# Apprentissage génératif à bas régime des données pour la segmentation d'images en oncologie

Guillaume Sallé

COLLEGE DOCTORAL BRETAGNE
SCIENCES DE LA VIE ET DE LA SANTE

Université de Bretagne Occidentale

# THÈSE DE DOCTORAT DE

L'UNIVERSITÉ DE BRETAGNE OCCIDENTALE

ÉCOLE DOCTORALE N° 637
*Sciences de la Vie et de la Santé*
Spécialité : *Analyse et Traitement de l'Information et des Images Médicales*

Par

**Guillaume SALLÉ**

## Apprentissage génératif à bas régime de données pour la segmentation d'images en oncologie

**Thèse présentée et soutenue à Brest, le 12/06/24**
**Unité de recherche : LaTIM, INSERM, UMR 1101**

### Rapporteurs avant soutenance :

| | |
|---|---|
| Benjamin LEMASSON | Chargé de recherche, U1216 Inserm Grenoble Institut des Neurosciences |
| Clovis TAUBER | Maître de conférences, Université de Tours, U1253 Inserm Imagerie & Cerveau |

### Composition du Jury :

| | | |
|---|---|---|
| Président : | Douraied BEN SALEM | Professeur des Universités Praticien Hospitalier, CHRU Brest, U1101 Inserm LaTIM |
| Examinateurs : | Douraied BEN SALEM | Professeur des Universités Praticien Hospitalier, CHRU Brest, U1101 Inserm LaTIM |
| | Nicolas BOUSSION | Radiophysicien, CHRU Brest, U1101 Inserm LaTIM |
| | Carole LARTIZIEN | Directrice de recherche CNRS, CREATIS |
| | Benjamin LEMASSON | Chargé de recherche, U1216 Inserm Grenoble Institut des Neurosciences |
| | Clovis TAUBER | Maître de conférences, Université de Tours, U1253 Inserm Imagerie & Cerveau |
| | Dimitris VISVIKIS | Directeur de recherche INSERM, U1101 Inserm LaTIM |
| Dir. de thèse : | Nicolas BOUSSION | Radiophysicien, CHRU Brest, U1101 Inserm LaTIM |

### Invité(s) :

| | | |
|---|---|---|
| Encadrant de thèse : | Vincent JAOUEN | Maître de conférences, IMT Atlantique, U1101 Inserm LaTIM |

# Remerciements

Dans ce long chemin non sans périples, de nombreuses personnes m'ont soutenu. Je pense tout d'abord à ma famille, notamment ma mère, mon père, ma sœur et tous mes autres proches. Je ressens votre fierté, et c'est pour moi la plus grande des réussites. Votre soutien et votre bienveillance sont sans faille et à jamais dans mon cœur. Merci également à Mathilde pour le début des travaux.

Ensuite, je me dois de remercier mes amis proches et alliés pendant ces 3 années, en particulier Adélie le sang, Lancine, Morgane, Joachim, Salomé, Flora et Eloi. La clique de la TP m'a également tiré vers le haut pendant de (trop) nombreuses heures, en me proposant un sac à dos des plus confortables, mais aussi en me confortant dans tous les moments. Je remercie ma famille allemande, dont Noah, Vera, Joscha, Kim et toute l'équipe, ils sont tous d'une gentillesse et bienveillance sans pareil. Je passe également une dédicace spéciale à CLAAP, nos échecs successifs m'ont façonné et je suis enfin fier de ce que je suis devenu, en partie grâce à vous. Je rajoute une pensée particulière à mon ami Julien, lui comme moi sommes destinés à impacter notre monde autant que nous le pourrons, et ce n'est que le début. J'en profite pour remercier toutes les personnes avec qui j'ai pu refaire le monde pendant quelques minutes ou de nombreuses heures, que nous soyons toujours en contact ou non. Toutes ces rencontres m'ont fait évoluer et ont affiné ma compréhension du monde, ce qui est et reste mon objectif premier. Je remercie donc Népal pour ses travaux et ses enseignements d'une richesse inouïe. Il m'a conseillé Siddhartha, à la fois au meilleur et pire moment, en plus d'avoir développé une plume tranchante et juste. L'inspiration qu'il a procuré est presque infinie et je découvre encore de nouveaux niveaux de lecture de ses textes avec les années.

J'adresse maintenant mes derniers remerciements à mes collègues, en particulier mon binôme de thèse préféré Yi-Heng, mais aussi Baptiste, Gustavo, Aziliz, Wistan, Isabelle, Amine, Joel et sa connaissance des bois de Plougastel, Thore et Sacha. Merci à Mathieu Lestum pour nos discussions constructives concernant l'urgence climatique et les moyens d'action. Merci à Pierre-Henri Conze, qui m'a fait découvrir la vision par ordinateur, domaine qui m'a passionné. En plus de sa présence dans la quasi-totalité des projets du LaTIM, il est un enseignant patient et compréhensif. Enfin, je remercie mon encadrant de

thèse, Vincent Jaouen, pour son encadrement d'une grande excellence académique, ainsi que pour toutes nos discussions scientifiques particulièrement riches et pertinentes. Même si nous nous sommes parfois (souvent) donnés du fil à retordre mutuellement, nous avons réussi à proposer des travaux et des résultats dont nous pouvons être fiers. Merci encore pour ton aide et ton temps.

# Résumé

En apprentissage statistique, la performance des modèles est influencée par divers biais inhérents aux données, comme la rareté des données (*data scarcity*) et le décalage de domaine (*domain shift*). Cette thèse s'intéresse à réduire leur influence dans le cadre de la segmentation de structures pathologiques en imagerie médicale. En particulier, notre objectif est de réduire les écarts de données au niveau des régions d'intérêt (ROI) entre les domaines source d'entraînement et cible de déploiement, qu'ils soient intrinsèques aux données ou induits par la faible quantité de données disponibles. Dans ce but, nous proposons une stratégie d'augmentation de données adaptative, basée sur l'analyse de la distribution des intensités des ROI dans le domaine de déploiement. Une première contribution, que nous qualifions d'augmentation naïve, consiste à altérer l'apparence des ROI du domaine d'entraînement pour mieux correspondre aux caractéristiques des ROI du domaine de déploiement. Une seconde étape, complétant la première, rend l'altération plus réaliste par rapport aux propriétés du domaine cible grâce à un modèle génératif d'harmonisation *one-shot*, applicable dans toutes les situations de disponibilité de données. De cette façon, nous renforçons la robustesse du modèle de segmentation en aval pour les ROI dont les caractéristiques étaient initialement sous-représentées à l'entraînement. Nous évaluons notre approche à différents régimes de données et divers contextes cliniques, notamment en IRM, TDM et radiographie pulmonaire. En outre, notre approche a montré des résultats impressionnants lors d'un challenge de segmentation de tumeurs à la conférence MICCAI 2022.

# Abstract

In statistical learning, the performance of models is affected by various biases present within the data, including data scarcity and domain shift. This thesis focuses on reducing their impact in the field of pathological structure segmentation in medical imaging. Our goal is to minimize data discrepancies at the region of interest (ROI) level between the training source domain and the target deployment domain, whether it is intrinsic to the data or caused by the limited data availability. To this end, we present an adaptive data augmentation strategy, based on the analysis of the intensity distribution of the ROIs in the deployment domain. A first contribution, which we call naive augmentation, consists of altering the appearance of the training ROIs to better match the characteristics of the ROIs in the deployment domain. A second augmentation, complementing the first, makes the alteration more realistic relative to the properties of the target domain by harmonizing the characteristics of the altered image. For this, we employ a generative model trained on a single unlabeled image from the deployment domain (one-shot approach), making this technique usable in any data regime encountered. In this way, we enhance the robustness of the downstream segmentation model for ROIs whose characteristics were initially underrepresented in the deployment domain. The effectiveness of this method is evaluated under various data regimes and in different clinical contexts (MRI, CT, CXR). Our approach demonstrated impressive results in a tumor segmentation challenge at MICCAI 2022.

# TABLE DES MATIÈRES

# Liste des figures et tableaux

# Liste des sigles et acronymes

**CBCT** Tomodensitométrie à faisceau conique

**DG** Généralisation de domaines *(domain generalization)*

**DL** Apprentissage profond *(deep learning)*

**GAN** Réseaux antagonistes génératifs *(generative adversarial networks)*. Composé d'un générateur et d'un discriminateur, ils permettent la synthèse d'images réalistes. Leur fonctionnement est détaillé au chapitre 2.

**GBA** Augmentation par mélange génératif *(generative blending augmentation)*, proposée au chapitre 3

**GPU** Carte graphique *(graphics processing unit)*

**GTV** Volume tumoral macroscopique *(gross target volume)*

**I2I** image-à-image, souvent utilisé dans le contexte "synthèse image-à-image", qui siginifie générer une image à partir d'une autre, permettant la synthèse d'un domaine vers un autre.

**KDE** Estimation par noyau *(kernel density estimation)*

**ROI** Régions d'intérêt *(regions of interest)*

**TDM** Scanner, tomodensitométrie, ou encore tomographie assistée par ordinateur

**TL** Apprentissage par transfert *(transfer learning)*

**UDA** Adaptation de domaine non-supervisée *(unsupervised domain adaptation)*. On cherche à apprendre une tâche sur un domaine cible non-annoté à partir d'un domaine source annoté.

**VAE** Auto-encodeur variationnel *(variational auto-encoder)*. Modèle caractérisé par un encodeur, un espace latent de caractéristiques, ainsi qu'un décodeur.

**VS** Schwannome vestibulaire *(vestibular schwannoma)*

# INTRODUCTION

Depuis la fin des années 2000, les avancées matérielles et l'amélioration des capacités de calcul ont favorisé l'émergence et le développement rapide de l'apprentissage profond (DL, *deep learning*). L'entraînement des réseaux de neurones artificiels étant désormais possible en un temps raisonnable, cette technologie a révolutionné de nombreux domaines, dans l'industrie comme dans le quotidien des individus. Diverses tâches, de la génération de texte à l'analyse de signaux, ont trouvé des applications dans presque tous les secteurs, et en particulier celui de la santé. En chirurgie assistée par ordinateur, en télémédecine, ou encore pour mieux comprendre certaines pathologies, les potentiels usages sont multiples. La vision par ordinateur en imagerie médicale a notamment pour objectif de permettre des traitements automatiques [1] :

— précis
— répétables et indépendants de l'observateur
— massifs (pour de grands volumes de données)
— rapides

Ainsi, les modèles de vision par ordinateur apportent des outils supplémentaires aux cliniciens, leur permettant de traiter plus de patients tout en réduisant les potentielles erreurs d'interprétation. Parmi les avancées méthodologiques les plus significatives en imagerie médicale, la segmentation automatique de structures volumétriques d'intérêt par apprentissage profond se démarque par ses performances nettement supérieures aux techniques précédentes dans de nombreux contextes cliniques [2]. En oncologie, domaine d'application de cette thèse, la segmentation est notamment utile à l'extraction de paramètres d'intérêt, par exemple pour évaluer certaines caractéristiques de tumeurs ou définir des volumes de traitement en radiothérapie. La segmentation automatique joue ainsi un rôle crucial afin d'identifier de manière fiable et indépendante de l'utilisateur les masses tumorales, et ce, dans différentes modalités d'imagerie [3].

Malgré ces avancées significatives, certains obstacles récurrents et transverses aux différentes applications cliniques limitent toujours les performances des modèles, et ainsi, leur utilisation en contexte clinique. En pratique, plusieurs biais sont induits par les données elles-mêmes [4]. Le premier est le manque de données disponibles au développement,

notamment à cause du coût d'acquisition et d'annotation des images, mais aussi du caractère confidentiel de ces données. La politique de partage des données patient est variable, ce qui limite leur diffusion (par exemple l'accès public aux données ou non). Les réseaux de neurones profonds ayant besoin de suffisamment d'échantillons pour apprendre correctement, cette limite est critique dans de nombreux contextes. Un deuxième biais majeur est dû à la différence entre les données disponibles en phase de développement et celles rencontrées lors de la phase de déploiement du modèle. Si cette différence est trop importante, le modèle risque de mal identifier les structures clés au sein des images en utilisation clinique et ainsi de fausser ses résultats. De plus, d'une modalité à une autre, et même d'une machine à une autre, les distributions des images acquises peuvent être suffisamment différentes pour réduire les capacités de généralisation d'un modèle et donc ses performances dans le domaine de déploiement. Ce décalage entre les données est donc critique pour l'utilisation finale des modèles dans leur pratique clinique. Il s'agit donc *in fine* d'évaluer dans quelle mesure un modèle peut être déployé sur de nouveaux patients en utilisation réelle, et de s'assurer de ses performances dans ce contexte.

Un certain nombre de stratégies d'apprentissage ont pour objectif de se prémunir de ces biais, notamment en fonction du *régime* de données rencontré (rareté ou relative abondance de données). L'apprentissage *few-shot*, un domaine encore récent où le nombre de données est volontairement très restreint à l'entraînement, est une solution particulièrement adaptée aux bas régimes de données. L'*adaptation de domaine* est un autre thème de recherche visant à conserver les performances d'entraînement dans le domaine de déploiement, améliorant donc la robustesse d'un modèle face à de nouvelles données potentiellement différentes. Parmi ces techniques, l'augmentation de données est particulièrement populaire. Il s'agit de l'ensemble des processus visant à altérer et diversifier les images d'entraînement de façon synthétique pour rendre le modèle plus robuste. Les transformations appliquées sur les images peuvent être assez simples, comme des modifications globales de la luminosité, des rotations, etc. D'autres techniques plus évoluées peuvent néanmoins être considérées, comme la génération d'image synthétiques réalistes par des modèles génératifs avancés. En employant avec précaution les bonnes augmentations, il devient possible de réduire la dépendance aux données réelles lors de l'entraînement, pour un certain niveau de performance. Dans le cadre de cette thèse, nous verrons comment corriger un décalage de données spécifique aux régions d'intérêt dont on souhaite apprendre la segmentation. Nous proposons une méthode d'augmentation réaliste, basée sur un modèle d'harmonisation *one-shot*, afin que celle-ci soit utilisable à tout régime de don-

nées. Nous l'employons ensuite dans différentes configurations et contextes cliniques, afin d'évaluer son potentiel, notamment à moyen et bas régime.

Cette thèse est structurée en cinq chapitres.

Le **premier chapitre** présente le contexte global de nos travaux. Les concepts clés et les différents biais liés aux données y sont explicités. Nous détaillons leur conséquence en segmentation automatique.

Le **deuxième chapitre** introduit les différents champs de recherche de l'adaptation de domaines, incluant notamment l'augmentation de données et l'apprentissage à nombre de données restreint. Un état de l'art sur leurs applications types en imagerie médicale complète ce chapitre.

Dans le **troisième chapitre**, nous présentons et décrivons la méthodologie d'augmentation proposée. Elle se décompose en deux augmentations, une naïve sans DL, et une basée sur le modèle *one-shot* SinGAN, appelée GBA. Une altération linéaire du contraste de la région d'intérêt combinée avec son harmonisation réaliste au sein de l'image d'origine, la méthode profonde génère des images réalistes permettant de suivre une distribution de tumeurs différente de celle d'origine pour réduire un décalage de données. Nous validons qualitativement la synthèse en IRM avec un écart de distribution entre deux centres d'acquisition.

Le **quatrième chapitre** porte sur une première application en segmentation intermodale et multi-centrique de schwannome vestibulaire en IRM. Dans le cadre du challenge MICCAI CrossMoDA 2022, nous proposons une méthode employant notre technique d'augmentation GBA pour diversifier les apparences de tumeurs et réduire un décalage de domaine. À l'issue du challenge, nous avons obtenu la première place sur la tâche de segmentation de tumeur, et troisième au classement général, classement combiné de la segmentation de la tumeur et de l'organe à risque associé. Ce travail a été décrit dans un article à paraître dans les *IEEE Transactions on Biomedical Engineering*.

Le **cinquième chapitre** présente deux applications plus exploratoires montrant le potentiel de nos augmentations pour l'imagerie médicale dans sa globalité, en segmentation de métastases cérébrales sur tomodensitométrie (TDM, *computed tomography*) à bas régimes de données, puis en détection de nodules sur radiographie pulmonaire. Dans le premier cas, nous cherchons à évaluer à quel point il est possible, avec nos augmentations, de réduire la taille du jeu de données d'entraînement sans perte de performance, et ce, dans un contexte de segmentation difficile. Dans le second, nous appliquons l'harmonisation *one-shot* permise par SinGAN dans un contexte où la méthode traditionnelle

*Poisson blending* est généralement employé : la synthèse de nodules 2D sur radiographie à partir de nodules 3D en TDM. Nous mesurons ainsi à quel point ce type d'approches peut modifier les pratiques actuelles.

Enfin, nous concluons ce travail en résumant les différentes contributions apportées. Nous mettons en perspective les conclusions tirées de chaque application pour préciser l'intérêt et les limites de notre augmentation.

Une annexe contenant la portée environnementale et le bilan carbone de cette thèse complète ce manuscrit. Elle fournit également quelques constats sur les émissions de gaz à effet de serre induites par l'apprentissage profond, ainsi que plusieurs pistes de réduction des émissions.

Chapitre 1

# Biais des données d'apprentissage en imagerie médicale

**Résumé**

Ce chapitre introduit plusieurs concepts élémentaires de l'apprentissage profond appliqué à l'imagerie médicale. Nous y présentons certains enjeux et limites associés aux données fréquemment rencontrés ayant motivé les développements méthodologiques de cette thèse. Parmi elles, on compte la rareté et le décalage de données, dont les causes sont multiples. Leur conséquence est généralement une dégradation des performances et de la généralisation des modèles, notamment en segmentation, notre contexte applicatif principal. Pour compenser cela, l'augmentation de données est une piste majeure pour améliorer de façon systématique les performances. En outre, nous montrons que l'intérêt d'augmentations conventionnelles peut s'avérer limité dans certains contextes, en présence de fort décalage de données par exemple.

## 1.1 Généralités et mesure de performance

L'apprentissage profond (DL, *deep learning*) repose sur l'utilisation de réseaux de neurones artificiels permettant d'apprendre automatiquement une représentation hiérarchique des données. Ils sont composés de couches successives de neurones, chacune extrayant des caractéristiques de plus en plus complexes des données d'entrée. La réponse des neurones, et donc cette représentation, est rendue non-linéaire grâce à des fonctions d'activations. Un modèle dit profond bénéficie d'un nombre de couches important (généralement admis

comme supérieur ou égal à 3), lui permettant d'apprendre une représentation latente élaborée des données. Sous certaines conditions, le réseau de neurones multi-couches peut être considéré dans le cas limite comme un *approximateur de fonctions universel* [5].

En pratique, dans les procédures d'apprentissage statistique, une fois la structure du réseau, la stratégie d'entraînement et les données définies, le modèle ajuste ses paramètres pour optimiser une métrique reflétant la tâche à résoudre sur ces données. Cependant, au-delà de l'objectif général d'obtenir un certain niveau de performance lié à cette métrique, il est impératif de s'assurer que le modèle reste fiable et pertinent face à la diversité des situations qu'il pourrait rencontrer en utilisation réelle, lors de la phase dite de *déploiement* de ce modèle. En plus des performances, il est donc nécessaire d'évaluer la *robustesse* statistique du modèle.

## 1.2 Robustesse et généralisation

La robustesse représente la capacité d'un modèle statistique à maintenir des performances élevées même en présence de perturbations ou de variations. Cela revient à rendre le modèle invariant à certaines transformations ou caractéristiques d'images. Ces transformations peuvent prendre différentes formes (par exemple : bruit, flou, changement de système d'acquisition, etc.) et induisent des variations de *distribution* des données. Un modèle robuste sera alors capable de produire des résultats fiables et cohérents dans des conditions différentes de celles sur lesquelles il a été entraîné [6]. Pour illustrer cette notion, une forme typique des courbes d'apprentissage obtenues pour des modèles robustes ou non est montrée sur la Fig. 1.1, où un modèle peu robuste perdra rapidement en performance sur de nouvelles données. En imagerie médicale, la robustesse revêt une importance capitale pour développer des modèles déployables sur plusieurs centres d'acquisition, ou encore applicables à plusieurs modalités d'imagerie. Un exemple de technique pouvant mettre en défaut la robustesse est l'attaque adversaire, consistant à trouver des perturbations minimales qui, une fois ajoutée à une donnée, trompent le modèle [7].

La notion de généralisation, au sens premier du terme, est souvent confondue avec la robustesse. Il s'agit de la capacité d'un réseau à générer des résultats cohérents pour des données n'ayant pas été fournies au modèle lors de l'apprentissage. En d'autres termes, un modèle ayant correctement généralisé n'a pas uniquement mémorisé le jeu de donnée d'entraînement et les résultats attendus associés (sur-apprentissage, *overfitting*) ; il a appris à reconnaître des motifs et des caractéristiques sous-jacentes dans les données, lui permet-

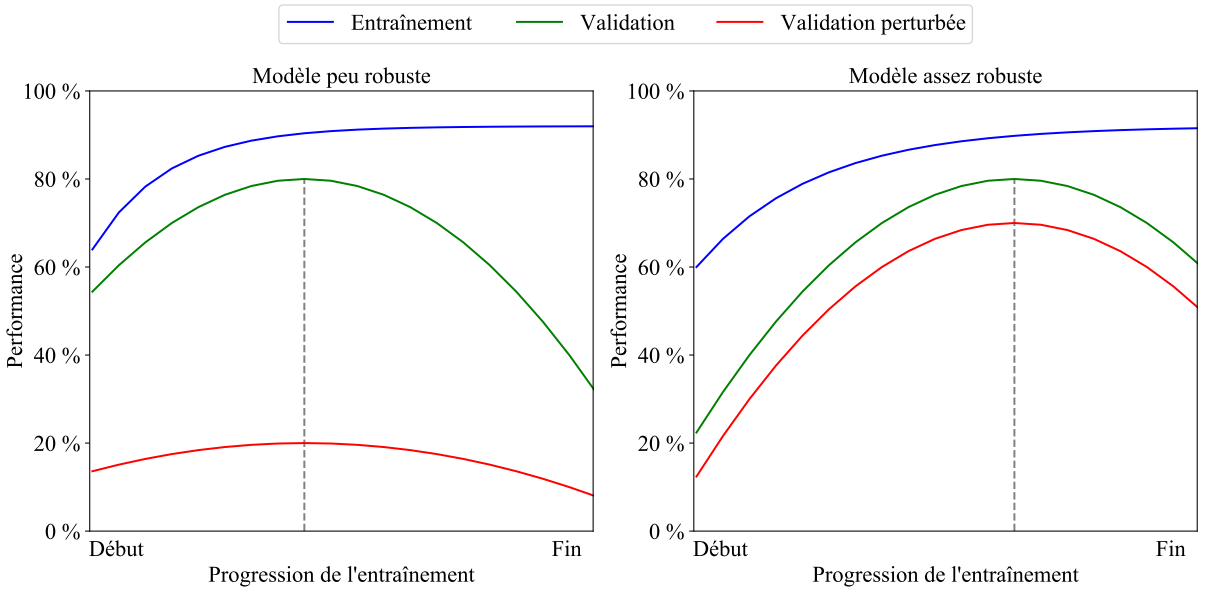



FIGURE 1.1 – Exemple typique de courbes d'apprentissage d'un modèle peu et assez robuste. Un modèle peu robuste montre une importante perte de performance sur des images dégradées ou perturbées, ne suivant donc pas la distribution d'entraînement. Cette perte est réduite lorsque le modèle est suffisamment robuste. Lorsque l'optimum (trait en pointillé) est dépassé, le modèle devient trop spécifique aux données d'apprentissage ; on parle alors de sur-apprentissage et de perte de généralisation.

tant d'effectuer des prédictions précises sur de nouvelles données qui partagent des similitudes avec l'ensemble d'apprentissage, sans perturbation ou variation particulière. Afin de s'assurer de la bonne généralisation du modèle, une pratique communément acceptée est de sous-diviser le jeu de données d'apprentissage en trois parties (ou sous-ensembles) distinctes : l'ensemble d'entraînement *(training)*, l'ensemble de validation *(validation)* et l'ensemble de test *(testing)*, comme présenté sur la Fig. 1.2. Des variations de ce principe avec la validation croisée *k-fold* sont également couramment employées.

Par abus de langage, la généralisation peut aussi représenter la capacité d'un modèle à se déployer sur des données en apparence similaires, mais qui suivent une distribution différente. Elle se confond donc parfois avec la notion de robustesse. La sous-division des jeux de données en ensembles d'entraînement, validation et test ne suffit alors plus pour rendre le modèle robuste à ces variations de distribution. D'autres techniques pour améliorer la généralisation sont alors couramment utilisés, telles que la réduction de la profondeur du réseau ou encore des mécanismes comme le *dropout* qui visent à réduire le

risque de sur-apprentissage en désactivant aléatoirement certains neurones entre chaque rétropropagation de gradient [8]. En imagerie médicale, la robustesse et la généralisation se confondent souvent et sont deux considérations essentielles au vu des contraintes inhérentes à ce domaine.

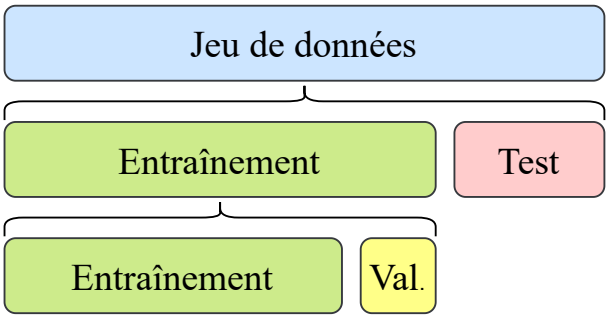



FIGURE 1.2 – Séparation d'un jeu de données en sous-ensembles d'entraînement, validation et test pour encourager la généralisation d'un modèle d'apprentissage profond.

## 1.3 Limitations en imagerie médicale

De nombreux champs de recherche du DL et de la vision par ordinateur trouvent des applications en imagerie médicale, par exemple la segmentation ou l'aide à la décision pour la planification de traitement. Cependant, malgré des avancées significatives, certains obstacles récurrents et transverses aux tâches cliniques en aval de ces traitements limitent les performances des modèles [4]. Ces obstacles sont propres à l'apprentissage statistique, et donc pas uniquement au DL. La Fig. 1.4 récapitule l'ensemble de ces biais en les répartissant dans deux catégories : le manque et le décalage de données, incluant le décalage de domaines.

### 1.3.1 Manque de données et représentativité

Le principal frein à la performance des modèles est le manque de disponibilité des données médicales. Ces données présentent un caractère sensible dû au secret médical et à la vie privée des patients. Les partager à d'autres centres de recherche, voire les rendre publiques, est souvent légalement restreint pour cette raison, limitant le nombre de jeux de données publiques et la quantité de données qu'ils contiennent. En parallèle, certaines pathologies sont particulièrement rares, rendant complexe le rassemblement d'images pour

celles-ci. Même si certains jeux de données conséquents existent, notamment dans le cadre de compétitions internationales (*challenges*), le manque de données reste un problème fréquemment rencontré [9].

Avec peu de données, l'homogénéité ou l'hétérogénéité au sein de celles-ci joue un rôle majeur. Si les données sont trop similaires, le modèle sur-apprend et devient trop spécifique, confondant les motifs généralisables d'une image à l'autre avec des caractéristiques propres aux images d'entraînement, particulièrement le bruit inclus dans leur distribution. Si les données sont trop différentes, le modèle risque de sous-apprendre (*underfitting*) et ne pas réussir à extraire des caractéristiques communes à ces images, particulièrement à partir de peu d'exemples. Un équilibre entre les deux est alors nécessaire, afin que le modèle apprenne la distribution sous-jacente des images, comme illustré sur la Fig. 1.3a, où la distribution d'entraînement associée à la probabilité $P_{\text{tr}}(X)$ ne couvre qu'une partie de la distribution de test $P_{\text{te}}(X)$. Les données doivent être représentatives de cette distribution afin d'en extraire des caractéristiques communes sans devenir trop spécifiques à certains types de tumeurs ou de morphologies par exemple. Chaque acquisition vue lors de l'apprentissage doit donc non seulement être informative en soi, mais doit également contribuer à couvrir, sans trop de redondance, l'ensemble de la variabilité intrinsèque à la tâche. En l'absence de représentativité, les performances risquent d'être limitées en contexte clinique, face à la variété d'images généralement rencontrées. Cette contrainte est suffisamment importante, car une sélection de données véritablement représentatives nécessite à la fois suffisamment de données en amont et une compréhension préalable de la distribution des données.

### 1.3.2 Coût et biais des annotations

Une seconde limitation majeure est la faible quantité de données annotées (labels) disponibles en imagerie médicale. En effet, l'annotation manuelle des images médicales est une tâche laborieuse et coûteuse, nécessitant une expertise spécialisée.

Au-delà du coût de ces annotations, deux biais majeurs sont induits lors du processus d'annotation par un ou plusieurs annotateurs. [4], [10]. Le premier est la variabilité intra-expert, ajoutant une première source de bruit dans les labels. Elle apparaît lorsqu'un même expert annote différemment des images similaires d'une image à une autre, par exemple sous l'effet de la fatigue. Ce biais devient d'autant plus important lorsque le jeu de données à annoter est volumineux. Ensuite, demander à plusieurs experts d'annoter un jeu de donnée introduit un autre biais, la variabilité inter-expert. Elle résulte des

différences d'expertise et d'interprétation entre les multiples experts qui doivent souvent s'accorder à une seule vérité-terrain (*ground-truth*) par image uniquement. Enfin, le ou les annotateurs peuvent être influencés par des informations annexes, comme des hypothèses pré-existantes sur une certaine pathologie ou encore des rapports médicaux apportant des connaissances supplémentaires sur un ou plusieurs patients. Pour limiter cette variabilité liée aux experts, des méthodes basées sur l'apprentissage semi-supervisé peuvent par exemple être employées [11]. En imagerie volumétrique, les experts peuvent ainsi annoter manuellement une coupe pour chaque patient du jeu de données et utiliser ce type de méthode pour étendre l'annotation aux autres coupes. Cependant, il reste nécessaire de vérifier et corriger ces segmentations extrapolées.

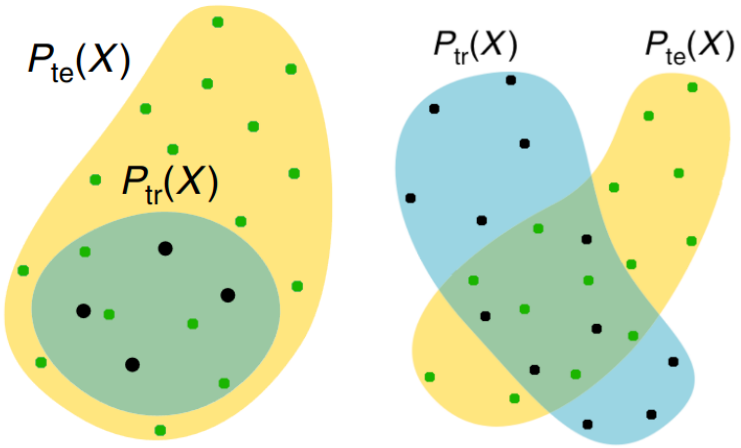



FIGURE 1.3 – Illustration du a) biais de sélection lié au manque de données et du b) décalage de domaine. $P_{te}(X)$ représente la distribution d'entraînement, et $P_{te}(X)$ la distribution test ou cible. Dans les deux cas, la distribution d'entraînement ne correspond pas à celle de test. Adapté de [4].

### 1.3.3 Décalage de données

Le décalage de données (*data mismatch*), illustré en Fig. 1.3, représente la situation où les conditions d'entraînement et de test sont significativement éloignées. Cette différence implique alors une dégradation des performances ainsi qu'un manque de robustesse des modèles sur ces données de test. En pratique, cette limitation est particulièrement présente car les images rencontrées en contexte clinique sont souvent beaucoup plus variables que

celles rencontrées lors du développement du modèle [4].

La première origine du décalage de données, mentionnée en section 1.3.1 et représenté en Fig. 1.3a, est le biais de sélection. À cause de leur rareté ou d'un manque d'hétérogénéité, les données sélectionnées pour l'entraînement ne représentent seulement qu'une partie de la distribution de test. Ce problème souligne le besoin en robustesse des modèles, car au-delà du manque de données, il n'existe aucune manière de s'assurer que les données d'apprentissage soient représentatives de toute la variété anatomique ou pathologique possible.

La seconde origine est l'écart ou décalage de domaine, illustré en Fig. 1.3b et souvent confondu avec le terme de décalage de données. Dans cette situation, la distribution d'entraînement diffère de celle de test sans être nécessairement incluse dans celle-ci, et cela dû à des facteurs extérieurs. Tout d'abord, un modèle entraîné pour une certaine population peut mal généraliser à d'autres populations, en fonction par exemple de l'âge, le sexe ou l'origine ethnique. Les variations morphologiques entre ces populations, d'un point de vue probabiliste, modifient profondément les distributions des images, ce qui suffit à créer un écart de domaines ; on parle alors de biais de population. Ensuite, les différences d'équipements et de protocoles cliniques peuvent également créer un décalage de données. D'un centre clinique à l'autre, les équipements diffèrent, avec des machines d'acquisition plus ou moins récentes et produites par différents constructeurs, résultant en une grande variété de contraste, résolution et quantité de bruit différents. Ainsi, même si ces différences constantes d'apparence ne présentent pas de frein particulier pour le personnel hospitalier, elles suffisent à détériorer les performances d'un modèle en créant un décalage de domaines ; on l'appelle biais d'acquisition. De plus, les avancées technologiques dans les mécanismes d'imagerie et leur manière d'acquérir une image permettent généralement d'améliorer certaines contraintes physiques (temps d'acquisition, résolution, etc.), créant une « nouvelle » modalité associée à un nouveau domaine. Pour toutes ces raisons, un même modèle basé apprentissage profond peut donc obtenir des performances élevées pour un centre et basse pour un autre, malgré un protocole d'acquisition similaire [12]. Enfin, certaines tumeurs, comme les métastases cérébrales, évoluent très rapidement et montrent différents stades d'évolution. Ainsi, une acquisition le jour J ou le jour J+3 peuvent mener à deux résultats très variables, modifiant potentiellement leur interprétation. L'impact d'un traitement modifie également l'apparence de la tumeur, plus ou moins rapidement. Plus généralement, si la variété des stades d'évolution d'une certaine pathologie n'est pas représentée dans le jeu de données, un décalage de domaine peut apparaître.

On nomme cette limite le biais d'évolution ou de manifestation d'une maladie.

En parallèle, les différences de prévalence d'une pathologie au sein d'un jeu de données peuvent aussi biaiser l'apprentissage vis-à-vis des données de test ou de la future utilisation clinique. Par exemple, pour la prédiction de la survie d'un patient dans le cas d'une certaine pathologie, un modèle est susceptible d'être influencé par la proportion de patients ayant survécu parmi les patients disponibles lors de l'entraînement si celle-ci est trop forte ou faible [13]. On parle alors de déséquilibre de classe, abordé plus en détail dans le cas de la segmentation en section 1.4.2.

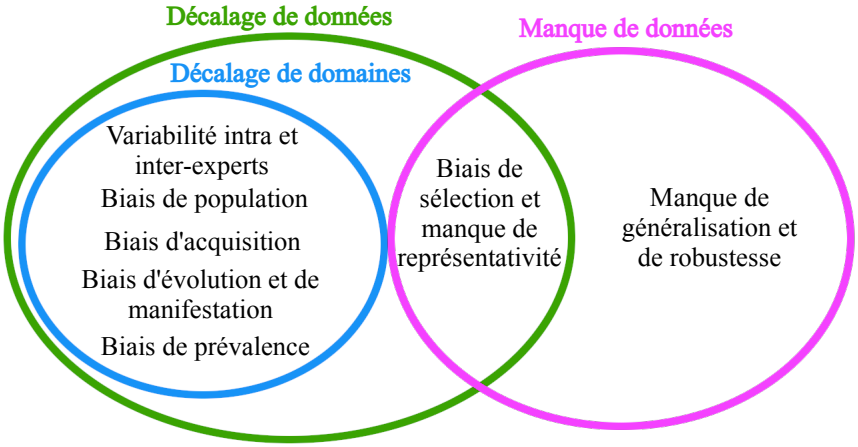



FIGURE 1.4 – Résumé des différents biais liés aux données découlant du manque et des décalages de données. Le biais de sélection est souvent induit par le manque de données et crée un décalage de données.

## 1.4 Le cas de la segmentation d'images

Dans le cadre de la segmentation sémantique d'images, le contexte applicatif principal de cette thèse, la conséquence principale des limites évoquées est la dégradation des labels prédits durant la phase de déploiement du modèle en contexte clinique, face à des données inédites. Plus précisément, le manque de données et d'annotations réduit la généralisation tout en augmentant l'incertitude associée aux masques prédits. Les erreurs répétées dues à la variabilité intra-expert peuvent également biaiser l'apprentissage, sans qu'une perte de performance n'apparaisse sur le jeu de test, créant un biais non quantifiable sans une étude de la robustesse du modèle. Le manque de représentativité et le décalage de

données génèrent des risques similaires. Ainsi, un modèle de segmentation développé pour un certain centre d'acquisition devra être adapté, dans les cas les plus simples, avec un réajustement des paramètres (*fine-tuning*), voire nécessiter un ré-entraînement complet du modèle, s'exposant alors aux potentiels biais présent dans ces nouvelles données.

En dépit des limitations inhérentes à l'apprentissage statistique, le DL domine largement le domaine de la segmentation depuis 2015. L'engouement pour ces modèles s'explique notamment par le succès de l'architecture convolutive U-Net [14], qui a permis des gains très significatifs en performance par rapport aux méthodes de traitement d'images alternatives de l'époque [15]. Un U-Net, représenté en Fig. 1.5, consiste en un réseau encodeur-décodeur extrayant de façon hiérarchique des relations de plus en plus globales dans l'image tout en compressant sa représentation spatiale, à la manière d'un auto-encodeur [16]. Contrairement à l'auto-encodeur, les cartes de caractéristiques obtenues par l'encodeur avant chaque opération d'agrégation (*pooling*) sont transférées par concaténation au décodeur. Ces connexions supplémentaires, appelées *skip connections*, facilitent la propagation d'information haute fréquence et ainsi l'obtention de cartes de segmentation précises.

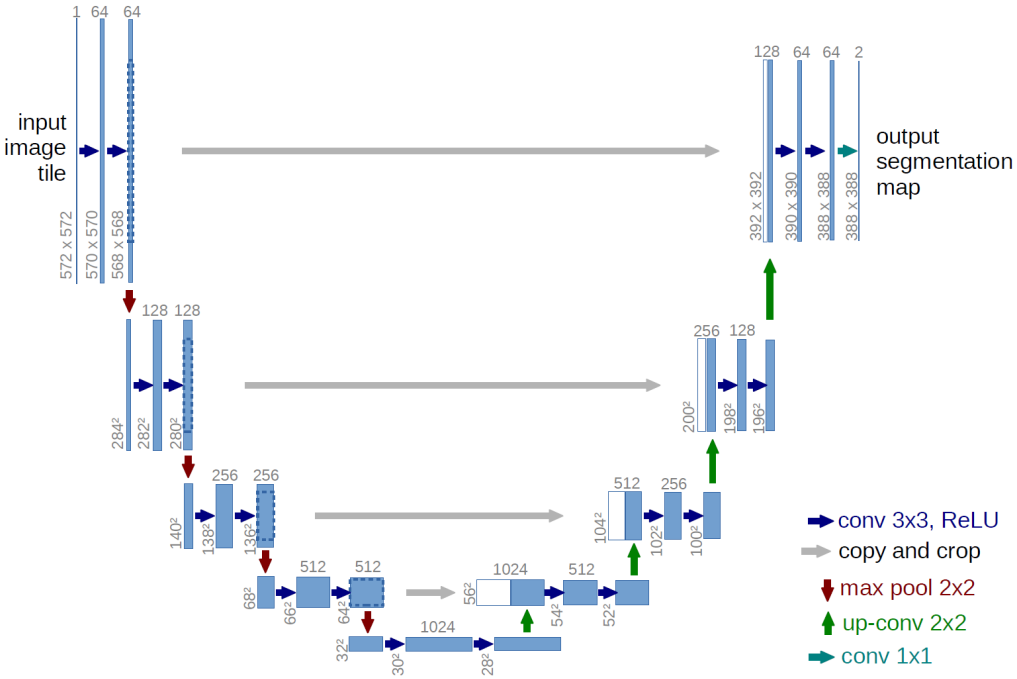



FIGURE 1.5 – Architecture U-Net standard. Extrait de [14].

La supériorité des U-Net en segmentation vis-à-vis d'autres approches a été démontrée

dans divers contextes [17], [18]. Leur utilisation à large échelle en imagerie médicale a été favorisée et démocratisée par de nombreuses victoires à des compétitions de segmentation, aussi bien à bas qu'à haut régime de données [19] ou encore en présence de décalage de données [20]. En particulier, une solution de segmentation générique basée U-Net, nommée nnU-Net (pour *no new U-Net*), a démontré sa robustesse vis-à-vis de modèles de segmentation plus spécifiques [21]. Ce *framework* utilise une architecture U-Net et optimise, en fonction des données fournies et grâce à différentes heuristiques, plusieurs paramètres et hyper-paramètres concernant le modèle et le pré-traitement des données. nnU-Net s'est rapidement imposé comme la *baseline* de référence en segmentation d'images médicales [22].

Nous distinguons trois principaux axes de recherche en segmentation d'images par DL, en fonction que les modifications soient liées 1) à l'architecture du réseau de neurones 2) à l'optimisation de l'entraînement, et principalement aux fonctions de coût adaptées, ou 3) au traitement de la donnée d'entraînement ou de test par des méthodes d'adaptation et/ou d'augmentation [23]. C'est ce troisième point qui nous intéresse en particulier dans le cadre de cette thèse.

### 1.4.1 Modifications architecturales

De nombreuses variantes des modèles U-Net ont été développés par la communauté pour s'adapter aux différentes tâches et contextes cliniques [24]. Plus récemment, des mécanismes d'attention comme les *vision transformers* et des approches basées diffusion ont également été employés en segmentation, sans toutefois montrer une amélioration significative dans la plupart des contextes [25]-[27]. Pour le moment, il n'existe aucune règle générale concernant le type d'approche ayant les meilleures performances pour une tâche donnée, car aucun gap conséquent n'a été observé [28]. Le choix de nnU-Net de se concentrer sur une architecture U-Net standard malgré le nombre de variantes et d'alternatives souligne l'efficacité et la polyvalence de cette architecture [18]. Ce choix réduit également l'intérêt du développement d'architectures plus complexes pour les situations génériques.

### 1.4.2 Fonctions de coût

En segmentation d'organes comme de tumeurs, plusieurs fonctions de coût sont communément employées. La plus utilisée est l'entropie croisée (*cross-entropy*) et représente

la classification de l'ensemble des pixels ou voxels d'une image [29] :

$$\mathcal{L}_{\text{Cross-entropy}} = -\sum_i [g_i \log(p_i) + (1 - g_i) \log(1 - p_i)] \tag{1.1}$$

où $p_i$ et $g_i$ représentent la prédiction et la vérité-terrain associés au pixel ou voxel $i$. Elle mesure ainsi la dissimilarité entre la distribution prédite par le modèle et la vérité-terrain. Cette fonction de coût pondère autant les pixels de chaque classe. Cependant, dans le cas de la segmentation tumorale, les tumeurs représentent généralement une très faible proportion des pixels ou voxels des images, créant un déséquilibre de classe. Cela peut affecter les algorithmes d'apprentissage classiques qui, entraînés sur de telles données, ont tendance à prédire la classe majoritaire, soit les pixels ou voxels non tumoraux, au détriment de la minoritaire. Cette prédominance amène donc à des faux négatifs, c'est-à-dire à la non-détection de certaines tumeurs, ce qui est critique en routine clinique. Pour résoudre ce déséquilibre (sans distinguer faux négatifs et faux positifs), on introduit souvent une fonction de coût associée au coefficient de Dice [30], [31] :

$$\mathcal{L}_{\text{Dice}} = 1 - \frac{2 \sum_i p_i g_i}{\sum_i p_i + \sum_i g_i} \tag{1.2}$$

avec les mêmes notations. Celle-ci prend en compte la différence spatiale entre les classes prédites, hors arrière-plan. Afin de pénaliser plus particulièrement les faux négatifs, l'indice de similarité de Tversky peut également être utilisé comme fonction de coût [32], [33] :

$$\mathcal{L}_{\text{Tversky}} = 1 - \frac{\sum p_i g_i}{\sum p_i g_i + \alpha \sum p_i (1 - g_i) + \beta \sum (1 - p_i) g_i} \tag{1.3}$$

où $\alpha$ et $\beta$ sont des poids relatifs associés respectivement aux faux positifs et faux négatifs. En adaptant ces deux paramètres, la capacité de détection du modèle peut être affinée.

