiquee, Ruibin Feng, Nima Tajbakhsh, Michael B Gotway et Jianming Liang (2019). « Models genesis : Generic autodidactic models for 3D medical image analysis ». In : *International Conference on Medical Image Computing and Computer-Assisted Intervention*. Springer, p. 384-393 (pages 25, 26, 43).

Zhu, Jun-Yan, T. Park, Phillip Isola et Alexei A. Efros (2017). « Unpaired Image-to-Image Translation Using Cycle-Consistent Adversarial Networks ». In : *IEEE International Conference on Computer Vision (ICCV)*, p. 2242-2251 (page 18).

Zou, Yang, Zhiding Yu, BVK Kumar et Jinsong Wang (2018). « Unsupervised domain adaptation for semantic segmentation via class-balanced self-training ». In : *Proceedings of the European conference on computer vision (ECCV)*, p. 289-305 (pages 16, 18).

Zwanenburg, Alex, Martin Vallières, Mahmoud A Abdalah, Hugo JWL Aerts, Vincent Andrearczyk, Aditya Apte, Saeed Ashrafinia, Spyridon Bakas, Roelof J Beukinga, Ronald Boellaard et al. (2020). « The image biomarker standardization initiative : standardized quantitative radiomics for high-throughput image-based phenotyping ». In : *Radiology* 295.2, p. 328-338 (pages 10, 57, 65, 97).

Annexe A

Publications

Articles publiés

— Vincent Couteaux, Salim Si-Mohamed, Olivier Nempont, Thierry Lefevre, Alexandre Popoff, Guillaume Pizaine, Nicolas Villain, Isabelle Bloch, Anne Cotten et Loïc Boussel (2019). « Automatic knee meniscus tear detection and orientation classification with Mask-RCNN ». *Diagnostic and interventional imaging* 100.4 p. 235-242.

— Vincent Couteaux, Salim Si-Mohamed, Raphaële Renard-Penna Olivier Nempont, Thierry Lefevre, Alexandre Popoff, Guillaume Pizaine, Nicolas Villain, Isabelle Bloch, Julien Behr, Marie-France Bellin, Catherine Roy, Olivier Rouviere, Sarah Montagne, Nathalie Lassau et Loïc Boussel (2019). « Kidney Cortex segmentation in 2D CT with U-Nets ensemble aggregation ». *Diagnostic and Interventional Imaging* 100, p. 211-217.

— Vincent Couteaux, Olivier Nempont, Guillaume Pizaine et Isabelle Bloch (2019). « Towards interpretability of segmentation networks by analyzing DeepDreams »*Interpretability of Machine Intelligence in Medical Image Computing and Multimodal Learning for Clinical Decision Support*. Springer, p. 56-63.

Article accepté

— Vincent Couteaux, Olivier Nempont, Guillaume Pizaine et Isabelle Bloch (2021). « Cooperating networks to enforce a similarity constraint in paired but unregistered multimodal liver segmentation ». *International Symposium on Biomedical Imaging*.

Soumission ArXiv

— Vincent Couteaux, Olivier Nempont, Guillaume Pizaine et Isabelle Bloch (2021). « Comparing Deep Learning strategies for paired but unregistered mul-

timodal segmentation of the liver in T1 and T2-weighted MRI ». *ArXiv : 2101.06979*.

Annexe B

Articles des Compétitions

La Société Francophone de Radiologie a organisé dans le cadre des Journées Françaises de Radiologie (JFR) en 2019 deux compétitions auxquelles mes collègues et moi avons participé.

La première consistait à détecter des fissures de ménisque du genou dans des coupes d'images IRM centrées sur le genou en précisant quel ménisque était touché (antérieur ou postérieur), et à classifier l'orientation de la fissure le cas échéant (horizontale ou verticale). L'équipe qui parvenait à concevoir un algorithme réalisant la meilleure précision de classification sur un jeu d'images inconnu remportait la compétition.

La seconde consistait à segmenter le cortex rénal dans des coupes d'images CT centrées autour du rein. Le but était d'obtenir l'algorithme le plus performant sur un ensemble d'images inconnues.

Nous avons remporté les deux compétitions, et j'ai publié les deux articles ci-après qui détaillent nos méthodes (Couteaux, Si-Mohamed, Nempont et al. 2019 ; Couteaux, Si-Mohamed, Renard-Penna et al. 2019).

Diagnostic and Interventional Imaging (2019) **100**, 211—217

ORIGINAL ARTICLE / *Computer developments*

Kidney cortex segmentation in 2D CT with U-Nets ensemble aggregation

V. Couteaux[a,b,*], S. Si-Mohamed[c,d], R. Renard-Penna[e], O. Nempont[a], T. Lefevre[a], A. Popoff[a], G. Pizaine[a], N. Villain[a], I. Bloch[b], J. Behr[f], M.-F. Bellin[g], C. Roy[h], O. Rouvière[i], S. Montagne[j], N. Lassau[k], L. Boussel[c,d]

[a] *Philips Research France, 33, rue de Verdun, 92150 Suresnes, France*
[b] *LTCI, Télécom ParisTech, Université Paris-Saclay, 75013 Paris, France*
[c] *CREATIS, CNRS UMR 5220, Inserm U1206, INSA-Lyon, Claude Bernard Lyon 1 University, 69100 Villeurbanne, France*
[d] *Department of Radiology, Hospices Civils de Lyon, 69002 Lyon, France*
[e] *Department of Radiology, Hôpital Tenon, AP—HP, GRC-UPMC n°5 Oncotype-URO, Sorbonne universités, 75020 Paris, France*
[f] *Department of Radiology, CHRU de Besançon, 25000 Besançon, France*
[g] *Department of Radiology, Hôpitaux Universitaires Paris Sud, 94270 Le Kremlin Bicêtre, France*
[h] *Department of Radiology, CHU de Strasbourg, Nouvel Hôpital Civil, 67000 Strasbourg, France*
[i] *Department of Uroradiology, Hospices Civils de Lyon, Faculté de Médecine Lyon Est, 69002 Lyon, France*
[j] *Department of Radiology, Hôpital Pitié Salpétrière, AP—HP, 75013 Paris, France*
[k] *Department of Radiology, Gustave Roussy, IR4M, UMR8081, CNRS, Université Paris-Sud, Université Paris-Saclay, 94805 Villejuif, France*

KEYWORDS
Renal cortex;
Image segmentation;
Artificial intelligence (AI);
Computed tomography (CT)

Abstract
Purpose: This work presents our contribution to one of the data challenges organized by the French Radiology Society during the *Journées Francophones de Radiologie*. This challenge consisted in segmenting the kidney cortex from coronal computed tomography (CT) images, cropped around the cortex.
Materials and methods: We chose to train an ensemble of fully-convolutional networks and to aggregate their prediction at test time to perform the segmentation. An image database was made available in 3 batches. A first training batch of 250 images with segmentation masks was provided by the challenge organizers one month before the conference. An additional training batch of 247 pairs was shared when the conference began. Participants were ranked using a Dice score.

* Corresponding author. Philips Research France, 33, rue de Verdun, 92150 Suresnes, France.
E-mail address: vincent.couteaux@telecom-paristech.fr (V. Couteaux).

https://doi.org/10.1016/j.diii.2019.03.001

Results: The segmentation results of our algorithm match the renal cortex with a good precision. Our strategy yielded a Dice score of 0.867, ranking us first in the data challenge.
Conclusion: The proposed solution provides robust and accurate automatic segmentations of the renal cortex in CT images although the precision of the provided reference segmentations seemed to set a low upper bound on the numerical performance. However, this process should be applied in 3D to quantify the renal cortex volume, which would require a marked labelling effort to train the networks.

Renal diseases are often associated with cortical morphological changes, such as volume reduction or notch defect. All these features are considered as surrogate markers of renal diseases and can be visible on imaging examinations, such as ultrasound, magnetic resonance imaging (MRI), or computed tomography (CT) [1,2]. Despite a well-established qualitative assessment of the renal cortex with these modalities, a quantitative approach helps improve the diagnostic work-up of renal diseases [3]. However, to date quantitative assessment of renal cortex is hampered by complex and time-consuming analyses such as semi-automated segmentations based on a pixel value threshold algorithm, region growing, appearance models combined with graph cuts or random forests [4—8]. The recent development of convolutional neural networks (CNN), as well as the access to very large imaging databases, could help overcome these limitations. Very promising results have recently been obtained in several applications such as the segmentation of cardiac chambers, and the brain [9,10]. However, the appropriate artificial intelligence (AI) tools for kidney analysis still need to be developed.

Fully-convolutional networks have drastically improved the state-of-the-art in image segmentation [11]. U-Nets are currently a standard approach for two-dimensional (2D) or three-dimensional (3D) medical image segmentation problems [12—18].

