Semantic segmentation of 3D medical images with deep learning

Olivier Petit

HESAM UNIVERSITÉ

le cnam

ÉCOLE DOCTORALE Sciences des Métiers de l'Ingénieur
Centre d'études et de recherche en informatique et communications

THÈSE

présentée par : **Olivier PETIT**

soutenue le : **17 décembre 2021**

pour obtenir le grade de : **Docteur d'HESAM Université**

préparée au : **Conservatoire national des arts et métiers**

Discipline : **Sciences et technologies de l'information et de la communication**
Spécialité : **Informatique**

Segmentation sémantique d'images médicales 3D par deep learning

THÈSE dirigée par :
Pr. THOME Nicolas CEDRIC, Cnam

Jury

M. Olivier BERNARD	Professeur, CREATIS, University de Lyon	Rapporteur
Mme. Caroline PETITJEAN	Professeure, LITIS, Université de Rouen Normandie	Rapporteur
M. Pierrick COUPÉ	Directeur de recherche CNRS, Université de Bordeaux	Examinateur
M. Michel CRUCIANU	Professeur, CEDRIC, CNAM Paris	Président
Mme Diana MATEUS	Professeure, LS2N, Centrale Nantes	Examinateur
M. Christian WOLF	Maître de conférence HDR, LIRIS, INSA de Lyon	Examinateur
M. Nicolas THOME	Professeur, CEDRIC, CNAM Paris	Directeur
M. Luc SOLER	Professeur, Visible Patient, Strasbourg	Co-Directeur

T H È S E

Affidavit

Je soussigné, Olivier Petit, déclare par la présente que le travail présenté dans ce manuscrit est mon propre travail, réalisé sous la direction scientifique de Pr. Nicolas Thome et de Pr. Luc Soler, dans le respect des principes d'honnêteté, d'intégrité et de responsabilité inhérents à la mission de recherche. Les travaux de recherche et la rédaction de ce manuscrit ont été réalisés dans le respect de la charte nationale de déontologie des métiers de la recherche. Ce travail n'a pas été précédemment soumis en France ou à l'étranger dans une version identique ou similaire à un organisme examinateur.

Fait à Paris, le 29/10/2021

Signature 

Affidavit

I, undersigned, Olivier Petit, hereby declare that the work presented in this manuscript is my own work, carried out under the scientific direction of Pr. Nicolas Thome and of Pr. Luc Soler, in accordance with the principles of honesty, integrity and responsibility inherent to the research mission. The research work and the writing of this manuscript have been carried out in compliance with the French charter for Research Integrity. This work has not been submitted previously either in France or abroad in the same or in a similar version to any other examination body.

Place Paris, date 29/10/2021

Signature 

Remerciements

Il me serait difficile de remercier individuellement toutes les personnes qui m'ont soutenu et grâce à qui j'ai pu mener cette thèse à son terme.

Je tiens tout d'abord à remercier vivement le professeur Nicolas Thome qui m'a accueilli au sein de son équipe de recherche au laboratoire CEDRIC du CNAM. Il m'a fait confiance et m'a accompagné tout au long de ces quatre années. Il s'est beaucoup investi pour la réussite de cette recherche même dans les moments où les échéances se rapprochaient et où il fallait donner le maximum.

Je remercie également le professeur Luc Soler sans qui cette thèse n'aurait pas pu exister et qui a eu confiance dans mon travail du début jusqu'à la fin. Je dois aussi remercier toutes les personnes chez Visible Patient pour leur accueil et plus particulièrement Arnaud Charnoz et Julien Weinzorn pour leur accompagnement et leur soutien mais aussi Adrien Heitz, Etienne Landure et Eric Alber.

Je tiens également à remercier mes camarades du CNAM Thi-Lam-Thuy Le, Laura Calem, Charles Corbière et Vincent Le Guen avec qui j'ai commencé ma thèse. Nous avons beaucoup partagé : les déménagements successifs du début et la crise sanitaire. Mais aussi plus récemment Loïc Themyr avec qui j'ai le plaisir de partager un projet.

Je remercie la professeure Carole Lartizien et le professeur Christian Wolf pour avoir participé aux différents comités de suivi, leurs conseils ont été essentiels pour la réalisation de cette thèse.

Évidement, je tiens à remercier l'ensemble des membres du jury pour l'intérêt qu'ils ont porté à mes travaux, pour avoir accepté de relire ma thèse puis d'assister à la soutenance qui finalise le travail de ces quatre années de recherche.

Pour finir, je veux remercier toute ma famille et mes amis dont le soutien et la présence dans les moments les plus difficiles m'ont permis d'aller jusqu'au bout. Tout particulièrement Ophélie Guenoux avec qui je partage ma vie mais aussi mes parents Jean-Paul Petit et Maryse Harvey ainsi que mes sœurs Marie-Hélène Petit et Amélie Petit.

Résumé

L'apprentissage profond a récemment montré des résultats impressionnants en vision par ordinateur en particulier avec les performances atteintes par les réseaux de neurones convolutifs. Ces méthodes ont redéfini l'état-de-l'art dans de nombreuses applications telles que la segmentation d'images médicales. Dans cette thèse nous abordons le problème de segmentation des organes de l'abdomen en utilisant des méthodes issues de l'apprentissage profond.

Premièrement, nous nous sommes intéressés à l'entraînement de réseaux de neurones convolutifs profonds avec des bases de données partiellement étiquetées. Les professionnels se concentrent souvent sur des régions anatomiques précises, ce qui a pour conséquence de constituer des bases de données hétérogènes et donc partiellement étiquetées. Malheureusement, entraîner un modèle de segmentation directement sur ces bases donne de très mauvais résultats à cause de la présence d'étiquettes erronées là où sont situés les organes manquants. Dans notre méthode, nous proposons un schéma d'entraînement qui utilise toutes les étiquettes disponibles sans être affecté par les mauvaises. De plus, nous proposons un schéma itératif permettant de progressivement étiqueter les organes manquants dans l'ensemble d'entraînement ce qui permet d'améliorer encore notre modèle.

Dans un second temps, nous avons étudié l'utilisation d'un a priori spatial sur la position absolue des organes afin d'améliorer la détection des structures et réduire les erreurs aberrantes de segmentation. Les réseaux convolutifs qui sont largement utilisés en classification ne permettent pas de capturer l'information de position spatiale absolue. Cependant, les images médicales sont très structurées et il y a des conventions sur les positions attendues des organes. Dans ces travaux nous proposons un a priori spatial 3D qui capture la position des organes et qui va explicitement biaiser le modèle grâce à une fonction d'activation « prior-driven ». En plus d'améliorer la segmentation des organes difficiles, nous montrons que l'utilisation de notre a priori spatial dans un schéma de pseudo-labeling permet d'obtenir de très bons résultats même avec peu de données étiquetées en empêchant l'ajout de faux positifs dans les données d'entraînement.

Pour finir, nous avons étudié les modèles Transformers qui permettent de modéliser des interactions à long terme entre les structures anatomiques dans un modèle de segmentation

classiquement utilisé en segmentation d'organes. Les réseaux convolutifs traditionnels ne permettent pas de capturer ces interactions globales principalement à cause de leur champ réceptif limité souvent plus petit que la taille de l'image d'entrée. Utiliser le mécanisme d'attention dense proposé dans les modèles Transformers permet de connecter tous les pixels entre eux, ce qui a pour conséquence de modéliser des interactions complexes entre les différentes parties de l'image. Nous avons montré qu'utiliser un tel mécanisme d'attention améliore significativement la qualité de la segmentation sur plusieurs bases de données avec des gains plus importants sur les petits organes ainsi que les plus difficiles.

Abstract

Deep Learning has recently shown impressive results in computer vision especially with the performances reached by convolutional neural networks (ConvNets). Those methods have redefined the state of the art in many applications such as medical image segmentation. In this thesis we address the problem of abdominal organ segmentation with deep learning models.

More precisely, we first tackle the issue of training deep ConvNets on partially labeled datasets. Professionals often focus on a specific anatomical region leading to heterogeneous datasets with partially labeled images. However, training a segmentation model directly on such data leads to very poor results due to the presence of wrong labels for the missing organs. In our method, we propose a training scheme that leverages all the labels without being affected by wrong labels. Moreover, we propose an iterative scheme for progressively relabeling the missing organs in the training set in order to further improve the segmentation model.

Secondly, we aim at using spatial prior about the position of the organs to improve the detection of structures and reduce outliers in the segmentation. It comes from the fact that ConvNets, which have been proposed for image classification, do not capture absolute spatial information. However, medical images are very structured and there are some conventions about the expected position of organs. In this work we propose a 3D spatial prior that captures the spatial position of organs and then explicitly biases the model through a prior-driven activation function. In addition to improving the segmentation of difficult organs, we show that using our spatial prior in a pseudo-labeling scheme preserves high performances even with few labeled images by mitigating the introduction of false positives.

Finally, we focus on Transformers models to model long range dependencies between anatomical structures in a classic segmentation model used for organ segmentation. Traditional ConvNets do not capture such interactions because of the receptive field which is often much smaller than the input image. Using dense attention introduced in the Transformer model however, allows to connect every pixel with each other and thus to model complex interactions on different parts of the input image. We show that it improves the quality of the segmentation on various datasets for every organ with a more interesting gain for difficult and complex organs.

Contents

List of Tables

List of Figures

Résumé de la Thèse

Segmentation sémantique d'images médicales 3D par deep learning

L'intelligence Artificielle (IA) est un domaine qui suscite depuis longtemps un grand intérêt. Le but étant de donner aux ordinateurs la capacité de réaliser des tâches de perception telles que comprendre le contenu d'une image. L'IA a fait de grandes avancées dans de nombreux domaines comme la vision par ordinateur avec la classification d'images, la détection d'objets ou la segmentation sémantique, mais aussi en traitement automatique des langues et dans les jeux-vidéos et la robotique grâce aux progrès de l'apprentissage par renforcement.

La vision par ordinateur est un des domaines qui a montré des avancées impressionnantes principalement grâce à l'évolution du matériel informatique et de la puissance de calcul. Ce domaine regroupe trois grands types de problèmes : la classification qui a pour but d'assigner une étiquette à une image ; la détection où l'on cherche à localiser les objets dans les images en y associant des boîtes englobantes ; et finalement la segmentation qui a pour but de prédire une carte dense avec une classification pour chacun des pixels. Cette dernière tâche est la plus complète et difficile car elle regroupe les compétences nécessaires pour réaliser les deux autres tâches.

Dans cette thèse nous abordons des problèmes en analyse d'images médicales. Plus précisément nous étudions des méthodes permettant de reconstruire numériquement en 3D et de façon automatique des organes et structures anatomiques en se basant sur des images médicales. Cependant, il n'est pas évident d'appliquer directement des méthodes de vision par ordinateur sur ces données et certaines particularités doivent être prises en compte. Les images médicales peuvent être de différentes modalités : IRM (Imagerie par Résonance Magnétique), scanner, échographie. Pour chaque modalité, de nombreux équipements différents sont utilisés et il n'y a pas de protocole standard pour l'acquisition ce qui rend chaque image unique. De plus il existe une quantité très importante de tâches si l'on considère l'opération visée (*e.g.* recalage, segmentation), les différentes modalités (*e.g.* IRM, scanner) et les différentes applications (*e.g.* diagnostic, détection de tumeurs, segmentation d'une partie précise du corps) ce qui implique que chaque problème nécessite une base de données suffisamment conséquente alors que ces même

données sont coûteuses et difficiles à obtenir.

Le principal problème abordé dans cette thèse est la segmentation d'organes de l'abdomen à partir d'images de scanner. Cette tâche consiste pour un volume donné à fournir un volume de même taille où pour chaque valeur est associé l'information du tissu (*e.g.* foie, pancréas, reins, etc). L'apprentissage profond a déjà été largement adopté pour cette tâche et des architectures adaptées ont été proposées à l'instar de U-Net [2]. Cependant, de nombreux challenges sont encore étudiés et il s'agit d'un des domaines les plus prometteurs pour l'IA.

Cette thèse est une CIFRE (Convention Industrielle de Formation par la REcherche) entre le Conservatoire National des Arts et Métiers et l'entreprise Visible Patient. Cette dernière développe un service permettant aux professionnels de santé de faire réaliser une modélisation 3D des structures anatomiques internes de patients à partir d'images médicales. Les modélisations sont aujourd'hui réalisées manuellement par des radiologistes professionnels. Ils emploient des outils semi-automatiques mais la plupart des cas nécessitent un travail précis au niveau pixel afin de fournir le meilleur résultat possible au client. Cette thèse s'inscrit dans la volonté de l'entreprise de proposer aux radiologistes des outils automatiques plus robustes qui permettront, à terme, de réduire le temps de traitement d'un cas et ainsi de pouvoir en traiter davantage et se concentrer sur les éléments les plus difficiles et critiques comme les tumeurs.

Dans un premier projet avant le début de la thèse, nous avons eu l'occasion de tester l'efficacité des modèles d'apprentissage profond. Pour cela, nous avons utilisé des méthodes d'entraînement standard qui ont mis en évidence un certain nombre de problèmes induits par la nature médicale des données.

Dans la première partie nous abordons le problème de la disponibilité de données complètement étiquetées. En effet, de nombreuses images sont disponibles mais le processus d'annotation est très coûteux ce qui a pour conséquence de générer des étiquettes partielles qui ne correspondent qu'au besoin pour lequel elles ont été réalisées. Nous proposons donc une méthode qui permet d'apprendre sur des données hétérogènes dans leurs étiquettes en se concentrant sur toutes celles disponibles. Puis, nous proposons un schéma itératif permettant de ré-étiqueter les étiquettes manquantes et ainsi améliorer significativement les performances de modèles disposant initialement de peu de données.

Dans un second temps, nous abordons un problème inhérent aux réseaux convolutifs qui est leur incapacité à utiliser l'information de position spatiale dans l'image. Pourtant, les images médicales sont très structurées et utiliser les connaissances *a priori* de position des organes pourrait permettre d'améliorer les performances. Nous proposons donc de modéliser cet *a priori* sous forme d'une carte de probabilités de présence d'un organe que nous utilisons pour biaiser la prédiction finale en l'intégrant de façon explicite dans la fonction d'activation finale.

La troisième partie aborde un autre problème induit par l'utilisation des convolutions qui est le manque d'information contextuelle pour prédire un pixel donné. Celui-ci est déterminé par le champ réceptif qui est souvent limité en particulier avec les réseaux de segmentation. Nous proposons donc l'utilisation de mécanismes d'attention venant des modèles Transformers qui ont la capacité de modéliser des interactions à long terme entre les caractéristiques apprises.

Entraîner un réseau de neurones entièrement convolutif (FCN) profond sur des données partiellement étiquetées pour la segmentation d'images médicales

L'un des principaux problèmes rencontrés afin d'entraîner des réseaux de neurones profonds pour la segmentation d'images médicales est certainement l'accès à suffisamment de données étiquetées. En effet, de nombreux problèmes en segmentation existent et chacun nécessite une base de données spécifique. De plus, le processus d'annotation doit être réalisé par des professionnels et prend un temps considérable rendant l'opération très coûteuse. Par conséquent, beaucoup de données sont disponibles, mais les étiquettes qui y sont associées sont hétérogènes dépendamment de la structure étudiée et du problème rencontré par le patient.

Dans cette partie de la thèse, nous étudions comment nous pouvons entraîner un réseau de neurones profonds sur des données partiellement étiquetées. Pour cela, nous proposons la méthode INERRANT qui dans un premier temps permet de se concentrer sur toutes les données étiquetées et d'ignorer les étiquettes ambiguës, *i.e.* les endroits où l'étiquette « fond » a été donnée par défaut à l'emplacement d'un organe que l'on souhaite segmenter. Dans un second temps, nous proposons un schéma d'entraînement itératif basé sur les idées venant du Curriculum Learning [12] ou Self -Paced Learning [13] où l'on va progressivement ré-étiqueter les organes manquants dans l'ensemble d'entraînement.

Premièrement, nous cherchons à nous entraîner sur toutes les étiquettes disponibles. Nous devons commencer par déterminer quels sont les organes manquants. Pour cela, nous partons du principe que tous les organes que nous souhaitons segmenter sont visibles dans les images d'entraînement, mais que l'étiquette qui y a été associée est erronée et qu'il s'agit de la classe par défaut « fond ». A partir du moment où un seul organe est manquant, toutes les étiquettes du fond sont alors considérées comme ambiguës ce qui empêche d'utiliser une fonction de coût multiclasses. Nous avons ainsi choisi de transformer notre problème de segmentation multi-classes à $(K+1)$ classes en K problèmes de segmentation binaire. De cette façon nous pouvons individuellement contrôler quel classifieur peut être appris pour quel exemple en se basant sur les organes disponibles. Cependant, pour garder le caractère exclusif apporté par les fonctions comme softmax nous ajoutons à la fin cette contrainte manuellement en choisissant l'étiquette finale en prenant uniquement la classe ayant obtenue le score maximal parmi toutes les autres.

Dans la seconde partie nous proposons un schéma d'apprentissage itératif basé sur le Curriculum Learning [12]. L'idée étant de s'entraîner d'abord sur des données faciles, ici uniquement les données étiquetées, puis de progressivement ajouter des données difficiles, des pseudo-labels qui sont des étiquettes produites par le modèle précédemment entraîné et qui sont ajoutés à la base d'entraînement. De cette façon, nous re-étiquetons progressivement les organes manquants. A chaque étape, nous sélectionnons des pseudo-labels qui seront ajoutés à l'ensemble d'entraînement. Pour cela, nous devons utiliser une mesure de confiance permettant de choisir les meilleurs candidats et d'éviter les erreurs. Nous testons deux solutions : la première est d'utiliser la probabilité de la classe prédite (MCP) ; la seconde consiste à entraîner un réseau dédié qui cherche à prédire la probabilité de la vraie classe. La seconde méthode est plus robuste car elle donne des probabilités en général plus faibles sur les erreurs et évite la sur-confiance que l'on peut observer avec la première. Nous montrons ensuite expérimentalement la supériorité de la seconde solution.

Pour évaluer notre méthode nous avons mené des expérimentations sur trois bases de données : LiTS pour la segmentation du foie, TCIA pour la segmentation du pancréas et IMO qui est une base interne pour la segmentation multi-organes de l'abdomen. La première étape a été de simuler avec ces bases des étiquettes manquantes afin de pouvoir ensuite évaluer la qualité du ré-étiquetage. Pour cela, nous avons retiré au hasard des organes pour garder une proportion fixée : 70%, 50%, 30%, 10%. Concernant le modèle de segmentation, nous avons choisi d'utiliser un U-Net qui est très répandu en segmentation d'images médicales et donne de très bonnes performances générales. Pour montrer que les étiquettes partielles peuvent réellement impacter très négativement l'entraînement du modèle nous avons d'abord entraîné sur les données partiellement étiquetées telles quelles. Sans surprise, les performances sont très mauvaises même avec des proportions élevées d'étiquettes. En utilisant INERRANT nous montrons que nous arrivons à garder de très bonnes performances même avec très peu de données étiquetées. De plus, le schéma itératif permet de booster encore plus les performances, par exemple pour la segmentation du pancréas où le ré-étiquetage fait gagner 10pts quand on dispose de 10% d'étiquettes. Concernant la base de données multi-organes, les gains les plus importants sont systématiquement observés sur les organes les plus difficiles en particulier avec l'étape de ré-étiquetage qui permet d'améliorer significativement les résultats sur la vésicule biliaire, le pancréas et l'estomac avec respectivement +9pts, +13.2pts et +12pts.

Nous avons également comparé notre méthode avec d'autres approches semi-supervisées. La première utilise un apprentissage adversaire, la seconde une fonction de coût de consistance (mean-teacher) et la dernière utilise également des pseudo-labels. Nous avons observé qu'avec les mêmes paramètres, INERRANT donne systématiquement de meilleurs résultats pour toutes

les proportions d'organes manquants.

Ensuite, nous proposons une analyse plus poussée de l'étape de ré-étiquetage. D'abord en évaluant les performances des différentes options pour mesurer la confiance puis en regardant plus en détail l'étape itérative basée sur le Curriculum Learning. Pour la première partie nous avons comparé l'utilisation de MCP avec un réseau de confiance appris. Pour cela, nous avons mesuré la capacité des méthodes à trier les prédictions durant le ré-étiquetage en regardant trois métriques : l'air sous la courbe ROC (AUC) et les Average Precision (AP) des succès et des erreurs. En se concentrant sur l'AUC ce qui donne une mesure globale du tri, nous avons observé que le réseau appris est systématiquement meilleur. Par exemple, à 10% nous obtenons une amélioration de +1.53pts en AUC allant de 68.68% pour MCP à 70.21% pour la confiance apprise. Globalement, la confiance apprise permet de mieux trier les bonnes prédictions donc de sélectionner de bonnes étiquettes et de mieux détecter les erreurs et ainsi diminuer l'ajout de faux positifs. Concernant l'apprentissage itératif, nous détaillons les performances à chaque étape en fixant un maximum de trois étapes de re-étiquetage (4 modèles entraînés). Nous avons observé que la deuxième étape donnait en général les meilleurs résultats car trop d'erreurs sont ensuite ajoutées.

Dans une dernière expérimentation nous avons combiné deux datasets ayant des étiquettes différentes. En particulier nous sommes partis de la base multi-organes interne et avons ajouté des exemples venant de la base TCIA avec uniquement des étiquettes du pancréas. Nous montrons alors une augmentation très importante des performances en combinant les deux bases avec INERRANT, +9.5pts pour la vésicule biliaire, +10.5pts pour l'estomac et +25.6pts pour le pancréas.

Dans cette partie nous proposons INERRANT qui est une méthode permettant d'apprendre sur des données partiellement étiquetées puis grâce à un schéma itératif de ré-étiquetage des organes manquants qui permet d'améliorer significativement les résultats. Les résultats expérimentaux montrent un gain systématique de la méthode même comparée à d'autres méthodes semi-supervisées de l'état-de-l'art pour toutes les proportions d'organes manquants.

Ajouter de l'information spatiale *a priori* dans l'entraînement de FCNs profonds

La segmentation d'organes n'est pas une tâche facile car elle doit précisément étiqueter des objets ayant des formes, des textures et des positions très diverses au niveau pixel. De plus, les organes que nous souhaitons segmenter sont des tissus mous ayant des valeurs dans les scanners très similaires. De nombreux cas restent très délicats si l'on se base uniquement sur l'information de contexte local. Le faible contraste entre les tissus ainsi que les ambiguïtés visuelles en particulier au niveau des frontières entre les organes rendent la tâche de segmen-

tation automatique particulièrement compliquée. Cependant, les images médicales sont très structurées et nous avons de forts *a priori* sur les différentes structures anatomiques ; plus précisément sur la position absolue des organes. Ces connaissances sont d'ailleurs largement utilisées par les professionnels qui ne se basent pas uniquement sur l'apparence visuelle locale qui est souvent insuffisante. Utiliser des connaissances externes pourrait effectivement permettre d'aider à entraîner des modèles plus robustes. Malheureusement, les modèles actuels qui sont les plus utilisés en segmentation d'images médicales sont des réseaux de neurones convolutifs qui sont par construction incapables d'apprendre des caractéristiques sur la position des objets.

Dans ce chapitre, nous étudions comment ajouter de l'information spatiale *a priori* dans des réseaux de neurones convolutifs afin d'améliorer la qualité de la segmentation. Nous proposons une méthode appelée STIPPLE qui a pour but de construire un *a priori* spatial 3D sous forme de carte de probabilités de la présence d'un organe à une position donnée. Cet *a priori* est ajouté explicitement à la fin d'un modèle de type réseau de neurones convolutifs grâce à une fonction de prédiction « prior-driven ». Nous montrons également que l'utilisation de cet *a priori* peut améliorer la sélection des pseudo-labels dans le contexte de la méthode présentée précédemment.

Notre méthode STIPPLE se base sur deux hypothèses principales : (1) le volume 3D a été réalisé suivant la direction axiale, le patient étant allongé sur le dos ; (2) il y a de fortes variations dans la position de l'organe en z mais les variations en (x, y) restent faibles. Pour cette dernière, nous l'avons largement observée sur de nombreux datasets, ces variations sont dues à l'acquisition qui commence et s'arrête à des positions variables suivant les patients.

Pour la construction de l'*a priori*, nous utilisons les étiquettes de la base de données d'entraînement afin de calculer une carte moyenne. Pour cela, nous découpons dans chaque volume une zone de taille fixe ($W_p \times H_p \times \Delta_z$) centrée sur l'organe que nous accumulons pour créer une représentation moyenne de l'organe. Les valeurs, W_p, H_p et Δ_z sont choisies de telle sorte que tous les organes de la base peuvent y entrer. Cette carte donne à la fois une information de position mais également sur la forme de l'organe. Cependant, du fait de la forte variabilité de la position des organes en z , nous discrétisons cet axe en un nombre de boîtes fixé B. Nous obtenons ainsi une carte de probabilité de taille ($W_p \times H_p \times B$).

Cet *a priori* est utilisé de façon explicite à la fin du modèle de segmentation au niveau de la fonction d'activation. Le but principal est d'influencer directement la prédiction et de s'assurer de sa participation dans le résultat final ainsi que pendant l'entraînement du modèle. Pour cela, nous utilisons une fonction d'activation qui est une généralisation de la fonction softmax qui est décrite dans l'équation suivante : Equation 4.2. Toutefois, il est nécessaire de positionner

notre *a priori* dans l'image car celui-ci doit être centré sur l'organe. Pour ce faire nous utilisons durant l'entraînement les positions venant de la réalité terrain, mais pour les images de test, nous utilisons la prédiction venant d'un premier modèle de segmentation peu précis permettant ainsi de bien localiser l'organe. Une étape supplémentaire est nécessaire pour ajuster la position de notre *a priori* afin qu'il soit utilisé au mieux.

Nous avons dans cette partie utilisé notre *a priori* dans deux contextes, d'abord avec une supervision complète puis dans un contexte avec peu de données étiquetées tel que défini dans la partie précédente. Cela permet de voir comment notre *a priori* améliore les performances quand peu de données sont disponibles. De plus, nous avons utilisé le même schéma d'apprentissage itératif que dans la partie précédente afin de regarder comment les deux éléments peuvent collaborer.

Concernant les résultats expérimentaux, nous avons testé notre méthode sur la tâche complexe de segmentation du pancréas avec la base TCIA. Nous testons comme dans la partie précédente avec des proportions d'étiquettes allant de 70% à 10% et évaluons également le contexte entièrement supervisé, *i.e.* proportion de 100%. Les gains en utilisant l'*a priori* seul par rapport à la baseline sont les suivants : +1.41pts à 100%, +2.90pts à 70%, +1.32pts à 50%, +1.50pts à 30%, et finalement +2.84pts à 10%. En ajoutant le schéma itératif de ré-étiquetage, ce gain devient encore plus important avec des gains de +4.0pts à 70%, +3.7pts à 50%, +5.9pts à 30% et +9.9pts à 10%.

La segmentation du pancréas est particulièrement difficile car cet organe a une forme complexe et ses bordures sont très souvent ambiguës. Les résultats expérimentaux montrent qu'utiliser notre *a priori* permet d'obtenir des gains dans deux situations, tout d'abord en l'utilisant tel quel mais également en l'utilisant dans un schéma d'apprentissage itératif. En effet, cet *a priori* renforce les probabilités dans la région la plus probable et permet de récupérer des prédictions qui auraient été manquées. Ensuite, pour le ré-étiquetage, il réduit l'ajout de faux positifs en nettoyant des erreurs aberrantes pour mieux concentrer la sélection des pseudo-labels dans la bonne région.

Nous avons comme dans la partie précédente évalué notre méthode contre d'autres méthodes semi-supervisées état-de-l'art mais également contre une méthode utilisant un mécanisme d'attention, Attention U-Net [11]. Pour ce dernier, en utilisant le même réseau de base, nous montrons que notre *a priori* permet d'avoir de meilleurs résultats avec un gain de +1.1pts dans le contexte entièrement supervisé. Cela peut s'expliquer par le fait que notre *a priori* exploite de l'information en 3D contrairement au modèle attentionnel qui est 2D. De plus, nous montrons que notre méthode est plus robuste quand le nombre de données annotées diminue.

Pour finir, nous proposons une analyse plus fine de deux éléments de la méthode qui sont

la discrétisation en B boîtes et le positionnement de l'*a priori*. Pour le premier, nous avons testé différentes valeurs de B avec différentes valeurs d'étiquettes manquantes. Nous montrons que la valeur standard de $B = 5$ est bonne en moyenne, mais aussi que quand B augmente les performances diminuent ce qui montre l'intérêt d'utiliser la discrétisation en z. Concernant le positionnement de l'*a priori*, cet élément est très important pour obtenir un résultat optimal. Nous montrons expérimentalement que l'étape qui nous permet de la raffiner est importante et qu'utiliser simplement la prédiction grossière n'est pas suffisant.

Dans cette partie nous avons proposé STIPPLE qui intègre un *a priori* sur la position absolue des organes pour la tâche de segmentation. Notre méthode donne de très bons résultats en particulier dans un contexte où la quantité d'étiquettes est limitée.

Transformers et self-attention pour la modélisation d'interactions à long terme

Les réseaux de neurones complètement convolutifs (FCNs) sont les modèles les plus utilisés en analyse d'images médicales. Cependant, ils ont un problème majeur qui est leur champ réceptif limité en particulier pour les modèles de segmentation. Ce champ réceptif permet de mesurer la vue d'un pixel de sortie sur l'ensemble des pixels d'entrées. Il est principalement défini par l'architecture. Cependant, en regardant le champ réceptif effectif, on s'aperçoit que la participation des pixels voisins décrit une gaussienne et donc que plus un pixel est éloigné de celui considéré, moins il va participer à la décision finale. Par conséquent, le champ réceptif est souvent limité et ne va pas permettre de capturer suffisamment d'information contextuelle qui est pourtant essentielle pour la tâche finale.

Dans cette partie, nous avons voulu utiliser les facultés des modèles Transformer à modéliser des interactions à long terme pour ainsi exploiter un maximum d'information contextuelle globale pour la tâche de segmentation d'organes. Pour cela, nous proposons U-Transformer, un modèle qui utilise le mécanisme d'attention proposée dans les Transformers dans un modèle de segmentation entièrement convolutif de type U-Net [2]. Les Transformers sont des modèles qui ont été initialement proposés en traitement du langage et permettent de modéliser des interactions à long terme entre les différents éléments d'une phrase. Ils se basent sur l'utilisation de modules d'attention qui connectent chaque élément d'entrée avec tous les autres contrairement aux modèles attentionnels précédemment proposés dans la littérature qui sont calculés localement au niveau d'un élément ou seulement quelques voisins.

Dans U-Transformer nous proposons d'utiliser deux modules : premièrement un module de self-attention ayant pour but de modéliser des interactions à long terme entre toutes les parties de l'image d'entrée qui est situé dans le *bottleneck* ; puis un module de cross-attention situé dans les différents blocs du décodeur qui permet de filtrer les features non-sémantiques venant

des skip-connections avec les features hautement sémantiques venant des précédents blocs dans le décodeur. L'architecture est présentée en Figure 5.4 en utilisant un U-Net mais les modules proposés peuvent être utilisés dans tous les FCN avec une architecture similaire à U-Net.

Le module de self-attention, MHSA, utilise le même calcul d'attention que celui des Transformers. L'entrée est envoyée dans trois matrices, Q, K et V. Les deux premières, Q et K sont utilisées pour calculer la matrice d'attention A. Cette dernière a une taille qui est le carré de la taille de l'entrée. Chaque ligne étant un vecteur donnant pour une feature la relation avec toutes les autres features. Cette matrice est utilisée pour transformer V et le résultat est renvoyé pour obtenir une carte de même taille que celle d'entrée.

Concernant le module de cross-attention, il reprend le même mécanisme que celui de la self-attention. Cependant, il joue un rôle de filtre pour la skip-connection. Pour cela, les matrices Q et K sont connectées au bloc du décodeur précédent et V est la skip-connection. Le résultat de la partie attentionnelle est ensuite transformé pour obtenir une carte de même taille que la skip-connection et sera utilisé pour la filtrer grâce à un produit terme-à-terme.

Expérimentalement, nous avons principalement utilisé deux bases de données : une pour la segmentation du pancréas et la seconde pour la segmentation de plusieurs organes de l'abdomen. Nous nous sommes également comparés à Attention U-Net [11] qui est un modèle attentionnel classique utilisant une attention locale au niveau pixel. Concernant la première base sur le pancréas, le gain observé est de +2.4pts au total. La segmentation de cet organe est particulièrement difficile et nous avons observé que U-Transformer réussissait mieux à détecter certaines parties que les autres modèles. Pour la base multi-organes, nous avons observé un gain moyen de 1.3pts. Cependant, quand on regarde le détail par organe on s'aperçoit que les gains les plus importants sont sur les plus petits et complexes : le pancréas +3.4pts, la vésicule biliaire +1.3pts et l'estomac +2.2pts. Cependant, même le foie avec un score de 96.40% de DSC voit une augmentation avec U-Transformer pour atteindre 97.03%. Cette tendance est d'ailleurs validée sur une autre base multi-organes dans des expérimentations récentes.

Nous proposons également une analyse ablative pour évaluer l'intérêt de chaque partie de notre modèle. Pour cela, nous avons testé l'impact de l'encodage positionnel ainsi que l'utilisation de la cross-attention à tous les niveaux. En considérant un modèle entraîné uniquement avec la self-attention, l'encodage positionnel permet de gagner +0.7pts sur TCIA et +0.6pts pour le multi-organes. Ensuite, en utilisant uniquement la cross-attention, l'encodage positionnel permet de gagner +1.7pts sur le pancréas et +0.6pts en multi-organes. Cela valide le fait que cet élément est particulièrement important pour faire fonctionner les Transformers. Ensuite, nous avons évalué un modèle utilisant uniquement un seul bloc de cross-attention et l'équivalent à chaque niveau ayant pour effet d'augmenter les performances de +1.8pts pour le

pancréas et +0.6pts pour le multi-organes.

Finalement nous avons voulu étudier comment nos modules peuvent s'adapter à d'autres architectures. Pour cela, nous les avons intégrés dans un nnU-Net [14] qui est aujourd'hui une méthode très puissante pour la segmentation et qui utilise une structure 3D. Grâce à notre méthode, nous avons pu obtenir un gain de +1pt qui est un gain conséquent connaissant le score obtenu par le modèle de base.

Dans cette partie nous avons présenté U-Transformer qui a pour but d'utiliser des mécanismes d'attention venant des Transformers dans des modèles de type U-net. Nous montrons expérimentalement des gains systématiques dans tous nos résultats, ce qui valide l'intérêt de modéliser des interactions globales en utilisant les Transformers.

Dans cette thèse, nous avons étudié trois problèmes majeurs en segmentation d'images médicales. Tout d'abord l'entraînement de réseaux de neurones profonds sur des données partiellement étiquetées. Ensuite, nous avons étudié l'utilisation de connaissances *a priori* sur la position absolue des organes en proposant un *a priori* 3D qui vient biaiser directement le modèle de façon explicite. Finalement, nous avons abordé la question du champ réceptif limité des réseaux complètement convolutifs et avons proposé d'utiliser les Transformers qui ont la capacité de modéliser des interactions globales dans l'image.

En perspective de futurs travaux, il serait intéressant de chercher à modéliser et utiliser des connaissances *a priori* plus complexes comme des relations explicites entre des organes. Cependant, le manque d'explicabilité dans les modèles de réseaux de neurones profonds s'avère être un frein, de nouvelles avancées dans ce domaine permettraient d'utiliser des *a priori* différents de façon plus efficace. Ensuite, il serait intéressant de regarder plus en détail l'entraînement sur des données hétérogènes. Pour cela, un domaine prometteur est l'adaptation de domaine et plus précisément la généralisation de domaine qui permettrait d'apprendre sur un type de données en l'appliquant directement sur d'autres types de données. Cela pourrait s'intégrer dans un contexte multi-modale où l'on cherche à entraîner des modèles fonctionnant sur plusieurs modalités, par exemple sur des scanner et IRM.