De nombreuses autres fonctions de coût ont été proposées par la communauté en fonction des défis spécifiques aux tâches ou aux différents jeux de données [34]. Elles sont généralement rassemblées en trois catégories : celles basées sur les différences de distribution entre labels comme l'entropie croisée, celles évaluant les recouvrements spatiaux de prédictions comme le Dice, ainsi que celles se concentrant sur les contours des masques. Parmi l'ensemble, les plus connues et utilisées restent celles d'entropie croisée et de Dice. Par défaut, nnU-Net utilise exclusivement ces deux fonctions de coût. Même si ce *framework* utilise plusieurs heuristiques basées sur les données d'entraînement pour déterminer certaines caractéristiques de son modèle, les poids relatifs de ces deux fonctions de coût

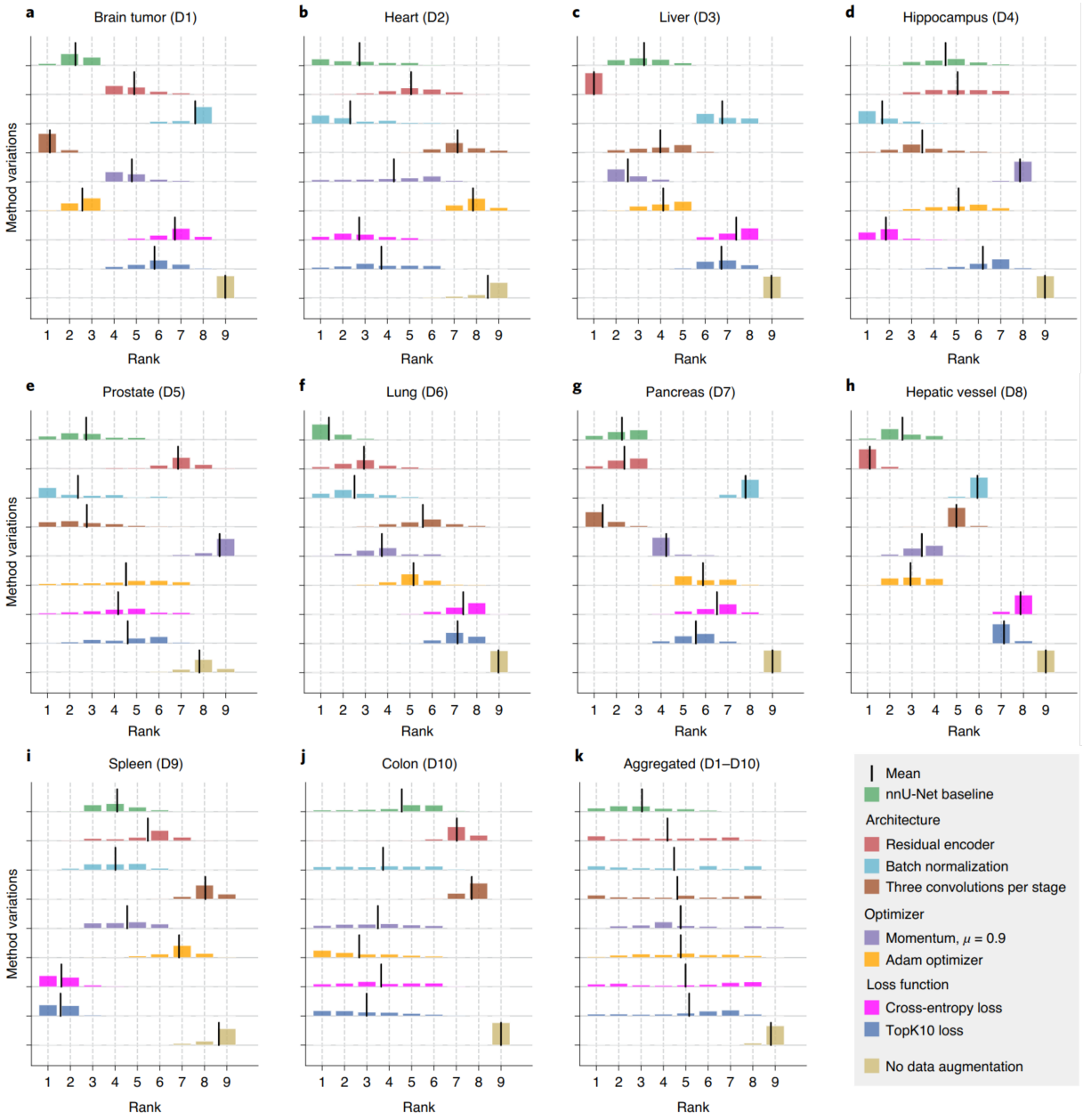



FIGURE 1.6 – Ablations de nnU-Net sur les dix jeux de données du décathlon de la segmentation médicale [19]. Le retrait des augmentations de données est critique et place cette ablation derrière toutes les autres. Extrait de [21].

sont fixes. Cela souligne que dans le cas le plus générique, affiner ou adapter les poids relatifs à chaque situation, voire ajouter d'autres fonctions de coût, ne semble permettre que de faibles gains de performance.

### 1.4.3 Données d'apprentissage

Afin d'uniformiser le format des données et d'encourager la généralisation, les images sont généralement normalisées et ré-échantillonnées vers un échantillonnage spatial (*spacing*) cible avant apprentissage. En fonction de la modalité d'imagerie, certaines étapes comme le fenêtrage ou le retrait des percentiles extrêmes sont également appliquées. Ces pratiques réduisent les différences de distribution des images sans modifier leur contenu, et donc sans résoudre les potentiels manques de données, d'annotations ou décalages de données. Pour cette raison, de nombreuses stratégies ont été développées et sont regroupées sous le terme d'adaptation de domaine. Il s'agit de l'ensemble des processus permettant d'ajuster un modèle entraîné sur une distribution source à mieux performer sur une distribution cible différente. L'adaptation de domaine inclut différentes techniques, comme la synthèse image-à-image ou encore l'apprentissage par transfert, décrites au chapitre 2. Ces méthodes nécessitent une attention particulière en amont, souvent avec l'entraînement d'autres réseaux de neurones. Ainsi, elles ne peuvent pas être simplement appliquées sur les images avant passage dans le réseau de segmentation.

La méthode la plus répandue d'adaptation de domaine est l'augmentation de données. Elle consiste à créer de nouveaux exemples d'entraînement synthétiques en appliquant des transformations aux données existantes. Parmi les exemples les plus conventionnels, on compte l'ajout de bruit gaussien, de flou linéaire, ou encore de rotations de l'image. Cela permet d'exposer le modèle à une plus grande diversité, et ainsi d'encourager sa capacité de généralisation. Les techniques d'augmentation peuvent par ailleurs être employées pour capturer l'incertitude liée aux données lors de l'étape d'inférence, en examinant la robustesse du modèle à des perturbations de l'entrée [35], [36]. Nous discutons plus en profondeur de méthodes d'augmentation au chapitre suivant, en section 2.2.

En segmentation d'images, l'augmentation de données a montré son intérêt aussi bien à bas qu'à haut régime de données [37]. Par exemple, au sein de nnU-Net, les augmentations conventionnelles employées (rotations, changement d'échelle, bruit gaussien, flou gaussien, changement de luminosité et contraste, diminution de résolution, correction gamma et symétrie miroir) semblent revêtir une importance capitale vis-à-vis de la performance [21]. En particulier, une large étude ablative, portant sur de nombreux challenges, a été

conduite pour identifier quels sont les choix les plus critiques entre architecture, optimisation et augmentation [21]. La Fig. 1.6 évalue plusieurs ablations du *framework*, comme le retrait des augmentations lors de l'apprentissage. Les retirer dégrade considérablement les résultats et place cette ablation derrière toutes les autres. En comparaison, les modifications d'architectures ou de fonctions de coût contribuent de manière plus marginale aux performances. Cela souligne l'aspect critique des augmentations sur les performances des modèles de segmentation.

Toutefois, l'impact des augmentations conventionnelles risque d'être plus réduit lorsqu'un fort décalage de données est présent entre les domaines d'entraînement et de déploiement [38]. L'utilisation d'augmentations conventionnelles n'est alors généralement pas suffisante pour permettre au réseau d'apprendre efficacement une représentation latente commune aux modalités. De plus, ces augmentations ont émergé pour l'imagerie naturelle afin de rendre le modèle plus robuste aux caractéristiques non pertinentes, comme le bruit ou le flou. En imagerie médicale, on souhaite également que le modèle soit robuste à d'autres caractéristiques, comme les artefacts propres aux différentes modalités, ou aux différences visuelles observées d'un centre clinique à l'autre [39].

## 1.5 Hypothèses de la thèse

L'augmentation de données conventionnelle, c'est-à-dire issue de traitements simples de l'image, permet des gains de performance significatifs en segmentation pour une grande variété de régimes de données. Il semble néanmoins pertinent de penser que des stratégies d'augmentation plus élaborées et dédiées à l'imagerie médicale permettraient des gains supplémentaires, en particulier à faible régime de données.

Dans un contexte de segmentation de tumeurs, nous formulons d'une part l'hypothèse qu'il est préférable de diversifier l'apparence des régions d'intérêt dans une proportion correspondant à la distribution d'apparence du domaine cible, au lieu d'augmentations aléatoires décorrélées de cette distribution. Une augmentation ajustable en fonction du décalage de données rencontré lors du déploiement permettrait ainsi d'augmenter la robustesse du modèle de manière plus efficace. D'autre part, nous pensons que cette diversification d'apparence doit être réaliste vis-à-vis des propriétés du domaine cible. Dans cette optique, les modèles génératifs profonds à bas régime de données offrent une piste technique prometteuse. Ainsi, nous proposons d'appliquer des transferts réalistes d'apparence des tissus altérés (c'est-à-dire, une harmonisation) entre les domaines d'entraînement et

de déploiement.

## 1.6 Conclusion

Nous avons vu dans ce chapitre les limites récurrentes fréquemment rencontrées en imagerie médicale, à savoir le manque de données, d'annotations et le décalage de données. Ces biais ont de multiples origines, liées au processus d'annotation ou encore aux différences de matériels d'acquisition et de protocoles d'un centre clinique à l'autre. Ils entraînent principalement des pertes de performances ou de généralisation des modèles, limitant leur utilisation potentielle en contexte clinique.

Dans le domaine de la segmentation, les changements de fonctions de coût ou de topologie du réseau peuvent montrer des niveaux de performances variables selon les situations d'entraînement et les applications. À l'inverse, l'augmentation de donnée, même conventionnelle, permet une amélioration marquée et constante des performances dans des régimes de données variés. En intégrant d'autres augmentations plus spécifiques à l'imagerie médicale, nous pensons pouvoir réduire l'impact des différents biais présents au sein des données.

Dans le chapitre suivant, nous détaillons différentes méthodes d'adaptation de domaine existantes et leurs limitations. Dans la suite de ce manuscrit, nous introduirons une méthode d'augmentation de données permettant d'altérer l'apparence des régions d'intérêt de manière réaliste vis-à-vis du domaine de déploiement.

Chapitre 2

# Adaptation de domaine : méthodes et limitations

**Résumé**

Ce chapitre décrit plusieurs approches d'adaptation de domaine, notamment à bas régime de données (synthèse, augmentation de données, transfert de connaissance). En particulier, nous mentionnons les techniques visant à générer des images réalistes et réduire le décalage de données, qui sont celles les plus proches des développements méthodologiques proposés dans cette thèse. Plusieurs de ces méthodes se fondant sur des modèles génératifs de type réseaux antagonistes génératifs (GAN, *Generative adversarial network*), leur principe y est également explicité. Nous évoquons leurs applications à l'imagerie médicale, ainsi que les limitations des techniques associées.

L'adaptation de domaine représente l'ensemble des méthodes visant à améliorer la performance d'un modèle lorsqu'il est appliqué à des données provenant d'un domaine (ou distribution) cible différent du domaine source sur lequel il a été initialement entraîné [40]. Les deux domaines sont différents, mais partagent leur contenu ainsi que certaines caractéristiques (anatomie, présence d'une même pathologie, etc.) tout en variant d'apparence et de style (par exemple montrant des variations de texture ou de contraste d'une même structure anatomique). Une part importante des techniques d'adaptation de domaine consiste ainsi à réduire les disparités entre les distributions des données d'entraînement et de test. Cela permet d'améliorer la généralisation, par extension la robustesse et la performance du modèle, sans modifier la tâche à apprendre. En imagerie médicale, de nombreux champs de recherche peuvent être associées à de l'adaptation de domaines,

comme certaines applications basées sur la synthèse image-à-image, la généralisation de domaines, l'apprentissage par transfert, ou encore l'augmentation de données.

## 2.1 Synthèse image-à-image et GAN

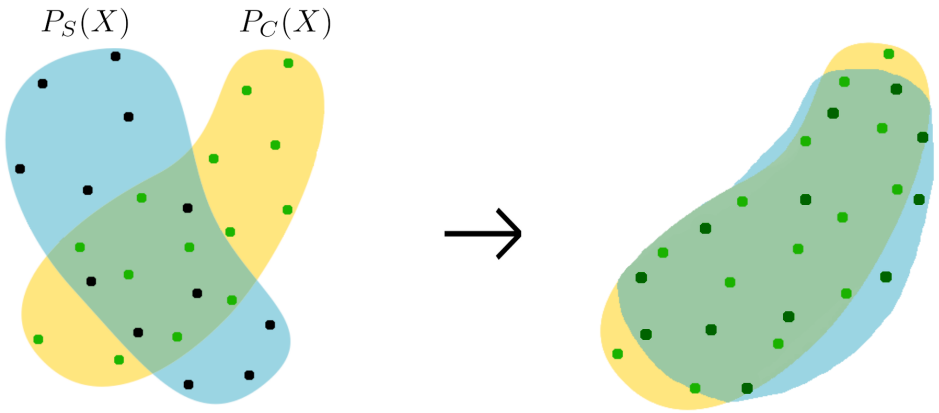



FIGURE 2.1 – Schéma de la synthèse image-à-image (I2I). $P_S(X)$ et $P_C(X)$ représentent respectivement les distributions source et cible. Le domaine source est alors adapté pour correspondre à la distribution cible. Il est alors communément appelé domaine pseudo-cible. Inspiré de [4]

La synthèse image-à-image (I2I, *image-to-image*) consiste à apprendre une application (*mapping*) entre deux domaines d'images *source* et *cible* différents. Plus précisément, chaque élément (pixel ou voxel) des images de la distribution source est traduit dans la modalité cible, de sorte que les images générées suivent globalement la distribution cible. Ce type d'approche permet donc de rapprocher deux distributions en générant un nouveau domaine, souvent appelé "pseudo-cible" (*pseudo-target*), censé suivre la distribution cible comme illustré sur la Fig. 2.1.

Les deux domaines peuvent représenter les données d'entraînement et de test, ou bien plus simplement deux distributions différentes. Dans ce cas, employer une approche d'I2I peut permettre non seulement de réduire un décalage de domaine, mais aussi de générer des images suivant une certaine distribution pour d'autres objectifs, par exemple pour l'augmentation de données. Deux configurations typiques d'approches I2I existent alors selon la disponibilité des données :

— la synthèse I2I supervisée, lorsqu'il existe une bijection d'un domaine à l'autre,

c'est-à-dire que chaque image du domaine source est alignée avec une image du domaine cible. On dira alors que les deux ensembles sont appairés ou alignés, et on travaillera avec des paires d'images évaluées par un coût de reconstruction de type norme $\ell_1$ ou $\ell_2$. C'est typiquement la configuration d'un modèle pix2pix [41].

— la synthèse I2I non supervisée, lorsque aucun appariement entre les domaines source et cible n'existe. Ce cas est plus général, car aucune hypothèse sur les données d'entraînement n'est nécessaire. En revanche, l'optimisation directe avec un coût de reconstruction est impossible, et des stratégies d'entrainement alternatives doivent être considérées, comme dans le modèle CycleGAN [42].

En imagerie médicale, la synthèse I2I supervisée est particulièrement adaptée à l'imagerie cérébrale, où les mouvements entre acquisitions sont limités, ou peuvent être rendus négligeables par recalage [43], [44]. En revanche, la synthèse non supervisée est souvent considérée pour d'autres localisations, par exemple en imagerie abdominale où les mouvements respiratoires des patients empêchent l'alignement fidèle entre modalités [45]. Les modèles de synthèse I2I les plus souvent considérés en imagerie médicale reposant en majorité sur les GAN [46], et en particulier des GAN conditionnels comme les modèles CycleGAN et pix2pix.

### 2.1.1 Principe des GAN

Les GAN sont une classe de méthodes d'apprentissage profond proposée pour la première fois en 2014 [47], [48]. Les modèles basés GAN ont émergé après avoir montré une capacité de génération de données particulièrement réalistes. Initialement non supervisée et non-conditionnelle pour l'imagerie naturelle, ce type d'approche a trouvé de nombreuses applications, y compris en imagerie médicale.

À l'origine, un GAN est constitué de deux réseaux de neurones distincts, appelés le générateur et le discriminateur. Le générateur cherche à produire des données qui suivent le plus possible la distribution de données réelles. Il prend en entrée un vecteur de bruit aléatoire et le transforme en une image synthétique. Le discriminateur, un classificateur binaire, est entraîné pour différencier les vraies données de celles synthétisées par le générateur. Ces deux réseaux sont entraînés simultanément dans un jeu compétitif à somme nulle de type *minimax*, d'où le terme "antagoniste" [49]-[51]. Chaque réseau cherche à minimiser sa propre fonction de coût tout en maximisant celle de son adversaire. Ainsi, au fur et à mesure de l'entraînement, le générateur s'améliore pour produire des données qui ressemblent de plus en plus à des données réelles, tandis que le discriminateur amé-

liore également sa capacité à distinguer les vraies données de celles générées. Le processus s'arrête idéalement lorsque le discriminateur ne parvient plus à distinguer les données réelles des données générées, montrant donc un taux de réussite de 50%, ou lorsque le générateur produit des données satisfaisantes suffisamment proches de la distribution des vraies données. Formellement, on peut définir cet apprentissage avec l'expression suivante [47] :

$$\min_G \max_D V(D,G) = \mathbb{E}_{x\sim p_{\text{data}}(x)}[\log D(x)] + \mathbb{E}_{z\sim p_z(z)}[\log(1 - D(G(z)))] \tag{2.1}$$

où $V(D,G)$ est la fonction de valeur du jeu minimax, $G$ et $D$ respectivement le générateur et le discriminateur, $x$ un échantillon de donnée réelle tiré de la distribution de données réelles $p_{\text{data}}$, et $z$ un échantillon de bruit tiré de la distribution de bruit $p_z$. On nomme communément la fonction de coût associée comme étant "adversaire" (*adversarial loss*).

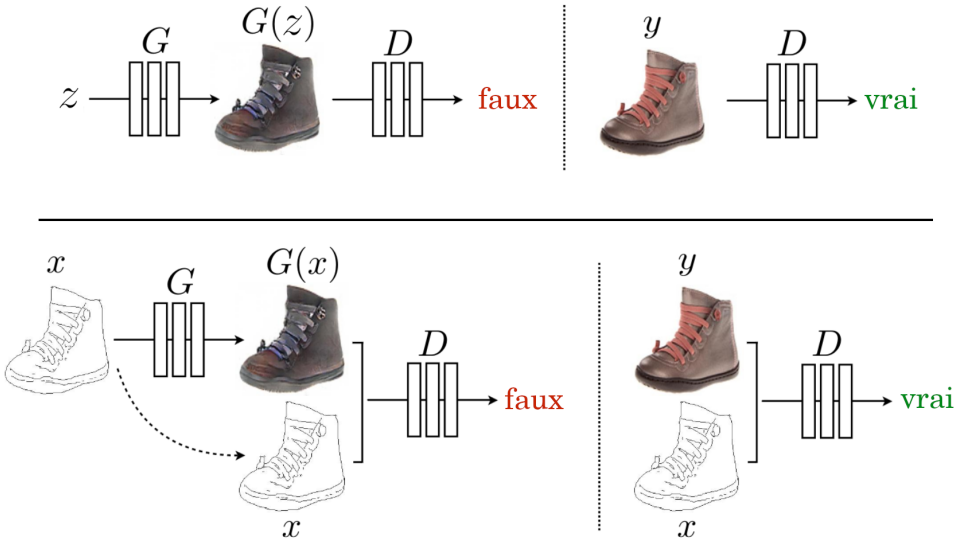



FIGURE 2.2 – Schéma comparant le fonctionnement d'un GAN non-conditionnel (en haut) et conditionnel (en bas). Le générateur $G$ cherche à synthétiser une image réaliste à partir de bruit ou d'informations, tandis que le discriminateur $D$ cherche à distinguer images réelles et générées. Inspiré de [41].

Ce principe a ensuite été adapté pour prendre en compte un vecteur d'informations supplémentaire ou une image avant la génération, avec le GAN conditionnel, présenté en Fig. 2.2. Dans cette configuration, ces informations sont fournies en entrée du générateur et parfois aussi du discriminateur, en supplément ou à la place du bruit. Ces informations peuvent représenter un label cible, afin de générer une sortie associée à celui-ci [52], ou

encore directement une image, permettant la génération d'une image à partir d'une autre, avec des architectures profondes et convolutives de GAN [53]. La formule 2.1 devient alors :

$$\min_G \max_D V(D,G) = \mathbb{E}_{x\sim p_x(x),y\sim p_y(y)}[\log D(x,y)] + \mathbb{E}_{z\sim p_z(z),y\sim p_y(y)}[\log(1 - D(G(z|y),y))] \tag{2.2}$$

où $y$ représente les informations conditionnelles, suivant la distribution $p_y$. Ainsi, de nombreuses méthodes de synthèse I2I sont basées sur les GAN, elles seront détaillées dans la section suivante.

Cependant, l'entraînement direct des GAN montre plusieurs limitations. La plus connue est l'effondrement des modes (*mode collapse*), se traduisant par un déséquilibre entre le générateur et le discriminateur [53], [54]. Idéalement, ces deux entités doivent se développer ensemble de manière équilibrée, mais si l'un des deux devient trop puissant par rapport à l'autre, l'équilibre peut être rompu. Ainsi, si le générateur est trop puissant, il peut générer certains échantillons (modes) avec précision, mais n'a plus besoin de couvrir l'ensemble de la distribution des données réelles pour tromper le discriminateur. À l'inverse, si le discriminateur est trop puissant, il peut être capable de repérer facilement des différences subtiles entre les vrais et les faux échantillons, ce qui conduit le générateur à se concentrer sur quelques modes précis de la distribution des données réelles pour maximiser ses chances de tromper le discriminateur. Ce biais crée donc un problème de non-convergence des GAN [54]. De plus, il est difficile, voire impossible, d'observer ces biais sur les courbes des fonctions de coût. La Fig. 2.3 montre des courbes typiques d'évolution lors de l'entraînement d'un GAN (ici un CycleGAN). Même si ces courbes peuvent être lissées, elles restent peu informatives.

Pour contrer ce phénomène, des améliorations ont été proposées par la communauté, notamment en utilisant la distance Wasserstein [55]-[57]. Celle-ci permet de réduire le risque d'effondrement des modes, tout en améliorant la stabilité et la lisibilité des courbes représentant l'évolution des fonctions de coût. Cependant, l'amélioration est mince et ce type de méthode dégrade parfois la qualité des images générées, nécessitant de trouver un équilibre entre stabilité et qualité [58]. Une alternative consiste à commencer l'entraînement à basse résolution puis à l'augmenter progressivement [59]-[61]. Cet entraînement multi-échelle permet au générateur de se concentrer sur la structure globale des données (comme l'anatomie en imagerie médicale) avant d'affiner la génération à haute résolution avec l'ajout de couches convolutionnelles dédiées. Cependant, l'entraînement est plus coûteux et nécessite un plus grand nombre de paramètres à définir en amont (quand changer

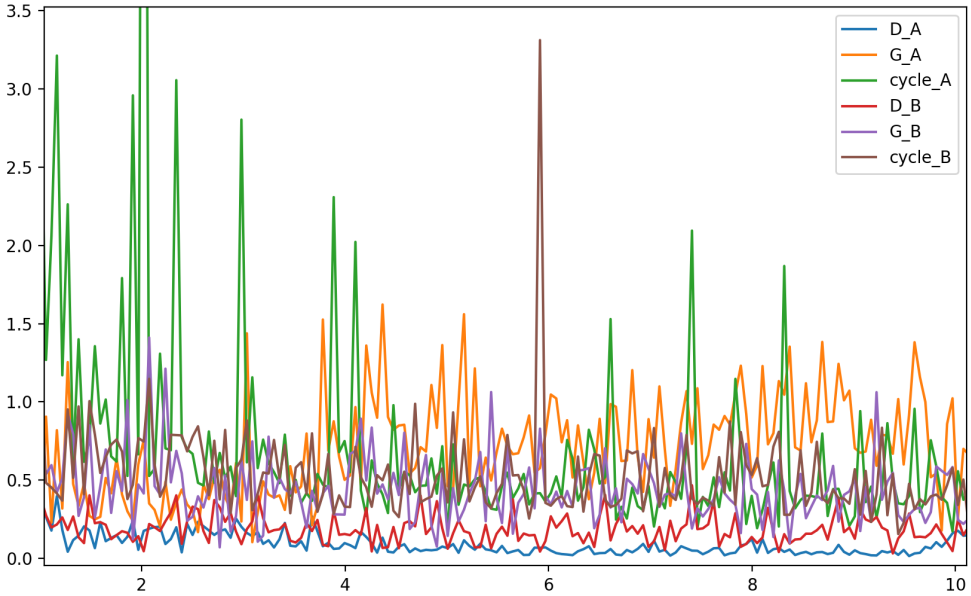



FIGURE 2.3 – Variation des fonctions de coût lors de l'entraînement d'un CycleGAN. $A$ et $B$ représentent les domaines source et cible. Les fonctions de coût $\{D\_X, G\_X, \text{cycle}\_X\}$ sont respectivement les fonctions de coût du discriminateur, du générateur et des images reconstruites dans le domaine $X$.

les résolutions, comment ajuster les poids entre les anciennes et les nouvelles couches convolutionnelles, etc).

De nombreuses modifications des GAN conditionnels ont également été proposées par la communauté, pour de nombreux cas d'applications, y compris médicales [62]-[64]. Parmi les évolutions notables, on peut compter l'ajout de mécanisme d'auto-attention (*self-attention*) [65]-[67], la combinaison avec un auto-encodeur variationnel (VAE, *variational auto-encoder*) [68], nommée StyleGAN et se concentrant sur la texture des images générées [69]-[71], ou encore l'utilisation d'un espace latent intermédiaire sémantique [72]. Les approches développées spécifiquement pour la segmentation, comme l'ajout d'un segmenteur en parallèle du générateur, seront détaillées en section 2.5.

### 2.1.2 Limite des GANs en imagerie médicale

De nombreuses modifications des GAN conditionnels ont été proposées par la communauté médicale, pour de nombreux cas d'applications [62]-[64]. En synthèse supervisée, les modèles pix2pix et pix2pixHD sont les plus connus et employés, calculant une fonction de coût de norme $\ell_1$ avec la vérité-terrain en plus d'une fonction de coût adversaire [41], [73],

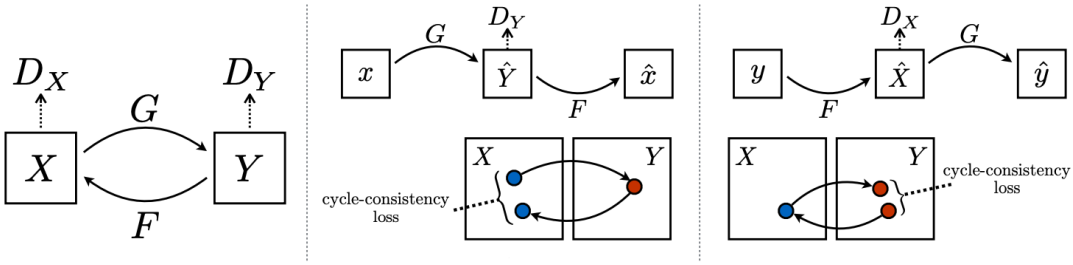



FIGURE 2.4 – Schéma illustrant la notion de cohérence cyclique et la fonction de coût associée. Une image réelle $x \in X$ est fournie au générateur $G$ pour obtenir son équivalent dans le domaine $Y$, avant d'être fournie au second générateur $F$ pour reconstruire l'image $\hat{x}$ et la comparer avec $x$. Idem avec $y \in Y$ en appliquant les générateurs $F$ puis $G$. Extrait de [42].

[74]. Une extension de ce modèle pour l'imagerie volumétrique, appelé vox2vox, permet à la fois la synthèse I2I et la segmentation en imagerie médicale [75]. Ce type d'approches est utilisé entre des domaines différents, pour la génération d'images inter-modalité, le débruitage ou encore la super-résolution [62].

Pour la synthèse non supervisée, une plus grande variété de modèles a été développée en fonction des applications et des jeux de données. Une partie des approches utilisent la notion de cohérence cyclique introduite avec le modèle CycleGAN, présenté en Fig. 2.4 [42]. Comme aucune paire d'images de chaque domaine n'est disponible pour l'apprentissage, cette méthode consiste à entraîner deux générateurs conditionnels, le premier traduisant une image du domaine $X$ à $Y$ et le second dans l'autre sens. Il est alors possible de reconstruire une image d'origine en appliquant les deux générateurs successivement et de mesurer la différence entre l'image d'origine et sa version reconstruite. Cette approche a révolutionné la synthèse I2I et de nombreuses évolutions ont été proposées, notamment pour l'imagerie médicale [76]-[78]. Les applications étendent celles de la synthèse I2I supervisée, avec par exemple l'harmonisation ou standardisation des centres, décrit dans la section 2.3.1.

Cependant, la notion de cycle suppose qu'une bijection parfaite existe entre les domaines, c'est-à-dire que chaque image d'un domaine possède un unique équivalent dans l'autre domaine. En pratique et notamment en imagerie médicale, les paramètres d'acquisition des images influent sur le contraste, le bruit et la résolution des images obtenues, tout en correspondant à une même image d'une autre modalité par exemple. Plutôt que la cohérence de cycle, un autre type d'approche utilise la représentation latente du contenu

et/ou du style de chaque domaine, comme proposé avec les modèles UNIT et MUNIT [79], [80]. Toutefois, en imagerie médicale, certaines modalités montrent du contenu proche avec des styles très variés, comme entre IRM et TDM par exemple. De plus, les sorties des réseaux sont moins diversifiées et peuvent ressembler à la "moyenne" du domaine cible. Enfin, ces modèles, tout comme CycleGAN, apprennent une application *mapping* global entre les intensités des deux domaines. Même si cela est acceptable en imagerie naturelle, cela signifie en imagerie médicale qu'on ne se concentre pas sur la préservation de l'anatomie ou sur celle des régions d'intérêts comme les tumeurs. Pourtant, ces dernières revêtent une importance capitale et sont généralement sous-représentées dans l'ensemble d'entraînement au vu de leur petite taille [81], [82]. Plus précisément, la proportion de chaque type d'image (par exemple saine et tumorale) modifie l'apparence des sorties, comme illustré sur la Fig. 2.5. Maintenir un équilibre entre ces proportions au sein des distributions source et cible devient alors crucial pour traduire avec précision ces régions. Toutefois, comme mentionné en section 1.3.1, cela nécessite des connaissances préalables de la composition des images de la distribution cible voire de l'ensemble de test, ce qui reste un problème ouvert en leur absence. Pour contrer cette limite, certains modèles d'I2I génèrent et segmentent directement l'image source, se rapprochant donc de la segmentation inter-modalité. Ces approches seront détaillées en section 2.5.

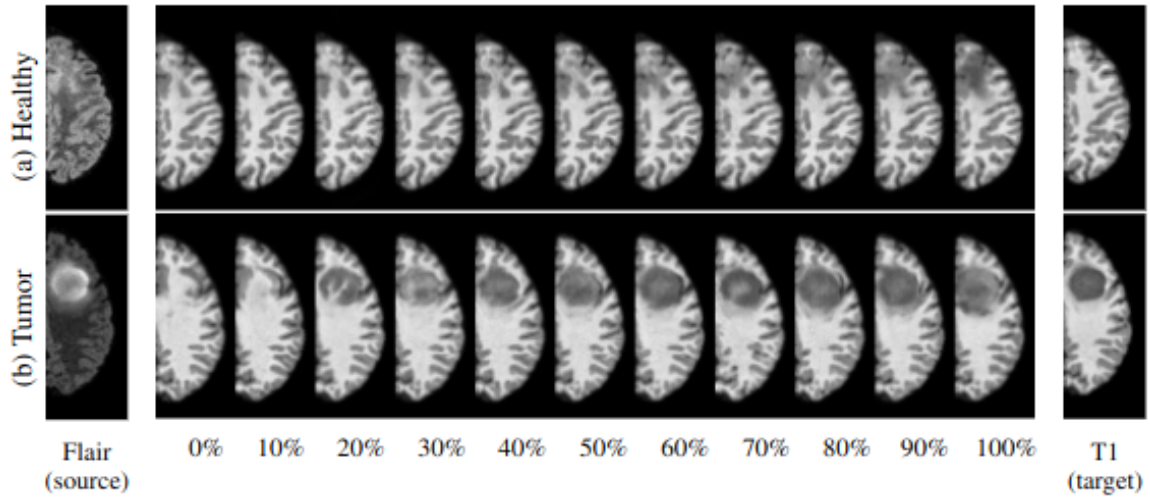



FIGURE 2.5 – Variation de l'apparence des images générées par CycleGAN entre IRM Flair et T1 en fonction de la proportion d'images tumorales dans le domaine d'arrivée, pour une image (a) saine et (b) tumorale. La proportion de ces images dans le domaine source est de 50%. Extrait de [82].

Bien que les GAN favorisent des sorties réalistes, de nouveaux paradigmes génératifs basés sur d'autres approches émergent également. En imagerie médicale, on observe par exemple de nouvelles approches basées sur l'apprentissage contrastif [83], [84], les architectures de *vision transformer* [85], [86], ou encore sur la diffusion stable [87], [88], chacune couplée ou non avec des fonctions de coût adversaires.

Toutefois, la plupart des méthodes existantes de synthèse I2I restent basées sur les GAN [38], [40]. En plus ou en conséquence des limites précédemment évoquées (notamment les problèmes de convergence), un aspect souvent négligé est la grande variabilité des résultats et la difficulté à reproduire les sorties [89], [90]. En effet, un même modèle entraîné plusieurs fois avec les mêmes données et paramètres de manière strictement identique peut ensuite générer, pour une même image, des textures et formes différentes. Pour contrer cela, il est courant de conserver le modèle le plus performant parmi plusieurs entraînements [91]. En parallèle, dans la configuration non supervisée, il existe peu de métrique reconnue pour l'évaluation de la qualité ou de la diversité des images générées par des GAN [91]. En pratique, les échantillons générés sont soit jugés qualitativement et donc subjectivement, soit employés pour une autre tâche en aval, en comparaison avec des données réelles afin de quantifier leur réalisme. Dans les deux cas, ces pratiques ne sont pas satisfaisantes en raison de leur nature non reproductible, c'est-à-dire de l'aléa induit par les ré-entraînements variables. Ainsi, dans le cas non supervisé, il reste particulièrement complexe d'évaluer à quel point les images générées sont proches des images réelles du domaine cible. Le décalage de domaine est potentiellement réduit mais toujours présent, ce qui peut encore affecter les performances des modèles en aval [81], [82].

## 2.2 Augmentation de données

L'augmentation de données, abordé en section 1.4.3, est une des approches les plus simples et efficaces pour améliorer la généralisation et la robustesse d'un modèle, peu importe le nombre de données disponibles. Pour cela, les images sont modifiées artificiellement afin d'étendre le jeu de données d'entraînement pour confronter le modèle à un éventail plus large de scénarios tout en couvrant différentes caractéristiques d'images, comme illustré schématiquement sur la Fig. 2.6. Un modèle peut ainsi être entraîné avec différentes versions d'une même image, ce qui l'encourage à prédire le même résultat indépendamment des perturbations appliquées. On dira alors qu'un modèle devient robuste ou invariant à certaines variations ou caractéristiques. Cela permet également de réduire

le sur-apprentissage.

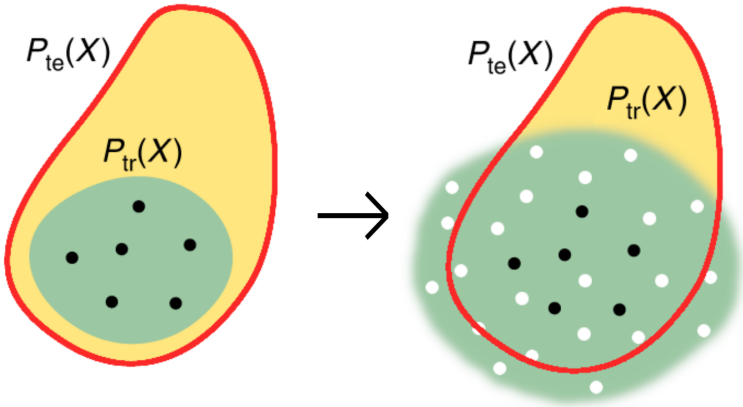



FIGURE 2.6 – Augmentation de données de manière schématique. L'augmentation de données vise à améliorer la généralisation, et par extension la robustesse d'un modèle, en générant des variations des images d'origines. Parmi les augmentations les plus conventionnelles, les images augmentées suivent très rarement la distribution de test. Inspiré de [4]

Selon les situations et en adaptant les paramètres d'augmentation, elles peuvent également encourager à réduire ou corriger un déséquilibre de classe [92]. Pour cette raison, de récentes démarches visent à développer les augmentations de données, notamment en imagerie médicale avec le challenge BraTs 2023, proposant une tâche d'augmentation à modèle fixé [1]. À l'inverse des approches traditionnelles où les participants proposent un modèle, cette initiative encourage à recentrer les recherches sur les données elles-mêmes afin de quantifier les potentiels gains obtenus grâce aux augmentations.

### 2.2.1 Augmentations sans apprentissage profond

Ces méthodes ont tout d'abord émergé pour l'imagerie naturelle, où les données disponibles présentent souvent une grande variabilité en termes d'éclairage, de fond, d'orientation, de bruit, etc. En enrichissant par exemple les données d'apprentissage avec des altérations d'intensité ou de flou, de nouvelles images sont générées de manière réaliste. Celles-ci représentent alors les variations potentielles rencontrées en conditions réelles. Parmi ces augmentations, on compte par exemple la rotation, l'effet miroir, les variations d'intensité ou de contraste, la correction gamma, l'ajout de bruit (gaussien ou autre),

1. `https://www.synapse.org/#!Synapse:syn51156910/wiki/622358`

l'ajout de flou, le changement de point de vue (zoom ou *crop*), les changements de résolution via interpolation, ou encore la dégradation de la résolution [14], [93]. La plupart de ces altérations ont un coût particulièrement faible, car elles sont appliquées via des opérations matricielles élémentaires très rapides sur carte graphique (GPU, *graphics processing unit*). Avec des paramètres aléatoires, elles permettent de générer un nombre infini de scénarios d'augmentations différents.

En plus de ces transformations élémentaires traditionnelles, l'augmentation *mix-up* a été particulièrement étudiée en raison de son très faible coût [94]-[96]. Elle consiste en la combinaison linéaire convexe de deux données réelles $(x_1, x_2)$ ainsi que leurs étiquettes $(y_1, y_2)$ selon un scalaire $\lambda \in [0, 1]$ :

$$\begin{aligned} x &= \lambda * x_1 + (1 - \lambda) * x_2 \\ y &= \lambda * y_1 + (1 - \lambda) * y_2 \end{aligned} \tag{2.3}$$

Les annotations ne sont alors plus binaires, encourageant le modèle à apprendre une fonction plus douce entre les étiquettes, ce qui peut conduire à une meilleure généralisation. Lors de la phase d'inférence sur image réelle, on applique généralement un seuil pour convertir ces prédictions continues en labels binaires.

D'autres méthodes proviennent également de l'imagerie naturelle, comme la déformation élastique. Cela consiste en une déformation sans contrainte, en comparaison avec les déformations affines traditionnelles [97], [98]. On peut ainsi modéliser facilement l'évolution d'une tumeur en la déformant et la grossissant artificiellement [99]. Sans effort particulier, les images générées peuvent manquer de réalisme [100]. Afin de s'assurer de la synthèse d'une image la plus cohérente et réaliste possible, l'application d'une fonction de déformation difféomorphique est à privilégier [101], [102]. De cette manière, la topologie de l'image est préservée tout en s'assurant de la continuité de l'image sans création d'artefacts. Ce type de déformation apparaît également dans plusieurs méthodes de recalage [103], [104]. Certains travaux ont également montré que des déformations non réalistes et intenses, modifiant beaucoup l'apparence des images d'origine, peuvent améliorer les performances, tout en améliorant la robustesse du modèle [105].

La dernière variation parfois employée en imagerie médicale est l'occlusion : une partie de l'image est remplacée par une valeur fixe d'intensité ou du bruit [106], [107]. Cela permet de rendre le réseau plus robuste aux occlusions ou aux artefacts cachant une partie de l'image, tout en encourageant le réseau à ne pas apprendre des motifs trop simplistes. On peut également décorréler la prédiction du réseau vis-à-vis de certaines zones, lorsqu'un

biais lié à une zone précise des images est présent au sein des données.

Même si l'ensemble de ces méthodes est globalement bénéfique pour les modèles, ces transformations sont simplistes et majoritairement issues de l'imagerie naturelle, ne représentant donc pas les variations traditionnellement observées en imagerie médicale [39]. Afin d'être plus spécifique aux contraintes et aux différentes modalités de l'imagerie médicales, des augmentations spécifiques ont été développées, souvent basées sur les GAN.

### 2.2.2 Augmentations basées GAN

Grâce à leur capacité générative, les GAN peuvent être entraînés sur des images médicales pour générer des échantillons adaptés à une modalité ou une pathologie spécifique [100], [108]-[111]. Ainsi, certaines méthodes utilisent les GAN pour de l'*inpainting* de patient ou de tumeur [64]. Cela consiste en l'occlusion d'une partie ou de l'image entière, avant l'emploi d'un modèle pour recréer cette partie de manière cohérente et réaliste. Il est alors possible de générer des tumeurs dans des images saines [112], illustré sur la Fig. 2.7, ou encore de re-synthétiser la tumeur d'une image vers une autre tout en ajustant l'anatomie de l'image cible [113]. Ces techniques diffèrent des méthodes de transfert de style, où l'on cherche à appliquer le style d'une image sur une autre sans modifier son contenu [114]. Une application est par exemple la simulation d'artefacts réalistes au sein des images d'entraînement [115], [116]. Dans ce cas précis, en plus de rendre le modèle plus robuste, cela peut permettre de résoudre un décalage de domaine dû à un plus grand nombre d'artefacts rencontrés en routine clinique qu'en entraînement.

De plus, les méthodes d'I2I peuvent également être utilisées pour l'augmentation de données [117], [118]. La synthèse I2I peut en effet être considérée comme du transfert de style, couramment employée pour l'augmentation de donnée. Ensuite, comme mentionné en section 2.1.2, l'intérêt de l'I2I en synthèse inter-modalité est souvent observée sur une tâche en aval telle que la segmentation, en raison du décalage de domaine ou du manque d'annotations par exemple. Dans ces cas, la génération d'images et leur utilisation pour l'entraînement du modèle en aval peut être considéré comme de l'augmentation de donnée.

### 2.2.3 Augmentations basées autre que GAN

Au-delà des nombreuses approches fondées sur les GAN, quelques autres méthodes existent. Parmi elles, les méthodes utilisant un atlas permettent d'établir des connaissances préalables sur la variabilité anatomique et tumorale des patients pour ensuite

générer des variations réalistes des images réelles [119]. On remarque également quelques méthodes de recalage employées pour l'augmentation de données [111]. Celles-ci ne permettent cependant pas de générer de nouvelles données réalistes, mais ont tout de même montré leur intérêt pour l'amélioration des performances et de la généralisation.