The *Journées Francophones de Radiologie* was held in Paris in October 2018. For the first time this year, the French Society of Radiology organized an AI competition. Teams of industrial researchers, students, and radiologists were invited to take part in five data challenges. In this paper, we present our approach to address the kidney cortex segmentation challenge aiming at segmenting the renal cortex on 2D coronal CT images.

Method

Kidney cortex segmentation challenge

An image database was made available in 3 batches. A first training batch of 250 images with segmentation masks was provided by the challenge organizers one month before the conference. An additional training batch of 247 pairs was shared when the conference began. Two days later, the teams were ranked on a test batch of 299 images.

CT images in the coronal plane, cropped and resized around the kidney (192×192 pixels with a pixel size of 1×1 mm and intensity in Hounsfield units [HU]) were provided (Fig. 1). The reference segmentation was provided as a binary mask for each image of the training set. Due to the usual difficulties of manual segmentation, in particular for irregularly shaped objects such as the renal cortex, several reference segmentations were debatable or even erroneous. We observed that a proportion of the pixels at the edge of the cortex were either left out when they should not have been, or mislabeled as cortex while clearly outside (Fig. 1c). Moreover, blood vessels inside the kidney were occasionally included in the reference segmentation, but this was inconsistent throughout the dataset (Fig. 1d). In fact, it can be hard to distinguish actual renal columns from some blood vessels. We clipped the image intensity values between −150 HU and 200 HU and rescaled them between 0 and 1. This range has been chosen manually to contain all the renal cortex dynamic and limit the influence of high values in the image, corresponding to bones, and very low values, corresponding to air.

To address the specific difficulties of this challenge, such as the imprecision of the reference segmentations, we adopted several popular strategies such as artificial data augmentation, meta parameter optimization, pre-training and post-processing with connected components analysis [19—22]. We also used ensemble aggregation, a standard machine learning technique frequently applied to deep learning [12,22,23].

Network architecture

We chose a U-Net architecture with 5 levels of depth, residual blocks, and rectified linear units (ReLU) activation functions, and added convolutions on the skip connections (Fig. 2) [18,24—26]. We set the meta-parameters using a Bayesian optimization approach [19,20]. We used artificial data augmentation during training to limit overfitting, by randomly applying translations, rotations, zooms, noise, brightness and contrast shifts to the input samples. The training was performed until convergence and lasted

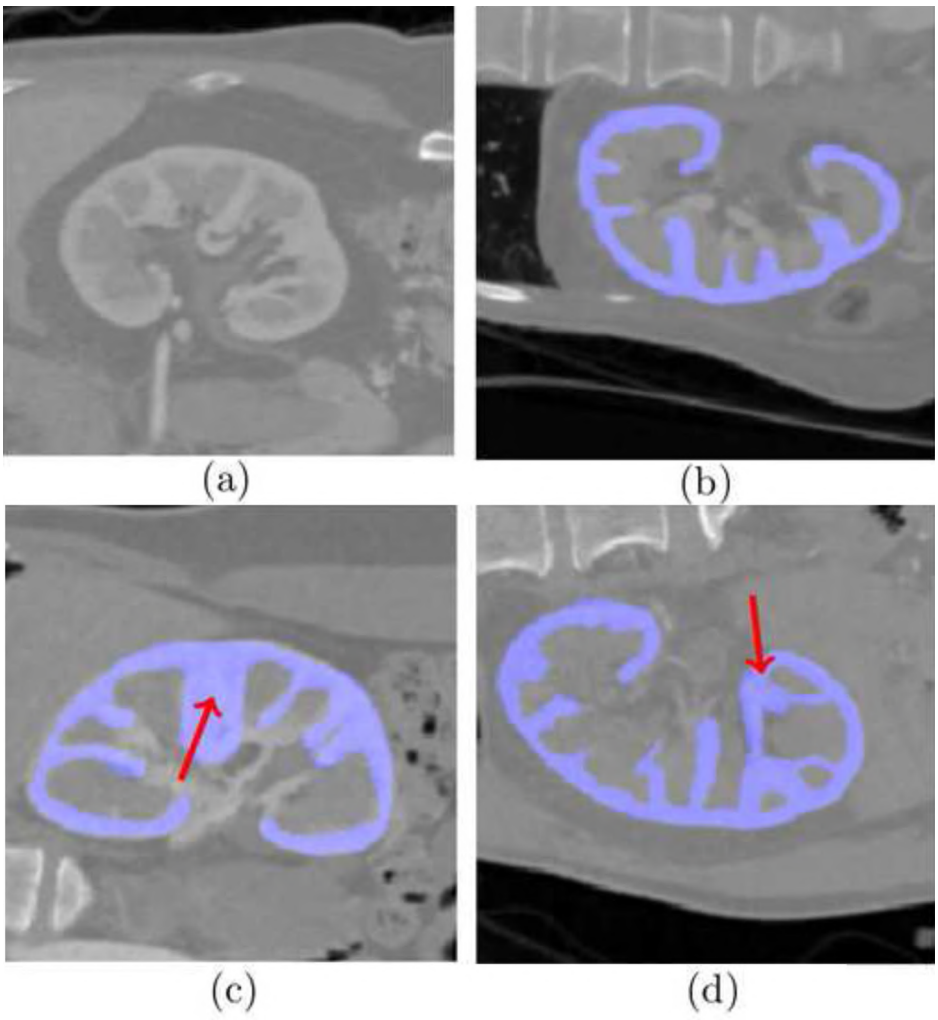


Figure 1. CT images of the kidney from the training set provided by the data challenge organizers. The reference segmentation is overlapped in blue; a: image only; b: correct segmentation; c: inaccurate segmentation and renal column clusters (arrow); d: blood vessels included in the segmentation (arrow).

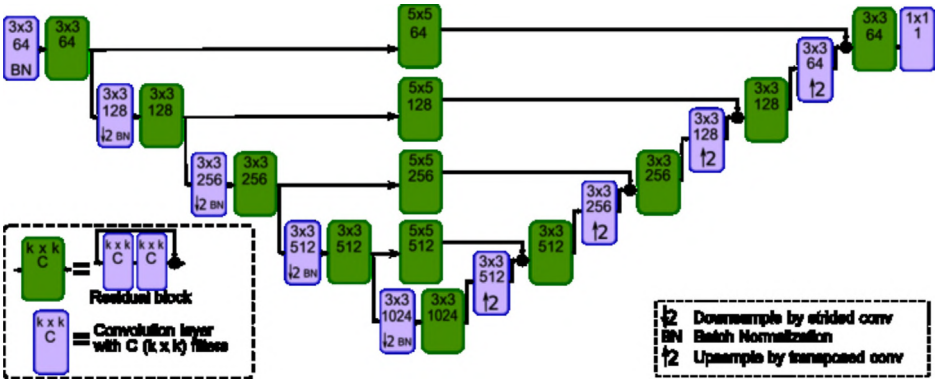


Figure 2. Selected network architecture to achieve the segmentation task. Green boxes are residual blocks, blue boxes are simple convolutional layers with ReLU activation. Batch normalization is applied after convolution and before activation.

between one and two hours. We used Adam optimizer with a learning rate of 1.10^{-4} on batches of 10 images.

Weight initialization and pre-training

Considering the low amount of data available for training following the popular practice initiated in [21], we considered that pre-training the network on a large and publicly-available dataset would be advantageous. We therefore pre-trained our U-Nets to segment persons, the common objects in context (COCO) dataset [26]. We compared training experiments using randomly initialized weights or pre-training (Fig. 3). Although the final score was similar, the training converges faster using a pre-trained network,

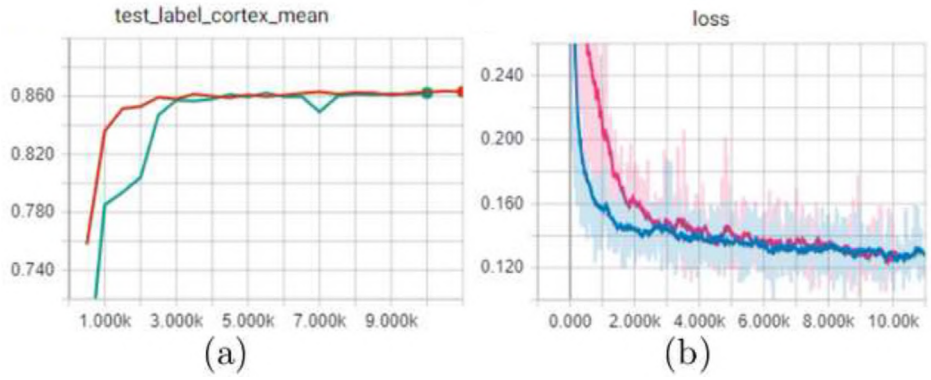


Figure 3. Impact of pre-training on the training procedure: a: evolution of the Dice score on the validation set during training (red is pre-trained, green is not); b: evolution of the binary cross-entropy on the training set (blue is pre-trained, pink is not). The *x*-axis represents the number of training steps.