Chapter 1

Introduction

Contents

1.1 Context

Artificial Intelligence (AI) has been a field of great interest for the last decades, aiming at developing new techniques that give computers the ability of performing perception tasks like image understanding or speech recognition. As shown in Figure 1.1, AI has made strong advances in Computer Vision (CV) with image classification, semantic segmentation or object detection. In Natural Language Processing (NLP) which includes text understanding or text generation. Even in games and robotics with the evolution of Reinforcement Learning (RL), *e.g.* in 2017, the AlphaGo program has defeated the number one ranked player of Go board game using AI models. All those advances open the path to new applications which go from virtual assistants to robot-assisted surgery and autonomous driving cars.

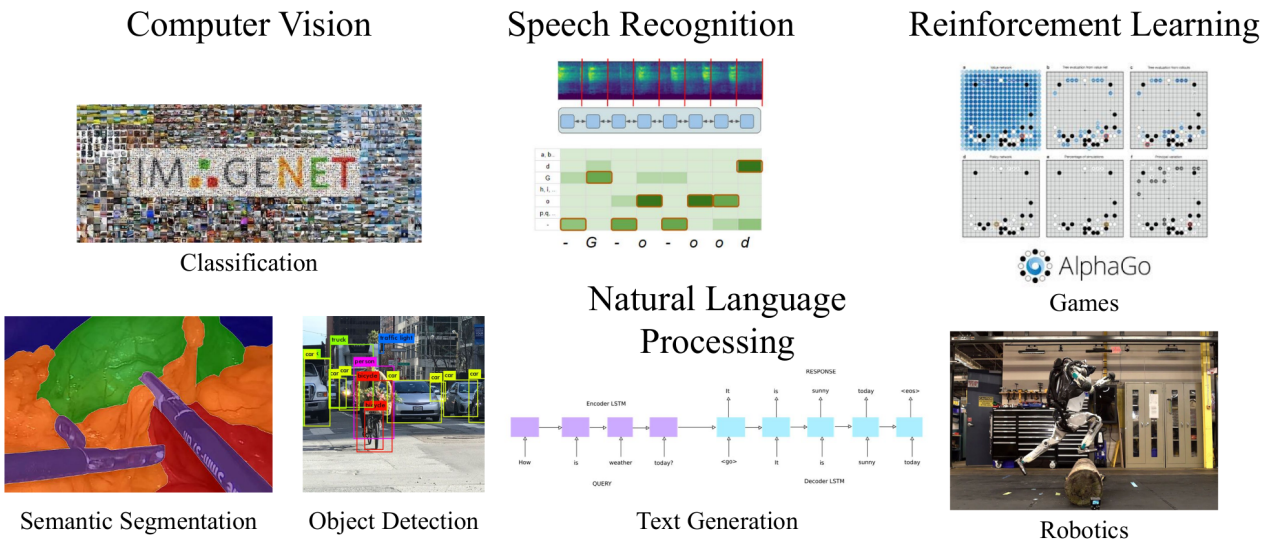


Figure 1.1: Applications where AI has showing impressive performances in the last decades.

Computer Vision (CV) is one of the fields that showed the most impressive advances with AI for automatically processing images and videos. This domain has a key role in many applications to keep up with the exponential growth of available data. The recent advances were driven by the evolution of computer hardware and software which are able to estimate a growing number of parameters in trainable models. Common tasks addressed by CV are: **classification** which aims at giving a label to an image, *e.g.* the object represented; **detection** where the objects in the image should be spatially localized with a bounding box and also classified; **segmentation** which goals is to predict a dense map with a pixel-level classification. The latter is the most complete but difficult task as it combines the skills of understanding the image, localizing the objects, delineating them and finally classifying them. CV has greatly benefited

from the latest advances in supervised algorithms from Machine Learning (ML). Since 2012, Deep Learning (DL) has shown impressive results which propelled it as the best methods currently available in ML. DL relies on the use of Artificial Neural Networks (ANN) which consist in a succession of layers with basic units which then form a deep network with a large amount of parameters. Beyond those fully connected architectures, specific design choices have been crucial to the success of DL. For example, the introduction of the convolution operation was extremely important for the processing of low-level signals (*e.g.* images, audio and time-series in general), since it induces equivariance to small transformations. In 2012, Convolution Neural Networks (ConvNets) started showing their impressive performances with the success of AlexNet [15] on the ImageNet Challenge. The years after, all the classic methods were outperformed by ConvNets which continued to progress and beat records on the ImageNet dataset which progressively became the DL's advances showcase with deeper and deeper models.

In this thesis, we address challenging problems in medical image analysis. More precisely, we study tools for modeling internal organs in 3D based on medical images. Today, a lot of solutions exist for natural image segmentation. However, adapting those algorithms is not straightforward. It is necessary to consider the particularities of the data to develop adapted solutions. Moreover, medical images are acquired with various modalities, *e.g.* CT, MRI, US, etc, and there are no standard acquisition protocol due to the large variations in equipment and scanning settings. Consequently, there is a certain domain shift between the data making the training of models difficult. Also, it is worth noting that a very large number of different tasks exist in medical image analysis. When taking into account the different goals (*e.g.* registration, segmentation, etc), the different modalities and the different applications (*e.g.* disease diagnosis, tumor segmentation, cranial vault registration, etc), there is a huge panel of complex tasks to be addressed. As a consequence, the lack of data is a major concern knowing that every task needs an appropriate dataset. Moreover, privacy rules regarding patient data and the institutions make it difficult to centralize complete datasets. In all the available modalities, Computerized Tomography (CT) scans are the most important imaging tool and the most widely used. For example, over 70 million CT scans were performed in the United States in 2007 [16]. Additionally, there are more conventions about the acquisition process than with other modalities. CT scans consist in sending X-ray beams from different angles around the body to create slices which are then assembled to create a volume. The final volume shows the internal structures and organs of the patient in three dimensions which is much more informative than plain X-ray images.

The main problem addressed in this thesis is abdominal organ segmentation in CT scans which consists in creating a dense 3D map of the internal organs from a given CT scan image,

Figure 1.2. Medical image analysis is without a doubt one of the most promising applications of AI and could tackle a lot of issues currently faced. For example, detecting small tumors more precociously, designing sophisticated tools for surgery, etc.

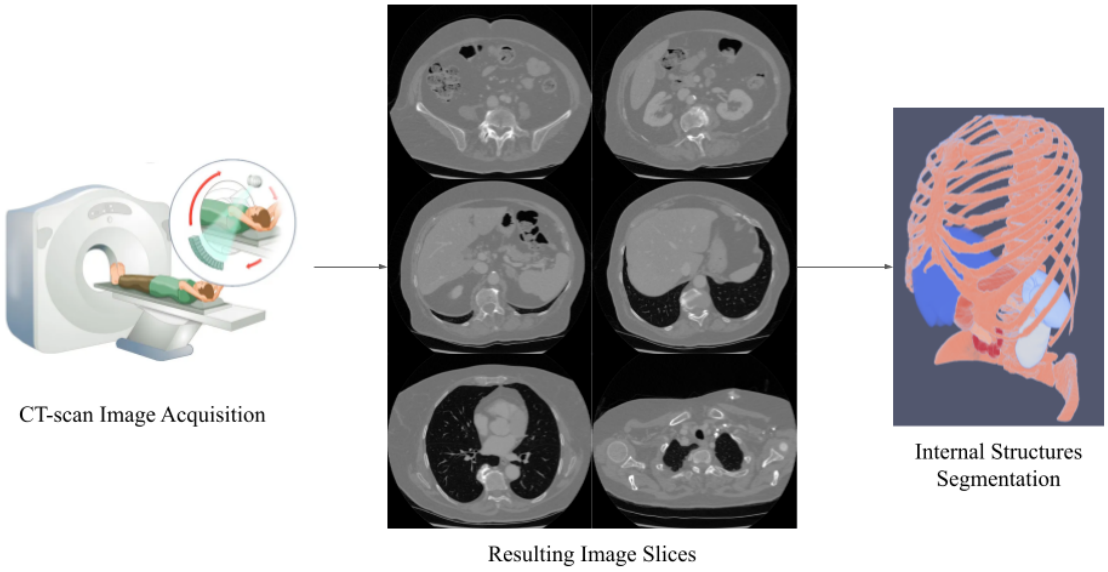


Figure 1.2: The main application addressed in this thesis is medical image segmentation which consists in modeling the internal structures of a patient given its medical images, *e.g.* CT-scans.

Industrial Application at Visible Patient (VP) This thesis is a CIFRE (Convention Industrielle de Formation par la REcherche) program between le Conservatoire National des Arts et Métiers (le CNAM) in Paris and the company Visible Patient (VP) in Strasbourg. VP provides a patient modeling service. Every patient is unique and has its own anatomical particularities. When it comes to planning intervention, images are crucial to understand the structures and to operate as precisely as possible. The introduction of scanners and MRI has been revolutionary in the surgery field by giving a precious tool for surgeons to anticipate unexpected findings. It has led to an important decrease in post-operative complications by allowing more precise operation and less invasive surgery. Thus, technology has continually improved the quality, accuracy and speed of image acquisition but a main constraint remains in the visualization of those images. The 3D volumes can only be seen as successive 2D slices which make it difficult to interpret small structures. VP addresses this problem by proposing a solution to visualize the patient in three dimensions, including the modelization of their organs. The virtual model of a patient gives the internal anatomy of the organs, vessels, lesions, etc from a medical image (*e.g.* CT scan) sent by a physician. The result can be visualized with a software (Visible Patient Planning) which also provides tools to interact with the structures.

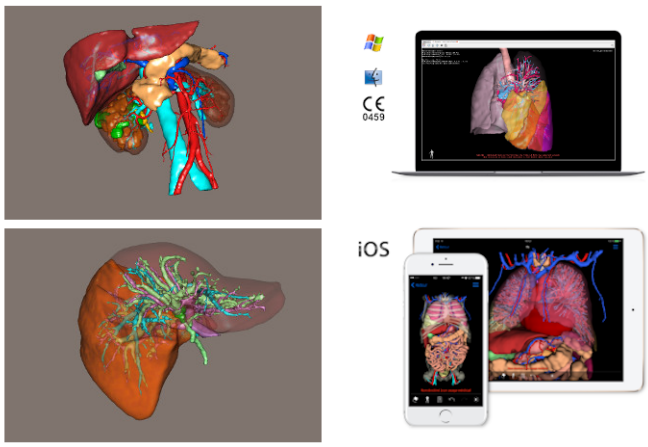


Figure 1.3: Visible Patient offers a 3D modeling service for medical images.

The organ modelization is performed by professional radiologists with semi-automated tools. The segmentation task is still very manual because the current algorithms are very low-level. Consequently, most of the segmentation is completely realized manually. The final segmentation is carefully verified by a second expert before being sent back to the client. This work is difficult and very time-consuming. Thus, VP is interested in introducing automated tools in the current solution. This way, radiologists could focus more on the quality and the critical regions (like the lesions). Moreover, the effectiveness of the solution will be boosted by drastically reducing the processing time for each patient.

1.2 Motivations

Lately, DL has deeply impacted the medical image analysis field especially in the task of organ segmentation. Specifically designed architecture has been proposed such as the well-known U-Net [2] in 2015. In a nine months project with VP before the beginning of this PhD, we had the opportunity to validate the effectiveness of DL models on internal data. We observed that standard architectures were already able to give great results for the automatic segmentation of various organs and tissues. We also successfully used standard training techniques such as **transfer learning** and **fine-tuning** which consists in first training the model on a large dataset, generally ImageNet, and then slowly tuning the parameters on the target dataset. However, those first results also brought out limitations and the need to address specific problems induced by the medical nature of the images.

Dealing with partially-labeled datasets A major concern with medical image segmentation is the availability of sufficiently large and exhaustive datasets. The labels are by essence sparse and noisy because labeling medical images requires highly-qualified professionals and is very time-consuming thus expensive. Moreover, different tasks require different labels which induce sparsity in the annotations. Finally, there is high inter-patient label inconsistency due to the involvement of different annotators in addition to the different acquisition settings. Most of the existing datasets focus on specific structures, thus two abdominal datasets could have few organs in common. This phenomenon was observed on a VP's internal dataset where a lot of images are in fact available but with an heterogeneous labeling as shown in Figure 1.4. Having large datasets is great but the partial-labeling is problematic and needs to be addressed by adapting the model training. The main challenge is to leverage the available labels without negatively impacting the training with wrong labels (*i.e.* pixels which are labeled as background while they are in fact an organ of interest).

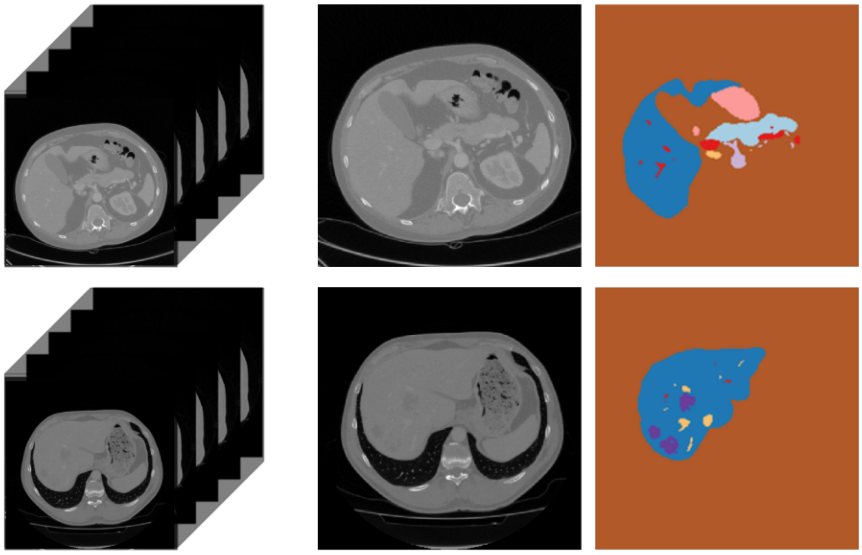


Figure 1.4: Labels are heterogeneous and depend on both the organ of interest and the granularity of the segmentation.

Visual ambiguities in medical images Organ segmentation is not a trivial task as it faces various shapes, textures and localization of the different structures. In fact, soft tissues have very similar values in CT-scans leading to low contrast between the abdominal organs. Moreover the borders of the objects are often ambiguous and not easily distinctable even for a human annotator. However, there is a strong knowledge about the expected position, shape of each individual organ or even relationships between them. Medical images are very structured which

means that there is some conventions about the organs. For example, the absolute position of an organ is fairly consistent between the patients. In Figure 1.5 we can see how the organs concentrate in a limited region of the image. However, this knowledge is unfortunately not exploited in ConvNet based models traditionally used for organ segmentation which are by construction incapable of learning knowledge about the absolute position of objects. Leveraging the prior knowledge on the spatial position and shape or even more complex information in a ConvNet could guide the training for a better segmentation result.

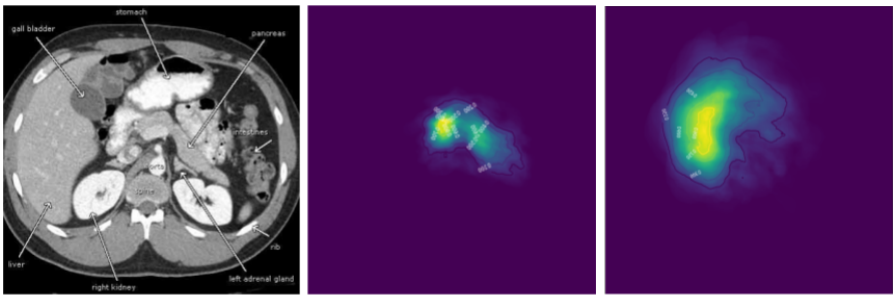


Figure 1.5: Organs have known positions which could be leveraged to bias the model and avoid segmentation errors caused by visual ambiguities.

Limited Receptive Field in ConvNets The view of a FCN on the input image is characterized by The Receptive Field (RF). More precisely, the view of an output unit on the input image is limited to a field of view determined by the network architecture, especially with the number of pooling operations. When considering segmentation networks, this RF is often limited because of the architecture which includes less pooling operations and no fully-connected layers. In practice, the RF is a Gaussian with a more important strength at the position of the considered pixel. For example in Figure 1.6 we have represented the Effective Receptive Field (ERF) of a U-Net and can see that it is limited to a small local region. However, in medical image segmentation, the contextual information is crucial as the local context is often ambiguous and insufficient to correctly decide for complex structures and tissues. Consequently, we need to find a way to leverage larger contextual information for a robust and precise segmentation result.

1.3 Contributions and Outline

To tackle the aforementioned issues, we develop in this thesis three main contributions:

- Chapter 3: Training Deep FCNs with Partial-Labels for Medical Image Segmentation

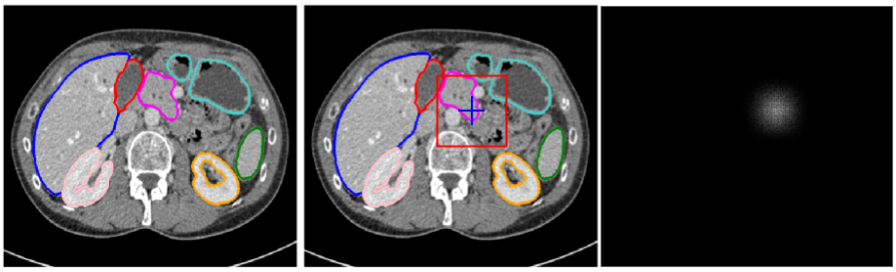


Figure 1.6: The limited Effective Receptive Field of a U-Net in the bottleneck.

Collecting pixel-level labels for medical image segmentation is very expensive and requires a high level of expertise. Thus, clinical experts often focus on specific anatomical structures leading to partially labeled images. In this chapter we detail the work presented in SMILE [1], later extended by INERRANT which addresses the problem of training deep FCNs with partial-labels for medical image segmentation. The first part aims at training a segmentation model on heterogeneous labels by focusing on all the available labels and ignoring ambiguous labels. For that we rely on the fact that in an abdominal image, all the organs should be visible, thus we can deduce which ones are unlabeled and then ignore the pixels corresponding to that region. Moreover, we propose with INERRANT, which further improves on the first part, an iterative pseudo-labeling scheme inspired by curriculum learning that progressively assigns new labels to missing organs based on a confidence measure. We first look at a simple yet effective measure based on the output probabilities and then propose a dedicated network that gives a confidence map and shows its superiority.

- Chapter 4: Incorporating Spatial Knowledge on Organ Positions when Training Deep FCNs

  To address the problem of visual ambiguities and improve deep FCNs for abdominal organ segmentation, we present in this second work a method for incorporating the strong prior about the absolute position of the organs in a DL model. To this end, we build a 3D spatial prior map based on the observed positions in the training set. This spatial prior gives the probability of the organ's presence at a given position in the input image. Then, we directly bias the output prediction with a prior-driven activation function. The spatial prior is explicitly added in the model and is directly interpretable. We experimentally show the relevance of the method for the challenging pancreas organ segmentation task. Moreover, we add this prior in a pseudo-labeling scheme as used in Chapter 3 and show that it could even improve the selection of pseudo-labels.

- Chapter 5: Transformers and Dense Attention for Modeling Long Range Interactions

  The limited RF in ConvNets is a problem when considering tasks such as semantic segmentation where the whole image context is important. It is especially the case for medical images where the images are very structured, *i.e.* the organs have a standard layout. Modeling global interactions between the structures is therefore important for this task. In this part we propose to leverage the strong abilities of Transformers models and their capacity for modeling long-range dependencies. Those models were first used in NLP [17] but they have been successfully applied in CV [9]. Using Transformers for medical image segmentation is particularly relevant as they could model relationships between different structures and thus use the complete image context without being limited by the RF. We propose to integrate the attention mechanisms from Transformers to integrate them into a U-shaped FCN at two levels: in the bottleneck with a self-attention that explicitly models full contextual information and long-range interactions and at each skip connection with a cross-attention that filters non-semantic features from the skip connections with features from the previous decoder block and thus improves spatial recovery.

Chapter 2 gives an overview of general Deep Learning (DL) methods and medical image segmentation. Then, it introduces the state-of-the-art for the three proposed contributions. Eventually, we will conclude this thesis and talk about interesting perspectives for future works in Chapter 6.

1.4 Related Publications

This thesis is based on the material published in the following papers:

SMILE	Olivier Petit, Nicolas Thome, Arnaud Charnoz, Alexandre Hostettler and Luc Soler. **"Handling Missing Annotations for Semantic Segmentation with Deep ConvNets."** *Medical Image Computing and Computer Assisted Intervention (MICCAI), workshop Deep Learning for Medical Imaging (DLMIA), 2018.*	Chapter 3
INERRANT	Olivier Petit, Nicolas Thome and Luc Soler. **"Iterative Confidence Relabeling with Deep ConvNets for Organ Segmentation with Partial Labels"**. *Computerized Medical Imaging and Graphics (CMIG), 2021.*	Chapter 3
	Olivier Petit, Nicolas Thome and Luc Soler. **"Biasing Deep ConvNets for Semantic Segmentation of Medical Images with a Prior-driven Prediction Function"**. *Medical Imaging with Deep Learning (MIDL), extended abstract, 2019.*	Chapter 4
STIPPLE	Olivier Petit, Nicolas Thome and Luc Soler. **"3D Spatial Priors for Semi-Supervised Organ Segmentation with Deep Convolutional Neural Networks"**. *International Journal of Computer Assisted Radiology and Surgery (IJCARS), 2021.*	Chapter 4
U-Transformer	Olivier Petit, Nicolas Thome and Luc Soler. **"U-Net Transformer: Self and Cross Attention for Medical Image Segmentation"**. *Medical Image Computing and Computer Assisted Intervention (MICCAI), workshop Machine Learning in Medical Imaging (MLMI), Oral Presentation, 2021.*	Chapter 5

Chapter 2

State of the Art in Medical Image Segmentation

Contents

Medical image segmentation has been extensively studied in the last decades [18, 19, 20]. It consists in assigning a tissue label to every voxel of the input image. In medical image analysis, there is five main tasks:

- *Reconstruction:* aims at creating a medical image based on the acquired signals (*e.g.* radiations). For instance, it is particularly important for reducing the amount of radiation received by the patient.
- *Enhancement:* aims at improving the visual aspect of images. It is often based on image denoising or super-resolution but recent works focus on style-transfer with for instance modality translation which aims with generative models to generate a synthetic image in one modality based on an image from another modality.
- *Registration:* aims to align different images in the same coordinate space. It was particularly used in label transfer for organ segmentation in atlas and multi-atlas methods.
- *Segmentation:* aims at labeling an image at a pixel level which requires to localize, delineate and classify every object. This task is important for planning intervention or to compare physical changes in response to a treatment. There is a vast literature on this task with a wide range of applications depending on the targeted structures.
- *Computer aided diagnosis:* aims at localizing lesions in images and classifying them as benign or malignant.

In this thesis, we focus on the segmentation of abdominal organs in CT-scans. Thus, the following sections present the literature related to this specific task. Firstly, in Section 2.1, we introduce the CT modality and model-based methods for automatic segmentation. Then, in Section 2.2, we dive into the recent emergence of DL models and how they are used for medical image segmentation. We discuss the general architectures and how to use them with medical images, what are the specific losses and metrics used to evaluate our models and finally Section 2.2.4 presents common datasets used in the different experiments of this thesis in addition to more recent challenges. Finally, in the three sections: Section 2.3, Section 2.4, Section 2.5, we detail the specific literature for each part of the thesis which are developed from Chapter 3 to 5.

2.1 Medical Image Segmentation

In this section, we detail specific characteristics related to the CT-scans which are the main modality used in this thesis. Then, we look into segmentation methods that were the most used

before the DL breakthrough.

2.1.1 CT-scan Image Acquisition and Characteristics

Medical imaging is the process of getting a view on the human body by using specific tools and techniques. It assists diagnosis and can be used to track ongoing treatments. Acquisition techniques, or modalities, include Magnetic Resonance Imaging (MRI), ultrasonography, radiography, etc. Some modalities are **tomographic** which means that the images are acquired by assembling thin 2D slices into a 3D volume. For example, MRI and CT-scans both output tomographic images. When analysing internal organs especially in the abdomen, the most frequently used modalities are MRI and CT-scans as they allow to visualize all the organs in three dimensions and thus gives an insight on the volume and layout of the structures.

In this thesis we use Computerized Tomography (CT) scans which is a tomographic modality using X-rays such as in radiography. It consists of a X-ray beam which spins around the patient. Then, the multiple responses captured from the different angles are interpreted by a computer which reconstructs an image corresponding to a 2D slice of the patient. This procedure is performed multiple times by moving to a small step corresponding to the inter-slice spacing. Finally, the slices are assembled to a final 3D volume representing the studied area.

The data is presented as a multidimensional array where the values represent the relative radiodensity observed at a given position (x,y,z). It uses the Hounsfield scale (in Hounsfield Units, HU) which spreads between +3,071 HU (most attenuating) to -1,024 HU (least attenuating). Dense structures such as bones have high values and air has low values. This scale is calibrated with air and water, respectively -1000 and 0 HU. Then, the values are determined by the formula given in Equation 2.1.

$$HU = 1000 \times \frac{\mu - \mu_{water}}{\mu_{water} - \mu_{air}} \tag{2.1}$$

Each image has the following meta-information in addition to the patient's information: the **origin** and a **direction** vector give together a landmark to read the data in three dimensions. A **voxel-sapcing** which specifies the distance in millimeters between two voxels in each direction. Numerically, a medical image is stored as a multidimensional array with each voxel being a value in HU. The image size in voxels differs from the true size in spatial coordinates as it depends on the voxel-spacing.

Table 2.1 gives some examples of HU value ranges for different tissues and substances. We can see that the major structures targeted in medical image segmentation are in the soft tissues whose values go from +100 to +300. In practice, some organs overlap which makes the segmentation challenging. In order to get a better visualization of the structures, a common

Substance	HU
Air	-1000
Lung	-700 to -600
Fat	-120 to -90
Water	0
Kidney	+20 to +45
Muscle	+35 to +55
Liver	+55 to +65
Bones	+300 to +1900

Table 2.1: Examples of HU value ranges for different substances. The organs of interest are often situated in the soft tissue range roughly between +100 and +300 HU.

practice consists in using a **windowing** which clips the values in a given range. In ML, it is integrated in the image preprocessing step. However, as the values are very close and overlap, there is in fact a very low contrast between some tissues making challenging the automatic and even manual delineation.

2.1.2 Model-Based Segmentation Methods

Automatic segmentation of objects in medical images has always been a subject of great interest and extensively studied in the literature [18, 19, 20]. Initially, the availability of manually segmented images was a critical point. Obtaining such data was very expensive and researchers often had to select a limited number of cases trying to be the most representative and thus transferable as possible.

The first attempts mainly focus on image driven methods were the segmentation is based on the image itself and predefined rules, *e.g.* thresholding [21, 22], region-growing [23, 24] or watershed [25, 26].

Then, model-based models emerged mainly with the growing number of labeled data. It includes deformable models [27, 28, 29, 30, 31] and atlas-based methods [32, 33, 34, 35, 36, 37]. The former aims at modifying a predefined curve to fit to the image based on an energy minimization algorithm. The former are very specific to medical images and use the fact that the organs are located at very similar positions across the patient and use label-transfer to give a label to the target volume based on the annotations from the dataset. In addition, Statistical Shape Models (SSMs) [38, 29, 39, 40] was often use to constraint the models with shape information extracted from labeled images.

Many classifications were proposed [18, 19, 20] for medical images segmentation techniques and an exhaustive description of each method is beyond the scope of this chapter. However,

we detail, more precisely, two very important methods that were the most used before the DL breakthrough: deformable models and Multi-Atlas Segmentation (MAS).

2.1.2.1 Deformable Models: active contours and level sets

Deformable models are physically motivated and model-based methods. The main idea is to deform a parametric curve based on the influence of internal and external forces. Those models relate to the **active contours** concept or *snakes* which formulate the problem as an energy minimization. However, a major drawback of those methods is the fact that they are quite sensitive to initial conditions. In fact, deformable models need to be properly initialized otherwise, if the initialization curve is not close enough to the actual boundary it easily converges to a wrong boundary.

The **level set** technique gives a new paradigm for expressing the curve in active contour problems. The concept was first introduced by Osher and Sethian in [41]. The initial curve is interpreted as the zero level curve of a function $\Phi(t, 0)$. Then, the evolution of the curve is determined by a Partial Differential Equation (PDE). Moreover, level set methods strongly rely on the construction of the speed function F which controls the movement of the curve.

Formulation With the level set method, the evolving curve Γ is defined by the zero level set of a function Φ. At each time t the evolving curve $\Gamma(t)$ is given by the level set function $\Phi(x, y, t)$ by taking the set of points which satisfy: $\{(x, y)|\phi(x, y, t) = 0\}$. Thus, the standard definition of level set is given by Equation 2.2.

$$\Phi(\Gamma(t), t) = 0 \tag{2.2}$$

As we said, the evolution of the contour is determined by a speed function F. It denotes the speed at which the curve evolves in its normal direction, Equation 2.3

$$\frac{\partial \Gamma}{\partial t}.N = F \tag{2.3}$$

where N represents the outwards normal, Equation 2.4

$$N = -\frac{\nabla \phi}{|\nabla \phi|} \tag{2.4}$$

Now we can express the evolution function for ϕ by differentiating Equation 2.2 with respect to t which gives Equation 2.5.

$$\frac{\partial \phi}{\partial t} + \nabla \phi \frac{\partial \Gamma}{\partial t} \tag{2.5}$$

Then, by using Equation 2.3 and Equation 2.4 into Equation 2.5 we can write:

$$\Phi_t + F|\nabla\Phi| = 0 \tag{2.6}$$

Now we can easily see that the evolution of the boundary is defined via a partial differential equation on the zero level set of Φ:

$$\frac{\partial\Phi}{\partial t} = -F|\nabla\Phi| \tag{2.7}$$

In practice, the function Φ is defined by the distance to the boundary, *e.g.* euclidean distance, in such a way that it is positive outside the boundary and negative inside.

This gives the general framework of level sets. Then the speed function plays a key role depending on the problem. A lot of works studies this function and how additional terms could improve the convergence or precision [27, 28]. However, the initialization still plays an important role. Thus, to segment an anatomical structure, having prior information about the expected shape can significantly help in the process [29, 30, 31].

2.1.2.2 Multi-Atlas Segmentation

An atlas refers to a pair of intensity image and its corresponding label map. They are the result of the manual localization and delineation of the anatomical structures in order to give a label for every intensity value of the image.

Atlas-based segmentation methods are example-based approaches where the main objective is to find a transformation that maps the spatial coordinates of an atlas to a target image. Then, the target image could be segmented by label transfer, *i.e.* the assigned labels correspond to the label map of the spatially transformed atlas. The transformation should thus map the different structures and adapt to each individual case. This mapping refers to the process called **registration** which formally aims at aligning spatially two images: a reference image which remains static and a moving image that is transformed according to some criteria.

The first atlas-based methods where **Single-Atlas Segmentation (SAS)** approaches. Which means that a single atlas was registered to the target image. Thus, one could use a criterion to select the best suited atlas depending on the examples or register every available atlas and keep the one that gave the best registration score.

Then, **Multi-Atlas Segmentation (MAS)** methods started to emerge especially with the growing hardware computing power. Thus, it was possible to register every atlas and then use a label fusion to create the final segmentation map. Initially the idea was to send every atlas in a common intermediate space. It gives a probabilistic atlas representing the average

representation of the structures. Then, the target image was sent into this intermediate space and the inverse transform was used to assign the labels to the input image.

Then, other works found that directly registering the atlas to the target image was accurate and more computationally efficient. However, they suggest using a first atlas selection step to choose a limited number of atlas which are best suited for the given target image.

Thus, a MAS method gather the following steps:

- *Atlas Generation*: typically a manual operation which consists in creating the dataset of atlases
- *Preprocessing*: the atlas are preprocess to improve the generalization by reducing noise and artifact which are induced by the acquisition
- *Registration*: the atlases are registered to the target image. It implies finding the transformation that maps the two intensity images
- *Atlas Selection*: given the registration results, the atlases are carefully selected
- *Label Fusion*: each atlas can give a segmentation of the target image. Thus, we use label fusion, *e.g.* majority voting, to fuse each prediction into a unique and robust segmentation. This could be seen as an ensembling technique where many weak classifiers are combined to create a unique but strong one.

2.2 Deep Learning for Medical Image Segmentation

Deep Learning (DL) consists in training deep artificial neural networks. It has brought impressive breakthroughs in various machine learning tasks. Especially in computer vision with the introduction of Convolutional Neural Networks (ConvNets). For instance, in image classification, ImageNet was a key challenge that showed the potential of deep neural networks thanks to the availability of a large amount of labeled data. Rapidly, a lot of other fields showed interest in those models from Natural Language Processing (NLP) to face recognition or object localization. DL has also deeply impacted the medical image analysis field by becoming the standard choice for many applications. For example, in medical image segmentation, which is one of the most studied tasks, DL models have largely imposed themselves in the literature by beating all state of the art methods.

In this section we give an overview of ConvNets and the standard architectures for semantic segmentation. Then, we discuss how DL is used in medical image segmentation, what are the

losses and metrics used to evaluate the models and introduce several datasets for abdominal organ segmentation challenges.

2.2.1 Convolutional Neural Networks and Segmentation Networks

Convolutional Neural Networks (ConvNets) are a type of feed forward ANN specialised in CV tasks. Those models have been popularised by AlexNet [15] which won the ImageNet Large Scale Visual Recognition Challenge (ILSVRC) in 2012 though they were already beating traditional methods in other challenges. They are commonly composed of stacked **convolutional layers**, replacing the fully connected layers used in MultiLayer Perceptrons (MLPs). Convolutions as opposed to fully connected layers, process 2D images with small patches that run through the complete image. Thus, only a limited number of parameters are needed per layer which do not change depending on the input image. The first layers usually extract simple patterns (borders, textures, simple shapes, etc). Then, by using **pooling operations**, the last layers have a larger Receptive Field (RF) making it possible to assemble complex and more semantic patterns [42]. In image classification the last layers consist in a MLP which can be seen as a classifier that uses the semantic features extracted by the convolutions to take the final decision. The VGG-16 [3] in Figure 2.1 shows a typical convolutional architecture for image classification.

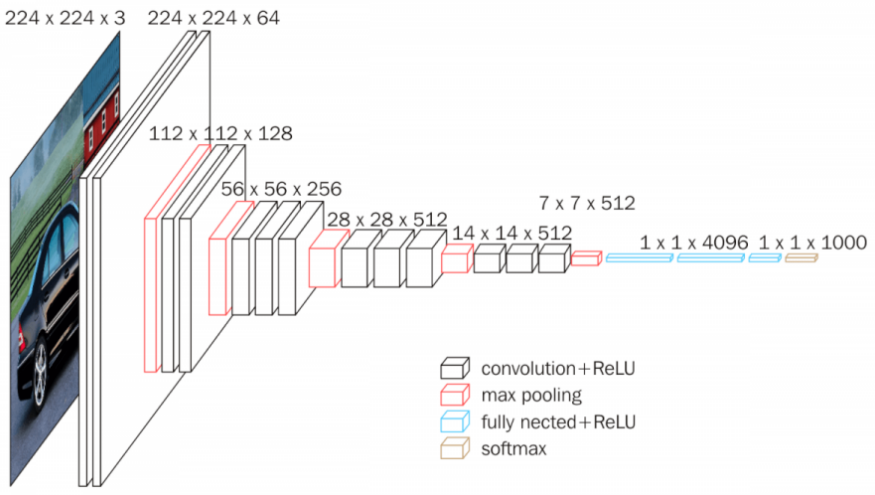


Figure 2.1: VGG-16 [3] architecture which uses a typical succession of convolution operations followed by fully connected layers for image classification.

