Technical University of Munich
TUM School of Computation, Information and Technology

Methods for the acquisition, learning-based segmentation, and quantitative analysis of ultrasound volumes

Vanessa Gonzalez Duque

Complete reprint of the dissertation approved by the TUM School of Computation, Information and Technology of the Technical University of Munich for the award of the

Doktorin der Naturwissenschaften (Dr. rer. nat.)

Chair: Prof. Dr. Bertrand Michel

Examiners:

1. Prof. Dr. Nassir Navab
   Prof. Dr. Diana Mateus
2. Assoc. Prof. Mohammad Yaqub
3. Assistant Prof. Maria Alejandra Zuluaga Valencia

The dissertation was submitted to the Technical University of Munich on 14.02.2024 and accepted by the TUM School of Computation, Information and Technology on 18.06.2024.

Declaration of Authorship

— I know that plagiarism is wrong. Plagiarism is to use another's work and pretend that it is one's own.

— I have used British English spelling and the Harvard convention for citation and referencing. Each contribution to, and quotation in, this report from the work(s) of other people has been attributed, cited, and referenced.

— This research report is my own work.

— I have not allowed, and will not allow, anyone to copy my work with the intention of passing it off as his or her own work.

Name: Vanessa Gonzalez Duque
Date: 10th March 2024

Acknowledgments

I would like to thank my thesis supervisors, Prof. Diana Mateus and Prof. Nassir Navab, for guiding me along this path, managing the funding, and directing the scientific content of this thesis. I also thank them for trusting me, and for spending week and evenings helping me with the papers and this manuscript. Your guidance taught me to investigate and open my eyes to the research world.

I thank deeply the jury for their comments that helped me to improve my manuscript. Thanks to my CSI Jury for these 4 years of follow-up. Thank you to Nantes University and the researchers in the Department of Sport and Performance for their ultrasound-labeled dataset. My friends in Centrale and those at TUM, thank you for making this journey enjoyable, and a big thank you to my best friends Ludivine and Corentin for their technical advice and their contribution to my research way of thinking.

One special thank you to my co-authors for all the discussions and their strong contribution to the papers. I would especially thank Dawood for his guidance. He taught me how to supervise students. In this order, I would like to thank the 23 students in projects that I had the pleasure to supervise during my thesis. Watching them grow and learn from scratch the basics of networks, allowed me to understand the beauty of teaching. Through their eyes, I question myself what I took for granted and easy. They force me to explain better.

I would like to dedicate this thesis to my mother, **Corina Duque**. Without her confidence in my skills, I would have stopped my research or I would have never started. She saw in me something that not even I see. A big thank you for believing in me even when I do not believe I could make it. My family had a great influence on the success of this thesis. I would like to especially thank my sister and my gran mother for being in my life in this precious moment, thank you grad-ma for fighting to be with us, even if probably you barely recognize me now, I know you are proud.

Abstract

Context: Ultrasound imaging is a non-invasive technique, offering an inexpensive alternative to other imaging modalities such as Magnetic Resonance Imaging (MRI) or Computed Tomography (CT). Although typically used as a 2D modality, imaging 3D volumes is possible with matricial transducers or through probe tracking (freehand ultrasound). Three-dimensional (3D) ultrasound images are useful in assessing various medical conditions and treatment planning, including Duchenne dystrophy, hyperthyroidism, prostate cancer, spine injections, and leprosy. In the context of these conditions, 3D ultrasound enables, among others, to identify the borders or compute the volume of organs or lesions, contributing with quantitative information regarding the diagnosis, disease progression, or treatment response. Manually segmenting such volumes is time-consuming and difficult even for experts, therefore, there is a need to develop automated 3D ultrasound segmentation approaches, which will be the central topic of this thesis, addressed through deep-learning methods.

Deep learning image segmentation algorithms have demonstrated fast computational times at inference, beneficial for deployment as expert support tools. Despite their potential, 3D deep learning methods for ultrasound face unique challenges. First, available 3D datasets are scarce and often contain incomplete annotations. Second, 3D ultrasound images suffer from acquisition variability and tracking errors. Differences in probe positioning and patient anatomy can lead to significant variations in image quality and appearance. Third, ultrasound images are characterised by speckles, shadows, noise, reverberation, and low contrast, which lead to annotations variability. However, accurate annotations are essential for the success of supervised deep-learning methods. But, when it comes to ultrasound images, annotations are prone to significant variability. All these challenges highlight the need to develop specialised deep-learning methods for 3D ultrasound segmentation.

This thesis consists of three parts: the first part focuses on generating high-fidelity 3D ultrasound image volumes and labels. The second part proposes two models to segment 3D ultrasound volumes from sparse and incomplete annotations. The third part focuses on experimental analysis and modelling of the performance of the 3D deep-learning method with emphasis on the quality of border predictions.

Methods: Our proposed methods for building annotated volumetric ultrasound datasets presented in Part I can be divided into two studies. The first study assesses various label interpolation methods for generating three-dimensional (3D) annotations from two-dimensional (2D) sparse manual annotations with tracking. The second study explores the possibility of reconstructing ultrasound volumes from images alone (sensorless compounding) or reconstructing the volume while correcting tracking errors from a freehand ultrasound sequence. We made two propositions that work similarly to the state-of-the-art methods.

Two novel methods for 3D ultrasound segmentation are presented in part II. The first approach, named UNet-S-R-CLSTM, divides the 3D images into sub-volumes processed sequentially to handle the large input size while keeping image resolution. The method also proposes a modification of the segmentation loss function with negative labels to handle incomplete annotations. The second proposed approach introduces a new architecture called "Interactive Few Shot-Siamese Network (IFSSnet)" that segments 3D ultrasound volumes. This architecture incorporates a recurrent loop that feeds the predictions of preceding sub-volumes as input, leading to predictions with smooth borders. Moreover, an adaptable loss was proposed to penalise precision and recall during training while a memory module retains information regarding key slices.

Finally, part III summarises two empirical studies that analyse the factors that affect the performance of deep learning methods for ultrasound segmentation. These studies leverage an experimental validation on several architectures and datasets. The first method provides the network with precomputed confidence maps, supplying prior information on the variability of the annotations and potentially highlighting areas of higher uncertainty. The second study separately evaluates the performance of different architectures on complete and incomplete borders.

Keywords: 3D ultrasound, label interpolation, confidence maps, volume compounding, labeling variability, negative labels, multi-task learning.

Zusammenfassung

Kontext: Ultraschallbildgebung ist eine nicht-invasive Technik, die eine kostengünstige Alternative zu anderen Bildgebungsmodalitäten wie der Magnetresonanztomographie (MRT) oder der Computertomographie (CT) bietet. Obwohl üblicherweise als 2D-Modalität eingesetzt, ist die Bildgebung von 3D-Volumen mit matrixförmigen Wandlern oder durch Tracking der Sonde (Freihand-Ultraschall) möglich. Dreidimensionale (3D) Ultraschallbilder sind nützlich bei der Bewertung verschiedener medizinischer Zustände, einschließlich Duchenne-Dystrophie, Hyperthyreose, Prostatakrebs, Wirbelsäuleninjektionen und Lepra. Im Kontext dieser Zustände ermöglicht der 3D-Ultraschall unter anderem die Identifizierung der Grenzen von bestimmter Anatomien oder die Berechnung des Volumens von Organen oder Läsionen, was quantitative Informationen bezüglich der Diagnose, Krankheitsprogression oder Behandlungsmethode liefert. Die manuelle Segmentierung solcher Volumen ist zeitaufwändig und selbst für Experten schwierig, daher besteht die Notwendigkeit, automatisierte 3D-Ultraschallsegmentierungsansätze zu entwickeln, die das zentrale Thema dieser Thesis sein werden, adressiert durch Deep-Learning-Methoden.

Deep-Learning-Bildsegmentierungsalgorithmen haben schnelle Rechenzeiten bei der Inferenz gezeigt, was für den Einsatz als Expertenunterstützungswerkzeuge vorteilhaft ist. Trotz ihres Potenzials stehen 3D-Deep-Learning-Methoden für Ultraschall vor einzigartigen Herausforderungen. Erstens sind verfügbare 3D-Datensätze knapp und enthalten oft unvollständige Annotationen. Zweitens leiden 3D-Ultraschallbilder unter Variabilität bei der Akquisition und Tracking-Fehlern. Unterschiede in der Sondenpositionierung und der Patientenanatomie können zu signifikanten Variationen in der Bildqualität und dem Erscheinungsbild führen. Drittens sind Ultraschallbilder durch Speckle, Schatten, Rauschreverberation und geringen Kontrast gekennzeichnet. Schließlich ist die Variabilität der Beschriftungen wichtig, wobei die Qualität der Annotationen üblicherweise die Anisotropie der Bildqualität widerspiegelt. All diese Herausforderungen unterstreichen die Notwendigkeit, spezialisierte Deep-Learning-Methoden für die 3D-Ultraschallsegmentierung zu entwickeln.

Diese Thesis besteht aus drei Teilen: Der erste Teil konzentriert sich auf die Erzeugung von sehr genauen 3D-Ultraschallbildvolumen und -segmentierungen. Der zweite Teil schlägt zwei Modelle zur Segmentierung von 3D-Ultraschallvolumen aus wenigen und unvollständigen Annotationen vor. Der dritte Teil konzentriert sich auf die experimentelle Analyse und Modellierung der Leistung der 3D-Deep-Learning-Methoden mit Schwerpunkt auf den Grenzvorhersagen.

Methoden: Unsere vorgeschlagenen Methoden zum Erstellen von annotierten volumetrischen Ultraschalldatensätzen, die in Teil I vorgestellt werden, können in zwei Studien unterteilt werden. Die erste Studie bewertet verschiedene Label-Interpolationsmethoden zur Erzeugung von dreidimensionalen (3D) Annotationen aus zweidimensionalen (2D) spärlichen manuellen Annotationen mit Tracking. Die zweite Studie erforscht die Möglichkeit, Ultraschallvolumen allein aus Bildern (sensorlose Compoundierung) zu rekonstruieren oder das Volumen während der Korrektur von Tracking-Fehlern aus einer Freihand - Ultraschallsequenz zu rekonstruieren. Wir haben zwei Vorschläge gemacht, die ähnlich wie die State-of-the-Art-Methoden funktionieren.

Der zweite Methoden Ansatz führt eine neue Architektur ein, die als ""Interactive Few-Shot-Siamese Network (IFSSnet)" bezeichnet wird, die 3D-Ultraschallvolumina segmentiert. Diese Architektur integriert eine rekurrente Schleife, die die Vorhersagen vorangegangener Teilvolumen als Eingabe verwendet, was zu Vorhersagen mit glatten Grenzen führt. Darüber hinaus wurde eine anpassbare Verlustfunktion, um Genauigkeit und Rückruf während des Trainings zu bestrafen, während ein Speichermodul Informationen über Schlüsselschnitte behält. Schließlich fasst Teil III zwei empirische Studien zusammen, die die Faktoren analysieren, die die Leistung von Deep-Learning-Methoden für die Ultraschallsegmentierung beeinflussen. Diese Studien nutzen Erfahrungen mit mehreren Architekturen und Datensätzen. Die erste Methode versorgt das Netzwerk mit zusätzlichen Vorabinformationen über die Variabilität der Annotation in Form von vorberechneten "confidence maps", die potenziell bestimmte Bereiche hervorheben. Die zweite Studie bewertet separat die Leistung verschiedener Architekturen bei vollständigen und unvollständigen Grenzen.

Schlüsselwörter: 3D-Ultraschall, Etiketteninterpolation, Vertrauenskarten, negative Etiketten, Multi-Task-Lernen, Volumenkompoundierung, Etikettierungsvariabilität,.

Résumé étendu

L'échographie est une modalité d'imagerie médicale largement répandue utilisée pour le diagnostic et le suivi de pathologies, telles que les calculs biliaires, la déchirure des tendons de la coiffe des rotateurs, les grossesses normales ou ectopiques ou les problèmes de valves cardiaques [1]. Elle est fréquemment utilisée chez les femmes enceintes et les patients dans les unités d'urgence des hôpitaux [2, 3] en raison de sa nature non irradiante, de son coût inférieur, de son confort supérieur pour le patient, et de sa plus grande accessibilité. Sa portabilité et sa capacité à accélérer la prise de décisions cliniques en font un outil important dans le diagnostic, par rapport à des techniques d'imagerie telles que l'imagerie par résonance magnétique (IRM) ou la tomographie computationel (CT). Alors que l'IRM et le CT offrent une vue complète du champ et un contraste élevé des tissus, l'échographie 2D ne dispose pas d'un champ de vue 3D complet.

Malgré les limitations actuelles, l'échographie 3D a montré son intérêt clinique dans le diagnostic et le suivi de l'hyperthyroïdie [4,5], de la dystrophie musculaire de Duchenne [6–8], de la lèpre [9] et du cancer de la prostate [10], parmi d'autres maladies. Par exemple, la segmentation 3D des muscles des membres inférieurs fournit des informations volumétriques importantes pour le suivi des traitements tels que la dystrophie de Duchenne. Dans cette thèse, nous nous concentrons sur le développement de méthodes d'apprentissage profond pour aider les médecins dans l'analyse quantitative des séquences et volumes échographiques, dans le contexte des pathologies mentionnées ci-dessus, avec un accent particulier sur la segmentation.

Les volumes échographiques 3D peuvent être créés pour améliorer le champ de vue et permettre l'analyse volumétrique des structures anatomiques. Pour construire de tels volumes, plusieurs voies sont possibles, parmi lesquelles l'utilisation de sondes 3D [11] et d'acquisitions 3D combinées à du compounding [12]. Bien que les sondes 3D soient coûteuses, le compounding 3D à partir d'images échographiques 2D nécessite simplement de suivre la sonde pour faciliter l'alignement spatial des images dans une grille 3D qui sera remplie avec les valeurs d'intensité des b-mode images correspondantes. Cependant, l'attribution précise de ces valeurs dépend grandement de la précision des mécanismes de suivi. Il reste encore des améliorations à apporter pour réduire les exigences de suivi [13–15] (par exemple, ligne de mire) et générer des volumes échographiques 3D de haute qualité.

La création de bases de données d'images est le point de départ pour la formation des approches d'apprentissage profond. Des méthodes assistées par ordinateur IA ont été créées pour aider les cliniciens pendant le diagnostic, la planification des traitements et même dans la réalisation de procédures complexes avec une grande précision et efficacité. Les approches d'apprentissage profond ont réussi pour la localisation de plans standard fœtaux [16]„ la classification des lésions du sein et du foie [17,18], la segmentation des muscles cervicaux [19], le suivi des repères dans les séquences hépatiques [8] ou le suivi du cartilage du genou [7]. La segmentation des images et des volumes est particulièrement importante pour l'analyse précise et l'interprétation des maladies telles que l'hyperthyroïdie, la lèpre et le cancer de la prostate; où la taille, la forme et la distribution des structures sont des biomarqueurs pour le diagnostic de la réponse au traitement. La tâche de segmentation consiste à délimiter les limites des structures d'intérêt. Sur l'échographie volumétrique, c'est une tâche difficile pour les médecins et les méthodes IA en raison des caractéristiques intrinsèques de la modalité.

Les défis de la segmentation IA en échographie sont principalement dus à la dépendance au patient, à l'opérateur et au scanner. Lorsqu'un échographiste tient la sonde échographique, il/elle détermine la configuration des paramètres de la machine, positionne la sonde à un angle/pression spécifique sur le patient, et exploite ses connaissances anatomiques pour acquérir ce qu'ils considèrent comme une image de haute qualité de l'anatomie pertinente. Malgré les meilleurs efforts de l'échographiste, l'échographie reste susceptible à des problèmes tels que les limites floues [20] et divers artefacts tels que le speckle, l'ombre, la réverbération et la diffraction [21]. La complexité inhérente à l'interprétation des images échographiques nécessite que les médecins se forment pendant plusieurs années pour obtenir et évaluer avec précision des images échographiques de haute qualité, détectant des motifs et interprétant des relations sur les images qui sont difficiles à articuler ou à expliquer de manière concise. Les méthodologies d'intelligence artificielle (IA) émulent ce processus d'apprentissage en s'entraînant sur des ensembles de données analogues à ceux utilisés dans la formation médicale, imitant ainsi la trajectoire de développement de l'expertise médicale.

Comme exprimé auparavant, les annotations d'experts sur les volumes prendent beaucoup de temps et dépendent de l'utilisateur. Elles nécessitent des ressources informatiques coûteuses et des entrées d'experts pour une segmentation et une analyse detaille des images, augmentant les coûts et les efforts de compilation des images et des annotations dans ce qui est appelé un "Jeu de données" pour la formation en apprentissage profond. De nos jours, la disponibilité des jeux de données 3D est limitée, et ils incluent souvent des annotations partielles effectuées sur certaines des images 2D [22,23]. La nature anisotrope inhérente de l'échographie complique davantage l'identification précise des objets d'intérêt, créant une grande variabilité dans les annotations changeant la position, la forme et l'apparence des structures. Tous ces défis doivent être pris en compte lors de la construction des méthodes de segmentation en apprentissage profond.

Dans l'état de l'art, nous avons trouvé qu'en 2023, plus de 100 réseaux avaient mis l'accent sur la segmentation des structures [24], que ce soit dans des images 2D, des vidéos 2.5D ou des volumes 3D. Parmi les principales applications, nous avons trouvé : la segmentation fetale [25–28], la segmentation cardiaque [29–31] et la détection du cancer du sein [32, 33]. Une analyse complète des réseaux neuronaux de pointe actuels pour la segmentation échographique révèle une tendance à atteindre des performances optimales : augmenter la complexité architecturale [34–36], s'appuyer sur de plus grands volumes de données [37, 38], s'appuyer sur de nouveaux modules de mémoire et d'attention pour gérer la consistence des annotations [39–41] et utiliser des processus de pointe pour gérer des données limitées [42, 43]. Chacune des architectures contient différentes méthodes pour aborder les tâches et surmonter les défis de la segmentation échographique.

Au lieu d'augmenter la complexité architecturale pour des volumes haute résolution, tels que DAF3D [44] et Attetion-3DUNet [45] pour la segmentation du pancréas en 3D, nous identifions des modules qui gèrent efficacement les données haute résolution sous forme de sous-volumes et nous concentrons sur l'utilisation de données avec des étiquettes manquantes et des variations le long des bordures [46]. Similaire à PG-NET [47] pour la propagation des étiquettes vidéo, nous nous concentrons sur la capacité à propager les annotations intelligemment, en nous concentrant sur la formation avec des ensembles de données limités, éventuellement par des techniques telles que l'apprentissage incrémental. Nous analysons l'influence de la variabilité des étiquettes sur les performances des réseaux pour des architectures simples comme UNet [48], des architectures avec des modules complexes tels que Attention-UNet [45] et des architectures complexes telles que UNet-transformer [49], nous nous concentrons sur l'étude des méthodes que les réseaux utilisent pour la segmentation des bordures dans la modalité d'échographie difficile avec des bordures floues [50].

Objectifs de la recherche:

Nos recherches et principaux objectifs étaient motivés par des lacunes dans la littérature et les défis de l'échographie tels que les images à faible contraste, les artefacts, la dépendance à l'opérateur, mais aussi les erreurs de formation des volumes, la dépendance angulaire et les annotations éparses en 3D. Nos contributions se concentrent principalement sur les défis tels que les petits ensembles de données disponibles, les annotations de données limitées, le déséquilibre des classes et les incertitudes inhérentes à l'imagerie. Elles peuvent être résumées en trois groupes principaux liés aux trois objectifs.

But principal : *Notre objectif est de développer des algorithmes d'apprentissage automatique pour acquérir et segmenter des images échographiques afin d'améliorer les* *mesures quantitatives en 2D et 3D. Nous visons la création d'un jeu de données de volumes échographiques 3D haute résolution et, de plus, le développement et la validation de méthodes de segmentation automatique pour l'application dans diverses maladies nécessitant le calcul du volume. L'objectif a été décomposé en trois objectifs décrits ci-dessous.*

Motivés par les problématiques mises en évidence dans l'introduction, et cherchant des méthodes qui permettent la création de volumes echographiques et leurs annotations de façon automatique, le but et les objectifs suivants ont été développés pour ce travail de thèse :

Objectif 1: *Création de données echographiques 3D de haute résolution.*

L'objectif implique le développement et la mise en oeuvre de techniques avancées pour la reconstruction des volumes échographiques et leurs annotations 3D respectives. Atteindre cet objectif est important pour plusieurs raisons. Premièrement, les données échographiques 3D haute résolution fournissent un champ de vue détaillé et complet des structures anatomiques internes, permettant une analyse plus précise et approfondie des tissus. Cela pourrait entraîner une réduction de la complexité de la segmentation des images échographiques. De plus, cela fournit aux professionnels de la santé des informations plus détaillées et fiables pour éclairer leur prise de décision. Il semble important d'évaluer la qualité du suivi de la sonde, car elle est responsable de la qualité des volumes. L'apprentissage de bonnes méthodes d'acquisition de données ou l'amélioration des méthodes actuelles pourraient être étudiées. D'autre part, du côté des annotations, nous pourrions enquêter et proposer des méthodes pour créer des annotations 3D à partir d'annotations 2D éparses avec suivi.