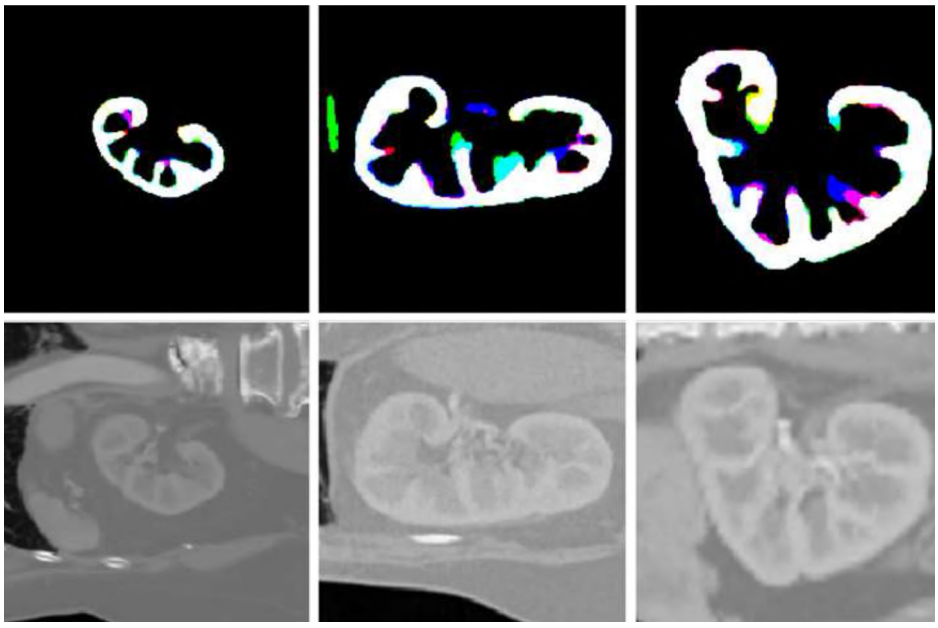


Figure 4. Top line: segmentation achieved by three networks trained on three different folds of the training database (each output is displayed on a different color channel, so that white represents a consensus for positively-labeled regions. We observe inconsistencies on the inner parts of the renal columns, and to a lesser extent on the outermost edge of the renal cortex). Bottom line: corresponding input CT images.

and was more stable overall. Therefore, we used pre-trained networks.

Post-processing and ensemble aggregation

We noticed that networks trained on different folds of the training database behave differently, especially on ambiguous pixels (Fig. 4). To improve the robustness and reduce the variability, we used ensemble aggregation.

We trained five networks on random folds of the training dataset, and two others on the complete training dataset. For each image at test time, we thus obtained seven segmentation masks taking pixel values in the interval ''[0,1]''. In each mask we only kept the largest connected component in order to remove obvious false positives (see, for instance Fig. 4, top middle: a blob is falsely labeled positively by one of the networks). Finally, we aggregated the results by taking the median value for each pixel, as it has shown to produce better results than the mean, by reducing the influence of extreme or outlier values.

Results

Participants were ranked using a Dice score: $S = \frac{2|P \cap T|}{|P|+|T|}$, where P is the predicted mask and T is the reference mask. We obtained a score of 0.867 on the test dataset and won the challenge by a narrow margin. The slight improvement obtained by the ensemble aggregation enabled us to win this challenge, as the second ranked team scored higher than our best network.

The segmentation results of our algorithm match the renal cortex with a good precision (Fig. 5). However, some of the flaws of the provided reference segmentations remain,

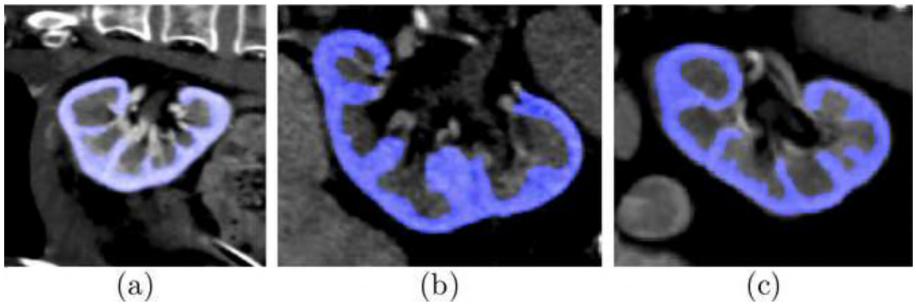


Figure 5. Illustration of automatic segmentation results obtained with the proposed approach (overlapped in blue on the input CT image); a: correct segmentation; b: cluster of renal columns; c: overextended segmentation.

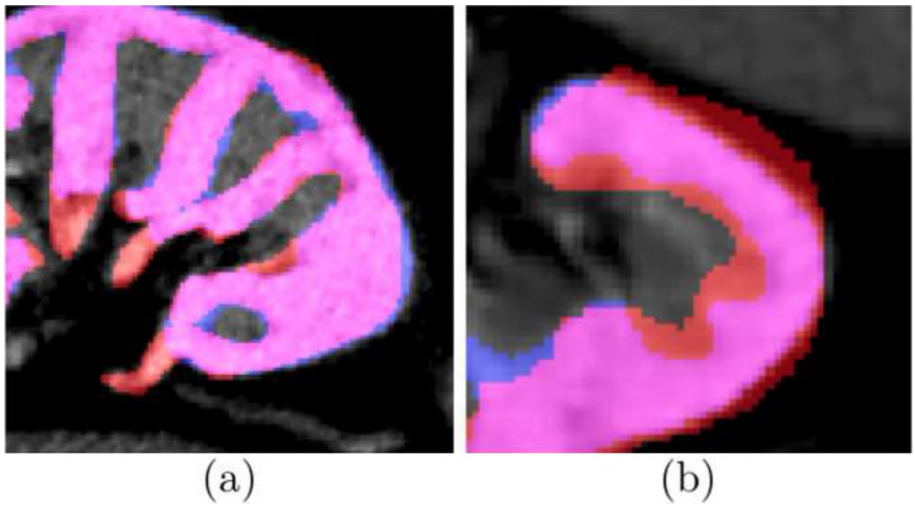


Figure 6. Illustration of test cases where the automatic segmentation results (blue) seem more accurate than the provided reference segmentation (red). Intersection in pink; a: vessels included in the reference mask but not in automatic segmentation result; b: reference segmentation obviously too wide.

such as the large clusters of renal columns, or when parts of the cortex are too widely segmented and join each other. Nonetheless, our algorithm seems to be less imprecise than the provided annotation, especially at the boundary of the cortex (Fig. 6).

Discussion

The state-of-the-art in image segmentation has improved greatly during the past five years, thanks to the progress accomplished in Deep Learning, to the point that some segmentation problems, which would have been considered a challenge ten years ago, now seem easy [27,28]. This is the case of renal cortex segmentation, where one can quickly achieve good results by training a UNet with any recent architecture found in the literature [18]. To the best of our knowledge, all the contestants chose a deep learning approach and the gap between participants was less than 0.03 Dice points.

The precision of the reference segmentations provided for this challenge seemed to set a low upper bound on the performance, as corroborated by the narrow gap between the first and second place (< 0.003 Dice points), and the gap between all the candidates (< 0.03 Dice points). As a consequence, the performance gain achieved by each of our algorithm details (image intensity scaling, data augmentation, pre-training, meta-parameter optimization, connected components analysis and ensemble aggregation) was difficult to quantify and barely significant if at all when considered alone, but enabled us, when added together, to improve the overall performance and win the challenge.

In conclusion, although 3D segmentation is useful clinically, the choice of 2D makes sense for a data challenge as it simplifies data collection, annotation, and storage [13,15—17]. Future research is needed to address the problem of renal cortex segmentation in 3D volumes.

Human and animal rights

The authors declare that the work described has been carried out in accordance with the Declaration of Helsinki of the World Medical Association revised in 2013 for experiments involving humans as well as in accordance with the EU Directive 2010/63/EU for animal experiments.

Informed consent and patient details

The authors declare that this report does not contain any personal information that could lead to the identification of the patient(s).

The authors declare that they obtained a written informed consent from the patients and/or volunteers included in the article. The authors also confirm that the personal details of the patients and/or volunteers have been removed.

Funding

This work received funding from Association Nationale de la Recherche et de la Technologie (Contract 2018/2439)

Author contributions

All authors attest that they meet the current International Committee of Medical Journal Editors (ICMJE) criteria for Authorship.

Credit author statement

Vincent Couteaux: conceptualization and design; data curation; writing-original draft preparation; review & editing.

Salim Si-Mohamed: conceptualization and design; data curation; supervision; resources; writing — original draft preparation; review & editing.

Raphaele Renard-Penna: conceptualization and design; resources; data curation; writing — original draft preparation; review & editing.

Olivier Nempont: conceptualization and design; data curation; writing — original draft preparation; review & editing.