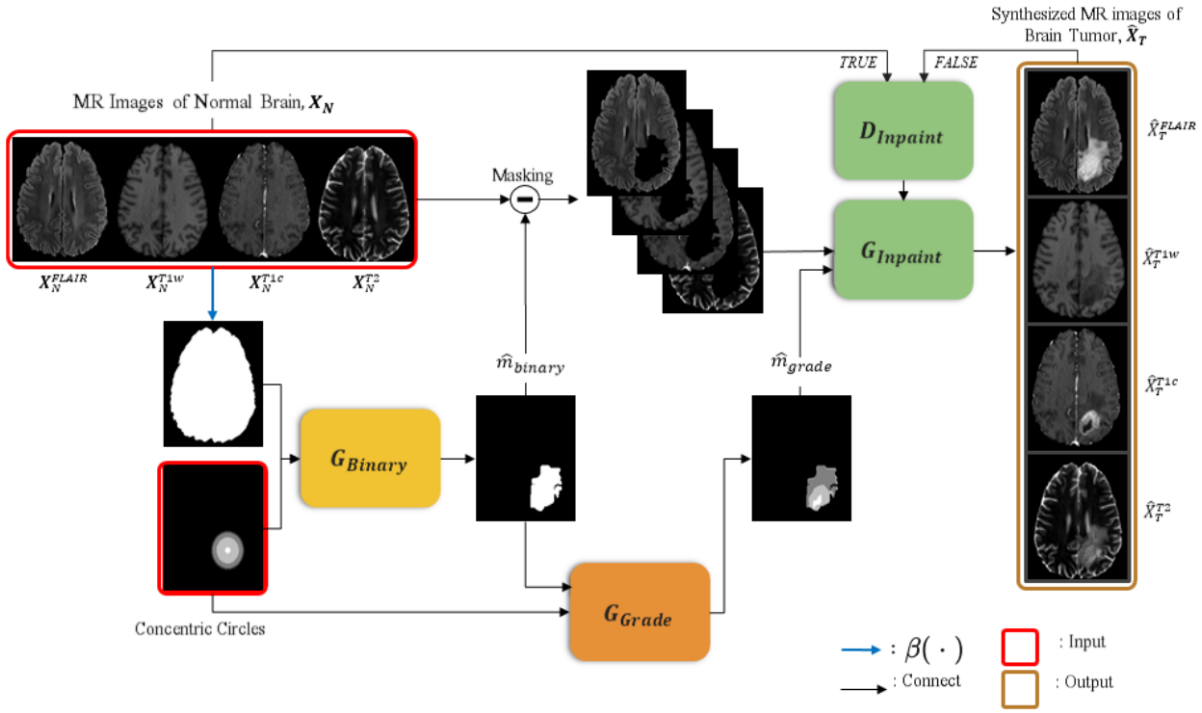



FIGURE 2.7 – Schéma illustrant une méthode d'inpainting de tumeur sur images saines. Plusieurs générateurs établissent la forme de l'œdème et de la tumeur nécrosée, avant d'occulter la partie correspondante de l'image et d'utiliser l'inpainting pour synthétiser la tumeur complète. Extrait de [112].

Dans cette même dynamique, certains se sont concentrés sur l'optimisation des techniques conventionnelles d'augmentation sans chercher à les adapter [120]-[122]. Ces méthodes, bien qu'optimale vis-à-vis des augmentations traditionnelles, nécessitent un temps d'entraînement particulièrement élevé, en raison du nombre de configurations à tester, et ce malgré les efforts pour diminuer ce temps. Ensuite, la différence entre deux configurations peut être très minime et non significative. De plus, la mise à jour des poids d'un modèle avec de nouvelles données sous-entend d'optimiser à nouveau ces augmentations. À l'inverse, d'autres choisissent de déterminer une fois pour toutes les paramètres des augmentations conventionnelles à utiliser, puis de fixer ces paramètres. C'est le choix effectué par nnU-Net comme présenté en section 1.4 [21]. Vis-à-vis des performances de

la seconde approche à de nombreux challenges et du coût calculatoire de la première, la seconde approche semble être à privilégier.

### 2.2.4 Limites

L'ensemble des méthodes décrites permettent d'augmenter les performances, ou encore, dans certains cas précis, de réduire le nombre de données nécessaires à l'entraînement sans dégrader les résultats, voire en les améliorant [123]. Cependant, les approches n'utilisant pas d'intelligence artificielle montre une amélioration limitée des performances, tandis que celles basées DL reposent généralement sur de grands ensembles de données d'entraînement pour permettre au générateur et au discriminateur d'apprendre la diversité de la distribution des données. Lorsque ces données sont disponibles, un modèle entraîné sans augmentation coûteuse peut déjà être suffisamment performant. Dans ce cas, l'utilisation de ces augmentations, si elle est bénéfique, permet une faible amélioration, pas nécessairement significative. À l'inverse, dans les scénarios où peu de données sont disponibles, ces techniques manquent généralement de robustesse. De plus, l'ensemble des limites induites par les GAN, comme l'effondrement des modes ou la perte de stabilité, restent présentes dans les augmentations associées. Enfin, en présence d'un décalage de données, la plupart de ces méthodes s'effondreront, car elles visent à générer des images suivant la distribution des données d'entraînement. Si la distribution des données de test ou des données obtenues en utilisation clinique diffèrent, l'augmentation écarte davantage les deux distributions.

## 2.3 Autres domaines assimilés à l'adaptation de domaines

Même si nos travaux se concentrent principalement sur la synthèse d'images pour l'augmentation de données, d'autres stratégies d'adaptation de domaine ont été proposées par la communauté. Par souci d'exhaustivité, nous présentons ici les paradigmes et les méthodes en généralisation de domaines (DG, *domain generalization*), harmonisation de données, apprentissage par transfert (TL, *transfer learning*) ainsi que leurs limites. Nous présentons également l'auto-supervision, qui sera utilisé dans ce travail.

### 2.3.1 Généralisation de domaines et harmonisation de données

La DG vise à former un modèle sur un ou plusieurs domaines de manière à ce qu'il puisse fonctionner efficacement sur des domaines non observés pendant la phase d'entraînement, comme illustré sur la Fig. 2.8. Le principal enjeu est alors d'assurer que les modèles soient suffisamment flexibles pour s'adapter à de nouvelles situations, tout en évitant d'être trop spécifiques aux données initiales. Pour cela, il doit apprendre à extraire des caractéristiques propres au contenu des images, et ce indépendamment du style de celles-ci. Sa capacité à bien généraliser à de nouveaux domaines non observés dépend en grande partie de sa résilience face au décalage de domaines. Lorsque ce décalage est significatif entre le domaine source et le domaine cible, il peut compromettre la performance du modèle, rendant ainsi la DG particulièrement difficile. À l'inverse, une bonne gestion

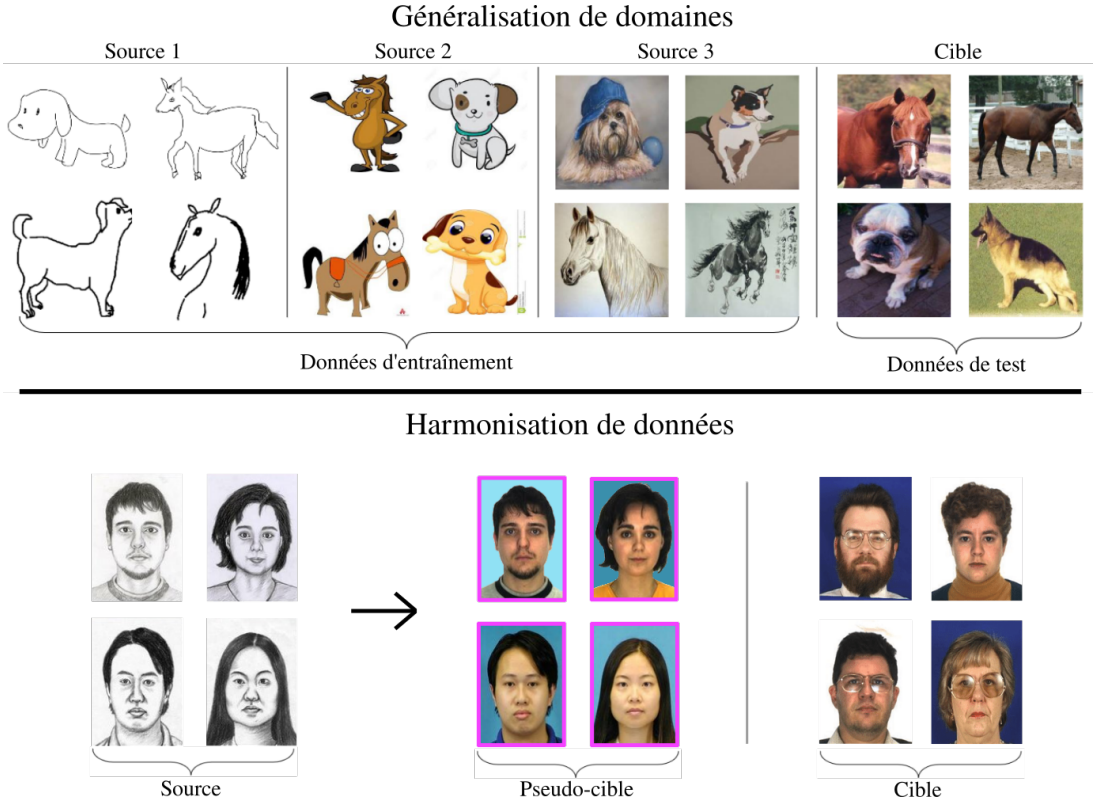



FIGURE 2.8 – Cas typiques de généralisation de domaines et d'harmonisation de données. Dans le second cas, les images "pseudo-cibles" représentent les images générées par le modèle d'harmonisation. Échantillons extraits des jeux de données PACS et CHUK, inspiré de [124].

du décalage de domaines favorise une généralisation plus robuste, permettant au modèle

de maintenir sa performance malgré les différences entre les distributions. Ainsi, ce type d'approche est à évaluer et à adapter au cas par cas, afin de s'assurer de l'efficacité transversale du modèle sur les différents domaines. Dans les contextes où la DG échoue, il est possible d'employer des modèles d'harmonisation de données. On cherche alors à rendre les différents ensembles de données plus proches et ressemblants, c'est-à-dire à rapprocher les distributions, afin de considérer les données de chaque distribution comme similaires et non hétérogènes. L'enjeu central est d'assurer que les variations inhérentes à chaque source de données n'introduisent pas de biais lors de l'entraînement du modèle. Cette étape est cruciale, notamment dans les contextes où le modèle n'arrive pas à généraliser aux domaines, par exemple dû à l'origine multi-centrique des données ou à leur faible nombre dans chaque sous-ensemble.

En DG, on distingue deux situations : lorsque le modèle est entraîné sur un ou plusieurs domaines sources. Le premier cas revient à rendre le modèle robuste aux *outliers*, aux images suivant une autre distribution que celle vue à l'entraînement. Pour cela, on peut par exemple utiliser de l'augmentation de données avec des fonctions de coût de reconstruction de type norme $\ell_1$ ou $\ell_2$. Dans le second cas, un plus grand nombre de méthodes existe [124], [125]. En imagerie médicale, certaines cherchent à aligner les caractéristiques extraites par un encodeur au sein de l'espace latent en fonction du résultat à prédire [126]. Il est alors possible d'utiliser des techniques de démêlage des caractéristiques (*feature disentanglement*) pour séparer les différentes informations au sein des caractéristiques latentes. D'autres méthodes visent à utiliser des notions de méta-apprentissage en considérant un des domaines sources comme étant le jeu de méta-test et les autres comme les jeux de méta-entraînement, afin de simuler le décalage de domaine [127]. Au vu de la grande variété d'approches possibles, le.a lecteur.trice intéressé.e peut s'intéresser aux différentes *reviews* du sujet [124], [125]. Cependant, les performances de ce type d'approches en imagerie médicale sont encore particulièrement limitées. Les domaines annotés sont généralement rassemblés pour créer un seul ensemble d'entraînement hétérogène, sans tenir compte des écarts potentiels entre les domaines. De manière assez surprenante, des résultats équivalents à l'état de l'art ont parfois été observés lorsqu'un grand nombre d'images sont disponibles à l'entraînement [21], [126], [128]. Avec peu de données dans chaque domaine, comme c'est le cas lorsque quelques données issues de plusieurs centres sont rassemblées, ces approches montrent des résultats encore assez décevants comparés à un modèle considérant toutes les données comme homogènes.

Il est également possible d'aligner non pas les caractéristiques latentes des images, mais

directement leur représentation dans leur espace d'origine, ce qui définit l'harmonisation de données. La standardisation ou la normalisation peuvent déjà être vu comme des méthodes élémentaires d'harmonisation, en unifiant la représentation mathématique des données, par exemple en les rendant respectivement centrées et réduites dans le cas de la standardisation *z-score*. Ces méthodes simplistes ne réduisent cependant pas les écarts de distributions trop importants, comme ceux typiques de l'imagerie médicale. Pour cette raison, d'autres techniques ont été développées, par exemple des approches statistiques comme ComBat [129]-[131], ou encore basées apprentissage profond [132], [133]. Parmi celles-ci, la plupart se basent sur des modèles I2I pour du transfert de style, et donc de domaines [134]-[137].

Bien que la DG offre des solutions prometteuses pour traiter les disparités inhérentes à l'imagerie médicale, celles-ci doivent être mises en œuvre avec précaution pour garantir la fiabilité et la pertinence des résultats obtenus. Par exemple, les méthodes de DG supposent que chaque domaine est homogène et équilibré. Cette hypothèse est souvent peu respectée en imagerie médicale, y compris au sein d'une même modalité, comme vu en section 1.3.1 et 1.3.3 à cause de la représentativité des données et de la multitude d'origines des décalages de domaines. L'harmonisation des données peut alors atténuer ces variations et rendre les domaines sources plus homogènes, mais cela engendre parfois une perte d'informations spécifiques à un domaine ou l'altération d'un autre. De plus, un nombre suffisant (et souvent important) d'images dans chaque domaine est nécessaire pour pouvoir caractériser le style du domaine afin d'en extraire le contenu uniquement, en DG comme en harmonisation de données [124], [125]. Lorsque ce nombre est trop faible, on se tourne généralement vers le TL ou les modèles performant avec peu de données.

### 2.3.2 Apprentissage par transfert et auto-supervision

Le TL vise à geler puis réutiliser un modèle déjà entraîné (ou une partie de celui-ci) pour une certaine tâche et à l'appliquer à une tâche différente. On distingue alors deux approches : avec et sans ajustement fin des paramètres (ou *fine-tuning*) [139]. Cette technique consiste à utiliser les poids du modèle entraîné comme initialisation avant un nouvel entraînement pour la tâche voulue. Ces poids étant déjà performants selon un ancien objectif, on suppose que cette initialisation est plus proche des poids finaux qu'une initialisation classique (fixée ou aléatoire par exemple). Cela permet alors de bénéficier de cet ancien entraînement, en affinant les poids pendant moins d'époques qu'un entraînement complet. Ce type d'approche est également employé dans les contextes où peu

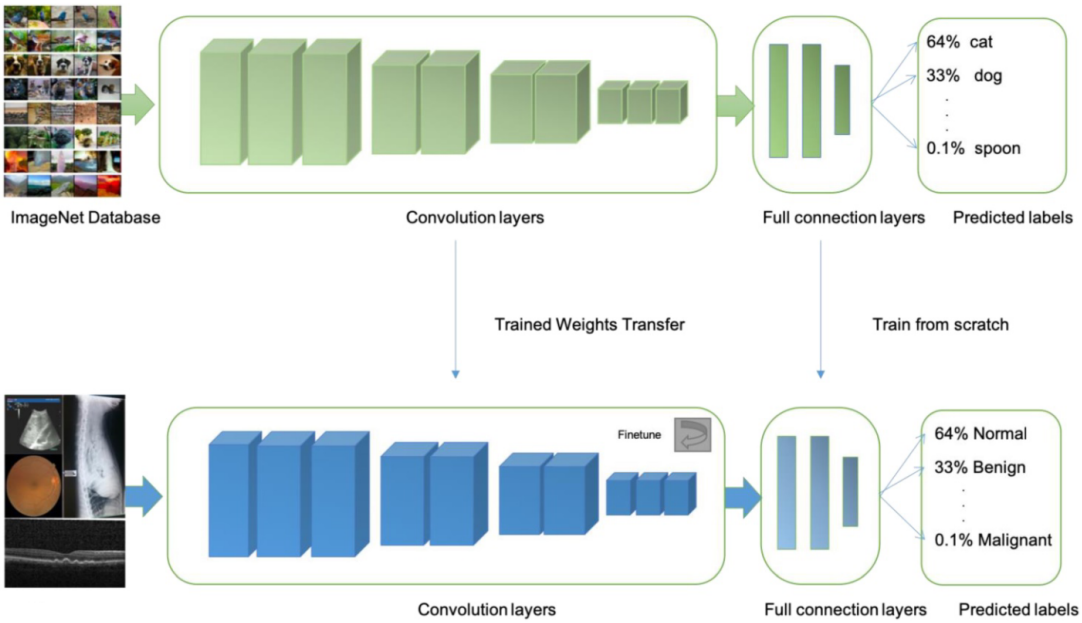



FIGURE 2.9 – Schéma illustrant une méthode de TL. Après l'apprentissage d'un CNN sur ImageNet, tous les poids sont gelés, à l'exception des dernières couches *fully-connnected* qui subissent un*fine-tuning* sur les données de la tâche cible. Extrait de [138].

de données sont disponibles. En effet, si les deux tâches sont suffisamment proches, un *fine-tuning* avec ces quelques données permet parfois d'obtenir un modèle performant et robuste sans risque de sur-apprentissage [139]. Il est également possible de réutiliser une partie d'un modèle, typiquement un encodeur, sans affiner ses poids et en les gardant fixés. L'encodeur est alors employé comme extracteur de caractéristiques pour un second modèle, un classificateur par exemple, ou pour définir une fonction de coût ou une métrique pour une grande variété de tâches, comme la *Fréchet inception distance* [140]. Ainsi, avec ou sans ajustement fin, des encodeurs entraînés sur de gros jeux de données, typiquement issus de l'imagerie naturelle comme ImageNet [141], peuvent être employé dans l'imagerie médicale [142], comme illustré en Fig. 2.9.

Que ces poids soient gelés ou non, il existe un grand nombre d'approches permettant un transfert. D'une manière plus générale, on parle de transfert de connaissance (*knowledge transfer*), même si les deux expressions sont confondues par abus de langage. Parmi elles, on trouve par exemple le paradigme enseignant/élève, initialement proposé pour la distillation de connaissance (*knowledge distillation*) afin de réduire la taille d'un réseau en entraînant un petit modèle et en bénéficiant d'un modèle plus grand initialement entraîné

[143]. Cette approche consiste à entraîner un modèle dit "enseignant", à le tester sur des données non-annotées, puis à utiliser ces prédictions pour le nouvel entraînement d'un modèle dit "élève". L'élève bénéficie alors des connaissances apprises en amont, et devient à son tour enseignant pour un nouvel entraînement d'un élève. Cette technique, appelée auto-entraînement (*self-training*), permet d'améliorer itérativement les performances. On peut également adapter ce paradigme directement aux données de test, dans le cadre d'un apprentissage transductif (*transductive learning*). Cependant, si les labels prédits par l'enseignant sont trop bruités voire complètement faux, l'erreur sera potentiellement répétée par l'élève, détériorant alors les performances à chaque itération [144]. Certains proposent alors de considérer directement les masques comme inexacts ou encore d'affecter un score de pertinence aux prédictions, afin d'exclure les plus bruitées d'entre elles [17], [145], [146].

Par extension, l'auto-entraînement a été élargi à l'apprentissage non-supervisé ; on parle alors d'auto-supervision (*self-supervision* ou *self-supervised learning*) [147], [148]. Avant l'entraînement, des pseudo-labels pour les données non annotées sont générées, permettant au modèle "enseignant" de s'entraîner à les prédire. Ainsi, cet entraînement préalable peut bénéficier au modèle "élève", qui apprendra la tâche cible. Cette approche se développe particulièrement récemment en imagerie médicale, notamment afin d'exploiter pleinement les données non annotées disponibles, en apprentissage semi-supervisé par exemple [149], [150].

Assez récemment également, de nouveaux modèles entraînés sur des millions d'images naturelles et des milliards de masques visent à segmenter n'importe quelle image sans aucun *fine-tuning*, comme le modèle SAM (*Segment anything model*) [151]. Cette tâche, appelée *zero-shot transfer*, est permise grâce à la quantité d'images variées vues à l'apprentissage. Même si son utilisation directe pour l'imagerie médicale semble limitée, comme illustré sur la Fig. 2.10, cette initiative montre qu'avec suffisamment d'exemples variés, un même modèle peut segmenter avec pertinence des images très différentes, y compris en imagerie médicale [105]. Cela ouvre donc la voie à la création de jeux de données larges, multi-modaux et multi-pathologies [152], [153].

D'une manière générale, les approches de TL sont sujettes au manque de cohérence entre images naturelles et médicales. En effet, l'écart entre ces deux ensembles étant conséquent, le bénéfice du premier entraînement est souvent limité voire complètement absent [155], [156]. Même sur des tâches similaires en imagerie naturelle, le TL montre parfois des performances plus basses qu'un ré-entraînement complet, avec des scores plus élevés et moins variables en initialisant les poids de manière aléatoire [147], [157]. D'un point

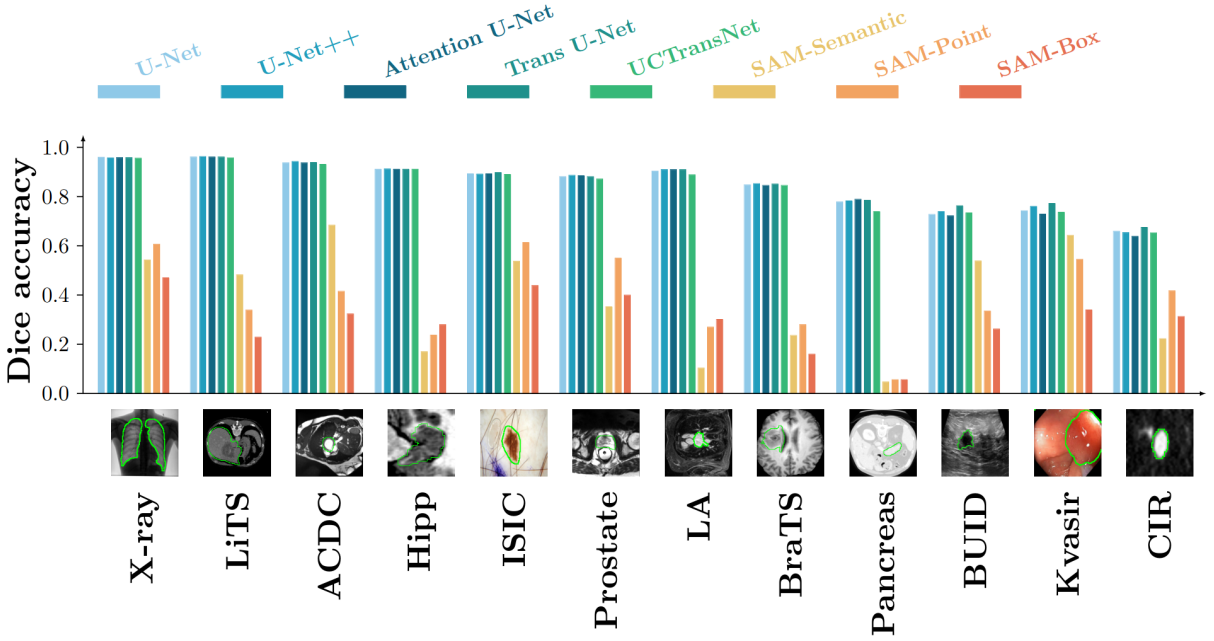



FIGURE 2.10 – Scores de segmentation entre des approches traditionnelles et SAM dans diverses modalités et applications. Extrait de [154].

de vue mathématique, cela est probablement dû à un minimum local atteint proche des poids du transfert avec un taux d'apprentissage (*learning rate*) trop faible. De plus, le personnel hospitalier peut percevoir ces modèles avec scepticisme, car il est intuitif de considérer qu'un réseau formé pour reconnaître des caractéristiques d'imagerie naturelle pourrait mal s'adapter à la complexité et à la spécificité de l'imagerie médicale. Ces approches peuvent donc entraver la confiance des professionnels de santé envers les résultats fournis par ces modèles.

En l'absence d'annotations, l'auto-supervision permet également d'apprendre une représentation des données dans un espace sémantique pour initialiser des poids avant un entraînement avec un objectif défini [158], [159]. Toutefois, ce type d'approches étant un sous-cas du TL, elles nécessitent souvent un grand nombre de données pour le pré-entraînement.

## 2.4 Entraînement avec peu de données

De manière transverse à plusieurs champs de la vision par ordinateur, dont l'adaptation de domaine, de nouveaux paradigmes ont émergé pour traiter les situations où

le nombre de données d'entraînement est limité. En effet, en imagerie médicale et par exemple pour les maladies rares, il est souvent complexe d'obtenir de gros jeux de données. Ainsi, les méthodes dites *few-shot* se concentrent sur quelques images uniquement pour apprendre des motifs généralisables parmi celles-ci. Même si les performances sont souvent moins élevées qu'un modèle plus traditionnel entraîné avec plus de données, ce champ de recherche est de plus en plus étudié pour répondre au manque de données.

Le risque principal est naturellement de sur-apprendre de ces quelques données. Certaines techniques permettent de réduire ce risque, par exemple en diminuant le champ réceptif du modèle (*receptive field*) [160]. Il s'agit de la taille de la région d'entrée qui affecte la sortie d'un neurone donné. Le champ réceptif d'un encodeur représente donc le nombre de pixels pris en compte dans le calcul d'une *feature* de sortie. Si ce champ est grand, il capturera beaucoup de contexte dans l'image d'entrée, permettant d'identifier des structures plus complexes. Les détails fins seront donc au fur et à mesure agrégés, réduisant leur importance et augmentant le risque de sur-apprentissage. À l'inverse, un petit champ réceptif permet de prendre en compte des détails plus fins, plus localement et donc avec moins de contexte, diminuant ce risque. Il est possible de réduire ce champ de plusieurs manières : en diminuant la profondeur du réseau, en augmentant le pas de convolution (*stride*), ou encore en réduisant la taille des noyaux de convolution [160]. Ainsi, la capacité d'apprentissage du modèle est réduite, limitant le sur-apprentissage. Si le champ réceptif est trop réduit, le modèle risque cependant de sous-apprendre.

L'apprentissage *few-shot* a été particulièrement étudié, principalement en imagerie naturelle, et parfois adapté pour l'imagerie médicale [161]-[164]. Les modèles proposées exploitent différents paradigmes afin d'apprendre à partir d'un nombre très restreints d'images. Parmi eux, on compte le méta-apprentissage et le TL, décrite en section précédente, ou encore les GAN via diverses méthodologies. En effet, les principes de méta-apprentissage sont approfondis et particulièrement exploités en contexte *few-shot*, y compris en imagerie médicale [165]. On retrouve, par exemple, des modèles prototypiques (*prototypical networks*) visant à regrouper les images dans l'espace latent [166], ou bien un méta-apprentissage indépendant du modèle et simulant des décalages de données [126], [167], [168]. En TL, les approches se concentrent sur l'apprentissage de caractéristiques génériques et leur représentation sur un jeu de données plus grand et diversifié, avant d'affiner les paramètres pour le petit jeu de données étudié.

Ensuite, de nombreux modèles *few-shot* voire *one-shot* (c'est-à-dire apprenant à partir d'une seule image) sont conçus avec des GAN et leurs évolutions, employant ou non

quelques techniques du TL. Parmi les modèles *one-shot* les plus connus, on trouve SinGAN [169] ou encore One-shot GAN [170]. Le premier apprend à reconstruire l'unique image


FIGURE 2.11 – Exemples d'applications de SinGAN en imagerie naturelle : génération aléatoire, *paint to image*, retouche, harmonisation, super-résolution, animation. Extrait de [169].

d'entraînement à différentes échelles, en se concentrant sur le style de l'image. Pour cela, plusieurs GAN sont entraînés successivement à chaque échelle, leur permettant d'être utilisé indépendamment les uns des autres. Ainsi, avec un unique entraînement sur une seule image, SinGAN peut réaliser différentes tâches illustrées sur la Fig. 2.11, comme la génération d'images aléatoire, le *paint-to-image* ou encore la super-résolution. Cela rend ce modèle très attractif à bas régime de données mais pas uniquement. SinGAN sera détaillé plus en détail dans le chapitre 3. Le second, One-shot GAN, vise à dissocier contenu et réalisme global (incluant la disposition du contenu) au sein du discriminateur. En rajoutant une régularisation sur le bruit fourni en entrée, la variabilité du générateur est

améliorée. Celui-ci n'a pour le moment pas été employé dans un contexte médical. En plus de la réduction du champ réceptif, les autres techniques et approches varient énormément, utilisant par exemple les *a priori* de formes au sein de l'imagerie médicale pour employer des techniques de recalage [119], ou encore plus récemment, des transformers pour de la diffusion *one-shot* [171]. Des évolutions de SinGAN ont également été proposées [67], [172]-[174]. Cependant, ces méthodes n'ont, à notre connaissance, été employées qu'une fois en imagerie médicale, pour de l'augmentation de données en entraînant un modèle SinGAN par image [175]. En plus d'être particulièrement coûteux en ressources et clairement sous-optimal dans la mesure où les connaissances apprises sur une image ne sont pas utilisées pour les autres, ce type d'approche semble être à bannir.

## 2.5 Applications types

Les méthodes d'adaptation de domaine trouvent plusieurs applications en imagerie médicale, notamment pour réduire les écarts présents entre les conditions d'entraînement et de déploiement. Au-delà des champs de recherche précédemment décrits, plusieurs contextes précis ont émergé, en fonction de la disponibilité des annotations au sein des données.

Tout d'abord, on parle d'adaptation de domaine supervisée lorsque les deux domaines source et cible sont annotés, généralement avec peu de données disponibles dans le domaine cible. Les deux domaines partageant des similarités de contenu, l'utilisation des données sources en plus des données cibles lors de l'entraînement peut permettre d'obtenir de meilleures performances qu'un modèle entraîné directement sur la modalité cible. Cependant, des labels sont nécessaires pour les deux ensembles de données, ce qui est particulièrement coûteux.

Ensuite, l'adaptation de domain non-supervisée (UDA, *unsupervised domain adaptation)*) représente la situation où les données sources et cibles sont respectivement annotées et non-annotées. Il est donc impossible d'entraîner un modèle directement sur les données cibles, et l'utilisation des données sources dans le processus d'apprentissage est nécessaire. Ce cas est plus général que le précédent car il nécessite moins d'annotations, mais il en devient plus complexe. L'UDA diffère de la DG car l'apprentissage est spécifique aux domaines source et cible, alors que seul le (ou les) domaines sources sont nécessaires en DG, peu importe le domaine cible [124]. Les méthodes de synthèse I2I permettent alors d'obtenir des données pseudo-cibles plus proches du vrai domaine cible que les données

sources.

Enfin, l'adaptation de domaine semi-supervisée est un contexte intermédiaire, où une partie seulement des données cibles sont annotées. La quantité de labels dans le domaine cible est généralement limitée, marquant la différence avec l'adaptation de domaine supervisée. Il devient alors indispensable de prendre en compte les données non annotées lors du processus d'entraînement.

En fonction de l'écart entre les domaines, certains termes supplémentaires sont employés. Plus particulièrement en imagerie médicale, lorsque les domaines représentent différentes modalités, par exemple en segmentation entre IRM et TDM, on parlera de segmentation inter-modalité ou segmentation inter-modale (*cross-modality segmentation* ou *cross-modal segmentation*).

Les algorithmes d'adaptation de domaine supervisée peuvent être vus comme un cas simple de DG, avec deux domaines sources et un domaine cible suivant la même distribution qu'un des domaines sources. Ainsi, les techniques mentionnées en section 2.3.1 sont également applicables, notamment à l'aide de synthèse I2I supervisée ou non [40], [176]. Le domaine cible étant connu, il est également possible d'utiliser le TL avec *fine-tuning*, comme décrit en section 2.3.2.

Une plus grande variété d'approches ont été proposées pour l'UDA en imagerie médicale grâce au grand nombre de cas d'applications possibles de ces méthodes, mélangeant des principes issus de tous les sous-domaines mentionnés dans ce chapitre [38], [40]. Ainsi, les méthodologies types consistent à aligner les caractéristiques extraites des images [177], ou directement les images via des modèles de synthèse I2I [178], voire les deux [179], avant d'apprendre la tâche cible en aval. Pour encourager la préservation de certaines structures d'intérêt, plusieurs travaux proposent de combiner les encodeurs-décodeurs utilisés dans les modèles d'I2I avec un segmenteur supplémentaire [50], [179]-[185]. Même pour la segmentation, un second modèle dédié à la tâche cible est souvent entraîné à l'aide des données pseudo-cibles générées, généralement appelées "pseudo-images", alignées aux masques du domaine source. Celui-ci peut être entraîné en même temps ou séparément des autres poids. Ensuite, l'auto-entraînement est également utilisé, afin de bénéficier, lors de l'apprentissage de la tâche cible, à la fois des pseudo-images associées aux vrais masques et des images cibles réelles avec des "pseudo-masques" générés par le premier modèle [38], [186]. Les différentes méthodes présentent cependant les mêmes défauts que les méthodes d'I2I, à savoir que les résultats sont souvent variables d'un apprentissage à l'autre, étant donné la difficulté de la tâche. En parallèle, un décalage de données peut

persister entre données réelles et générées. De plus, les limites habituelles sont présentes, notamment celles liées à la quantité de données homogènes nécessaires et à la qualité des annotations. Néanmoins, les algorithmes d'UDA nécessitent des annotations dans un seul domaine seulement, les rendant plus accessibles pour un grand nombre de situations.

## 2.6 Conclusion

Dans ce chapitre, nous avons fait un état de l'art couvrant différents aspects de l'adaptation de domaine. Un certain nombre d'approches ont montré leur intérêt en amont d'un modèle de segmentation tumorale. Parmi elles, l'utilisation de GAN permet de générer des images réalistes, en présence de décalage de données ou non, via les modèles de synthèse I2I. Ces images bénéficient alors à la segmentation mono-modale en tant qu'augmentation de données spécifique, ou pour l'UDA face à une situation inter-modalité.

Plusieurs précautions doivent néanmoins être prises. L'apprentissage des GAN est assez instable, avec plusieurs biais identifiés comme la variabilité des résultats en sortie de réseau ou l'effondrement des modes. Par ailleurs, les modèles de synthèse I2I nécessitent un grand nombre de données d'entraînement pour généraliser correctement. Pour ces raisons, un décalage de données peut alors persister au sein d'une modalité entre images réelles et générées, notamment au niveau des tumeurs, limitant les performances de la segmentation en aval.

Dans le chapitre suivant, nous décrivons notre contribution méthodologique consistant à exploiter une méthode *one-shot* d'augmentation de données basée sur un modèle GAN auto-supervisé, visant à élargir la distribution de tumeurs couvertes pour mieux prévenir le décalage de domaine lors du déploiement.

Chapitre 3

# Augmentation par altération de contraste et harmonisation des régions d'intérêt

**Résumé**

Ce chapitre présente une nouvelle méthode d'augmentation de données pour la segmentation, axée sur l'altération des régions d'intérêt (ROI, *regions of interest*). Nous favorisons l'adéquation de l'apparence des ROI du domaine source d'entraînement avec celle du domaine cible de déploiement.

Notre contribution est double. D'une part, nous proposons une stratégie de diversification du contraste des ROI, par des transformations linéaires tenant compte des particularités du domaine cible. Nous qualifions cette approche de "naïve", car elle ne permet aucune harmonisation réaliste de l'apparence vis-à-vis du domaine cible. D'autre part, nous proposons une harmonisation réaliste de ces ROI altérées par l'emploi d'un modèle génératif de type SinGAN, entraîné dans le domaine cible. Nous nommons cette méthode augmentation par mélange génératif (GBA, *generative blending augmentation*). Nous évaluons chacun de ces deux aspects de manière qualitative pour dégager les grands principes qui guident les choix méthodologiques exposés dans le chapitre suivant, consacré à la solution proposée pour la segmentation de tumeurs cross-modale en IRM.

Nous avons vu que le décalage de domaine nuit à la généralisation et la robustesse des modèles de segmentation. En particulier, les différences entre les ROI observées en phase de développement et en phase de déploiement risquent d'entraîner des pertes de performance [187]. Dans cette situation, le manque de variabilité des ROI visibles à l'entraînement, ne couvrant alors qu'une partie de leur distribution globale, entraîne une sous-représentation de certaines caractéristiques de ROI. La correction de cette distribution en augmentant artificiellement les ROI serait donc bénéfique, afin de l'élargir, voire de la rapprocher de celle des ROI rencontrées en contexte clinique. C'est la raison pour laquelle nous proposons une méthode d'augmentation de données basée sur l'estimation de la distribution des ROI en phase de déploiement. Elle diversifie le contraste des ROI afin de réduire le décalage de distribution par rapport aux données d'apprentissage, et les intègre avec réalisme dans l'image d'origine grâce à un modèle d'harmonisation. Les détails d'implémentations sont précisés dans un dépôt GitHub dédié [1].

## 3.1 Méthode

La méthode proposée est découpée en deux étapes : altération linéaire du contraste des ROI par mise à l'échelle linéaire de l'intensité des voxels, puis harmonisation réaliste de l'image altérée avec un modèle génératif SinGAN. Nous considérons que la première étape constitue une première méthode d'augmentation de données dite "naïve". La méthode complète, appelée augmentation par mélange génératif (GBA, *generative blending augmentation*), est une extension de la version naïve.

### 3.1.1 Méthode naïve

Soient $x$ l'image du jeu d'entraînement contenant une ROI dont on souhaite apprendre la segmentation, et $y$ son masque de segmentation. Pour altérer le contraste d'une ROI une fois l'image normalisée dans $[0, 1]$, il suffit de sélectionner un paramètre scalaire $\lambda$ et de multiplier l'intensité des voxels de ROI avec cette valeur :

$$x_\lambda = \begin{cases} x & \text{en dehors de } y, \\ \lambda \cdot x & \text{au sein de } y \end{cases} \tag{3.1}$$

Ainsi, si $\lambda < 1$, le signal de la ROI est artificiellement abaissé et devient davantage

1. `https://github.com/GuigzoS/crossmodalGBA`

hyposignal, c'est-à-dire plus sombre, par rapport aux tissus environnants. À l'inverse, $\lambda > 1$ la rend davantage hypersignal, c'est-à-dire plus claire. En choisissant soigneusement les images à augmenter et les valeurs $\lambda$, on peut réduire un décalage de domaine entre diverses caractéristiques de ROI afin de s'adapter aux particularités du domaine cible.

Avec cette augmentation, il devient possible de diversifier le jeu d'entraînement de manière naïve, dans l'objectif de rendre le modèle plus robuste aux variations d'intensité des ROI. Cependant, cette augmentation relativement simpliste introduit un changement brusque de contraste au niveau des contours des ROI altérées.

### 3.1.2 Modèle SinGAN et méthode profonde

Afin de rendre les contours plus réalistes, mais aussi pour varier la texture et le contenu de la ROI, un modèle d'harmonisation est ensuite employé. Nous avons choisi le modèle *one-shot* SinGAN pour sa capacité à apprendre le style d'un domaine cible à partir d'une unique image [169]. Son utilisation pour l'harmonisation de ROI définit alors la version profonde de notre augmentation, GBA, applicable à tout régime de données grâce à son aspect *one-shot*.

SinGAN est un modèle 2D multi-échelle. Le réseau se caractérise par sa structure pyramidale, composée d'une série de $K+1$ GAN $(G_k, D_k)_{0 \leq k \leq K}$ où chacun apprend la distribution de l'image à une échelle spécifique. L'échelle $k$ correspond à $k$ sous-échantillonnages successifs par un facteur $r$ ($r < 1$ ; notée $\downarrow_{r^k}$). Ainsi, $K$ fait référence à l'échelle la plus basse, et 0 à l'échelle d'origine. Avant l'apprentissage, des sous-images (généralement appelées patchs ou imagettes) sont extraites à chaque niveau, de 0 à $K$.

L'entraînement de SinGAN, illustré en Fig. 3.1, débute à l'échelle la plus basse $K$ avec un GAN non-conditionnel $(G_K, D_K)$ prenant en entrée du bruit gaussien et apprenant la composition de l'image à cette très faible résolution, de l'ordre de $25 \times 25$ pixels. $D_K$ vise alors à reconnaître si un patch de cette échelle est généré ou issu de l'image d'origine. Une fois ce niveau entraîné, les sorties sont sur-échantillonnées par $r$ pour atteindre l'échelle associée à l'entier $K-1$ avant d'être concaténées avec du bruit pour l'entraînement du GAN conditionnel $(G_{K-1}, D_{K-1})$. À cette échelle, comme à celles $k \in [0, K-1]$, le générateur $G_k$ apprend à ajouter les détails typiques de la résolution $k$ afin que ses sorties suivent la distribution de l'échelle $k$ pour tromper le discriminateur $D_k$. Ainsi, avec un sur-échantillonnage par $r$ (noté $\uparrow_r$) entre chaque niveau pour que les résolutions correspondent à chaque niveau $k$, le réseau apprend à reconstruire l'image d'origine en rajoutant itérativement les détails de plus en plus fins typiques de chaque échelle.

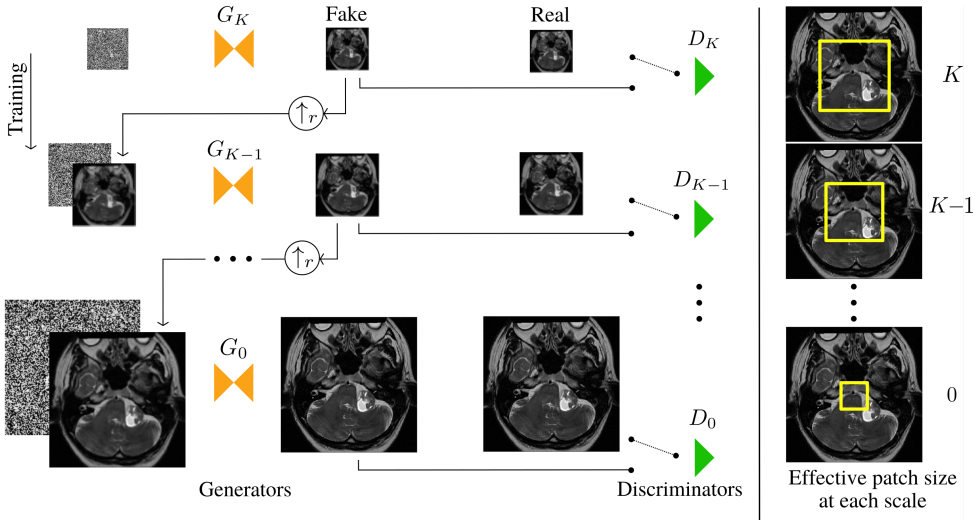



FIGURE 3.1 – Apprentissage du modèle SinGAN. Adapté de [169].

L'architecture des différents $(G_k, D_k)$ est identique, avec deux patchGAN composés de cinq blocs convolutifs de la forme convolution $3 \times 3$, *batch normalisation*, *LeakyReLU* [41]. Ils sont chacun entraînés durant 2000 époques. Le premier niveau de GAN contient 32 noyaux (*kernel*) dans chaque sous-bloc. Ce nombre est ensuite multiplié par 2 tous les 4 niveaux de GAN. Le nombre d'échelles $K$ est déterminé par le facteur $r$ afin que la dimension de l'image à l'échelle $K$ soit de 25 pixels. Pour éviter le sur-apprentissage, le champ réceptif, décrit en section 2.4, est également réduit à la taille $11 \times 11$.

Ce modèle ainsi décomposé en plusieurs niveaux de GAN permet de réaliser différentes tâches de vision par ordinateur, notamment la super-résolution, la génération d'image non-conditionnelle ou encore l'harmonisation d'objets. C'est cette dernière application qui est mise en œuvre au sein de GBA, résumée sur la Fig. 3.2. Nous entraînons donc SinGAN sur une coupe (ou *slice*) 2D du domaine de déploiement, contenant une région d'intérêt et aléatoirement choisie dans le jeu de données. À l'inférence, une fois $x_\lambda$ généré en suivant la méthode naïve, une échelle $k^*$ est sélectionnée et les générateurs liés aux échelles plus fines $(G_k)_{0 \leq k \leq k^*}$ sont appliqués successivement, avec des sous et sur-échantillonnages afin que les échelles correspondent :

$$\begin{aligned} G_k^{up} &= \uparrow_r \circ G_k \\ \psi_{k^*}(x_\lambda) &= (G_0 \circ G_1^{up} \circ ... \circ G_{k^*}^{up} \circ \downarrow_{r^{k^*}})(x_\lambda) \end{aligned} \tag{3.2}$$

Chaque coupe où des ROI apparaissent est alors harmonisée successivement dans le

plan axial. Les détails liés aux échelles les plus fines sont ajoutées sur les ROI pour améliorer leur réalisme. Cette étape modifie cependant l'entièreté de l'image alors que nous souhaitons créer des variations typiques uniquement au niveau des ROI. Pour cela, le masque binaire $y$ est dilaté et convolué avec un filtre gaussien dans le plan axial. Ce nouveau masque non-binaire $\hat{y}$ représente alors un masque de pondération des pixels utilisé pour fusionner l'image d'origine et l'image augmentée :

$$\psi_{\text{GBA}}(x_\lambda, \hat{y}, k^*) = \hat{y} \cdot \psi_{k^*}(x_\lambda) + (1 - \hat{y}) \cdot x_\lambda. \tag{3.3}$$

Enfin, le volume 3D est reconstruit pour générer l'image altérée par GBA. Ainsi, celle-ci présente les mêmes caractéristiques anatomiques que l'image d'origine tout en altérant l'apparence des ROI. Pour la segmentation ou toute autre application, il suffit alors de mélanger les images altérées par GBA aux réelles au sein du jeu de données d'entraînement du modèle dédié à la tâche cible.

