Titre: Title:	Modélisation du mouvement tridimensionnel du foie à partir d'images échographiques 2D par auto-encodeurs convolutionnels
Auteur: Author:	Tal Mezheritsky
Date:	2021
Type:	Mémoire ou thèse / Dissertation or Thesis
Référence: Citation:	Mezheritsky, T. (2021). Modélisation du mouvement tridimensionnel du foie à partir d'images échographiques 2D par auto-encodeurs convolutionnels [Mémoire de maîtrise, Polytechnique Montréal]. PolyPublie. https://publications.polymtl.ca/6292/

POLYTECHNIQUE MONTRÉAL
affiliée à l'Université de Montréal

Modélisation du mouvement tridimensionnel du foie à partir d'images échographiques 2D par auto-encodeurs convolutionnels

TAL MEZHERITSKY
Institut de génie biomédical

Mémoire présenté en vue de l'obtention du diplôme de *Maîtrise ès sciences appliquées*
Génie biomédical
Mai 2021

POLYTECHNIQUE MONTRÉAL
affiliée à l'Université de Montréal

Ce mémoire intitulé :

Modélisation du mouvement tridimensionnel du foie à partir d'images échographiques 2D par auto-encodeurs convolutionnels

présenté par **Tal MEZHERITSKY**
en vue de l'obtention du diplôme de *Maîtrise ès sciences appliquées*
a été dûment accepté par le jury d'examen constitué de :

Jean PROVOST, président
Samuel KADOURY, membre et directeur de recherche
Hassan RIVAZ, membre externe

DÉDICACE

À tous ceux qui m'ont soutenu.

REMERCIEMENTS

J'aimerais d'abord remercier mon directeur de recherche Samuel Kadoury de m'avoir donné l'opportunité de travailler sur ce projet de maîtrise complexe et stimulant. Je le remercie également de m'avoir encouragé à repousser mes limites et à développer mon autonomie en tant que chercheur.

Je remercie les Drs Hassan Rivaz et Jean Provost d'avoir accepté de faire partie de mon jury.

Merci à ma mentore, collègue et amie, Liset, de m'avoir accompagné à chaque étape de ce projet. Je n'aurais pas pu accomplir tout ce que j'ai accompli au cours de ces deux dernières années sans son soutien et ses conseils. Son éthique de travail exceptionnelle, ainsi que sa rigueur, m'ont toujours inspiré à donner le meilleur de moi-même et de garder la tête haute même dans les situations difficiles.

Merci à tous les membres présents et passés du laboratoire MedICAL. Je garderai de bons souvenirs de mes conversations et interactions avec vous, malgré cette fin de maîtrise à distance.

J'aimerais remercier également le CRSNG et le FRQNT pour le soutien financier qui m'a permis de me concentrer pleinement sur mon projet de maîtrise.

Merci à ma famille et mes proches qui m'ont offert leur encouragement et support inconditionnel tout au long de ce chapitre de ma vie.

Finalement, j'aimerais remercier mon amoureuse d'avoir été à mes côtés à travers les hauts et les bas. Merci de m'avoir soutenue et d'avoir cru en moi, même quand j'en étais incapable.

RÉSUMÉ

Le mouvement engendré par la respiration est un facteur de complication pour les procédures de traitement du cancer du foie comme la radiothérapie externe. Afin d'assurer la livraison de l'irradiation au tissu cancéreux, la position en 3D de la tumeur et des structures à risque doit être suivie au cours de la procédure. Des solutions de guidage par imagerie ont été utilisées en clinique afin de compenser le mouvement respiratoire en ajustant la trajectoire du faisceau d'irradiation. Toutefois, la plupart des systèmes commerciaux utilisent des modalités d'imagerie telles que les rayons X ou la tomographie à faisceau conique qui appliquent une dose supplémentaire d'irradiation au patient en plus d'avoir des fréquences d'acquisition trop lentes pour les applications en temps réel. L'imagerie par US, quant à elle, permet d'effectuer des acquisitions en 2D, 3D et 4D pour un coût relativement bas, le tout sans appliquer d'irradiation ionisante. Pour ces raisons, l'US est une modalité intéressante pour les systèmes de radiothérapie guidée par l'image. Cependant, l'US 2D ne permet pas de suivre les tumeurs qui se déplacent en 3D tandis que l'US 3D possède un temps d'acquisition et de traitement trop long. Par conséquent, une approche hybride qui permet de générer des volumes d'US 3D à partir d'images d'US 2D acquises en temps réel peut constituer une solution pour les systèmes de radiothérapie guidée par l'US.

Les récents développements en apprentissage profond ont accéléré l'innovation dans le domaine de l'analyse d'images médicales. Des approches de modélisation du mouvement basées sur l'apprentissage profond ont été développées pour des modalités telles que l'IRM et le CT. Toutefois, peu de travaux ont été axés sur la génération de volume d'US 3D.

Dans le présent mémoire, nous présentons une méthode de modélisation du mouvement basée sur l'apprentissage profond qui permet de générer des volumes d'US 3D mis à jour à partir d'un nombre limité de volumes de prétraitement ainsi qu'une séquence d'images d'US 2D. L'étude de différentes architectures a souligné le potentiel des autoencodeurs convolutionels pour accomplir notre tâche. Notre modèle apprend une représentation de basse dimension commune entre des champs de mouvement en 3D et des images d'US 2D correspondantes. Les champs de mouvements 3D générés à partir des images 2D permettent de déformer le volume de référence de prétraitement et d'obtenir un nouveau volume du foie en 3D, même pour de nouveaux sujets. Nous introduisons des améliorations à l'autoencodeur convolutionel qui l'adaptent pour la tâche de modélisation du mouvement respiratoire. Le modèle est validé sur un ensemble de données de 20 volontaires à l'aide de métriques de similarité d'image et de suivi de cibles. Une erreur de localisation de cible de 3.5 ± 2.4 mm est rapportée, montrant le fort potentiel de notre méthode pour les applications en radiothérapie guidée par l'US.

ABSTRACT

Respiratory motion is a complicating factor for liver cancer treatments such as external beam radiotherapy. In order to ensure proper delivery of radiation to the treatment site, the 3D position of the tumoral volume and organs at risk must be tracked at all times during the procedure. Image-guided solutions have been used clinically to account for respiratory motion and adjust the beam trajectory to optimize treatment delivery. However, most commercial solutions rely on imaging modalities such as X-ray imaging or cone-beam computed tomography which apply an additional radiation dose to the patient and have acquisition times that are too long for real-time applications. Contrastingly, ultrasound (US) imaging offers 2D, 3D and 4D imaging capabilities for a relatively low cost without imparting ionizing radiation to the patient. US imaging is therefore a viable candidate for image-guided radiation therapy systems. However, 2D US can only provide in-plane guidance while 3D US volumes provide complete spatial information but take longer to acquire and to analyze afterwards. Consequently, a hybrid approach which generates 3D US volumes from 2D US images acquired in real-time could constitute a viable solution for US-guided radiation therapy systems.

In recent years, developments in deep learning for computer vision have accelerated innovation in the medical image analysis field. Approaches for deep learning based motion models have been proposed for modalities such as MRI and CT however very few works focus on 3D US generation from 2D images.

In the present work, we propose a motion modelling method based on deep learning which allows to generate up-to-date 3D US volumes by using a limited number of pre-treatment volumes and a sequence of 2D US images. A study of different model architectures revealed the potential of convolutional autoencoder based models to complete the aforementioned task. Our proposed model learns a common low-dimensional representation between 3D US motion fields and corresponding 2D US images. The model is then able to generate a 3D motion field from an input 2D US image, even for previously unseen subjects. The generated motion field is used to deform a reference volume acquired before treatment, thus providing an updated 3D representation of the liver. We introduce improvements over the traditional convolutional autoencoder which adapt it to the specific task of respiratory motion modelling. We validate our method on a dataset of 20 healthy volunteers using image similarity and target tracking metrics achieving a mean target reconstruction error of 3.5 ± 2.4 mm. The reported results showcase the strong potential of our method for US guided radiotherapy applications.

TABLE DES MATIÈRES

LISTE DES TABLEAUX

LISTE DES FIGURES

LISTE DES SIGLES ET ABRÉVIATIONS

US	Ultrason
IRM	Imagerie par résonance magnétique
CBCT	Tomographie à faisceau cubique
RTE	Radiothérapie externe
SI	Supérieur-inférieur
AP	Antérieur-postérieur
LR	Gauche-droite
CHC	Carcinome hépatocellulaire
TACE	Chimioembolisation transartérielle
LINAC	Accélérateur linéaire
CT	Tomographie calculée
PET	Tomographie par émission de positrons
GTV	Volume brut de la tumeur
CTV	Volume clinique de la tumeur
ITV	Volume tumoral interne
PTV	Volume de planification
OAR	Organe à risque
RTGI	Radiothérapie guidée par l'imagerie
PCA	Analyse par composantes principales
CNN	Réseaux de neurones convolutifs
AE	Autoencodeur
CAE	Autoencodeur contractif
DAE	Autoencodeur de débruitage
VAE	Autoencodeur variationnel
CVAE	Autoencodeur variationnel conditionnel
STN	Réseau de transformation spatiale
HIFU	Échographie focalisée haute intensité

CHAPITRE 1 INTRODUCTION

Selon l'Organisation mondiale de la santé, le cancer du foie est le troisième cancer le plus mortel en 2020, causant plus d'un demi-million de morts chaque année à travers le monde [1]. Parmi les traitements disponibles, la radiothérapie externe (RTE) est une modalité de traitement du cancer qui applique une dose d'irradiation ionisante au tissu tumoral dans le but de causer la mort cellulaire. Les avantages de cette modalité de traitement sont sa capacité d'appliquer de hautes doses d'irradiation de manière très sélective en évitant d'endommager des tissus sains. L'administration d'un tel traitement requiert beaucoup de planification afin de définir l'emplacement de la cible et d'organes à risque environnants. Le principal défi de la planification de la radiothérapie externe est de quantifier l'incertitude quant à la position de la tumeur lors de l'intervention. Dans le cas des organes abdominaux comme le foie, c'est le mouvement généré par la respiration du patient qui cause les plus grands déplacements de la tumeur. Il est donc nécessaire d'utiliser des stratégies de gestion de la respiration afin d'optimiser la livraison de la radiation au site tumoral.

Mise à part l'augmentation des marges de traitement, différentes techniques de respiration qui soit limitent l'amplitude du mouvement respiratoire ou qui l'éliminent temporairement peuvent être utilisées. Elles comportent malgré tout certaines limites liées à leur reproductibilité et la physiologie des patients atteints de cancer. L'imagerie médicale est utilisée dans le cadre de la RTE pour plusieurs étapes incluant la planification du traitement et la vérification de la position du patient le jour du traitement. Elle peut aussi être utilisée afin de déterminer la position de la cible de traitement lors de l'intervention. Des modalités comme l'imagerie par rayons X ou la tomographie à faisceau cubique (CBCT) sont couramment utilisées dans ce contexte. Toutefois, leur nature ionisante constitue une limitation à leur utilisation. L'imagerie par ultrason (US), quant à elle, est basée sur l'envoi et la recapture d'ondes acoustiques ultrasonores dans le but d'imager le corps humain. Les avantages principaux de cette modalité sont sa nature non ionisante, sa portabilité et son faible coût.

Une majorité des systèmes de suivi de cibles pour la RTE sont basés sur l'imagerie en 2D. Toutefois, les tumeurs du foie subissent des déplacements complexes en 3D. Bien que l'imagerie 3D est possible avec des modalités comme la CBCT ou l'US, leurs temps d'acquisition et temps de traitement ne sont pas compatibles avec le suivi en temps réel. Pour cette raison, des approches permettant d'obtenir de l'information en 3D sur l'emplacement de la tumeur à partir d'information en 2D sont considérées comme des pistes de solution pour l'amélioration de l'efficacité de la RTE.

Les modèles de mouvement tentent d'estimer le mouvement subit par tout l'organe imagé, donnant de l'information sur l'emplacement de la cible de traitement, mais aussi sur celui d'organes à risque. Le principe de base derrière les modèles de mouvement est l'établissement d'une correspondance entre des champs de mouvement denses en 3D avec des signaux substituts qui sont simples à acquérir et qui peuvent être représentés en une ou deux dimensions. Un modèle de mouvement peut être spécifique à un seul patient ou être constitué de données de plusieurs patients. Malgré les résultats prometteurs de ces modèles sur des données patients, ils comportent tout de même des limitations importantes. Notamment, la construction d'un modèle à patient unique requiert l'acquisition de données 4D pour chaque nouveau patient et les modèles de population nécessitent une étape de préparation de données coûteuse en temps.

Avec l'essor de l'apprentissage profond, de nouvelles manières de construire des modèles de mouvement ont été proposées. Des modèles d'apprentissage non supervisé comme les autoencodeurs (AE) convolutifs permettent de représenter des données d'imagerie complexes d'une manière compacte qui se porte bien pour des applications de modélisation de mouvement. L'avantage des approches d'apprentissage profond par rapport aux modèles de mouvement conventionnels est leur capacité d'apprentissage avec très peu d'intervention humaine. Ces approches ne requièrent pas de données annotées ou préalignées pour être optimisées. Ceci rend les modèles d'apprentissage profond très flexibles et puissants lorsque la quantité de données d'entraînement est suffisante.

Toutefois, très peu de travaux se sont concentrés sur la construction de modèles de mouvement pour l'imagerie par US 3D à partir d'un signal substitut constitué d'images d'US 2D. De plus, l'application des notions récentes en apprentissage profond pour la modélisation de mouvement n'a pas été suffisamment étudiée pour introduire ces techniques en clinique. Le but principal de ce projet de maîtrise est donc le développement de modèles basés sur l'apprentissage profond dans le but d'inférer le mouvement 3D subit par le foie à partir d'images d'US 2D.

1.1 Plan du mémoire

Ce mémoire de maîtrise est organisé de la manière suivante. Dans le chapitre **2**, une revue de la littérature sur les différents aspects pertinents pour ce projet de maîtrise est effectuée. Les sujets abordés sont le foie, la radiothérapie, l'imagerie par US et les méthodes d'estimation du mouvement en 3D à partir d'images en 2D. Le chapitre **3**, présente la méthodologie du travail de recherche incluant l'acquisition des données. Les chapitres **4** et **5** présentent deux modèles de génération de volumes d'US 3D à partir d'images 2D. Le chapitre **6** présente une discussion générale sur la solution proposée, ses limites et sa pertinence clinique avant de conclure ce travail au chapitre **7**.

CHAPITRE 2 REVUE DE LITTÉRATURE

2.1 Anatomie et physiologie du foie

Dans cette section, nous introduirons l'organe sur lequel se concentre ce projet de maîtrise, le foie. D'abord, une description de l'organe, de son emplacement et de ses fonctions principales est fournie. Ensuite, les causes et caractéristiques principales du mouvement du foie sont présentées. Finalement, les circonstances menant au cancer du foie ainsi que les modalités de traitement existantes sont décrites.

2.1.1 Le foie

Le foie est le plus massif des organes abdominaux chez l'humain pesant en moyenne 1.5 kg. Le foie se trouve dans la partie supérieure droite de l'abdomen, à droite de l'estomac et au-dessus du duodénum. Logé sous la cage thoracique, le foie est séparé des poumons et du coeur par le diaphragme. Le foie est entièrement recouvert d'une membrane fibreuse appelée «capsule de Glisson» et est partiellement protégé par les côtes [2, 3]. La figure 2.1 montre l'emplacement du foie dans le corps humain. Ayant une forme asymétrique, le foie

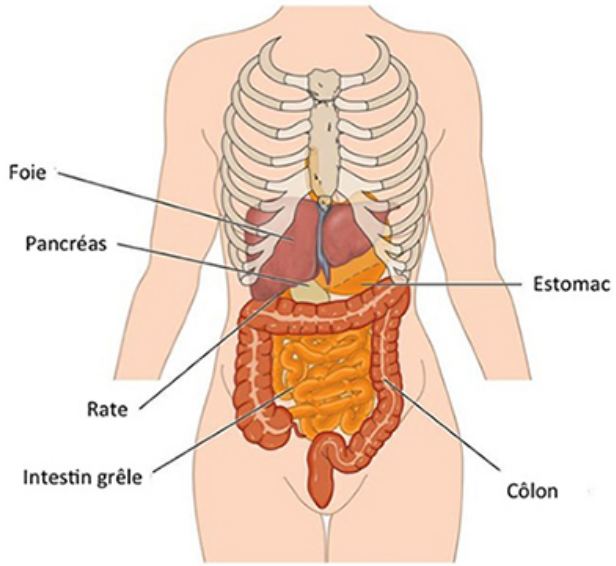


Figure 2.1 L'emplacement du foie dans le corps humain. Image tirée de [4]

peut être divisé en deux lobes de taille inégale, soit le lobe gauche et le lobe droit. Le foie est un organe très vascularisé et possède deux sources d'approvisionnement en sang. Du sang

riche en oxygène provient du coeur par l'artère hépatique, puis du sang riche en nutriments provient du système digestif par la veine porte. Les veines hépatiques permettent au sang de quitter le foie par la veine cave inférieure et retourner vers le coeur [3].

Le foie est aussi l'organe responsable du plus grand nombre de réactions chimiques dans le corps humain et remplit plusieurs fonctions essentielles au bon fonctionnement de l'organisme. Notamment, le foie gère le métabolisme de plusieurs nutriments issus de la digestion comme les glucides et lipides. Il est responsable de la synthèse de la majorité des protéines sanguines telle que l'hémoglobine. Les cellules du foie s'occupent également de la dégradation d'une grande variété de substances toxiques comme l'éthanol et l'ammoniaque en substances non toxiques qui sont ensuite éliminées par voie rénale ou par les selles [4].

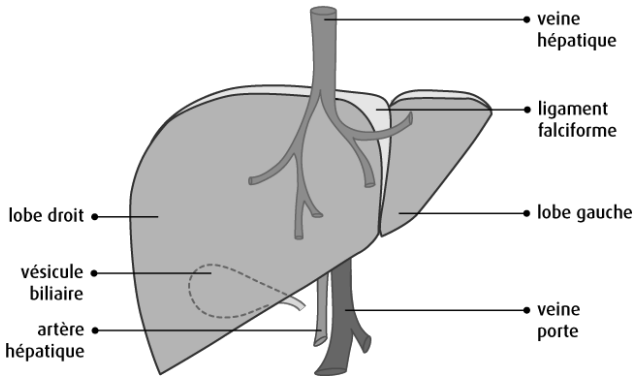


Figure 2.2 Structure globale du foie humain et ses vaisseaux principaux. Image tirée de [3]

2.1.2 Mouvement du foie

La respiration est un processus biomécanique qui est composé de deux phases, l'inspiration et l'expiration. Ce processus quasi périodique permet l'échange de gaz (oxygène et dioxyde de carbone) entre le corps et son environnement. L'inspiration requiert la contraction du diaphragme ainsi que des muscles intercostaux situés entre les côtes. La contraction du diaphragme le déplace dans la direction inférieure, créant ainsi une pression négative dans la cage thoracique qui force l'entrée d'air dans les poumons. Les muscles intercostaux, quant à eux, déplacent les côtes inférieures dans les directions antérieure et supérieure, contribuant ainsi à l'augmentation du volume pulmonaire. Lors de l'expiration, le diaphragme et les muscles intercostaux se décontractent permettant au volume pulmonaire de diminuer de manière passive due à l'élasticité des poumons et de la cage thoracique. Il est important à noter qu'il existe un phénomène d'hystérésis entre le volume et la pression des poumons. C'est-à-dire que le volume pulmonaire est différent entre les phases d'inspiration et d'expiration pour la

même pression d'air. Une multitude de variables affectent l'amplitude ainsi que la fréquence de la respiration chez un même individu. Notamment, la position du corps, le temps passé depuis le dernier repas ainsi que l'état émotionnel [5, 6].

Dû à la proximité des poumons à l'abdomen, les mouvements engendrés par la respiration décrits ci-dessus affectent également la position d'organes comme le foie, l'estomac et le pancréas. Selon l'Association américaine des Physiciens médicaux, le déplacement des organes abdominaux dans la direction supérieure inférieure (SI) varie en moyenne entre 10-25 mm [7]. Les déplacements dans les directions antérieures postérieures (AP) et gauches droites (LR) ont une amplitude moindre qui ne dépasse pas 2 mm en général. Dans le cas du foie, il existe également des variations dans la direction et l'amplitude du mouvement de ses différents segments [8].

2.1.3 Le cancer du foie

Le cancer du foie primaire est le sixième cancer le plus diagnostiqué au monde et cause plus d'un demi-million de morts chaque année à travers le monde. Selon l'Organisation mondiale de la santé, le cancer du foie est le quatrième cancer le plus mortel après le cancer des poumons, le cancer colorectal et le cancer de l'estomac. L'incidence du cancer du foie varie selon le sexe étant plus fréquent chez les hommes [9]. Une variation géographique existe également, atteignant un maximum dans l'Asie de l'Est. Le pronostic de survie est mauvais avec seulement 19% de survie après 5 ans en moyenne [10].

Le cancer du foie peut être divisé en deux types, les cancers primaires et secondaires. Dans le cas des cancers primaires, la tumeur cancéreuse a comme origine le foie lui-même. Le type de cancer primaire le plus fréquent est le carcinome hépatocellulaire (CHC). Le CHC est un cancer agressif qui se développe plus fréquemment dans les foies cirrhotiques [11]. Dans le cas des cancers secondaires, la tumeur se développe dans un autre organe initialement puis se répand au foie par le biais de métastases. La propagation métastatique survient fréquemment durant le développement de tumeurs solides [12]. Lorsqu'une tumeur atteint une taille suffisante, certaines de ses cellules peuvent s'en détacher puis se propager au foie par le sang ou le système lymphatique [13]. Les cancers les plus susceptibles de se propager vers le foie sont le cancer colorectal, le cancer des poumons et le cancer du sein [14].

Plusieurs traitements ont été proposés afin de guérir le cancer du foie ou dans les cas plus avancés, prolonger la vie du patient. Pour les cancers du foie primaires et secondaires, le retrait de la tumeur par intervention chirurgicale est une des solutions donnant le meilleur pronostic de survie. Par contre, très peu de patients sont admissibles à être opérés. Un autre traitement curatif est la transplantation du foie qui offre un taux de survie de plus de 70%.

Toutefois, il n'est pas toujours possible de trouver un donneur compatible. Étant donné le taux de réponse très bas (<20%) à la chimiothérapie systémique dans le cas du CHC, des traitements locaux, non chirurgicaux ont été proposés [14]. Parmi ces traitements, la chimioembolisation transartérielle (TACE) est un traitement qui vise à administrer un agent de chimiothérapie directement au site de la tumeur par le biais de ses artères principales. Ensuite, un agent d'embolisation vient bloquer l'alimentation en sang de la tumeur, la privant d'oxygène et de nutriments. Ceci entraîne la nécrose du tissu cancéreux [11]. Une autre approche de traitement localisée est l'ablation par radiofréquences. Un courant alternatif de haute fréquence est utilisé pour générer de la chaleur qui est appliquée au tissu cancéreux afin de causer la mort cellulaire [14]. Toutefois, lorsque la tumeur se trouve à proximité de vaisseaux principaux du foie ou près de la vésicule biliaire, l'ablation par radiofréquences ainsi que la TACE sont déconseillées [15]. Dans les dernières années, la radiothérapie externe, livrée en une dose ou par fractions, est émergée comme une solution de rechange aux traitements chirurgicaux et interventionnels pour les métastases du foie [12, 15] et le CHC [16]. Les avancées technologiques ont permis de mieux cibler l'administration de la radiothérapie au site tumoral sans endommager les tissus sains environnants.

2.2 Traitement par radiothérapie

Dans cette section, la radiothérapie externe en tant que thérapie pour le cancer du foie sera présentée. D'abord, le but général de ce type de traitement sera expliqué. Ensuite, les principes physiques derrière la génération et l'administration de l'irradiation ionisante seront exposés. Pour poursuivre, les étapes de planification et d'administration du traitement seront présentées. Finalement, les effets de la respiration sur la RTE ainsi que les approches de gestion de la respiration seront décrits.

2.2.1 La radiothérapie externe

Le principe fondamental de la RTE est d'appliquer une dose concentrée d'irradiation ionisante aux tissus cancéreux de manière très précise. La dose d'irradiation peut être appliquée en une séance ou divisée en plusieurs fractions [17]. Le but de la RTE est de détruire le matériel génétique des cellules cancéreuses. Ce faisant, l'apoptose de celles-ci est induite, ce qui freine la progression du cancer ou l'élimine complètement. Toutefois, l'irradiation utilisée peut aussi être dommageable aux cellules saines environnantes [18]. Pour cette raison, l'administration de la RTE requiert non seulement la connaissance de l'emplacement de la tumeur à traiter, mais également la connaissance de la position des tissus et organes qui l'entourent. Le défi principal de la RTE est donc de maximiser la dose d'irradiation à la tumeur tout en

minimisant la dose appliquée aux tissus sains environnants.

2.2.2 Principes de fonctionnement

La RTE peut être administrée avec différents appareils spécialisés tels que le CyberKnife, le scalpel gamma ou l'accélérateur linéaire (LINAC). Le LINAC est utilisé dans la plupart des traitements de RTE [19]. La plupart des appareils de RTE requièrent l'aménagement d'une salle de traitement spécialisée où se trouve le système de traitement, des systèmes de positionnement et d'imagerie. Afin de permettre la surveillance du déroulement du traitement par le personnel soignant, une salle de contrôle adjacente à la salle de traitement est aménagée avec une vue sur le patient. La salle de contrôle comporte également tout le matériel informatique pour administrer le traitement selon le plan établi ainsi qu'un système de communication pour donner des instructions vocales au patient.

La figure 2.3 montre une vue d'ensemble des composantes principales d'un accélérateur linéaire utilisé pour les traitements de RTE. Le statif est l'élément qui fixe le LINAC au sol. Le bras du LINAC est fixé au statif par un axe de rotation qui lui permet de repositionner la tête de l'accélérateur autour de l'isocentre de l'appareil. À la sortie de la tête, le faisceau de radiation passe par une série de collimateurs afin d'adapter la forme de celui-ci à la tumeur traitée. Finalement, la table de traitement permet d'installer le patient lors du traitement. Dans la plupart des systèmes, la position de la table peut être ajustée avec plusieurs degrés de liberté afin de faire correspondre la position du patient à celle établie lors de la planification du traitement [20].

Deux types de rayonnements peuvent être utilisés dans le cadre de la RTE, soit les photons rayons X ou des électrons. Le processus de base pour générer le faisceau d'irradiation se résume comme suit. Dans le canon à électrons, un filament est chauffé afin de produire un faisceau d'électrons par émission thermoïonique. Les électrons émis sont ensuite acheminés vers le guide d'onde où les électrons sont accélérés par l'application d'un champ de radiofréquences de haute intensité jusqu'à l'atteinte du niveau d'énergie cinétique requis (entre 4 et 35 MeV). Suite à l'accélération, le faisceau d'électrons est dirigé vers la cible de rayons X [20]. Traditionnellement, la cible est composée de wolfram pour son haut nombre atomique (Z=74) et son haut point de fusion (3370°C) [21]. Plusieurs autres matériaux sont utilisés pour la production de rayons X, leur choix dépend principalement du niveau d'énergie désiré à la sortie. Dans le cas ou des faisceaux d'électrons sont utilisés pour le traitement, la cible de rayons X est tout simplement retirée. Avant d'atteindre le patient, l'étendue et la forme du faisceau d'irradiation sont conformées par une série de collimateurs intégrés à la tête du LINAC. Les collimateurs primaire et secondaire définissent une zone rectangulaire de taille variable qui

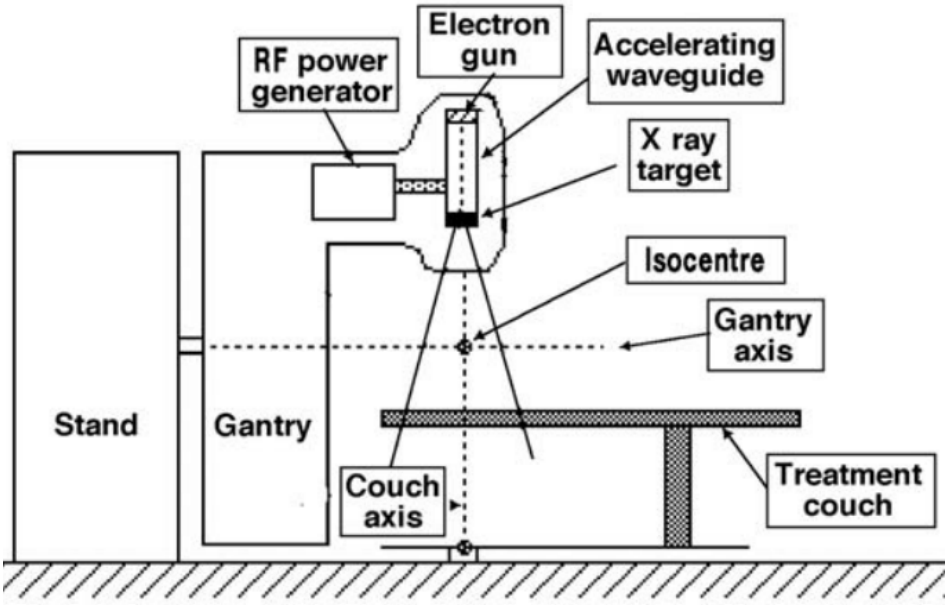


Figure 2.3 Vue d'ensemble des composantes d'un LINAC. Image tirée de [20]

sera irradiée par le faisceau. Suite aux collimateurs principaux, un collimateur multilames (MLC) permet d'appliquer des conformations plus complexes au faisceau d'irradiation afin d'appliquer une dose seulement au tissu à traiter [20].

2.2.3 Étapes du traitement

Un patient atteint du cancer du foie qui se voit prescrire un traitement par RTE va passer par différentes étapes de préparation et de planification avant de recevoir son traitement. La première étape de prétraitement consiste à établir la position de référence du patient et acquérir des données d'imagerie de la zone à traiter. Ensuite, une étape de planification de traitement va identifier les endroits où la radiothérapie devra être appliquée, ainsi que les endroits qui seront à éviter. L'étape de planification comprend aussi la validation de tous les paramètres de traitement incluant le positionnement du patient, l'orientation des faisceaux et les doses prescrites. Lors de chaque jour de traitement, l'équipe médicale commence par s'assurer que tous les paramètres établis lors de la planification sont respectés. Une fois que tout est en place, le traitement est administré au patient sous la supervision du personnel soignant qui peut arrêter le traitement à tout moment si une déviation du plan est détectée.