One of the earliest success of ConvNets is in the works of LeCun [43, 44] who proposed LeNet-5[45] for handwritten digit recognition (MNIST). Next, more sophisticated datasets emerged such as CIFAR-10 and larger with millions of examples such as ImageNet. The amount of data

was a critical point to train ConvNets. Moreover, the hardware exponential evolution allowed us to store and train deeper models in a reasonable amount of time. That is why, since 2012, increasing the depth of the models was a major element of improvement. Table 2.2 shows an overview of the evolution of ConvNets in image classification over the years. We can see that one of the most important elements that improved the score in the ImageNet challenge was the increasing of the depth.

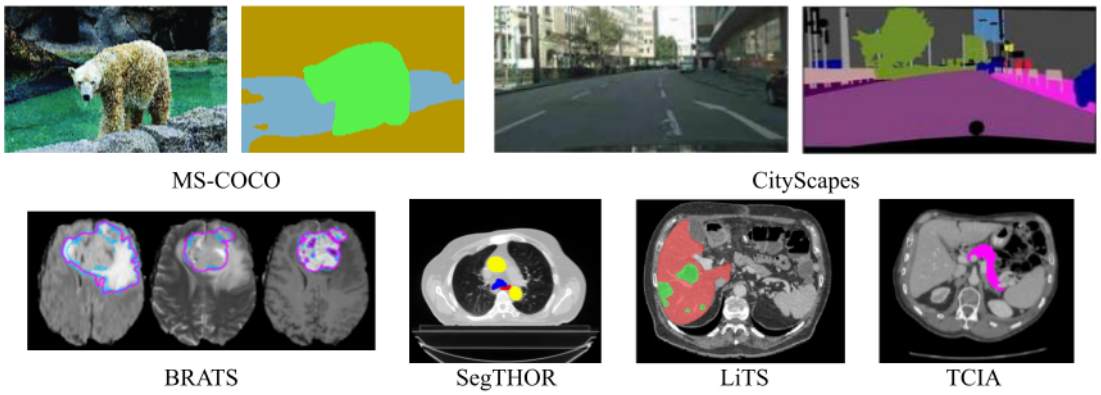


Figure 2.2: Examples of segmentation for various types of images. In the first row, natural images such as in MS-COCO and CityScapes datasets. In the second row, medical image segmentation datasets with from left to right: brain tumor segmentation (BRATS), SegTHOR, LiTS and TCIA pancreas.

Network	Year	Top-5 error on ImageNet	Number of layers	Number of parameters
AlexNet	2012	16.4	8	60M
VGG	2014	7.3	19	138M
Inception-V1 (GoogLeNet)	2015	6.7	22	4M
Inception-V3	2015	3.5	159	23.6M
ResNet	2016	3.6	152	25.6M
ResNeXt	2017	4.4	101	68.1M
Xception	2017	5.5	126	22.8M

Table 2.2: Overview of the best ConvNets over the years with the associated score on ImageNet, the number of layers and parameters for each architecture.

After the huge performance gains brought by ConvNets in image classification, other CV tasks were addressed by adapting those models accordingly, *e.g.* for the task of semantic segmentation. Image segmentation consists in assigning a label to every pixel of an input image. The main goal is then to detect, delineate and classify every object in the image and finally

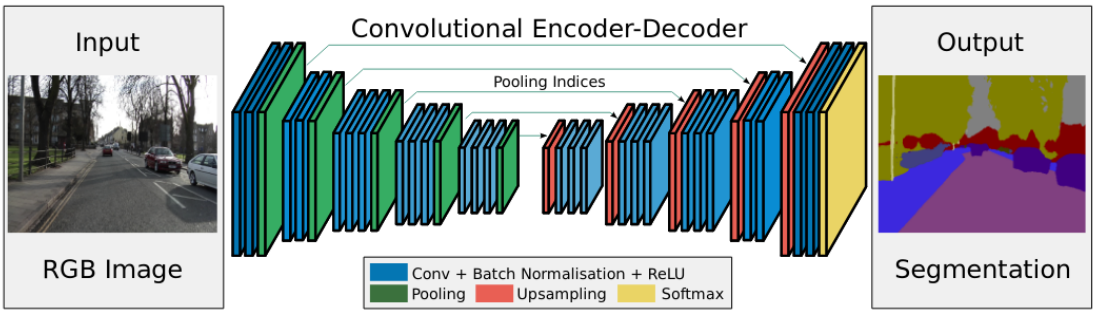


Figure 2.3: The SegNet [4] network from the original paper *"SegNet: A Deep Convolutional Encoder-Decoder Architecture for Image Segmentation"*

produce a dense prediction map. Figure 2.2 shows multiple examples of segmentation examples in various types of images.

The first successful result of a ConvNet in semantic segmentation was introduced in Long et al. [46]. The proposed architecture relies on VGG but the authors propose a Fully Convolutional Neural Network (FCN). It consists in a ConvNet where the fully connected layers are discarded and a final operation outputs a dense prediction, *e.g.* bilinear interpolation [46]. Thus, those models do not include other layers than convolutional layers.

The main issue when using directly the output of a network initially design for classification is the reduced spatial resolution. The information is thus insufficient to correctly localized the object at a pixel-level. To address this problem, one can use a second step that refine the segmentation map. For instance, the first version of DeepLab [47] propose to use a Conditional Random Field (CRF) at the end. Then, the authors further improved the proposed method by using atrou convolutions which allows to increase the field of view of filters without increasing the number of parameters. Another solution is to add a **decoder** which is designed to progressively recover the spatial information by using successive upsampling operations (*e.g.* deconvolution, bilinear interpolations) and convolutional layers. Thus, the **encoder** part is composed of convolutions and pooling operations which reduces the spatial resolution but increases the receptive field which is crucial for robust pixel classification. A typical example is SegNet [4] shown in Figure 2.3 based on VGG the authors proposed a symmetric architecture where the decoder is as deep as the encoder contrary to [46]. The DeepLab team adopted an encoder-decoder strategy in their latest version DeepLabv3+ [48] making this kind of model standard in image segmentation.

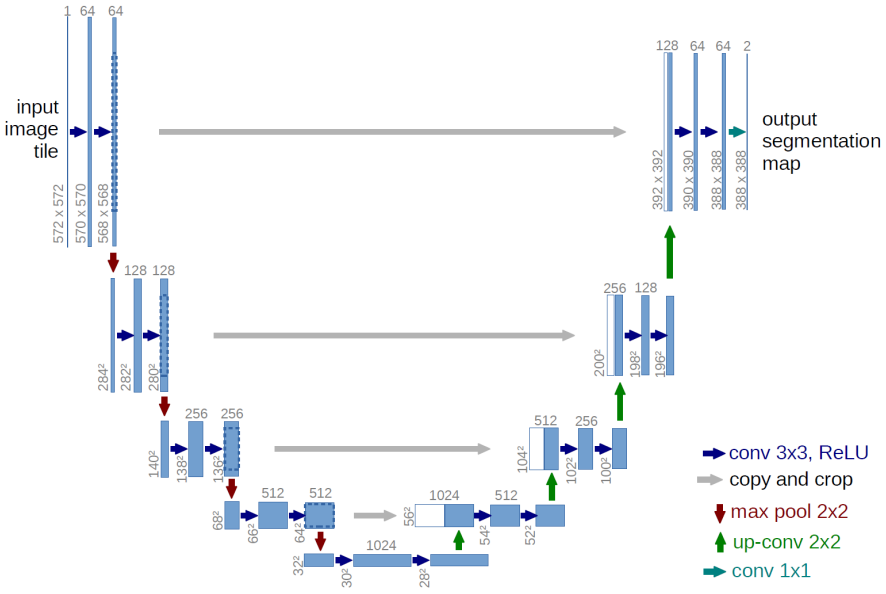


Figure 2.4: The U-Net architecture [2] presented in *"U-Net: Convolutional Networks for Biomedical Image Segmentation"*

2.2.2 Segmentation of Medical Images

In the previous section, we introduced architectures used for image classification and segmentation of natural images. In fact, ConvNets were at first designed for this task but other fields quickly started to use DL for addressing specific tasks. In medical images analysis, it started to gain a lot of interest after 2015 and the introduction of U-Net [2]. This model is the most popular encoder-decoder architecture which was in fact inspired by [49]. It has been developed for biomedical image segmentation in parallel with the models used for natural image segmentation. The architecture is adapted to numerous segmentation tasks but mostly finds its way in the medical image analysis community. It quickly imposed itself in the literature and was declined in numerous ways [50, 51, 52, 14, 53, 11]. The U-shaped architecture is completely symmetric such that the decoder part has the same number of layers as the encoder part, see Figure 2.4. Moreover, **skip connections** link the different levels to improve spatial recovery. This element is crucial and has given its popularity to U-Net.

Working on volumetric images Segmentation of volumetric images implies particular challenges. The main issue being the need of processing large volumetric images under memory constraints. When training a ConvNets one should keep in memory all the intermediate features in addition to the network parameters. With large input images, it quickly overflows current GPU cards. To address this problem, one should make a compromise on the input resolution,

the input size or the number of network parameters.

A first strategy consists in processing smaller images by using image patches to reduce the input size. Patch-based segmentation could be 2D slices along one or multiple views [54] or 3D patches cropped from the full volume [11, 55]. Using small patches reduces substantially the needed memory but at the cost of a reduced spatial context. Another approach is to reduce the spatial resolution of the input image by drastically downsampling the image so it can fit in memory [50, 51, 11]. More contextual information is preserved but it becomes tenuous to localize small objects and correctly delineate complex boundaries. It is also possible to combine the two solutions like in [56] which uses a patch-based network on 2D slices along with a coarse 3D segmentation network. The combination of the two segmentations gives the final result.

A second solution is to reduce the network's depth by using less convolutional layers. However, it mechanically reduces the RF. For example, the original U-Net [2] has 19 convolutional layers but processes 2D images compared to the 3D U-Net [50] which has 15 layers but processes 3D volumes. To mitigate this effect, it is possible to increase the size of the convolutions kernels but then the number of parameters increase rapidly especially with 3D kernels.

2.2.3 Metrics and Losses

Working on medical images implies dealing with specific characteristics. Moreover, some conventions differ from the analysis of natural images. In this section we will look at specific elements we need to consider when training models on medical images for organ segmentation.

Model Evaluation In semantic image segmentation, the evaluation of the model performances relies almost exclusively on the mean Intersection Over Union (mIOU) measure, also called Jaccard index. It gets a score by observing the overlapping of the prediction over the ground truth for each class. Then, the average over the classes gives the mIOU.

In the medical analysis literature, the mIOU is nearly never used. However, the Sørensen–Dice coefficient also known as the Dice Similarity Coefficient (DSC) is preferred. Equation 2.8 is the formulation of this metric which is very close to the Jaccard index. In fact, one could find the DSC from the Jaccard index with the following equation: $DSC = 2J/(1 + J)$.

$$DSC = \frac{2|X \cup Y|}{|X| + |Y|} \tag{2.8}$$

In addition to the overlapping measure, it is common in medical image analysis to also report surface distances. The two mostly used distance metrics are the Hausdorff Distance (HD) and the Average Surface Distance (ASD). Let us consider M as the set of all the voxels belonging to the boundary of the model segmentation (thus, at least one neighboring voxel is outside the

segmentation), and G the set of the voxel belonging to the boundary of the ground truth. Then the HD is the maximum distance observed between the two sets as shown in Equation 2.9.

$$HD(M,G) = \max\left\{ \sup_{m\in M} d(m,G), \sup_{g\in G} d(M,g) \right\} \tag{2.9}$$

On the other hand, the ASD is the average of the distances between the sets as shown in Equation 2.10.

$$ASD(M,G) = \frac{\sum_{m\in M} d(m,G) + \sum_{g\in G} d(M,g)}{|M| + |G|} \tag{2.10}$$

The DSC is the standard metric used in medical image segmentation but has the major drawback of being sensible to the size of the evaluated object [57]. An error of one pixel in a small object impacts considerably the DSC compared to the same error on a large object. Moreover, this metric does not give insight on the shape and spatial position coherence of the prediction. On the other hand, surface distances are able to detect critical outliers and how the prediction was able to correctly reconstruct the shape of the considered organ.

Overlapping measures (DSC) and surface distances are complementary and give specific insights on the prediction quality. However, the DSC is more general and thus much more used and analyzed in the literature.

Specific Losses When training a ConvNet, one should choose the appropriate loss function. The most used is cross-entropy (CE) (Equation 2.11) which again comes from image classification. In segmentation this loss shows very decent performances. Thus, it has been used largely. However, it does not take into account the class imbalance which is a problem in medical images segmentation. This may be a problem because the large majority of a medical image when segmenting a specific organ is the default "background" label. Thus, an easy solution for the cross-entropy is to predict only "background" which is a completely unsatisfactory solution given the initial problem.

$$CE = -\sum_{i=1}^{N} y_i log(\hat{y}_i) \tag{2.11}$$

To address this issue, the first and most straightforward solution is to weight the loss [46, 2] depending on the occurrence of the labels. However, the choice of the weighing could also cause over-segmentation for small objects.

Using cross-entropy for image segmentation also raises the question about coherence with the evaluation metric. In fact, cross-entropy is derived from the accuracy metric from image classification. However, as we presented in the last paragraph, we evaluate our model by using

the DSC. In [51], the authors thus propose to use a loss derived from the evaluated score, *i.e.* DSC. They propose the Dice loss in Equation 2.12.

$$d_{\text{loss}} = 1 - \frac{2\sum_i^N y_i \hat{y}_i}{\sum_i^N y_i^2 + \sum_i^N \hat{y}_i^2} \tag{2.12}$$

However, this loss could be unstable [57] and is often mixed with the CE loss [11].

2.2.4 Abdominal Organ Segmentation Datasets

Today, a huge amount of medical images are acquired daily. However, interpreting such data requires a strong knowledge and professional radiologists are needed. Thus, there is a limited number of labeled datasets and they usually focus on a limited number of structures. In this section we detail the datasets used in the following works and recent challenges that enrich the landscape of organ segmentation. We also include a private dataset provided by Visible Patient (VP) for abdominal multi-organ segmentation extracted from internal data.

- **Liver Tumor Segmentation Challenge - LiTS**

  The LiTS dataset is well known as it is the most exhaustive source of data for liver tumors segmentation. It was proposed for challenges in ISBI 2017 and MICCAI 2017. The total dataset contains 131 CT-scans with the segmentation of livers and tumors. In this thesis we focus on the segmentation of organs and not the specific task of tumor segmentation. Thus, in our experiments we use the segmentation of the liver and do not evaluate on tumors.

  Each CT-scan is composed of $74 \sim 987$ slices of 512×512 pixels and a voxel spatial resolution of $([0.56 \sim 1.0] \times [0.56 \sim 1.0] \times [0.70 \sim 5.0])\text{mm}^3$.

- **TCIA Pancreas [58]**

  The TCIA dataset is a very challenging problem. It contains 82 CT-scans with the pancreas completely labeled. However, the pancreas is a very difficult organ to delineate, thus there is a significant inter-annotator variation leading to heterogeneous labeling. However, the large number of cases made this dataset very interesting to evaluate DL models.

  Each CT-scan is composed of $181 \sim 466$ slices of 512×512 pixels and a voxel spatial resolution of $([0.66 \sim 0.98] \times [0.66 \sim 0.98] \times [0.5 \sim 1.0])\text{mm}^3$.

  Multiorgan Other organs in this dataset have been manually segmented in [59] (the spleen, left kidney, gallbladder, esophagus, liver, stomach, pancreas and duodenum). This dataset

is composed of the enriched TCIA dataset and the Beyond the Cranial Vault (BTCV) Abdomen dataset.

- **Internal Multi-Organ dataset - IMO**

  In addition to the two above mentioned public datasets, we used a private multi-organ dataset. The main advantage of it is the number of cases with many labeled abdominal organs. It is composed of 90 CT-scans where the following organs are completely labeled: liver, gallbladder, pancreas, spleen, right and left kidneys and stomach.

  Each CT-scan is composed of $57 \sim 500$ slices of 512×512 pixels and a voxel spatial resolution of $([0.42 \sim 0.98] \times [0.42 \sim 0.98] \times [0.63 \sim 4.00])\text{mm}^3$.

- **Combined Healthy Abdominal Organ Segmentation - CHAOS [60]**

  This dataset combined CT-scans and MRI with the segmentation of liver, kidneys and spleen. It was part of an ISBI challenge in 2019 and aimed at evaluating the ability of segmentation methods to deal with data from two different modalities which is also called "cross-modality". Five main tasks were proposed to the participants including the segmentation of the liver with single modality, *i.e.* CT then MRI, and with both modalities, *i.e.* CT and MRI. Then adding the kidneys and the spleen, two other tasks include the multi-organ segmentation with cross-modality, *i.e.* CT and MRI, and with a single modality, *i.e.* MRI.

  A first dataset includes 40 CT-scans with a resolution of 512×512 and $77 \sim 105$ slices. The voxel spatial resolution is $([0.7 \sim 0.8] \times [0.7 \sim 0.8] \times [3.0 \sim 3.2])\text{mm}^3$.

  The second dataset is composed of 120 MRI images with a resolution of 256×256 and $26 \sim 50$ slices with a spatial resolution of $([1.36 \sim 1.89] \times [1.36 \sim 1.89] \times [5.5 \sim 9.0])\text{mm}^3$.

- **Segmentation of THoracic Organs at Risk in CT images - SegTHOR [61]**

  SegTHOR provides a dataset of CT images to address the problem of Organs At Risk segmentation (OAR) in lung and esophageal cancer. Thus, four organs are manually delineated: the heart, aorta, trachea and esophagus. They are particularly difficult to segment especially in the case of tubular organs, *e.g.* esophagus. The dataset is composed of 40 training scans and 20 for testing.

  CT-scans are composed of $150 \sim 284$ slices with a resolution of 512×512 pixels with a mean voxel spatial resolution of $(0.98 \times 0.98 \times [2.0 \sim 3.7])\text{mm}^3$.

- **Medical Segmentation Decathlon - MSD [62, 63]**

The decathlon dataset was built with the main purpose of giving a fully open source and comprehensive benchmark to evaluate challenging problems in medical image segmentation through organized challenges. It consists of ten different tasks each having a specific dataset and test set. It goes from abdominal organ segmentation, *e.g.* liver, pancreas spleen, to brain or even cardiac segmentation. Some problems imply difficulties such as the segmentation of tumors or tubular structures like vessels.

2.3 Semi-supervised Learning and Partial-Labels

In the first part of this thesis we were interested in training a deep FCN with partial-labels. This problem comes from the fact that the available data have heterogeneous labeling due to the region of interest studied for each case.

When compiling data with heterogeneous labeling, some organs are unlabeled in some volumes and labeled in others which leads to a wrong labeling of the unlabeled cases where the voxels associated with the organ are in fact labeled as the default "background" value. When looking at a specific organ and ignoring the others, we have a set of labeled images and unlabeled images. Consequently, our problem could be seen as a Semi-Supervised Learning (SSL) problem [64].

In recent literature, SSL approaches in medical image segmentation can be classified into generative models, teacher-student networks and pseudo-labeling methods.

Generative models can be leveraged to incorporate training signals on unlabeled data for medical image segmentation. For example, [65] uses a variational autoencoder (VAE) to learn representations on all images, and then train a decoder only on labeled data. In the same idea, [66] applies a generative model based on a VAE, where the encoder is trained to reconstruct input images, and the decoder to reconstruct unpaired segmentation masks. Adversarial training [67] is another appealing direction for semi-supervised semantic segmentation. The overall idea initially applied to generalist images in [68], is to consider the segmentation network as a conditional generator given input images, whose output distribution should be similar to the ground truth distribution of segmentation masks. The appealing feature in SSL is that this adversarial loss can be applied on unlabeled data to improve segmentation performances. Recently, the approach has also been successfully applied for medical image segmentation [5, 69]. In those methods, when an image is unlabeled the output of the discriminator is used as a confidence map to compute a segmentation loss between the encoder prediction and its binarized counterpart for the most confident pixels. An example is given for ASDNet in Figure 2.5.

Teacher-student networks have also been used in SSL to enforce desirable behaviours on the segmentation models, where the teacher is trained only on the labeled data and the student is

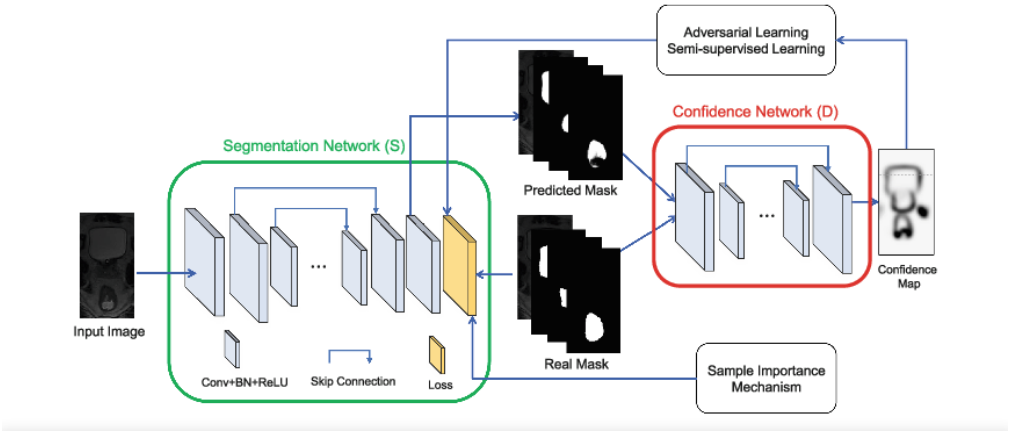


Figure 2.5: The ASDNet architecture [5] presented in *"ASDNet: Attention Based Semi-supervised Deep Networks for Medical Image Segmentation"* which uses an adversarial training for a semi-supervised problem.

subsequently trained on all data. Some methods introduce an auxiliary task that does not need the segmentation ground truth. In [70] and [71], the authors proposed to regress the region size and use a consistency term that penalizes non-realistic sizes. In the same way, the fact that the same image under different transformations should get the same output is used to create a consistency term. For example in [72, 73, 74, 75] this idea is applied by defining two losses, the first is the classic segmentation loss and the second the consistency loss which does not need ground truth labels. In [75], the authors proposed an advanced method by introducing a confidence estimation based on monte carlo dropout to select the most certain predictions in the consistency term for the unlabeled images.

Although these SSL methods show good results, the incorporation of the unlabeled data in the final results is implicit. **Pseudo-labeling** [76, 77] consists in using the model's predictions as ground truth training signals on unlabeled data. In the context of partial labels, the goal of these approaches is to automatically relabel unlabeled data from a model trained on a labeled set. Recently, this strategy has been extensively applied for semi-supervised semantic segmentation, [6, 78, 79], leading to state-of-the-art performances. For example, Figure 2.6 is an example of a method which uses pseudo-labels for training a segmentation model in a domain adaptation problem. Then, this strategy has also recently been applied for medical image segmentation [80, 81, 82]: the idea is to first learn a model on the labeled data. Then, enlarge the training set with the union of the labeled data and the model's predictions for the unlabeled data. Finally, either the same model or a new model is trained on the new training set.

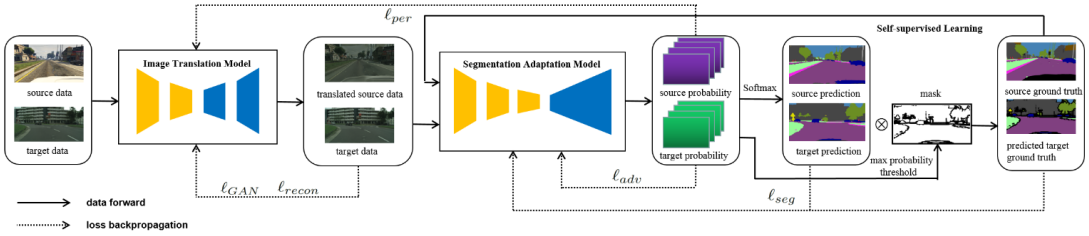


Figure 2.6: Self-supervised training with pseudo-labeling presented in *"Bidirectional Learning for Domain Adaptation of Semantic Segmentation"* [6] for domain adaptation.

2.4 Integrating Prior Knowledge

Medical imaging and tomography are incredibly powerful tools that give a view of the internal body without invasive procedure. It is known for a long time that organs have a standard spatial layout. Spatial representations of human anatomy have had a key role in the communication and the development of knowledge about biological mechanisms. The first attempt of labeling anatomical regions based on the spatial information could be with the work of neurosurgeon Jean Talairach with the *Talairach Atlas* in 1967. He created a coordinate system that adapts to brains of different sizes and can be used with the *Brodmann areas* to label the brain regions. Then, with the emergence of computers, automatic segmentation methods were developed such as atlas-based methods which rely on the spatial location of structures and try to register "standard" atlases to a target image. Early methods for medical image segmentation used spatial information as it is a key characteristic to localise structures. Even professional practitioners rely on their strong knowledge of the human anatomy to manually label the organs.

Lately, the great success of DL in medical image segmentation has seen a lot of end-to-end trained models based on the use of annotated data. The most used models being FCNs. However, it ignores an important property of the convolution which is the equivariance to small transformations. Thus, FCNs are by design unable to directly encode spatial location information to bias predictions as shown in [83]. The authors show that FCNs are unable to model a coordinate transform task such as regressing the coordinates of a 1 in a grid of zeros.

Models and Architectural Solutions A solution could be the Locally Connected Networks (LCNs) which are able to model spatial information. LCNs learn prediction models specific to each spatial position, and have been successfully applied to face recognition, e.g. DeepFace [84]. However, LCNs significantly increase the number of parameters of the model (compared to their convolutional counterparts), and thus require huge labeled datasets to be robust to

overfitting. LCNs are consequently not adapted for medical image segmentation where only few labeled data are available.

In the medical image analysis literature, cascaded networks [85, 86, 87, 88, 89, 7] include spatial information by relying on the selection of a Region of Interest (RoI) by a first model, which is subsequently refined by a second one which performs a more accurate segmentation. For example, Figure 2.7 illustrates the proposed method introduced in [7] where we can see how a fine-scale segmentation is performed based on a coarse segmentation of the pancreas. Although these approaches are efficient, they are intrinsically limited by the quality of the first RoI selection step. Some works simply take cropped images of the expected RoI [90, 11] which is in fact a very strong prior about the organ position. However, it does not use the whole image and is very limited to the selected region. Thus, each class should be learned independently [90] which drastically increases the model complexity and computational burden.

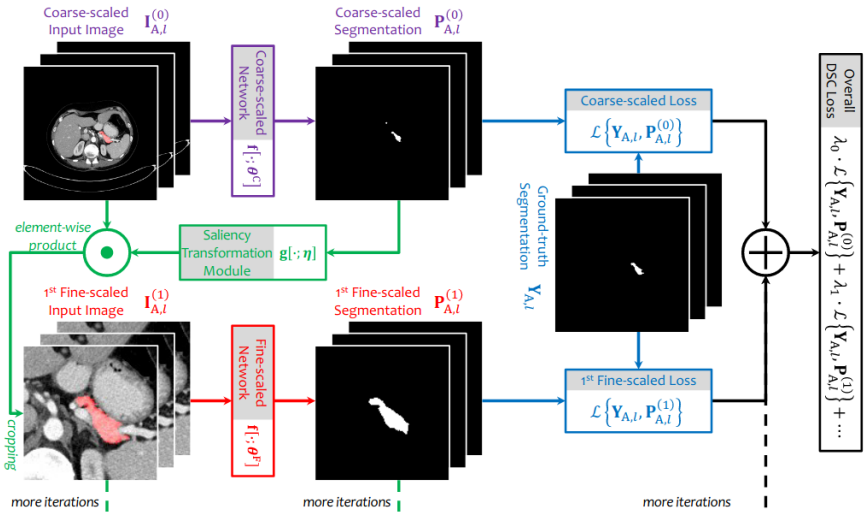


Figure 2.7: The presented cascaded scheme proposed in *"Recurrent Saliency Transformation Network: Incorporating Multi-Stage Visual Cues for Small Organ Segmentation"* [7].

Other methods try to incorporate spatial prior information by biasing the learning of internal deep representations in an implicit manner [91, 66, 92]. In the same way, attention mechanisms have gained popularity in the last few years. New parameters bias the intermediate representations to focus on a specific region of an image. For example, in [11] the method integrates an additive attention block in the decoder part of a U-Net model. The attention coefficients are learned during training and are completely implicit. Thus, we cannot assure that the model actually learns a prior on the spatial position. Moreover, despite the reasonable improvements shown by these methods in fully-supervised settings, they are intrinsically limited to 2D spatial

information, which may arguably be inaccurate for organ segmentation with a complex shape varying in 3D.

Biasing the Training with the Loss Function Some works try to constrain the learning by using prior knowledge in the loss [93]. For example in [94], the authors proposed a constraint on the organ size. By observing the training set, they get the expected organ size and then add a term in the loss which penalizes predictions that do not respect this constraint. In [69], a probabilistic spatial prior is used to weight the pixels depending on the difficulty in the sense that regions where the organ appears only a few times are critical. This loss is directly inspired by the focal loss but with a weighting set by the spatial prior. Another idea is to use a regularization term which penalizes unrealistic predictions like in [8]. In this work the authors proposed a star shape prior for skin lesion segmentation. They use the fact that a lesion should be a star shape which means that if c is the center of an object O, O is a star shape object if, for any point p inside O, all the pixels q lying on the straight line segment connecting p to c are inside O, as illustrated in Figure 2.8 left image (a). Then, every violation of this rule Figure 2.8 (b) and (c) adds a cost to the loss.

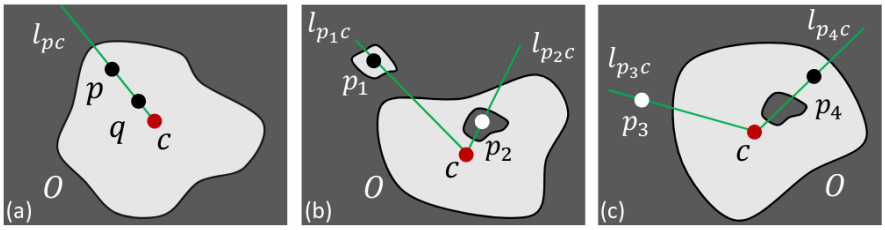


Figure 2.8: The star shape regularized loss presented in *"Star Shape Prior in Fully Convolutional Networks for Skin Lesion Segmentation"* [8].

2.5 Leveraging Contextual Information

As we saw in the previous section, Section 2.4, ConvNets have a major limitation which is their inability to model spatial absolute positions of objects in images. Another determinant point is the fact that ConvNets are unable to model relationships between objects. Moreover, the Receptive Field (RF) is often limited and the participation of a pixel in the final decision decreases with respect to the distance of the considered pixel. Thus, ConvNets do not correctly leverage contextual information which is crucial in medical images segmentation, especially when considering the difficulty of some organs.

Some works have tried to address this problem, often indirectly. This is the case of the cascaded networks [85, 86, 87, 88, 89, 7] as discussed in the previous section, Section 2.4. Using a multi-step segmentation allows to converge to a finer solution by focusing on a limited and small region (RoI). However, it does not try to increase the use of the overall context but avoids it by an architectural trick. Moreover, this solution is computationally inefficient and leads to parameter redundancy and excessive computational resources.

In fact, learning to leverage contextual information is not straightforward. Lately, attention models have gained a lot of interest and try to make the model focus on specific regions making use of a maximum of relevant context.

Attention Models Attention mechanisms in deep learning models aim at focusing on local regions based on local features and filtering irrelevant ones based on information from global features. These models have become important in various tasks initially in NLP with machine translation [95, 96] but has also been successfully applied in vision problems such as image captioning [97, 98] or image classification [99, 100, 101].

Despite the popularity of attention models in computer vision, they are a relatively recent problem in medical imaging and only few works using simple attention modules have emerged [102, 103, 104, 5, 105, 11, 106]. In [102], the authors propose a multi-resolution attention module which combines local deep attention features with a global context. For that, they use a simple attention module which consists of three convolution layers followed by a softmax which outputs the attention map. The same attention module is also used in [103]. Then, in Attention U-Net [11], additive attention gates are introduced in the decoder of a U-Net for combining the up-sampled features and the skip connections. This model is further detailed in Chapter 5. However, those solutions still have an important drawback, they compute weights on local features and do not model interactions with other pixels. Recent works in [106, 107] address this problem through a Dual attention proving the importance of full range attention but to the cost of large parameter overhead and multiple concurrent loss functions.

Transformers Differently, Transformer models [17] tackle the problem of local attention by proposing a self-attention mechanism which aims at modeling long-term dependencies. They have brought a lot of interest in the CV field and have witnessed increasing success in the last few years. Those models were initially introduced in Natural Language Processing (NLP) with text embeddings [17, 108]. A pioneer use of transformers in CV is non-local networks [109], which combines self-attention with a convolutional backbone. The attention modules of Transformers compute similarities between a given feature with all the other features which give them a good ability to model long-range dependencies. The complete Transformer block is shown in

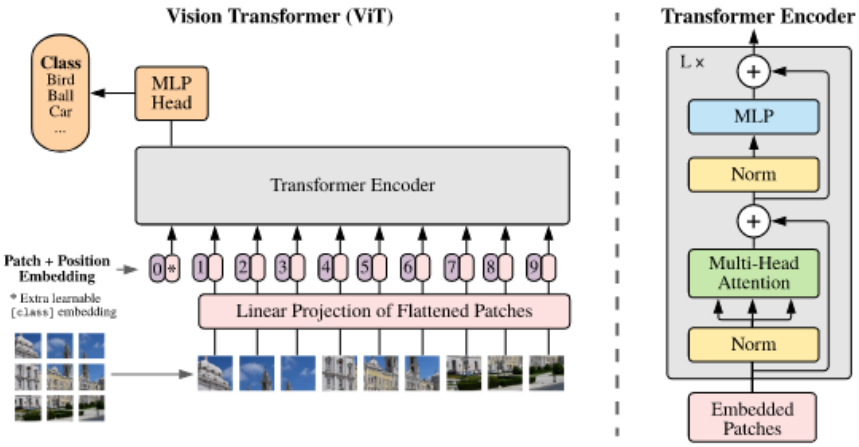


Figure 2.9: The Transformer encoder from [9]: *"An Image Is Worth 16X16 Words: Transformers For Image Recognition At Scale.* The encoder is composed of L Transformer blocks with Layer Normalizations followed by a Multi-Head Attention, another Layer Normalizations followed by a MLP and residual connections.

Figure 2.9 as used in Vision Transformer (ViT) [9]. Transformers were quickly adapted to other tasks such as object detection, *e.g.* DeTr [110] or image classification, *e.g.* ViT [9]. In semantic segmentation, initial attempts include [111, 112], and more recent but promising works are SeTr [113] or Swin Transformer [114].

Transformers are undoubtedly promising models especially in vision where leveraging global interactions is crucial. In Chapter 5, we propose to use Transformers for medical image segmentation and a solution to adapt them in the context of organ segmentation.

