Titre: Title:	Méthodes faiblement supervisées d'apprentissage profond pour la segmentation inter-modalités de tumeurs en imagerie médicale
Auteur: Author:	Malo Alefsen de Boisredon d'Assier
Date:	2023
Type:	Mémoire ou thèse / Dissertation or Thesis
Référence: Citation:	Alefsen de Boisredon d'Assier, M. (2023). Méthodes faiblement supervisées d'apprentissage profond pour la segmentation inter-modalités de tumeurs en imagerie médicale [Master's thesis, Polytechnique Montréal]. PolyPublie. https://publications.polymtl.ca/54396/

POLYTECHNIQUE MONTRÉAL
affiliée à l'Université de Montréal

Méthodes faiblement supervisées d'apprentissage profond pour la segmentation inter-modalités de tumeurs en imagerie médicale

MALO ALEFSEN DE BOISREDON D'ASSIER

Institut de génie biomédical

Mémoire présenté en vue de l'obtention du diplôme de *Maîtrise ès sciences appliquées*

Génie biomédical

Juillet 2023

POLYTECHNIQUE MONTRÉAL
affiliée à l'Université de Montréal

Ce mémoire intitulé :

Méthodes faiblement supervisées d'apprentissage profond pour la segmentation inter-modalités de tumeurs en imagerie médicale

présenté par **Malo ALEFSEN DE BOISREDON D'ASSIER**
en vue de l'obtention du diplôme de *Maîtrise ès sciences appliquées*
a été dûment accepté par le jury d'examen constitué de :

Benjamin DE LEENER, président
Samuel KADOURY, membre et directeur de recherche
Laurent LÉTOURNEAU-GUILLON, membre externe

DÉDICACE

À Karma et Polka,
à ma famille,
à mes amis...

REMERCIEMENTS

En premier lieu je tiens à chaleureusement remercier mon directeur de recherche, Samuel Kadoury, pour m'avoir supervisé et permis de progresser en autonomie au sein de son laboratoire. Son exigence et sa volonté de pousser chacun au maximum de son potentiel m'auront servi de carburant durant ces deux années de maitrise.

Merci également à Eugene Vorontsov et William Trung-Le pour leur disponibilité et leur bienveillance. Leur rigueur scientifique et leurs conseils furent d'une aide précieuse et guidèrent mes réflexions au cours de mon projet.

Je remercie mes collègues du laboratoire MedICAL, avec qui j'ai partagé des moments studieux intenses, pour avoir fourni une ambiance chaleureuse et propice au développement de soi. Je leur souhaite une bonne continuation et le meilleur pour leur recherche.

Merci à ma famille de m'avoir soutenu depuis la France.

Finalement, je remercie tous les amis que j'ai pu rencontrer à Montréal. Merci à eux d'avoir fait de ces deux merveilleuses années ce qu'elles ont été.

RÉSUMÉ

Permettant de fournir des indications précieuses pour un diagnostic clinique plus rapide et plus précis, les méthodes de vision par ordinateur pour l'analyse d'images médicales ont suscité beaucoup d'intérêt au cours de la dernière décennie dans la communauté de la recherche médicale. En particulier, parmi les différents domaines de l'analyse d'images médicales, la segmentation de tumeurs a fait continuellement objet d'études. Dans ce contexte, les méthodes par apprentissage profond ont démontré un potentiel conséquent pour segmenter des tumeurs dans des clichés médicaux, et ce de manière rapide, précise et automatisée. Cependant, le succès de la plupart des méthodes existantes repose sur la disponibilité d'annotations précises, qui sont pénibles et coûteuses en temps à produire. En outre, malgré les efforts inlassables des chercheurs, il est reconnu que les algorithmes d'apprentissage automatique - y compris les méthodes d'apprentissage profond - ne sont pas adaptés aux changements de distributions. Ces lacunes constituent un obstacle au déploiement de ces méthodes à plus grande échelle. En particulier, la généralisation inter-modalités est essentielle lorsqu'une modalité d'imagerie manque de données annotées pour l'entraînement d'un réseau. Par exemple, elle pourrait être nécessaire dans les tâches de segmentation IRM où les annotations seraient disponibles en nombre suffisant pour des séquences T1, mais manqueraient pour les séquences FLAIR. Ce défi peut être relevé grâce à des approches semi-supervisées d'adaptation de domaines, qui exploitent les annotations disponibles dans la modalité "source" pour apprendre la distribution des tumeurs dans la modalité d'imagerie "cible" manquant d'annotations.

Dans cette thèse, nous présentons M-GenSeg et MoDATTS, deux nouvelles stratégies d'apprentissage semi-supervisé pour la segmentation de tumeurs inter-modalités sur des ensembles de données bi-modales non pairées. En apprenant à translater les images entre les différentes modalités, les annotations disponibles dans la modalité source sont exploitées par ces modèles pour apprendre la distribution des tumeurs dans la modalité cible non annotée. En outre, avec l'ajout d'images connues comme étant saines, un objectif non supervisé encourage M-GenSeg et MoDATTS à retirer les tumeurs de l'arrière-plan, ce qui est similaire à une tâche de segmentation.

M-GenSeg, le premier modèle que nous proposons, est une méthode convolutionnelle 2D. Nous avons évalué ses performances sur un jeu de données de segmentation de tumeurs cérébrales en IRM multi-paramétrique provenant du challenge BraTS 2020, pour lequel nous proposons une version adaptée au problème de segmentation entre modalités. Des ensembles de données pour la tâche d'adaptation de domaine inter-modale ont été constitués à partir de

chaque paire des quatre contrastes IRM disponibles (T1, T2, T1ce et FLAIR). Nous avons ainsi pu expérimenter avec douze combinaisons différentes de modalités source/cible non appariées. En moyenne sur toutes ces paires, et pour une tâche d'adaptation de domaine standard (respectivement 100% et 0% d'annotations dans les modalités source et cible), M-GenSeg démontre un gain absolu en score de segmentation de Dice de respectivement 0.04 et 0.08 par rapport à AttENT et AccSegNet, deux méthodes de pointe pour l'adaptation de domaine entre modalités. De plus, contrairement aux méthodes antérieures, M-GenSeg offre la possibilité de s'entraîner avec une modalité source partiellement annotée. En conséquence, sur la tâche d'adaptation T1 $\rightarrow$ T2, M-GenSeg maintient un score de Dice de 0.734 ± 0.008 (contre 0.478 ± 0.011 et 0.423 ± 0.007 pour AttENT et AccSegNet) dans la modalité cible (T2) non annotée avec seulement 1% d'annotations disponibles dans la modalité source (T1). De fait M-GenSeg s'illustre comme une méthode permettant de réduire efficacement la dépendance des réseaux à la disponibilité de données annotées.

Les réseaux de segmentation les plus performants étant des modèles basés sur des convolutions 3D, nous avons exploré dans une deuxième étude l'utilisation d'images 3D complètes plutôt que des coupes 2D, ainsi que des architectures plus optimales. Nous proposons donc MoDATTS, un modèle palliant aux défauts de M-GenSeg en permettant l'entraînement sur des volumes complets. MoDATTS profite aussi d'une nouvelle architecture intégrant des blocs transformers, ainsi que d'une stratégie itérative d'auto-entrainement. Sur le jeu de données de tumeurs cérébrales mentionné précédemment MoDATTS montre un gain de performance considérable par rapport à M-GenSeg. En effet pour une adaptation de domaine standard sur BraTS, 95% de la performance d'un modèle entrainé avec des données cibles annotées a été atteint par MoDATTS contre 87% pour M-GenSeg. De plus, sur une tâche de segmentation de schwannomes vestibulaires entre modalités T1ce et hrT2 (T2 haute résolution), nous relevons un score de segmentation de Dice de 0.87 ± 0.04, surpassant la performance des participants au challenge d'adaptation de domaine CrossMoDA 2022. Pour une adaptation entre modalités T1 $\rightarrow$ T2 sur BraTS, alors qu'aucune image n'est annotée dans la modalité cible (T2), 1% des annotations fournies dans la modalité source (T1) suffisent à atteindre 88% de la performance d'un modèle entrainé avec toutes les annotations de la modalité cible (T2), démontrant ainsi la robustesse de MoDATTS face à un déficit important en annotations. De plus amples expériences ont montré que 99% de cette performance maximale (obtenue par supervision dans la modalité cible) peut être atteinte en annotant seulement 20% des données cibles, additionnellement à celles fournies dans la modalité source (à hauteur de 100%). MoDATTS ouvre ainsi la voie vers une diminution considérable des coûts d'annotation, que l'on peut réduire de 80% en exploitant une modalité déjà annotée, et sans dégradation conséquente de la performance (1% environ).

ABSTRACT

As it can provide insightful guidance for faster and more accurate clinical diagnosis, computer vision for medical image analysis has fostered lots of interest among the medical research community in the last decade. Notably, among various medical image analysis challenges, tumor segmentation has been unceasingly studied. In this context, deep learning methods have demonstrated their substantial potential for automated, fast and accurate medical image tumor segmentation. However, the success of most existing pipelines rely on the availability of pixel-level annotations, which are time-consuming and difficult to produce. Besides, in spite of tireless efforts of researchers, it is acknowledged that machine learning algorithms - deep learning methods included - are not adequately fitted for distribution shifts. These shortcomings constitute a barrier to the deployment of such methods at a larger scale. In particular, cross-modality generalization is key when one imaging modality lacks annotated training examples. For instance, it could be needed in MRI segmentation tasks where annotations are sufficiently available for T1-weighted, but lacking for FLAIR sequences. This challenge can be tackled through semi-supervised domain-adaptive approaches, which leverage the pixel-level annotations available in the "source" modality to learn the tumor distribution in the "target" imaging modality lacking annotations.

In this thesis we present M-GenSeg and MoDATTS, two new semi-supervised training strategy for cross-modality tumor segmentation on unpaired bi-modal datasets. By leveraging available pixel-level annotations from the source modality through cross-modality image generation, both models enable to segment tumors in the unannotated target modality. Besides, with the addition of known healthy images, an unsupervised objective encourages the model to disentangle tumors from the background, which parallels the segmentation task.

M-GenSeg, the first model that we propose, is a 2D convolutional method. We evaluated its performance on a multi-parametric MRI brain tumor segmentation dataset from the BraTS 2020 challenge, for which we provide an adapted version for cross-modality segmentation. Datasets for inter-modality domain adaptation tasks were created from each pair of the four available MRI contrasts (T1, T2, T1ce, and FLAIR). This allowed us to experiment with twelve different combinations of unpaired source/target modalities. On average across all these pairs, and for a standard domain adaptation task (100% and 0% annotations in the source and target modalities, respectively), M-GenSeg demonstrates an absolute gain in Dice segmentation score of 0.04 and 0.08 in comparison to AttENT and AccSegNet, two state-of-the-art methods for cross-modality domain adaptation. Furthermore, unlike previous

methods, M-GenSeg offers the ability to train with partially annotated source modality. As a result, in the T1 $\rightarrow$ T2 adaptation task, M-GenSeg maintains a Dice score of 0.734 ± 0.008 (compared to 0.478 ± 0.011 and 0.423 ± 0.007 for AttENT and AccSegNet) in the unannotated target modality (T2) with only 1% annotations available in the source modality (T1). Thus, M-GenSeg proves to be an effective method to reduce the dependency of neural networks to the availability of annotated data.

As top performing methods are 3D-based models, in a second study we explored the use of full 3D images rather than 2D slices, along with more optimal architectures. We thus propose MoDATTS, an improved version of M-GenSeg, conceived to allow training on full volumes. MoDATTS benefits from powerful architectures based on visual transformers, and introduces an iterative self-training strategy. On the aforementioned brain tumor dataset, MoDATTS shows a considerable performance improvement over M-GenSeg. Indeed, for a standard domain adaptation on BraTS, MoDATTS achieves 95% of the performance of a model trained with annotated target data, compared to 87% for M-GenSeg. Furthermore, in a vestibular schwannoma segmentation task between T1ce $\rightarrow$ hrT2 (high-resolution T2) modalities, we achieve a Dice segmentation score of 0.87 ± 0.04, surpassing the performance of participants in the CrossMoDA 2022 domain adaptation challenge. For a domain adaptation task between T1 $\rightarrow$ T2 modalities on BraTS, with no annotations in the target modality (T2), 1% of the annotations provided in the source modality (T1) are sufficient to achieve 88% of the performance of a model trained with all annotations in the target modality (T2), demonstrating the robustness of MoDATTS when facing a significant annotation deficit. Further experiments have shown that 99% of this maximum performance (achieved by supervision in the target modality) can be attained by annotating only 20% of the target data, in addition to those provided in the source modality (100% of annotations available). MoDATTS thus paves the way for a substantial reduction in annotation costs, which can be reduced by 80% by leveraging an already annotated modality, and with minimal degradation in performance (approximately 1%).

TABLE DES MATIÈRES

LISTE DES TABLEAUX

LISTE DES FIGURES

LISTE DES SIGLES ET ABRÉVIATIONS

IRM	Imagerie par Résonance Magnétique
CT	Computed Tomography
TE	Temps d'Echo
TR	Temps de Répétition
FLAIR	Fluid-Attenuated Inversion Recovery
T1ce	T1 Contrast-Enhanced
PET	Tomographie par Emission de Positrons
FDG	Fluorodésoxyglucose
DSC	Dice Similiraty Coefficient
JSC	Jaccard Similiraty Coefficient
TP	True Positive
FP	False Positive
FN	False Negative
RVD	Relative Volume Difference
HD	Hausdorff Distance
ASSD	Average Symetric Surface Distance
kNN	k-Nearest Neighbors
SVM	Support Vector Machine
MLP	Multi Layer Perceptron
GPU	Graphical Processing Unit
ROI	Region Of Interest
RNN	Recurrent Neural Network
FCN	Fully Convolutional Network
GAN	Generative Adversarial Network
VS	Schwannome Vestibulaire

CHAPITRE 1 INTRODUCTION

Selon l'Organisation Mondiale de la Santé, le cancer est la deuxième cause de décès dans le monde, avec environ 10 millions de morts chaque année. Pouvant affecter n'importe quel organe du corps et se présenter sous de nombreuses formes différentes, le traitement de cette pathologie est rendu complexe et coûteux. En plus des coûts directs liés au traitement, le cancer peut également avoir des répercussions financières importantes sur les patients et leur famille, en raison des coûts indirects tels que les pertes de revenus. La prévention et la détection précoce de tumeurs, permettant de considérablement améliorer les chances de guérison pour les patients, sont donc des enjeux majeurs pour la santé publique.

L'imagerie médicale joue un rôle crucial dans la détection de ces tumeurs. Les techniques d'imagerie médicale telles que la tomodensitométrie (CT) ou l'imagerie par résonnance magnétique (IRM) sont utilisées pour visualiser les tissus internes du corps humain. Ces techniques permettent de détecter les anomalies, y compris les tumeurs, pouvant être difficiles à détecter à l'examen clinique. La segmentation de lésions tumorales sur des clichés médicaux est une étape cruciale dans le processus de diagnostic et de traitement. En effet, délimiter la tumeur permet d'accéder à des informations précieuses sur la taille, la forme et la position de la tumeur, ainsi que sur la présence de métastases. Ces données sont essentielles pour évaluer l'étendue du cancer et déterminer un plan de traitement adapté, et personnalisé, pour chaque patient. Selon la nature de la tumeur, sa dimension et sa localisation le médecin pourra se diriger vers une chirurgie ablative, une chimio-thérapie ou encore une radio-thérapie.

Le procédé de segmentation requiert un haut niveau d'expertise, rendant la tâche difficilement délégable. De plus la tâche est coûteuse en temps, elle peut ainsi mobiliser l'attention d'un professionnel de santé pendant plusieurs dizaines de minutes. A noter également que chaque opération humaine est associée un certain biais, ce qui peut être problématique quand la tumeur évolue dans un environnement sensible.

Avec les avancées récentes dans les techniques d'apprentissage machine et de vision par ordinateur, les deux dernières décennies ont vu apparaitre des méthodes de segmentation nécessitant une interaction avec le médecin, puis des algorithmes totalement automatisés. Ces nouveaux outils, offrant un gain de temps considérable ainsi qu'une objectivité accrue, ouvrent la voie vers une amélioration significative de la qualité des soins de santé pour les patients atteints de cancer.

Actuellement, les meilleures dispositifs pour le traitement d'images - données médicales y comprises - reposent sur des réseaux de neurones et l'apprentissage profond, où des millions

de poids organisés en couche successives sont optimisés pour réaliser une tâche donnée (recalage, segmentation, etc.). Ces réseaux se démarquent par leur modularité et leur performance impressionantes en comparaison aux méthodes d'apprentissage machine traditionnelles. Cependant, le succès de ces méthodes repose sur la disponibilité d'annotations précises, qui sont pénibles et coûteuses en temps à produire. De plus, les réseaux neuronaux profonds généralisent mal à des distributions auxquelles ils n'ont pas été exposés durant leur entrainement. Ces limitations sont un défi majeur au déploiement de ces outils à plus grande échelle. Dans le domaine médical en particulier, de nombreux facteurs peuvent être la source d'un décalage entre la distribution des données d'entrainement du réseau et la distribution d'application de ce dernier. Le simple fait de changer de population (variablité anatomique et pathologique différente), de site d'acquisition (paramètres d'acquisition ou modèles de machine différents) ou encore de méthode d'acquisition (modalités différentes) peut fortement impacter la performance d'un modèle lors de l'inférence. De fait, développer un modèle permettant la généralisation, sans annotations supplémentaires, à des données présentant un tel décalage constitue un enjeu central.

Dans ce projet de maîtrise on se limite au problème de la généralisation entre modalités pour la segmentation de lésions tumorales. Répondre à cette problématique revient à réaliser une adaptation de domaine entre différentes modalités i.e. exploiter les annotations disponibles dans une modalité source pour apprendre la distribution des tumeurs dans une modalité cible non annotée. Une telle stratégie est nécessaire lorsqu'une modalité d'imagerie manque de données annotées pour l'entraînement d'un réseau. Par exemple, elle pourrait être utile dans les tâches de segmentation IRM où les annotations seraient disponibles en nombre suffisant pour les séquences T1, mais manqueraient pour les séquences T2. De même, si l'on manque de données d'entrainement pour segmenter des acquisitions CT, une adaptation inter-modalités à partir de données IRM serait un gain majeur.

Ainsi, dans ce mémoire on propose de développer une pipeline d'apprentissage profond semi-supervisée permettant de segmenter des tumeurs sur des données bi-modales lorsqu'une des deux modalités présente un fort déficit en annotations. Il s'agira d'implémenter un modèle génératif basé sur la translation d'images entre modalités. On évaluera de plus le potentiel d'une approche permettant d'étendre l'entraînement à des images connues comme étant saines, ainsi que d'une stratégie d'auto-entrainement.

1.1 Plan du mémoire

Le chapitre 2 présente la revue de littérature. Les sujets sont répartis entre imagerie et diagnostics de tumeurs, notions d'apprentissage profond et applications pour l'adaptation de domaine à travers les modalités. Les notions médicales abordées comprennent quelques concepts d'imagerie ainsi que son intêret pour la détection et la segmentation de tumeurs. Les notions d'apprentissage profond abordées comprennent des concepts généraux d'apprentissage machine, le perceptron multicouche, les réseaux de neurones convolutifs, les réseaux de type transformer, les réseaux encodeurs-décodeurs, les réseaux antagonistes génératifs et l'apprentissage multi-tâches. Les applications de l'apprentissage profond étudiées comprennent la translation de modalité et l'adaptation de domaine pour la segmentation de tumeurs inter-modalités.

Le troisième chapitre expose la méthodologie mise en place dans le cadre de la recherche et détaille les différentes étapes impliquées dans la conception des modèles proposés : M-GenSeg et MoDATTS. On y présente rapidement les deux approches développées, ainsi que la méthode d'évaluation de ces dernières. De plus, ce chapitre énumère les publications incorporées dans les chapitres 4 et 5.

Le chapitre 4 présente le premier article portant sur M-GenSeg, une méthode de segmentation de tumeurs pour des données multi-modales où peu d'annotations sont disponibles dans une modalité (ou aucune). En agençant de multiples encodeurs et décodeurs convolutifs 2D, deux approches génératives de translation d'images (entre modalités et entre domaines sain/malade) y sont combinées pour permettre la segmentation de tumeurs dans la modalité cible manquant d'annotations. On y explore aussi l'utilisation de mécanismes d'attention pour les différentes composantes de segmentation et de translation de modalités.

Le chapitre 5 présente le deuxième article portant sur MoDATTS, une version étendue et améliorée de M-GenSeg (développée dans le premier article). Afin de pallier les limites d'une segmentation réalisée sur des coupes 2D, MoDATTS permet un entrainement sur des volumes 3D complets. De plus, l'utilisation d'une architecture basée sur des blocs transformers est investiguée dans cette seconde étude. Enfin, le potentiel d'une stratégie d'auto-entrainement y est aussi explorée.

Le chapitre 6 conclut ce mémoire avec une discussion des méthodes développées, en présentant les difficultés rencontrées, les solutions mises en place pour y faire face, les limitations qui peuvent persister ainsi que les possibilités d'amélioration futures.

CHAPITRE 2 REVUE DE LITTÉRATURE

2.1 Imagerie, modalités et détection de tumeurs

Une tumeur peut affecter n'importe quel organe du corps et se présenter sous de nombreuses formes différentes, faisant de la détection de cette dernière une tâche complexe. L'imagerie médicale joue alors un rôle crucial dans la diagnostic de ces tumeurs. Diverses méthodes d'imagerie sont aujourd'hui utilisées pour visualiser les tissus internes du corps humain de manière non-invasive. Ces dernières permettent de détecter des anomalies - dont les tumeurs - pouvant être difficiles à détecter à l'examen clinique. Le choix de l'imagerie dépendra de plusieurs facteurs, notamment la localisation possible de la tumeur, sa nature supposée, la disponibilité de l'appareil d'imagerie et les préférences du médecin traitant. Les modalités les plus utilisées en clinique restent cepdendant l'imagerie par résonance magnétique et l'imagerie par tomodensimétrie.

2.1.1 Imagerie par résonance magnétique

Pour entrevoir l'anatomie interne du corps humain, l'imagerie par résonance magnétique constitue l'une des modalités d'imagerie les plus complètes. Cette technologie se base sur les propriétés magnétiques des atomes d'hydrogène que le corps humain, constitué à 65% d'eau, contient en grande quantité. Les protons H+ des noyaux des atomes d'hydrogène, dont le spin est constamment en mouvement, peuvent être considérés comme de petits aimants. Lorsqu'un patient est placé dans un champ magnétique intense, les spins s'alignent, créant une aimantation mesurable. Cette aimantation peut être représentée par un vecteur $\overrightarrow{M}$ qu'on décompose en une composante transversale M_{xy} et longitudinale M_z. Sans perturbation les spins nucléaires sont dirigés dans la direction du champ stable et sont en phase (i.e. $M_{xy} = 0$ et $\overrightarrow{M} = M_z \cdot \overrightarrow{z}$).

Une impulsion d'ondes éléctromagnétiques peut être utilisée pour exciter les protons, provoquant une rotation de leur aimantation. Comme montré en Fig. 2.1, modifier les paramètres de l'impulsion (durée et amplitude) permet de rotationner le champ d'un angle α choisi. En IRM, une pulsation à 90° place le champ dans le plan (xy). Après la perturbation, les protons reviennent à leur état stable et relâchent l'énergie accumulée, on parle de relaxation. On peut alors introduire deux temps de relaxation caractéristiques : (i) T1, correspondant au temps de rétablissement de la valeur d'équilibre de la composante longitudinale M_z ; (ii) et T2, correspondant à l'affaiblissement de la magnétisation transversale M_{xy}. Après une

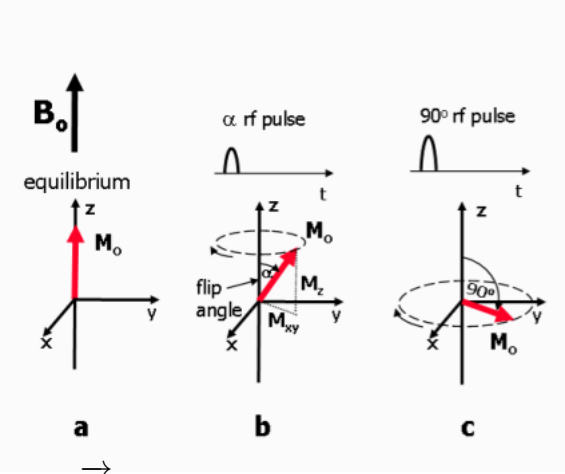
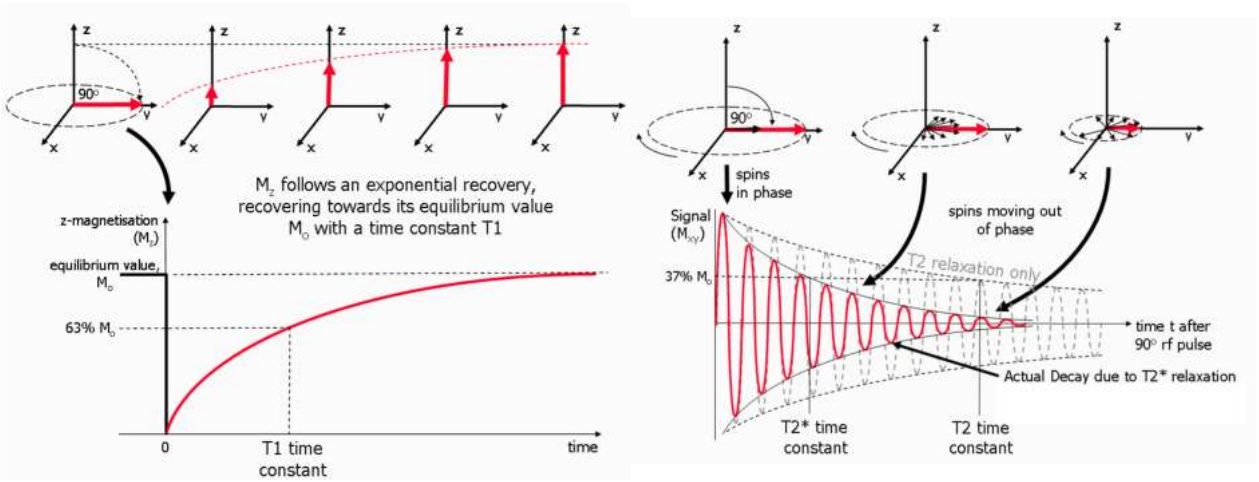


FIGURE 2.1 (a) Aimantation $\overrightarrow{M}$: (a) à l'équilibre, (b) après une impulsion d'angle α, (c) après une impulsion d'angle 90°. Image extraite de [1].

impulsion unique les spins se retrouvent rapidement en phase et la composante transversale devient donc rapidement nulle (i.e. $M_{xy} = 0$). Ce temps de relaxation T2* est trop court pour pouvoir mesurer correctement le signal correspondant. Pour pallier cela, on génère une impulsion à 180° un temps $\frac{TE}{2}$ après l'impulsion à 90°, remettant les spins en phase et provoquant un écho dans l'aimantation transversale ($M_{xy} \neq 0$). Le maximum de rephasage de la composante transversale est alors obtenu au temps d'écho TE. En répétant cet écho avec une période TE on génère une nouvelle enveloppe décroissante, de temps de relaxation T2 (>T2*). Ces variables sont illustrées en Fig. 2.2 et 2.3. On nomme aussi temps de répétition (TR) le temps entre deux pulsations à 90°.



(a) (b)

FIGURE 2.2 (a) Temps de relaxation T1. (b) Temps de relaxation T2* et T2. Extrait de [1].

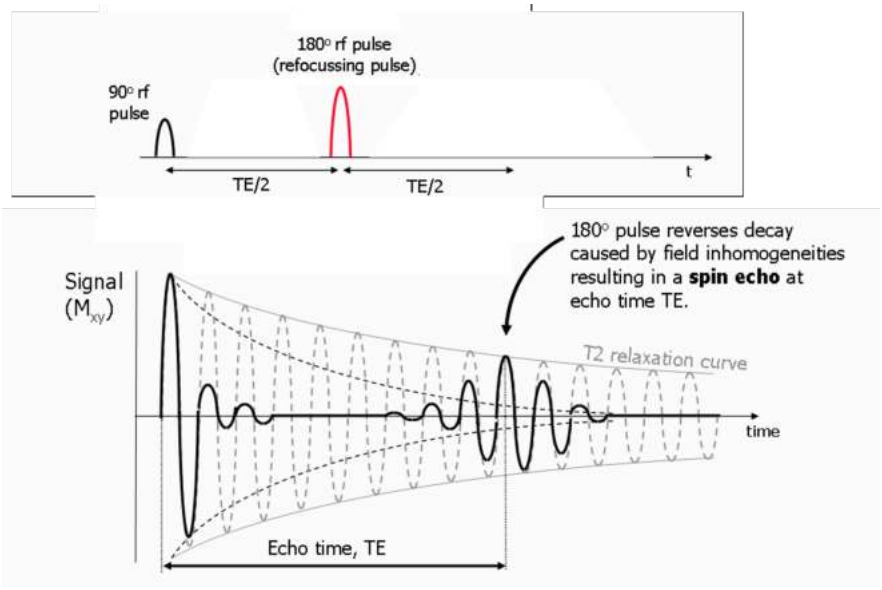


FIGURE 2.3 Génération d'un écho par application successive d'une pulsation à 90° puis 180°. Adapté de [1].

Chaque tissu possède des temps caractéristiques de relaxation T1 et T2 différents, selon sa composition et son agencement moléculaire [26]. La graisse a par exemple un temps de relaxation T1 très faible, alors qu'un tissu essentiellement aqueux possède un temps de relaxation T1 long. En jouant sur les paramètres impulsionels TR et TE, il est possible d'atténuer séparement l'impact de T1 et T2 sur le signal acquis. Considérons deux tissus A et B ayant des temps de relaxation T1 et T2 distincts. En utilisant un TR long, l'impact de T1 sur le signal est négligeable, car entre deux excitations la composante longitudinale de l'aimantation M_z des deux tissus a eu le temps de revenir à l'équilibre. En revanche avec un TR court, l'aimantation M_z de A et de B est dans un état de rétablissement différent, provoquant une différence de contraste entre les deux tissus. De la même manière, en utilisant un TE court i.e. en appliquant la pulsation à 180° directement après la pulsation à 90°, le différentiel de relaxation des composantes transversales M_{xy} entre A et B est négligeable. En revanche, avec un TE long la composante transversale de l'aimantation n'est pas la même au moment de l'écho pour les tissus A et B [1].

En utilisant un TR court (200-500 ms) et un TE court (15-30 ms), on obtient ainsi une séquence pondérée en T1, dont le contraste est relatif au différentiel de relaxation T1 entre les tissus. Les séquences T1 sont souvent utilisées comme images fonctionnelles, dans la mesure où la détection de l'eau mobile (extracellulaire ou intravasculaire) y est favorisée.

Inversement, avec un TR long (2500 ms) et un TE long (100-200 ms), on observe donc des

contrastes uniquement dépendants de la caractéristique T2 des tissus. On parle de séquences pondérées en T2. Elles sont souvent utilisées pour visualiser l'anatomie, puisqu'on y favorise la détection de l'eau peu mobile, c'est à dire intracellulaire (e.g. signal élevé pour la substance grise et faible pour les os).

En imagerie cérébrale, il peut être difficile de différencier une lésion du liquide cérébrospinal sur des séquences T2. En choisissant un TR et TE encore plus long que pour les séquences T2, on atténue le signal lié au liquide. Ces séquences FLAIR (Fluid-Attenuated Inversion Recovery) permettent donc de mieux distinguer l'anomalie, puisque cette dernière reste hyper-intense (par rapport au liquide cérébrospinal).

Pour améliorer le contraste de ces images, des agents peuvent être injectés par voie veineuse. Ces derniers vont modifier localement les paramètres électromagnétiques intrinsèques des tissus qu'ils vont intégrer. Le gadolinium par exemple, en s'accumulant dans les tissus pathologiques (tumeurs, zones d'inflammation/infection) peut induire des signaux d'hyper-intensité en imagerie pondérée T1 (séquence T1ce) [27].

Par essence, une tumeur prolifère dans un tissu sain et sa composition diffère de celle de son environnement. De fait, l'IRM est parfaitement adaptée à sa détection. Elle constitue par exemple la modalité la plus sensible pour détecter des tumeurs du cerveau, ce dernier étant composé à 90% d'eau [28]. Comme montré en Fig. 2.4, il est possible de combiner les quatre modalités introduites ci-dessus pour obtenir une vue plus détaillée de la tumeur. Les séquences T1 vont notamment permettre de réaliser une analyse structurale et distinguer les tissus sains. En modalité T1ce les frontières de la tumeur apparaissent hyper-intenses, et permettent de différencier la partie nécrotique de la tumeur de sa composante active. L'œdème apparait quant-à lui comme une région brillante en séquence T2, et est différenciable du liquide cérébro-spinal grâce à la séquence FLAIR.

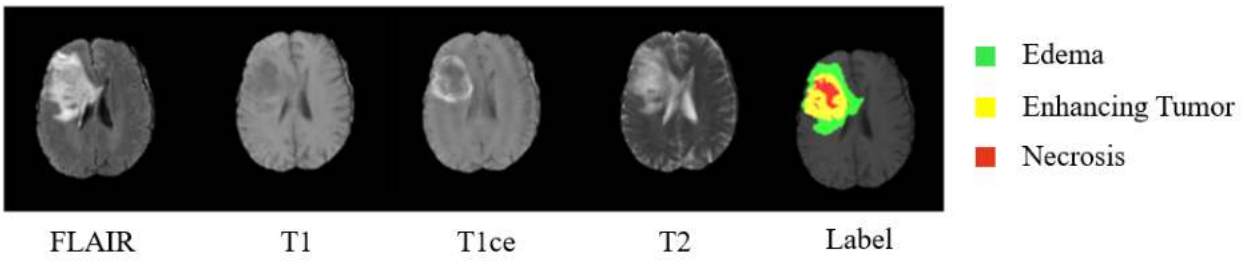
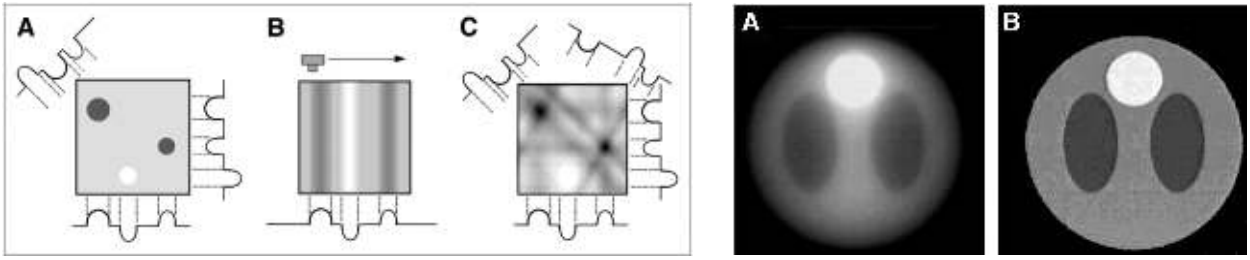


FIGURE 2.4 Segmentation de tumeurs en IRM cérébral, adapté de [2].

2.1.2 Tomodensimétrie

Le scanner CT aussi appelé imagerie par tomodensimétrie est de même une modalité d'imagerie standard, mais reste par ailleurs moins coûteux que l'IRM [29]. Pour créer des images tridimensionnelles du corps la technologie repose sur des rayons X, ce qui peut permettre de détecter des tumeurs qui ne sont pas visibles sur d'autres types d'images. L'objet d'intérêt est analysé par tranches successives, sur lesquelles des rayons sont projetés selon plusieurs directions à 360°. Chaque direction de mesure correspond à une mesure d'atténuation. En accumulant ces mesures, il est possible de déterminer la valeur de l'atténuation propre à chaque élément de volume. Ce principe, appelé reconstruction tomographique [30] est illustré en Fig. 2.5a. Comme visible en Fig. 2.5b, l'image ainsi recueillie est souvent floutée et des méthodes algorithmiques de filtrage sont nécessaires pour reconstruire l'image le plus fidèlement possible [3].



(a) (b)

FIGURE 2.5 (a) Des rayons X sont envoyés dans une direction (**B**) sur un fantôme contenant 3 objets aux valeurs d'atténuation différentes (**A**). L'accumulation du signal sur diverses directions permet de reconstruire l'image (**C**). (b) Reconstructions d'une image avec (**A**) et sans (**B**) filtrage. La version filtrée présente des bords plus affinés. Extraits de [3].

Ainsi, l'intensité lumineuse relative de l'image finale reflète l'absorption spécifique de chaque élément de volume. Une tumeur, ayant des caractéristiques d'absorption différentes des tissus sains est donc distinguable de son environnement sur une image CT. Cela fait de la tomodensimétrie une modalité de choix pour imager et détecter des tumeurs pulmonaires [31], pancréatiques [5], hépatiques [32] ou encore reinales [7] (voir Fig. 2.6).

Un scanner CT peut aussi être couplé à de la tomographie par émission de positrons (TEP) pour faire un suivi plus approndi de tumeurs (voir Fig. 2.7). En effet, lors d'un scanner PET un produit radio-actif injecté dans le patient permet de détecter les zones à haute activité métabolique, où cet agent va s'accumuler. Les tumeurs étant de grandes consommatrices de glucose, en utilisant des molécules de glucose marquées radio-activement (FDG) on peut obtenir des informations précises sur la localisation de ces lésions cancéreuses [33].

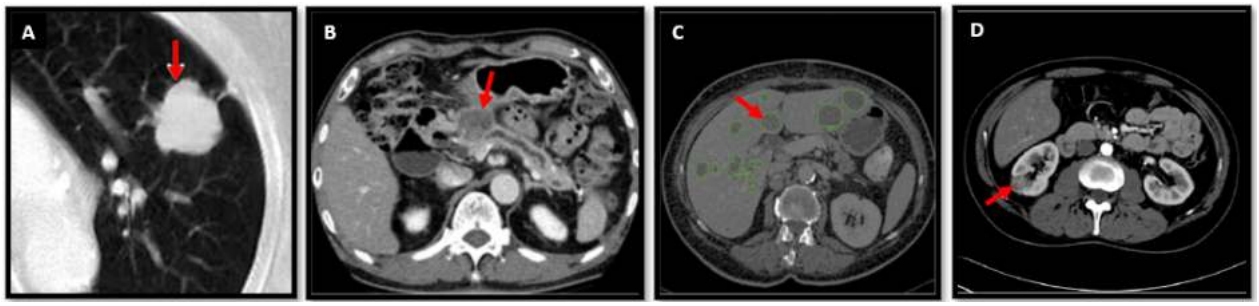
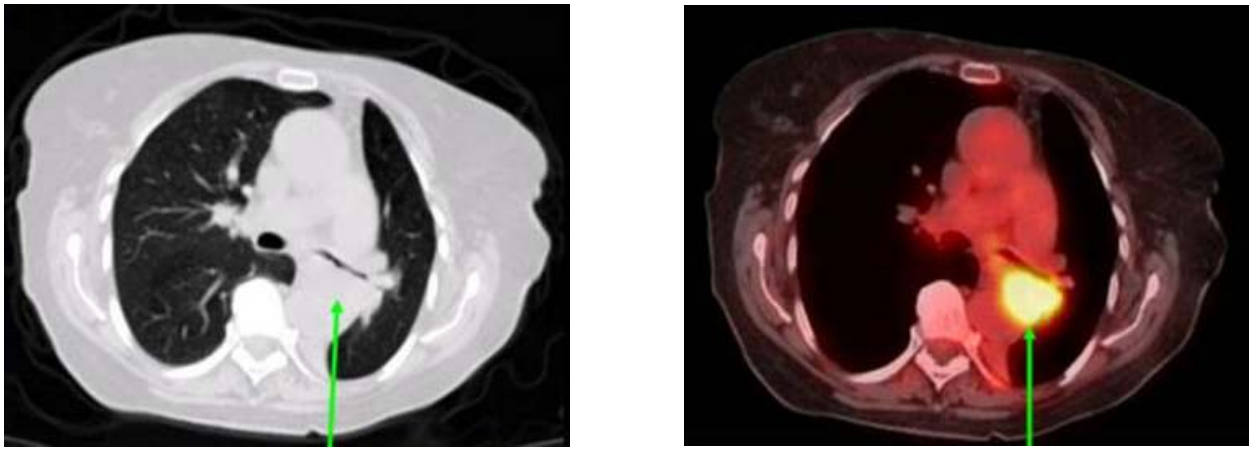


FIGURE 2.6 Images CT de poumon (A), de pancréas (B), de foie (C) et du rein (D). Diverses tumeurs sont localisées par des flèches rouges. Images tirées de [4–7].