L'étape de prétraitement a comme but de fournir au radiooncologue les données d'imagerie nécessaires à la planification du traitement. Pour ce faire, l'équipe de radio-oncologie établit d'abord une position de référence qui peut être reproduite de manière fiable lors de chacun des

traitements. De l'équipement de positionnement et d'immobilisation est utilisé pour tenter de minimiser toute source de mouvement du patient dû à la respiration par exemple [17]. Une fois qu'une position de référence convenable est établie, la partie du corps à traiter est imagée. Des images de tomographie calculée (CT) en 3D avec ou sans agent de contraste sont acquises afin de visualiser les tissus cancéreux ainsi que les tissus et organes environnants. Afin d'évaluer le mouvement causé par la respiration du patient, des acquisitions en 4D peuvent être effectuées. Dans certains cas, des acquisitions par résonance magnétique (IRM) ou de tomographie par émission de positrons (PET) sont requises afin d'améliorer la définition de la cible de traitement [15].

Avec les données d'imagerie, il est possible d'entamer l'étape de planification du traitement. Tout d'abord, le radiooncologue va définir tous les volumes d'intérêt à la planification. Le volume brut de la tumeur (GTV) représente l'étendue visible de la tumeur selon les données d'imagerie disponibles. Le volume clinique de la tumeur (CTV) ajoute une marge fixe ou variable au GTV afin de tenir compte des parties de la tumeur qui sont invisibles sur les images de prétraitement. Ensuite, le volume tumoral interne (ITV) étend les contours du CTV pour tenir compte du changement de taille ou de position que peut subir la tumeur. Le volume de planification (PTV) tient compte des incertitudes liées au positionnement du patient, les erreurs des machines, ainsi que les variations entre les traitements. Finalement, le radiooncologue identifie les organes à risque (OAR) autour de la tumeur. Les OAR sont des organes qui ont une basse tolérance à la radiation. Il faut donc éviter de dépasser leur seuil d'irradiation limite [22]. La figure 2.4 montre un schéma de la relation entre tous les volumes d'intérêt. Suite à l'identification des volumes d'intérêt, les trajectoires des faisceaux d'irradiation sont établies de manière à livrer la dose prescrite par le radiooncologue à la tumeur tout en minimisant la dose reçue par les tissus sains et les OAR.

Une fois la planification terminée, l'équipe médicale procède à une simulation du traitement pour assurer entre autres que les faisceaux planifiés livreront la dose requise [15]. Dans le cas où le LINAC utilisé est monté sur un bras robotique comme le système CyberKnife et que du matériel d'immobilisation est requis, il faut s'assurer qu'aucune collision ne surviennent durant le déplacement du LINAC.

Finalement, le traitement peut être administré en une séance ou divisé en fractions. À chaque séance, le personnel doit s'assurer que la position du patient correspond à celle établie lors de l'étape de prétraitement afin de respecter le plan de traitement. Ceci est accompli avec l'aide d'équipement d'immobilisation. Il est également possible d'utiliser l'imagerie médicale dans la salle de traitement afin de comparer la position du patient avec celle imagée lors du prétraitement.

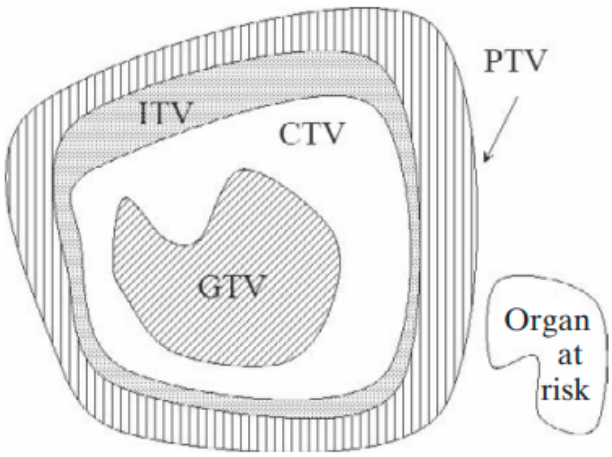


Figure 2.4 Représentation schématique des volumes d'intérêt identifiés pour la planification du traitement par RTE. Image tirée de [22]

2.2.4 Gestion de la respiration en RTE

Dans le contexte d'un traitement du cancer du foie par RTE, le mouvement du foie causé par la respiration du patient représente une source d'erreur lors de l'administration du traitement. Principalement, le mouvement important de la tumeur lors du traitement entraîne une réduction de la dose reçue par le tissu cancéreux et une augmentation de la dose des tissus sains. Ces déviations de la dose planifiée peuvent causer des maladies du foie radio-induites lorsque du tissu hépatique sain reçoit une dose d'irradiation significativement supérieure à son seuil de tolérance radiobiologique [8].

Plusieurs solutions ont été proposées afin de gérer le mouvement dû à la respiration durant les traitements de radiothérapie. Ces solutions peuvent être séparées en deux approches, les approches non adaptatives et adaptatives [14]. Les approches non adaptatives sont plus simples à implémenter, mais comportent plusieurs limitations. Une première approche non adaptative propose de tout simplement augmenter les marges de traitement selon l'amplitude du mouvement de la tumeur observée durant les acquisitions de planification [7]. Selon [8], l'augmentation des marges ajoutées au CTV afin de tenir compte du mouvement de respiration sont d'au moins 2.5 mm, 2.5 mm et 5 mm dans les directions LR, AP et SI, respectivement. Le désavantage de cette solution est qu'en augmentant les marges de traitement, les tissus sains seront plus exposés à la radiation.

D'autres solutions non adaptatives comme le «respiratory gating» appliquent le faisceau d'irradiation seulement lorsque le patient se trouve dans une position précise du cycle respiratoire. Cette technique prolonge la durée du traitement en plus de nécessiter l'utilisation

de signaux respiratoires externes où des marqueurs qui sont implantés dans le corps du patient [7]. Dans le premier cas, les signaux externes peuvent parfois être peu représentatifs de l'état de l'organe à l'intérieur du corps. Dans le deuxième cas, l'implantation de marqueurs requiert une intervention chirurgicale.

La technique «breath hold» se fie sur le patient pour retenir son souffle de manière reproductible afin d'administrer la dose d'irradiation seulement pendant cette période. Dans ce cas, le foie est immobile et se trouve à la même position qu'à la planification. Les désavantages de cette approche sont, entre autres, la nécessité d'entraîner les patients à retenir leur souffle de manière fiable et reproductible. Il y a donc un élément d'erreur humaine qui découle directement du patient [7].

Une manière de restreindre l'amplitude de respiration du patient est d'appliquer une pression sur son abdomen. C'est le principe derrière l'approche «forced shallow breathing» où une plaque est utilisée afin de réduire la distance parcourue par la tumeur en permettant quand même au patient de respirer minimalement. Malgré l'efficacité de cette technique [23], certains patients ont des limitations physiologiques qui ne permettent pas l'utilisation de cette approche.

Les approches adaptatives, quant à elles, tentent d'adapter l'administration du traitement selon la position changeante de la tumeur. Dans le cas idéal, ces approches permettraient de suivre la tumeur, repositionner le faisceau d'irradiation et d'adapter la dose administrée, le tout en temps réel. Pour suivre la tumeur en temps réel, il est possible soit d'imager la tumeur elle-même ou utiliser un marqueur implanté au site de la tumeur [7]. On parle dans ce cas de radiothérapie guidée par l'imagerie (RTGI).

Parmi les modalités d'imagerie utilisée pour la RTGI, on retrouve l'imagerie par rayons X en 2D. Il est possible d'utiliser des structures comme les vertèbres pour estimer le mouvement du foie dans les directions AP et LR. L'imagerie par rayons X peut être utilisée autant en respiration libre qu'en «breath hold». L'imagerie en 3D comme le CBCT permettent d'imager l'anatomie traitée en 3D, par contre ces acquisitions sont plus longues et le mouvement respiratoire peut introduire des artefacts qui dégradent la qualité de l'image. Il est donc recommandé de faire des acquisitions CBCT avec «breath hold» lorsque l'amplitude de respiration du patient dépasse 5 mm [24]. Étant des modalités d'imagerie ionisantes, leur utilisation doit être limitée afin de respecter le dosage que le patient peut recevoir.

Des systèmes comme le CyberKnife de Accuray utilisent une stratégie où un modèle de correspondance entre un signal respiratoire externe et la position de la tumeur est préparé avant le traitement [25]. Ainsi, durant le traitement il est possible d'estimer la position de la tumeur. Toutefois, comme pour la technique de «respiratory gating», les signaux externes ne

Figure 2.5 Le système de traitement de radiothérapie CyberKnife de la compagnie Accuray. On aperçoit la tête du LINAC montée sur un bras robotique à 6 axes. Le système d'imagerie par rayons X se trouve au plafond avec le détecteur placé sur le plancher. Image tirée de Wikimedia Commons 2021

représentent pas toujours la position interne de la tumeur de manière fiable.

D'autres modalités d'imagerie comme l'IRM peuvent également être utilisées pour la RTGI. Avec le développement de systèmes combinés de LINAC et résonance magnétique, il sera possible d'acquérir des images IRM du patient en temps réel tout en administrant le traitement par RTE. Par contre, ces systèmes seront très coûteux et peu accessibles [26] . Des systèmes de RTGI utilisant l'imagerie par ultrasons existent également. Ces systèmes seront abordés à la section 2.3.2

2.3 Imagerie par ultrasons

Dans cette section, il s'agira de présenter la modalité d'imagerie utilisée dans le cadre de ce projet, l'imagerie par US. Pour commencer, les aspects théoriques et physiques de l'imagerie par US seront expliqués. Ensuite, des détails quant aux modes d'acquisitions seront fournis. Finalement, les applications cliniques de l'US telles que la radiothérapie seront présentées.

2.3.1 Principes physiques

L'imagerie par US est une modalité d'imagerie non invasive et non ionisante. Les appareils permettant l'acquisition d'images US sont portables et relativement peu coûteux rendant l'imagerie par US accessible. Le principe fondamental de l'imagerie par US est l'envoi d'ondes ultrasonores dans le tissu et la capture des ondes réfléchies par les différentes interfaces rencontrées, ce qui permet d'imager le corps humain. Les ondes utilisées sont des ondes ayant

des fréquences allant de 1-10 MHz. Les ondes acoustiques sont produites par un transducteur piézoélectrique qui convertit des impulsions électriques en vibrations mécaniques. Lorsqu'une onde pénètre dans le tissu humain, elle est transmise à une vitesse c qui dépend de la compressibilité κ et la densité ρ du tissu :

$$c = \sqrt{\frac{1}{\kappa\rho}} \tag{2.1}$$

Lors de la propagation de l'onde acoustique, l'interaction avec le tissu cause une atténuation de l'onde due à trois facteurs principaux. Premièrement, l'absorption de l'énergie mécanique de l'onde par le tissu que l'onde fait vibrer. Ensuite, des phénomènes de diffraction causés par des particules de taille proche de la longueur d'onde de l'onde acoustique. Finalement, le changement de mode de l'onde longitudinale en onde transverse. L'atténuation de l'onde acoustique est décrite par une relation exponentielle décroissante entre la distance parcourue par l'onde et son intensité :

$$I_{(z)} = I_{(z=0)} \exp^{-\mu z} \tag{2.2}$$

où $I_{(z=0)}$ désigne l'intensité initiale de l'onde et μ est un coefficient d'atténuation propre au tissu. Une autre propriété importante des tissus dans le cadre de l'imagerie par US est l'impédance acoustique Z. En faisant l'analogie avec les notions des circuits électriques, l'impédance acoustique d'un tissu représente sa résistance à la transmission de l'onde acoustique. Z est obtenue à partir des caractéristiques du tissu de la manière suivante :

$$Z = \sqrt{\frac{\rho}{\kappa}} = \rho c \tag{2.3}$$

Lorsqu'une onde d'US arrive à une interface entre deux tissus avec des impédances acoustiques différentes ($Z_1 \neq Z_2$), l'intensité de l'onde sera divisée en deux parties. Une partie de l'onde sera transmise et continuera de se propager dans le tissu, tandis que l'autre partie sera réfléchie et amorcera son retour vers le transducteur. L'angle de réflexion de l'onde réfléchie est égal à l'angle d'incidence ($\theta_i = \theta_r$) tandis que l'angle de l'onde transmise est régi par la loi de Snell :

$$\frac{sin(\theta_i)}{sin(\theta_t)} = \frac{c_1}{c_2} \tag{2.4}$$

où c_1 et c_2 sont les vitesses de propagation de l'onde sonore dans les tissus composant l'interface. La figure 2.6 montre un schéma de la réflexion et la transmission de l'onde acoustique entre deux tissus ayant des impédances acoustiques différentes. En plus de changer de direction, lors du changement d'interface, l'intensité de l'onde incidente est divisée entre l'onde réfléchie et l'onde transmise. On définit le coefficient de l'intensité de réflexion R_I et de trans-

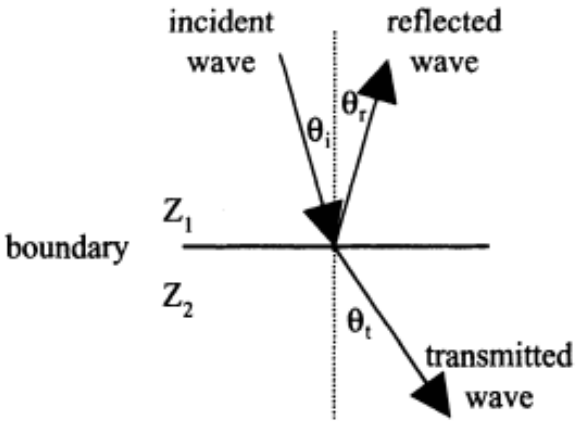


Figure 2.6 Représentation schématique de la réflexion et la transmission d'une onde acoustique à l'interface entre deux tissus ayant des impédances acoustiques différentes. Image tirée de [27]

mission T_I selon l'intensité de l'onde incidente (I_i), l'onde réfléchie (I_r) et l'onde transmise (I_t).

$$R_I = \frac{I_r}{I_i} = \frac{(Z_2 cos(\theta_i) - Z_1 cos(\theta_t))^2}{(Z_2 cos(\theta_i) + Z_1 cos(\theta_t))^2} \tag{2.5}$$

$$T_I = \frac{I_t}{I_i} = \frac{4 Z_2 Z_1 cos^2(\theta_i)}{(Z_2 cos(\theta_i) + Z_1 cos(\theta_t))^2} \tag{2.6}$$

Grâce au phénomène de réflexion, les ondes initialement envoyées par le transducteur produisent plusieurs ondes réfléchies à chacune des interfaces rencontrées. Ces ondes réfléchies retournent vers le transducteur où l'énergie mécanique des ondes fait vibrer l'élément piézoélectrique qui émet un signal électrique. Une fois amplifié et filtré, le signal électrique est numérisé. En connaissant la vitesse de propagation de l'onde dans le tissu ainsi que le temps écoulé entre l'émission et la réception de l'onde, il est possible de reconstruire la position des interfaces tissulaires dans le corps imagé [27].

En imagerie médicale, les artefacts se manifestent par l'apparition d'une structure sans que cette structure existe réellement dans le tissu imagé. Dans le cas de l'US, il existe plusieurs mécanismes physiques qui mènent à l'apparition d'artefacts. Notamment, la réverbération acoustique survient lorsque l'onde acoustique rencontre deux interfaces parallèles avec un coefficient de réflexion élevé. L'onde est alors réfléchie plusieurs fois par les deux interfaces, causant des retours multiples au transducteur qui les interprète comme étant des interfaces se trouvant à différentes profondeurs dans le corps. Un autre artefact qui survient lors de la présence d'interfaces très réfléchissantes est l'ombrage acoustique. Puisque très peu d'ondes se rendent plus loin que cette interface, le transducteur n'acquiert aucune information quant

à ce qui se trouve plus loin que celle-ci. Par conséquent, cette section de l'image apparaît noire. Ceci survient fréquemment aux interfaces gaz/tissu et lors de la présence d'os. Le phénomène inverse survient lors de la présence d'une région de basse atténuation dans un milieu homogène. Cette partie de l'image apparaît comme ayant une haute intensité (hyper-échogène). La réfraction subie par l'onde transmise peut entraîner une déviation importante de l'angle de propagation de l'onde acoustique. En conséquence, lors du retour de l'onde vers le transducteur, l'interface sera reconstruite au mauvais endroit, ce qui peut causer une mauvaise interprétation de l'image [27]. Finalement, une particularité des images d'US est l'apparition de structures granuleuses appelées «speckle» [28]. Le «speckle» est causé par l'interférence d'ondes acoustiques entre deux diffuseurs d'ultrasons qui sont trop près pour être résolus spatialement [29]. Plusieurs approches d'acquisition et de filtrage d'images ont été proposées pour réduire le «speckle» [30]. Il est possible de suivre des tissus de manière précise en suivant les motifs de «speckle» uniques à ces tissus. Par contre, lorsque le tissu subit des mouvements et des rotations importantes, cette approche n'est plus valable [29].

Il existe trois principaux modes d'acquisition en imagerie par US. Le choix du mode à utiliser dépend de l'application clinique souhaitée. Le mode amplitude (A) est le mode d'acquisition le plus simple. Dans ce mode, on enregistre l'intensité réfléchie d'une ou plusieurs ondes émises par un seul transducteur. On obtient donc un signal en une dimension qui désigne les interfaces tissulaires se trouvant dans le chemin de l'onde émise. Le mode mouvement (M) effectue plusieurs acquisitions de mode A dans le temps. Ce type d'acquisition permet de suivre la position de structures dans le temps. Finalement, le mode brillance (B) va plutôt combiner les acquisitions de type A de manière spatiale. C'est-à-dire qu'un arrangement séquentiel de plusieurs transducteurs émet des ondes acoustiques de manière à préserver une cohérence spatiale entre les fronts d'onde. Suite à la réception des ondes réfléchies par tous les transducteurs, une image en 2D en tons de gris peut être reconstruite. Les interfaces plus réfléchissantes apparaissent en blanc tandis que les interfaces peu réfléchissantes restent sombres [27, 31].

Dû au manque d'information quant à l'orientation des images d'US en mode B, les cliniciens doivent user de leur expérience pour imaginer la composition en 3D des structures imagées [32]. Afin de fournir des volumes d'US 3D, beaucoup d'efforts ont été investis dans le développement de sondes 3D. Initialement, il était possible de reconstruire des volumes 3D en acquérant plusieurs images parallèles en mode B, de manière manuelle ou automatique, et en les combinant par la suite. Aujourd'hui, avec des sondes matricielles comme la X6-1 de Phillips, il est possible d'acquérir des volumes 3D et même des séquences de volumes en 4D en temps réel sans recourir à un balayage. Les sondes matricielles comportent plusieurs éléments piézoélectriques disposés dans une matrice 2D. Avec la miniaturisation électronique, il

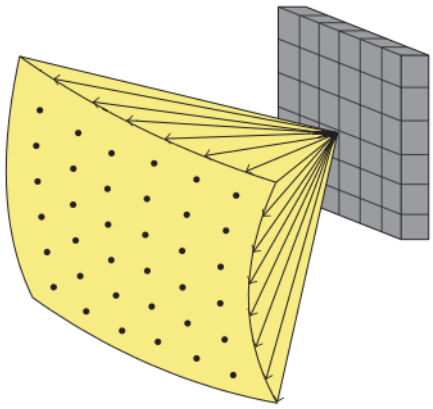


Figure 2.7 Forme d'une acquisition d'US en 3D. Image tirée de [32]

est possible de contrôler chaque transducteur individuellement. Ce contrôle individuel permet de changer le point de convergence des fronts d'onde en azimut et en élévation pour venir imager différents chemins de propagation [32]. Le volume résultant possède la forme d'une pyramide rectangulaire telle que montrée à la figure 2.7.

2.3.2 Application clinique

Le champ d'application clinique de l'imagerie par US est très large grâce au faible coût et la nature non ionisante de cette modalité. L'US est utilisé en gynécologie et obstétrique dans le but de suivre le développement du foetus au cours de la grossesse. On l'utilise aussi en cardiologie pour inspecter le fonctionnement des valves cardiaques et pour l'évaluation du flux sanguin à l'aide de l'effet Doppler. On retrouve des systèmes d'imagerie par US dans des salles de chirurgie afin de localiser des outils au cours d'une intervention. On en retrouve même dans les ambulances afin de pouvoir évaluer rapidement l'état d'un patient. Par contre, l'application clinique de l'US sur laquelle nous nous pencherons dans cette sous-section est celle de la RTE.

Comme décrit à la section 2.2.4, plusieurs modalités d'imagerie sont utilisées pour toutes les étapes du traitement de radiothérapie. Par exemple le CT, l'IRM et le PET pour la planification ou bien l'imagerie par rayons X avec marqueurs et le CBCT pour le positionnement du patient et le suivi de tumeurs. L'imagerie par US présente plusieurs avantages par rapport à ces modalités d'imagerie. Ceci a poussé l'inclusion de cette modalité dans le processus de la RTE. Les avantages principaux de l'US sont son temps d'acquisition très court et sa nature non ionisante. En effet, les images acquises peuvent être reconstruites et visualisées en temps réel. Puisque les ondes ultrasonores ne peuvent pas se propager à travers des interfaces de

haute impédance acoustique telles que les os, l'US est principalement utilisé pour la visualisation de tissus mous dans des organes comme la prostate, le pancréas, les reins et le foie. L'équipement nécessaire pour l'imagerie par US est relativement peu coûteux et portable ce qui rend cette modalité très accessible aux établissements médicaux.

Lors de l'étape de planification, l'utilisation principale de l'US est l'identification des contours de la tumeur. L'US est souvent utilisé pour la planification de traitements de la prostate puisque celle-ci est bien visible lorsque la vessie est remplie. Il est également possible d'utiliser l'imagerie par US en tant que complément à d'autres modalités afin de mieux définir les contours de la tumeur. Enfin, avec la disponibilité accrue aux systèmes d'acquisition d'US 4D, cette modalité est de plus en plus utilisée pour quantifier l'amplitude de mouvement que subit la tumeur ce qui améliore la définition de volumes d'intérêt comme le ITV [33].

L'US peut aussi être utilisé lors du positionnement du patient le jour du traitement afin de pallier les changements anatomiques entre les fractions. Des contours obtenus le jour du traitement par imagerie US peuvent être comparés à des contours obtenus lors des acquisitions de prétraitement soit par une autre modalité comme le CT (système intermodalité) ou l'US (système intramodalité). Dans le cas des systèmes intermodalité, l'établissement de correspondances entre l'US et la modalité de planification peut s'avérer complexe dû à la différence entre les méthodes d'acquisition. Du côté des systèmes intramodalité, les étapes de mise en correspondance peuvent également être coûteuses en temps sans formation appropriée [33].

Une autre étape importante où l'US peut contribuer à améliorer la livraison du traitement de RTE est lors de l'administration du traitement. Il est important de tenir compte des mouvements de la tumeur et des organes environnants durant la livraison de l'irradiation au site tumoral. Des études ont démontré de manière expérimentale que l'utilisation de l'imagerie par US pour des traitements de RTGI permet de compenser pour le mouvement de la cible à traiter et de ce fait, réduire l'étendue du PTV [34,35]. L'intégration de l'US dans cette étape du traitement est encore peu répandue pour des raisons telles que l'interférence de la sonde US avec la distribution de la dose ou bien la fréquence temporelle insuffisante de certains systèmes US 3D. Le premier système de suivi de tumeur basé sur l'US est Clarity Autoscan de Elekta [29]. Ce système intègre un transducteur qui est capable d'acquérir des volumes d'US 3D à l'aide d'un mécanisme de balayage. Le système a été optimisé pour les traitements de la prostate et est compatible lors des étapes de prétraitement et de traitement. Avant le début de l'intervention, un volume de référence est acquis par la sonde, puis, durant le traitement les volumes sont acquis avec une fréquence temporelle de 0.4 Hz et recalés au volume de référence par corrélation [36]. Des études ont rapporté des erreurs de moins de 1.2 mm pour le système Clarity. Toutefois, le mouvement subi par la prostate reste relativement

petit comparé aux organes abdominaux. De plus, la très basse fréquence d'acquisition limite l'utilisation de ce système pour d'autres applications qui ont besoin de capacités d'analyse en temps réel [29].

Plusieurs avenues de recherche et de développement pourraient rendre l'imagerie par US plus intégrée dans le processus de la RTE. D'abord, il est nécessaire de développer des techniques d'estimation du mouvement qui sont effectuées en temps réel. Malgré que l'imagerie US 3D peut être effectuée avec une résolution temporelle suffisante en utilisant des sondes matricielles, le traitement de ces images par la suite peut rendre le processus trop lent. Une autre avenue de recherche importante traite de l'élargissement du nombre de sites anatomiques où l'US peut être utilisé. Pour l'instant, l'estimation du mouvement est limitée à la prostate, mais d'autres sites comme le foie, les reins et le pancréas peuvent également bénéficier de l'imagerie par US pour cette tâche [29].

2.4 Méthodes d'estimation du mouvement en 3D à partir d'images 2D

L'estimation de l'apparence en 3D d'un objet à partir d'une représentation 2D de celui-ci est un objectif de longue date au sein de la communauté de vision par ordinateur [37]. Plusieurs approches ont été proposées afin de permettre aux ordinateurs de comprendre la structure 3D de divers objets comme des voitures, des meubles ou même le visage humains [38–40].

Dans le cas de la RTGI, les étapes de la simulation et de planifications se servent de volumes en 3D, mais lorsque vient le temps de traiter le patient, le personnel de santé est limité à utiliser de l'imagerie en 2D pour des applications nécessitant des performances en temps réel comme le suivi de la tumeur à traiter. Par contre, l'imagerie en 2D ne permet pas de suivre des cibles qui ont des trajectoires en 3D comme les tumeurs hépatiques. Dans le cas où l'imagerie 3D soit utilisée, sa fréquence d'acquisition est trop lente et ne permet pas de suivi en temps réel de la tumeur. Il serait donc utile de pouvoir estimer la position de la tumeur et des organes à risque en 3D lors du traitement en utilisant seulement des données en 2D qui sont accessibles cliniquement.

Avec l'accessibilité accrue aux données d'imagerie médicale en 4D pour la plupart des modalités utilisées en RTE, plusieurs nouvelles approches pour l'estimation du mouvement de cibles en 3D ont vu le jour. De plus, avec l'essor de l'apprentissage profond appliqué à la vision par ordinateur, de nouvelles notions d'analyse d'images médicales ont pu être intégrées dans le cadre d'interventions médicales comme la RTE.

Dans cette section, nous présenterons d'abord les deux types d'approches existantes de modélisation du mouvement respiratoire. Les modèles de suivi locaux et globaux. Ensuite, les

notions de base de l'apprentissage profond seront expliquées. Finalement, nous présenterons les plus récents travaux utilisant ces notions pour l'estimation du mouvement en 3D dans l'imagerie médicale pour la RTGI.

2.4.1 Modèles de suivi locaux

Les modèles de suivi locaux ont comme but de suivre une cible anatomique choisie à travers le temps. L'accomplissement de cette tâche pour les séquences d'US en 2D et 3D a généré un intérêt considérable. Ceci a mené à la création de compétitions ouvertes comme CLUST15 [41]. En ayant accès à des ensembles de données d'US 2D et 3D communs, il était possible de directement comparer les différentes solutions proposées. Les approches principales utilisées pour le suivi local sont entre autres les filtres de particules [42], le flux d'optique [43], le recalage d'images [44, 45] ou le «template matching» [46]. Dans [47] une approche de suivi composée de plusieurs étapes de correspondance par blocs est présentée pour les séquences d'US 2D. Chacune des étapes était responsable de produire un niveau de mouvement de plus en plus raffiné en comparant une image de référence et l'image courante. Dans [48] une technique de suivi basée sur les caractéristiques de supports appelées, «supporters» est proposée. Les «supporters» sont des traits de l'image qui se trouvent à proximité de la cible suivie. En suivant les «supporters», il était possible de suivre la cible avec plus de précision. Cette approche améliorait également la performance lorsque la visibilité de la cible n'était pas optimale. Ces deux approches ont atteint des erreurs sous-millimétriques, mais elles ont été testées seulement sur les séquences d'US 2D.

Quant aux méthodes testées sur des séquences d'US 3D, dans [49] une approche par deux étapes de recalage rigide a été employée. Premièrement, un ensemble de points couvrant l'entièreté du volume sont recalés par correspondance de bloc. Ensuite, un ensemble de points à proximité de la cible sont utilisés pour effectuer un deuxième recalage qui retourne la nouvelle position de la cible. Dans ţteRoyer2017, la cible de suivi est représentée par un modèle de cellules tétraédriques. Les mouvements externes et internes de chaque cellule ont été estimés par un modèle mécanique suivi d'une approche basée sur les intensités de voxels.

En général, les approches de suivi locales ont des temps de calcul très courts et fournissent une bonne précision. Par contre, ces approches partagent le désavantage de fournir seulement la nouvelle position de la cible. Aucune information sur le mouvement des tissus environnants n'est fournie, ce qui est de l'information utile dans le cas de la RTGI à des fins d'estimation de dose par exemple.