Cette thèse contribue avec un jeu de données open-source haute résolution de volumes échographiques 3D du membre inférieur avec trois muscles segmentés. Nous mettons en open-source le jeu de données avec des modèles 3D complets des muscles : Gastrocnemius Medialis, Gastrocnemius Lateralis, et Solius. Il contient des données de 44 participants de l'étude de Crouzier et al [51]. Notre contribution était une méthode de segmentation semi-automatique pour créer des annotations 3D à partir d'étiquettes 2D éparses avec suivi optique. En termes d'erreur volumétrique moyenne, cette méthode atteint une erreur de performance similaire pour les annotateurs interopératoires. Notre méthode utilise l'enregistrement d'images 3D-3D pour fusionner efficacement les étiquettes de scans de qualité variable, améliorant la cohérence et la fluidité des annotations. Nous avons comparé cette approche avec des méthodes de segmentation non basées sur l'apprentissage profond et l'interpolation linéaire. À notre connaissance, c'est le plus grand jeu de données multi-étiquettes d'échographie 3D mis à disposition de la communauté pour la recherche et la comparaison de méthodes.

***Duque, V. G.**, Alchanti, D., Crouzier, M., Nordez, A., Lacourpaille, L., Mateus, D. (2020, November). lower limb muscles segmentation in 3D freehand ultrasound using non-learning methods and label transfer. In 16th International Symposium on Medical Information Processing and Analysis (Vol. 11583, pp. 154-163). SPIE.*

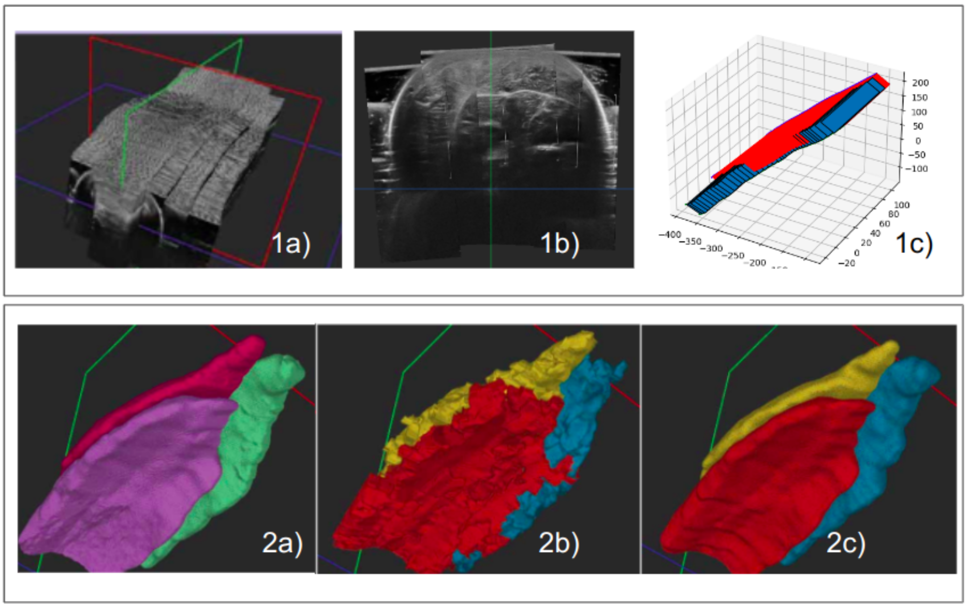


Figure 1 – 1a)3D ultrasound volume construct with tracking 1b) Cross-correlation view 1c)Sensorless positioning. 2a) Expert labels of the lower limb 2b)Grow from seeds method for label creation 2c)Our ZOI labels.

***Objective 2**: Développement de méthodes de segmentation automatique pour les muscles des membres inférieurs dans des volumes échographiques 3D.*

Cet objectif vise à créer des algorithmes précis et efficaces pour identifier et isoler de manière fiable des organes spécifiques (ici des muscles) dans des images échographiques tridimensionnelles. Transcendant les limites des techniques de segmentation manuelles ou semi-automatiques, ce travail se concentre sur l'offre de solutions plus rapides, plus objectives et reproductibles pour l'analyse quantitative des images échographiques, en particulier des mesures de volume musculaire. Une segmentation précise des muscles peut améliorer la compréhension anatomique et pathologique. De plus, les biomarqueurs issus de la segmentation peuvent informer et guider le diagnostic, le traitement et le suivi de diverses affections musculaires, des blessures d'athlètes sportifs aux maladies neuromusculaires telles que la dystrophie de Duchenne ou la sarcopénie.

Nous proposons des sub-volumes pourun réseau récurrent et de nous appuyer sur une stratégie de pseudo-étiquetage séquentiel pour gérer les annotations éparses. En pratique, nous proposons deux architectures, UNet-S-R-CLSTM et IFSS-Net, pour la segmentation des muscles des membres inférieurs en échographie. UNet-S-R-CLSTM est une architecture à un encodeur et deux décodeurs, utilisant un réseau Long-short Term Memory convolutionnel (CLSTM) pour la segmentation binaire des volumes échographiques à main levée de faible résolution. Cette approche tire parti de l'apprentissage multitâche pour améliorer l'estimation géométrique des formes de masques et adopte une stratégie d'apprentissage avec des étiquettes faibles en raison de l'ensemble de données comprenant des annotations de tranches 2D éparses. D'autre part, le IFSS-Net est une architecture à deux encodeurs-un décodeur avec un CLSTM bidirectionnel pour la segmentation des muscles dans des volumes échographiques volumétriques haute résolution. Cette méthode atteint une erreur volumétrique faible, comparable aux standards intra-opératoires, et introduit une mise à jour décrémentale de la fonction objectif pour faciliter la convergence du modèle avec des données annotées limitées. Pour aborder le déséquilibre des classes, nous proposons une fonction de perte de Tversky paramétrique, pénalisant adaptivement les faux positifs et les faux négatifs.

Gonzalez Duque, V., *Al Chanti, D., Crouzier, M., Nordez, A., Lacourpaille, L., Mateus, D. (2020). Spatio-temporal consistency and negative label transfer for 3D freehand US segmentation. In Medical Image Computing and Computer Assisted Intervention (MICCAI) 2020: 23rd International Conference, Lima, Peru, October 4–8, 2020, Proceedings, Part I 23 (pp. 710-720). Springer International Publishing.*

Al Chanti, D., ***Duque, V. G.***, *Crouzier, M., Nordez, A., Lacourpaille, L., Mateus, D. (2021). Interactive few-shot siamese network (IFSS-Net): for faster muscle segmentation and propagation in volumetric ultrasound. Institute of Electrical and Electronics Engineers (IEEE) transactions on medical imaging, 40(10), 2615-2628.*

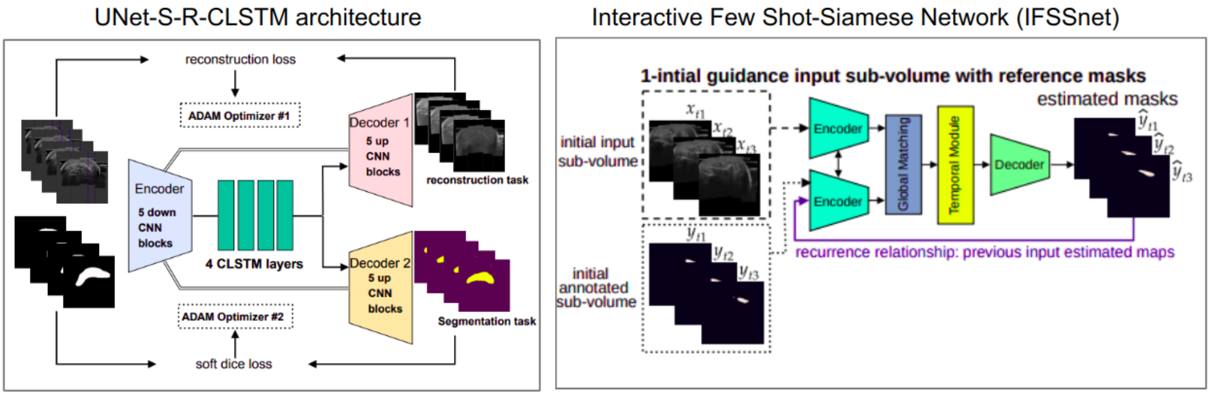


Figure 2 – UNet-S-R-CLSTM architecture et Interactive Few Shot-Siamese Network (IFSSnet)

***Objective 3**: Contribuer à la compréhension de la variabilité des performances des réseaux neuronaux d'apprentissage profond pour la segmentation d'échographies de pointe.*

Comprendre comment les médecins et les réseaux détectent les bordures dans les images échographiques est important pour relever les défis spécifiques à cette modalité, tels que les bords flous et l'anisotropie dans les étiquettes. Il est alors devenu important de fournir au réseau des informations sur la variabilité des étiquettes. La segmentation des bordures par les experts en échographie est influencée par des caractéristiques anisotropes et des facteurs dépendant de la position, découlant des principes physiques fondamentaux de l'imagerie par ultrasons. Les annotations sont susceptibles de présenter des variations substantielles le long des bords et des divergences significatives, dues aux bords flous ou aux faibles valeurs de signal. Certaines parties des bordures sont plus faciles à segmenter que d'autres, pour les médecins comme pour les réseaux. Par conséquent, nous cherchons à comprendre comment les bordures sont identifiées et à fournir des informations sur la variabilité des annotations.

Nous proposons les "Ultrasound Confidence Maps" (CM) comme outil pour l'estimation de la variabilité des étiquettes et adaptons les cartes d'activation (par exemple, Grad-Cam) pour soutenir l'analyse. Lorsque des méthodes d'IA sont appliquées à des images naturelles, IRM ou CT, les bordures sont plus faciles à définir par rapport aux images échographiques, où certaines bordures peuvent être évidentes tandis que d'autres doivent être interpolées, générant de la variabilité dans les annotations. Nous avons exploré l'utilisation de "Confidence Maps" en échographie dans les réseaux neuronaux pour identifier les régions incertaines de l'image et améliorer la segmentation. Cette méthode propose l'intégration novatrice de CMs comme une seconde entrée de canal ou dans la fonction de perte, améliorant les prédictions dans les zones d'incertitude physique inhérente à l'imagerie par ultrasons. D'autre part, nous avons proposé de diviser l'analyse des métriques de bordure pour l'échographie en bordures évidentes et complétées afin d'analyser expérimentalement la variabilité des performances des architectures existantes.

***Duque, V. G.**, Zirus, L., Velikova, Y., Navab, N., Mateus, D. (2023). Can ultrasound Confidence Maps predict expert labels' variability? In ASMUS workshop at MICCAI 2023: 26th International Conference, Vancouver, Canada, October 8–14, 2023, Proceedings pp. 100-120). Springer International Publishing.*

***Duque, V. G.**, Marquardt, A., Velikova, Y., Lacourpaille, L., Nordez, A., Crouzier, M., Lee H.J., Mateus, D., Navab N., (2023). Ultrasound Segmentation Analysis via Distinct and Completed Anatomical Borders. 15th International Conference on Information Processing in Computer-Assisted Interventions(IPCAI). International Journal of Com-*

puter Assisted Radiology and Surgery (IJCARS).

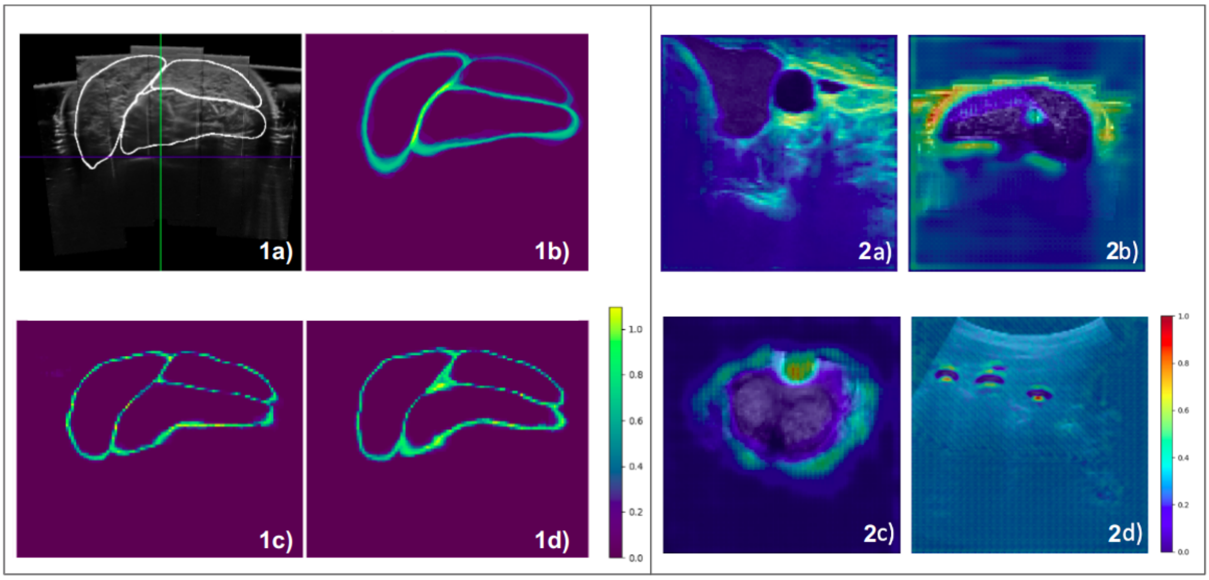


Figure 3 – 1a)Ultrasound cross-correlation view 1b) Entropy of experts annotations 1c)Entropy of UNet with 1 channel 1d) Entropy of UNet using Confidence maps; UNet Grad-cam of the background of the 2a) Thyroid 2b)lower limb muscles 2c)Prostate 2d)Spine bones.

Conclusion

En conclusion, cette thèse développe des algorithmes d'apprentissage automatique pour acquérir et segmenter des images échographiques afin d'améliorer les mesures quantitatives en 2D et 3D. Elle abordé les défis spécifiques à l'échographie inhérent à la modalité, tels que la gestion des annotations éparse, le faible contraste des images, les artefacts, et la incertitude des annotations. Nous avons créé un jeu de données de volumes échographiques 3D haute résolution, nous avons développé et validé des méthodes de segmentation automatique pour l'application dans diverses maladies nécessitant le calcul du volume. Nous avons etudie la variabilité des annotations en utilisant des techniques telles que les confidence maps et Seg-grad-cam. Finalement, nous avons evalue l'addaptation des mesures quantitatives especifiques pour l'ultrason. Cette recherche contribue à la littérature existante en offrant un ensemble de données open-source pour la segmentation musculaire en 3D, et en proposant des méthodes innovantes pour la segmentation et l'annotation automatisées. Ces contributions ouvrent des voies prometteuses pour de futures recherches.

Financement

La cotutelle a été établie entre la chaire CAMP de l'Université Technique de Munich (Allemagne) et l'équipe SIMS du laboratoire LS2N a l'École Centrale de Nantes (France). Ce travail a été soutenu d'une part par une bourse d'un an a la TUM et d'une autre par

Fonds Européen de développement régional, la région Pays de la Loire, Nantes Métropole (Convention 2017-10470), sur le programme Connect Talent MILCOM (Imagerie multimodale et apprentissage pour la médecine basée sur le calcul).

Table of Contents

List of Figures

Abstract

Abstract: Zusammenfassung

Abstract: Résumé étendu

Building 3D ultrasound annotated datasets

Deep neural networks for 3D ultrasound segmentation

Analysis of ultrasound segmentation architectures

List of Tables

Notations

General

S	Ultrasound raw signa
N	Number of items
n	Ithem number n^{th}
d	Distance
c	speed of sound
u	columns in a matrix
u	rows in a matrix
R	Rotation matrix
T	Transformation
r_{ij}	elements of a matrix
α	pitch angle
β	yaw angle
γ	roll angle
V	Stack of 2D images
L	2D Annotations
$\hat{Y}$	Segmentation map
X	Input elements
K	Number of channels of the input image
C	Number of channels equivalent to the number of segmentation classes
Θ	Biases
$\mathcal{L}$	Loss function
δ	Euclidean distance
$T_{\alpha,\beta}$	Learnable Tversky Loss
m	Voxels inherent to an image
CM	Confidence Map

Abbreviations

MRI Magnetic Resonance Imaging

CT Computed Tomography

AI Artificial intelligence

ZOI Zero Order Interpolation Method

IFSSnet Interactive Few Shot-Siamese Network

CLSTM Convolutional Long-short Term Memory

IFSS-Net Interactive few-shot siamese network

MICCAI Medical Image Computing and Computer Assisted Intervention

IEEE Institute of Electrical and Electronics Engineers

BICLSTM Bidirectional Convolutional long-short term memory

CM Confidence Map

DOFs Degrees of freedom

ZOI zero-order interpolation

MONAI Medical open network for AI

US Ultrasound

SNR signal-to-noise ratio

CSA cross-sectional area

MONAI Medical Open Network for AI

SciPy Science Python Library

FCNN Fully Convolutional Neural Network

DAS Delay-and-Sum

Dice Sørensen–Dice index

IoU Intersection over Union

PPV Precision or positive predictive value

TPR True Positive Rate

FNR False Negative Rate

HD Hausdorff distance

ASD Average surface distance

ASD Average surface distance

DF-FCN Direction-Fused Fully Connected Network

RNN Recurrent neuronal networks

FBS Fill Between Slices

GFS Grow from seeds

WS Watershed

GL Gastrocnemius Lateral

GM Gastrocnemius Medial

SOL Solius

mIoU mean Intersection over Union

Introduction and context

Ultrasound is a widely spread medical imaging modality used for diagnosis and follow-up of pathologies, such as Gallstone, Rotator Cuff Tendon Tear, Normal or Ectopic Pregnancy or Heart Valve Problems [1]. It is frequently used on pregnant women and patients in hospital emergency units [2, 3] due to its non-irradiation nature, lower cost, higher patient comfort, and greater accessibility. Its portability and ability to accelerate clinical decision-making make it an important tool in diagnosis, compared to imaging techniques such as magnetic resonance imaging (MRI) or computed tomography (CT). While MRI and CT provide a full field of view and high bone contrast, 2D ultrasound does not provides a full 3D field of view.

Despite current limitations, 3D ultrasound has shown its clinical advantages in the diagnosis and follow-up of hyperthyroidism [4, 5], Duchenne muscular dystrophy [6–8], Leprosy [9], and Prostate Cancer [10], among other diseases. For example, 3D segmentation of lower limb leg muscles provides important volumetric information for the follow-up of treatments such as Duchenne Dystrophy. In this thesis, we focus on the development of deep-learning methods to assist physicians in the quantitative analysis of ultrasound sequences and volumes, in the context of the above pathologies, with a major focus on segmentation.

3D ultrasound volumes can be created in order to increase the field of view and allow for the volumetric analysis of anatomical structures. In order to build such volumes, multiple avenues are possible, among which are 3D probes [11] and 3D acquisitions combined with compounding [12]. While 3D probes are expensive, 3D compounding from 2D ultrasound images requires tracking the probe to facilitate spatial alignment of the images in a 3D grid that will be filled with the corresponding B-mode intensity values. However, the accurate attribution of these values is highly dependent on the accuracy of tracking mechanisms. There is still room for improvement to reduce the tracking requirements [13–15] (e.g. line of sight) and generate 3D ultrasound high-quality volumes.

The creation of image databases is the starting point for the training of deep-learning approaches. AI computer-assisted methods have been created to help clinicians during diagnosis, treatment planning, and even in performing complex procedures with great accuracy and efficiency. Deep learning approaches have been successful for fetal standard plane localisation [16], breast and liver lesion classification [17, 18], cervical muscle segmentation [19], landmark tracking in liver sequences [8], or knee cartilage tracking [7]. Image and volume segmentation, in particular, are important for the precise analysis and interpretation of diseases like hyperthyroidism, leprosy, and prostate cancer; where the size, shape, and distribution of the structures are biomarkers for diagnosis of treatment response. The segmentation task consists in delineating the boundaries of the structures of interest. On volumetric ultrasound, this is a challenging task for physicians and AI methods due to the intrinsic characteristics of the modality.

Ultrasound AI-segmentation challenges are mainly due to patient dependency, operator dependency, and scanner dependency. When sonographer hold the ultrasound probe, they determine the machine's parameter configuration, position the probe at a specific angle/pressure over the patient, and leverage anatomical knowledge to acquire what they consider a high-quality image of the relevant anatomy. Despite the sonographers best efforts, ultrasound remains susceptible to issues like blurred boundaries [20] and various artefacts such as speckle, shadowing, reverberation, and diffraction [21]. The complexity inherent in interpreting ultrasound imagery necessitates that physicians train for several years to accurately obtain and evaluate high-quality ultrasound images detecting patterns and interpreting relationships on the images that are challenging to articulate or explain concisely. AI methodologies should emulate the learning process of medical professionals, like sonographers, by training on diverse datasets that reflect real-world clinical scenarios. This involves using a wide range of ultrasound images and expert annotations to enable the AI to generalize effectively. The training should be iterative, incorporating feedback from experienced practitioners to refine the model's performance, and follow a developmental trajectory that starts with simpler cases before advancing to more complex situations. By emulating these aspects of human learning, AI can enhance its effectiveness in assisting healthcare professionals and improving patient outcomes.

As expressed before, expert annotations on volumes are time-consuming and user-dependent. They require expensive computational resources and expert input for careful segmentation and analysis of the images, escalating the costs and efforts of compiling the images and annotation in what is called a "Dataset" for deep-learning training. Nowadays, the availability of 3D datasets is limited, and they often include partial annotations performed on some of the 2D images [22, 23]. The inherent anisotropic nature of ultrasound further complicates the accurate identification of objects of interest, creating high label variability in the annotations providing imprecise position, shape, and appearance of structures. All these challenges should be taken into account when building deep-learning segmentation methods.