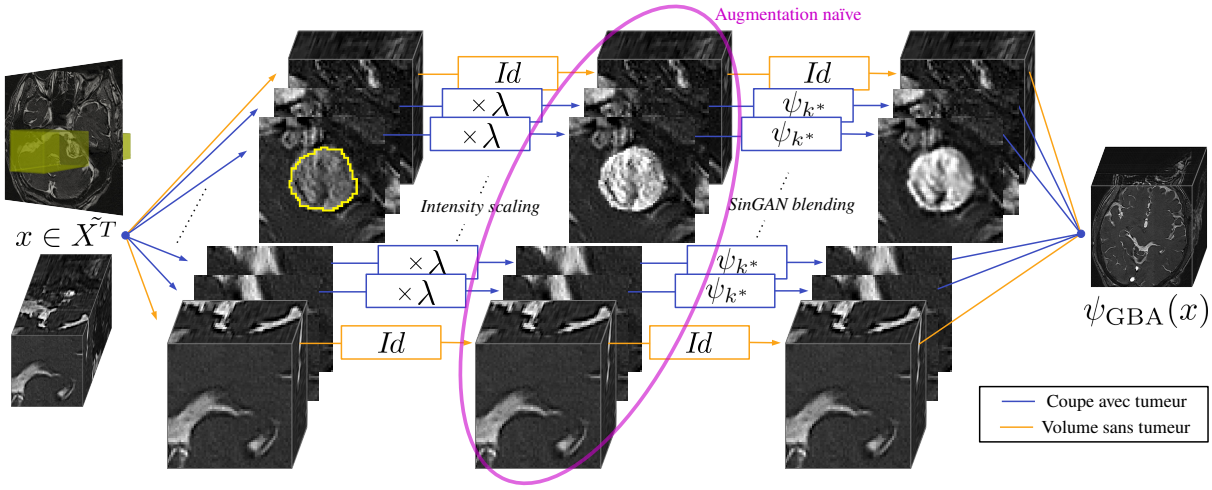



FIGURE 3.2 – Schéma de GBA illustrée sur un exemple en IRM.

## 3.2 Contexte de validation et pré-traitement

Le nombre total de paramètres à optimiser est trop élevé pour effectuer un ajustement complet via une recherche par quadrillage (*grid search*). Il faudrait en effet déterminer précisément quelles images augmenter, avec quels paramètres et combien de fois, car une même image peut être augmentée plusieurs fois avec différents $\lambda$ ou $k^*$ par exemple.

Néanmoins, nous évaluons qualitativement les images générées ainsi que les distributions engendrées, avec et sans notre augmentation. Nous comparons également la qualité de l'harmonisation à une méthode traditionnelle sans DL : le *Poisson blending* [188]. Il s'agit d'une technique de fusion d'images visant à intégrer un élément d'une image source vers une image cible en créant une transition fluide et la plus réaliste possible entre l'élément et l'image cible. Elle repose sur la résolution d'une équation de Poisson visant à fusionner les gradients des deux images en minimisant les variations brusques de gradients au sein de l'image résultante. Cette technique d'harmonisation, dite de clonage sans couture ("*seamless cloning*"), est utilisé en tant qu'augmentation dans certains contextes de l'imagerie naturelle comme médicale [189], [190].

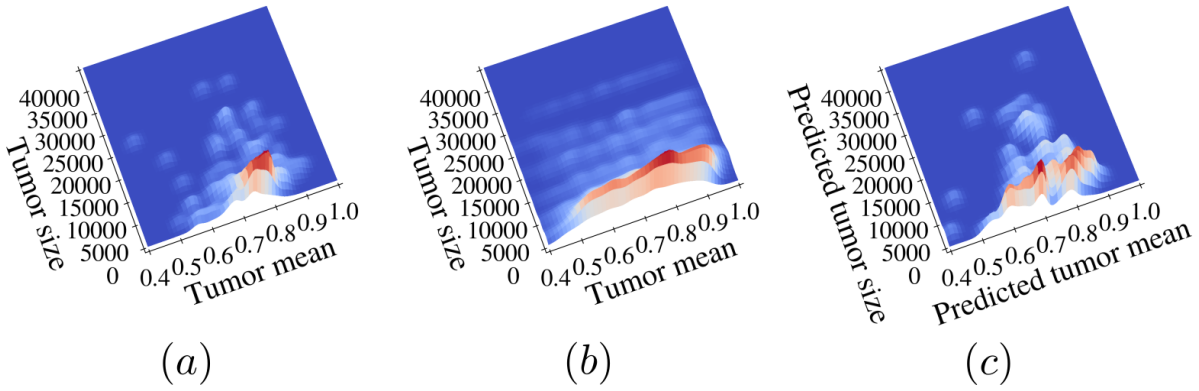



FIGURE 3.3 – Histogramme représentant (a) la distribution des tumeurs du centre $A$, (b) la distribution obtenue en augmentant ce centre avec GBA ($\lambda \in \{0.7, 0.8, 0.9, 1.1, 1.2, 1.3\}$), et (c) la distribution cible, issue du centre $B$.

Le jeu de donnée choisi pour ces évaluations est partiellement extrait du challenge CrossMoDA 2022 pour la segmentation de schwannome vestibulaire. Le jeu de donnée complet ainsi que le contexte clinique seront décrits en détail au chapitre 4. Les IRMs étudiées sont pondérées en T1 contrastées (ceT1, *contrast-enhanced T1*) et issues de deux centres $A$ et $B$. Le centre $A$ est composé de 105 ceT1 ayant un *spacing* axial de $0.4 \times 0.4$ mm$^2$ et un écart inter-slice variant de 1.0 à 1.5 mm. Le centre $B$ fournit 105 ceT1 ayant une résolution de $0.8 \times 0.8 \times 1.5$ mm$^3$. Les masques de segmentation des tumeurs ont été labellisés manuellement sur quelques coupes 2D à l'aide des ceT1 et d'images pondérée en T2 haute résolution (hrT2, *high-resolution T2*). Ils ont ensuite été étendus en 3D à l'aide d'un logiciel semi-automatique de segmentation. Afin d'harmoniser les formats des

deux centres, nous avons échantillonné toutes les images vers un *spacing* intermédiaire $0.6 \times 0.6 \times 1.0$ mm$^3$. Nous avons ensuite extrait une sous-image de taille $256 \times 256 \times z$ centrée sur le cerveau, plus précisément centrée sur la localisation moyenne des pixels ayant une intensité supérieure au $75^{\text{ème}}$ percentile.

Après pré-traitement, les Fig. 3.3a et 3.3c montrent les distributions des ROI de chaque centre, ici des tumeurs bénignes, les schwannomes vestibulaires. Ces estimation par noyau (KDE, *kernel density estimation*) témoignent du décalage de domaine au niveau des tumeurs entre les deux centres. Nous avons notamment observé une variété d'apparence de tumeurs plus large dans le second centre, à la fois plus hypo- et hypersignal par rapport au premier. En termes de taille, les tumeurs des deux ensembles étaient relativement équivalentes. Sans comparer les textures propres aux deux centres, cet histogramme dénotait un décalage intra-modalité.

## 3.3 Résultats

### 3.3.1 Apports de l'altération de contraste et de l'harmonisation

La Fig. 3.4 compare les images générées à partir d'un patient du centre $A$, présentée sur la Fig. 3.4a, par augmentation naïve et par GBA en fonction de diverses valeurs de $\lambda$. GBA est entraîné sur l'image issue du centre $B$ (figure 3.4b) , avec $K = 16, k^* = 3$. Pour chaque $\lambda$, le changement de contraste était notable pour les deux augmentations. Ensuite, en fonction des apparences de tumeur et de la qualité des masques, les contours du masque de la tumeur étaient souvent visibles sur les augmentations naïves, nuisant au réalisme global. Plus précisément, plus les valeurs de $\lambda$ étaient éloignées de 1.0, plus les contours étaient identifiables à l'œil nu, à cause des gradients irréalistes induits par ces $\lambda$. L'apport de GBA via SinGAN était alors double : il a permis de lisser les contours pour mieux fondre les tumeurs altérées dans l'image, mais aussi d'adapter la texture de la tumeur de manière non-linéaire en la diversifiant pour qu'elle soit plus réaliste vis-à-vis du contraste choisi. Ces variations étaient également plus prononcées dans les cas $\lambda \in \{0.7, 1.3\}$ que $\lambda \in \{0.9, 1.1\}$.

D'un point de vue distribution, ces augmentations ont permis de modifier la distribution globale des tumeurs du jeu de données. Par exemple, en augmentant 6 fois chaque image d'entraînement avec $\lambda \in \{0.7, 0.8, 0.9, 1.1, 1.2, 1.3\}$ ainsi que les mêmes $K$ et $k^*$, nous avons obtenu l'histogramme de tumeurs présenté sur la Fig. 3.3b. Ce dernier a cou-

vert une distribution plus large de schwannomes, permettant d'entraîner un modèle de segmentation en aval sur des tumeurs d'intensités et d'apparences plus variables que dans l'ensemble d'origine.

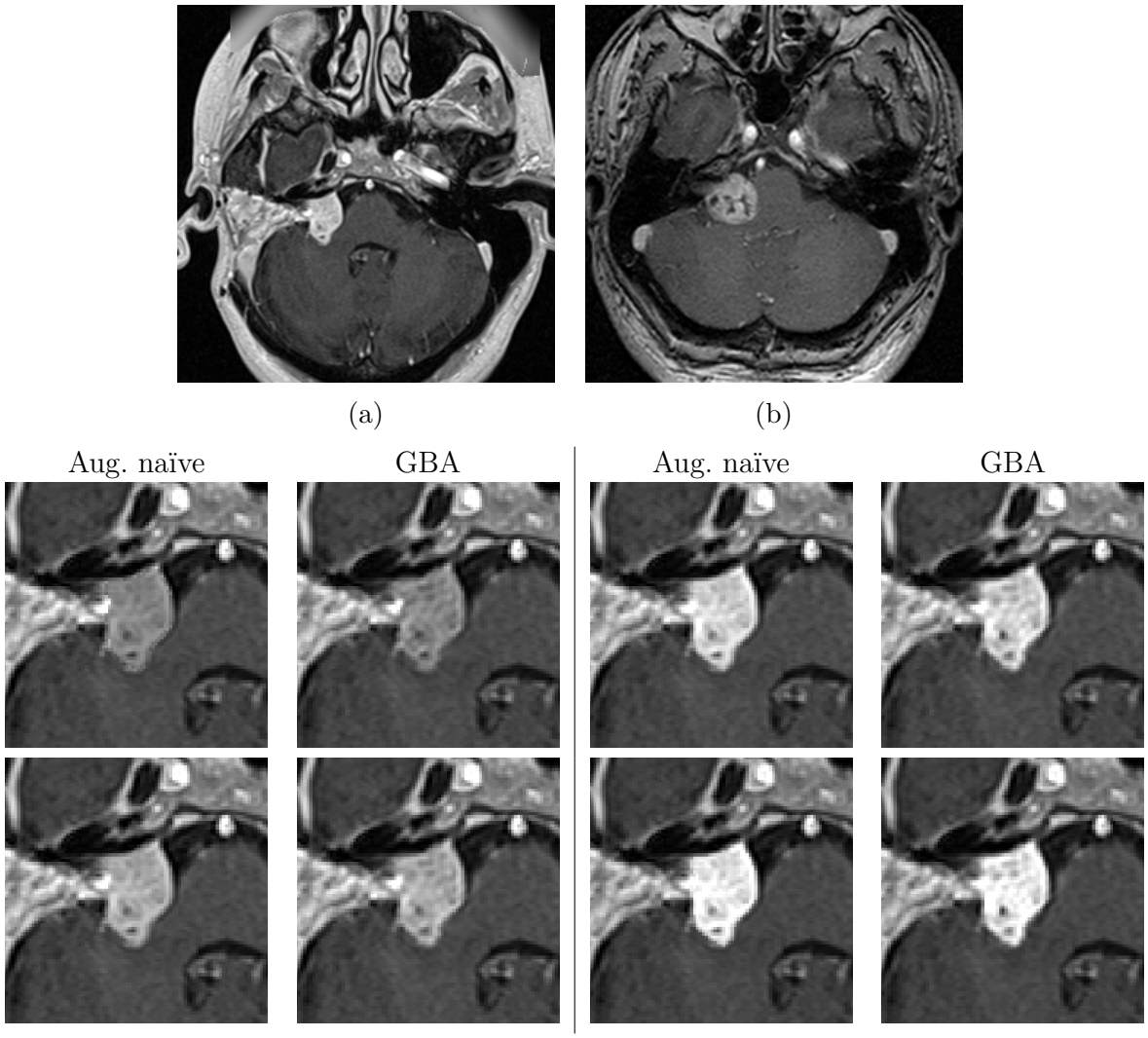
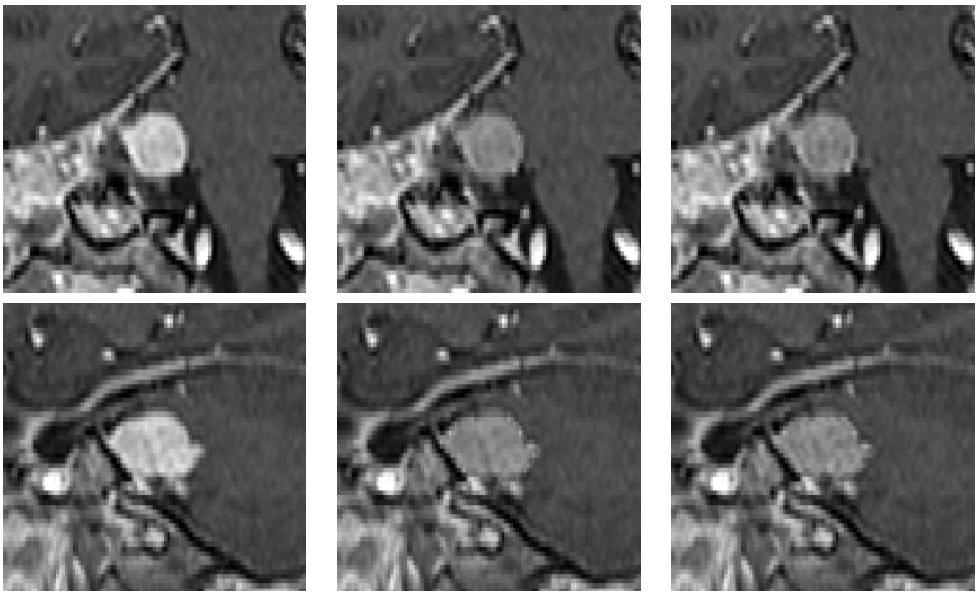



FIGURE 3.4 – Exemples qualitatifs de nos augmentations. (a) image d'origine, (b) image d'entraînement de GBA, puis résultats naïfs et profonds, respectivement avec $\lambda \in \{0.7, 0.9\}$ (hyposignal) à gauche et $\lambda \in \{1.1, 1.3\}$ (hypersignal) à droite.

Afin de vérifier la cohérence entre les coupes axiales, la Fig. 3.5 présente une coupe sagittale et coronale issues du même patient que la figure précédente. Comme précédemment, les contours du masque de segmentation étaient visibles après l'augmentation naïve. De la même manière, l'utilisation de SinGAN a permis de mieux fondre la tumeur. Le réalisme de l'image générée a donc également été conservé dans ces deux plans, et ce

même après un apprentissage 2D axial.



(a) (b) (c)

FIGURE 3.5 – Résultats dans les plans coronaux (ligne 1) et sagittaux (ligne 2) en fonction de chaque augmentation. (a) image d'origine, (b) image augmentée naïvement ($\lambda = 0.7$), et (c) image augmentée avec GBA.

### 3.3.2 Expérimentations spécifiques à la méthode profonde

Plusieurs paramètres propres à SinGAN, et donc à GBA, doivent également être sélectionnés avec précaution. Ainsi, à défaut d'un ajustement fin de ces paramètres avec une recherche par quadrillage, nous avons évalué qualitativement leur impact sur les images générées.

#### Influence de l'image d'apprentissage et du choix de l'échelle

Les Tab. 3.6 et 3.7 présentent 4 patients et leurs augmentations en fonction de trois ré-entraînements de SinGAN avec différents $k^*$ (à $K = 16$ fixé, sur deux images du centre $A$, "$A_1$" et "$A_2$", et une du centre $B$, "$B_1$"). Le premier tableau a été obtenu avec $\lambda = 0.7$ et le second avec $\lambda = 1.3$.

Tout d'abord, les entraînements de GBA $A_1$ et $A_2$ ont montré quelques différences subtiles, en particulier avec $\lambda = 0.7$. Les images générées via $A_1$ ont légèrement amélioré

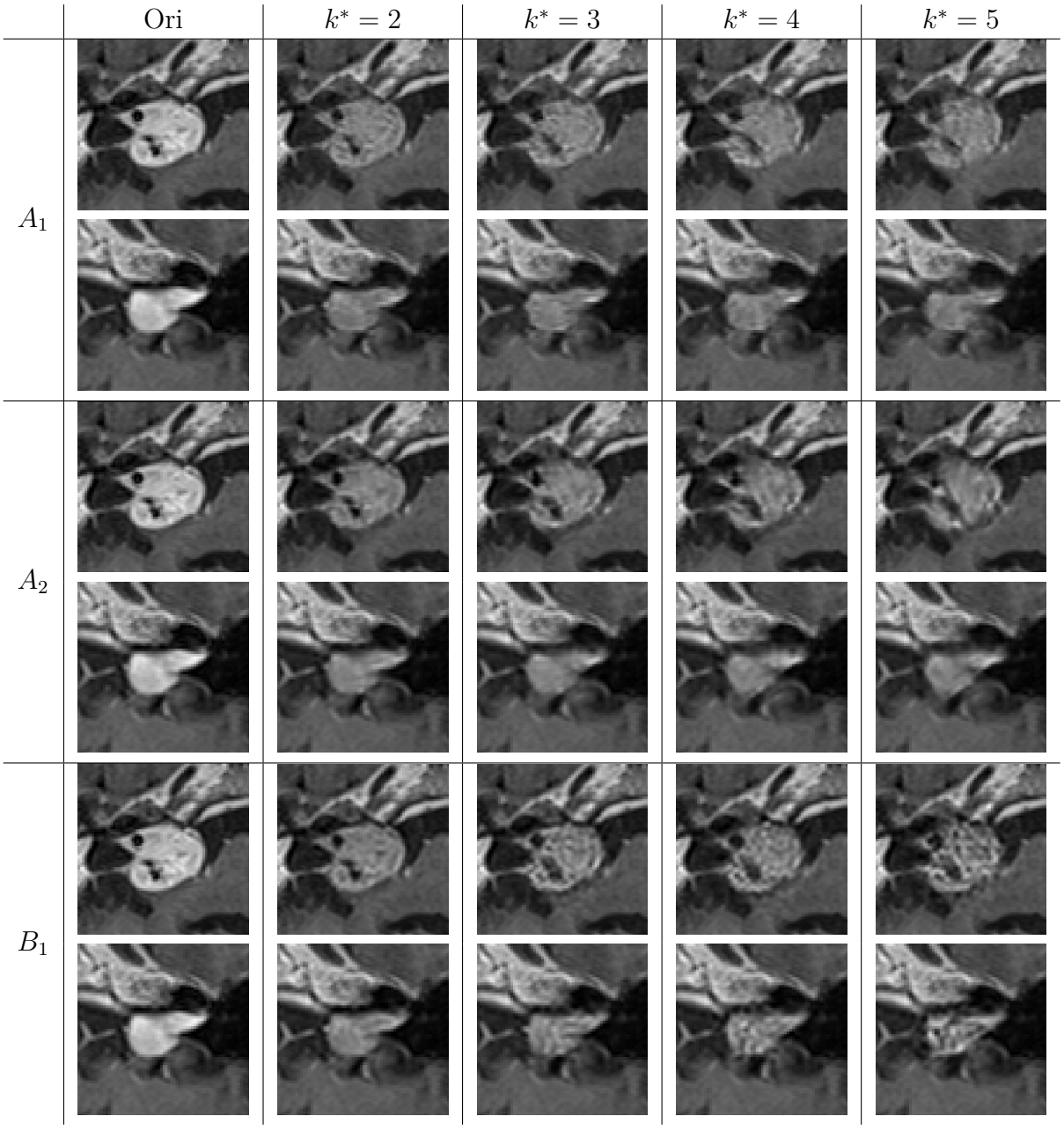



TABLE 3.6 – Variations de $k^*$ et de l'image d'apprentissage, avec $\lambda = 0.7$.

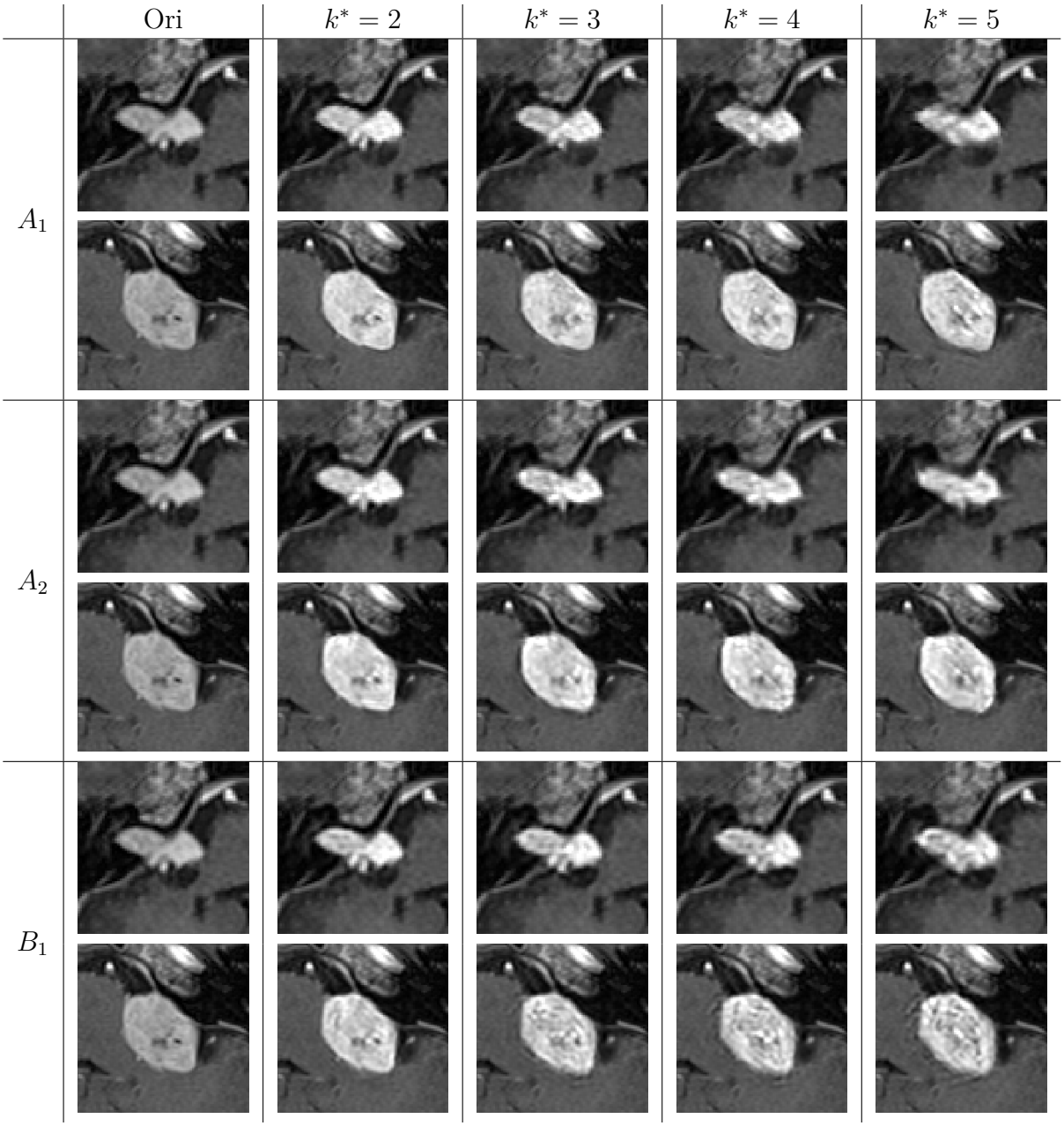


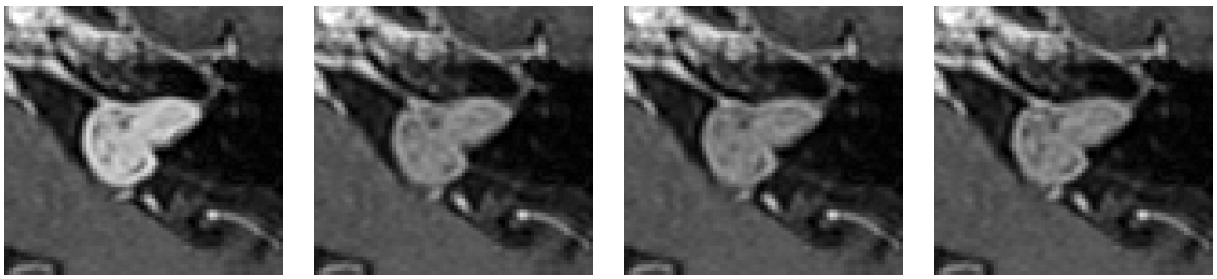

TABLE 3.7 – Variations de $k^*$ et de l'image d'apprentissage, avec $\lambda = 1.3$.

la netteté, particulièrement lorsque $k^*$ est haut, alors que l'entraînement $A_2$ a plutôt eu tendance à ajouter un léger flou réaliste sur certaines parties du schwannome. Dans ces deux cas, lorsque $k^* < 3$, les images générées étaient assez similaires. Cependant, les différences avec $B_1$ étaient plus visibles. Ce dernier apprentissage a principalement ajouté du bruit plus ou moins réaliste sur les tumeurs. Cela s'explique par la différence d'apparence des images issues des deux centres.

Pour les 4 patients, que ce soit avec $\lambda = 0.7$ ou $\lambda = 1.3$, choisir $k^* = 3$ a permis d'améliorer les contours des tumeurs sans particulièrement diversifier le contenu des tumeurs. Les premières modifications sont apparues à partir de $k^* = 3$ puis de manière croissante lorsque $k^*$ augmentait. Ainsi, utiliser un grand nombre de couches de SinGAN a permis de modifier plus profondément les schwannomes. Toutefois, lorsque $k^*$ était particulièrement élevé, par exemple $k^* = 5$, nous avons observé des modifications parfois irréalistes des tumeurs. En effet, les contours du schwannome devenaient plus flous et moins identifiables. Parfois, des artefacts ont été complètement hallucinés par GBA, éloignant alors les masques de segmentation des images augmentées.

À ce stade, il est légitime de se demander quelle influence a la coupe choisie au sein de l'image d'apprentissage. En pratique, nous n'avons observé aucune différence notable d'un patient à l'autre.

**Influence du nombre total de GAN**



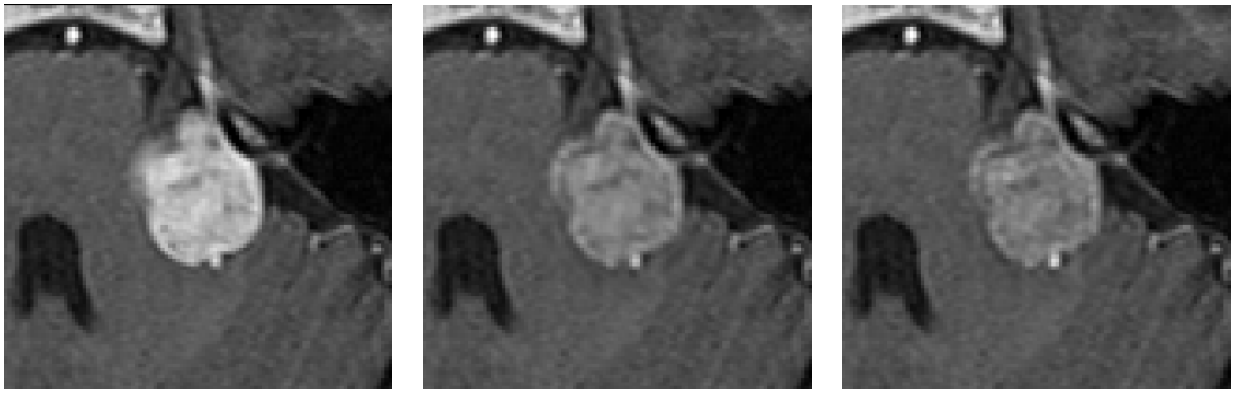
(a) (b) (c) (d)

FIGURE 3.8 – Variations de $r$ et donc de $K$ pour un même patient, en prenant $k^* = \lceil 0.2 \times K \rceil$. (a) l'image d'origine, (b) GBA avec $r = 0.65$, $K = 7$, (c) GBA avec $r = 0.75$, $K = 10$ et (d) GBA avec $r = 0.85$, $K = 16$.

La Fig. 3.8 présente le résultat de 3 ré-entraînements de GBA avec différents facteurs d'échelle $r$ et donc différents nombres de niveaux de GAN $K$. L'échelle choisie dans chaque cas pour l'inférence est alors $k^* = \lceil 0.2 \times K \rceil$. Visuellement, les 3 images générées étaient

pleinement cohérentes et relativement similaires. Nous avons remarqué sensiblement plus de détails ajoutés par GBA sur l'image associée à $K = 16$ (Fig. 3.8c). En pratique, ces 3 versions de GBA offraient des résultats similaires. L'entraînement avec un $K$ suffisamment élevé permettait néanmoins un plus grand nombre de choix pour $k^*$, afin d'affiner le rendu souhaité.

**Influence de la taille des noyaux de convolution**



(a) (b) (c)

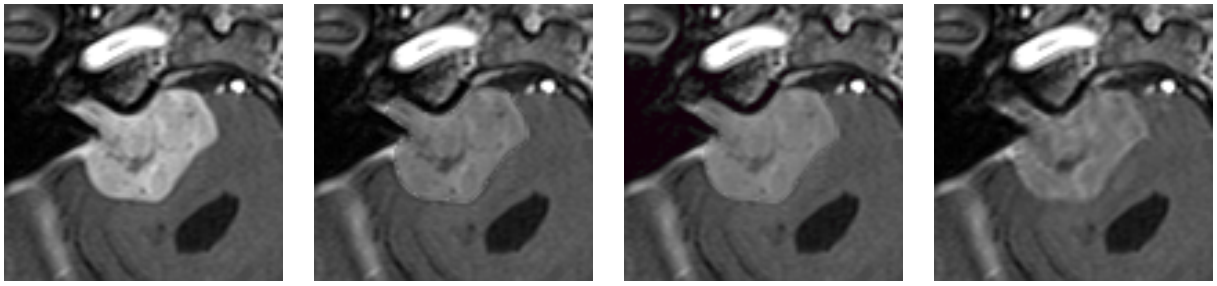
FIGURE 3.9 – Variations de la taille des noyaux de convolution au sein de SinGAN. (a) image d'origine, (b) GBA avec des noyaux $3 \times 3$, et (c) GBA avec des noyaux $5 \times 5$.

La Fig. 3.9 montre le résultat de deux entraînements de GBA en variant la taille des noyaux de convolutions, à $K = 16$ et $k^* = 3$. En utilisant des noyaux de taille $3 \times 3$, l'image augmentée présentait un léger effet de flou sur l'ensemble de la tumeur, et ce, peu importe l'image d'apprentissage de SinGAN ou le choix de $k^*$. Cela réduisait donc la variété de tumeur générée tout en nuisant au réalisme des images générées. De plus, cela ajoutait un motif reconnaissable au sein des images augmentées, visible par le réseau en aval et créant potentiellement un biais supplémentaire. En choisissant des noyaux de taille $5 \times 5$, ce flou était réduit, permettant de générer un éventail plus large de tumeurs, avec parfois des tumeurs plus nettes et mieux définies.

### 3.3.3 Comparaison avec le *Poisson blending*

La Fig. 3.10 évalue les différences entre l'harmonisation GBA et une tumeur harmonisée avec la méthode de *Poisson blending* [188]. Qualitativement, les contours de la tumeur

étaient un peu plus fondus que l'augmentation naïve tout en restant partiellement identifiable à l'œil nu. Les gradients au sein de celle-ci ont également été réduits, donnant une texture et un contraste de tumeur plus uniformes. La méthode GBA a proposé une adaptation plus cohérente des textures, générant une image globalement plus réaliste sans aucun contour apparent.



(a) (b) (c) (d)

FIGURE 3.10 – Comparaison de GBA et de *Poisson blending* pour harmoniser la tumeur altérée. (a) image d'origine, (b) image altérée naïvement ($\lambda = 0.7$), (c) par *Poisson blending* [188] et (d) par GBA.

## 3.4 Discussion

Les images générées par mélange génératif montrent un potentiel intéressant pour l'augmentation de données en segmentation. En présentant à un modèle dédié une plus grande variété de caractéristiques de ROI, suivant une distribution élargie, le modèle de segmentation pourrait bénéficier de cette palette pour accroître sa robustesse, en phase de test comme en contexte clinique. En choisissant avec précaution les paramètres ou en utilisant un critère de sélection des tumeurs à augmenter, il devient possible de générer des tumeurs ayant des caractéristiques particulières pour adapter plus finement les attributs auxquels le modèle doit être robuste. L'analyse de la distribution des ROI, par exemple au travers de KDE permettent alors d'évaluer la variété de ROI observées par le modèle de segmentation en aval. On pourrait alors imaginer prendre en compte un plus grand nombre de paramètres radiomiques pour préciser les distributions tumorales.

Visuellement, la version naïve de GBA altère le contraste global des ROI sans adaptation particulière de celle-ci. Les contours n'étant pas fondus dans l'image d'origine, ils créent un gradient peu réaliste dans l'image résultante, risquant d'ajouter un biais au modèle de segmentation. De plus, l'image altérée montre parfois un manque global de

vraisemblance dû à l'hypo- ou hypersignal créé. Comme discuté en section 2.2.1, certains travaux tendent à montrer que ce type d'artefacts ne sont pas nécessairement un frein pour l'apprentissage d'un modèle de segmentation et peuvent participer à leur robustesse [105]. Même s'il reste intéressant d'apprendre à intégrer les ROI de manière plus réaliste, voire également à diversifier son contenu pour encore améliorer les performances, l'augmentation naïve présentée offre à elle seule un potentiel clair en segmentation.

Ensuite, GBA, par l'intermédiaire de SinGAN, permet une harmonisation auto-supervisée, non-linéaire et cohérente des ROI altérées. On observe alternativement des ajouts de flou et de bruit réalistes, mais aussi des améliorations de netteté, selon les paramètres et les ROI, dans le plan d'entraînement axial tout comme dans les plans sagittaux et coronaux. Cette augmentation adapte les ROI plus localement en appliquant le style d'une image issue du domaine cible de déploiement, préalablement choisie. Ainsi, nous diversifions également le contenu des ROI en apportant des variations réalistes. Le choix du nombre de couches de GAN à employer à l'inférence, $k^*$, affecte directement la quantité de détails à ajouter. Plus ce nombre est élevé, plus l'harmonisation sera puissante, amenant parfois à halluciner des structures. Ces artefacts sont généralement réalistes, mais risquent d'impacter la forme globale des ROI. Cette limite risque d'être critique pour la segmentation, car elle est équivalente à ajouter du bruit et de l'incertitude aux labels. Une solution consiste donc à employer diverses valeurs de $k^*$ afin d'atténuer ce risque tout en bénéficiant des variations de textures de tumeurs.

Le choix des hyper-paramètres d'altération de contraste $\lambda$ joue un rôle essentiel dans la distribution des ROI obtenue. Leur sélection étant manuelle et arbitraire, on peut facilement imaginer des variations conceptuelles afin d'affiner cette sélection. Par exemple, les $\lambda$ pourraient suivre une loi de probabilité particulière afin de s'approcher au mieux de la distribution cible. On obtiendrait théoriquement l'ensemble de ROI le plus représentatif pour l'apprentissage d'un modèle vis-à-vis d'un certain domaine cible, car on aurait optimisé la réduction du décalage de données vis-à-vis des caractéristiques de ROI. Si l'ensemble des ROI cibles est suffisamment représentatif de la variété des ROI existantes, on peut alors théoriquement améliorer les performances globales en segmentation pour les caractéristiques de ROI initialement sous-représentées, encourageant donc le déploiement clinique du modèle entraîné. Si cet ensemble n'est représentatif que d'une partie des ROI réellement observées, on peut tout de même évaluer la robustesse du modèle sur ces caractéristiques, et même étendre les paramètres d'augmentation pour élargir davantage la distribution couverte.

Le principal risque lié à la sélection arbitraire de ces hyper-paramètres est de nuire à l'apprentissage de la représentation latente des ROI. En effet, des augmentations générant des contrastes trop différents des ROI réelles risquent d'encourager la prédiction de faux positifs. Inversement, des augmentations trop proches des ROI d'origine sont superflues, ne contribuant pas à réduire le risque de faux négatifs. Le choix des $\lambda$, et plus précisément l'étendue du spectre des valeurs choisies, impacte donc directement la représentativité des ROI, et ainsi les performances du modèle.

## 3.5 Conclusion

Dans ce chapitre, nous avons présenté la méthode profonde GBA et sa version naïve, deux techniques d'augmentation de données pour la segmentation, utilisables à tous les régimes de données et visant à diversifier les apparences des ROI. En altérant leur contraste via une transformation linéaire, puis en l'harmonisant de manière réaliste au sein de l'image d'origine, nous générons une nouvelle image conservant l'anatomie du patient tout en variant les caractéristiques des ROI au sein de l'ensemble d'entraînement.

En choisissant précautionneusement les valeurs d'altération de contraste, nous modifions la distribution des ROI visibles à l'entraînement pour l'étendre ou la rapprocher d'une autre distribution, celle des ROI du domaine cible par exemple. Cela rend l'augmentation adaptable aux différents décalages de données possibles au niveau de ces régions, au sein d'une même modalité tout comme dans un contexte inter-modale, en fonction du domaine cible de déploiement.

Dans le chapitre suivant, nous présentons une première application concrète de ces augmentations en IRM, dans le contexte de segmentation inter-modale via le challenge CrossMoDA 2022.

Chapitre 4

# Segmentation inter-modale de schwannome vestibulaire en IRM

**Résumé**

Ce chapitre présente une première application des techniques d'augmentation proposées au chapitre précédent pour la segmentation inter-modale non-supervisée du schwannome vestibulaire, dans le cadre du challenge MICCAI CrossMoDA 2022. Deux modalités sont considérées : la source (IRM T1 contrastée, annotée) et la cible (IRM T2 haute résolution, non-annotée) dans un contexte multi-centrique. Afin d'apprendre à segmenter la modalité cible, nous proposons d'employer une approche d'UDA standard : un modèle d'I2I suivi d'un schéma de segmentation itératif avec auto-entraînement. Après synthèse I2I traditionnelle, un décalage de données dû aux différences inter-centriques est observé. Nous corrigeons ce décalage en combinant notre approche d'augmentation par mélange génératif à une stratégie d'auto-entrainement, où nous réutilisons de façon itérative les pseudo-labels prédits. Nous obtenons un net gain de performance vis à vis d'augmentations conventionnelles typiquement utilisées en DL. À l'issue du challenge, notre solution a atteint la première place sur la tâche de segmentation de tumeur, ainsi que la 3ème place au classement général incorporant une autre tâche que nous n'avons pas optimisée.

Comme vu au chapitre 2, l'UDA, et plus particulièrement la segmentation inter-modale, est particulièrement étudiée en raison de sa difficulté et du nombre conséquent d'applications potentielles [38], [40]. C'est dans ce contexte que s'inscrit le challenge MIC-

CAI CrossMoDA 2022, proposant de mélanger un décalage de domaines entre modalité avec une contrainte multi-centrique. Dans cette situation, un modèle de synthèse I2I puis de segmentation sont généralement employés [38]. Nous proposons une méthode similaire utilisant GBA afin de réduire davantage le décalage. Ce travail a été décrit dans un article à paraître dans les *IEEE Transactions on Biomedical Engineering*. L'image Docker présentée dans le cadre du challenge est disponible via notre GitHub [1].

## 4.1 Contexte clinique et problématique des agents de contraste

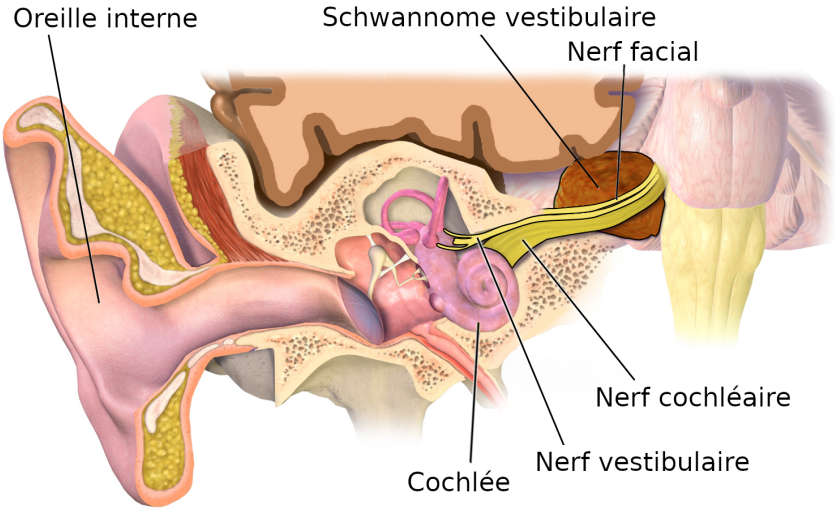



FIGURE 4.1 – Schéma du schwannome vestibulaire et des nerfs proches. Adapté de [191].

Le schwannome vestibulaire (VS, *vestibular schwannoma*) est une tumeur bénigne de la gaine du nerf vestibulaire, responsable de l'audition et de l'équilibre, dont l'incidence est estimée à 1 sur 1000 [192]. Les symptômes sont une perte d'audition du côté de la tumeur, des troubles de l'équilibre, voire une paralysie faciale due à la proximité du nerf facial. L'évolution de la VS est mesurée par la classification de Koos, basée principalement sur le volume de la tumeur et son impact sur les structures environnantes telles que la cochlée et le tronc cérébral [193].

1. `https://github.com/GuigzoS/crossmodalGBA`

Dans la routine clinique la plus courante pour la planification de traitement par radiothérapie du VS, la tumeur et le principal organe à risque, la cochlée, sont respectivement segmentés sur des IRM pondérées en T1 avec renforcement du contraste (ceT1) et des IRM hautes résolutions pondérées en T2 (hrT2) [194], [195]. En d'autres termes, deux modalités d'IRM dont une avec injection de produit de contraste, souvent du gadolinium, sont acquises pour cette pathologie.

Le gadolinium est un métal paramagnétique généralement toléré par le corps humain (hors allergie particulière). Lorsque le corps est placé dans le champ magnétique lors de l'IRM, les protons des molécules d'eau s'alignent avec ce champ. L'émission d'une onde électromagnétique à fréquence radio perturbe alors cet alignement. Une fois celle-ci éteinte, les protons reviennent à leur alignement initial, un phénomène appelé relaxation. Ce temps de relaxation varie selon les différents types de tissus, fournissant ainsi divers contrastes au sein des IRMs. En présence d'un agent de contraste comme le gadolinium, ces temps de relaxation, en particulier en pondération T1, sont réduits, accentuant ainsi certains contrastes et permettant une meilleure identification des zones pathologiques [196]. Pour les lésions les plus subtiles telles que les micro-angiomes ou les petites métastases cérébrales de moins de 5 mm, le gadolinium est indispensable pour les distinguer et les observer. À l'inverse, les tumeurs ayant un plus gros volume, comme les méningiomes, pouvant atteindre 10 cm de diamètre, peuvent souvent être distingué sans utilisation d'agent de contraste. Bien que l'administration de gadolinium dans de telles situations permette d'améliorer les détails de ces tumeurs, son utilité est réduite.

En parallèle, l'utilisation de gadolinium augmente le coût financier de l'IRM mais aussi certains risques [197], [198]. Tout d'abord, celui-ci est principalement évacué par le corps par le biais des urines. Par conséquent, il se retrouve dans nos systèmes d'épuration des eaux, puis potentiellement dans l'environnement, avec des conséquences écologiques encore mal comprises. Ensuite, dans certains cas comme chez les patients atteints d'insuffisance rénale chronique, l'injection de produits de contraste peut provoquer une fibrose systémique néphrogénique, impactant d'abord la peau puis potentiellement d'autres organes [199]. Enfin, certaines études ont montré que le gadolinium s'accumule à long terme dans certains tissus cérébraux [200]. Bien que les conséquences cliniques soient pour le moment méconnues, cette rétention reste préoccupante.

Au vu de ces éléments, il paraît naturel de réduire l'utilisation du gadolinium aux cas où celui-ci est strictement nécessaire. Pour le VS, des séquences IRM dédiées telles que les images hrT2 suscite un intérêt grandissant pour sa segmentation afin de réduire les coûts

globaux et l'utilisation d'agents de contraste [197], [198]. Sur cette modalité, la majorité des VS sont en hyposignal, et donc moins visible à l'œil nu, même si certains sont très hétérogènes et en hypersignal. Dans les deux cas, les progrès en segmentation automatique encouragent à l'utilisation de cette modalité pour segmenter à la fois le VS et la cochlée [38], réduisant ainsi coût et risques pour le patient. En raison du temps et du coût de production d'annotations sur des acquisitions hrT2, les modèles d'UDA inter-modalités permettent de réutiliser des séquences ceT1 précédemment étiquetées afin d'entraîner un modèle de segmentation VS sur hrT2 [201].