2.4.2 Modèles de suivi globaux

Afin de pallier les désavantages des approches locales, il est possible d'utiliser des modèles de suivi globaux. Les approches de suivi globales permettent de déterminer la nouvelle position de la structure anatomique suivie en générant les déplacements subis par les tissus environnants et l'organe dans son entièreté. La sortie attendue dans ce cas est un champ de mouvement qui couvre tout le champ de vue de l'image. Ce champ de mouvement peut être utilisé non seulement pour suivre la cible dans l'espace, mais également pour l'ajustement des plans de traitement, ainsi que le calcul de la dose d'irradiation dans le contexte d'un traitement par RTE. L'obtention de champs de mouvements complexes en 3D en utilisant des données de dimension inférieure comme signal substitut a été communément accomplie dans le contexte de la RTGI par des approches de modélisation du mouvement [50]. Comme montré à la figure 2.8, un modèle de mouvement est formé en faisant l'acquisition simultanée de données d'imagerie et d'un signal substitut dans le temps. Suite au calcul des champs de mouvement par recalage des données 4D, le signal substitut est lié aux champs de mouvement correspondants. Une fois formé, le modèle permet d'obtenir des champs de mouvement à partir d'une nouvelle acquisition du signal substitut, lors d'un traitement par exemple. Les signaux substitut peuvent être des données en 1D comme la spirométrie [51] ou le suivi de la surface de la peau [52]. Les images en 2D acquises durant les traitements par RTE peuvent également être utilisées comme signal substitut pour inférer les champs de mouvement en 3D de l'organe traité [50]. L'application des modèles de mouvement dans le cadre de traitements de RTE a été étudiée pour une variété de modalités d'imagerie. Toutefois, très peu de travaux se sont concentrés sur l'utilisation de l'imagerie par US dus aux difficultés inhérentes présentées par cette modalité telles que la présence d'artefacts uniques [53]. Néanmoins, les approches de modélisation de mouvement restent flexibles par rapport aux choix de la modalité d'imagerie. Dans certaines applications, la position de marqueurs imagés par US 2D a été utilisée comme signal substitut pour générer le mouvement dans des volumes d'IRM [54].

Dans la littérature, on distingue deux types de modèles de mouvement, les modèles à patient unique et les modèles de population. Le premier type est spécifique à un seul patient tandis que le deuxième permet de modéliser la respiration d'une population de patients. Les techniques principales utilisées pour construire ces modèles sont les modèles biomécaniques [55,56] et les modèles statistiques [57, 58]. Toutefois, les modèles statistiques sont utilisés plus fréquemment. Les paragraphes qui suivent présentent ces deux types de modèles de mouvement plus en détail.

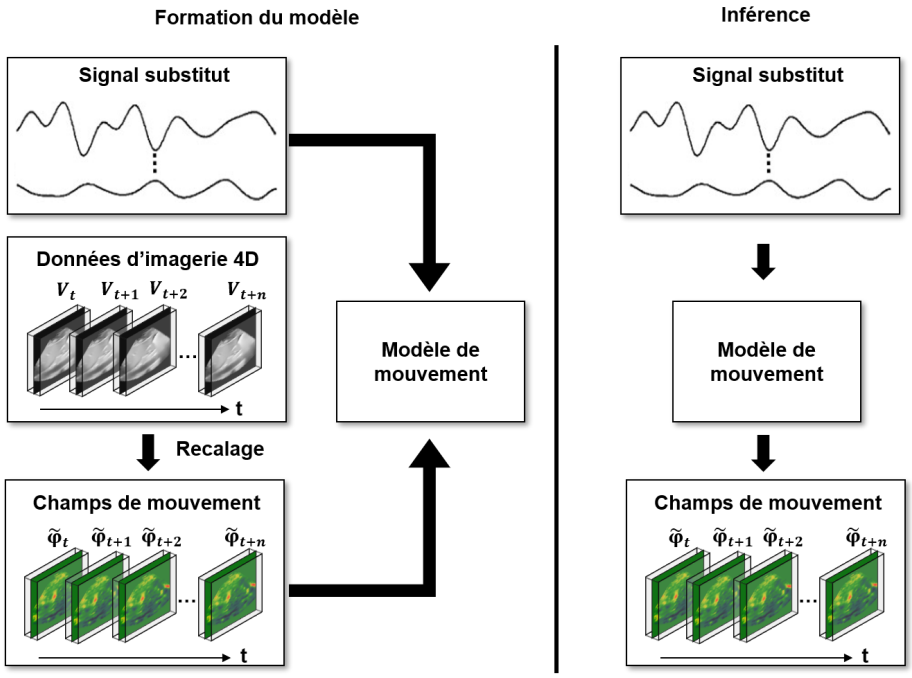


Figure 2.8 Diagramme de la formation et l'inférence d'un modèle de mouvement. Lors de la formation du modèle, un signal substitut est lié à des champs de mouvement correspondants. Lors de l'inférence, le modèle permet d'obtenir des champs de mouvement à partir d'un nouveau signal substitut. Inspiré de [50]

Modèles à patient unique

Une première famille de modèles de mouvement globaux est celle des modèles qui sont construits uniquement à partir de données d'un patient spécifique. Leur construction se fait en acquérant des données 4D du patient ainsi que des signaux de substitution avant le traitement. Ensuite, les champs de mouvement 3D peuvent être obtenus en recalant les volumes de la séquence 4D à un volume de référence choisi à une certaine phase respiratoire. Finalement, le signal substitut est lié aux champs de mouvement calculés. Dans [59], le mouvement du foie est modélisé en acquérant des volumes IRM en expiration et inhalation complète ainsi que 8 autres phases respiratoires intermédiaires. Les champs de mouvement ont été décomposés en un mouvement global et local calculé par recalage rigide et déformable respectivement. Toutefois, cette approche ne propose pas de moyen pour lier un signal substitut aux champs de déformations afin de pouvoir les récupérer. Plusieurs approches visant à créer une correspondance entre les signaux substitut et les champs de mouvement ont été proposées. Dans [60], un atlas de champs de mouvement est créé à partir de données d'IRM 4D qui ont été récupérées à partir d'un signal respiratoire acquis durant le traitement. Dans [61], des images parallèles d'IRM 2D temporelles acquises à 6 positions à travers le foie ont été recalées avec

les tranches correspondantes d'un volume de référence afin d'extrapoler les champs de mouvement complets à travers le temps. Ainsi, il était possible de créer une table de recherche liant les images d'IRM 2D avec les champs de mouvement extrapolés. Dans [62], une tentative d'unification des étapes de calculs des champs de mouvement et de mise en correspondance avec les signaux substituts est présentée. Ceci a été accompli en développant un modèle de recalage basé sur un modèle linéaire. Les paramètres du modèle combinent le signal substitut de manière à obtenir un champ de déformation. Cette procédure d'optimisation permet de lier les signaux substitut au champ de déformation optimisés d'un seul coup. Leur approche généralisée a présenté plusieurs avantages, par contre, des temps de calcul élevés ont limité son utilisation dans des applications nécessitant des performances en temps réel.

Parmi les travaux sur les modèles de mouvement à patient unique, l'analyse par composantes principales (PCA) [63] est une approche fréquemment utilisée. L'approche par PCA permet d'effectuer une décomposition linéaire des champs de mouvement du patient. Suite à la décomposition linéaire, les champs de mouvements D peuvent être récupérés avec une combinaison linéaire des vecteurs propres D_i des N premières composantes principales obtenues :

$$D = D_{moy} + \sum_{i=1}^{N} w_i D_i \tag{2.7}$$

C'est donc en variant les coefficients de combinaison w_i qu'il est possible d'obtenir des champs de déformation correspondants à différentes phases respiratoires. Le lien entre le signal substitut et les champs des mouvements se fait par ces coefficients. Dans [57] cette approche a été implémentée afin de reconstruire des champs de déformation en 3D à partir de projections 2D d'une acquisition CBCT d'un fantôme. Dans [64], une étude a été menée sur 8 patients ayant un cancer des poumons et qui avait des données d'imagerie CT en 4D. Ils ont conclu que l'utilisation de deux composantes principales permet de représenter une variabilité suffisante pour modéliser le mouvement de la plupart des patients étudiés. Les chercheurs ont utilisé la position de marqueurs artificiels implantés afin de dériver les valeurs des coefficients de la PCA. Il est également possible d'utiliser des images navigatrices en 2D afin de trouver les coefficients de combinaison optimaux. Dans [58], des images d'IRM 2D sont utilisées afin d'effectuer l'inférence de leur modèle de mouvement. Ils ont également implémenté un mécanisme qui suggère la position optimale où acquérir le navigateur. Enfin, ils ont proposé un algorithme d'évaluation du modèle de mouvement par le navigateur 2D afin de savoir si l'acquisition de plus de données volumétriques est nécessaire. D'autres manières de récupérer les coefficients linéaires de la PCA incluent la maximisation de la similarité entre l'image substitut et l'image du volume de référence déformé par le champ de mouvement généré [65–67] ou bien la correspondance par blocs [68].

Les modèles à patient uniques sont reconnus pour avoir une plus haute précision en termes de localisation de cibles comparée aux modèles de population. Ceci est dû au fait que le modèle est optimisé pour chaque patient individuellement. Toutefois, le désavantage principal de ces modèles est la nécessité d'acquérir des données 4D du patient afin de pouvoir modéliser ses patrons de mouvements, ce qui n'est pas possible dans toutes les institutions hospitalières. De plus, dans le cas où l'anatomie du patient change au cours d'un traitement avec plusieurs fractions de dose, le modèle peut devenir erroné et nécessiter une mise à jour, ce qui peut devenir coûteux en temps.

Modèles de population

Le deuxième groupe de modèles de mouvement globaux, appelés modèles de population, vise à tenir compte d'une plus grande variété de champs de mouvement en traitant un ensemble de données issu d'une population de patient. Ces modèles sont souvent considérés comme des modèles représentant seulement la moyenne du mouvement respiratoire de l'ensemble de données. Toutefois, les modèles de population ont le potentiel d'être appliqués à de nouveaux individus sans requérir à l'acquisition de nouvelles données [53]. Dans [69], le concept de modèles exemplaires est introduit. Premièrement, un modèle à patient unique est conçu pour chaque patient de l'ensemble de données. Ensuite, le signal substitut d'un nouveau patient est comparé aux modèles exemplaires afin de déterminer la combinaison linéaire optimale de ces derniers. Une autre approche qui combine plusieurs modèles à patient uniques a été proposé dans [70]. Suite à la construction des modèles à patient uniques, un atlas de la forme et l'intensité moyenne des poumons est créé comme référence. Un modèle de population moyen est finalement obtenu par le recalage des modèles à patient uniques à l'atlas de référence. Dans [71], un modèle de mouvement statistique est utilisé afin de générer le mouvement des poumons en apprenant une régression linéaire multivariée entre les éléments à prédire et l'ensemble de données d'entraînement CT en 4D de 10 patients. Dans [72], une approche de reconstruction bayésienne a été utilisée pour prédire le mouvement du foie à partir de données partielles. Le modèle statistique construit à partir de données CT de 12 patients utilise un volume CT préopératoire et la position de marqueurs internes pour reconstruire les champs de déformation. Dans [73], un modèle de mouvement global est proposé. Le modèle peut inférer directement le champ de mouvement 3D complet en extrapolant le recalage de deux images d'IRM perpendiculaires avec les tranches correspondantes dans un volume de référence. Comme dans le cas des modèles à patient unique, la PCA est fréquemment utilisée pour la construction de modèles de population [74–77]. Dans [54], la conception d'un modèle de mouvement qui combine l'information tirée d'images d'US 2D avec la PCA est présentée. Cette approche permet de prédire le mouvement tridimensionnel du foie acquis à

l'aide d'imagerie IRM à l'aide d'un signal substitut provenant d'images US.

Les modèles de population sont considérés comme étant plus flexibles que les modèles à patient unique puisqu'ils peuvent être appliqués à des données de patients qui n'ont pas été utilisées lors de la construction du modèle. Ils permettent également d'améliorer leur performance avec l'ajout de plus de données de patients au fil du temps. Toutefois, l'utilisation de modèles statistiques comme la PCA pour les modèles de population présente un désavantage considérable. En effet, il est nécessaire d'établir des correspondances entre les patients afin de normaliser les composantes principales extraites des champs de mouvement. Ceci est un processus qui s'avère souvent inexact et très coûteux en temps puisqu'il s'allonge avec le nombre de patients dans l'ensemble de données.

2.4.3 Méthodes d'apprentissage profond

Perceptron multi-couches

Le perceptron multi-couches [78] représente la forme de base d'un réseau de neurones profond. L'unité fondamentale de ce réseau est le perceptron. Cette unité est aussi appelée neurone puisque le fonctionnement de celle-ci a été inspiré par la manière dont les neurones du cerveau communiquent. Chaque neurone possède un poids, un biais ainsi qu'une fonction d'activation. Ces paramètres régissent le comportement du neurone en décidant si ce dernier émet, ou non, un signal selon l'entrée reçue. Lorsque les neurones sont organisés dans un réseau de couches comme dans la figure 2.9, il s'agit d'un perceptron multi-couches. Au minimum, ce réseau comporte trois couches, une couche d'entrée, une couche cachée ainsi qu'une couche de sortie. À chaque couche, l'entrée x subit une opération matricielle affine avant de passer dans la fonction d'activation non linéaire f. Ainsi, l'opération à chaque couche peut être représentée par $f(W^T x + b)$ où W est la matrice des poids et b est le vecteur de biais des neurones composant la couche.

Le problème d'apprentissage des paramètres du réseau pour une tâche donnée est souvent posé comme un problème d'optimisation où une fonction de perte $\mathcal{L}$ adaptée au problème à accomplir est minimisée. Dû au grand nombre de paramètres et à la complexité des réseaux, une approche d'optimisation par descente de gradient stochastique est utilisée. À chacune des itérations d'entraînement, une partie de l'ensemble de données est utilisé afin d'estimer le gradient de la fonction de perte selon les poids du réseau. Les gradients sont estimés par rétropropagation de l'erreur de prédiction commise par le réseau à l'itération courante. En ayant les gradients, il est possible de mettre à jour les paramètres du réseau de la manière

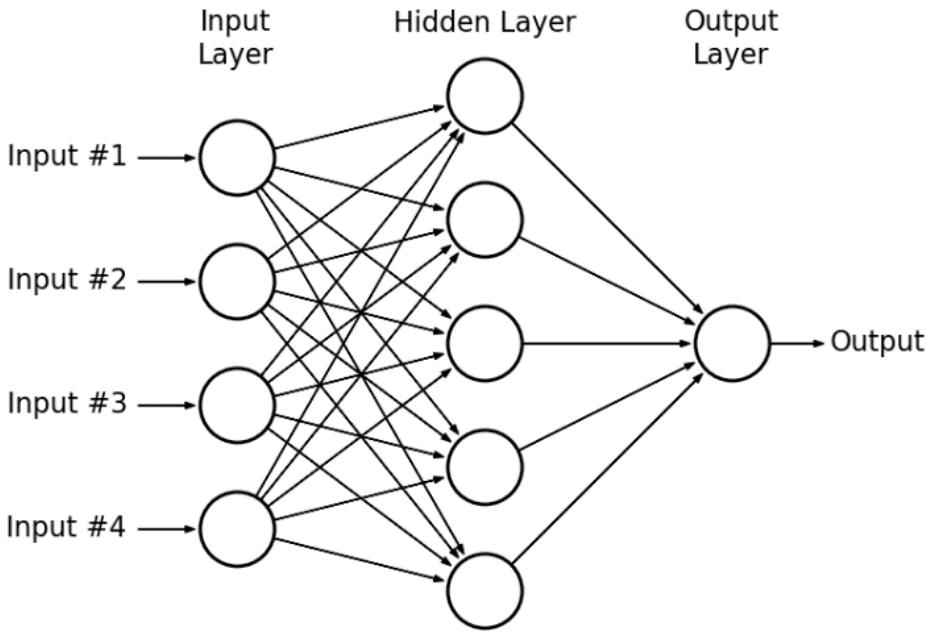


Figure 2.9 Représentation schématique d'un perceptron multi-couches. On y identifie les 3 couches minimales soit la couche d'entrée, couche cachée et couche de sortie. Image tirée de [79]

suivante :

$$w_{t+1} = w_t + \gamma(-\nabla\mathcal{L}(w_t)) \tag{2.8}$$

où w_t est la valeur d'un poids du réseau, γ est le taux d'apprentissage et $\nabla\mathcal{L}(w_t)$ est le gradient de la fonction de perte selon les poids du réseau.

Réseaux de neurones convolutifs

Les réseaux de neurones convolutifs (CNN) [80] reprennent l'organisation par couches introduite par les perceptrons multi-couches. Toutefois, une différence fondamentale se trouve dans la manière dont l'information est traitée à l'intérieur de chaque couche. Ceci rend les CNN particulièrement utiles pour les applications de vision par ordinateur. Comme le nom l'indique, les CNN font appel à l'opération de convolution pour remplacer l'opération matricielle utilisée par les neurones. Chaque couche est composée d'un ensemble de filtres possédant des paramètres uniques et ajustables. Lors de la propagation avant, chaque filtre est appliqué à l'entrée de la manière suivante $f(W * x + b)$ où $*$ dénote l'opération de la convolution de l'image d'entrée par les paramètres W du filtre. Le but ultime de ce type de réseau est d'optimiser de manière automatique les poids de tous les filtres afin d'acquérir une compréhension

générale des éléments qui permettent de reconnaître un objet dans une image, peu importe l'éclairage, l'orientation ou la taille de celui-ci [81]. Dans le cas des réseaux de perceptrons, les entrées sont toutes traitées comme vecteurs unidimensionnels. En utilisant des convolutions, il est possible de maintenir les relations spatiales présentes dans les images d'entrée et réduire le nombre de paramètres à optimiser. Pour cette raison, les CNN sont considérés plus efficaces que les réseaux de perceptrons [82]. Suite à plusieurs succès dans des compétitions de vision par ordinateur telles que ImageNet [83], la popularité des CNN n'a cessé d'augmenter dans plusieurs domaines d'application tels que l'analyse d'images médicales [84].

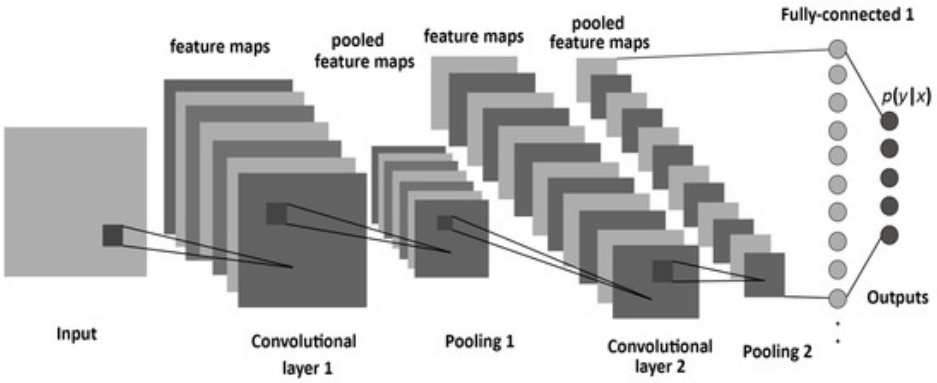


Figure 2.10 Représentation schématique d'un réseau de neurones convolutif. Image tirée de [85]

Les autoencodeurs

Le terme autoencodeur (AE) désigne un réseau de neurones qui a pour but de reproduire son entrée à sa sortie. Les opérations principales d'un AE sont l'encodage de l'entrée x vers un code z ($z = f(x)$) et le décodage du code z vers la sortie reconstruite r ($r = g(z)$). De prime abord, un modèle qui apprend simplement à reconstruire son entrée semble peu utile, mais l'utilité des AE réside dans le code z. En effet, le but de l'AE est de concevoir un code z qui représente les attributs qui sont utiles à la reconstruction de l'ensemble des objets sur lesquels l'AE est entraîné. L'objectif d'entraînement d'un AE peut être résumé comme suit :

$$arg\min_{\theta,\omega} \ \mathcal{L}(x,\ g_\omega(f_\theta(x)) \tag{2.9}$$

où θ et ω sont les paramètres de l'encodeur et du décodeur respectivement et $\mathcal{L}$ est une mesure de la similarité entre l'entrée et la sortie reconstruite par l'AE. De cette manière, suite à l'entraînement, l'AE permet de décrire un ensemble de données selon une représentation compacte entreposée dans le code z [82]. La représentation cachée z obtenue suite à

l'entraînement d'un AE est analogue à la décomposition obtenue suite à l'application de la PCA à un ensemble de données. Par contre, la capacité de représentation de l'AE surpasse celle de la PCA en utilisant des fonctions d'activation non linéaires. Les AE sont donc vus comme une alternative non linéaire à la PCA [86]. Un des désavantages des AE par rapport à la PCA est que la représentation créée par l'AE ne garde pas une relation linéairement indépendante entre les caractéristiques entreposées dans le code z. Ainsi, cette représentation est plus difficilement interprétable que les composantes principales obtenues par PCA.

Lors de la conception d'un AE, le choix de la taille de la représentation z est crucial. En effet, si la taille du code z est similaire ou même plus grande que la taille des objets donnés en entrée, le réseau tendra à apprendre tout simplement une fonction d'identité, sans extraire d'attributs qui décrivent l'ensemble d'entraînement. Dans ce scénario, l'objectif décrit à l'équation 2.9 sera atteint, mais le code z sera inutile. Dans le cas contraire, si une taille trop petite est choisie pour le vecteur z, il sera impossible pour l'AE de représenter toute la variance de l'ensemble d'entraînement et donc l'objectif de reconstruction ne sera pas atteint. Il faut donc laisser une capacité de représentation suffisante à l'AE tout en le forçant à générer une représentation utile [87].

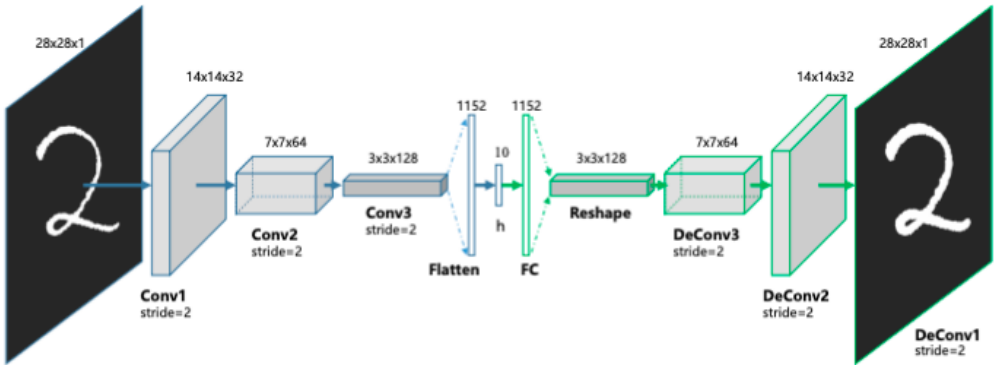


Figure 2.11 Représentation schématique d'un autoencodeur convolutif à trois couches. Image tirée de [88]

Plusieurs approches de régularisation ont été proposées afin d'assurer l'obtention d'un code z non trivial. Les AE contractifs (CAE) assurent que la taille du code z au centre de l'AE est suffisamment inférieure à la taille de l'entrée. Les AE de débruitage (DAE) changent l'objectif initial de l'entraînement en ajoutant du bruit aléatoire aux données d'entrée. Ceci empêche l'AE de simplement apprendre une fonction d'identité puisque la sortie doit être débruitée. Cette approche à l'avantage de forcer l'AE de se concentrer sur les attributs globaux des objets d'entraînement plutôt qu'aux détails ou bruits présents dans l'entrée [87]. D'autres variantes comme les AE variationnels (VAE) [89], imposent au code z de prendre la forme d'une

distribution gaussienne avec une moyenne de 0 et variance de 1. Cette restriction a permis d'utiliser le VAE comme modèles génératifs, c'est-à-dire utiliser le code z non seulement pour décrire les attributs de l'ensemble d'entraînement, mais pour produire de nouveaux objets qui n'existaient pas lors de l'entraînement. Les AE variationnels conditionnels (CVAE) ajoutent la possibilité de conditionner le processus de génération du décodeur en donnant une condition quant à la classe de l'objet à générer.

À l'origine, les AE sont conçus avec des couches de perceptrons, mais dans les applications de vision par ordinateur les AE convolutifs sont devenus plus utilisés pour des raisons similaires à celles mentionnées plus tôt [90]. Un AE convolutif peut être schématisé comme à la figure 2.11. On remarque que l'encodeur est composé de plusieurs couches convolutives, qui réduisent graduellement la taille de l'image d'entrée, et d'une couche de perceptrons qui produisent la représentation cachée z. Le décodeur possède la plupart du temps une structure symétrique à celle de l'encodeur.

Module de transformation spatiale

Les modules de transformation spatiale (STN) ont été initialement proposés dans [91], afin d'augmenter la robustesse des CNN envers des variations spatiales, telles que la translation et la rotation, dans leurs images d'entrée. L'inclusion de ce module permet aux CNN d'appliquer des transformations aux représentations intermédiaires générées dans chaque couche. Le STN prend en entrée un volume ou une image U puis détermine les paramètres de transformation θ qui permettent de normaliser son apparence par rapport à l'ensemble de données. Le module va ensuite rééchantillonner l'image d'entrée selon une grille transformée $\mathcal{T}_\theta(G)$ par les paramètres trouvés. Ainsi, on obtient l'image transformée V à la sortie. La figure 2.12 montre le module STN et ses composantes. Une propriété importante de ce module est l'uti-

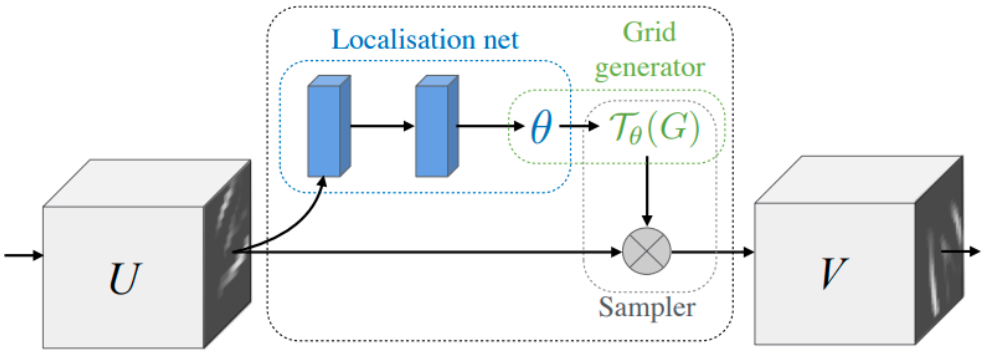


Figure 2.12 Représentation schématique d'un module de transformation spatiale. Image tirée de [91]

lisation d'opérations dérivables. Ainsi, il est possible d'intégrer ce module dans tout réseau de neurones entraîné par rétropropagation de bout en bout.

Depuis leur introduction, les STN ont également été utilisés dans des travaux de recalage et modélisation du mouvement avec des réseaux de neurones [92, 93]. Dans ces applications, le STN est seulement utilisé afin d'appliquer les déformations prédites par le réseau à l'image ou le volume d'entrée. Le volume déformé peut ensuite être comparé à la sortie attendue par similarité d'image ou toute autre métrique de comparaison. L'ajout de ce module dans les réseaux génératifs comme les autoencodeurs convolutifs permet de les optimiser pour la génération de champs de mouvement plutôt que de générer des intensités de voxels.

Applications à la modélisation du mouvement

Les premières applications des approches d'apprentissage profond à la modélisation du mouvement se concentraient sur des images des scènes naturelles. Dans [94], un réseau convolutif entraîné sur des milliers de séquences d'images naturelles est présenté. Le réseau permet de prédire le mouvement que subira l'objet ou l'humain filmé. En prenant une seule image en entrée, le modèle produit un champ de déformation dense indiquant la direction et l'amplitude du mouvement que subira chaque pixel de l'image. Les auteurs ont défini 40 types de mouvement possibles pour chaque pixel individuel. Ainsi, le problème a été formulé comme une tâche de classification plutôt qu'une tâche de régression. L'approche proposée dans [95] utilise une architecture d'autoencodeur récurrent afin de prédire les mouvements subits par des objets sur de longues séquences de vidéo. Le modèle prend en entrée une paire d'images de la même scène et retourne un champ de mouvement en 3D permettant de prédire l'évolution de la distance de l'objet à la caméra au cours de la vidéo. Ces approches ont montré des résultats prometteurs. Toutefois, elles ont été évaluées sur des ensembles de données montrant des patrons de mouvements de basse complexité comme le KTH action dataset [96] et le NTU RGB+D dataset [97].

L'apprentissage profond a permis d'explorer de nouvelles solutions au problème de suivi de cibles et de modélisation du mouvement en RTGI. Dans le cas des solutions de suivi locales, des approches en 2D [98, 99] et en 3D [100] ont réussi à atteindre une précision millimétrique en utilisant les CNN sur des séquences d'US. En particulier, dans [98], un module d'attention est utilisé afin d'ignorer les portions statiques des images ainsi que des termes de perte adaptés à la tâche de suivi local. Dans [100], le modèle de suivi est composé de deux parties. Un premier module, basé sur les CNN, extraits des caractéristiques des volumes d'US. Le second module analyse les caractéristiques avec des couches de perceptrons afin de rendre une décision quant à la nouvelle position de la cible.

Des modèles de suivi globaux basés sur l'apprentissage profond ont également été développés. Dans [101], un CVAE est utilisé afin de modéliser le mouvement du coeur dans des séquences d'IRM 2D. La formulation probabiliste du modèle leur permettait de simuler et d'interpoler des patrons de mouvement réalistes suite à l'entraînement du modèle. Un modèle à patient unique basé sur les réseaux génératifs advérsariels conditionnels a été présenté dans [102]. Le modèle a été entraîné pour la prédiction de champs de mouvement sur des volumes d'IRM à partir d'un signal substitut d'US 2D qui a été acquis simultanément avec une sonde adaptée à l'IRM. Cette méthode a cependant été validée sur 3 sujets seulement. Dans [93], un modèle de population basé sur les CNN, les STN et des unités de mémoire à court long terme convolutives est introduit. Ce réseau permet de prédire le champ de mouvement du foie en 2D avec une avance atteignant cinq pas de temps. L'évaluation a été effectuée sur trois modalités d'imagerie, soit IRM, CT et US. Par contre, cette approche n'a pas été appliquée à l'imagerie en 3D.

L'avantage principal des approches par apprentissage profond est leur capacité à extraire des caractéristiques d'images pertinentes à la tâche d'estimation de mouvement sans intervention humaine. Contrairement aux modèles de population, il n'est pas nécessaire d'établir des correspondances entre les patients pour l'ensemble de données pour entraîner un réseau de neurones. Ceci réduit significativement la préparation des données et procure une plus grande flexibilité au concepteur de ces modèles. De plus, l'adaptation d'un modèle d'apprentissage profond pour un site anatomique différent requiert moins d'effort que pour les modèles statistiques. Néanmoins, les modèles d'apprentissage profond requièrent de grandes quantités de données pour être capables de généraliser l'accomplissement de leur tâche pour une population de patients. Contrairement aux images naturelles, les images de nature médicale ne sont pas abondantes pour les chercheurs, ce qui ralentit le développement des techniques d'apprentissage profond dans ce domaine.