In the state of the art, we found that by 2023, more than 100 methods had focused on the segmentation of structures [24], whether in 2D images, 2.5D videos, or 3D volumes. Among the main applications we found: fetus segmentation [25–28], heart segmentation [29–31] and breast cancer detection [32, 33]. A comprehensive analysis of the current state-of-the-art neural networks for ultrasound segmentation reveals a trend to achieve optimal performance: Increasing architectural complexity [34–36], relying on larger volumes of data [37,38], hinging on new memory and attention modules for handling smoothness [39–41] and using cutting-edge process for handling limited data [42,43]. Each one of the architectures contains different methods to address the tasks and overcome the ultrasound segmentation challenges.

Instead of increasing architectural complexity for high-resolution volumes, such as DAF3D [44] and attetion-3DUNet [45] for 3D pancreas segmentation, we identify modules that effectively handle high-resolution data as sub-volumes and focus on using data with missing labels and variations along borders [46]. Similar to PG-NET [47] for video label propagation, we focus on the ability to propagate annotations intelligently, focusing on training with limited datasets, possibly through techniques like incremental learning. We analyse the influence of label variability on networks' performance for simple architectures like UNet [48], architectures with complex modules such as attention-UNet [45] on 3D and complex architectures such as UNet-transformer [49], we focus the study on understanding the methods networks use for border segmentation in ultrasound, which is a challenging modality with blurred borders [50].

Our research and main objectives were motivated by gaps in the literature and the ultrasound challenges such as low contrast images, artefacts, operator-dependency, training errors, angular dependency and sparse annotations in 3D. **Our contributions** mainly focused on addressing challenges such as small available datasets, limited data annotations, class imbalance, and inherent imaging uncertainties. They can be summarised into three main groups correlated to the three objectives.

Our goal is to develop machine learning algorithms to acquire and segment ultrasound images in order to enhance quantitative measurements in 2D and 3D, aiming at the creation of a dataset of high-resolution 3D ultrasound volumes and, moreover, the development and validation of automatic segmentation methods for application in various diseases requiring volume calculation. The goal has been broken down into three objectives described below.

Objective 1*: Creation of high-resolution 3D ultrasound datasets.*

The objective involves the development and implementation of advanced techniques for the reconstruction of ultrasound volumes and their respective 3D annotations. Achieving this goal is important for several reasons. First, high-resolution 3D ultrasound data provides a detailed and complete field of view of internal anatomical structures, allowing for a more accurate and thorough analysis of tissues. This could result in a reduction in the complexity of segmenting ultrasound images. Additionally, it provides healthcare professionals with more detailed and reliable information to inform their decision-making. It seems important to evaluate the quality of the probe tracking, as it is directly correlated with the quality of the volumes. Learning good data acquisition methods or improving the present ones could be studied. On the other hand, on the annotation side, we investigated and propose methods for creating 3D annotations from 2D sparse annotations with tracking.

This thesis contributes with a high-resolution open-source dataset of 3D ultrasound volumes of the lower limb with three muscles segmented. We open-source the dataset with complete 3D models of the muscles: Gastrocnemius Medial (GM), Gastrocnemius Lateral (GL), and Solius (SOL). It contains data from 44 participants from the study of Crouzier *et al.* [51]. Participants were prone with the leg in a custom bath to prevent pressure dependency of the images. 4 to 6 sweeps were recorded from the knee to the ankle, for 15 of the participants a second recorded with a higher frequency was needed. Such 2D sweeps data and annotations done using Stradwin software were the input of our method. Our contribution was a semi-automatic segmentation method for creating 3D annotations from 2D sparse labels with optical tracking. In terms of mean volumetric error, this method achieves similar performance error for inter-operative annotators. Our method employs 3D-3D image registration to effectively merge labels from varying quality scans, enhancing annotation consistency and smoothness. We compared this approach with non-deep-learning segmentation methods and linear interpolation. To the best of our knowledge, this is the biggest 3D Ultrasound multi-label dataset of the leg made available to the community for research and methods comparison.

***Duque, V. G.**, Alchanti, D., Crouzier, M., Nordez, A., Lacourpaille, L., Mateus, D. (2020, November). lower limb muscles segmentation in 3D freehand ultrasound using non-learning methods and label transfer. In 16th International Symposium on Medical Information Processing and Analysis (Vol. 11583, pp. 154-163). SPIE.*

***Objective 2**: Development of automatic segmentation methods for low limb muscles in 3D ultrasound volumes.*

This objective aims to create accurate and efficient algorithms to reliably identify and isolate specific organs (here muscles) within three-dimensional ultrasound images. Transcending the limitations of manual or semi-automatic segmentation techniques, this work focuses on offering faster, more objective, and reproducible solutions for the quantitative analysis of ultrasound images, in particular, muscle volume measurements. A precise segmentation of muscles can improve anatomical and pathological understanding. Moreover, biomarkers from the segmentation can inform and guide the diagnosis, treatment, and monitoring of various muscular conditions, from sports athlete injuries to neuromuscular diseases like Duchenne dystrophy or sarcopenia.

We propose to leverage the processed volume with a recurrent network and rely on a sequential pseudo-labelling strategy to deal with sparse annotations. In practice, we propose two architectures, UNet-S-R-CLSTM and IFSS-Net architectures, for lower limb muscle ultrasound segmentation. UNet-S-R-CLSTM is a one-encoder and two-decoder architecture, utilising a Convolutional Long-short Term Memory (CLSTM) network for binary segmentation of freehand low-resolution ultrasound volumes. This approach leverages multitask learning to improve the geometrical estimation of mask shapes and adopts a weak-label learning strategy due to the dataset comprising sparse 2D slice annotations. On the other hand, the IFSS-Net is a two-encoders-one-decoder architecture with a Bidirectional CLSTM for muscle segmentation in high-resolution volumetric ultrasound volumes. This method achieves low volumetric error, comparable to intra-operative standards, and introduces a decremental update for the objective function to facilitate model convergence with limited annotated data. To address class imbalance, we propose a parametric Tversky loss function, adaptively penalizing false positives and negatives.

***Gonzalez Duque, V.**, Al Chanti, D., Crouzier, M., Nordez, A., Lacourpaille, L., Mateus, D. (2020). Spatio-temporal consistency and negative label transfer for 3D freehand US segmentation. In MICCAI 2020: 23rd International Conference, Lima, Peru, October 4–8, 2020, Proceedings, Part I 23 (pp. 710-720). Springer International Publishing.*

*Al Chanti, D., **Duque, V. G.**, Crouzier, M., Nordez, A., Lacourpaille, L., Mateus, D. (2021). IFSS-Net: for faster muscle segmentation and propagation in volumetric ultrasound. IEEE transactions on medical imaging, 40(10), 2615-2628.*

Scientific background

Abstract

In the following part, I will introduce the background concepts serving as a foundation for the development presented later in this thesis. The part is composed of two chapters. The first chapter explains the physical principles of ultrasound imaging, first in 2D, and then in 3D. Section 1.1 focuses on 2D ultrasound, presenting its advantages and applications, the physical principles, the potential artefacts, and the challenges for image analysis. Section 1.2 covers the medical acquisition protocols and the existing datasets used later for the experimental validation. Section 1.3 describes the 3D ultrasound acquisition process, emphasising the importance of tracking systems. methods and challenges. Section 1.4 presents a conclusion about 2D and 3D challenges, including their limitations and complexities.

Chapter two outlines the application of artificial intelligence to ultrasound image analysis. It covers computer-aided diagnosis application scenarios (Section 2.1) and a detailed explanation of segmentation metrics (Section 2.2). Section 2.3 thoroughly explains segmentation methods in ultrasound imaging exploring various ultrasound segmentation architectures, for 2D ultrasound images (Section 2.3.1), 2.5D ultrasound Videos(Section 2.3.2), and 3D ultrasound volumes (Section 2.3.3), emphasizing their design and implementation challenges. This overview serves as a foundation for understanding the complexities and advancements in ultrasound.

Chapter 1

Medical ultrasound imaging

1.1 2D Ultrasound imaging

1.1.1 Advantages and applications

Ultrasound is a valuable medical imaging technique known for its safety, as it is non-invasive and does not use ionizing radiation. It offers real-time imaging capabilities, which are crucial for observing organ function and guiding procedures like biopsies. Cost-effective and portable, it is widely used in settings ranging from hospitals to remote areas. Ultrasound serves multiple purposes, from monitoring pregnancies fetal health in obstetrics to diagnosing heart conditions in cardiology; it is useful in evaluating abdominal organs, musculoskeletal injuries, or vascular diseases, but also in emergency medicine and image-guided interventions. Such versatility makes it a fundamental tool in diverse medical fields [2,3]. Nowadays, ultrasound has become the standard modality for the detection and follow-up of hyperthyroidism, prostate cancer, breast cancer, and fetal development [53].

Studies continuously propose shifting from modalities such as Computer Tomography (CT) or Magnetic Resonance Imaging (MRI) to ultrasound for several diagnosis tasks. For instance, Jahanshir *et al.*. [54] proposed the use of ultrasound for the diagnosis of blunt chest trauma patients who suffer from ascites, pleural effusion, pericardial effusion, and pneumothorax. Similarly, Yassa *et al.* [55] used Ultrasound for the diagnosis of coronavirus disease over MRI. Also, Nordez *et al.*. [56] propose 3D ultrasound as a reliable tool for measuring muscle volume compared to MRI. Reasons for proposing such a modality shift include improving the safety for a broader range of patients, e.g. children and elderly subjects with metal implants or pacemakers; increasing accessibility during pandemics, as ultrasound's portability allows for bedside exams, which can reduce virus spread. Additionally, ultrasound is cost-effective and faster compared to MRI, importants factor in resource-limited scenarios. After motivating the use of ultrasound in medicine, we will describe the physical principles of the modality responsible for its low cost, portability, real-time, and safety advantages.

1.1.2 Ultrasound acquisition principle

The discovery of ultrasound imaging in medicin is attributed to Donald *et al.*. in 1958 [57]. It is considered a safe modality since it does not require irradiation. Its basic physical principle relies on a piezoelectric material, that generates and receives high-frequency waves imperceptible to the human ear. These piezoelectric semiconductors are housed in a probe, which is known as a transducer. The transducer first emits a pulse wave that propagates to the observed tissues. The probe is then switched to the receiver mode to collect the reflected sound waves during a period of time. Under the assumption of a constant speed of sound, the conversion of reflecting sounds to B-mode images is done by calculating their time of flight and the intensity with which the sound returns to the probe. This process is known as beamforming. All of the datasets used in this thesis rely on the widely-used Delay-and-Sum (DAS) beamforming method [58].

DAS performs constructive addition of signals at a particular point in the image, enhancing signal reception or transmission in a specified direction while suppressing noise and signals from other directions. This process significantly improves the resolution and quality of the received signal. The technique involves two main steps: delaying and summing the signals received by an array of sensors. Each sensor in the array receives a signal that has travelled over a different distance, resulting in a time delay associated with each signal. This delay is based on the geometry of the array and the direction of the incoming signal. After applying these calculated delays to the respective signals, the next step is to sum the time-aligned signals to obtain the output of the beamformer. The resultant beamformed signal $S(t)$ for any scan line can be expressed as a single equation that incorporates both the delay and summation steps:

$$S(t) = \sum_{n=1}^{N} p_n \left(t - \frac{d_n}{c} \right) \tag{1.1}$$

In this equation, N represents the number of sensors in the array, $p_n(t - \frac{d_n}{c})$ is the signal received by the n^{th} piezoelectric sensor delayed by the time $\tau_n = \frac{d_n}{c}$, where d_n is the distance from the signal source to the n^{th} sensor, and c is the speed of sound in the medium. After beamforming, the signal follows envelope detection, log compression and other post-processing steps. The process is repeated for several scan lines to build the final B-mode image. See Figure 1.1 for a graphical explanation of the image creation process.

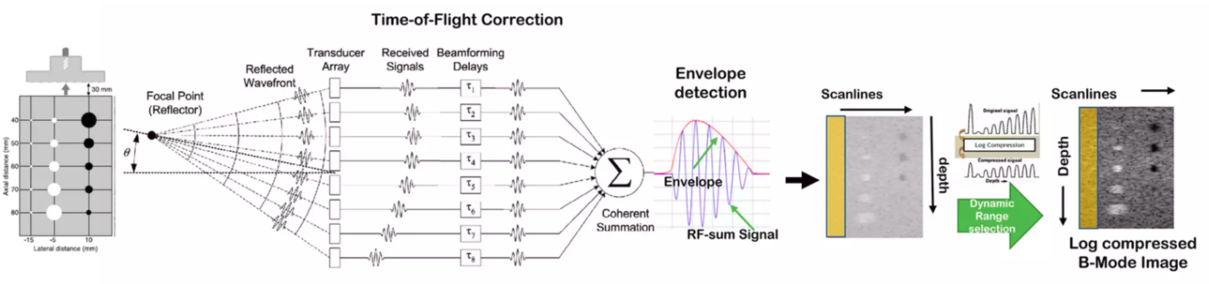


Figure 1.1 – Ultrasound beamforming, image taken from [59].

1.1.3 Ultrasound image artifacts

B-mode images are influenced by the interactions of sound with the traversed tissues. Although ideally, sound is expected to be reflected, it can be attenuated, absorbed, refracted, or scattered. The most common artifacts are shadowing, reverberation, diffraction, mirroring artifacts, and posterior acoustic enhancements [21].

- *Shadowing* occurs when the ultrasound beam is completely absorbed or reflected by a highly attenuating structure (like bones or gallstones), resulting in a dark or hypoechoic area on the image beyond that structure (See Figure 1.2-a).
- *Mirroring* artefacts occur when the ultrasound beam reflects off a strong reflector (like the diaphragm), creating a mirror-like duplicate image of structures on the opposite side of the reflector, leading to the potential misinterpretation of the anatomy (See Figure 1.2-b).
- *Reverberation* happens when the ultrasound waves bounce back and forth between two highly reflective surfaces, like gas, creating multiple, equally spaced echoes that appear as a series of parallel lines on the image (See Figure 1.2-c).
- *Posterior acoustic enhancement* is seen as an area of increased brightness or echogenicity behind structures that are less attenuating to the ultrasound beam, such as fluid-

filled cysts, indicating that more sound waves are passing through and reaching the deeper tissues (See Figure 1.2-d).

— *Diffraction* artifacts arise when the ultrasound beam spreads out after passing through a small aperture or around the edges of structures, potentially causing a blurring or smearing of the image at the edges.

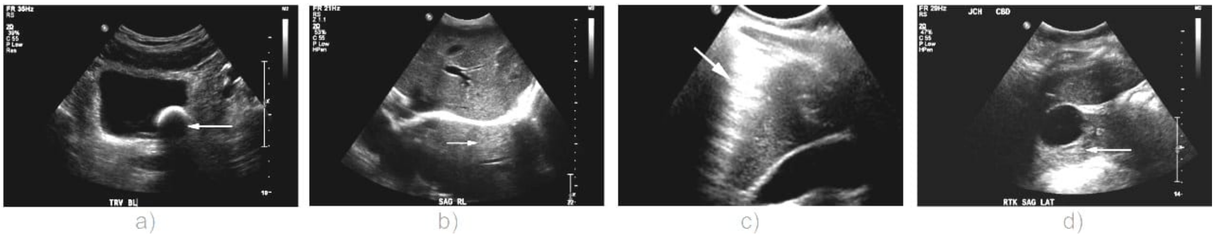


Figure 1.2 – Images taken from [21]. Ultrasound artifacts: a) Shadowing below a calculus in the urinary bladder, b) Mirroring artifact on the diaphragm, c) Posterior Acoustic Enhancement in the Renal Parenchyma, d) Reverberation artifact from Pneumoperitoneum against the Liver edge.

Sometimes, artefacts patterns provide useful information to the sonographer about the content in the image, nevertheless, more often, they decrease the accuracy of the content of deeper structures and make image content less reliable.

1.1.4 2D ultrasound challenges

Beyond artefacts, 2D ultrasound image quality depends on various other factors, such as patient variability, operator dependency, and scanner set-up. Operators' skills and experience significantly influence the choice of various parameters, including frequency, depth, focus, angle, and pressure, which vary depending on the patient's body and compliance. The visibility of specific zones of interest is highly dependent on the angle and pressure applied when the ultrasound probe contacts the skin. As the modality is dynamic, it requires physicians to use their anatomical knowledge to move the probe over the patient's body to find the plane of interest. Ultrasound images in 2D are view-dependent and often contain blurry contours, noise, and artefacts or have a limited field of view, making area measurements and structure localisation challenging. Additionally, the underlying physics of ultrasound results in the presence of speckle noise, which tends to reduce image resolution and contrast. The manual steering of the ultrasound probe to achieve the correct perspective and appropriate settings also contributes to operator dependency, leading to image variability and results that are not always directly comparable.

Ultrasound advantages encourage non-radiology-trained physicians to perform ultrasound examinations without a standardised training system or extensive quality assurance, in contrast to conventional radiology practices [60]. This leads to a substantial difference in image quality, further extending the operator dependency. To reduce such dependency, the medical community relies on manuals and protocols. These manuals define parameters such as the orientation of the probe, the frequency ranges, the structures that must be present in the images, and the analysis to be performed by the physician (e.g. landmark placement or volume calculation). Kim *et al.* [61] recommended a utopic system, where institutional practice guidelines set the scope of individuals who perform or interpret ultrasound examinations. But until this happens, physicians and artificial intelligence algorithms must deal with images with high variability. Operator, patient and scanner variability is a major challenge for deep-learning methods, which often require building sets of standardised images or datasets.

1.2 Datasets

Acquisition of medical ultrasound datasets is challenging compared to other types of modalities due to several reasons. Firstly, a protocol must fix a range for the parameters and restrict the high variation in the quality of the images. Secondly, medical image annotations demand specialised expertise. Consequently, the annotation process is usually expensive and sparse. Third, being medical data regulated by data protection, free, open-source datasets are still rare, although in constant growth.

1.2.1 Acquisition protocols

Ultrasound data acquisition protocols define the standardised processes and guidelines utilised in the gathering, processing, and transmission of data during ultrasound imaging procedures. There are various components and considerations involved in ultrasound data acquisition protocols [62]:

- **Transducer selection and handling**: Choosing the right transducer type (linear, phased, or curvilinear) and frequency is fundamental to optimize image quality based on depth and resolution.
- **Imaging modes selection**: is crucial for visualization of specific features in the images. B-Mode for anatomical clarity, M-Mode for motion analysis, and Doppler for blood flow assessment. All provide comprehensive diagnostic capabilities.
- **Image acquisition**: Tailored scanning techniques are adopted to accurately capture the targeted anatomical structures. Among the view directions, we can find transverse, longitudinal, or oblique techniques, which provide different perspectives.
- **Documentation and reporting**: is essential for ensuring that the Physician's findings are well-documented and effectively communicated among healthcare providers.
- **Data processing**: techniques are needed for image quality enhancement for precise diagnosis. Some existing techniques are noise filtering or contrast adjustment

— **Data Storage and transmission** is essential to handling secure and efficient medical data. Formats boost interoperability between diverse systems, make it easier to share and access data across different platforms, and streamline the diagnostic and treatment processes in healthcare settings. Different formats exist to save images and additional data information, between which we found: Digital Imaging and Communications in Medicine (DICOM), Meta Image (MHA/MHD), Joint Photo-graphic Experts Group (JPEG), Tag Image File Format (TIFF), Portable Network Graphic (PNG), Bitmap Image file (BMP), raw, Neuroimaging Informatics Technology Initiative (NIFTI), Hierarchical Data Format Version 5 (HDF5).

— **Safety and compliance**: follow the ALARA principle (As Low As Reasonably Achievable), which is a commitment to minimizing patient exposure time to ultrasound waves. Additionally, it is recommended to maintain equipment quality through regular checks. This dual approach ensures patient safety and reliable diagnostic data, fostering responsible healthcare practices.

By adhering to established protocols, healthcare professionals can ensure the uniformity and quality of ultrasound data acquisition, which is critical for accurate diagnostics, patient care, and collection of datasets for neural network training.

1.2.2 2D Datasets used in this thesis

As mentioned in the introduction, we focus in this thesis on the segmentation task in ultrasound images and volumes. In this subsection, I will describe the 2D datasets used in this thesis, along with their purpose and acquisition protocol. 3D datasets will be introduced in section 1.1.

	Contributor	Description	Size of set	Resolution	Format	Ref
2D	**UTP university**	Nerves: Median, ulnar, ciatic & femoral	1857 images	360 × 279	png 4ch files	[63]
2D	**Oxford university**	Spine bones	8 participants, 3292 images	128 × 128	PNG tracking	[64]

Table 1.1 – Overview of the 2D open-source datasets used in this thesis

The first dataset was recorded by an anaesthesiologist with ultrasound experience from the Universidad Tecnológica de Pereira (**UTP**) who acquired the 2D ultrasound dataset of the nerves in 2021 at Santa Mónica Hospital in Dosquebradas-Colombia. He used a SONOSITE Nano-Maxx device at a fixed resolution of 640 × 480 pixels, but the final network images were cropped to the region of interest to a maximal resolution of 360 × 279 pixels. The dataset comprises 691 images of nerves: 287 images from the sciatic nerve, 221 from the ulnar nerve, 41 from the median nerve, and 70 from the femoral nerve.

The second dataset was performed by an experienced physician at **Oxford** who set image ultrasound parameters to a depth of 90 mm for all 6 six scoliotic participants. He used a point-of-care ultrasound machine, MicrUs EXT-1H (Telemed Medical Systems, Milano, Italy). He recorded sagittal images with horizontal sweep movement from upper thoracic to lower lumbar levels, recorded at 10 frames per second over 2–3 minutes. Data annotations involved segmenting visible bone contours, employing a paintbrush-style tool within the 3D Slicer application, with approximately one frame segmented every 0.5 seconds.