(a) (b)

FIGURE 2.7 (a) Tumeur pulmonaire difficilement délimitable sur un scan CT. (b) Même tumeur, gourmande en FGD, mise en évidence par une fusion PET-CT. Images extraites de [8].

2.2 Segmentation de tumeurs

2.2.1 Une étape essentielle pour le traitement

Une fois la tumeur identifiée et imagée, établir un traitement nécessite plusieurs étapes importantes, comprenant la segmentation de cette lésion. Cette étape permet d'obtenir des informations précieuses sur la tumeur comme son volume ou sa localisation précise par rapport aux structures environnantes potentiellement sensibles. Si le médecin pense qu'une chirurgie ablative est envisageable, les contours de la tumeur sont utilisés pour planifier la résection (i.e. enlever la tumeur en entier ainsi qu'une marge de tissu sain) [34, 35]. Si la tumeur n'est pas opérable, la segmentation permet d'établir des plans de dose pour une radio-thérapie [36, 37]. Par la suite les segmentations sont utilisées pour faire un suivi de l'évolution de la pathologie

en réponse au traitement [38].

En pratique cette opération est réalisée à la main, ce qui est coûteux en temps et fastidieux. Un autre problème avec la segmentation manuelle est qu'elle est soumise à des variations, pour un même annotateur et aussi entre annotateurs. On parle de variabilité intra et inter observateur [39, 40]. Les méthodes de contours assistées ont pour objectif d'uniformiser et de faciliter la production des contours [41, 42].

2.2.2 Evaluer une segmentation

Tous les algorithmes de segmentation d'images médicales doivent être validés et comparés. Diverses métriques d'évaluation populaires existent dans la littérature, mais chacune est sensible à un type d'erreur de segmentation différent (taille, localisation et forme) [43].

Métriques de superposition

Une première classe de métriques mesure le chevauchement entre la segmentation de la tumeur prédite et celle de référence. Parmi ces dernières on compte :

- La précision qui calcule le volume de la lésion correctement segmentée par rapport au volume total prédit. Elle ne prend pas en compte les erreurs de sous-segmentation.
- Elle est souvent utilisée de pair avec le rappel [44], qui comptabilise le volume de la région correctement segmentée par rapport au volume total de la vérité de terrain. Le rappel ne prend pas en compte les erreurs de sous-segmentation.
- L'indice de Sørensen–Dice, qui est l'indice le plus répandu. Plus connu sous le nom de coefficient de similarité de Dice (DSC), il prend ses valeurs dans l'intervalle [0,1] où 1 signifie qu'il y a un chevauchement parfait et 0 correspond à des segmentations disjointes [45, 46]. On peut le définir en termes de classification par pixel de vrais positifs (TP), faux positifs (FP) et faux négatifs (FN) :

$$DSC = \frac{2 \cdot TP}{2 \cdot TP + FP + FN}$$

- Le coefficient de similarité de Jaccard (JSC) [47], directement relié au coefficient de Dice par la relation :

$$JSC = \frac{DSC}{2 - DSC}$$

Métrique volumique

La différence de volume relative (RVD) mesure la différence absolue de volumes entre la prédiction et la vérité de terrain, comme fraction du volume de la segmentation de référence. Elle est couramment utilisée en combinaison avec d'autres mesures [45, 48–50].

Métriques de distance à la frontière

La distance entre un voxel x et un volume A peut s'écrire $d(x, A) = \min_{y \in A} \|x - y\|_1$. Basé sur cette définition de la distance, plusieurs métriques servent à quantifier la dissimilarité entre la vérité de terrain et la prédiction :

- La distance de Hausdorff (HD) correspond à la plus grande distance de tous les points d'un ensemble A à leur point le plus proche dans un ensemble B. Elle est régulièrement utilisée pour la segmentation d'organes ou de lésions tumorales [47, 51].

$$d_{HD}(A, B) = \max\{\max_{a \in A} d(a, B), \max_{b \in B} d(A, b)\}$$

  .

- La distance de surface symétrique (ASSD) [48–50] calcule la distance moyenne entre chaque point d'une surface A et le point le plus proche d'une autre surface B.

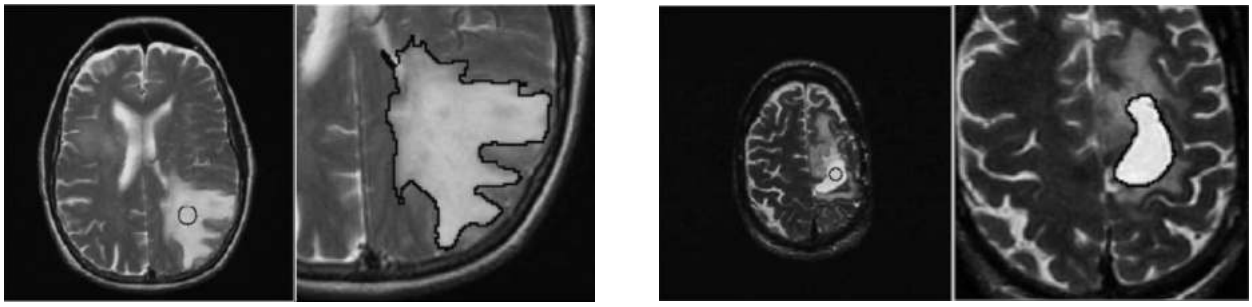
$$d_{ASSD}(A, B) = \frac{\sum_{a \in A} d(a, B) + \sum_{b \in B} d(A, b)}{|A| + |B|}$$

2.2.3 Méthodes de segmentation

Les méthodes algorithmiques de segmentation peuvent êtres séparées en deux classes (i) les méthodes semi-automatiques, autrement appelées méthodes interactives, reposant sur l'intervention de l'utilisateur pour réaliser la segmentation, (ii) et les méthodes totalement automatisées, développées pour éliminer toute interaction humaine. Qu'elles soient interactives ou totalement automatiques, ces méthodes peuvent être regroupées en différentes classes selon le modèle utilisé : modèles déformables, de regroupement de pixels ou de classification.

Modèles déformables

Les modèles déformables impliquent l'utilisation d'une interface de propagation (une courbe fermée en 2D et une surface fermée en 3D). A partir d'un contour initialisé par l'utilisateur autour de la structure à délimiter (ici la tumeur), le contour est modifié de manière itérative en appliquant des opérations de rétrecissement/expansion visant à minimiser une fonction



(a) (b)

FIGURE 2.8 Segmentation d'une tumeur cérébrale sur deux séquences T2 avec un modèle déformable. Les deux cas (a) et (b) montrent le contour utilisé pour initialiser le modèle (gauche) et la segmentation finale obtenue après minimisation de la fonction d'énergie (droite). Dans le cas (b) la méthode ne permet pas de segmenter la globalité de la tumeur (œdème exclu). Images extraites de [9].

d'énergie (voir Fig. 2.8). Cette dernière tient compte des propriétés locales et globales de l'image, servant de contraintes pour le déplacement de l'interface. Formulés sous forme de modèles paramétriques (*snakes*) initialement introduits par Kass et al. [52], ou sous forme de modèles géométriques (*level set*) dont la théorie a été développée par Osher et al. [53], les modèles déformables ont été amplement exploités et adaptés pour la segmentation en imagerie médicale [9, 54–60].

Pour rendre le procédé totalement automatisé, plusieurs méthodes ont par la suite été développées afin d'initialiser le contour de la tumeur. Par exemple, en segmentation de tumeur cérébrale, Khotanlou et al. [61] ont initialisé leur modèle en exploitant la symétrie du cerveau. En supposant que la lésion n'apparait que d'un côté, ils ont en effet localisé la tumeur en comparant les histogrammes de l'hémisphère gauche et droit. Ho et al. [62] ont utilisé la différence entre des séquences IRM T1 et T1ce comme carte de caractéristiques pour un modèle de mélange gaussien (GMM). Ils ont ainsi obtenu une carte de probabilité correspondant à la tumeur, ensuite utilisée pour initialiser la segmentation. De la même manière, Prastawa et al. [63] ont initialisé le contour d'un modèle déformable en se basant sur des cartes de probabilités d'anomalie, obtenues en recalant les volumes cérébraux à un atlas.

Cependant, les modèles déformables dépendent fortement des gradients de l'image et si les frontières de la tumeur à segmenter ne sont pas assez bien définies, la méthode est fortement susceptible d'échouer. De plus, si l'environnement de la tumeur présente des gradients de forte intensité, il y a un risque majeur que le contour actif soit dirigé dans la mauvaise direction. Notons aussi qu'il n'est pas trivial d'intégrer plusieurs modalités dans ces algorithmes. De plus, ce sont fondamentalement des méthodes de segmentation à deux classes (background vs

objet d'intérêt), il est donc difficile de réaliser une segmentation multi-classes, comme requis par exemple pour les tumeurs cérébrales.

Regroupement non supervisé de pixels

Des méthodes non supervisées de *clustering* - où les pixels d'une image sont regroupés en sous-ensembles homogènes (cluster) de pixels en fonction de leurs caractéristiques (intensité, texture, etc.) - ont aussi été adaptées pour la segmentation de tumeurs en imagerie médicale. Les pixels d'une même tumeur sont supposés constituer un de ces clusters. Par exemple, Linguraru et al. [64] ont adopté la méthode de "normalized cut" pour segmenter des tumeurs du foie. On y construit un graphe où chaque noeud correspond à un voxel et chaque arête est associée à la mesure de similarité entre les deux voxels connectés. En se basant sur la théorie spectrale des graphes on peut réaliser une partition de ce dernier telle que la similarité des pixels au sein d'un même sous-ensemble soit maximale, mais minimale entre les noeuds de deux sous-ensembles distincts. La méthode est cependant interactive car l'utilisateur doit finaliser la segmentation en choisissant quel sous-ensemble correspond à la tumeur. La même méthode a été adaptée par Archip et al. [65] pour la segmentation de tumeurs du cerveau. Pour la même application Capelle et al. [66] ont aussi exploité la théorie des graphes en modélisant la probabilité conditionnelle d'appartenance de chaque pixel d'une IRM à des clusters, grâce aux champs aléatoires de Markov.

Du fait que ces méthodes soient non supervisées, plusieurs défauts majeurs y sont associés [67]. En effet, les contours peu définis et l'inhomogénéité de certaines tumeurs comprenant plusieurs régions (œdème, nécrose, etc.) en réduisent fortement la performance. De plus, le nombre de régions servant à subdiviser l'image doit souvent être spécifié en amont.

Classification de pixels

Les modèles de classification se basent principalement sur l'apprentissage machine, où des images préalablement segmentées (partiellement ou totalement) sont explorées et exploitées par un algorithme afin d'attribuer à chaque pixel d'une image une classe d'appartenance (tissu sain, œdème, tumeur active, etc.). Dans le cas semi-automatique, l'utilisateur est amené à annoter quelques pixels de chaque classe sur une image à segmenter. L'algorithme utilise alors ces pixels comme données d'entraînement, afin de propager la segmentation au reste de l'image. Vinitski et al. [68] par exemple utilisent un algorithme des k plus proches voisins (kNN) basé sur quelques voxels annotés en entrée par l'utilisateur. Au moment de la prédiction, pour chaque voxel à segmenter l'algorithme trouve les k voxels les plus proches (ici en terme d'intensité), la classe majoritaire parmi ces k voxels est alors attribuée au

voxel d'entrée. L'algorithme de fuzzy c-means dans sa version semi-supervisée (seeded fuzzy c-means [69]), a par la suite surpassé la méthode basée sur les plus proches voisins [70]. A partir de l'initialisation, l'algorithme calcule le centroïde de chaque classe, et les pixels ayant un degré de proximité au centre d'une classe supérieur à un certain seuil prédéfini sont alors associés à cette dernière. La méthode est ainsi réitérée à partir de ce nouvel état, jusqu'à ce que la totalité de l'image soit segmentée. De même, basés sur la sélection de quelques voxels d'une tumeur, Zhang et al. [71] ont entrainé une machine à vecteurs de support (SVM) pour discriminer entre le tissu sain et la lésion. Des caractéristiques spatiales peuvent aussi être ajoutées aux intensités de l'image pour complexifier le modèle de segmentation [72, 73].

Il est possible de se débarrasser de toute interaction humaine si le modèle de classification a déjà été entraîné en amont. Cela requiert notamment d'avoir préalablement accès à des données annotées pour superviser l'algorithme. La première étape de ces méthodes consiste à construire un espace de caractéristiques à plusieurs dimensions regroupant diverses informations spécifiques relatives à chaque pixel (intensités, textures, informations de voisinage, etc.). On parle plus communément d'extraction des *features*. En Fig. 2.9 on montre un exemple de features de texture extraits d'une séquence cérébrale FLAIR [10]. L'algorithme défini se base alors sur l'ensemble de ces features pour son entraînement, puis par la suite pour l'inférence de segmentations sur des images externes. Basés sur des séquences T1, T2 et FLAIR, Pereira et al. [74] ont par exemple extrait 300 features différents pour entraîner une forêt d'arbres décisionnels (*random forest* [75]) à segmenter des tumeurs cérébrales. D'autres algorithmes, comme des classifieurs bayésiens [76] ou des SVM [77] ont aussi été exploités. Le choix des caractéristiques jouant un rôle essentiel dans la capacité des méthodes à performer, une étape de sélection des features peut être ajoutée au procédé. Pour réduire la dimension de l'ensemble des caractéristiques construit, il est en effet possible de sélectionner les features qui expliquent le mieux - pour un pixel - le fait d'appartenir aux différentes classes de segmentation. Simonetti et [78] al. et Luts et al. [79] ont comparé certaines de ces méthodes de sélection sur une tâche de segmentation de tumeurs en spectroscopie par IRM.

La nécessité de devoir définir manuellement des caractéristiques ou des descripteurs pertinents est cependant un facteur limitant. L'apprentissage profond, en plein essor depuis une dizaine d'années, a apporté une solution à cette contrainte. En effet, les réseaux de neurones profonds peuvent apprendre ces représentations de manière implicite et automatique, en identifiant les motifs et les structures les plus adaptés pour résoudre la tâche donnée. Cette capacité à apprendre des représentations de données hautement abstraites et complexes a dès lors rendu les méthodes d'apprentissage machine conventionnelles obsolètes [80].

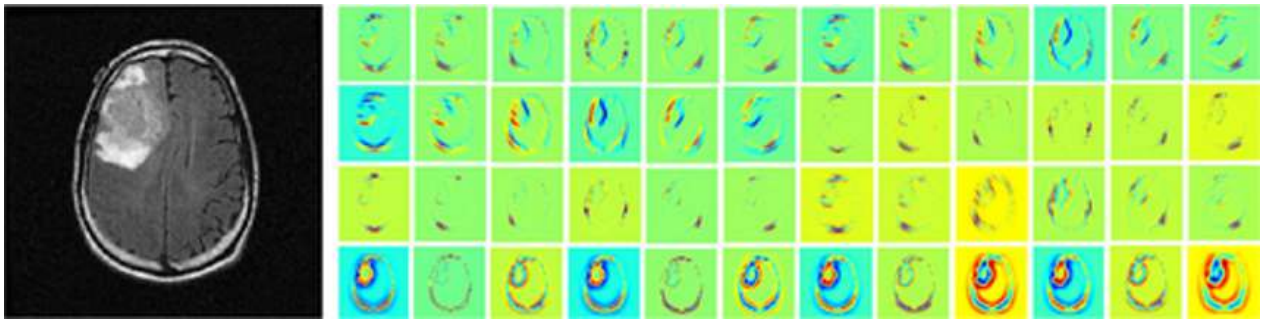


FIGURE 2.9 Génération de 48 features de texture à partir de la banque de filtres Leung Malik sur une séquence cérébrale FLAIR. Associés à d'autres cartes de caractéristiques, ils peuvent être utilisés pour segmenter des tumeurs [10].

2.3 Notions d'apprentissage profond

2.3.1 Entraîner un réseau simple : le perceptron multi-couches

Le réseau de neurones le plus simple que l'on puisse constituer est le perceptron multi-couches (MLP) [81]. Formulée sous forme de vecteur, une entrée x est intégrée par une couche de neurones via la multiplication matricielle avec son vecteur de poids W et l'ajout de son biais b. La sortie y est finalement obtenue par l'application d'une fonction d'activation σ :

$$y = \sigma(W^T x + b) \tag{2.1}$$

Comme illustré en Fig. 2.10, cette opération de propagation vers l'avant (*feed forward*) peut être répétée successivement sur l'ensemble des couches qui constituent le réseau.

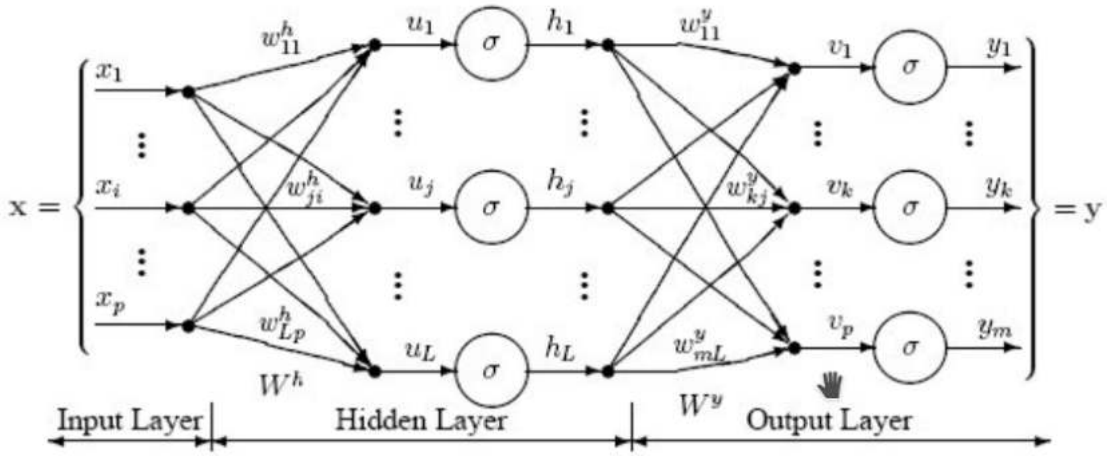


FIGURE 2.10 Perceptron à 2 couches. Image extraite de [11].

Pour guider le réseau dans l'apprentissage d'une tâche donnée, une fonction de perte L est utilisée. Cette fonction représente l'erreur commise par le réseau sur sa prédiction vis-à-vis de la réponse réelle attendue. Les poids du réseau sont alors optimisés pour minimiser L. Pour cela, le gradient de la fonction de coût par rapport aux paramètres du réseau est calculé successivement depuis la couche de sortie jusqu'à la couche d'entrée du réseau. Cette rétropropagation permet alors d'ajuster les poids et les biais de chaque neurone dans la direction opposée du gradient. On parle plus communément de descente du gradient [82]. En notant λ le taux d'apprentissage (un hyperparamètre défini par l'utilisateur), la mise à jour w' d'un poids du réseau w est obtenue par l'opération suivante :

$$w' = w - \lambda \frac{\partial L}{\partial w} \tag{2.2}$$

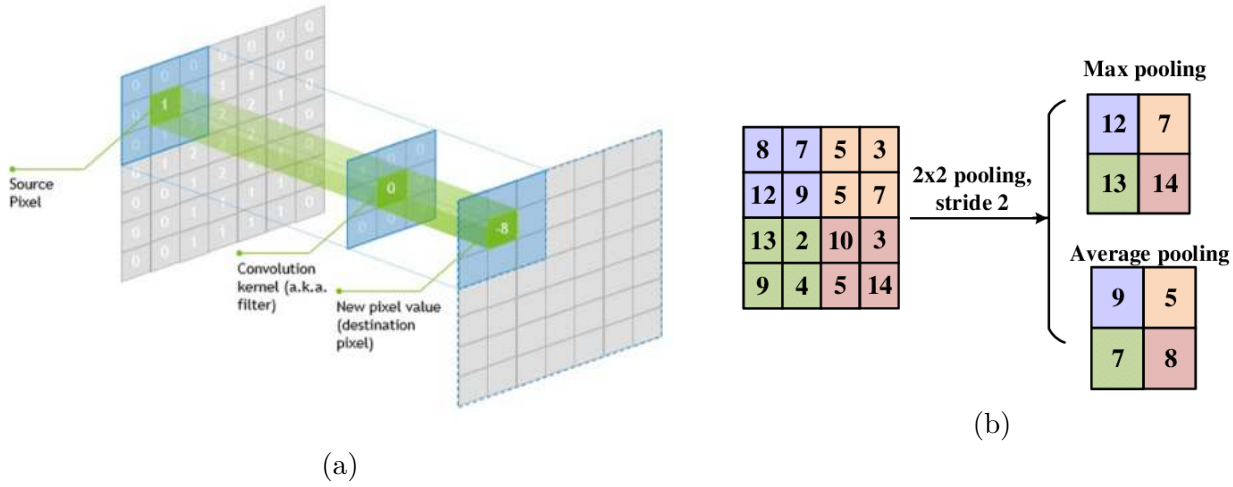
Le processus est répété jusqu'à ce que la fonction de perte atteigne son minimum, ou qu'un nombre maximum d'itérations soit atteint.

2.3.2 Réseaux convolutifs

Spécifiquement adaptés pour traiter des données ordonnées spatialement ou temporellement, les réseaux de neurones convolutifs (CNN) ont largement été exploités pour la manipulation d'images médicales [83]. Une architecture CNN repose essentiellement sur l'application d'une succession de filtres de convolution (voir Fig. 2.11), dont les poids sont optimisables via une descente de gradient. Tout comme le MLP, un CNN est organisé en couches, donnant lieu à une structure hiérarchique dans les représentations obtenues à partir de l'entrée. A chaque convolution les données sont parcourues localement par un kernel de taille fixe, chaque couche est de fait amenée à extraire des caractéristiques depuis son prédecesseur dans la hiérarchie. Ces architectures convolutives donnent lieu à des représentations hautement complexes de l'image d'entrée, et contrairement au MLP, intègrent un contexte spatial essentiel en traitement d'images. De plus, par sa nature convolutive, un CNN est invariant aux translations et peut détecter des motifs indépendamment de leurs positions dans l'image. Notons qu'une image à plusieurs canaux peut être mise en entrée d'un CNN, offrant donc la possiblité d'intégrer plusieurs modalités au modèle (e.g. plusieurs séquences d'IRM). Aussi, comme illustré en Fig. 2.12 un MLP peut être branché à la sortie des couches de convolution pour profiter des représentations obtenues. Cela est notamment utile sur des tâches de classification (e.g. prédiction de types [84] ou de grades [85] de tumeurs) à partir de clichés médicaux.

Des opérations de sous-échantillonnage (voir Fig. 2.11), souvent basées sur du regroupement par maximum (max-pooling) ou moyennage (average-pooling), sont insérées entre les opérations de convolution. Visant à réduire la dimension des cartes de caractéristiques, elles

servent à réduire le nombre de paramètres totaux du modèle et ainsi que les coûts computationnels. Le réseau est par ailleurs amené à apprendre des caractéristiques plus générales et à être moins sensible aux variations mineures des données. Une alternative aux méthodes par regroupement est d'utiliser un pas (ou *stride*) différent de un pour l'application des convolutions.



(a) (b)

FIGURE 2.11 Opérations de (a) convolution et (b) de pooling. Extraits de [12] et [13].

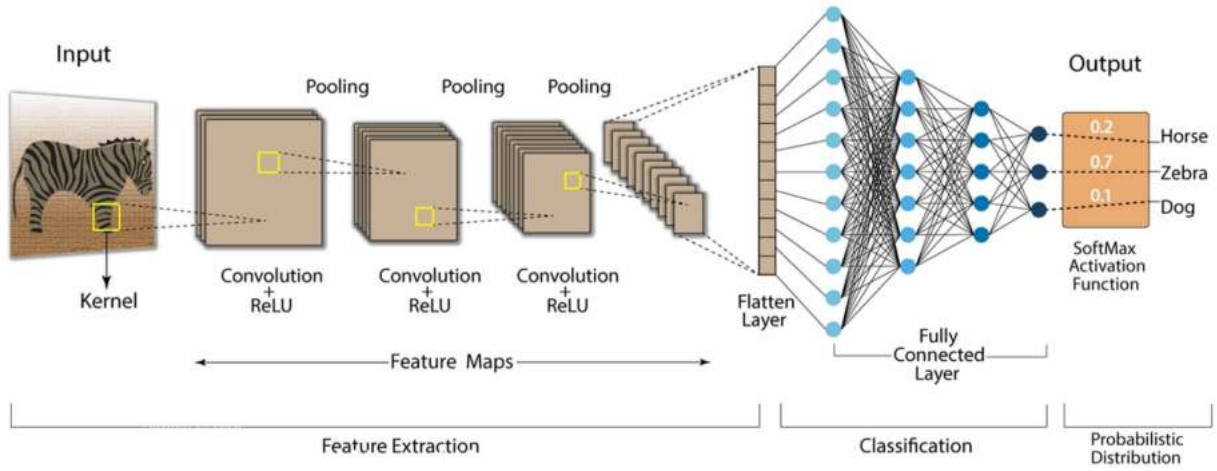


FIGURE 2.12 Exemple de CNN. Des cartes de caractéristiques sont extraites par trois couches alternées de convolution et de max-pooling. Un MLP traite ces représentations pour une tâche de classification. Image extraite de [14].

Néanmoins les calculs de gradients pour des réseaux convolutifs sont computationnellement coûteux. Des processeurs graphiques (GPU) sont couramment utilisés pour accélerer l'entraînement de ces réseaux et utiliser des architectures plus larges et plus performantes [86].

Malgré des progrès considérables ces GPUs peuvent parfois possèder une mémoire limitante, notamment si les images utilisées sont des volumes 3D. Pour remédier à cela, des stratégies de traitement multiple [87] peuvent être mises en place en divisant et exécutant la tâche sur plusieurs processeurs en parallèle. Par ailleurs, couper l'image sur la région d'intérêt (ROI) ou réduire la résolution des images sont aussi des solutions envisageables.

2.3.3 Architectures de type encodeur-décodeur

Espace latent et objectif d'optimisation

Pour des tâches spécifiques de vision (génération d'images, segmentation, débruitage, etc.), les réseaux sont parfois organisés sous forme encodeur-décodeur. Etant donnée une image x en entrée, un encodeur e permet d'extraire une représentation riche et hautement abstraite $e(x)$ de cette dernière. Le décodeur d est alors chargé d'interpréter cet encodage pour générer une sortie $d(e(x))$ la plus proche possible d'une réponse y attendue (par rapport à une fonction de perte L). Comme précédemment, on optimise ces deux composantes de manière conjointe avec une descente du gradient. La paire encodeur/décodeur obtenue est donc la solution de :

$$\arg\min_{e,d} L(d \circ e(x), y) \tag{2.3}$$

Envoyer l'image vers un espace latent de plus faible dimension via l'encodage est en réalité similaire à l'action d'extraction et de sélection des features dans les modèles d'apprentissage machine conventionnels. La capacité des architectures de type encodeur-décodeur à compresser l'information contenue dans une image est cependant plus puissante et plus souple que ses prédecesseurs [88].

Application à la segmentation

Les architectures encodeur-décodeur sont au coeur des tâches de segmentation. Le décodeur est alors un réseau en quelque sorte symétrique à l'encodeur. A travers des opérations successives de déconvolution et de sur-échantillonage, il interprète la représentation latente de l'image et reformate progressivement les cartes de caractéristiques afin de produire une sortie aux mêmes dimensions que l'image d'entrée (voir Fig. 2.13). Ces réseaux de segmentation sont essentiellements constitués de couches de convolutions et ne présentent aucune couche totalement connectée (couche standard du MLP), on parle de FCN (*Fully Convolutional Network*) [89]. Pour produire la carte de segmentation d'une image, on prédit pour tous les pixels x_i un score d'appartenance $p_n(x_i)$ à chaque classe n donnée parmi les N du problème.

Une fonction d'activation softmax [90] permet de s'assurer que ces scores soient en réalité des probabilités ($p_n(x_i) \in [0,1]$ et $\sum_{n=1}^{N} p_n(x_i) = 1$).

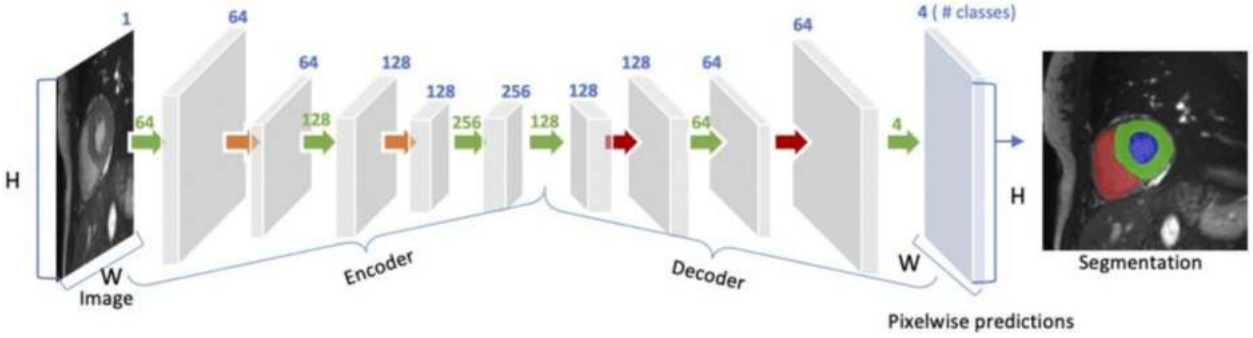


FIGURE 2.13 Exemple de FCN pour une tâche de segmentation à 3 classes. Extrait de [15].

Lors de l'entraînement, différentes fonctions de perte peuvent être appliquées pour guider le réseau dans sa segmentation :

- La fonction d'entropie croisée [91], mesurant l'erreur de prédiction par pixel.
- La fonction de perte de Dice. S'écrivant $L_{Dice} = \sum_{n=1}^{N}(1 - DSC_n)$ où DSC_n correspond au score de Dice relatif à la classe n, elle mesure une erreur de superposition plus globale que l'entropie croisée.
- En cas de déséquilibre fort des classes (sur-représentation de l'une d'entre elles), une fonction de Dice généralisée [92] peut être mise en place. En corrigeant la contribution de chaque classe par l'inverse de sa fréquence on peut en effet limiter le potentiel biais d'entraînement en faveur de la classe prépondérante.
- La fonction de perte de Tversky [93] ou encore sa variante focale [94] peuvent servir à améliorer la segmentation de cas difficiles.
- D'autres fonctions combinant l'entropie croisée, la fonction de perte de Dice et/ou leurs variants [91, 95, 96] sont aussi envisageables.

Connexions raccourcies

Un simple FCN avec une structure encodeur-décodeur (comme en Fig. 2.13) reste cependant limité pour formuler une segmentation précise. Du fait de l'encodage et du sous-échantillonage, certaines caractéristiques de l'image peuvent être eclipsées. En 2015 Ronneberg et al. [16] ont donc proposé un nouveau type d'architecture pour la segmentation d'images médicales : les réseaux U-Net. Mettant en place des connexions raccourcies longue distance (*long skip connections*) entre l'encodeur et le décodeur, l'information spatiale perdue au cours de l'encodage est mieux reconstituée. Comme illustré en Fig. 2.14, des opérations

de concaténation de features interviennent à chaque niveau de résolution du U-Net.

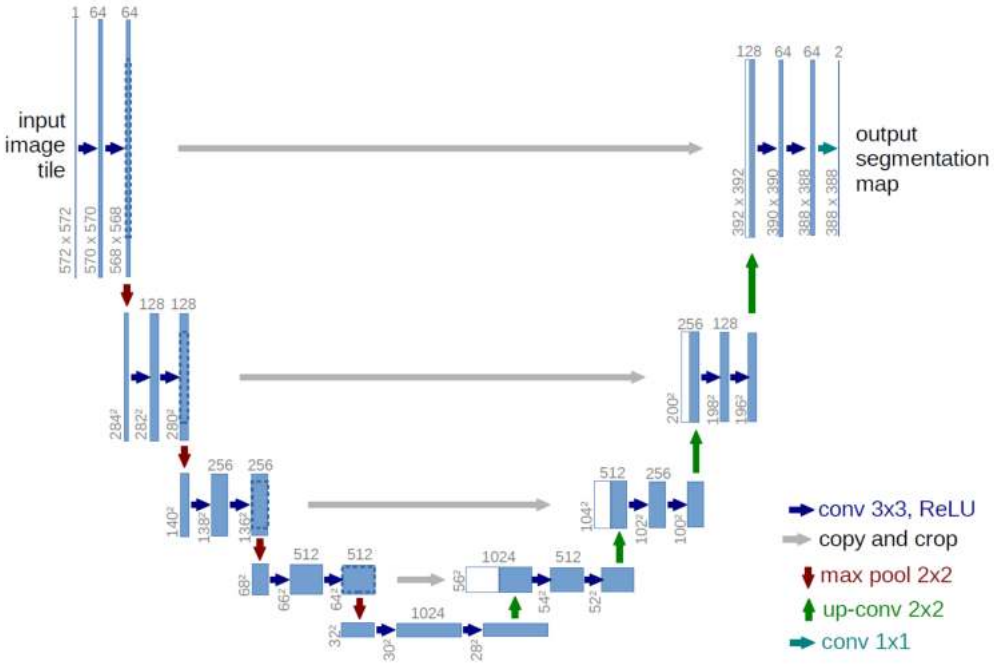


FIGURE 2.14 Architecture U-Net. Des connexions raccourcies entre l'encodeur et le décodeur du réseau de segmentation permettent de mieux reconstituer l'information spatiale perdue lors de l'encodage. Image extraite de [16].

Le U-Net standard, tel qu'il a été conçu initialement est néanmoins limité en performance. Sa profondeur fixe le rend peu flexible et réduit sa capacité à bien se généraliser à des jeux de données complexes. Augmenter le nombre de paramètres du réseau et sa profondeur peut par ailleurs être suboptimal du fait du problème de disparition du gradient. En effet, dans des réseaux très profonds le gradient peut devenir infiniment petit par multiplication successive (loi de la chaîne [97]) en se rétro-propageant vers les couches hautes. Cela a pour conséquence de ne plus optimiser les poids dans les couches supérieures [98].

Les connexions raccourcies résiduelles [99] ou denses [17] constituent une solution au problème. Ces connexions courtes (*short skip connections*) connectent des couches proches du réseau par des opérations de concaténation ou de somme de features (voir Fig. 2.15). En conséquence, lors de la rétropropagation, le gradient dans les couches hautes ne s'exprime plus uniquement comme une suite de termes multipliés mais comme une somme de gradients provenant des couches basses, limitant ainsi le phénomène de disparition. Ces connections, une fois implémentées dans des architectures plus profondes (ResNet et DenseNet) permettent de considérablement accroître la capacité de réprésentation et de généralisation des modèles.

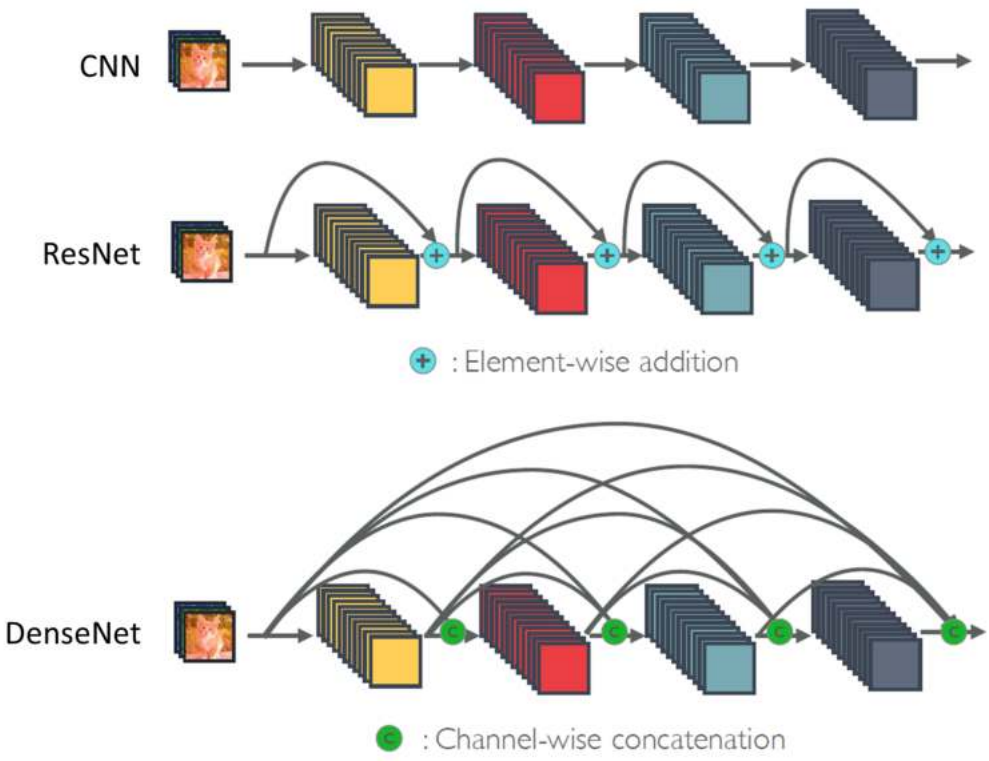


FIGURE 2.15 Illustrations de connexions standards (CNN), résiduelles (ResNet) et denses (DenseNet). Image extraite de [17].

Portes d'attention

Pour beaucoup d'images médicales, les structures à segmenter ne représentent qu'une partie mineure de l'ensemble des pixels/voxels. La segmentation peut dès lors être imprécise puisque beaucoup d'activations sont perdues à traiter de l'information non pertinente. Des méthodes choisissent de préalablement extraire une région d'intérêt pour recentrer le problème sur la structure à segmenter. Cela peut être fait en proposant des cadres de délimitation via des réseaux de détection [100–102]. Néanmoins, de par l'ajout d'une étape supplémentaire, ces méthodes sont plus lourdes.

Une alternative plus optimale est d'utiliser des portes d'attention. Introduites en segmentation d'images par Oktay et al. [103], ces *attention gates* offrent au réseau la possibilité d'apprendre à ignorer les régions non pertinentes tout en mettant l'accent sur les caractéristiques essentielles à la tâche. Comme montré en Fig. 2.16, ces unités d'attention sont intégrées à chaque niveau du décodeur et génèrent des grilles (dont les éléments sont à valeurs dans $[0, 1]$) qui permettent de repondérer les éléments des features passés dans les connexions raccourcies. La nature différentiable de ces portes d'attention permet leur entrainement implicite

lors de la descente de gradient. La Fig. 2.17 montre l'intérêt de telles unités dans une tâche de segmentation multi-classes sur des scanners CT abdominaux.

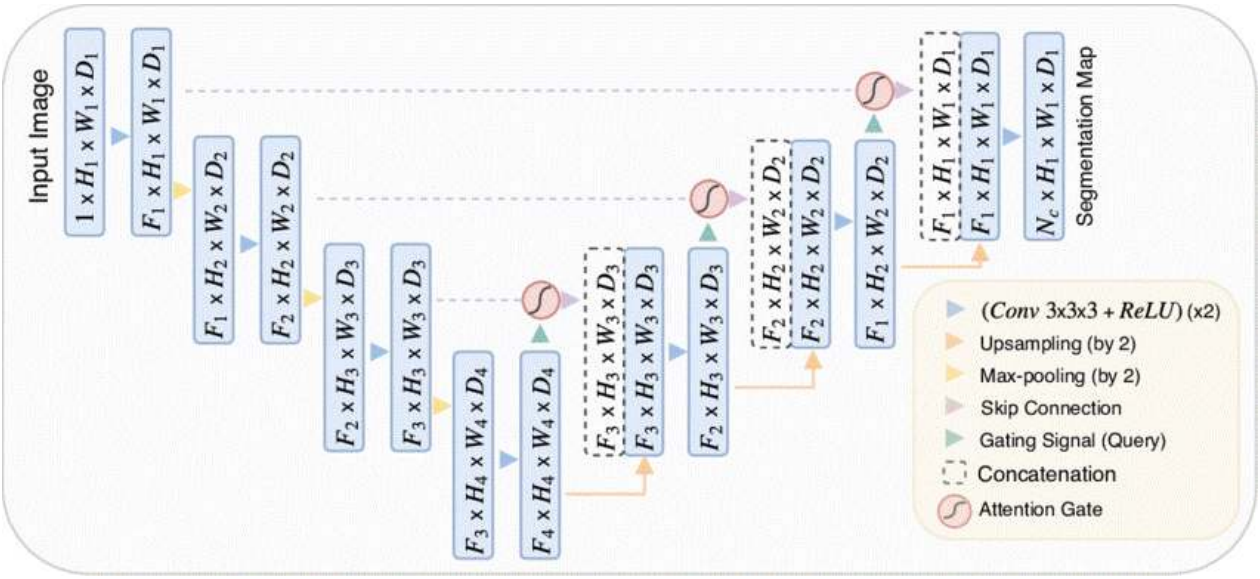


FIGURE 2.16 Attention U-Net. Des portes d'attention sont introduites dans les connexions raccourcies.