2.5 Mot de synthèse

La revue de littérature a permis d'exposer les différents aspects composant la problématique à laquelle ce projet de maîtrise tente de répondre. En effet, cette analyse a permis de réaliser que le traitement du cancer du foie par la radiothérapie est un processus complexe nécessitant des approches guidées par l'imagerie médicale comme l'US. Toutefois, il n'existe aucun système commercial utilisant l'US pour le suivi du foie en temps réel dans le cadre des traitements par RTE. Les modèles de mouvement sont un concept qui peut répondre à ce besoin, mais leur construction est coûteuse en temps. Avec l'utilisation de techniques d'apprentissage profond, il est possible de construire ces modèles de mouvement de manière plus efficace et avec moins

d'intervention humaine. Par contre, peu de travaux de modélisation du mouvement avec des modèles d'apprentissage profond ont été effectués avec l'imagerie par US en 3D. Une avenue de recherche qui permettrait d'adresser ces manquements est de développer des approches de modélisation de mouvement basées sur l'apprentissage profond qui sont adaptées au suivi de foie sur des séquences d'US en 3D.

CHAPITRE 3 MÉTHODOLOGIE DU TRAVAIL DE RECHERCHE

Suite à l'identification de la problématique au chapitre précédent, il est possible de définir l'objectif de ce projet de recherche. Le but principal est de développer une méthode d'estimation du mouvement du foie dans les séquences d'US 3D pour des applications en RTE. La solution proposée doit prendre en compte les limitations liées au flot de travail courant en RTE comme le type et le nombre d'acquisitions qu'il est possible d'obtenir avant et durant le traitement.

Le point de départ de ce projet est donc de concevoir un modèle qui permet de prendre en entrée de l'information 3D statique acquise avant le traitement et de l'information en 2D courante sur l'état du foie au cours du traitement. En sortie, ce modèle donne de l'information en 3D quant à la forme du foie mise à jour, permettant de déterminer la nouvelle position de cibles anatomiques ainsi que d'autres structures d'intérêt.

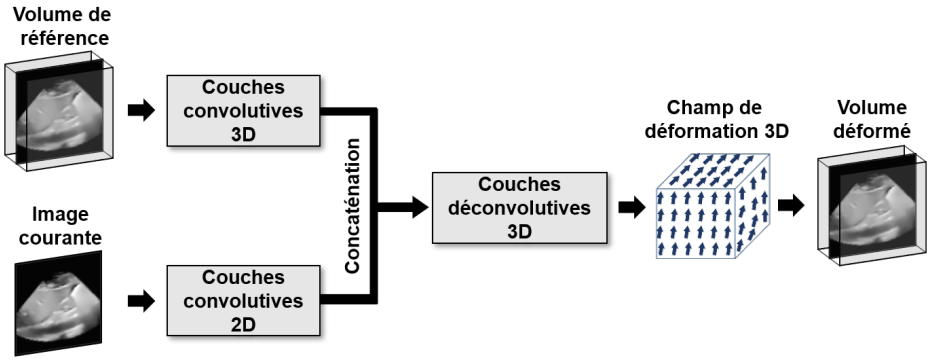


Figure 3.1 Architecture du modèle basé sur les CNN

Inspirés par les récents travaux appliquant les principes d'apprentissage profond à l'imagerie médicale, nous avons exploré les moyens d'adapter ces modèles à la tâche d'estimation du mouvement dans les séquences d'US en 3D. Une première piste de solution se basait sur les capacités d'extraction de caractéristiques des CNN. Cette première approche est présentée à la figure 3.1. Le modèle était conçu de deux branches d'entrée. La première analysait un volume de référence en 3D du foie et la deuxième analysait une image en 2D représentant le foie dans sa position courante. Chaque branche produisait des cartes de caractéristiques à l'aide de couches convolutives en 3D ou en 2D selon le type de l'entrée. Ces caractéristiques étaient ensuite combinées puis envoyées dans une séquence de couches déconvolutives qui permettaient de récupérer le champ de déformation à appliquer au volume de référence afin de reproduire la position courante du foie en 3D. Le développement de cette méthode a mené

à la publication d'un article à la conférence *International Symposium on Biomedical Imaging* (ISBI) en avril 2020. L'évaluation a été effectuée sur un ensemble de données publiques fournies par le concours CLUST15 [41]. Cet article est présenté au chapitre 4.

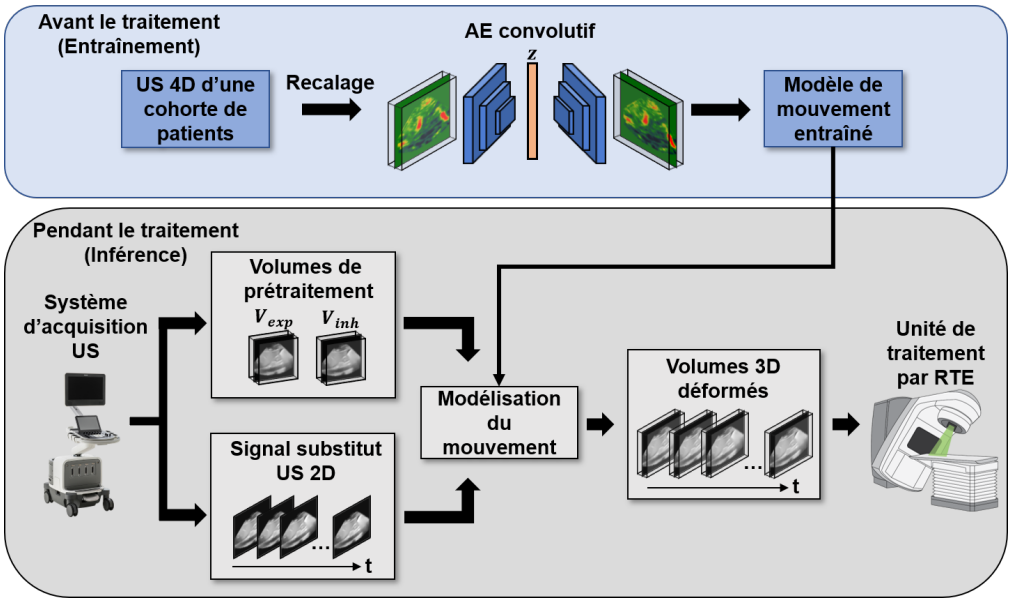


Figure 3.2 Schéma d'entraînement et d'inférence du modèle basé sur l'AE convolutif

Malgré la validité de ce premier modèle en théorie, l'évaluation de ce dernier a permis de révéler des limitations pratiques. Notamment, le modèle était chargé d'inférer une déformation 3D sans connaître la nature d'un champ de déformation valide. Ceci limitait la performance du modèle pour des données exclues de l'ensemble d'entraînement. Pour pallier cette limitation, les approches de modélisation du mouvement basées sur les AE convolutifs ont été explorées. La stratégie d'entraînement était de premièrement apprendre les caractéristiques principales qui décrivent les champs de mouvement pour ensuite apprendre à les générer à partir d'un signal substitut en 2D. La figure 3.2 décrit cette stratégie d'entraînement et l'utilisation du modèle entraîné lors d'un traitement.

Différentes variations de l'AE convolutif ont été considérées et entraînées pour la génération de champs de mouvement 3D à partir d'images 2D. Cette démarche a souligné la nécessité d'unifier de l'information provenant d'une population de patients avec de l'information unique à un seul patient lors de la génération du mouvement. Il s'est également avéré que modéliser le mouvement avec deux composantes, une rigide et une déformable, permet d'améliorer les prédictions du modèle. C'est avec ces constatations en tête qu'un second article, soumis au journal *Medical Image Analysis*, a été rédigé. Cet article présente une évaluation sur des données acquises auprès de 20 volontaires en santé. L'acquisition de ces données est

décrite à la section 3.1. Cet ensemble de données a été choisi à la place de l'ensemble de données publique puisqu'il comportait moins de variations quant au champs de vue du foie. Il comprenait également plus d'annotations de cibles anatomiques pour des fins d'évaluation de suivi. Le chapitre 5 présente les détails de ce deuxième article.

Voici la liste complète des publications réalisées dans le cadre de ce projet de maîtrise :

— T. Mezehritsky, L. Vázquez-Romaguera, S. Kadoury, "3D ULTRASOUND GENERATION FROM PARTIAL 2D OBSERVATIONS USING FULLY CONVOLUTIONAL AND SPATIAL TRANSFORMATION NETWORKS", *IEEE International Symposium on Biomedical Imaging*, publié, Avril 2020

— T. Mezehritsky, L. Vázquez-Romaguera, W. Le, S. Kadoury, "POPULATION-BASED 3D MOTION MODELLING FROM CONVOLUTIONAL AUTOENCODERS FOR 2D ULTRASOUND-GUIDED RADIOTHERAPY", *Medical Image Analysis*, soumis, Mars 2021

3.1 Acquisition des données US 4D

Figure 3.3 Position de la sonde lors des acquisitions d'US 4D

Un ensemble de données composé de séquences d'US 4D en respiration libre a été acquis auprès de 20 volontaires en santé ayant fourni leur consentement écrit. Les acquisitions ont été faites avec le système Philips EPIQ 7G muni de la sonde matricielle Philips X6-1. Au cours de l'acquisition, la sonde a été placée sous le sternum le long du plan sagittal (voir figure 3.3). La profondeur a été fixée à 12 cm. Le focus et le contraste ont été ajustés pour chaque volontaire afin de mieux voir les bordures du foie et ses vaisseaux. Avec une fenêtre temporelle d'acquisition de 15 secondes, il était possible d'imager jusqu'à 3 cycles respiratoires avec une

résolution temporelle de 250 ms. Chaque séquence est donc composée d'approximativement 60 volumes par volontaire, donnant un total de 1200 volumes pour l'ensemble de données. Les volumes bruts ont été filtrés avec un filtre de moyennes non locales bayésien [103] afin de réduire la présence de «speckle». Ensuite, les volumes ont été échantillonnés avec une résolution spatiale de 2.0×2.0 mm^2 dans le plan sagittal et une épaisseur de coupe de 1.0 mm. Finalement, les volumes ont été recadrés à une taille de $64 \times 64 \times 32$ voxels (rangées $\times$ colonnes $\times$ tranches). Pour chaque séquence, 4 cibles anatomiques telles que des vaisseaux ou des bordures du foie ont été identifiées manuellement au cours d'un cycle respiratoire.

3.2 Validation des solutions

Afin de maximiser la taille de l'ensemble d'entraînement et d'utiliser chacun des 20 cas dans l'ensemble de test, une stratégie d'entraînement «leave-one-out» a été utilisée. Avec cette stratégie de validation croisée, 20 modèles sont entraînés avec un cas de test différent pour chacun. De cette manière, la performance du modèle est évaluée pour tous les cas de l'ensemble des données tout en assurant une bonne séparation entre les données d'entraînement et de test.

La similarité entre les volumes générés par le modèle et les volumes de la séquence 4D a été évaluée selon trois métriques de similarité d'image. La moyenne de l'erreur au carré (MSE), la corrélation croisée normalisée (NCC) et l'index de similarité structurelle (SSIM) [104]. Ces métriques permettent d'évaluer globalement la qualité des volumes produits par le modèle.

Dans le but d'évaluer la qualité des champs de déformation produits par le modèle, le déterminant Jacobien de ces derniers a été calculé. Un déterminant Jacobien supérieur à 1 indique une expansion tandis qu'un déterminant positif, mais inférieur à 1 indique une contraction. Enfin, un déterminant négatif indique un repliement dans le champ de déformation, ce qui le rend anatomiquement impossible puisque les tissus sont considérés comme incompressibles. En quantifiant le pourcentage des voxels avec un déterminant Jacobien négatif, il est possible d'évaluer la qualité des champs de déformation produits.

La précision quant à l'emplacement des cibles anatomiques suivies dans les volumes générés a été quantifiée pour évaluer la capacité du modèle à suivre des cibles telles que des vaisseaux ou des bordures du foie. Les cibles anatomiques identifiées sur les volumes 4D ont été annotées de façon manuelle dans les volumes générés. L'erreur entre la position attendue de la cible et sa position générée a été calculée par distance euclidienne.

Plus de détails sur l'implémentation et l'utilisation de ces métriques se trouvent dans les chapitres 4 et 5.

CHAPITRE 4 ARTICLE 1 : 3D ULTRASOUND GENERATION FROM PARTIAL 2D OBSERVATIONS USING FULLY CONVOLUTIONAL AND SPATIAL TRANSFORMATION NETWORKS

Cet article accepté à la conférence ISBI 2020 présente notre première implémentation d'un modèle d'apprentissage profond pour la génération de volumes US 3D à partir d'un volume de référence et des images 2D. L'article présente des résultats de similarité d'images pour différentes configurations du modèle proposé ainsi que des résultats de suivi de cibles sur les données du concours public CLUST15 [41].

Auteurs

Tal Mezheritsky[1], Liset Vázquez Romaguera[1], Samuel Kadoury[1,2]

Affiliations

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

[2] CHUM Research Center, Montréal, Canada

4.1 Abstract

External beam radiation therapy (EBRT) is a therapeutic modality often used for the treatment of various types of cancer. EBRT's efficiency highly depends on accurate tracking of the target to be treated and therefore requires the use of real-time imaging modalities such as ultrasound (US) during treatment. While US is cost effective and non-ionizing, 2D US is not well suited to track targets that displace in 3D, while 3D US is challenging to integrate in real-time due to insufficient temporal frequency. In this work, we present a 3D inference model based on fully convolutional networks combined with a spatial transformative network (STN) layer, which given a 2D US image and a baseline 3D US volume as inputs, can predict the deformation of the baseline volume to generate an up-to-date 3D US volume in real-time. We train our model using 20 4D liver US sequences taken from the CLUST15 3D tracking challenge, testing the model on image tracking sequences. The proposed model achieves a normalized cross-correlation of 0.56 in an ablation study and a mean landmark location error of 2.92 $\pm$ 1.67mm for target anatomy tracking. These promising results demonstrate the potential of generative STN models for predicting 3D motion fields during EBRT.

Keywords 3D volume inference, liver cancer, ultrasound, deep learning, spatial transformation networks

4.2 Introduction

Every year, more than 14 million new cases of cancer are diagnosed world-wide, and it is estimated that 50 percent of cancer patients can benefit from radiotherapy to control and manage their disease [105]. External beam radiation therapy (EBRT) is a particular therapeutic modality that applies a dose of ionising radiation to a region of cancerous tissues within the patient's body using external collimators. In order to avoid damage to healthy

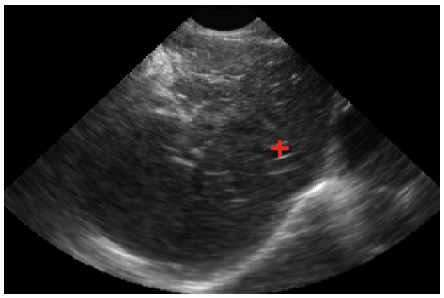


Figure 4.1 Sample 2D US image of a liver from the CLUST15 challenge 3D dataset. The vessel to be tracked is indicated by a red cross.

tissues and precisely target the tumor, the treated organ needs to be constantly imaged and located during treatment [106]. Two-dimensional ultrasound (2D US) imaging in EBRT has been extensively studied [29, 33, 106] and is preferred to X-ray based tracking methods since it is cost effective and non-ionizing. However, 2D US fails to provide information outside of the imaged plane, making it difficult to track targets that move in three dimensions such as tumors or vessels. Three-dimensional (3D) US on the other hand, is able to provide the out-of-plane information missing in 2D US but with a much lower temporal resolution making it inefficient for real-time tracking tasks.

An efficient way to combine the strengths of both 2D and 3D US would be to perform 2D image-based 3D inference of the US volumes. This task is an ill-posed problem that has been tackled in computer vision and deep learning in recent years for object tracking. In [107], a conditional variational autoencoder (REC-CVAE) network was proposed to perform the 3D reconstruction of the fetal skull surface from 2D US standard planes of the head. Another work by [73] used orthogonal interleaved 2D cine-MRI sequences and a reference 3D MRI volume to extrapolate the full 3D deformation field from the partial interleaved 2D deformation fields obtained by performing deformable image registration (DIR) between the cine-MRI slices and corresponding slices in the reference 3D volume.

In this work, we present a network that predicts a liver 3D US volume at any time t_n during free-breathing interventions, using a prior baseline 3D US volume taken at time t_0 and a single 2D US image taken at a given time t_n. Our generative network is inspired by [108] uses fully convolutional layers to extract features from the reference 3D volume and the 2D image. The features are then combined, upsampled and fed to a spatial transformer network (STN) layer [91] to generate the deformation that needs to be applied to the reference volume in order to obtain the new volume at time t_n. The network is evaluated on the dataset from the CLUST15 MICCAI challenge [41] using spatial correspondence and landmark tracking metrics.

4.3 Materials and Methods

4.3.1 Dataset and Setup

The CLUST15 dataset [41] was originally used for 2D and 3D liver vessel tracking challenges. Out of the 22 available 4D US sequences only 20 were used. From those, 11 sequences were acquired on a Philips iU22 system with a X6-1 probe and 9 sequences were acquired on a GE E9 system with a 4V-D probe. We discarded 2 cases due to inconsistent field of view. Sequences acquired on the iU22 system have a spatial resolution of $1.14 \times 0.59 \times 1.19$

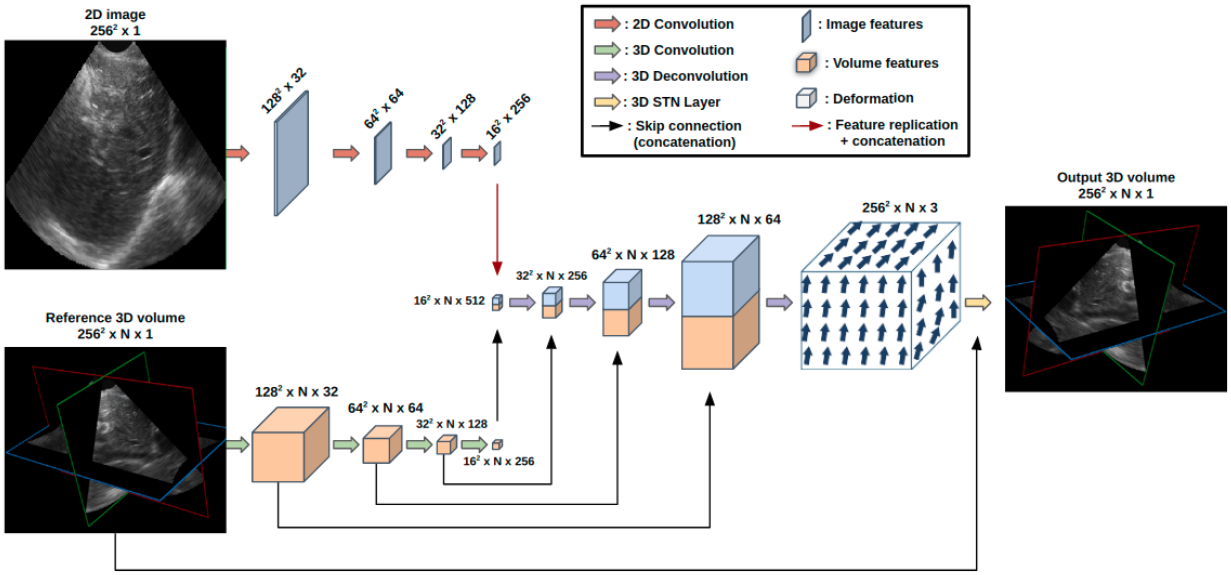


Figure 4.2 Schematic representation of the proposed model (n represents the thickness of the volumes).

mm and a temporal resolution of 6 Hz while the sequences acquired on the GE system have an isometric spatial resolution of 0.7 mm and a temporal resolution of 8 Hz. Each 4D ultrasound sequence is composed of 100 3D US volumes on average, providing approximately 2000 distinct volumes to train, validate and test our network.

Out of the 20 sequences, six provide manual annotations of anatomical landmarks such as vessels and bifurcations (Figure 4.1). The number of landmarks per sequence varies between 1 and 4 while the number of annotations per landmark also varies from 7 up to 100.

For our experiments, we needed to generate a dataset of 3D volumes with corresponding 2D images. Therefore, for each sequence, the first volume was used as the reference volume at t_0 and subsequent volumes were used as ground truth volumes at t_n. Finally, the input images at t_n were obtained by extracting the middle slice from each ground truth volume at the corresponding time t_n. Input and ground truth volumes were cropped to have a thickness of n slices.

4.3.2 Proposed FCN-STN Model

Our architecture is composed of two fully convolutional encoders, one fully convolutional decoder and a STN layer. The encoders act as feature extractors for the 2D image and the reference 3D volume. Figure 4.2 shows a schematic representation of the proposed model.

Each encoder uses four strided convolutional layers followed by batch normalization and ReLU activation to gradually reduce the dimensions of the inputs. Once both inputs are encoded, the features from the 2D image are replicated to match the size of the features from the reference 3D volume and both sets of features are concatenated along the channel dimension. Every convolution of the 2D image encoder uses kernel size = 4, stride = 2 and padding = 1. For the 3D volume encoder, every convolution uses kernel size = (4,4,3), stride = (2,2,1) and padding = 1 to only reduce the size of the first 2 dimensions and leave the size of the third dimension the same.

The decoder uses four strided deconvolutional layers followed by batch normalization and leakyReLU activations with $\alpha = 0.2$ to upsample the features back to the original dimension of the reference volume with 3 channels. Every deconvolutional layer uses the same kernel, stride and padding parameters as the volume encoder. Before every deconvolutional layer, features from the reference 3D volume encoder are concatenated to the decoded features through skip connections. This forces the network to use information from the reference volume during the decoding of the features.

STNs have been recently proposed as a module that can spatially transform an image or feature map by learning appropriate transformations. The transformation matrices may include both affine and non-rigid deformations. Furthermore, it presents the advantage of being differentiable, allowing for end-to-end trainable models using standard back-propagation. We placed this layer at the end of the model in order to take the output of the decoder and generate the non-rigid deformation field that need to be applied to the reference volume to obtain the new volume at time t_n. We used the spatial transformation function implemented by [109] inspired by STN.

4.3.3 Training Protocol

Data augmentation : In order to improve the quality of the generated volumes, a data augmentation strategy was used to increase the dataset size. During regular training, the input 2D image was extracted from the center of the ground-truth 3D volume since it is the slice that contains the least 0-intensity voxels within the volume. While other slices in the volume contain less voxel information, they can still be used for training. Therefore, we increased the dataset size by using input 2D images extracted in off-center positions from the ground-truth volume. Due to hardware constraints we limited ourselves to a five-fold dataset augmentation meaning that we used 4 additional off-center positions in the ground-truth volume to extract input 2D images. The off-center positions were sampled at at regular interval of 5 slices and were oriented the same way as the central slice.

Data preprocessing : We applied mean centering and standard deviation normalization to all input images and volumes. Zero-padding was used to set all the inputs with a size of 256×256 along the first two dimensions.

Train/Validation/Test split : We used a 60/10/30 split to be able to use all the annotated sequences (6 out of 20) in the test set. Thus, the model was tested with unseen subjects.

Hyperparameters and optimization : We used a mean squared error (MSE) loss, computed on image intensities, with an Adam optimizer [110]. The initial learning rate was set to 10^{-4} and was reduced by a factor of 2 after every 10 epochs without improvement to the loss on the validation set. The minimum acceptable learning rate was 10^{-10}. Training was performed with batch size of 4 to benefit from GPU acceleration during training. Models were left to train until the validation loss stabilized for 30 epochs. On average, the models converged after 100 epochs.

First, we evaluated the proposed model in an ablation study, comparing the effect of each component on the model's overall performance. Our baseline model, hereinafter referred as the Encoder/Decoder network (EDNet), is only composed of the encoder and the decoder described in section 4.3.2. Its output is the inferred voxel intensities of the volume at time t_n. The next model adds the skip connections between the volume encoder features and the decoder. This model is referred as EDNet + skip. The third model adds the STN layer discussed in 4.3.2 at the end of the decoder, meaning that it is the first model that does not directly infer voxel intensities. Instead, it learns to infer the deformation that needs to be applied to the reference volume. We refer to this model as EDNet + skip + STN. Finally, the proposed model is obtained by adding the data augmentation strategy presented in 4.3.3 to the training of the EDNet + skip + STN model. All the networks mentioned above were trained to generate volumes with thickness $n = 15$. For the last two models that we evaluated, we decided to increase the volume thickness to $n = 30$ and $n = 45$ to measure the proposed model's ability to generate increasingly larger volumes.

4.4 Results and discussion

We evaluated all versions of the model using 2 metrics. First, the spatial correspondence between the generated volumes and the ground-truth volumes was assessed using normalized cross-correlation (NCC). Then, manual landmark annotations available from the CLUST15 challenge were compared to the predicted landmark positions in the generated volumes.

Figure 4.3 shows the comparison of the six models based on the NCC metric. Being a normalized metric, the closer the NCC is to 1, the more similar the generated volume is to

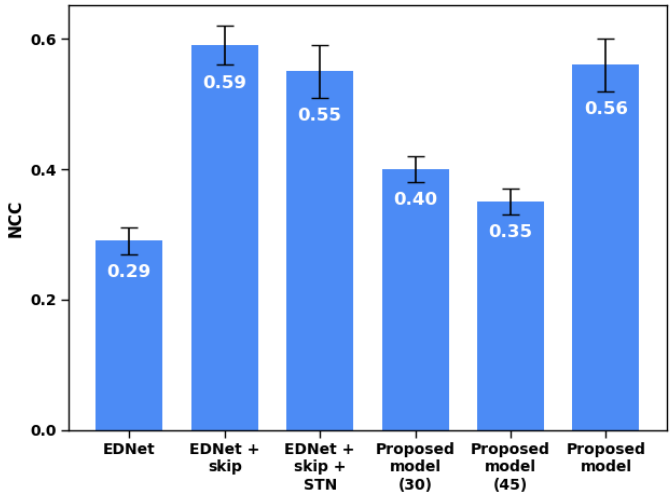


Figure 4.3 NCC results for the evaluated models. Values show the mean NCC between the generated and target volumes of the test set. All means are statistically different with $\alpha < 0.01$.

Tableau 4.1 Tracking performance of the trained models based on average landmark location error (LLE).

Model	**Mean ± STD LLE (mm)**	**Maximum LLE (mm)**
EDNet + skip	4.10 ± 2.70	24.00
EDNet + skip + STN	2.99 ± 1.91	17.85
Proposed model	**2.92 ± 1.67**	**15.35**

the ground-truth. For the models of the ablation study, we observed that the addition of each component (skip connections, STN layer and data augmentation strategy) increases the performance of the base EDNet model.

The component that has the strongest impact is the addition of the skip connections. Without skip connections, the basic EDNet tends to ignore the features from the reference volume, and mostly replicates the input image throughout the output volume. By injecting the features from the reference volume representations during the decoding process, the model is forced to use this information. Adding the STN layer slightly reduces the performance of the model. However, the data augmentation strategy improves the quality of the generated volumes since the model has access to a larger variety of training examples. As for the generation of thicker volumes, the performance is reduced when the model is asked to learn to generate thicker

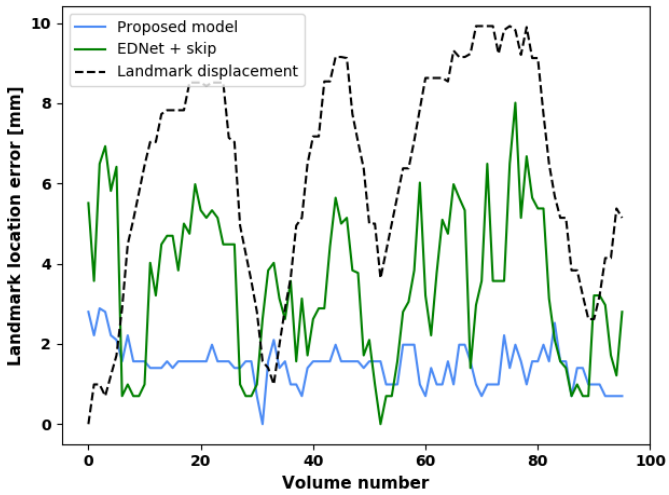


Figure 4.4 Comparison of a vessel bifurcation tracking in a sample sequence of volumes generated by different models. The landmark displacement plot is calculated using the landmark's ground truth positions with respect to its initial position.

volumes ($n > 15$). Since all the models were trained with early stopping, volume thickness is one of the limitations.

For the landmark tracking evaluation, three different models were studied. We compared the proposed model to its versions without the data augmentation strategy and without the STN layer. Table 4.1 shows the mean landmark location error (LLE) in millimeters while Figure 4.4 shows the LLE value for a sample sequence of volumes. As mentioned in section 4.3.1, only landmarks that had ground truth annotations were tracked in the corresponding generated volumes. The proposed model and its version without data augmentation produce similar results on average, showing that even without the improvement in image quality with data augmentation, it is still possible to identify the different targets in the image. In the case of the EDNet + skip model, the mean LLE is significantly greater. Upon visual inspection of the generated volumes, we can observe that without the STN layer the model only learns to deform the reference volume around the center (where the up-to-date information from the input image is available). The further the generated slice is from the center of the volume, the more it resembles the reference volume, thus not showing proper movement. This is supported by the EDNet + skip plot in Figure 4.4, that exhibits a cyclical behaviour, increasing as the landmark moves away from its position in the reference volume. With the addition of the STN layer, the model can learn to deform the entire reference volume and although the image quality is slightly reduced (Figure 4.3), the model generates volumes that give better tracking performances, which is the primary focus for medical applications.

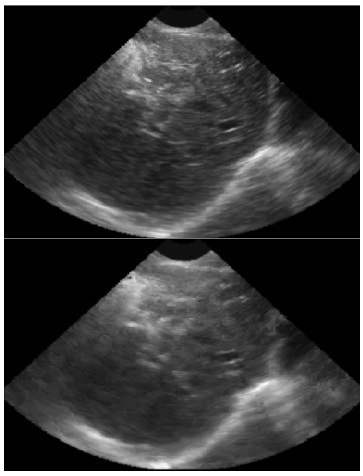


Figure 4.5 Example of a liver from a volume generated by the proposed model (bottom) with its respective ground truth image (top).