Thierry Lefevre: conceptualization and design; data curation; writing — original draft preparation; review & editing.

Alexandre Popoff: conceptualization and design; data curation; writing — original draft preparation; review & editing.

Guillaume Pizaine: conceptualization and design; data curation; writing — original draft preparation; review & editing.

Nicolas Villain: conceptualization and design; data curation; writing — original draft preparation; review & editing.

Isabelle Bloch: conceptualization and design; data curation; writing — original draft preparation; review & editing.

Julien Behr: conceptualization and design; resources; data curation; writing — original draft preparation; review & editing.

Marie-France Bellin: conceptualization and design; resources; data curation.

Catherine Roy: conceptualization and design; resources; data curation; writing — original draft preparation; review & editing.

Olivier Rouviere: conceptualization and design; resources; data curation; writing — original draft preparation; review & editing.

Sarah Montagne: conceptualization and design; resources; data curation.

Nathalie Lassau: conceptualization and design; resources; data curation; writing — original draft preparation; review & editing.

Anne Cotten: conceptualization and design; data curation; resources; review & editing.

Loïc Boussel: conceptualization and design; supervision; writing — original draft preparation; review & editing.

Disclosure of interest

The authors declare that they have no competing interest.

References

[1] van den Dool SW, Wasser MN, de Fijter JW, Hoekstra J, van der Geest RJ. Functional renal volume: quantitative analysis at gadolinium-enhanced MR angiography–feasibility study in healthy potential kidney donors. Radiology 2005;236:189—95.

[2] Gandy SJ, Armoogum K, Nicholas RS, McLeay TB, Houston JG. A clinical MRI investigation of the relationship between kidney volume measurements and renal function in patients with renovascular disease. Br J Radiol 2007;80:12—20.

[3] Grantham JJ, Torres VE, Chapman AB, Guay-Woodford LM, Bae KT, King Jr BF, et al. Volume progression in polycystic kidney disease. N Engl J Med 2006;354:2122—30.

[4] Chen X, Summers RM, Cho M, Bagci U, Yao J. An automatic method for renal cortex segmentation on CT images: evaluation on kidney donors. Acad Radiol 2012;19:562—70.

[5] Halleck F, Diederichs G, Koehlitz T, Slowinski T, Engelken F, Liefeldt L, et al. Volume matters: CT-based renal cortex volume measurement in the evaluation of living kidney donors. Transpl Int 2013;26:1208—16.

[6] Jin C, Shi F, Xiang D, Jiang X, Zhang B, Wang X, et al. 3D fast automatic segmentation of kidney based on modified AAM and random forest. Trans Med Imaging 2016;35:1395—407.

[7] Pohle R, Toennies KD. A new approach for model-based adaptive region growing in medical image analysis. Computer Analysis of Images and Patterns Springer 2001;2124:238—46.

[8] Torimoto I, Takebayashi S, Sekikawa Z, Teranishi J, Uchida K, Inoue T. Renal perfusional cortex volume for arterial input function measured by semiautomatic segmentation technique using MDCT angiographic data with 0.5-mm collimation. AJR Am J Roentgenol 2015;204:98—104.

[9] Akkus Z, Galimzianova A, Hoogi A, Rubin DL, Erickson BJ. Deep learning for brain MRI segmentation: state of the art and future directions. J Digit Imaging 2017;30:449—59.

[10] Avendi MR, Kheradvar A, Jafarkhani H. Automatic segmentation of the right ventricle from cardiac MRI using a learning-based approach. Magn Reson Med 2017;78:2439—48.

[11] Shelhamer E, Long J, Darrell T. Fully convolutional networks for semantic segmentation. IEEE Trans Pattern Anal Mach Intell 2017;39:640—51.

[12] Chen Y, Shi B, Wang Z, Zhang P, Smith CD, Liu. J.Hippocampus segmentation through multi-view ensemble ConvNets. 2017. p. 192—6.

[13] P.F. Christ, F. Ettlinger, F. Grün, M.E.A. Elshaera, J. Lipkova, S. Schlecht, et al. Automatic liver and tumor segmentation of CT and MRI volumes using cascaded fully convolutional neural networks. https://arxiv.org/abs/1702.05970 [Accessed on March 20, 2019].

[14] Çiçek Ö, Abdulkadir A, Lienkamp SS, Brox T, Ronneberger O. 3D U-Net: learning dense volumetric segmentation from sparse

annotation. Medical Image Computing and Computer-Assisted Intervention. MICCAI, 9901. Cham: Springer; 2016 [Lecture Notes in Computer Science].
[15] Dong H, Yang G, Liu F, Mo Y, Guo Y. Automatic brain tumor detection and segmentation using U-Net based fully convolutional networks. In: Valdés Hernández M, González-Castro V, editors. Medical Image Understanding and Analysis. MIUA. Communications in Computer and Information Science, 723. Cham: Springer; 2017.
[16] Erden B, Gamboa N, Wood S. 3D convolutional neural network for brain tumor segmentation. Computer Science. Stanford University; 2017 http://cs231n.stanford.edu/reports/2017/pdfs/526.pdf.
[17] F. Milletari, N. Navab, SA. Ahmadi. V-Net: Fully convolutional neural networks for volumetric medical image segmentation. 3D Vision. IEEE 2016:565-71 [Accessed on March 20, 2019].
[18] Ronneberger O, Fischer P, Brox T. U-Net: convolutional networks for biomedical image segmentation. Medical Image Computing and Computer-Assisted Intervention; 2015. p. 234—41.
[19] Bertrand H, Ardon R, Perrot M, Bloch I. Hyperparameter optimization of deep neural networks: combining hyperband with bayesian model selection. France: CAP; 2017.
[20] Bertrand H, Perrot M, Ardon R, Bloch I. Classification of MRI data using deep learning and Gaussian process-based model selection. Biomedical Imaging. IEEE; 2017. p. 745—8.
[21] Oquab M, Bottou L, Laptev I, Sivic J. Learning and transferring mid-level image representations using convolutional neural networks. Computer Vision and Pattern Recognition. IEEE; 2014. p. 1717—24.
[22] Rokach L. Ensemble-based classifiers. Artificial Intelligence Review 2009;33:1—39.
[23] Marmanis D, Wegner JD, Galliani S, Schindler K, Datcu M, Stilla U. Semantic segmentation of aerial images with an ensemble of CNNs, ISPRS Annals of the Photogrammetry. Remote Sens Spatial Info Sci 2016;III:473—80.
[24] He K, Zhang X, Ren S, Sun J. Deep residual learning for image recognition. Computer Vision and Pattern Recognition. IEEE; 2016. p. 770—8.
[25] Peng C, Zhang X, Yu G, Luo G, Sun J. Large kernel matters improve semantic segmentation by global convolutional network. Computer Vision and Pattern Recognition. IEEE; 2017. p. 1743—51.
[26] Lin TY, Maire M, Belongie SJ, Bourdev LD, Girshick RB, Hays J, et al. Microsoft COCO: common objects in context. European Conference on Computer Vision; 2014. p. 740—55.
[27] Garcia-Garcia A, Orts S, Oprea S, Villena-Martinez V, Rodríguez JG. A review on deep learning techniques applied to semantic segmentation. Computer Vision and Pattern Recognition. Cornell University; 2017.
[28] LeCun Y, Bengio Y, Hinton G. Deep learning. Nature 2015;521:436—44.

Diagnostic and Interventional Imaging (2019) **100**, 235—242

ORIGINAL ARTICLE / *Computer developments*

Automatic knee meniscus tear detection and orientation classification with Mask-RCNN

V. Couteaux [a,b,*], S. Si-Mohamed [c,d], O. Nempont [a], T. Lefevre [a], A. Popoff [a], G. Pizaine [a], N. Villain [a], I. Bloch [b], A. Cotten [e], L. Boussel [c,d]

[a] *Philips Research France, 33, rue de Verdun, 92150 Suresnes, France*
[b] *LTCI, Télécom ParisTech, université Paris-Saclay, 46, rue Barrault, 75013 Paris, France*
[c] *Inserm U1206, INSA-Lyon, Claude-Bernard-Lyon 1 University, CREATIS, CNRS UMR 5220, 69100 Villeurbanne, France*
[d] *Department of Radiology, hospices civils de Lyon, 69002 Lyon, France*
[e] *Department of Musculoskeletal Radiology, CHRU de Lille, 59000 Lille, France*

KEYWORDS
Knee meniscus;
Artificial intelligence;
Mask region-based convolutional neural network (R-CNN);
Meniscal tear detection;
Orientation classification

Abstract
Purpose: This work presents our contribution to a data challenge organized by the French Radiology Society during the *Journées Francophones de Radiologie* in October 2018. This challenge consisted in classifying MR images of the knee with respect to the presence of tears in the knee menisci, on meniscal tear location, and meniscal tear orientation.
Materials and methods: We trained a mask region-based convolutional neural network (R-CNN) to explicitly localize normal and torn menisci, made it more robust with ensemble aggregation, and cascaded it into a shallow ConvNet to classify the orientation of the tear.
Results: Our approach predicted accurately tears in the database provided for the challenge. This strategy yielded a weighted AUC score of 0.906 for all three tasks, ranking first in this challenge.
Conclusion: The extension of the database or the use of 3D data could contribute to further improve the performances especially for non-typical cases of extensively damaged menisci or multiple tears.