Both datasets were used in this thesis for the validation and evaluation of networks and metrics in section 2.4

1.3 3D Ultrasound acquisition

Conventional 2D ultrasound imaging, while widely used in medical diagnostics, faces a significant limitation due to its restricted field of view. This constraint often hampers the ability to fully visualise and understand complex big anatomical structures and dynamic physiological processes. To solve this challenge, 3D ultrasound probes were invented to provide volumetric imaging, which enables a more comprehensive visualisation of tissues and organs. However, 3D probes are expensive compared to 2D probes, and they are not as commonly found in medical settings as their 2D counterparts, primarily due to the need for specialised training to interpret 3D images effectively.

Nevertheless, the clinical motivations for adopting 3D ultrasound are strong. The enhanced imaging capabilities of 3D ultrasound can lead to better patient outcomes through more accurate diagnoses and targeted treatment plans. Due to the clear advantages of 3D ultrasound, the limitations of 3D ultrasound probes and the high availability of 2D ultrasound probes, compounding methods of 3D volumes using 2D images were proposed in 1997 by Rohling *et al.* [12]. They created 3D ultrasound volumes from 2D tracked B-scans, also called "ultrasound compounding". In order to understand in detail what makes ultrasound compounding have good quality, it is necessary to go into detail about the tracking systems and the ultrasound compounding methods. We explain then the main components of the ultrasound compounding system in the following sections.

1.3.1 Tracking systems

Freehand scanners consist of a conventional transducer equipped with passive or active markers tracked from an external system. As the transducer moves following the surface of the scanned object, the system records the six (6) degrees of freedom (DOF)

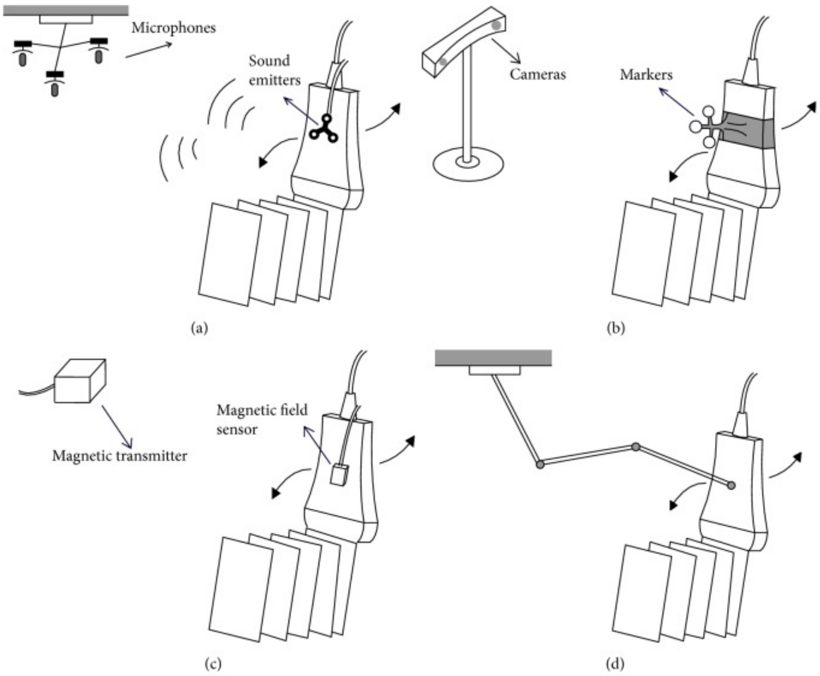


Figure 1.3 – Schematic structure of three types of position sensor: (a) acoustic sensor; (b) optimal positioner; (c) magnetic field sensor; (d) articulated arm positioner. Image taken from [11]

position of the probe over time. It is important, therefore, that the tracking system does not interfere with the trajectory and provides the freedom to scan the region of interest(ROI) from any angle or position. Several types of tracking systems exist, following the classification of Huang *et al.* [11](See Figure 1.3).

- **Acoustic tracking** [65]: In this type of system, the transducer holds three fixed sound-emitting components, while a series of microphones are distributed in the room. Position is calculated from the recorded time-of-flight from each sound emitter to the microphones. To maintain a satisfactory signal-to-noise ratio (SNR) of the sound-tracking signal, it is crucial to position the microphones near the patient and ensure the area between the emitters and microphones is clear of obstructions.
- **Optical tracking** [66]: This method involves a handheld transducer with an optical positioning system featuring either passive or active targets attached to the transducer and at least two cameras tracking these targets. The positional and orientation information is derived from the 2D images of targets, considering their relative positions. Optical systems are divided into passive stereo-vision systems (using three or more matte objects as targets) and active marker systems (using multiple known-frequency infrared diodes as markers). While recognised for its stability and precision, this setup allows accurate scanning with a handheld transducer and optical position, requiring a direct line of sight between the markers and the tracking cameras.

- **Magnetic Field Sensor tracking** [13,67]: In this setup, a transducer pairs with a magnetic field sensor comprising a magnetic transmitter positioned near the patient and a receiver with three orthogonal coils mounted on the transducer. This receiver measures the magnetic field intensity in three perpendicular directions to ascertain the transducer's position and orientation essential for 3D reconstruction. This compact and adaptable system does not require a clear line of sight. However, its efficacy can be hampered by electromagnetic interference and the presence of metal objects, which might distort readings and lessen tracking accuracy. To circumvent these issues, it is recommended to increase the magnetic field sampling rate.
- **Robotic ultrasound tracking** [5]: This strategy involves attaching the transducer to an articulated arm equipped with several movable joints, permitting clinicians to manoeuvre the transducer into any orientation. To enhance accuracy, it is advisable to keep the arm segments short.
- **Image-Based tracking** [68]: This method uses image characteristics like speckles to determine relative positioning, eliminating the need for additional sensors. Based on the principle of speckle decorrelation, the gap between two nearby images can be inferred from the correlation variations. When employing this method, operators are advised to move the transducer at a steady pace, either linearly or rotational, to maintain proper intervals, although this method might fall in terms of accuracy.

In this work, our datasets used mainly the optical tracking system and the magnetic tracking system. More details about the 3D datasets used in this thesis and their tracking system can be found below.

1.3.2 3D Datasets used in this thesis

Of the 3D datasets used in this thesis, the first one was open-source, the second one was made free in this thesis, and the third one is still private. They were used for the validation and evaluation of networks and metrics in section 2.4.

Thyroid Dataset: Presented by Kronke *et al.*. [5] offer 32 3D volumes from 16 individuals, capturing both the left and right neck regions. 3D Annotations include the thyroid gland, the jugular vein, and the carotid artery. Volume's pixel resolution is 380 × 330 × 300 and a voxel spacing of 0.12mm, acquired with a 3D curvilinear probe boasting 64 channels and a magnetic tracking system known as the "PIUR tUS tracking system."
Prostate 3D dataset: is an in-house dataset tailored specifically for the comprehensive examination of prostate health in the context of cancer suspicion. It comprises 40 3D volumes have a pixel resolution of 230 × 230 × 70 with a voxel spacing of 0.27mm. Each patient underwent both ultrasound and magnetic resonance-T1 (MRI) scans. The

ultrasound volumes were documented using a Wisonic Endokavitär Sonde EV10-4 rectal ultrasound probe in conjunction with the Acuson Juniper Siemens ultrasound apparatus. Prostate labels were derived from annotations on the MRI scans after the registration of the ultrasound volumes.

	Contributor	Description	Size of set	Resolution	Format	Ref
3D	**TUM**	Thyroid, carotid artery & jugular vein	16 participants: 32 volumes	380 × 330 × 300	Nifty files	[5]
3D	**Univ-Nantes**	Low limb Muscles: GM, GL, SOL	44 participants: 44 volumes	564 × 632 × 1443	mha with tracking	[51]
3D	**TUM**	Prostate	40 Participants	230 × 230 × 70	DICOM	in house

Table 1.2 – Overview of the 3D open-source datasets used in this thesis

The Low-limb dataset recorded 44 participants evenly split between males and females, who had an average age of 26±6 years, stood at a height of 173±11 cm, and weighed around 64.3±12.4 kg. Data was recorded as a part of the research on Achilles tendinopathy by Crouzier *et al.* [51] with ethical clearance from the local ethics board (Rennes Ouest V – CPP-MIP-010), fully conforming to the ethical standards laid out in the Declaration of Helsinki. None of these participants faced significant lower limb issues that required medical attention in the preceding six months.

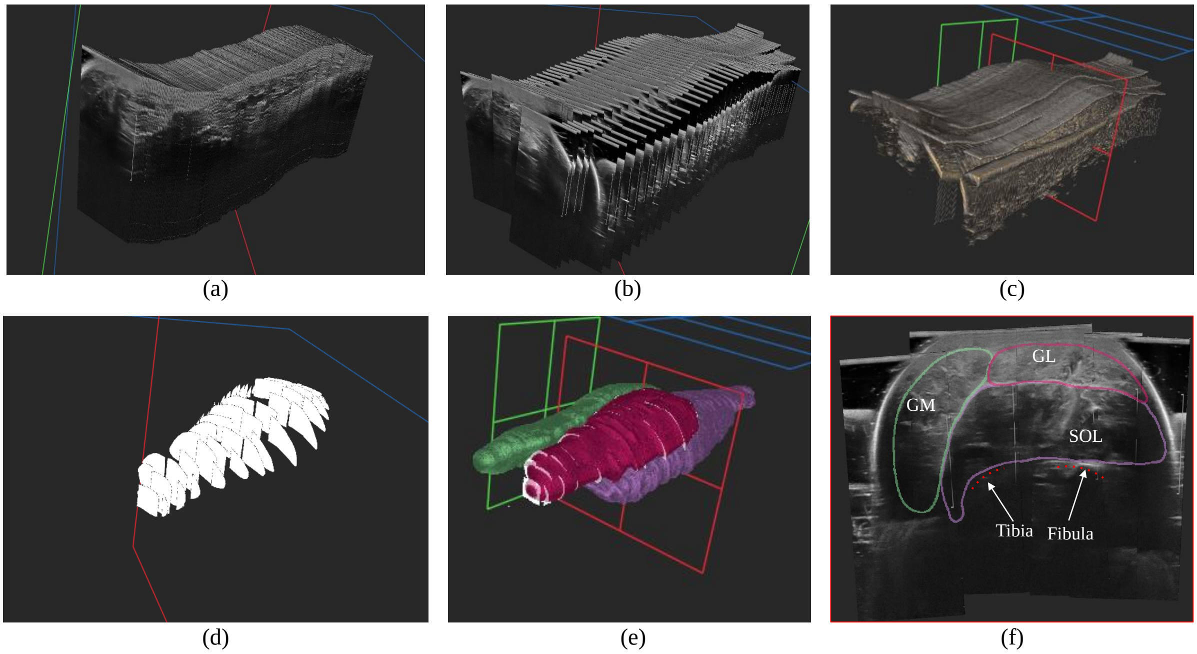


Figure 1.4 – LEG-3D-US dataset: a) Single ultrasound sweep, b) 5 sweeps, c) 3D volumes reconstructed, d) Sparse annotations e) Interpolations f) Cross-sectional view Solius (SOL), Gastrocnemius Lateralis (GL), Gastrocnemius Medialis (GM).

To capture these images, each participant's lower limb was positioned in a specially designed water tank (as shown in Figure 1.4-a) to prevent muscle distortion from the pressure of the ultrasound probe. A freehand ultrasound approach was used, employing a system of six optical tracking cameras (Optitrack, Natural point, USA) to generate the 3D volume scans. These images span from the knee to the ankle and were captured in 4-6 contiguous sweeps (illustrated in Figure 1.4-b) with the help of a 50-mm linear ultrasound probe (frequency range 4-15 MHz; Aixplorer, Supersonic Imagine, Aix-en-Provence, France). The 3D volumes were compiled using the compound volume algorithm within the ImFusion Suite software[1], employing a Gaussian kernel with a five-pixel spread (refer to Figure 1.4-c). The resulting volumes formed a voxel grid measuring $564\times632\times1443\pm(49\times38\times207)$ with an average voxel spacing of 0.276993 mm^3 ± 0.015 mm^3.

Muscular structures were sparsely annotated on 2D B-mode high-resolution images using Stradwin Software [22] by two separate annotators, achieving an intra-operator precision of 4%. The gastrocnemius medialis (GM), gastrocnemius lateralis (GL), and soleus (SOL) muscles were the primary focus of the segmentation. Three-dimensional muscle models were then created employing the "Zero order interpolation (ZOI) method" [69], and for a subset of 15 participants, these models were further refined by an expert. The LEG-3D-US dataset is now made publicly accessible for research and development purposes. Chapter 2.4 consider this dataset as one of the main contributions of the section.

1.3.3 From probe positions to image values

A graphic representation of the positioning of the coordinate reference systems is presented in Figure 1.5a. We observe in Figure 1.5-b that accurate tracking enables preservation of the continuity of edges between overlapping images, while tracking errors generate structure miss-alignment, as presented in Figure 1.5-c.

Formally tracking consists of finding the 6 Degrees of freedom (DOFs) position of the transducer **R** with respect to a global origin coordinate system **C**. A calibration step is required to relate the sensor receiver **O** to the global origin **C** and between the transducer position and the image plane **P**. It should be noted that accurate pre-calibrated tracking is required for volume reconstruction. As a result, they are two main sources of tracking errors: calibration errors affecting the transformation between the image and the probe and inaccuracies inherent to the tracking system relating **R** to **O**.

The 3D volume **V** is defined as a function assigning an intensity to every voxel **v** in a volumetric grid, $\mathbf{V} : \mathbf{v} \in \mathbb{R}^3 \to \mathbb{R}$. **V** is placed in the global coordinate system **C**, defined as a 3 x 3 orthonormal matrix. During the compounding reconstruction step, the volume

[1] ImFusion GmbH, Munich, Germany

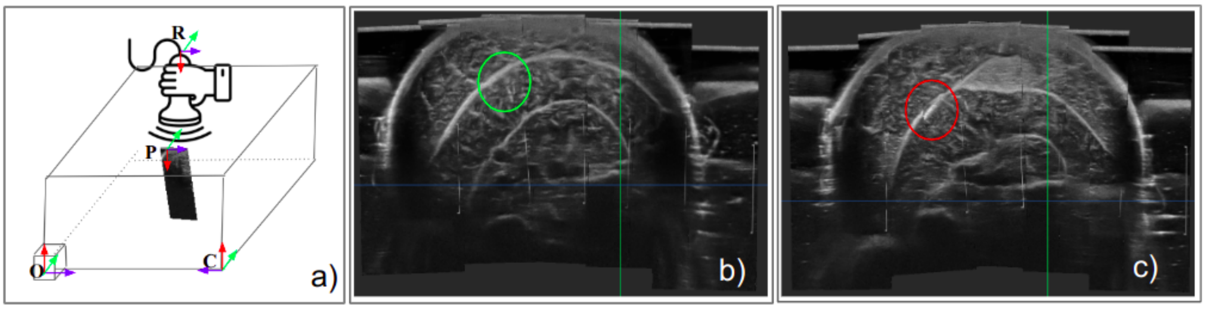


Figure 1.5 – Tracking coordinate reference systems and examples of 3D volumes: a)Reference coordinate systems, b) Correct tracking assuring smooth surface transitioning c)Patient movement affecting tracking and smoothness of the surface transitioning.

is filled with the pixel values of the B-scans passing through each voxel **v**. This is achieved by calculating the transformation of the image pixel in **P** to the corresponding voxel in **V** using relative transformation matrices, as expressed in equation 1.2.

Transformations (**T**) are modelled with 4 × 4 homogeneous matrices containing the translation, rotation, and scaling information between two coordinate systems. The objective is to express a pixel from one B-mode image in terms of the global reference frame **C** at the origin of the volume. Any pixel $x = \{\mathbf{u}, \mathbf{v}, 0, 1\}^T$ in the image (with $\mathbf{u}, \mathbf{v}$ its 2D coordinates), is first scaled to the metric system of the B-scan plane ${}^{P}\mathbf{T}_{x}$ (i.e. going from pixels to millimetres). A second transform ${}^{R}\mathbf{T}_{P}$ expresses the origin of the image plane with respect to the coordinate system of the ultrasound probe **R**. The third step corresponds to the transformation ${}^{O}\mathbf{T}_{T}$ going from the probe to the origin of the tracking system **O**. Finally, the last calibration matrix ${}^{C}\mathbf{T}_{O}$ express the pixel from the origin to the coordinate system of the volume. In summary:

$$ {}^{C}\underline{\mathbf{x}} = {}^{C}\mathbf{T}_{O}{}^{O}\mathbf{T}_{R}{}^{R}\mathbf{T}_{P}{}^{P}\mathbf{T}_{x} \tag{1.2} $$

P is the coordinate system of the B-scan plane, with an origin in the top left-hand corner of the cropped image. The y-axis is conventionally defined in the beam direction, and the x-axis in the lateral direction. The z-axis is in the elevational direction, orthogonal to the B-Scan plane. Finally, to build a US volume, we need to collect and fusion the information from a sequence of N B-mode images $\{I_1, I_2, ..., I_i, ...I_N\}$. It is thus possible to find the global coordinates of a point in image i using the absolute transformation ${}^{C}\mathbf{T}_{p_i}$, or compounding through multiplication the relative transformation between subsequent images ${}^{p_{i-1}}\mathbf{T}_{p_i}$ such that:

$$ \begin{aligned} {}^{C}\underline{\mathbf{x_i}} &= {}^{C}\mathbf{T}_{p_i}{}^{P_i}\underline{\mathbf{x_i}} \\ {}^{C}\underline{\mathbf{x_i}} &= {}^{C}\mathbf{T}_{p_1}{}^{p_1}\mathbf{T}_{p_2}...{}^{p_{i-1}}\mathbf{T}_{p_i}{}^{p_i}\underline{\mathbf{x_i}} \end{aligned} \tag{1.3} $$

Utilising relative transformations in tracking representations offers several advantages over absolute tracking. Relative tracking enhances robustness by localising errors to specific image pairs, reducing overall error accumulation compared to absolute tracking. This approach simplifies calculations, as it is easier to compute transformations between consecutive images rather than recalculating positions against a fixed global reference for each image. Relative tracking is also more efficient for analysing local movements and changes between frames, providing direct and relevant information for motion analysis.

1.3.4 Alternative rotation representations

As explained above, transformations are rigid homogeneous matrices and 2 characteristics: First, the last row is filled with 3 zeros and one 1, second the translation is described by a vector $[t_x, t_y, t_z]^T$ and the rotation matrix $\mathbf{R_9} \in \mathrm{SO}(3)$, has elements r_{ij} the elements of the matrix. Combining the translations and rotations in a single matrix we have:

$$\begin{bmatrix} & & & t_x \\ & \mathbf{R_9} & & t_y \\ & & & t_z \\ 0 & 0 & 0 & 1 \end{bmatrix} = \begin{bmatrix} r_{11} & r_{12} & r_{13} & t_x \\ r_{21} & r_{22} & r_{23} & t_y \\ r_{31} & r_{32} & r_{33} & t_z \\ 0 & 0 & 0 & 1 \end{bmatrix} \tag{1.4}$$

Representing the rotation with a SO(3) matrix will be called the $\mathbf{R_9}$ representation. We describe in the following other common ways to represent rotation matrices, which will be exploited in chapter 2 to define loss functions over rotations.

The $\mathbf{R_6}$ representation consists of using just 2 columns of the $\mathbf{R_9}$ matrix since the third column can be obtained with the cross-product of the first two, eliminating the redundancy of the $\mathbf{R_9}$ representation.

Then $\mathbf{R_3}$ representation, also called "Euler representation" [70] uses only 3 values (α, β, γ) to represent pitch, yaw, and roll of the axes in a fixed coordinate system. Euler representations are known to suffer from discontinuities. It is possible to convert the $\mathbf{R_9}$ representation into the Euler representation following equations 1.5.

$$\begin{gathered} \gamma = arctan\left(\frac{r_{21}}{r_{11}}\right), \\ \beta = arctan\left(\frac{-r_{31}}{\sqrt{1-r_{31}^2}}\right), \alpha = arctan\left(\frac{r_{32}}{r_{33}}\right) \end{gathered} \tag{1.5}$$

The $\mathbf{R_4}$ representation does not suffer from discontinuities, and it is commonly called the "quaternions representation". It is a 4-dimensional vector that can represent 3D rotations as:

$$\mathbf{q} = a + b\hat{i} + c\hat{j} + d\hat{k} \tag{1.6}$$

Where a is the scalar part and $\hat{\imath}, \hat{\jmath}, \hat{k}$ are unitary vectors. To convert from a rotation matrix $\mathbf{R_9}$ to quaternions $\mathbf{R_4}$, one can use the following expressions and normalise the resulting values by dividing each component by $\sqrt{a^2+b^2+c^2+d^2}$ to ensure the quaternions has a unit form.:

$$\begin{aligned} a &= \frac{1}{2}\sqrt{1+r_{11}+r_{22}+r_{33}} \\ b &= \frac{1}{2}\sqrt{1+r_{11}-r_{22}-r_{33}} \\ c &= \frac{1}{2}\sqrt{1-r_{11}+r_{22}-r_{33}} \\ d &= \frac{1}{2}\sqrt{1-r_{11}-r_{22}+r_{33}} \end{aligned} \tag{1.7}$$

Euler angles offer interpretability advantages but suffer from discontinuities. In contrast, $\mathbf{R_4}$, $\mathbf{R_6}$, $\mathbf{R_9}$, provide mathematical robustness, and smooth interpolation.