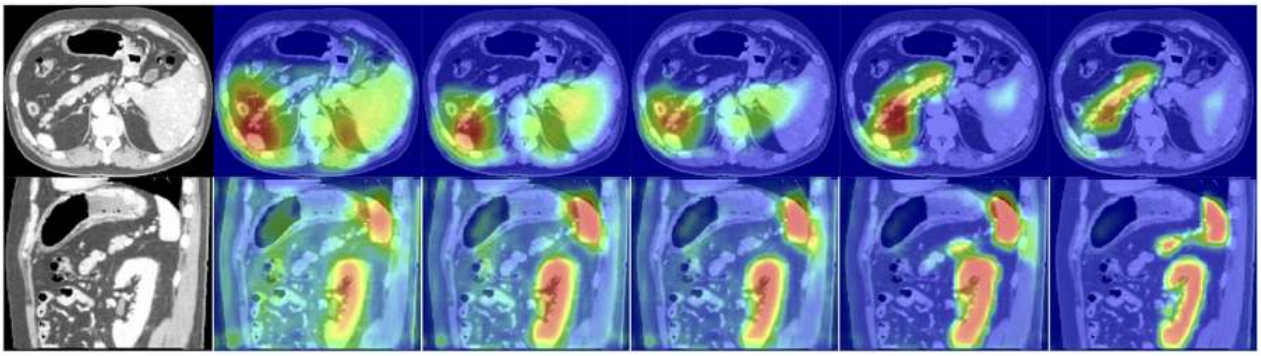


FIGURE 2.17 Coefficients d'attention d'un attention U-Net visualisés à 3, 6, 10 et 150 époques pour une tâche de segmentation multi-classes de scanners abdominaux. Les zones en rouge correspondent aux régions revalorisées favorablement par les portes d'attention. Le modèle apprend progressivement à se concentrer sur le pancréas, le rein et la rate.

2.3.4 Réseaux transformers

"Attention is all you need"

Initialement introduits pour le traitement du langage comme remplaçants aux réseaux de neurones récurrents (RNN) [104], les réseaux transformers reposent essentiellement sur le

principe d'attention [19]. Comme introduit précédemment, les mécanismes d'attention constituent l'ensemble des techniques visant à améliorer l'information contextuelle dans un réseau en pondérant favorablement les composantes clés de la tâche en cours. Dans le cas des transformers on parle plus spécifiquement de self-attention, puisqu'on capture les dépendances entre les différents éléments d'une même séquence d'entrée.

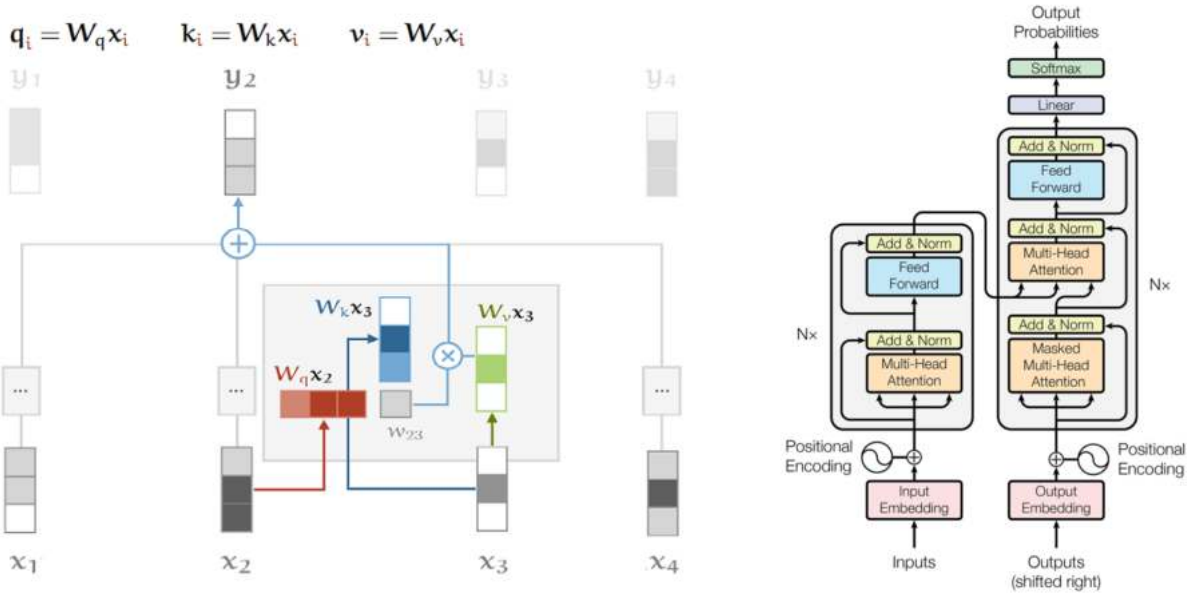


FIGURE 2.18 (a) Illustration du principe de self-attention. Les matrices W_q, W_k et W_v, respectivement associées au vecteur de requête, de clé et de valeur, constituent des poids que le réseau est amené à apprendre. Image extraite de [18]. (b) Architecture transformer pour la traduction séquence à séquence. Tirée de [19].

Pour un mot donné d'une phrase (préalablement encodée), un score d'attention est obtenu par un produit scalaire entre son vecteur de requête (*query* Q) et le vecteur clé (*key* K) de chaque mot de la séquence. Après application d'une fonction softmax, ces scores (à valeurs entre 0 et 1) permettent de calculer une somme pondérée des vecteurs de valeur (*value* V) de l'ensemble des mots de la phrase, résultant en un vecteur de contexte qui contient chaque composante pertinente de la séquence considérée. Ces opérations, illustrées en Fig. 2.18, se résument en l'équation suivante, où d_k est un facteur de normalisation :

$$Attention(Q, K, V) = softmax\left(\frac{QK^T}{\sqrt{d_k}}\right)V \tag{2.4}$$

Ce principe est intégré dans des architectures encodeur-décodeur où s'alternent des calculs

de self-attention dans des blocs à plusieurs têtes mises en parallèle (*multi-head attention*) et des opérations de traitement simples dans des blocs dits de *feed forward* (voir Fig. 2.18).

"An image is worth 16×16 words"

Devant le succès des transformers en traitement du langage, le principe de self-attention fut par la suite adapté pour les images et les tâches de vision. Dans ces nouveaux modèles (*vision transformer*), l'image d'entrée est découpée en plusieurs patchs de taille fixe. L'image, dès lors transformée en une séquence, peut être traitée par des architectures transformers classiques. Seul un encodage de la position des patchs est ajouté afin de retenir la structure spatiale.

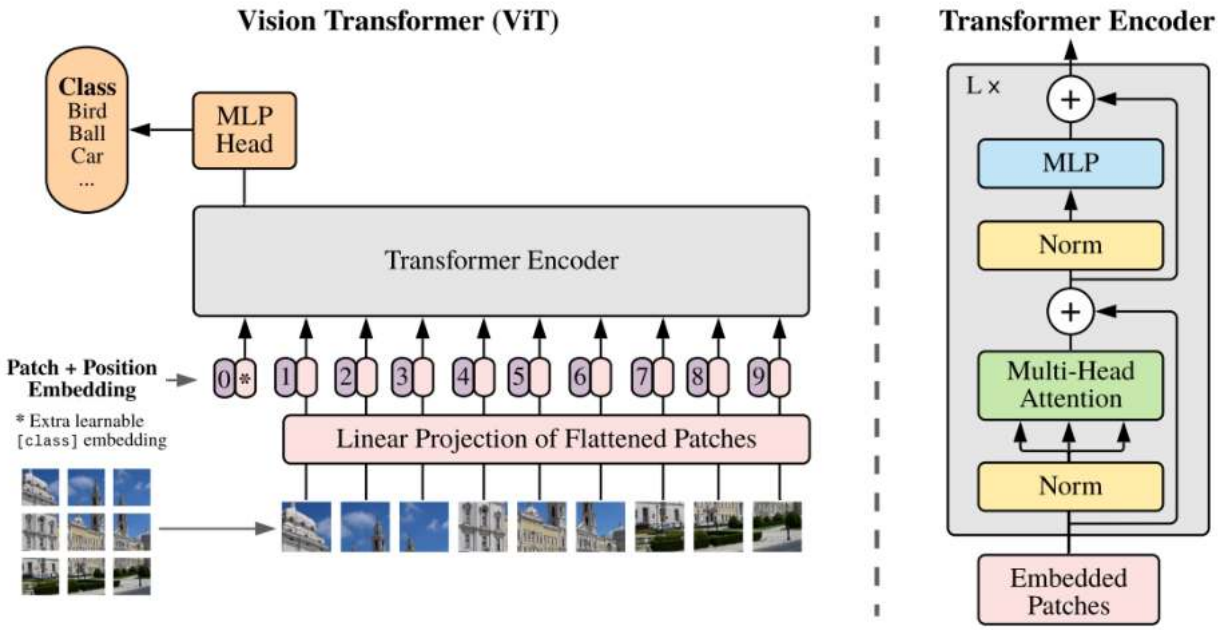


FIGURE 2.19 Exemple de transformer pour la vision. Ici, la représentation de l'image extraite par un transformer est utilisée par un MLP pour de la classification. Un encodage de la position des patchs est ajouté pour intégrer la composante spatiale au modèle. Extrait de [20].

Couplés à des MLPs les transformers se sont révélés être plus puissants que les CNNs sur une tâche de reconnaissance d'images [20]. Dès lors, ils ont par la suite été adoptés en traitement d'images médicales pour de la classification (e.g. fracture du fémur [105] ou cancer de la peau [106]). Implémentés dans des architectures de type encodeur-décodeur (voir Fig. 2.20), les tranformeurs se sont aussi révélés très efficaces dans des tâches de segmentation d'images médicales [21, 107–109].

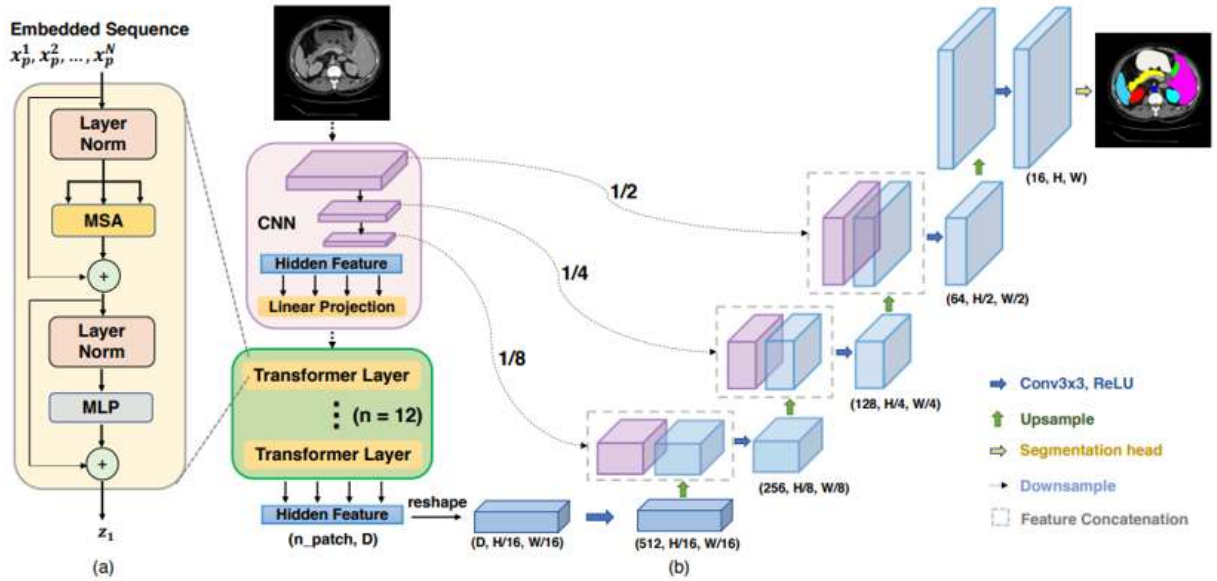
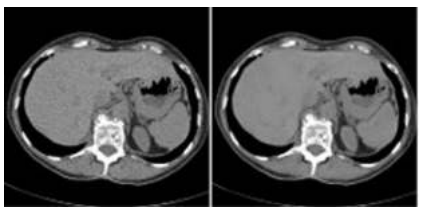


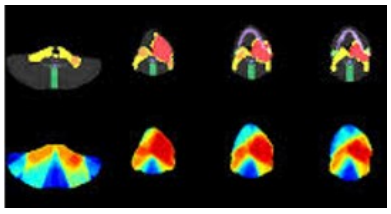
FIGURE 2.20 Exemple d'architecture transformer pour la segmentation d'images médicales. Ici le TransUnet [21], un modèle hybride convolution/transformer. Dans l'encodeur se succédent un module convolutif et un ensemble de blocs transformers. Le décodeur, pleinement convolutif, permet de générer la segmentation de l'image d'entrée.

2.3.5 Réseaux antagonistes génératifs

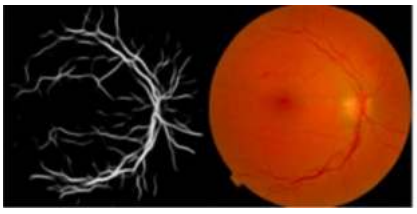
Introduits par Goodfellow et al. [110], les réseaux antagonistes génératifs (GAN) sont une classe de modèles permettant la génération de données. De par leurs applications multiples, ils ont suscité beaucoup d'intérêt au sein de la communauté des chercheurs. Comme illustré en Fig. 2.21, ils ont notamment grandement profité au traitement d'images médicales (synthèse d'images pour l'augmentation de données, segmentation, génération de plans de dose, etc.) [22].



(a) Débruitage de scanners CT à faible dose.



(b) Génération de plans de dose pour la radiothérapie.



(c) Synthèse de fonds d'oeil à partir de vaisseaux.

FIGURE 2.21 Exemples d'application des GAN en imagerie médicale. Extraits de [22].

Un GAN est décomposable en deux sous composantes : un générateur et un discriminateur.

Ces deux réseaux sont entraînés simultanément dans un jeu de concurrence : le générateur cherche à produire des images suffisamment réalistes pour tromper le discriminateur, tandis que le discriminateur cherche à distinguer les images réelles des images synthétisées par son antagoniste. Pour modéliser mathématiquement l'apprentissage d'un réseau antagoniste, on définit donc :

- Le réseau générateur $G(.)$ qui prend en entrée des données z de densité p_z et a pour sortie $x_g = G(z)$.
- Le réseau discriminateur $D(.)$ qui prend une entrée x pouvant être réelle (x_t de densité p_t) ou synthétique (x_g de densité $p_g = G(p_z)$), et retourne la probabilité $D(x)$ que x soit une donnée réelle.

L'espérance de l'erreur commise par le discriminateur D pour un générateur G donné peut alors s'écrire :

$$\begin{aligned} E(G, D) &= \mathbb{E}_{x \sim p_t}[1 - D(x)] + \mathbb{E}_{z \sim p_z}[D(G(x))] \\ &= \mathbb{E}_{x \sim p_t}[1 - D(x)] + \mathbb{E}_{x \sim p_g}[D(x)] \end{aligned} \tag{2.5}$$

Quand on entraine le générateur, on veut maximiser cette erreur (i.e. tromper le discriminateur) alors qu'on cherche à la minimiser pour le discriminateur (i.e. correctement distinguer les données générées et réelles). L'objectif d'optimisation s'écrit donc :

$$\max_G \left(\min_D E(G, D) \right) \tag{2.6}$$

Après entraînement, les données x_g doivent théoriquement suivre la distribution cible p_t.

2.3.6 Apprentissage et optimisation

Normalisation

La normalisation des features est un concept essentiel en apprentissage profond. C'est en effet un moyen simple d'améliorer la régularisation des réseaux de neurones au cours de leur entrainement. La normalisation par batch est la normalisation la plus prépondérante, et présente plusieurs avantages [111] :

- En normalisant chaque feature de telle manière à englober la contribution des autres échantillons du batch, on limite le potentiel biais lié à l'apparition de valeurs hautes dans les features.
- Elle réduit le décalage de covariance interne (*internal covariate shift*) i.e. les changements de distributions des activations des couches cachées lors de l'entrainement.

Avec la normalisation, on maintient des distributions plus stables, ce qui a pour effet d'accélerer la convergence du modèle.

- Elle rend la surface de la fonction de perte plus lisse (les magnitudes des gradients sont contraintes à un intervalle réduit), ce qui facilite la descente de gradient.
- Elle permet de partiellement endiguer le problème de disparition du gradient.

D'autres types de normalisation (voir Fig. 2.22) sont aussi envisageables. Le choix de l'une ou l'autre est un hyperparamètre qu'il convient de définir en fonction du jeu de données utilisé.

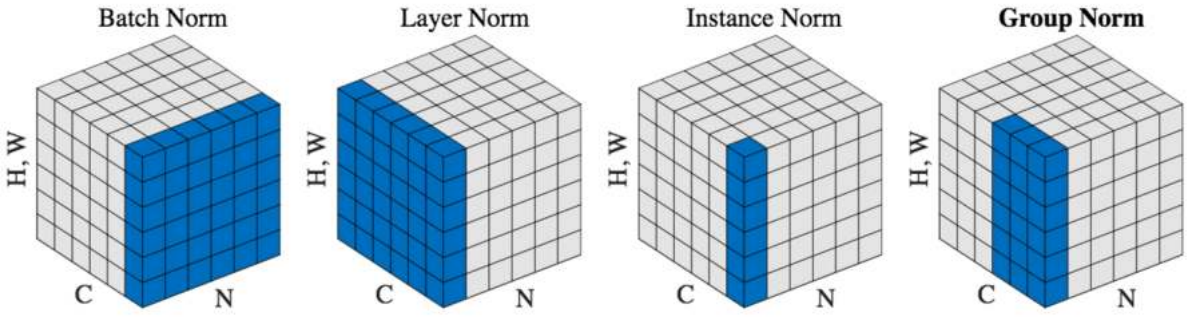


FIGURE 2.22 Différents types de normalisation. Chaque subplot montre un vecteur de features, avec N le batch size, C le nombre de canaux, H et W les dimensions spatiales. Les pixels bleus sont normalisés par la variance et la moyenne calculées sur ces derniers. Illustration extraite de [23].

Augmentation de données

Afin de limiter le surapprentissage d'un réseau i.e. empêcher qu'il ne se spécialise trop sur les données d'entrainement, il est courant d'utiliser des stratégies d'augmentation de données. En imagerie médicale notamment, l'augmentation de données est absolument nécessaire car avoir accès à un large ensemble de données peut s'avérer complexe [112, 113]. Ces techniques consistent à artificiellement augmenter la quantité de données d'entraînement. En appliquant une série de transformations aléatoires (rotations, zooms, translations, changements de contraste, etc.) aux données d'entraînement existantes, on peut créer de nouvelles données légérement différentes des données déjà existantes. Une autre option est d'utiliser des GANs pour synthétiser de nouvelles données à partir du jeu d'entrainement. On expose ainsi le modèle à plus d'images, et avec une plus grande variabilité.

Apprentissage multi-tâches

En apprentissage profond, il est souvent intéressant de réaliser de l'apprentissage multi-tâches i.e. entraîner un modèle à effectuer plusieurs tâches simultanément (e.g. segmentation et

classification de tumeurs) tout en partageant les connaissances et les représentations apprises entre ces tâches. L'idée est qu'un modèle formé simultanément sur deux tâches est contraint à apprendre des représentations plus générales et pertinentes que si l'entrainement se faisait indépendamment. Cela peut conduire à des modèles plus performants et robustes [114].

En pratique une telle stratégie requiert un unique encodeur partagé entre les différentes tâches. L'espace latent qui en découle peut alors être interprété différemment par les diverses branches mises en place (voir Fig. 2.23). Aussi, pour entrainer le modèle on associe à chaque tâche t une fonction de perte L_t qui rend compte de l'erreur de prédiction spécifique à cette dernière. Le modèle est alors entraîné par descente de gradient sur une fonction de perte globale L_g constituée d'une somme pondérée des fonctions de perte L_t :

$$L_g = \sum_t \lambda_t L_t \tag{2.7}$$

Les poids λ_t constituent alors des hyperparamètres supplémentaires qu'il convient d'affiner pour optimiser la performance du modèle.

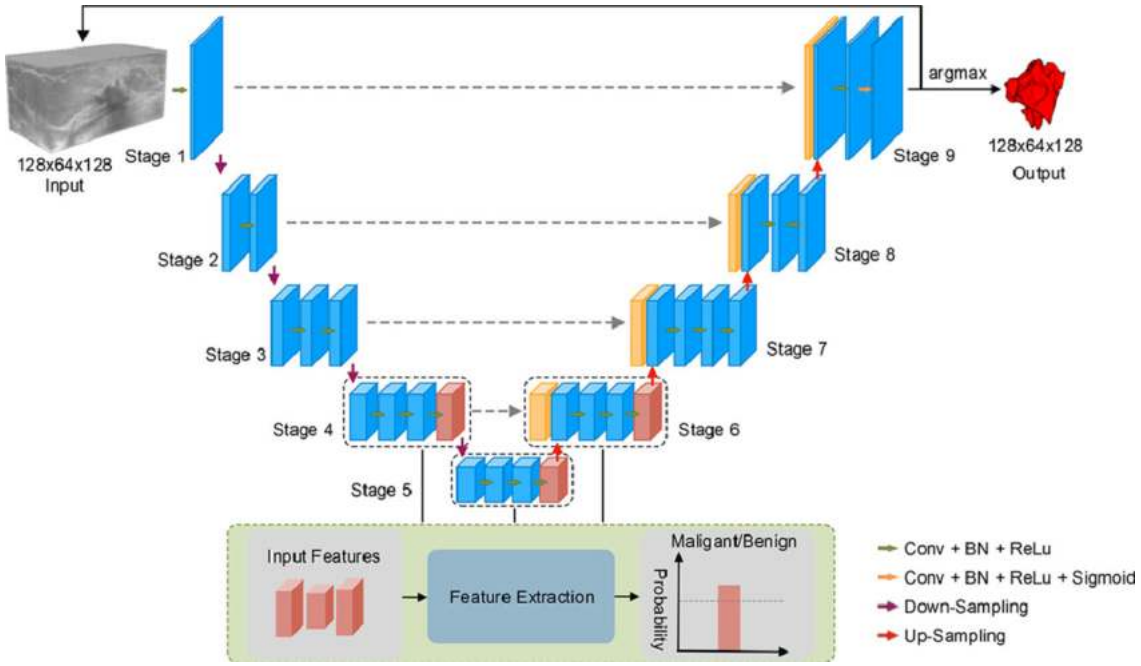
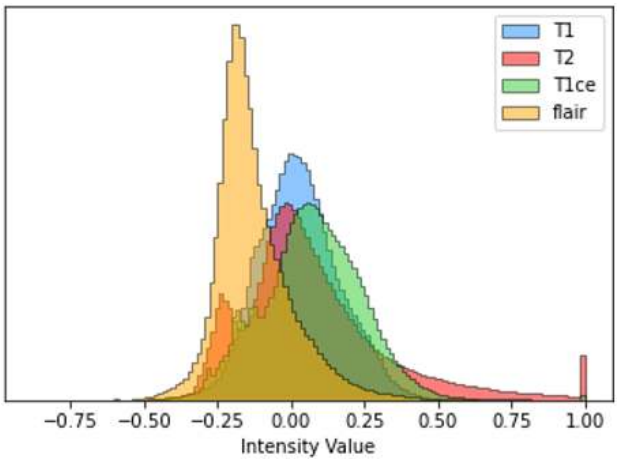


FIGURE 2.23 Illustration de l'apprentissage multi-tâches. La classification et la segmentation de tumeurs sur des ultrasons du sein sont conjointement entrainées. L'espace latent est simultanément interprété par un décodeur pour la segmentation, et par un MLP pour la classification. Extrait de [24].

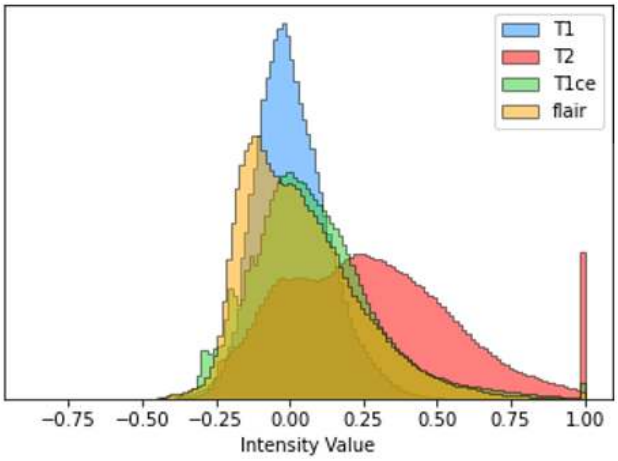
2.4 Segmentation à travers les modalités

2.4.1 Décalage de distributions

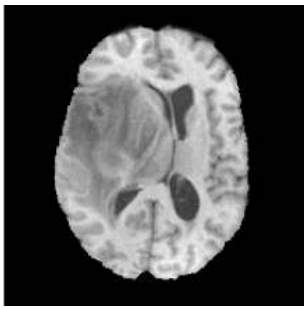
Dans la section précédente, on a introduit les concepts essentiels de l'apprentissage profond et montré que, en comparaison aux méthodes de segmentation traditionnelles, les réseaux de neurones se démarquent par leur modularité et leur performance impressionante. Cependant, le succès de la plupart de ces méthodes repose sur la disponibilité d'annotations précises, qui sont pénibles et coûteuses en temps à produire. En outre, les algorithmes d'apprentissage automatique ne sont pas adaptés aux changements de distributions. Ces décalages pouvant être majeurs entre modalités d'imagerie (voir Fig. 2.24), les limitations émises ci-dessus sont donc un problème au déploiement des modèles d'apprentissage profond à travers les modalités. De fait, un modèle entrainé à segmenter des tumeurs sur une modalité donnée verra sa performance chuter sur une autre modalité (relativement à l'ampleur de la disparité).



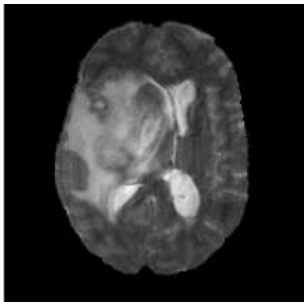
(a) Histogramme des intensités du tissu cérébral.



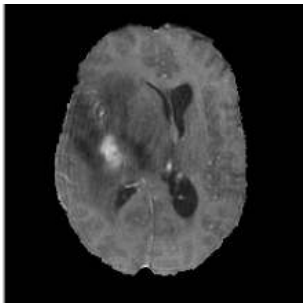
(b) Histogramme des intensités du tissu tumoral.



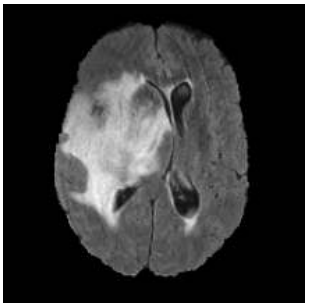
(c) T1.



(d) T2.



(e) T1ce.



(f) FLAIR.

FIGURE 2.24 Illustration du décalage de distributions entre séquences IRM. Les histogrammes sont calculés sur l'ensemble du jeu de données de segmentation de tumeurs cérébrales BraTS [25].

2.4.2 Adaptation de domaine

Ce défi de la généralisation inter-modalités peut être relevé, sans ajout d'annotations supplémentaires, grâce à des approches semi-supervisées d'adaptation de domaines. Ces techniques consistent à utiliser les connaissances disponibles dans un jeu de données (ou domaine) source possédant des annotations pour apprendre une tâche (segmentation, classification, etc.) sur un ensemble de données cible non annotées. L'apparence visuelle des images (ou style) peut facilement varier en imagerie médicale selon le site d'acquisition (paramètres d'acquisition ou modèles de machine différents) ou la modalité d'imagerie utilisée. L'exploitation des méthodes d'adaptation de domaine y est donc courant [115].

Alignement des espaces latents

L'adaptation en segmentation peut se faire par alignement des espaces latents des domaines source et cible. Des méthodes basées sur la divergence ou sur un entrainement antagoniste sont envisageables (voir Fig. 2.25).

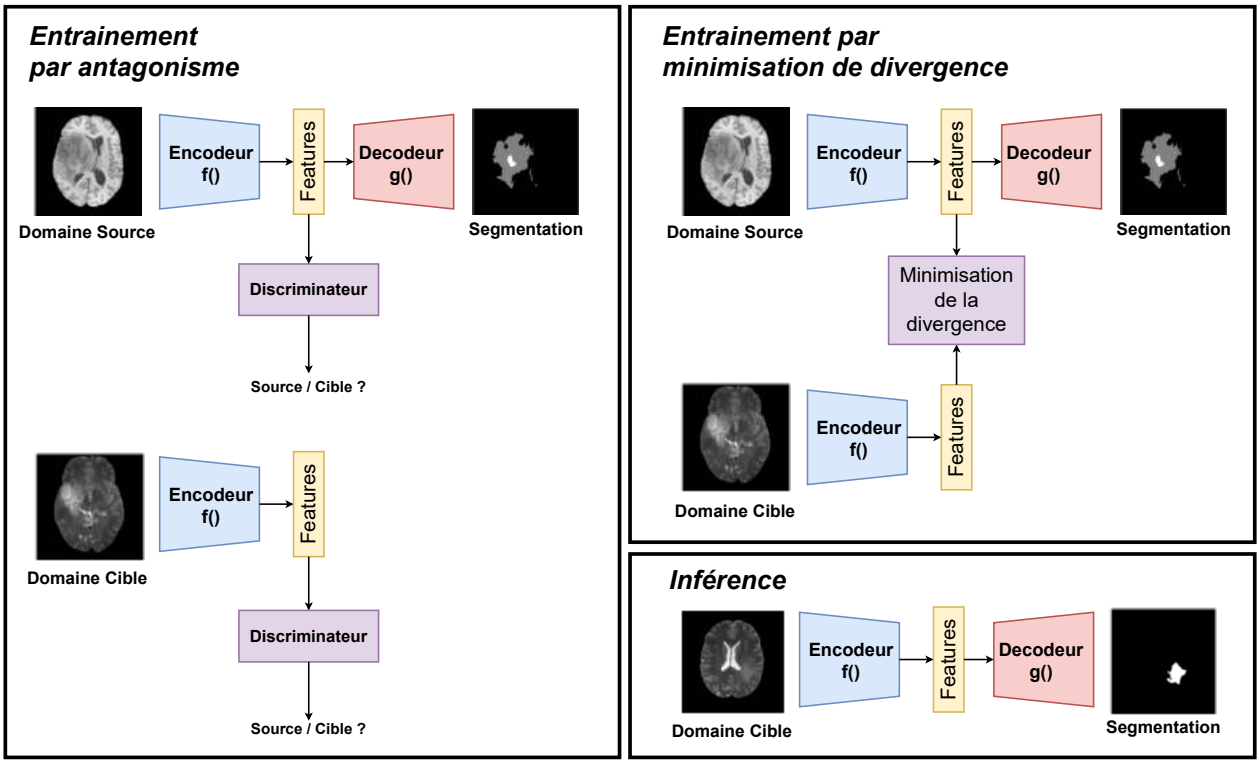


FIGURE 2.25 Adaptation de domaine par alignement des espaces latents.

Dans le premier cas, une fonction de perte basée sur la divergence (Wasserstein, Maximum Mean Discrepancy, etc. [116]) assure alors que les features extraits par un encodeur soient

invariants aux domaines. De plus, par ajout d'une loss de segmentation, les poids de l'encodeur et d'un décodeur sont conjointement optimisés pour assurer une segmentation cohérente des structures d'intérêt. Lors de l'inférence, on peut alors simplement passer une image du domaine cible au couple encodeur-décodeur, supposé être invariant après entrainement.

La deuxième option est relativement similaire mais profite des connaissances sur l'apprentissage par antagonisme pour extraire des représentations communes aux deux domaines. Un discriminateur permet de différencier les représentations extraites d'une image source des représentations extraites d'une image cible. L'encodeur commun aux deux domaines est alors entrainé à aligner les espaces latents pour que les représentations soient indistinugables.

L'adaptation par alignement des features semble avoir préférentiellement été adopté pour des tâches de segmentation où la structure des objets d'intérêt est similaire entre les patients (e.g. segmentation des structures cardiaques [117–120]).

Translation d'images

Une autre option consiste à exploiter des techniques de translation d'image-à-image. Comme illustré en Fig. 2.26, des réseaux antagonistes génératifs sont capables de transférer des images du domaine cible dans le domaine source, en modifiant uniquement leur apparence visuelle et en préservant leur structure. Ainsi, un réseau segmenteur entrainé sur le domaine source peut être appliqué sur des images cible préalablement translatées dans le domaine source (ou vice-versa).

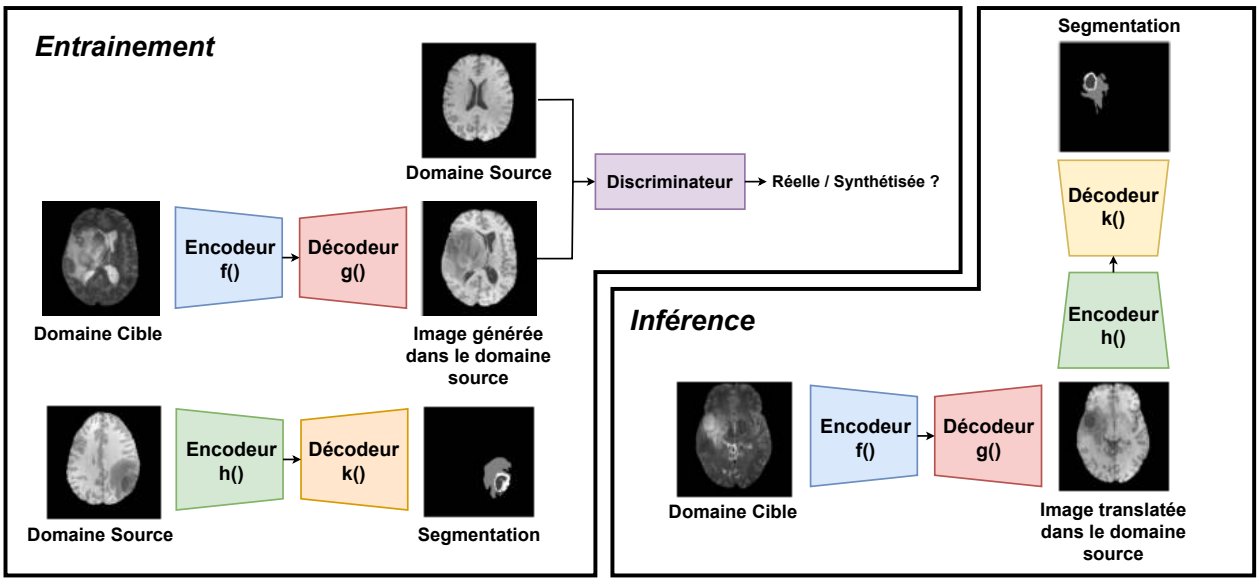


FIGURE 2.26 Adaptation de domaine par translation d'images.

Pour réaliser les translations, de multiples modèles sont envisageables [121]. Néanmoins, un modèle simple et très répandu - notamment en translation de modalités - est le CycleGAN [122]. Deux réseaux de type encodeur-décodeur y apprennent simultanément à translater les images d'un domaine à l'autre. Pour assurer la cohérence des transformations, une fonction de perte de constance cyclique est ajoutée. Cette fonction garantit que translater d'un domaine à l'autre puis dans le sens inverse résulte en une image similaire à l'entrée.

Notons qu'il existe des méthodes combinant translation de domaines et alignement des espaces latents [123–126]. En effet, en translatant entre les différents domaines on peut apprendre au modèle à dissocier les représentations de chaque image en une composante spécifique au domaine (textures, intensités, etc.) et une composante invariante aux deux domaines (structures anatomiques). Un décodeur est alors entrainé à générer les segmentations à partir de la représentation invariante des images du domaine source.

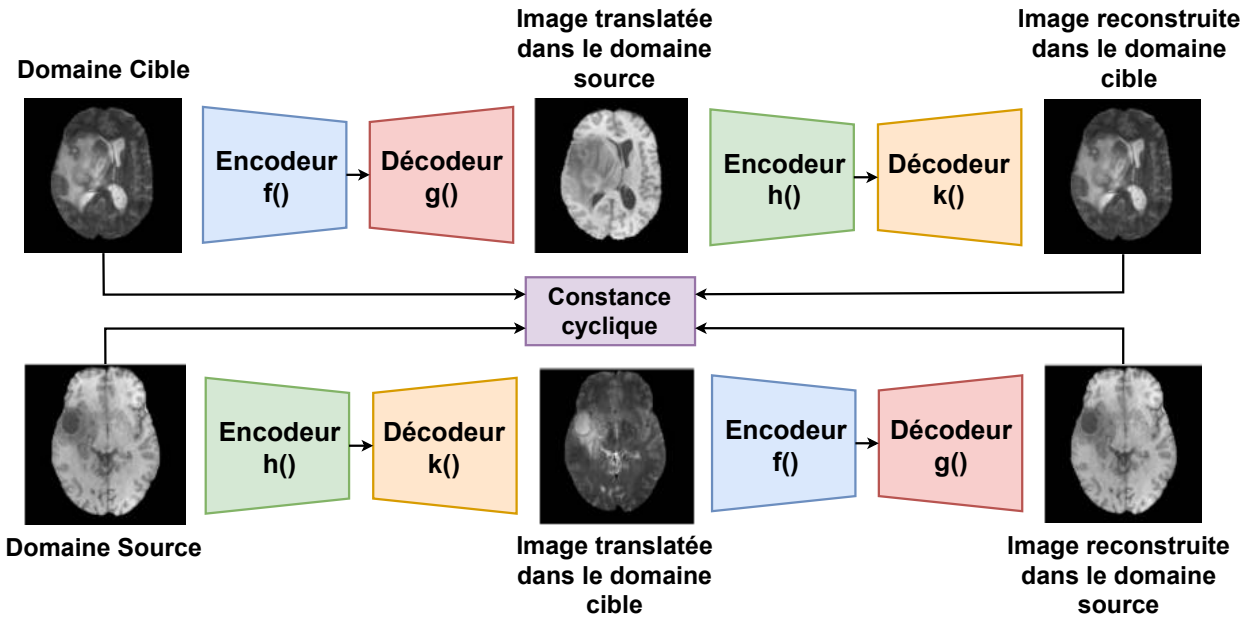


FIGURE 2.27 Principe du CycleGAN. Pour la lisibilité, les discriminateurs ne sont pas montrés.

2.5 Mot de synthèse

Nous avons détaillé dans cette revue de littérature tous les concepts élémentaires essentiels à la compréhension de la problématique de ce projet. Les techniques d'imagerie sont des atouts majeurs pour la détection de tumeurs. Et, la segmentation des lésions - étape essentielle pour la planification du traitement - peut être accélérée et facilitée par les méthodes de

traitement par ordinateur. En particulier l'apprentissage profond, en se démarquant par ses performances et sa flexibilité, constitue une alternative viable aux segmentations réalisées de manière manuelle ou interactive par des experts médicaux. Toutefois ces méthodes ne sont pas adaptées aux changements de distribution, ce qui les rend peu généralisables entre modalités d'imagerie. Les techniques d'adaptation de domaine sont donc une opportunité considérable pour la généralisation inter-modalités, sans le coût supplémentaire que représente la production d'annotations. On souligne que la segmentation d'organes entre modalités est un sujet largement traité, les études portant sur la segmentation de tumeurs restent cependant minoritairement représentées.