* Corresponding author. Philips Research France, 33, rue de Verdun, 92150 Suresnes, France.
E-mail address: vincent.couteaux@telecom-paristech.fr (V. Couteaux).

https://doi.org/10.1016/j.diii.2019.03.002

Introduction

Meniscal lesions are a frequent and common cause of knee pain, responsible for approximately 700,000 arthroscopic partial meniscectomies per year in the United States [1]. They are defined as a tear within the meniscus, and can lead to articular cartilage degeneration over time, further necessitating surgical treatment. Magnetic resonance imaging (MRI) plays a central role in the diagnosis of meniscus lesions, the preoperative planning and the postoperative rehabilitation of the patient [2,3]. As meniscal lesions are very frequent, their diagnosis could certainly benefit from a quantitative and automated solution giving more accurate results in a faster way. Computer-aided detection systems for meniscal tears were thus proposed whereby regions of interest in the image are extracted and classified based on handcrafted image features [4–8].

The *Journées Francophones de Radiologie* was held in Paris in October 2018. For the first time, the French Society of Radiology organized an artificial intelligence (AI) competition involving teams of industrial researchers, students and radiologists.

This paper presents our contribution to the knee meniscus tear challenge, where participants had to classify sagittal MRI slices cropped around the knee depending on the presence of tears in anterior and posterior menisci and on their orientation (horizontal or vertical). We proposed a method that takes advantage of recent advances in deep learning [9,10]. More precisely, we propose to localize, segment, and classify healthy and torn menisci using a mask region-based convolutional neural network (R-CNN) approach that is cascaded into a shallow ConvNet to classify tear orientation.

Method

Knee meniscus tear challenge

Sagittal MR images centered around the knee were provided with the following annotations:

- position of the image (medial or lateral);
- presence of a tear in the posterior meniscus;
- presence of a tear in the anterior meniscus;
- orientation of the tear in the posterior meniscus (if any);
- orientation of the tear in the anterior meniscus (if any).

Two training batches were provided; the first made of 257 images was shared one month before the conference and the other, made of 871 images, 2 days before the end of the challenge. The first batch contained 55/257 (21.4%) images with horizontal posterior tears, 46/257 (17.9%) with vertical posterior tears, 13/257 (5.1%) with horizontal anterior tears and 8/257 (3.1%) with vertical anterior tears. The second batch contained 107/871 (12.3%) images with horizontal posterior tears, 60/871 (6.9%) with vertical posterior tears, 8/871 (0.9%) with horizontal anterior tears and 3/871 (0.3%) with vertical anterior tears. The classes were imbalanced, with horizontal tears and posterior meniscus tears being more frequent, and a low number of anterior tears were available for training. We reviewed the database and removed any ambiguous annotations from the training set.

Images of size 256×256, either of the medial or of the lateral plane of the knee, were provided, as illustrated in Fig. 1a, b. The femur was always on the left and the tibia on the right, with the anterior meniscus at the top and the posterior meniscus at the bottom of the image. Horizontal tears appeared vertical and vice versa. The grey level scale was in an arbitrary unit scaled between 0 and 1, and the

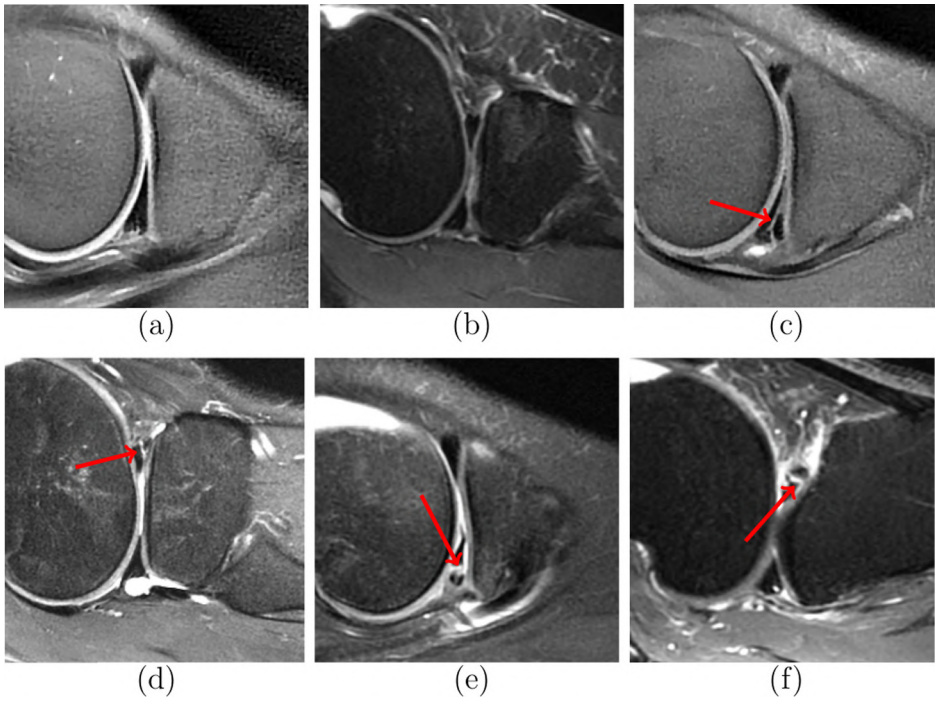


Figure 1. Database contains either medial e.g. (a) or lateral e.g. (b) MR images of the knee. (a–b) MR images shows healthy menisci. (c–f) MR images shows examples of tears as present in the database. (c) Horizontal tear in posterior meniscus, (d) Horizontal tear in anterior meniscus, (e) Vertical tear in posterior meniscus and (f), Vertical tear in anterior meniscus. Arrows point out tears.

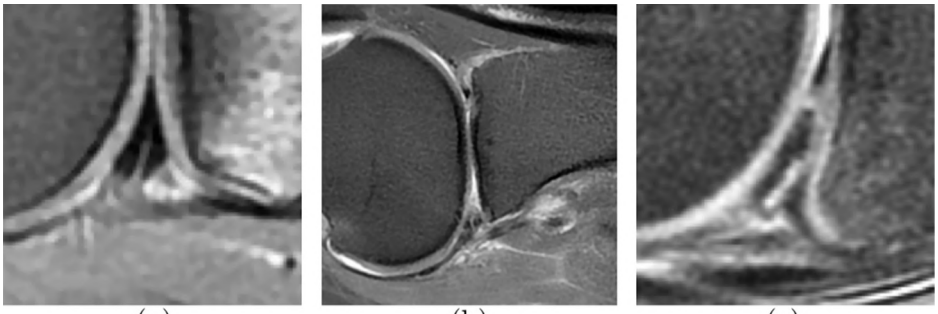


Figure 2. MR images illustrate challenging cases. (a) Potentially misleading lesion. (b) Barely visible meniscus. (c) Multiple tears in the same meniscus.

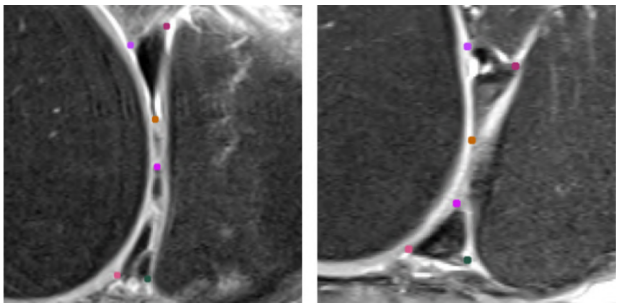


Figure 3. Figure shows two MR images from the training database illustrating the menisci annotation process. Each colored dot is a vertex of a triangle that approximates the segmentation of a meniscus.

images did not have consistent brightness and contrast.

On MRI, meniscal tears appear as thin, hyperintense lines that cut across the menisci, but we can observe various hyper-intense signals in the menisci (Fig. 1). For instance, hyper-intense lines that only partially cut across the menisci should not be classified as tears (Fig. 2a). Moreover, the menisci may be barely visible (Fig. 2b) or have multiple tears (Fig. 2c). In the latter case, the provided orientation may be ill-defined.

Menisci are small structures and tears present as thin abnormalities within menisci on MRI. To facilitate tear detection, we first localized the menisci. However, the localization of the menisci was not part of the provided annotations. To efficiently perform this annotation, we chose to approximate menisci by triangles resulting in a coarse segmentation of menisci (Fig. 3).