Finally, Figures 4.5 and 4.6 show qualitative results of the proposed model. Figure 4.5 shows a slice from the generated volume (thickness $n = 15$) and Figure 4.6 shows a comparison between the input reference volume, the generated volume and the ground-truth volume when viewed along the thickness dimension of size $n = 30$. The latter shows that the model is able to infer the motion that needs to be applied to the entire reference volume rather than just the slices that are closest to the input 2D image.

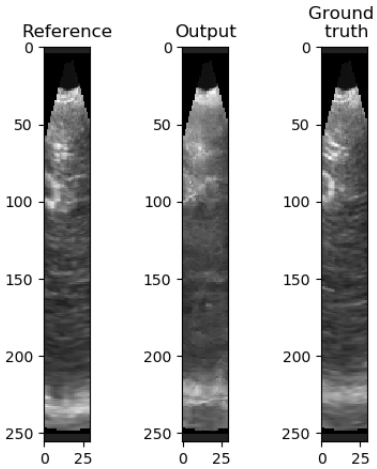


Figure 4.6 Comparison of the input reference volume, the output generated volume and the ground-truth volume when viewed along the thickness dimension ($n = 30$).

4.5 Conclusion

In this work, we presented a 3D inference model using as inputs a real-time 2D US image taken at time t_n and a reference 3D US volume taken at baseline to generate an updated 3D US volume at time t_n. The performed ablation study allowed us to identify the skip connections between the reference volume encoder and the decoder as a crucial component for the generation of good quality volumes. The landmark tracking evaluation showed that the final STN layer allows the model to learn a complete deformation of the reference volume, yielding improved tracking performance. The main limitations of this work are the reduced image resolution compared to ground truth and the limited thickness of the generated volumes.

CHAPITRE 5 ARTICLE 2 : POPULATION-BASED 3D MOTION MODELLING FROM CONVOLUTIONAL AUTOENCODERS FOR 2D ULTRASOUND-GUIDED RADIOTHERAPY

Cet article soumis au journal Medical Image Analysis (MedIA) présente un système de modélisation de mouvement pour séquences d'US 3D basé sur l'apprentissage profond. Le système est composé d'un module d'alignement rigide ainsi qu'un module de mouvement déformable. Le modèle proposé est évalué sur l'ensemble de données US 4D décrit à la section 3.1. L'article rapporte des résultats de similarité d'images et de suivi de cibles pour différentes configurations du modèle ainsi que pour des méthodologies comparables dans la littérature.

Auteurs

Tal Mezheritsky[1], Liset Vázquez Romaguera[1], William Le[2], Samuel Kadoury[1,2]

Affiliations

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

[2] CHUM Research Center, Montréal, Canada

5.1 Abstract

Radiotherapy is a widely used treatment modality for various types of cancers. A challenge for precise delivery of radiation to the treatment site is the management of internal motion caused by the patient's breathing, especially around abdominal organs such as the liver. Current image-guided radiation therapy (IGRT) solutions rely on ionising imaging modalities such as X-ray or CBCT, which do not allow real-time target tracking. Ultrasound imaging (US) on the other hand is relatively inexpensive, portable and non-ionising. Although 2D US can be acquired at a sufficient temporal frequency, it doesn't allow for target tracking in multiple planes, while 3D US acquisitions are not adapted for real-time. In this work, a novel deep learning-based motion modelling framework is presented for ultrasound IGRT. Our solution includes an image similarity-based rigid alignment module combined with a deep deformable motion model. Leveraging the representational capabilities of convolutional autoencoders, our deformable motion model associates complex 3D deformations with 2D surrogate US images through a common learned low dimensional representation. The model is trained on a variety of deformations and anatomies which enables it to generate the 3D motion experienced by the liver of a previously unseen subject. During inference, our framework only requires two pre-treatment 3D volumes of the liver at extreme breathing phases and live 2D surrogate images representing the current state of the organ. In this study, the presented model is evaluated on a 4D US data set of 20 volunteers based on image similarity as well as anatomical target tracking performance. We report results that surpass comparable methodologies in both metric categories with a mean tracking error of 3.5±2.4 mm, demonstrating the potential of this technique for IGRT.

Keywords Motion modelling, Ultrasound-guided radiotherapy, Deformable registration, Liver cancer, Convolutional autoencoders

5.2 Introduction

Radiation therapy (RT) is used in more than 50% of cancer patients to treat and control disease progression [105]. External beam radiotherapy (EBRT), a specific modality of RT, uses an external radiation source and collimators to deliver precise doses of radiation to the tumor site from different orientations around the patient's body. The goal of EBRT is to deliver enough radiation to damage the genetic material of cancerous cells, thus disabling them from dividing and growing the cancerous tumor further [111]. However, radiation is not only harmful to cancerous cells, it can also damage healthy cells [18], making the precision of RT delivery systems crucial, especially for organs at risk. In the case of EBRT, the most

complex sites to treat are the ones that experience severe motion induced by the patient's breathing. Indeed, respiratory motion poses great challenges to the administration during EBRT due to large motion organs such as the liver [112]. This forces radio-oncologists to increase the treatment margins to reduce the probability of cancer recurrence, thus increasing toxicity to healthy tissues [7]. In an attempt to minimize the negative effects of respiratory motion on the efficiency of EBRT, a variety of respiratory motion management techniques have been proposed and used in clinical settings.

For respiratory gating approaches, the treatment is administered only within a predefined range of the patient's respiratory cycle using breath-holds. On the other hand, the forced shallow breathing technique does not require the patient to temporarily stop breathing, however it reduces the amplitude of respiratory motion by applying pressure to the patient's abdomen [7]. While improving the precision of EBRT, the aforementioned techniques bear limitations such as increased treatment time and physical discomfort to the patient. For image-guided radiotherapy (IGRT), the aim is to use imaging to track the treatment target at all times during the administration of radiation to the tumor site. As the target is tracked, the delivery system adjusts its beam to account for the displacement of the tumor inside the patient's body. Therefore, IGRT has the potential of reducing the amount of damage caused to healthy tissues due to large treatment margins, all while allowing the patient to breath freely during the procedure [113].

A wide range of imaging modalities can be used in the context of IGRT. X-ray imaging with or without the implantation of fiducial markers is often used in clinical practice, however the additional radiation dose it imparts reduces the imaging frame-rate that can be used. Similarly, cone-beam computed tomography (CBCT) cannot be used in real time during treatment due to significant exposure to ionizing radiation. In recent years systems that use MRI for IGRT have emerged, however they aren't widely available yet [26]. In contrast, ultrasound (US) is a non-ionizing, portable and inexpensive medical imaging modality that circumvents most of the disadvantages of other imaging modalities within the scope of IGRT. As current US system are capable of 2D, 3D and 4D imaging, they can be used both in the planning and treatment stages of the RT workflow [33].

However current US-based IGRT systems rely on 2D imaging to track targets during imaging, even though targets are known to experience complex 3D trajectories especially in organs such as the liver and lungs [7]. Therefore, 3D US imaging can be useful in IGRT applications. Clinically available 3D US matrix-array probes provide complete anatomical information of the tissues surrounding the tumor target, still the acquisition frame-rate is significantly lower than in 2D US imaging and the considerable storage size of 3D volumes significantly

increases processing and computing times, making it difficult to use for real-time IGRT applications [26]. As such, we present a hybrid solution employing both 2D and 3D US for US-guided EBRT, by learning the relationship between 2D images and 3D deformation fields for real-time inference of volumetric US imaging.

5.2.1 Related works

The task of tracking anatomical targets in 3D US has generated significant interest, leading to open challenges like CLUST15 [41], providing a common datasets to compare solutions both on 2D and 3D temporal US sequences. [47] proposed a block matching multi-step tracking approach where each step accounted for an increasingly finer level of motion. [48] proposed a tracking technique based on supporter features surrounding the tracking target. By tracking the supporters, the tracking accuracy of the desired target was improved. Both approaches achieved sub-millimeter performance, however they were only tested on 2D images. Methods tested on 3D US data included [49], registering a global point set across temporal volumes using block matching, followed with a 3D registration of a local point set around the anatomical landmark, while [114] represented the 3D target as a model of tetrahedral cells and vertices. The internal and external motion of the target mesh were estimated using a mechanical model and an intensity based approach respectively. In general, local tracking methods share a common disadvantage in failing to provide information about the motion of surrounding tissues which could be useful for dose re-planning [53].

Global tracking solutions, on the other hand, attempt to determine the new position of a target by providing the motion experienced by its surroundings and the treated organ as a whole. The expected output becomes a motion field that spans the entire volume, which can be used not only to track treatment targets but also to adjust treatment planning and dose calculation. Obtaining complex 3D motion fields by leveraging inputs of a lower dimension has been commonly achieved in the context of IGRT through the use of motion modelling [50]. Surrogate signals can be obtained through 1D signals such as spirometry, skin surface motion tracking or 2D images of the treated organ, which can be used during treatment to infer the 3D motion field experienced at the time of the procedure [50]. However, very few works focused on motion modelling for 3D US due to the inherent difficulties such as low image quality and presence of unique artifacts [53]. Nevertheless, motion modelling remains flexible in terms of modality choice, even allowing to use one modality as a surrogate for another in certain applications [54].

A first group of global tracking solutions based on motion modeling are patient-specific, where before treatment, the acquisition of 4D data along with surrogate signals is performed

on the patient. The 3D motion is then obtained through registration of the 4D data to a reference volume chosen at a certain respiratory phase. Several approaches establishing a correspondence between surrogate signals and motion fields have been proposed. [60] created an atlas of motion from 4D MRI data, which is recovered using a respiratory signal acquired during treatment. [61] acquired cine MRI slices at 6 positions across the liver and registered them to a reference 3D MRI volume to obtain a lookup table of extrapolated 3D deformation fields corresponding to a variety of liver states. Among the works on patient-specific motion models, principal component analysis (PCA) stands out as a reference, using a linear decomposition of the patient-specific 3D motion fields, which is recovered using a surrogate signal, such as a 2D navigator [58]. Other means to obtain PCA combination coefficients include maximizing image similarity between acquired surrogate and deformed reference volume slice [65–67] or sparse block matching [68]. [62] proposed to unify the steps of motion calculation and establish the surrogate correspondences which are usually separated. Their general framework showed many advantages, however high computation time limits its use in real-time applications. The main drawback with these approaches is that patient-specific 4D data is needed in order to model the motion patterns, which is far from being widely accessible in all institutions.

The second group of global motion models, referred to as population-based or cross-population models, aims to capture a wider variety of motion fields by capturing motion variability across a population of patients. [69] introduced the concept of exemplar models, where each patient in the data set was used to fit exemplar patient-specific models. For new patients, the obtained surrogate is compared to the exemplar models and an optimized linear combination of all the patient-specific models is obtained. [73] proposed a global motion model that directly infers the complete 3D deformation field by extrapolating the registration of interleaved 2D MRI surrogates with the planning MRI volume. Just as subject-specific models, PCA is also widely used when constructing population based motion models [75–77]. [54] proposed to combine information from 2D US images with a PCA motion model to predict the 3D motion of the liver acquired using MRI. However, the main drawback of using PCA is the requirement of establishing inter-patient correspondences, which is a time-consuming and often inaccurate process.

Deep learning has allowed to explore new solutions for the problem of target tracking in IGRT. Local tracking solutions in 2D [98, 99] and 3D [100] using convolutional neural networks (CNN) have archived accuracies within 3-4mm. [102] proposed a subject-specific motion model based on conditional generative adversarial networks (cGAN). The cGAN learned to predict 3D deformations of MRI based on a simultaneously acquired 2D US surrogate. However this method was only validated on three subjects. [93] introduced a global motion

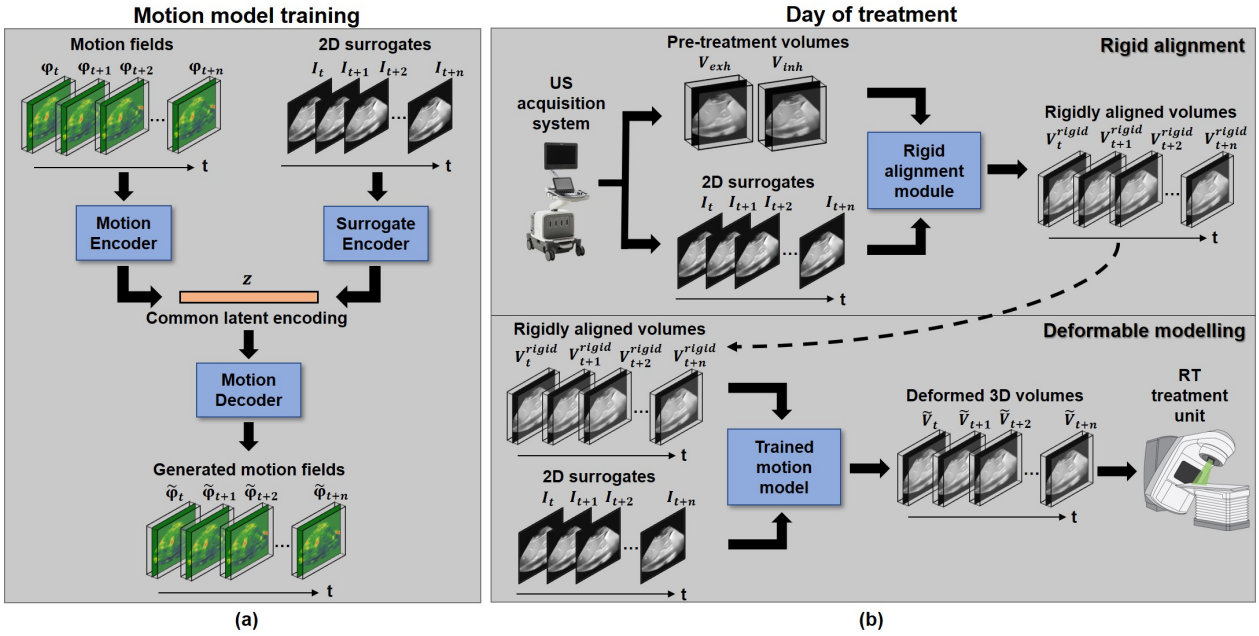


Figure 5.1 Overall training and clinical workflow for the proposed motion modelling framework. (a) A deep motion model is first trained to recover complex 3D motion fields, which are associated to 2D image surrogates through a common latent encoding from a population of subjects. (b) On the day of treatment, two pre-treatment volumes (inhale and exhale) are first acquired before treatment for an unseen subject. During treatment, the real-time 2D images and pre-treatment volumes are fed into the proposed framework, generating deformed 3D volumes of the imaged organ which can be used to adjust the administration of radiotherapy in real-time.

model based on CNN and convolutional long short-term memory (CLSTM) units to perform in-plane target tracking with up to 5 timesteps prediction. The model was validated on 3 imaging modalities (MRI, CT and US), however it can only be applied to 2D images. [115] proposed a model to generate up-to-date US volumes by combining image features from a reference 3D volume and a current 2D US image. While the model showed promise, its validation was limited to a small testing set and tracking of a single anatomical landmark.

5.2.2 Contributions

In this work, a novel motion modelling framework is presented. As shown in Figure 5.1, the deep motion model first learns to link complex 3D motion fields with 2D image surrogates through a common latent encoding. The model also learns to recover the 3D motion fields from the latent encoding. On the day of treatment, the proposed framework composed of a rigid alignment module and the trained deep motion model, is able to process 2D US

acquisitions of previously unseen cases in real time to provide three-dimensional information about the state of the liver. Once sent to the treatment unit, this information can be used to adjust radiation delivery as needed.

The proposed deep motion model is able to capture a wide variety of motion patterns while also taking into account subject-specific anatomical information to improve its prediction. Our proposed framework does not require prior 4D acquisitions for new subjects and removes the need to establish inter-subject correspondences within the training 4D data set, an important advantage over previously presented global motion models.

As such, our main contributions are :

— A novel real-time motion modelling framework composed of a rigid alignment module and a deep deformable model evaluated on 20 free-breathing subjects.

— A convolutional autoencoder motion model which learns to recover complex 3D deformations for a previously unseen subject with only a pair of pre-treatment volumes and a sequence of 2D image.

— The introduction of an image similarity-based rigid alignment strategy to cope with large displacements of the treated organ.

5.3 Methods

In this section, we present our motion modelling framework. We first formally define the problem at hand. Next, we describe in detail each module composing the proposed motion modelling framework as well as its training procedure. Finally, details of the framework's implementation are provided.

5.3.1 Problem formulation

We consider a dataset of free-breathing 4D US acquisitions of the liver from a population of N individuals. For each subject $s_i \in (s_1, s_2, ...s_n)$, a sequence of 3D US volumes $\boldsymbol{V} = (V_1, V_2, ..., V_t)$ is defined, spanning a given time period $[0, t]$. To obtain a temporal sequence of 2D surrogate images $\boldsymbol{I} = (I_1, I_2, ..., I_t)$, the central slice of each volume $V_t \in \boldsymbol{V}$ is extracted from the chosen anatomical plane. In each sequence $\boldsymbol{V}$, two reference volumes are identified at exhale (V_{exh}) and inhale (V_{inh}) respiratory phases. The rationale for this choice is to cover the entire range of variation during a breathing cycle. The motion observed in the volume sequence can be measured by performing rigid and non-rigid registration between V_{ref} and the current volume $V_t \in \boldsymbol{V}$. Since the exhale phase is a more easily reproducible position for the liver, V_{exh} is chosen as V_{ref}. Hence, the sequence of rigid transformations

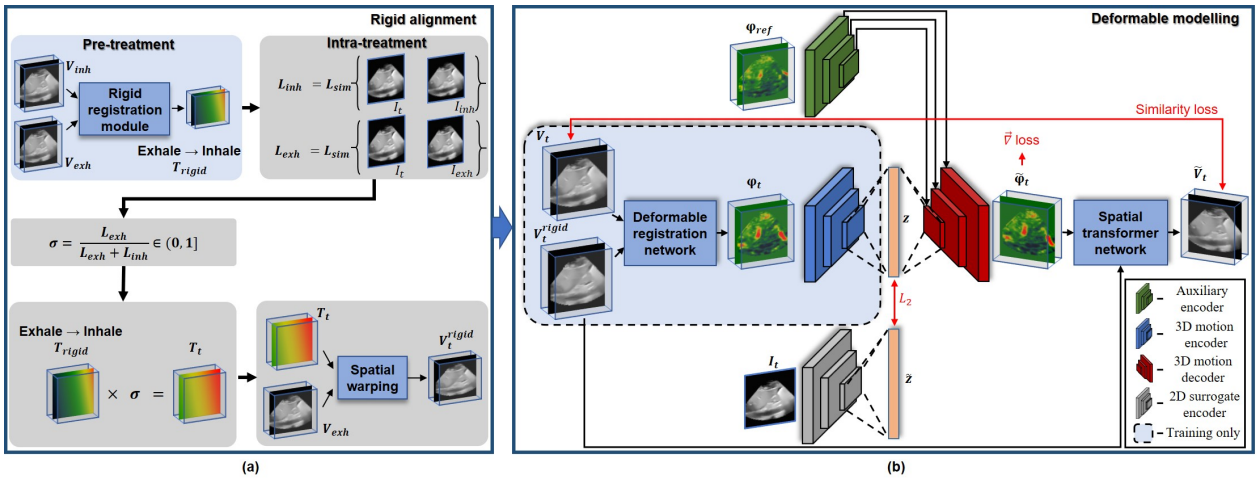


Figure 5.2 Schematic representation of the proposed motion modelling framework. (a) First, a rigid transformation is applied to the reference volume in order to coarsely align it with the current state of the liver. The transformation is based on the similarity of the current surrogate 2D image I_t to the central slices of two pre-treatment volumes acquired at exhale I_{exh} and inhale I_{inh}. (b) Once the rigid alignment is performed, the motion autoencoder receives the registration field between V_t and V_t^{rigid} computed by the deformable registration network. The motion field is compressed into the latent vector z and then recovered with the use of prior subject-specific features from the auxiliary encoder. To be able to generate motion fields in the absence of the motion encoder, the 2D surrogate encoder learns to replicate the latent encoding z from the surrogate 2D image. The generated motion field is used to warp V_t^{rigid} through the spatial transformer network (STN) thereby generating the predicted volume $\widetilde{V}_t$.

$\boldsymbol{T} = (T_1, T_2, ..., T_t)$ and deformation vector fields (DVF) $\boldsymbol{\Phi} = (\phi_1, \phi_2, ..., \phi_t)$ individually represent the deformations that need to be applied to V_{ref} in order to obtain the corresponding volume V_t. The first step is to compute a 3D rigid transformation between V_{ref} and V_t using solely a 2D US image I_t, V_{exh} and V_{inh}. Having the rigidly aligned reference volume ($V_t^{rigid} = \mathcal{T}(V_{exh}, T_t)$), the second step is to learn the deformable component to be applied on V_t^{rigid} to match V_t. Therefore, the prediction of each temporal volume is based only on V_{exh}, V_{inh} and I_t as inputs.

5.3.2 Proposed framework

In the following subsections, we present details about each component of our proposed solution. As shown in Figure 5.2, our solution is composed of 2 main components : a rigid alignment module and a deformable motion model, generating 3D volumes in real-time. The rigid alignment module applies an initial rigid displacement to the reference volume in order

to coarsely align it with the current position of the liver. The rigidly aligned volume is then fed to the deformable motion model which applies finer localized deformations. The deformable motion model generates its output from a learned low-dimensional encoding of the organ's deformation field and subject-specific features included as skip connections.

Rigid alignment

Figure 5.2a illustrates the proposed approach to rigidly align V_{ref} to V_t during treatment by using two pre-treatment volumes at the extreme respiratory phases and a single 2D US image I_t. Before treatment, two volumes acquired at exhale (V_{exh}) and inhale (V_{inh}) phases are rigidly registered. It is assumed that during treatment, the liver will be located within the exhale-inhale range obtained before treatment. Since the pre-treatment volume at exhale corresponds to V_{ref}, the rigid transformations that will be required to align V_{ref} during treatment are bound between the null transformation and the exhale-inhale transformation. To identify the respiratory phase in which the liver is located during treatment, the current 2D US frame I_t is compared to the corresponding central slices of the pre-treatment volumes I_{exh} and I_{inh} using an image similarity metric $\mathcal{L}_{sim}$. The similarity measures $\mathcal{L}_{exh}$ and $\mathcal{L}_{inh}$ are used to compute a scaling factor σ, as follows :

$$\sigma = \frac{\mathcal{L}_{exh}}{\mathcal{L}_{exh} + \mathcal{L}_{inh}} \in (0, 1]. \tag{5.1}$$

This factor tends to 0 when I_t is similar to I_{exh} and dissimilar to I_{inh} and tends to 1 in the opposite scenario. In this manner, when the current state of the liver is close to the reference volume (i.e. exhale), a relatively small displacement is applied. As the state of the liver approaches the one in I_{inh}, σ gradually increases and so does the amplitude of the displacement. The obtained value is applied to the exhale-inhale transformation through element-wise multiplication to produce a scaled version which is finally used to generate V_t^{rigid} by deforming V_{ref}. For this work, the Mean Squared Error (MSE) was chosen as $\mathcal{L}_{sim}$, as it was found to be the more efficient when comparing mono-modal images. It is assumed that I_t, V_{exh} and V_{inh} were all acquired in approximately the same orientation and anatomical location.

Deformable motion modelling

Once rigidly aligned, V_t^{rigid} is fed to the deep deformable motion model shown in Figure 5.2b. The goal of this step is to apply a non-rigid 3D deformation ϕ_t to V_t^{rigid} in order to obtain the final 3D output volume $\widetilde{V}_t$ which represents the current state of the imaged organ

($\widetilde{V}_t = \mathcal{T}(V_t^{rigid}, \phi_t)$). First, a pre-trained deformable registration neural network, described in detail in Section 5.3.3, is used to generate the deformation field ϕ_t between V_t^{rigid} and the current volume V_t. The convolutional motion autoencoder is then trained to compress each 3D motion field $\phi_t \in \mathbf{\Phi}$ into a corresponding low dimensional latent encoding z. The compression is followed by the recovery of the input motion fields from the obtained latent encoding. Subject-specific information is also incorporated through skip connections which originate from a separate auxiliary encoder. Since the 3D motion fields $\mathbf{\Phi}$ that were used as inputs to the motion autoencoder are not available during inference, a separate 2D surrogate encoder is trained to predict the latent encoding z with a different input. Specifically, the 2D surrogate encoder learns to obtain a latent encoding $\widetilde{z}$ as similar as possible to z by only using a single 2D image of the liver's current state as input. By creating this shared latent representation between the 3D motion autoencoder and the 2D surrogate encoder, the model is capable to infer the complete 3D motion field ϕ_t with only a 2D image as input. To obtain the final predicted volume $\widetilde{V}_t$, a spatial transformation network (STN) warps V_t^{rigid} with the generated 3D motion field $\widetilde{\phi}_t$.

Motion autoencoder The central module of the deformable motion model is the 3D motion autoencoder. Its role is to learn how to compress and recover the input ϕ_t so that during inference, only the latent representation $\widetilde{z}$ is needed to obtain $\widetilde{\phi}_t$. The main components of the autoencoder are the 3D motion encoder and decoder. Both are fully convolutional networks that use strided downsampling operations to reduce the spatial dimension of the input. At the bottleneck of the autoencoder, a latent vector z of size 3072 is obtained by passing the 3D motion encoder's output through one fully connected layer. It is important to note that the dimension of the latent vector z should be determined empirically for this application. An excessively small latent dimension might limit the autoencoder's representational capabilities, while a too large of a latent dimension could lead to an over-parametrization of the network. To recover $\widetilde{\phi}_t$, z is reshaped and passed to the 3D motion decoder. The 3D motion decoder uses transposed convolutions to gradually upsample z back to its original size. Detailed information about the implementation of the network's components is presented in Section 5.3.3.

Auxiliary autoencoder As the compression of ϕ_t inevitably involves loss of information, the 3D motion decoder bears the complicated task of recreating that lost information using the latent vector z. This is especially challenging when attempted on a previously unseen anatomy during inference. Therefore, to improve the decoder's performance, subject-specific anatomical information is provided at the decoding stage through the use of skip connections [116] which carry features from a reference DVF (ϕ_{ref}). Figure 5.3 shows how ϕ_{ref} is obtained at any time t. First, I_t is replicated along the third dimension to match the size of V_t^{rigid}. The

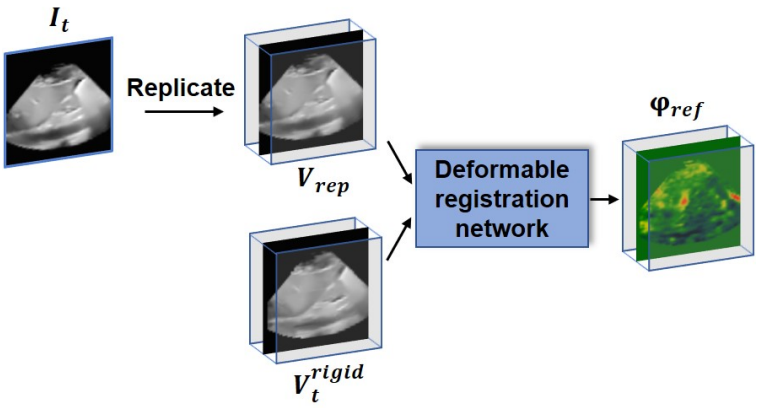


Figure 5.3 Schematic representation of the generation of ϕ_{ref}. The surrogate 2D image is replicated to match the dimensions of V_t^{rigid}. Both volumes are then processed by the deformable registration network to obtain an approximation of the DVF to predict.

resulting volume is denoted as V_{rep}. Then, to obtain ϕ_{ref}, the deformable registration network aligns V_t^{rigid} with V_{rep}. Before being included in the decoding process, ϕ_{ref} is processed by the auxiliary encoder which has an identical architecture as the 3D motion encoder. Once encoded, features from each layer of the auxiliary encoder can be concatenated with the features of the analogous decoding layer of the 3D motion decoder. Those skip connections provide the 3D motion decoder with an approximation of the main features of the expected output DVF for the novel unseen anatomy. It is important to note that the number of skip connections can greatly change the behaviour of the network. The optimal number of skip connections to use in this application is determined empirically and shown in Section 5.4.

Surrogate encoder The surrogate encoder aims to regress a latent representation $\tilde{z}$ as similar to z from a surrogate image I_t. Using this scheme, the 3D deformations are associated with partial observations through their common latent representation. This is achieved by minimizing the following expression :

$$arg\min ||z, \tilde{z}||_2^2 = arg\min_{\theta,\omega} ||f_\theta(\phi_t), g_\omega(I_t)||_2^2 \tag{5.2}$$

where f_θ and g_ω are functions that parameterize the 3D motion encoder and the surrogate encoder, respectively. The surrogate encoder learns to regress the desired latent encoding z, learned during the autoencoder training, by using the surrogate images provided during treatment. The architecture of the 2D encoder is composed of five 2D convolutional layers, all using strided downsampling layers to reduce the dimension of I_t while gradually increasing the number of channels. Finally, $\tilde{z}$ is obtained at the output of two fully connected layers.

Once the common latent representation is established, the surrogate encoder can replace the 3D motion encoder during inference. In this manner, the full deformation field ϕ_t can be recovered using only the 2D image I_t as input.

Spatial transformer network The STN module was originally proposed by [91] to increase the robustness of image registration using convolutional neural networks with respect to spatial variations in their inputs. Since then, it has been used to provide models with the ability to perform spatial warping operations on images and volumes [92, 93]. It is comprised entirely of differentiable operations, which is an important property when used in end-to-end trained models. In this work, the STN is used to warp V_t^{rigid} with the predicted deformation fields ϕ_t to obtain the predicted volume $\widetilde{V_t}$, thereby enabling computing the similarity to the true V_t. By using the STN, the motion autoencoder is optimized to predict deformation fields instead of attempting to directly regress the voxel intensities of V_t.