CHAPITRE 3 MÉTHODOLOGIE DU TRAVAIL DE RECHERCHE

Le chapitre précédent a établi un historique des méthodes d'apprentissage automatique pour la segmentation de tumeurs. Il a de même permis de mettre en exergue le besoin de la généralisation inter-modalités pour les modèles d'apprentissage profond, ainsi que de suggérer quelques solutions à explorer. L'objectif de ce projet est alors de trouver une méthode efficace pour réduire la charge que représentent les annotations dans les tâches de segmentation inter-modalités. La solution de segmentation proposée doit permettre l'adaptation de domaine entre différentes modalités, c'est-à-dire apprendre la distribution des tumeurs dans le domaine cible non annoté à partir des segmentations fournies dans un domaine source. L'approche développée doit être robuste et compétitive sur plusieurs pathologies et jeux de données. Nous nous sommes portés vers une méthode générative basée sur la translation d'images entre modalités. A cet effet, nous avons décidé d'exploiter le CycleGAN (introduit en section 2.4.2), en y apportant des modifications pour le rendre plus adapté à la génération d'images tumorales. Nous avons de plus exploré le potentiel d'un objectif semi-supervisé exploitant des images connues comme étant saines et encourageant le modèle à dissocier les tumeurs du fond des images. Finalement, nous avons examiné le potentiel d'un modèle 3D plutôt que 2D, l'utilisation d'architectures plus puissantes intégrant des transformers, ainsi que l'apport d'une stratégie itérative d'auto-entrainement.

Voici l'ensemble des articles réalisés dans le cadre de ce mémoire de maîtrise :

- M. Alefsen de Boisredon d'Assier, E. Vorontsov and S. Kadoury, "M-GenSeg : Domain Adaptation For Target Modality Tumor Segmentation With Annotation-efficient Supervision", Medical Image Computing and Computer Assisted Intervention – MICCAI 2023, accepté, Juin 2023.
- M. Alefsen de Boisredon d'Assier, E. Vorontsov, W. Le and S. Kadoury, "Image-Level Supervision And Self-Training For Transformer-Based Cross-Modality Tumor Segmentation", Medical Image Analysis, soumis, Juin 2023.

3.1 Modélisation générale de l'approche (M-GenSeg et MoDATTS)

M-GenSeg et MoDATTS reposent sur deux étapes majeures (illustrées en Fig. 3.1) : (1) L'entrainement d'un réseau permettant la translation entre les différentes modalités dans le but de générer des images synthétiques de la modalité cible à partir d'images de la modalité source ; et (2) l'entrainement d'un modèle de segmentation apprenant à délimiter les tumeurs dans la modalité cible à partir des images synthétisées à l'étape précédente. De vraies

images de la modalité cible - dépourvues d'annotations - sont additionnellement intégrées à l'apprentissage via le framework GenSeg [127]. Ce dernier intègre un objectif semi-supervisé apprenant au modèle à réaliser des translations entre images saines et images tumorales en détachant les tumeurs du background. Bien que reposant sur ces deux principes clés, notons que les modèles proposés (M-GenSeg et MoDATTS) sont bien deux approches distinctes. Les spécificités de chacun sont présentées dans les sections 3.2 et 3.3 suivantes.

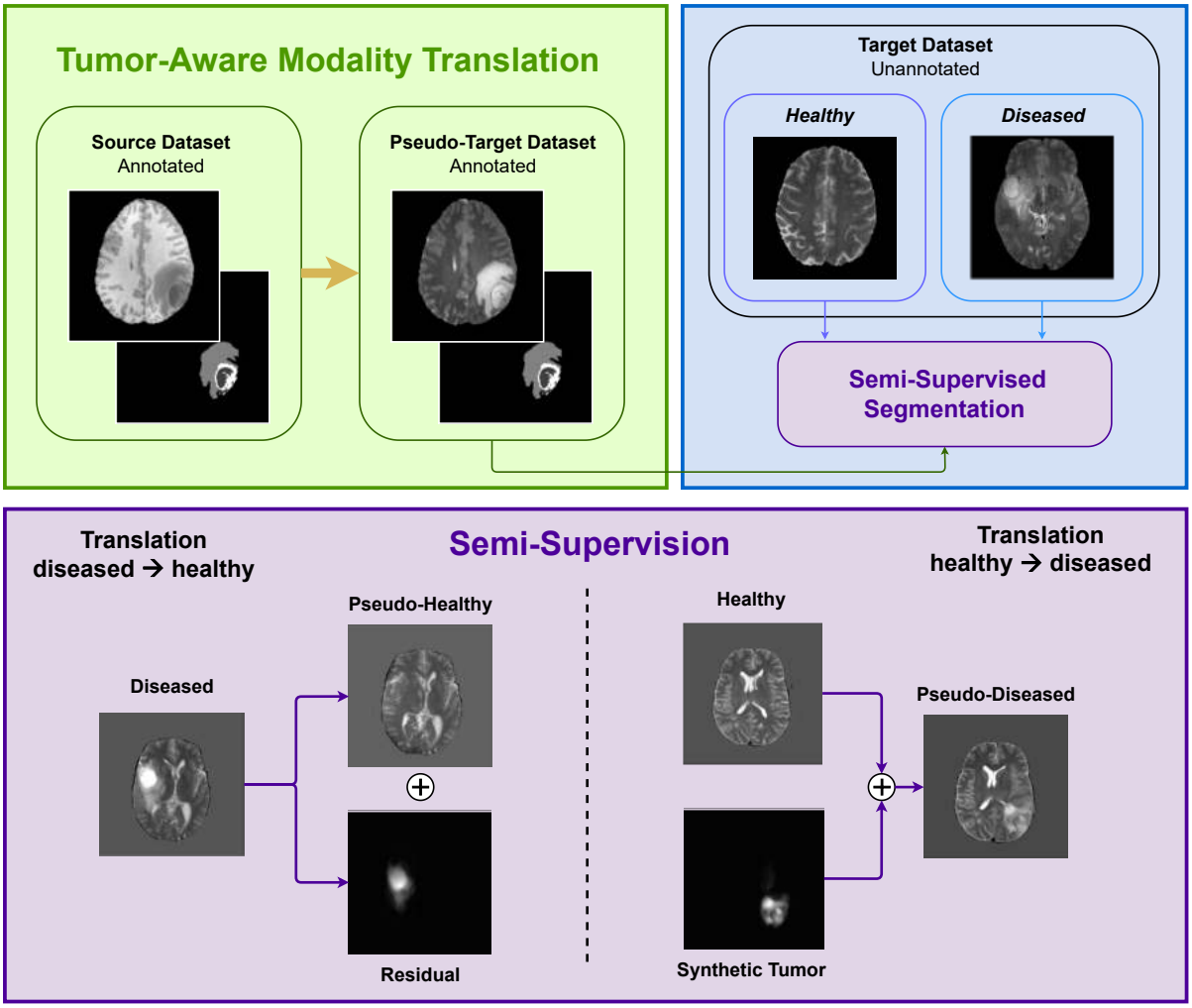


FIGURE 3.1 Principe général de M-GenSeg et MoDATTS. La première étape (verte) consiste à générer des images réalistes dans la modalité cible à partir des données sources en entraînant une translation cyclique inter-modale. Dans la deuxième étape (bleue), la segmentation est entraînée en utilisant une combinaison d'images synthétiques et réelles de la modalité cible. Le modèle de segmentation intègre une composante semi-supervisée (violet), basée sur l'idée qu'une image tumorale se décompose en une composante saine et une image résiduelle (équivalente à la segmentation de la tumeur). En exploitant des labels "faibles" indiquant si une image contient une tumeur (malade) ou pas (saine), le modèle peut apprendre à délimiter les lésions tumorales en translatant entre ces deux domaines (sain vs malade).

3.2 M-GenSeg : Modèle 2D d'adaptation de domaine

3.2.1 Défis de conception et détails de l'approche

Nous développons dans cette première étape M-GenSeg, une approche 2D pour l'adaptation de domaine entre modalités. La difficulté dans la conception de ce modèle relevait notamment du choix d'entrainer les deux étapes de génération d'images à travers les modalités et de segmentation semi-supervisée de bout-en-bout (*end-to-end method*). En effet, l'interaction entre les différents encodeurs et décodeurs qu'impliquent la combinaison des deux objectifs de translation d'image-à-image (translations entre modalités et entre domaines sain/malade) est hautement complexe. De plus, la mise en place de M-GenSeg requiert de multiples fonctions de perte qu'il convient de sélectionner avec attention. A cette fonction de perte multiple est associée une multitude d'hyper-paramètres qu'il est nécessaire d'optimiser, sans quoi M-GenSeg serait non fonctionnel ou instable.

Nous soulignons que, additionnellement à l'entrainement semi-supervisé dans la modalité cible, un module GenSeg est aussi entrainé sur les images sources afin que le modèle apprenne l'apparence des tumeurs dans cette modalité même lorsqu'elle est peu annotée. Cela permet de préserver les structures tumorales lors de la génération des images synthétiques de modalité cible, ce qui constitue une amélioration par rapport aux méthodes existantes. Par ailleurs, nous avons réussi à combiner efficacement le module de translation entre les modalités et le module GenSeg de sorte à ce que leurs représentations latentes et la mise à jour de leurs poids soient partagées. Finalement, nous avons intégré des portes d'attention dans les connexions raccourcies entre les encodeurs et décodeurs de M-GenSeg afin d'améliorer l'interprétabilité du modèle.

3.2.2 Evaluation de la méthode

Nous évaluons d'abord M-GenSeg sur un jeu de données issu du challenge BraTS 2020 [25,128, 129]. Ce dernier propose 369 IRMs multi-paramétriques de volumes cérébraux présentant des gliomes (tumeurs cérébrales), acquises avec des protocoles et des machines différentes dans 19 sites distincts. Chaque acquisition dispose de quatre contrastes différentes : T1, T2, T1ce et FLAIR. Chaque volume a été uniformisé pour présenter des dimensions de $240 \times 240 \times 155$ avec une résolution de $1 \times 1 \times 1$ mm. Nous proposons une version adaptée au problème de segmentation de tumeurs entre modalités, où les images sont connues comme étant "malades" (présence de tumeurs) ou "saines" (absence de tumeurs). Parmi les 369 volumes cérébraux disponibles dans BraTS, 37 ont été alloués à l'étape de validation (10%) et à l'étape de test (10%), tandis que les 295 restants ont été utilisés pour l'entraînement des modèles (80%).

Les ensembles de données ont ensuite été constitués à partir de chaque paire distincte des quatre contrastes IRM disponibles (T1, T2, T1ce et FLAIR) pour la tâche d'adaptation entre modalités. Pour constituer des données d'entraînement non appariées, nous n'avons utilisé qu'un seul contraste spécifique (source ou cible) par volume d'entraînement. Nous avons donc pu expérimenter avec douze combinaisons différentes de modalités source/cible non appariées. Bien que l'ensemble de données offre plusieurs classes de segmentation (tumeur en croissance, œdème péri-tumoral, tumeur nécrotique et non croissante), notez que nous ne considérons que la globalité de chaque tumeur comme notre objectif de segmentation. Aussi, bien qu'il ne soit pas cliniquement utile d'apprendre la segmentation inter-séquences si des acquisitions multi-paramétriques sont effectuées, comme c'est le cas dans BraTS, nous soutenons que cette version modifiée offre un excellent cas d'étude pour évaluer les performances réelles de toute méthode d'adaptation de domaine pour la segmentation de tumeurs entre modalités. De plus amples informations sont disponibles au chapitre 4.

M-GenSeg a ainsi été testé avec de multiples contrastes IRMs et nous obtenons une amélioration du score de segmentation de Dice sur toutes les paires de modalités disponibles en comparaison à deux approches de l'état de l'art. Plusieurs expériences d'ablation ont de même été menées sur le modèle proposé pour confirmer nos choix de conception.

3.3 MoDATTS : Modèle 3D d'adaptation de domaine

3.3.1 Nouveautés et détails de l'approche

Dans une deuxième étape, nous avons cherché à améliorer M-GenSeg en conservant ses composantes essentielles, mais en palliant les limites liées à une segmentation réalisée sur des coupes 2D. Du fait des contraintes en ressources de calcul, il n'était pas faisable d'étendre directement la version précédente en une version 3D. Dès lors, nous avons fait le choix de dissocier l'étape de translation de modalité et l'étape de segmentation (*two-stage method*). Bien que les mises à jour du modèle ne soient alors pas partagées entre les deux tâches, les bénéfices sont multiples : (1) la segmentation peut être réalisée sur des volumes 3D complets ; (2) l'optimisation du modèle est simplifiée du fait de la réduction des interactions entre ses différentes composantes ; (3) chaque étape peut désormais profiter d'architectures plus larges et plus performantes.

La modèle de translation de modalité reste basé sur un CycleGAN 2D, auquel on intégre des décodeurs de segmentation de sorte à conserver l'information tumorale au cours de la génération d'images. Cependant on utilise cette fois une architecture TransUNet [21] hybride, basée sur des blocs transformers et convolutifs.

L'étape de segmentation conserve les bénéfices d'une approche semi-supervisée (translation entre images saines et malades) mais profite cette fois d'une architecture transformer 3D basée sur le modèle Medformer [109], un réseau capable de traiter des données à grande échelle et qui surpasse le nnU-Net [130] ainsi que d'autres transformers pour la vision [131, 132] sur plusieurs tâches de segmentation d'images médicales. Enfin, une stratégie itérative d'auto-entrainement est aussi explorée afin d'affiner les segmentations générées dans la modalité cible.

3.3.2 Evaluation de la méthode

MoDATTS est évalué dans un premier lieu sur le même jeu de données issu de BraTS que lors de la première étude (voir section 3.2.2). Dans un second temps l'approche est mise à l'épreuve sur le jeu de données issu du challenge CrossMoDA 2022 [133, 134]. Le challenge met en place une tâche de segmentation de schwannomes vestibulaires (VS) entre séquences d'IRMs T1ce et T2 haute résolution (hrT2). L'ensemble d'apprentissage est composé de 210 volumes pondérés en T1 avec les segmentations des VS fournies, ainsi que de 210 volumes hrT2 non annotés. Ces volumes proviennent de patients différents et ne sont donc pas appariés. Un ensemble de validation supplémentaire, composé de 64 images hrT2 non annotées, est disponible pour évaluer les performances des modèles sur la plateforme du challenge. Les images ont été acquises de manière équilibrée dans deux centres distincts, *London* et *Tilburg*, et présentent des résolutions et des tailles différentes, référencées dans la table 3.1 ci-dessous. De plus amples informations sur le traitement de ces données sont disponibles au chapitre 5.

Site	In-plane matrix		In-plane resolution		Slice thickness	
	T1ce	hrT2	T1ce	hrT2	T1ce	hrT2
London	512×512	384×384 448×448	0.4×0.4 mm	0.5×0.5 mm	1-1.5 mm	1-1.5 mm
Tilburg	256×256	384×384	0.8×0.8 mm	0.4×0.4 mm	1.5 mm	1 mm

Tableau 3.1 Caractéristique des données IRMs du challenge CrossMoDA.

MoDATTS a démontré une performance grandement accrue sur le jeu de données BraTS en comparaison à M-GenSeg. Par ailleurs, elle s'est démontrée compétitive dans le contexte du challenge CrossMoDA 2022 où le score de segmentation atteint celui de l'équipe ayant gagné le challenge. De plus, plusieurs expériences prouvent que MoDATTS peut être utilisé pour réduire le fardeau lié au besoin d'annotations dans l'entrainement des réseaux de neurones.

CHAPITRE 4 ARTICLE 1 : M-GENSEG: DOMAIN ADAPTATION FOR TARGET MODALITY TUMOR SEGMENTATION WITH ANNOTATION-EFFICIENT SUPERVISION

Cet article accepté pour la conférence MICCAI 2023 présente une méthode 2D d'adaptation de domaine pour la segmentation tumorale entre modalités lorsque l'une d'entre elles manque d'annotations. L'article décrit l'approche mise en place, à savoir l'utilisation d'un modèle génératif de translation entre modalités, et un objectif de segmentation semi-supervisé intégrant des images non annotées à l'entrainement.

Auteurs

Malo Alefsen de Boisredon d'Assier[1], Eugene Vorontsov[2], Samuel Kadoury[1,3]

Affiliations

[1] MedICAL Laboratory, Polytechnique Montréal, Montréal, Canada
[2] Paige, Montréal, Canada
[3] Centre de recherche du CHUM (CrCHUM), Montréal, Canada

Contribution

Recherche bibliographique, définition de la méthode, expérimentation, analyse et rédaction. Le pourcentage de la contribution réelle par rapport à celle des autres coauteurs est évaluée à 80%.

Soumission

23 février 2023 à la conférence MICCAI 2023.

4.1 Abstract

Automated medical image segmentation using deep neural networks typically requires substantial supervised training. However, these models fail to generalize well across different imaging modalities. This shortcoming, amplified by the limited availability of expert annotated data, has been hampering the deployment of such methods at a larger scale across modalities. To address these issues, we propose M-GenSeg, a new semi-supervised generative training strategy for cross-modality tumor segmentation on unpaired bi-modal datasets. With the addition of known healthy images, an unsupervised objective encourages the model to disentangling tumors from the background, which parallels the segmentation task. Then, by teaching the model to convert images across modalities, we leverage available pixel-level annotations from the source modality to enable segmentation in the unannotated target modality. We evaluated the performance on a brain tumor segmentation dataset composed of four different contrast sequences from the public BraTS 2020 challenge data. We report consistent improvement in Dice scores over state-of-the-art domain-adaptive baselines on the unannotated target modality. Unlike the prior art, M-GenSeg also introduces the ability to train with a partially annotated source modality.

Keywords
Image Segmentation, Semi-supervised Learning, Unpaired Image-to-image Translation.

4.2 Introduction

Deep learning methods have demonstrated their tremendous potential when it comes to medical image segmentation. However, the success of most existing architectures relies on the availability of pixel-level annotations, which are difficult to produce [135]. Furthermore, these methods are known to be inadequately equipped for distribution shifts. Therefore, cross-modality generalization is needed when one imaging modality has insufficient training data. For instance, conditions such as Vestibular Schwannoma, where new hrT2 sequences are set to replace ceT1 for diagnosis to mitigate the use of contrast agents, is a sample use case [134]. Recently Billot et al. [136] proposed a domain randomisation strategy to segment images from a wide range of target contrasts without any fine-tuning. The method demonstrated great generalization capability for brain parcellation, but the model performance when exposed to tumors and pathologies was not quantified. This challenge could also be addressed through unsupervised domain-adaptive approaches, which transfer the knowledge available in the "source" modality S from pixel-level labels to the "target" imaging modality T lacking annotations [115].

Several strategies can be employed to address the modality adaptation problem. For instance, generative models attempt to generalize to a target modality by performing unsupervised domain adaptation through image-to-image translation and image reconstruction. In [126], by learning to translate between CT and MR cardiac images, the proposed method jointly disentangles the domain specific and domain invariant features between each modality and trains a segmenter from the domain invariant features. Other methods [137–143] also integrate this translation approach, but the segmenter is trained in an end-to-end manner on the synthetic target images generated from the source modality using a CycleGAN [122] model. These methods perform well but do not explicitly use the unannotated target modality data to further improve the segmentation, and rely only on the availability of pixel-level annotations of source modality images.

In this paper, we propose M-GenSeg, a novel training strategy for cross-modality domain adaptation, as illustrated in Fig. 4.1. This work leverages and extends GenSeg [127], a generative method that uses image-level "diseased" or "healthy" labels for semi-supervised segmentation. Given these labels, the model imposes an image-to-image translation objective between the image domain presenting tumor lesions and the domain corresponding to an absence of lesions. Therefore, like in low-rank atlas based methods [144–146] the model is taught to find and remove a lesion, which acts as a guide for the segmentation. We incorporate cross-modality image segmentation with an image-to-image translation objective between source and target modalities. We hypothesize both objectives are complementary since GenSeg helps localizing the tumors on unannotated target images, while modality translation enables fine-tuning the segmenter on the target modality by displaying annotated pseudo-target images. We evaluate M-GenSeg on a modified version of the BraTS 2020 dataset, in which each type of sequence (T1, T2, T1ce and FLAIR) is considered as a distinct modality. We demonstrate that our model can better generalize than other state-of-the-art methods to the target modality.

4.3 Methods

4.3.1 M-GenSeg : semi-supervised segmentation

Healthy-diseased translation. We propose to integrate image-level supervision to the cross-modality segmentation task with GenSeg, a model that introduces translation between domains with a presence (P) or absence (A) of tumor lesions. Leveraging this framework has a two-fold advantage here. Indeed, *(i)* training a GenSeg module on the source modality makes the model aware of the tumor appearances in the source images even with limited

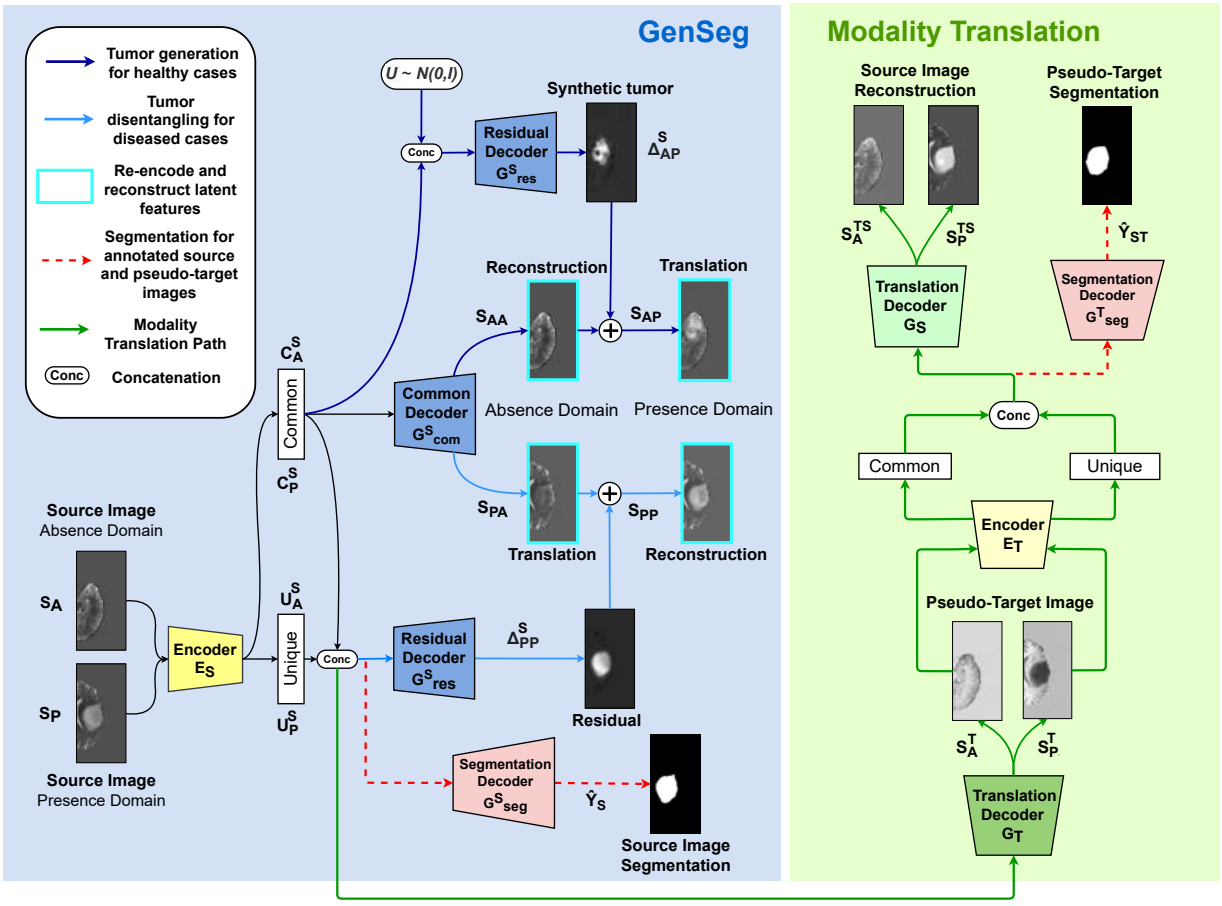


Figure 4.1 M-GenSeg: Latent representations are shared for simultaneous cross-modality translation (green) and semi-supervised segmentation (blue). Source images are passed through the source GenSeg module and the S→T*→S modality translation cycle. Domain adaptation is achieved when training the segmentation on annotated pseudo-target T* images ($\mathbf{S_P^T}$). It is not shown but, symmetrically, target images are treated in an other branch to train the T→S→T cyclic translation, and the target GenSeg module to further close the domain gap.

source pixel-level annotations. This helps to preserve tumor structures during the generation of pseudo-target samples (see section 4.3.1). Furthermore, *(ii)* training a second GenSeg module on the target modality allows to further close the domain gap by extending the segmentation objective to unannotated target data.

In order to disentangle the information common to A and P, and the information specific to P, we split the latent representation of each image into a common code $\mathbf{c}$ and a unique code $\mathbf{u}$. Essentially, the common code contains information inherent to both domains, which represents organs and other structures, while the unique code stores features like tumor shapes and location. In the two following paragraphs, we explain P→A and A→P translations for

source images. The same process is applied for target images by replacing S with T.

Presence to absence translation. Given an image $\mathbf{S_P}$ of modality S in the presence domain P, we use an encoder E_S to compute the latent representation $[\mathbf{c_P^S}, \mathbf{u_P^S}]$. A common decoder G_{com}^S takes as input the common code $\mathbf{c_P^S}$ and generates a healthy version $\mathbf{S_{PA}}$ of that image by removing the apparent tumor region. Simultaneously, both common and unique codes are used by a residual decoder G_{res}^S to output a residual image $\mathbf{\Delta_{PP}^S}$, which corresponds to the additive change necessary to shift the generated healthy image back to the presence domain. In other words, the residual is the disentangled tumor that can be added to the generated healthy image to create a reconstruction $\mathbf{S_{PP}}$ of the initial diseased image:

$$\mathbf{S_{PA}} = G_{com}^S(\mathbf{c_P^S}) \text{ and } \mathbf{\Delta_{PP}^S} = G_{res}^S(\mathbf{c_P^S}, \mathbf{u_P^S}) \text{ and } \mathbf{S_{PP}} = \mathbf{S_{PA}} + \mathbf{\Delta_{PP}^S} \tag{4.1}$$

Absence to presence translation. Concomitantly, a similar path is implemented for images in the healthy domain. Given an image $\mathbf{S_A}$ of modality S in domain A, we generate a translated version in domain P. To do so, a synthetic tumor $\mathbf{\Delta_{AP}^S}$ is generated by sampling a code from the normal distribution $\mathcal{N}(0, \mathbf{I})$ and replacing the encoded unique code for that image. The reconstruction $\mathbf{S_{AA}}$ of the original image in domain A and the synthetic diseased image $\mathbf{S_{AP}}$ in domain P are computed from the encoded features $[\mathbf{c_A^S}, \mathbf{u_A^S}]$ as follows:

$$\mathbf{S_{AA}} = G_{com}^S(\mathbf{c_A^S}) \quad \text{and} \quad \mathbf{S_{AP}} = \mathbf{S_{AA}} + G_{res}^S(\mathbf{c_A^S}, \mathbf{u} \sim \mathcal{N}(0, \mathbf{I})) \tag{4.2}$$

Like approaches in [147–149] we therefore generate diseased samples from healthy ones for data augmentation. However, M-GenSeg aims primarily at tackling cross-modality lesion segmentation tasks, which is not addressed in these studies. Furthermore, note that these methods are limited to data augmentation and do not incorporate any unannotated diseased samples when training the segmentation network, as achieved by our model with the P→A translation. initial image.

Modality translation. Our objective is to learn to segment tumor lesions in a target modality by reusing potentially scarce image annotations in a source modality. Note that for each modality $m \in \{S, T\}$, M-GenSeg holds a segmentation decoder G_{seg}^m that shares most of its weights with the residual decoder G_{res}^m, but has its own set of normalization parameters and a supplementary classifying layer. Thus, through the Absence and Presence translations, these segmenters have already learned how to disentangle the tumor from the background. However, supervised training on a few example annotations is still required to learn how to transform the resulting residual representation into appropriate segmentation maps. While

this is a fairly straightforward task for the source modality using pixel-level annotations, achieving this for the target modality is more complex, justifying the second unsupervised translation objective between source and target modalities. Based on the CycleGan [122] approach, modality translations are performed via two distinct generators that share their encoder with the GenSeg task. More precisely, combined with the encoder E_S a decoder G_T enables performing S→T modality translation, while the encoder E_T and a second decoder G_S perform the T→S modality translation. To maintain the anatomical information, we ensure cycle-consistency by reconstructing the initial images after mapping them back to their original modality. We note $\mathbf{S_d^T} = G_T \circ E_S(\mathbf{S_d})$ and $\mathbf{S_d^{TS}} = G_S \circ E_T(\mathbf{S_d^T})$, respectively the translation and reconstruction of $\mathbf{S_d}$ in the S→T→S translation loop, with domain $\mathbf{d} \in \{A, P\}$ and $\circ$ the composition operation. Similarly we have $\mathbf{T_d^S} = G_S \circ E_T(\mathbf{T_d})$ and $\mathbf{T_d^{ST}} = G_T \circ E_S(\mathbf{T_d^S})$ for the T→S→T cycle.

Note that to perform the domain adaptation, training the model to segment only the pseudo-target images generated by the S→T modality generator would suffice (in addition to the diseased/healthy target translation). However, training the segmentation on diseased source images also imposes additional constraints on encoder E_S, ensuring the preservation of tumor structures. This constraint proves beneficial for the translation decoder G_T as it generates pseudo-target tumoral samples that are more reliable. Segmentation is therefore trained on both diseased source images $\mathbf{S_P}$ and their corresponding synthetic target images $\mathbf{S_P^T}$, when provided with annotations $\mathbf{y_S}$. To such an extent, two segmentation masks are predicted : $\hat{\mathbf{y}}_\mathbf{S} = G_{seg}^S \circ E_S(\mathbf{S_P})$ and $\hat{\mathbf{y}}_\mathbf{ST} = G_{seg}^T \circ E_T(\mathbf{S_P^T})$.

4.3.2 Loss functions

Segmentation Loss. For the segmentation objective, we compute a soft Dice loss [150] on the predictions for both labelled source images and their translations:

$$\mathcal{L}_{seg} = Dice\left(\mathbf{y_S}, \hat{\mathbf{y}}_\mathbf{S}\right) + Dice\left(\mathbf{y_S}, \hat{\mathbf{y}}_\mathbf{ST}\right) \tag{4.3}$$

Reconstruction Losses. $\mathcal{L}_{cyc}^{mod}$ and $\mathcal{L}_{rec}^{Gen}$ respectively impose pixel-level image reconstruction constraints on modality translation and GenSeg tasks. Note that $\mathcal{L}_1$ refers to the standard L1 norm:

$$\begin{aligned}
\mathcal{L}_{cyc}^{mod} &= \mathcal{L}_1\left(\mathbf{S_A^{TS}}, \mathbf{S_A}\right) + \mathcal{L}_1\left(\mathbf{T_A^{ST}}, \mathbf{T_A}\right) + \mathcal{L}_1\left(\mathbf{S_P^{TS}}, \mathbf{S_P}\right) + \mathcal{L}_1\left(\mathbf{T_P^{ST}}, \mathbf{T_P}\right) \\
\mathcal{L}_{rec}^{Gen} &= \mathcal{L}_1\left(\mathbf{S_{AA}}, \mathbf{S_A}\right) + \mathcal{L}_1\left(\mathbf{S_{PP}}, \mathbf{S_P}\right) + \mathcal{L}_1\left(\mathbf{T_{AA}}, \mathbf{T_A}\right) + \mathcal{L}_1\left(\mathbf{T_{PP}}, \mathbf{T_P}\right)
\end{aligned} \tag{4.4}$$

Moreover, like in [127] we compute a loss $\mathcal{L}_{lat}^{Gen}$ that ensures that the translation task holds

the information relative to the initial image, by reconstructing their latent codes with the L1 norm. It also enforces the distribution of unique codes to match the prior $\mathcal{N}(0, \mathbf{I})$ by making $\mathbf{u_{AP}}$ match $\mathbf{u}$, where $\mathbf{u_{AP}}$ is obtained by encoding the fake diseased sample $\mathbf{x_{AP}}$ produced with random sample $\mathbf{u}$.

Adversarial Loss. For the healthy-diseased translation adversarial objective, we compute a hinge loss $\mathcal{L}_{adv}^{Gen}$ as in GenSeg, learning to discriminate between pairs of real/synthetic images of the same output domain and always in the same imaging modality, e.g. $\mathbf{S_A}$ vs $\mathbf{S_{PA}}$. In the modality translation task, the $\mathcal{L}_{adv}^{mod}$ loss is computed between pairs of images of the same modality without distinction between domains A and P, e.g. $\{\mathbf{S_A}, \mathbf{S_P}\}$ vs $\{\mathbf{T_A^S}, \mathbf{T_P^S}\}$

Overall Loss. The overall loss for M-GenSeg is a weighted sum of the aforementioned losses. These are tuned separately. All weights sum to 1. First, λ_{adv}^{Gen}, λ_{rec}^{Gen}, and λ_{lat}^{Gen} weights are tuned for successful translation between diseased and healthy images. Then, λ_{adv}^{mod} and λ_{cyc}^{mod} are tuned for successful modality translation. Finally, λ_{seg} is tuned for segmentation performance.

$$\begin{aligned}\mathcal{L}_{Total} &= \lambda_{seg}\mathcal{L}_{seg} + \lambda_{adv}^{mod}\mathcal{L}_{adv}^{mod} + \lambda_{cyc}^{mod}\mathcal{L}_{cyc}^{mod} \\ &+\lambda_{adv}^{Gen}\mathcal{L}_{adv}^{Gen} + \lambda_{rec}^{Gen}\mathcal{L}_{rec}^{Gen} + \lambda_{lat}^{Gen}\mathcal{L}_{lat}^{Gen}\end{aligned} \tag{4.5}$$

4.3.3 Implementation Details

Training and hyper-parameters. All models are implemented using PyTorch and are trained on one NVIDIA A100 GPU with 40 GB memory. We used a batch size of 15, an AMSGrad optimizer ($\beta_1 = 0.5$ and $\beta_2 = 0.999$) and a learning rate of 10^{-4}. Our models were trained for 300 epochs and weights of the segmentation model with the highest validation Dice score were saved for evaluation. The same on-the-fly data augmentation as in [127] was applied for all runs. Each training experiment was repeated three times with a different random seed for weight initialization. The performance reported is the mean of all test Dice scores, with standard deviation, across the three runs. The following parameters yielded both great modality and absence/presence translations : $\lambda_{adv}^{mod} = 3$, $\lambda_{cyc}^{mod} = 20$, $\lambda_{adv}^{Gen} = 6$, $\lambda_{rec}^{Gen} = 20$ and $\lambda_{lat}^{Gen} = 2$. Note that optimal λ_{seg} varies depending on the fraction of pixel-level annotations provided to the network for training.

Architecture. One distinct encoder, common decoder, residual/segmentation decoder, and modality translation decoder are used for each modality. The architecture used for encoders, decoders and discriminators is the same as in [127]. However, in order to give insight on the model's behaviour and properly choose the semantic information relevant for each objective, we introduced attention gates [103] in the skip connections. Fig. 4.2a shows

the attention maps generated for each type of decoder. As expected, residual decoders focus towards tumor areas. More interestingly, in order not to disturb the process of healthy image generation, common decoders avoid lesion locations. Finally, modality translators tend to focus on salient details of the brain tissue, which facilitates contrast redefinition needed for accurate translation.

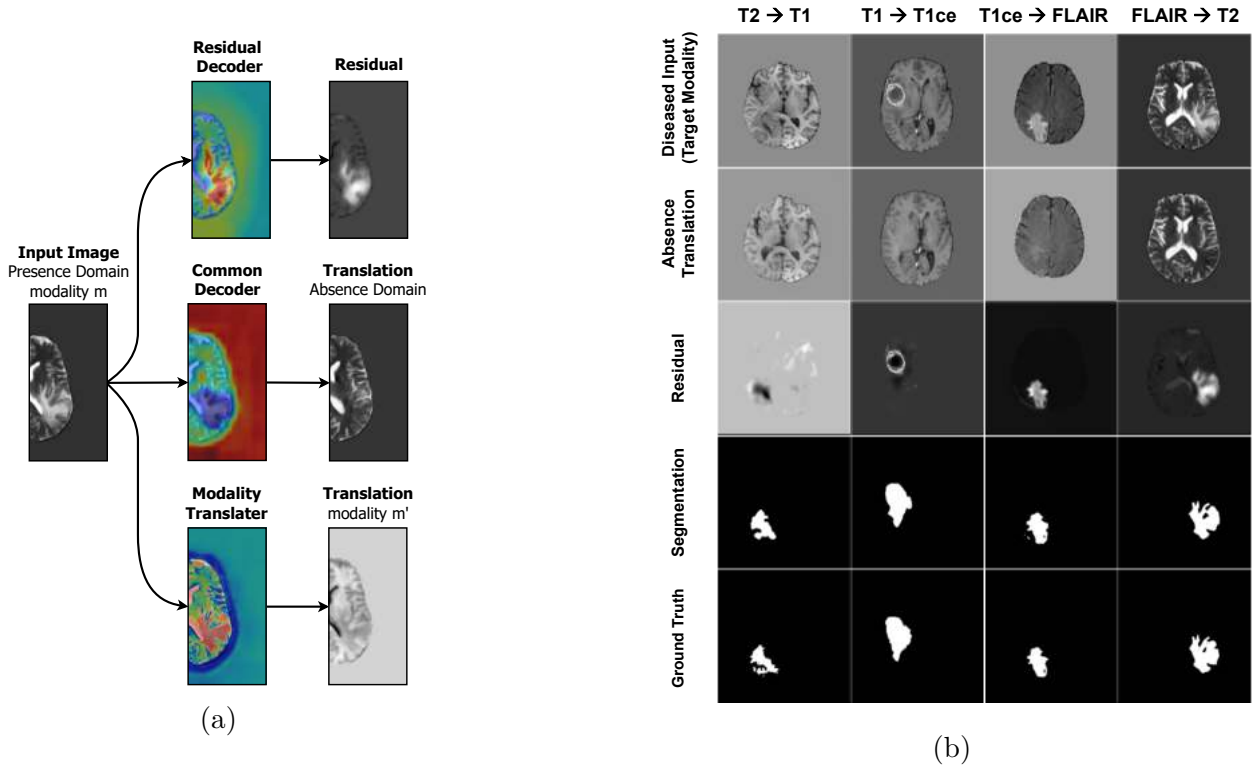


Figure 4.2 (a) Attention maps for Presence → Absence and modality translations. Red indicates areas of focus while dark blue correspond to locations ignored by the network. (b) Examples of translations from Presence to Absence domains and resulting segmentation. Each column represents a domain adaptation scenario where target modality had no pixel-level annotations provided.

4.4 Experimental Results

4.4.1 Datasets

Experiments were performed on the BraTS 2020 challenge dataset [25, 128, 129], adapted for the multi-modal brain tumor segmentation problem where images are known to be diseased (presence of tumors) or healthy (absence tumors). Amongst the 369 brain volumes available in BraTS, 37 were allocated each for validation and test steps, while the 295 left were used

for training. We split the 3D brain volumes into 2 hemispheres and extracted 2D axial slices. Any slices with at least 1% tumor by brain surface area were considered diseased. Those that didn't show any tumor lesion were labelled as healthy images. Datasets were then assembled from each distinct pair of the four MRI contrasts available (T1, T2, T1ce and FLAIR) for the modality adaptation task. To constitute unpaired training data, we used only one modality (source or target) per training volume.

4.4.2 Model Evaluation

Domain adaptation. We compared M-GenSeg with AccSegNet [141] and AttENT [137], two high performance models for domain-adaptative medical image segmentation. To that extent, we performed domain-adaptation experiments with source and target modalities drawn from T1, T2, FLAIR and T1ce. We used available GitHub code for the two baselines and performed fine-tuning on our data. For each possible source/target pair, pixel-level annotations were only retained for the source modality. We show in Fig. 4.2b several presence to absence translations and segmentation examples on different target modality images. As shown in the figure, although no pixel-level annotations were provided for the target modality, tumors were well disentangled from the brain, resulting in a successful presence to absence translation, as well as segmentation. Note that for T1 and T1ce sequences, where lesions are hypo-intense, M-GenSeg still manages to convert complex residuals into consistent segmentation maps. We plot in Fig. 4.3 the Dice performance on the target modality for *(i)* supervised segmentation on source data without domain adaptation, *(ii)* domain adaptation methods and *(iii)* UA-GAN [151], a model designed for unpaired multi-modal datasets, trained on all source and target data. Over all modality pairs our model shows an absolute Dice score increase of 0.04 and 0.08, respectively, compared to AccSegNet and AttENT.