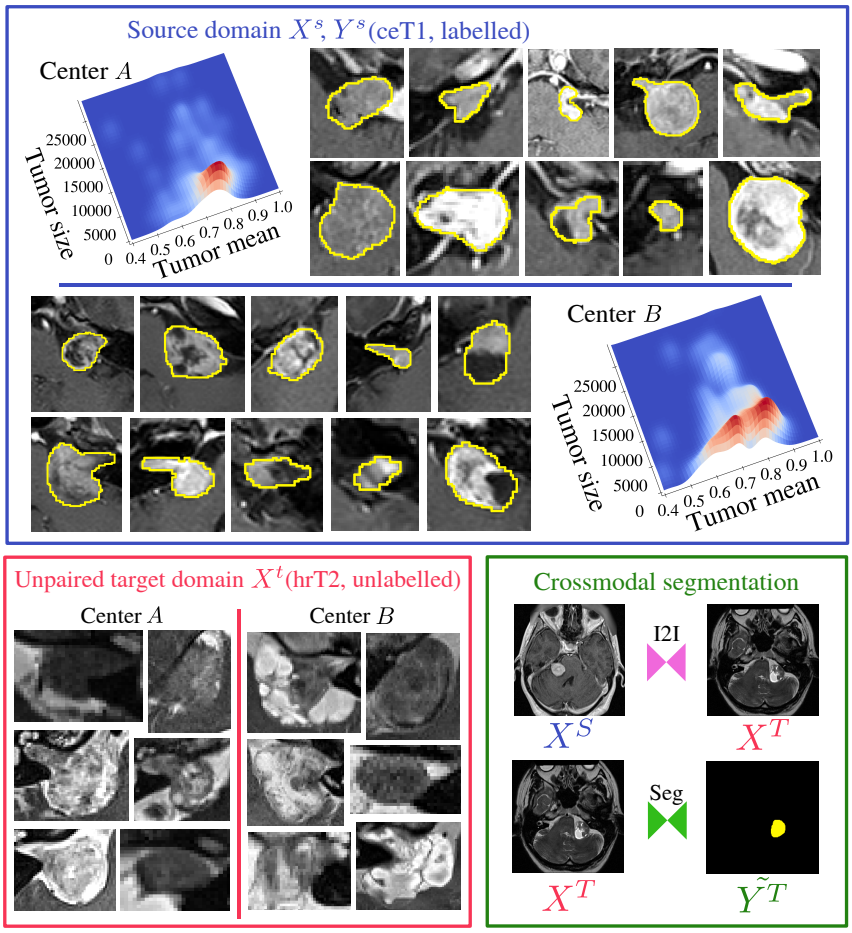



FIGURE 4.2 – Scénario du contexte du challenge CrossMoDA 2022 en segmentation de schwannome vestibulaire.

Le challenge CrossMoDA propose donc, à partir d'un ensemble d'images ceT1 annotées et hrT2 non-annotées, d'apprendre un modèle de segmentation de VS et de cochlée pour cette seconde modalité. Ce schéma est résumé en Fig. 4.2 et montre quelques exemples

de tumeurs. Dans l'édition 2022, chaque ensemble d'images est fourni par deux centres d'acquisition. De plus, la cochlée étant déjà segmentée sur hrT2, cette tâche est un faux problème, car il existe déjà des bases de données d'IRM hrT2 avec annotations de celle-ci. Pour cette raison, nous présentons ici uniquement notre solution pour la segmentation inter-modale de VS.

## 4.2 Méthode de segmentation inter-modale

Notre méthodologie se décompose en 3 étapes : la synthèse I2I, la génération d'exemples augmentées avec GBA et la segmentation itérative avec auto-entraînement. Afin de présenter une solution générique à la segmentation inter-modale entre d'autres modalités sources et cibles, introduisons quelques notations.

Soient $(X^S, Y^S)$ les images et masques de la modalité source (dans notre cas, ceT1) et $X^T$ les images de la modalité cible (dans notre cas, hrT2). Chaque ensemble étant composé de deux centres, on les note $A$ et $B$ tels que $X^S = X^S_A \cup X^S_B$ et $X^T = X^T_A \cup X^T_B$. À ce stade, le décalage entre les deux centres est potentiellement conséquent, rendant impossible l'apprentissage direct d'une synthèse $S \to T$ avec les deux centres simultanément.

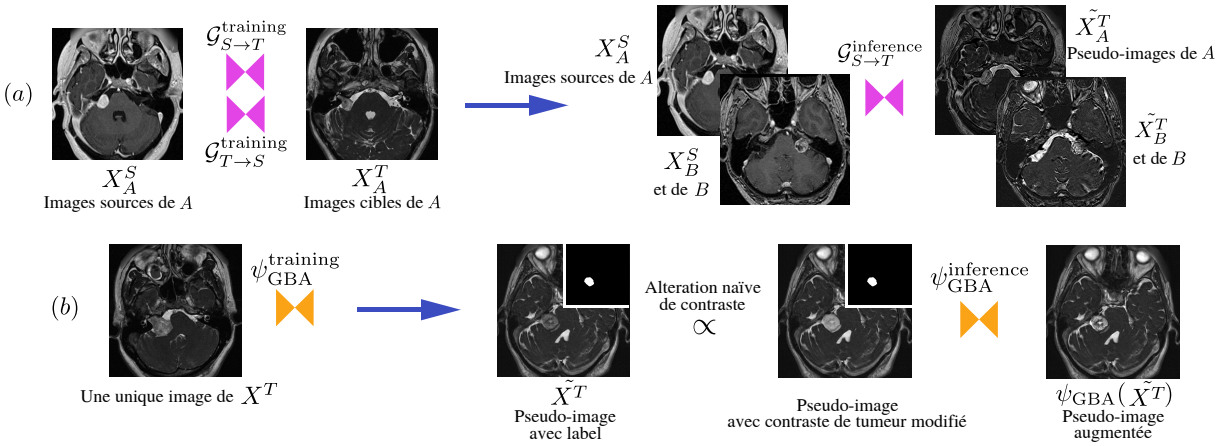



FIGURE 4.3 – Schéma récapitulatif de (a) la synthèse I2I et de (b) GBA, dans le *workflow* global pour la segmentation inter-modale.

### 4.2.1 Synthèse image-à-image et augmentation générative

Tout d'abord, nous apprenons une fonction (*mapping*) global d'intensité entre les images sources $X_A^S$ et cibles $X_A^T$ à l'aide d'un modèle CycleGAN de synthèse I2I non-supervisée, illustré sur la Fig. 4.3a. Cet apprentissage est restreint à un des deux centres seulement à cause de l'écart intra-modalité trop important qui, si combinés à l'entraînement, pourrait nuire au *mapping* appris. Nous entraînons donc deux générateurs $(\mathcal{G}_{S\to T}, \mathcal{G}_{T\to S})$ et deux discriminateurs $(\mathcal{D}_S, \mathcal{D}_T)$ avec une contrainte adversaire et de cohérence de cycle. La fonction de coût adversaire est définie par :

$$\begin{aligned}\mathcal{L}_{adv}(\mathcal{G}_{S\to T}, \mathcal{D}_T, X^S, X^T) = \mathbb{E}_{x^T\sim X^T}[log(\mathcal{D}_T(x^T))] \\ + \mathbb{E}_{x^S\sim X^S}[log(1-\mathcal{D}_T(\mathcal{G}_{S\to T}(x^S)))]\end{aligned} \tag{4.1}$$

et celle de cohérence de cycle par :

$$\begin{aligned}\mathcal{L}_{cyc}(\mathcal{G}_{S\to T}, \mathcal{G}_{T\to S}, X^S, X^T) = \\ \mathbb{E}_{x^S\sim X^S}[\|\mathcal{G}_{T\to S}(\mathcal{G}_{S\to T}(x^S)) - x^S\|_1] \\ +\mathbb{E}_{x^T\sim X^T}[\|\mathcal{G}_{S\to T}(\mathcal{G}_{T\to S}(x^T)) - x^T\|_1].\end{aligned} \tag{4.2}$$

La fonction de coût total est donc :

$$\begin{aligned}\mathcal{L}(\mathcal{G}_{S\to T}, \mathcal{G}_{T\to S}, \mathcal{D}_S, \mathcal{D}_T, X^S, X^T) = \\ \mathcal{L}_{adv}(\mathcal{G}_{S\to T}, \mathcal{D}_T, X^S, X^T) \\ +\mathcal{L}_{adv}(\mathcal{G}_{T\to S}, \mathcal{D}_S, X^T, X^S) \\ +\mu_{cyc}\cdot\mathcal{L}_{cyc}(\mathcal{G}_{S\to T}, \mathcal{G}_{T\to S}, X^S, X^T)\end{aligned} \tag{4.3}$$

Seul le générateur $\mathcal{G}_{S\to T}$ est ensuite utilisé pour les deux centres, afin de générer des pseudo-images $\tilde{X^T}$, alignées aux $X^S$ mais ayant les caractéristiques du domaine cible :

$$\tilde{X^T} = \mathcal{G}_{S\to T}(X^S). \tag{4.4}$$

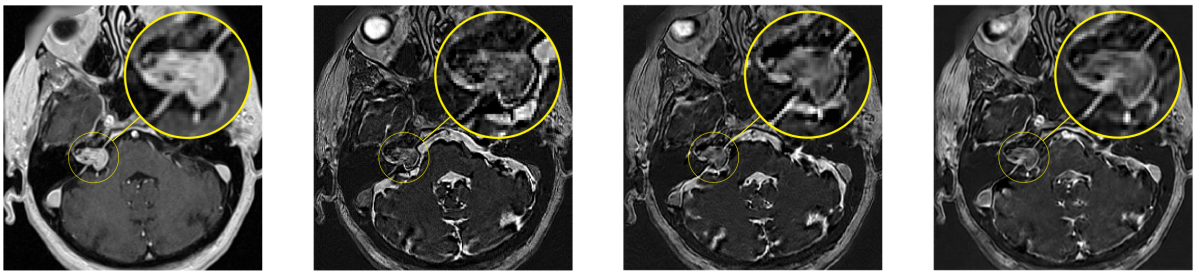
Même si l'apparence globale des images générées $\tilde{X^T}$ est similaire aux $X_A^T$, comme illustré sur la Fig. 4.4, plusieurs limites critiques sont identifiées après la synthèse. Tout d'abord, en analysant certaines caractéristiques locales des $\tilde{X^T}$, en particulier au niveau des tumeurs, les textures et formes des VS synthétisés ont été altérées, et leurs contrastes réduits en comparaison aux images réelles $X^T$. On observe ces différences sur les zooms

de la Fig. 4.4. Un effet de flou est également parfois constaté. Cela illustre les tendances, décrites au chapitre 2, des CycleGAN à ignorer les détails à basse échelle [82]. Ensuite, les images générées sont particulièrement variables d'un apprentissage à l'autre, même avec des ré-entraînements identiques, comme également discuté au chapitre 2 et observé sur cette même figure. Ces variations nuisent directement à la tâche de segmentation en aval. Enfin, comme le modèle a été entraîné sur les images du centre $A$ seulement, celles du centre $B$ n'ont pas été fidèlement synthétisées. La distribution des $X_B^S$ n'est en effet pas couverte par celle des $X_A^S$. Cela a également pour conséquence que certaines caractéristiques typiques de tumeurs de $X_B^T$ n'ont pas été générées et ne sont donc pas représentées parmi les $\tilde{X^T}$.

Pour ces raisons, il est nécessaire d'augmenter la diversité des VS au sein des images générées $\tilde{X^T}$ avant la segmentation. Plus précisément, nous cherchons à augmenter les pseudo-images afin d'étendre la distribution couverte de tumeurs pour la tâche en aval. Cela permet notamment de 1) corriger les apparences de tumeurs au sein des pseudo-images, 2) compenser l'aléa induit par CycleGAN et 3) réduire les potentiels écarts entre domaines dus aux différences entre $\tilde{X^T}$ et $X_B^T$. À ce stade, GBA semble particulièrement approprié, car nous constatons qualitativement que les contrastes des tumeurs de $X_B^T$ sont plus variables que celles de $\tilde{X^T}$, parfois en hypo ou hypersignal comparé aux tissus environnants. L'hypothèse sous-jacente est donc qu'en générant des variantes de tumeurs ayant des caractéristiques différentes de celles générées par CycleGAN, le modèle de segmentation sera plus robuste et segmentera mieux les tumeurs les plus atypiques. Ainsi, nous entraînons GBA, comme décrit au chapitre 3 et rappelé sur la Fig. 4.3, à l'aide d'un modèle SinGAN représenté par les équations 3.1, 3.2 et 3.3 afin d'obtenir $\psi_{\text{GBA}}$. Les paramètres d'augmentation seront détaillés en section 4.3.2.

### 4.2.2 Segmentation itérative et auto-entraînement

Afin d'entraîner un modèle de segmentation pour les images réelles $X^T$ à partir des images générées $\tilde{X^T}$, un schéma de segmentation itérative avec auto-entraînement est employé, inspiré par les méthodes les plus performantes de l'édition 2021 de CrossMoDA [38], [186], [202]. Ce schéma est une évolution du paradigme enseignant/élève, abordé en section 2.3.2. La Fig. 4.5 présente cette partie du *workflow* global. L'objectif est d'utiliser à l'entraînement les images générées avec les vrais labels, mais aussi les vraies images avec des pseudo-labels générés par le modèle enseignant. Nous combinons cette technique avec les images augmentées par GBA.



(a) (b) (c) (d)

FIGURE 4.4 – Variabilité de CycleGAN, avec un zoom sur le VS. (a) Image source, (b,c,d) images générées après 3 ré-entraînements avec paramètres identiques.

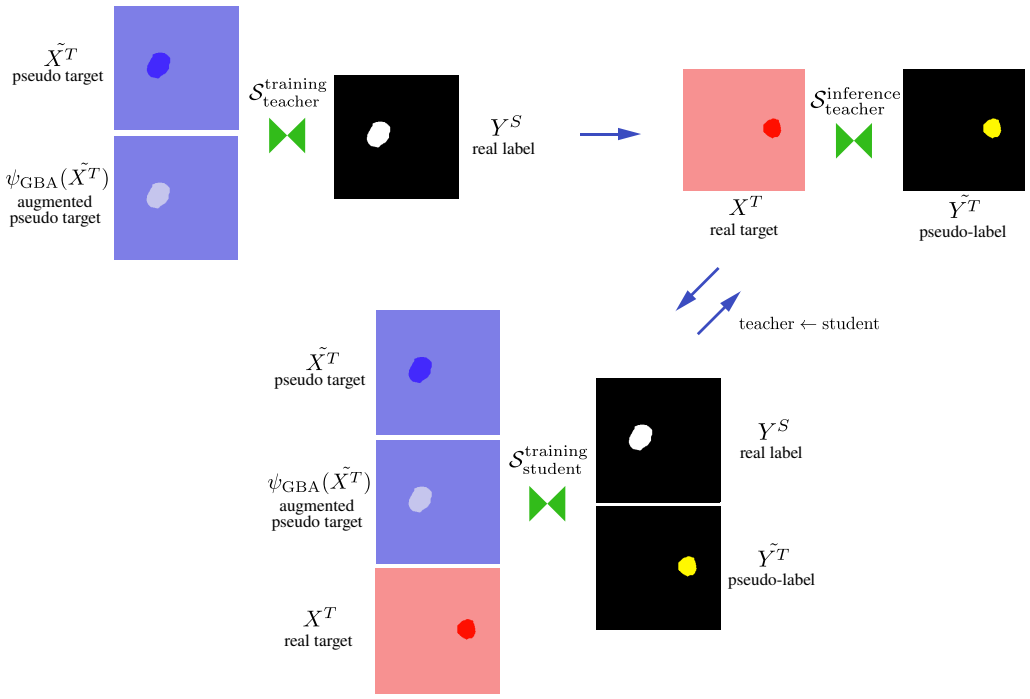



FIGURE 4.5 – Schéma récapitulatif de la segmentation itérative avec auto-entraînement.

Plus précisément, un premier modèle de segmentation $\mathcal{S}_{\text{teacher}}$ est entraîné sur les sorties de la synthèse I2I et leur version augmentée par GBA $(\tilde{X^T} \cup \psi_{\text{GBA}}(\tilde{X^T}), Y^S)$. Les images réelles $X^T$ sont ensuite fournies à l'inférence au modèle afin de produire des pseudo-labels $\tilde{Y^T}$ alignés aux $X^T$. Notons que ces derniers sont potentiellement bruités à cause du décalage de domaines entre données réelles et générées. À ce stade, les paires image-masque disponibles sont divisées en, d'un côté, images générées et vrais masques $(\tilde{X^T} \cup \psi_{\text{GBA}}(\tilde{X^T}), Y^S)$, et de l'autre, images réelles et masques estimés, notées $(X^T, \tilde{Y^T})$. Ces ensembles sont alors combinés pour un nouvel entraînement, formant l'ensemble $(\tilde{X^T} \cup \psi_{\text{GBA}}(\tilde{X^T}) \cup X^T, Y^S \cup \tilde{Y^T})$, afin d'améliorer la robustesse du modèle de segmentation pour ensuite mettre à jour les labels $\tilde{Y^T}$. Nous répétons ensuite ce principe plusieurs fois en remplaçant les pseudo-masques jusqu'à ce qu'aucune amélioration de Dice liée à ces derniers ne soit observée. En pratique, cette convergence est obtenue après trois itérations.

## 4.3 Contexte de validation

Dans cette section, plusieurs aspects pratiques sont décrits, notamment le jeu de données, les pré-traitements appliqués, les métriques utilisées ainsi que la stratégie d'augmentation établie pour les différentes expérimentations.

### 4.3.1 Jeux de données et métriques

Le jeu de données d'entraînement de CrossMoDA 2022 est composé de 210 images ceT1 annotées et de 210 images hrT2 non appariées et non annotées, réparties équitablement entre deux centres cliniques : le Centre $A$ (*Londres*) et le Centre $B$ (*Tilbourg*) [195]. La Fig. 4.6 représente les distributions normalisées des deux modalités pour chaque centre, soulignant les écarts de domaines entre centres et entre modalités. Les images ceT1 issues de A présentent une résolution axiale de $0.4 \times 0.4$ mm$^2$ associée à des matrices de taille $512 \times 512$, avec un écart inter-slice variant de 1.0 à 1.5 mm. Les images hrT2 de ce centre ont une résolution axiale de $0.5 \times 0.5$ associées à des matrices de $384 \times 384$ ou $448 \times 448$, avec un écart inter-slice variant également de 1.0 à 1.5 mm. Concernant le centre $B$, les images ceT1 présentent une résolution axiale de $0.8 \times 0.8$ mm$^2$ associée à des matrices de taille $256 \times 256$, avec un écart inter-slice de 1.5 mm. Les images hrT2 de ce centre montrent une résolution axiale de $0.4 \times 0.4$ mm, associées à des matrices de taille $512 \times 512$, avec un écart inter-slice de 1.0 mm. Tous les masques de segmentation de VS ont été délimités

manuellement sur des coupes axiales sélectionnées en consensus par un neurochirurgien et un oncologue sur les images ceT1 en utilisant les modalités ceT1 et hrT2 comme guides. Un logiciel de segmentation semi-automatique 2D a ensuite été utilisé pour produire des masques de segmentation 3D [2]. Aucun label de référence n'a été fournie pour aucune image cible.

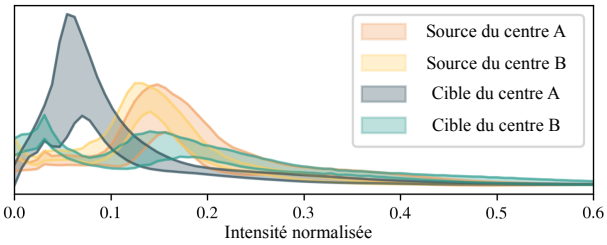



FIGURE 4.6 – Distribution normalisée des intensités de chaque modalité en fonction du centre. Les images ont été rééchelonnées dans l'intervalle $[0, 1]$ et les voxels d'arrière-plan ont été exclus. Le milieu des bandes représente la moyenne ; celles-ci s'étendent d'un écart-type au-dessus et en dessous de la moyenne.

Un ensemble de validation est également accessible, contenant 32 images hrT2 non annotées de chaque centre, soit 64 au total. Pour évaluer la méthode pendant et après le challenge sans donner accès aux labels vérité-terrain, la validation n'a été possible qu'en ligne avec une soumission quotidienne autorisée renvoyant deux métriques pour chaque patient : le coefficient de similarité de Dice (DSC) et la distance moyenne symétrique de surface (ASSD), définis par les formules suivantes :

$$D(\tilde{y}, y) = \frac{2|\tilde{y} \cap y|}{|\tilde{y}| + |y|} \tag{4.5}$$

$$\mathrm{ASSD}(\tilde{y}, y) = \frac{1}{2}\left(\frac{1}{|\tilde{y}|}\sum_i \min_j \mathrm{dist}(\tilde{y}_i, y_j) + \frac{1}{|y|}\sum_j \min_i \mathrm{dist}(y_j, \tilde{y}_i)\right) \tag{4.6}$$

où $\tilde{y}$ représente une prédiction, $y$ la vérité-terrain associée, $\tilde{y}_i$ le pixel prédit de tumeur numéroté $i$ et $y_j$ le pixel réel de tumeur numéroté $j$. En pratique, l'ASSD est un indicateur du nombre de VS non détectés par le modèle de segmentation.

De plus, 270 patients (137 de *Londres*, 133 de *Tilbourg*) sont restés privés afin de tester les différentes méthodes proposées. Dans le cadre de la compétition, chaque méthode a été classée pour chacun des 270 patients, pour la cochlée et le VS. Un classement moyen

2. Aucune précision supplémentaire n'a été fournie par les auteurs du challenge.

pour tous les patients et pour les deux classes a alors été calculé, établissant le classement final du *challenge*.

### 4.3.2 Détails d'implémentations

#### Stratégie d'augmentations

La logique des choix des paramètres de GBA a été déterminée afin de couvrir la distribution des tumeurs en hrT2 observée sur les deux centres. Pour guider ce choix, en particulier celui des scalaires d'ajustement de contraste $\lambda$ et du nombre d'augmentations, nous avons estimé la distribution réelle des tumeurs au sein des images $X^T$ à l'aide des masques prédits $\tilde{Y^T}$ obtenus après une première segmentation. La Fig. 4.7 compare les KDE des distributions de tumeurs générées en termes de taille et d'intensité, avant et après GBA, avec celles estimées de tumeurs réelles sur hrT2 après la dernière itération (obtenue *a posteriori*). Empiriquement, nous avons orienté nos augmentations pour combler deux manques :

— Le centre $B$, *Tilbourg*, présentent plus de grosses tumeurs hétérogènes que le centre $A$, *Londres*. La proportion de celles-ci a donc été augmentée. Tous les VS de $\tilde{X_B^T}$ ayant un volume supérieur à 2000 mm$^3$ et un écart-type supérieur à 0.10 (29 patients au total) ont donc été augmentés. 3 valeurs $\lambda$ ont été employées : 0.7, 1.2 et 1.5. Le modèle SinGAN a été entraîné avec $K = 16$, soit 17 échelles de GAN au total. Pour chaque $\lambda$ et chaque image sélectionnée, 2 versions de la tumeur ont été générées, en faisant varier le choix de la première échelle utilisée dans SinGAN : $k^* = 1$ et $k^* = 3$. Au total, $29 \times 6 = 174$ images ont été générées.

— Le contraste des petites tumeurs a été réduit. Les images des deux centres $\tilde{X^T}$ ayant une tumeur de moins de 300 mm$^3$, soit 19 images au total, ont été augmentées avec des valeurs de $\lambda$ globalement plus basses : 0.6, 0.8 et 1.2. Les mêmes paramètres de SinGAN ($K$ et $k^*$) ont été choisis, soit 6 augmentations par image ou 114 images augmentées.

Ainsi, la distribution élargie des ROI, dont la KDE est représentée sur la Fig. 4.7b, se rapproche davantage de la distribution cible, représentée sur la Fig. 4.7c.

#### Traitement des données

Avant tout entraînement, les images des deux modalités ont été ré-échantillonnées vers un *spacing* inter-voxels de $0.6 \times 0.6 \times 1.0$ mm$^3$. Ensuite, pour chaque image, nous avons

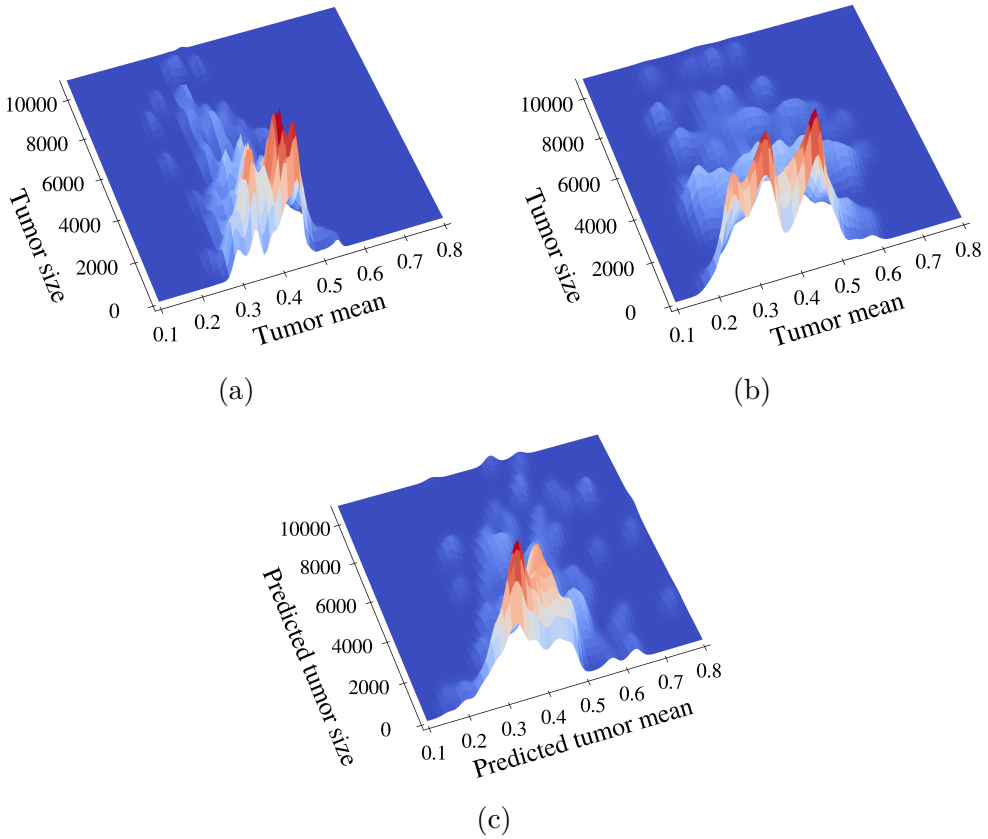



FIGURE 4.7 – KDE du volume et de l'intensité des tumeurs au sein des images normalisées du challenge CrossMoDA 2022. (a) pseudo-images hrT2 sans augmentation, (b) pseudo-images hrT2 et leurs augmentations par GBA, et (c) images hrT2 réelles. Les masques réels n'étant pas disponibles, la KDE (c) est estimé à l'aide de $\tilde{Y^T}$.

déterminé les coordonnées axiales $(x, y)$ du centre du cerveau comme étant la localisation moyenne des voxels ayant une intensité supérieure au $75^{\text{ème}}$ percentile, comme l'a proposé *Choi et al.* [203]. Un sous-volume de taille $256 \times 256 \times Z$ centré en $(x, y)$ a ensuite été extrait, où Z représente le nombre de slices axiales. Après synthèse I2I, à cause d'un léger effet de flou causé par CycleGAN, les pseudo-images générées ont été déconvoluées par l'algorithme de déconvolution itératif de Van Cittert, avec une PSF de $1 \times 1 \times 2.5$ mm$^3$ et pendant 15 itérations, qui a montré une légère amélioration sur le jeu de validation. Enfin, avant la dernière itération de segmentation, les images ont été ré-échantillonées à un *spacing* plus fin : $0.4 \times 0.4 \times 1.0$ mm$^3$.

**Autres paramètres**

Concernant le modèle CycleGAN et l'équation 4.3, nous avons utilisé comme poids pour le coût du cycle $\mu_{cyc} = 10$. Concernant la segmentation, nous avons employé le *framework* nnU-Net en 3D haute résolution, avec une validation croisée à 5 plis (*5-fold ensembling*) [21]. Les voxels mal classifiés pouvant impacter significativement l'entraînement, cela permet d'obtenir les pseudo-labels les plus fiables possibles. En incluant les 3 itérations d'auto-entraînement, nous avons au total entraîné 20 plis. Les augmentations conventionnelles par défaut de nnU-Net ont été employées en plus de GBA, à savoir des rotations, redimensionnements, ajout de bruit gaussien, flou gaussien, changement de luminosité, changement de contraste, aliasing, correction gamma et effet miroir. Nous avons entraîné nos modèles pendant 500 époques excepté pour la dernière itération, entraîné durant 1000 époques. Entre chaque segmentation et après la dernière, seule la plus grande composante principale après activation softmax est conservée et considérée comme masque de segmentation de VS.

### 4.3.3 Expérimentations

En parallèle des résultats finaux du challenge, nous avons mené une série d'expériences afin de quantifier l'intérêt de GBA et sa version naïve, mais aussi de la stratégie d'auto-apprentissage. Pour cela, nous avons soumis une variété de résultats de segmentation intermédiaires au tableau de validation. Tout d'abord, nous avons ré-entraîné plusieurs modèles CycleGAN avec des paramètres identiques pour mesurer l'instabilité intrinsèque à son entraînement ainsi que son impact sur la segmentation en aval. Pour étudier l'influence de l'effet des deux centres, nous avons soumis des résultats obtenus en utilisant des modèles CycleGAN formés sur 1) uniquement *Londres* ($X_A^S, X_A^T$), 2) uniquement *Tilbourg* ($X_B^S, X_B^T$), ou 3) les deux centres ($X^S, X^T$). Nous avons ensuite évalué l'intérêt de nos augmentations spécifiques pour chaque ré-entraînement. Nous avons également analysé l'apport de SinGAN dans GBA par rapport à la version naïve. Pour éviter une consommation d'énergie inutile, les synthèses I2I ayant conduit à de très mauvaises performances de segmentation sans GBA, c'est-à-dire un score Dice moyen $\leq 0.4$, n'ont cependant pas été ré-entraînées avec GBA ; ces CycleGAN sont considérés comme des échecs, et leurs synthèses comme inutilisables pour n'importe quelle tâche de segmentation.

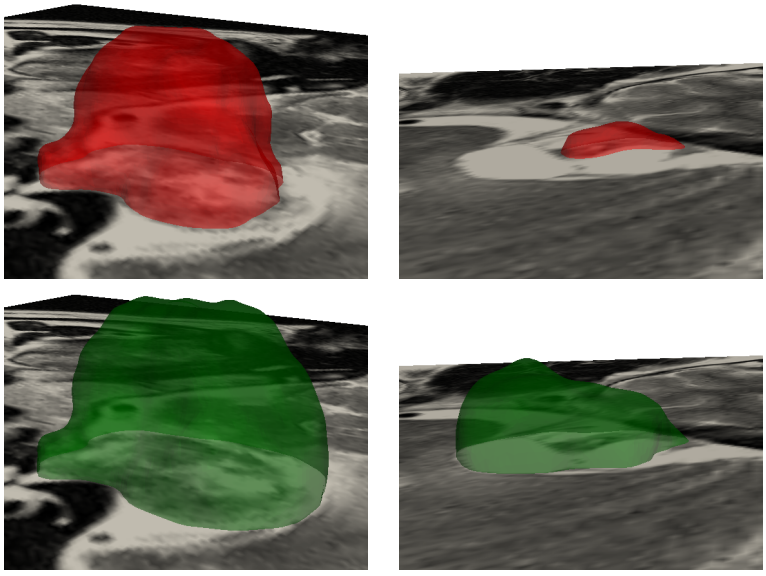
Nous avons également suivi l'évolution des scores de segmentation tout au long de l'auto-entraînement. Pour évaluer l'apport de l'harmonisation via SinGAN dans GBA,

nous avons également ré-entraîné le schéma itératif avec le même modèle CycleGAN et en utilisant notre augmentation naïve sans SinGAN. Les scores obtenus pour les plus petits et les plus gros VS ont enfin été précisés.

## 4.4 Résultats

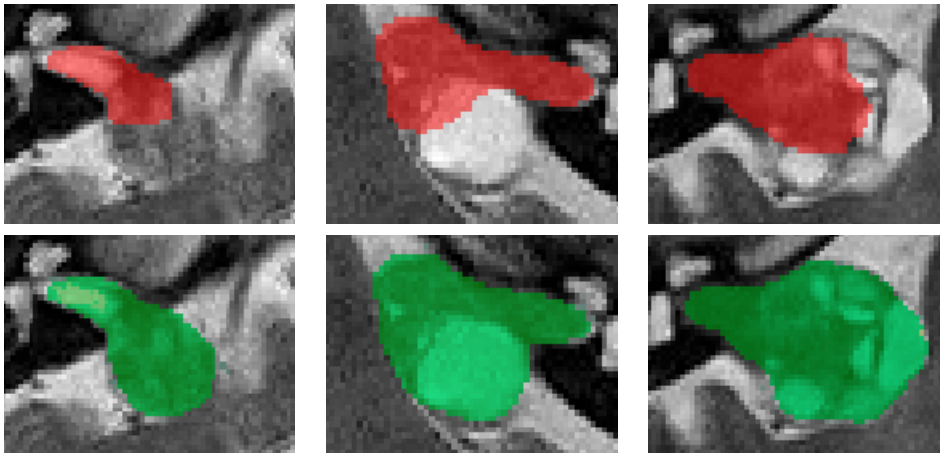
Les conclusions visuelles de notre augmentation naïve ou de GBA sont identiques à celles présentées sur ces mêmes données au chapitre précédent, en section 3.3.

Les Fig.s 4.8 et 4.9 montrent respectivement des représentations surfaciques des masques de segmentation et les masques 2D sur coupes axiales, obtenus avec ou sans GBA pour divers patients de l'ensemble de validation et après la première étape de segmentation. Sans GBA, des portions importantes du volume tumoral étaient perdues, contrairement à la segmentation basée sur GBA qui a capturé le VS de manière plus complète.



(a) (b)

FIGURE 4.8 – Rendu 3D des masques de segmentation prédits pour deux patients du jeu de validation, après le premier modèle de segmentation. Sans GBA en rouge, avec GBA en vert.



(a) (b) (c)

FIGURE 4.9 – Masques de segmentation 2D prédits pour trois patients du jeu de validation, après le premier modèle de segmentation. Sans GBA en rouge, avec GBA en vert.

| | Aug. de nn-UNet | | + aug. naïve | | + GBA | |
|---|---|---|---|---|---|---|
| Centres | DSC ↑ | ASSD ↓ | DSC ↑ | ASSD ↓ | DSC ↑ | ASSD ↓ |
| Londres seul #1 | $0.679 \pm 0.341$ | $7.853 \pm 16.750$ | $0.789 \pm 0.241$ | $3.267 \pm 10.738$ | $\mathbf{0.832 \pm 0.159}$ | $\mathbf{1.315 \pm 6.235}$ |
| #2 | $0.668 \pm 0.280$ | $3.653 \pm 10.857$ | $0.647 \pm 0.293$ | $2.185 \pm 6.341$ | $\mathbf{0.752 \pm 0.202}$ | $\mathbf{1.998 \pm 4.196}$ |
| #3 | $0.711 \pm 0.289$ | $2.856 \pm 8.662$ | $0.812 \pm 0.146$ | $0.667 \pm 0.887$ | $\mathbf{0.816 \pm 0.144}$ | $\mathbf{0.638 \pm 0.850}$ |
| Tilbourg seul #1 | $0.290 \pm 0.338$ | $21.060 \pm 34.855$ | N/A | N/A | N/A | N/A |
| #2 | $0.284 \pm 0.370$ | $38.803 \pm 44.943$ | N/A | N/A | N/A | N/A |
| #3 | $0.312 \pm 0.321$ | $18.989 \pm 33.111$ | N/A | N/A | N/A | N/A |
| Lon. + Til. #1 | $0.243 \pm 0.354$ | $27.559 \pm 22.062$ | N/A | N/A | N/A | N/A |
| #2 | $0.743 \pm 0.233$ | $2.116 \pm 7.419$ | $\mathbf{0.811 \pm 0.150}$ | $\mathbf{0.641 \pm 0.617}$ | $0.800 \pm 0.173$ | $1.688 \pm 8.628$ |
| #3 | $0.474 \pm 0.338$ | $7.714 \pm 15.357$ | $0.594 \pm 0.304$ | $\mathbf{2.743 \pm 7.347}$ | $\mathbf{0.612 \pm 0.304}$ | $4.060 \pm 10.349$ |

TABLE 4.10 – Résultats des segmentations sur le jeu de validation en fonction des configurations de données, des ré-entraînements et des augmentations. N/A indique que le score de la *baseline* est trop bas (Dice$\leq 0.4$ avec les augmentations de nnU-Net seules) et que la qualité de ces images n'est pas suffisante pour entraîner un modèle de segmentation pertinent.

### 4.4.1 Évaluation de la variabilité de CycleGAN

Le Tab. 4.10 résume les scores de segmentations atteints après le premier réseau de segmentation, entraîné avec pseudo-images et vrais masques seulement, en fonction des différents centres utilisés et dans 3 configurations (correspondant aux colonnes) différentes : 1) avec les augmentations de nnU-Net seules, 2) avec ces mêmes augmentations ainsi que l'altération naïve de contraste de tumeur, et 3) avec ces mêmes augmentations ainsi que GBA. Dans chaque cas, CycleGAN est ré-entraîné trois fois. Sans augmentation spécifique, c'est en utilisant les données du centre de *Londres* ($X_A$) seules que les résultats les plus stables après ré-entraînement ont été obtenus, le plus mauvais Dice étant de $0.668 \pm 0.280$. À l'inverse, avec les données provenant de *Tilbourg*, la synthèse apprise était erronée, générant des images irréalistes provoquant une mauvaise segmentation (DSC$\simeq$ 0.3, ASSD$\simeq$ 20 mm). En combinant les deux centres, les résultats étaient instables, avec des variations observées d'un ré-entraînement à l'autre allant jusqu'à 0.5 de Dice. Cela justifie *a posteriori* notre choix d'utiliser uniquement les données provenant de *Londres* pour l'apprentissage de CycleGAN. Cela confirme également qu'un décalage de domaines conséquent existe entre les deux centres, comme observé sur la Fig. 4.6, empêchant un modèle standard d'I2I d'apprendre cette synthèse. Avec nos augmentations, l'emploi de GBA ou de sa version naïve a systématiquement amélioré les deux métriques, avec des gains de Dice allant de 5 à 15%, excepté pour l'expérience "Londres seul #2" avec l'augmentation naïve, montrant une baisse de 2.1% de Dice. Ainsi, le Dice le plus haut a été atteint avec GBA durant l'essai "Londres seul #1", avec DSC=$0.832 \pm 0.159$ et ASSD=$1.315 \pm 6.235$, tandis que la meilleure ASSD durant l'essai "Londres seul #3", également avec GBA, avec ASSD=$0.638 \pm 0.850$ et DSC=$0.816 \pm 0.144$. Entre les deux augmentations, GBA a montré des performances légèrement supérieures pour les expériences utilisant les données de Londres seules, avec une amélioration allant de 0.4% à 10.5% de Dice et de 0.029 à 1.952 mm d'ASSD.

### 4.4.2 Segmentation itérative avec GBA

Le Tab. 4.11 évalue l'amélioration des scores de segmentation au fil des itérations d'auto-apprentissage avec pseudo-labels pour chaque métrique et dans les trois mêmes configurations d'augmentations. Les scores obtenus sur les 15 plus petites et les 15 plus grosses tumeurs (respectivement $< 720\text{mm}^3$ et $> 5400\text{mm}^3$) sont également précisés. Lors de la segmentation post-I2I (avant les itérations et avec seulement les images pseudo-

| Tumeurs de validations - Étape | Aug. de nn-UNet DSC ↑ | Aug. de nn-UNet ASSD ↓ | + aug. naïve DSC ↑ | + aug. naïve ASSD ↓ | + GBA DSC ↑ | + GBA ASSD ↓ |
|---|---|---|---|---|---|---|
| Toutes tumeurs - Post-I2I | $0.774 \pm 0.220$ | $2.673 \pm 9.795$ | $0.839 \pm 0.126$ | $0.555 \pm 0.586$ | $\mathbf{0.847 \pm 0.114}$ | $\mathbf{0.514 \pm 0.510}$ |
| Itération 1 | $0.780 \pm 0.219$ | $3.647 \pm 14.086$ | $0.859 \pm 0.067$ | $0.459 \pm 0.192$ | $\mathbf{0.862 \pm 0.065}$ | $\mathbf{0.444 \pm 0.182}$ |
| Itération 2 | $0.771 \pm 0.237$ | $4.711 \pm 16.324$ | $0.860 \pm 0.066$ | $0.455 \pm 0.189$ | $\mathbf{0.865 \pm 0.062}$ | $\mathbf{0.438 \pm 0.184}$ |
| Itération 3 | $0.823 \pm 0.132$ | $0.598 \pm 0.541$ | $0.864 \pm 0.062$ | $0.447 \pm 0.186$ | $\mathbf{0.868 \pm 0.061}$ | $\mathbf{0.431 \pm 0.177}$ |
| Petites tumeurs - Post-I2I | $0.763 \pm 0.100$ | $0.441 \pm 0.192$ | $0.762 \pm 0.112$ | $0.501 \pm 0.337$ | $\mathbf{0.781 \pm 0.080}$ | $\mathbf{0.400 \pm 0.144}$ |
| Itération 3 | $0.748 \pm 0.141$ | $0.482 \pm 0.303$ | $0.801 \pm 0.071$ | $0.366 \pm 0.123$ | $\mathbf{0.806 \pm 0.068}$ | $\mathbf{0.349 \pm 0.114}$ |
| Grosses tumeurs - Post-I2I | $0.786 \pm 0.174$ | $1.096 \pm 0.798$ | $0.890 \pm 0.043$ | $0.585 \pm 0.232$ | $\mathbf{0.893 \pm 0.038}$ | $\mathbf{0.571 \pm 0.198}$ |
| Itération 3 | $0.802 \pm 0.183$ | $0.991 \pm 0.867$ | $0.901 \pm 0.037$ | $0.537 \pm 0.206$ | $\mathbf{0.905 \pm 0.035}$ | $\mathbf{0.513 \pm 0.192}$ |

TABLE 4.11 – Scores de segmentation sur le jeu de validation aux différentes itérations de l'auto-entraînement. Parmi ces patients, les scores moyens obtenus pour les 15 tumeurs les plus petites ($< 720\text{mm}^3$) et les plus grosses ($> 5400\text{mm}^3$) sont également précisés.

cibles et les vraies annotations comme données d'entraînement), les augmentations naïves (DSC=0.839±0.126, ASSD=0.555±0.586, p-value = 0.044) et surtout GBA (DSC=0.847± 0.114, ASSD=0.514±0.510, p-value = 0.021) ont entraîné une nette amélioration des performances par rapport à la *baseline* nnU-Net (DSC=$0.774 \pm 0.220$). Ces scores restaient même supérieurs à la *baseline* après 3 itérations (DSC=0.823±0.132, ASSD=0.598±0.541, p-value = 0.57). Avec nos augmentations, les résultats ont été davantage améliorés pendant les 3 itérations de manière constante, alors que la *baseline* a perdu en performance entre les différentes itérations. Concernant nos deux augmentations, GBA a constamment obtenu de meilleurs résultats comparés à sa version naïve sans SinGAN, même si l'écart à chaque itération s'était réduit. Par exemple, les scores de Dice ont été améliorés de +0.8% après la première étape de segmentation et de +0.4% de Dice après la dernière itération d'auto-apprentissage pour un Dice maximal de $0.868 \pm 0.061$ et une ASSD minimale de $0.431 \pm 0.177$ mm. La différence entre les deux méthodes n'était cependant pas statistiquement significative selon un t-test appairé : p-value = 0.7. Des tendances similaires ont été observées sur les sous-ensembles des 15 plus petites ou plus grosses tumeurs, avec un gain de Dice supérieur à 10% pour les plus grosses tumeurs en utilisant une de nos augmentations et sans itération. Les différences de performances pour les 15 plus petites tumeurs sont plus fines avant itération, mais ont montré respectivement environ 5% et 6% de Dice en utilisant l'augmentation naïve et GBA après les trois itérations.