Training procedure and inference

Training The proposed deformable motion model is trained in 3 steps. First, the autoencoder is trained independently, using the 3D motion fields $\phi_t \in \mathbf{\Phi}$ generated from the registration of V_t^{rigid} and V_t by the deformable registration network. Second, the weights of the autoencoder are fixed while the surrogate encoder is trained to replicate the latent representation of the autoencoder. Finally, all the weights are freed and the entire network is trained together as a final fine-tuning step. During the first step the network is optimized using the first 2 terms of the following loss function :

$$\mathcal{L} = \mathcal{L}_{sim}(\widetilde{V}_t, V_t) + \beta \mathcal{L}_{grad}(\widetilde{\phi}_t) + ||z, \widetilde{z}||_2^2 \tag{5.3}$$

where the first term ($\mathcal{L}_{sim}$) represents the similarity between the predicted volume $\widetilde{V_t}$ and the true current volume V_t. The second term ($\mathcal{L}_{grad}$), weighted by the parameter β, is a gradient penalty for $\widetilde{\phi}_t$ which encourages the generation of smooth and diffeomorphic deformation fields [92]. During the second step, the final loss term in Equation 5.3 is used. It represents the L_2 norm between the autoencoder's latent vector z and the surrogate encoder vector $\widetilde{z}$. Finally, in the last step all the terms of Equation 5.3 are used to fine-tune all of the network components.

Inference Once trained, the deformable motion model is used without the motion encoder, as shown in Figure 5.2b. The inputs during inference are I_t and V_t^{rigid}, which are used to estimate ϕ_{ref}. I_t is used to obtain $\widetilde{z}$, which is passed to the 3D motion decoder. V_t^{rigid} is also used at the last step when it is deformed by the STN to obtain the network's output $\widetilde{V}_t$.

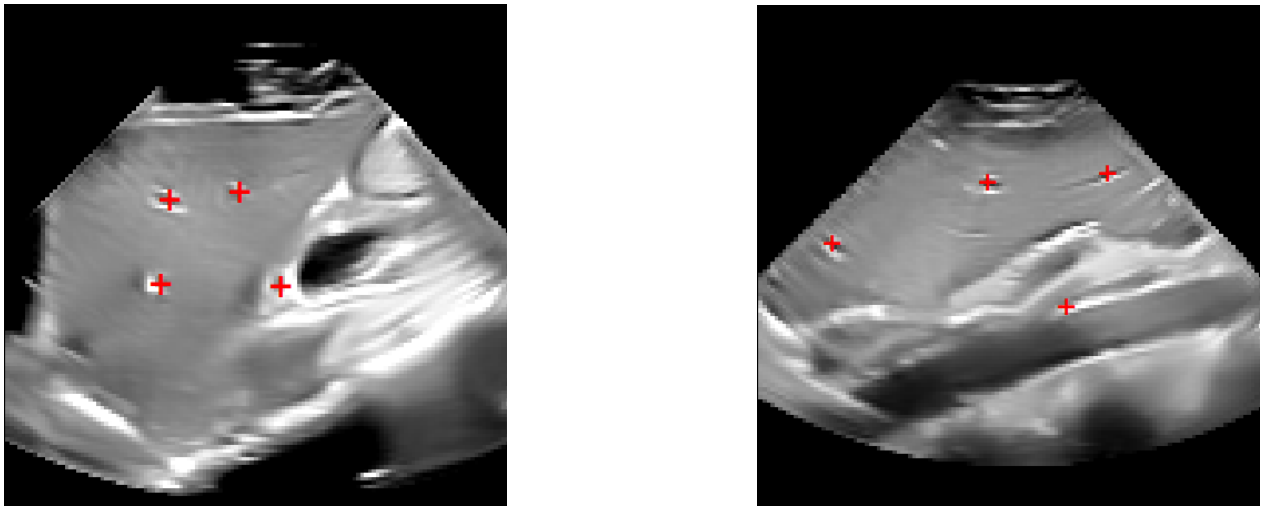
5.3.3 Implementation details

The proposed model was implemented using PyTorch 1.7.0 [117]. The motion encoder is implemented with a 6-layer fully convolutional network. The first three layers include strided downsampling operations with a rate of 2. The number of channels was progressively changed over each layer in the following order $[64, 128, 256, 128, 64, 24]$. The kernel size for all 3D convolutions was $3 \times 3 \times 3$ and the stride and padding were adjusted depending on whether the layer was used for downsampling or not. Each convolutional layer was followed by batch normalization and a ReLU activation layer. The motion decoder is the mirror image of the motion encoder except that all convolutions were replaced by transposed convolutions. Moreover, Leaky ReLU activations with a slope of 0.2 were used for the decoder. The auxiliary encoder has the same architecture as the motion encoder. The surrogate encoder is a fully convolutional network as well. It is comprised of five 2D convolutional layers, four of which use strided downsampling operations. The convolution parameters are the same as for the motion encoder except for the number of channels that was set to $[64, 128, 256, 256, 384]$ to match the dimension of the latent vector. The convolutional layers are followed by two fully-connected layers to regress $\widetilde{z}$.

The Adam optimizer [110] was used with an initial learning rate of 10^{-4} which was halved when the validation loss stopped decreasing for 15 epochs. The stopping criteria for step 1 was met when $\mathcal{L}_{sim}(\widetilde{V}_t, V_t)$ did not improve by 0.01 for 10 epochs. The weighting term β in Equation 5.3 was set to 0.01. Training in step 2 was stopped when $\mathcal{L}_2(\widetilde{z}, z)$ did not improve by more than 0.01 for 10 epochs. Finally, the stopping criteria for the final training step was the same as step 1 but the threshold was decreased to 10^{-3} to allow for fine-tuning.

For the similarity loss $\mathcal{L}_{sim}$, the MSE loss was slightly adapted for the use with US images. Since US acquisitions appear as a conical shape on a black background, there is a large portion of the voxels that contain no information. A mask representing only non-empty voxels was applied to ignore those regions when registering two volumes or computing image similarity.

A leave-one-out validation scheme was employed to evaluate the network's performance on each unseen subject. The deformable registration network used in this work is the U-Net like model proposed by [92]. It is important to note that any deep learning based deformable registration network can be used within our framework. Since this module is used during inference to generate ϕ_{ref}, it was also trained using the leave-one-out scheme to ensure no data leakage between the model components. Finally, the exhale-inhale transformation used in the proposed rigid alignment module was computed with the widely used medical image registration library Elastix [118].



(a) (b)

Figure 5.4 Examples of expert-annotated landmarks placed in the 4D US dataset used for evaluation.

5.4 Experiments and results

In this section, we present the experimental setup used to evaluate the motion modelling framework, with comparisons to state-of-the-art methods. We first present the 4D US dataset that was used to train and test the framework. A first set of experiments is presented to analyze the individual contribution of each component to the framework's overall performance. This is achieved through an ablation study and experiments focusing on individual components such as the rigid alignment module, auxiliary encoder and surrogate encoder. Finally, a second set of experiments is conducted to compare our method to other related approaches based on image similarity and target tracking metrics. Results were tested for statistical significance using the Wilcoxon signed rank test with significance level $\alpha = 1\%$. Effect size was measured using Pearson correlation (ρ).

5.4.1 4D US dataset

A dataset of free-breathing 4D US sequences was acquired from 20 healthy volunteers, who provided their written consent. The acquisitions were performed using a Philips EPIQ 7G ultrasound system with a X6-1 matrix array transducer. During acquisition, the ultrasound probe was placed under the sternum along the sagittal plane, capturing a cross section of the left liver lobe. The imaging depth was set to 12cm. Focus and contrast were adjusted to provide the best visualization of the liver and its vessels. Using a 15 seconds acquisition

window, up to 3 respiratory cycles were captured with a 250ms temporal resolution, producing sequences of around 60 volumes per volunteer. This yielded the total amount of 1200 volumes in the dataset. The acquired volumes were first pre-processed by applying a Bayesian non-local means filter [103] for speckle removal. Then, the volumes were resampled to a 2.0×2.0 mm^2 spatial resolution in the sagittal plane and a slice thickness of 1.0 mm. Finally the volumes were cropped to a size of $64 \times 64 \times 32$ (rows $\times$ columns $\times$ slices). For each sequence, between 4 and 5 anatomical landmarks such as vessels or liver boundaries were manually annotated by an expert on each temporal volume through one respiratory cycle (see Figure 5.4).

5.4.2 Proposed framework analysis

Ablation study In order to better understand the role and contribution of each component of our framework, an ablation study was performed. Different configurations of the proposed deformable model were compared based on the image similarity between ground-truth and predicted volumes. MSE, normalized cross-correlation (NCC) and structural similarity (SSIM) were used as similarity metrics exclusively on the portion of the volumes covered by the mask described in Section 5.3.3.

The baseline version of the model includes an autoencoder and the surrogate encoder without the rigid alignment module. This means that the model attempts to learn how to directly generate volumes by regressing the voxel intensities instead of deformations. To generate deformation fields instead of voxel intensities, the deformable registration network and the STN are added to the baseline. Next, the auxiliary encoder is introduced to assist the model during the decoding stage. Following that, the rigid alignment module from Section 5.3.2 is included upstream to the model, thereby completing all the model components. The proposed model was evaluated when using sagittal and axial orientations for the surrogate image.

Table 5.1 shows the results of the ablation study, evaluating each configuration based on the similarity metrics. It can be seen that the successive addition of each component allows to improve the output volumes across all similarity metrics. The large improvement from Baseline to Baseline + STN shows that the deformable motion model performs better when it is optimized to generate deformation fields instead of voxel intensities. The addition of the skip connections (extracted from ϕ_{ref}) further improves the output's quality by providing patient-specific information to the decoder. Finally, the addition of the rigidly aligned input gives an additional improvement to the appearance of the output volumes by reducing the amount of motion that needs to be represented by the autoencoder. This shifts the focus of the deformable motion model on more localized motion patterns. Results also show that

Tableau 5.1 Resulting image similarity metrics for different model configurations leading to the proposed model. Values are mean ± std.

Model	**MSE**	**NCC**	**SSIM**
Baseline	0.15 ± 0.04	0.42 ± 0.05	0.29 ± 0.05
Baseline + STN	0.10 ± 0.06	0.57 ± 0.10	0.54 ± 0.11
Baseline + STN + ϕ_{ref}	0.07 ± 0.04	0.62 ± 0.10	0.60 ± 0.10
Rigid	0.10 ± 0.06	0.61 ± 0.11	0.60 ± 0.12
Proposed (axi.)	0.07 ± 0.04	0.63 ± 0.10	0.61 ± 0.10
Proposed (sag.)	**0.06 ± 0.03**	**0.66 ± 0.09**	**0.65 ± 0.08**

the model performs better when the sagittal view images are used as surrogate ($\alpha < 0.01$, $\rho > 0.9$). Presumably, this is because the sagittal view covers a larger liver area than the axial view for sequences acquired under the sternum.

Rigid alignment module Our next experiment aimed at validating the robustness of the rigid alignment mechanism when the chosen pre-treatment volumes do not represent the full range of motion of the liver during intervention, measuring the robustness towards variation between baseline and online acquisitions. As stated in Section 5.3.2, it is assumed that the range of motion of the liver is bound by the acquired pre-treatment volumes V_{exh} and V_{inh}. Generally, the former can be chosen reliably since the exhale position is easily reproducible. On the other hand, ensuring that V_{inh} represents the deepest breathing amplitude for the entire sequence is not trivial. To measure the effect of incorrectly choosing V_{inh} on the performance of the rigid alignment module, we replaced the true V_{inh} by volumes that are adjacent to it in the temporal sequence. We quantified their distance (in mm) to the actual inhale position through rigid registration. Each inhale volume was used by the rigid alignment module to generate a set of rigid transformations covering one respiratory cycle for each case of the data set. Using the same approach as before, the displacement applied by the rigid transformations was computed. The resulting displacement values were then split into 3 respiratory phase groups (exhale, mid-cycle and inhale), each representing 1/3 of the respiratory cycle.

Table 5.2 shows the displacement introduced by the rigid alignment module at each phase as a function of the average V_{inh} selection error. It can be observed that the overall effect of increasing the selection error induces a decrease in the generated rigid motion amplitude. This effect is most prominent in the phases closest to inhale where the decrease in displacement is almost equal to the shift from the true inhale position. In contrast, volumes at exhale and mid-cycle phases are less affected by the selection error. In summary, the error in the choice

Tableau 5.2 Displacement (in mm) applied by the rigid alignment module in different respiratory phases with respect to the distance of the chosen inhale volume to the true inhale position. Values are mean ± std.

V_{inh} **selection error**	**Exhale**	**Mid-cycle**	**Inhale**	**Overall**
0.0	2.9 ± 0.9	7.0 ± 2.4	12.0 ± 1.3	7.2 ± 3.8
1.5 ± 0.2	2.8 ± 0.9	6.7 ± 2.3	10.5 ± 0.9	6.7 ± 3.3
3.4 ± 0.4	2.5 ± 0.8	6.1 ± 2.0	8.3 ± 0.7	5.8 ± 2.6
5.2 ± 0.7	2.3 ± 0.8	5.3 ± 1.5	6.4 ± 0.4	4.9 ± 2.0
7.4 ± 0.9	2.1 ± 0.7	4.5 ± 1.1	4.6 ± 0.2	4.0 ± 1.5

of either V_{exh} or V_{inh} has a direct effect on the maximum displacement yielded by the rigid module.

Auxiliary encoder During the motion generation stage, the motion decoder gets information from two sources ; the latent vector z and the skip connections from the auxiliary encoder. In order to better understand how the model uses both sources of information, the number of skip connections varied from 1 to 5, starting from the highest resolution layer and going towards the bottleneck of the autoencoder. In essence, allowing for more skip connections means that the model has more information or features from ϕ_{ref}. This can ultimately lead to ignoring completely the information contained in z. To detect when this occurs, the model's autoencoder was tested both with a learned z vector and with a randomly generated vector z_{rand} of the same size as z. Hence, for each configuration, we evaluate whether the information from the latent vector contributes to the model's performance. Figure 5.5 presents MSE values between ground-truth and predicted volumes for each model configuration. The dotted horizontal line indicates the similarity of the volume obtained by directly applying ϕ_{ref} to V_t^{rigid} without going through the autoencoder. It is noticeable that for models using 2 or more skip connections, the output's similarity when generated using either z_{rand} or z is practically the same. Indeed, all pairs of results for models with more than one skip connection are statistically the same. This allows us to identify the configuration with one skip connection, at the layer of highest resolution, as the optimal way to introduce patient-specific information from ϕ_{ref}. If more than one skip connection is used, the model tends to ignore the information contained in z and only focuses on the features carried by the skip connections. In that scenario, the model does not take into account any information about the current state of the organ given by I_t.

To further analyze the contribution of the single skip connection and vector z, the image similarity is evaluated at different portions of the output volume. The volumes were split into

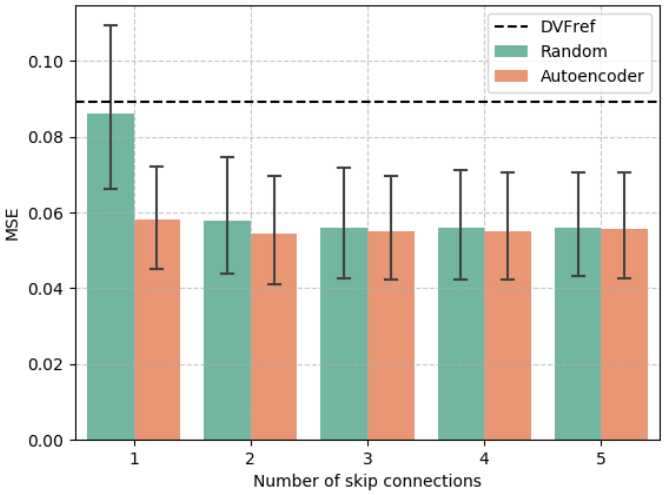


Figure 5.5 Motion autoencoder performance with learned and random latent vectors when varying the number of skip connections sent from the auxiliary encoder. The dotted horizontal line indicates the similarity of the volume obtained by directly applying ϕ_{ref} to V_t^{rigid} without going through the autoencoder.

5 sub-volumes along the right-left axis. The similarity was evaluated in 4 scenarios : after rigid alignment (V_t^{rigid}), after warping with ϕ_{ref} only, after warping with a DVF obtained using z_{rand} and after warping with the DVF obtained using the true z vector. Figure 5.6 shows the MSE between ground-truth and predicted sub-volumes across the different positions. The rigid input volumes V_t^{rigid} show a stable mean similarity across all positions within the complete volume. For volumes warped with ϕ_{ref}, the similarity is better at the center of the volume (i.e near the surrogate image position) and becomes increasingly worse as the sub-volume gets further from the center. This is expected as ϕ_{ref} is generated by registering a volume where I_t is replicated across all slices. Consequently, the most accurate registration is obtained at the center, which is the correct position for I_t. As we moving further away from the center, the less accurate the registration becomes. As for volumes warped with a DVF obtained using z_{rand}, a similar conclusion can be made from the skip connections experiment. When one skip connection is used, the model performs worse when provided with random information from the bottleneck. Overall, the best performance was shown by the proposed model that uses the true latent vector z, especially at the edges of the volumes.

Surrogate encoder Since US acquisitions are often performed using hand-held probes, there is a possibility the probe is not positioned at the same location at every fraction of the radiotherapy treatment. Therefore, it is necessary to evaluate the robustness of the model to

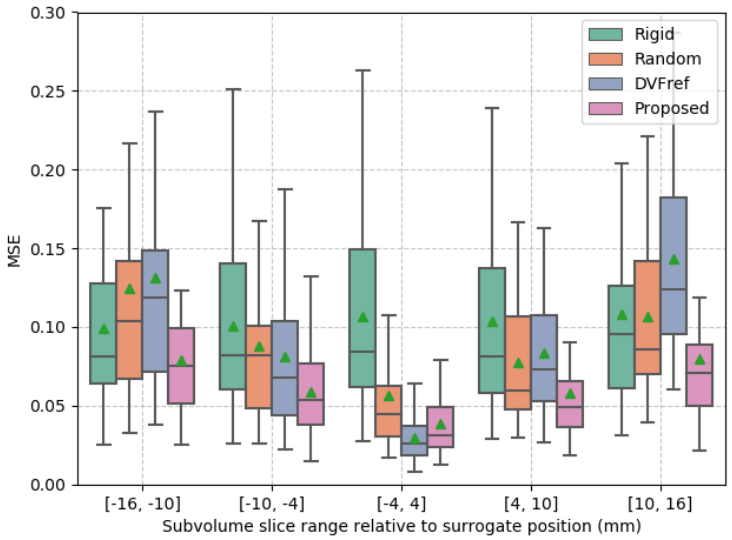


Figure 5.6 MSE value distributions between ground-truth and predicted sub-volumes along the right-left axis. Mean values are indicated by the green triangles.

potential shifts in the position of the 2D surrogate image. To do so, the deformable motion model portion of the framework was first trained to perform predictions based on input 2D images taken from the center of the ground-truth volumes. Then, during inference, the surrogate location was changed by taking slices from different positions along the right-left axis. We varied the shift in position from -15mm to 15mm with respect to the center, covering the entire volume. Figure 5.7 shows the NCC between ground-truth and generated volumes across the entire data set when varying the surrogate slice location. It can be observed that, as the shift in the position of the surrogate image increases, the image similarity decreases reaching its minimum when the surrogate is taken from the edges of the volume. The mean difference in NCC does not exceed 0.01 when the shift remain between -4 and 4mm. Therefore, the model can be considered capable to cope with slight changes to the position of the surrogate image.

Real-time application compatibility Finally, to assess the compatibility with real-time applications, the inference time of the proposed framework was evaluated. The total time to process the rigid and deformable steps of the framework was 0.47 $\pm$ 0.04 s when executed on CPU and 0.09 $\pm$ 0.01 s when executed on a NVIDIA Titan X GPU with 12 GB of RAM. This shows that the required time to generate motion predictions is sufficiently short to be included within a real-time radiotherapy workflow.

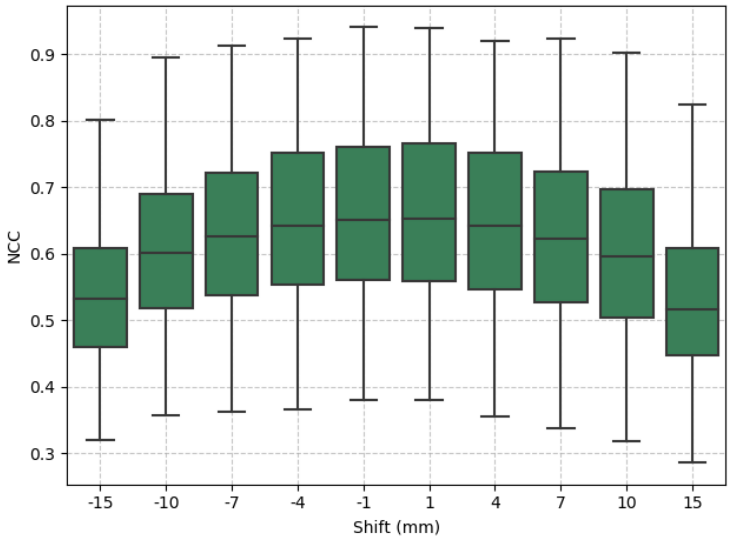


Figure 5.7 Image similarity for the entire data set when shifting the position of the surrogate image I_t from -15 mm to 15 mm.

5.4.3 Comparative results

The next set of experiments compared the performance of the proposed framework to related approaches for 3D motion modelling and target tracking in US. Namely, we compare the proposed approach to two other methods described by [73] and [115] in the context of image-guided radiation treatments. In the former case, two orthogonal 2D slices extracted from the reference volume (V_{ref}) are registered to the corresponding orthogonal 2D slices in the in-room volume (V_t). Subsequently, the partial 2D motion fields are combined to extrapolate the entire 3D motion. We will refer to this approach as motion extrapolation (ME). As for the model from [115], the approach consists of predicting ϕ_t by combining 3D features from V_{ref}

Tableau 5.3 Image similarity metrics between ground-truth and predicted volumes for different comparative methods. Values are mean ± std.

Model	**MSE**	**NCC**	**SSIM**
Unregistered	0.09 ± 0.06	0.59 ± 0.11	0.55 ± 0.13
Rigid	0.10 ± 0.06	0.61 ± 0.11	0.60 ± 0.12
ME [73]	0.21 ± 0.08	0.59 ± 0.08	0.53 ± 0.10
FC [115]	0.09 ± 0.04	0.57 ± 0.09	0.54 ± 0.10
FC + Rigid	0.08 ± 0.05	0.63 ± 0.10	**0.63 ± 0.10**
Proposed	**0.06 ± 0.03**	**0.66 ± 0.09**	**0.65 ± 0.08**

Tableau 5.4 3D tracking performance (in mm) of the compared approaches based on local TRE at different phases. Values are mean ± std. ($\mu \pm \sigma$) and 95^{th} percentile (P95).

Model	**Exhale**		**Mid-cycle**		**Inhale**		**Overall**
	$\boldsymbol{\mu \pm \sigma}$	$\boldsymbol{P_{95}}$	$\boldsymbol{\mu \pm \sigma}$	$\boldsymbol{P_{95}}$	$\boldsymbol{\mu \pm \sigma}$	$\boldsymbol{P_{95}}$	
Unregistered	—	—	9.8 ± 8.2	20.2	18.0 ± 13.4	31.4	10.7 ± 9.7
Rigid	3.5 ± 1.3	7.6	3.9 ± 1.7	6.9	6.3 ± 4.3	12.6	4.6 ± 3.2
ME [73]	**2.7 ± 1.4**	**6.1**	5.9 ± 2.8	13.3	10.9 ± 7.9	23.5	6.5 ± 6.4
FC [115]	5.0 ± 3.3	9.7	7.9 ± 4.3	15.4	13.8 ± 10.7	27.5	8.9 ± 7.5
FC + Rigid	**3.1 ± 0.5**	**6.8**	4.5 ± 2.2	6.5	7.2 ± 4.4	10.8	4.9 ± 3.9
Proposed	**2.8 ± 1.6**	**5.6**	**3.2 ± 0.8**	**5.1**	**4.5 ± 2.5**	**9.5**	**3.5 ± 2.4**

and 2D features from I_t. The model is comprised of a 2D encoder for I_t, a 3D encoder for V_{ref} and a 3D decoder coupled with a STN to generate ϕ_t and apply it to V_{ref}. We will refer to this approach as feature combination (FC). All three approaches (ME, FC and the proposed framework) aim to generate the motion field corresponding to the respiratory state indicated by the surrogate 2D information. We first compare their performances based on the similarity metrics used in Section 5.4.2. We also compute the global and local target registration error (TRE) using 3D deformable image registration (DIR) between ground-truth and predicted volumes, and manual landmark annotations, respectively.

Image similarity Table 5.3 shows the similarity metrics for the different compared methodologies. As a reference, in the first row of the table, we report the result when there is no motion compensation (Unregistered). The second row represents the values measured when only the rigid alignment is applied on the reference volume. Overall the proposed approach showed the best performance for all metrics except for SSIM where it was statistically equivalent to FC with rigid pre-alignment ($\alpha = 0.4$, $\rho = 0.38$). Although the rigid pre-alignment was designed to be used in the proposed model, it was able to significantly improve the results for the FC model ($\alpha < 0.01, \rho > 0.9$), showing its usability as an independent rigid alignment module. The worst similarity results were obtained by ME, which was designed for local modeling. In consequence, there is a poor overall similarity between ground-truth and predicted volumes.

Target tracking Table 5.4 compares the methods based on local TRE for different respiratory phases and for the entire respiratory cycle overall. The results were obtained by manually tracking each of the identified landmarks and averaging the difference between the ground-truth and predicted landmark positions. The results obtained on the reference volume were excluded whenever it was part of the analyzed respiratory cycle. It can be observed that

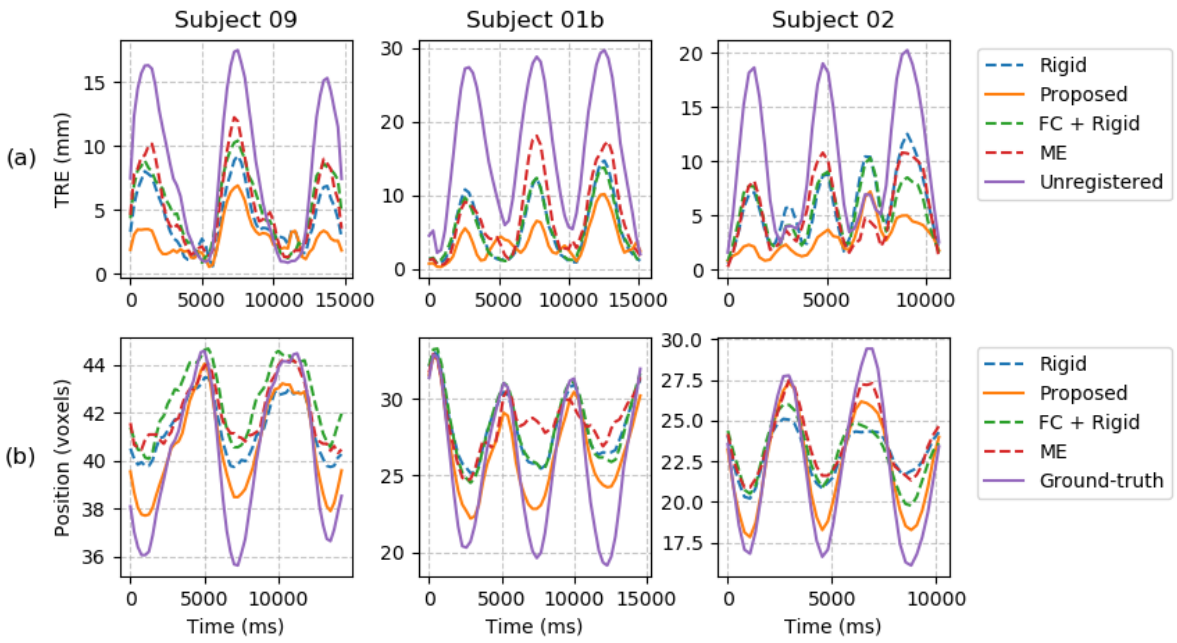


Figure 5.8 (a) Evolution of TRE through time and (b) target trajectories for 3 cases. Landmarks were tracked for all 3 acquired breathing cycles.

all models were able to improve over the unregistered volumes for all respiratory phases. Moreover, the errors are larger for phases that are temporally further away from the reference respiratory phase. The largest improvement in TRE came from the rigid alignment of the volumes. This is expected as this step aims to represent the general motion of the organ, which includes the largest displacement. Further improvements in the predicted landmark positions result from the proposed deformable modelling. Three models (ME, FC + rigid and proposed) performed similarly at exhale, however as phases get closer to the inhale phase, the proposed approach shows significantly lower errors achieving the best local TRE result overall.

To better visualize the tracking performance of each method, Figure 5.8 shows how the tracking error as well as the vessel trajectory evolve over time for each model during 3 respiratory cycles for 3 subjects within the 4D US data set. As previously observed in Table 5.4, the largest tracking errors occur at the inhale respiratory phase. Figure 5.8a shows that the proposed framework is able to maintain the lowest error throughout all respiratory cycles compared to other models. The plots in Figure 5.8b show that the proposed model is able to follow the ground-truth trajectory better than the comparative approaches.

In addition to local TRE, the global displacement error of the proposed framework was evaluated by applying 3D DIR between the ground-truth and generated volumes. By converting

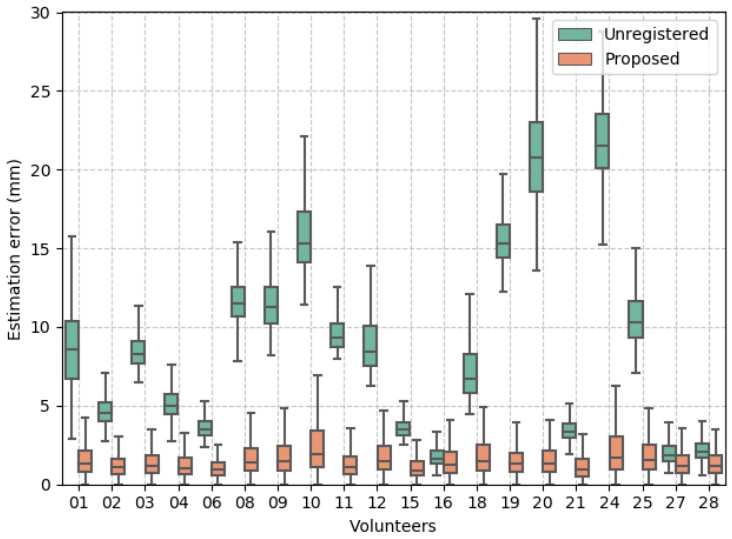


Figure 5.9 Estimation errors for each test case, calculated by 3D DIR for unregistered volumes and volumes generated by the proposed model.

the obtained displacement fields to displacement magnitudes and then averaging over the entire volume, we obtain the average estimation error over all voxels in the generated volume. The same procedure was applied to the unregistered volumes, however the displacement from the rigid transformation calculated by the rigid module was also taken into account. Figure 5.9 shows the calculated global estimation error distribution for each case in the US dataset. By observing the different value distributions of the unregistered volumes, it is noticeable that the data set presents a wide variety of motion amplitudes. For all cases, the proposed model is able to reduce the global estimation error. The mean global error was reduced from 8.7 mm (Unregistered) to 1.7 mm with the proposed solution.