Reaching supervised performance. As shown in Fig. 4.3, M-GenSeg performs well when the target modality is completely unannotated. Further experiments showed that with a fully annotated source modality, it is sufficient to annotate 25% of the target modality to reach 99% of the performance of fully-supervised UAGAN (e.g. M-GenSeg : 0.861 ± 0.004 vs UAGAN : 0.872 ± 0.003 for T1 $\rightarrow$ T2 experiment). Thus, the annotation burden could be reduced with M-GenSeg.

Annotation deficit. M-GenSeg introduces the ability to train with limited pixel-level annotations available in the source modality. We show in Fig. 4.4 the Dice scores for models trained when only 1%, 10%, 40%, or 70% of the source T1 modality and 0% of the T2

target modality annotations were available. While performance is severely dropping at 1% of annotations for the baselines, our model shows in comparison only a slight decrease. We thus claim that M-GenSeg can yield robust performance even when a small fraction of the source images is annotated.

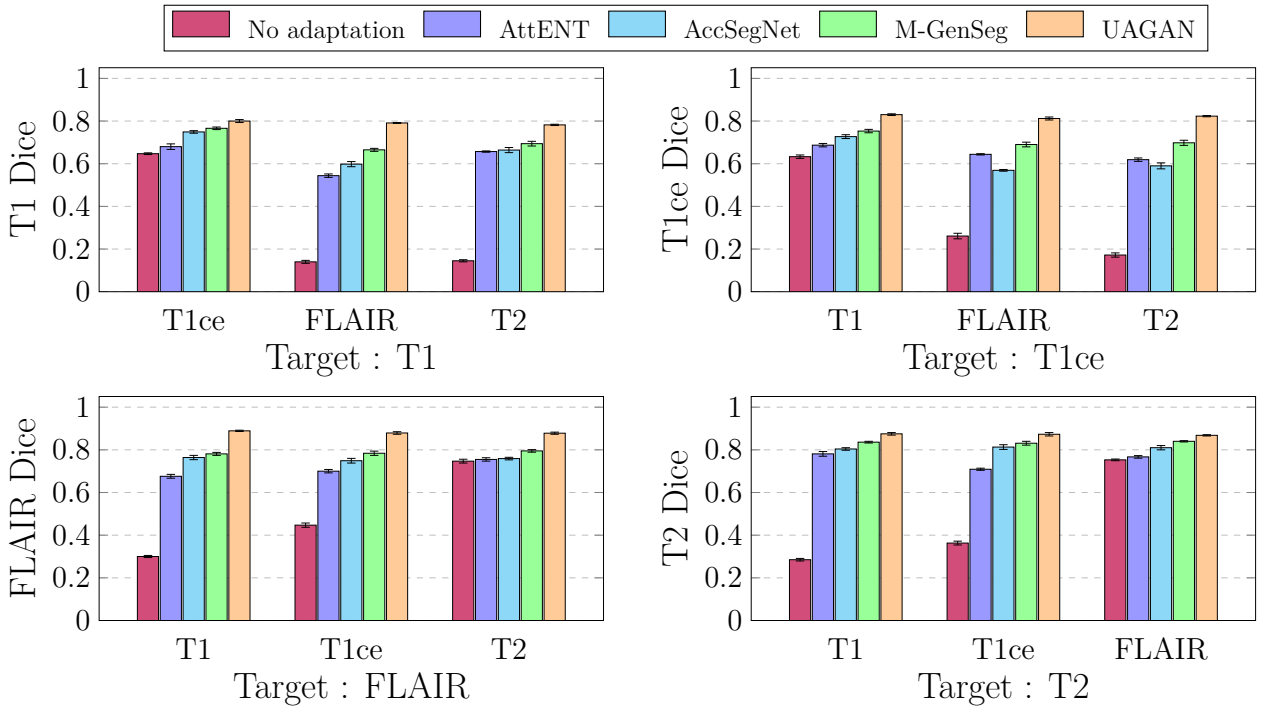


Figure 4.3 Dice performance on the target modality for each possible source modality. We compare results for M-GenSeg with AccSegNet and AttENT baselines. For reference we also show Dice scores for source supervised segmentation (No adaptation) and UAGAN trained with all source and target annotations.

4.4.3 Ablation Experiments

We conducted ablation tests to validate our methodological choices. We report in Table 4.1 the relative loss in Dice scores on target modality as compared to the proposed model. We assessed the value of doing image-level supervision by setting all the λ^{Gen} loss weights to 0 (○). Also, we showed that training modality translation only on diseased data is sufficient (○). However, doing it for healthy data as well provides additional training examples for this task. Likewise, performing translation from absence to presence domain is not necessary (○) but makes more efficient use of the data. Finally, we evaluated M-GenSeg with separate latent spaces (○) for the image-level supervision and modality translation, and we contend

that M-GenSeg efficiently combines both tasks when the latent representations share model updates.

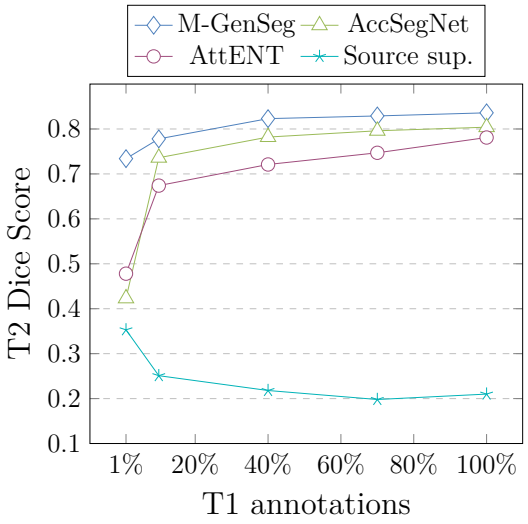


Figure 4.4 T2 domain adaptation with T1 annotation deficit.

Ablation	Mean		Std	
No image-level supervision	-8.22	±	2.71	%
No healthy modality translation	-2.41	±	1.29	%
No absence to presence translation	-3.84	±	1.71	%
Unshared latent spaces	-4.39	±	1.91	%

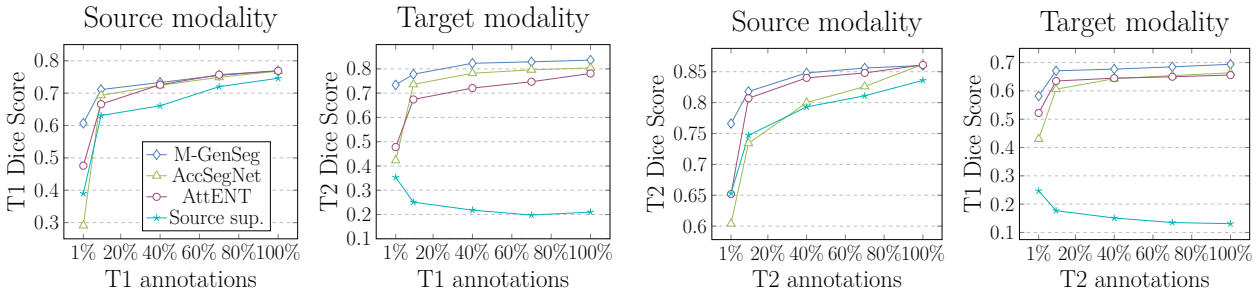
Table 4.1 Ablation studies : relative Dice change on target modality.

4.5 Conclusion

We propose M-GenSeg, a new framework for unpaired cross-modality tumor segmentation. We show that M-GenSeg is an annotation-efficient framework that greatly reduces the performance gap due to domain shift in cross-modality tumor segmentation. We claim that healthy tissues, if adequately incorporated to the training process of neural networks like in M-GenSeg, can help to better delineate tumor lesions in segmentation tasks. However, top performing methods on BraTS are 3D models. Thus, future work will explore the use of full 3D images rather than 2D slices, along with more optimal architectures.

M-GenSeg 2D : Supplementary Materials

Annotation Deficit.



(a) T1 to T2 adaptation (b) T2 to T1 adaptation

Figure 4.5 Dice scores with T1/T2 pair when using reference annotations for 0% of target data and various fractions (1%, 10%, 40%, 70% and 100%) of source data during training. For readability, standard deviations across the runs are not shown. While performance is severely dropping at 1% of annotations for the baselines, M-GenSeg shows in comparison only a slight decrease. "Source sup." refers to a model with the same architecture than M-GenSeg, trained on source data without domain adaptation.

Fully-supervised results.

Method \ Modality pair	T1/T1ce	T1/FLAIR	T1/T2
Supervised Unet (M-GenSeg backbone)	T1 : 0.766 ± 0.003 T1ce : 0.812 ± 0.004	T1 : 0.719 ± 0.005 FLAIR : 0.834 ± 0.012	T1 : $0.708 \pm 0.0.005$ T2 : 0.837 ± 0.003
UAGAN	T1 : 0.8 ± 0.007 T1ce : 0.83 ± 0.004	T1 : 0.791 ± 0.002 FLAIR : 0.889 ± 0.003	T1 : $0.782 \pm 0.0.003$ T2 : 0.875 ± 0.006
M-GenSeg	T1 : 0.791 ± 0.010 T1ce : 0.820 ± 0.006	T1 : 0.781 ± 0.008 FLAIR : 0.874 ± 0.008	T1 : $0.776 \pm 0.0.005$ T2 : 0.872 ± 0.003

Method \ Modality pair	T1ce/FLAIR	T1ce/T2	FLAIR/T2
Supervised Unet (M-GenSeg backbone)	T1ce : 0.757 ± 0.010 FLAIR : $0.838 \pm 0.0.005$	T1ce : 0.764 ± 0.004 T2 : 0.844 ± 0.004	FLAIR : 0.835 ± 0.009 T2 : 0.838 ± 0.008
UAGAN	T1ce : 0.812 ± 0.007 FLAIR : $0.879 \pm 0.0.006$	T1ce : 0.823 ± 0.003 T2 : 0.873 ± 0.008	FLAIR : 0.878 ± 0.005 T2 : 0.868 ± 0.003
M-GenSeg	T1ce : 0.810 ± 0.004 FLAIR : $0.871 \pm 0.0.004$	T1ce : 0.812 ± 0.007 T2 : 0.869 ± 0.003	FLAIR : 0.883 ± 0.007 T2 : 0.868 ± 0.004

Table 4.2 segmentation test Dice scores for fully-supervised experiments. For each modality pair, all the images of both modalities were provided with pixel-level annotations. M-GenSeg is semi-supervised but still outperforms a Unet with the same backbone, and almost achieves similar performance than UAGAN, a fully supervised method specifically designed for unpaired multi-modal datasets.

Domain adaptation absolute Dice scores.

Source \ Target	T1	T1ce	Flair	T2
T1	X	Sup : 0.633 ± 0.008 Att : 0.687 ± 0.008 Acc : 0.727 ± 0.009 Mgen : 0.753 ± 0.008	Sup : 0.300 ± 0.005 Att : 0.676 ± 0.009 Acc : 0.764 ± 0.010 Mgen : 0.781 ± 0.007	Sup : 0.285 ± 0.006 Att : 0.781 ± 0.011 Acc : 0.804 ± 0.006 Mgen : 0.836 ± 0.004
T1ce	Sup : 0.647 ± 0.004 Att : 0.680 ± 0.013 Acc : 0.749 ± 0.006 Mgen : 0.766 ± 0.006	X	Sup : 0.447 ± 0.010 Att : 0.700 ± 0.008 Acc : 0.749 ± 0.011 Mgen : 0.784 ± 0.010	Sup : 0.363 ± 0.009 Att : 0.709 ± 0.005 Acc : 0.813 ± 0.011 Mgen : 0.831 ± 0.009
FLAIR	Sup : 0.140 ± 0.007 Att : 0.544 ± 0.008 Acc : 0.598 ± 0.012 Mgen : 0.665 ± 0.007	Sup : 0.261 ± 0.013 Att : 0.644 ± 0.003 Acc : 0.569 ± 0.004 Mgen : 0.690 ± 0.011	X	Sup : 0.753 ± 0.004 Att : 0.767 ± 0.006 Acc : 0.810 ± 0.010 Mgen : 0.850 ± 0.003
T2	Sup : 0.145 ± 0.005 Att : 0.657 ± 0.003 Acc : 0.664 ± 0.012 Mgen : 0.694 ± 0.011	Sup : 0.172 ± 0.010 Att : 0.619 ± 0.008 Acc : 0.590 ± 0.014 Mgen : 0.698 ± 0.012	Sup : 0.747 ± 0.009 Att : 0.755 ± 0.008 Acc : 0.759 ± 0.006 Mgen : 0.795 ± 0.006	X

Table 4.3 Target modality segmentation test Dice scores for domain adaptation experiments (100% source annotations and 0% target annotations). 'Acc', 'Att' and 'Mgen' respectively stand for evaluated models AccSegNet, AttENT and M-GenSeg (our method). 'Sup' refers to a supervised Unet without any domain adaptation, with the same backbone architecture as the one used for M-GenSeg.

CHAPITRE 5 ARTICLE 2 : IMAGE-LEVEL SUPERVISION AND SELF-TRAINING FOR TRANSFORMER-BASED CROSS-MODALITY TUMOR SEGMENTATION

Cet article soumis au journal Medical Image Analysis présente une méthode 3D d'adaptation de domaine pour la segmentation de tumeurs visant à combler le fossé entre une modalité source annotée (ou partiellement annotée) et une modalité cible non annotée.. L'article décrit l'approche mise en place, à savoir (1) l'utilisation d'un modèle 2D de translation entre modalités capable de conserver les structures tumorales lors de la synthétisation d'images cibles; (2) un modèle de segmentation 3D intégrant une composante semi-supervisée afin d'exploiter des images connues comme étant saines; et (3) l'intégration d'une stratégie itérative d'auto-entrainement pour augmenter la taille du jeu de données d'entrainement lors de l'étape de segmentation.

Auteurs

Malo Alefsen de Boisredon d'Assier[1], Eugene Vorontsov[2], William Trung Le[1,3], Samuel Kadoury[1,3]

Affiliations

[1] MedICAL Laboratory, Polytechnique Montréal, Montréal, Canada

[2] Paige, Montréal, Canada

[3] Centre de recherche du CHUM (CrCHUM), Montréal, Canada

Contribution

Recherche bibliographique, définition de la méthode, expérimentation, analyse et rédaction. Le pourcentage de la contribution réelle par rapport à celle des autres coauteurs est évaluée à 80%.

Soumission

22 juin 2023 au journal Medical Image Analysis.

5.1 Abstract

Deep neural networks are commonly used for automated medical image segmentation, but models will frequently struggle to generalize well across different imaging modalities. This issue is particularly problematic due to the limited availability of annotated data, making it difficult to deploy these models on a larger scale. To overcome these challenges, we propose a new semi-supervised training strategy called MoDATTS. Our approach is designed for accurate cross-modality 3D tumor segmentation on unpaired bi-modal datasets. An image-to-image translation strategy between imaging modalities is used to produce annotated pseudo-target volumes and improve generalization to the unannotated target modality. We also use powerful vision transformer architectures and introduce an iterative self-training procedure to further close the domain gap between modalities. MoDATTS additionally allows the possibility to extend the training to unannotated target data by exploiting image-level labels with an unsupervised objective that encourages the model to perform 3D diseased-to-healthy translation by disentangling tumors from the background. The proposed model achieves superior performance compared to other methods from participating teams in the CrossMoDA 2022 challenge, as evidenced by its reported top Dice score of 0.87 ± 0.04 for the VS segmentation. MoDATTS also yields consistent improvements in Dice scores over baselines on a cross-modality brain tumor segmentation task composed of four different contrasts from the BraTS 2020 challenge dataset, where 95% of a target supervised model performance is reached. We report that 99% and 100% of this maximum performance can be attained if 20% and 50% of the target data is additionally annotated, which further demonstrates that MoDATTS can be leveraged to reduce the annotation burden.

Keywords
Tumor Segmentation, Semi-supervised Learning, Domain adaptation, Self-training.

5.2 Introduction

Deep learning has shown outstanding performance and potential in various medical image analysis applications [152]. Notably, it has been successfully leveraged in medical image segmentation, showing equivalent accuracy to manual expert annotations [153]. However, these breakthroughs are tempered by the issue of performance degradation when models face data from an unseen domain [154]. This problem is particularly important in medical imaging, where distribution shifts are common. Annotating data from all domains would be inefficient and intractable, notably in image segmentation where expert pixel-level labels are expensive and difficult to produce [135]. Building models that can generalize well across domains without any additional annotations is thus a challenge that needs to be addressed.

Specifically, cross-modality generalization is a key contribution towards the reduction of the data dependency and the usability of deep neural networks at a larger scale. Applications of such models are manifold, as it is common that one imaging modality lacks annotated training examples. For instance, acquisition of contrast-enhanced T1-weighted (T1ce) MR images is the most commonly used protocol for Vestibular Schwannoma (VS) detection. Accurate diagnosis and delineation of VS is of considerable importance to avoid boundless tumor growth, which can lead to irreversible hearing loss. However, in order to reduce scan time in T1ce imaging and alleviate the risks associated with the use of Gadolinium contrast agent, high resolution T2-weighted (hrT2) has recently gained popularity in clinical workflows [155]. Existing annotated T1ce databases can thus be leveraged to address the lack of training data for VS segmentation on hrT2 images. Furthermore, such models could be used for anomaly detection across modalities and pathologies, if the associated lesions show similar patterns. An example is to use pixel-level annotations of brain gliomas in MRIs to learn the distribution of intraparenchymal hemorrhages on CT scans [156].

Recently, a method to segment images from a wide range of contrasts without any retraining or fine-tuning was proposed by Billot et al. [136]. Using a generative approach conditioned on segmentations they synthetically generate images of random contrasts and resolutions, used at a later stage to train a segmentation network robust to highly heterogeneous data. However, although this domain randomisation strategy demonstrates improved generalization capability for brain parcellation, the model performance when exposed to tumours and pathologies was not quantified. More commonly, the key challenge of cross-modality generalization can be tackled through unsupervised domain adaptative (UDA) methods, which aim at leveraging the information learned from a "source" domain with abundant labeled data to improve the performance of a model on a "target" domain where labeled data is scarce or unavailable [157]. In medical imaging applications, several UDA strategies are based on feature space alignment and have been widely adopted for cross-modality organ segmentation (e.g. delineation of cardiac structures [117–119]). More widespread UDA models are based on generative strategies and tackle the issue by teaching the model to perform image-to-image translations [158] between modalities. Through adversarial training, annotated synthetic pseudo-target images can be generated from annotated source modality images and used to train a segmentation network. These methods demonstrate satisfactory results but solely rely on pixel-level annotations for source modality images. Furthermore, due to dataset and training resource constraints, these end-to-end models tend to be limited to 2D. While modality translation can be performed in 2D without performance loss, segmentation tasks highly benefit from computations on 3D volumes rather than 2D slices.

Hence we propose in this paper MoDATTS, a new **M**odality **Do**main **A**daptation **T**ransformer-

based pipeline for **T**umor **S**egmentation which aims at bridging the gap between a partially annotated source modality and an unannotated target modality. As illustrated in Fig. 5.1, our model comprises two stages for training. In the first stage a 2D network is taught to translate between imaging modalities, to eventually generate pseudo-target images from the source brain volumes. The translation generators are bounded to preserve the tumor information during the modality transfer by sharing the latent representations with segmentation decoders (see Fig. 5.2). The resultant annotated synthesized target images are then used in a second stage to teach a 3D network to perform the segmentation task (see Fig. 5.3). To alleviate the need for source annotations and extend the training to original target images, we incorporate an unsupervised anomaly detection objective on the target modality. This is done by leveraging a 2D generative strategy (GenSeg) that uses image-level "diseased" or "healthy" labels for semi-supervised segmentation [127]. Similarly to low-rank atlas based methods [144–146] the model is taught to find and remove the lesions, which acts as a guide for the segmentation. An iterative self-training procedure is also implemented to further close the gap between source and target modalities. Finally, MoDATTS leverages powerful vision transformer architectures to enhance the modality translation and segmentation.

Considering the challenge of cross-modality tumor segmentation, our main contributions are as follow :

1. We propose a tumor-aware modality translation training procedure that can accurately retain the shape of lesions.
2. We develop a 3D segmentation network that can leverage volumes known to be healthy, and explore its potential for unsupervised tumor delineation on cross-modality segmentation tasks.
3. We build our domain adaptation framework with effective vision transformer architectures.
4. The proposed model has the ability to augment the training set using pseudo-labeling and self-training mechanisms.

MoDATTS is evaluated on two distinct cross-modality tumor segmentation tasks: (i) a customized version of the BraTS 2020 dataset [25, 128, 129], where each of the four contrast sequences (T1, T2, T1ce, and FLAIR) were treated as separate modalities, and (ii) the CrossMoDA 2022 data challenge [133, 134]. We demonstrate that our model can better generalize than other state-of-the-art methods to the target modality and yield robust performance even with few source modality annotations.

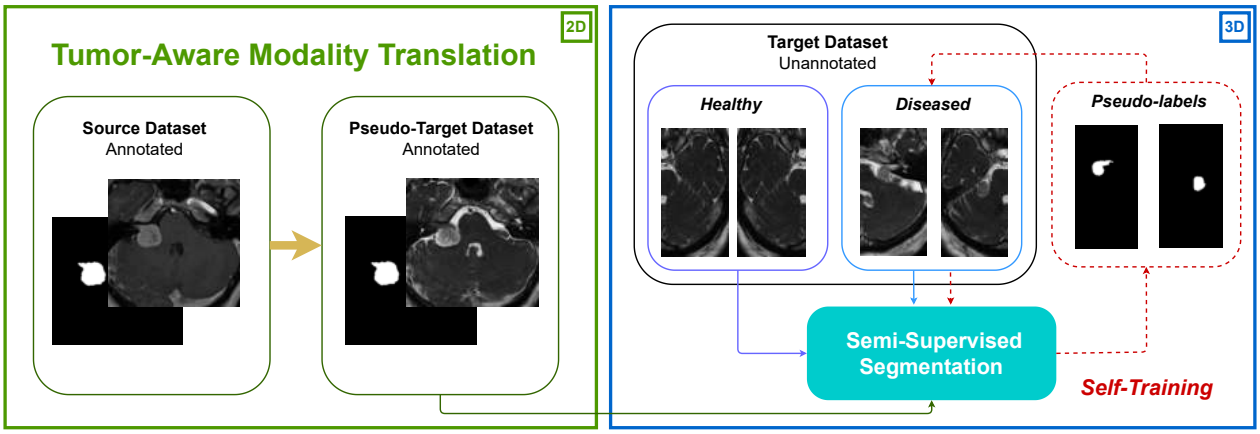


Figure 5.1 Overview of MoDATTS. Stage 1 (green) consists in generating realistic target images from source data by training cyclic cross-modality translation. In stage 2 (blue), segmentation is trained in a semi-supervised approach on a combination of synthetic and original target modality images. Finally, pseudo-labeling is performed and the segmentation model is refined through several self-training iterations.

5.3 Related Works

5.3.1 Unsupervised domain adaptation

Domain adaptation has emerged as a popular solution to address the common issue of domain shifts and heterogeneity in medical imaging. By minimizing distribution differences between related but different domains, UDA methods facilitate the use of machine learning models across varied medical image datasets [115]. Latent space alignment strategies have been widely adopted in different applications to deal with heterogeneity between sets of images from different centers (e.g. knee tissue segmentation [159] or mass detection on mammograms [160]) or with different modalities (e.g. cardiac structures segmentation on MRI using CT scans [119]). In these models, an encoder is trained to learn modality invariant representations of the images either through divergence minimization of the feature distributions or adversarial training on the latent spaces. A segmentation decoder trained on annotated source data is then bounded to produce consistent segmentation maps for the target images. For tumor segmentation tasks, generative approaches based on cross-modality translation are more frequent and will be reviewed in next section. To alleviate the need for source data availability during the adaptation stage, some source-free domain adaptation methods have been developed. Using a source segmentation model, Yang et al. [161] proposed to transform target images into high-quality source-like images with batch norm constraints. As low-frequency components in the Fourier domain can represent style information, refinement of the generated images is achieved with the mutual Fourier Transform.

Feature-level and output-level alignment is then performed based on the generated paired source-like and target images. Liu et al. [162] used a pre-trained tumor segmentation model on source T2-weighted MRI brain images and fine-tuned its parameters on the target domain (T1, T1-weighted or FLAIR) by explicitly enforcing high order batch statistics consistency and minimizing the self-entropy of predictions on the target distribution. The adaptation phase proposed by Bateson et al. [163] involved minimizing a loss function that incorporates the Shannon entropy of predictions and a prior based on the class-ratio in the target domain. However, these approaches under-perform in comparison to state-of-the-art generative methods and often relies on image-level labels incurring substantial annotation costs [163].

5.3.2 Style transfer and cross-modality segmentation

Style transfer neural networks, which involve transferring the visual appearance (or style) of one image to another while preserving the content of the latter, were first introduced by Gatys et al. [164]. Their approach enables the generation of novel images with high perceptual quality that combine the content of any given photograph with the visual style of various well-known artworks. Such models can be leveraged for domain adaptation purposes in medical image applications by generating synthetic images in the target domain to supervise a target modality segmentation model. Notably, the CycleGAN model proposed by Zhu et al. [122] became the standard for transfers between imaging modalities. CycleGAN is an unpaired bidirectional image translation network based on generative adversarial training, and preserves content specific information through cyclic pixel-level reconstruction constraints. Several works proposed a domain adaptation framework based on a CycleGAN-like approach to perform modality translation [137–141, 143]. Segmentation is jointly trained in an end-to-end manner on the labeled synthetic target images translated from the annotated source modality. Alternatively modality translation can be combined with latent space alignment to further regularize the model. Pei et al. [126] retained the principle of cyclic modality translations but proposed to jointly disentangle the domain specific and domain invariant features between each modality and train a segmenter on top of the domain invariant features. Similarly, Ouyang et al. [165] proposed a VAE-based feature prior matching mechanism to learn domain invariant features while training for modality translation and segmentation.

Note that in these methods the modality translation networks are able to maintain the structures of interest (e.g. the tumors) by integrating features from the segmentation network. Due to memory constraints, performing segmentation end-to-end with modality translation on full 3D volumes is not tractable. Thus, performing domain adaptation in a two-stage manner with 2D tumor-aware modality translation followed by 3D segmentation is adequate.

5.3.3 Self-training

Self-training is a semi-supervised learning technique with pseudo-labeling. A teacher model trained on labeled data is used to produce pseudo-labels on unlabeled data. Only pseudo-labels that the model predicts with a high probability are retained. This process can be iterated on the expanded label set, generating additional pseudo-labels. Augmenting the training set with pseudo-labeled examples increases the model's robustness towards out-of-distribution data [166]. It was shown to have great potential in leveraging unlabeled data in semantic segmentation applications [167, 168]. It is also a suitable candidate for improvements in domain adaptation tasks [169–172]. Self-training was introduced for cross-modality segmentation in the context of the CrossMoDA 2021 domain adaptation challenge [173]. In the reiteration of the challenge in 2022, self-training was a core strategy among the top ranked methods [174, 175].

5.4 Methods

Let us consider the scenario where we have a set of images X_T without pixel-level tumor annotations for a "target" modality T. The objective of this work is to learn consistent segmentations on the target data using a second set of images X_S of the "source" modality S, that is partially or totally annotated with labels Y_S. Note that the datasets are considered to be unpaired.

5.4.1 Tumor-aware cross-modality translation

The first phase of our model consists in augmenting the source images into realistic pseudo-target images (see Fig. 5.2), so that the pixel-level annotations available in the source modality can be reused to train a segmentation network on the target modality. Based on the CycleGan model [122], we perform modality translations via two distinct encoder-decoder networks (see Fig. 5.2). Encoders E_S and E_T are used to encode source and target modality images, respectively. Combined with E_S, a decoder G_T enables performing S→T modality translation, while E_T and a second decoder G_S performs the T→S modality translation. To preserve the anatomical contents, the model is forced to reconstruct the original images after mapping back to the original modality. This is referred to as cycle-consistency. We note $X_{S'} = G_T \circ E_S(X_S)$ and $X_{S''} = G_S \circ E_T(X_{S'})$, respectively the translation and reconstruction of X_S in the S→T→S translation loop. Similarly we have $X_{T'} = G_S \circ E_T(X_T)$ and $X_{T''} = G_T \circ E_S(X_{T'})$ for the T→S→T cycle. We specify that $\circ$ is the composition operation.

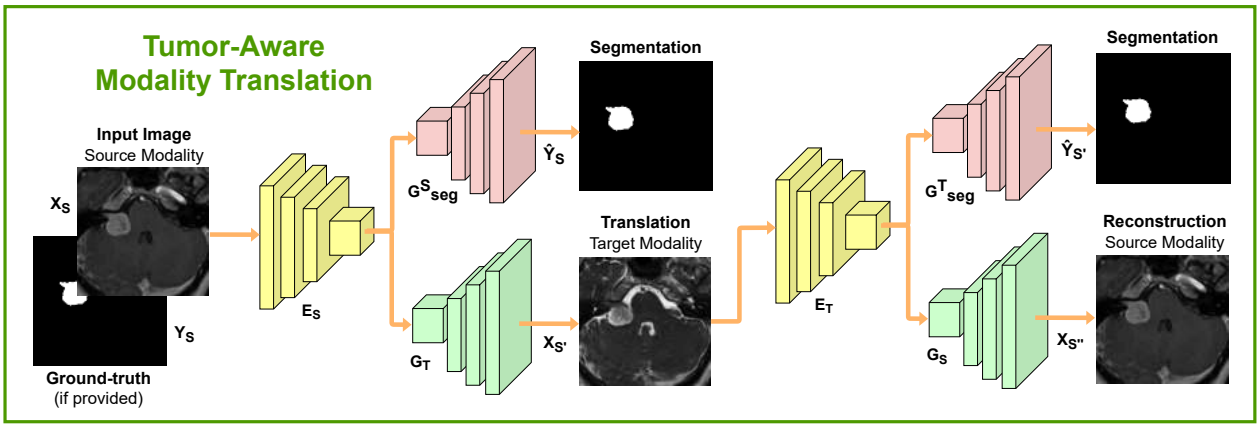


Figure 5.2 Overview of the proposed tumor-aware cross-modality translation. The model is trained to translate between modalities in a CycleGAN approach. $G_T \circ E_S$ and $G_S \circ E_T$ encoder-decoders respectively perform S $\rightarrow$ T and T $\rightarrow$ S modality translations. Same latent representations are shared with co-trained segmentation decoders G^S_{seg} and G^T_{seg} to preserve the semantic information related to the tumors. Note that the T $\rightarrow$ S $\rightarrow$ T translation loop is not represented for readability. We precise that the latter does not yield segmentation predictions since we assume no annotations are provided for the target modality.

Because image reconstruction from cycle-consistency is imperfect in practice [176], the model is guided to specifically retain detailed geometrical tumor structures by incorporating two segmentation decoders G^S_{seg} and G^T_{seg}. This has shown to be efficient for two-stage domain adaptation methods [177]. The segmentation decoders share the same latent input representation as the modality decoders, which constrains the encoders to learn features that encompass the tumors information. For each annotated source image, the model outputs a segmentation map of the original image ($\hat{Y}_S = G^S_{seg} \circ E_S(X_S)$) and the image's translation to the target domain ($\hat{Y}_{S'} = G^T_{seg} \circ E_T(X_{S'})$).

The loss function for this stage is therefore composed of‘ three terms :

1. An **adversarial objective** based on the hinge loss that aims at discriminating between real and generated images of the same modality:

$$\begin{aligned}\mathcal{L}^{mod}_{adv} = \sum_{m\in\{S,T\}} \min_G \max_D \mathbb{E}_{X_m}\left(\min\left(0, D_m(X_m) - 1\right)\right) \\ - \mathbb{E}_{\hat{X}_m}\left(\min\left(0, -D_m(\hat{X}_m) - 1\right)\right) - \mathbb{E}_{\hat{X}_m} D_m(\hat{X}_m)\end{aligned} \tag{5.1}$$

2. A **reconstruction loss** enforcing cycle consistency:

$$\mathcal{L}_{cyc} = \|X_S - X_{S''}\|_1 + \|X_T - X_{T''}\|_1 \tag{5.2}$$

3. A **segmentation objective**, based on a differentiable soft Dice loss like in [150]:

$$\mathcal{L}_{seg}^{mod} = Dice\left(Y_S, \hat{Y}_S\right) + Dice\left(Y_S, \hat{Y}_{S'}\right) \tag{5.3}$$

The overall translation loss $\mathcal{L}_{trans}$ is a weighted sum of the aforementioned terms:

$$\mathcal{L}_{trans} = \lambda_{seg}^{mod}\mathcal{L}_{seg}^{mod} + \lambda_{adv}^{mod}\mathcal{L}_{adv}^{mod} + \lambda_{cyc}\mathcal{L}_{cyc} \tag{5.4}$$

Note that to facilitate the hyper-parameter search, weights are normalized so that their sum always equals to 1.

5.4.2 Target modality segmentation

Supervision from pseudo-target data Once modality translation is learned, we yield a dataset of pseudo-target images X_{pT} and their corresponding pixel-level annotations Y_{pT}. Like in state-of-the-art methods, prior to self-training iterations we train a 3D segmentation network by teaching an encoder E and a decoder G_{seg} the segmentation task on this synthetic data. Based on images X_{pT}, we predict segmentation maps $\hat{Y}_{pT} = G_{seg} \circ E(X_{pT})$ that can be compared to the ground-truths Y_{pT} for model optimization. The corresponding loss function is termed :

$$\mathcal{L}_{seg}^{pT} = Dice\left(Y_{pT}, \hat{Y}_{pT}\right) \tag{5.5}$$

Semi-supervised segmentation Unlike prior methods that simply train the segmentation network on the annotated pseudo-target images before performing self-training, we propose a semi-supervised approach (see Fig. 5.3). By using the GenSeg training strategy [127], we believe the model can better fit the target modality distribution than with only supervision from the pseudo-target data which may still suffer from a small distribution shift. This also allows us to model relevant tumor representations even when only few source images have pixel-level annotations.

To localize lesions, we use image-level labels that describe whether an image contains a lesion or not. These "diseased" and "healthy" labels can be efficiently leveraged by a generative model by translating between presence and absence (of tumor lesions) domains, referred to as P and A. In this setup, we seek to separate the information that is shared between the two domains (A and P) from the information that is specific to domain P (that is, separate out the lesions). We therefore divide the latent representation of each image into two distinct codes: c and u. The common code c contains information that is inherent to both domains, such as organs and other structures, while the unique code u stores features specific to domain P, such as tumor shapes and localization.

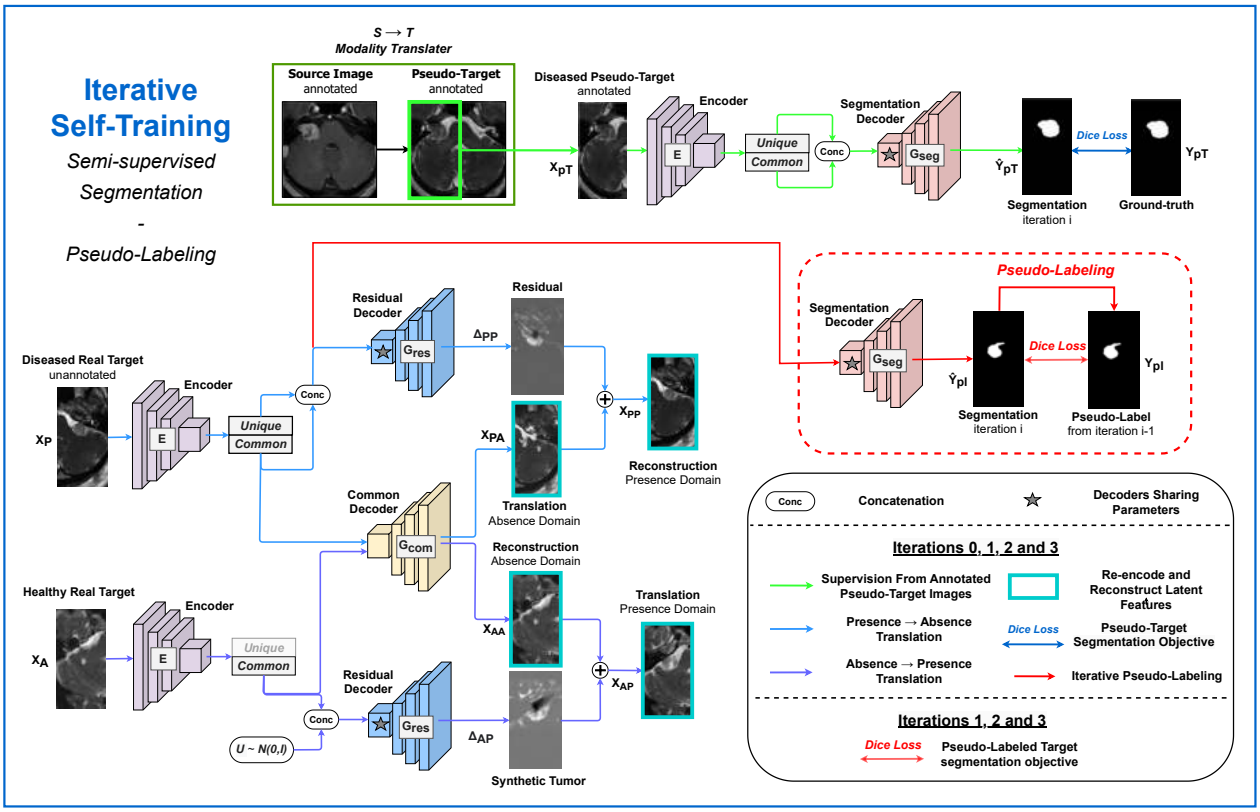


Figure 5.3 Overview of the segmentation stage. A segmentation decoder G_{seg} is trained for tumor segmentation on annotated pseudo-target images resulting from stage 1. Simultaneously a common decoder G_{com} and a residual decoder G_{res} are jointly trained for unsupervised tumor delineation on real target images by performing diseased → healthy and healthy → diseased translations. The segmentation decoder shares parameters with the residual decoder to benefit from this unsupervised objective. Finally, The segmentation encoder-decoder $G_{seg} \circ E$ is used to generate pseudo-labels for unannotated target images. The model is then further refined through several self-training iterations.

Presence to absence translation Given an original image of the target modality in the presence domain X_P, we use the encoder E to compute its latent representation $[c_P, u_P]$. A common decoder G_{com} interprets the common code c_P and generates a healthy version X_{PA} of that image by removing the apparent tumor region. At the same time, a residual decoder G_{res} employs both common and unique codes to produce a residual image Δ_{PP}, which represents the additive modification required to shift the generated healthy image back to the presence domain. In other words, the residual is the disentangled tumor that can be added to the generated healthy image to create a reconstruction X_{PP} of the initial diseased image:

$$X_{PA} = G_{com}(c_P), \tag{5.6}$$

$$\Delta_{PP} = G_{res}(c_P, u_P), \tag{5.7}$$

$$X_{PP} = X_{PA} + \Delta_{PP}. \tag{5.8}$$

Absence to presence translation In parallel, a similar process is implemented for images in the healthy domain. Given an original target image X_A in the absence domain A, a translated version in domain P is generated. Hence, a synthetic tumor Δ_{AP} is created by sampling a code from a prior distribution $\mathcal{N}(0, I)$ and substituting the encoded unique code for that image. The reconstruction X_{AA} of the original image in domain A and the synthetic diseased image X_{AP} in domain P are calculated from the encoded features $[c_A, u_A]$ in the following manner:

$$X_{AA} = G_{com}(c_A), \tag{5.9}$$

$$X_{AP} = X_{AA} + G_{res}(c_A, u \sim \mathcal{N}(0, I)). \tag{5.10}$$

A tumor can have various appearances, which means that translating from the absence to the presence domain requires a one-to-many mapping. For this reason, the unique code is replaced by a code sampled from a normal distribution $\mathcal{N}(0, I)$. Each different sampled unique code can then be interpreted by the residual decoder as a different tumor. Additionally, we reconstruct the latent representations of the generated images in both translation directions to ensure that the information from the original image is preserved. Note that in the absence-to-presence direction this enforces the distribution of unique codes to match the prior $\mathcal{N}(0, I)$. Indeed, we make u_{AP} match u, where u_{AP} is obtained by encoding the fake diseased sample X_{AP} produced with random sample u. It is worth noting that translating from the absence to the presence domain indirectly augments the target modality data, which in turn improves the domain adaptation.