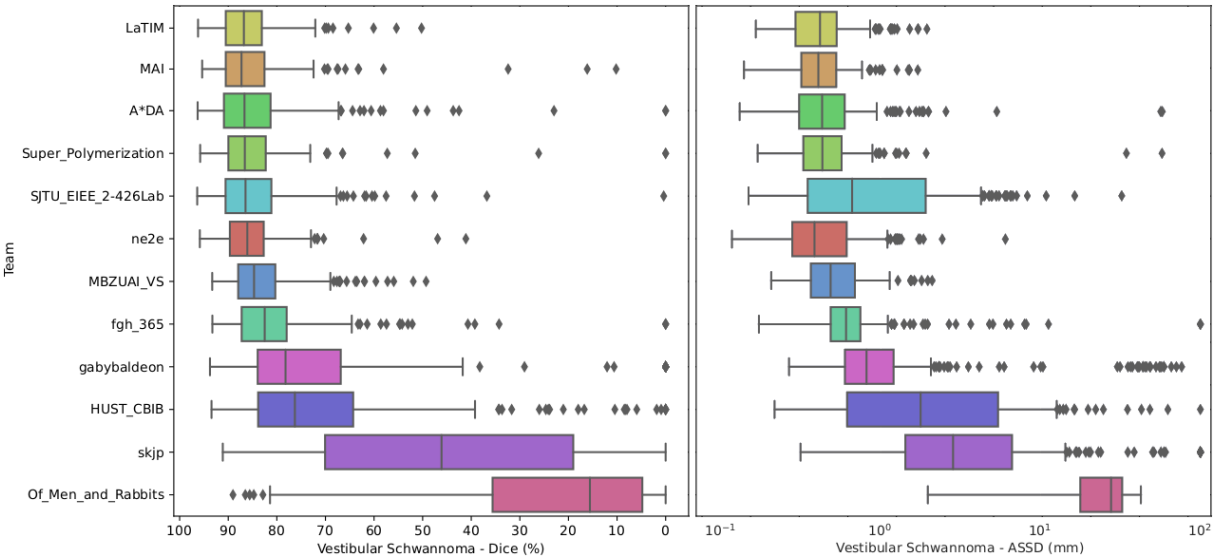



FIGURE 4.12 – Distribution des DSC et ASSD obtenus par les différents participants de CrossMoDA 2022. Fourni par les organisateurs [38], [195].

### 4.4.3 Comparaison avec les autres participants sur les données de test

La Fig. 4.12 présente les scores finaux des différents participants pour la tâche de segmentation de VS seule. Notre équipe, "LaTIM", a terminé à la première place de ce classement, avec un rang moyen de 3.661 associé à un Dice moyen de $0.859 \pm 0.066$ et une ASSD moyenne de $0.459 \pm 0.252$. Selon cette même métrique, la seconde place a été obtenue par l'équipe "MAI-LAB", avec un classement moyen de 3.862 associé à un Dice moyen de $0.852 \pm 0.094$ et une ASSD moyenne de $0.455 \pm 0.215$. Enfin, l'équipe "ne2e" a obtenu la troisième place sur cette tâche (classement moyen de 3.948, Dice moyen de $0.856 \pm 0.064$, ASSD moyenne de $0.515 \pm 0.469$). Au classement final de la compétition, prenant en compte la tâche de segmentation de VS mais aussi de cochlée, nous avons terminé à la troisième place, derrière les équipes "ne2e" et "MAI".

Les métriques quantitatives associées à notre méthode sur ces données de test sont résumées dans le Tab. 4.13. Le patient le moins bien segmenté a obtenu un DSC$\geq$ 0.5 et une ASSD$\leq$ 2 mm, ce qui indique que l'approche proposée a été capable d'identifier l'emplacement du VS au moins partiellement pour tous les sujets. Comparée aux autres participants, notre méthode semble donc plus robuste grâce à notre augmentation.

| LaTIM | Dice ↑ | ASSD (mm) ↓ |
|---|---|---|
| Moyenne | $0.859 \pm 0.066$ | $0.459 \pm 0.252$ |
| Pire | 0.502 | 1.945 |
| $1^{\text{er}}$ perc. | 0.637 | 1.573 |
| $5^{\text{ème}}$ perc. | 0.747 | 0.934 |
| $25^{\text{ème}}$ perc. | 0.831 | 0.530 |
| $50^{\text{ème}}$ perc. | 0.868 | 0.417 |
| $75^{\text{ème}}$ perc. | 0.905 | 0.293 |
| $95^{\text{ème}}$ perc. | 0.937 | 0.209 |
| $99^{\text{ème}}$ perc. | 0.947 | 0.178 |
| Meilleur | 0.962 | 0.165 |

TABLE 4.13 – Résultats finaux de notre méthode, obtenus sur le jeu de test de CrossMoDA 2022 (270 patients), incluant diverses statistiques.

## 4.5 Discussion

L'objectif principal de ce travail était d'étudier l'intérêt de GBA dans un contexte de décalage de domaines. La logique sous-jacente était d'exposer le réseau à un large éventail de tumeurs, avec des variations de contraste et de texture réalistes, afin de diminuer le décalage de domaine lors du potentiel déploiement du modèle. Cette approche semble donc particulièrement utile dans les scénarios où la distribution des tumeurs disponibles à l'entraînement risque de ne pas couvrir les différentes caractéristiques de tumeurs rencontrées au déploiement.

En segmentation inter-modale, les méthodes conventionnelles, basées sur la synthèse I2I suivie de la segmentation, sont généralement difficiles à entraîner en raison de la présence de décalages de domaine au sein des domaines source et cible eux-mêmes, comme c'est le cas dans le challenge CrossMoDA 2022. Dans ce contexte, les données des deux centres ne peuvent pas être considérés comme homogènes, rendant impossible l'apprentissage sensé d'une synthèse I2I avec des modèles conventionnels comme CycleGAN. En plus du décalage de domaines entre les deux centres, la faible quantité de coupes axiales 2D disponibles pour le second centre (14000 coupes pour *Tilbourg*, 30000 pour *Londres*) a probablement nuit à la capacité de généralisation apprise pour ce centre. Pour résoudre cela, nous avons limité la synthèse de domaine à un sous-domaine spécifique, dans notre cas le centre de *Londres A*, pour laquelle une synthèse satisfaisante pouvait être obtenue.

Nous avons ensuite appliqué une augmentation appropriée, GBA, afin d'exposer le réseau de segmentation en aval à diverses apparences de tumeurs, en particulier celles

attendues dans le second sous-domaine $B$ et manquante à la synthèse I2I. Naturellement, d'autres améliorations auraient pu être envisagées, qu'elles soient liées à la synthèse ou à notre manière d'augmenter les pseudo-images. Par exemple, le développement d'architectures de réseaux plus complexes pour la synthèse I2I pourrait permettre de réduire davantage le décalage de domaine et ainsi faciliter la segmentation, voire en les entraînant de bout en bout (*end-to-end*). Des augmentations spécifiques durant la synthèse pourraient également permettre cela. D'autres équipes performantes du challenge ont effectivement eu recours à des modèles I2I plus complexes, basés par exemple sur CUT [83], NICE-GAN [204], ou encore sur l'ajout de segmenteurs à l'intérieur du modèle de synthèse de domaines [180]). En théorie, notre méthode GBA pourrait également être combinée à ces approches. Cependant, notre objectif était d'explorer l'intérêt de GBA de manière indépendante en utilisant des modèles de synthèse et de segmentation conventionnels connus de la communauté. De ce point de vue, l'augmentation proposée a conduit à des améliorations significatives des performances de segmentation lorsqu'elle est employée en combinaison avec des méthodes d'augmentation de données conventionnelles, utilisées dans des modèles reconnus tels que nnU-Net. La version naïve de GBA, une altération simple mais grossière des contrastes des tumeurs sans variation réaliste, permettait déjà d'améliorer les performances de segmentation. Le fait de diversifier davantage les tumeurs altérées au sein de l'image cible en utilisant SinGAN a ensuite amélioré de manière constante les performances de segmentation en aval, bien que la différence n'était pas significative selon un test t apparié sur l'ensemble de validation constitué de 64 patients.

Nous avons soumis un certain nombre d'expériences à la validation en ligne pour confirmer l'importance de GBA dans notre approche. En raison du nombre limité de soumissions (une par jour) et du coût en ressources, et comme expliqué au chapitre 3, de nombreux paramètres ont été déterminés empiriquement afin d'optimiser qualitativement les images générées puis les distributions des tumeurs présentées sur les KDE de la Fig. 4.7. Même si plusieurs choix ont été décidés par rapport aux scores obtenus sur la validation, un ajustement plus fin (*fine-tuning*) des paramètres aurait probablement conduit à de meilleurs résultats. Par exemple, nous avons sélectionné nos valeurs scalaires d'altération de contraste $\lambda$ en inspectant la distribution des tumeurs dans l'ensemble cible à l'aide de pseudo-labels après la première étape de segmentation et à l'issue d'un entraînement particulier de CycleGAN, soit "Londres seul #1" dans le Tab. 4.10, ce qui est clairement sous-optimal. Cela explique dans ce tableau pourquoi cette expérience montre le plus gros gain avec GBA. D'une manière générale, trouver une meilleure heuristique pour adapter

ces paramètres en temps réel au cours de la procédure d'auto-entraînement permettrait probablement d'obtenir de meilleurs résultats.

Enfin, bien qu'une stratégie itérative auto-supervisée avec des pseudo-labels ait considérablement amélioré les résultats de segmentation, cette pratique est souvent limitée. Premièrement, le coût en ressources est fortement augmenté lors de l'utilisation d'auto-entraînement itératif. Avec GBA, nous avons considérablement amélioré les performances avant toute itération, réduisant le bénéfice de l'auto-entraînement grâce à GBA, une méthode d'augmentation bien moins coûteuse en ressources. Deuxièmement, les prédictions erronées réutilisées en aval, où les tumeurs sont partiellement ou totalement manquées, peuvent nuire aux performances du modèle final. Pour résoudre ce point, attribuer différents niveaux de confiances aux pseudo-labels pourrait réduire le risque d'employer des pseudo-annotations trop bruitées lors de l'apprentissage [146], [205]. C'est précisément cette limite qui peut expliquer l'instabilité de l'auto-entraînement lorsque seules les augmentations nnU-Net étaient utilisées dans le Tab. 4.11. Dans notre cas, l'altération naïve des tumeurs, tout comme GBA, ont chacune entraîné une augmentation statistiquement significative des performances de segmentation après le premier passage de segmentation par rapport aux augmentations nn-UNet. Bien que les performances entre GBA et sa version naïve n'étaient pas statistiquement significatives sur l'ensemble de validation constitué de seulement 64 patients, nous postulons que le fait d'exposer le réseau à une plus grande diversité de tumeurs réalistes à chaque itération a été déterminante au succès de notre approche.

Même si cela n'a pas été quantifié dans nos travaux, l'utilisation de GBA risque également d'amener à des faux positifs. Plus particulièrement, si le nombre d'augmentations est trop élevé ou que les paramètres $\lambda$ sont trop étendus, le modèle pourrait étendre sa représentation latente de ce qu'est une tumeur, et ainsi inclure d'autres structures lors de sa segmentation. Garder la plus grande composante principale permet alors de retirer les potentiels faux positifs, mais ce problème reste ouvert si plusieurs tumeurs peuvent exister au sein d'une même image.

## 4.6 Conclusion

Dans ce chapitre, nous avons présenté une application de notre méthode d'augmentation GBA dans le contexte du challenge CrossMoDA 2022 en segmentation inter-modale non-supervisée de tumeurs entre IRM ceT1 annotés et hrT2 non-annotées, tous deux d'ori-

gine multi-centrique. La méthode proposée, composée d'un modèle de synthèse I2I suivi de segmentation avec auto-entraînement itératif, s'est placé en première place du challenge CrossMoDA 2022 sur la tâche de segmentation de tumeurs. L'utilisation de GBA lors de la segmentation a permis de réduire le décalage de domaine tout en diversifiant les apparences de tumeurs après la synthèse. La différence entre GBA et sa version naïve est sensible, avec un avantage faible mais constant pour la version profonde. L'utilisation de SinGAN a donc été un facteur clé de notre réussite à ce challenge.

Dans le chapitre suivant, nous abordons d'autres applications de GBA dans des travaux exploratoires, en segmentation de métastases cérébrales sur images TDM ainsi qu'en détection de nodules sur radiographie pulmonaire.

Chapitre 5

# Applications exploratoires en tomodensitométrie cérébrale et radiographie thoracique

**Résumé**

Ce chapitre présente deux applications exploratoires montrant le potentiel de GBA et de l'harmonisation *one-shot*. Tout d'abord, nous évaluons à quel point il est possible de réduire le nombre d'images d'entraînement sans dégradation des résultats. Pour cela, nous appliquons GBA et sa version naïve pour la segmentation de métastases cérébrales en imagerie TDM, une tâche difficile, à très bas régime de données. Nous montrons qu'il est possible de réduire drastiquement le nombre d'images à l'entraînement sans perte de performances à l'aide de GBA. Ensuite, dans le cadre du challenge NODE21, nous tâchons de synthétiser des nodules sur radiographie pulmonaire 2D à partir de patchs de nodules 3D en TDM pour améliorer leur détection. Après projection des coupes coronales et adaptation des contrastes, la méthode traditionnelle propose d'employer *Poisson blending* pour intégrer le nodule dans une image donnée. Nous remplaçons cette technique par l'harmonisation utilisée dans GBA afin d'intégrer de manière plus réaliste la tumeur. Même si la méthode de référence a fourni de meilleures résultats, l'emploi d'un modèle *one-shot* conserve un potentiel intérêt.

## 5.1 Segmentation à bas régime de données de métastases cérébrales sur imagerie TDM

Nous avons vu dans le chapitre précédent une première application de GBA en segmentation inter-modale afin de réduire un écart de domaines. Au vu du réalisme des images générées avec cette méthode, nous postulons qu'il est possible de réduire le nombre d'images disponibles à l'entraînement sans perte de performance en employant GBA. En effet, en variant les apparences et contrastes des ROI, il est possible de synthétiser une variété conséquente d'images réalistes représentant la distribution de ROI souhaitée. Même si notre méthode n'inclut pas de variations de l'anatomie de chaque patient, nous posons l'hypothèse qu'à bas régime de données, il est plus intéressant de varier les tumeurs que les anatomies. Ainsi, nous évaluons, dans cette section, notre augmentation dans un contexte de segmentation difficile, pour lequel peu d'images pertinentes sont disponibles à l'entraînement : la segmentation de métastases cérébrales en imagerie TDM. Ce travail a été présenté à la conférence *IEEE NSS-MIC 2022*.

### 5.1.1 Contexte clinique et problématique

Une métastase cérébrale est une tumeur maligne. Elle apparaît dans le cerveau lorsqu'un autre organe est atteint d'un cancer, créant une tumeur via le processus de cascade métastatique [206]. Les symptômes diffèrent en fonction de la zone du cortex atteinte, comme des maux de tête persistants, nausées, crises d'épilepsie ou encore troubles fonctionnels et comportementaux. En présence de métastases cérébrales, la routine clinique pour un traitement par radiothérapie stéréotaxique est généralement l'acquisition d'une IRM pour la segmentation des tumeurs, d'un scan TDM pour le calcul de dose à délivrer, ainsi que d'une tomodensitométrie à faisceau conique (CBCT, *cone beam computed tomography*) avant chaque traitement pour positionner le patient [207]. Le scan TDM étant nécessaire pour mesurer l'atténuation des rayons dans les différents tissus, nous proposons d'essayer de segmenter les tumeurs sur cette même modalité afin de simplifier le parcours patient. Cependant, les métastases peuvent être très difficiles à identifier et à délimiter sur cette modalité, notamment en raison de leur taille parfois très réduite, de leur emplacement ou de leur composition tissulaire. En parallèle, d'autres sont parfaitement identifiables sur cette modalité, comme présenté sur la Fig. 5.1. En pratique, l'apparence des métastases sur imagerie TDM est très variable d'un patient à l'autre.

Étant donné la difficulté de la tâche, les modèles de détection ont principalement été évalués par la communauté. Sur images TDM avec injection de contraste, les performances sont encourageantes, avec des résultats acceptables pour les lésions de plus de 3 mm de diamètre [208]. Cependant, sans injection, un certain nombre de tumeurs de diamètre inférieur à 6 mm ne sont pas reconnus [209]. Pour cette raison, l'IRM reste privilégiée pour la segmentation des métastases [207]. Cependant, la segmentation sur images TDM permettrait, pour certains patients, de se passer d'acquisition IRM. De plus, les métastases évoluant rapidement, la segmentation pourrait aussi être adaptée ou corrigée si un temps conséquent s'est écoulé entre l'acquisition IRM et l'imagerie TDM. À terme, la segmentation précise de tumeurs sur CBCT permettrait d'affiner la localisation de la tumeur au moment du traitement, tendant donc vers la radiothérapie adaptative.

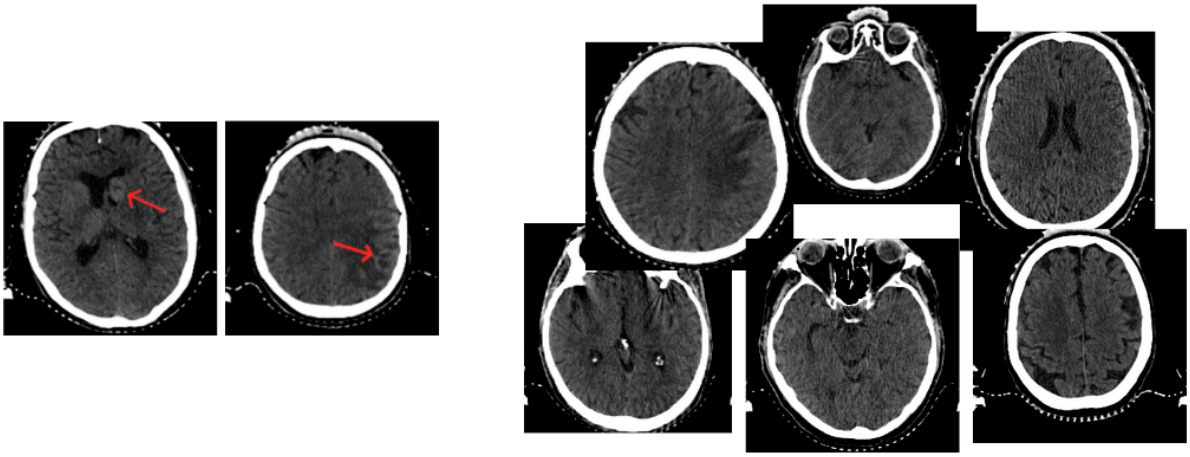


FIGURE 5.1 – Exemple d'images TDM où la tumeur est visible (à gauche) ou non (à droite) à l'œil humain.

### 5.1.2 Augmentation réaliste

Dans ce contexte, nous disposons d'un certain nombre d'acquisitions TDM pour l'apprentissage. Sur une majorité d'entre elles, les métastases ne semblent pas visibles à l'œil nu. Nous choisissons donc de restreindre l'apprentissage à des patients pour lesquelles les tumeurs sont clairement identifiables. Ces images sont ainsi sélectionnées manuellement avant entraînement d'un modèle de segmentation, réduisant radicalement le nombre d'images disponibles à l'apprentissage. Nous cherchons alors à tester le potentiel de GBA

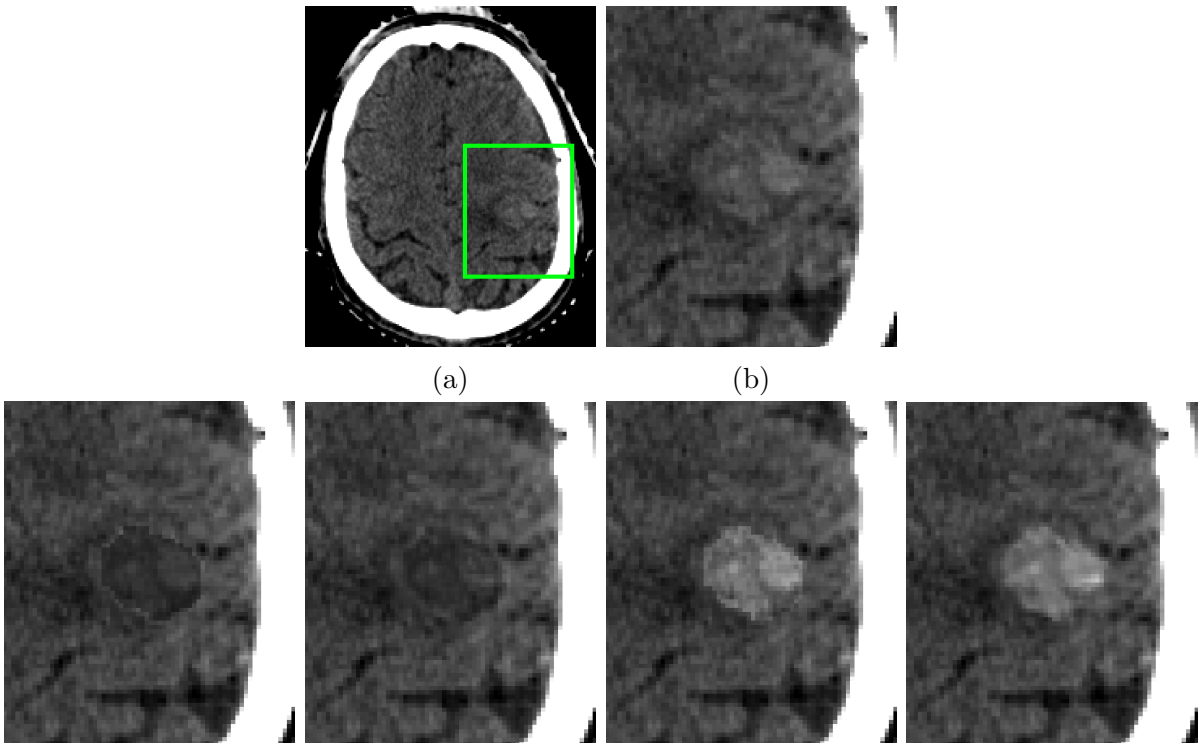
et de sa version naïve pour diversifier ces tumeurs et améliorer la robustesse et la généralisation du modèle de segmentation. Nous testons ainsi deux régimes de données, $N = 10$ et $N = 32$ images, sans puis avec GBA ou sa version naïve.

### 5.1.3 Contexte de validation

Le jeu de données est privé et composé de 184 scans TDM de patients atteints de métastases cérébrales [210]. Ces images, collectées dans le contexte de radiothérapie stéréotaxique au CHRU de Brest de 2014 à 2018, sont manuellement segmentées sur des IRM co-enregistrées et alignées aux TDM afin de produire des masques du volume tumoral macroscopique (GTV, *gross target volume*). Les types histologiques des patients sont principalement le cancer du sein (61%), le cancer des poumons (16%) et le mélanome (11%). 62% des patients présentent une seule métastase, les autres en présentant entre 2 et 6. L'échantillonnage spatial des images TDM d'origine est compris entre $0.7 \times 0.7 \times 1.5$ mm$^3$ et $1.0 \times 1.0 \times 1.5$ mm$^3$, avec une taille axiale de $512 \times 512$ pixels. Avant augmentation ou passage dans le modèle, les images sont seuillées selon le fenêtrage $[0, 150]$ unité Hounsfield pour améliorer le contraste des tissus mous. Une sous-image (ou *crop*) centré sur le cerveau est ensuite extraite, de taille $256 \times 256 \times z$ où $z$ représente le nombre de coupes axiales 2D.

Le modèle de segmentation est nnU-Net, entraîné pendant 500 époques sur un unique pli (ou *fold*), avec l'emploi d'augmentations de données conventionnelles [21]. Au vu du nombre d'images présentes et du caractère expérimental de ces travaux, l'entraînement de cinq *folds* avec *ensembling* n'est pas adapté. [1] La métrique d'évaluation choisie est le Dice. Dans chaque régime de données ($N = 10$ et $N = 32$), la totalité des images ont été augmentées via GBA ou sa version naïve quatre fois, avec $\lambda \in \{0.6, 0.8, 1.2, 1.4\}$, ainsi que $K = 12$ et $k^* = 3$ dans le cas de GBA. Ces paramètres ont été manuellement choisis afin de générer des tumeurs réalistes à l'œil humain. Nous utilisons les 152 images restantes pour la validation. Afin d'évaluer la segmentation uniquement pour les patients où la métastase est identifiable, nous retirons toutes les images n'atteignant jamais 0.2 de Dice. Cette sélection a donc réduit l'évaluation à un ensemble de 61 images.

1. Nous discutons de ce choix dans l'annexe Évaluation environnementale et bilan carbone.



(a) (b)

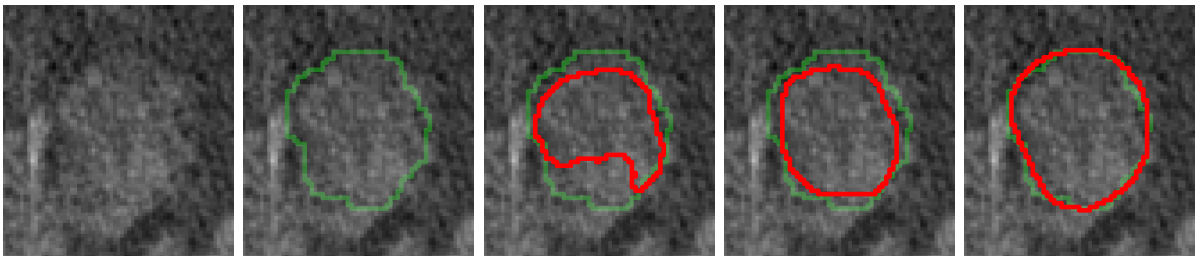
(c) Naive, $\lambda = 0.6$ (d) GBA, $\lambda = 0.6$ (e) Naive, $\lambda = 1.4$ (f) GBA, $\lambda = 1.4$

FIGURE 5.2 – Résultats de l'augmentation naïve et profonde de métastase cérébrale sur images TDM. (a) Image TDM (b) zoom sur la tumeur (c,d,e,f) tumeurs augmentées avec les deux méthodes et différents $\lambda$.

### 5.1.4 Résultats

La Fig. 5.2 présente une tumeur d'une acquisition TDM du jeu d'entraînement et ses versions augmentées avec $\lambda = 0.6$ et $\lambda = 1.4$. Les images générées par GBA étaient globalement réalistes, à la fois dans le cas $\lambda < 1$ et $\lambda > 1$, malgré de légers artefacts en hypersignal induit par SinGAN, comme sur l'image représentée sur la Fig. 5.2f. Concernant l'augmentation naïve, l'image 5.2e était également assez réaliste, tandis que les contours du masque de segmentation étaient identifiables sur l'image de la Fig. 5.2c, nuisant au réalisme de celle-ci.

La Fig. 5.3 montre un résultat de segmentation sur un patient de l'ensemble de test, sans et avec GBA pour les deux régimes de données. Pour ce patient, le modèle entraîné avec 10 images sans GBA n'avait pas reconnu la tumeur. L'emploi de GBA avait alors permis sa reconnaissance partielle, avant d'être nettement amélioré avec 32 images à l'entraînement.



(a) (b) (c) (d) (e)

FIGURE 5.3 – Résultats qualitatifs de segmentation, extraits de l'ensemble de test, en fonction des différents régimes de données. (a) scan TDM d'origine, (b) $N = 10$ sans GBA, (c) $N = 10$ avec GBA, (d) $N = 32$ sans GBA, (e) $N = 32$ avec GBA. Les contours verts représentent la vérité-terrain, tandis que les contours rouges correspondent aux différentes méthodes.

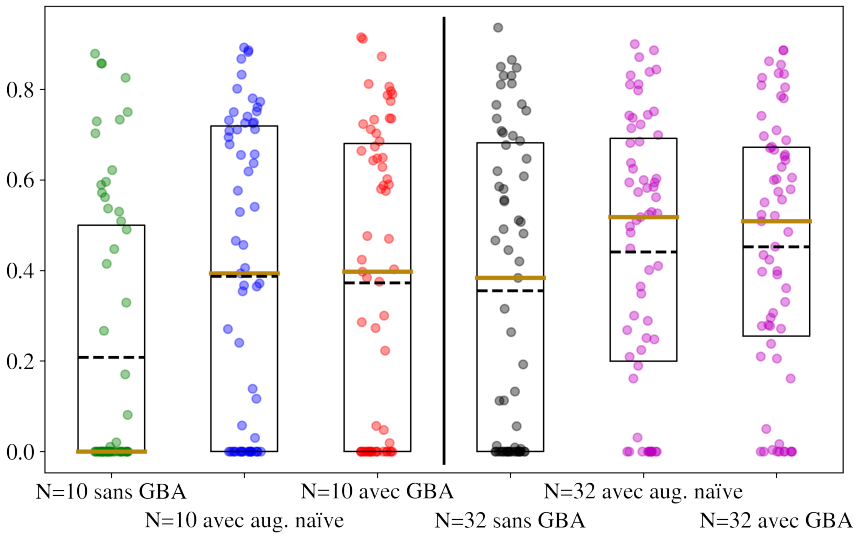



FIGURE 5.4 – Diagramme moustache des scores de segmentation en fonction des différents régimes de données. Dans chaque sous-diagramme, le trait en pointillé et le trait plein jaune représente respectivement la moyenne et la médiane des scores obtenus.

La Fig. 5.4 présente un diagramme moustache en fonction des deux régimes de données et des différentes augmentations appliquées : avec les augmentations conventionnelles de nnU-Net seules, avec l'ajout de notre augmentation naïve, ou l'ajout de GBA. Le Tab. 5.5 résume les scores obtenus dans chaque cas. Tout d'abord, la meilleure moyenne de Dice obtenue était de $0.46 \pm 0.28$, avec $N = 32$ et GBA. Nous avons également observé un certain nombre de patients ayant un Dice supérieur à 0.8, témoignant d'une segmentation réussie. Ensuite, dans les deux régimes de données, l'emploi d'une des versions de GBA a considérablement amélioré les performances de segmentation et réduit le nombre de cas d'échec, passant de 40 à 23 dans le cas $N = 10$, et de 27 à 13, voire 11 selon l'augmentation. Plus précisément, les performances obtenues avec $N = 10$ avec l'une de nos augmentations étaient similaires à celles obtenues avec $N = 32$ sans augmentation spécifique. Enfin, les résultats obtenus avec la version naïve ou profonde de GBA étaient très proches, avec une médiane plus élevée et une moyenne plus basse pour GBA dans le cas $N = 10$, et inversement dans le cas $N = 32$.

| Configuration | Dice (25%, 50%, 75%) | Nombre d'échecs |
|---|---|---|
| $N = 10$ avec aug. de nnU-Net | $0.21 \pm 0.30$ (0, 0, 0.51) | 40 |
| $N = 10$ + aug. naïve | $0.39 \pm 0.32$ (0, 0.39, **0.73**) | 23 |
| $N = 10$ + GBA | $0.38 \pm 0.33$ (0, 0.40, 0.69) | 23 |
| $N = 32$ avec aug. de nnU-Net | $0.37 \pm 0.33$ (0, 0.42, 0.69) | 27 |
| $N = 32$ + aug. naïve | $0.44 \pm 0.30$ (0.20, **0.53**, 0.71) | 13 |
| **N = 32 w/ aug (ours)** | **0.46 ± 0.28 (0.28,** 0.52**,** 0.67**)** | **11** |

TABLE 5.5 – Scores de segmentation sur l'ensemble de test.

### 5.1.5 Discussion

La tâche de segmentation de métastases cérébrales en imagerie TDM est particulièrement complexe. À travers ces travaux, nous avons cherché à évaluer la pertinence de cette tâche et l'intérêt de diversifier les ROI. Les performances globales sont assez basses, avec des moyennes de Dice inférieure à 0.5. Pour un nombre restreint de patients, nous avons observé des Dices supérieurs à 0.8, témoignant d'une segmentation assez fidèle à la vérité-terrain pour ces patients. La méthode de validation que nous avons employée sous-évalue leur proportion, car nous avons préalablement retiré les 32 patients où les tumeurs étaient les mieux reconnaissables à l'œil nu. Parmi les 152 patients de validation, on dénote principalement des tumeurs assez difficiles à segmenter. En approfondissant

ces travaux et en associant des scores de confiance aux segmentations générées, on peut imaginer un affinement de la segmentation des métastases juste avant le traitement par dose. Cependant, parmi ces mêmes 152 patients, 91 d'entre eux ont été exclus, car ils n'atteignaient jamais 0.2 de Dice dans au moins une expérience. L'intérêt clinique de cette tâche serait donc, en l'état, plus que limité. Le nombre d'images à l'entraînement devrait tout d'abord être augmenté, et les paramètres d'utilisation de GBA devraient être affinés en fonction.

L'utilisation des versions naïve comme profonde de GBA montrent tout de même un gros intérêt pour améliorer les performances de segmentation dans un contexte difficile en imagerie TDM. Comme observé sur la Fig. 5.4, nous montrons qu'il est possible de réduire le nombre de données lors de l'entraînement, de 32 à 10, lorsqu'une de nos augmentations est employée. GBA permet donc de réduire la dépendance en données dans les cas où leur nombre est particulièrement limité, en employant 80% d'images augmentées dans l'ensemble d'entraînement. Cependant, la différence entre les versions naïve et profonde est très faible. Aucune des méthodes ne semble dépasser l'autre au vu des résultats obtenus. Cela montre qu'à bas régime de données, la qualité de l'harmonisation n'a qu'une très faible influence sur la qualité des segmentations générées.

## 5.2 Synthèse et détection de nodule pulmonaire sur radiographie

Dans le chapitre 3, nous avons comparé la méthode d'harmonisation traditionnelle sans DL *Poisson blending* et l'harmonisation via SinGAN pour l'intégration de tumeur dont le contraste a été altéré. Nous avons observé que le *Poisson blending* ne modifiait pas le contenu des tumeurs et était à peine plus réaliste qu'une intégration naïve sans harmonisation. Nous postulons donc que la méthode de *Poisson blending* peut être remplacée par SinGAN dans les contextes où elle est communément employée [189], [190]. C'est notamment le cas en simulation de nodules sur radiographie pulmonaire à partir de patchs 3D de nodules en imagerie TDM [211]. Nous évaluons donc, dans cette section, le potentiel de l'harmonisation *one-shot* de tumeurs dans le cadre de la synthèse de nodules. Cette méthode a été proposée pour la compétition NODE21.

### 5.2.1 Contexte clinique

Un nodule est une tumeur, bénigne ou maligne, localisée dans les poumons [212]. Sa détection doit être la plus précoce possible pour traiter un potentiel cancer des poumons et augmenter les chances de guérison. Elle est généralement effectuée sur image TDM ou radiographie pulmonaire. En raison de son faible coût et de la faible dose reçue par le patient, cette seconde modalité est privilégiée. Elle est également plus courante que l'examen TDM. Cependant, la détection de nodule est plus difficile sur radiographie thoracique que sur acquisitions TDM [213]. La Fig. 5.6 montre quelques exemples types de nodules sur radiographies. Des algorithmes de détections, basés par exemple sur *Faster R-CNN*, ont donc été développés, principalement en imagerie TDM avec injection de produit de contraste ainsi que sur radiographie [214]-[216]. Afin d'améliorer les performances en radiographie, la synthèse de nodules sur radiographie à partir de nodules TDM et sans DL est employée en tant qu'augmentation de données [211]. Cette technique a été définie comme *baseline* dans la tâche de génération du challenge NODE21 afin d'encourager la communauté à faire évoluer cette méthode.

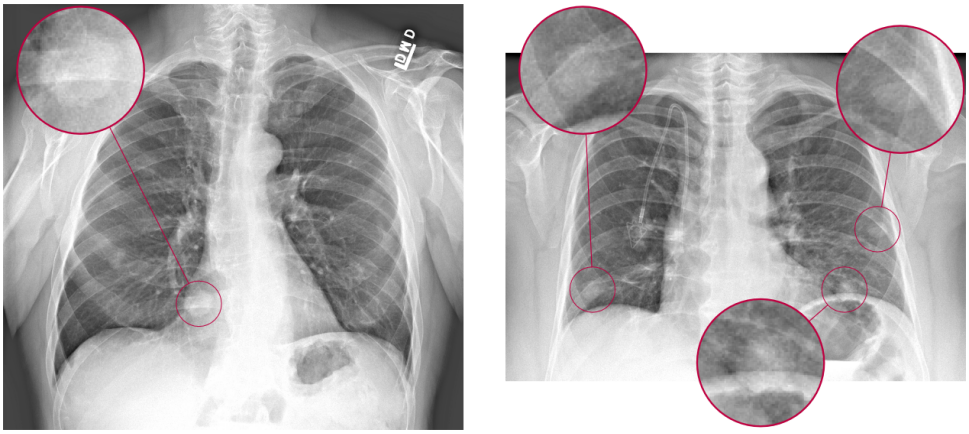


FIGURE 5.6 – Deux exemples de radiographies présentant des nodules pulmonaires, avec zoom sur ces derniers.

### 5.2.2 Méthodologie

Dans ce contexte, nous proposons une variation de la méthode de référence, basée sur *Poisson blending*, en appliquant le modèle *one-shot* SinGAN pour l'harmonisation.

### Méthode de référence

La méthode de référence, sans DL, part de l'hypothèse qu'une image TDM projetée dans le plan coronal est similaire à une radiographie frontale [211]. Pour chaque radiographie à augmenter en simulant un nodule dans une certaine boîte englobante (*bounding box*), un patch de nodule TDM 3D est sélectionné parmi ceux ayant un rayon similaire à la taille de la boîte. Après ré-échantillonage isotropique et sans distorsion, le nodule est projeté dans le plan 2D frontal. Le contraste global est ensuite modifié afin que les pixels correspondent aux valeurs types rencontrées sur radiographe pulmonaire et que le nodule présente un contraste compatible avec l'arrière-plan de la boîte englobante. Le nodule 2D est enfin intégré dans la boîte englobante à l'aide de la méthode de *Poisson blending* pour intégrer le nodule dans l'image saine [188].

### Emploi de SinGAN

Nous proposons d'adapter cette méthode en remplaçant *Poisson blending* par l'harmonisation effectué par SinGAN dans notre méthode d'augmentation GBA. Avant cette étape, en raison de la nature projective de la radiographie par rayon X, nous mélangeons le nodule à l'information déjà présente dans l'image. Ainsi, les tissus superposés au nodule sont atténués, mais restent présents dans la radiographie augmentée. Toutes les autres étapes sont identiques à la méthode de référence. Ainsi, le contraste souhaité est choisi selon la méthode de référence, différant de la méthode GBA proposée au chapitre 3.

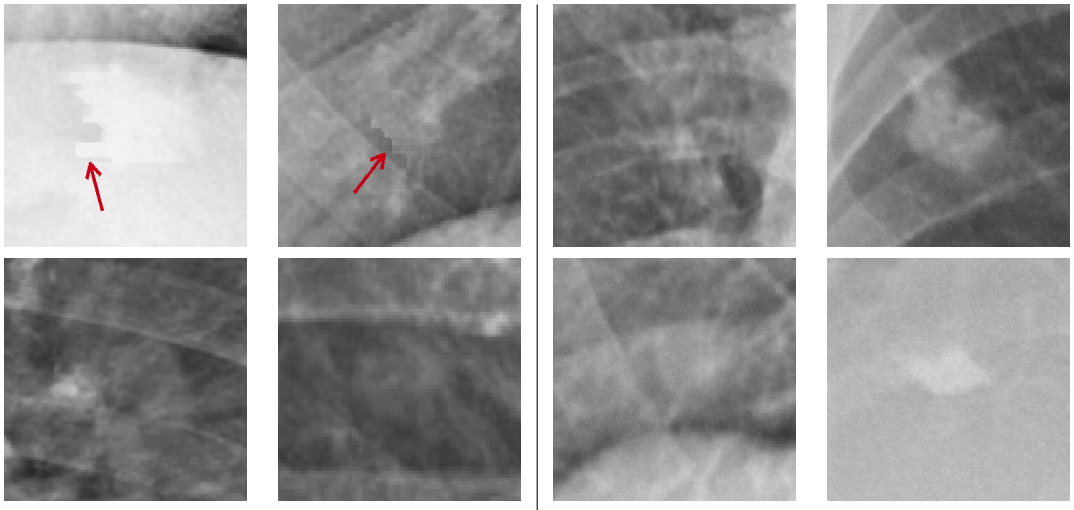
Après évaluation visuelle des nodules réels, nous observons qu'une grande variété d'apparence de nodules est présente. Nous avons notamment remarqué que ceux situés dans les régions inférieures des poumons, ayant subi une atténuation plus élevée, présentent généralement un hypersignal spécifique à ces régions. Nous distinguons donc deux catégories de nodules, associées à deux entraînements de SinGAN distincts. La décision du choix des poids à utiliser repose alors sur un seuillage Otsu en 4 classes, une méthode de *clustering* traditionnelle [217]. À partir de l'histogramme de l'image, les pixels sont répartis dans chaque classe en minimisant la variance intra-classe. Les poids du modèle SinGAN d'intensité plus élevée sont alors sélectionnés lorsque la majorité des voxels de la boîte englobante appartiennent à la classe d'intensité supérieure.

### 5.2.3 Contexte de vadidation

Les organisateurs du challenge NODE21 ont fourni 4882 radiographies, dont 3748 images saines, toutes de taille $1024 \times 1024$ pixels. Un ensemble de 1186 nodules TDM 3D et leurs masques de segmentation correspondants était également disponible. Visuellement, les radiographies montrent des contrastes et bruits très différents d'une image à l'autre. Dans le cadre de la compétition, les images à augmenter et les coordonnées des boîtes englobantes ont été figées. Ainsi, 1200 images saines ont été augmentées à l'aide de notre méthode.

Afin de classer les méthodes, un réseau de détection est entraîné pour chaque participant à la compétition. Même si nous postulons qu'il est nécessaire de mélanger images réelles et générées pour limiter le risque de sur-apprentissage, chaque modèle a été entraîné uniquement avec des images générées, sans aucune image réelle. Ainsi, les organisateurs ont choisi de séparer ces 1200 images en deux groupes : 1000 pour l'entraînement, ainsi que 200 pour la sélection des meilleurs poids lors de l'entraînement. Afin d'évaluer les différentes méthodes pendant la compétition, un ensemble d'images composé de 200 radiographies saines et de 200 contenant de vrais nodules, toutes différentes de celles vues à l'entraînement, sont utilisées pour évaluer le modèle après entraînement. 298 radiographies sont également restées privées pour comparer et classer les différents participants à l'issue de la compétition, sans précision sur la proportion d'images saines et malsaines. Les participants ont été classés en fonction de deux métriques, liées à la classification de chaque image comme saine ou non (75% de la métrique finale), et à la précision des boîtes englobantes prédites (25% de la métrique finale). Les deux tâches étant liées, nous évaluons ici uniquement les performances de la première métrique, l'aire sous la courbe ROC (AUC, *area under the ROC curve*).

SinGAN a été entraîné de manière équivalente à son utilisation dans GBA, deux fois sur deux sous-images de taille $256 \times 256$ pour éviter les dépassements de capacité des GPU, avec $r = 0.85, K = 16, k^* = 3$. Afin de renforcer le réalisme des nodules générés tout en réduisant le risque de sur-apprentissage, nous avons ajouté du bruit gaussien avec $\mu = 0, \sigma = 3$ après harmonisation. Le *Faster R-CNN* utilisé pour évaluer la méthode a été entraîné durant 1000 époques. Les poids finalement conservés sont ceux ayant obtenu l'AUC la plus élevée sur le sous-ensemble de validation.



(a) (b) (c) (d)

FIGURE 5.7 – Résultats visuels de la méthode de référence (colonnes (a) et (b)) et de la variante basée sur SinGAN (colonnes (c) et (d)).

### 5.2.4 Résultats

La Fig. 5.7 présente quelques images augmentées par la méthode de référence et celle proposée. Qualitativement, la *baseline* a parfois généré plusieurs artefacts, pointés par les flèches rouges sur la figure. Ces derniers étaient dûs aux contours de certains nodules combinés à une mauvaise intégration par *Poisson blending*. Aucun artefact de ce type n'a été observé dans les images générées par notre méthode. Notre synthèse présentait des nodules globalement mieux fondus, sans aucun contours apparents. En dehors de cet aspect, les deux augmentations ont généré une large palette de nodules réalistes, autant en termes de forme que de contraste.

| | Harmo. avec *Poisson blending* | Harmo. avec SinGAN |
|---|---|---|
| AUC | $\mathbf{0.71 \pm 0.03}$ | $0.68 \pm 0.02$ |

TABLE 5.8 – Scores de classification de la méthode de référence et de la variante proposée incluant SinGAN, obtenus sur images réelles.

Cependant, les résultats quantitatifs obtenus sur images réelles, présenté dans le Tab. 5.8, étaient plus faibles avec notre méthode. Nous avons observé une différence d'AUC de 0.03 sur un total de 400 images. Ces performances étaient inattendues au vu des résultats visuels précédemment décrits. Nos performances se sont ensuite effondrées sur le jeu de test resté privé, avec une AUC de 0.58. De plus, lors des expérimentations, nous avons

observé une grande instabilité d'apprentissage, résultant en des AUC très variables.