To detect tears in both menisci and identify their orientation, we opted for a cascaded approach. First, menisci were localized and tears were identified. Then the orientation of torn menisci was classified. For both tasks, we applied a morphological pre-processing, as described below, to enhance the relevant structures in the image.

Morphological pre-processing of images

In a first attempt to better understand the dataset and the classification task at hand, we trained simple neural network classifiers on the training dataset in order to classify the posterior (resp. anterior) meniscus into healthy and torn cases. Both networks were based on a simplified VGG-like architecture: the whole image is taken as the input, on which four 2D-convolution/ReLu/Maxpool layers are applied followed by two dense layers and a final softmax-activated output layer [11]. Note that these neural networks were only meant to explore the given dataset.

An accuracy of 83% could readily be obtained (precision, 0.76; recall, 0.8). However, when analyzing the interpretation of the ConvNet classification, it appeared that the network used non-relevant features in the image to provide the result (Fig. 4). This phenomenon was consistently observed on other images in the dataset and is probably attributed to the variability of the images. Given that the dataset is small, the network is not able to properly generalize on so few samples. This prompted us to consider pre-processing the images in order to bring robustness to the classification. The approach we propose is close to that for detecting white matter hyper-intensities in brain MRI [12].

Since meniscus tears are primarily characterized by distinct morphological features, it seems reasonable to assume that any image pre-processing step that would retain these characteristics while limiting the influence of other structures may be beneficial. Fig. 4c shows the result of applying a black top-hat morphological filter on the image of Fig. 4a. A black top-hat filter outputs an image wherein the bright regions correspond to regions in the original image which are smaller than the structuring element and darker than their surroundings. As can be clearly observed, the torn posterior meniscus can still be identified. A ConvNet classifier trained on black top-hat filtered images shows performance similar to the initial one (accuracy, 83%; precision, 86%; recall, 67%) but is able to focus precisely on the meniscus region (Fig. 4d). In this case, the saliency map clearly indicates that only the meniscus region is relevant for the classification.

Meniscus localization and tear detection

To localize both menisci and identify tears in each meniscus, we used the Mask R-CNN framework, a state-of-the-art approach for Instance Segmentation. It performs object detection, segmentation and classification in a single forward pass [13]. We trained the model to detect and segment four objects:

- healthy anterior meniscus;
- torn anterior meniscus;
- healthy posterior meniscus;
- torn posterior meniscus.

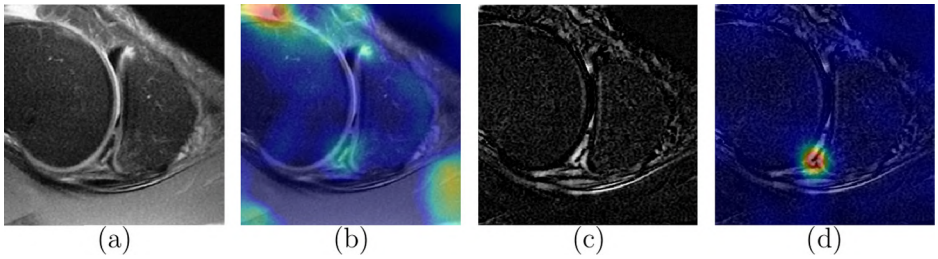


Figure 4. (a) Image from the training database with a clearly visible torn posterior meniscus that was correctly classified by the ConvNet. (b) Same image with a superimposed saliency mask indicating that the network focuses on non-relevant regions and barely considers the posterior meniscus itself. (c) Image (a) after applying a black top-hat filter with a disk structuring element of radius 5 pixels. (d) Saliency map for the processed image.

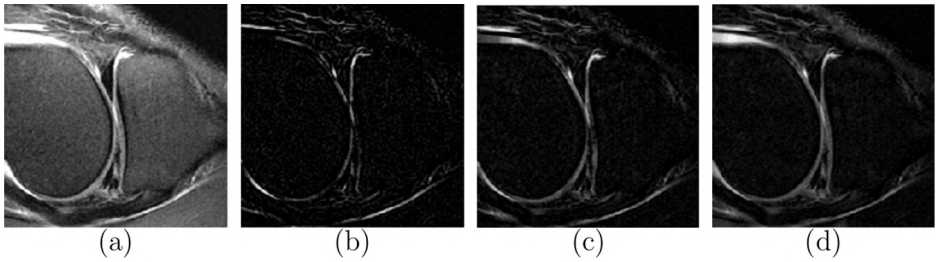


Figure 5. Pre-processing of the data used as input of the Mask R-CNN. (a) Original image. (b) 5×5 white top-hat. (c) 11×11 white top-hat. (d) 21×21 white top-hat.

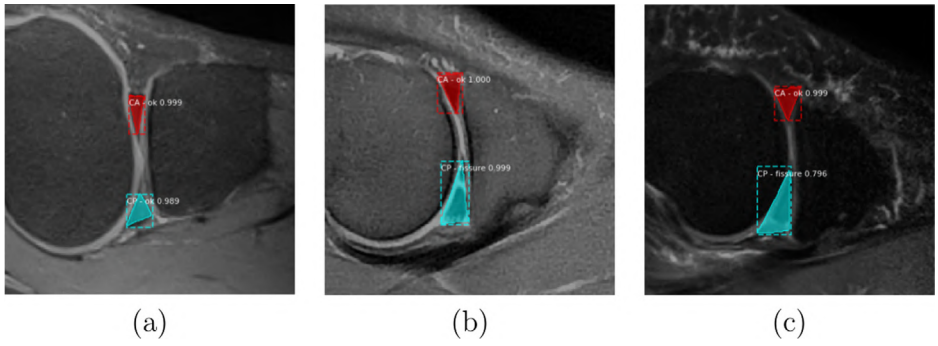


Figure 6. Output of Mask R-CNN. (a—b) Correct results. (c) Posterior meniscus incorrectly segmented and labeled as torn.

In this way, we obtained the localization of each meniscus, the classification of healthy vs. torn, and a classification score. We chose to perform the classification of tear orientation independently on the segmented meniscus region only, as explained below because the classes would have been too imbalanced otherwise (only 11 vertical tears in the anterior meniscus for instance).

We used a Mask R-CNN model pre-trained on the common object in context (COCO) dataset [14] whose input is a three channel image. We applied three white top-hat filters (the dual of the black top-hat filters described above) on original MRI slices with square structuring elements of size 5×5, 11×11 and 21×21 (Fig. 5) to generate network inputs. Note that we did not constrain the model to return exactly one result for each meniscus because the two menisci were correctly detected in almost all cases. We illustrate in Fig. 6 the output of the Mask R-CNN. In Fig. 6a, the two healthy menisci are properly detected. In Fig. 6b, the posterior meniscus is appropriately identified as torn. However, the posterior meniscus is too widely segmented and incorrectly labeled as torn (Fig. 6).

Training

We fine-tuned a Mask R-CNN with a ResNet-101 backbone, pretrained on COCO dataset) [13—15]. The training was done using an Adam optimizer, 1.10 −3 learning rate and batches of 8 images, during 1000 epochs of 100 batches.

Ensemble aggregation

To improve the robustness of our model, we applied ensemble aggregation. We trained five models on random folds of the full training data set (1128 images) and retained five additional models trained on random folds of the first training batch only (the first 257 images). We aggregated the

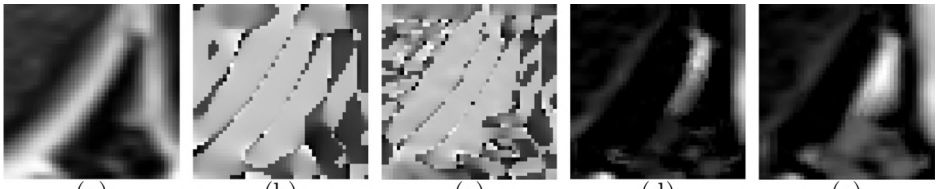


Figure 7. Patch extraction for orientation classification. (a) Extracted patch, resized to 47 × 47. (b) Local orientation map, $\sigma = 3$. (c) Local orientation map, $\sigma = 1$. (d) Black top-hat, r = 4. (e) Black top-hat, r = 8.

results differently for anterior and posterior menisci. We classified the anterior meniscus as torn when at least one network had detected a torn anterior meniscus, with a probability $P_{ant}(F)$ equal to the mean classification score of all detected torn anterior menisci by the ensemble. We classified the posterior meniscus as torn when the strict majority of the networks had detected a torn posterior meniscus. The probability $P_{post}(F)$ is equal to the mean classification score of all detected torn posterior menisci by the ensemble. We used different aggregation methods as a large majority of anterior menisci are healthy. Some networks may not have seen enough torn anterior menisci in order to recognize them.