Chapter 3

Training Deep FCNs with Partial-Labels for Medical Image Segmentation

Contents

Abstract

Training deep ConvNets requires large labeled datasets. However, in medical image analysis, there is plenty of tasks to be addressed and each problem needs a dataset with a specific type of images, a different modality or certain targeted structures. Thus, compiling a sufficiently large and exhaustive dataset is hardly conceivable. However, there is a huge amount of heterogeneously labeled data. For example in organ segmentation, some images only have a single labeled organ when others have multiple ones. Thus, using those partially labeled data could benefit in the process of training deep neural networks. In this chapter, we introduce INERRANT a method for training deep FCNs on partially-labeled images. Firstly, we introduce a specific loss function which aims at training the model

only on correct labels and ignore the background labels which are assigned by default to the missing organs. Secondly, we propose an iterative pseudo-labeling scheme inspired by curriculum learning where we progressively relabel the missing organs by selecting the most confident predictions and add them to the training set. Moreover, we propose a dedicated confidence network which learns an advanced confidence estimation for a better label selection. We show experimentally on three datasets the relevance of our method and propose a deeper analysis of the different parts. Finally we show a practical use case where a limited number of completely labeled data are enriched by publicly available but partially labeled data.

3.1 Introduction

As detailed in Chapter 1, training deep ConvNets requires large datasets of fully-labeled images. However, in medical image analysis, there is a wide range of problems with various contexts which need to be addressed. For example, in medical image segmentation one could study the segmentation of the liver when another the kidneys. Each problem needs an adapted dataset and building a complete and exhaustive one is not conceivable. Moreover, in medical image segmentation the labeling process is extremely time-consuming and requires highly qualified professionals. As a consequence, large-scale and clean medical image datasets are rarely available and the manual labeling process often focuses on specific anatomical structures, *e.g.* in Figure 1.4, the first image show an image with multiple labeled organs when the second has only the liver and tumors probably because this second case aimed at analysing only the evolution of the lesions. Thus, large datasets containing partially-labeled images are easier to obtain by aggregating smaller labeled datasets with different amounts of labels compared to a complete dataset containing all the abdominal organs.

The disposal of large-scale labeled and publicly available datasets has increased recently, *e.g.* CT-ORG [115]. Having access to such large-scale public datasets is very valuable and can help to provide more powerful prediction models. However, collecting large-scale datasets that are "universal" and could be useful for any medical image segmentation task arguably remains elusive. For example, the IMO dataset used in our paper contains 90 CT-scans with 7 abdominal organs, while CT-ORG is larger in terms of cases (140 CT-scans) but with fewer labeled organs: 6 organ classes, and only 3 in common with ours (liver, gallbladder and kidney). This illustrates the challenge addressed in this paper: despite the existence of massively annotated datasets, it is very difficult to compile a complete, exhaustive and homogeneous dataset for any medical problem. Heterogeneity in medical imaging can have various sources. Firstly, granularity between studies might substantially differ: datasets on the entire body will focus on large structures (*e.g.* bones, lungs, liver), a study focusing precisely on the abdomen will try to get finer structures (*e.g.* . pancreas, spleen, stomach), while finer tasks could even include the vascularisation with vein/artery networks. Secondly, there are commonly strong variabilities in the acquisition process between studies: images are acquired with different devices and different protocols: images depend on the injection time of the contrast media which is chosen depending on the targeted structure [116, 117, 118].

In this chapter, we address the important issue of training deep ConvNets with noisy and partially-labeled datasets. Our training context is illustrated in Figure 3.1: in this example, the input slice is partially labeled with 3 organ classes out of 7 for the unknown complete labeling. As we verify experimentally, naively applying state-of-the-art models such as U-Net to these

partial labels leads to bad performances, since it includes wrongly labeled background pixels for missing organs.

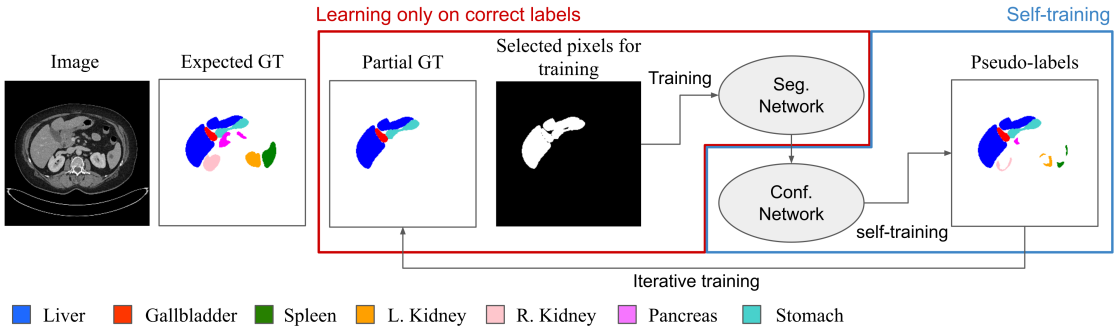


Figure 3.1: The 3D CT-scan is partially labeled: in this slice, only 3 out of 7 organs are labeled. Naively using such partial ground-truth (GT) labels is inappropriate since it includes wrong background labels for missing organs. INERRANT is based on identifying pixels for which labels are correct, and ignoring others. The segmentation network is trained on those data and a confidence network outputs confidence scores for each pixel to incrementally add pseudo-labels to the training set and recover the unknown complete ground-truth labels.

To specifically handle the partial labeling problem, we introduce a method which could be decoupled into two parts: Firstly, we propose a specific loss to train the segmentation network dedicated to include only correct labels, *i.e.* it selects pixels that could be learned and those that should be ignored during training (white *vs* black pixels in Figure 3.1). The general motivation is to eliminate all pixels that are wrongly labeled as background for missing organs. Secondly, we propose a self-supervised scheme to iteratively relabel the missing organs by introducing pseudo-labels into the training set, in order to estimate the unknown complete ground-truth labels. For that, we add a confidence network that helps to select the best pseudo-labels and thus reduce the introduction of wrong predictions. The overall approach is called INERRANT, Iterative coNfidencE Relabeling of paRtial ANnoTations.

Our proposed method strongly relies on the medical nature of the considered images. We leverage import priors about the organ labels such that every organ is present, even if non-labeled in the input volume at only one place (a class is assigned to only one object in the volume). It allows us to deduce the unlabeled organs for each volume and thus, which classes should be ignored during training. Moreover, our method is complementary with the access of large datasets by leveraging various types of labels and granularities to build a more exhaustive dataset and thus a more robust segmentation model. It opens up the possibility to add a new organ class which is less represented in public and private datasets to enrich an existing one.

The contributions of this work are as follows:

1. We propose a specific loss to train a segmentation network only on correct labels. Relying on the fact that all organs should be visible in the volume, we deduce and thus select pixels that could be learned and those that should be ignored.

2. Then, we propose an iterative pseudo-labeling scheme based on Curriculum Learning [12] for automatically relabeling missing organs in the training set. To further improve the pseudo-label selection, we introduce a confidence estimation learned via a dedicated network best suited for distinguishing errors from correct predictions, which enables to maximise the number of correct labels introduced during pseudo-labeling.

3. We provide a thorough evaluation by reporting performances on two public datasets (TCIA and LiTS) and one internal dataset containing seven organ classes (IMO). Moreover, we give a comparison with recent state-of-the-art semi-supervised methods for learning with partial labels and an ablation study highlighting the importance of the iterative pseudo-labeling process and the confidence measure.

4. Finally, we show a practical use case where we combine a IMO dataset with a single-organ public dataset (TCIA). This shows that we can exploit large amounts of labeled images by gathering heterogeneous data.

In this chapter we start in Section 3.2 with a review of the specific state-of-the-art for this method. A more general review is given in Chapter 2, Section 2.3. We then present in Section 3.3 the INERRANT method for training a deep FCN on a partially labeled dataset and then how we progressively incorporate pseudo-labels using an advanced confidence measure. The experiments and results are given in Section 3.4. It shows how INERRANT gives better results than state-of-the-art semi-supervised methods on three different datasets. Moreover a detailed analysis of the confidence evaluation is presented. Eventually, a conclusion is given in Section 3.5.

3.2 Related Work

Iterative Pseudo-Labeling In Section 2.3, we saw that our partial-labeled problem could be seen as a semi-supervised problem. Thus, we introduced several approaches for including unlabeled images in the training of DL models.

We choose to build our method with a pseudo-labeling technique by using ideas from Curriculum Learning (CL) [12] which states that a machine learning model reaches higher accuracy if easy examples are presented first and hard examples near the end of the training. For that,

they used a trained model which sorted the data by difficulty, *i.e.* which examples are hard and which are easy. Then, a new model is trained by using first the easy examples and then progressively adding the hard examples. They experimentally showed that the second model has a higher accuracy than the first one.

A more recent paper introduced a method with a similar idea, the Self-Paced Learning (SPL) [13]. With the same motivations as CL, [13] ordered the examples on-the-fly, which means that a "hard" example at step t could become "easy" at step $t + n$. The hard examples are then dropped by using a threshold and the model is trained only on the easy examples at each step. When the model becomes more accurate, some hard examples become easy and are thus integrated to the training. SPL technique stabilizes the gradient descent and allows a quicker and higher convergence.

In both methods, one should quantify the difficulty of each example. The most straightforward solution is to use the output probabilities of the ML model.

With INERRANT, we adopt an iterative pseudo-labeling scheme where for each step we add new examples based on the predictions from the previous step. The first selected pseudo-labels are the most confident predictions and are seen as "easy" examples. It is even more the case for the labeled examples that were used to train the model. Then, at each iteration we add new pseudo-labels in the same way as in SPL where "hard" examples become "easy" examples when the model improves.

In the iterative scheme, we need to properly select the pseudo-labels to add to the training set. Thus, we must rank the best candidates to mitigate the selection of false positives. For that we need to measure the **confidence** of the network in a given prediction.

Confidence Estimation with Deep Learning Confidence estimation in deep learning is a crucial yet complex problem. The most naive confidence estimation for deep neural networks consists in using the probability of the predicted class, *i.e.* the Maximum Class Probability (MCP) [119]. Although this baseline is widely used in practice, it also suffers from fundamental drawbacks, *e.g.* the probabilities are known to be non-calibrated [120]. In the last few years, there has been an extensive revival of Bayesian deep learning, especially by the connections drawn between variational inference and stochastic regularization in deep learning, *e.g.* Monte-Carlo Dropout [121]. However, this confidence measure is computationally demanding since it requires several forward passes, and does not yield accurate uncertainty measures when aleatoric uncertainty is crucial. In contrast, misclassification approaches design confidence estimates targeted to properly separate correct predictions from errors, *e.g.* trust score [122] or ConfidNet [123].

In pseudo-labeling, the chosen confidence measure should prevent incorporating wrong la-

bels to improve the final prediction. It is worth mentioning that most recent approaches for semantic segmentation rely on MCP for selecting target labeled pixels [6, 78, 79, 80, 81, 82], although MCP by design assigns overestimated confidence values to prediction errors. In this paper, we train an auxiliary network to design a relevant confidence measure, which is based on misclassification detection and explicitly assigns low confidence values to prediction errors. We verify experimentally that this confidence measure leads to better final segmentation performances than MCP.

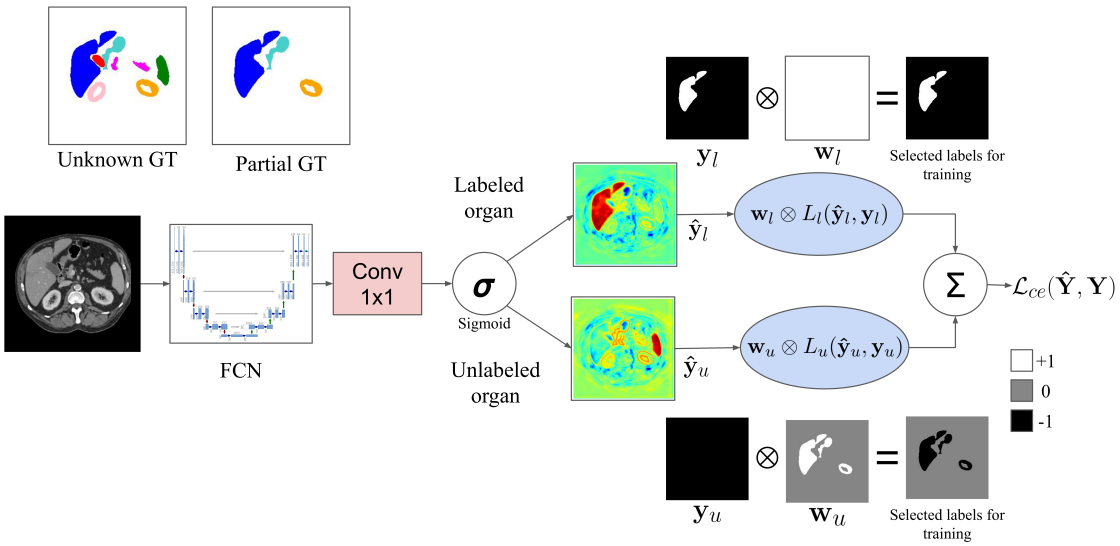


Figure 3.2: Training INERRANT on a partially labeled dataset. Each organ is predicted by a common FCN. Depending on the missing organs deduced by the available labels, an ambiguity map $\mathbf{w}_k$ is created to ignore potential wrong labels in the loss. It acts as a weighting in the final loss function.

3.3 Training from partial labels with INERRANT

In this section we detail INERRANT for training deep ConvNets on partially labeled data.

Firstly, we introduce in section 3.3.1 a learning scheme that only leverages correct labels. More precisely, INERRANT is trained not only with the true positives (TPs), *i.e.* the positive labels which are actually positive in the complete ground truth, but also with the true negatives (TNs), *i.e.* the background labels which are actually background in the ground truth. As mentioned in section 3.1, a naive method that learns directly with partial labels incorporates false negatives (FNs), which negatively impact performances. We also provide a statistical analysis of the ratio of correct labels used by our method *vs* the naive baseline.

Then, we introduce in section 3.3.2 a self-supervised scheme which iteratively adds pseudo-labels for the missing organs, in order to recover the missing ground-truth labels. Since the pseudo-labeling is automatic, the challenge is to maximise the number of correct label predictions, denoted as true positives (TPs), while minimizing the number of wrong predictions denoted as false positives (FPs). Ultimately, we aim at maximizing the relabel precision $TP/(TP + FP)$.

Since our method is iterative, INERRANT^0 is the first step which consists in learning on the partially labeled data without relabeling, and INERRANT corresponds to the method after training the model on the incorporated pseudo-labels.

3.3.1 Learning on a partially labeled dataset

We address the issue of learning on partially labeled data by a simple yet effective method, which is shown in Figure 3.2. The first step consists in extracting the maximum of information from the partially labeled data, by deducing from the labeled organs where there are ambiguities that should be handled.

Training exclusively with correct labels We know by construction that if an organ is unlabeled, then it is the case for the entire volume, *i.e.* no intermediate slice contains this label. Thus, we can deduce beforehand the missing classes for every patient. However, we do not know where they are located and thus where the wrong labels are.

However, if we want to exclusively use correct labels, we cannot use a classic softmax activation function and a multiclass loss. Indeed, in that configuration when only one organ is missing no background label can be used. To address this problem, we transform the $(K + 1)$ multiclass classification problem into K binary classification problems where each organ is learned independently. The rationale behind this is to control the classes that are labeled and can be learned and those that are unlabeled and have to be ignored. By doing that we can learn features from the labeled classes for both the positives (the organs) and the negatives (the background) whereas for the unlabeled classes, both positives and negatives are ignored.

In practice, we replace the final softmax by a sigmoid activation function in the last 1×1 convolution layer. However, we still want to keep the exclusive aspect of the softmax, *i.e.* only one class is predicted for a given voxel. Thus, our class prediction is computed by taking, for each voxel, the class with the highest probability among all K classes - and the background label is assigned if all probabilities are lower than 0.5.

Training K binary classifiers requires adjustments, especially on the loss function. Actually, we have K losses, one for each class. We choose the binary cross entropy to train our model

defined in Equation 3.1 for each voxel i and class k:

$$l_{i,k}(\hat{y}_{i,k}, y_{i,k}) = -(y_{i,k}\ log(\hat{y}_{i,k}) + (1 - y_{i,k})\ log(1 - \hat{y}_{i,k})) \tag{3.1}$$

Let us denote as $\hat{\mathbf{Y}} \in \mathbb{R}^{H,W,K}$ the dense prediction of our model and $\mathbf{Y} \in \mathbb{R}^{H,W,K}$ as the ground truth. Then the K losses are aggregated to obtain one final loss in Equation 3.2:

$$\mathcal{L}_{ce}(\hat{\mathbf{Y}}, \mathbf{Y}) = \sum_{k=1}^{K} \sum_{i=1}^{N} w_{i,k}\ l_{i,k}(\hat{y}_{i,k}, y_{i,k}) \tag{3.2}$$

where $\mathbf{W} \in \mathbb{R}^{H,W,K}$ composed of K maps $\mathbf{w_k} \in \mathbb{R}^{H,W}$, is a binary matrix which selects or discards the voxels that should be learned for class k, for which back-propagation is applied.

$\mathbf{W}$ is an ambiguity map since it represents the pixels' location where we cannot decide if the label is correct or not. $\mathbf{W}$ is built beforehand based on the missing organs of each patient. As shown in Figure 3.2, if an organ is labeled, we fill $\mathbf{w}_k$ with ones to learn the associated model. On the other hand, when an organ is missing $\mathbf{w}_k$ is set to zeros to ignore this organ during training. However, we can still use extra information from other organs, which are assigned as negative labels.

In the example of Figure 3.2, three organs are labeled. However, when learning a missing organ like the spleen (bottom branch), we use an ambiguity map containing zeros everywhere except where the other organs. In that case the label of the organ is used to fill the ambiguity map of the spleen with ones.

Statistical analysis of the training labels To quantify the quality of the labels used during training, let us consider the binary classification problem for the k^{th} organ class. We denote as β_k the number of pixels for this organ and α the ratio of missing organs on the whole dataset. Table 3.1 shows confusion matrices for two different methods: *naive* consists in learning directly with the partial labels, and our method *INERRANT*^{0}. We can see that the naive method has $\alpha \cdot \beta_k$ FNs. Meanwhile, INERRANT^0 completely discards FNs but also reduces the number of TNs.

The naive approach learns with $(1 - \beta_k)$ TNs whereas INERRANT^0 learns with $(1-\alpha)(1-\beta_k) + \epsilon$ TNs, where $\epsilon = \sum_{k' \neq k} \beta_{k'}$ corresponds to the other organ labels. In medical image segmentation, organs represent usually a small proportion of the total volume of labels, which induces a high class imbalance between positives and negatives, such that $\beta << 1$, *e.g.* $\beta = 0.05$. As a consequence, we still have enough information to properly learn the background class with *INERRANT*^{0}.

Table 3.1: TP/ FP training label analysis

(a) Naive

GT \ Used	Pos	Neg
Pos	$(1-\alpha)\cdot\beta_k$	$\alpha\cdot\beta_k$
Neg	0	$1-\beta_k$

(b) INERRANT^0

GT \ Used	Pos	Neg
Pos	$(1-\alpha)\cdot\beta_k$	0
Neg	0	$(1-\alpha)\cdot(1-\beta_k)+\epsilon$

3.3.2 Self-supervision and pseudo-labeling

The number of TPs linearly decreases with the ratio of missing organs α. To recover missing labels in training images, we propose to iteratively add new positive labels $y_{i,t}=1$ in an image with missing labels $\mathbf{x}_i$ for each class k[1], using a curriculum strategy [12].

Iterative relabeling Initially, the model is trained on all correct labels that can be regarded as "easy positive samples". Let us denote as $\hat{y}_i^+$, the pixels predicted as positive for a given unlabeled image $\mathbf{x}_i$. The idea of INERRANT is to recover positive labels, $y_{i,t}^+$ by selecting the top scoring pixels among $\hat{y}_i^+$. Then, the model is retrained with the new labels added to the training set.

This procedure is iteratively performed T times, by selecting a ratio $\gamma_t = \frac{t}{T}\gamma_{max}$ of top scoring pixels among the positives. The pseudo-labels incorporated at each step are the "hard examples" since they come from a pseudo-labeling scheme that could introduce errors.

Uncertainty estimate for collecting pseudo-labels Our pseudo-labeling approach in Algorithm 1 is based on selecting the most confident pixels of the segmentation model. We therefore seek an accurate confidence criterion for our deep FCN in semantic segmentation.

Measuring model uncertainty in deep learning is an open and difficult problem, as detailed in Section 3.2. Although the Maximum Class Probability (MCP [119]) gives decent performances in practice, it also suffers from important conceptual limitations (see Section 3.2). Especially, misclassified pixels (failures) receive an unjustified high confidence. In our pseudo-labeling approach, this presents the risk of including wrong labels and negatively impacting performances.

[1]We drop the dependence of class in $y_{i,t}$ for clarity.

Algorithm 1: Training INERRANT for class k

Data: $\{(x_i, y_i)\}$, γ_{max}, T, m_0
Result: m_T
$N_u \leftarrow$ number of unlabeled images;
$y_{i,0} = y_i$;
for $t \leftarrow 1$ *to* T **do**
  $\gamma_t = \frac{t}{T}\gamma_{max}$;
  for $i = 1$ *to* N_u **do**
    $\hat{y}_i^+ \leftarrow m_t(x_i)$ // Take predicted $\oplus$;
    $y_{i,t}^+ \leftarrow s(\hat{y}_i^+, \gamma_t)$ // Assign new $\oplus$ target labels;
    $y_{i,t} = y_{i,t-1} \cup y_{i,t}^+$ // Augment training set;
  $m_t = train(\{(x_i, y_{i,t})\})$ // Re-train model

Therefore, we propose to use a more relevant uncertainty measure. Our target confidence criterion is the True Class Probability (TCP [123]), from which guarantees can be derived for discriminating correct from incorrect predictions (TCP is able to assign small confidence values to misclassifications). Since TCP requires the knowledge of the ground truth class for each pixel, which is not accessible at test time, we need an auxiliary network specifically dedicated to predict the TCP value computed by our segmentation model, e.g. U-Net. For each pixel $i \in \{1, ..., N\}$ and each class $k \in \{1, ..., K\}$, we want the predicted confidence $\hat{c}_{i,k}$ to match $TCP_{i,k} = c_{i,k}$: learning the confidence network is a regression task where we use the following L2 loss:

$$\mathcal{L}_{conf} = \frac{1}{K}\frac{1}{N}\sum_{k}^{K}\sum_{i}^{N}(\hat{c}_{i,k} - c_{i,k})^2 \tag{3.3}$$

The confidence network is illustrated in Figure 3.3. It is attached to the segmentation model in order to leverage latent representations learned for the segmentation task. In practice, we connect it to the antepenultimate layer, i.e. before the final 1×1 convolutional layer.

The confidence network is thus initialized with parameters from the segmentation model. During training, we can freeze these parameters or fine-tune them, which we find superior in practice. If the entire model is fine-tuned, a duplicate of the original FCN allows to keep the same segmentation predictions.

The confidence network is trained before relabeling and after training the segmentation network. Algorithm 2 shows the different steps of training our model by using pseudo-labels generated iteratively.

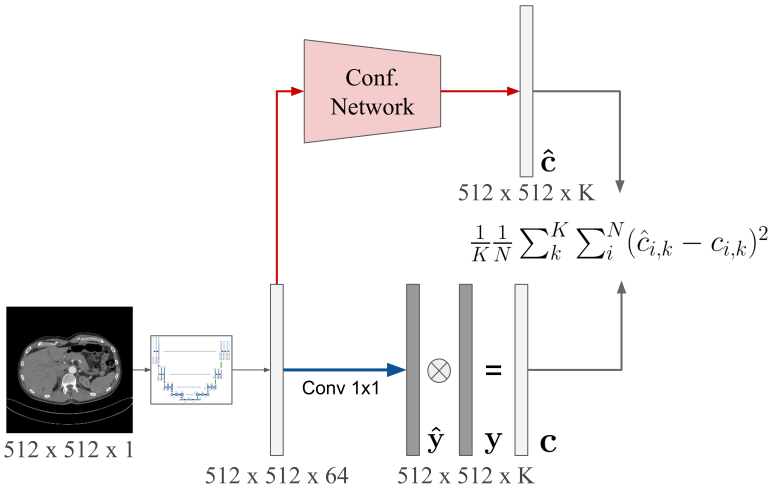


Figure 3.3: The confidence network part is included at the end of the segmentation network by taking the features before the final 1×1 convolutional layer.

Algorithm 2: Relabeling the missing organs with the confidence network.

Train the FCN on partially labeled data;
for $t \leftarrow 1$ *to* T **do**
    Train the confidence network;
    Relabel the $\frac{t}{T}\gamma_{max}$ pixels with the highest confidence score;
    Fine-tune the initial FCN with the new labels;

3.4 Experiments and Results

3.4.1 Experimental setup

For this work we test our proposed method on three abdominal organ segmentation datasets. It includes LiTS, TCIA and the IMO datasets, all three described in Section 2.2.4.

Simulating partially labeled datasets Large datasets for organ segmentation are tedious and expensive to obtain. Depending on the medical center and the patient's pathology only some organs are labeled for a given case. Consequently, it is easier to gather data with heterogeneous labels but the resulting dataset will be partially labeled. To reproduce this context and analyse how the model performs under different amounts of missing labels, we start from fully labeled datasets and randomly remove the labels at a volume level. Thus, we reproduce real clinical conditions and keep control over the exact quantity of available information. Moreover we can evaluate the method on a completely labeled test set. The proportion of labeled organs in each

Table 3.2: Quantitative results for the TCIA pancreas dataset. The scores are the mean DSC (± std) for every missing label proportion (α). In bold the highest results that pass a t-test with p-value < 0.05 compared to the other methods.

Proportion (α)	100%	70%	50%	30%	10%
Naive	76.13 (± 0.94)	49.75 (± 5.58)	28.99 (± 6.07)	10.75 (± 5.71)	1.16 (± 0.77)
INERRANT^0	-	72.12 (± 2.01)	70.43 (± 3.38)	64.48 (± 2.13)	44.57 (± 5.24)
INERRANT	-	**75.52** (± 1.74)	**74.23** (± 2.50)	**71.10** (± 1.52)	**56.19** (± 6.22)
INERRANT^0 3D	78.76 (± 1.91)	77.22 (± 2.41)	75.59 (± 1.69)	71.73 (± 1.93)	52.98 (± 8.83)
INERRANT 3D	-	**77.35** (± 1.67)	**76.02** (± 0.88)	**73.41** (± 1.00)	**57.77** (± 7.53)

Table 3.3: Quantitative results for the LiTS dataset. The scores are the mean DSC (± std) for every missing label's proportions (α). In bold the highest results that pass a t-test with p-value < 0.05 compared to the other methods.

Proportion (α)	100%	30%	10%	5%	1%
Naive	94.72 (± 1.22)	14.10 (± 6.28)	0.41 (± 0.18)	1.14 (± 2.47)	0.31 (± 0.53)
INERRANT^0	-	93.12 (± 1.41)	89.70 (± 2.51)	88.22 (± 2.87)	51.08 (± 13.80)
INERRANT	-	**93.51** (± 1.15)	**90.05** (± 1.41)	**88.88** (± 2.48)	**58.76** (± 10.94)

Table 3.4: Quantitative results on IMO. The scores are the mean DSC (± std) for every missing label proportion (α). In bold the highest results that pass a t-test with p-value < 0.05 compared to the other methods.

Proportion (α)	100%	70%	50%	30%	10%
Naive	86.03 (± 2.16)	66.85 (± 4.89)	45.32 (± 2.67)	19.51 (± 2.39)	2.82 (± 1.30)
INERRANT^0	-	84.19 (± 2.85)	81.25 (± 5.51)	76.58 (± 7.15)	67.69 (± 5.34)
INERRANT	-	**85.36** (± 2.70)	**84.43** (± 3.56)	**82.60** (± 3.40)	**73.49** (± 3.08)

volume is denoted as α. When $\alpha = 100\%$, all the organs are labeled and when $\alpha = 0\%$ no label is available in each volume.

For the IMO dataset, the label proportion α is applied to every organ, independently. It means that $\alpha\%$ of the cases have a labeled liver, $\alpha\%$ a labeled spleen, etc. Thus, a case could have between 0 and 7 labeled organs. Moreover, we paid attention to incrementally remove labels. The same labeled organs are found through the different proportions, i.e. with $\alpha = 70\%$, the dataset contains all the labels of a dataset with $\alpha = 50\%$ but with more of them. In the labels point of view, we can say that $D(10\%) \subset D(30\%) \subset D(50\%) \subset D(70\%)$. This allows fair comparisons between the different proportions as they are trained with the same labeled images.

Implementation details We use a U-Net as our main FCN which is well-known for 2D medical image segmentation. This model is still extensively used as it gives competitive results though it requires reasonable memory cost and can be trained on standard GPUs.

The standard U-Net used in our experiments is around 31M parameters. The confidence network only adds 0.8M parameters but this network is only used for the relabeling step and is discarded for the final prediction network which is simply the U-Net, thus our method does not add any computational nor memory overhead compared to the baseline in test (See Table A.1 for a detailed overview of the network used in the study including the layers' parameters). The models are trained with the Adam optimizer and an initial learning rate of 10^{-4} which exponentially decreases to 10^{-5} at the end of the training. Standard data augmentation techniques are used including random translations, random rotations and random scales. The models are implemented with the Tensorflow library and the training is performed on RTX 2080Ti GPUs. We perform 5 fold cross-validation for every dataset and proportion. The results shown in section 3.4.2 give the mean Dice Similarity Coefficient (DSC) and standard deviation across the folds.

The overall quantitative evaluation carried out in section 3.4.2 gives the results for the naive baseline, *i.e.* when the model is trained directly on the data. Then with the proposed INERRANT^0. And finally with INERRANT that iteratively adds pseudo-labels using the previously introduced confidence network. Then, a finer analysis of the impact of curriculum iterations and confidence measures is provided in section 3.4.3.

3.4.2 Quantitative results

To highlight the problem of training on partially labeled data, we evaluate the naive approach which consists in learning on the partially labeled data with the background label

assigned to missing pixel labels. Then, we show the results using our method, first with the ambiguity map only (INERRANT^0) and then using pseudo-labels (INERRANT).

TCIA pancreas Results for the TCIA pancreas dataset are given in Table 3.2. As we can see, the naive approach quickly deteriorates when the number of missing labels increases, i.e. α decreases. For example, with $\alpha = 70\%$, we already observe a drop of about 26.4pts in DSC. By assigning the background label to missing organ labels, this naive baseline makes the model trained with many wrong labels of an already over-represented class. So, it naturally tends to predict "background" for the entire image.

INERRANT^0 gives better results as the model is trained only on correct labels. We can see that even with $\alpha = 30\%$, which is less than the third of the labels, we lose 11.6pts when the naive baseline is at less than 11% in DSC.

Next, INERRANT which introduces pseudo-labels helps to improve the mean DSC for every proportion. We can even see that the gain increases when α decreases. At $\alpha = 10\%$ INERRANT has improved the results by 10pts. The gains are significant and show the relevance of the proposed method and how using pseudo-labels can improve the final scores.

Finally, Table 3.2 also reports results with a 3D backbone. In INERRANT^0 3D and INERRANT 3D we replace the 2D U-Net with its 3D counterpart to show that our method is agnostic to the chosen backbone FCN. In this setup, we have an input patch size of $144 \times 144 \times 96$ which is cropped in the center of the image. This could also explain the performance boost compared to the 2D U-Net, however this method could not be applied to the multi-organ setup where one should perform predictions with, for example, a sliding window. Nevertheless, the same trends are observable: the relabeling step INERRANT 3D outperformed INERRANT^0 3D for every proportion and the highest gain is at $\alpha = 10\%$ with +4.79pts.

LiTS Contrary to the pancreas, the liver is easier to segment, since it is one of the largest organs in the abdomen, leading to more pixel labels. In addition, its boundaries are less ambiguous.

Table 3.3 shows the results on the LiTS dataset. We observe that the performance of the baseline U-Net for $\alpha = 100\%$ is high, i.e. more than 94% DSC. It is worth mentioning that for $\alpha = 30\%$, the naive baseline already gives terrible results.

The interesting point here is the fact that INERRANT^0 gives very high results even with very few examples. As we can see the result with $\alpha = 5\%$ loses only 6.5pts compared to the model trained on 100% of data. In this dataset, $\alpha = 5\%$ corresponds to only 5 labeled cases which correspond to a reduction in labels by a factor of 20.

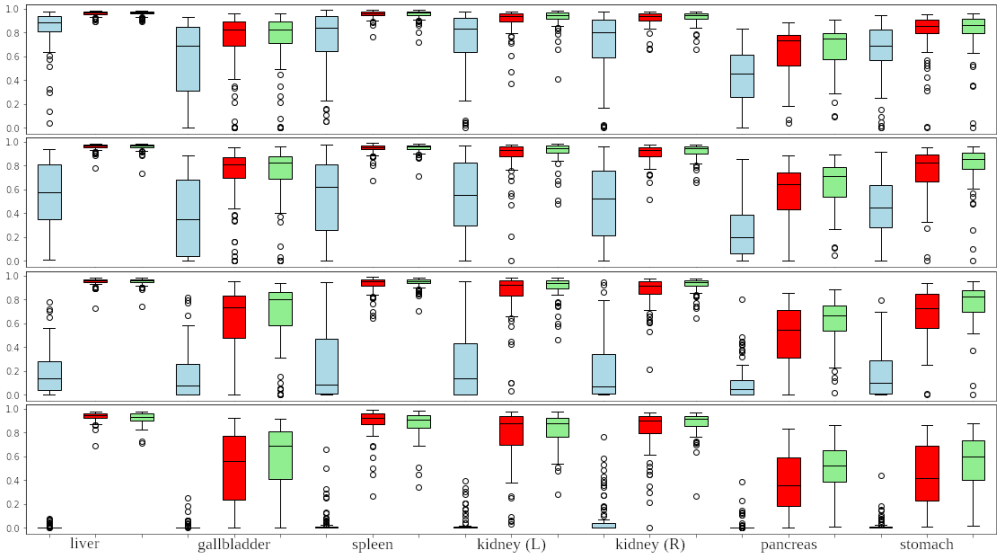


Figure 3.4: Per patient DSC scores analysis for the IMO dataset. First row with $\alpha = 70\%$, second $\alpha = 50\%$, third $\alpha = 30\%$ and fourth $\alpha = 10\%$. In blue the naive method, red INERRANT^0 and green INERRANT with pseudo-labeling.

Moreover, the relabeling step helps to consistently improve the results. The most important gain is again with the lower α (*i.e.* $\alpha = 1\%$) with a difference of 7.7pts.

With this dataset, the overall conclusion is similar to TCIA, but the regime is different. As described above, the scores of our approach without relabeling are very high even with few labels. However, introducing pseudo-labels still improves the model, with the largest gain at $\alpha = 1\%$.

Table 3.5: State-of-the-art comparison on the TCIA pancreas dataset

Proportion (α)	70%	50%	30%	10%
Naive	49.75 ($\pm$ 5.58)	28.99 ($\pm$ 6.07)	10.75 ($\pm$ 5.71)	1.16 ($\pm$ 0.77)
INERRANT^0	72.12 ($\pm$ 2.01)	70.43 ($\pm$ 3.38)	64.48 ($\pm$ 2.13)	44.57 ($\pm$ 5.24)
Pseudo-labels ([80])	**75.12** ($\pm$ 1.91)	**73.71** ($\pm$ 2.59)	69.00 ($\pm$ 2.04)	51.91 ($\pm$ 7.77)
Adversarial ([5])	**75.41** ($\pm$ 1.78)	**73.91** ($\pm$ 2.27)	67.60 ($\pm$ 1.84)	52.09 ($\pm$ 6.00)
Consistency ([75])	**74.53** ($\pm$ 2.10)	**72.68** ($\pm$ 3.05)	66.99 ($\pm$ 1.38)	46.04 ($\pm$ 3.70)
INERRANT (Ours)	**75.52** ($\pm$ 1.74)	**74.23** ($\pm$ 2.50)	**71.10** ($\pm$ 1.52)	**56.19** ($\pm$ 6.22)

IMO dataset Figure 3.4 shows the results on this dataset detailed per organ and Table 3.4 the average DSC for all proportions and the different methods (scores per organ are detailed in Table 3.6). As we can see, all the methods give better results compared to LiTS and TCIA.