### 5.2.5 Discussion

La raison principale de cette instabilité et de la perte de performance est probablement la grande variété de bruit existant dans les radiographies pulmonaires à augmenter. SinGAN étant entraîné sur une unique image, il n'a pas appris les différents bruits et textures présents dans l'ensemble des données. Il est donc incapable de reproduire avec précision la diversité du bruit présent dans tout l'ensemble, ni de l'adapter au cas par cas. On peut par exemple observer cette différence sur la deuxième image de 5.7d, où le nodule est moins bruité que les projections environnantes. De plus, le modèle de détection étant entraîné uniquement sur images augmentées, i.e. aucun nodule réel, le risque de sur-apprentissage est considérablement plus élevé qu'un entraînement mélangeant les deux. Nous pensons donc que le modèle a davantage appris à reconnaître les variations de bruit au sein des images d'entraînement plutôt que les caractéristiques propres aux nodules. En pratique, l'entraînement de la détection avec nos augmentations convergeait plus rapidement que celui de la *baseline*, ce qui semble confirmer cette hypothèse. Dans ce contexte, la méthode de *Poisson blending* semble donc plus appropriée qu'un modèle *one-shot* comme SinGAN. Les artefacts générés par la *baseline* n'ont a priori pas nui à l'entraînement de la détection.

Une première solution serait d'entraîner le modèle sur un mélange de nodules réels et générés afin de réduire le risque de sur-apprentissage. Cependant, dans le cadre de la compétition, le modèle de détection est entraîné par les organisateurs, ne nous permettant pas d'influer sur les images employées à cette étape. Nous pensons tout de même que cette modification de validation pourrait avoir un impact significatif sur les performances de notre méthode. Une seconde solution pourrait être d'augmenter le nombre d'entraînements de SinGAN, en fonction des différents types de bruit observés. Cela mènerait à une grande distinction de cas, qui risque toujours de ne pas particulièrement profiter au modèle. De plus, le coût en temps et ressources d'un plus grand nombre d'entraînements de SinGAN, comme abordé en section 2.4, ne représente pas la démarche de nos travaux. Enfin, une dernière solution serait d'entraîner un unique modèle sur l'ensemble des images malsaines disponibles. L'équipe remportant le challenge a d'ailleurs choisi cette approche, employant un modèle d'*inpainting* [218]. Leur méthode consiste à apprendre le remplissage réaliste des boîtes englobantes en contraignant la texture générée [219]. Ils n'ont ainsi pas utilisé les patchs de nodules TDM fournis et ont ajouté des images publiques de radiographies

thoraciques externes au challenge durant l'entraînement.

L'utilisation de SinGAN pour l'harmonisation d'une structure sur une modalité d'image projective, où *Poisson Blending* est communément employé, montre une première limite du modèle. Même si le protocole de validation montre plusieurs faiblesses, la disparition du bruit spécifique aux images est critique, notamment lorsque aucun nodule réel n'est employé à l'entraînement.

## 5.3 Conclusion

Nous avons présenté dans ce chapitre deux travaux exploratoires, sur deux modalités différentes, afin d'évaluer divers cas d'usages potentiels de GBA et ses limites.

Tout d'abord, la segmentation de métastases cérébrales en imagerie TDM tend à montrer qu'il est possible de réduire le nombre de données d'entraînement sans perte de performance. Les performances des versions naïves et profondes de GBA étant très proches, l'harmonisation n'a que peu d'influence à bas régime de données. Une altération des intensités des voxels tumoraux sans *blending* particulier permet alors une amélioration conséquente des performances.

Ensuite, la détection de nodules pulmonaires sur radiographie pulmonaire témoigne du besoin de mélanger images réelles et générées, afin de limiter les biais liés à la validation croisée. La variété de bruit présent dans les images d'entraînement et de test est également à prendre en compte lors de l'utilisation de GBA ou plus généralement de la fonction d'harmonisation de SinGAN.

# Conclusion générale et perspectives

Dans ce travail de thèse, nous avons proposé une stratégie d'augmentation de données pour la segmentation d'images biomédicales, GBA, se décomposant 1) en une méthode naïve d'altération des contrastes pathologiques suivant la distribution attendue dans le domaine de déploiement, et 2) en une harmonisation fine de l'apparence de ces structures altérées, adaptée au domaine de déploiement par un modèle génératif profond de type SinGAN. Ces contributions permettent de réduire l'impact de plusieurs biais intrinsèques aux données pour l'apprentissage profond, notamment le manque et le décalage de données au niveau des ROI. Nous favorisons l'adéquation de l'apparence des ROI du domaine source d'entraînement avec celle du domaine cible en altérant l'intensité et le contraste des ROI. Nous rapprochons ainsi les caractéristiques des ROI disponibles à l'entraînement, avec celles vues lors du déploiement. Nous insistons sur le réalisme des images augmentées pour qu'elles soient fidèles aux particularités du domaine cible grâce à l'harmonisation *one-shot* auto-supervisée des tissus altérés. Ce faisant, nous encourageons ainsi la généralisation et la robustesse du modèle de segmentation en aval, quelle que soit la quantité de données disponibles à l'entraînement.

Au travers des différents scénarios cliniques présentés dans ce manuscrit, nous avons évalué l'influence des techniques d'augmentation proposées pour différents régimes de données et sur plusieurs modalités d'imagerie. L'apport des transformations linéaires de contrastes des ROI (approche dite naïve) permet des gains de performance conséquents vis-à-vis d'augmentations conventionnelles, avec une complexité faible. Nous avons notamment observé des améliorations de l'ordre de 5 à 10 points de Dice, en segmentation directe de ROI ou en segmentation inter-modale. À plus haut régime de données, harmoniser de manière plus réaliste et non-linéaire les ROI par un modèle SinGAN a permis d'améliorer les performances de cette stratégie naïve, avec un gain supplémentaire de l'ordre de un point de Dice. L'harmonisation par mélange génératif permet à la fois de mieux réintégrer la ROI dans l'image d'origine, tout en adaptant et diversifiant son apparence de manière réaliste par rapport au domaine cible. À bas régime de données, cette

étape montre toutefois un intérêt plus limité, avec des performances similaires pour les deux cas d'augmentations considérées (altération naïve ou mélange génératif).

À partir de ces observations, nous pouvons déterminer des heuristiques, des bonnes pratiques à adopter en fonction du régime de données rencontré. Avec un nombre restreint d'images, altérer naïvement les ROI semble suffisant. Cela permet d'étendre la distribution couverte et ainsi de corriger un potentiel écart de domaines au niveau des ROI. À plus haut régime de données, comme celui rencontré dans le challenge CrossMoDA, appliquer une harmonisation adaptée et réaliste semble améliorer plus sensiblement les performances. L'emploi d'un modèle *one-shot* de type SinGAN pour apprendre l'harmonisation semble aussi particulièrement pertinent quand images réelles et augmentées composent le jeu d'entrainement, comme dans le cas de configurations d'auto-entrainement itératif avec pseudo-labels. Dans le cas du mélange génératif, il convient également d'évaluer les caractéristiques du domaine de déploiement : s'il est trop hétérogène, le risque de surapprentissage des caractéristiques d'une seule image est plus grand. Nous nous sommes en effet restreints dans ce travail au cas où une seule image du domaine de déploiement (dans le cas de CrossMoDA ou des métastases sur TDM) ou deux (dans le cas de NODE21) sont employées pour l'apprentissage du modèle de mélange génératif. On pourrait néanmoins envisager d'apprendre sur davantage d'images pour mieux couvrir les caractéristiques du domaine de déploiement, avec un coût calculatoire proportionnel au nombre d'images.

Dans les différents scénarios évalués, un grand nombre d'hyper-paramètres ont été manuellement sélectionnés afin de générer des ROI d'apparence globale réaliste pour l'œil humain. Si cette approche empirique n'est pas optimale, comme discuté au chapitre 3, elle permet néanmoins des gains de performance sans utilisation de ressources d'optimisation superflues liées à l'ajustement de ces paramètres, ce qui nous semble un point important à l'heure où la ressource calculatoire est devenue un aspect énergétique critique [2]. De plus, dans le contexte inter-modal du chapitre 4, le gain entre GBA et sa version naïve étant faible, statistiquement non significatif, nous pensons qu'une exploration exhaustive de l'espace des paramètres n'aurait pu permettre qu'une amélioration marginale des résultats obtenus sur ce cas.

D'autres méthodes d'harmonisation pourraient également être appliquées aux contextes cliniques décrits dans cette thèse. Nous avons notamment vu au chapitre 5 que la méthode de *Poisson Blending* reste pertinente, bien qu'elle ne permette pas des modifications structurelles des régions à harmoniser. Ensuite, des modèles profonds peuvent également,

2. Nous dressons un bilan énergétique en annexe Évaluation environnementale et bilan carbone

avec des tâches auto-supervisées comme la nôtre ou via l'apprentissage par transfert par exemple, apprendre à harmoniser les images altérées, que ce soit en modifiant la ROI ou directement l'ensemble de l'image. Enfin, récemment, des approches basées patch, non basées sur l'apprentissage, ont été proposées pour émuler les propriétés de SinGAN, réduisant considérablement les coûts calculatoires engendrés par l'emploi de modèles profonds [220]. Ce type d'approches alternatives mériterait certainement une évaluation dans un contexte de mélange génératif pour l'augmentation.

Dans les différentes applications cliniques décrites, nous avons montré que des augmentations dédiées à l'imagerie médicales permettent des gains de performance vis-à-vis de méthodes d'augmentations conventionnelles appliquées de manière aléatoire. Nous nous appuyons ainsi sur des modèles efficaces et relativement économes en paramètres, comme les U-Nets [14], à partir desquels nous démontrons des améliorations significatives de performance.

Au-delà de la segmentation, ces travaux s'inscrivent dans le cadre d'une approche plus frugale, contrairement aux approches privilégiant des architectures plus complexes d'un point de vue calculatoire, ou plus consommatrices de données, pour un gain de performance parfois marginal. Nous avons ainsi mis les données et leur représentativité au centre de nos réflexions, plutôt que l'architecture des modèles ou leur optimisation. Les méthodes auto-supervisées que nous avons mises en place, comme SinGAN ou l'auto-entraînement itératif, incarnent cette vision en tirant le meilleur parti des données existantes, sans nécessiter de nouvelles annotations. Leur adoption croissante dans des domaines variés, tels que le traitement du langage, témoigne de leur capacité à produire des résultats de haute qualité.

Notre approche d'augmentation *one-shot* se distingue également par son faible coût calculatoire. Dans un contexte d'urgence climatique, notre méthode est ainsi pertinente pour améliorer l'empreinte carbone de l'apprentissage profond, par l'usage de modèles plus sobres. L'annexe Évaluation environnementale et bilan carbone apporte une réflexion supplémentaire sur l'intérêt environnemental de nos travaux et présente le bilan carbone associé à cette thèse. Nous proposons plusieurs pistes pour la réduction des émissions liées à l'apprentissage profond et son utilisation pour la recherche publique.

# ÉVALUATION ENVIRONNEMENTALE ET BILAN CARBONE

**Résumé**

Cette annexe présente les aspects environnementaux liés à ce projet, incluant le bilan carbone de cette thèse ainsi qu'un ensemble de réflexions concernant l'apprentissage profond dans le contexte du réchauffement climatique. En effet, le domaine du numérique et les technologies de l'information sont particulièrement émetteurs de gaz à effet de serre, et connaissent une croissance exponentielle depuis plusieurs années. Après estimation du bilan carbone, la quantité de gaz équivalent $CO_2$ émis par cette thèse est de 2,14 tonnes sur trois ans, en excluant les émissions produites par les conférences elles-mêmes (en présentiel ou distanciel ; chauffage, matériel informatique sur place, site web, etc). Au total, ce projet a nécessité environ 650 jours de temps de calcul sur GPU. Nous présentons ensuite plusieurs pistes de réflexions visant à réduire les émissions dues à l'apprentissage profond, et plus largement dues à la recherche publique et son fonctionnement.

D'après le rapport du GIEC 2023 [3], les activités humaines ont indéniablement provoqué le réchauffement climatique, par l'émission de différents gaz à effet de serre. Afin d'atteindre les accords de Paris et de limiter le réchauffement en dessous de 2°C, il est nécessaire de réduire les émissions mondiales de 5 % supplémentaires chaque année, et ce à partir de 2018 [221]. En France, on estime que cela implique de diminuer considérablement le bilan carbone annuel par habitant, d'environ 10 tonnes de gaz équivalent $CO_2$ émis [4]

3. https://www.ipcc.ch/report/ar6/syr/downloads/report/IPCC_AR6_SYR_SPM.pdf

4. Le $CO_2$ n'étant pas le seul gaz à effet de serre émis par les activités humaines, les autres gaz sont convertis en "équivalent $CO_2$", noté "eq$CO_2$", pour les unir dans une même métrique.

en 2019, à 2 tonnes à l'horizon 2050 [5]. Ainsi, en plus des individus eux-mêmes, chaque secteur doit repenser son fonctionnement et ses objectifs pour un avenir plus durable.

Dans ce contexte, cette annexe présente l'intérêt environnemental ainsi que le bilan carbone de ce doctorat dans son entièreté. Plusieurs constats et pistes pour réduire l'empreinte carbone associée à l'apprentissage profond ou à la recherche complètent cette annexe. Nous ne questionnerons pas le besoin ou la nécessité de ce travail dans le contexte d'urgence climatique.

## Numérique et santé

Parmi les domaines en pleine expansion, les technologies du numérique, demandant une quantité croissante de ressources rares et d'énergie, amènent déjà à des émissions de carbone considérables, de l'ordre de 3 à 4% des émissions mondiales et 2% des émissions nationales en 2020 d'après l'Arcep. Parmi tous les systèmes conçus, un grand nombre d'entre eux sont basés sur l'intelligence artificielle. En particulier, l'apprentissage profond est un gros consommateur énergétique en raison des milliers, millions, voire milliards de paramètres à optimiser. Son utilisation finale est parfois même plus énergivore que son développement [222].

Prenons l'exemple du modèle GPT3, utilisé dans une des premières versions de ChatGPT. La consommation énergétique de son entraînement a été estimée à 1287 MWh, soit environ 552 tonnes eq$CO_2$ émis [223], [224]. En supposant que cette étape ait été réalisée aux États-Unis (les valeurs varient selon les pays et les énergies utilisées), cela implique également la consommation de 5.4 millions de litres d'eau potable, incluant l'évaporation de 700000 litres [225]. À l'inférence du modèle, on considère qu'une requête consomme environ 4 Wh d'électricité, 10 fois plus qu'une recherche Google. Sachant qu'environ 10 millions de requêtes sont effectuées chaque jour, l'utilisation du modèle depuis son entraînement a déjà consommé 14600 MWh, soit une émission de plus de 6200 tonnes eq$CO_2$ et l'utilisation de 61 millions de litres d'eau potable.

Le secteur de la santé, comme nombre d'autres secteurs, devrait être particulièrement affecté par le changement climatique dans les années à venir [6]. En France et en 2021, 46 millions de tonnes eq$CO_2$ ont été émises dans ce secteur, soit 8% des émissions nationales

5. Cet objectif reflète partiellement la situation. Pour en savoir plus : `https://bonpote.com/objectif-2-tonnes-vrai-defi-ou-mauvaise-cible/#Dou_vient_le_chiffre_2_tonnes`

6. Décarboner la santé pour soigner durablement, par *The shift project* : `https://theshiftproject.org/wp-content/uploads/2021/11/TSP_Sante%CC%81_Synthe%CC%80se_DEF.pdf`

[221]. Parmi ces émissions, 87% sont indirectes, dues par exemple au transport des salariés et patients, à l'achat de médicament ou de dispositifs médicaux, etc. Il est donc nécessaire d'œuvrer pour la réduction des émissions dans ce secteur comme dans tous les autres, sans impacter la qualité de la prévention ni des soins.

## Portée environnementale de la thèse

Dans ce manuscrit, nous avons proposé une augmentation *one-shot* permettant de réduire un décalage de domaine, en particulier au niveau des régions d'intérêt (ROI, *regions of interest*) et pour la segmentation. Au lieu d'entraîner encore et encore un même modèle avec de subtiles variations, via une recherche par quadrillage (*grid search*) par exemple, nous mettons l'accent sur les données et leur représentativité par rapport aux images cibles sur lesquelles le modèle sera employé en utilisation clinique réelle. Ainsi, en générant des tumeurs ayant des caractéristiques n'apparaissant pas dans les images d'entraînement, l'importance de l'affinement des hyper-paramètres est réduite, économisant donc des ré-entraînements couteux en énergie, ou alors permettra d'encore améliorer les performances.

La méthode d'augmentation profonde utilise le modèle one-shot SinGAN, apprenant une distribution à partir d'une seule image. Afin de limiter le sur-apprentissage, l'entraînement est plus limité qu'un modèle classique entraîné sur un jeu de données plus ou moins conséquent, et est donc à la fois plus rapide à entraîner et moins énergivore. Comme discuté en conclusion générale, un apprentissage sur un plus grand nombre d'images pourrait permettre d'encore mieux fondre les tumeurs, mais ce dernier serait plus énergivore et montrerait probablement une amélioration des performances très faible et non significative.

Dans les contextes de segmentation intermodalité comme celui présenté au chapitre 4, plusieurs réseaux de segmentation sont généralement entraînés successivement afin d'améliorer itérativement les performances. Grâce à notre augmentation, un seul apprentissage de segmentation a permis d'obtenir des performances supérieures à de multiples itérations sans cette augmentation. Ainsi, notre méthode permet d'économiser les émissions carbones liées à ces nombreux entraînements successifs.

Enfin, notre méthode a réduit l'intérêt de l'acquisition d'IRM avec injection de produit de contraste dans le cas de la segmentation de schwannome vestibulaire. Comme décrit dans la section 4.1, l'usage de ces produits, tels que le gadolinum, a des répercussions

néfastes sur nos écosystèmes. Avec la méthode de segmentation inter-modale proposée dans cette thèse, nous encourageons à réduire son utilisation.

## Bilan carbone

Cette section présente le bilan carbone de cette thèse, estimé sur trois ans. Les différents postes d'émission sont estimés afin de donner un ordre de grandeur le plus précis possible. À cause d'un manque d'informations fournies par leurs organisateurs, les émissions liées à mes participations en conférence nationales et internationales sont exclues. Le bilan est donc sous-évalué par rapport aux émissions réelles.

L'empreinte carbone totale est de $2,14$ t eq$CO_2$. Les travaux de thèse ont émis $1,82$ t eq$CO_2$ soit 85% du bilan total, tandis que la soutenance seule correspond à l'émission de $0,32$ t eq$CO_2$ soit 15%.

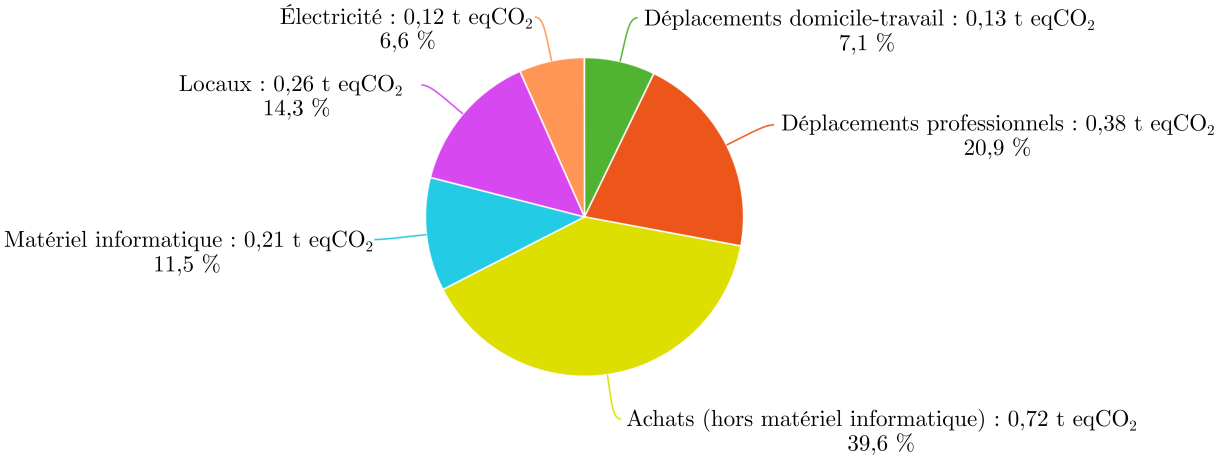



FIGURE A.1 – Bilan carbone de la thèse hors soutenance, représentant l'émission de $1,82$ t eq$CO_2$.

La Fig. A.1 détaille les six postes d'émissions des travaux de thèse hors soutenance. Trois d'entre eux sont induits par le laboratoire et son fonctionnement : les émissions liées aux locaux, à la fabrication du matériel informatique utilisé, ainsi qu'aux autres achats, incluant les émissions des principaux prestataires liés à l'INSERM ou l'université. Leurs émissions sont réparties sur l'ensemble du personnel. Les trois postes restants sont variables selon les membres du laboratoire et dépendent donc directement de mon activité : les déplacements domicile-travail, professionnels, et l'électricité consommée.

## Électricité et utilisation de GPU

Ma part de la consommation d'électricité du laboratoire est estimée à $0,12$ t eq$CO_2$, partagée à 26% par mon ordinateur de travail et à 74% par mon utilisation quotidienne d'un cluster de machines de calculs GPU (respectivement 5250 et 14880 kWh). La consommation moyenne des membres du laboratoire est également de $0,12$ t eq$CO_2$. La consommation principale étant ce cluster et en considérant que tous les membres du laboratoire ne l'utilisent pas (secrétariat, chercheurs n'utilisant pas l'apprentissage profond, etc.), ma consommation électrique due au cluster est en dessous de celle de la moyenne des chercheurs l'utilisant. À noter que l'électricité utilisée est d'origine nucléaire, soit environ 6 g eq$CO_2$ émis pour 1 kWh, expliquant les faibles émissions associées. À titre d'exemple, si elle provenait d'une centrale à charbon, ces émissions auraient été de l'ordre de $5,5$ t eq$CO_2$.

Plus précisément, le nombre d'expériences lancées sur le cluster de machine est d'environ 2400, incluant prétraitement de données, entraînements, inférences et post-traitement. Au total, j'ai utilisé des cartes GPU sur un temps cumulé de 650 jours, sachant qu'un modèle de segmentation nnU-Net à 5 plis (*fold*) correspond à 5 à 8 jours de temps GPU selon le type de cartes disponibles. En incluant 4 itérations, et donc 4 ré-entraînements complets, d'un modèle de segmentation pour mener les expériences du chapitre 4 par exemple, 650 jours d'utilisation cumulée semblent donc être un temps relativement raisonnable comparé aux pratiques d'autres chercheurs, comme l'affinement intensif de paramètres parmi un grand spectre de valeurs.

## Déplacements domicile-travail et professionnels

Mes déplacements domicile-travail étaient essentiellement à pied, avec l'utilisation de la voiture environ une fois par semaine. J'ai également donné des cours dans une école d'ingénieur loin du laboratoire, nécessitant également l'utilisation de la voiture pour m'y rendre. Même si ces émissions ne sont que $7,1$% de mon bilan final, ces dernières auraient pu être encore réduites en me déplaçant plus souvent à vélo ou en transport en commun.

Durant l'ensemble de mes déplacements professionnels, je n'ai effectué qu'un seul aller-retour en avion, ayant émis environ 0.28 t eqCO2, soit près de 75% des émissions liées à mes déplacements professionnels. Celui-ci aurait pu être évité, car une liaison en train existait pour cette destination. D'autres membres de mon laboratoire se rendant à cette conférence en avion, j'ai opté pour ce mode de transport en suivant le groupe. Un effort

collectif pourrait donc être effectué en encourageant le train malgré la fatigue induite par les longs trajets, réduisant ainsi le bilan carbone global du laboratoire.

## Empreinte carbone de la soutenance

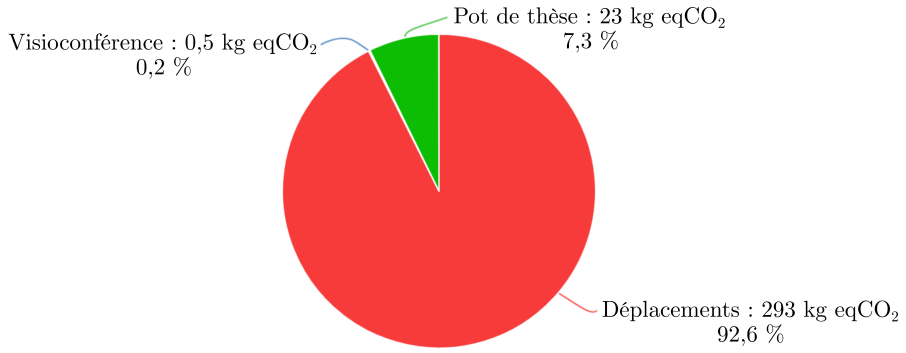



FIGURE A.2 – Bilan carbone de la soutenance, d'environ 316 kg $eqCO_2$.

La Fig. A.2 présente le bilan carbone de la soutenance seule, estimé à $0,32$ t $eqCO_2$, et représente les émissions dues aux déplacements des membres du jury et de mes proches, au système de visioconférence utilisé, ainsi qu'au pot de thèse. Les déplacements sont responsables de la majorité des émissions, dû à 85% à un unique trajet en avion d'un membre du jury, étant le seul à s'être déplacé depuis une autre région [7]. N'ayant pas choisi les moyens de transport utilisés par les membres du jury, mon impact sur la soutenance est limité. Afin de réduire autant que possible les émissions, mes proches assistants à la soutenance se sont déplacés en train. J'ai également choisi un traiteur proposant des produits locaux et principalement végétarien pour le pot de thèse, ce qui a réduit ce poste d'émission de 5 à 10 kg $eqCO_2$.

Proscrire l'avion pourrait ainsi réduire considérablement les émissions de la soutenance. Cela questionne le besoin de se déplacer pour assister physiquement à une soutenance, plutôt qu'en visioconférence. Dans les cas où la présence physique est jugée indispensable, il serait bénéfique de sélectionner les membres de jury en fonction de leur localisation ou

7. Cette annexe vise à informer et quantifier l'impact environnemental de cette thèse dans son ensemble, sans jugement sur les actions et choix individuels de chacun.

du moyen de transport qu'ils utiliseraient. Par ailleurs, les écoles doctorales pourraient envisager des contraintes de composition de jury tout aussi rigoureuses académiquement, mais limitant les déplacements couteux en carbone, voire plus globalement, les émissions de l'ensemble de la thèse.

## Quelques recommandations en apprentissage profond

Afin de limiter les émissions liées à l'apprentissage profond, certaines précautions peuvent être prises, sans nuire à la qualité de la recherche, voire en l'améliorant. Nous proposons ici quelques réflexions allant dans ce sens.

La piste que nous jugeons la plus prometteuse est d'encourager une planification plus précise de la stratégie de recherche et des expériences à effectuer. Cette piste paraît évidente, mais certains chercheurs ont tendance à expérimenter, avec de nombreux tests très exploratoires, avant d'établir précisément ce qu'ils cherchent à prouver ou à obtenir, en précisant les hypothèses de travail *a posteriori*. Nous pensons que le nombre de ces expériences peut être réduit avec des réflexions plus approfondies en amont. Cette approche est déjà présente dans nombre de domaines de recherche, en biologie par exemple, où les consommables à usage unique sont plus coûteux. De plus, le biologiste les jette lui-même, tandis que les consommables du chercheur en apprentissage profond, du temps GPU et de l'électricité, sont plus indirects, voire sont invisibles pour l'utilisateur s'il ne les quantifie pas lui-même. [8]. Il semble donc nécessaire de responsabiliser davantage les chercheurs en apprentissage profond concernant ce type de pratiques et les émissions engendrées.

Cette réflexion en amont permet de se poser certaines questions sur la finalité du projet. Cherche-t-on, par exemple, à établir une preuve de concept d'apprentissage profond ? Ou bien à développer un modèle pour une utilisation en contexte clinique ? Selon les situations, l'ordre de grandeur des performances souhaitées est potentiellement très différent. On peut ensuite choisir le modèle le plus adapté *a priori*, en prenant notamment en compte sa taille et son nombre de paramètres. Ainsi, on limite la surconsommation énergétique liée à l'entraînement de modèles coûteux en ressources, *i.e.* ayant beaucoup de paramètres à entraîner, lorsque leur utilisation n'est pas nécessaire au vu des performances visées. De la même manière, on peut questionner le choix d'un modèle très profond, ayant beaucoup

8. Les cartes GPU sont également des consommables, mais durent généralement 5 à 10 ans avant d'être remplacées. En pratique, c'est très rarement le chercheur qui change lui-même les GPU d'un cluster de machine.

de paramètres, lorsqu'un modèle plus simple permet d'obtenir des performances très légèrement inférieures. À noter que plusieurs techniques permettent de réduire le nombre de paramètres, et donc l'impact écologique des modèles, comme le *pruning* ou encore la distillation (*knowledge distillation*). Par exemple, le modèle de traitement du langage naturel BERT a également une version plus légère, DistillBERT, utilisant 40% de paramètres de moins et s'entraînant 60% plus rapidement tout en conservant 95% des performances de BERT [9]. En parallèle, la stratégie d'entraînement joue également un rôle crucial et nécessite d'être établie en amont. Il paraît par exemple curieux d'effectuer une recherche exhaustive par quadrillage si on cherche à développer une preuve de concept.

Enfin, le partage du code, mais aussi du poids des modèles, semble être une pratique à encourager. En effet, pour comparer un modèle en développement avec l'état de l'art proposé par la communauté, les chercheurs sont parfois amenés à reproduire de manière identique des expériences issues d'articles. La publication des poids permet alors de l'éviter, économisant de l'électricité, et ne nécessitant que peu d'efforts pour les auteurs. Cela encourage également une uniformisation des méthodes de travail et des *framework*, via l'utilisation d'outils communs, bénéfique à la communauté. Plusieurs initiatives de chercheurs vont déjà dans ce sens, avec le partage de code pour encourager la reproducibilité, ou avec le rassemblement d'outils au sein de bibliothèques par exemple [226], [227]. De la même manière, les poids des modèles pourraient être plus souvent associés au code d'entraînement.

## Diminution des émissions dans la recherche publique

Afin d'encourager la réduction des émissions à plus grande échelle, il est essentiel d'inciter les chercheurs et les structures à intégrer les considérations environnementales dans leurs travaux. Les financements publics pourraient par exemple être davantage conditionnés par la prise en compte de critères environnementaux dans les projets de recherche.

Plus largement, l'organisation et le fonctionnement de la recherche publique devraient également être questionnées. Par exemple, comme précisé avant le bilan carbone, la plupart des conférences ne mesurent pas leurs émissions directes ni indirectes. De plus, l'intérêt concret des conférences pour la recherche (élargissement de son réseau, collaborations, etc.) devrait également être quantifié, afin d'évaluer leur bénéfice réel à court, moyen et long terme. Cela permettra ainsi une meilleure optimisation des ressources allouées à

9. `https://www.lemagit.fr/conseil/IA-frugale-trois-regles-simples-pour-optimiser-les-algorithmes`

l'événement. Faut-il alors privilégier des conférences en présentiel, visioconférence [10], ou hybride ? Répondre à ces questions encourage ensuite la prise d'initiatives effectives pour réduire les émissions. On peut imaginer par exemple, pour les conférences en présentiel, l'interdiction de l'utilisation de l'avion, ou la mise en place d'un quota de participants venant en avion.

Afin d'encourager une démarche de réduction des émissions à plus grande échelle, il est nécessaire que chaque laboratoire calcule son bilan carbone. Cette initiative semble de plus en plus encouragée par les différents organismes tels que le CNRS ou l'INSERM. Nous pensons que le bilan devrait alors être rendu public, ainsi que les mesures mises en place pour limiter les émissions. Chaque laboratoire pourrait alors bénéficier des réflexions et bonnes pratiques des autres, favorisant ainsi une culture de responsabilité environnementale au sein de la communauté scientifique. Une évaluation plus systématique des émissions carbones au sein des différents travaux pourraient également être demandée, que ce soit au sein des publications scientifiques ou des manuscrits académiques comme celui-ci.

# Conclusion

Dans cette annexe, nous avons présenté le bilan carbone (de l'ordre de $2,14$ t eq$CO_2$) et l'intérêt environnemental de cette thèse, notamment en encourageant les approches basées sur les données et leur analyse plutôt que sur l'optimisation des modèles. Nous avons discuté de la nécessité d'intégrer des considérations environnementales, en apprentissage profond mais aussi dans la recherche en général, afin d'assurer un développement technologique plus durable. Cela passe par la responsabilisation des chercheurs.ses, incluant la quantification de leur impact et la remise en question de leurs pratiques [11]. Les institutions jouent alors un rôle majeur pour fournir un cadre réglementaire incitant ces pratiques.

10. Le matériel nécessaire émet des gaz à effet de serre, à sa fabrication comme à son utilisation.

11. Nous encourageons le.a chercheur.se à se questionner, ainsi qu'à questionner ses pairs : connaissez-vous la puissance moyenne de votre GPU, ses émissions sur une année, ou bien le nombre d'heures de GPU qu'a nécessité votre dernier article ?

# COMMUNICATIONS

## Challenges

— **G. Sallé**, G. Andrade-Miranda, P. H. Conze, N. Boussion, J. Bert, D. Visvikis, V. Jaouen. *MICCAI CrossMoDA 2022 challenge, segmentation task* : 3ème au classement général, 1er sur la tâche de segmentation de schwannome vestibulaire.

## Revues internationales

— **G. Sallé**, G. Andrade-Miranda, P. H. Conze, N. Boussion, J. Bert, D. Visvikis, V. Jaouen. *Cross-modal tumor segmentation using generative blending augmentation and self-training.* IEEE Transactions on biomedical engineering, 2024 (accepté et à paraître)

## Conférences internationales

— **G. Sallé**, P. H. Conze, N. Boussion, J. Bert, D. Visvikis, V. Jaouen. *Synthetic tumor insertion using one-shot generative learning for cross-modal image segmentation.* IEEE NSS-MIC, 2021 (présentation orale, distanciel)

— **G. Sallé**, P. H. Conze, N. Boussion, J. Bert, D. Visvikis, V. Jaouen. *Tumor blending augmentation using one-shot generative learning for crossmodal MRI segmentation.* Medical Image Computing and Computer Assisted Intervention (MICCAI), Joint workshop session, 2022 (présentation orale, distanciel)

— **G. Sallé**, V. Bourbonne, P. H. Conze, N. Boussion, J. Bert, D. Visvikis, V. Jaouen. *Tumor blending augmentation using one-shot generative learning for brain CT tumor segmentation.* IEEE NSS-MIC, 2022 (présentation orale, présentiel)

## Conférences nationales

— **G. Sallé**, P. H. Conze, N. Boussion, J. Bert, D. Visvikis, V. Jaouen. *Synthèse de tumeurs par modèle génératif one-shot pour la segmentation inter-modale en imagerie médicale.* ORASIS, 2021 (poster, présentiel)

— **G. Sallé**, P. H. Conze, N. Boussion, J. Bert, D. Visvikis, V. Jaouen. *Synthetic tumor insertion using one-shot generative learning for cross-modal image segmentation.* Recherche en Imagerie et Technologies pour la Santé (RITS), 2022 (présentation orale, présentiel)

— **G. Sallé**, P. H. Conze, N. Boussion, J. Bert, D. Visvikis, V. Jaouen. *Altération de tumeurs par modèle génératif one-shot pour la segmentation intermodale en IRM.* IABM, 2023 (poster, présentiel)

## Workshop et école d'été

— **G. Sallé**, P. H. Conze, N. Boussion, J. Bert, D. Visvikis, V. Jaouen. *Synthetic tumor insertion using one-shot generative learning for cross-modal image segmentation.* Cancéropôle Grand Ouest (CGO), 2021 (présentation orale, présentiel)

— **G. Sallé**, V. Bourbonne, P. H. Conze, N. Boussion, J. Bert, D. Visvikis, V. Jaouen. *Tumor blending augmentation using one-shot generative learning for brain CT tumor segmentation.* IEEE EMBS-SPS International Summer School on Biomedical Imaging, 2022 (poster, présentiel)

# BIBLIOGRAPHIE


[1] N. Sharma et L. M. Aggarwal, « Automated medical image segmentation techniques », *Journal of medical physics/Association of Medical Physicists of India*, t. 35, *1*, p. 3, 2010.

[2] X. Liu, L. Song, S. Liu et Y. Zhang, « A review of deep-learning-based medical image segmentation methods », *Sustainability*, t. 13, *3*, p. 1224, 2021.

[3] H. Lin, H. Xiao, L. Dong et al., « Deep learning for automatic target volume segmentation in radiation therapy : a review », *Quantitative Imaging in Medicine and Surgery*, t. 11, *12*, p. 4847, 2021.

[4] D. C. de Castro, I. Walker et B. Glocker, « Causality matters in medical imaging », *Nature Communications*, 2019.

[5] K. Hornik, M. Stinchcombe et H. White, « Multilayer feedforward networks are universal approximators », *Neural networks*, t. 2, *5*, p. 359-366, 1989.

[6] M. Paschali, S. Conjeti, F. Navarro et N. Navab, « Generalizability vs. robustness : adversarial examples for medical imaging », *arXiv preprint arXiv :1804.00504*, 2018.

[7] K. D. Apostolidis et G. A. Papakostas, « A survey on adversarial deep learning robustness in medical image analysis », *Electronics*, t. 10, *17*, p. 2132, 2021.

[8] N. Srivastava, G. Hinton, A. Krizhevsky, I. Sutskever et R. Salakhutdinov, « Dropout : a simple way to prevent neural networks from overfitting », *The journal of machine learning research*, t. 15, *1*, p. 1929-1958, 2014.

[9] A. K. Upadhyay et A. K. Bhandari, « Advances in Deep Learning Models for Resolving Medical Image Segmentation Data Scarcity Problem : A Topical Review », *Archives of Computational Methods in Engineering*, p. 1-19, 2023.

[10] A. Jungo, R. Meier, E. Ermis et al., « On the effect of inter-observer variability for a reliable estimation of uncertainty of medical image segmentation », in *Medical Image Computing and Computer Assisted Intervention–MICCAI 2018 : 21st International Conference, Granada, Spain, September 16-20, 2018, Proceedings, Part I*, Springer, 2018, p. 682-690.

[11] H. McGrath, P. Li, R. Dorent et al., « Manual segmentation versus semi-automated segmentation for quantifying vestibular schwannoma volume on MRI », *International Journal of Computer Assisted Radiology and Surgery*, t. 15, p. 1445-1455, 2020.

[12] M. Liu, P. Maiti, S. Thomopoulos et al., « Style transfer using generative adversarial networks for multi-site mri harmonization », in *Medical Image Computing and Computer Assisted Intervention–MICCAI 2021 : 24th International Conference, Strasbourg, France, September 27–October 1, 2021, Proceedings, Part III 24*, Springer, 2021, p. 313-322.

[13] E. Hallaji, R. Razavi-Far, V. Palade et M. Saif, « Adversarial learning on incomplete and imbalanced medical data for robust survival prediction of liver transplant patients », *IEEE Access*, t. 9, p. 73 641-73 650, 2021.

[14] O. Ronneberger, P. Fischer et T. Brox, « U-net : Convolutional networks for biomedical image segmentation », in *International Conference on Medical image computing and computer-assisted intervention*, Springer, 2015, p. 234-241.

[15] A. E. Kavur, N. S. Gezer, M. Barış et al., « Comparison of semi-automatic and deep learning-based automatic methods for liver segmentation in living liver transplant donors », *Diagnostic and Interventional Radiology*, t. 26, *1*, p. 11, 2020.

[16] D. P. Kingma et M. Welling, « Auto-encoding variational bayes », *arXiv preprint arXiv :1312.6114*, 2013.

[17] N. Tajbakhsh, L. Jeyaseelan, Q. Li, J. N. Chiang, Z. Wu et X. Ding, « Embracing imperfect datasets : A review of deep learning solutions for medical image segmentation », *Medical Image Analysis*, t. 63, p. 101 693, 2020.

[18] R. Azad, E. K. Aghdam, A. Rauland et al., « Medical image segmentation review : The success of u-net », *arXiv preprint arXiv :2211.14830*, 2022.

[19] M. Antonelli, A. Reinke, S. Bakas et al., « The medical segmentation decathlon », *Nature communications*, t. 13, *1*, p. 4128, 2022.

[20] F. Isensee, C. Ulrich, T. Wald et K. H. Maier-Hein, « Extending nnU-Net is all you need », in *BVM Workshop*, Springer, 2023, p. 12-17.

[21] F. Isensee, P. F. Jaeger, S. A. Kohl, J. Petersen et K. H. Maier-Hein, « nnU-Net : a self-configuring method for deep learning-based biomedical image segmentation », *Nature methods*, t. 18, *2*, p. 203-211, 2021.

[22] F. Isensee, P. F. Jäger, P. M. Full, P. Vollmuth et K. H. Maier-Hein, « nnU-Net for brain tumor segmentation », in *Brainlesion : Glioma, Multiple Sclerosis, Stroke and Traumatic Brain Injuries : 6th International Workshop, BrainLes 2020, Held in Conjunction with MICCAI 2020, Lima, Peru, October 4, 2020, Revised Selected Papers, Part II 6*, Springer, 2021, p. 118-132.

[23] P.-H. Conze, G. Andrade-Miranda, V. K. Singh, V. Jaouen et D. Visvikis, « Current and emerging trends in medical image segmentation with deep learning », *IEEE Transactions on Radiation and Plasma Medical Sciences*, 2023.

[24] N. Siddique, S. Paheding, C. P. Elkin et V. Devabhaktuni, « U-net and its variants for medical image segmentation : A review of theory and applications », *Ieee Access*, t. 9, p. 82 031-82 057, 2021.

[25] A. Hatamizadeh, Y. Tang, V. Nath et al., « Unetr : Transformers for 3d medical image segmentation », in *Proceedings of the IEEE/CVF winter conference on applications of computer vision*, 2022, p. 574-584.

[26] H. Xiao, L. Li, Q. Liu, X. Zhu et Q. Zhang, « Transformers in medical image segmentation : A review », *Biomedical Signal Processing and Control*, t. 84, p. 104 791, 2023.

[27] J. Wu, R. Fu, H. Fang et al., « Medsegdiff : Medical image segmentation with diffusion probabilistic model », in *Medical Imaging with Deep Learning*, PMLR, 2024, p. 1623-1639.

[28] G. Andrade-Miranda, V. Jaouen, O. Tankyevych, C. C. Le Rest, D. Visvikis et P.-H. Conze, « Multi-modal medical Transformers : A meta-analysis for medical image segmentation in oncology », *Computerized Medical Imaging and Graphics*, t. 110, p. 102 308, 2023.

[29] J. Lin, « Divergence measures based on the Shannon entropy », *IEEE Transactions on Information theory*, t. 37, *1*, p. 145-151, 1991.

[30] L. R. Dice, « Measures of the amount of ecologic association between species », *Ecology*, t. 26, *3*, p. 297-302, 1945.

[31] F. Milletari, N. Navab et S.-A. Ahmadi, « V-net : Fully convolutional neural networks for volumetric medical image segmentation », in *2016 fourth international conference on 3D vision (3DV)*, Ieee, 2016, p. 565-571.

[32] A. Tversky, « Features of similarity. », *Psychological review*, t. 84, *4*, p. 327, 1977.

[33] S. S. M. Salehi, D. Erdogmus et A. Gholipour, « Tversky loss function for image segmentation using 3D fully convolutional deep networks », in *International workshop on machine learning in medical imaging*, Springer, 2017, p. 379-387.

[34] J. Ma, J. Chen, M. Ng et al., « Loss odyssey in medical image segmentation », *Medical Image Analysis*, t. 71, p. 102 035, 2021.

[35] G. Wang, W. Li, M. Aertsen, J. Deprest, S. Ourselin et T. Vercauteren, « Aleatoric uncertainty estimation with test-time augmentation for medical image segmentation with convolutional neural networks », *Neurocomputing*, t. 338, p. 34-45, 2019.

[36] Z. Wang, J. Yang, H. Jiang et X. Fan, « CNN training with twenty samples for crack detection via data augmentation », *Sensors*, t. 20, *17*, p. 4849, 2020.

[37] F. Garcea, A. Serra, F. Lamberti et L. Morra, « Data augmentation for medical imaging : A systematic literature review », *Computers in Biology and Medicine*, t. 152, p. 106 391, 2023.

[38] R. Dorent et al., « CrossMoDA 2021 challenge : Benchmark of cross-modality domain adaptation techniques for vestibular schwannoma and cochlea segmentation », *Medical Image Analysis*, t. 83, p. 102 628, 2023.

[39] C. Bowles, L. Chen, R. Guerrero et al., « Gan augmentation : Augmenting training data using generative adversarial networks », *arXiv preprint arXiv :1810.10863*, 2018.

[40] H. Guan et M. Liu, « Domain adaptation for medical image analysis : a survey », *IEEE Transactions on Biomedical Engineering*, t. 69, *3*, p. 1173-1185, 2021.