Weight sharing We use a configuration where the segmentation decoder G_{seg} shares most of its weights with the residual decoder G_{res} and only differs from the latter by a distinct set of normalization parameters and the addition of a classifying layer. Therefore, through the Absence and Presence translations the segmentation decoder is implicitly learning how to disentangle the tumors from the background on original target modality samples. Additional supervision from the pseudo-target data is still required to teach the segmentation decoder how to transform the resulting residual representation into appropriate segmentation maps. However the requirement for source pixel-level annotations is reduced in comparison to usual domain adaptation methods.

Loss function To enforce the diseased-healthy translation we rely on the three components exposed below. Note that the synthetic images X_{pT} are excluded from these terms:

1. A healthy-diseased translation **adversarial loss**. We build a hinge loss $\mathcal{L}_{adv}^{gen}$ like in Eq. 5.1 aiming at discriminating between pairs of real/synthetic images of the same output domain i.e. X_A vs X_{PA} and X_P vs X_{AP}.
2. A pixel-level **image reconstruction loss** $\mathcal{L}_{rec}$ to regularize the translation between A and P domains :
$$\mathcal{L}_{rec} = \|X_{AA} - X_A\|_1 + \|X_{PP} - X_P\|_1 \tag{5.11}$$
3. A **latent code reconstruction loss** $\mathcal{L}_{lat}$ that forces the model to preserve information, enforcing a one-to-one correspondence between latent codes and their corresponding images.

The image and latent code reconstruction losses together prevent mode collapse and makes sure that when a tumor is added or removed, the background tissue remains the same. To train the 3D segmentation model, we define a global weighted sum $\mathcal{L}_{Seg}^{init}$ that encompass the diseased-healthy translation losses along with the synthetic supervision term $\mathcal{L}_{seg}^{pT}$ (Eq. 5.5):

$$\mathcal{L}_{Seg}^{init} = \lambda_{adv}^{gen}\mathcal{L}_{adv}^{gen} + \lambda_{lat}\mathcal{L}_{lat} + \lambda_{rec}\mathcal{L}_{rec} + \lambda_{seg}^{pT}\mathcal{L}_{seg}^{pT} \tag{5.12}$$

In the same way as for modality translation in the first phase (Sec. 5.4.1), weights are normalized so that their is sum equal to 1 in order to ease hyper-parameter tuning.

Self-training At this stage the model has already been trained on real target modality images through the diseased-healthy translation objective. However, the segmentation decoder would specifically benefit from tuning on the original data as it was essentially trained on the synthetic pseudo-target images. We thus further include non-annotated original data to the segmentation objective with a self-training procedure as Shin et al. did [177]. To do so, the segmentation model is used to output probability maps for each target domain images. These are then thresholded with a value α to keep only the predictions in which the model has a high confidence. The resulting pseudo-labels Y_{pl} are considered as new ground-truth annotations for the unannotated target images for finetuning the segmentation model. This procedure can be repeated k times to iteratively refine the pseudo-labels and improve the model segmentation performance on the unannotated modality. During the i^{th} self-training iteration we thus compute predictions $\hat{Y}_{pl}$ that can be compared to the pseudo-labels Y_{pl}^{i-1} resulting from the $(i-1)^{th}$ training stage. This is done with an additional self-training segmentation term $\mathcal{L}_{seg}^{st} = Dice\left(Y_{pl}^{i-1}, \hat{Y}_{pl}\right)$. Hence we obtain the following global loss for

self-training iterations:

$$\mathcal{L}_{Seg}^{ST} = \mathcal{L}_{Seg}^{init} + \lambda_{seg}^{st}\mathcal{L}_{seg}^{st} \tag{5.13}$$

Here, we set $\lambda_{seg}^{st} = \lambda_{seg}^{pT}$ for a balanced segmentation objective between the annotated pseudo-target images and the pseudo-labeled original target images.

5.5 Experiments and results

5.5.1 Experimental settings

Datasets

BraTS We first evaluate MoDATTS on the BraTS 2020 challenge dataset [25, 128, 129], adapted for the cross-modality brain tumor segmentation problem where images are known to be diseased (presence of tumors) or healthy (absence tumors). Amongst the 369 brain volumes available in BraTS, 37 volumes were allocated to each of the validation (10%) and test (10%) sets. The 295 volumes left were used for training (80%). Based only on brain tissue, each volume was mean-centered, divided by five times the standard deviation and clipped to the [-1,1] interval. Datasets were then assembled from each distinct pair of the four MRI contrasts available (T1, T2, T1ce and FLAIR) for the modality adaptation task. To constitute unpaired training data, we used only one specific contrast (source or target) per training volume. We could therefore experiment with twelve different combinations of unpaired source/target modalities. Even though it is not clinically useful to learn cross-sequence segmentation if multi-parametric acquisitions are performed as is the case in BraTS, this modified version of the dataset provides an excellent study case to assess the actual performance of any modality adaptation method for tumor segmentation. Although the dataset offers several segmentation classes (enhancing tumor, peritumoral edema, necrotic and non-enhancing tumor core), note that we only consider the whole tumors as our segmentation objective.

CrossMoDA We also used the dataset from the 2022 CrossMoDA domain adaptation challenge [133, 134] for a cross-modality vestibular schwannoma segmentation task. The training dataset is composed of 210 contrast-enhanced T1-weighted MR volumes with pixel-level annotations and 210 unannotated unpaired high-resolution T2-weighted MR volumes. An additional test set of 64 unannotated hrT2 images was available for performance evaluation of the models on the data challenge platform. The images were equally acquired in 2 distinct centers, *London* and *Tilburg*, and showed different resolutions and sizes. To mitigate these disparities, all the 3D images were first resampled to a spacing of $0.41 \times 0.41 \times 1$. Then, to align the volumes and later facilitate the split into known healthy and diseased samples, we

selected a hrT2 image as an atlas to perform inter and intra modality affine registrations. We used the Advanced Normalization Tools module [178], and employed the mutual information loss for T1ce images and the normalized cross-correlation loss for hrT2 volumes. Images were then cropped to an ROI of $256 \times 256 \times 60$. Each volume was finally mean-centered, divided by five times the standard deviation and clipped to the [-1,1] interval.

Tumor-aware cross-modality translation

2D slicing Due to resource constraints, prior to 3D segmentation, MoDATTS achieves 2D cross-modality translation to generate pseudo-target samples from the source modality. Therefore, CrossMoDA and BraTS volumes were respectively split into full 256×256 and 240×240 2D slices before being fed to the modality translation network.

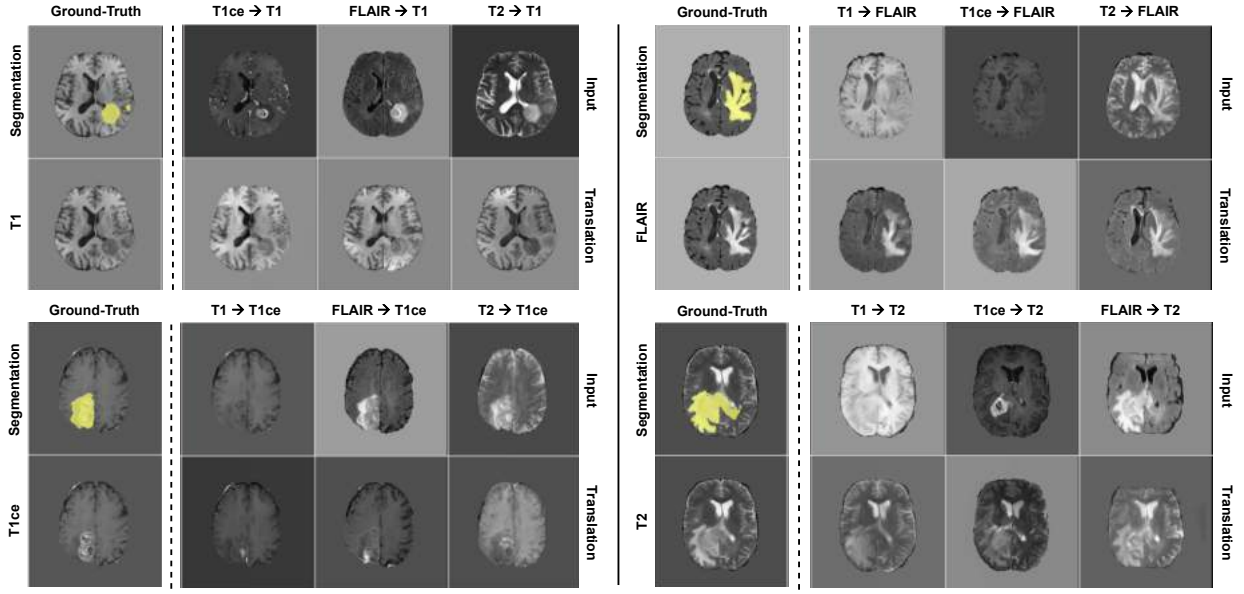


Figure 5.4 Cross-modality translation examples for the BraTS dataset. Each group of images represents all possible translations towards a specific modality (T1, T1ce, FLAIR or T2) for one case. For visual evaluation of the method we also show the corresponding ground truth target images and the whole tumor segmentations. The resulting visual appearances differ depending on the source modality, but tumor information is retained across all translations.

Architecture For the architecture, we leverage the recent works on vision transformers [20], which were shown to be suitable for translation tasks [179] when combined with fully-convolutional discriminators. We exploit the TransUnet model [21], a powerful 2D U-shaped network for medical images that has shown great performance on several segmentation tasks.

The architecture of our generators is based on the hybrid "R50-ViT" TransUnet configuration that combines a ResNet-50 and a ViT model with 12 transformer layers. The encoder backbones were pre-trained on ImageNet [180]. For each of the two modality generators, the TransUnet decoder is duplicated producing a segmentation decoder with a sigmoid output activation and a translation decoder with a *tanh* output activation. As for discriminators, we use a convolutional multi-scale architecture as in [181] that averages output values across several scales. Further details on the different layers are showed in Table 5.1. We use leaky ReLU with a slope of 0.2 as the non-linear activation function.

Discriminators			
Layer	Channels	Kernel size	Stride
C	60	4×4	1
LN+LR+C	60	4×4	2
LN+LR+C	120	4×4	2
LN+LR+C	240	4×4	2
LN+LR+C	480	4×4	2
C	1	1×1	1

Table 5.1 Discriminator architecture used for the modality translation stage. **LN** = Layer Normalization, **LR** = Leaky ReLU activation, **C** = Convolution. The same architecture is used to train the diseased-healthy translation in the second stage, but kernels are extended to 3D.

Training Our modality translation model was trained for 200 epochs. We used a batch size of 15 with the AMSGrad optimizer with $\beta_1 = 0.5$, $\beta_2 = 0.999$, and a learning rate of 0.0001. Pixel-level ground-truth annotations were provided only for the source modality slices. The same on-the-fly 2D data augmentation as in [127] was applied. The following loss parameters, defined in section 5.4.1, yielded great translation and preserved tumor appearance across modalities for both datasets : $\lambda_{adv}^{mod} = 1$, $\lambda_{seg}^{mod} = 1$ and $\lambda_{cyc} = 10$.

Synthetic target data generation Once the translation model was trained, 2D source slices were augmented into pseudo-target images using the last state of the translation model. The latter were then assembled back to constitute the synthetic target 3D volumes required for the segmentation stage.

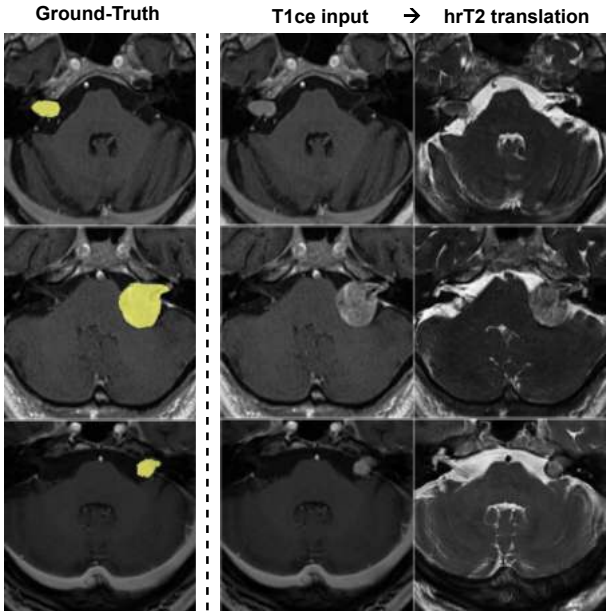


Figure 5.5 Cross-modality translation examples for the CrossMoDA dataset. We display several T1ce → hrT2 translations along with the VS segmentation ground truth. Tumor structures are preserved in the pseudo hrT2 images.

Semi-supervised target modality segmentation

Diseased/Healthy labeling For the segmentation task, we created the sets of known healthy and diseased volumes. Each pseudo and real target 3D brain volumes were split into two hemispheres. For BraTS we attributed to each hemisphere the label P (presence of tumor) if any of its pixels was indicated to be part of a tumor by the ground truth segmentation, or the label A (absence of tumor) otherwise. Since Vestibular Schwannoma (VS) segmentations were available for T1ce in the CrossMoDA dataset, the same process was applied for the synthetized hrT2 images. However as the ground truth VS segmentations were not provided for original hrT2 data, manual labels were added to left or right hemispheres containing the VS for each of the 210 hrT2 volumes. We specify that in all the experiments the images are provided with absence/presence weak labels, distinct from the pixel-level annotations that we provided only to a subset of the data.

Architecture Integrating the diseased/healthy translation to the domain adaptation framework requires a common and a residual decoder in addition to the standard segmentation encoder-decoder. As stated before, weights between the segmentation decoder and the residual decoder are shared so that segmentation is implicitly learned from the unsupervised objective. Therefore our model actually involves one unique residual/segmentation decoder but with two sets of normalization parameters. The latter also contains two distinct output layers, with tanh activation to generate residuals and sigmoid activation to yield segmentation maps. Our encoder and decoder architectures are based on a 3D Medformer [109], a recent data-scalable transformer architecture that outperformed the nnU-Net [130] and other vision transformers [131, 132] on several medical image segmentation tasks. Note that we propose a self-supervised setup which only includes the supervision over pseudo-target samples and self-training (no common/residual decoders), and a semi-supervised setup which additionally performs the diseased/healthy translation on original target images. The number of weights per encoder/decoder for the semi-supervised variant had to be decreased in comparison to the self-supervised variant due to memory constraints. Further details on the different encoder and decoder layers are provided in Table 5.2. In the semi-supervised variant, we also introduce two discriminators that are responsible for discriminating between real and generated diseased/healthy samples. Their architecture is the same as in the modality translation stage (see Table 5.1).

Encoder						
	Ch. sf.	Ch. sm.	Conv.	Trans.	Heads	Kernel
Conv. Stem	16	32	1	0	0	$3 \times 3 \times 3$
	32	64	2	0	0	$\dagger\, 3 \times 3 \times 3$
B-MDH blocks	64	128	0	2	2	$\dagger\, 3 \times 3 \times 3$
	128	256	0	4	8	$\dagger\, 3 \times 3 \times 3$
	256	320	0	6	10	$\dagger\, 3 \times 3 \times 3$

Decoders						
	Ch. sf.	Ch. sm.	Conv.	Trans.	Heads	Kernel
B-MDH blocks	128	256	0	4	8	$\ddagger\, 3 \times 3 \times 3$
	64	128	0	2	4	$\ddagger\, 3 \times 3 \times 3$
Conv. Decoder	32	64	2	0	0	$\ddagger\, 3 \times 3 \times 3$
	16	32	2	0	0	$\ddagger\, 3 \times 3 \times 3$
	1	1	1	0	0	$\star\, 1 \times 1 \times 1$

Table 5.2 Architectures used for training at the segmentation stage. Except for the common decoder for which we add an extra convolution layer at the bottleneck with kernel size $1 \times 1 \times 1$ to map the common code back to the encoder output channel number, we use identical architectures for all decoders. Ch. sm. and Ch. sf. respectively refers to the number of channels for the semi-supervised and self-supervised variants. At each layer we show the number of convolution blocks (Conv.), and the number of B-MDH - Bidirectional Multi-Head Attention - blocks (Trans.) along with the number of heads for each of these blocks (Heads). $\dagger = \mathbf{2}\times$ down-scaling with tri-linear interpolation and passing semantic maps to multi-scale fusion module. $\ddagger = \mathbf{2}\times$ up-sampling with tri-linear interpolation and long skip connection from multi-scale fusion module. $\star$ indicates which layer is duplicated in the shared residual/segmentation decoder.

Training All segmentation models were trained for 300 epochs. We then performed three self-training iterations of 150 epochs each. For all runs, we applied 3D nnU-Net on-the-fly data augmentation and weights with the highest validation Dice score were saved for the next step. We used a batch-size of 2, and the same optimizer parameters as in the modality translation phase. The threshold α that defines the level of confidence required to keep the pseudo-labels was set to 0.6 as in [182]. For BraTS, each training experiment was repeated three times, with a different random seed for weight initialization. We therefore report the mean of all test Dice scores with standard deviation across the three runs. Specifically in the segmentation stage, we performed 5-fold cross-validation on the training set for each CrossMoDA experiment. Ensembling was achieved with the resulting models for performance evaluation on the test dataset. As we were limited for quantitative evaluation on the online CrossMoDA data challenge platform, each experiment was evaluated only once. Note that for VS segmentation we applied up-sampling on large heterogeneous and small-sized tumors

along with tumor intensity augmentation, as encouraged by Sallé et al. [175].

Hyper-parameter search Our approach involved the following strategy : (1) increasing the weights of the reconstruction terms λ_{rec} and λ_{lat} relatively to the adversarial term λ_{adv}^{gen} until mode dropping stops occuring; and (2) subsequently determining the optimal weight λ_{seg}^{pT} for supervision from the synthetic data. We found that the following parameters (normalized to equal 1) yielded great diseased/healthy translations for BraTS : $\lambda_{adv}^{gen} = 5$, $\lambda_{rec} = 50$, and $\lambda_{lat} = 5$. For CrossMoDA, stronger reconstruction constraints were required as $\lambda_{rec} = 75$ and $\lambda_{lat} = 10$ yielded visually better results. In a standard domain adaptation scenario where 100% of source data is provided with pixel-level annotations and all the target images are unnanotated, we found that $\lambda_{seg}^{pT} = 100$ was optimal. Note that, during the first training of the segmentation model (prior to self-training), increasing λ_{seg}^{pT} involves higher dependence on the synthetic target data. When lowering the number of samples provided with pixel-level annotations in the source modality, (1) the tumor appearances in the generated pseudo-target images are likely to be less accurate and (2) the segmentation model is fed with fewer annotated synthetic samples. This requires adjusting λ_{seg}^{pT} down, so that the segmentation model relies more on the unsupervised tumor delineation objective rather than on the supervision from the synthetic data. When using 70%, 40%, 10% and 1% of the source modality annotations, we set λ_{seg}^{pT} to respectively 50, 25, 1 and 0.1. These λ_{seg}^{pT} values were tuned on CrossMoDA and reused for BraTS.

Model comparison

When evaluating our model on BraTS, we compare the performance of the proposed approach against state-of-the-art domain-adaptive medical image segmentation models AccSegNet [141] and AttEnt [137]. We used available GitHub code for the two baselines and performed fine-tuning on our data. Because these two methods are 2D, for fair comparison, we also evaluate a 2D version of MoDATTS. Note that unless "2D" is specified, MoDATTS refers to the 3D version of our model. For CrossMoDA, we compare the performance of MoDATTS against the top 4 teams in the validation phase of the data challenge. The VS Dice scores and Average Symmetric Surface Distances (ASSD) for these methods were provided in the leader-board. For further experiments, the team Super-Poly [183] was the only one to make its code available. Their MSF-nnU-Net model ranked 4^{th} in the data challenge but we believe their approach constitutes a reasonable baseline as it used a nnU-Net in the segmentation stage, similarly to the other top methods.

5.5.2 Domain adaptation

The first set of experiments consists in evaluating MoDATTS in a standard domain adaptation scenario where all of the source data is provided with pixel-level annotations and all the target images are unannotated. This is the standard scenario for CrossMoDA as hrT2 segmentations are not available. As for BraTS we drew all possible source and target modality pairs from T1, T2, FLAIR and T1ce, and pixel-level annotations were only retained for the source modality.

Qualitative results We show in Fig. 5.4 several generated samples of pseudo-target brain images after training of the modality translation model, when all the source samples were provided with pixel-level annotations. Interestingly, each source modality leaves its own style footprint in the generated images as the tumor appearances in the resulting target translations differ accordingly. An interesting feature of our model is that the tumor structures seem to be visually preserved across the modality translation. For instance, note that the different substructures of the tumor can still be differentiated in the FLAIR $\rightarrow$ T1 modality translation. Also notice that the translation model can successfully augment hypo-intense tumors that are hardly distinguishable from the background (e.g. T1ce $\rightarrow$ FLAIR or T1 $\rightarrow$ FLAIR). Similarly we show in Fig. 5.5 several T1ce $\rightarrow$ hrT2 VS translations. This further proves successful maintenance of tumor layouts during the pseudo-target sample generations, even for small lesions.

An illustration of several translations from the diseased to the healthy domain for the brain tumor and VS segmentation tasks are displayed in Fig. 5.6. As depicted in the figure, even without any pixel-level annotations for the target modality, the tumors were effectively separated from the brain, leading to a successful translation from the presence to the absence domain, as well as accurate segmentation. It is worth noting that even for lesions appearing hypo-intense, as in brain T1 and T1ce sequences, MoDATTS can effectively handle complex residuals and alternatively convert them into reliable segmentation results.

Quantitative results We present the resulting absolute Dice scores for MoDATTS and each evaluated baseline on the brain tumor dataset in Fig. 5.7. Note that the results for MoDATTS correspond to the self-supervised variant as it was more effective than the semi-supervised variant when 100% of the source data was annotated (see section 5.5.4). Additionally we display the results obtained by a supervised Medformer model with the same backbone architecture as the self-supervised variant of MoDATTS, trained on the one hand with source data without any domain adaptation strategy and on the other hand with fully annotated target dat, which respectively act as lower and upper bounds of the domain adapta-

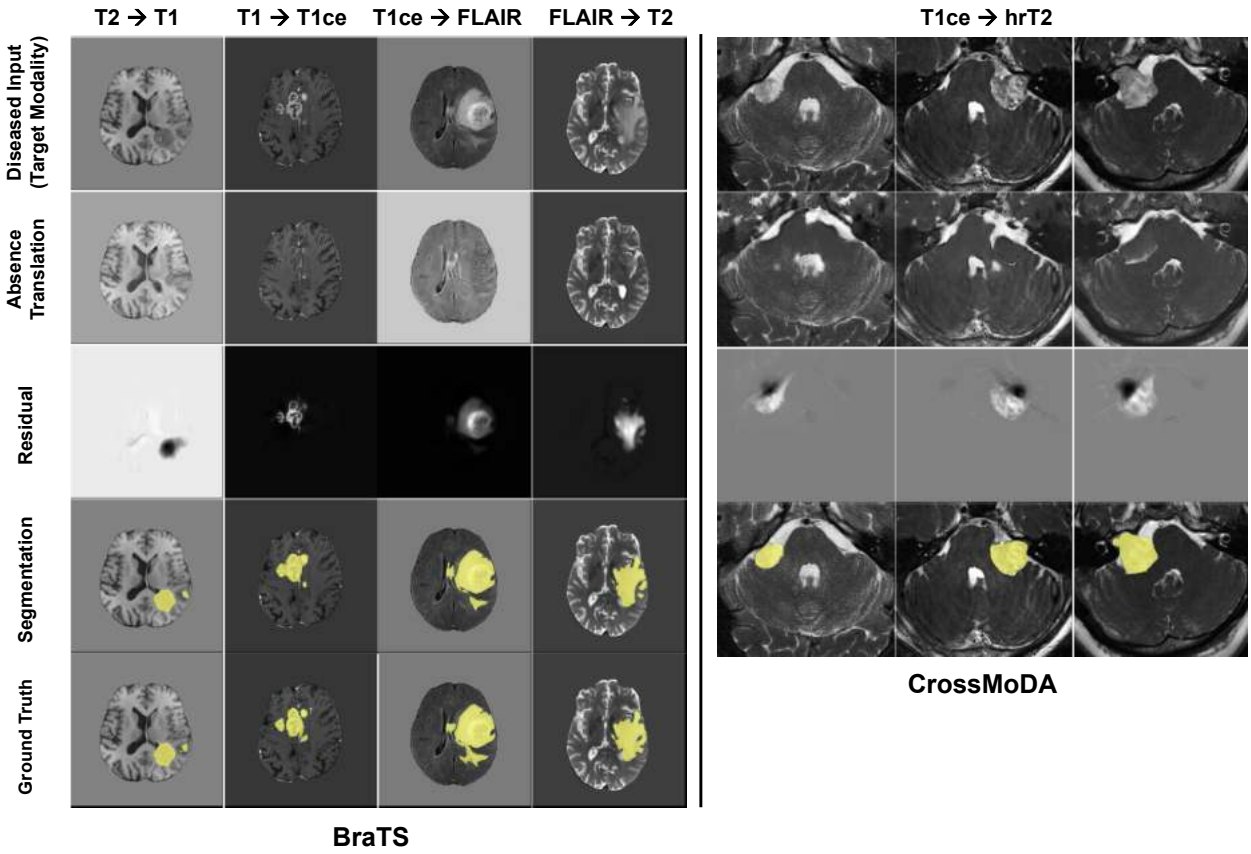


Figure 5.6 Examples of translations from Presence to Absence domains and resulting segmentation in a domain adaptation application where target modality had no pixel-level annotations provided. For BraTS, we show in each column a different source → target scenario. We do not display ground truth VS segmentations for CrossMoDA as hrT2 segmentation maps were not provided in the data challenge.

tion task. As expected, models without domain adaptation approaches trained on the source data fail to properly segment tumors on the target modality, particularly for modality pairs that show high domain shifts (e.g. T1/FLAIR or T1ce/T2). Note that MoDATTS shows great performance as it outperforms AttENT and AccSegNet with a considerable margin on the target modality. On average, over the 12 different domain adaptation experiments we report that 3D MoDATTS reaches 95% of the target supervised model performance. These results demonstrate that our transformer-based modality translation approach is effective and is able to produce reliable pseudo-target images to train a segmentation model to delineate tumors in the target modality. In comparison AttENT and AccSegNet reached 79% and 82%, respectively. The 2D version of MoDATTS, reached 87% of the target supervised model performance. This demonstrates that the superior performance of our model is not solely attributable to working in 3D.

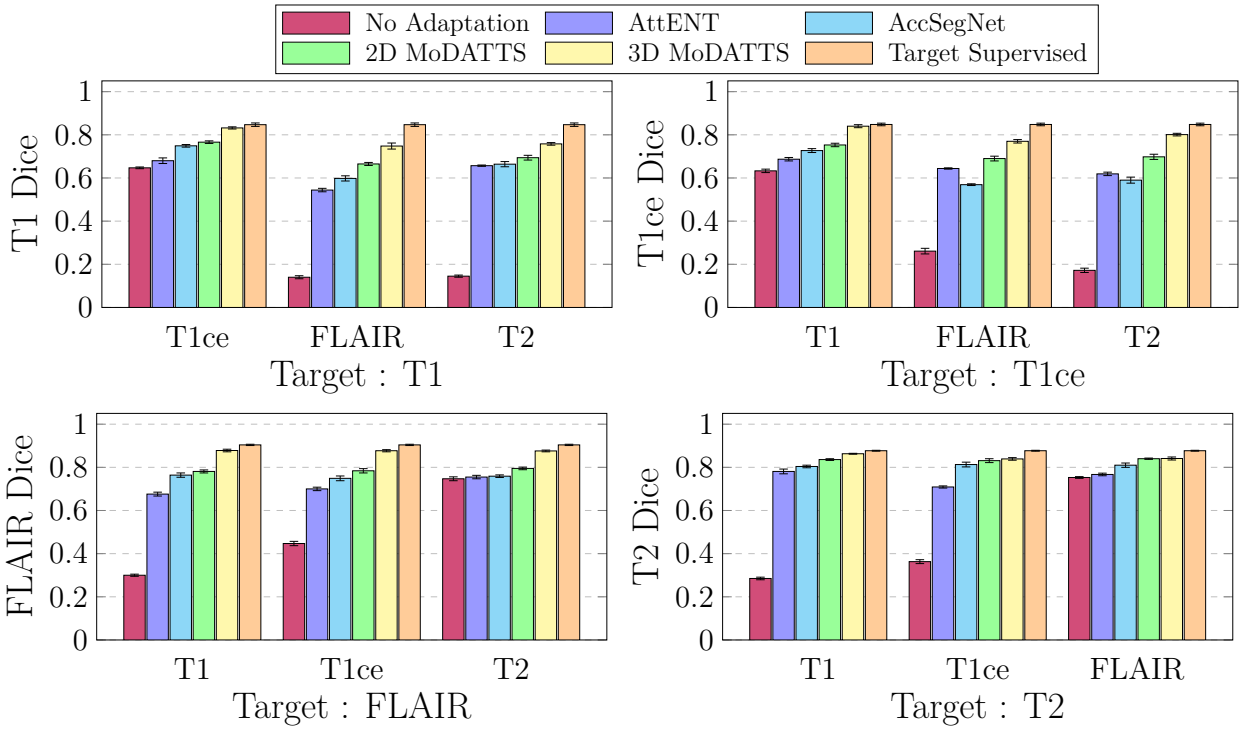


Figure 5.7 Dice performance on the target modality for each possible modality pair. Pixel-level annotations were only provided in the source modality indicated on the x axis. We compare results for MoDATTS (2D and 3D) with AccSegNet and AttENT domain adaptation baselines. We also show Dice scores for supervised segmentation models respectively trained with all annotations on source data (No adaptation) and on target data (Target supervised) as for lower and upper bounds of the cross-modality segmentation task.

We also report in Table 5.3 the VS Dice scores and ASSD on the CrossMoDA dataset. MoDATTS shows similar VS segmentation performance as team LaTIM who ranked first in the data challenge, and outperforms runner-up entries. This further proves that our method is effectively able to reduce the performance gap due to domain shifts in cross-modality tumor segmentation. Although the performance gains are limited, note that these approaches were specifically designed to perform on the CrossMoDA challenge and may not generalize well to other datasets. In contrast, our results on BraTS and CrossMoDA indicate that MoDATTS is competitive in several domain adaptation tumor segmentation tasks. We further note that the only competing approach (LaTIM) requires training a SinGAN [184] purposely adapted for CrossMoDA to augment and diversify the target VS appearances, in addition to the conventional modality translation and segmentation models. MoDATTS also achieves such data augmentation through the healthy $\rightarrow$ diseased translation objective. However it is encompassed in the segmentation model, therefore mitigating the need for an additional step.

Model	VS Dice ↑	VS ASSD ↓
[182] ne2e	0.847 ± 0.063	0.551 ± 0.303
[183] Super-Poly	0.849 ± 0.068	0.520 ± 0.229
[174] MAI	0.852 ± 0.089	0.475 ± 0.207
[175] LaTIM	0.868 ± 0.060	$\mathbf{0.430 \pm 0.178}$
MoDATTS (Ours)	$\mathbf{0.870 \pm 0.048}$	0.432 ± 0.175

Table 5.3 Dice score and ASSD for the VS segmentation on the target hrT2 modality in the CrossMoDA challenge. We compare our performance with the top 4 teams in the validation phase. The standard deviations reported correspond to the performance variation across the 64 test cases.

5.5.3 Reaching supervised performance

As mentionned in the previous section, MoDATTS performs well when the target modality is completely unannotated, as on average 95% of the target supervised model performance is reached on BraTS. With the aim of determining the fraction of target modality annotations required to match the performance of a target supervised model, we trained MoDATTS (self-supervised) with a fully annotated source modality and increasing fractions of target annotations (0%, 10%, 20%, 30% and 50%) on the BraTS T2 → T1ce domain adaptation task. Results are provided in Table 5.4. We show that with a fully annotated source modality, it is sufficient to annotate 20% of the target modality to reach 99% (T1ce : 0.839 ± 0.005) of the target supervised model performance (T1ce : 0.848 ± 0.006). This emphasizes that the annotation burden could be reduced with our approach.

T1ce annotations	0%	10%	20%	30%	50%
T1ce Dice Score	0.801 ± 0.006	0.826 ± 0.006	0.839 ± 0.005	0.844 ± 0.007	0.847 ± 0.005
% of TSMP	94.5%	97.5%	99%	99.5%	100%

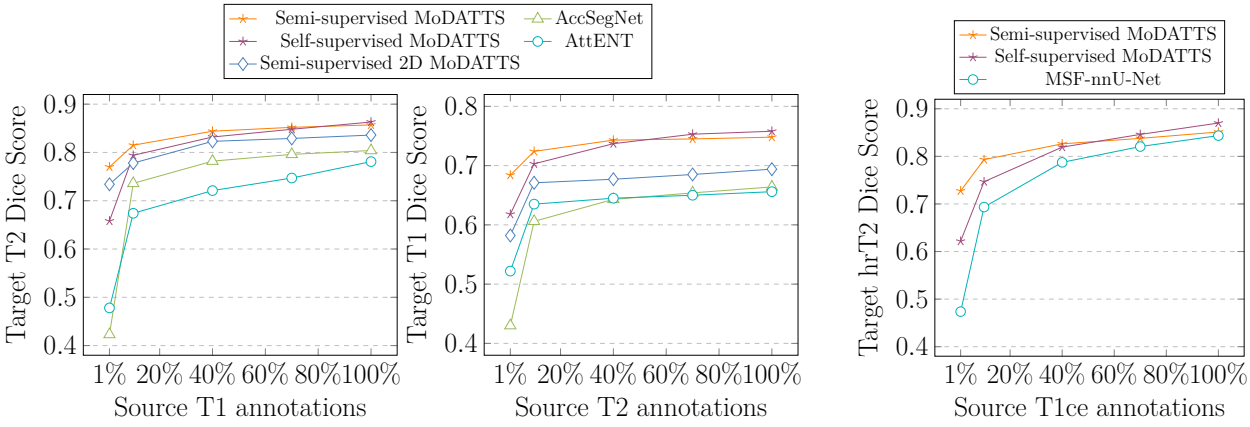
Table 5.4 Brain tumor segmentation Dice scores when using reference annotations for 100% of source T2 data and various fractions (0%, 10%, 30% and 50%) of target T1ce data during training. We also show the % of the target supervised model performance (TSMP) that is reached by MoDATTS for the corresponding fractions of T1ce annotations.

5.5.4 Semi-supervision and annotation deficit

MoDATTS introduces the ability to train with limited pixel-level annotations available in the source modality, a distinct feature over previous baselines. We show in Fig. 5.8 the Dice scores for models trained when 1%, 10%, 40%, 70% or 100% of the source modality's annotations were available combined with 0% for the target modality. Note that the modality translation networks were retrained accordingly. While the performance of the baselines and the self-

supervised variant of MoDATTS show a significant drop with fewer source annotations, the semi-supervised variant exhibits consistent performance with only slight degradation. Notably, the semi-supervised variant outperforms the self-supervised variant when less than 40% of the source samples are annotated. For instance for BraTS T1→T2 domain adaptation, semi-supervised MoDATTS with 1% source annotations still reaches 88% of the performance of a target (T2) supervised model, while the self-supervised variant only achieves 75%. These findings validate that MoDATTS has the potential achieve robust performance even with a small fraction of annotated source images.

Note that when most of the source data is annotated, the performance gap between the self-supervised and semi-supervised variants remains small. When 100% of source data is annotated, we report (self-supervised vs semi-supervised) target modality Dice scores of 0.863 vs 0.857 for T1 → T2 (BraTS), 0.758 vs 0.748 for T2 → T1 (BraTS), and 0.870 vs 0.851 for T1ce → hrT2 (CrossMoDA). This indicates that supervision from highly reliable synthetic data combined with self-training provide enough information to the segmentation model to close the domain gap. Finally, this small gap may be closed or reduced with access to better hardware since we had to reduce the size of the segmentation encoder-decoder in the semi-supervised variant from 38 million parameters (self-supervised) to 8 million parameters (semi-supervised).



(a) BraTS Dataset (b) CrossMoDA Dataset

Figure 5.8 Dice scores when using reference annotations for 0% of target data and various fractions (1%, 10%, 40%, 70% and 100%) of source data during training. For BraTS, we picked the T1/T2 modality pair and ran the experiments in both T1 → T2 and T2 → T1 directions. For CrossMoDA, hrT2 annotations are not available so the experiment is run only in the T1ce → hrT2 direction. While performance is dropping at low % of annotations for the baselines, semi-supervised MoDATTS shows in comparison only a slight decrease. For readability, standard deviations across the runs are not shown.

5.5.5 Attention in MoDATTS

Attention-based networks like transformers allow us to interrogate the model by analyzing the learned attention mechanisms. As suggested by Voita et al. [185], we computed the "confidence" of each attention head in the model as the average of its maximum attention weight. We show in Fig. 5.9 the attention maps generated by the most confident heads in MoDATTS on CrossMoDA. A confident head can be interpreted as one that assigns high attention values to specific regions of the image. We note that the transformer component (encoder) in the modality translation network tends to focus on global anatomical details of the image. Interestingly, it is also highlighting the VS, which emphasizes its ability to preserve tumor structures during the pseudo-target image synthesis. In the segmentation phase, the heads of the common and residual/segmentation decoders have different behaviors. Interestingly, the common decoder seems to avoid the tumor location, as a way to focus on the anatomical and healthy content of the image. As expected, the joint residual/segmentation decoder focuses on tumor areas in order to generate accurate residuals and segmentation maps. It also looks beyond the tumor. This is not surprising because the network has to compare the tumor to background tissue; also, a tumor can impact surrounding structures and the way the tumor appears in the image depends on the rest of the tissue.

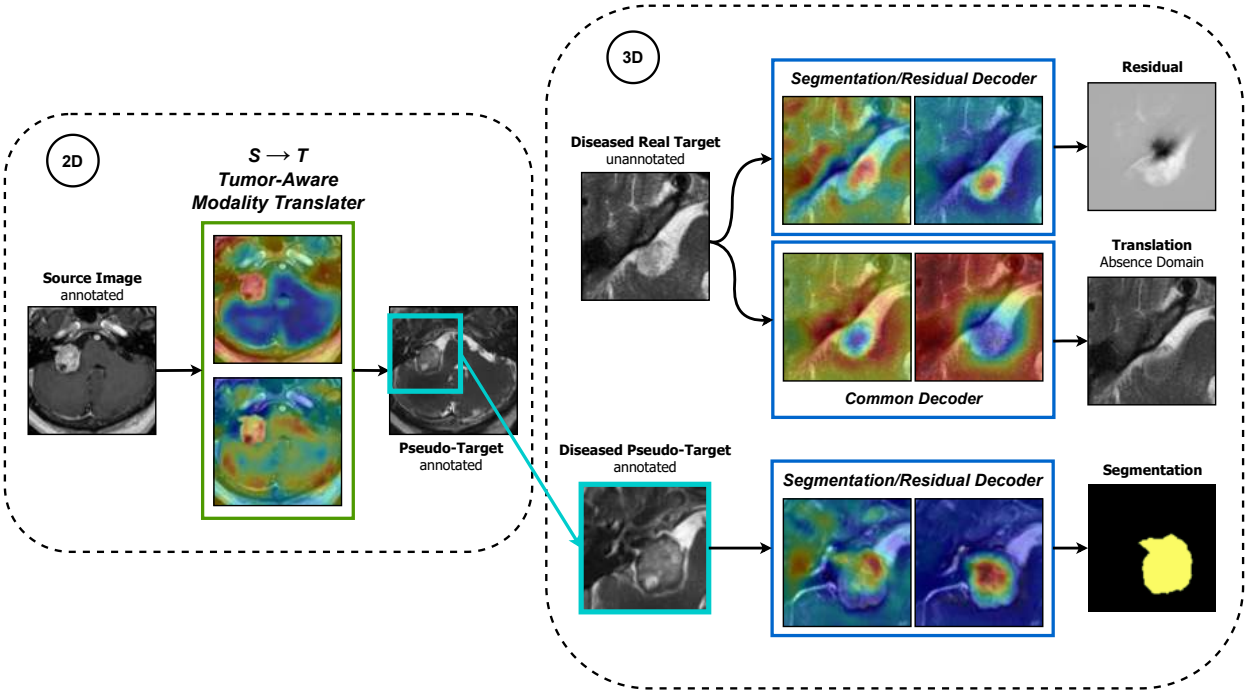


Figure 5.9 Attention maps yielded by the most confident transformer heads in MoDATTS. Red color indicates areas of focus while dark blue corresponds to locations ignored by the network. Note that the maps presented in the modality translation stage are produced by the encoder of the network as the decoder does not contain transformer blocks.