This can be explained by two important points. Firstly, the background class is less represented because we have multiple organs. Secondly, considering INERRANT, for one particular case only 1 or 2 organs could be unlabeled especially for high proportions like 70%. It implies that a lot of background labels could be correctly learned even without the organ label thanks to the other organs. This shows that our method is actually strengthened in the case of multi-organ with missing labels. We can see in Figure 3.2 an example of an ambiguity map for a missing organ (bottom branch) in a case where some labels are available. We can notice that a wide part of the image can be used to learn a negative label where all the other organs are located.

Considering the naive baseline, as for the two previous datasets, the scores quickly fall until reaching a very low value of 2.8% when $\alpha = 10\%$. For our method, however, it gives good performances even with few labels. But depending on the organ, the behavior is different. The liver, spleen and kidneys, stay with high scores even with few labels with an impressive result for the liver that only loses 3pts between 100% and 10%.

On the other hand, the gallbladder, the pancreas and the stomach fall more quickly than the other organs. Those organs are the smallest and in general more difficult to segment. For instance with the gallbladder, the segmentation model tends to segment it as the liver because it is located close to it in addition to being very small. The pancreas is also difficult to segment due to its complex boundaries and pixel intensities which are very close to the connected structures. Finally, the stomach is difficult to segment because of its shape, size and position variability in addition to the presence of air which makes holes in the structure that add randomness about the organ visibility.

INERRANT^0 gives 48.96% for the gallbladder, 37.04% for the pancreas and 44.05% for the stomach at $\alpha = 10\%$. But after adding the pseudo-labels the most important gains are with those 3 organs. Respectively, 57.93% (+9pts), 50.25% (+13.2pts) and 56.03% (+12pts).

The curriculum learning approach combined with the learned confidence boosts the results for every organ and every proportion. The most impressive gains are for the most difficult organs which are the gallbladder, pancreas and stomach. It could be explained by the fact that those organs need more labels due to their complexity, and we show that our pseudo-labeling approach greatly helps to comply with this requirement.

State-of-the-art comparison We compare INERRANT with three other semi-supervised methods representing three different types of approaches. Firstly, [80] which consists in using all the predictions as pseudo-labels. Then, [5] which is an adversarial training where the output of the discriminator allows to select pseudo-labels on the fly by adding them to the segmentation loss during training. And finally [75] which is a mean teacher model based on [72] that uses unlabeled data through a consistency loss.

We implemented the above-mentioned methods with the same backbone FCN (*i.e.* 2D U-Net) to segment the pancreas from the TCIA segmentation dataset. Each experiment is evaluated with a 5-fold cross-validation. The models are trained with the same procedure, *i.e.* with the same dataset, the same folds, the same missing labels and with the appropriate hyper-parameters. Table 3.5 shows the results for every approach compared to the baselines that didn't use unlabeled images.

With [80] all the predictions for the missing organs are used as pseudo-labels. It gives better results than INERRANT^0 as it injects more information with the correct pseudo-labels. However, though it performs well with high α values, it tends to add a lot of wrong labels with low α values which reduces the gains. We can see that using a better pseudo-label selection scheme, we can prevent this effect while preserving the performances with high α values.

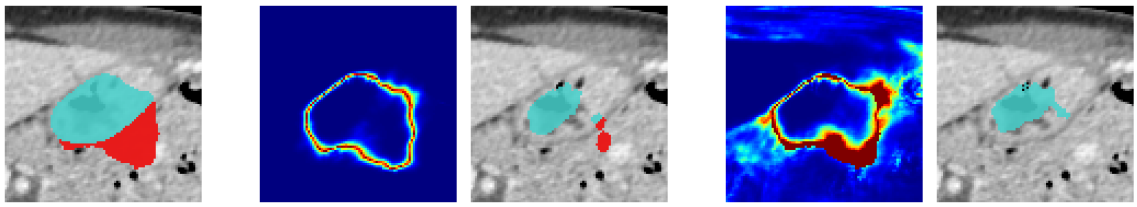
Concerning the adversarial training [5], the method gives comparable results to [80]. We can see that the results are better than INERRANT^0 but INERRANT still outperformed it for every proportion. This model can leverage a meaningful loss applicable to unlabeled data, but is hard to train due to instabilities in the adversarial approach.

The consistency method based on mean teacher [75] still improves over INERRANT^0, but is not the best performing strategy for handling unlabeled data in our context. For $\alpha = 10\%$, the performance drop is significant compared to the other approaches. It can be explained by the fact that the loss function does not explicitly exploit predicted segmentation masks on unlabeled data.

In all cases, we can see that INERRANT performs better than the other methods, with a gain being more pronounced for low α proportions.

Table 3.6: Results on the IMO dataset detailed per organ

Method	Liver	Gallbladder	Spleen	Kidney (L)	Kidney (R)	Pancreas	Stomach
70%							
Naive	83.37 (± 5.64)	58.36 (± 5.22)	74.60 (± 12.05)	72.64 (± 13.93)	71.11 (± 12.60)	43.26 (± 4.76)	64.63 (± 6.45)
INERRANT^0	96.14 (± 0.45)	72.25 (± 8.90)	95.31 (± 0.70)	90.33 (± 2.97)	91.83 (± 2.07)	64.06 (± 6.16)	79.42 (± 8.30)
INERRANT	**96.22** (± 0.50)	**72.95** (± 9.91)	**95.37** (± 1.12)	**92.51** (± 1.95)	**92.69** (± 1.49)	**67.25** (± 4.32)	**80.57** (± 8.23)
50%							
Naive	57.23 (± 9.00)	35.87 (± 10.13)	54.90 (± 12.90)	52.61 (± 13.05)	48.66 (± 6.89)	24.36 (± 6.27)	43.65 (± 5.36)
INERRANT^0	95.81 (± 0.62)	70.09 (± 9.77)	94.27 (± 0.84)	87.76 (± 5.83)	90.16 (± 3.33)	55.59 (± 16.37)	75.05 (± 9.56)
INERRANT	**95.93** (± 0.79)	**72.75** (± 9.54)	**94.99** (± 1.17)	**91.59** (± 3.26)	**92.14** (± 1.75)	**64.15** (± 8.23)	**79.49** (± 8.80)
30%							
Naive	19.07 (± 4.66)	15.51 (± 3.53)	27.48 (± 9.63)	24.95 (± 9.08)	21.36 (± 10.75)	10.26 (± 4.35)	17.95 (± 2.75)
INERRANT^0	95.34 (± 0.79)	60.75 (± 15.37)	92.56 (± 1.91)	84.53 (± 7.69)	86.81 (± 5.66)	48.78 (± 13.83)	67.26 (± 9.01)
INERRANT	**95.38** (± 0.81)	**67.23** (± 11.34)	**94.57** (± 0.97)	**90.69** (± 1.89)	**92.09** (± 1.58)	**61.99** (± 7.10)	**76.25** (± 5.67)
10%							
Naive	0.56 (± 0.58)	1.12 (± 0.64)	4.03 (± 3.28)	3.41 (± 1.49)	7.03 (± 9.12)	1.62 (± 1.37)	1.99 (± 1.25)
INERRANT^0	**93.56** (± 1.07)	48.96 (± 10.03)	**89.41** (± 2.82)	78.00 (± 14.13)	82.84 (± 9.87)	37.04 (± 5.87)	44.05 (± 11.67)
INERRANT	92.45 (± 1.35)	**57.93** (± 11.59)	87.20 (± 3.87)	**82.12** (± 7.06)	**88.46** (± 3.18)	**50.25** (± 3.63)	**56.03** (± 9.57)



(a) Prediction (b) MCP (c) Learned Conf.

Figure 3.5: Confidence maps for MCP and our confidence network for the stomach. The prediction in (a) gives the TPs in cyan and FPs in red. For both MCP (b) and the learned confidence (c), a confidence map is given with values between 0.5 (red) and 1.0 (blue) and the selected pseudo-labels with the TPs in cyan and the FPs in red. In (b), MCP gives low confidence only at the boundaries. As a contrary in (c) the confidence network gives low confidence values to the model errors and thus prevents relabeling wrong predictions.

3.4.3 Model analysis

This section aims to provide an analysis of the relabeling. First, we discuss the differences between the uncertainty evaluation methods and how they impact the relabeling and final score. Then, we show the impact of the curriculum learning and how it behaves depending on the number of performed relabeling steps.

Uncertainty methods evaluation We evaluate the performances of the proposed confidence network as described in section 3.3.2, and compared it with MCP, which corresponds to the previous work of SMILE [1], on the TCIA pancreas dataset.

The confidence network can be trained with two different configurations, by transfer learning: the U-Net weights are frozen during the confidence training, or by fine-tuning: the U-Net and the confidence network are trained together. For the last configuration, it is necessary to duplicate the U-Net part of the model which adds complexity to the final model. However, we found in practice that fine-tuning gives better results, thus the following results are obtained with this method.

A detailed analysis of the impact of the two uncertainty estimation methods is provided in Table 3.7 on the TCIA pancreas dataset (Additional results for the IMO dataset could be find in Table 3.10). We evaluate how the confidence score ranks the pixels considered for relabel (We relabel only the positives and never the background). Three metrics are shown: the AUC (area under the ROC curve), the Average Precision of the success (AP_success), and the AP of the errors (AP_error). The first metric gives a measure of the overall ranking of the predictions. The second, measures the method's capacity of assigning high values to the correct predictions. Finally, the AP error gives a measure of the method's capacity of assigning low values to the

Table 3.7: Analysis of ranking metrics for uncertainty estimation with MCP, equivalent to SMILE [1], and the learned confidence method. The metrics are computed only on the pixels that are considered for relabeling, *i.e.* predicted as positive and not already relabeled. The values are percentages.

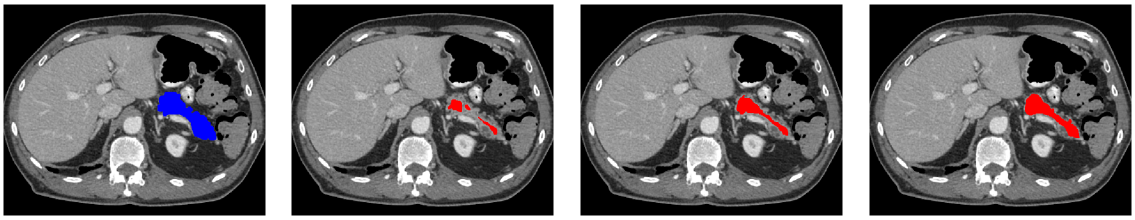
Method	AUC	AP_success	AP_error	Final DSC
70%				
MCP	73.86 (± 1.02)	92.00 (± 0.71)	34.99 (± 2.34)	73.97 (± 1.28)
L. conf.	**75.50** (± 1.77)	**92.66** (± 0.83)	**38.17** (± 3.86)	**75.52** (± 1.74)
50%				
MCP	72.67 (± 1.05)	90.51 (± 1.97)	36.50 (± 3.06)	73.82 (± 2.15)
L. conf.	**73.94** (± 0.94)	**91.06** (± 1.60)	**38.69** (± 4.55)	**74.23** (± 2.50)
30%				
MCP	71.55 (± 1.95)	90.58 (± 1.43)	34.29 (± 2.68)	69.72 (± 1.75)
L. conf.	**73.06** (± 2.00)	**91.25** (± 1.24)	**36.80** (± 4.11)	**71.10** (± 1.52)
10%				
MCP	68.68 (± 2.28)	84.97 (± 3.91)	41.11 (± 4.85)	54.66 (± 6.53)
L. conf.	**70.21** (± 3.46)	**85.72** (± 4.09)	**43.76** (± 7.70)	**56.19** (± 6.22)

wrong predictions.

Table 3.7 shows significant improvements for all the metrics and for the different proportions. At 10%, the relative gain is the most important. We observe an improvement of 1.53pts in AUC, 0.75pt in AP_success and 2.65pts in AP_error. It means that we have a better ranking of the candidates in addition to a better error detection which translates into an improvement of the final DSC after training the model on the pseudo-labels of 1.5pts. At this proportion, the absolute gain is equivalent to the one at 70% but the relative gain is higher in the way that it will impact much more the final results.

Qualitatively, Figure 3.5 shows uncertainty maps for both methods. We can notice that the learned confidence has a more detailed result than MCP. In fact, MCP concentrates the low confidence values at the border whereas the confidence network assigns lower confidence values to the model errors. In this example a part of the segmentation, at the bottom right, is wrong and we can see that the confidence network has assigned lower values at this place than MCP. This illustrates how our confidence network helps to prevent the relabeling of wrong predictions and thus the incorporation of errors in the training set.

Curriculum learning analysis Curriculum learning consists in introducing easy examples before adding more complex ones. In our application, the easy examples are the available labels and the more complex, the pseudo-labels which contain wrong labels. The pseudo-labels are introduced incrementally by first taking the most confident predictions and ending by the less



(a) Ground truth (b) $t = 0$ (c) $t = 1$ (d) $t = 2$

Figure 3.6: Complete relabeling of a pancreas with INERRANT, $T = 3$ iterations, $\gamma_{max} = 1.0$ and $\alpha = 50\%$.

Table 3.8: Complete organ relabel detailed for 3 steps on the IMO dataset. Information given are the percentage of added pixels, the relabeling precision and recall and the final DSC after training on the updated dataset. Values are percentages.

α/step	Added pixels	Relal. P	Relab. R	Final DSC
50%				
0	0%	-	-	79.81
1	33%	98.14	19.99	84.93
2	66%	95.11	25.89	85.15
3	100%	89.04	25.87	**85.53**
30%				
0	0%	-	-	72.27
1	33%	96.62	18.37	82.17
2	66%	93.62	25.58	**83.79**
3	100%	85.20	25.81	83.30
10%				
0	0%	-	-	58.98
1	33%	94.44	15.57	72.89
2	66%	81.80	23.70	**74.26**
3	100%	65.09	23.96	72.01

confident that would by definition contain more wrong labels.

As we can see in Figure 3.6, using an iterative approach allows us to relabel progressively the missing organ from the center to the border. In fact, we noticed that the most certain predictions were located in the center and that the confidence decreases as we move closer to the border (see Figure 3.5).

Table 3.8 presents a quantitative evaluation of the iterative relabeling, and shows the relabeling precision and recall for each step of the curriculum learning method with the final DSC after fine-tuning the model on them. Overall, using $T = 2$ relabeling iterations is the best strategy, although differences can be observed for different levels of missing labels α. For $\alpha > 50\%$,

the relabeling precision is higher and thus the best results are with the last step. On the other hand, with $\alpha < 30\%$, the best results are for an intermediate step because performing the last iterations adds too many wrong predictions and thus deteriorates the model performances. However, it is worth noting that for every proportion the relabeling improves the final score.

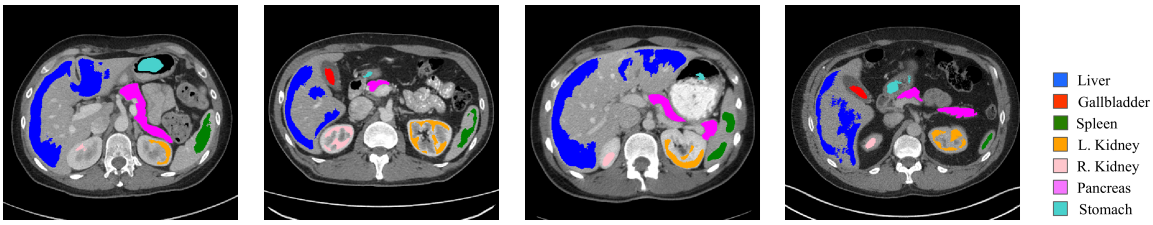
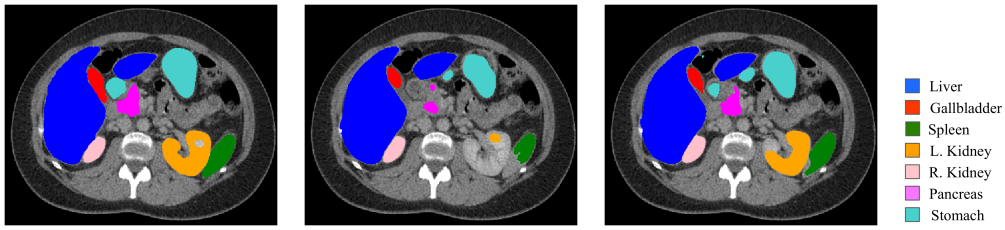


Figure 3.7: Relabeling of TCIA images with a model trained with only 9 completely labeled images from the IMO dataset.



(a) Ground Truth (b) INERRANT^0 (c) INERRANT

Figure 3.8: Segmentation results for INERRANT^0 and INERRANT, $\alpha = 30\%$.

Method	Multi-organ	TCIA	Liver	Gallbladder	Spleen	Kidney (L)	Kidney (R)	Pancreas	Stomach	Avg.
INERRANT^0	9	0	89.43	48.09	84.72	78.39	80.78	32.55	48.13	66.02
INERRANT	9	82	89.85	57.63	87.46	85.22	85.33	58.15	58.85	74.64

Table 3.9: Results in DSC (%) when combining 9 completely labeled examples from the IMO dataset with the 82 partially labeled examples (only the pancreas) of the TCIA dataset with INERRANT. The models are evaluated on the remaining 81 multi-organ examples.

Qualitative results To illustrate the previous results, Figure 3.8 shows an example of a segmentation result for the IMO dataset for INERRANT^0 and after training on the pseudo-labels, INERRANT. We can notice that INERRANT helps by segmenting more pixels and thus fill organs that have been missed by INERRANT^0.

Figure 3.6 is an example of a complete relabel of a missing pancreas. It illustrates how the method progressively adds more pixels from the most certain (in the center) to the least certain (at the border).

3.4.4 Fusion of heterogeneous data from multiple datasets

Completely labeled data for abdominal organ segmentation are expensive and tedious to obtain. In this experiment, we show that with INERRANT we can build a good segmentation model by starting with few completely labeled examples and leveraging public datasets with few labeled organs. Thus, we use 9 cases from the IMO dataset with the 7 organs completely labeled and add the 82 cases from the TCIA datasets which are partially-labeled compared to the multi-organ cases (only the pancreas is available). Then, we evaluate the remaining 81 multi-organ examples. In Table 3.9, we evaluate a model trained only on the completely labeled 9 cases. Then we add the 82 cases from TCIA and follow the INERRANT method. We can see a large improvement for every organ, especially for the small ones, *i.e.* the gallbladder, +9.5pts, the stomach, +10.5pts and obviously the pancreas, +25.6pts. It is worth noting that even if both datasets are abdominal CT-scans, there is a slight domain shift. In fact, if we have two different sources that acquire data under different parameters, then the quality of the annotations could be very different. For instance in TCIA we can assume that the pancreas' annotations are more precise as they focus on this very organ. A qualitative evaluation is provided in Figure 3.7. It shows how we relabel the TCIA examples based on a model trained only on 9 completely labeled images.

This experiment points out that even with a little domain shift we can build a better model by enriching a small dataset with external sources of images.

3.5 Conclusion

In this chapter, we studied the challenging problem of learning with partial labels. To address that issue, we proposed INERRANT, a method based on a specifically designed loss for ignoring ambiguous labels coupled with an iterative pseudo-labeling scheme. Moreover, we introduce a confidence network that learns an uncertainty criterion leveraged by the relabeling process which iteratively adds new labels to the training set. In our experiments we show very good results on three abdominal organ segmentation datasets. Moreover, we observed that our method is even more relevant and efficient with low label proportions.

We show the good performances obtained by INERRANT compared to state-of-the-art semi-supervised methods. Last but not least, we provide a showcase illustrating INERRANT's capacity to combine real datasets with different labeling and how it improves segmentation performances.

In the next chapter, Chapter 4, we study how we could incorporate prior knowledge about the absolute position of organs. This spatial prior gives the probability of presence of an organ

Table 3.10: Analysis of ranking metrics for uncertainty estimation with MCP and the learned confidence method. Results are given per organ for the IMO dataset in average across the folds.

Method		Liver	Gallbladder	Spleen	Kidney (L)	Kidney (R)	Pancreas	Stomach
70%								
	AUC	86.09	79.37	91.89	83.69	86.46	75.55	75.70
MCP	AP_success	99.15	97.26	99.48	98.03	98.81	94.87	95.25
	AP_error	27.21	23.63	**29.17**	30.80	29.96	28.13	25.32
	AUC	**89.99**	**81.83**	**92.78**	**87.41**	**88.18**	**77.97**	**81.82**
Learned conf.	AP_success	**99.43**	**97.72**	**99.67**	**98.35**	**98.91**	**95.53**	**96.22**
	AP_error	**33.89**	**29.22**	26.69	**33.31**	**32.58**	**32.26**	**37.65**
50%								
	AUC	87.55	82.93	93.61	78.26	83.84	72.66	71.82
MCP	AP_success	99.24	97.91	99.81	97.36	98.59	93.84	94.69
	AP_error	29.20	27.70	23.64	24.80	24.69	27.35	22.33
	AUC	**91.11**	**85.76**	**94.38**	**83.96**	**85.59**	**77.38**	**81.24**
Learned conf.	AP_success	**99.49**	**98.37**	**99.87**	**98.03**	**98.73**	**95.03**	**96.71**
	AP_error	**39.60**	**39.65**	**27.17**	**37.12**	**31.08**	**37.67**	**44.28**
30%								
	AUC	87.04	**83.04**	**91.27**	79.91	83.21	69.06	**75.07**
MCP	AP_success	99.04	97.27	99.69	97.33	98.80	90.27	96.11
	AP_error	31.50	**36.41**	**20.92**	25.66	20.62	30.94	20.50
	AUC	**90.89**	82.57	90.92	**84.30**	**87.49**	**72.21**	74.51
Learned conf.	AP_success	**99.39**	**97.31**	**99.72**	**98.10**	**99.12**	**91.55**	**96.20**
	AP_error	**39.29**	34.34	18.15	**31.70**	**28.82**	**36.08**	**21.71**
10%								
	AUC	85.41	**76.35**	87.34	**82.75**	85.58	68.59	**71.49**
MCP	AP_success	98.41	**96.42**	98.79	96.39	98.42	83.85	**90.31**
	AP_error	31.40	24.69	27.49	**39.79**	30.63	42.57	32.98
	AUC	**88.50**	75.61	**89.00**	82.34	**86.05**	**69.48**	70.81
Learned conf.	AP_success	**98.88**	96.41	**99.09**	**96.83**	**98.56**	**84.39**	90.19
	AP_error	**34.21**	**27.39**	**32.15**	39.07	**34.16**	**44.77**	**32.32**

at a given absolute position and directly bias the end of the FCN. We show that in a similar partially-labeled context, this prior could improve the pseudo-label step.

Chapter 4

Incorporating Spatial Knowledge on Organ Positions when Training Deep FCNs

Contents

Abstract

In medical image segmentation, using prior knowledge about the target structures has already been a subject of interest. For example, in deformable models, the initial curve needs to be positioned sufficiently close to the object and thus needs to leverage prior knowledge. With the arrival of deep learning models, FCNs have become the standard choice for organ segmentation. However, those models are by design incapable of modeling spatial information and fail at realizing simple coordinate transforms. In this chapter, we

study how we could integrate a prior knowledge about the 3D absolute position of an organ which could be crucial for proper labeling in challenging contexts. For that, we introduce STIPPLE, a method that combines a model representing prior probabilities of an organ position in 3D with visual FCN predictions by means of a generalized prior-driven prediction function. Then, we also use our 3D spatial prior in a self-labeling process such as in Chapter 3 in order to improve the quality of the pseudo-label selection. We experimentally show on the TCIA dataset the relevance of the proposed method and how STIPPLE outperforms state-of-the-art semi-supervised segmentation methods by leveraging the 3D spatial prior information.

4.1 Introduction

Modern DL models and FCNs brought huge performance gains in medical image segmentation. However it remains a challenging task due to low contrasts between organs, and visual ambiguities. In many cases, the local visual context of an image is insufficient to perform a clear decision and external knowledge is required.

In this chapter, we study how we can include a prior knowledge about the 3D absolute spatial position of the organs to improve the quality of the segmentation. It is a particularly strong and relevant prior for medical images since there are some conventions on how the image should be due to the fixed position of the patients, Figure 1.5. For example, the liver is situated on the left side of the image and the kidneys are at the bottom on each side of the spine. Moreover, some organs are visible only on limited slices in the depth and others like the pancreas may largely vary in position, Figure 4.5. Using prior knowledge is in fact common for practitioners, which perform segmentation not only by using the visual appearance of medical images, but also leverage their strong knowledge on the absolute position of organs or relative layout between them.

The proposed method is STIPPLE for SpaTIal Priors and Pseudo LabEls. We introduce a 3D spatial prior which is a probability map of the organ presence at a given position. This map is merged with the visual information extracted by the FCN through a prior-driven prediction function, Section 4.2.2. We also propose a semi-supervised extension based on the work presented in Chapter 3 with an iterative self-labeling process. It forms a virtuous circle where the 3D prior is leveraged for selecting relevant pseudo-labels, leading to refined interactions between visual and prior predictions.

We perform experiments on a challenging pancreas segmentation dataset and show that our method outperforms the performances of other state-of-the-art approaches for both semi-supervision and integration of position information.

The main contributions of this chapter are as follows:

- We introduce STIPPLE, a 3D spatial prior that explicitly incorporates knowledge in a deep FCN for medical image segmentation. The prior is added in the final activation function via a prior-driven softmax.

- We show the relevance of such a prior in a fully-supervised setting and how it could be leveraged for semi-supervised within a pseudo-labeling scheme. For the latter, our prior helps to select new labels by limiting the incorporation of wrong predictions, especially outliers that could ruin the training.

- Experiments show that our prior is particularly powerful when very few labels are available. Moreover, compared to other state-of-the art methods, STIPPLE shows better results for every proportion of missing labels.

In this chapter we first present in Section 4.2, the proposed method and how we compute and integrate our spatial prior in a FCN through a prior-driven activation function. In Section 4.3, we give the experimental setup and the results for the segmentation of the pancreas and a deeper analysis of the impact of the parameters and other elements of the method. Discussion, especially about the limitations of the method is then given in Section 4.4. Eventually, a conclusion is given in Section 4.5.

4.2 Organ segmentation with 3D spatial priors and pseudo-labeling

In this section, we introduce our STIPPLE model dedicated to leverage spatial priors and pseudo-labeling for semantic segmentation of medical images. The overall prediction model of STIPPLE is depicted in Figure 4.1.

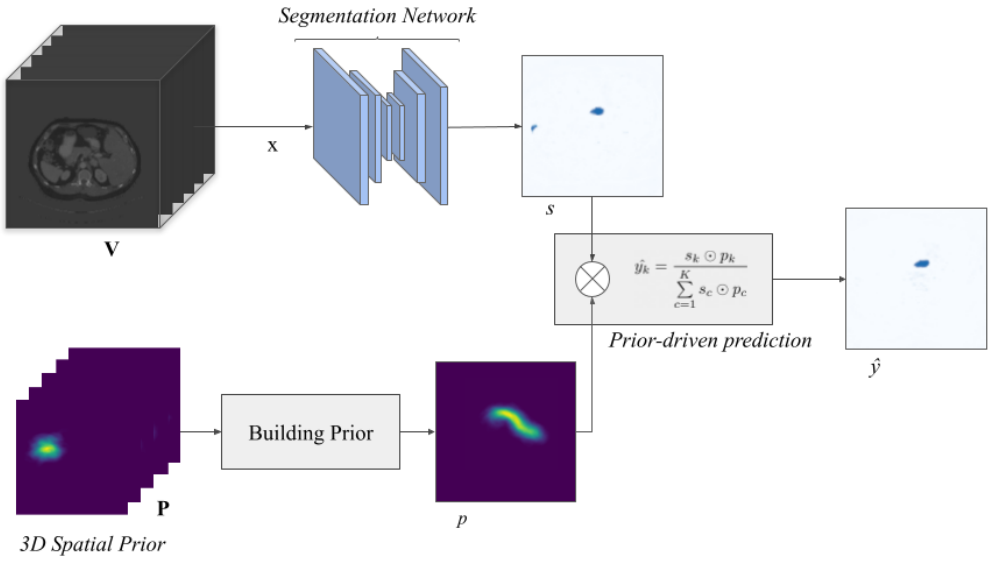


Figure 4.1: The input volume $\mathbf{V}$ is sliced along the axial view. The segmentation network outputs a visual prediction $\mathbf{S}$. The 3D spatial prior $\mathbf{P}$ is aligned to the slice before being combined through a prior-driven prediction function. The result is the final prediction $\hat{\mathbf{Y}}$.

A given input volume $\mathbf{V}$ is processed by the backbone FCN segmentation model which

outputs a probability prediction volume $\mathbf{S} = \{s_k\}_{k\in\{1;K\}}$ where K in the number of classes. Our approach is agnostic to the choice of the FCN: in our experiments we use 2D U-Net [2] due to hardware limitations and for experiment efficiency, but it could easily extend to 3D models [50].

Formally, let us consider a volume $\mathbf{V} \in \mathbb{R}^{W\times H\times Z}$ composed of Z axial slices, *i.e.* $\mathbf{V} = \{x_z\}_{z\in\{1;Z\}}$, with $x_z \in \mathbb{R}^{W\times H}$. The semantic segmentation problem consists in predicting a label among K organ classes (including the background) for each voxel of the volume $\mathbf{V}(w,h,z)$[1]. The FCN segmentation network computes posterior probabilities:

$$s(w,h)_{z,k} = \mathbf{Pr}\left(\mathbf{Y}_{w,h,z} = k \mid N(x(w,h)_z), \mathbf{W}\right)$$

for our case with a 2D model, where $\mathbf{W}$ represents the model parameters and $N(x(w,h)_z)$ is the voxel neighborhood in a given slice z, characterized by the FCN receptive field.

As previously mentioned, the computation of $s(w,h)_{z,k}$ doesn't incorporate any absolute position information. We propose to define a 3D spatial prior $\mathbf{P}$ which represents the probability of an organ presence given its 3D position. The final prediction of STIPPLE $\hat{\mathbf{Y}}$ consists in merging $\mathbf{P}$ and $\mathbf{S}$, as described in section 4.2.2.

4.2.1 3D spatial prior design and computation

To overcome the lack of absolute position information encoded in our FCN predictions $s(w,h)_{z,k} = \mathbf{Pr}\left(\mathbf{Y}_{w,h,z} = k \mid N(x(w,h)_z), \mathbf{W}\right)$, we propose to model the prior probabilities of the organ position, *i.e.* with $\mathbf{P} = \{p_k\}_{k\in\{1;K\}}$, $p(w,h)_{z,k} = \mathbf{Pr}\left(\mathbf{Y}_{w,h,z} = k \mid (w,h,z)\right)$, independently of the visual input $N(x(w,h)_z)$ and model parameters $\mathbf{W}$.

The construction of the proposed 3D spatial prior is based on the following assumptions: (1) the 3D volumes are given in the axial direction (z), with the patient lying on the back ; (2) In the axial (z) direction, there might be strong variations in the organ position, i.e. the $[z_{min}; z_{max}]$ interval where the organ is visible might significantly change. On the other hand, the variability in the (w,h) plane for a given z value is supposed to be much smaller, such that we can accumulate the organ positions in this plane across the dataset to obtain relevant statistics of organ position.

Note that these assumptions are valid in many clinical cases, since acquisitions in the axial direction are common. Moreover, it is also common for anatomical structures to be visible in variable $[z_{min}; z_{max}]$ values in the z direction because of differences in acquisition procedures.

[1]Here we choose to designate the coordinates with (w,h,z) so it is a different notation than the model's output and input, x and y.

Our prior $\mathbf{P}$ is estimated on a training dataset of labeled organs $\{\mathbf{Y}_i\}_{i\in\{1;N\}}$ where N is the number of examples, by computing statistics of the organ presence in a 3D rectangular volume of size $(W_p \times H_p \times \Delta_z)$ with W_p, H_p and Δ_z being respectively the width, the height and depth of the rectangular volume. This size is determined by taking the maximum width, height and depth of the considered organ in the training set such that every example fits into it. We observed that the position of the organs are relatively stable in the (w, h) coordinates, but may largely vary in the z direction. So we decide to discretize the prior over the z axis such that the prior $\mathbf{P}$ itself is of size $(W_p \times H_p \times B)$; where B bins aggregate the Δ_z slices, with $B < \Delta_z$ to gain invariance with respect to misalignment of organs in the z direction, but $B > 1$ to capture organ shape variations. Eventually, $p(w, h)_{z,k}$ is estimated from the full training dataset by a non-parametric estimation, *i.e.* histogram estimation:

$$p(w,h)_{z,k} = \mathbf{Pr}\left(\mathbf{Y}_{w,h,z} = k \mid (w,h,z)\right) = \frac{1}{Z_{tot}} \sum_{z=1}^{Z_{tot}} \mathbb{1}(\mathbf{Y}_{w,h,z} = k) \tag{4.1}$$

where Z_{tot} is the total number of slices in a given bin b.

In practice, the training volumes are first aligned with the center of the organ segmentation masks and then a sub-volume of size $(W_p \times H_p \times \Delta_z)$ is cropped around this center.

The prior computation is illustrated in Algorithm 3. An example of a 3D prior map with $B = 3$ bins is shown in Figure 4.2. We can see that each bin results in an average of multiple neighboring slices from the input volume. The bin (1) corresponds to the top of the segmentation mask whereas the bin (3) is the bottom of the pancreas. For those two bins the corresponding probabilities are localised in very different regions.

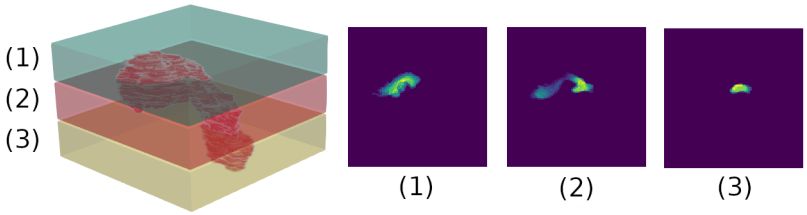


Figure 4.2: Prior computation visualisation on one volume with $B = 3$ bins in the z axis.

4.2.2 Prior-driven prediction function

The prior probabilities are introduced through a prior-driven prediction function which explicitly integrates our 3D spatial prior in a late fusion manner. For the sake of clarity we remove the notation of the dependency in (w,h,z). The main intuition which is presented in Figure 4.1 is to take the visual predictions of the FCN $\mathbf{S} \in \mathbb{R}^{W,H,Z,K}$ where K in the number

Algorithm 3: Prior construction for a given organ. y_i designates a volume label and N the total number of training volumes. Then B is the number of expected bins for the final prior and W_p, H_p are respectively the prior's Width and Height when Δ_z is the maximum depth observed for the organ in the training set.