1.3.5 Compounding methods

From a sequence of image $\{\mathbf{I}_1, ..., \mathbf{I}_N\}$ and their corresponding absolute 3D positions $\{\mathbf{P}_1, ..., \mathbf{P}_N\}$, compounding or reconstruction step searches to fill the intensities of voxels $\mathbf{v}$ in a volume $\mathbf{V}$, with $\mathbf{V} \in \mathbb{R}^3$. Each image contributes with a set of 3D points, $\{\mathbf{x}_1, ..., \mathbf{x}_M\}$. These pixels are mapped to 3D points following Equation 1.3. Such that we obtain I_i : $\{{}^C\mathbf{x}_{i1}, ..., {}^C\mathbf{x}_{iM}\}$.

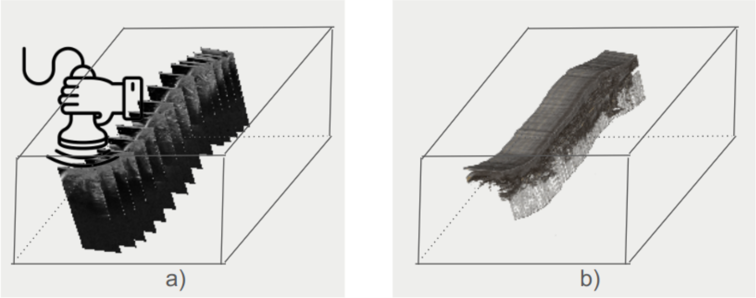


Figure 1.6 – From Sweep to volume: a) Ultrasound sweep, b)Compounded volume.

Since the location of the transformed points does not match voxel centers, compounding algorithms propose different strategies to interpolate the available data. Reconstruction algorithms can be classified into 3 types based on their implementation: Voxel-based methods (VBMs), Pixel-based methods (PBMs), and Function-based methods (FBMs). Figure 1.7 presents an example for each method, with only 2D projections of the 3D case. Rays represents images and circles represent pixel points $\{\mathbf{x_{i1}}, ..., \mathbf{x_{iM}}\}$.

Voxel-based methods (VBM) utilize the intensities of one or multiple image points to assign values to empty voxels. For instance, the voxel Nearest Neighbor (VNN) method

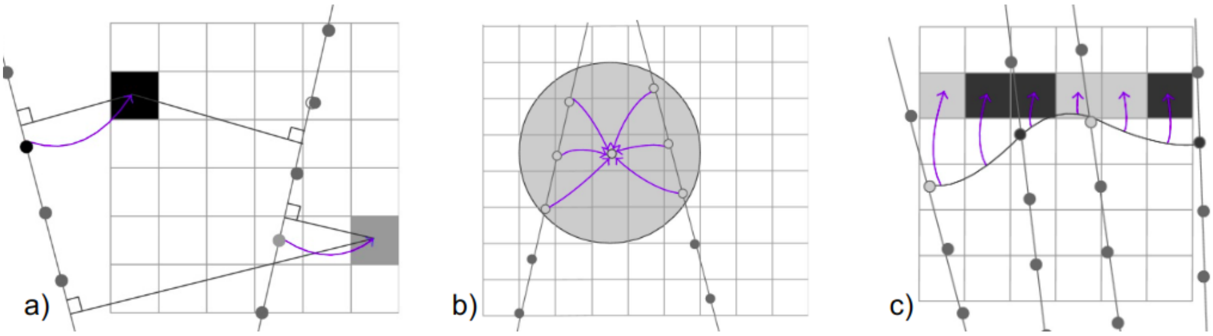


Figure 1.7 – Types of compounding methods, images taken from [11]: a) Voxel nearest neighbor, b)Squared distance weighted interpolation, c) Functional interpolation.

introduced by Gee *et al.* [22] assigns values to voxels based on the nearest 2D image pixels. Voxel-based methods with Interpolation (VBMI) proposed by Trobaugh *et al.* [71] determine voxel values through the interpolation of several nearby pixels. The intensity of a voxel is computed as the weighted average of neighbouring pixels with a distance-weight technique, with the weight being the inverse of the distance from the pixel to the voxel. Coupé *et al.*. [72] introduced an enhancement of VBM by estimating the probe trajectory, thus identifying and weighting intersecting points between the nearest B-scans in time to allocate intensity values to voxels.

Pixel-based methods (PBMs) consist of a distribution stage and a Gap-filling stage. The distribution stage interpolates from pixels in the images to voxels in the volume. The Gap-filling stage interpolates from filled voxels in the volume to empty voxels as can be represented with Figure 1.7-b.

First in the distribution stage, we find early methods like the Pixel Nearest Neighbor Interpolation (PNN) [73] or more advanced methods such as the Pixel trilinear interpolation (PTL) [74], the square distance weighted [75] method, the adaptive squared-distance-weighted [76] method, and the Gaussian distance weighted method [77] that include kernels and median-filter strategies for noise reduction and better edge preservation to enhance image quality. In the second stage, the Gap-filling, we find strategies using bilinear interpolation between non-empty voxels or applying various shaped kernels (sphere, ellipsoid [78], etc.) to either filled or empty voxels. Techniques range from simple methods, such as filling with the nearest nonempty voxel or averaging, to more computationally intensive methods involving the computation of weighted averages [75] or utilising normalised convolution with adaptable kernels [79]. Deciding the optimal kernel size remains a critical aspect to prevent over-smoothing or persisting gaps in the final volume.

Function-Based Methods (FBMs) fill the missing voxel intensities with finer details interpolating functions, see Figure 1.7-c. The Radial Basis Function [80], for example, assumes data smoothness over several B-scans, giving high values to points close to the centre and low values to points farther away. The RBF starts with a central point, and its value decreases as one moves away from that point, following a certain mathematical pattern (usually a Gaussian Bell curve). The Rayleigh interpolation technique [81] employs a Bayesian statistical method to fill the voxels, enhancing the overall image resolution progressively by incorporating finer details incrementally.

The volumes utilised in Chapter 2 were generated employing a pixel-based approach, featuring trilinear interpolation complemented by a Gaussian kernel filling method with a size parameter of 5. Among the alternatives presented above, this approach exhibited superior quality, as assessed by expert evaluations.

1.3.6 Challenges specific to 3D ultrasound

Challenges specific to 3D ultrasound imaging arise from the nature of this technology. In 3D ultrasound imaging, the complexity lies in handling volumetric data and accurately localising structures within the volume. Similar to 2D ultrasound, 3D ultrasound images can exhibit issues such as blurred boundaries, noise, and artefacts, but the challenges are amplified in three-dimensional space. Furthermore, the presence of speckle noise in 3D ultrasound makes image interpretation and analysis more intricate. While there are datasets with images and tracking already acquired, inherent sources of error, like tracking, cannot be eliminated. Ultrasound has a high reliance on operators with varying levels of expertise resulting in substantial differences in image quality, leading to challenges in standardisation and consistency. 3D ultrasound datasets are typically small, not open-source, and contain few annotations, necessitating more representative and generalised samples. Datasets are often annotated by a single expert, introducing potential bias and favouring the selection of high-quality images, exacerbating image quality operator dependency. Addressing these challenges is crucial for improving the accuracy and reliability of 3D ultrasound datasets needed for training AI methods.

1.4 Conclusion

In conclusion, this chapter delved into various aspects of 2D ultrasound imaging generation, highlighting its advantages, applications, acquisition principles, and potential challenges. We also explored the importance of datasets, acquisition protocols, and open-source datasets for training and research purposes. The transition to 3D ultrasound acquisition was discussed, covering tracking systems, methods for tracking ultrasound probes, alternative rotation representations, and compounding techniques. Moreover, we addressed the specific challenges associated with 3D ultrasound imaging. This comprehensive examination provides the foundation for understanding the complexities and nuances of ultrasound imaging, setting the stage for further research and exploration in the field.

Chapter 2

Deep-learning approaches for ultrasound image analysis

2.1 Computer-aided ultrasound image analysis

Deep-learning techniques utilised in the field of medicine have emerged as valuable assets for medical professionals. Nowadays, radiologists employ Computer-Aided-Diagnosis (CAD) tools to enhance performance feedback [82] and/or personalise patient care [83]. Deep-learning algorithms have been developed for ultrasound to enhance image acquisition, evaluate image quality, offer an objective diagnosis, and optimize clinical workflows [24]. Some tools focusing on decreasing the time to perform certain medical procedures [84], can be easily integrated in the routine, and some have received FDA approval [85]. However, for clinical use, the validation or correction of experts is still often recommended, handing over the final decision and responsibility to the radiologist [82].

Deep learning methods for ultrasound image analysis have been applied to at least six different tasks, following the classification of Liu et al. [24]: i) segmentation, ii) detection/Localisation, iii) classification, iv) registration, v) image enhancement, and vi) 3D reconstruction. In terms of medical applications, there has been significant interest in the detection of nodule lesions in the thyroid, breast, and prostate due to the potential of such algorithms to assist in early and accurate diagnosis, non-invasive screening, promising improved patient outcomes and personalised medicine approaches. Anatomies under study include fetuses, soft tissues such as the breast, liver, heart, small structures like nerves, veins, arteries, and even bone surfaces for needle insertion. Clinical tasks on the above anatomies can be divided into bio-metric measurements, therapy follow-up, computer diagnosis, and image-guided interventions.

More concretely, in the past years, ultrasound deep learning methods have shown effectiveness in different clinical tasks such as the diagnosis of hepatic fibrosis [86, 87], focal liver lesions detection [88], spine vertebra guided intervention [89], diagnosis of fatty liver disease [90], classification of benign or malignant tumours [91], identification of plaque obstruction in the carotid artery [92], identification of anatomical plane for abdomen [93], and selection of a quality frame in videos [94]. During this thesis, we focused on the segmentation methods for diseases requiring volumetric measurements, such as Duchenne muscular dystrophy and Hyperthyroidism.

2.2 Semantic segmentation with deep-learning

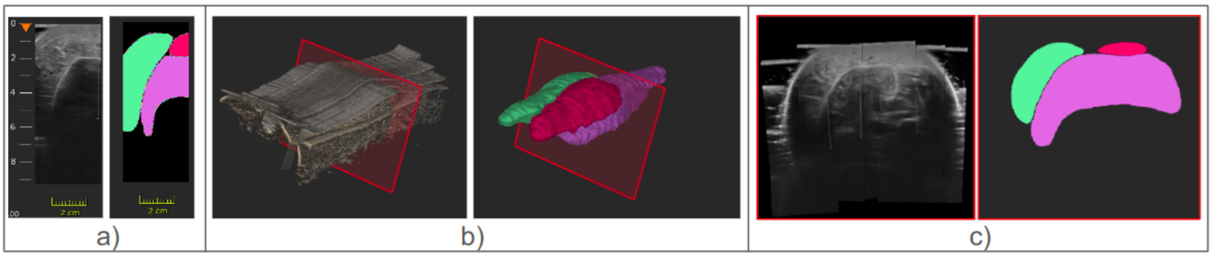


Figure 2.1 – Semantic segmentation of US volumes: a) 2D ultrasound images and annotations, b) ultrasound volumes and labels, c) volumes represented as a stack of 2D ultrasound slices with annotations.

Semantic segmentation involves assigning a specific label or category to each pixel in an image, thereby partitioning the image into semantically meaningful regions, as can be observed in Figure 2.1-a). When applied to 3D volumes, voxels get assigned values, see Figure 2.1-b). Volumes and annotations can also be represented as a stack of 2D images ($\mathbf{V} = \{\mathbf{X_1}, ..., \mathbf{X_N}\}$) with 2D annotations ($\mathbf{L} = \{\mathbf{Y_1}, ..., \mathbf{Y_N}\}$) as presented in Figure 2.1-c).

In the following, we formalise the segmentation problem for 2D images. We will then review existing methods for 2D, 2.5D, and 3D ultrasound images. In the context of semantic segmentation with deep learning, a neural network can be represented as a function that takes as input image $\mathbf{X}$ and produces as output a segmentation map $\hat{\mathbf{Y}}$. The function $\mathcal{F}$ simultaneously learns to extract features from images and to perform the segmentation task. With $\mathbf{X} \in \mathcal{R}^{W,H,K}$ and $\hat{\mathbf{Y}} \in \mathcal{R}^{W,H,C}$, where $\mathbf{K}$ represents the number of channels of the input image (e.g. three for RGB) and **C** represents the number of channels equivalent to the number of segmentation classes (e.g. one for binary classification). The network function $\mathcal{F}$ is parameterized by a set of weights and biases Θ, which are learned during the training process. These learned weights and biases define the network's architecture and its ability to map input data to the desired output.

$$\hat{\mathbf{Y}} = \mathcal{F}(\mathbf{X}; \Theta) \tag{2.1}$$

Training involves two main steps: the forward pass and the backward pass.

The Forward Pass captures features from the input image and makes segmentation predictions. Many segmentation networks encode the features by reducing the image size while increasing the channel dimensionality. They decode the features and extract meaningful information at different scales. The most common architecture includes encoding and decoding blocks, each composed of convolutional or linear layers, activation functions, residual connections, gated layers and pooling layers. Examples of such layer blocks are presented in the well-known UNet architecture [48], with one of the variants Attention UNet [45] presented in Figure 2.2.

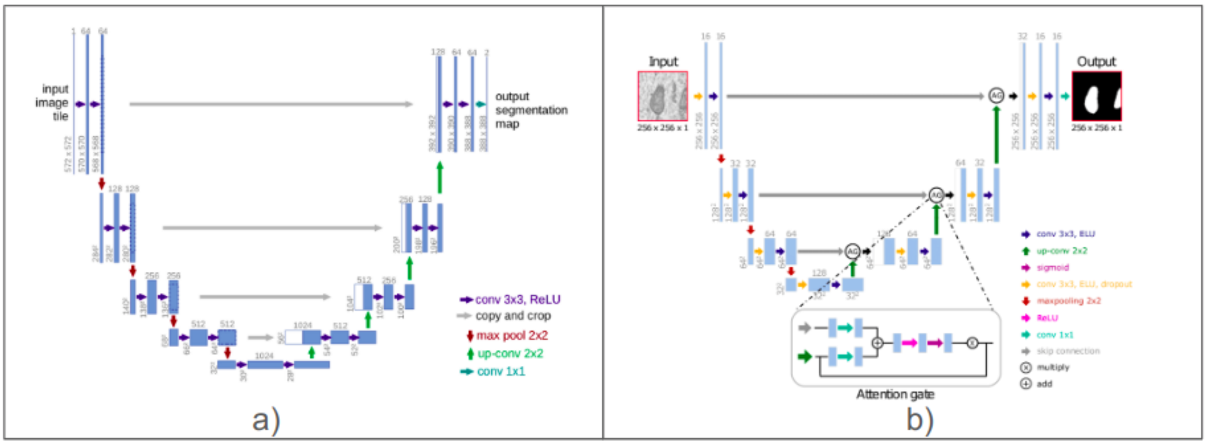


Figure 2.2 – Deep-learning layers: a)UNet architecture [48], b) Attention UNet [45], image taken from [95].

The backward pass consists of training the network to learn the optimal values for its weights and biases (Θ) by minimising a loss function. This is typically done through back-propagation and gradient descent optimisation algorithms, such as Adaptive Moment Estimation (ADAM) [96], Stochastic Gradient Descent (SGD) [97], Mini-Batch Gradient Descent [98], etc. A loss function ($\mathcal{L}$) is calculated between the predicted segmentation map ($\hat{\mathbf{Y}} \in \mathcal{R}^{W,H,C}$) and the ground truth segmentation map ($\mathbf{Y}_{gt} \in \mathcal{R}^{W,H,C}$), with (i, j) representing the pixel coordinates and **c** representing the channel. The loss is aggregated for more than one image in the forward pass. We call the group of images a batch. One common loss for semantic segmentation tasks is the pixel-wise cross-entropy loss, calculated as:

$$\mathcal{L}_{i,j} = -(\frac{1}{C})\sum_{c=1}^{C} \mathbf{Y}_{gt_{i,j,c}} \cdot \log(\hat{\mathbf{Y}}_{i,j,c}) \tag{2.2}$$

In a nutshell, the roles of different components in an UNet are:

- *Convolutional layers* perform feature extraction by applying learned filters to input data.
- *Activation functions* introduce non-linearities to the neural networks, enabling them to model intricate relationships within the data. Common activation functions like ReLU, sigmoid, and tanh determine whether neurons activate or not based on their input.
- *Pooling layers* downsample feature maps to reduce computational complexity, expand the receptive field, and enhance robustness to variations.
- *Batch normalization layer* are used to improve the training speed and stability by normalizing the inputs of each layer within a mini-batch, reducing internal covariate shift and enabling faster convergence.
- *Up-convolutional layers* increase the spatial resolution of feature maps effectively expanding their size to match that of lower-resolution feature maps from earlier layers.
- *Skip connection concatenation* preserves semantic information while up-sampling.

The above operations are repeated for multiple layers in the encoder and the decoder, gradually decreasing and increasing the spatial dimensions to match the original image size. The final layer produces the predicted segmentation map $\hat{\mathbf{Y}}$, which is often post-processed with thresholding or morphological operations to produce better binary masks.

2.3 Segmentation of ultrasound images architectures

The following subsections review existing deep learning models addressing the semantic segmentation problem on 2D or 3D medical ultrasound images.

2.3.1 Early 2D Segmentation methods

We first review some early works on 2D segmention. One of the first papers on 2D Ultrasound (US) segmentation was presented by Zhang *et al.* [99], focusing on the pixel-wise detection of lymph nodes in the neck. The model consisted of two fully connected layers trained and evaluated on 80 2D ultrasound images. Wu *et al.* [37] addressed the segmentation of fetal head and liver relying on a dataset of size 900 fetal head and 688 abdominal images. In 2018, this work was extended [34] to the fetal brain using a hybrid of UNet and a SegNet [100], compared with three more classical pipelines of three segmentation architectures: UNet [48], SegNet [100] and Pix2Pix [101]. Using only 337 images, Kumar *et al.* [43] segmented breast masses with an ensemble of 10-UNets trained with different initialisation and data orderings. Most of the early works rely on the variants of the UNet, focus on 2D ultrasound images and use datasets of modest size for training.

Regarding the segmentation of muscles, which is one of the main focuses of this thesis, Cunningham *et al.* [38] proposed segmenting cervical muscles during head motion in 2019. It contains around 1100 hand-segmented images in 14 categories (e.g., Skin, Trapezius, Splenius). However, datasets of this size are expensive to acquire and rarely open-sourced. The acquisition and annotation of ultrasound datasets are major limitations for training 2D and 3D muscle segmentation architectures. This is why clinical and sports studies still rely on manual segmentation. Given such limitations, some works address the problem of segmenting ultrasound images with small training datasets [42], synthesising US images from various segmentation masks with a generator or with transformers blocks on 2D patches of the images [39]. We develop the state of the art regarding this challenge and propose an innovative solution in chapter 2.

2.3.2 2.5D deep-learning ultrasound segmentation methods

Given the dynamic nature of ultrasound, several researchers have explored using videos to enforce segmentation with smoother frame transitions. In 2018, Mishra *et al.*. [40] proposed a unique training scheme and fusion layer tailored to prevent fragmented boundaries. The team trained the architecture on two distinct tasks: the segmentation of lumen regions and a vessel segmentation dataset comprised of 69 US images. This approach was further extended in 2019 [102], wherein instead of videos, the training and evaluation of the network was performed on 144 Carotid 3D ultrasound volumes processed in a sliding window manner. This approach facilitated the quantification of carotid plaque and enabled dynamic fine-tuning [1].

In a similar vein, other works have also explored viewing 3D volumes as a series of sequential slices. Pourtaheiran *et al.*. [103] enhanced needle detection on 2D patches of the 3D volume, performing classification and semantic segmentation. Yang *et al.* [35] introduced a Direction-Fused Fully Connected Network (DF-FCN) architecture for enhanced catheter detection in cross sections, leveraging the Quicknat [104] architecture. The network extracts feature maps from three distinct directions of the volume (axial, sagittal,

[1] Fine-tuning involves the process of retraining a pre-existing neural network model on the new dataset with a relatively small learning rate, allowing the model to adapt its parameters to the specific characteristics and patterns present in the new data while preserving the knowledge and features learned from the original training dataset.

and coronal) and fuses the three predictions, capitalizing on the robust capabilities of the pre-trained 2D model, thereby achieving superior results in medical imaging tasks.

Similar to them, in this thesis, we evaluate handling high-resolution volumes as sequential slices, enforcing smoothness of labels with different losses, reducing computation burden and accelerating inference. More details are presented in chapter 2.

2.3.3 3D ultrasound segmentation methods

By 2020 [2], many of the papers used pure 2D images, and architectures were normally Fully convolutional networks or encoder-decoder architectures. However, 3D ultrasound volumes have the potential to increase the field of view. These volumes can be obtained with different acquisition methods explained in Chapter 1.3.5.