5.5.6 Ablation studies

In the scenario where all the source samples have pixel-level annotations and none are available in the target modality, we conduct the following ablation experiments and report the results in Table 5.5. Specifically, we focus here on the self-supervised variant of MoDATTS (●), as it exhibited the highest segmentation performance in this particular setup. Note that we chose T1 and T2 to be respectively the source and target modalities for the ablations on BraTS.

Self-training We evaluate the performance of MoDATTS before performing iterative self-training (●). We notice an improvement of around +0.02 Dice score in the segmentation performance after the process for VS and Brain tumor segmentation. Qualitative evaluation of the impact of self-training in the domain adaptation task for BraTS and CrossMoDA datasets is also provided in Fig. 5.10. The latter shows that the tumors on the test set are either filled or refined after self-training.

Tumor-aware modality translation To assess the value of additional tumor supervision in the modality translation stage, we retrained the translation model with $\lambda_{seg}^{mod} = 0$. Iterative self-training was not applied in the segmentation stage (●). As expected, the target modality segmentation performance dropped as compared to the previous ablation (BraTS : -0.027 and CrossMoDA : -0.025 in Dice). This implies that the joint tumor supervision in the modality translation stage actually helps to retain detailed lesion structures and provide more accurate pseudo-target images.

The next ablations focus on the semi-supervised variant of MoDATTS (●) and the contribution of the diseased-healthy translation when few annotated source samples are provided. We experiment with 1% of annotated source data, as it is where the semi-supervised variant is the most relevant. Values are reported in Table 5.5, and interpretations are provided below.

Image-level supervision We notice that only training the translation from diseased to healthy domains (P $\rightarrow$ A) suffices (●), as 82% of a target supervised model performance is reached on BraTS. But teaching the model to perform healthy to diseased (A $\rightarrow$ P) yields better performance (BraTS : +0.055 and CrossMoDA : +0.061 in Dice) by making more efficient use of the data. Note that when the whole diseased-healthy unsupervised objective is removed (●), the segmentation performance on the target modality is on par with the one achieved by the self-supervised variant.

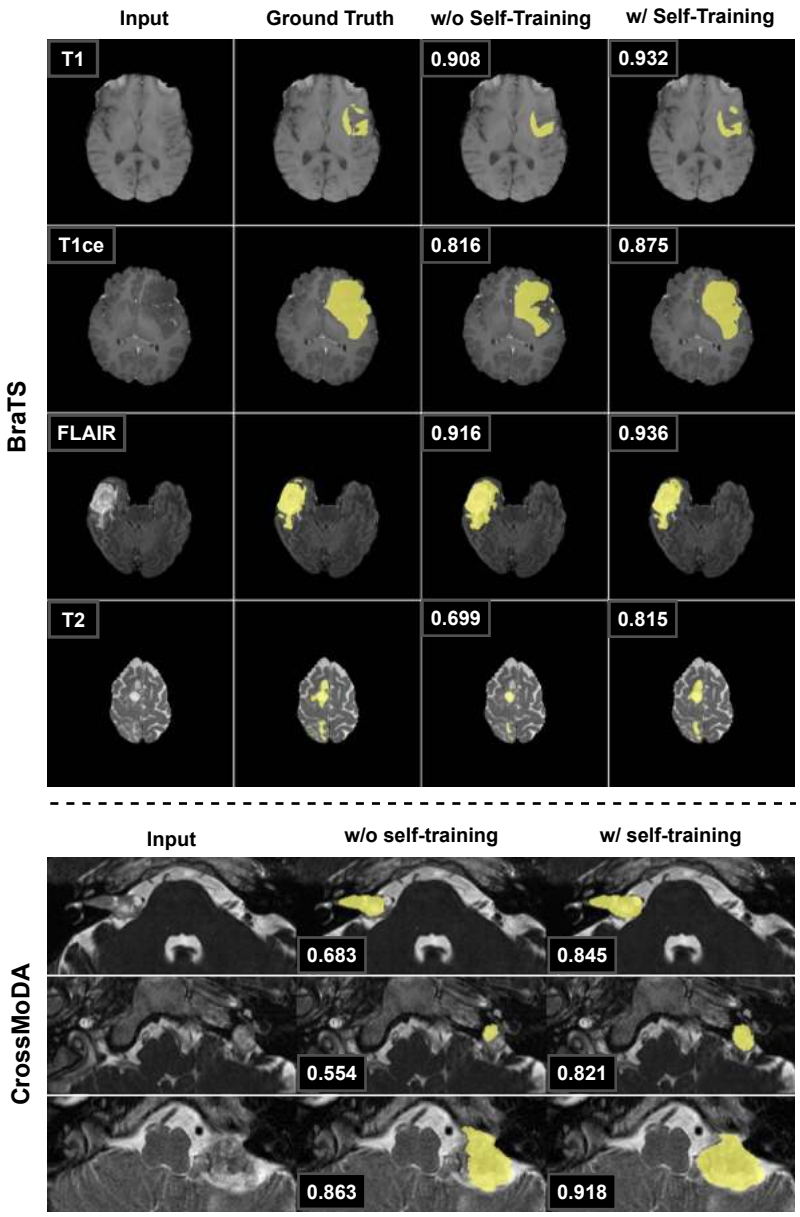


Figure 5.10 Qualitative evaluation of the impact of self-training on the test set for brain tumor and VS segmentation tasks. Last two columns show, respectively, the segmentation map for the same model before and after the self-training iterations. For BraTS, each row illustrates a different scenario where the target modality - indicated in the first column - was not provided with any annotations. We also show ground truth tumor segmentations to visually assess the improvements of the model when self-training is performed, along with the dice score obtained on the whole volume. For CrossMoDA the ground truth VS segmentations on target hrT2 MRIs were not provided. Self-training helps to better leverage unnanotated data in the target modality and acts as a refining of the segmentation maps, which helps to reach better performance.

Separate residual and segmentation decoders Finally, we explored the effect of the decoders by employing a separate segmentation decoder instead of sharing the residual and segmentation weights (●). This separate version shows lower performance on the brain (−0.31 Dice) and VS (−0.20 Dice) datasets. This observation emphasizes that disentangling tumors from the background to perform diseased to healthy translations is similar to a segmentation objective, and therefore is beneficial for accurate cross-modality tumor segmentation when few source samples are annotated.

Source annotations	Ablations	BraTS : T1 → T2	CrossMoDA : T1ce → hrT2
100% (Self-supervised Variant)	● w/o TAMT w/o ST	0.817 ± 0.005 (93%)	0.826 ± 0.100
	● w/o ST	0.844 ± 0.001 (96%)	0.851 ± 0.051
	● **Proposed (Self-sup.)**	$\mathbf{0.863 \pm 0.002}$ (98%)	$\mathbf{0.870 \pm 0.048}$
1% (Semi-supervised Variant)	● w/o P→A and A→P	0.658 ± 0.009 (75%)	0.621 ± 0.269
	● w/o A → P	0.715 ± 0.012 (82%)	0.666 ± 0.222
	● w/o dual-use Res./Seg.	0.739 ± 0.008 (84%)	0.707 ± 0.177
	● **Proposed (Semi-sup.)**	$\mathbf{0.770 \pm 0.005}$ (88%)	$\mathbf{0.727 \pm 0.173}$

Table 5.5 Ablation studies : absolute Dice scores on the target modality. For BraTS, we selected T1 and T2 to be respectively the source and target modalities for these experiments. Ablations of the tumor-aware modality translation (TAMT) and self-training (ST) were performed when 100% of the source data was annotated with the self-supervised variant of MoDATTS, as it yielded the best performance. Alternatively, the ablations related to the diseased-healthy translation were performed on the semi-supervised variant when 1% of the source data was annotated. We also report for BraTS the % of a target supervised model performance that is reached by the model (values indicated in parenthesis). Note that for CrossMoDA the standard deviations reported correspond to the performance variation across the 64 test cases.

5.6 Applications and extensions

We have introduced a domain adaptation method to segment tumors on unnanotated target modality datasets from annotated or partially annotated source modality images. We have demonstrated the competitiveness and robustness of MoDATTS on cross-modality brain tumor and vestibular schwannoma MR sequences.

Self-supervised vs semi-supervised variant The proposed model offers (1) a self-supervised variant that achieves supervision over pseudo-target samples and self-training; and (2) a semi-supervised variant that further includes real target modality images provided with diseased or healthy labels through unsupervised tumor disentanglement. Training the semi-supervised model requires more memory and expensive computation. Due to our limited resources, the semi-supervised model is equipped with fewer parameters. As a consequence, when enough pixel-level annotations are available in the source modality, the self-supervised

model outperforms the semi-supervised model. This implies that a larger and more optimal segmentation network, relying solely on supervision from annotated synthetic pseudo-target data generated by our modality translation network, yields better performance than a smaller model provided with additional weak labels. This observation raises that there is a trade-off between training a bigger model and training a semi-supervised model. However, when annotated source data is scarce (less than 50% of annotated samples), the semi-supervised variant enables to preserve consistent segmentations and outperforms the self-supervised variant, even with a smaller segmentation network. Indeed, in this scenario training the model on actual target images becomes crucial as it has access to fewer synthetic samples and the generated pseudo-target images may be less reliable. Therefore we claim that MoDATTS has the potential to alleviate the annotation burden in cross-modality segmentation tasks.

Although producing the diseased and healthy image-level labels for 3D images does not add substantial annotation cost, it still represents a limitation when compared to unsupervised domain adaptation methods that rely only on pixel-level annotations available in the source modality. We thus specify that the self-supervised variant of MoDATTS does not require these weak labels to be trained, which is an advantage over the semi-supervised variant. In the end, the choice of either variant depends on the proportion of unannotated samples that the concerned dataset holds and the computational resources available.

Limitations As the pathologies we studied in this article (brain tumor and VS) mostly showed lesions on one side of the volumetric images, we were able to yield healthy and diseased samples by splitting the data into hemispheres. In real scenarios, full volumes showing healthy conditions should be collected to train the model. However, even though we experimented outside this setting we believe our results are representative of the actual behavior of the different models.

MoDATTS has demonstrated encouraging performance in the challenging task of cross-modality domain adaptation. However, it remains a heavy 3D method that requires training two distinct models (modality translation and segmentation). This represents high training times and consequent computational resources, specifically for the semi-supervised variant that contains several decoders and discriminators. Furthermore, unlike source-free models that solely rely on a segmentation model trained on the source modality for target adaptation, MoDATTS requires the presence of source images during training. However, it is worth noting that source-free methods under-perform in comparison to generative methods like MoDATTS and sometimes rely on image-level labels incurring substantial annotation costs [163].

Extensions We have tested MoDATTS on CrossMoDA and a modified version of BraTS, which both offer an ideal environment to test any domain adaption strategy for cross-modality segmentation. However they remain limited to segmentation between different MR contrasts. Further work will explore MR to CT adaptation. We even consider evaluating MoDATTS on a cross-pathology and cross-modality task. Specifically, we consider that leveraging annotated gliomas in FLAIR sequences from BraTS to segment intraparenchymal hemorrhages on CT scans [156] is in the range of applications of our model. We believe that the semi-supervised variant of MoDATTS might provide benefits to mitigate the shift between these two conditions.

5.7 Conclusion

MoDATTS is a new 3D transformer-based domain adaptation framework to handle unpaired cross-modality medical image segmentation when target modality lacks annotated samples. We propose a self-supervised variant relying on the supervision from generated pseudo-target images and self-training, bridging the performance gap related to domain shifts in cross-modality tumor segmentation and outperforming other baselines in such scenarios. We offer as well a semi-supervised variant that additionally leverages diseased and healthy weak labels to extend the training to unannotated target images. We show that this annotation-efficient setup helps to maintain consistent performance on the target modality even when source pixel-level annotations are scarce. MoDATTS's ability to achieve 99% and 100% of a target supervised model performance when respectively 20% and 50% of the target data is annotated further emphasizes that our approach can help to mitigate the lack of annotations. The evaluation of MR to CT adaptation tasks will provide further insights into the potential applications of our approach.

Acknowledgments

This research has been funded by the Natural Sciences and Engineering Research Council of Canada (NSERC), and the Canada Research Chair. We thank Compute Canada for providing the essential computational resources to complete this study.

CHAPITRE 6 DISCUSSION GÉNÉRALE

6.1 Synthèse des travaux

Nous avons présenté dans les chapitres précédents les deux modèles développés au cours de ce projet de maîtrise, ayant pour objectif de répondre à la problématique de généralisation entre modalités des réseaux neuronaux pour la segmentation de tumeurs. L'étude de M-GenSeg et MoDATTS a démontré le potentiel des méthodes génératives adversariales pour l'adaptation de domaine entre modalités. Dans ce dernier chapitre, on s'applique à discuter des deux méthodes proposées en mettant en exergue leurs avantages et leurs limitations. On présentera enfin les potentielles améliorations et extensions possibles à notre étude.

6.1.1 M-GenSeg : modèle 2D d'adaptation de domaine

La version 2D de notre modèle (M-GenSeg) est une méthode complexe d'adaptation de domaine. Cette première étude a permis d'explorer :

- L'utilisation d'une approche adversariale permettant la génération d'images à travers les modalités, afin de pallier au manque d'annotations dans une modalité cible.
- L'intégration d'un second objectif semi-supervisé, permettant d'étendre l'entrainement d'un réseau segmenteur à des images de la modalité cible dépourvues d'annotations. En effet, en apprenant au modèle à réaliser des translations entre images saines et images tumorales (en détachant les tumeurs du background), le réseau de segmentation peut profiter des représentations apprises par les décodeurs annexes (*common decoder* et *residual decoder*), ainsi que de la mise à jour de leurs poids.

Combiner ces deux composantes en un seul modèle entrainé de bout-en-bout fût une tâche complexe. Cela a notamment demandé une optimisation réfléchie de la fonction de perte (à multiples composantes) et de comprendre les interactions possibles entre les différents encodeurs et décodeurs du réseau. Par ailleurs, nous avons réussi à combiner efficacement les deux modules de sorte à ce que leurs représentations latentes et la mise à jour de leurs poids soient partagées. Du fait de cet apprentissage multi-tâches, le réseau de translation inter-modalités est capable de préserver les structures tumorales lors de la génération des images synthétiques de modalité cible, ce qui constitue une amélioration par rapport aux méthodes existantes. De plus, en étendant de même l'objectif semi-supervisé aux images de modalité source, la propriété précédente est maintenue même dans un cadre où peu d'anntotations source sont disponibles. Enfin, l'apport de portes d'attention dans les connexions raccour-

cies de notre modèle entre les encodeurs et décodeurs de M-GenSeg a permis d'améliorer légèrement la performance du modèle, mais surtout de visualiser le comportement interne du réseau. Ce gain en interprétabilité est essentiel en apprentissage profond où les réseaux sont souvent considérés comme des "boites noires".

M-GenSeg s'est démarqué de deux méthodes de pointe (AttENT [137] et AccSegNet [141]) par son gain de performance sur toutes les paires de modalités source/cible du jeu de données issu de BraTS développé spécifiquement pour notre problématique. De plus l'approche proposée s'est révélée robuste à un déficit important d'annotations (e.g. 1% de données annotées dans la modalité source et aucune donnée annotée dans la modalité cible). Malgré ces résultats encourageants, la méthode reste limitée par plusieurs aspects (voir section 6.1.2 suivante).

6.1.2 MoDATTS : modèle 3D d'adaptation de domaine

MoDATTS a été développé pour pallier aux défauts de M-GenSeg, à savoir :

- Une performance de segmentation limitée par son utilisation sur des coupes 2D plutôt que des volumes 3D.
- L'attribution des labels sain (absence de tumeurs) et malade (présence de tumeur) représente un coût d'annotation conséquent quand on traite avec des slices 2D.
- Le modèle est gourmand en mémoire et complexe à optimiser, ce qui rend l'utilisation d'architectures plus larges et optimales impossible avec les ressources de calcul dont nous disposions.

Ainsi, MoDATTS conserve les composantes essentielles de M-GenSeg (i.e. la génération d'images tumorales réalistes dans la modalité cible et la composante semi-supervisée) mais en corrigeant les limites liées à une segmentation réalisée sur des coupes 2D. En raison des contraintes de ressources de calcul, il n'était pas faisable de maintenir un modèle 3D complet entrainé de bout-en-bout. Dès lors, nous avons fait le choix de dissocier l'étape de translation de modalité et l'étape de segmentation, faisant de MoDATTS un modèle en 2 étapes (*two-stage method*). Bien que les mises à jour du modèle ne soient alors pas partagées entre les deux tâches, les bénéfices sont multiples : (1) la segmentation peut être réalisée sur des volumes 3D complets; (2) l'optimisation des réseaux est simplifiée du fait de la réduction des interactions entre ses différentes composantes; (3) chaque étape peut désormais profiter d'architectures plus performantes.

Le réseau de translation de modalité de MoDATTS reste un module 2D basé sur le CycleGAN. En y attachant des modules de segmentation nous avons prouvé cependant que les structures tumorales peuvent être maintenues lors de la translation, tout comme dans M-GenSeg. Cela est mis en exergue par la visualisation des modules d'attention intégrés dans la nouvelle

architecture hybride convolutive/transformer utilisée, où on constate que le réseau revalorise positivement les composantes tumorales.

De même, MoDATTS conserve la composante semi-supervisée de M-GenSeg (translation entre images saines et malades) mais le modèle peut désormais utiliser une architecture transformer 3D. Une stratégie itérative d'auto-entrainement a de plus été efficace pour affiner les segmentations générées dans la modalité cible et se rapprocher de la performance d'un modèle entrainé avec des données cibles annotées.

L'étude de MoDATTS a démontré que la performance de son variant "self-supervised" dépasse celle de son variant "semi-supervised" dans des conditions où 50% (ou plus) des annotations de la modalité source sont fournies, alors que ce dernier profite pourtant d'un accès à des labels faibles (images saines vs malades). On attribue ce déficit au fait que le réseau de segmentation du variant "semi-supervised" possède un nombre de paramètres plus limité pour permettre la translation entre les domaines sain et malade. Cela met en exergue que pour des ressources de calcul données, il existe un compromis entre entrainer un modèle semi-supervisé (comme MoDATTS dans cette configuration) et entrainer un modèle de taille plus importante. A moins de disposer de ressources de calcul considérables (e.g. grandes entreprises disposant de clusters de GPUs pour faire du multi-processing) ou d'avoir un déficit important en annotations (moins de 50% dans la modalité source), la version "self-supervised" de MoDATTS reste à favoriser.

En outre MoDATTS a montré un gain de performance considérable par rapport à M-GenSeg sur le jeu de données issu de BraTS, où 95% de la performance d'un modèle entrainé avec des données cibles annotées a été atteint (contre 87% pour M-GenSeg). MoDATTS a aussi obtenu une performance supérieure aux méthodes participantes au data challenge CrossMoDA 2022 pour la segmentation de schwanommes vestibulaires. Cela prouve que notre approche est compétitive sur plusieurs tâches d'adaptation de domaine.

Additionnellement, MoDATTS s'est démarqué par sa robustesse face à un déficit important d'annotations. Nous rappelons en effet que, sur une adaptation T1 $\rightarrow$ T2 (BraTS), 1% des annotation fournies dans la modalité source (T1) et aucune annotation dans la modalité cible (T2) suffisent à atteindre 88% de la performance d'un modèle entrainé avec toutes les annotations de la modalité cible (T2). Ce résultat est impressionant et important du point de vue de l'expérimentation car il montre le potentiel de MoDATTS d'atténuer les baisses de performance liées au déficit d'annotations. En revanche, il faut comprendre qu'un tel scénario n'est pas réaliste en pratique (1% revient à utiliser 3 ou 4 volumes annotés pour l'entrainement dans notre cas). De plus, une baisse de 12% de la performance par rapport à la performance maximale atteignable n'est pas souhaitable en application clinique.

En revanche, de plus amples expériences (menées sur BraTS) ont montré que 99% et 100% de la performance maximale (obtenue par supervision dans la modalité cible) peut être atteinte en annotant 20% et 50% des données cibles (additionnellement à celles fournies dans la modalité source). Cette dernière statistique est centrale à notre étude, car elle montre que l'on peut économiser 80% du coût d'annotation dans la modalité cible en gardant une dégradation de la performance de seulement 1%. Si cette légère baisse de performance n'est pas acceptable (selon les critères définis pour une potentielle application clinique), le coût peut être réduit à 50% sans voir quelconque réduction de performance. Nous pensons que si MoDATTS (ou M-genSeg) a vocation à être utilisé de manière totalement automatisée, c'est dans ce scénario où une fraction des images de la modalité cible est tout de même annotée.

On pourrait malgré cela imaginer utiliser notre approche dans un cadre où aucune annotation n'est disponible dans la modalité cible, mais les segmentations obtenues devraient être impérativement vérifiées et affinées par un agent humain. Une fois ces annotations corrigées elles pourraient permettre d'entrainer un autre réseau de manière supervisée, et ainsi produire un modèle plus fiable. Le gain de temps par rapport à des segmentations obtenues manuellement de bout-en-bout resterait considérable.

6.2 Limitations des solutions proposées

Nous avons étudié des pathologies (tumeurs cérébrales et schwannomes vestibulaires) dont les lésions ne sont souvent présentes que d'un seul côté de l'image. De fait nous avons pu extraire des échantillons sains et malades en séparant les images tumorales en 2 hémisphères. Dans un scénario idéal, des volumes sains ne présentant pas de pathologie devraient être rassemblées pour entrainer M-GenSeg et MoDATTS sur des images entières. Malgré cela, nous pensons que les résultats présentés sont représentatifs des comportements réels des modèles étudiés.

M-GenSeg et MoDATTS ont montré des perfomances encourageantes dans des tâches d'adaptation de domaine. Elles restent cependant des méthodes lourdes demandant d'entrainer 2 réseaux (translation de modalité et segmentation). Cela représente des temps d'entrainement conséquents, et requiert des ressources de calcul importantes. Par exemple M-GenSeg demande 60h d'entrainement pour expérimenter avec une unique paire de modalités sur BraTS avec un GPU de 40 Go. En comparaison l'entrainement de MoDATTS demande 36h pour la même expérience (sans les étapes de self-training). Du fait des coûts financiers importants que représente l'achat de cartes graphiques de mémoire importante, la plupart des ressources en recherche académique atteignent rarement la capacité des GPUs utilisés lors de nos expérimentations. Or les calculs sont déjà réalisés sur des images de dimensions réduites (séparation en hémisphère et découpe autour de la région d'intérêt) et avec des batchs de petite taille. A

moins de réduire la résolution des images utilisées (ce qui s'accompagne souvent d'une perte de performance), les possibilités de réutilisation de MoDATTS ou M-GenSeg dans un cadre de recherche académique sont plutôt limitées.

Il est aussi nécessaire de préciser que M-GenSeg et MoDATTS reposent sur l'accès aux images de la modalité source pour l'adaptation à la modalité cible, ce qui n'est pas toujours offert en pratique (confidentialité des données, etc.). Des modèles dits "*source-free*" pallient à ce problème en reposant uniquement sur un modèle de segmentation pré-entrainé sur la modalité source (mis à disposition à la place des images), et allégent aussi l'entrainement en se débarrassant de l'étape générative de translation entre modalités [161–163]. Néanmoins, bien que présentant les avantages ci-dessus, ces approches donnent lieu à des performances réduites par rapport aux modèles génératifs comme M-GenSeg ou MoDATTS [163].

Notons de même que M-GenSeg et MoDATTS ne permettent qu'une adaptation entre deux modalités différentes, alors que des modèles proposent de réaliser des translations entre de multiples domaines/modalités [186–188]. Il serait envisageable de remplacer nos modules de translation de modalités par de tels approches, mais on ne sait pas si les problèmes de mémoire évoqués ci-dessus seraient un facteur limitant à cette extension du modèle.

6.3 Applications futures

Les jeux de données utilisés pour évaluer M-GenSeg et MoDATTS (BraTS et CrossMoDA) fournissent d'excellents environnements pour tester toute méthode d'adaptation de domaine pour la segmentation tumorale. Cependant on reste limité à des applications de segmentation entre contrastes d'IRM. De fait, il serait souhaitable de tester MoDATTS sur une tâche d'adaptation entre modalités IRM et CT, ce qui sera réalisé dans des travaux ultérieurs. Peu de jeux de données CT/IRM existent cependant pour la segmentation de tumeurs. On s'intéressera donc plutôt à une tâche d'adaptation entre modalités et pathologies. Nous pensons qu'exploiter les annotations de gliomes (tumeurs cérébrales) dans les séquences FLAIR de BraTS pour segmenter des hémorragies intraparenchymateuses sur des scanners CT est dans le champ d'application de notre modèle [156].

CHAPITRE 7 CONCLUSION

L'imagerie médicale joue un rôle crucial dans la détection de ces tumeurs, et la segmentation de ces lésions sur des clichés médicaux est une étape cruciale dans le processus de diagnostic et de traitement. Les réseaux de neurones profonds sont un moyen efficace de réduire les coûts liés à une segmentation réalisée manuellement par des experts, ainsi que d'en améliorer l'objectivité. Leur succès repose cependant sur la disponibilité d'annotations, coûteuses et pénibles à produire. De plus ces modèles généralisent mal entre modalités d'imagerie. L'objectif de ce projet aura donc été de proposer une méthode d'adaptation de domaine permettant d'exploiter des annotations disponibles dans une modalité (source) pour ne pas réitérer le travail d'annotations dans une autre modalité (cible).

A cet effet, nous avons exploré des modèles d'apprentissage profond génératifs permettant la synthétisation d'images tumorales dans la modalité cible depuis la modalité source. Cela a permis de réutiliser les annotations disponibles dans cette dernière pour entrainer un réseau de segmentation capable de détecter des tumeurs dans la modalité cible. Nous avons aussi évalué l'apport d'un module semi-supervisé dans le réseau de segmentation, permettant d'ajouter à l'entrainement des images connues comme étant saines.

M-GenSeg, le premier modèle développé (2D) s'est démarqué face à deux approches de l'état de l'art sur une tâche de segmentation de gliomes entre plusieurs contrastes d'IRM. De part sa composante semi-supervisée, M-GenSeg s'est aussi montré robuste face à un déficit important d'annotations.

MoDATTS, le second modèle développé dans ce projet, permet de remédier aux défauts de M-GenSeg en étendant son application à des volumes 3D. Cette nouvelle approche utilise aussi des architectures puissantes basées sur des transformers, et intégre additionnellement une stratégie itérative d'auto-entrainement. MoDATTS présente un gain important en performance par rapport à M-GenSeg sur notre jeu de segmentation issu de BraTS, et dépasse le score de Dice des participants du data challenge CrossMoDA 2022 sur une tâche de segmentation de schwannomes vestibulaires. Par ailleurs, plusieurs expériences ont montré que MoDATTS peut efficacement réduire la dépendance des réseaux à la disponibilité de données annotées.

L'évaluation de MoDATTS sur une tâche de segmentation entre modalités IRM et CT fournira de nouvelles perspectives sur les applications potentielles de notre approche.

RÉFÉRENCES

[1] J. P. Ridgway, "Cardiovascular magnetic resonance physics for clinicians: part I," *Journal of Cardiovascular Magnetic Resonance*, vol. 12, nº. 1, p. 71, 2010. [En ligne]. Disponible: https://jcmr-online.biomedcentral.com/articles/10.1186/1532-429X-12-71

[2] X. Cheng *et al.*, "Memory-Efficient Cascade 3D U-Net for Brain Tumor Segmentation," dans *Brainlesion: Glioma, Multiple Sclerosis, Stroke and Traumatic Brain Injuries*. Springer International Publishing, 2020, vol. 11992, p. 242–253. [En ligne]. Disponible: http://link.springer.com/10.1007/978-3-030-46640-4_23

[3] L. W. Goldman, "Principles of CT and CT Technology," *Journal of Nuclear Medicine Technology*, vol. 35, nº. 3, p. 115–128, 2007. [En ligne]. Disponible: http://tech.snmjournals.org/cgi/doi/10.2967/jnmt.107.042978

[4] N. C. Purandare et V. Rangarajan, "Imaging of lung cancer: Implications on staging and management," *Indian Journal of Radiology and Imaging*, vol. 25, nº. 02, p. 109–120, 2015. [En ligne]. Disponible: http://www.thieme-connect.de/DOI/DOI?10.4103/0971-3026.155831

[5] E. S. Lee, "Imaging diagnosis of pancreatic cancer: A state-of-the-art review," *World Journal of Gastroenterology*, vol. 20, nº. 24, p. 7864, 2014. [En ligne]. Disponible: http://www.wjgnet.com/1007-9327/full/v20/i24/7864.htm

[6] P. Bilic, P. Christ et H. B. Li, "The Liver Tumor Segmentation Benchmark (LiTS)," *Medical Image Analysis*, vol. 84, p. 102680, 2023. [En ligne]. Disponible: https://www.sciencedirect.com/science/article/pii/S1361841522003085

[7] N. Heller *et al.*, "The state of the art in kidney and kidney tumor segmentation in contrast-enhanced CT imaging: Results of the KiTS19 challenge," *Medical Image Analysis*, vol. 67, p. 101821, 2021. [En ligne]. Disponible: https://linkinghub.elsevier.com/retrieve/pii/S1361841520301857

[8] N. I. of Biomedical Imaging et Bioengineering, "Nuclear medicine," July. 2016. [En ligne]. Disponible: https://www.nibib.nih.gov/science-education/science-topics/nuclear-medicine

[9] H.-H. Chang et D. J. Valentino, "An electrostatic deformable model for medical image segmentation," *Computerized Medical Imaging and Graphics*, vol. 32, nº. 1, p. 22–35, 2008. [En ligne]. Disponible: https://linkinghub.elsevier.com/retrieve/pii/S0895611107001322

[10] J. Zhang *et al.*, "A fully automatic extraction of magnetic resonance image features in glioblastoma patients," *Medical Physics*, vol. 41, nº. 4, p. 042301, 2014. [En ligne]. Disponible: https://onlinelibrary.wiley.com/doi/10.1118/1.4866218

[11] "Réseaux de neurones et réseaux convolutifs," dans *Cours Cnam US330X*, 2018. [En ligne]. Disponible: http://cedric.cnam.fr/~thomen/cours/US330X/tpbackprop.html

[12] Nvidia, "How does a convolutional neural network work?" dans *Convolutional Neural Networks*, 2019. [En ligne]. Disponible: https://www.nvidia.com/en-us/glossary/data-science/convolutional-neural-network/

[13] H. Yingge, I. Ali et K.-Y. Lee, "Deep Neural Networks on Chip - A Survey," dans *2020 IEEE International Conference on Big Data and Smart Computing (BigComp)*. Busan, Korea (South): IEEE, 2020, p. 589–592. [En ligne]. Disponible: https://ieeexplore.ieee.org/document/9070243/

[14] "Cnn architectures: Lenet, alexnet, vgg, googlenet, resnet and more," dans *coderz.py*, 2020. [En ligne]. Disponible: https://coderzpy.com/cnn-architectures-lenet-alexnet-vgg-googlenet-resnet-and-more/

[15] C. Chen *et al.*, "Deep Learning for Cardiac Image Segmentation: A Review," *Frontiers in Cardiovascular Medicine*, vol. 7, p. 25, 2020. [En ligne]. Disponible: https://www.frontiersin.org/article/10.3389/fcvm.2020.00025/full

[16] 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*. Cham: Springer International Publishing, 2015, vol. 9351, p. 234–241. [En ligne]. Disponible: http://link.springer.com/10.1007/978-3-319-24574-4_28

[17] G. Huang *et al.*, "Densely Connected Convolutional Networks," dans *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. Honolulu, HI: IEEE, 2017, p. 2261–2269. [En ligne]. Disponible: https://ieeexplore.ieee.org/document/8099726/

[18] "Attention is all you need: Discovering the transformer paper," dans *Towards Data Science*, 2020. [En ligne]. Disponible: https://towardsdatascience.com/attention-is-all-you-need-discovering-the-transformer-paper-73e5ff5e0634

[19] A. Vaswani *et al.*, "Attention is all you need," *arXiv*, 2017. [En ligne]. Disponible: http://arxiv.org/abs/1706.03762

[20] A. Dosovitskiy *et al.*, "An image is worth 16x16 words: Transformers for image recognition at scale," *arXiv*, 2020. [En ligne]. Disponible: https://arxiv.org/abs/2010.11929

[21] J. Chen *et al.*, "Transunet: Transformers make strong encoders for medical image segmentation," *arXiv*, 2021. [En ligne]. Disponible: https://arxiv.org/abs/2102.04306

[22] X. Yi, E. Walia et P. Babyn, "Generative adversarial network in medical imaging: A review," *Medical Image Analysis*, vol. 58, p. 101552, 2019. [En ligne]. Disponible: https://linkinghub.elsevier.com/retrieve/pii/S1361841518308430

[23] Y. Wu et K. He, "Group normalization," dans *Proceedings of the European Conference on Computer Vision (ECCV)*, September 2018.

[24] Y. Zhou *et al.*, "Multi-task learning for segmentation and classification of tumors in 3D automated breast ultrasound images," *Medical Image Analysis*, vol. 70, p. 101918, 2021. [En ligne]. Disponible: https://linkinghub.elsevier.com/retrieve/pii/S1361841520302826

[25] B. H. Menze *et al.*, "The Multimodal Brain Tumor Image Segmentation Benchmark (BRATS)," *IEEE Trans. Med. Imaging*, vol. 34, nº. 10, p. 1993–2024, 2015.

[26] S. M. Havaei, "Machine learning methods for brain tumor segmentation," thèse de doctorat, Dép. de génie informatique, Université de Sherbroooke, 2017.

[27] Y.-D. Xiao *et al.*, "MRI contrast agents: Classification and application (Review)," *International Journal of Molecular Medicine*, vol. 38, nº. 5, p. 1319–1326, 2016. [En ligne]. Disponible: https://www.spandidos-publications.com/10.3892/ijmm.2016.2744

[28] V. Kuperman, *Magnetic resonance imaging: physical principles and applications*, ser. Electromagnetism. San Diego, USA: Academic Press, 2000.

[29] C. J. Garvey, "Computed tomography in clinical practice," *BMJ*, vol. 324, nº. 7345, p. 1077–1080, 2002. [En ligne]. Disponible: https://www.bmj.com/lookup/doi/10.1136/bmj.324.7345.1077

[30] A. C. Kak et M. Slaney, *Principles of computerized tomographic imaging*, ser. Classics in applied mathematics. Society for Industrial and Applied Mathematics, 2001, nº. 33.

[31] P. M. Boiselle, "Computed Tomography Screening for Lung Cancer," *JAMA*, vol. 309, nº. 11, p. 1163, 2013. [En ligne]. Disponible: http://jama.jamanetwork.com/article.aspx?doi=10.1001/jama.2012.216988

[32] L. R. Roberts *et al.*, "Imaging for the diagnosis of hepatocellular carcinoma: A systematic review and meta-analysis," *Hepatology*, vol. 67, nº. 1, p. 401–421, 2018. [En ligne]. Disponible: https://journals.lww.com/01515467-201801000-00034

[33] R. Fonti, M. Conson et S. Del Vecchio, "PET/CT in radiation oncology," *Seminars in Oncology*, vol. 46, nº. 3, p. 202–209, 2019. [En ligne]. Disponible: https://linkinghub.elsevier.com/retrieve/pii/S009377541930079X

[34] M. Ahmed *et al.*, "Image-guided Tumor Ablation: Standardization of Terminology and Reporting Criteria—A 10-Year Update," *Radiology*, vol. 273, n°. 1, p. 241–260, 2014. [En ligne]. Disponible: http://pubs.rsna.org/doi/10.1148/radiol.14132958

[35] S. N. Goldberg *et al.*, "Image-guided Tumor Ablation: Proposal for Standardization of Terms and Reporting Criteria," *Radiology*, vol. 228, n°. 2, p. 335–345, 2003. [En ligne]. Disponible: http://pubs.rsna.org/doi/10.1148/radiol.2282021787

[36] C. R. Hansen *et al.*, "Plan quality in radiotherapy treatment planning – Review of the factors and challenges," *Journal of Medical Imaging and Radiation Oncology*, vol. 66, n°. 2, p. 267–278, 2022. [En ligne]. Disponible: https://onlinelibrary.wiley.com/doi/10.1111/1754-9485.13374

[37] G. Delaney *et al.*, "The role of radiotherapy in cancer treatment: Estimating optimal utilization from a review of evidence-based clinical guidelines," *Cancer*, vol. 104, n°. 6, p. 1129–1137, 2005. [En ligne]. Disponible: https://onlinelibrary.wiley.com/doi/10.1002/cncr.21324

[38] S. Stieb *et al.*, "Imaging for Response Assessment in Radiation Oncology," *Hematology/Oncology Clinics of North America*, vol. 34, n°. 1, p. 293–306, 2020. [En ligne]. Disponible: https://linkinghub.elsevier.com/retrieve/pii/S088985881930125X

[39] E. C. Covert *et al.*, "Intra- and inter-operator variability in MRI-based manual segmentation of HCC lesions and its impact on dosimetry," *EJNMMI Physics*, vol. 9, n°. 1, p. 90, 2022. [En ligne]. Disponible: https://ejnmmiphys.springeropen.com/articles/10.1186/s40658-022-00515-6

[40] G. Corrao *et al.*, "Intra- and inter-observer variability in breast tumour bed contouring and the controversial role of surgical clips," *Medical Oncology*, vol. 36, n°. 6, p. 51, 2019. [En ligne]. Disponible: http://link.springer.com/10.1007/s12032-019-1273-1

[41] J. Dinkel *et al.*, "Inter-observer reproducibility of semi-automatic tumor diameter measurement and volumetric analysis in patients with lung cancer," *Lung Cancer*, vol. 82, n°. 1, p. 76–82, 2013. [En ligne]. Disponible: https://linkinghub.elsevier.com/retrieve/pii/S0169500213003152

[42] C. Schumann *et al.*, "State of the Art in Computer-Assisted Planning, Intervention, and Assessment of Liver-Tumor Ablation," *Critical Reviews™ in Biomedical Engineering*, vol. 38, n°. 1, p. 31–52, 2010. [En ligne]. Disponible: http://www.dl.begellhouse.com/journals/4b27cbfc562e21b8,41996af914259394,70ee48b85b5e4f16.html

[43] V. Yeghiazaryan et I. Voiculescu, "Family of boundary overlap metrics for the evaluation of medical image segmentation," *Journal of Medical Imaging*, vol. 5, n°. 01, p. 1, 2018. [En ligne]. Disponible:

https://www.spiedigitallibrary.org/journals/journal-of-medical-imaging/volume-5/issue-01/015006/Family-of-boundary-overlap-metrics-for-the-evaluation-of-medical/10.1117/1.JMI.5.1.015006.full

[44] P. Campadelli, E. Casiraghi et S. Pratissoli, "A segmentation framework for abdominal organs from CT scans," *Artificial Intelligence in Medicine*, vol. 50, n°. 1, p. 3–11, 2010. [En ligne]. Disponible: https://linkinghub.elsevier.com/retrieve/pii/S0933365710000539

[45] M. G. Linguraru *et al.*, "Statistical 4D graphs for multi-organ abdominal segmentation from multiphase CT," *Medical Image Analysis*, vol. 16, n°. 4, p. 904–914, 2012. [En ligne]. Disponible: https://linkinghub.elsevier.com/retrieve/pii/S1361841512000291

[46] ——, "Renal tumor quantification and classification in contrast-enhanced abdominal CT," *Pattern Recognition*, vol. 42, n°. 6, p. 1149–1161, 2009. [En ligne]. Disponible: https://linkinghub.elsevier.com/retrieve/pii/S0031320308003750

[47] Y. Liu *et al.*, "An Effective Approach of Lesion Segmentation Within the Breast Ultrasound Image Based on the Cellular Automata Principle," *Journal of Digital Imaging*, vol. 25, n°. 5, p. 580–590, oct. 2012. [En ligne]. Disponible: http://link.springer.com/10.1007/s10278-011-9450-6

[48] T. Heimann *et al.*, "Comparison and Evaluation of Methods for Liver Segmentation From CT Datasets," *IEEE Transactions on Medical Imaging*, vol. 28, n°. 8, p. 1251–1265, 2009. [En ligne]. Disponible: http://ieeexplore.ieee.org/document/4781564/

[49] Xinjian Chen *et al.*, "Medical Image Segmentation by Combining Graph Cuts and Oriented Active Appearance Models," *IEEE Transactions on Image Processing*, vol. 21, n°. 4, p. 2035–2046, 2012. [En ligne]. Disponible: http://ieeexplore.ieee.org/document/6144010/

[50] L. Ruskó, G. Bekes et M. Fidrich, "Automatic segmentation of the liver from multi- and single-phase contrast-enhanced CT images," *Medical Image Analysis*, vol. 13, n°. 6, p. 871–882, 2009. [En ligne]. Disponible: https://linkinghub.elsevier.com/retrieve/pii/S1361841509000644

[51] Y. Chen *et al.*, "The domain knowledge based graph-cut model for liver CT segmentation," *Biomedical Signal Processing and Control*, vol. 7, n°. 6, p. 591–598, 2012. [En ligne]. Disponible: https://linkinghub.elsevier.com/retrieve/pii/S1746809412000493

[52] M. Kass, A. Witkin et D. Terzopoulos, "Snakes: Active contour models," *International Journal of Computer Vision*, vol. 1, n°. 4, p. 321–331, 1988. [En ligne]. Disponible: http://link.springer.com/10.1007/BF00133570

[53] 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, nov. 1988. [En ligne]. Disponible: https://linkinghub.elsevier.com/retrieve/pii/0021999188900022

[54] L. Suhuai, L. R et O. Sebastien, “A new deformable model using dynamic gradient vector flow and adaptive balloon forces,” 2003. [En ligne]. Disponible: https://staff.itee.uq.edu.au/lovell/aprs/wdic2003/CDROM/9.pdf

[55] Tao Wang, I. Cheng et A. Basu, “Fluid Vector Flow and Applications in Brain Tumor Segmentation,” *IEEE Transactions on Biomedical Engineering*, vol. 56, n°. 3, p. 781–789, 2009. [En ligne]. Disponible: http://ieeexplore.ieee.org/document/4760239/

[56] A. Law, F. Lam et F. Chan, “A fast deformable region model for brain tumor boundary extraction,” dans *Proceedings of the Second Joint 24th Annual Conference and the Annual Fall Meeting of the Biomedical Engineering Society] [Engineering in Medicine and Biology*, vol. 2. Houston, TX, USA: IEEE, 2002, p. 1055–1056. [En ligne]. Disponible: http://ieeexplore.ieee.org/document/1106273/

[57] D. J. Valentino, “Image segmentation using a charged fluid method,” *Journal of Electronic Imaging*, vol. 15, n°. 2, p. 023011, 2006. [En ligne]. Disponible: http://electronicimaging.spiedigitallibrary.org/article.aspx?doi=10.1117/1.2199555

[58] A. Yezzi *et al.*, “A geometric snake model for segmentation of medical imagery,” *IEEE Transactions on Medical Imaging*, vol. 16, n°. 2, p. 199–209, 1997. [En ligne]. Disponible: http://ieeexplore.ieee.org/document/563665/

[59] A. E. Lefohn, J. E. Cates et R. T. Whitaker, “Interactive, GPU-Based Level Sets for 3D Segmentation,” dans *Medical Image Computing and Computer-Assisted Intervention - MICCAI 2003*. Berlin, Heidelberg: Springer Berlin Heidelberg, 2003, vol. 2878. [En ligne]. Disponible: http://link.springer.com/10.1007/978-3-540-39899-8_70

[60] K. Xie *et al.*, “Semi-automated brain tumor and edema segmentation using MRI,” *European Journal of Radiology*, vol. 56, n°. 1, p. 12–19, 2005. [En ligne]. Disponible: https://linkinghub.elsevier.com/retrieve/pii/S0720048X05001518

[61] H. Khotanlou *et al.*, “3D brain tumor segmentation in MRI using fuzzy classification, symmetry analysis and spatially constrained deformable models,” *Fuzzy Sets and Systems*, vol. 160, n°. 10, p. 1457–1473, 2009. [En ligne]. Disponible: https://linkinghub.elsevier.com/retrieve/pii/S0165011408005368

[62] S. Ho, E. Bullitt et G. Gerig, “Level-set evolution with region competition: automatic 3-D segmentation of brain tumors,” dans *Object recognition supported by user interaction*

for service robots, vol. 1. Quebec City, Que., Canada: IEEE Comput. Soc, 2002, p. 532–535. [En ligne]. Disponible: http://ieeexplore.ieee.org/document/1044788/

[63] M. Prastawa *et al.*, "Robust Estimation for Brain Tumor Segmentation," dans *Medical Image Computing and Computer-Assisted Intervention - MICCAI 2003*. Berlin, Heidelberg: Springer Berlin Heidelberg, 2003, vol. 2879, p. 530–537. [En ligne]. Disponible: http://link.springer.com/10.1007/978-3-540-39903-2_65

[64] M. G. Linguraru *et al.*, "Tumor Burden Analysis on Computed Tomography by Automated Liver and Tumor Segmentation," *IEEE Transactions on Medical Imaging*, vol. 31, nº. 10, p. 1965–1976, 2012. [En ligne]. Disponible: http://ieeexplore.ieee.org/document/6262482/

[65] N. Archip, F. A. Jolesz et S. K. Warfield, "A Validation Framework for Brain Tumor Segmentation," *Academic Radiology*, vol. 14, nº. 10, p. 1242–1251, 2007. [En ligne]. Disponible: https://linkinghub.elsevier.com/retrieve/pii/S1076633207004072

[66] A. Capelle *et al.*, "Unsupervised segmentation for automatic detection of brain tumors in MRI," dans *Proceedings 2000 International Conference on Image Processing (Cat. No.00CH37101)*, vol. 1. Vancouver, BC, Canada: IEEE, 2000, p. 613–616. [En ligne]. Disponible: http://ieeexplore.ieee.org/document/901033/

[67] M. Schmidt, "Automatic brain tumor segmentation," thèse de doctorat, University of Aberta, 2005.