Data: $\{(y_i)\}_N, B, W_p, H_p, \Delta_z$
Result: $Prior$
$N \leftarrow$ number of label maps y;
$Prior \leftarrow \text{zeros}(w, h, B)$;
for $i \leftarrow 1$ *to* N **do**
  $c_w, c_h, c_z \leftarrow get\_organ\_center(y_i)$;
  $w_{min} \leftarrow c_w - W_p/2$;
  $w_{max} \leftarrow c_w + W_p/2$;
  $h_{min} \leftarrow c_h - H_p/2$;
  $h_{max} \leftarrow c_h + H_p/2$;
  $z_{min} \leftarrow c_z - \Delta_z/2$;
  $z_{max} \leftarrow c_z + \Delta_z/2$;
  for $s = z_{min}$ *to* z_{max} **do**
    $idx\_in\_prior \leftarrow \frac{B \times (s - z_{min})}{z_{max} - z_{min}}$;
    $Prior[:, :, idx\_in\_prior] += y_i[x_{min} : x_{max}, y_{min} : y_{max}, s]$;
$Prior \leftarrow normalize\_bins(Prior) \leftarrow$ normalize the values between 0 and 1 by dividing by the number of slices added in a given bin;

of classes, so $\mathbf{S} = \{s_k\}_{k\in\{1;K\}}$ and apply a Hadamard product with the prior probabilities $\mathbf{P} = \{p_k\}_{k\in\{1;K\}}$. Then we normalize to rescale the values between 0 and 1.

When combining those operations, the final formulation (Equation 4.2) is denoted as a "prior-driven softmax", which outputs $\hat{\mathbf{Y}} = \{\hat{y_k}\}_{k\in\{1;K\}}$.

$$\hat{y_k} = \frac{s_k \odot p_k}{\sum_{c=1}^{K} s_c \odot p_c} = \frac{e^{\tilde{s}_k}\ p_k}{\sum_{c=1}^{K} e^{\tilde{s_c}}\ p_c} = \frac{e^{\tilde{s}_k + ln(p_k)}}{\sum_{c=1}^{K} e^{\tilde{s}_c + ln(p_c)}} \tag{4.2}$$

$\tilde{\mathbf{S}} = \{\tilde{s_k}\}_{k\in\{1;K\}}$ are the values before activation, usually denoted as "logits".

Interestingly, we can notice that our prediction function in Equation 4.2 is a consistent generalization of the standard softmax, since it reduces to it when the prior is uniformly distributed through the classes, *i.e.* when $p_k = p_c = \frac{1}{K}\ \forall k \in \{1..K\}$.

When the prior $\mathbf{P}$ is not uniform, it can be used to bias the prediction of a given class k based on its visual input $e^{\tilde{s}_k}$, depending on its spatial location. For example, if p_k is close to 1 (resp. 0), the prediction of class k is made close to 1 (resp. 0) whatever the $e^{\tilde{s}_k}$ value. Our prior-driven softmax prediction function in Equation 4.2 can thus be leveraged to overcome visual ambiguities between organs and the background.

This formulation is obviously applicable in binary segmentation using a sigmoid (σ) as shown in Equation 4.3. It becomes a "prior-driven sigmoid".

$$\hat{y_k} = \frac{s_k \odot p_k}{s_k \odot p_k + (1 - s_k) \odot (1 - p_k)} = \sigma(\tilde{s}_k - ln(1 - p_k) + ln(p_k)) \tag{4.3}$$

Positioning the prior in a volume During training, we can use the position of the organ label to position the prior in the image. However, for unlabeled volume and test volumes we need to find the position. We first take the output probabilities of a segmentation network on the target (unlabeled) volume, which gives a first but coarse position of the organ. Then, a reference volume is randomly selected among the labeled volumes in the training set. For that volume, we have a segmentation map and the true position of the considered organ. With that, we compute the KL divergence between the two with different small translations applied to the probabilities obtained on the target volume. We can finally keep the translation that gives the lowest KL divergence value and adjust the position of the organ for the target volume.

4.2.3 Integration in a semi-supervised context

To further evaluate STIPPLE, we propose a semi-supervised extension of our model, dedicated to leverage unlabeled data. We use a self-training strategy based on pseudo-labeling

such as the one proposed in Chapter 3. With STIPPLE, we use the MCP uncertainty measure. Concretely, we consider that a prediction with a high probability is more certain than another with a lower probability. Then, for a given volume, we select among the predictions of the organ the top-k most confident voxels that will be selected as pseudo-labels. Our STIPPLE method actually provides a "prior-driven uncertainty measure", in the sense that our 3D prior is leveraged to improve the selection of pseudo-labels by using 3D absolute position information.

4.3 Experiments and Results

4.3.1 Experimental setup

Evaluation dataset We evaluate our method on the publicly available dataset TCIA [58] for pancreas segmentation in CT-scans (see Section 2.2.4). In all our experiments, we performed 5 fold cross-validation and reported the standard deviation between the folds. For each fold, a different spatial prior is computed.

Implementation Details We carried out experiments in a semi-supervised setting. Similar to INERRANT (Chapter 3), we randomly removed labels (uniform sampling without replacement) at a patient level to reach proportions (α) like 70%, 50%, 30% and 10% of labeled volumes in the training set such that the test set remains the same across the experiments. We also report the results for a fully-supervised setting, *i.e.* a label proportion of 100%. In practice we use one step of relabeling for the low proportions from 50% to 10% and two steps at 70%.

The input volumes are preprocessed by clipping the Hounsfield Units (HU) values in the abdominal organ range [-160,300]. Then the values are normalized to have zero mean and unit variance. In all the experiments, we use a backbone 2D U-Net. The models are trained using the Adam optimizer with standard data augmentation techniques, *i.e.* random translations, random rotations.

The spatial prior is estimated with the available training examples only. We choose $B = 5$ for every proportion, and study its impact in section 4.3.5.

4.3.2 Pancreas segmentation results

The results on the TCIA pancreas dataset are given in Figure 4.3. STIPPLE is compared with a U-Net baseline for every proportion. In each case, our method shows significant gains which are validated with a paired t-test, see Table 4.1. At a label proportion of 100%, we see an improvement of +1.4pts, at 70%: +4.0pts, at 50%: +3.7pts, at 30%: +5.9pts and finally at 10%: +9.9pts. The gains are more pronounced when the proportion α is low. It is validated

by the $p-$values shown in Table 4.1. The gains increase and the $p-$values decrease when α decreases.

Table 4.1: $p-$values given by a paired t-test between the baseline and STIPPLE.

Proportion (α)	100%	70%	50%	30%	10%
p values	4.51%	3.00×10^{-4}%	5.53×10^{-2}%	6.40×10^{-6}%	2.60×10^{-7}%

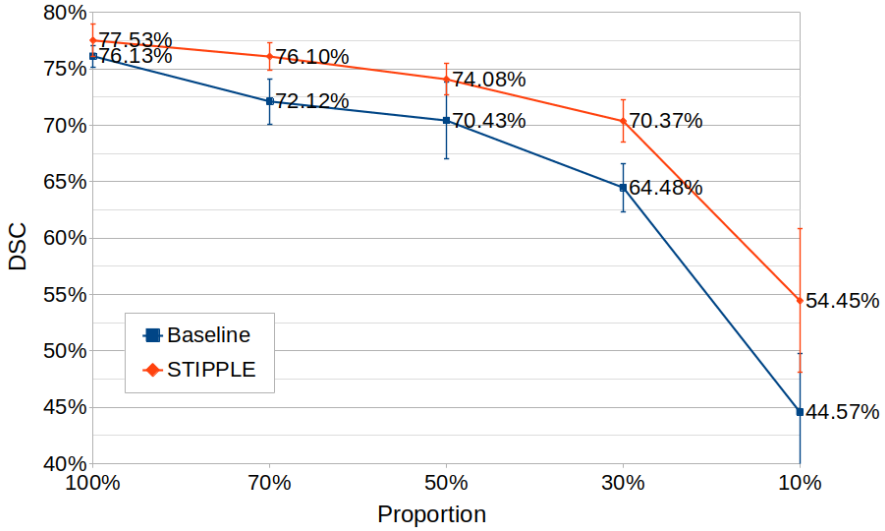


Figure 4.3: Segmentation results for STIPPLE ($B = 5$) compared to the baseline. Values are Dice Scores (DSC) for every proportion of missing labels from 100% (every image is labeled) to 10% (only 10% of the images are labeled). Error bars show the standard deviations of the results between the folds.

The images could be ambiguous due to the low contrast between the objects and because of the reduced size of the organ region. In medical image segmentation, it is common that the local visual content is insufficient, such that one needs external knowledge for proper segmentation. Moreover, the low balance of labeled pixels makes the model naturally under-segment, and this effect is exacerbated when very few labeled images are provided.

All this causes multiple kinds of errors which are addressed by the prior. Firstly, it reinforces the probabilities in the most probable region and allows to recover missed predictions. Secondly, it reduces false positives by cleaning out errors far from the region of interest. Finally, the prior stabilizes the relabeling step by selecting only the pixels in the correct region which avoid potential errors that could cause drops in performances.

To illustrate how the spatial prior acts on the predictions, we show in Figure 4.4 two examples. The first row is a missed prediction which has been correctly recovered thanks to the

prior. In that case the visual prediction has been reinforced by the spatial prior shown in the last column. The second row shows how the prior removes improbable segmentation and more generally false positives out of the organ region. We see that the wrong prediction of the baseline is out of the high prior probabilities in the last column. The visual prediction was not sufficient to correctly decide in this area but with STIPPLE the prior has removed the ambiguity and filtered out those errors. In this case, the prior combined with the visual prediction reduces the false positives and has a positive impact on the relabeling step by preventing adding errors.

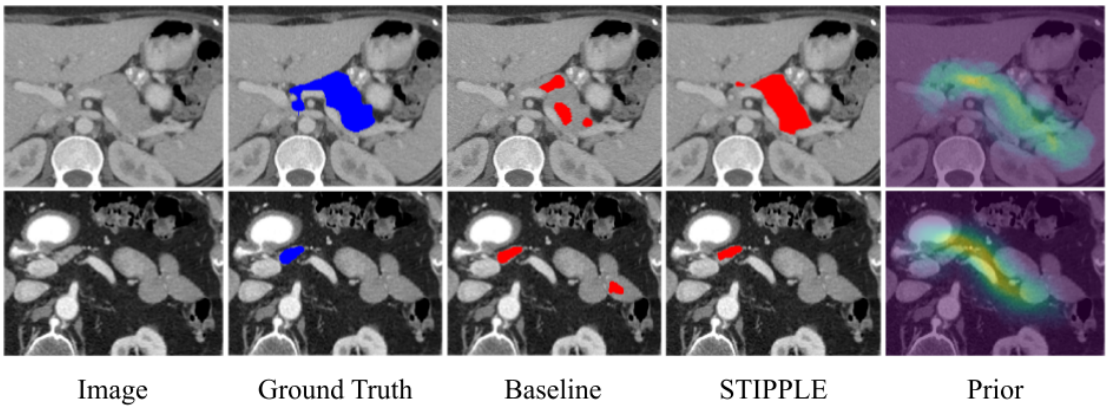


Figure 4.4: Examples of two behaviours induced by the spatial prior. First row: recovery of a missed prediction. Second row: cleaning of a wrong prediction in an unexpected area. The last column represents the spatial prior on top of the input image to illustrate where the prior influences the prediction.

To show that our method is agnostic to the choice of the backbone, we carry out experiments using a patch-based 3D U-Net. We choose a fixed fold and add the prior using the same method as explained. At 50%, we observe an improvement for the baseline of +3pts from 68% in DSC to 71% for the 3D U-Net. Then, with the spatial prior we observe an improvement of +1pt validating the relevance of our method. At 10%, our spatial prior with a 3D U-Net gets a 58% DSC outperforming both the baseline (+6pts) and our prior (+3pts) with the 2D U-Net. Our method can easily be extended to other backbones and our 3D spatial prior still improves the final results even with a strong baseline, e.i. 3D U-Net.

4.3.3 Ablation study

To understand how the different parts of STIPPLE act on the final performance, we show in Table 4.2 an ablation study of the method. The results are given for the different stages: the 2D U-Net baseline which is also the backbone in our experiments; after adding the 3D prior but without relabeling; the complete method, including the prior and the relabeling step.

Table 4.2: Ablation study of STIPPLE. The reported values are Dice Similarity Scores (DSC,%).

Proportion (α)	100%	70%	50%	30%	10%
Baseline	76.13 ($\pm$ 0.94)	72.12 ($\pm$ 2.01)	70.43 ($\pm$ 3.38)	64.48 ($\pm$ 2.13)	44.57 ($\pm$ 5.24)
STIPPLE w/o relab	77.53 ($\pm$ 1.44)	75.02 ($\pm$ 2.21)	71.74 ($\pm$ 2.02)	65.99 ($\pm$ 1.71)	47.41 ($\pm$ 8.40)
Baseline w relab	-	75.12 ($\pm$ 1.91)	73.71 ($\pm$ 2.59)	69.00 ($\pm$ 2.04)	51.91 ($\pm$ 7.77)
STIPPLE	**77.53** ($\pm$ 1.44)	**76.10** ($\pm$ 1.23)	**74.08** ($\pm$ 1.39)	**70.37** ($\pm$ 1.88)	**54.45** ($\pm$ 6.37)

Adding the prior alone outperforms the baseline for every proportion. The relative gains are +1.41pts at 100%, +2.90pts at 70%, +1.32pts at 50%, +1.50pts at 30%, and finally +2.84pts at 10%. The information brought by the spatial prior allows to increase the results consistently through the proportions. This shows the relevance of exploiting the absolute position for organ segmentation. Then, the relabeling step boosts the performances as we can see in the last row. This step is particularly interesting for the low proportions. As discussed in section 4.3.2, the gains are more and more important when the proportion α is decreasing.

Using a prior impacts positively the performances in the two contexts: with or without relabeling. We can also notice that the relabeling step boosts the results especially for the low αs.

4.3.4 State-of-the art comparison

We compare our method with other semi-supervised approaches in addition to a method that includes an attention mechanism. In [80], the unlabeled images are completely relabeled before training a new model. [5] propose an adversarial training to incorporate unlabeled images during training. Finally, [75], use a mean teacher method where the unlabeled images are used through the consistency loss. We also compare our method with an attention model from [11]. It uses an additive attention gate in the decoder part of the U-Net before the concatenation of the skip-connections.

Table 4.3: State-of-the-art comparison on TCIA.

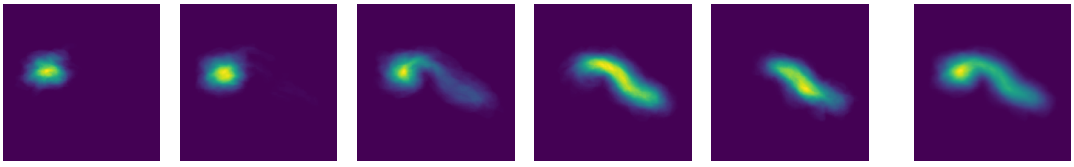
Proportion (α)	100%	70%	50%	30%	10%
Baseline	76.13 ($\pm$ 0.94)	72.12 ($\pm$ 2.01)	70.43 ($\pm$ 3.38)	64.48 ($\pm$ 2.13)	44.57 ($\pm$ 5.24)
Pseudo-labels ([80])	-	75.12 ($\pm$ 1.91)	73.71 ($\pm$ 2.59)	69.00 ($\pm$ 2.04)	51.91 ($\pm$ 7.77)
Adversarial ([5])	-	75.41 ($\pm$ 1.78)	73.91 ($\pm$ 2.27)	67.60 ($\pm$ 1.84)	52.09 ($\pm$ 6.00)
Consistency ([75])	-	74.53 ($\pm$ 2.10)	72.68 ($\pm$ 3.05)	66.99 ($\pm$ 1.38)	46.04 ($\pm$ 3.70)
Attention U-Net [11]	76.38 ($\pm$ 1.27)	74.18 ($\pm$ 1.57)	71.37 ($\pm$ 1.73)	64.25 ($\pm$ 2.49)	41.28 ($\pm$ 6.47)
STIPPLE (Ours)	**77.53** ($\pm$ 1.44)	**76.10** ($\pm$ 1.23)	**74.08** ($\pm$ 1.39)	**70.37** ($\pm$ 1.88)	**54.45** ($\pm$ 6.37)

Table 4.3 shows the results of the comparison. For every row, we implement the method with the same backbone 2D U-Net. STIPPLE shows better results for every proportion with

a more pronounced gain in the low αs, *e.g.* at 10%, STIPPLE is better by 2.4pts than the best method (the adversarial). The pseudo-labels method [80] is the closest to ours but we see that STIPPLE stays above for every proportion thanks to the spatial prior and the progressive adding of pseudo-labels.

Concerning the attention model in [11], we can see that compared to the baseline, it helps consistently from $\alpha = 100\%$ to $\alpha = 50\%$. Then, the scores drop below the baseline. STIPPLE is better for every proportion and especially for the low αs. It could be explained by the fact that our prior exploits the three dimensions unlike the attention module which is 2D. Moreover it is built beforehand by following a specific method which is adapted to low label proportions.

4.3.5 Further Analysis



(a) $b = 0$ (b) $b = 1$ (c) $b = 2$ (d) $b = 3$ (e) $b = 4$ (f) 2D prior

Figure 4.5: Visualization of a spatial prior with $B = 5$. We can see how it captures the depth information compared to (f) which is a 2D prior.

Impact of the prior size B. The number of bins, B, of the prior impacts the final results and the best value may depend on the available data. As an example, Figure 4.5 shows a spatial prior with $B = 5$ and $B = 1$, *i.e.* 2D prior. At $B = 5$, we can see how the spatial position evolves through the 3D prior bins. As a contrary, the 2D prior ($B = 1$) doesn't encode the depth information and is thus less informative.

We evaluate STIPPLE without relabeling with different B values (1, 2, 5, 7, 10 and 90) at 10% and 70% of labeled images, see Figure 4.6. $B = 90$ means that there is no discretization in z, *i.e.* the spatial prior is complete.

We observe that the best value at 70% is 5 but for every B there is a significant improvement compared to the baseline. At 10%, the best results are given for 5, 7 and 10 with an optimal value at 7. In our experiments in section 4.3.3, we choose a standard value of $B = 5$. Though it is good in practice, it means that we could get better results by increasing B for lower proportions.

For both proportions, we can see that the prior has better results than the baseline. Using a 2D prior ($B = 1$) is effective but using more bins boosts the performances. Then, with a complete prior, $B = 90$, the scores decrease which shows that discretizing the z axis is relevant.

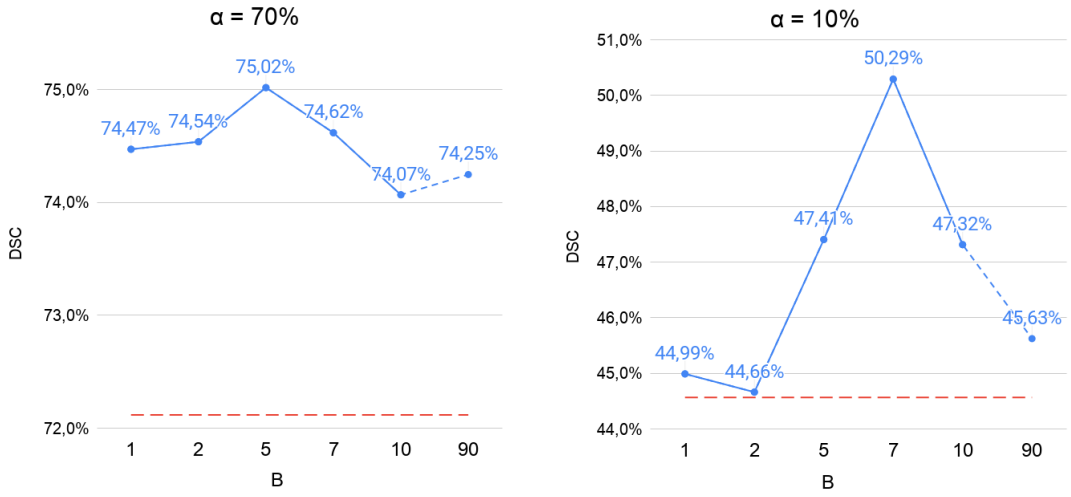


Figure 4.6: Dice score versus the number of bins B at 70% and 10% of labeled images. In blue, STIPPLE without relabeling. In dotted red, the baseline.

Impact of the prior positioning As explained in section 4.2.2, the prior has to be positioned in the test volumes. We use the predicted position refined by an adjustment step. Table 4.4 shows the results with the naive method of using only the center given by the segmentation model. Then, with the adjustment step used in STIPPLE.

As we can see the naive approach is not sufficient and alters the final results. The adjustment step is necessary and allows to reach optimal results comparable to the one obtained by using the true organ position.

Table 4.4: Impact of the prior positioning on the final results.

Proportion	100%	70%	50%	30%	10%
Naive	74.48 (± 2.53)	72.84 (± 3.15)	69.90 (± 1.85)	61.82 (± 3.49)	41.80 (± 9.94)
Ours	**77.53** (± 1.44)	**75.02** (± 2.21)	**71.74** (± 2.02)	**65.99** (± 1.71)	**47.41** (± 8.40)

4.4 Discussion and Limitations

Our STIPPLE method aims at integrating prior knowledge about the absolute position of the organs in a deep FCN. Contrarily to other works, we formulate an explicit 3D spatial prior which is completely integrated in the model in a late fusion manner.

As discussed in Chapter 4, some approaches use implicit mechanisms to add a prior knowledge such as using the loss function to bias the training [93, 94, 69]. Those approaches influence the training stage by forcing the network to reach a different minima that is expected to be

better. However, when the training is done, the prior is no longer used thus it could be more seen as a regularization term that smooths the learning but we can't tell that the prior is integrated in the network. In STIPPLE, the prior is directly integrated into the network in a late-fusion manner within the final activation function. We can thus assure that our spatial prior explicitly participates in the final decision.

Another way of seeing the incorporation of prior knowledge is via attention modules such as in Attention U-Net [11]. Those modules learn to focus on certain regions of the image by computing attention weights. Thus, attention models are trained end-to-end and learn how to focus using the same loss as the segmentation itself. However, it is really different than explicitly integrating a prior. The attention weights are situated in intermediate feature maps and have a filtering action which eliminates irrelevant values. Thus, we can't actually tell that it is going to learn complex spatial information and interactions contrary to STIPPLE.

STIPPLE and its explicit formulation of a spatial prior has another important property that we show in our experiments. It could benefit from a semi-supervised pseudo-labeling scheme (Table 4.3). In fact, in a semi-supervised context, our spatial prior is particularly relevant as it concentrates the observed position and shape of the organ into a probabilistic map which directly biases the model. Thus, when very few labeled data are available it gives an explicit direction and guides the learning. Moreover, we propose a pseudo-labeling scheme and observe that our spatial prior gives a better relabeling by eliminating irrelevant predictions and concentrates the selected pseudo-labels in the most probable area.

Limitations STIPPLE relies on assumptions such that the position of an organ in (w,h) varies slightly compared to the variations in z. Thus, there could be an issue when strong rotations (*e.g.* of the patient) occur, or for data mixing various acquisition directions (axial/coronal/sagittal). In this case, our approach would require a (manual or automatic) method to register with respect to those variations.

A second problem could emerge for atypical cases. For example, for patients with *situs inversus* where the major abdominal organs are reversed from their normal positions. With STIPPLE we define a spatial prior which translates the observed average position of the organs. However, with certain conditions, it could not apply and a human professional is needed. We must point out that those conditions represent a fraction of the cases and most of the available segmentation datasets do not contain any atypical cases.

However, our method could be adapted to other imaging modalities by adapting the prior computation or the prior positioning depending on the problem. The main idea is the same when a segmentation dataset with dense labels is provided.

4.5 Conclusion and perspectives

In this chapter, we studied how we could integrate prior knowledge about the absolute spatial position of the organs in the task of segmentation. For that, we introduced STIPPLE, a method that integrates a 3D spatial prior in a partially-labeled setting similar to that in Chapter 3. STIPPLE shows very important gains especially when few images are available which is particularly relevant in the medical field where labeled data are limited and very expensive to obtain. Comparisons with state-of-the-art methods further highlight the relevance of our method compared to attention models and other semi-supervision techniques.

In the next chapter, Section 5, we look at attention networks and more specifically Transformers [17]. We study how we could learn spatial information instead of using an explicit prior as in STIPPLE. For that we propose an architecture based on U-Net that uses self-attention at different levels of the network. It aims at leveraging long range dependencies between the objects but could also model the absolute position information thanks to the positional encoding.

Chapter 5

Transformers and Dense Attention for Modeling Long Range Interactions

Contents

Abstract

FCNs suffer from conceptual limitations especially for the task of segmentation. In fact, segmentation networks have a limited receptive field which does not capture sufficient contextual information which is however crucial to correctly address difficult and complex structures. In this chapter, we introduce the U-Transformer network, which combines a U-shaped architecture for image segmentation with Transformers. For that, we propose two modules with a self and cross attention. The self-attention aims at explicitly model full contextual information and long range interactions between anatomical structures. Then, the cross-attention works as a filtering operation that allows a fine spatial recovery in the decoder by filtering non-semantic features from the skip connections. Those attention modules are different in nature from previous works which use local attention mechanisms that take one or only few pixels into account. Experiments on TCIA and the IMO datasets show the large performance gain brought out by U-Transformer compared to U-Net and Attention U-Net which uses a pixel-level attention gate mechanism.

5.1 Introduction

In Chapter 4 we proposed an explicit way of introducing a spatial prior in a deep FCN. We also show that it could both improve the segmentation results on difficult organs such as the pancreas and reduce the FPs introduced with the pseudo-labeling scheme developed in Chapter 3. In STIPPLE we choose to explicitly bias the model with a late fusion scheme. It has the major advantage of being directly interpretable. For instance, biasing the loss function with prior knowledge [69] does not assure that the absolute spatial information is taken into account. However, this solution does not add global contextual information and is thus limited by the local RF of the ConvNet.

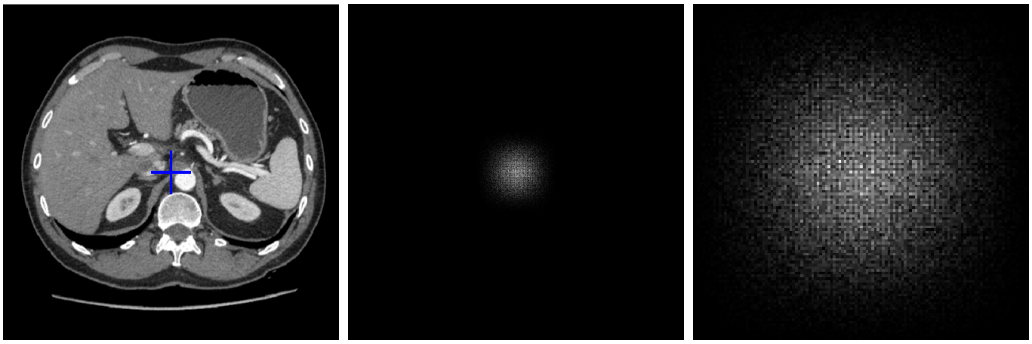
In this chapter we address the problem of explicitly learning contextual information. Contrary to STIPPLE, we want to learn, in addition to the spatial locations of the organs, interactions between structures and thus leverage complex contextual information.

We should note that the response for a given pixel uses the contextual information from the receptive field. However, this RF is often limited, especially in segmentation networks where the encoder part is limited because of the compromise induced by the memory cost. In Figure 5.1 a) for segmenting the blue cross corresponding to the pancreas with U-Net: the limited Receptive Field (RF) framed in red does not capture sufficient contextual information, making the segmentation difficult, see Figure 5.1 c).

Figure 5.1: Global context is crucial for complex organ segmentation but cannot be captured by vanilla U-Nets with a limited receptive field, *i.e.* blue cross region in a) with failed segmentation in c). The proposed U-Transformer network represents full image context by means of attention maps b), which leverage long-range interactions with other anatomical structures to properly segment the complex pancreas region in d).

The Receptive Field (RF) in a ConvNet designates the area of the input image reached by a unit at the end of the network. It can be obtained theoretically by looking at the convolution and

pooling operations. In our work, we use 512x512 input images and the Theoretical Receptive Field (TRF) of a standard U-Net is small (140x140) which do not enable to model full contextual information. Although the TRF is larger for deeper networks (*e.g.* nnU-Net), the TRF often over-estimates the actual contextual information that the network could handle. This has been studied in [10] where the authors introduced the notion of Effective Receptive Field (ERF). The proposed method consists in putting a gradient of one at the end of the bottleneck for the central unit and set the other gradients to zeros. Then, we propagate the gradients with the back-propagation and get the values assigned to the network's input. Thus, we obtain an array with the same size as the input as shown in Figure 5.2. We can see that the gradients describe a Gaussian centered on the image's center with the values quickly decreasing till reaching zero. With that Gaussian, we can get the Effective Receptive Field (ERF) as formulated in [10]. We can easily imagine that the ERF is considerably smaller than the TRF and for instance the nnU-Net, Figure 5.2c, which has a large TRF gets a ERF of about 200x200, which is much lower.



(a) CT slice (b) U-Net (c) nnU-Net

Figure 5.2: The Effective Receptive Field as formulated in [10]. We put a gradient of one at the end of the encoder and propagate it to the input. The figures show high gradient values in white and zero gradients in black. We analyse the U-Net and nnU-Net architectures and observe that the final ERF is much smaller than the TRF.

In this chapter, we introduce the U-Transformer network, which leverages the strong abilities of transformers [17] to model long-range interactions and spatial relationships between anatomical structures. U-Transformer keeps the inductive bias of convolution by using a U-shaped architecture, but introduces attention mechanisms at two main levels, which help to interpret the model decision. Firstly, a self-attention module leverages global interactions between semantic features at the end of the encoder. This module uses the dense attention mechanism from Transformers where for each pixel a weight is given for every other pixel. It thus explicitly models full contextual information and long-range dependencies. Secondly, we

introduce a cross-attention module which is used in the skip connections to filter non-semantic features. It uses the output of the previous decoder block to compute the attention matrix which is then used to transform the skip connection. This module allows a fine spatial recovery in the U-Net decoder and reduces the segmentation of wrong structures.

Figure 5.1 b) shows a cross-attention map induced by U-Transformer, which highlights the most important regions for segmenting the blue cross region in Figure 5.1 a): our model leverages the long-range interactions with respect to other organs (liver, stomach, spleen) and their positions to properly segment the whole pancreas region, see Figure 5.1 d). Quantitative experiments conducted on two abdominal CT-image datasets show the large performance gain brought out by U-Transformer compared to U-Net and to the local attention in [11].

The main contributions brought by this work are as follows:

- We propose U-Transformer, a U-shaped FCN which uses attention mechanisms from Transformers.
  - We first introduce a self-attention module which leverages global interactions between features in the bottleneck. It allows to model long-range dependencies thanks to the attention mechanisms from Transformers which compute weights between all pairs of input pixels.
  - Then we propose a cross attention module positioned in the skip connections. It aims at filtering non-semantic features from the skip connections based on the previous decoder block allowing a finer spatial recovery.
- Experimental results show that using self and cross attention gives systematic gains on different datasets and outperforms other attention models especially on challenging organs. Moreover, it could be easily adapted to other FCNs backbone and adds only a limited memory overhead. Finally, we give qualitative results to illustrate how the attention participates in the decision.

In this chapter we present the U-Transformer model. Firstly, in Section 5.2, we propose a specific review of the related works. Then, in Section 5.3 we dive into the U-Transformer model and explain how we integrate the self-attention module and the cross-attention modules at each decoder level in a U-Net model. Next, we show the experimental results in Section 5.4 on TCIA and IMO dataset which show the relevance of the proposed method and how it improves the segmentation of small and difficult organs. Eventually, in Section 5.5 we will discuss the method limitations and the perspectives for future works.

5.2 Related Work

Attention Models As we said in Section 2.5, only few works have been proposed and use simple attention modules [102, 103, 104, 5, 105, 11, 106]. Attention U-Net [11, 104] is one of those models and introduces an additive attention gate which aim at filtering the features coming from the skip connections, as shown in Figure 5.3. The attention weights are computed from the gating signal coming from the previous level of the decoder and the skip connection. At the bottom we can see the detailed attention gate and how the weights are computed. The final attention is very local because every operation is done at a pixel-level.

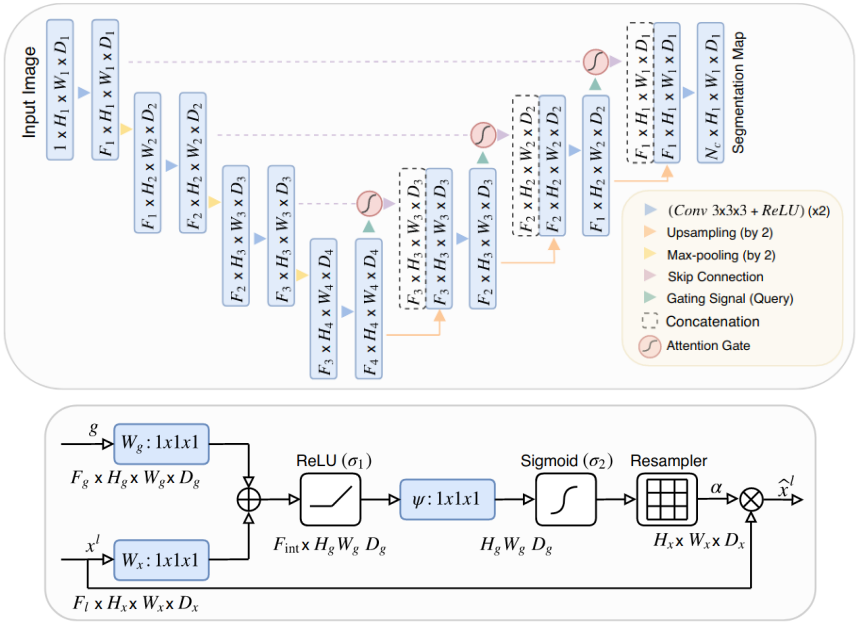


Figure 5.3: The Attention U-Net as proposed in [11]: *"Attention U-Net: Learning Where to Look for the Pancreas"*. The top image is the overall architecture with Attention Gates (AGs) at each skip connection. The bottom image is the attention gate mechanism with g being the gating signal (from the previous decoder block) and x the input signal (the skip connection).

In U-transformer, we introduce a cross-attention module in the decoder. It shares the same motivation of filtering out the skip connections based on more semantic features than in Attention U-Net. However, the attention gate shares the same limitation of local attention than the other models. On the other hand, our cross-attention is based on Transformers [17] and is able to model long-range interactions. A detailed description of this module is given in Section 5.3.2 and a schema in Figure 5.6. Moreover, our MHCA is original in its design since the keys and the queries are computed from the high-level features. It differs from the standard way cross-attention is used in [17]. In our case, we are not trying to express similarities between

the different U-Net levels but rather to filter the skip-connections based on the self-similarity of more semantic features. On top of that, we propose to add a Multi-Head Self-Attention (MHSA) in the bottleneck which further enforces the modelization of global interactions in our model, which are not leveraged in Attention U-Net.

Discussion on Concurrent Works Transformer networks have not been extensively studied in medical image analysis. However, there have been several attempts in the last few months [124, 53, 125, 126, 127]. In TransUnet [53], the authors propose a method inspired from DeTr [110] integrated into a U-Net model. It could be seen as using only self-attention in the bottleneck as compared to our model which also adds cross-attention mechanisms in the skip connections. In the TransFuse model [125], the attention module is inspired from SeTr [113] where the image is first divided into patches which are then considered as tokens. Using this approach reduces considerably the input information contrary to our model which uses the complete image and could model finer global interactions. In CoTr [127], the model is based on Deformable DeTr [128] which is a very specific method aiming at reducing the memory needed by Transformers by using a "deformable" Transformer that do not compute the complete attention matrix. Instead, they use a limited number of reference points which point with an offset vector to the most important tokens but not all of them. It allows the processing of multi-scale and high resolution features. Despite those attempts, none of them propose to use a cross-attention in a U-shape FCN to improve the spatial recovery in the decoder, contrary to U-Transformer.