Tear orientation classification

To classify the orientation of torn menisci as horizontal or vertical, we trained a neural network on images cropped to the bounding boxes of detected torn menisci, resized to 47 × 47 pixels. This network was fed with pre-processed patches, each input having five channels illustrated in Fig. 7:

- unprocessed patch;
- local orientation map, computed with $\sigma = 3$ (see below);
- local orientation map, computed with $\sigma = 1$;
- black top-hat transform, with a disk structuring element of radius 4 pixels;
- black top-hat transform, with a disk structuring element of radius 8 pixels.

The local orientation map represents the angle of the smallest eigenvector of the Hessian matrix at each pixel. The Hessian matrix was computed with the second derivative of a Gaussian kernel, whose standard deviation σ is a parameter.

Only 300 torn menisci were provided for training. Therefore, we trained a very shallow CNN based on a VGG-like architecture:

- Convolution, 3 × 3 kernel, 8 filters, ReLU activation;
- Max-pooling, 2 × 2;
- Convolution, 3 × 3 kernel, 16 filters, ReLU activation;
- Max-pooling, 2 × 2;
- Convolution, 3 × 3 kernel, 32 filters, ReLU activation;
- Max-pooling, 2 × 2;
- Dense Layer with 1024 units, ReLU activation, $P = 0.5$ dropout;
- Dense Layer with 1024 units, ReLU activation, $P = 0.5$ dropout;
- Dense Layer with 2 units and a softmax activation.

We trained this network on 246 torn menisci of the training database with a Stochastic Gradient Descent, 1.10 −3 learning rate and batches of 32 images, during 800 epochs (approximately 5 min). We validated the method on the remaining 54 cases and selected the model with the highest validation accuracy.

Results

Score and ranking

Teams were ranked according to a weighted average of the area under the ROC curves (AUC) of the tear detection task *Det* (tear in any meniscus), the tear localization task *Loc* (anterior or posterior) and the orientation classification task *Or* (horizontal or vertical), according to Eq. 1 (E1):

$$Score = 0.4 \times AUC\,(Det) + 0.3 \times AUC\,(Loc) + 0.3 \times AUC\,(Or) \quad \text{(E1)}$$

The organizers therefore removed from the database cases where both menisci had tears and the following values were submitted for each image:

- Probability of a tear in any meniscus *P(F)*;
- Probability that the tear (if any) is in the anterior meniscus *P(Ant)*;
- Probability that the tear (if any) is horizontal *P(H)*.

The Mask R-CNN ensemble outputs a probability $P_{ant}(F)$ that the anterior meniscus is torn, and a probability $P_{post}(F)$ that the posterior meniscus is torn, both being independent a priori. This results in Eq. 2 (E2)

$$P\,(F) = P_{post}\,(F) + P_{ant}\,(F) - P_{post}\,(F)\,P_{ant}\,(F) \quad \text{(E2)}$$

where *P(Ant)* is defined by Equation 3 (E3)

$$P\,(Ant) = P_{ant}\,(F) \,/\, \left(P_{ant(F)} + P_{post}\,(F)\right) \quad \text{(E3)}$$

To obtain *P(H)*, we applied the orientation classifier on the anterior meniscus when *P(Ant)* > 0.5 and on the posterior meniscus otherwise.

A test set of 700 images was used for ranking. We obtained a score of 0.906 and shared the first place with another team (score 0.903).

Visual inspection

In most cases, the prediction was in line with our interpretation as illustrated in Fig. 8, but a few cases seemed suspicious. The resulting classification scores were almost binary, either very close to 1 or very close to 0, especially *P(F)*. However, for some images, the predictor returned classification scores close to 0.5 (Fig. 9).

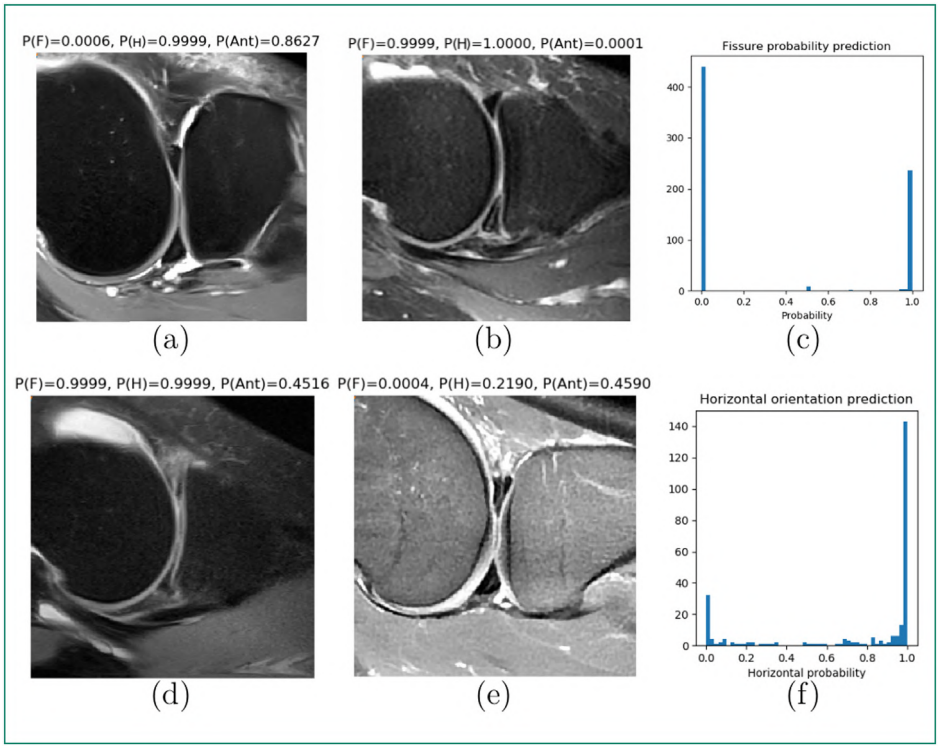


Figure 8. Prediction results on the testing batch. Most results seem correct, *e.g.* (a—b). However, some predictions are suspicious, e.g. (d—e). (a) No tear. (b) Horizontal tear on the posterior meniscus. (d) *P(Ant)* $\sim$ 0.45 but the anterior meniscus looks torn. (e) *P(F)* $\sim$ 0 but a tear is visible in the anterior meniscus. (c) Distribution of *P(F)*. (f) Distribution of P (H) for cases satisfying *P(F)* > 0.5.

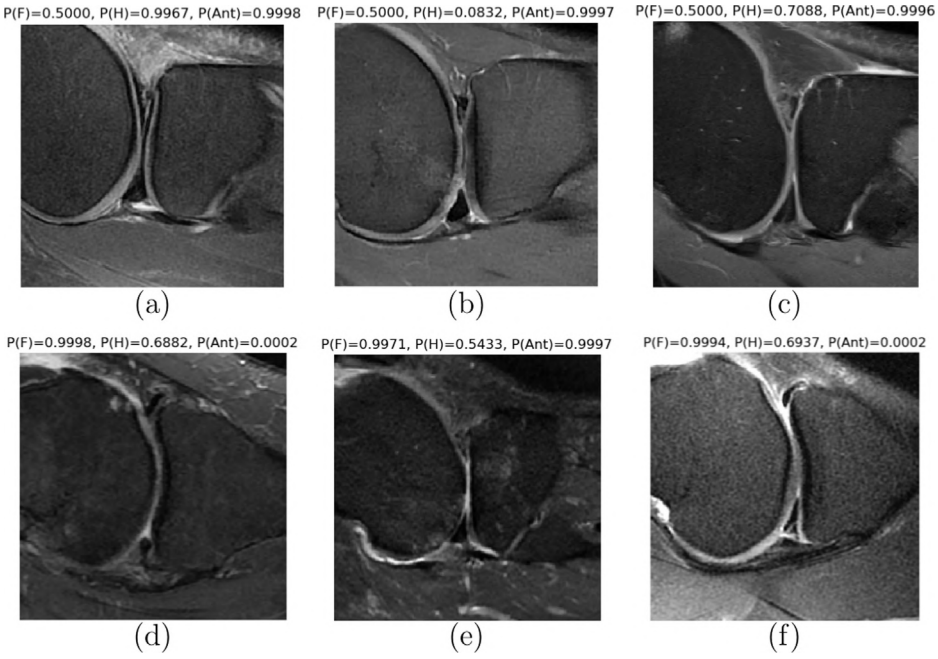


Figure 9. Cases for which P (F) (a—c) or P (H) (d—f) were close to 0.5. (a) Tear on the anterior meniscus but a slice where the menisci are connected was selected which does not meet the inclusion criteria. (b) Damaged anterior meniscus, but the presence of a tear is unclear. Yet the algorithm focused on the anterior meniscus: P (Ant) > 0.99. (c) Untypical lesion on the anterior meniscus. (d—e) Extensively damaged meniscus. (f) Several tears in one meniscus.

Discussion

The knee meniscus tear challenge posed an image classification problem. Image classification tasks in computer vision aim to discriminate many classes from the prominent object in the image [16]. However, in this problem, the classification result should be based on thin details at a specific location in the image. Moreover, only a small database was available. Training a standard classifier from the image may therefore result in sub-optimal performances as observed in our initial experiments.