[68] S. Vinitski *et al.*, "Fast tissue segmentation based on a 4D feature map in characterization of intracranial lesions," *Journal of Magnetic Resonance Imaging*, vol. 9, nº. 6, p. 768–776, 1999. [En ligne]. Disponible: https://onlinelibrary.wiley.com/doi/10.1002/(SICI)1522-2586(199906)9:6<768::AID-JMRI3>3.0.CO;2-2

[69] L. Santos *et al.*, "Medical Image Segmentation Using Seeded Fuzzy C-means: A Semi-supervised Clustering Algorithm," dans *2018 International Joint Conference on Neural Networks (IJCNN)*. IEEE, 2018, p. 1–7. [En ligne]. Disponible: https://ieeexplore.ieee.org/document/8489401/

[70] M. Vaidyanathan *et al.*, "Comparison of supervised MRI segmentation methods for tumor volume determination during therapy," *Magnetic Resonance Imaging*, vol. 13, nº. 5, p. 719–728, 1995. [En ligne]. Disponible: https://linkinghub.elsevier.com/retrieve/pii/0730725X95000126

[71] J. Zhang *et al.*, "Tumor segmentation from magnetic resonance imaging by learning via one-class support vector machine," 2004.

[72] M. Havaei, P.-M. Jodoin et H. Larochelle, "Efficient Interactive Brain Tumor Segmentation as Within-Brain kNN Classification," dans *2014 22nd International*

Conference on Pattern Recognition. IEEE, 2014, p. 556–561. [En ligne]. Disponible: http://ieeexplore.ieee.org/document/6976816/

[73] M. Kaus *et al.*, "Adaptive Template Moderated Brain Tumor Segmentation in MRI," dans *Bildverarbeitung für die Medizin 1999*. Springer Berlin Heidelberg, 1999, p. 102–106. [En ligne]. Disponible: http://link.springer.com/10.1007/978-3-642-60125-5_19

[74] S. Pereira *et al.*, "Automatic brain tissue segmentation of multi-sequence mr images using random decision forests," *The MIDAS Journal*, 2013. [En ligne]. Disponible: https://www.semanticscholar.org/paper/Automatic-Brain-Tissue-Segmentation-of-MR-Images-Pereira-Festa/ffc40cb5c34de3977324d96adfe9ef03cefa1549

[75] "Understanding random forest," dans *Towards Data Science*, June 2019. [En ligne]. Disponible: https://towardsdatascience.com/understanding-random-forest-58381e0602d2

[76] N. K. Subbanna *et al.*, "Hierarchical Probabilistic Gabor and MRF Segmentation of Brain Tumours in MRI Volumes," dans *Advanced Information Systems Engineering*. Berlin, Heidelberg: Springer Berlin Heidelberg, 2013, vol. 7908, p. 751–758, series Title: Lecture Notes in Computer Science. [En ligne]. Disponible: http://link.springer.com/10.1007/978-3-642-40811-3_94

[77] M. Schmidt *et al.*, "Segmenting Brain Tumors using Alignment-Based Features," dans *Fourth International Conference on Machine Learning and Applications (ICMLA'05)*. Los Angeles, CA, USA: IEEE, 2005, p. 215–220. [En ligne]. Disponible: http://ieeexplore.ieee.org/document/1607453/

[78] A. W. Simonetti *et al.*, "Combination of feature-reduced MR spectroscopic and MR imaging data for improved brain tumor classification," *NMR in Biomedicine*, vol. 18, n°. 1, p. 34–43, 2005. [En ligne]. Disponible: https://onlinelibrary.wiley.com/doi/10.1002/nbm.919

[79] J. Luts *et al.*, "A combined MRI and MRSI based multiclass system for brain tumour recognition using LS-SVMs with class probabilities and feature selection," *Artificial Intelligence in Medicine*, vol. 40, n°. 2, p. 87–102, 2007. [En ligne]. Disponible: https://linkinghub.elsevier.com/retrieve/pii/S0933365707000188

[80] Y. LeCun, Y. Bengio et G. Hinton, "Deep learning," *Nature*, vol. 521, n°. 7553, p. 436–444, 2015. [En ligne]. Disponible: http://www.nature.com/articles/nature14539

[81] F. Murtagh, "Multilayer perceptrons for classification and regression," *Neurocomputing*, vol. 2, n°. 5-6, p. 183–197, 1991. [En ligne]. Disponible: https://linkinghub.elsevier.com/retrieve/pii/0925231291900235

[82] "Optimization: Stochastic gradient descent," dans *Stanford Course CS231n : Convolutional Neural Networks for Visual Recognition*, 2023. [En ligne]. Disponible: https://cs231n.github.io/optimization-1/

[83] S. M. Anwar *et al.*, "Medical Image Analysis using Convolutional Neural Networks: A Review," *Journal of Medical Systems*, vol. 42, n°. 11, p. 226, 2018. [En ligne]. Disponible: http://link.springer.com/10.1007/s10916-018-1088-1

[84] Y. Chang *et al.*, "Classification of parotid gland tumors by using multimodal MRI and deep learning," *NMR in Biomedicine*, vol. 34, n°. 1, janv. 2021. [En ligne]. Disponible: https://onlinelibrary.wiley.com/doi/10.1002/nbm.4408

[85] J. W. Yoo *et al.*, "Deep learning diagnostics for bladder tumor identification and grade prediction using RGB method," *Scientific Reports*, vol. 12, n°. 1, p. 17699, 2022. [En ligne]. Disponible: https://www.nature.com/articles/s41598-022-22797-7

[86] L. Alzubaidi *et al.*, "Review of deep learning: concepts, CNN architectures, challenges, applications, future directions," *Journal of Big Data*, vol. 8, n°. 1, p. 53, 2021. [En ligne]. Disponible: https://journalofbigdata.springeropen.com/articles/10.1186/s40537-021-00444-8

[87] "Parallelization with multiprocessing in python," dans *Towards Data Science*, 2022. [En ligne]. Disponible: https://towardsdatascience.com/parallelization-w-multiprocessing-in-python-bd2fc234f516

[88] G. E. Hinton et R. R. Salakhutdinov, "Reducing the Dimensionality of Data with Neural Networks," *Science*, vol. 313, n°. 5786, p. 504–507, 2006. [En ligne]. Disponible: https://www.science.org/doi/10.1126/science.1127647

[89] J. Long, E. Shelhamer et T. Darrell, "Fully convolutional networks for semantic segmentation," dans *2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. Boston, MA, USA: IEEE, 2015, p. 3431–3440. [En ligne]. Disponible: http://ieeexplore.ieee.org/document/7298965/

[90] T. Wood, "Softmax function," dans *DeepAI*. [En ligne]. Disponible: https://deepai.org/machine-learning-glossary-and-terms/softmax-layer

[91] S. Jadon, "A survey of loss functions for semantic segmentation," dans *2020 IEEE Conference on Computational Intelligence in Bioinformatics and Computational Biology (CIBCB)*. Via del Mar, Chile: IEEE, 2020, p. 1–7. [En ligne]. Disponible: https://ieeexplore.ieee.org/document/9277638/

[92] C. H. Sudre *et al.*, "Generalised Dice Overlap as a Deep Learning Loss Function for Highly Unbalanced Segmentations," dans *Deep Learning in Medical Image Analysis*

and Multimodal Learning for Clinical Decision Support, M. J. Cardoso *et al.*, édit. Springer International Publishing, 2017, vol. 10553, p. 240–248. [En ligne]. Disponible: http://link.springer.com/10.1007/978-3-319-67558-9_28

[93] S. S. M. Salehi, D. Erdogmus et A. Gholipour, "Tversky Loss Function for Image Segmentation Using 3D Fully Convolutional Deep Networks," dans *Machine Learning in Medical Imaging*, Q. Wang *et al.*, édit. Springer International Publishing, 2017, vol. 10541, p. 379–387. [En ligne]. Disponible: http://link.springer.com/10.1007/978-3-319-67389-9_44

[94] N. Abraham et N. M. Khan, "A Novel Focal Tversky Loss Function With Improved Attention U-Net for Lesion Segmentation," dans *2019 IEEE 16th International Symposium on Biomedical Imaging (ISBI 2019)*. Venice, Italy: IEEE, 2019, p. 683–687. [En ligne]. Disponible: https://ieeexplore.ieee.org/document/8759329/

[95] M. Yeung *et al.*, "A mixed focal loss function for handling class imbalanced medical image segmentation," *ArXiv*, 2021.

[96] S. A. Taghanaki *et al.*, "Combo loss: Handling input and output imbalance in multi-organ segmentation," *Computerized Medical Imaging and Graphics*, vol. 75, p. 24–33, 2019. [En ligne]. Disponible: https://linkinghub.elsevier.com/retrieve/pii/S0895611118305688

[97] "Backpropagation in neural networks," dans *Towards Data Science*, 2021. [En ligne]. Disponible: https://towardsdatascience.com/backpropagation-in-neural-networks-6561e1268da8

[98] "The vanishing gradient problem," dans *Towards Data Science*, 2019. [En ligne]. Disponible: https://towardsdatascience.com/the-vanishing-gradient-problem-69bf08b15484

[99] K. He *et al.*, "Deep Residual Learning for Image Recognition," dans *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. Las Vegas, NV, USA: IEEE, 2016, p. 770–778. [En ligne]. Disponible: http://ieeexplore.ieee.org/document/7780459/

[100] R. Girshick *et al.*, "Rich Feature Hierarchies for Accurate Object Detection and Semantic Segmentation," dans *2014 IEEE Conference on Computer Vision and Pattern Recognition*. Columbus, OH, USA: IEEE, 2014, p. 580–587. [En ligne]. Disponible: http://ieeexplore.ieee.org/document/6909475/

[101] R. Girshick, "Fast R-CNN," dans *2015 IEEE International Conference on Computer Vision (ICCV)*. Santiago, Chile: IEEE, 2015, p. 1440–1448. [En ligne]. Disponible: http://ieeexplore.ieee.org/document/7410526/

[102] K. He *et al.*, “Mask R-CNN,” dans *2017 IEEE International Conference on Computer Vision (ICCV)*. Venice: IEEE, 2017, p. 2980–2988. [En ligne]. Disponible: http://ieeexplore.ieee.org/document/8237584/

[103] O. Oktay *et al.*, “Attention U-Net: Learning Where to Look for the Pancreas,” *ArXiv*, 2018. [En ligne]. Disponible: https://arxiv.org/abs/1804.03999

[104] A. Sherstinsky, “Fundamentals of recurrent neural network (RNN) and long short-term memory (LSTM) network,” *arXiv*, 2018. [En ligne]. Disponible: http://arxiv.org/abs/1808.03314

[105] L. Tanzi *et al.*, “Vision Transformer for femur fracture classification,” *Injury*, vol. 53, n°. 7, p. 2625–2634, 2022. [En ligne]. Disponible: https://linkinghub.elsevier.com/retrieve/pii/S0020138322002868

[106] C. Xin *et al.*, “An improved transformer network for skin cancer classification,” *Computers in Biology and Medicine*, vol. 149, p. 105939, 2022. [En ligne]. Disponible: https://linkinghub.elsevier.com/retrieve/pii/S0010482522006746

[107] Y. Liu *et al.*, “3D Deep Attentive U-Net with Transformer for Breast Tumor Segmentation from Automated Breast Volume Scanner,” dans *2021 43rd Annual International Conference of the IEEE Engineering in Medicine & Biology Society (EMBC)*. Mexico: IEEE, 2021, p. 4011–4014. [En ligne]. Disponible: https://ieeexplore.ieee.org/document/9629523/

[108] Z. Yang et S. Li, “Dual-path Network for Liver and Tumor Segmentation in CT ImagesUsing Swin Transformer Encoding Approach,” *Current Medical Imaging Formerly Current Medical Imaging Reviews*, vol. 19, n°. 10, p. 1114–1123, 2023. [En ligne]. Disponible: https://www.eurekaselect.com/210033/article

[109] Y. Gao *et al.*, “A data-scalable transformer for medical image segmentation: Architecture, model efficiency, and benchmark,” 2023. [En ligne]. Disponible: https://arxiv.org/abs/2203.00131

[110] I. Goodfellow *et al.*, “Generative Adversarial Nets,” dans *Advances in Neural Information Processing Systems*, vol. 27. Curran Associates, Inc., 2014. [En ligne]. Disponible: https://proceedings.neurips.cc/paper_files/paper/2014/file/5ca3e9b122f61f8f06494c97b1afccf3-Paper.pdf

[111] A. Bindal, “Normalization techniques in deep neural networks,” dans *Medium*, February 2019. [En ligne]. Disponible: https://medium.com/techspace-usict/normalization-techniques-in-deep-neural-networks-9121bf100d8

[112] F. Garcea *et al.*, "Data augmentation for medical imaging: A systematic literature review," *Computers in Biology and Medicine*, vol. 152, p. 106391, janv. 2023. [En ligne]. Disponible: https://linkinghub.elsevier.com/retrieve/pii/S001048252201099X

[113] P. Chlap *et al.*, "A review of medical image data augmentation techniques for deep learning applications," *Journal of Medical Imaging and Radiation Oncology*, vol. 65, n°. 5, p. 545–563, août 2021. [En ligne]. Disponible: https://onlinelibrary.wiley.com/doi/10.1111/1754-9485.13261

[114] S. Ruder, "An overview of multi-task learning in deep neural networks," *ArXiv*, 2017.

[115] H. Guan et M. Liu, "Domain Adaptation for Medical Image Analysis: A Survey," *IEEE Transactions on Biomedical Engineering*, vol. 69, n°. 3, p. 1173–1185, mars 2022. [En ligne]. Disponible: https://ieeexplore.ieee.org/document/9557808/

[116] M. Harsh, "Understanding domain adaptation," dans *Towards Data Science*, May 2021. [En ligne]. Disponible: https://towardsdatascience.com/understanding-domain-adaptation-5baa723ac71f

[117] Q. Dou *et al.*, "Pnp-adanet: Plug-and-play adversarial domain adaptation network with a benchmark at cross-modality cardiac segmentation," *ArXiv*, 2018. [En ligne]. Disponible: https://arxiv.org/abs/1812.07907

[118] F. Wu et X. Zhuang, "Cf distance: A new domain discrepancy metric and application to explicit domain adaptation for cross-modality cardiac image segmentation," *IEEE Transactions on Medical Imaging*, vol. 39, n°. 12, p. 4274–4285, 2020. [En ligne]. Disponible: https://ieeexplore.ieee.org/abstract/document/9165963

[119] ——, "Unsupervised Domain Adaptation With Variational Approximation for Cardiac Segmentation," *IEEE Transactions on Medical Imaging*, vol. 40, n°. 12, p. 3555–3567, 2021. [En ligne]. Disponible: https://ieeexplore.ieee.org/document/9459711/

[120] F. Garcea *et al.*, "Data augmentation for medical imaging: A systematic literature review," *Computers in Biology and Medicine*, vol. 152, p. 106391, janv. 2023. [En ligne]. Disponible: https://linkinghub.elsevier.com/retrieve/pii/S001048252201099X

[121] Y. Pang *et al.*, "Image-to-Image Translation: Methods and Applications," *IEEE Transactions on Multimedia*, vol. 24, p. 3859–3881, 2022. [En ligne]. Disponible: https://ieeexplore.ieee.org/document/9528943/

[122] J.-Y. Zhu *et al.*, "Unpaired Image-to-Image Translation Using Cycle-Consistent Adversarial Networks," dans *2017 IEEE International Conference on Computer Vision (ICCV)*. IEEE, oct. 2017, p. 2242–2251. [En ligne]. Disponible: http://ieeexplore.ieee.org/document/8237506/

[123] J. Yang *et al.*, “Unsupervised Domain Adaptation via Disentangled Representations: Application to Cross-Modality Liver Segmentation,” dans *Medical Image Computing and Computer Assisted Intervention – MICCAI 2019*. Cham: Springer International Publishing, 2019, vol. 11765, p. 255–263. [En ligne]. Disponible: https://link.springer.com/10.1007/978-3-030-32245-8_29

[124] C. Yang *et al.*, “Mutual-Prototype Adaptation for Cross-Domain Polyp Segmentation,” *IEEE Journal of Biomedical and Health Informatics*, vol. 25, p. 3886–3897, oct. 2021. [En ligne]. Disponible: https://ieeexplore.ieee.org/document/9423517/

[125] J. Yang *et al.*, “Domain-Agnostic Learning With Anatomy-Consistent Embedding for Cross-Modality Liver Segmentation,” dans *2019 IEEE/CVF International Conference on Computer Vision Workshop (ICCVW)*. Seoul, Korea (South): IEEE, oct. 2019, p. 323–331. [En ligne]. Disponible: https://ieeexplore.ieee.org/document/9022229/

[126] C. Pei *et al.*, “Disentangle domain features for cross-modality cardiac image segmentation,” *Medical Image Analysis*, vol. 71, p. 102078, juill. 2021. [En ligne]. Disponible: https://linkinghub.elsevier.com/retrieve/pii/S1361841521001249

[127] E. Vorontsov *et al.*, “Towards annotation-efficient segmentation via image-to-image translation,” *Medical Image Analysis*, vol. 82, p. 102624, 2022. [En ligne]. Disponible: https://linkinghub.elsevier.com/retrieve/pii/S1361841522002523

[128] S. Bakas *et al.*, “Advancing The Cancer Genome Atlas glioma MRI collections with expert segmentation labels and radiomic features,” *Sci Data*, vol. 4, nº. 1, p. 170117, 2017. [En ligne]. Disponible: https://www.nature.com/articles/sdata2017117

[129] ——, “Identifying the Best Machine Learning Algorithms for Brain Tumor Segmentation, Progression Assessment, and Overall Survival Prediction in the BRATS Challenge,” 2018. [En ligne]. Disponible: https://arxiv.org/abs/1811.02629

[130] F. Isensee *et al.*, “nnU-Net: a self-configuring method for deep learning-based biomedical image segmentation,” *Nature Methods*, vol. 18, nº. 2, p. 203–211, 2021. [En ligne]. Disponible: https://www.nature.com/articles/s41592-020-01008-z

[131] A. Hatamizadeh *et al.*, “Swin UNETR: Swin Transformers for Semantic Segmentation of Brain Tumors in MRI Images,” dans *Brainlesion: Glioma, Multiple Sclerosis, Stroke and Traumatic Brain Injuries*, A. Crimi et S. Bakas, édit. Cham: Springer International Publishing, 2022, vol. 12962, p. 272–284. [En ligne]. Disponible: https://link.springer.com/10.1007/978-3-031-08999-2_22

[132] H.-Y. Zhou *et al.*, “nnformer: Interleaved transformer for volumetric segmentation,” 2022.

[133] J. Shapey *et al.*, "Segmentation of vestibular schwannoma from MRI, an open annotated dataset and baseline algorithm," *Scientific Data*, vol. 8, n°. 1, p. 286, 2021. [En ligne]. Disponible: https://www.nature.com/articles/s41597-021-01064-w

[134] R. Dorent *et al.*, "CrossMoDA 2021 challenge: Benchmark of cross-modality domain adaptation techniques for vestibular schwannoma and cochlea segmentation," *Medical Image Analysis*, vol. 83, p. 102628, 2023. [En ligne]. Disponible: https://www.sciencedirect.com/science/article/pii/S1361841522002560

[135] L. M. Prevedello *et al.*, "Challenges related to artificial intelligence research in medical imaging and the importance of image analysis competitions," *Radiology: Artificial Intelligence*, vol. 1, n°. 1, p. e180031, 2019. [En ligne]. Disponible: https://pubs.rsna.org/doi/full/10.1148/ryai.2019180031

[136] B. Billot *et al.*, "SynthSeg: Segmentation of brain MRI scans of any contrast and resolution without retraining," *Medical Image Analysis*, vol. 86, p. 102789, 2023. [En ligne]. Disponible: https://www.sciencedirect.com/science/article/pii/S1361841523000506

[137] C. Li *et al.*, "Attent: Domain-adaptive medical image segmentation via attention-aware translation and adversarial entropy minimization," dans *2021 IEEE International Conference on Bioinformatics and Biomedicine (BIBM)*. IEEE, 2021, p. 952–959. [En ligne]. Disponible: https://ieeexplore.ieee.org/document/9669620

[138] Y. Huo *et al.*, "Synseg-net: Synthetic segmentation without target modality ground truth," *IEEE Transactions on Medical Imaging*, vol. 38, n°. 4, p. 1016–1025, 2019. [En ligne]. Disponible: https://ieeexplore.ieee.org/document/8494797/

[139] C. Chen *et al.*, "Synergistic image and feature adaptation: Towards cross-modality domain adaptation for medical image segmentation," vol. 38, 2019, p. 865–872. [En ligne]. Disponible: https://aaai.org/papers/00865-synergistic-image-and-feature-adaptation-towards-cross-modality-domain-adaptation-for-med

[140] J. Jiang *et al.*, "Self-derived organ attention for unpaired CT-MRI deep domain adaptation based MRI segmentation," *Phys. Med. Biol.*, vol. 65, n°. 20, p. 205001, 2020. [En ligne]. Disponible: https://iopscience.iop.org/article/10.1088/1361-6560/ab9fca

[141] B. Zhou, C. Liu et J. S. Duncan, "Anatomy-Constrained Contrastive Learning for Synthetic Segmentation Without Ground-Truth," dans *Medical Image Computing and Computer Assisted Intervention*. Springer, 2021, vol. 12901, p. 47–56. [En ligne]. Disponible: https://link.springer.com/chapter/10.1007/978-3-030-87193-2_5

[142] J. Hoffman *et al.*, "CyCADA: Cycle-Consistent Adversarial Domain Adaptation," 2017. [En ligne]. Disponible: https://arxiv.org/abs/1711.03213

[143] Y. Zhang *et al.*, "Task Driven Generative Modeling for Unsupervised Domain Adaptation: Application to X-ray Image Segmentation," dans *Medical Image Computing and Computer Assisted Intervention – MICCAI 2018*. Springer International Publishing, 2018, vol. 11071, p. 599–607. [En ligne]. Disponible: https://link.springer.com/chapter/10.1007/978-3-030-00934-2_67

[144] X. Liu *et al.*, "Low-Rank Atlas Image Analyses in the Presence of Pathologies," *IEEE Trans. Med. Imaging*, vol. 34, nº. 12, p. 2583–2591, 2015. [En ligne]. Disponible: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC4707015/

[145] C. Lin *et al.*, "Low-Rank Based Image Analyses for Pathological MR Image Segmentation and Recovery," *Front. Neurosci.*, vol. 13, nº. 13, p. 333, 2019. [En ligne]. Disponible: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC6465608/

[146] S. Changfa *et al.*, "Multi-slice low-rank tensor decomposition based multi-atlas segmentation: Application to automatic pathological liver ct segmentation," *Medical Image Analysis*, vol. 73, p. 102152, 2021. [En ligne]. Disponible: https://www.sciencedirect.com/science/article/pii/S1361841521001985

[147] H.-C. Shin *et al.*, "Medical image synthesis for data augmentation and anonymization using generative adversarial networks," dans *Simulation and Synthesis in Medical Imaging*. Cham: Springer International Publishing, 2018, p. 1–11. [En ligne]. Disponible: https://link.springer.com/chapter/10.1007/978-3-030-00536-8_1

[148] T. C. W. Mok et A. C. S. Chung, "Learning data augmentation for brain tumor segmentation with coarse-to-fine generative adversarial networks," dans *Brainlesion: Glioma, Multiple Sclerosis, Stroke and Traumatic Brain Injuries*. Cham: Springer International Publishing, 2019, p. 70–80. [En ligne]. Disponible: https://link.springer.com/chapter/10.1007/978-3-030-11723-8_7

[149] S. Kim, B. Kim et H. Park, "Synthesis of brain tumor multicontrast mr images for improved data augmentation," *Medical Physics*, vol. 48, nº. 5, p. 2185–2198, 2021. [En ligne]. Disponible: https://aapm.onlinelibrary.wiley.com/doi/10.1002/mp.14701

[150] M. Drozdzal *et al.*, "The Importance of Skip Connections in Biomedical Image Segmentation," dans *Deep Learning and Data Labeling for Medical Applications*. Springer, 2016, vol. 10008, p. 179–187. [En ligne]. Disponible: https://link.springer.com/chapter/10.1007/978-3-319-46976-8_19

[151] W. Yuan *et al.*, "Unified Attentional Generative Adversarial Network for Brain Tumor Segmentation from Multimodal Unpaired Images," dans *Medical Image Computing and Computer Assisted Intervention*. Springer, 2019, vol. 11766, p. 229–237. [En ligne]. Disponible: https://link.springer.com/chapter/10.1007/978-3-030-32248-9_26

[152] X. Chen *et al.*, “Recent advances and clinical applications of deep learning in medical image analysis,” *Medical Image Analysis*, vol. 79, p. 102444, 2022. [En ligne]. Disponible: https://www.sciencedirect.com/science/article/pii/S1361841522000913

[153] S. Minaee *et al.*, “Image Segmentation Using Deep Learning: A Survey,” *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2021. [En ligne]. Disponible: https://ieeexplore.ieee.org/document/9356353

[154] A. Torralba et A. A. Efros, “Unbiased look at dataset bias,” dans *CVPR 2011*. Colorado Springs, CO, USA: IEEE, 2011, p. 1521–1528. [En ligne]. Disponible: https://ieeexplore.ieee.org/document/5995347

[155] L. Dang, N. C.-y. Tu et E. Y. Chan, “Current imaging tools for vestibular schwannoma,” *Current Opinion in Otolaryngology & Head & Neck Surgery*, vol. 28, nº. 5, p. 302–307, 2020. [En ligne]. Disponible: https://journals.lww.com/co-otolaryngology/Abstract/2020/10000/Current_imaging_tools_for_vestibular_schwannoma.8.aspx

[156] D. Dong *et al.*, “An unsupervised domain adaptation brain CT segmentation method across image modalities and diseases,” *Expert Systems with Applications*, vol. 207, p. 118016, 2022. [En ligne]. Disponible: https://www.sciencedirect.com/science/article/pii/S0957417422012337

[157] M. Wang et W. Deng, “Deep visual domain adaptation: A survey,” *Neurocomputing*, vol. 312, p. 135–153, 2018. [En ligne]. Disponible: https://www.sciencedirect.com/science/article/pii/S0925231218306684

[158] H. Hoyez *et al.*, “Unsupervised Image-to-Image Translation: A Review,” *Sensors*, vol. 22, p. 8540, 2022. [En ligne]. Disponible: https://www.mdpi.com/1424-8220/22/21/8540

[159] E. Panfilov *et al.*, “Improving Robustness of Deep Learning Based Knee MRI Segmentation: Mixup and Adversarial Domain Adaptation,” dans *2019 IEEE/CVF International Conference on Computer Vision Workshop (ICCVW)*. Seoul, Korea (South): IEEE, 2019, p. 450–459. [En ligne]. Disponible: https://ieeexplore.ieee.org/document/9022164

[160] R. Shen *et al.*, “Unsupervised domain adaptation with adversarial learning for mass detection in mammogram,” *Neurocomputing*, vol. 393, p. 27–37, 2020. [En ligne]. Disponible: https://www.sciencedirect.com/science/article/pii/S0925231220301570

[161] C. Yang *et al.*, “Source free domain adaptation for medical image segmentation with fourier style mining,” *Medical Image Analysis*, vol. 79, p. 102457, juill. 2022. [En ligne]. Disponible: https://www.sciencedirect.com/science/article/pii/S1361841522001049

[162] X. Liu *et al.*, "Adapting Off-the-Shelf Source Segmenter for Target Medical Image Segmentation," dans *Medical Image Computing and Computer Assisted Intervention – MICCAI 2021*. Cham: Springer International Publishing, 2021, vol. 12902, p. 549–559. [En ligne]. Disponible: https://link.springer.com/chapter/10.1007/978-3-030-87196-3_51

[163] M. Bateson *et al.*, "Source-free domain adaptation for image segmentation," *Medical Image Analysis*, vol. 82, p. 102617, nov. 2022. [En ligne]. Disponible: https://www.sciencedirect.com/science/article/pii/S1361841522002456

[164] L. A. Gatys, A. S. Ecker et M. Bethge, "Image Style Transfer Using Convolutional Neural Networks," dans *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. Las Vegas, NV, USA: IEEE, juin 2016, p. 2414–2423. [En ligne]. Disponible: https://ieeexplore.ieee.org/document/7780634

[165] C. Ouyang *et al.*, "Data Efficient Unsupervised Domain Adaptation For Cross-modality Image Segmentation," dans *Medical Image Computing and Computer Assisted Intervention – MICCAI 2019*. Cham: Springer International Publishing, 2019, vol. 11765, p. 669–677, series Title: Lecture Notes in Computer Science. [En ligne]. Disponible: https://link.springer.com/chapter/10.1007/978-3-030-32245-8_74

[166] Q. Xie *et al.*, "Self-Training With Noisy Student Improves ImageNet Classification," dans *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. Seattle, WA, USA: IEEE, juin 2020, p. 10 684–10 695. [En ligne]. Disponible: https://ieeexplore.ieee.org/document/9156610

[167] Y. Zou *et al.*, "Pseudoseg: Designing pseudo labels for semantic segmentation," dans *International Conference on Learning Representations*, 2021. [En ligne]. Disponible: https://openreview.net/forum?id=-TwO99rbVRu

[168] Y. Zhu *et al.*, "Improving Semantic Segmentation via Efficient Self-Training," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, p. 1–1, 2021. [En ligne]. Disponible: https://ieeexplore.ieee.org/document/9663011

[169] Y. Zou *et al.*, "Unsupervised Domain Adaptation for Semantic Segmentation via Class-Balanced Self-training," dans *Computer Vision – ECCV 2018*, V. Ferrari *et al.*, édit. Cham: Springer International Publishing, 2018, vol. 11207, p. 297–313, series Title: Lecture Notes in Computer Science. [En ligne]. Disponible: https://link.springer.com/chapter/10.1007/978-3-030-01219-9_18

[170] A. Kumar, T. Ma et P. Liang, "Understanding self-training for gradual domain adaptation," dans *Proceedings of the 37th International Conference on Machine*

Learning, ser. Proceedings of Machine Learning Research, vol. 119. PMLR, 2020, p. 5468–5479. [En ligne]. Disponible: https://proceedings.mlr.press/v119/kumar20c.html

[171] H. Liu, J. Wang et M. Long, "Cycle Self-Training for Domain Adaptation," dans *Advances in Neural Information Processing Systems*, vol. 34. Curran Associates, Inc., 2021, p. 22 968–22 981. [En ligne]. Disponible: https://proceedings.neurips.cc/paper/2021/hash/c1fea270c48e8079d8ddf7d06d26ab52-Abstract.html

[172] F. Yu *et al.*, "DAST: Unsupervised Domain Adaptation in Semantic Segmentation Based on Discriminator Attention and Self-Training," *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 35, nº. 12, p. 10 754–10 762, 2021. [En ligne]. Disponible: https://ojs.aaai.org/index.php/AAAI/article/view/17285

[173] H. Shin *et al.*, "Self-Training Based Unsupervised Cross-Modality Domain Adaptation for Vestibular Schwannoma and Cochlea Segmentation," 2021. [En ligne]. Disponible: https://arxiv.org/abs/2109.10674

[174] B. Kang *et al.*, "Multi-view cross-modality mr image translation for vestibular schwannoma and cochlea segmentation," 2023. [En ligne]. Disponible: https://arxiv.org/abs/2303.14998

[175] G. Sallé *et al.*, "Cross-modal tumor segmentation using generative blending augmentation and self training," 2023. [En ligne]. Disponible: https://arxiv.org/abs/2304.01705

[176] J. P. Cohen, M. Luck et S. Honari, "Distribution Matching Losses Can Hallucinate Features in Medical Image Translation," dans *Medical Image Computing and Computer Assisted Intervention – MICCAI 2018*. Cham: Springer International Publishing, 2018, vol. 11070, p. 529–536, series Title: Lecture Notes in Computer Science. [En ligne]. Disponible: http://link.springer.com/10.1007/978-3-030-00928-1_60

[177] H. Shin *et al.*, "Cosmos: Cross-modality unsupervised domain adaptation for 3d medical image segmentation based on target-aware domain translation and iterative self-training," 2022. [En ligne]. Disponible: https://arxiv.org/abs/2203.16557

[178] B. B. Avants *et al.*, "Advanced normalization tools (ants)," *Insight j*, vol. 2, nº. 365, p. 1–35, 2009.

[179] S. R. Dubey et S. K. Singh, "Transformer-based generative adversarial networks in computer vision: A comprehensive survey," 2023. [En ligne]. Disponible: https://arxiv.org/abs/2302.08641

[180] J. Deng *et al.*, "Imagenet: A large-scale hierarchical image database," dans *2009 IEEE Conference on Computer Vision and Pattern Recognition*, 2009, p. 248–255. [En ligne]. Disponible: https://ieeexplore.ieee.org/document/5206848

[181] T.-C. Wang *et al.*, “High-resolution image synthesis and semantic manipulation with conditional gans,” dans *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2018, p. 8798–8807. [En ligne]. Disponible: https://ieeexplore.ieee.org/document/8579015

[182] H. Dong *et al.*, “Unsupervised domain adaptation in semantic segmentation based on pixel alignment and self-training,” 2021. [En ligne]. Disponible: https://arxiv.org/abs/2109.14219

[183] L. Han *et al.*, “Unsupervised cross-modality domain adaptation for vestibular schwannoma segmentation and koos grade prediction based on semi-supervised contrastive learning,” 2022. [En ligne]. Disponible: https://arxiv.org/abs/2210.04255

[184] T. Shaham, T. Dekel et T. Michaeli, “Singan: Learning a generative model from a single natural image,” dans *2019 IEEE/CVF International Conference on Computer Vision (ICCV)*. Los Alamitos, CA, USA: IEEE Computer Society, 2019, p. 4569–4579. [En ligne]. Disponible: https://doi.ieeecomputersociety.org/10.1109/ICCV.2019.00467

[185] E. Voita *et al.*, “Analyzing multi-head self-attention: Specialized heads do the heavy lifting, the rest can be pruned,” dans *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*. Florence, Italy: Association for Computational Linguistics, 2019, p. 5797–5808. [En ligne]. Disponible: https://aclanthology.org/P19-1580

[186] Y. Choi *et al.*, “Stargan: Unified generative adversarial networks for multi-domain image-to-image translation,” dans *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2018, p. 8789–8797. [En ligne]. Disponible: https://ieeexplore.ieee.org/document/8579014

[187] H. Yang *et al.*, “A unified hyper-gan model for unpaired multi-contrast mr image translation,” dans *Medical Image Computing and Computer Assisted Intervention – MICCAI 2021*, M. de Bruijne *et al.*, édit. Cham: Springer International Publishing, 2021, p. 127–137. [En ligne]. Disponible: https://link.springer.com/chapter/10.1007/978-3-030-87199-4_12

[188] W. Yuan *et al.*, “Unified generative adversarial networks for multimodal segmentation from unpaired 3d medical images,” *Medical Image Analysis*, vol. 64, p. 101731, 2020. [En ligne]. Disponible: https://www.sciencedirect.com/science/article/pii/S1361841520300955