In 2015, Ghesu *et al.* [30] proposed the Marginal Space Deep Learning framework (MSDL) to perform anatomical pose estimation and boundary delineation on 3D ultrasound volumes of the aortic valve. Such pipeline first classifies subvolumes as containing or not the anatomy of interest, and second localizes with a bounding box the sought structure to finally estimate the non-rigid object boundary with an active Deep-learning shape model. Their extensive dataset contains 2891 volumes from 869 patients with a 3D ultrasound probe. In 2017, Yang *et al.* [105] segmented 17 trans-rectal ultrasound volumes: using Recurrent neuronal networks (RNN) and shape priors to improve boundary inference. The method was compared against a VGG16 2014 [106] model pre-trained on Imagenet [107]. The same year, Yang *et al.*extended their work to 104 prenatal volumes [108] with simultaneous semantic segmentation of the fetus, the gestational sac, and the placenta. The approach was compared against other methods that had been used for 3D segmentation but not for ultrasound: Auto-context: a stack of 3-fully connected networks [109] and conditional Generative Adversarial Networks (GCN) [110]). More recently, Lei *et al.*presented the DaF3D architecture for prostate segmentation [111]. The method is applied to a dataset containing 44 3D ultrasound volumes and compared against V-net(2016) of Milletari *et al.* [112]. In 2023, Li *et al.* [36] presented their ATTransUNet architecture for 3D thyroid segmentation. Their major contribution is the attention module using the entropy of the probability map between the classes. The method was compared against: UNet [48](2015), Unet++ [113](2018), Axial-Attention [114](2019), transUnet [115] (2021), UTNet [116](2021), Swin-UNet [117] (2023).

Among the most recent 3D ultrasound segmentation works, we find models based on transformers. Transformers blocks consist of a stack of self-attention layers, where each layer processes input data by computing weighted combinations of all input elements, allowing the model to capture complex relationships and dependencies in the data, both locally and globally. In recent years, many state-of-the-art segmentation architectures have adopted hybrid designs that combine traditional Convolutional Neural Network (CNN) structures, such as UNet, with Transformer blocks. While UNet primarily relies on convolutional layers to process spatial information hierarchically, transformer blocks excel in modelling long-range dependencies by utilising self-attention mechanisms. This hybrid approach marries the strengths of both architectures, enabling segmentation models to simultaneously capture local and global context.

In this thesis, we proposed two methods in 2020 and 2021: The UNet-S-R-CLSTM and IFSSnet. Both methods perform muscle segmentation in 3D volumes of the lower limb; more details can be found in chapter 2. We compare our methods with 3D UNet [48](2015), V-Net [112](2016), DAF3D [44](2019) and PG-Net [118] (2018): An architecture usually used for label video propagation on natural images. We additionally perform some studies of label variability and border completition in the UNET [48], attention-UNet [45], and UNet transformer [115] architectures.

2.4 Challenges of deep-learning-based segmentation methods on 3D ultrasound images

3D ultrasound imaging has several advantages. Firstly, 3D increases the field of view compared to 2D ultrasound, which captures a limited plane at various angles, making it challenging to reproduce the same plane for follow-up studies. Secondly, 3D volumes are easier to interpret and do not require the operator's mental integration of multiple images. Lastly, 3D allows accurate estimation of organ or tumour volume for diagnosis and treatment decisions. Despite the advantages, it is worth noting that 3D ultrasound segmentation is challenging. Compared to CT and MRI, which are standardized modalities with well-defined edges, patient position-independent, and do not struggle to scan regions of the body containing gas and bones, ultrasound volumes impose specific challenges to deep-learning-based segmentation methods.

First, images suffer from operator dependency, leading to variations in image appearance due to differences in scanning techniques and operator expertise. Second, Datasets are small and normally not open-source compared to natural image datasets. Acquiring large and diverse datasets for training deep learning models is challenging, as medical data is limited and subject to strict privacy regulations. Nevertheless, variability is needed for the models to be generalised. Normally, models suffer to detect outliers or images on which they were not trained. Third, annotations are challenging, expensive, and time-consuming because ultrasound images often exhibit lower quality compared to other modalities, with issues such as noise, artefacts, and blurry boundaries, making precise segmentation more challenging. Datasets are normally annotated by one single expert, generating a high possibility of bias and potentially skewed selection of the best-quality images for the training set, making image quality operator-dependent. Fourth, 3D segmentation architectures for comparison are often not designed for ultrasound, or when designed, models work on specific ultrasound datasets that rarely are made public, making results difficult to replicate.

2.5 Segmentation metrics

In this thesis, mainly seven segmentation metrics were used [119]: Sørensen–Dice index (Dice), Intersection over Union (IoU), Positive Predicted Value (PPV), recall or True Positive Rate (TPR), Miss Rate (False Negative Rate (FNR)), Hausdorff distance (HD) and Average surface distance (ASD). These metrics evaluate the closeness of the network prediction to the expert's segmentation, either in terms of area or volume or by evaluating their contours. HD and ASD metrics are better with values closer to 0 [e.g. millimetres], while all the other metrics present their minimum value at 0 and their best value at 1 [pixel percentage].

The **Sørensen–Dice index** and the **Intersection over union IoU** measure the overlap between the predicted and ground truth regions. IoU is primarily used to evaluate detection results. They quantify the similarity as the proportion of pixels in the overlapped region with their sum (Dice) or their union (IoU). Being $\mathbf{Y}$ the ground truth label and $\hat{\mathbf{Y}}$ the prediction, the scores can be measured as:

$$\begin{aligned} DSC &= \frac{2|\mathbf{Y} \cap \hat{\mathbf{Y}}|}{|\mathbf{Y}| + |\hat{\mathbf{Y}}|} \\ IoU &= \frac{|\mathbf{Y} \cap \hat{\mathbf{Y}}|}{|\mathbf{Y} \cup \hat{\mathbf{Y}}|} = \frac{|\mathbf{Y} \cap \hat{\mathbf{Y}}|}{|\mathbf{Y}| + |\hat{\mathbf{Y}}| - |\mathbf{Y} \cap \hat{\mathbf{Y}}|} \end{aligned} \tag{2.3}$$

Here, $|\mathbf{Y}_{gt} \cap \hat{\mathbf{Y}}|$ represents the intersection between the ground truth and the prediction. A perfect DSC score of 1 indicates that the prediction precisely matches the ground truth.

Class imbalance and small regions pose significant challenges when evaluated with DSC and IoU. Class imbalance occurs when one class vastly outnumbers the others, making it difficult for the metrics to accurately assess model performance. In such cases, even if a model performs exceptionally well on the majority class, it may fail to adequately capture the minority class, resulting in an artificially high overall score. Small regions, on the other hand, can disproportionately influence the Dice Score and IoU to be highly sensitive to minor spatial discrepancies. Consequently, it is crucial to be mindful of these challenges when interpreting these metrics.

Precision (Positive Predictive Value Precision or positive predictive value (PPV)), **Recall** (True Positive Rate TPR), and **Miss Rate** (false negative rate FNR) are calculated using a confusion matrix [120] at a 0.5 threshold of the probabilistic predictions. Precision evaluates the correct proportion of positive predictions; Recall the proportion of true positives successfully detected by the model relative to the total number of positive instances; and the Miss-rate measures the proportion of false negatives in the prediction. Formally, these scores can be calculated as follows:

$$\begin{aligned}\text{Precision} &= \frac{\mathbf{TP}}{\mathbf{TP}+\mathbf{FP}},\\ \text{Recall} &= \frac{\mathbf{TP}}{\mathbf{TP}+\mathbf{FN}},\\ \text{Miss}-\text{rate} &= \frac{\mathbf{FN}}{\mathbf{FN}+\mathbf{TP}}\end{aligned} \tag{2.4}$$

Where **TP** represents the sum of true positive pixels (correctly predicted positives), **TN** the sum of true negatives pixels (incorrectly predicted negatives), **FP** the sum of false positives pixels (incorrectly predicted positives), and **FN** the sum of false negatives pixels (incorrectly predicted negative).

Hausdorff distance (HD) and **Average surface distance (ASD)** measures the maximum and average distance between the contour points of the ground truth $\mathbf{Y_c}$ and the predicted label $\hat{\mathbf{Y}}_\mathbf{c}$, in 2D (respectively the surface in 3D).

4D provides valuable insights into the border dissimilarities and can help identify outliers or significant disparities. A lower HD and ASD value indicates a better segmentation result. ASD provides an alternative to HD evaluation criteria, considering the distances over the entire border rather than just the maximum distance HD. With δ corresponding to the Euclidean distance between boundary pixels, formally the HD and ASD are defined as:

$$\begin{aligned}HD\left(\mathbf{Y_c},\hat{\mathbf{Y}}_\mathbf{c}\right) &= \max\left\{\begin{array}{c}\sup_{a\in\mathbf{Y_c}}\delta\left(a,\hat{\mathbf{Y}}_\mathbf{c}\right),\\ \sup_{b\in\hat{\mathbf{Y}}_\mathbf{c}}\delta\left(\mathbf{Y_c},b\right)\end{array}\right\}.\\ ASD\left(\mathbf{Y_c},\hat{\mathbf{Y}}_\mathbf{c}\right) &= \frac{\sum_{a\in\mathbf{Y_c}}\delta\left(a,\hat{\mathbf{Y}}_\mathbf{c}\right)+\sum_{b\in\hat{\mathbf{Y}}_\mathbf{c}}\delta\left(\mathbf{Y_c},b\right)}{\left|\mathbf{Y_c}\right|+\left|\hat{\mathbf{Y}}_\mathbf{c}\right|}\end{aligned} \tag{2.5}$$

2.6 Conclusion

This chapter presents the state-of-the-art 2D, 2.5D and 3D ultrasound deep-learning segmentation networks, with a description of the architectures, the metrics used for evaluation and their contribution. We study the models for ultrasound medical applications, particularly in areas like the segmentation of muscle tissues in diseases requiring volumetric assessments. We observe the need for extensive datasets of high-quality and accurately annotated data as a challenge in common for all the architectures. We used this knowledge as a baseline for our research, which also needed to solve the training of networks with small ultrasound datasets and sparse annotations for muscles segmentation.

Part I

Building 3D ultrasound annotated datasets

Abstract

This part aims to assist in creating reliable freehand annotated ultrasound volumes suitable for training deep-learning segmentation methods, thereby accelerating 2D and 3D quantitative ultrasound measurements such as the volume. To this end, we propose to rely on sparse annotations, which significantly reduces annotation time. These annotations are done on the original 2D B-mode images with higher resolutions, but despite their partial field of view, we proceed to compound both 2D images and annotations into volumes later.

The next two chapters describe two approaches towards improving the creation of such 3D training datasets. The first study revolves around the exploration of interpolation methods aimed at generating smooth 3D labels from sparse 2D annotations. This work involves various non-deep-learning-based interpolation and seed propagation techniques, and it was published in SIPAIM symposium in 2020 [69]. In our second study, we shifted our focus to the enhancement of freehand ultrasound acquisitions, searching to reduce the need for tracking systems and to improve the overall quality of ultrasound volumes. This study involves evaluating a range of methods to learn to predict the probe motion from a sequence of 2D US images.

Both studies are experimental. The results of the first study have helped us to create an ultrasound volumetric dataset characterised by high-resolution and precise ultrasound volumes with annotations, and we will be later in this thesis in chapter 1.5. The results of the sensorless tracking approach were not sufficiently conclusive to gather quantitative biomarkers, so for the rest of the thesis, we will consider ultrasound volumes reconstructed with the help of a tracking system.

Clinical motivation

As the introduction mentions, quantitative volume measurements are important in several clinical and sports applications [121]. Applications include diagnosis and monitoring of tendon swelling [4], fluid accumulation [8], and tumour growth [5, 6]. In kids suffering from Duchenne muscular dystrophy [122], volume calculation provides a tool for follow-up on the progress of treatment. In this part of the thesis, we focus on data from subjects in sports, where the volume is known to be a biomarker for detecting improvement of Achilles tendon treatment [51] because physicians rely on 3D freehand ultrasound sequences to perform such volume measurements. This process still requires manual, tedious, and time-consuming annotations. In order to create an ultrasound dataset of the lower limb with high-resolution images and accurate annotations. In this order, we performed 2 different studies addressing specific challenges.

The first study employs sparse annotations on high-resolution 2D B-mode images, which are then expertly compounded into 3D volumes. Unlike static 2D annotations, 3D models provide a comprehensive view of the structure's volume, offering quantitative insights which could be exploited for treatment planning. However, generating 3D labels from freehand ultrasound presents unique challenges. It's important that annotations align seamlessly with the ultrasound image, maintaining model smoothness and accuracy. Our research focuses on methods that balance image adherence with interpolation smoothness.

The second study concentrates on enhancing ultrasound acquisition searching to reduce the need for tracking systems and to improve the overall quality of ultrasound volumes. In fact, for large structures such as the lower limb muscles, the freehand sequences consist of several sweeps, which can accumulate tracking errors, preventing the correct overlap of anatomical structures during compounding. Accurate probe tracking is important for producing clear, consistent ultrasound images, capturing precise details of anatomical structures, and defining borders accurately. Currently, optical tracking systems require patients to be brought to specially equipped rooms, limiting one of ultrasound's key advantages—portability. Our goal is to restore this portability without compromising image quality. In section 2, we investigate learning base approaches to improve acquisitions or remove the need for it to restore portability.

Through these studies, we search to obtain high-quality, reliable ultrasound datasets, ensuring higher resolution, accuracy, and patient comfort, thereby contributing significantly to the advancement of medical diagnostics and treatment strategies.

Chapter 1

High-resolution lower limb ultrasound dataset

This chapter focuses on the creation of a 3D ultrasound high-resolution dataset from B-mode images with optical tracking. The original data contains 2D sparse annotations of three lower limb muscles: the gastrocnemius medialis (GM), the gastrocnemius lateralis (GL), and the Soleus (SOL). This chapter aims to provide 3D smooth and consistent muscle annotations. Different non-deep-learning techniques were used to propagate the seeds of the sparse labels, and results were evaluated in terms of volumetric error, Dice score, and Hausdorff distance. Among the evaluated methods, the "zero-order interpolation (ZOI)" leads to the best results. Our ZOI method further relies on a 3D-3D image registration approach for merging labels from distinct ultrasound scans of the same leg, each with varying quality. Deeper details can be found next in this chapter.

1.1 Related work: label-transferring and seed interpolation

A first challenge with our dataset is the availability of only 2D sparse manual annotations. Indeed, the muscle boundaries can be outlined on either high-resolution partial B-mode images or 2D slices derived from a 3D image after compounding. However, meticulously delineating all slices within a full sequential acquisition or even in a volume proves to be excessively time-consuming and unfeasible, particularly when dealing with numerous patients. Experts prefer to rely on label interpolation techniques to reduce the annotation time. Numerous software solutions provide interactive or semi-automatic segmentation tools that enable the propagation of initial annotations or seed points. However, despite these tools, the involvement of clinical experts remains necessary to initiate the seeds and subsequently refine the segmentation post-automatic propagation. Software platforms like Slicer [23], Imfusion [123] and Stradwin [124], include non-learning-based built-in functionalities such as "Fill Between Slices," "Grow from seeds," and "Watershed" for label propagation.

Propagation methods can be categorised into two groups. The first group involves propagation exclusively over binary label masks, while the second group incorporates the image content in conjunction with the label seeds. Among label-interpolation strategies, the "Fill Between Slices" technique [125] from Slicer3D stands out as an iterative morphological contour interpolator. This method employs morphological dilation operations to create a gradual alteration of the binary object's mask. Correspondingly, the "maximal disc-guided interpolation" method [126] in Stradwin facilitates interpolation by establishing a surface through sparse, non-parallel labels. Both of these methods [125,126] conduct interpolation on the 3D volume post-reconstruction. In contrast, and following an experimental comparison, our proposition consists of a simple zero-order interpolation between two partial expert annotations carried out on 2D B-mode images, spanning across the sweep to achieve masks with higher resolution.

In the second group, we encounter methods that incorporate the image content. Beyond ensuring a seamless transition between annotated slices, these methods search to align the boundaries of the segmented region with the contours present in the image. Included within this category are approaches such as watershed [127] and graph-cut [128]. The majority of these embedded techniques assume uniform areas of interest and well-defined image contours. However, such assumptions do not hold true for muscle segmentation within ultrasound (US) images, leading to issues like leakage. In the absence of specialised adjustments, the standard built-in implementation necessitates a significant number of labelled background and foreground seeds. In order to build an accurate 3D fully annotated dataset for training a deep neural network for segmentation, we perform a series of experiments comparing the above tools and methods, as we describe in the following sections.

The second main challenge of our dataset comes from visualising the boundaries of superficial and deeper muscles, which require different acquisition parameters, like frequency, point of focus, or windowing, due to their anatomical differences. Superficial muscles closer to the skin require high-resolution imaging with less penetration, as they are less obscured by tissues above. In contrast, deeper muscles, buried beneath layers of tissue and surrounded by bones or organs, need stronger penetration for effective imaging. In practice, two sweeps with different frequencies are required, and only partial annotations (some muscles) are available for each acquisition. An additional challenge is, therefore, to build a dataset of 3D volumes with full annotations (all muscles in all slices). Few works have discussed the problem of annotations across different frequencies. Inhatsenka *et al.* [129] evaluate the efficacy of annotating ultrasound acquisitions from different parameters. Yoshizumi *et al.* [130] introduced a strategy involving multiple-frequency ultrasonic imaging to enhance image resolution by combining and blending diverse images with varying frequencies. This approach leads to increased visibility of additional structures, thereby aiding the segmentation task. While effective, this method alters the content of the image itself. Instead, we choose the most suitable frequency for manual segmentation of each individual muscle. To facilitate the transfer of annotations between the two sequences, we rely on mono-modal image-based rigid registration [131]. Our proposal entails utilising 3D image registration to transfer manual annotations conducted on the most evident sequence to the other acquisition, where the muscle is less distinct yet still present. Consequently, we can create a volume with full labels for each acquisition, potentially enhancing the database variability.

1.2 Dataset of sparse annotations

The dataset employed in our study originates from the work of Crouzier *et al.* [51]. In this dataset, participants assume a prone position with their lower limb immersed in a custom-designed water bath (Figure 1.3-a). During the recordings, two sets of US freehand image sequences are captured with varying parameters. To cover the region from the knee to the ankle, four to six parallel sweeps are executed, and the probe's movement is tracked using optically reflective markers. As the probe moves orthogonal to the image plane, B-mode images (with dimensions 3.9 cm in width and 9.5 cm in depth ±1.2) are recorded with a 5 mm displacement interval at a low speed. By utilizing the tracking matrices of the probe, 3D ultrasound volumes were compounded, resulting in a voxel grid (Figure 1.3-b) of $564 \times 632 \times 1443 \pm (49 \times 38 \times 207)$, with a pixel size of 0.276993 mm/pixel ± 0.015. Sparse annotations of each muscle were done in 2D images (Figure 1.3-c). From the original pool of 44 participants, we select a subset of 15 participants, each with two recordings denoted as x1 and x2. For this specific group, we ask an expert to perform full 3D muscle annotations (Figure 1.3-d), which we use to evaluate our methods.

1.3 Method

Here, we describe the retained processing pipeline to build our annotated dataset from sparse and partial annotations. In the experiments subsection, we will show the quantitative evaluations that guide our choice. An overview of the method is presented in Figure 1.2. The initial step involves 3D-3D monomodal image-based ultrasound rigid registration at different frequencies (Figure 1.2-a). The second step gathers partial annotations on each acquisition from specific high-resolution 2D B-mode images (Figure 1.2-b). The third step consists of a zero-order label propagation, copying the smallest mask onto

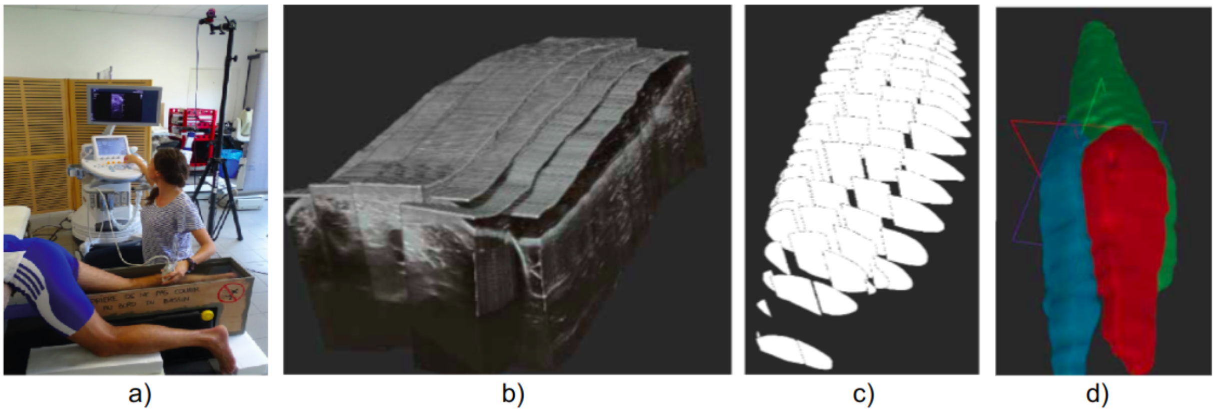


Figure 1.1 – 3D ultrasound dataset acquisition: a) Custom bad setup with cameras for optical tracking b) 3D ultrasound volume reconstruction using optical tracking c)Sparse 2D muscle seeds d) Full 3D expert annotations

all the images that lack labelling situated between the two masks (Figure 1.2-c). The fourth step applies a 7x7 Gaussian smoothing filter to the compounded interpolated US volume, culminating in a refined muscle mask (Figure 1.2-d).

By isolating the process of annotating organs situated at distinct depths and subsequently transferring these annotations from the more suitable acquisition to the less, we obtain better 3D masks. The proposed approach, consisting of several simple stages, has the advantages of being fast to implement and highly reproducible while substantially alleviating the load for experts. Also, as we later demonstrate, the resultant datasets are suitable for training deep learning models.