Deformation quality When generating motion from volumetric images, deformation fields are expected to be diffeomorphic to ensure physically plausible displacements. We evaluated the smoothness of the deformation fields produced by our deformable motion model by calculating the Jacobian matrix determinant ($|J|$) over the entire motion field. The average $|J|$ for the entire dataset was 0.97 ± 0.43 with only 1.1% of negative values, indicating that the model produces smooth and plausible deformations with very few foldings.

Qualitative results Figure 5.10 illustrates the generated and ground-truth volumes at three respiratory positions (mid-inhale, inhale and mid-exhale) along two imaging planes (sagittal and axial) for one example case. Difference maps with respect to the ground-truth at the inhale phase are presented as well. A general observation is that the proposed framework is able to infer motion outside of the surrogate plane since motion is visible in both perpen-

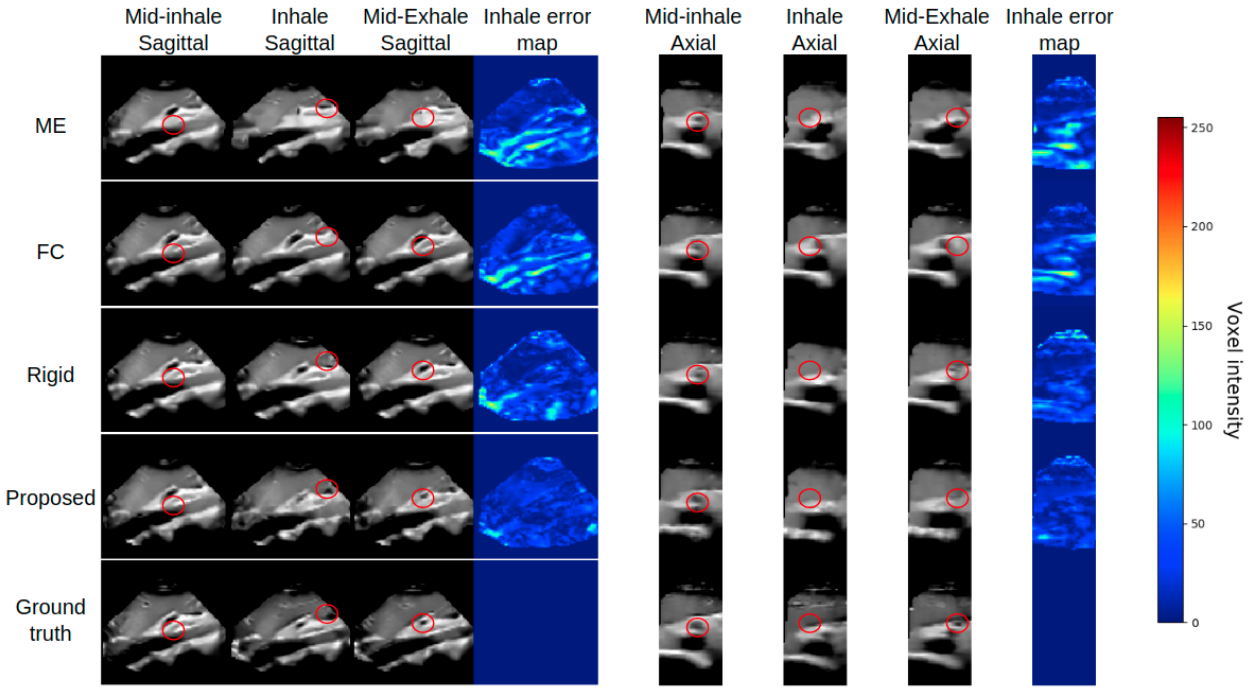


Figure 5.10 Qualitative results for all compared methods. For both sagittal and axial planes, the central slice of the volume is shown at mid-inhale, inhale and mid-exhale respiratory phases. For the inhale phase, an error map is calculated and shown. Red circles are included to highlight differences between the displayed approaches.

dicular planes. Furthermore, the difference maps in both planes showcase that the proposed approach generates the lowest voxel intensity errors at inhale. Also, the proposed framework reproduces small features like vessels and liver borders better than the ME and FC approaches as highlighted by the red circles. It is also noticeable that the application of the deformable motion component over the output of the rigid alignment module improves local correspondences with the ground truth volume, highlighting the importance of the second step of the proposed framework.

Finally, Figure 5.11 presents generated and ground truth slices at 5 different phases between exhale and inhale. To display the true and generated deformation fields, green and yellow arrows were overlaid on the generated volume slices. Small sections of the deformation fields were increased in size for better visualisation. At the reference phase, the generated deformation field is essentially null. As the the phases get closer to inhale, the amplitude of motion applied to the reference volume is gradually increased. It is noticeable that for the majority of positions, the generated motion field follows the expected motion field well. In general the motion is oriented in the inferior direction as expected during inhalation. More localized motion patterns, representing the deformable components of motion, are also present. They

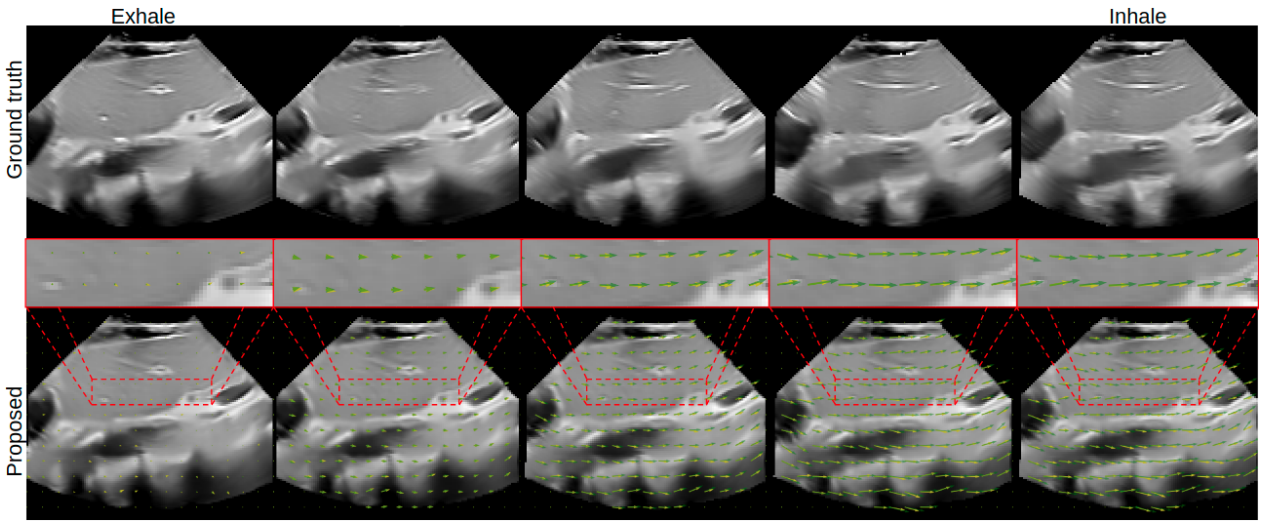


Figure 5.11 Qualitative results from exhale to inhale phases with overlaid ground-truth (green) and predicted (yellow) displacement fields.

can be seen at the left of the images where the heart is visible, as well as in the bottom section of the images. Additional qualitative results can be found in the supplementary materials.

5.5 Discussion

This work presented a novel 3D motion modelling framework that includes advantages from both population-based and patient-specific models for US-guided radiotherapy procedures. Being based on a deep learning model allows reducing the amount of manual preparation and data manipulation required to construct the motion model while also allowing for real-time inference capabilities. The proposed model has demonstrated promising results for image similarity metrics and target tracking when compared to both traditional and deep learning based approaches.

The ablation study has revealed that the inclusion of features through skip connections was the component with greater contribution to the deformable motion modelling component of the framework. By skipping relevant features to the decoder, the model leverages patient-specific information as it attempts to generate a deformation field for an anatomy it has not seen during training. A limitation that is often attributed to population-based models is that by fitting them to a large amount of anatomies, they ultimately learn to represent an average motion field without being able to properly model the motion of each individual subject [53]. However, our model is able to avoid this by leveraging the patient-specific features it receives during the motion field generation. Furthermore, the experiments showed the importance of

controlling the amount of patient-specific information that is provided to the model. If too many features from ϕ_{ref} are skipped through, the latent representation of the model collapses and no longer captures meaningful information about the current state of the organ. In this case, the model would merely learn to refine ϕ_{ref}. Therefore, the combination of both skip connections with the latent vector z must be optimized to use both sources of information.

When comparing our framework to statistical global model solutions such as in [54], our framework presents improvements over the motion model construction step. Indeed, when constructing a population model based on PCA, an inevitable step is the establishment of inter-subject correspondences. Usually, this process is done manually or semi-automatically which adds a significant amount of time to the data preparation and model construction. In addition, this limits the type of data that the model can operate on. If the training data doesn't include the inter-subject correspondences established during the model construction, the model's performance will decrease. By employing a deep learning framework for the population based motion model construction, we allow the network to learn those inter-subject correspondences implicitly. It is assumed however that during inference, the provided inputs show the same field-of-view as the ones used during training. However, fine-tuning the model to a new anatomy remains simpler when using a deep network instead of a global statistical model. A final advantage our approach presents over [54] is the fact that the model analyzes the entire surrogate image as it comes from the acquisition system. This means true 2D surrogate signals are used instead of tracking a fiducial marker within the surrogate images to drive the motion model.

During treatment planning, radio-oncologists add a margin of at least 5mm around the region to be treated to account for respiratory motion. This is done even in the presence of breath hold techniques [24]. Since our model achieved an average TRE of 3.5mm it could allow to reduce the extent of the added margins, thereby sparing healthy tissues from an unnecessary radiation dose. Although the main focus of this paper is the tracking capabilities of our network, its use is not limited to 3D target tracking. As shown by the image similarity and global TRE experiments, our model can also estimate the global and deformable motions of the organ along all the slices in the volume. This information can be useful for radiotherapy applications, in particular for dose delivery estimation. In addition, by requiring a surrogate image in only one imaging plane, the surrogate acquisition time is reduced without sacrificing the capability of the model to predict the motion outside of that plane.

The local TRE experiment has shown the flexibility of the proposed rigid alignment module as it has improved the performance of 2 different deep learning-based models by providing them with rigidly aligned input volumes. However, this approach is not exempt of limitations. As

explained in Section 5.3.2, it is assumed that the state of the liver is bound between the exhale and inhale positions acquired before treatment. In the event where the liver exceeds those bounds, σ no longer describes the position of the liver relative to the pair of pre-treatment volumes, thereby providing less accurate alignment between the central slice of the reference volume and the surrogate image. The first experiment of Section 5.4.2 has demonstrated that an error in one of the pre-treatment volumes will only affect the accurate alignment of volumes close to the faulty pre-treatment volume. Therefore it is important to insure that the pair of pre-treatment volumes accurately present the full range of motion of the liver. It is also recommended that the pre-treatment volumes are acquired before each treatment to ensure that changes in anatomy over the course of treatment do not negatively affect the rigid alignment.

An important feature that motion models need to have is the ability to predict and anticipate the motion the target will experience in real time. This is necessary because the adjustment of the treatment plan and delivery to a new position of the target isn't instantaneous and bears a latency that cannot be ruled out [7]. Several works on motion modelling have presented ways to include motion prediction within their framework [54, 65, 93]. While in this work the model doesn't present motion prediction capabilities, the framework is capable of including a temporal prediction module. Specifically, the surrogate branch of the motion model can learn to predict the future latent representation of the organ, thus generating the future anticipated motion field. While this addition is crucial for the applicability in a clinical setting, it is out of the scope of this work and needs to be validated in future studies.

Another limitation common to several motion models which hinders the transfer of those approaches to the treatment room, is the amount of subjects used for validation. In this work, the data acquired from the 20 subjects presented a good variability in anatomical appearance. However, the acquisition time for each sequence (15 seconds) has limited the amount of breathing variability that was captured. Also, long-term effects such as exhale drift [119] could not be taken into account either. Since the 4D dataset was acquired on healthy subjects only, the proposed solution was not evaluated on liver cancer patients undergoing radiotherapy treatment. As those cases can present higher variability in anatomical appearance and breathing patterns, due to the presence of tumors or other pathologies, the robustness of our framework needs to be validated on this type of data in future studies. Moreover, as explained in Section 5.3.1, this study assumes that the reference volumes used in our experiments are directly taken from the acquired 4D sequences. The surrogate 2D images are also assumed to be the central slices of the volumes within the 4D sequences. In a clinical setting, this wouldn't be the case as there would be no prior 4D acquisition. However, we do not believe the framework's performance will be affected as long as the V_{ref} and surrogate image show

the same anatomical location and field-of-view. This aspect would need to be validated in future experiments.

Future studies will include the addition of a temporal prediction mechanism, thus increasing the horizon for temporal sequences, with the evaluation on longer sequences, application for different imaging modalities as well as general improvements to individual components such as the rigid alignment module and motion modelling network. A prospective study with radiotherapy patients is planned to further evaluate in a clinical context.

CHAPITRE 6 DISCUSSION GÉNÉRALE

Les chapitres précédents ont présenté les deux solutions développées au cours de ce projet de maîtrise dans le but de répondre à la problématique identifiée au chapitre 3. Les deux articles montrent le cheminement effectué afin de développer un modèle permettant de générer des volumes d'US 3D mis à jour à partir d'images d'US 2D et d'un nombre limité de volumes de prétraitement. Dans le présent chapitre, il s'agira de discuter de ce cheminement et de souligner les leçons retenues. Les limitations de la méthodologie proposée ainsi que les perspectives futures de développement seront également présentées avant de conclure ce mémoire au chapitre suivant.

Le modèle initial présenté dans l'article publié à ISBI a permis d'évaluer la capacité des CNN à inférer un champ de déformation 3D à partir d'information en 3D et 2D. Dans ce modèle, le volume de référence servait comme élément de conditionnement par rapport à l'anatomie 3D du patient pour lequel le champ de déformation est généré. Le rôle des images 2D était de donner de l'information quant à la déviation par rapport à la référence. Le réseau avait donc la tâche d'estimer le mouvement qui s'est produit entre le volume 3D et l'image 2D et de générer le champ de mouvement correspondant. Avec rétrospective, cette approche semble ambitieuse. On se fit sur la grande capacité de représentation du réseau pour apprendre à générer des champs de déformation en 3D, sans que le modèle ait appris ce que constitue un champ de déformation 3D valide. Il doit l'apprendre implicitement par le biais de la fonction de similarité entre le volume déformé et le volume attendu. En théorie, cette approche est peut-être valable, mais étant donné la quantité limitée de données et leur inhomogénéité, la capacité de généralisation du réseau est limitée. Par conséquent, le modèle peine à bien déformer les volumes de référence pour une anatomie qui n'est pas incluse dans l'ensemble de données d'entraînement.

Un ajustement dans la méthodologie était nécessaire. La première modification était d'effectuer la génération du champ de mouvement de manière séparée afin d'alléger la tâche du réseau qui modélise le mouvement. Vu l'existence d'un grand éventail de modèles de recalage 3D basés sur l'apprentissage profond, un modèle développé par [92] a été choisi. Suite à l'entraînement de ce modèle, il était possible d'obtenir le champ de mouvement entre deux volumes d'US de manière très rapide. Ce modèle a donc pu être intégré en amont à la composante de modélisation du mouvement.

Suite à cet ajout, il fallait trouver un moyen d'établir une correspondance entre les images 2D et le champ de mouvement 3D. En s'inspirant des AE et leurs variations telles que les VAE

et CVAE, il semblait que la correspondance pourrait se faire par le biais de la représentation cachée z de l'AE. Ceci est une approche également utilisée par les modèles basés sur la PCA où seuls les coefficients de combinaisons doivent être changés pour obtenir un champ de déformation différent. C'est le même raisonnement pour les AE où il est possible de changer le vecteur z pour obtenir le champ de déformation désiré. De là est arrivée l'idée de l'entraînement par trois étapes présentée à la section 5.3.2.

Cette nouvelle approche d'entraînement semblait fonctionner. Toutefois, l'évaluation de ces modèles a révélé une grande dégradation de la performance de reconstruction pour l'ensemble de tests, montrant une généralisation sous-optimale de la part du réseau. C'est d'ailleurs une des critiques principales des modèles de population. En tenant compte de plusieurs anatomies, ils sont vus comme des modèles qui apprennent des patrons de mouvement moyens qui sont peu précis pour de nouvelles anatomies. En s'inspirant des "skip connections" utilisées dans notre premier article, une approche similaire a été employée lors de l'étape du décodage de l'AE. L'ajout des "skip connections" a joué un rôle important dans l'amélioration de la performance du réseau sur les cas exclus de l'ensemble d'entraînement. L'ajout de cette composante au modèle de mouvement a permis de joindre les avantages des modèles à patient unique à ceux des modèles de population.

Les évaluations subséquentes ont permis de déterminer un autre aspect à améliorer, la performance du modèle pour les phases près de l'inhalation. En visualisant les volumes générés par le modèle il était clair que la qualité de génération se dégradait près de la phase d'inhalation complète. De plus, les champs de mouvement générés par le modèle ne permettaient pas d'atteindre une amplitude du mouvement suffisante pour représenter le foie à cette phase respiratoire. Après investigation, il s'est avéré que la dégradation du volume était due à la qualité des champs de mouvement produit par le réseau de recalage. Puisque l'amplitude du mouvement du foie était si grande, la génération d'un champ de déformation lisse était difficile pour le modèle. La manière choisie pour remédier à ce problème était de séparer les composantes rigides et déformables du mouvement. Le patron de mouvement général de l'organe pouvait être représenté par une transformation rigide et puis les déformations plus localisées pouvaient être prises en compte par le modèle de recalage 3D. C'est ainsi que le module d'alignement rigide a été ajouté au modèle d'estimation du mouvement. Il a permis d'améliorer la qualité des volumes générés en plus d'aider à réduire l'erreur de suivi de cibles. De plus, ce module est très flexible et peut être jumelé avec d'autres modèles d'estimation du mouvement.

En comparaison aux approches utilisant des modèles statistiques comme la PCA, la solution proposée présente plusieurs avantages. Premièrement, nous avons démontré que le modèle

est capable d'inférer le mouvement pour des cas non inclus dans l'ensemble d'entraînement. Dans le cas des modèles à patient unique, de nouvelles données sont requises pour adapter le modèle à un nouveau patient. Deuxièmement, la préparation des données pour l'entraînement n'a requis aucune annotation manuelle de marqueurs ou de segmentation de surfaces communes à tous les volontaires. Ceci est au contraire des modèles de population qui requièrent l'établissement de correspondances entre tous les sujets composant l'ensemble des données. En utilisant l'apprentissage profond, le modèle établit par lui-même ces correspondances sans intervention humaine, ce qui sauve du temps lors de sa construction. Poursuivant sur ce point, la préparation de données minimale requises par les modèles d'apprentissage profond permet de les adapter à de nouveaux sites anatomiques plus facilement que les modèles statistiques. Considérant la multitude de sites anatomiques où la RTGI est utilisée, ceci est un avantage important d'un point de vue clinique.

Enfin, bien que la modalité d'imagerie au centre de ce mémoire est l'US, la solution développée n'est pas limitée à ce type d'images. En théorie, toutes les composantes du modèle permettent d'utiliser celui-ci pour des images d'autres modalités ou même deux modalités différentes comme l'US et l'IRM par exemple. L'unique condition est que les images utilisées comme signal substitut soient représentatives du mouvement observé dans les volumes 4D. Il faut seulement changer les données d'entraînement pour optimiser le réseau aux nouvelles modalités d'imagerie. De plus, le modèle proposé peut potentiellement être utilisé pour d'autres applications cliniques ablatives comme le TACE ou bien l'échographie focalisée à haute intensité (HIFU). En fait, toute intervention nécessitant la localisation de cible ou d'organes dans un espace 3D en temps réel peut bénéficier d'un modèle de mouvement comme celui proposé dans ce mémoire.

Ces affirmations devront être validées expérimentalement dans des études subséquentes.

6.1 Limitations

La première limitation de ce travail est la taille de l'ensemble de données utilisé pour l'entraînement et l'évaluation de notre méthode. C'est le cas de bien de modèles qui tentent d'appliquer les notions d'apprentissage profond aux données médicales. Puisque la performance des réseaux de neurones est intimement liée à la quantité de données sur lesquelles ils sont entraînés, le transfert de modèles opérant sur des images naturelles vers les images médicales est complexe et demande des adaptations méthodologiques importantes. L'autre conséquence d'un jeu de données limité est qu'il est difficile de quantifier la performance du réseau sur des données se trouvant en dehors de l'ensemble de données utilisé. Certes, si un plus grand ensemble de données est disponible le modèle deviendra, en théorie, plus général

et dépendra moins de l'ensemble de données utilisé. Cette affirmation reste tout de même théorique et doit être démontrée expérimentalement pour la tâche étudiée dans le cadre de ce travail. Enfin, la longueur de 15 secondes des acquisitions constituant l'ensemble de données n'a pas permis d'évaluer des effets de mouvement se manifestant sur une fenêtre de temps plus long.

Une seconde limitation du modèle proposé est l'absence de mécanisme pour la prédiction du mouvement de l'organe. Les techniques de RTGI doivent en premier lieu permettre de suivre la cible de traitement à travers le temps afin de connaître son emplacement. Mais une fonctionnalité essentielle à l'intégration des techniques de RTGI dans la pratique clinique est la capacité de prédire le mouvement que subira la tumeur. Ceci est nécessaire puisque le temps d'ajustement de la trajectoire du faisceau d'irradiation n'est pas négligeable. Le modèle présenté ne permet pas d'accomplir ceci dans sa forme actuelle. Toutefois, l'architecture choisie pour le modèle peut permettre l'ajout de cette fonctionnalité, mais ceci sort du cadre de ce projet.

Une dernière limitation concerne les champs de mouvements utilisés pour l'entraînement du modèle. Cette limitation affecte les modèles de mouvement en général puisque la manière d'obtenir les véritables champs de mouvement d'un organe n'est pas définie de manière formelle. Par conséquent, la performance du modèle d'estimation du mouvement est limitée par la qualité des champs de déformations utilisés lors de la formation du modèle. Dans la plupart des cas, une combinaison de recalages rigides et déformables est utilisée. Ceci permet d'avoir une bonne estimation du mouvement observé, mais demeure une estimation.

6.2 Pertinence clinique

Le but ultime de tout projet de nature biomédical est le transfert de la technologie développée en clinique. En tenant compte des éléments discutés ci-dessus, la question suivante se pose : «Où se trouve la solution proposée dans ce mémoire par rapport à l'application clinique visée ?»

La revue de littérature a bien démontré les avantages de l'imagerie par US par rapport à des modalités d'imagerie ionisante ou difficilement accessible. Par contre, très peu de solutions commerciales et de travaux de recherche se penchent sur l'utilisation de l'US pour la RTGI à l'aide de notions innovatrices comme l'apprentissage profond. Il faut donc souligner que ce travail de mémoire peut servir de catalyseur pour d'autres travaux de recherche et développement pour l'application de l'imagerie par US dans le cadre de la RTE.

Par contre, l'état actuel de ce projet est bien loin de l'application clinique convoitée. Princi-

palement dû aux limitations évoquées ci-dessus, mais également pour des raisons techniques qui sortent du cadre de ce projet de maîtrise. Ces raisons techniques incluent entre autres la nécessité de développer un système mécanique pour maintenir la sonde d'US en place lors du traitement puisqu'aucun membre du personnel ne peut la tenir au cours de la procédure. De plus, le placement de la sonde peut potentiellement affecter la livraison du traitement en limitant les trajectoires d'irradiation disponibles ou en risquant une collision avec des systèmes LINAC montés sur bras robotique. Enfin, les coordonnées de la cible anatomique identifiée sur les volumes d'US doivent être interprétées par le système de traitement. Il faut donc développer une procédure de calibration robuste pour permettre au système d'ajuster la livraison du traitement de manière précise. Tous ces aspects techniques et bien d'autres devront être adressés par des projets futurs dans le but d'amener cette technologie vers les cliniques où elles pourront contribuer à améliorer l'efficacité des traitements par RTE.

CHAPITRE 7 CONCLUSION

Ce mémoire s'est amorcé avec une présentation du contexte au sein duquel la problématique centrale de ce projet de maîtrise s'inscrit. À travers d'une revue de littérature détaillée, nous avons exposé les manquements des méthodes courantes pour le suivi de cibles au cours du traitement du cancer par RTGI. Cette revue a également permis d'explorer de nouvelles pistes de solution, notamment les approches par apprentissage profond pour la modélisation du mouvement. Ainsi, l'objectif principal de ce projet a pu être formulé, soit la génération de volume d'US 3D à partir d'image US 2D en respectant des contraintes quant aux acquisitions de prétraitement et de performance en temps réel.

Nous avons exploré une diversité d'architectures de modèles de mouvement basés sur l'apprentissage profond qui ont mené à la rédaction de deux articles présentant le cheminement effectué. Nous avons basé notre méthode sur l'architecture des AE convolutifs due à leur capacité de représenter des données complexes de manière compacte. Ceci a permis de faire correspondre des entrées en 2D avec les sorties souhaitées en 3D. Nous avons bonifié l'architecture initiale avec l'ajout d'information anatomique spécifique au patient, ainsi que l'ajout d'un module d'alignement rigide. Une validation sur un ensemble de données de 20 volontaires en santé a permis d'évaluer la précision de notre méthode à 3.5 ± 2.4 mm pour le suivi de cibles anatomique. Ceci démontre le fort potentiel de cette approche pour la modélisation du mouvement sur des séquences d'US 3D à partir d'images d'US 2D.

Toutefois, notre méthode n'est pas sans faille et comporte plusieurs limitations. En particulier, l'absence d'un module pour prédire le mouvement de la tumeur afin de permettre d'ajuster le faisceau d'irradiation à temps. De plus, la non-exhaustivité de l'ensemble de données et l'approximation des champs de mouvements par recalage font en sorte que cette approche demeure une preuve de concept plutôt qu'une solution clinique prête à être déployée.

Enfin, nous espérons que des avancées technologiques futures dans le domaine de l'apprentissage profond et de l'imagerie par US permettront de résoudre les défis d'intégration clinique de solutions de RTGI basées sur l'imagerie par US.

RÉFÉRENCES

[1] Organisation mondiale de la santé. (2021) Cancer. [En ligne]. Disponible : https://www.who.int/news-room/fact-sheets/detail/cancer

[2] O. Ciacio et D. Castaing. (2015) Le foie et les voies biliaires : Anatomie. [En ligne]. Disponible : https://www.centre-hepato-biliaire.org/maladies-foie/anatomie-foie.html

[3] Société canadienne du cancer. Le foie. [En ligne]. Disponible : https://www.cancer.ca:443/fr-ca/cancer-information/cancer-type/liver/liver-cancer/the-liver/?region=on

[4] C. Mony et J.-C. Duclos-Vallée. (2014) Les fonctions du foie. [En ligne]. Disponible : https://www.centre-hepato-biliaire.org/maladies-foie/anatomie-foie.html

[5] H. Shirato, Y. Seppenwoolde, K. Kitamura, R. Onimura et S. Shimizu, "Intrafractional tumor motion : lung and liver," *Seminars in Radiation Oncology*, vol. 14, n°. 1, p. 10–18, 2004, high-Precision Radiation Therapy of Moving Targets. [En ligne]. Disponible : https://www.sciencedirect.com/science/article/pii/S1053429603000894

[6] F. Preiswerk, "Modelling and reconstructing the respiratory motion of the liver," Thèse de doctorat, University of Basel, 2013.

[7] P. J. Keall, G. S. Mageras, J. M. Balter, R. S. Emery, K. M. Forster, S. B. Jiang, J. M. Kapatoes, D. A. Low, M. J. Murphy, B. R. Murray, C. R. Ramsey, M. B. Van Herk, S. S. Vedam, J. W. Wong et E. Yorke, "The management of respiratory motion in radiation oncology report of aapm task group 76a)," *Medical Physics*, vol. 33, n°. 10, p. 3874–3900, 2006.

[8] Y.-L. Tsai, C.-J. Wu, S. Shaw, P.-C. Yu, H.-H. Nien et L. T. Lui, "Quantitative analysis of respiration-induced motion of each liver segment with helical computed tomography and 4-dimensional computed tomography," *Radiation oncology (London, England)*, vol. 13, n°. 1, p. 59, 2018.

[9] F. Bray, J. Ferlay, I. Soerjomataram, R. L. Siegel, L. A. Torre et A. Jemal, "Global cancer statistics 2018 : Globocan estimates of incidence and mortality worldwide for 36 cancers in 185 countries," *CA : a cancer journal for clinicians*, vol. 68, n°. 6, p. 394–424, 2018.

[10] P. Dasgupta, C. Henshaw, D. R. Youlden, P. J. Clark, J. F. Aitken et P. D. Baade, "Global trends in incidence rates of primary adult liver cancers : A systematic review and meta-analysis," *Frontiers in oncology*, vol. 10, p. 171, 2020.

[11] C.-Y. Liu, K.-F. Chen et P.-J. Chen, "Treatment of liver cancer," *Cold Spring Harbor perspectives in medicine*, vol. 5, n°. 9, p. a021535, 2015.

[12] N. H. Andratschke, C. Nieder, F. Heppt, M. Molls et F. Zimmermann, "Stereotactic radiation therapy for liver metastases : factors affecting local control and survival," *Radiation oncology (London, England)*, vol. 10, p. 69, 2015.