We chose to localize and segment the menisci and perform the classification within the anterior and posterior menisci. We opted for a Mask R-CNN approach as it can perform both tasks jointly [13]. Due to the class imbalance, we did not classify the tear orientation using this approach, but cascaded the Mask R-CNN into a shallow ConvNet to classify tear orientation [17]. Moreover, we provided pre-processed images to the networks to focus on relevant parts of the images by enhancing the tears.

In conclusion, our approach ranked first in the challenge by predicting accurately tears in the database provided for the challenge. The extension of the database or the use of 3D data could contribute to further improve the performances especially on untypical cases such as very damaged menisci or multiple tears.

Human and animal rights

The authors declare that the work described has been carried out in accordance with the Declaration of Helsinki of the World Medical Association revised in 2013 for experiments involving humans as well as in accordance with the EU Directive 2010/63/EU for animal experiments.

Informed consent and patient details

The authors declare that this report does not contain any personal information that could lead to the identification of the patient(s).

The authors declare that they obtained a written informed consent from the patients and/or volunteers included in the article. The authors also confirm that the personal details of the patients and/or volunteers have been removed.

Funding

This work has been partially funded by the ANRT (*Association nationale de la recherche et de la technologie*)

Author contributions

All authors attest that they meet the current International Committee of Medical Journal Editors (ICMJE) criteria for Authorship.

Credit author statement

Vincent Couteaux: conceptualization and design; data curation; writing — original draft preparation; review & editing.

Salim Si-Mohamed: conceptualization and design; data curation; writing — original draft preparation; review & editing.

Olivier Nempont: conceptualization and design; data curation; writing — original draft preparation; review & editing.

Thierry Lefevre: conceptualization and design; data curation; writing — original draft preparation; review & editing.

Alexandre Popoff: conceptualization and design; data curation; writing — original draft preparation; review & editing.

Guillaume Pizaine: conceptualization and design; data curation; writing — original draft preparation; review & editing.

Nicolas Villain: conceptualization and design; data curation; writing — original draft preparation; review & editing.

Isabelle Bloch: conceptualization and design; data curation; writing — original draft preparation; review & editing.

Anne Cotten: conceptualization and design; data curation; resources; review & editing.

Loïc Boussel: conceptualization and design; supervision; writing — original draft preparation; review & editing.

Disclosure of interest

The authors declare that they have no competing interest.

References

[1] Cullen KA, Hall MJ, Golosinskiy A. Ambulatory surgery in the United States, 2006. Natl Health Stat Rep 2009;11:1—25.

[2] Pache S, Aman ZS, Kennedy M, Nakama GY, Moatshe G, Ziegler C, et al. Meniscal root tears: current concepts review. Bone Joint Surg 2018;6:250—9.

[3] Lecouvet F, Van Haver T, Acid S, Perlepe V, Kirchgesner T, Vande Berg B, et al. Magnetic resonance imaging (MRI) of the knee: Identification of difficult-to-diagnose meniscal lesions. Diagn Interv Imaging 2018;99:55—64.

[4] Boniatis I, Panayiotakis GS, Panagiotopoulos E. A computer-based system for the discrimination between normal and degenerated menisci from magnetic resonance images. Int Workshop Imaging Syst Techn IEEE 2008:335—9.

[5] Köse C, Gençalioğlu O, Şevik UU. An automatic diagnosis method for the knee meniscus tears in MR images. Expert Syst Appl 2009;36:1208—16.

[6] Ramakrishna B, Liu W, Saiprasad G, Safdar N, Chang CI, Siddiqui K, et al. An automatic computer-aided detection system for meniscal tears on magnetic resonance images. IEEE Trans Med Imaging 2009;28:1308—16.

[7] Saygili A, Albayrak S. Meniscus segmentation and tear detection in the knee MR images by fuzzy c-means method. Signal Processing and Communications Applications Conference (SIU); 2017. p. 1—4.

[8] Saygili A, Albayrak S. Meniscus tear classification using histogram of oriented gradients in knee MR images. Signal Processing and Communications Applications Conference (SIU); 2018. p. 1—4.

[9] Garcia-Garcia A, Orts S, Oprea S, Villena-Martinez V, Rodríguez JG. A review on deep learning techniques applied to semantic segmentation. Computer Vision and Pattern Recognition. Cornell University; 2017.
[10] LeCun Y, Bengio Y, Hinton G. Deep learning. Nature 2015;521:436—44.
[11] Simonyan K, Zisserman A. Very deep convolutional networks for large-scale image recognition. arXiv preprint arXiv :1409.1556; 2014.
[12] Xu Y, Géraud T, Puybareau E, Bloch I, Chazalon J. White matter hyperintensities segmentation in a few seconds using fully convolutional network and transfer learning, brain lesion: glioma, multiple sclerosis, stroke and traumatic brain injuries. Lect Notes Comp Sci 2017:1067.
[13] He K, Gkioxari G, Dollár P, Girshick R. Mask R-CNN. International Conference on Computer Vision (ICCV). IEEE; 2017. p. 2980—8.
[14] Lin TY, Maire M, Belongie SJ, Bourdev LD, Girshick RB, Hays J, et al. Microsoft COCO: Common objects in context. European Conference on Computer Vision (ECCV); 2014. p. 740—55.
[15] He K, Zhang X, Ren S, Sun J.Deep residual learning for image recognition. 2016. p. 770—8.
[16] Russakovsky O, Deng J, Su H, Krause J, Satheesh S, Ma S, et al. Imagenet large scale visual recognition challenge. Int Comput Vision 2015;115:211—52.
[17] Krizhevsky A, Sutskever I, Hinton GE. Imagenet classification with deep convolutional neural networks. Adv Neural Info Proc Syst 2012:25.

ECOLE DOCTORALE

Titre : Apprentissage profond pour la segmentation et la détection automatique en imagerie multi-modale. Application à l'oncologie hépatique

Mots clés : Segmentation, Détection d'objets, Interprétabilité, Apprentissage profond, IRM, Lésions hépatiques

Résumé : Pour caractériser les lésions hépatiques, les radiologues s'appuient sur plusieurs images acquises selon différentes modalités (différentes séquences IRM, tomodensitométrie, etc.) car celles-ci donnent des informations complémentaires. En outre, les outils automatiques de segmentation et de détection leur sont d'une grande aide pour la caractérisation des lésions, le suivi de la maladie ou la planification d'interventions. A l'heure où l'apprentissage profond domine l'état de l'art dans tous les domaines liés au traitement de l'image médicale, cette thèse vise à étudier comment ces méthodes peuvent relever certains défis liés à l'analyse d'images multi-modales, en s'articulant autour de trois axes : la segmentation automatique du foie, l'interprétabilité des réseaux de segmentation et la détection de lésions hépatiques.

La segmentation multi-modale dans un contexte où les images sont appariées mais pas recalées entre elles est un problème peu abordé dans la littérature. Je propose une comparaison de stratégies d'apprentissage proposées pour des problèmes voisins, ainsi qu'une méthode pour intégrer une contrainte de similarité des prédictions à l'apprentissage.

L'interprétabilité en apprentissage automatique est un champ de recherche jeune aux enjeux particulièrement importants en traitement de l'image médicale, mais qui jusqu'alors s'était concentré sur les réseaux de classification d'images naturelles. Je propose une méthode permettant d'interpréter les réseaux de segmentation d'images médicales.

Enfin, je présente un travail préliminaire sur une méthode de détection de lésions hépatiques dans des paires d'images de modalités différentes.

Title : Deep Learning for automatic segmentation and detection in multi-modal imaging. Application to hepatic oncology

Keywords : Segmentation, Object detection, Interpretability, Deep learning, MRI, Liver lesions

Abstract : In order to characterize hepatic lesions, radiologists rely on several images using different modalities (different MRI sequences, CT scan, etc.) because they provide complementary information. In addition, automatic segmentation and detection tools are a great help in characterizing lesions, monitoring disease or planning interventions. At a time when deep learning dominates the state of the art in all fields related to medical image processing, this thesis aims to study how these methods can meet certain challenges related to multi-modal image analysis, revolving around three axes : automatic segmentation of the liver, the interpretability of segmentation networks and detection of hepatic lesions.

Multi-modal segmentation in a context where the images are paired but not registered with respect to each other is a problem that is little addressed in the literature. I propose a comparison of learning strategies that have been proposed for related problems, as well as a method to enforce a constraint of similarity of predictions into learning.

Interpretability in machine learning is a young field of research with particularly important issues in medical image processing, but which so far has focused on natural image classification networks. I propose a method for interpreting medical image segmentation networks.

Finally, I present preliminary work on a method for detecting liver lesions in pairs of images of different modalities.

Institut Polytechnique de Paris
91120 Palaiseau, France