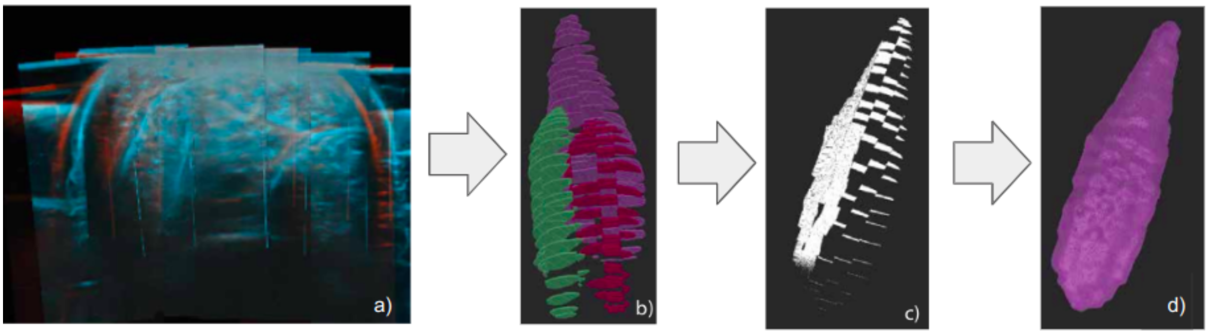


Figure 1.2 – a)3D-3D Ultrasound registration. b) Partial annotations from different acquisitions. c) zero-order label propagation d) Gaussian smoothing filtering.

Image-based Rigid Registration was employed to align the two reconstructed ultrasound (US) volumes derived from a single participant but acquired with different frequencies. This alignment ensures that similar structures are positioned within a unified grid in the same spatial location. The registration procedure was executed using ImFusion software with rigid registration, although alternative comparable tools can also serve this purpose, such as ITK or Slicer3D. Upon successful registration, the muscle structures overlap alignment, allowing for segmentation to be carried out interchangeably on either of the registered volumes.

The Ground truth test dataset was done with full manual expert annotations, consisting of a slice-by-slice segmentation of the cross-sectional area (CSA). Positioned on the "sagittal" plane, the annotator scrolls through each structure in an ascending and descending manner, aiming to comprehend how muscles evolve and establish connections with anatomical knowledge.

To ensure consistency across masks generated by different annotators, the following guidelines were adhered to:

- Identify the brightest structures, such as bones (fibula and tibia) and ligaments.
- Rely on the knowledge that muscles detach from the bone around two-thirds down the low limb.
- Begin annotations on slices with a high level of certainty, employing a 3D spherical brush.
- Whenever feasible, maintain smooth annotations on adjacent slices.
- When encountering unclear boundaries, draw upon shape cues from adjacent anatomical structures for context.
- Trace these structures up to the problematic slice to minimise ambiguities.

Annotators training with the software necessitates approximately 5 hours of adaptation, and segmenting three muscles in ten patients takes around 25 hours.

1.4 Experiments and results of non-deep-learning seeds propagation methods

1.4.1 Qualitative analysis

The implementation of the compared semi-automatic segmentation algorithms, namely Fill Between Slices (FBS), Grow from seeds (GFS), and Watershed (WS), was conducted relying on the open-source software Slicer 3D. FBS complete data between adjacent slices of a 3D volume, creating a continuous and coherent three-dimensional representation from sparse cross-sectional images. Unlike other methods, it requires full annotations of the object in the slides, not just seeds. GFS, on the other hand, is a segmentation technique in image processing where initial "seeds" points are iteratively expanded based on their neighbouring pixels' intensity or colour, effectively segmenting an image into regions based on these starting points. At each iteration, a propagation rule determines whether a neighbouring pixel is to be included in the segmented region or not. In our case, the rule is based on an intensity difference between pixels. The propagation stops when no more pixels can be added to the area. WS treats pixel intensities as topographical features, segmenting the image into distinct regions by simulating how water would flow and accumulate in the landscape. It starts from the lowest intensity pixels, simulating how water fills basins, and creates dividing lines at places where waters from different basins meet, thus defining boundaries or 'watershed lines'. It is applied to the gradient of the image and requires pre-processing steps in the image, like noise reduction, especially in ultrasound, to mitigate common challenges like over-segmentation. The outcomes obtained from these techniques highlight experts' recurrent difficulties when tasked with manual or semi-automatic annotations of ultrasound images. The inherent absence of clearly defined edges and the limited contrast observed between regions of interest contribute to the challenge of accurately delineating segmentation mask boundaries, which are difficulties faced by both experts and the employed methods. Our "ZOI" method, on the other hand, annotates ultrasound datasets using sparse annotations. It begins with 3D-3D rigid registration of ultrasound volumes at different frequencies to align similar structures. Partial annotations are then extended across images using zero-order label propagation, refined using a 7x7 Gaussian smoothing filter and finished with expert corrections.

The qualitative visualization in Fig 1.3 underscores the limitations encountered by the GFS and WS methods, notably leakage issues. These methods struggle to establish coherent boundaries for seed propagation, given the indistinct image contours present in ultrasound images. Meanwhile, the FBS method achieves seamless transitions, but its utilization relies on and necessitates parallel annotations, a requirement absent in our particular scenario for freehand ultrasound, with freedom in the rotation. Our dataset encompasses 44 participants and 59 volumes. Our simpler ZOI method, in contrast, effectively addresses leakage concerns. However, it tends to overly smooth the outcomes, potentially leading to the omission of finer border details.

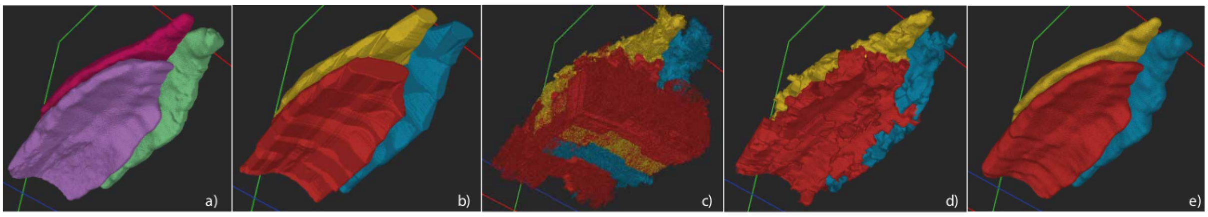


Figure 1.3 – Qualitative results of different interpolation methods: a) Ground truth b) Fill between slices(FBS) c) Grow from seeds (GFS) d) Watershed (WS) e)Our method (ZOI).

1.4.2 Quantitative evaluation

After registration, we quantitatively compounded the introduced mask-based (ZOI-FBS) and image-based (GFS, WS) label propagation techniques. We compare the inter-

Table 1.1 – Quantitative Results averaged over 10 participants

		DICE			IoU			Vol error		
Algo	Min	GL	GM	SOL	GL	GM	SOL	GL	GM	SOL
ZOI	90	**0.937**	**0.946**	0.841	**0.882**	**0.898**	0.791	**2.61**	**4.98**	**4.93**
FBS	96	0.909	0.924	**0.919**	0.835	0.859	**0.851**	10.00	10.70	13.03
GBS	44	0.818	0.755	0.763	0.699	0.615	0.620	24.02	10.61	15.53
WS	120	0.742	0.764	0.803	0.591	0.619	0.671	7.60	7.99	10.52

polated results to those of the fully labelled datasets in terms of three scores: the Dice, the mean Intersection over Union (mIoU), and the volumetric error. The results are reported in table 1.1. The ZOI approach performs better in cases involving smaller muscles, specifically GL and GM. For the Soleus muscle, however, the FBS method yields better results, followed by our approach. Across all muscles, the retained method attains a Dice of 0.908 ± 0.04 and a mIoU of 0.877 ± 0.02.

Comparatively, our semi-automatic method demonstrates a faster execution in comparison to both, the full slice-by-slice segmentation and the FBS method, as evident from Table 1.1. We therefore use this streamlined process as an initialization for the remaining 34 acquisitions with only partial and sparse annotations. The application of the method, followed by expert refinements, requires roughly 50 minutes per muscle. In contrast, this time frame proves significantly more efficient when weighed against the duration demanded by the slice-by-slice technique. Consequently, our approach significantly reduces the time investment required.

Regarding the volumetric accuracy, our method presents a smaller error of 4.17% on average across all the evaluated semi-automatic segmentation methodologies applied to 3D ultrasound volumes. Particularly concerning the deeper soleus muscle, inaccuracies along the segmentation borders exert only a minor influence on the estimated volume. Despite instances of leakage and inadequate muscle delimitation in some methods, the volumetric error remains relatively low due to substantial overlap with the ground truth. However, for reference, inter-expert variability is between 5% and 10% of volumetric error [4]. In this context, it is prudent to consider additional assessments to evaluate the obtained labels' smoothness.

1.5 Conclusion and perspectives

This chapter provided a simple, cost-effective solution, primarily relying on open-source software, for producing high-quality, fully annotated ultrasound 3D datasets. This proposed approach significantly reduces the amount of manual intervention while yielding reasonably accurate outcomes for segmentation and volume computation. The resultant dataset will help us in future chapters in the development of automated machine learning-based segmentation techniques.

Chapter 2

Sensorless freehand ultrasound

A second important aspect in building 3D freehand ultrasound datasets is the fusion of multiple 2D B-mode images into a geometrically unified volume. As introduced in section 1.3.5, this process, known as compounding, relies on the 6 DOFs tracking the probe during a sequential acquisition. Although reliable tracking systems exist today, they can be expensive and more importantly, they reduce the portability of ultrasound imaging. In this chapter, we explore the possibility of reconstructing 3D US volumes from a sequence of images alone, without the need of tracking. This type of acquisition is known as sensorless ultrasound. It offers several advantages. First, enhances cost-efficiency and accessibility by eliminating the need for additional equipment. Secondly, the absence of extra hardware preserves ultrasound portability and ease of use, making it suitable for point-of-care situations, including emergency rooms and remote locations. Finally, the ability to create 3D US images at a reduced cost, improving the probe's original field of view, could also improve guidance for diagnosis interventional procedures.

Before benefiting from these advantages, it is however crucial to ensure the accuracy and reliability of the compounded volumes. Indeed, the absence of sensors can be expected to impact the precision and quality of 3D image reconstruction. The main objective of our work in this direction was to evaluate the possibility of using the Sensorless Freehand 3D ultrasound based on deep-learning approaches to i) Compound large volumes without the need for tracking systems to improve portability, ii) Correct potential tracking errors for predicting a more accurate tracking for the sweeps, to obtain better 3D ultrasound volumes of the low limb. As we will see in the experimental section, our different propositions reached a performance close to the state of the art but were not accurate enough to build reliable datasets for quantitative volume segmentation and quantification.

2.1 Related works on sensorless freehand ultrasound reconstruction

There are four main types of approaches for reconstructing volumetric data from untracked sweeps. Non deep-learning methods [132–135], Deep-learning based methods relying on image pair inputs [14,136–140] against those using temporal smoothing losses or recurrent architectures [141–143], differential rendering methods [144] and deep-learning methods using low-demand tracking data from an inertial measurement unit [15,145–147]. Next, we describe each type in more detail, citing relevant prior work but without being exhaustive.

Among the **non-deep learning methods**, we find speckle tracking methods and methods using computer vision algorithms. Computer vision methods rely on camera information. Sun *et al.* [132] attached a small camera over the probe to perform patient-related localisation by extracting and matching feature points from the observed skin. The main advantage of this method is its robustness to rigid patient motion, but it needs correct camera calibration. Busam *et al.* [133] proposed using a camera facing the room rather than the patient, which allows rotations to be better calculated using orb-slam [148]. Such SLAM methods predict the camera position by reconstructing the room scene. In 2021, Cai *et al.* [134] introduced a hemispherical rigid body with passive non-coplanar markers, addressing self-occlusion issues in traditional designs. This increased the rotational range, giving sonographers more flexibility. Figure 2.1 presents an overview of the low-cost sensor and markers associated with the above-described computer vision approaches.

On the other hand, speckle correlation methods are imaging techniques measuring deformations by analysing the change in scattering patterns of image patches. Gee *et al.* [135] presented one of the first probe-position prediction methods using speckle correlation. Proposed in 2006, the method predicts in-plane motion thanks to 2D image registration algorithms. The more complex out-of-plane estimation relies on modelling an elevational decorrelation curve. Such curves describe how the correlation of corresponding patches in two images decreases as their separation grows. The curves computed during a calibration

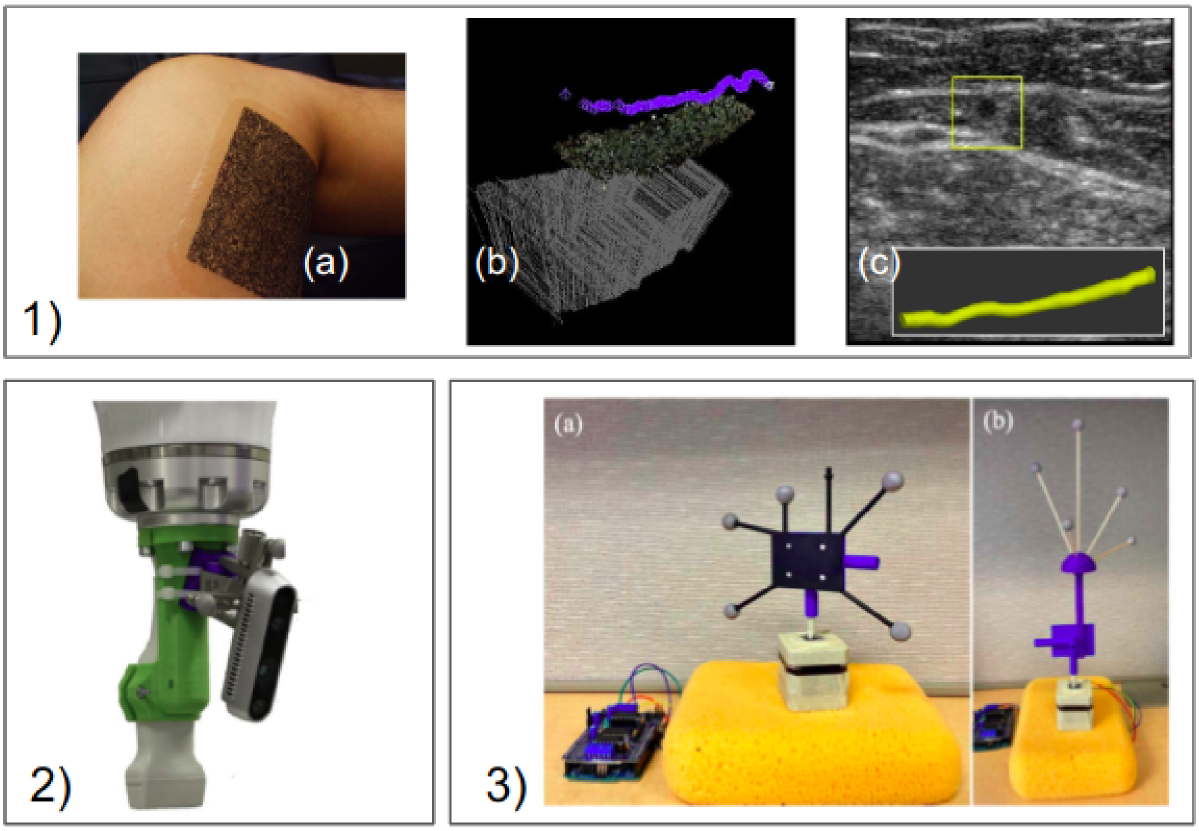


Figure 2.1 – SLAM-based probe position prediction set-ups: 1) Sun *et al.* [132] setup with the camera pointing to the patient (a), the 3D images placed in a 3D space (b) and the reconstruction of an artery in yellow (c), 2) Busman *et al.* [133] setup with a robot holding the ultrasound probe and the camera pointing to the room 3)Cai *et al.* [134] method with the sensor attached to non-coplanar markers (b) compared with co-planar markers (a).

stage are later used as lookup tables, where for a given correlation value, one can obtain the corresponding elevational translation. The method in [135] further considers an empirical adaptation scheme to adjust the curves to different tissue types, as it was shown that elevational correlation curves are tissue-dependent (see Figure 2.2). In practice, the calibration elevational curve is obtained by calculating the Pearson correlation coefficient (ρ) between a pair corresponding patches P_1 and P_2 from two images from different values d. Given at least three con-colinear patches with their correlation, it is possible to estimate the 3 DOF out-of-plane motion, including the translation in the elevational direction or tilt around the lateral axis and yaw around the axial axis. Despite the adaptation to real tissues proposed in [135], speckle correlation methods rely on the detection of patches with fully developed speckles and assume elevational motion is the only source of decorrelation. Leading to biased distance estimates [13, 126, 135, 149].

Deep learning methods analyse image sequences to predict the movements of objects within them or to estimate the probe position between frames only from the image content. Such image-based methods have been developed since 2017. Prevost *et al.*. [14]

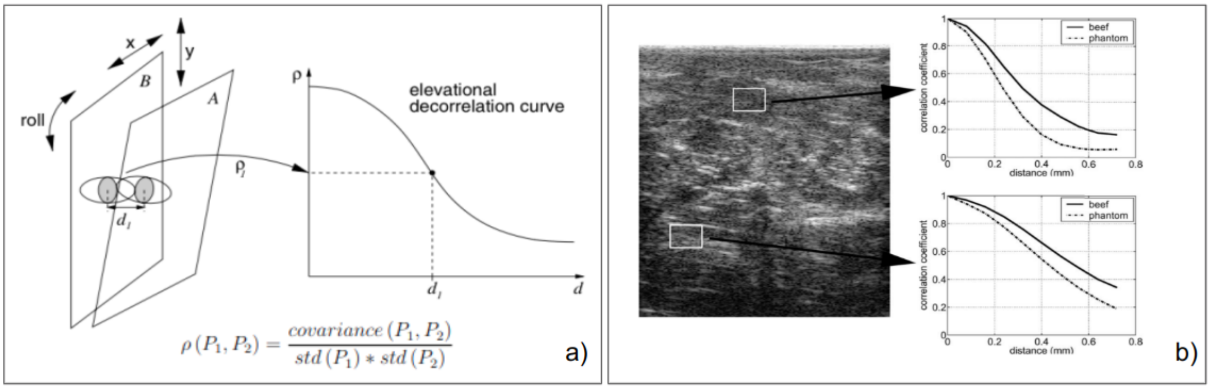


Figure 2.2 – Speckle correlation curves: a)Correlation curve for roll motion obtained for a patch location evaluating the correlation ρ, b)Correlation curves for a calibration phantom and real tissue phantom from a beef. As it can be seen, the resultant elevational decorrelation curves are tissue-dependent. Images taken from Gee *et al.* [135].

proposed the first image-based sensorless 3D freehand ultrasound method relying on a convolutional deep neural network. The method receives as input a pair of neighbouring frames plus the optical flow and predicts as output the relative translation and rotation using the Euler representation in 6 values (see Figure 2.3-a). Similarly, Miura *et al.* [139] (see Figure 2.3-b) and Xie *et al.* [136] (see Figure 2.3-c) add a second branch to process as additional input the optical flow between images. Features from the two branches are later fused through concatenation [139] or attention blocks [136]. The predicted output is again the 6 DOF Euler transformation relating to the input images.

However, more complex strategies rather than fusing image and optical flow features are needed to predict linear sweeps accurately. To improve probe predictions under more complex and diverse motions, several approaches have been proposed, which focus on changing the architectures or the loss function. Guo *et al.* [137] emphasise the benefits of utilising sub-sequences of 2D frames during sweeps rather than just a pair of images. Their proposed architecture integrates a speckle attention module, trained with a correlation loss that forces the predicted motion to be similar along different regions in the sweep. For this loss to be effective, the approach assumes linear sweeps where physicians do not speed or tilt suddenly the ultrasound probe. In a follow-up approach, Guo *et al.* [138] include a contrastive margin ranking loss for harder types of sweeps to enhance the feature similarity between US clips with similar motion trajectories, making trajectory predictions more sensitive to sudden changes in the probe's speed and orientation. An improvement of Miura *et al.* [140] method estimates probe motion from two US images by bifurcating the prediction into in-plane motion and out-of-plane motion estimation sections using a dual loss function. However, such methods based on feature analysis in adjacent frames suffer from the integration of the error for longer sweeps.

To reduce the error in longer sweeps, other works studied the use of memory architecture modules in the networks tomake use of the temporal information, provided by continuing images in sweeps. In this direction, Miura *et al.* [141] and Li *et al.* [142] advocated for integrating recurrent neural networks (RNNs). Ning *et al.* [143] used a hybrid transformer memory encoder. Luo *et al.* [144], include convolution long-shot-term-memory-modules (CLSTM) and shape priors (see Figure 2.3-d). Memory-based methods potentially extract temporal correlation in sequences but suffer from a high-computational demand and easily overfit in the trajectory when datasets are not big enough.

While understanding redundancy and computing the optical flow between frames improves the prediction of in-plane motion, out-of-plane motion remains challenging. An alternative is the use of additional sensor information. **Deep learning methods with low cost sensors**, such as Inertial Magnetic Unit (IMU), have proved to significantly improve the tracking prediction accuracy. Prevost *et al.*. [145], Luo *et al.*. [147] and Mikaeili *et al.*. [146] combined images with IMU sensor data, addressing elevational displacements and large cumulative drifts. Subsequently, Luo *et al.*. improve the method in [15] by adding data from multiple IMUs to enhance volume reconstruction. Their OSCNet method focuses ondiminishing inconsistencies between reconstructions from individual IMUs and ensuring consistency across the scanning sequence.