5.3 The U-Transformer Network

As mentioned in Section 5.1, encoder-decoder U-shaped architectures lack global context information to handle complex medical image segmentation tasks. We introduce the U-Transformer network, which augments U-Nets with attention modules built from multi-head transformers. U-Transformer models long-range contextual interactions and spatial dependencies by using two types of attention modules, see Figure 5.4: Multi-Head Self-Attention (MHSA) and Multi-Head Cross-Attention (MHCA). Both modules are designed to express a new representation of the input based on its self-attention in the first case (Section 5.3.1) or on the attention paid to higher level features in the second (Section 5.3.2).

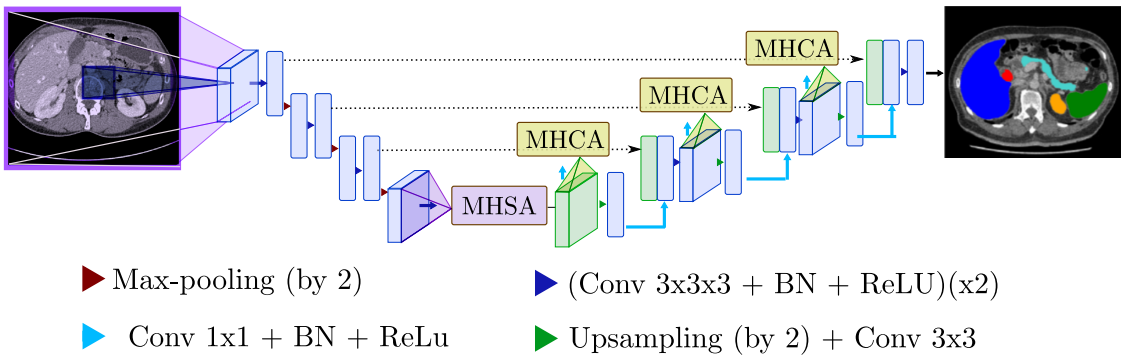


Figure 5.4: **U-Transformer** augments U-Nets with transformers to model long-range contextual interactions. The Multi-Head Self-Attention (MHSA) module at the end of the U-Net encoder gives access to a receptive field containing the whole image (shown in purple), in contrast to the limited U-Net receptive field (shown in blue). Multi-Head Cross-Attention (MHCA) modules are dedicated to combine the semantic richness in high level feature maps with the high resolution ones coming from the skip connections.

5.3.1 Self-attention

The MHSA module is designed to extract long range structural information from the images. To this end, it is composed of multi-head self-attention functions as described in [17] positioned at the bottom of the U-Net as shown in Figure 5.4. The main goal of MHSA is to connect every element in the highest feature map with each other, thus giving access to a receptive field including all the input image. The decision for one specific pixel can thus be influenced by any input pixel. The attention formulation is given in Equation 5.1. A self-attention module takes three inputs, a matrix of queries $\boldsymbol{Q} \in \mathbb{R}^{n \times d_k}$, a matrix of keys $\boldsymbol{K} \in \mathbb{R}^{n \times d_k}$ and a matrix of values $\boldsymbol{V} \in \mathbb{R}^{n \times d_k}$.

$$\text{Attention}(\boldsymbol{Q}, \boldsymbol{K}, \boldsymbol{V}) = \text{softmax}(\frac{\boldsymbol{Q}\boldsymbol{K}^T}{\sqrt{d_k}})\boldsymbol{V} = \boldsymbol{A}\boldsymbol{V} \tag{5.1}$$

A line of the attention matrix $\boldsymbol{A} \in \mathbb{R}^{n \times n}$ corresponds to the similarity of a given element in $\boldsymbol{Q}$ with respect to all the elements in $\boldsymbol{K}$. Then, the attention function performs a weighted average of the elements of the value $\boldsymbol{V}$ to account for all the interactions between the queries and the keys as illustrated in Figure 5.5. In our segmentation task, $\boldsymbol{Q}$, $\boldsymbol{K}$ and $\boldsymbol{V}$ share the same size and correspond to different learnt embedding of the highest level feature map denoted by $\boldsymbol{X}$ in Figure 5.5. The embedding matrices are denoted as $\boldsymbol{W}_q$, $\boldsymbol{W}_k$ and $\boldsymbol{W}_v$. The attention is calculated separately in multiple heads before being combined through another embedding. Moreover, to account for absolute contextual information, a positional encoding is added to the input features. It is especially relevant for medical image segmentation, where the different

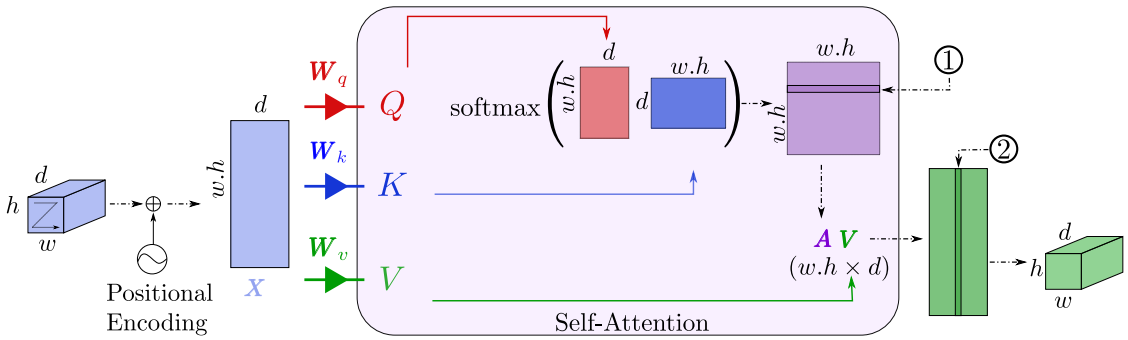


Figure 5.5: **MHSA module**: the input tensor is embedded into a matrix of queries $\boldsymbol{Q}$, keys $\boldsymbol{K}$ and values $\boldsymbol{V}$. The attention matrix $\boldsymbol{A}$ in purple is computed based on $\boldsymbol{Q}$ and $\boldsymbol{K}$. (1) A line of $\boldsymbol{A}$ corresponds to the attention given to all the elements in $\boldsymbol{K}$ with respect to one element in $\boldsymbol{Q}$. (2) A column of the value $\boldsymbol{V}$ corresponds to a feature map weighted by the attention in $\boldsymbol{A}$.

anatomical structures follow a fixed spatial position. The positional encoding can thus be leveraged to capture absolute and relative position between organs in MHSA.

5.3.2 Cross-attention

The MHSA module connects every element in the input with each other. Attention may also be used to increase the U-Net decoder efficiency and in particular enhance the lower level feature maps that are passed through the skip connections. Indeed, if these skip connections ensure to keep a high resolution information they lack the semantic richness that can be found deeper in the network. The idea behind the MHCA module is to turn off irrelevant or noisy areas from the skip connection features and highlight regions that present a significant interest for the application. Figure 5.6 shows the cross-attention module. The MHCA block is designed as a gating operation of the skip connection $\boldsymbol{S}$ based on the attention given to a high level feature map $\boldsymbol{Y}$. The computed weight values are then re-scaled between 0 and 1 through a sigmoid activation function. The resulting tensor, denoted $\boldsymbol{Z}$ in Figure 5.6, is a filter where low magnitude elements indicate noisy or irrelevant areas to be reduced. A cleaned up version of $\boldsymbol{S}$ is then given by the Hadamard product $\boldsymbol{Z} \odot \boldsymbol{S}$. Finally, the result of this filtering operation is concatenated with the high level feature tensor $\boldsymbol{Y}$. Here, the keys and queries are computed from the same source as we are designing a filtering operation whereas for NLP tasks, having homogeneous keys and values may be more meaningful. This configuration proved to be empirically more effective.

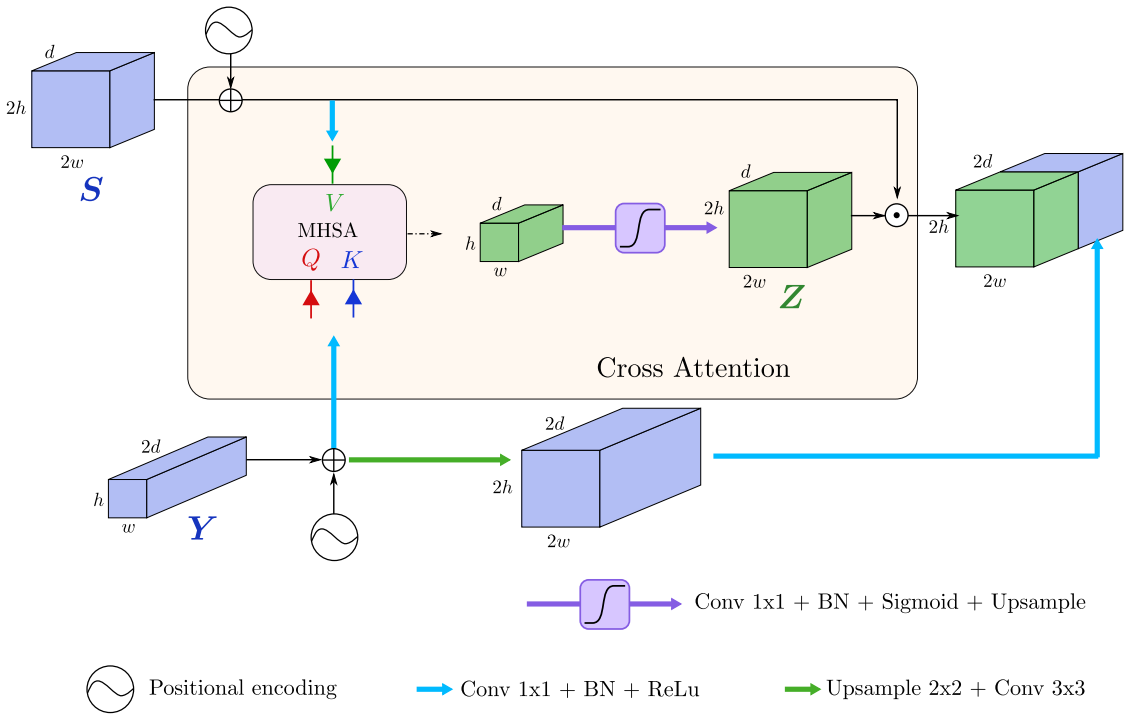


Figure 5.6: **MHCA module**: the value of the attention function corresponds to the skip connection $\boldsymbol{S}$ weighted by the attention given to the high level feature map $\boldsymbol{Y}$. This output is transformed into a filter $\boldsymbol{Z}$ and applied to the skip connection.

5.4 Experiments

We evaluate U-Transformer for abdominal organ segmentation on the TCIA pancreas public dataset, and the internal multi-organ dataset (IMO), Section 2.2.4.

Accurate pancreas segmentation is particularly difficult, due to its small size, complex and variable shape, and because of the low contrast with the neighboring structures, see Figure 5.1. In addition, the multi-organ setting assesses how U-transformer can leverage attention from multi-organ annotations.

Experimental setup All experiments follow a 5-fold cross validation, using 80% of images in training and 20% in test. We use the Tensorflow library to train the model, with Adam optimizer (10^{-4} learning rate, exponential decay scheduler).

We compare U-Transformer to the U-Net baseline [2] and Attention U-Net [11] with the same convolutional backbone for fair comparison. We also report performances with self-attention only (MHSA, section 5.3.1), and the cross-attention only (MHCA, section 5.3.2). U-Net has $\sim$ 30M parameters, the overhead from U-transformer is limited (MHSA $\sim$ 5M, each MHCA

block $\sim$ 2.5M).

Table 5.1: Results for each method in Dice similarity coefficient (DSC, %)

Dataset	U-Net [2]	Attn U-Net [11]	MHSA	MHCA	U-Transformer
TCIA	76.13 ($\pm$ 0.94)	76.82 ($\pm$ 1.26)	77.71 ($\pm$ 1.31)	77.84 ($\pm$ 2.59)	**78.50** ($\pm$ 1.92)
IMO	86.78 ($\pm$ 1.72)	86.45 ($\pm$ 1.69)	87.29 ($\pm$ 1.34)	87.38 ($\pm$ 1.53)	**88.08** ($\pm$ 1.37)

5.4.1 U-Transformer performances

Table 5.1 reports the performances in Dice averaged over the 5 folds, and over organs for IMO. U-Transformer outperforms U-Net by 2.4pts on TCIA and 1.3pts for IMO, and Attention U-Net by 1.7pts for TCIA and 1.6pts for IMO. The gains are consistent on all folds, and paired t-tests show that the improvement is significant with $p-$values $< 3\%$ for every experiment.

Figure 5.7 provides qualitative segmentation comparison between U-Net, Attention U-Net and U-Transformer. We observe that U-Transformer performs better on difficult cases, where the local structures are ambiguous. For example, in the second row, the pancreas has a complex shape which is missed by U-Net and Attention U-Net but U-Transformer successfully segments the organ.

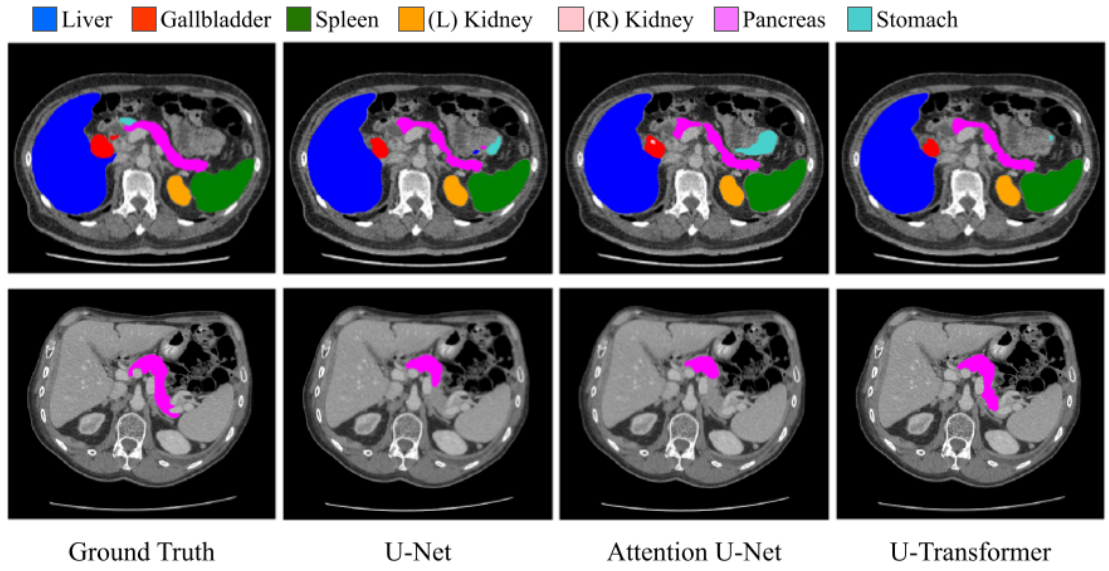


Figure 5.7: Segmentation results for U-Net [2], Attention U-Net [11] and U-Transformer on the multi-organ IMO dataset (first row) and on TCIA pancreas (second row).

In Table 5.1, we can see that the self-attention (MHSA) and cross-attention (MCHA) alone already outperform U-Net and Attention U-Net on TCIA and IMO. Since MCHA and Attention

U-Net apply attention mechanisms at the skip connection level, it highlights the superiority of modeling global interactions between anatomical structures and positional information instead of the simple local attention in [11]. Finally, the combination of MHSA and MHCA in U-Transformer shows that the two attention mechanisms are complementary and can collaborate to provide better segmentation predictions.

Table 5.2 details the results for each organ on the multi-organ IMO dataset. This further highlights the interest of U-Transformer, which significantly outperforms U-Net and Attention U-Net for the most challenging organs: pancreas: +3.4pts, gallbladder: +1.3pts and stomach: +2.2pts. This validates the capacity of U-Transformer to leverage multi-label annotations to drive the interactions between anatomical structures, and use easy organ predictions to improve the detection and delineation of more difficult ones. We can note that U-Transformer is better for every organ, even the liver which has a high score $> 95\%$ with U-Net.

Table 5.2: Results on IMO in Dice similarity coefficient (DSC, %) detailed per organ.

Organ	U-Net [2]	Attn U-Net [11]	MHSA	MHCA	U-Transformer
Pancreas	69.71 (± 3.74)	68.65 (± 2.95)	71.64 (± 3.01)	71.87 (± 2.97)	**73.10** (± 2.91)
Gallbladder	76.98 (± 6.60)	76.14 (± 6.98)	76.48 (± 6.12)	77.36 (± 6.22)	**78.32** (± 6.12)
Stomach	83.51 (± 4.49)	82.73 (± 4.62)	84.83 (± 3.79)	84.42 (± 4.35)	**85.73** (± 3.99)
Kidney(R)	92.36 (± 0.45)	92.88 (± 1.79)	92.91 (± 1.84)	92.98 (± 1.70)	**93.32** (± 1.74)
Kidney(L)	93.06 (± 1.68)	92.89 (± 0.64)	92.95 (± 1.30)	92.82 (± 1.06)	**93.31** (± 1.08)
Spleen	95.43 (± 1.76)	95.46 (± 1.95)	95.43 (± 2.16)	95.41 (± 2.21)	**95.74** (± 2.07)
Liver	96.40 (± 0.72)	96.41 (± 0.52)	96.82 (± 0.34)	96.79 (± 0.29)	**97.03** (± 0.31)

Additional results on the TCIA multiorgan dataset

To further analyse the method performances, we use the multiorgan extension of the TCIA dataset. Table 5.3 shows experiments with multiple organs annotations in addition to the pancreas. It shows the same trends as on our private dataset, with an improvement of 1.3pt on average Dice and very important gains for small, difficult organs: pancreas +4.7pts, duodenum +1.9pts, stomach +1.8pts, gallbladder +1.6pts.

5.4.2 U-Transformer analysis and properties

Positional encoding and multi-level MHCA The Positional Encoding (PE) allows to leverage the absolute position of the objects in the image. Table 5.4 shows an analysis of its impact, on one fold on both datasets. For MHSA, the PE improves the results by +0.7pt for TCIA and +0.6pt for IMO. For MHCA, we evaluate a single level of attention with and without PE.

Table 5.3: Results on the TCIA multiorgan dataset in Dice similarity coefficient (DSC, %) detailed per organ.

Organ	U-Net	MHSA	MHCA	U-Transformer
Spleen	96.21 (± 1.49)	96.32 (± 1.06)	96.53 (± 1.05)	**96.81** (± 0.94)
Kidney(L)	95.69 (± 0.58)	95.86 (± 0.65)	95.72 (± 0.64)	**96.03** (± 0.52)
Gallbladder	77.01 (± 7.88)	**80.20** (± 6.50)	78.24 (± 7.71)	78.65 (± 7.90)
Esophagus	**66.92** (± 4.59)	64.51 (± 5.79)	65.20 (± 6.87)	66.04 (± 6.49)
Liver	95.84 (± 0.35)	96.14 (± 0.39)	96.18 (± 0.20)	**96.33** (± 0.26)
Stomach	87.91 (± 2.63)	89.30 (± 3.21)	88.89 (± 2.78)	**89.70** (± 2.69)
Pancreas	70.86 (± 4.60)	74.38 (± 3.59)	74.13 (± 3.16)	**75.54** (± 3.44)
Duodenum	57.11 (± 4.77)	58.79 (± 4.25)	57.63 (± 4.23)	**59.01** (± 4.30)
Avg	80.94 (± 2.19)	81.52 (± 2.18)	81.57 (± 2.20)	**82.26** (± 2.16)

Table 5.4: Ablation study on the positional encoding and multi-level on one fold of TCIA and IMO.

			MHSA		MHCA		
	U-Net	Attn U-Net	wo PE –	w PE	1 lvl wo PE –	1 lvl w PE –	multi-lvl w PE
TCIA	76.35	77.23	78.17	**78.90**	77.18	78.88	**80.65**
IMO	88.18	87.52	88.16	**88.76**	87.96	88.52	**89.13**

We can observe an improvement of +1.7pts for TCIA and +0.6pt for IMO between the two versions.

Table 5.4 also shows the favorable impact of using multi *vs* single-level attention for MHCA: +1.8pts for TCIA and +0.6pt for IMO. It is worth noting that Attention U-Net uses multi-level attention but remains below MHCA with a single level. Figure 5.8 shows attention maps at each level of U-Transformer: level 3 corresponds to high-resolution features maps, and tends to focus on more specific regions compared to the first levels.

Further analysis To further analyse the behaviour of U-Transformer, we evaluate the impact of the number of attention heads for MHSA Figure 5.9: more heads lead to better performances, but the biggest gain comes from the first head (*i.e.* U-Net to MHSA). Finally, the evaluation of U-Transformer with respect to the Hausdorff distance Table 5.5 follows the same trend as with the Dice score. This highlights the capacity of U-Transformer to reduce prediction artefacts by means of self- and cross-attention. In addition, we have evaluated our method on a TCIA multiorgan extension which gives the same trends that with our IMO Table 5.3. Finally we trained another state-of-the-art architecture, nnU-Net [14], and also observed a significant gain

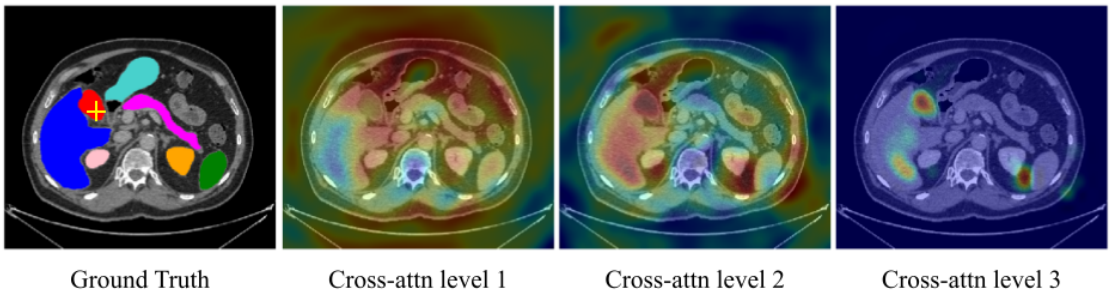


Figure 5.8: Cross-attention maps for the yellow-crossed pixel (left image).

Table 5.6.

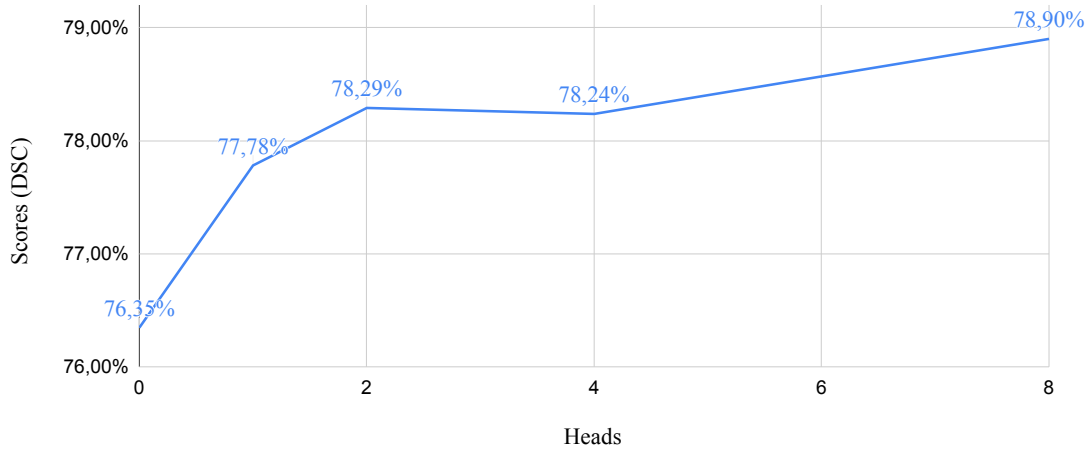


Figure 5.9: Evolution of the Dice Score on TCIA (fold 1) when the number of heads varies between 0 and 8 in MHSA.

Table 5.5: Hausdorff Distances (HD) for the different models

Dataset	U-Net	Attn U-Net	U-Transformer
TCIA	13.61 (± 2.01)	12.48 (± 1.36)	**12.34** (± 1.51)
IMO	12.06 (± 1.65)	12.13 (± 1.58)	**12.00** (± 1.32)

Ablation with nnU-Net We evaluated our approach using the nnU-Net architecture as our backbone on TCIA. The results show a gain of 1pt in Dice, which is a large improvement given the strong baseline. Moreover this result is statistically significant with a paired t-test (p=0.025). It highlights that our MHSA/MHCA modules improve performances over state-of-the-art convolutional models.

Table 5.6: Results by using nnU-Net as our baseline in Dice similarity coefficient (DSC, %).

Dataset	nnU-Net [14]	MHSA	MHCA	U-Transformer
TCIA	83.09 (± 1.01)	83.78 (± 0.91)	83.38 (± 0.91)	**84.08** (± 0.96)

5.5 Conclusion

In this chapter, we introduced the U-Transformer network which augments a U-shaped FCN based on U-Net [2] with Transformers [17]. We propose to use self and cross-attention modules to model long-range interactions and spatial dependencies. Moreover, the cross-attention module allows a finer spatial recovery in the decoder by filtering the skip-connections based on the semantic features from the previous decoder block. We highlight the relevance of the approach for abdominal organ segmentation, especially for small and complex organs such as the pancreas, gallbladder and stomach. U-Transformer has a lot of potential for medical image segmentation and could open the path to new network architectures using dense attention mechanisms from Transformers which have already shown great performances in other fields.

Chapter 6

Conclusions and Perspectives

Contents

6.1 Contributions

In this thesis, we addressed three major problems for medical image segmentation. Firstly, we studied how to train DL models with partially-labeled datasets. Most of the available datasets are not exhaustive, moreover, professionals often focus on specific structures leading to partially-labeled images. Then, we have studied how we could leverage the strong prior knowledge about the absolute position of the organs. For that, we proposed a prior which explicitly biases the final prediction and which is thus integrated into the network. Finally, we talked about the recent Transformer models and how to adapt them in the context of organ segmentation. FCNs have a limited RF which mechanically does not capture sufficient contextual information. We thus propose to use Transformers to leverage their capacity of modeling long range dependencies.

Partially-labeled data We first start with INERRANT in Chapter 3 which extends the work proposed in SMILE [1]. In this chapter, we proposed a method which trains DL models on partially-labeled data by focusing on the available labels. We strongly rely on the assumption that every organ of interest is visible in the input image, we can thus deduce which ones are unlabeled and ignore the associated voxels in the training loss. We further improve the method with a pseudo-labeling step which aims at relabeling the missing organs and improving the overall performances. The selection of the pseudo-labels are a critical step and having a good confidence measure is important. Thus, we used a dedicated network that learns to predict a confidence map. We experimentally show the relevance on organ segmentation datasets. Moreover, we compare with semi-supervised methods and show the superiority of the proposed solution.

Integrating Spatial Prior Knowledge In a second chapter, Chapter 4, we addressed a major problem of ConvNets which is their inability of modeling absolute spatial positions. However, organs in medical images have known positions which could be leveraged especially to improve the segmentation of difficult organs. Thus, we studied how to explicitly integrate a 3D spatial prior into a ConvNet. Based on the spatial positions of the organs observed in the training set, we build a 3D spatial prior which is then used to directly bias the ConvNet in an explicit manner. We show the relevance of this 3D spatial prior in the challenging pancreas segmentation task. Moreover, we studied how this prior could help in a pseudo-labeling scheme as presented in Chapter 3 and show that the prior is particularly relevant with small datasets.

Transformers for Organ Segmentation In the third chapter, Chapter 5, we proposed a network architecture, UNet-Transformer which combines the attention mechanisms from Transformers with a U-shaped FCN. Transformers allow to model long-range interactions and thus leverage complex relationships between structures. We proposed two main modules: a self-attention block in the bottleneck of the FCN which aims at modeling global interactions between semantic features; and a cross attention block which aims at filtering non-semantic features coming from the skip-connections based on high-level ones in the decoder. The proposed U-Transformer network shows important gains in organ segmentation even for the pancreas and also in a multi-organ setup. We observed the major gains on difficult organs such as the pancreas, the gallbladder and the stomach.

6.2 Perspectives for Future Works

Although we provide contributions at several levels, it remains room for improvement especially to generalize and further validate the works detailed in this thesis. In our experiments we mostly used the U-Net model as our backbone model. However, other architectures have been proposed lately and gives stronger baselines. For example the nnU-Net [14] pipeline is a very interesting work that redefines the training of U-Net like architectures. By using predefined rules, the proposed framework deduces rules for training the network and the hyper-parameters which should are best suited. We tried this framework with U-Transformer in Chapter 5 and show that we significantly boost the performances of our model, thus it could be a great baseline in future works.

Then, we mainly focused on the segmentation of abdominal organs in CT-scans. However, the proposed works are not limited to this context. It was more a choice of circumstances based on the available data, however it could be very interesting to evaluate our works on other contexts: with a different modality, MRI or US; with a different anatomical region: brain, heart, lungs, ... We should mentioned that evaluating in various context is one of the goal of the Decathlon dataset [63] presented in Section 2.2.4.

INERRANT In this work we proposed a way of training deep ConvNets on partially-labeled data. We found that it could be linked to semi-supervised learning and proposed a pseudo-labeling technique. However other methods exists as discussed in Section 2.3 and combining our proposed method with other approaches such as using adversarial training [5] or mean-teacher [75] would further validate our work. Moreover, the proposed solution is generic and could be easily transferred to other tasks or modalities. For example on MRI where datasets are even smaller than CT-scans datasets. Concerning the confidence estimation, we tried two

measures but the literature is vast in the area. Thus, other confidence measures could be used. We choose to train a dedicated network based on recent works [123] but this is an approach used for miss-classification detection. However, it could be very interesting to couple this with epistemic uncertainty which measures points that are far from the train set like in active learning [129].

STIPPLE As said before, MRI datasets are even smaller and few are publicly available. Knowing that, an interesting point that STIPPLE could bring is it capacity of being cross-modal. In fact, the spatial prior is computed based on the available annotations but are independent of the images. Thus, with adjustments especially on the voxel spacing which could differ, one could compute a spatial prior based on data from one modality and use it to train a model for another modality where the number of labeled data is limited. For example, a spatial prior could be learned on a large CT-scan dataset and used to train a model on MRI images which are often limited.

Another interesting point which could be studied is the impact of this spatial prior if used as different stages of the network. For example in the bottleneck or in the input. Moreover is could be used in conjunction of the CoordConv layers [83, 130] that shows it capacity of leveraging coordinates when given as a feature map. This way the two would collaborate to take the most of our spatial prior and directly encode it into the model still in an explicit manner.

U-Transformer In this work we proposed a U-shaped based architecture using attention mechanisms from Transformers. However, the paper [17] introduced a complete block integrating residual connections, normalizations and fully connected layers. In our experiments we found that using directly the attention gives better results and mitigate the parameter overhead. Following the works in vision [114, 9] using the complete block could gives better results is correctly tuned. Moreover, further experiments could be done using 3D networks. An important issue faced by Transformers is the need of large amount of memory which is also an important issue for 3D networks. Yet, using Transformers in a 3D FCN which could take full 3D volumes as input sounds like a perfect solution. To approach this solution, we could used an hybrid model with a FCN that encodes 3D patches of the input volume and reduces their dimensions. Then, a Transformer network could encode interactions between the patches and thus process the complete volume.

Longer-term Perspectives Beyond the individual contributions, more global perspectives could be considered. For example, to further develop the integration of prior knowledge one could think of more complex and richer information such as explicit dependencies and relationships

between organs. Moreover, the lack of explainability in the DL models is an important obstacle for efficiently integrating prior knowledge. With advances in this domain, new ways of using prior knowledge could be considered. Thus, coupling both model explainability and integration of prior knowledge seems crucial for future works.

On the other hand, in the perspective of training models on heterogeneous sources of data, a promising field is data adaptation and data generalization. Studying this task could allow us to train models on a source of data and apply the learned model on another one. Current solutions in medical image analysis do not generalize well as the image acquisition procedures are numerous and highly depend on the case. Finally, another promising path which is more and more studied is the combination of multiple modalities or data types. For example, cross-modal problems aim at training a model on multiple modalities (*e.g.* CT + MRI). Or it could also intend to leverage information coming from diagnosis or patient's history and physical examinations in a multi-model network.

Bibliography

[1] O. Petit, N. Thome, A. Charnoz, A. Hostettler et L. Soler, "Handling missing annotations for semantic segmentation with deep convnets," dans *Deep Learning in Medical Image Analysis and Multimodal Learning for Clinical Decision Support (DLMIA workshop MICCAI)*. Springer, 2018, p. 20–28.

[2] O. Ronneberger, P. Fischer et T. Brox, "U-net: Convolutional networks for biomedical image segmentation," dans *Medical Image Computing and Computer-Assisted Intervention – MICCAI 2015*, N. Navab, J. Hornegger, W. M. Wells et A. F. Frangi, édit. Cham: Springer International Publishing, 2015, p. 234–241.

[3] K. Simonyan et A. Zisserman, "Very deep convolutional networks for large-scale image recognition," dans *International Conference on Learning Representations (ICLR)*, 2015.

[4] V. Badrinarayanan, A. Kendall et R. Cipolla, "Segnet: A deep convolutional encoder-decoder architecture for image segmentation," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 39, n°. 12, p. 2481–2495, 2017.

[5] D. Nie, Y. Gao, L. Wang et D. Shen, "Asdnet: Attention based semi-supervised deep networks for medical image segmentation," dans *Medical Image Computing and Computer Assisted Intervention*, 2018, p. 370–378.

[6] Y. Li, L. Yuan et N. Vasconcelos, "Bidirectional learning for domain adaptation of semantic segmentation," dans *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2019, p. 6936–6945.

[7] Q. Yu, L. Xie, Y. Wang, Y. Zhou, E. K. Fishman et A. L. Yuille, "Recurrent saliency transformation network: Incorporating multi-stage visual cues for small organ segmentation," dans *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2018, p. 8280–8289.

[8] Z. Mirikharaji et G. Hamarneh, "Star shape prior in fully convolutional networks for skin lesion segmentation," dans *Medical Image Computing and Computer Assisted Intervention – MICCAI 2018*, A. F. Frangi, J. A. Schnabel, C. Davatzikos, C. Alberola-López et G. Fichtinger, édit. Cham: Springer International Publishing, 2018, p. 737–745.

[9] A. Dosovitskiy, L. Beyer, A. Kolesnikov, D. Weissenborn, X. Zhai, T. Unterthiner, M. Dehghani, M. Minderer, G. Heigold, S. Gelly, J. Uszkoreit et N. Houlsby, "An image is worth 16x16 words: Transformers for image recognition at scale," dans *International Conference on Learning Representations*, 2021.

[10] W. Luo, Y. Li, R. Urtasun et R. Zemel, "Understanding the effective receptive field in deep convolutional neural networks," dans *International Conference on Neural Information Processing Systems*, 2016, p. 4905–4913.

[11] O. Oktay, J. Schlemper, L. L. Folgoc, M. Lee, M. Heinrich, K. Misawa, K. Mori, S. McDonagh, N. Y. Hammerla, B. Kainz *et al.*, "Attention u-net: learning where to look for the pancreas," *MIDL*, 2018.

[12] Y. Bengio, J. Louradour, R. Collobert et J. Weston, "Curriculum learning," dans *Proceedings of the 26th Annual International Conference on Machine Learning*, ser. ICML '09, 2009, p. 41–48.

[13] M. P. Kumar, B. Packer et D. Koller, "Self-paced learning for latent variable models," dans *Advances in Neural Information Processing Systems 23*, J. D. Lafferty, C. K. I. Williams, J. Shawe-Taylor, R. S. Zemel et A. Culotta, édit. Curran Associates, Inc., 2010, p. 1189–1197.