[41] P. Isola, J.-Y. Zhu, T. Zhou et A. A. Efros, « Image-to-image translation with conditional adversarial networks », in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2017, p. 1125-1134.

[42] J.-Y. Zhu, T. Park, P. Isola et A. A. Efros, « Unpaired image-to-image translation using cycle-consistent adversarial networks », in *Proceedings of the IEEE international conference on computer vision*, 2017, p. 2223-2232.

[43] K. Armanious, C. Jiang, M. Fischer et al., « MedGAN : Medical image translation using GANs », *Computerized medical imaging and graphics*, t. 79, p. 101 684, 2020.

[44] F. P. Oliveira et J. M. R. Tavares, « Medical image registration : a review », *Computer methods in biomechanics and biomedical engineering*, t. 17, *2*, p. 73-93, 2014.

[45] J. Chen, S. Chen, L. Wee, A. Dekker et I. Bermejo, « Deep learning based unpaired image-to-image translation applications for medical physics : a systematic review », *Physics in Medicine & Biology*, 2023.

[46] A. Alotaibi, « Deep generative adversarial networks for image-to-image translation : A review », *Symmetry*, t. 12, *10*, p. 1705, 2020.

[47] I. Goodfellow, J. Pouget-Abadie, M. Mirza et al., « Generative adversarial nets », *Advances in neural information processing systems*, t. 27, 2014.

[48] I. Goodfellow, J. Pouget-Abadie, M. Mirza et al., « Generative adversarial networks », *Communications of the ACM*, t. 63, *11*, p. 139-144, 2020.

[49] K. Fan, « Minimax theorems », *Proceedings of the National Academy of Sciences*, t. 39, *1*, p. 42-47, 1953.

[50] J. Jiang, Y.-C. Hu, N. Tyagi et al., « Tumor-aware, adversarial domain adaptation from CT to MRI for lung cancer segmentation », in *Medical Image Computing and Computer Assisted Intervention–MICCAI 2018*, Springer, 2018, p. 777-785.

[51] A. S. Fard, D. C. Reutens et V. Vegh, « From CNNs to GANs for cross-modality medical image estimation », *Computers in Biology and Medicine*, p. 105 556, 2022.

[52] M. Mirza et S. Osindero, « Conditional generative adversarial nets », *arXiv preprint arXiv :1411.1784*, 2014.

[53] A. Radford, L. Metz et S. Chintala, « Unsupervised representation learning with deep convolutional generative adversarial networks », *arXiv preprint arXiv :1511.06434*, 2015.

[54] M. M. Saad, R. O'Reilly et M. H. Rehmani, « A survey on training challenges in generative adversarial networks for biomedical image analysis », *Artificial Intelligence Review*, t. 57, *2*, p. 19, 2024.

[55] M. Arjovsky, S. Chintala et L. Bottou, « Wasserstein generative adversarial networks », in *International conference on machine learning*, PMLR, 2017, p. 214-223.

[56] I. Gulrajani, F. Ahmed, M. Arjovsky, V. Dumoulin et A. C. Courville, « Improved training of wasserstein gans », *Advances in neural information processing systems*, t. 30, 2017.

[57] Q. Wang, X. Zhou, C. Wang et al., « WGAN-based synthetic minority oversampling technique : Improving semantic fine-grained classification for lung nodules in CT images », *IEEE Access*, t. 7, p. 18 450-18 463, 2019.

[58] S. Adiga, M. A. Attia, W.-T. Chang et R. Tandon, « On the tradeoff between mode collapse and sample quality in generative adversarial networks », in *2018 IEEE global conference on signal and information processing (GlobalSIP)*, IEEE, 2018, p. 1184-1188.

[59] T. Karras, T. Aila, S. Laine et J. Lehtinen, « Progressive growing of gans for improved quality, stability, and variation », *arXiv preprint arXiv :1710.10196*, 2017.

[60] A. Beers, J. Brown, K. Chang et al., « High-resolution medical image synthesis using progressively grown generative adversarial networks », *arXiv preprint arXiv :1805.03144*, 2018.

[61] J. Liang, X. Yang, Y. Huang et al., « Sketch guided and progressive growing GAN for realistic and editable ultrasound image synthesis », *Medical Image Analysis*, t. 79, p. 102 461, 2022.

[62] N. K. Singh et K. Raza, « Medical image generation using generative adversarial networks : A review », *Health informatics : A computational perspective in healthcare*, p. 77-96, 2021.

[63] X. Yi, E. Walia et P. Babyn, « Generative adversarial network in medical imaging : A review », *Medical image analysis*, t. 58, p. 101 552, 2019.

[64] R. Osuala, K. Kushibar, L. Garrucho et al., « Data synthesis and adversarial networks : A review and meta-analysis in cancer imaging », *Medical Image Analysis*, p. 102 704, 2022.

[65] H. Zhang, I. Goodfellow, D. Metaxas et A. Odena, « Self-attention generative adversarial networks », in *International conference on machine learning*, PMLR, 2019, p. 7354-7363.

[66] H. Lan, A. D. N. Initiative, A. W. Toga et F. Sepehrband, « SC-GAN : 3D self-attention conditional GAN with spectral normalization for multi-modal neuroimaging synthesis », *BioRxiv*, p. 2020-06, 2020.

[67] X. Chen, H. Zhao, D. Yang, Y. Li, Q. Kang et H. Lu, « SA-SinGAN : self-attention for single-image generation adversarial networks », *Machine Vision and Applications*, t. 32, p. 1-14, 2021.

[68] A. B. L. Larsen, S. K. Sønderby, H. Larochelle et O. Winther, « Autoencoding beyond pixels using a learned similarity metric », in *International conference on machine learning*, PMLR, 2016, p. 1558-1566.

[69] T. Karras, S. Laine, M. Aittala, J. Hellsten, J. Lehtinen et T. Aila, « Analyzing and improving the image quality of stylegan », in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2020, p. 8110-8119.

[70] T. Karras, S. Laine et T. Aila, « A style-based generator architecture for generative adversarial networks », in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2019, p. 4401-4410.

[71] S. Hong, R. Marinescu, A. V. Dalca et al., « 3d-stylegan : A style-based generative adversarial network for generative modeling of three-dimensional medical images », in *Deep Generative Models, and Data Augmentation, Labelling, and Imperfections : First Workshop, DGM4MICCAI 2021, and First Workshop, DALI 2021, Held in Conjunction with MICCAI 2021, Strasbourg, France, October 1, 2021, Proceedings 1*, Springer, 2021, p. 24-34.

[72] E. Richardson, Y. Alaluf, O. Patashnik et al., « Encoding in style : a stylegan encoder for image-to-image translation », in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2021, p. 2287-2296.

[73] T.-C. Wang, M.-Y. Liu, J.-Y. Zhu, A. Tao, J. Kautz et B. Catanzaro, « High-resolution image synthesis and semantic manipulation with conditional gans », in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2018, p. 8798-8807.

[74] M. Maspero, M. H. Savenije, A. M. Dinkla et al., « Dose evaluation of fast synthetic-CT generation using a generative adversarial network for general pelvis MR-only radiotherapy », *Physics in Medicine & Biology*, t. 63, *18*, p. 185 001, 2018.

[75] M. D. Cirillo, D. Abramian et A. Eklund, « Vox2Vox : 3D-GAN for brain tumour segmentation », in *Brainlesion : Glioma, Multiple Sclerosis, Stroke and Traumatic Brain Injuries : 6th International Workshop, BrainLes 2020, Held in Conjunction with MICCAI 2020, Lima, Peru, October 4, 2020, Revised Selected Papers, Part I 6*, Springer, 2021, p. 274-284.

[76] H. Yang, J. Sun, A. Carass et al., « Unpaired brain MR-to-CT synthesis using a structure-constrained CycleGAN », in *Deep Learning in Medical Image Analysis and Multimodal Learning for Clinical Decision Support : 4th International Workshop, DLMIA 2018, and 8th International Workshop, ML-CDS 2018, Held in Conjunction with MICCAI 2018, Granada, Spain, September 20, 2018, Proceedings 4*, Springer, 2018, p. 174-182.

[77] Y. Hiasa, Y. Otake, M. Takao et al., « Cross-modality image synthesis from unpaired data using CycleGAN : Effects of gradient consistency loss and training data size », in *Simulation and Synthesis in Medical Imaging : Third International Workshop, SASHIMI 2018, Held in Conjunction with MICCAI 2018, Granada, Spain, September 16, 2018, Proceedings 3*, Springer, 2018, p. 31-41.

[78] Y. Liu, A. Chen, H. Shi et al., « CT synthesis from MRI using multi-cycle GAN for head-and-neck radiation therapy », *Computerized medical imaging and graphics*, t. 91, p. 101 953, 2021.

[79] M.-Y. Liu, T. Breuel et J. Kautz, « Unsupervised image-to-image translation networks », *Advances in neural information processing systems*, t. 30, 2017.

[80] X. Huang, M.-Y. Liu, S. Belongie et J. Kautz, « Multimodal unsupervised image-to-image translation », in *Proceedings of the European conference on computer vision (ECCV)*, 2018, p. 172-189.

[81] J. P. Cohen, M. Luck et S. Honari, « Distribution matching losses can hallucinate features in medical image translation », in *Medical Image Computing and Computer Assisted Intervention–MICCAI*, Springer, 2018, p. 529-536.

[82] J. P. Cohen, M. Luck et S. Honari, « How to Cure Cancer (in images) with Unpaired Image Translation », in *MIDL 2018*.

[83] T. Park, A. A. Efros, R. Zhang et J.-Y. Zhu, « Contrastive learning for unpaired image-to-image translation », in *European conference on computer vision*, Springer, 2020, p. 319-345.

[84] J. W. Choi, « Using out-of-the-box frameworks for contrastive unpaired image translation for vestibular schwannoma and cochlea segmentation : An approach for the crossmoda challenge », in *International MICCAI Brainlesion Workshop*, Springer, 2021, p. 509-517.

[85] S. Yan, C. Wang, W. Chen et J. Lyu, « Swin transformer-based GAN for multimodal medical image translation », *Frontiers in Oncology*, t. 12, p. 942 511, 2022.

[86] K. He, C. Gan, Z. Li et al., « Transformers in medical image analysis : A review », *Intelligent Medicine*, 2022.

[87] M. Özbey, S. U. Dar, H. A. Bedel et al., « Unsupervised medical image translation with adversarial diffusion models », *arXiv preprint arXiv :2207.08208*, 2022.

[88] H. Sasaki, C. G. Willcocks et T. P. Breckon, « Unit-ddpm : Unpaired image translation with denoising diffusion probabilistic models », *arXiv preprint arXiv :2104.05358*, 2021.

[89] L. Huang, Z. Zhou, Y. Guo et Y. Wang, « A stability-enhanced CycleGAN for effective domain transformation of unpaired ultrasound images », *Biomedical Signal Processing and Control*, t. 77, p. 103 831, 2022.

[90] J. Liu, J. He, Y. Xie et al., « Illumination-invariant flotation froth color measuring via Wasserstein distance-based CycleGAN with structure-preserving constraint », *IEEE transactions on cybernetics*, t. 51, *2*, p. 839-852, 2020.

[91] A. Borji, « Pros and cons of gan evaluation measures », *Computer vision and image understanding*, t. 179, p. 41-65, 2019.

[92] G. Mariani, F. Scheidegger, R. Istrate, C. Bekas et C. Malossi, « Bagan : Data augmentation with balancing gan », *arXiv preprint arXiv :1803.09655*, 2018.

[93] C. Shorten et T. M. Khoshgoftaar, « A survey on image data augmentation for deep learning », *Journal of big data*, t. 6, *1*, p. 1-48, 2019.

[94] H. Zhang, M. Cisse, Y. N. Dauphin et D. Lopez-Paz, « mixup : Beyond empirical risk minimization », *arXiv preprint arXiv :1710.09412*, 2017.

[95] J. Cao, M. Luo, J. Yu, M.-H. Yang et R. He, « ScoreMix : A Scalable Augmentation Strategy for Training GANs with Limited Data », *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2022.

[96] E. Panfilov, A. Tiulpin, S. Klein, M. T. Nieminen et S. Saarakkala, « Improving robustness of deep learning based knee MRI segmentation : Mixup and adversarial domain adaptation », in *Proceedings of the IEEE/CVF International Conference on Computer Vision Workshops*, 2019, p. 0–0.

[97] P. Y. Simard, D. Steinkraus, J. C. Platt et al., « Best practices for convolutional neural networks applied to visual document analysis. », in *Icdar*, Edinburgh, t. 3, 2003.

[98] P. R. Lorenzo, J. Nalepa, B. Bobek-Billewicz et al., « Segmenting brain tumors from FLAIR MRI using fully convolutional neural networks », *Computer methods and programs in biomedicine*, t. 176, p. 135-148, 2019.

[99] X. Zhang, C. Liu, N. Ou et al., « CarveMix : a simple data augmentation method for brain lesion segmentation », *NeuroImage*, t. 271, p. 120 041, 2023.

[100] T. C. Mok et A. Chung, « Learning data augmentation for brain tumor segmentation with coarse-to-fine generative adversarial networks », in *international MICCAI brainlesion workshop*, Springer, 2019, p. 70-80.

[101] J. Nalepa, M. Marcinkiewicz et M. Kawulok, « Data augmentation for brain-tumor segmentation : a review », *Frontiers in computational neuroscience*, t. 13, p. 83, 2019.

[102] J. Nalepa, G. Mrukwa, S. Piechaczek et al., « Data augmentation via image registration », in *2019 IEEE International Conference on Image Processing (ICIP)*, IEEE, 2019, p. 4250-4254.

[103] B. B. Avants, C. L. Epstein, M. Grossman et J. C. Gee, « Symmetric diffeomorphic image registration with cross-correlation : evaluating automated labeling of elderly and neurodegenerative brain », *Medical image analysis*, t. 12, *1*, p. 26-41, 2008.

[104] A. Sotiras, C. Davatzikos et N. Paragios, « Deformable medical image registration : A survey », *IEEE transactions on medical imaging*, t. 32, *7*, p. 1153-1190, 2013.

[105] B. Billot, D. N. Greve, O. Puonti et al., « SynthSeg : Segmentation of brain MRI scans of any contrast and resolution without retraining », *Medical image analysis*, t. 86, p. 102 789, 2023.

[106] M. Kompanek, M. Tamajka et W. Benesova, « Volumetrie data augmentation as an effective tool in mri classification using 3d convolutional neural network », in *2019 international conference on systems, signals and image processing (IWSSIP)*, IEEE, 2019, p. 115-119.

[107] F. Zhang, B. Pan, P. Shao et al., « A single model deep learning approach for Alzheimer's disease diagnosis », *Neuroscience*, t. 491, p. 200-214, 2022.

[108] F. Garcea, A. Serra, F. Lamberti et L. Morra, « Data augmentation for medical imaging : A systematic literature review », *Computers in Biology and Medicine*, p. 106 391, 2022.

[109] M. Cossio, « Augmenting Medical Imaging : A Comprehensive Catalogue of 65 Techniques for Enhanced Data Analysis », *arXiv preprint arXiv :2303.01178*, 2023.

[110] A. Kebaili, J. Lapuyade-Lahorgue et S. Ruan, « Deep Learning Approaches for Data Augmentation in Medical Imaging : A Review », *Journal of Imaging*, t. 9, *4*, p. 81, 2023.

[111] P. Chlap, H. Min, N. Vandenberg, J. Dowling, L. Holloway et A. Haworth, « A review of medical image data augmentation techniques for deep learning applications », *Journal of Medical Imaging and Radiation Oncology*, t. 65, *5*, p. 545-563, 2021.

[112] S. Kim, B. Kim et H. Park, « Synthesis of brain tumor multicontrast MR images for improved data augmentation », *Medical Physics*, t. 48, *5*, p. 2185-2198, 2021.

[113] Q. Li, Z. Yu, Y. Wang et H. Zheng, « TumorGAN : A multi-modal data augmentation framework for brain tumor segmentation », *Sensors*, t. 20, *15*, p. 4203, 2020.

[114] G. Sallé, P.-H. Conze, N. Boussion, J. Bert, D. Visvikis et V. Jaouen, « Fake tumor insertion using one-shot generative learning for a cross-modal image segmentation », *IEEE NSS-MIC*, 2021.

[115] M. S. Graham, I. Drobnjak et H. Zhang, « Realistic simulation of artefacts in diffusion MRI for validating post-processing correction techniques », *NeuroImage*, t. 125, p. 1079-1094, 2016.

[116] C. Chen, C. Qin, H. Qiu et al., « Realistic adversarial data augmentation for MR image segmentation », in *International Conference on Medical Image Computing and Computer-Assisted Intervention*, Springer, 2020, p. 667-677.

[117] V. Sandfort, K. Yan, P. J. Pickhardt et R. M. Summers, « Data augmentation using generative adversarial networks (CycleGAN) to improve generalizability in CT segmentation tasks », *Scientific reports*, t. 9, *1*, p. 16 884, 2019.

[118] J. Jiang, Y.-C. Hu, N. Tyagi et al., « Cross-modality (CT-MRI) prior augmented deep learning for robust lung tumor segmentation from small MR datasets », *Medical physics*, t. 46, *10*, p. 4392-4404, 2019.

[119] A. Zhao, G. Balakrishnan, F. Durand, J. V. Guttag et A. V. Dalca, « Data augmentation using learned transformations for one-shot medical image segmentation », in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2019, p. 8543-8553.

[120] E. D. Cubuk, B. Zoph, D. Mane, V. Vasudevan et Q. V. Le, « Autoaugment : Learning augmentation strategies from data », in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2019, p. 113-123.

[121] D. Yang, H. Roth, Z. Xu, F. Milletari, L. Zhang et D. Xu, « Searching learning strategy with reinforcement learning for 3D medical image segmentation », in *Medical Image Computing and Computer Assisted Intervention–MICCAI 2019 : 22nd International Conference, Shenzhen, China, October 13–17, 2019, Proceedings, Part II 22*, Springer, 2019, p. 3-11.

[122] J. Xu, M. Li et Z. Zhu, « Automatic data augmentation for 3D medical image segmentation », in *Medical Image Computing and Computer Assisted Intervention–MICCAI 2020 : 23rd International Conference, Lima, Peru, October 4–8, 2020, Proceedings, Part I 23*, Springer, 2020, p. 378-387.

[123] G. Sallé, V. Bourbonne, P.-H. Conze et al., « Tumor blending augmentation using one-shot generative learning for brain CT tumor segmentation », *IEEE NSS-MIC*, 2022.

[124] J. Wang, C. Lan, C. Liu et al., « Generalizing to unseen domains : A survey on domain generalization », *IEEE Transactions on Knowledge and Data Engineering*, 2022.

[125] K. Zhou, Z. Liu, Y. Qiao, T. Xiang et C. C. Loy, « Domain generalization : A survey », *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2022.

[126] Q. Dou, D. Coelho de Castro, K. Kamnitsas et B. Glocker, « Domain generalization via model-agnostic learning of semantic features », *Advances in neural information processing systems*, t. 32, 2019.

[127] Q. Liu, Q. Dou et P.-A. Heng, « Shape-aware meta-learning for generalizing prostate MRI segmentation to unseen domains », in *Medical Image Computing and Computer Assisted Intervention–MICCAI 2020 : 23rd International Conference, Lima, Peru, October 4–8, 2020, Proceedings, Part II 23*, Springer, 2020, p. 475-485.

[128] B. Konkel, J. Macdonald, K. Lafata et al., « Systematic Analysis of Common Factors Impacting Deep Learning Model Generalizability in Liver Segmentation », *Radiology : Artificial Intelligence*, t. 5, *3*, e220080, 2023.

[129] W. E. Johnson, C. Li et A. Rabinovic, « Adjusting batch effects in microarray expression data using empirical Bayes methods », *Biostatistics*, t. 8, *1*, p. 118-127, 2007.

[130] R. Da-Ano, I. Masson, F. Lucia et al., « Performance comparison of modified ComBat for harmonization of radiomic features for multicenter studies », *Scientific Reports*, t. 10, *1*, p. 10 248, 2020.

[131] T. K. Bell, K. J. Godfrey, A. L. Ware, K. O. Yeates et A. D. Harris, « Harmonization of multi-site MRS data with ComBat », *NeuroImage*, t. 257, p. 119 330, 2022.

[132] G. Kumar, S. Basri, A. A. Imam, S. A. Khowaja, L. F. Capretz et A. O. Balogun, « Data harmonization for heterogeneous datasets : A systematic literature review », *Applied Sciences*, t. 11, *17*, p. 8275, 2021.

[133] S. Cackowski, E. L. Barbier, M. Dojat et T. Christen, « comBat versus cycleGAN for multi-center MR images harmonization », 2021.

[134] C. Hognon, F. Tixier, O. Gallinato, T. Colin, D. Visvikis et V. Jaouen, « Standardization of multicentric image datasets with generative adversarial networks », in *IEEE Nuclear Science Symposium and Medical Imaging Conference 2019*, 2019.

[135] Y. Liu, G. R. Kirk, B. M. Nacewicz et al., « Harmonization and targeted feature dropout for generalized segmentation : Application to multi-site traumatic brain injury images », in *Domain Adaptation and Representation Transfer and Medical Image Learning with Less Labels and Imperfect Data : First MICCAI Workshop, DART 2019, and First International Workshop, MIL3ID 2019, Shenzhen, Held in Conjunction with MICCAI 2019, Shenzhen, China, October 13 and 17, 2019, Proceedings 1*, Springer, 2019, p. 81-89.

[136] G. Modanwal, A. Vellal, M. Buda et M. A. Mazurowski, « MRI image harmonization using cycle-consistent generative adversarial network », in *Medical Imaging 2020 : Computer-Aided Diagnosis*, SPIE, t. 11314, 2020, p. 259-264.

[137] V. M. Bashyam, J. Doshi, G. Erus et al., « Deep generative medical image harmonization for improving cross-site generalization in deep learning predictors », *Journal of Magnetic Resonance Imaging*, t. 55, *3*, p. 908-916, 2022.

[138] J. Xu, K. Xue et K. Zhang, « Current status and future trends of clinical diagnoses via image-based deep learning », *Theranostics*, t. 9, *25*, p. 7556, 2019.

[139] H. E. Kim, A. Cosa-Linan, N. Santhanam, M. Jannesari, M. E. Maros et T. Ganslandt, « Transfer learning for medical image classification : a literature review », *BMC medical imaging*, t. 22, *1*, p. 69, 2022.

[140] M. Heusel, H. Ramsauer, T. Unterthiner, B. Nessler et S. Hochreiter, « Gans trained by a two time-scale update rule converge to a local nash equilibrium », *Advances in neural information processing systems*, t. 30, 2017.

[141] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li et L. Fei-Fei, « Imagenet : A large-scale hierarchical image database », in *2009 IEEE conference on computer vision and pattern recognition*, Ieee, 2009, p. 248-255.

[142] Z. N. K. Swati, Q. Zhao, M. Kabir et al., « Brain tumor classification for MR images using transfer learning and fine-tuning », *Computerized Medical Imaging and Graphics*, t. 75, p. 34-46, 2019.

[143] G. Hinton, O. Vinyals et J. Dean, « Distilling the knowledge in a neural network », *arXiv preprint arXiv :1503.02531*, 2015.

[144] K. Kamnitsas, S. Winzeck, E. N. Kornaropoulos et al., « Transductive image segmentation : Self-training and effect of uncertainty estimation », in *Domain Adaptation and Representation Transfer, and Affordable Healthcare and AI for Resource Diverse Global Health : Third MICCAI Workshop, DART 2021, and First MICCAI Workshop, FAIR 2021, Held in Conjunction with MICCAI 2021, Strasbourg, France, September 27 and October 1, 2021, Proceedings 3*, Springer, 2021, p. 79-89.

[145] D. Karimi, H. Dou, S. K. Warfield et A. Gholipour, « Deep learning with noisy labels : Exploring techniques and remedies in medical image analysis », *Medical image analysis*, t. 65, p. 101 759, 2020.

[146] H. Liu, Z. Xu, R. Gao et al., « COSST : Multi-organ Segmentation with Partially Labeled Datasets Using Comprehensive Supervisions and Self-training », *arXiv preprint arXiv :2304.14030*, 2023.

[147] B. Zoph, G. Ghiasi, T.-Y. Lin et al., « Rethinking pre-training and self-training », *Advances in neural information processing systems*, t. 33, p. 3833-3845, 2020.

[148] M.-R. Amini, V. Feofanov, L. Pauletto, E. Devijver et Y. Maximov, « Self-training : A survey », *arXiv preprint arXiv :2202.12040*, 2022.

[149] X. Liu, F. Xing, M. Stone et al., « Generative self-training for cross-domain unsupervised tagged-to-cine mri synthesis », in *Medical Image Computing and Computer Assisted Intervention–MICCAI 2021 : 24th International Conference, Strasbourg, France, September 27–October 1, 2021, Proceedings, Part III 24*, Springer, 2021, p. 138-148.

[150] K. Chaitanya, E. Erdil, N. Karani et E. Konukoglu, « Local contrastive loss with pseudo-label based self-training for semi-supervised medical image segmentation », *Medical Image Analysis*, t. 87, p. 102 792, 2023.

[151] A. Kirillov, E. Mintun, N. Ravi et al., « Segment anything », *arXiv preprint arXiv :2304.02643*, 2023.

[152] V. I. Butoi, J. J. G. Ortiz, T. Ma, M. R. Sabuncu, J. Guttag et A. V. Dalca, « Universeg : Universal medical image segmentation », *arXiv preprint arXiv :2304.06131*, 2023.

[153] J. Ma et B. Wang, « Segment anything in medical images », *arXiv preprint arXiv :2304.12306*, 2023.

[154] S. He, R. Bao, J. Li, P. E. Grant et Y. Ou, « Accuracy of segment-anything model (sam) in medical image segmentation tasks », *arXiv preprint arXiv :2304.09324*, 2023.

[155] Y. Wen, L. Chen, Y. Deng et C. Zhou, « Rethinking pre-training on medical imaging », *Journal of Visual Communication and Image Representation*, t. 78, p. 103 145, 2021.

[156] S. Abnar, M. Dehghani, B. Neyshabur et H. Sedghi, « Exploring the limits of large scale pre-training », *arXiv preprint arXiv :2110.02095*, 2021.

[157] K. He, R. Girshick et P. Dollár, « Rethinking imagenet pre-training », in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2019, p. 4918-4927.

[158] X. Liu, F. Zhang, Z. Hou et al., « Self-supervised learning : Generative or contrastive », *IEEE transactions on knowledge and data engineering*, t. 35, *1*, p. 857-876, 2021.

[159] A. Chowdhury, J. Rosenthal, J. Waring et R. Umeton, « Applying self-supervised learning to medicine : review of the state of the art and medical implementations », in *Informatics*, MDPI, t. 8, 2021, p. 59.

[160] W. Luo, Y. Li, R. Urtasun et R. Zemel, « Understanding the effective receptive field in deep convolutional neural networks », *Advances in neural information processing systems*, t. 29, 2016.

[161] Y. Wang, Q. Yao, J. T. Kwok et L. M. Ni, « Generalizing from a few examples : A survey on few-shot learning », *ACM computing surveys (csur)*, t. 53, *3*, p. 1-34, 2020.

[162] J. Kotia, A. Kotwal, R. Bharti et R. Mangrulkar, « Few shot learning for medical imaging », *Machine learning algorithms for industrial applications*, p. 107-132, 2021.

[163] N. Catalano et M. Matteucci, « Few Shot Semantic Segmentation : a review of methodologies and open challenges », *arXiv preprint arXiv :2304.05832*, 2023.

[164] M. Abdollahzadeh, T. Malekzadeh, C. T. Teo, K. Chandrasegaran, G. Liu et N.-M. Cheung, « A Survey on Generative Modeling with Limited Data, Few Shots, and Zero Shot », *arXiv preprint arXiv :2307.14397*, 2023.

[165] P. Khandelwal et P. Yushkevich, « Domain generalizer : A few-shot meta learning framework for domain generalization in medical imaging », in *Domain Adaptation and Representation Transfer, and Distributed and Collaborative Learning : Second MICCAI Workshop, DART 2020, and First MICCAI Workshop, DCL 2020, Held in Conjunction with MICCAI 2020, Lima, Peru, October 4–8, 2020, Proceedings 2*, Springer, 2020, p. 73-84.

[166] Q. Yu, K. Dang, N. Tajbakhsh, D. Terzopoulos et X. Ding, « A location-sensitive local prototype network for few-shot medical image segmentation », in *2021 IEEE 18th international symposium on biomedical imaging (ISBI)*, IEEE, 2021, p. 262-266.

[167] C. Finn, P. Abbeel et S. Levine, « Model-agnostic meta-learning for fast adaptation of deep networks », in *International conference on machine learning*, PMLR, 2017, p. 1126-1135.

[168] R. Khadka, D. Jha, S. Hicks et al., « Meta-learning with implicit gradients in a few-shot setting for medical image segmentation », *Computers in Biology and Medicine*, t. 143, p. 105 227, 2022.

[169] T. R. Shaham, T. Dekel et T. Michaeli, « Singan : Learning a generative model from a single natural image », in *ICCV*, 2019, p. 4570-4580.

[170] V. Sushko, J. Gall et A. Khoreva, « One-shot gan : Learning to generate samples from single images and videos », in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2021, p. 2596-2600.

[171] G. Giannone, D. Nielsen et O. Winther, « Few-shot diffusion models », *arXiv preprint arXiv :2205.15463*, 2022.

[172] T. Hinz, M. Fisher, O. Wang et S. Wermter, « Improved techniques for training single-image gans », in *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, 2021, p. 1300-1309.

[173] J. Yoo et Q. Chen, « SinIR : Efficient general image manipulation with single image reconstruction », in *International Conference on Machine Learning*, PMLR, 2021, p. 12 040-12 050.

[174] V. Kulikov, S. Yadin, M. Kleiner et T. Michaeli, « Sinddm : A single image denoising diffusion model », in *International Conference on Machine Learning*, PMLR, 2023, p. 17 920-17 930.

[175] V. Thambawita et al., « SinGAN-Seg : Synthetic training data generation for medical image segmentation », *PloS one*, t. 17, *5*, 2022.

[176] J. Cai, Z. Zhang, L. Cui, Y. Zheng et L. Yang, « Towards cross-modal organ translation and segmentation : A cycle-and shape-consistent generative adversarial network », *Medical image analysis*, t. 52, p. 174-184, 2019.

[177] K. Kamnitsas, C. Baumgartner, C. Ledig et al., « Unsupervised domain adaptation in brain lesion segmentation with adversarial networks », in *Information Processing in Medical Imaging : 25th International Conference, IPMI 2017, Boone, NC, USA, June 25-30, 2017, Proceedings 25*, Springer, 2017, p. 597-609.

[178] H. Liu, Y. Fan, C. Cui, D. Su, A. McNeil et B. M. Dawant, « Cross-modality domain adaptation for vestibular schwannoma and cochlea segmentation », *arXiv preprint arXiv :2109.06274*, 2021.

[179] C. Chen, Q. Dou, H. Chen, J. Qin et P. A. Heng, « Unsupervised bidirectional cross-modality adaptation via deeply synergistic image and feature alignment for medical image segmentation », *IEEE transactions on medical imaging*, t. 39, *7*, p. 2494-2505, 2020.

[180] Y. Huo, Z. Xu, H. Moon et al., « Synseg-net : Synthetic segmentation without target modality ground truth », *IEEE transactions on medical imaging*, t. 38, *4*, p. 1016-1025, 2018.

[181] J. Hoffman, E. Tzeng, T. Park et al., « Cycada : Cycle-consistent adversarial domain adaptation », in *International conference on machine learning*, Pmlr, 2018, p. 1989-1998.

[182] S. Fu, Y. Lu, Y. Wang et al., « Domain adaptive relational reasoning for 3d multi-organ segmentation », in *Medical Image Computing and Computer Assisted Intervention–MICCAI 2020*, Springer, 2020, p. 656-666.

[183] Z. Zhao, K. Xu, S. Li, Z. Zeng et C. Guan, « MT-UDA : Towards unsupervised cross-modality medical image segmentation with limited source labels », in *Medical Image Computing and Computer Assisted Intervention–MICCAI*, Springer, 2021, p. 293-303.

[184] Z. Zhao, F. Zhou, K. Xu, Z. Zeng, C. Guan et S. K. Zhou, « LE-UDA : Label-efficient unsupervised domain adaptation for medical image segmentation », *IEEE Transactions on Medical Imaging*, 2022.

[185] Z. Zhao, K. Xu, H. Z. Yeo, X. Yang et C. Guan, « MS-MT : Multi-Scale Mean Teacher with Contrastive Unpaired Translation for Cross-Modality Vestibular Schwannoma and Cochlea Segmentation », *arXiv preprint arXiv :2303.15826*, 2023.

[186] H. Shin, H. Kim, S. Kim, Y. Jun, T. Eo et D. Hwang, « COSMOS : Cross-Modality Unsupervised Domain Adaptation for 3D Medical Image Segmentation based on Target-aware Domain Translation and Iterative Self-Training », *arXiv preprint arXiv :2203.16557*, 2022.

[187] E. Kondrateva, M. Pominova, E. Popova, M. Sharaev, A. Bernstein et E. Burnaev, « Domain shift in computer vision models for MRI data analysis : an overview », in *Thirteenth International Conference on Machine Vision*, SPIE, t. 11605, 2021, p. 126-133.

[188] P. Pérez, M. Gangnet et A. Blake, « Poisson image editing », in *ACM SIGGRAPH 2003 Papers*, 2003, p. 313-318.

[189] H. Wang, Y. Zhou, J. Zhang et al., « Anomaly segmentation in retinal images with poisson-blending data augmentation », *Medical Image Analysis*, t. 81, p. 102 534, 2022.

[190] W.-H. Shen et M.-L. Li, « Copy-Paste Image Augmentation with Poisson Image Editing for Ultrasound Instance Segmentation Learning », *arXiv preprint arXiv :2308.14772*, 2023.

[191] B. Medical, « Medical gallery of Blausen medical 2014 », *WikiJournal of Medicine*, t. 1, *2*, p. 1-79, 2014.

[192] D. G. R. Evans, A. Moran, A. King, S. Saeed, N. Gurusinghe et R. Ramsden, « Incidence of vestibular schwannoma and neurofibromatosis 2 in the North West of England over a 10-year period : higher incidence than previously thought », *Otology & neurotology*, t. 26, *1*, p. 93-97, 2005.

[193] W. T. Koos, J. D. Day, C. Matula et D. I. Levy, « Neurotopographic considerations in the microsurgical treatment of small acoustic neurinomas », *Journal of neurosurgery*, t. 88, *3*, p. 506-512, 1998.

[194] E. A. Vokurka, A. Herwadkar, N. A. Thacker, R. T. Ramsden et A. Jackson, « Using Bayesian tissue classification to improve the accuracy of vestibular schwannoma volume and growth measurement », *American journal of neuroradiology*, t. 23, *3*, p. 459-467, 2002.

[195] J. Shapey, A. Kujawa, R. Dorent et al., « Segmentation of vestibular schwannoma from MRI, an open annotated dataset and baseline algorithm », *Scientific Data*, t. 8, *1*, p. 1-6, 2021.

[196] D. Carr, J. Brown, G. Bydder et al., « Intravenous chelated gadolinium as a contrast agent in NMR imaging of cerebral tumours », *The Lancet*, t. 323, *8375*, p. 484-486, 1984.

[197] D. H. Coelho, Y. Tang, B. Suddarth et M. Mamdani, « MRI surveillance of vestibular schwannomas without contrast enhancement : clinical and economic evaluation », *The Laryngoscope*, t. 128, *1*, p. 202-209, 2018.

[198] G. Wang et al., « Automatic segmentation of vestibular schwannoma from T2-weighted MRI by deep spatial attention with hardness-weighted loss », in *Medical Image Computing and Computer Assisted Intervention–MICCAI*, Springer, 2019, p. 264-272.

[199] D. R. Broome, M. S. Girguis, P. W. Baron, A. C. Cottrell, I. Kjellin et G. A. Kirk, « Gadodiamide-associated nephrogenic systemic fibrosis : why radiologists should be concerned », *American Journal of Roentgenology*, t. 188, *2*, p. 586-592, 2007.

[200] R. J. McDonald, D. Levine, J. Weinreb et al., « Gadolinium retention : a research roadmap from the 2018 NIH/ACR/RSNA workshop on gadolinium chelates », *Radiology*, t. 289, *2*, p. 517-534, 2018.

[201] R. Dorent et al., « Scribble-based domain adaptation via co-segmentation », in *Medical Image Computing and Computer Assisted Intervention–MICCAI*, Springer, 2020, p. 479-489.

[202] Q. Xie, M.-T. Luong, E. Hovy et Q. V. Le, « Self-training with noisy student improves imagenet classification », in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2020, p. 10 687-10 698.

[203] J. W. Choi, « Using Out-of-the-Box Frameworks for Unpaired Image Translation and Image Segmentation for the crossMoDA Challenge », *arXiv e-prints*, 2021.

[204] R. Chen, W. Huang, B. Huang, F. Sun et B. Fang, « Reusing discriminators for encoding : Towards unsupervised image-to-image translation », in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2020, p. 8168-8177.

[205] J. Su, Z. Luo, S. Lian, D. Lin et S. Li, « Mutual learning with reliable pseudo label for semi-supervised medical image segmentation », *Medical Image Analysis*, p. 103 111, 2024.

[206] K. Pantel et R. H. Brakenhoff, « Dissecting the metastatic cascade », *Nature reviews cancer*, t. 4, *6*, p. 448-456, 2004.

[207] O. Charron, A. Lallement, D. Jarnet, V. Noblet, J.-B. Clavier et P. Meyer, « Automatic detection and segmentation of brain metastases on multimodal MR images with a deep convolutional neural network », *Computers in biology and medicine*, 2018.

[208] S. Amemiya, H. Takao, S. Kato, H. Yamashita, N. Sakamoto et O. Abe, « Automatic detection of brain metastases on contrast-enhanced CT with deep-learning feature-fused single-shot detectors », *European Journal of Radiology*, 2021.

[209] S. e. a. Kato, « Automated detection of brain metastases on non-enhanced CT using single-shot detectors », *Neuroradiology*, 2021.

[210] V. Bourbonne, V. Jaouen, C. Hognon et al., « Dosimetric Validation of a GAN-Based Pseudo-CT Generation for MRI-Only Stereotactic Brain Radiotherapy », *Cancers*, 2021.

[211] G. J. Litjens, L. Hogeweg, A. M. Schilham, P. A. de Jong, M. A. Viergever et B. van Ginneken, « Simulation of nodules and diffuse infiltrates in chest radiographs using CT templates », in *Medical Image Computing and Computer-Assisted Intervention–MICCAI 2010 : 13th International Conference, Beijing, China, September 20-24, 2010, Proceedings, Part II 13*, Springer, 2010, p. 396-403.

[212] D. E. Ost et M. K. Gould, « Decision making in patients with pulmonary nodules », *American journal of respiratory and critical care medicine*, t. 185, *4*, p. 363-372, 2012.

[213] J. R. Muhm, W. Miller, R. Fontana, D. Sanderson et M. Uhlenhopp, « Lung cancer detected during a screening program using four-month chest radiographs. », *Radiology*, t. 148, *3*, p. 609-615, 1983.

[214] S. Ren, K. He, R. Girshick et J. Sun, « Faster R-CNN : Towards real-time object detection with region proposal networks », *Advances in neural information processing systems*, t. 28, 2015.

[215] C. C. Nguyen, G. S. Tran, J.-C. Burie, T. P. Nghiem et al., « Pulmonary nodule detection based on faster R-CNN with adaptive anchor box », *Ieee Access*, t. 9, p. 154 740-154 751, 2021.

[216] X. Li, L. Shen, X. Xie et al., « Multi-resolution convolutional networks for chest X-ray radiograph based lung nodule detection », *Artificial intelligence in medicine*, t. 103, p. 101 744, 2020.

[217] N. Otsu, « A threshold selection method from gray-level histograms », *IEEE transactions on systems, man, and cybernetics*, t. 9, *1*, p. 62-66, 1979.

[218] E. Marcus, S. Papa, J.-J. Sonke et J. Teuwen, « Node21-Generative Inpainting »,

[219] Y. Zeng, Z. Lin, H. Lu et V. M. Patel, « CR-Fill : Generative image inpainting with auxiliary contextual reconstruction », in *Proceedings of the IEEE/CVF international conference on computer vision*, 2021, p. 14 164-14 173.

[220] N. Granot, B. Feinstein, A. Shocher, S. Bagon et M. Irani, « Drop the gan : In defense of patches nearest neighbors as single image generative models », in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2022, p. 13 460-13 469.

[221] The shift project, *Climat, crises : Le plan de transformation de l'économie française*. Odile Jacob, 2022.

[222] R. Desislavov, F. Martinez-Plumed et J. Hernández-Orallo, « Trends in AI inference energy consumption : Beyond the performance-vs-parameter laws of deep learning », *Sustainable Computing : Informatics and Systems*, t. 38, p. 100 857, 2023.

[223] D. Patterson, J. Gonzalez, Q. Le et al., « Carbon emissions and large neural network training », *arXiv preprint arXiv :2104.10350*, 2021.

[224] D. Patterson, J. Gonzalez, U. Hölzle et al., « The carbon footprint of machine learning training will plateau, then shrink », *Computer*, t. 55, *7*, p. 18-28, 2022.

[225] P. Li, J. Yang, M. A. Islam et S. Ren, « Making AI Less" Thirsty" : Uncovering and Addressing the Secret Water Footprint of AI Models », *arXiv preprint arXiv :2304.03271*, 2023.

[226] M. J. Cardoso, W. Li, R. Brown et al., « Monai : An open-source framework for deep learning in healthcare », *arXiv preprint arXiv :2211.02701*, 2022.

[227] F. Pérez-García, R. Sparks et S. Ourselin, « TorchIO : a Python library for efficient loading, preprocessing, augmentation and patch-based sampling of medical images in deep learning », *Computer Methods and Programs in Biomedicine*, t. 208, p. 106 236, 2021.

COLLEGE DOCTORAL BRETAGNE SCIENCES DE LA VIE ET DE LA SANTE

Université de Bretagne Occidentale

**Titre :** Apprentissage génératif à bas régime de données pour la segmentation d'images en oncologie

**Mot clés :** Apprentissage profond, Augmentation de données, Écart de domaine, Apprentissage few-shot, Segmentation

**Résumé :** En apprentissage statistique, la performance des modèles est influencée par divers biais inhérents aux données, comme la rareté des données (*data scarcity*) et le décalage de domaine (*domain shift*). Cette thèse s'intéresse à réduire leur influence dans le cadre de la segmentation de structures pathologiques en imagerie médicale. En particulier, notre objectif est de réduire les écarts de données au niveau des régions d'intérêt (ROI) entre les domaines source d'entraînement et cible de déploiement, qu'ils soient intrinsèques aux données ou induits par la faible quantité de données disponibles. Dans ce but, nous proposons une stratégie d'augmentation de données adaptative, basée sur l'analyse de la distribution des intensités des ROI dans le domaine de déploiement. Une première contribution, que nous qualifions d'augmentation naïve, consiste à altérer l'apparence des ROI du domaine d'entraînement pour mieux correspondre aux caractéristiques des ROI du domaine de déploiement. Une seconde étape, complétant la première, rend l'altération plus réaliste par rapport aux propriétés du domaine cible grâce à un modèle génératif d'harmonisation *one-shot*, applicable dans toutes les situations de disponibilité de données. De cette façon, nous renforçons la robustesse du modèle de segmentation en aval pour les ROI dont les caractéristiques étaient initialement sous-représentées à l'entraînement. Nous évaluons notre approche à différents régimes de données et divers contextes cliniques, notamment en IRM, TDM et radiographie pulmonaire. En outre, notre approche a montré des résultats impressionnants lors d'un challenge de segmentation de tumeurs à la conférence MICCAI 2022.

**Title:** Generative learning at low data regime for image segmentation in oncology

**Keywords:** Deep learning, Data augmentation, Domain shift, Few-shot learning, Segmentation

**Abstract:** In statistical learning, the performance of models is affected by various biases present within the data, including data scarcity and domain shift. This thesis focuses on reducing their impact in the field of pathological structure segmentation in medical imaging. Our goal is to minimize data discrepancies at the region of interest (ROI) level between the training source domain and the target deployment domain, whether it is intrinsic to the data or caused by the limited data availability. To this end, we present an adaptive data augmentation strategy, based on the analysis of the intensity distribution of the ROIs in the deployment domain. A first contribution, which we call naive augmentation, consists of altering the appearance of the training ROIs to better match the characteristics of the ROIs in the deployment domain. A second augmentation, complementing the first, makes the alteration more realistic relative to the properties of the target domain by harmonizing the characteristics of the altered image. For this, we employ a generative model trained on a single unlabeled image from the deployment domain (one-shot approach), making this technique usable in any data regime encountered. In this way, we enhance the robustness of the downstream segmentation model for ROIs whose characteristics were initially underrepresented in the deployment domain. The effectiveness of this method is evaluated under various data regimes and in different clinical contexts (MRI, CT, CXR). Our approach demonstrated impressive results in a tumor segmentation challenge at MICCAI 2022.