[13] Société canadienne du cancer. Qu'est-ce que le cancer métastatique ? [En ligne]. Disponible : https://www.cancer.ca/fr-ca/cancer-information/cancer-type/metastatic-cancer/metastatic-cancer/?region=qc

[14] R. M. Goldstein, B. D. Berger et J. K. O'Connor, *Multidisciplinary Overview of Local-Regional Therapies for Liver Malignancies*. Berlin, Heidelberg : Springer Berlin Heidelberg, 2007, p. 205–215. [En ligne]. Disponible : https://doi.org/10.1007/978-3-540-69886-9_21

[15] M. Scorsetti, E. Clerici et T. Comito, "Stereotactic body radiation therapy for liver metastases," *Journal of gastrointestinal oncology*, vol. 5, n°. 3, p. 190–7, 2014.

[16] A. Takeda, N. Sanuki, T. Eriguchi, T. Kobayashi, S. Iwabutchi, K. Matsunaga, T. Mizuno, K. Yashiro, S. Nisimura et E. Kunieda, "Stereotactic ablative body radiotherapy for previously untreated solitary hepatocellular carcinoma," *Journal of Gastroenterology and Hepatology*, vol. 29, n°. 2, p. 372–379, 2014. [En ligne]. Disponible : https://onlinelibrary.wiley.com/doi/abs/10.1111/jgh.12350

[17] L. Potters, B. Kavanagh, J. M. Galvin, J. M. Hevezi, N. A. Janjan, D. A. Larson, M. P. Mehta, S. Ryu, M. Steinberg, R. Timmerman, J. S. Welsh et S. A. Rosenthal, "American society for therapeutic radiology and oncology (astro) and american college of radiology (acr) practice guideline for the performance of stereotactic body radiation therapy," *International Journal of Radiation Oncology*Biology*Physics*, vol. 76, n°. 2, p. 326–332, 2010. [En ligne]. Disponible : https://www.sciencedirect.com/science/article/pii/S0360301609033045

[18] E.-L. Dormand, P. E. Banwell et T. E. Goodacre, "Radiotherapy and wound healing," *International Wound Journal*, vol. 2, n°. 2, p. 112–127, 2005.

[19] Société canadienne du cancer. Radiothérapie externe. [En ligne]. Disponible : https://www.cancer.ca/fr-ca/cancer-information/diagnosis-and-treatment/radiation-therapy/external-radiation-therapy/?region=on

[20] E. B. Podgorsak *et al.*, "Treatment machines for external beam radiotherapy," dans *Radiation Oncology Physics*. INTERNATIONAL ATOMIC ENERGY AGENCY, 2005, p. 123–160.

[21] F. Khan et J. Gibbons, *Khan's the Physics of Radiation Therapy*. Lippincott Williams & Wilkins, 2014. [En ligne]. Disponible : https://books.google.ca/books?id=cvtPBAAAQBAJ

[22] W. Parker et H. Patrocinio, “Clinical treatment planning in external photon beam radiotherapy,” dans *Radiation Oncology Physics*. INTERNATIONAL ATOMIC ENERGY AGENCY, 2005, p. 219–272.

[23] Y. Negoro, Y. Nagata, T. Aoki, T. Mizowaki, N. Araki, K. Takayama, M. Kokubo, S. Yano, S. Koga, K. Sasai *et al.*, “The effectiveness of an immobilization device in conformal radiotherapy for lung tumor : reduction of respiratory tumor movement and evaluation of the daily setup accuracy,” *International Journal of Radiation Oncology* Biology* Physics*, vol. 50, n°. 4, p. 889–898, 2001.

[24] K. K. Brock, “Imaging and image-guided radiation therapy in liver cancer,” *Seminars in Radiation Oncology*, vol. 21, n°. 4, p. 247–255, 2011, radiation Therapy of Primary and Metastatic Liver Tumors. [En ligne]. Disponible : https://www.sciencedirect.com/science/article/pii/S1053429611000439

[25] J. B. West, J. Park, J. R. Dooley et C. R. Maurer, *4D Treatment Optimization and Planning for Radiosurgery with Respiratory Motion Tracking*. Berlin, Heidelberg : Springer Berlin Heidelberg, 2007, p. 249–264. [En ligne]. Disponible : https://doi.org/10.1007/978-3-540-69886-9_25

[26] C. Western, D. Hristov et J. Schlosser, “Ultrasound imaging in radiation therapy : From interfractional to intrafractional guidance,” *Cureus*, vol. 7, n°. 6, 2015.

[27] A. G. Webb, *Ultrasonic Imaging*. Wiley, 2003, p. 107–156.

[28] C. B. Burckhardt, “Speckle in ultrasound b-mode scans,” *IEEE Transactions on Sonics and ultrasonics*, vol. 25, n°. 1, p. 1–6, 1978.

[29] T. O’Shea, J. Bamber, D. Fontanarosa *et al.*, “Review of ultrasound image guidance in external beam radiotherapy part II : intra-fraction motion management and novel applications,” *Physics in Medicine and Biology*, vol. 61, n°. 8, p. R90–R137, mar 2016. [En ligne]. Disponible : https://doi.org/10.1088%2F0031-9155%2F61%2F8%2Fr90

[30] J. Park, J. B. Kang, J. H. Chang et Y. Yoo, “Speckle reduction techniques in medical ultrasound imaging,” *Biomedical Engineering Letters*, vol. 4, n°. 1, p. 32–40, 2014.

[31] J. H. STEWART et M. GRUBB, “Understanding vascular ultrasonography,” dans *Mayo Clinic Proceedings*, vol. 67, n°. 12. Elsevier, 1992, p. 1186–1196.

[32] Q. Huang et Z. Zeng, “A review on real-time 3d ultrasound imaging technology,” *BioMed research international*, vol. 2017, 2017.

[33] D. Fontanarosa, S. van der Meer, J. Bamber *et al.*, “Review of ultrasound image guidance in external beam radiotherapy : I. treatment planning and inter-fraction motion management,” *Physics in Medicine and Biology*, vol. 60, n°. 3, p. R77–R114,

jan 2015. [En ligne]. Disponible : https://doi.org/10.1088%2F0031-9155%2F60%2F3%2Fr77

[34] J. Schwaab, M. Prall, C. Sarti, R. Kaderka, C. Bert, C. Kurz, K. Parodi, M. Günther et J. Jenne, "Ultrasound tracking for intra-fractional motion compensation in radiation therapy," *Physica Medica*, vol. 30, nº. 5, p. 578–582, 2014, particle Radiosurgery Conference. [En ligne]. Disponible : https://www.sciencedirect.com/science/article/pii/S112017971400043X

[35] L.-L. Ting, H.-C. Chuang, A.-H. Liao, C.-C. Kuo, H.-W. Yu, Y.-L. Zhou, D.-C. Tien, S.-C. Jeng et J.-F. Chiou, "Experimental verification of a two-dimensional respiratory motion compensation system with ultrasound tracking technique in radiation therapy," *Physica Medica*, vol. 49, p. 11–18, 2018. [En ligne]. Disponible : https://www.sciencedirect.com/science/article/pii/S1120179718304526

[36] M. Lachaine et T. Falco, "Intrafractional prostate motion management with the clarity autoscan system," *Medical physics international*, vol. 1, 2013.

[37] L. G. Roberts, "Machine perception of three-dimensional solids," Thèse de doctorat, Massachusetts Institute of Technology, 1963.

[38] J. Xiao, B. Russell et A. Torralba, "Localizing 3d cuboids in single-view images," *Advances in Neural Information Processing Systems*, 2012.

[39] V. Blanz et T. Vetter, "A morphable model for the synthesis of 3d faces," dans *Proceedings of the 26th annual conference on Computer graphics and interactive techniques*, 1999, p. 187–194.

[40] A. Kar, S. Tulsiani, J. Carreira et J. Malik, "Category-specific object reconstruction from a single image," dans *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2015, p. 1966–1974.

[41] V. D. Luca, T. Benz, S. Kondo *et al.*, "The 2014 liver ultrasound tracking benchmark," *Physics in Medicine and Biology*, vol. 60, nº. 14, p. 5571–5599, jul 2015.

[42] A. E. Bourque, S. Bedwani, J.-F. Carrier, C. Ménard, P. Borman, C. Bos, B. W. Raaymakers, N. Mickevicius, E. Paulson et R. H. Tijssen, "Particle filter–based target tracking algorithm for magnetic resonance–guided respiratory compensation : Robustness and accuracy assessment," *International Journal of Radiation Oncology*Biology*Physics*, vol. 100, nº. 2, p. 325–334, 2018. [En ligne]. Disponible : https://www.sciencedirect.com/science/article/pii/S0360301617339652

[43] C. Zachiu, N. Papadakis, M. Ries, C. Moonen et B. D. De Senneville, "An improved optical flow tracking technique for real-time mr-guided beam therapies in moving organs," *Physics in Medicine & Biology*, vol. 60, nº. 23, p. 9003, 2015.

[44] M. Seregni, C. Paganelli, D. Lee, P. Greer, G. Baroni, P. Keall et M. Riboldi, "Motion prediction in mri-guided radiotherapy based on interleaved orthogonal cine-mri," *Physics in Medicine & Biology*, vol. 61, nº. 2, p. 872, 2016.

[45] E. Tryggestad, A. Flammang, R. Hales, J. Herman, J. Lee, T. McNutt, T. Roland, S. M. Shea et J. Wong, "4d tumor centroid tracking using orthogonal 2d dynamic mri : implications for radiotherapy planning," *Medical physics*, vol. 40, nº. 9, p. 091712, 2013.

[46] T. Bjerre, S. Crijns, P. M. af Rosenschöld, M. Aznar, L. Specht, R. Larsen et P. Keall, "Three-dimensional mri-linac intra-fraction guidance using multiple orthogonal cine-mri planes," *Physics in Medicine & Biology*, vol. 58, nº. 14, p. 4943, 2013.

[47] A. Shepard, B. Wang, T. Foo et B. Bednarz, "A block matching based approach with multiple simultaneous templates for the real-time 2d ultrasound tracking of liver vessels," *Medical Physics*, vol. 44, 09 2017.

[48] E. Ozkan, C. Tanner, M. Kastelic, O. Mattausch, M. Makhinya et O. Goksel, "Robust motion tracking in liver from 2d ultrasound images using supporters," *International Journal of Computer Assisted Radiology and Surgery*, vol. 12, 06 2017.

[49] J. Banerjee, C. Klink, E. Vast, W. Niessen, A. Moelker et T. van Walsum, "A combined tracking and registration approach for tracking anatomical landmarks in 4d ultrasound of the liver," dans *MICCAI Workshop : Challenge on Liver Ultrasound Tracking*, 2015, p. 36–43.

[50] J. McClelland, *Estimating Internal Respiratory Motion from Respiratory Surrogate Signals Using Correspondence Models*. Berlin, Heidelberg : Springer Berlin Heidelberg, 2013, p. 187–213.

[51] J. D. Hoisak, K. E. Sixel, R. Tirona, P. C. Cheung et J.-P. Pignol, "Correlation of lung tumor motion with external surrogate indicators of respiration," *International Journal of Radiation Oncology*Biology*Physics*, vol. 60, nº. 4, p. 1298–1306, 2004. [En ligne]. Disponible : https://www.sciencedirect.com/science/article/pii/S0360301604020681

[52] S. Hughes, J. McClelland, S. Tarte, D. Lawrence, S. Ahmad, D. Hawkes et D. Landau, "Assessment of two novel ventilatory surrogates for use in the delivery of gated/tracked radiotherapy for non-small cell lung cancer," *Radiotherapy and Oncology*, vol. 91, nº. 3, p. 336–341, 2009. [En ligne]. Disponible : https://www.sciencedirect.com/science/article/pii/S0167814009001406

[53] J. McClelland, D. Hawkes, T. Schaeffter et A. King, "Respiratory motion models : A review," *Medical Image Analysis*, vol. 17, nº. 1, p. 19 – 42, 2013.

[54] F. Preiswerk, V. De Luca, P. Arnold, Z. Celicanin, L. Petrusca, C. Tanner, O. Bieri, R. Salomir et P. Cattin, "Model-guided respiratory organ motion prediction of the liver from 2d ultrasound," *Medical Image Analysis*, vol. 18, 07 2014.

[55] K. K. Brock, M. B. Sharpe, L. A. Dawson, S. M. Kim et D. A. Jaffray, "Accuracy of finite element model-based multi-organ deformable image registration," *Medical Physics*, vol. 32, n°. 6Part1, p. 1647–1659, 2005. [En ligne]. Disponible : https://aapm.onlinelibrary.wiley.com/doi/abs/10.1118/1.1915012

[56] M. Velec, J. L. Moseley, S. Svensson, B. Hårdemark, D. A. Jaffray et K. K. Brock, "Validation of biomechanical deformable image registration in the abdomen, thorax, and pelvis in a commercial radiotherapy treatment planning system," *Medical Physics*, vol. 44, n°. 7, p. 3407–3417, 2017. [En ligne]. Disponible : https://aapm.onlinelibrary.wiley.com/doi/abs/10.1002/mp.12307

[57] R. Li, X. Jia, J. H. Lewis, X. Gu, M. Folkerts, C. Men et S. B. Jiang, "Real-time volumetric image reconstruction and 3d tumor localization based on a single x-ray projection image for lung cancer radiotherapy," *Medical Physics*, vol. 37, n°. 6Part1, p. 2822–2826, 2010. [En ligne]. Disponible : https://aapm.onlinelibrary.wiley.com/doi/abs/10.1118/1.3426002

[58] A. King, C. Buerger, C. Tsoumpas, P. Marsden et T. Schaeffter, "Thoracic respiratory motion estimation from mri using a statistical model and a 2-d image navigator," *Medical Image Analysis*, vol. 16, n°. 1, p. 252 – 264, 2012.

[59] T. Rohlfing, C. R. Maurer Jr., W. G. O'Dell et J. Zhong, "Modeling liver motion and deformation during the respiratory cycle using intensity-based nonrigid registration of gated mr images," *Medical Physics*, vol. 31, n°. 3, p. 427–432, 2004. [En ligne]. Disponible : https://aapm.onlinelibrary.wiley.com/doi/abs/10.1118/1.1644513

[60] P. Arnold, F. Preiswerk, B. Fasel, R. Salomir, K. Scheffler et P. C. Cattin, "3d organ motion prediction for mr-guided high intensity focused ultrasound," dans *Medical Image Computing and Computer-Assisted Intervention – MICCAI 2011*, G. Fichtinger, A. Martel et T. Peters, édit. Berlin, Heidelberg : Springer Berlin Heidelberg, 2011, p. 623–630.

[61] Y. Noorda, L. Bartels, M. Viergever et J. Pluim, "Subject-specific four-dimensional liver motion modeling based on registration of dynamic mri," *Journal of Medical Imaging*, vol. 3, p. 015002, 02 2016.

[62] J. R. McClelland, M. Modat, S. Arridge, H. Grimes, D. D'Souza, D. Thomas, D. O'Connell, D. A. Low, E. Kaza, D. J. Collins *et al.*, "A generalized framework unifying image registration and respiratory motion models and incorporating image reconstruction, for

partial image data or full images," *Physics in Medicine & Biology*, vol. 62, n°. 11, p. 4273, 2017.

[63] I. Jolliffe et B. Morgan, *Principal Component Analysis and Factor Analysis*. New York, NY : Springer New York, 2002, p. 150–166. [En ligne]. Disponible : https://doi.org/10.1007/0-387-22440-8_7

[64] L. Ruijiang, J. H. Lewis, X. Jia, T. Zhao, W. Liu, S. Wuenschel, J. Lamb, D. Yang, D. A. Low et S. B. Jiang, "On a PCA-based lung motion model," *Physics in Medicine and Biology*, vol. 56, n°. 18, p. 6009–6030, aug 2011. [En ligne]. Disponible : https://doi.org/10.1088/0031-9155/56/18/015

[65] W. Harris, L. Ren, J. Cai, Y. Zhang, Z. Chang et F.-F. Yin, "A technique for generating volumetric cine mri (vc-mri)," *International Journal of Radiation Oncology*Biology*Physics*, vol. 95, 02 2016.

[66] B. Stemkens, R. H. Tijssen, B. D. De Senneville, J. J. Lagendijk et C. A. Van Den Berg, "Image-driven, model-based 3d abdominal motion estimation for mr-guided radiotherapy," *Physics in Medicine & Biology*, vol. 61, n°. 14, p. 5335, 2016.

[67] J. Pham, W. Harris, W. Sun, Z. Yang, F.-F. Yin et L. Ren, "Predicting real-time 3d deformation field maps (dfm) based on volumetric cine mri (vc-mri) and artificial neural networks for on-board 4d target tracking : a feasibility study," *Physics in Medicine and Biology*, vol. 64, 07 2019.

[68] I. Y. Ha, M. Wilms, H. Handels et M. P. Heinrich, "Model-based sparse-to-dense image registration for realtime respiratory motion estimation in image-guided interventions," *IEEE Transactions on Biomedical Engineering*, vol. 66, n°. 2, p. 302–310, 2019.

[69] G. Samei, C. Tanner et G. Székely, "Predicting liver motion using exemplar models," dans *International MICCAI Workshop on Computational and Clinical Challenges in Abdominal Imaging*, 10 2012, p. 147–157.

[70] J. Ehrhardt, R. Werner, A. Schmidt-Richberg et H. Handels, "Statistical modeling of 4d respiratory lung motion using diffeomorphic image registration," *IEEE Transactions on Medical Imaging*, vol. 30, n°. 2, p. 251–265, 2011.

[71] T. Klinder, C. Lorenz et J. Ostermann, "Prediction framework for statistical respiratory motion modeling," dans *Medical Image Computing and Computer-Assisted Intervention – MICCAI 2010*, T. Jiang, N. Navab, J. P. W. Pluim et M. A. Viergever, édit. Berlin, Heidelberg : Springer Berlin Heidelberg, 2010, p. 327–334.

[72] F. Preiswerk, P. Arnold, B. Fasel et P. C. Cattin, "A bayesian framework for estimating respiratory liver motion from sparse measurements," dans *Abdominal Imaging. Com-*

putational and Clinical Applications, H. Yoshida, G. Sakas et M. G. Linguraru, édit. Berlin, Heidelberg : Springer Berlin Heidelberg, 2012, p. 207–214.

[73] C. Paganelli, D. Lee, J. Kipritidis, B. Whelan, P. Greer, G. Baroni, M. Riboldi et P. Keall, "Feasibility study on 3d image reconstruction from 2d orthogonal cine-mri for mri-guided radiotherapy," *Journal of Medical Imaging and Radiation Oncology*, vol. 62, 02 2018.

[74] H. Fayad, J. F. Clément, T. Pan, C. Roux, C. C. Le Rest, O. Pradier et D. Visvikis, "Towards a generic respiratory motion model for 4d ct imaging of the thorax," dans *2009 IEEE Nuclear Science Symposium Conference Record (NSS/MIC)*, 2009, p. 3975–3979.

[75] D. Boye, G. Samei, J. Schmidt, G. Székely et C. Tanner, "Population based modeling of respiratory lung motion and prediction from partial information," dans *Medical Imaging 2013 : Image Processing*, vol. 8669. International Society for Optics and Photonics, 2013, p. 86690U.

[76] C. Tanner, Y. Zur, K. French, G. Samei, J. Strehlow, G. Sat, H. Donald-Simpson, J. Houston, S. Kozerke, G. Székely, A. Melzer et T. Preusser, "In vivo validation of spatio-temporal liver motion prediction from motion tracked on mr thermometry images," *International journal of computer assisted radiology and surgery*, vol. 11, 04 2016.

[77] C. Jud, P. C. Cattin et F. Preiswerk, "Chapter 14 - statistical respiratory models for motion estimation," dans *Statistical Shape and Deformation Analysis*, G. Zheng, S. Li et G. Székely, édit. Academic Press, 2017, p. 379 – 407.

[78] F. Rosenblatt, "Principles of neurodynamics. perceptrons and the theory of brain mechanisms," Cornell Aeronautical Lab Inc Buffalo NY, Rapport technique, 1961.

[79] H. Hassan, A. Negm, M. Zahran et O. Saavedra, "Assessment of artificial neural network for bathymetry estimation using high resolution satellite imagery in shallow lakes : Case study el burullus lake." *International Water Technology Journal*, vol. 5, 12 2015.

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

[81] Y. LeCun, K. Kavukcuoglu et C. Farabet, "Convolutional networks and applications in vision," dans *Proceedings of 2010 IEEE International Symposium on Circuits and Systems*, 2010, p. 253–256.

[82] I. Goodfellow, Y. Bengio et A. Courville, *Deep Learning*. MIT Press, 2016, http://www.deeplearningbook.org.

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

[84] A. S. Lundervold et A. Lundervold, "An overview of deep learning in medical imaging focusing on mri," *Zeitschrift für Medizinische Physik*, vol. 29, nº. 2, p. 102–127, 2019, special Issue : Deep Learning in Medical Physics. [En ligne]. Disponible : https://www.sciencedirect.com/science/article/pii/S0939388918301181

[85] S. Albelwi et A. Mahmood, "A framework for designing the architectures of deep convolutional neural networks," *Entropy*, vol. 19, nº. 6, 2017. [En ligne]. Disponible : https://www.mdpi.com/1099-4300/19/6/242

[86] S. Ladjal, A. Newson et C.-H. Pham, "A pca-like autoencoder," 2019.

[87] Y. Bengio, A. Courville et P. Vincent, "Representation learning : A review and new perspectives," *IEEE transactions on pattern analysis and machine intelligence*, vol. 35, nº. 8, p. 1798–1828, 2013.

[88] Towards Data Science. (2019) Convolutional autoencoders for image noise reduction. [En ligne]. Disponible : https://towardsdatascience.com/convolutional-autoencoders-for-image-noise-reduction-32fce9fc1763

[89] D. P. Kingma et M. Welling, "An introduction to variational autoencoders," *Foundations and Trends® in Machine Learning*, vol. 12, nº. 4, p. 307–392, 2019. [En ligne]. Disponible : http://dx.doi.org/10.1561/2200000056

[90] M. Chen, X. Shi, Y. Zhang, D. Wu et M. Guizani, "Deep features learning for medical image analysis with convolutional autoencoder neural network," *IEEE Transactions on Big Data*, p. 1–1, 2017.

[91] M. Jaderberg, K. Simonyan, A. Zisserman *et al.*, "Spatial transformer networks," dans *Advances in neural information processing systems*, 2015, p. 2017–2025.

[92] G. Balakrishnan, A. Zhao, M. R. Sabuncu, J. Guttag et A. V. Dalca, "Voxelmorph : a learning framework for deformable medical image registration," *IEEE transactions on medical imaging*, vol. 38, nº. 8, p. 1788–1800, 2019.

[93] L. V. Romaguera, R. Plantefève, F. P. Romero, F. Hébert, J.-F. Carrier et S. Kadoury, "Prediction of in-plane organ deformation during free-breathing radiotherapy via discriminative spatial transformer networks," *Medical Image Analysis*, vol. 64, p. 101754, 2020.

[94] J. Walker, A. Gupta et M. Hebert, "Dense optical flow prediction from a static image," dans *Proceedings of the IEEE International Conference on Computer Vision*, 2015, p. 2443–2451.

[95] Z. Luo, B. Peng, D.-A. Huang, A. Alahi et L. Fei-Fei, "Unsupervised learning of long-term motion dynamics for videos," dans *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2017, p. 2203–2212.

[96] C. Schuldt, I. Laptev et B. Caputo, "Recognizing human actions : a local svm approach," dans *Proceedings of the 17th International Conference on Pattern Recognition, 2004. ICPR 2004.*, vol. 3. IEEE, 2004, p. 32–36.

[97] A. Shahroudy, J. Liu, T. Ng et G. Wang, "NTU RGB+D : A large scale dataset for 3d human activity analysis," *CoRR*, vol. abs/1604.02808, 2016. [En ligne]. Disponible : http://arxiv.org/abs/1604.02808

[98] P. Huang, G. Yu, H. Lu, D. Liu, L. Xing, Y. Yin, N. Kovalchuk, L. Xing et D. Li, "Attention-aware fully convolutional neural network with convolutional long short-term memory network for ultrasound-based motion tracking," *Medical Physics*, vol. 46, nº. 5, p. 2275–2285, 2019.

[99] F. Liu, D. Liu, J. Tian, X. Xie, X. Yang et K. Wang, "Cascaded one-shot deformable convolutional neural networks : Developing a deep learning model for respiratory motion estimation in ultrasound sequences," *Medical Image Analysis*, vol. 65, p. 101793, 2020.

[100] J. He, C. Shen, Y. Huang et J. Wu, "Siamese spatial pyramid matching network with location prior for anatomical landmark tracking in 3-dimension ultrasound sequence," dans *Pattern Recognition and Computer Vision*, Z. Lin, L. Wang, J. Yang, G. Shi, T. Tan, N. Zheng, X. Chen et Y. Zhang, édit. Cham : Springer International Publishing, 2019, p. 341–353.

[101] J. Krebs, T. Mansi, N. Ayache et H. Delingette, "Probabilistic motion modeling from medical image sequences : Application to cardiac cine-mri," dans *Statistical Atlases and Computational Models of the Heart. Multi-Sequence CMR Segmentation, CRT-EPiggy and LV Full Quantification Challenges*, M. Pop, M. Sermesant, O. Camara, X. Zhuang, S. Li, A. Young, T. Mansi et A. Suinesiaputra, édit. Cham : Springer International Publishing, 2020, p. 176–185.

[102] A. Giger, R. Sandkühler, C. Jud, G. Bauman, O. Bieri, R. Salomir et P. Cattin, "Respiratory motion modelling using cgans," dans *MICCAI*, 2018.

[103] P. Coupe, P. Hellier, C. Kervrann et C. Barillot, "Nonlocal means-based speckle filtering for ultrasound images," *IEEE Transactions on Image Processing*, vol. 18, nº. 10, p. 2221–2229, 2009.

[104] Z. Wang, A. C. Bovik, H. R. Sheikh et E. P. Simoncelli, "Image quality assessment : from error visibility to structural similarity," *IEEE transactions on image processing*, vol. 13, nº. 4, p. 600–612, 2004.

[105] D. A. Jaffray, *Radiation Therapy for Cancer*, 2015, ch. 14, p. 239–247. [En ligne]. Disponible : https://elibrary.worldbank.org/doi/abs/10.1596/978-1-4648-0349-9_ch14

[106] S. M. Camps, D. Fontanarosa, P. H. N. de With *et al.*, "The use of ultrasound imaging in the external beam radiotherapy workflow of prostate cancer patients," *BioMed Research International*, p. 16, 2018. [En ligne]. Disponible : https://doi.org/10.1155/2018/7569590

[107] J. J. Cerrolaza, Y. Li, C. Biffi *et al.*, "3d fetal skull reconstruction from 2dus via deep conditional generative networks," dans *Medical Image Computing and Computer Assisted Intervention – MICCAI 2018*. Cham : Springer International Publishing, 2018, p. 383–391.

[108] A. Kurenkov, J. Ji, A. Garg *et al.*, "Deformnet : Free-form deformation network for 3d shape reconstruction from a single image," *CoRR*, vol. abs/1708.04672, 2017. [En ligne]. Disponible : http://arxiv.org/abs/1708.04672

[109] A. V. Dalca, G. Balakrishnan, J. Guttag *et al.*, "Unsupervised learning for fast probabilistic diffeomorphic registration," *Lecture Notes in Computer Science*, p. 729–738, 2018. [En ligne]. Disponible : http://dx.doi.org/10.1007/978-3-030-00928-1_82

[110] D. P. Kingma et J. Ba, "Adam : A method for stochastic optimization," 2014.

[111] M. Baumann, M. Krause et R. Hill, "Exploring the role of cancer stem cells in radioresistance," *Nature Reviews Cancer*, vol. 8, n°. 7, 2008.

[112] D. Hawkes, D. Barratt, J. Blackall, C. Chan, P. Edwards, K. Rhode, G. Penney, J. McClelland et D. Hill, "Tissue deformation and shape models in image-guided interventions : a discussion paper," *Medical Image Analysis*, vol. 9, n°. 2, p. 163 – 175, 2005, medical Simulation - Delingette.

[113] K. K. Brock et L. A. Dawson, "Adaptive management of liver cancer radiotherapy," *Seminars in Radiation Oncology*, vol. 20, n°. 2, p. 107 – 115, 2010, adaptive Radiotherapy.

[114] L. Royer, A. Krupa, G. DARDENNE, A. Le Bras, E. Marchand et M. Marchal, "Real-time Target Tracking of Soft Tissues in 3D Ultrasound Images Based on Robust Visual Information and Mechanical Simulation," *Medical Image Analysis*, vol. 35, p. 582 – 598, janv. 2017.

[115] T. Mezheritsky, L. V. Romaguera et S. Kadoury, "3d ultrasound generation from partial 2d observations using fully convolutional and spatial transformation networks," dans *2020 IEEE 17th International Symposium on Biomedical Imaging (ISBI)*, 2020, p. 1808–1811.

[116] M. Drozdzal, E. Vorontsov, G. Chartrand, S. Kadoury et C. Pal, "The importance of skip connections in biomedical image segmentation," dans *Deep Learning and Data Labeling for Medical Applications*. Springer, 2016, p. 179–187.

[117] A. Paszke, S. Gross, F. Massa, A. Lerer, J. Bradbury, G. Chanan, T. Killeen, Z. Lin, N. Gimelshein, L. Antiga, A. Desmaison, A. Kopf, E. Yang, Z. DeVito, M. Raison, A. Tejani, S. Chilamkurthy, B. Steiner, L. Fang, J. Bai et S. Chintala, "Pytorch : An imperative style, high-performance deep learning library," dans *Advances in Neural Information Processing Systems 32*, H. Wallach, H. Larochelle, A. Beygelzimer, F. Alché-Buc, E. Fox et R. Garnett, édit. Curran Associates, Inc., 2019, p. 8024–8035.

[118] S. Klein, M. Staring, K. Murphy, M. A. Viergever et P. P. Josien, "elastix : a toolbox for intensity-based medical image registration," *IEEE Transactions on Medical Imaging*, vol. 29, n°. 1, p. 196 – 205, January 2010.

[119] M. von Siebenthal, G. Szekely, U. Gamper, P. Boesiger, A. Lomax et P. Cattin, "4d mr imaging of respiratory organ motion and its variability," *Physics in Medicine & Biology*, vol. 52, n°. 6, p. 1547, 2007.