[14] F. Isensee, P. F. Jaeger, S. A. Kohl, J. Petersen et K. H. Maier-Hein, "nnu-net: a self-configuring method for deep learning-based biomedical image segmentation," *Nature methods*, vol. 18, n°. 2, p. 203–211, 2021.

[15] A. Krizhevsky, I. Sutskever et G. E. Hinton, "Imagenet classification with deep convolutional neural networks," dans *NeurIPS*, 2012, p. 1097–1105.

[16] S. Armato, G. McLennan, L.Bidaut, M. McNitt-Gray, C. Meyer, A. Reeves et et al., "The lung image database consortium (lidc) and image database resource initiative (idri): a completed reference database of lung nodules on ct scans," *Medical Physics*, vol. 38, p. 915–931, 2011.

[17] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Ł. Kaiser et I. Polosukhin, "Attention is all you need," dans *NeurIPS*, 2017, p. 5998–6008.

[18] D. L. Pham, C. Xu et J. L. Prince, "Current methods in medical image segmentation," *Annual Review of Biomedical Engineering*, vol. 2, n°. 1, p. 315–337, 2000.

[19] N. Sharma et L. M. Aggarwal, "Automated medical image segmentation techniques," *Journal of medical physics/Association of Medical Physicists of India*, vol. 35, n°. 1, p. 3, 2010.

[20] D. Withey et Z. Koles, "Three generations of medical image segmentation: Methods and available software," *International Journal of Bioelectromagnetism*, vol. 9, n°. 2, p. 67–68, 2007.

[21] N. Otsu, "A threshold selection method from gray-level histograms," *IEEE Transactions on Systems, Man, and Cybernetics*, vol. 9, n°. 1, p. 62–66, 1979.

[22] M. Sezgin et B. Sankur, "Survey over image thresholding techniques and quantitative performance evaluation," *Journal of Electronic Imaging*, vol. 13, n°. 1, p. 146 – 165, 2004.

[23] R. Pohle et K. D. Toennies, "Segmentation of medical images using adaptive region growing," dans *Medical Imaging 2001: Image Processing*, M. Sonka et K. M. Hanson, édit., vol. 4322, International Society for Optics and Photonics. SPIE, 2001, p. 1337 – 1346.

[24] M. Dabass, S. Vashisth et R. Vig, "Effectiveness of region growing based segmentation technique for various medical images - a study," dans *Data Science and Analytics*, B. Panda, S. Sharma et N. R. Roy, édit., 2018, p. 234–259.

[25] C. Jia-xin et L. Sen, "A medical image segmentation method based on watershed transform," dans *The Fifth International Conference on Computer and Information Technology (CIT'05)*, 2005, p. 634–638.

[26] W. E. Higgins et E. J. Ojard, "Interactive morphological watershed analysis for 3d medical images," *Computerized Medical Imaging and Graphics*, vol. 17, n°. 4, p. 387–395, 1993.

[27] T. Kohlberger, M. Sofka, J. Zhang, N. Birkbeck, J. Wetzl, J. Kaftan, J. Declerck et S. K. Zhou, "Automatic multi-organ segmentation using learning-based segmentation and level set optimization," dans *Medical Image Computing and Computer-Assisted Intervention – MICCAI 2011*, G. Fichtinger, A. Martel et T. Peters, édit. Berlin, Heidelberg: Springer Berlin Heidelberg, 2011, p. 338–345.

[28] D. Cremers, O. Fluck, M. Rousson et S. Aharon, "A probabilistic level set formulation for interactive organ segmentation," dans *Medical Imaging 2007: Image Processing*, J. P. W. Pluim et J. M. Reinhardt, édit., vol. 6512, International Society for Optics and Photonics. SPIE, 2007, p. 304 – 312.

[29] A. Saito, S. Nawano et A. Shimizu, "Joint optimization of segmentation and shape prior from level-set-based statistical shape model, and its application to the automated segmentation of abdominal organs," *Medical Image Analysis*, vol. 28, p. 46 – 65, 2016.

[30] A. Wimmer, G. Soza et J. Hornegger, "A generic probabilistic active shape model for organ segmentation," dans *Medical Image Computing and Computer-Assisted Intervention – MICCAI 2009*, G.-Z. Yang, D. Hawkes, D. Rueckert, A. Noble et C. Taylor, édit. Berlin, Heidelberg: Springer Berlin Heidelberg, 2009, p. 26–33.

[31] T. Kohlberger, M. G. Uzunbaş, C. Alvino, T. Kadir, D. O. Slosman et G. Funka-Lea, "Organ segmentation with level sets using local shape and appearance priors," dans *Medical Image Computing and Computer-Assisted Intervention – MICCAI 2009*, G.-Z. Yang, D. Hawkes, D. Rueckert, A. Noble et C. Taylor, édit. Berlin, Heidelberg: Springer Berlin Heidelberg, 2009, p. 34–42.

[32] A. Klein, B. Mensh, S. Ghosh, J. Tourville et J. Hirsch, "Mindboggle: automated brain labeling with multiple atlases," *BMC medical imaging*, vol. 5, n°. 1, p. 7, 2005.

[33] Hyunjin Park, P. H. Bland et C. R. Meyer, "Construction of an abdominal probabilistic atlas and its application in segmentation," *IEEE Transactions on Medical Imaging*, vol. 22, n°. 4, p. 483–492, 2003.

[34] E. Schreibmann, D. Marcus et T. Fox, "Multiatlas segmentation of thoracic and abdominal anatomy with level set-based local search," *Journal of applied clinical medical physics / American College of Medical Physics*, vol. 15, p. 4468, 09 2014.

[35] R. Wolz, C. Chengwen, K. Misawa, M. Fujiwara, K. Mori et D. Rueckert, "Automated abdominal multi-organ segmentation with subject-specific atlas generation," *IEEE transactions on medical imaging*, vol. 32, p. 1723–1730, 06 2013.

[36] J. E. Iglesias et M. R. Sabuncu, "Multi-atlas segmentation of biomedical images: A survey," *Medical Image Analysis*, vol. 24, n°. 1, p. 205 – 219, 2015.

[37] H. Wang, J. W. Suh, S. R. Das, J. B. Pluta, C. Craige et P. A. Yushkevich, "Multi-atlas segmentation with joint label fusion," *IEEE transactions on pattern analysis and machine intelligence*, vol. 35, n°. 3, p. 611–623, 2012.

[38] A. Neumann et C. Lorenz, "Statistical shape model based segmentation of medical images," *Computerized Medical Imaging and Graphics*, vol. 22, n°. 2, p. 133 – 143, 1998.

[39] T. Okada, M. G. Linguraru, M. Hori, R. M. Summers, N. Tomiyama et Y. Sato, "Abdominal multi-organ segmentation from ct images using conditional shape–location and unsupervised intensity priors," *Medical Image Analysis*, vol. 26, n°. 1, p. 1–18, 2015.

[40] M. Hammon, A. Cavallaro, M. Erdt, P. Dankerl, M. Kirschner, K. Drechsler, S. Wesarg, M. Uder et R. Janka, "Model-based pancreas segmentation in portal venous phase contrast-enhanced ct images," *Journal of digital imaging*, vol. 26, n°. 6, p. 1082–1090, 2013.

[41] S. Osher et J. A. Sethian, "Fronts propagating with curvature-dependent speed: Algorithms based on hamilton-jacobi formulations," *Journal of computational physics*, vol. 79, n°. 1, p. 12–49, 1988.

[42] M. D. Zeiler et R. Fergus, "Visualizing and understanding convolutional networks," dans *Computer Vision – ECCV 2014*, D. Fleet, T. Pajdla, B. Schiele et T. Tuytelaars, édit. Cham: Springer International Publishing, 2014, p. 818–833.

[43] Y. Cun, I. Guyon, L. Jackel, D. Henderson, B. Boser, R. Howard, J. Denker, W. Hubbard et H. Graf, "Handwritten digit recognition: Applications of neural network chips and automatic learning," *IEEE Communications Society Magazine*, vol. 27, n°. 11, p. 41–46, nov. 1989.

[44] Y. LeCun, B. Boser, J. S. Denker, D. Henderson, R. E. Howard, W. Hubbard et L. D. Jackel, "Backpropagation Applied to Handwritten Zip Code Recognition," *Neural Computation*, vol. 1, n°. 4, p. 541–551, 12 1989.

[45] Y. Lecun, L. Bottou, Y. Bengio et P. Haffner, "Gradient-based learning applied to document recognition," *Proceedings of the IEEE*, vol. 86, n°. 11, p. 2278–2324, 1998.

[46] J. Long, E. Shelhamer et T. Darrell, "Fully convolutional networks for semantic segmentation," dans *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2015, p. 3431–3440.

[47] L.-C. Chen, G. Papandreou, I. Kokkinos, K. Murphy et A. Yuille, "Semantic image segmentation with deep convolutional nets and fully connected crfs," *International Conference on Learning Representations*, 2015.

[48] L.-C. Chen, Y. Zhu, G. Papandreou, F. Schroff et H. Adam, "Encoder-decoder with atrous separable convolution for semantic image segmentation," dans *Proceedings of the European conference on computer vision (ECCV)*, 2018, p. 801–818.

[49] D. Ciresan, A. Giusti, L. Gambardella et J. Schmidhuber, "Deep neural networks segment neuronal membranes in electron microscopy images," *Advances in neural information processing systems*, vol. 25, p. 2843–2851, 2012.

[50] Ö. Çiçek, A. Abdulkadir, S. S. Lienkamp, T. Brox et O. Ronneberger, "3d u-net: Learning dense volumetric segmentation from sparse annotation," dans *Medical Image Computing and Computer-Assisted Intervention – MICCAI 2016*, 2016, p. 424–432.

[51] F. Milletari, N. Navab et S. Ahmadi, "V-net: Fully convolutional neural networks for volumetric medical image segmentation," dans *2016 Fourth International Conference on 3D Vision (3DV)*, 2016, p. 565–571.

[52] Z. Zhou, M. M. Rahman Siddiquee, N. Tajbakhsh et J. Liang, "Unet++: A nested u-net architecture for medical image segmentation," dans *Deep Learning in Medical Image Analysis and Multimodal Learning for Clinical Decision Support*, 2018, p. 3–11.

[53] J. Chen, Y. Lu, Q. Yu, X. Luo, E. Adeli, Y. Wang, L. Lu, A. L. Yuille et Y. Zhou, "Transunet: Transformers make strong encoders for medical image segmentation," *arXiv preprint arXiv:2102.04306*, 2021.

[54] H. R. Roth, L. Lu, A. Farag, H.-C. Shin, J. Liu, E. B. Turkbey et R. M. Summers, "Deeporgan: Multi-level deep convolutional networks for automated pancreas segmentation," dans *Medical Image Computing and Computer-Assisted Intervention – MICCAI 2015*, N. Navab, J. Hornegger, W. M. Wells et A. Frangi, édit. Cham: Springer International Publishing, 2015, p. 556–564.

[55] E. Gibson, F. Giganti, Y. Hu, E. Bonmati, S. Bandula, K. Gurusamy, B. Davidson, S. P. Pereira, M. J. Clarkson et D. C. Barratt, "Automatic multi-organ segmentation on abdominal ct with dense v-networks," *IEEE Transactions on Medical Imaging*, vol. 37, n°. 8, p. 1822–1834, 2018.

[56] X. Li, H. Chen, X. Qi, Q. Dou, C.-W. Fu et P.-A. Heng, "H-denseunet: Hybrid densely connected unet for liver and tumor segmentation from ct volumes," *IEEE transactions on medical imaging*, vol. 37, n°. 12, p. 2663–2674, 2018.

[57] A. Reinke, L. Maier-Hein *et al.*, "Common limitations of performance metrics in biomedical image analysis," *MIDL - Short Paper*, 2021.

[58] H. R. Roth, A. Farag, E. B. Turkbey, L. Lu, J. Liu et R. M. Summers, "Data from pancreas-ct," dans *The Cancer Imaging Archive, (TCIA)*, 2016.

[59] E. Gibson, F. Giganti, Y. Hu, E. Bonmati, S. Bandula, K. Gurusamy, B. Davidson, S. P. Pereira, M. J. Clarkson et D. C. Barratt, "Multi-organ abdominal ct reference standard segmentations," *Dataset*, févr. 2018.

[60] A. E. Kavur, M. A. Selver, O. Dicle, M. Barış et N. S. Gezer, "CHAOS - Combined (CT-MR) Healthy Abdominal Organ Segmentation Challenge Data," *Dataset*, avr. 2019.

[61] Z. Lambert, C. Petitjean, B. Dubray et S. Ruan, "Segthor: Segmentation of thoracic organs at risk in ct images," *Dataset*, nov. 2020.

[62] A. L. Simpson, M. Antonelli, S. Bakas, M. Bilello, K. Farahani, B. Van Ginneken, A. Kopp-Schneider, B. A. Landman, G. Litjens, B. Menze *et al.*, "A large annotated medical image dataset for the development and evaluation of segmentation algorithms," *arXiv preprint arXiv:1902.09063*, 2019.

[63] M. Antonelli, A. Reinke, S. Bakas, K. Farahani, B. A. Landman, G. Litjens, B. Menze, O. Ronneberger, R. M. Summers, B. van Ginneken *et al.*, "The medical segmentation decathlon," *arXiv preprint arXiv:2106.05735*, 2021.

[64] O. Chapelle, B. Schölkopf et A. Zien, *Semi-Supervised Learning (Adaptive Computation and Machine Learning)*. The MIT Press, 2006.

[65] S. Sedai, D. Mahapatra, S. Hewavitharanage, S. Maetschke et R. Garnavi, "Semi-supervised segmentation of optic cup in retinal fundus images using variational autoencoder," dans *International Conference on Medical Image Computing and Computer-Assisted Intervention*. Springer, 2017, p. 75–82.

[66] A. V. Dalca, J. Guttag et M. R. Sabuncu, "Anatomical priors in convolutional networks for unsupervised biomedical segmentation," dans *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2018, p. 9290–9299.

[67] I. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville et Y. Bengio, "Generative adversarial nets," dans *Advances in Neural Information Processing Systems*, Z. Ghahramani, M. Welling, C. Cortes, N. Lawrence et K. Q. Weinberger, édit., vol. 27. Curran Associates, Inc., 2014, p. 2672–2680.

[68] W.-C. Hung, Y.-H. Tsai, Y.-T. Liou, Y.-Y. Lin et M.-H. Yang, "Adversarial learning for semi-supervised semantic segmentation," dans *Proceedings of the British Machine Vision Conference (BMVC)*, 2018.

[69] H. Zheng, L. Lin, H. Hu, Q. Zhang, Q. Chen, Y. Iwamoto, X. Han, Y.-W. Chen, R. Tong et J. Wu, "Semi-supervised segmentation of liver using adversarial learning with deep atlas prior," dans *International Conference on Medical Image Computing and Computer-Assisted Intervention*, 2019, p. 148–156.

[70] H. Kervadec, J. Dolz, É. Granger et I. B. Ayed, "Curriculum semi-supervised segmentation," dans *International Conference on Medical Image Computing and Computer-Assisted Intervention*. Springer, 2019, p. 568–576.

[71] Y. Zhou, Z. Li, S. Bai, X. Chen, M. Han, C. Wang, E. Fishman et A. Yuille, "Prior-aware neural network for partially-supervised multi-organ segmentation," dans *2019 IEEE/CVF International Conference on Computer Vision (ICCV)*, 2019, p. 10 671–10 680.

[72] A. Tarvainen et H. Valpola, "Mean teachers are better role models: Weight-averaged consistency targets improve semi-supervised deep learning results," dans *Advances in neural information processing systems*, 2017, p. 1195–1204.

[73] D. Berthelot, N. Carlini, I. Goodfellow, N. Papernot, A. Oliver et C. A. Raffel, "Mixmatch: A holistic approach to semi-supervised learning," dans *Advances in Neural Information Processing Systems*, 2019, p. 5049–5059.

[74] G. Bortsova, F. Dubost, L. Hogeweg, I. Katramados et M. de Bruijne, "Semi-supervised medical image segmentation via learning consistency under transformations," dans *International Conference on Medical Image Computing and Computer-Assisted Intervention*. Springer, 2019, p. 810–818.

[75] L. Yu, S. Wang, X. Li, C.-W. Fu et P.-A. Heng, "Uncertainty-aware self-ensembling model for semi-supervised 3d left atrium segmentation," dans *International Conference on Medical Image Computing and Computer-Assisted Intervention*, 2019, p. 605–613.

[76] Y. Grandvalet et Y. Bengio, "Semi-supervised learning by entropy minimization," dans *Advances in neural information processing systems*, 2005, p. 529–536.

[77] D.-H. Lee, "Pseudo-label: The simple and efficient semi-supervised learning method for deep neural networks," dans *Workshop on challenges in representation learning, ICML*, vol. 3, n°. 2, 2013.

[78] Y. Zou, Z. Yu, B. Vijaya Kumar et J. Wang, "Unsupervised domain adaptation for semantic segmentation via class-balanced self-training," dans *Proceedings of the European conference on computer vision (ECCV)*, 2018, p. 289–305.

[79] Y. Zou, Z. Yu, X. Liu, B. Kumar et J. Wang, "Confidence regularized self-training," dans *Proceedings of the IEEE International Conference on Computer Vision*, 2019, p. 5982–5991.

[80] W. Bai, O. Oktay, M. Sinclair, H. Suzuki, M. Rajchl, G. Tarroni, B. Glocker, A. King, P. M. Matthews et D. Rueckert, "Semi-supervised learning for network-based cardiac mr image segmentation," dans *Medical Image Computing and Computer-Assisted Intervention - MICCAI 2017*, 2017, p. 253–260.

[81] Y. Zhou, Y. Wang, P. Tang, S. Bai, W. Shen, E. Fishman et A. Yuille, "Semi-supervised 3d abdominal multi-organ segmentation via deep multi-planar co-training," dans *2019 IEEE Winter Conference on Applications of Computer Vision (WACV)*. IEEE, 2019, p. 121–140.

[82] Y.-X. Zhao, Y.-M. Zhang, M. Song et C.-L. Liu, "Multi-view semi-supervised 3d whole brain segmentation with a self-ensemble network," dans *International Conference on Medical Image Computing and Computer-Assisted Intervention*. Springer, 2019, p. 256–265.

[83] R. Liu, J. Lehman, P. Molino, F. P. Such, E. Frank, A. Sergeev et J. Yosinski, "An intriguing failing of convolutional neural networks and the coordconv solution," dans *NeurIPS*, 2018.

[84] Y. Taigman, M. Yang, M. Ranzato et L. Wolf, "Deepface: Closing the gap to human-level performance in face verification," dans *IEEE CVPR*, June 2014.

[85] H. R. Roth, L. Lu, N. Lay, A. P. Harrison, A. Farag, A. Sohn et R. M. Summers, "Spatial aggregation of holistically-nested convolutional neural networks for automated pancreas localization and segmentation," *Medical image analysis*, vol. 45, p. 94–107, 2018.

[86] H. Kakeya, T. Okada et Y. Oshiro, "3d u-japa-net: Mixture of convolutional networks for abdominal multi-organ ct segmentation," dans *MICCAI*, 2018, p. 426–433.

[87] H. R. Roth, H. Oda, X. Zhou, N. Shimizu, Y. Yang, Y. Hayashi, M. Oda, M. Fujiwara, K. Misawa et K. Mori, "An application of cascaded 3d fully convolutional networks for medical image segmentation," *Computerized Medical Imaging and Graphics*, vol. 66, p. 90–99, 2018.

[88] P. F. Christ, M. E. A. Elshaer, F. Ettlinger, S. Tatavarty, M. Bickel, P. Bilic, M. Rempfler, M. Armbruster, F. Hofmann, M. D'Anastasi *et al.*, "Automatic liver and lesion segmentation in ct using cascaded fully convolutional neural networks and 3d conditional random fields," dans *International Conference on Medical Image Computing and Computer-Assisted Intervention*. Springer, 2016, p. 415–423.

[89] M. Tang, Z. Zhang, D. Cobzas, M. Jagersand et J. L. Jaremko, "Segmentation-by-detection: A cascade network for volumetric medical image segmentation," dans *2018 IEEE 15th International Symposium on Biomedical Imaging (ISBI 2018)*. IEEE, 2018, p. 1356–1359.

[90] X. Feng, K. Qing, N. J. Tustison, C. H. Meyer et Q. Chen, "Deep convolutional neural network for segmentation of thoracic organs-at-risk using cropped 3d images," *Medical Physics*, vol. 46, n°. 5, p. 2169–2180, 2019.

[91] R. Trullo, C. Petitjean, B. Dubray et S. Ruan, "Multiorgan segmentation using distance-aware adversarial networks," *Journal of Medical Imaging*, vol. 6, n°. 1, p. 014001, 2019.

[92] O. Oktay, E. Ferrante, K. Kamnitsas, M. P. Heinrich, W. Bai, J. Caballero, S. A. Cook, A. de Marvao, T. Dawes, D. P. O'Regan, B. Kainz, B. Glocker et D. Rueckert, "Anatomically constrained neural networks (acnns): Application to cardiac image enhancement and segmentation," *IEEE Transactions on Medical Imaging*, vol. 37, p. 384–395, 2018.

[93] R. El Jurdi, C. Petitjean, P. Honeine, V. Cheplygina et F. Abdallah, "High-level prior-based loss functions for medical image segmentation: A survey," *Computer Vision and Image Understanding*, vol. 210, p. 103248, 2021.

[94] H. Kervadec, J. Dolz, M. Tang, E. Granger, Y. Boykov et I. Ben Ayed, "Constrained-cnn losses for weakly supervised segmentation," *Medical Image Analysis*, vol. 54, p. 88–99, 2019.

[95] D. Bahdanau, K. Cho et Y. Bengio, "Neural machine translation by jointly learning to align and translate," *arXiv preprint arXiv:1409.0473*, 2014.

[96] T. Luong, H. Pham et C. D. Manning, "Effective approaches to attention-based neural machine translation," dans *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, sept. 2015, p. 1412–1421.

[97] P. Anderson, X. He, C. Buehler, D. Teney, M. Johnson, S. Gould et L. Zhang, "Bottom-up and top-down attention for image captioning and visual question answering," dans *CVPR*, 2018.

[98] K. Xu, J. Ba, R. Kiros, K. Cho, A. Courville, R. Salakhudinov, R. Zemel et Y. Bengio, "Show, attend and tell: Neural image caption generation with visual attention," dans *Proceedings of the 32nd International Conference on Machine Learning*, vol. 37, 2015, p. 2048–2057.

[99] F. Wang, M. Jiang, C. Qian, S. Yang, C. Li, H. Zhang, X. Wang et X. Tang, "Residual attention network for image classification," dans *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2017, p. 3156–3164.

[100] S. Jetley, N. A. Lord, N. Lee et P. Torr, "Learn to pay attention," dans *International Conference on Learning Representations*, 2018.

[101] J. Hu, L. Shen et G. Sun, "Squeeze-and-excitation networks," dans *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2018, p. 7132–7141.

[102] Y. Wang, Z. Deng, X. Hu, L. Zhu, X. Yang, X. xu, P.-A. Heng et D. Ni, "Deep attentional features for prostate segmentation in ultrasound," dans *Medical Image Computing and Computer Assisted Intervention – MICCAI 2018*, 09 2018.

[103] C. Li, Q. Tong, X. Liao, W. Si, Y. Sun, Q. Wang et P.-A. Heng, "Attention based hierarchical aggregation network for 3d left atrial segmentation," dans *Statistical Atlases and Computational Models of the Heart. Atrial Segmentation and LV Quantification Challenges*, 2019, p. 255–264.

[104] J. Schlemper, O. Oktay, M. Schaap, M. Heinrich, B. Kainz, B. Glocker et D. Rueckert, "Attention gated networks: Learning to leverage salient regions in medical images," *Medical Image Analysis*, vol. 53, p. 197–207, 2019.

[105] A. G. Roy, N. Navab et C. Wachinger, "Concurrent spatial and channel squeeze & excitation in fully convolutional networks," dans *MICCAI*, vol. abs/1803.02579, 2018.

[106] A. Sinha et J. Dolz, "Multi-scale self-guided attention for medical image segmentation," *IEEE Journal of Biomedical and Health Informatics*, p. 1–1, 2020.

[107] J. Fu, J. Liu, H. Tian, Y. Li, Y. Bao, Z. Fang et H. Lu, "Dual attention network for scene segmentation," dans *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2019.

[108] J. Devlin, M.-W. Chang, K. Lee et K. Toutanova, "Bert: Pre-training of deep bidirectional transformers for language understanding," *arXiv preprint arXiv:1810.04805*, 2018.

[109] X. Wang, R. Girshick, A. Gupta et K. He, "Non-local neural networks," dans *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2018, p. 7794–7803.

[110] N. Carion, F. Massa, G. Synnaeve, N. Usunier, A. Kirillov et S. Zagoruyko, "End-to-end object detection with transformers," dans *European Conference on Computer Vision*. Springer, 2020, p. 213–229.

[111] L. Ye, M. Rochan, Z. Liu et Y. Wang, "Cross-modal self-attention network for referring image segmentation," dans *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2019, p. 10 502–10 511.

[112] H. Wang, Y. Zhu, B. Green, H. Adam, A. Yuille et L.-C. Chen, "Axial-deeplab: Stand-alone axial-attention for panoptic segmentation," dans *European Conference on Computer Vision*, 2020, p. 108–126.

[113] S. Zheng, J. Lu, H. Zhao, X. Zhu, Z. Luo, Y. Wang, Y. Fu, J. Feng, T. Xiang, P. H. Torr *et al.*, "Rethinking semantic segmentation from a sequence-to-sequence perspective with transformers," dans *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2021, p. 6881–6890.

[114] Z. Liu, Y. Lin, Y. Cao, H. Hu, Y. Wei, Z. Zhang, S. Lin et B. Guo, "Swin transformer: Hierarchical vision transformer using shifted windows," *arXiv preprint arXiv:2103.14030*, 2021.

[115] B. Rister, D. Yi, K. Shivakumar, T. Nobashi et D. L. Rubin, "Ct-org, a new dataset for multiple organ segmentation in computed tomography," *Dataset*, vol. 7, n°. 1, p. 1–9, 2020.

[116] X. Zhu, Z. Cheng, S. Wang, X. Chen et G. Lu, "Coronary angiography image segmentation based on pspnet," *Computer Methods and Programs in Biomedicine*, vol. 200, p. 105897, 2021.

[117] T. Küstner, S. Müller, M. Fischer, J. Weiß, K. Nikolaou, F. Bamberg, B. Yang, F. Schick et S. Gatidis, "Semantic organ segmentation in 3d whole-body mr images," dans *2018 25th IEEE International Conference on Image Processing (ICIP)*, 2018, p. 3498–3502.

[118] C. Ouyang, K. Kamnitsas, C. Biffi, J. Duan et D. Rueckert, "Data efficient unsupervised domain adaptation for cross-modality image segmentation," dans *Medical Image Computing and Computer Assisted Intervention – MICCAI 2019*, D. Shen, T. Liu, T. M. Peters, L. H. Staib, C. Essert, S. Zhou, P.-T. Yap et A. Khan, édit., 2019, p. 669–677.

[119] D. Hendrycks et K. Gimpel, "A baseline for detecting misclassified and out-of-distribution examples in neural networks," *Proceedings of International Conference on Learning Representations*, 2017.

[120] C. Guo, G. Pleiss, Y. Sun et K. Q. Weinberger, "On calibration of modern neural networks," dans *Proceedings of Machine Learning Research*, D. Precup et Y. W. Teh, édit., vol. 70. International Convention Centre, Sydney, Australia: PMLR, 06–11 Aug 2017, p. 1321–1330.

[121] Y. Gal et Z. Ghahramani, "Dropout as a bayesian approximation: Representing model uncertainty in deep learning," dans *international conference on machine learning*, 2016, p. 1050–1059.

[122] H. Jiang, B. Kim, M. Guan et M. Gupta, "To trust or not to trust a classifier," dans *Advances in Neural Information Processing Systems 31*, S. Bengio, H. Wallach, H. Larochelle, K. Grauman, N. Cesa-Bianchi et R. Garnett, édit. Curran Associates, Inc., 2018, p. 5541–5552.

[123] C. Corbière, N. Thome, A. Bar-Hen, M. Cord et P. Pérez, "Addressing failure prediction by learning model confidence," dans *Advances in Neural Information Processing Systems*, 2019, p. 2902–2913.

[124] J. M. J. Valanarasu, P. Oza, I. Hacihaliloglu et V. M. Patel, "Medical transformer: Gated axial-attention for medical image segmentation," *arXiv preprint arXiv:2102.10662*, 2021.

[125] Y. Zhang, H. Liu et Q. Hu, "Transfuse: Fusing transformers and cnns for medical image segmentation," *arXiv preprint arXiv:2102.08005*, 2021.

[126] A. Hatamizadeh, D. Yang, H. Roth et D. Xu, "Unetr: Transformers for 3d medical image segmentation," *arXiv preprint arXiv:2103.10504*, 2021.

[127] Y. Xie, J. Zhang, C. Shen et Y. Xia, "Cotr: Efficiently bridging cnn and transformer for 3d medical image segmentation," *arXiv preprint arXiv:2103.03024*, 2021.

[128] X. Zhu, W. Su, L. Lu, B. Li, X. Wang et J. Dai, "Deformable {detr}: Deformable transformers for end-to-end object detection," dans *International Conference on Learning Representations*, 2021.

[129] Y. Gal, R. Islam et Z. Ghahramani, "Deep bayesian active learning with image data," dans *International Conference on Machine Learning*. PMLR, 2017, p. 1183–1192.

[130] R. El Jurdi, C. Petitjean, P. Honeine et F. Abdallah, "Coordconv-unet: Investigating coordconv for organ segmentation," *IRBM*, 2021.

Appendix A

U-Net architecture

Table A.1: Details of the U-Net's blocks and layers used in the thesis. This architecture comes from U-Net [2]. Convolutions are given by conv(kernel_size, filters). The final two blocks: output_probabilities and confidence_network, are connected to the last block of the network, *i.e.* final_prediction. The overall number of parameters reaches 32M parameters including the confidence network which is around 0.8M parameters.

block name	output size	layer's parameters
input	$512 \times 512 \times 1$	
encoder_block_1	$256 \times 256 \times 64$	conv(3×3, 64) + relu conv(3×3, 64) + BN + relu $\rightarrow$ res_1 max_pool(2×2)
encoder_block_2	$128 \times 128 \times 128$	conv(3×3, 128) + relu conv(3×3, 128) + BN + relu $\rightarrow$ res_2 max_pool(2×2)
encoder_block_3	$64 \times 64 \times 256$	conv(3×3, 256) + relu conv(3×3, 256) + BN + relu $\rightarrow$ res_3 max_pool(2×2)
encoder_block_4	$32 \times 32 \times 512$	conv(3×3, 512) + relu conv(3×3, 512) + BN + relu $\rightarrow$ res_4 max_pool(2×2)
decoder_block_4	$64 \times 64 \times 1024$	conv(3×3, 1024) + relu conv(3×3, 1024) + BN + relu upsampling(2×2) conv(2×2, 512) + BN + relu concat(res_4)
decoder_block_3	$128 \times 128 \times 512$	conv(3×3, 512) + relu conv(3×3, 512) + BN + relu upsampling(2×2) conv(2×2, 256) + BN + relu concat(res_3)
decoder_block_2	$256 \times 256 \times 256$	conv(3×3, 256) + relu conv(3×3, 256) + BN + relu upsampling(2×2) conv(2×2, 128) + BN + relu concat(res_2)
decoder_block_1	$512 \times 512 \times 128$	conv(3×3, 128) + relu conv(3×3, 128) + BN + relu upsampling(2×2) conv(2×2, 64) + BN + relu concat(res_1)
final_prediction	$512 \times 512 \times 64$	conv(3×3, 64) + relu conv(3×3, 64) + relu
output_probabilities	$512 \times 512 \times nb\_classes$	conv(1×1, nb_classes) + {softmax;sigmoid}
confidence_network	$512 \times 512 \times nb\_classes$	conv(3×3, 400) + relu conv(3×3, 120) + relu conv(3×3, 64) + relu conv(3×3, 64) + relu conv(1×1, nb_classes) + sigmoid

le cnam

Olivier PETIT

Segmentation sémantique d'images médicales 3D par deep learning

HESAM UNIVERSITÉ

Résumé : L'apprentissage profond a récemment montré des résultats impressionnants en vision par ordinateur. En particulier avec les réseaux de neurones convolutifs (ConvNets) qui ont redéfini l'état-de-l'art dans de nombreuses applications telles que la segmentation d'images médicales. Dans cette thèse nous abordons des problèmes en segmentation d'organes de l'abdomen en utilisant ces modèles. Premièrement, nous nous sommes intéressés à l'entraînement de ConvNets avec des bases de données partiellement étiquetées. Les professionnels se concentrant souvent sur des régions anatomiques précises, les bases de données sont de ce fait hétérogènes et partiellement étiquetées. Entraîner un modèle de segmentation directement donne de très mauvais résultats. Nous proposons donc un schéma d'entraînement qui utilise toutes les étiquettes disponibles sans être affecté par les mauvaises. De plus, un schéma itératif permet de progressivement ré-étiqueter les organes manquants ce qui permet d'améliorer encore notre modèle. La seconde partie étudie l'utilisation d'un *a priori* spatial sur la position absolue des organes afin d'améliorer la détection des structures et réduire les erreurs aberrantes. Les ConvNets sont par construction incapables de capturer l'information de position spatiale absolue. Cependant, les images médicales sont très structurées et les positions des organes sont connues. Dans ces travaux nous proposons d'utiliser un a priori spatial 3D qui capture la position des organes et qui va explicitement biaiser le modèle grâce à une fonction d'activation « prior-driven ». Pour finir, nous étudions les Transformers qui permettent de modéliser des interactions à long terme entre les structures anatomiques dans un modèle de segmentation. Les ConvNets ne permettent pas de capturer ces interactions globales principalement à cause de leur champ réceptif limité. Utiliser le mécanisme d'attention proposé dans les Transformers permet de connecter tous les pixels entre eux, ayant pour effet de modéliser des interactions complexes. Nous proposons le modèle U-Transformer et montrons qu'il améliore la qualité de la segmentation sur plusieurs bases de données.

Mots clés : segmentation sémantique ; apprentissage profond ; imagerie médicale

Abstract : Deep Learning has recently shown impressive results in computer vision. Especially with the Convolutional Neural Networks (ConvNets) which have redefined the state of the art in many applications such as medical image segmentation. In this thesis we address problems in the task of abdominal organ segmentation using deep learning models. In the first part, we address the issue of training deep ConvNets on partially labeled data. Professionals often focus on specific anatomical regions leading to heterogeneous datasets with partially labeled images. Training a model directly on such data leads to very poor results. Thus, we propose a training scheme that leverages all the labels without being affected by the missing ones. Moreover, an iterative scheme relabels the missing organs of the training set which further improves the segmentation model. The second part aims at using spatial prior about the position of the organs to improve the detection of structures and reduce outliers in the segmentation. ConvNets by construction, does not capture absolute spatial information. However, medical images are very structured and there are conventions about the expected position of organs. Thus, we propose a 3D spatial prior that captures the spatial position of organs and then explicitly biases the model through a prior-driven activation function. Finally, we propose to use Transformers to model long range dependencies between anatomical structures in a segmentation model used for organ segmentation. ConvNets do not capture such interactions because of the receptive field which is often limited. Using dense attention introduced in Transformers allows to connect every pixel with each other and thus to model complex interactions on different parts of the input image. We propose U-Transformer and show that it improves the quality of the segmentation on various datasets.

Keywords: semantic segmentation; deep learning; medical imaging