In terms of the architectures presented in Figure 2.3, the architecture (a) from Guo *et al.* [137] is a feed forward network with parallel residual blocks to account for speckle patterns at different scales. Architecture (b) from Miura *et al.* [139] utilises a ResNet34 for static feature extraction and an encoder for capturing motion, focusing on analysing dynamic content. Architecture (c) from Xie *et al.* [136] emphasises on feature fusion from two branches through channel and spatial attention, using a wrapping layer and a pyramidal structure for multi-scale processing. Lastly, architecture (d) from Ning *et al.* [143] joins a CNN backbone with a transformer encoder, indicating an approach designed to handle long-range dependencies and complex patterns within sequences, which could be essential for detailed temporal analysis.

Despite the advances introduced by the above-cited approaches to solve sensorless or low-cost tracking problems, there is still no optimal solution for a dataset with multiple overlapping sweeps. In addition, most of the methods were evaluated on synthetic datasets and lacked real validation on real acquisitions. We present next the methods and adaptations explored in this thesis to address the sensorless tracking problem, introduced in chapter 2. We also generate synthetic motions from our real reconstructed volumes for a controlled validation (see section 2.5).

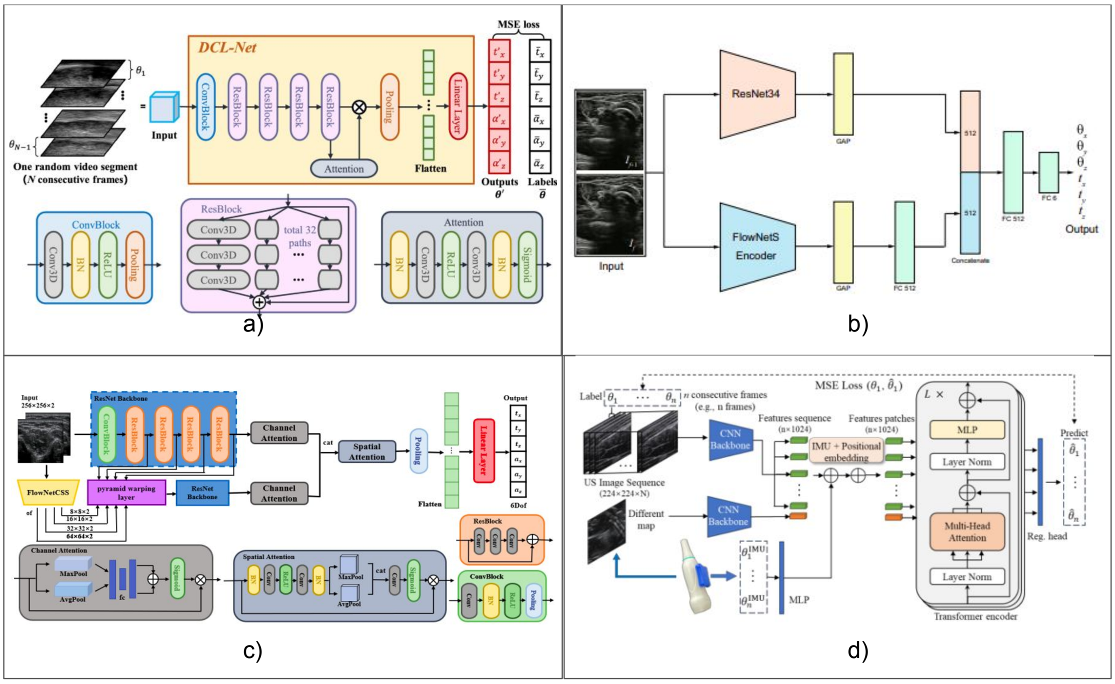


Figure 2.3 – Freehand ultrasound reconstruction methods based on deep-learning: **a)**method relying on image pairs only, from Guo *et al.* [137], **b-d)**methods considering two branches one for processing 2D US image pairs and a second one considering optical flow. The information of the two branches is fussioned by simple concatenation in the work of Miura *et al.* [139](**b**), while it relies on attention blocks in the case of Xie *et al.* [136](**c**) or Ning *et al.* [143](**d**). The later also considers additional position information from an IMU unit.

2.2 Methodology

Previous architectures focus on sweep-tracking prediction for linear movements. Guo *et al.* [137] and Luo *et al.* [144] highlight the difficulty of predicting more complex motion patterns like loops or sweeps with speed changes. Our dataset contains overlapped sweeps at different speeds. In order to understand how previous methods could improve the tracking of our dataset, we reduced the problem to sweeps and created a control dataset with different types of movements similar to the ones used in previous works, as we describe in section 2.2.1. We present in detail the different studied deep-learning approaches in section 2.2.2.

2.2.1 Dataset generation

The 3D sweep dataset contains 2D ultrasound (US) image sequences with 6DOFs tracking for the analysis of large structures.

The dataset was acquired during the study led by Crouzier *et al.* [51], which scanned the low limb of 44 individuals as described in chapter 1.3.3 section 1.1. The scanning pro-

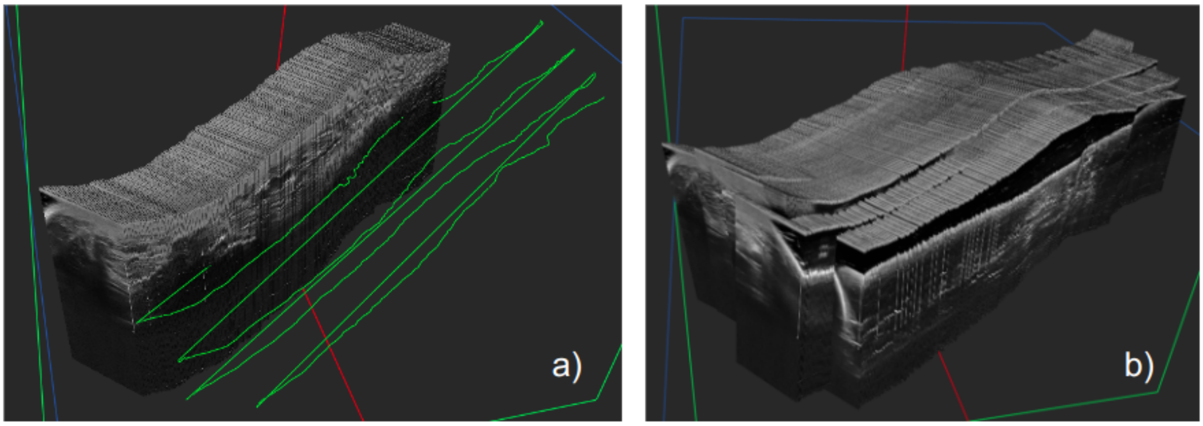


Figure 2.4 – Real sweeps from the low limb dataset: a)images from a single sweep with the hole probe tracking trajectories of the 5 sweeps in green, b) images from 5 different overlapping sweeps.

cess involved four to six sweeps from the knee to the ankle, as illustrated in Figure-2.4-b, where each sweep consisted of a recording of ultrasound images while tracking the probe optically[1], as shown in Figure 2.4-a, where the tracking is presented in green.

Single sweeps were extracted, sampling them from multiple sweeps per participant for a total of 230 sweeps. After splitting the dataset into individual sweeps, we proceed to calculate the relative transforms between adjacent frames(${}^{p_{i-1}}\mathbf{T}_{p_i}$) with a $\mathbf{R_9}$ rotation representation, from the recorded absolute transforms (${}^{O}\mathbf{T}_R$), defining the position of the probe(O) with respect to the receiver tracker(R) (see section 1.3.3). For N images, we have N-1 relative translations. We proceed then to rewrite each 4x4 matrix in 6 values: 3 translations (t_x, t_y, t_z), and 3 rotations in Euler angle representation ($\mathbf{R_3} = \{\sphericalangle x, \sphericalangle y, \sphericalangle z\}$). The dataset's mean and standard deviation per degree of freedom are presented in Table 2.1. We can observe smaller variability in yaw ($\sphericalangle z$) in comparison with $\sphericalangle y$ and $\sphericalangle x$, similar in t_y in comparison with t_x and t_z.

	t_x $[cm]$	t_y $[cm]$	t_z $[cm]$	$\sphericalangle z$ $[°]$	$\sphericalangle y$ $[°]$	$\sphericalangle x$ $[°]$
Mean	-0.68	-0.108	-0.853	-0.108	-0.853	-0.00935
Std	0.466	0.360	0.557	0.360	0.557	0.0172

Table 2.1 – Statistics of the real single sweep dataset: Mean and standard deviation of the 6 DOFs representation of the relative transforms.

Simulated sweeps from real data are created from a controlled sampling of the 3D reconstructed US volumes, following the compounding method described in chapter I in section 1.3.5. This dataset was designed to precisely gauge the efficacy of methodologies

[1] Tracking of the ultrasound probe was performed with 6 cameras, using the Optitrack system-Natural point.

predicting the spatial positioning of the probe as it provides access to precise and noiseless tracking data. Drawing upon the motion paradigms introduced by Luo *et al.*. [144], we curated four distinct types of sweeps: linear, fast and slow, sector, and loop, as depicted in Figure 2.5.

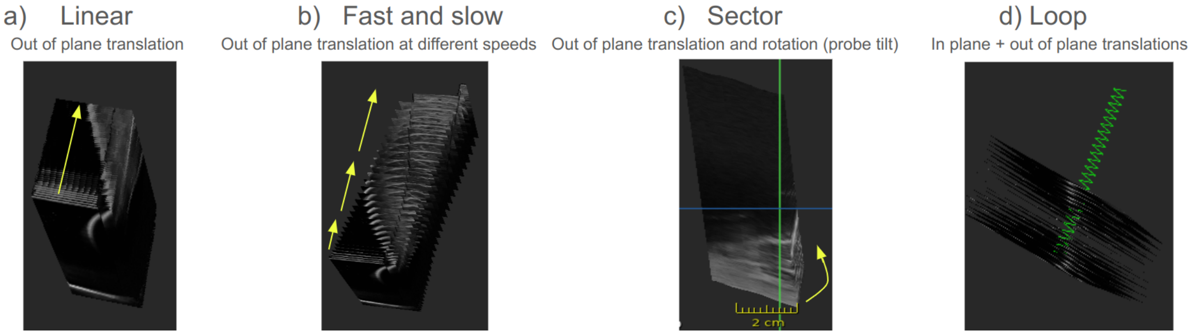

a) Linear
Out of plane translation
b) Fast and slow
Out of plane translation at different speeds
c) Sector
Out of plane translation and rotation (probe tilt)
2 cm
d) Loop
In plane + out of plane translations

Figure 2.5 – Simulated sweeps dataset: a)Linear Trajectory, b)Slow and Fast Trajectory, c)Sector Trajectory, d) Loop Trajectory.

To define the transformation matrix, we place the tracked position on the centre of the image, with the X-axis pointing to the right, the Y-axis to the bottom, and the Z-axis out of the plane. For selecting the "speed" of change per axis, or better called the amount of change per DOF in centimetres, we sample from a Gaussian distribution centred at 0 ($\mu = 0$) and distributed in values around $\pm 1cm$ ($\sigma = 1cm$). A *linear sweep* refers to a constant speed probe motion in 1 DOF: the out-of-plane motion axis (the Z-axis). *Fast and slow* refers to a changing speed probe motion in 1 DOF: the out-of-plane motion axis (the Z-axis). *Sector* refers to a speed variant probe movement in 2 DOFs: the in-plane motion axis (the X-axis) and the Z-axis. *Loop* refers to a constant speed probe motion in 4 DOFs: 3 translation directions and a rotation around the X-axis. Once the probe trajectory has been defined, an image sequence is "resampled" from the volume at the selected locations and orientations.

2.2.2 Problem statement

Assume a sequential freehand acquisition $\mathcal{P}$ with several k overlapping sweeps $\mathcal{P} = \{S_1, ..., S_k\}$. We further assume that the frames belonging to individual sweeps have been isolated and subdivided into smaller batches. Hereafter, we consider a dataset of N tracked batches $\mathbf{B}_i$ and their corresponding 3D tracking $\mathbf{T}_i$. Let the dataset be $\{\mathbf{B}_i, \mathbf{T}_i\}_{i=1,}^{N}$ where $\mathbf{B}_i$ is a 3D tensor containing the images composing a sweeps batch over time, and $\mathbf{T}_i$ the corresponding sequence of 6D transformation describing the position of a sweep frame at a given time $\mathbf{t}$. The problem we address is how to train a DNN to predict the transformations $\hat{\mathbf{T}}_{new}$ for a new unseen sweep $\mathbf{B}_{new}$.

$$\hat{\mathbf{T}}_{new} = DNN(\mathbf{B}_{new}, \theta) \tag{2.1}$$

To predict the position of the ultrasound probe from an US image sequence, we consider as input to a deep neural network a sub-set of n continuous B-mode images of a sweep, $\mathrm{I} = \{\mathrm{i}_1, \ldots, \mathrm{i}_n\}$. We let the network predict $n-1$ vectors $\mathrm{m} =$. Where $\mathrm{m} = [\mathrm{r}, \mathrm{t}]^T$ with $\mathrm{t} = [t_x, t_y, t_z]^T$ and $\mathrm{r} = [\sphericalangle x, \sphericalangle y, \sphericalangle z]^T$ in Euler angle representation(refer to section 2.13).

2.2.3 Deep Learning sensorless tracking methods

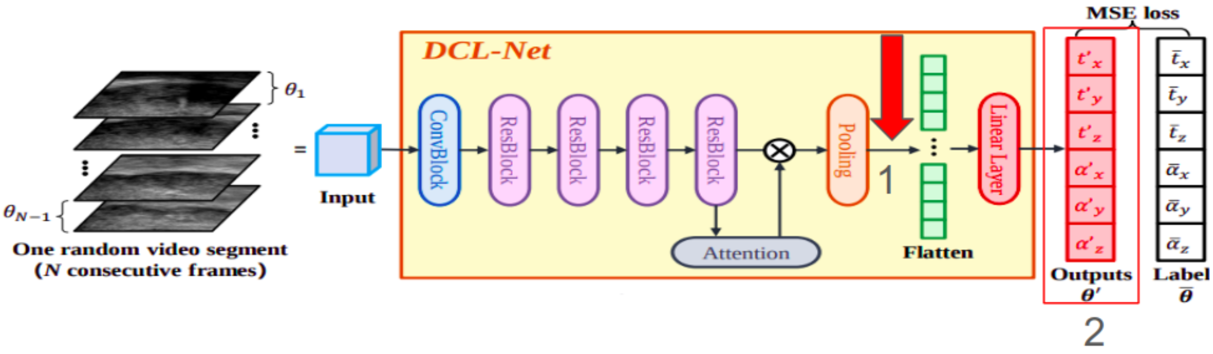


Figure 2.6 – DCL-Net architecture with modifications points: 1. Place for the CLSTM or the DF modules, 2. Angle representations.

Since this work was explored early in the thesis. We relied on one of the few available models at that time, the DCL-Net [137] architecture introduced by Guo *et al.*. We made five main modifications to the DCL-Net architecture in order to evaluate different hypotheses to improve predictions:

- We enforce rigid motion with an intermediate representation that reduces dimensionality in a smooth manner.
- We add a memory module like those within a CLSTM to learn temporal information.
- We use data augmentation techniques to increase the variability of the dataset.
- We split the learning of translation and rotations into 2 different architectures.
- We change the angle representation to other representations that are potentially easier to learn by the network.

Rigid motion enforcement: DCL-Net performs a large dimensionality reduction when passing from the pooling layer of dimension 1 × 1286 to the 6 output values representing a rigid transform between images, see Figure 2.6. We, instead, propose a smoother dimensionality reduction, mapping the poling layer first to a 7 × 7 × 3 space and then continuing to the 6 values representation. We hypothesise, that we can force this intermediate space to represent the 3 channels of a rigid displacement field. A displacement field between a pair of images is a map that describes how each point in one image has moved or shifted to match the corresponding point in the other image, facilitating analysis of changes or motion between them. We expect the conversion from displacement field to 6 DOFs will be more easily learned by the architecture. As we assume rigid transforms between images, such intermediate space should have an additional rigid loss.

To address this regression problem, we evaluate loss functions for a subset of images at two points in the architecture: at the intermediate layer representing the displacement field, being of size $7{\times}7{\times}3$, (Figure 2.6-position 1) and at the final output layer (Figure 2.6-position 2), being of size 1×6, with 6 transforms values per relative transform.

At position 2, we used two key loss functions the mean squared error ($\mathcal{L}mse$) and the correlation loss ($\mathcal{L}corr$). $\mathcal{L}mse$ calculates the average squared difference between estimated and actual values, effectively highlighting larger errors and making it suitable for precise predictions in high-dimensional outputs like a 6-value position vector. Conversely, $\mathcal{L}corr$ assesses how well-predicted values match the patterns of actual values, crucial for maintaining realistic relationships among components, such as orientation angles. We additionally propose two more losses calculated in the displacement field intermediate representation (DF): the mean square error of the DF ($\mathcal{L}_{mse-DF}$) and a rigid motion loss ($\mathcal{L}_{rigid}$). We rely on the mean square error of the displacement field (mse_{DF}) to penalise the predictions concerning the ground truth displacement field. Second, we propose to penalise the displacement of pixels that behave too differently from the others. Here, we exploit the fact that a rigid transform preserves the shape and size of an object, or in other words, the relative distances between points on the object remain unchanged. Expressed in an equation as:

$$Loss = ((1-\alpha) \times \mathcal{L}_{mse}) + \alpha \times \mathcal{L}_{corr} + \beta \times \mathcal{L}_{mse-DF} + \gamma \times \mathcal{L}_{rigid} \tag{2.2}$$

Learning temporal information: Inspired by Tan *et al.* [150], we add a 5-layer CLSTM after the pooling layer and before the flattening layer to learn temporal information. When evaluating, we provide batches with overlapping images for a sliding window inference or without repeating images.

Data augmentation techniques were proposed to increase the variability of the dataset and create architectures independent of the speed of acquisition. Inspired by Housden *et al.*. [149], we first add a sub-sampling interpolation strategy, skipping frames, in order to improve the learning of the out-of-plane motion t_z. Second, we provide the sweeps in the backward and forward order.

Rotational representation Zhou *et al.* [151] showed that alternative angle representations to Euler angles are easier to predict by a position network due to the representation continuity. In the chapter "Tracking Fundamentals: Representations" 2.13, we can find the equations for calculating the different types of representations for the rotation matrices. For our experiments, we proceed to train the DCL-Net architecture on predicting the $\mathbf{R_4}$ representation (Quaternion or 4 Values), the $\mathbf{R_6}$ representation with 6 Values, and the $\mathbf{R_9}$ with the full complete rotation matrices (9 Values).

2.2.4 Metrics

To calculate the error between the ground truth ${}^{C}\mathbf{t}_i$ and the estimated positions ${}^{C}\mathbf{\hat{t}_i}$, for $i = 1, ..., n$ positions, we compare the predicted translations and rotations ${}^{C}\mathbf{\hat{t}_i} = [\hat{t_x}, \hat{t_y}, \hat{t_z}, \sphericalangle\hat{x}, \sphericalangle\hat{y}, \sphericalangle\hat{z}]^T$ with respect to the ground truth ${}^{C}\mathbf{t}_i = [t_x, t_y, t_z, \sphericalangle x, \sphericalangle y, \sphericalangle z]^T$.

The **drift** [137, 144, 145, 152] is the most common metric for the accuracy of full sweep predictions. It calculates the maximum linear integrated error over the sequence. It is defined as the distance between the real absolute position (${}^{C}\mathbf{t}_n$) and the estimated absolute position (${}^{C}\mathbf{\hat{t}}_n$)of the central point of the final image i_n in the sweep, defined with the equation 2.3. Where C denotes the global coordinate system. Formally, the drift is computed as follows:

$$\text{Drift} = \left| \frac{\sqrt{({t_x}^2 - \hat{t_x}^2) + ({t_y}^2 - \hat{t_y}^2) + ({t_z}^2 - \hat{t_z}^2)}}{\sqrt{{t_x}^2 + {t_y}^2 + {t_z}^2}} \right| \tag{2.3}$$

Luo *et al.*. [144] introduced additional metrics related to the length of the sweep: The final drift rate (**FDR**), the average drift rate (**ADR**), the maximum drift (**MD**), the sum of drift (**SD**) and the bidirectional Hausdorff distance (**HD**). Similar to them, we calculate these metrics in our experiments, section 2.3

The geodesic error θ [144] is a measure of the difference between two rotations represented by unit quaternions. It specifically captures the shortest path on the unit quaternion sphere, corresponding to the actual rotational difference.

Given two unit quaternions, $\mathbf{R}_i$ and $\hat{\mathbf{R}}_i$, which represent rotations in the same coordinate frame, the quaternion r, defined as the product $\mathbf{R}_i\hat{\mathbf{R}}_i^*$ (where $\hat{\mathbf{R}}_i^*$ is the conjugate of $\hat{\mathbf{R}}_i$), captures the rotation from $\hat{\mathbf{R}}_i$ to $\mathbf{R}_i$. The angle of this difference rotation can be extracted from r, since r has components $\left(\cos\left(\frac{\theta}{2}\right), u \sin\left(\frac{\theta}{2}\right)\right)$, where θ is the angle of rotation and u is the axis of rotation in three-dimensional space. The first component of r, which corresponds to $\cos\left(\frac{\theta}{2}\right)$, is given by the dot product of $\mathbf{R}_i$ and $\hat{\mathbf{R}}_i$, namely $\mathbf{R}_{i1}\hat{\mathbf{R}}_{i1} + \mathbf{R}_{i2}\hat{\mathbf{R}}_{i2} + \mathbf{R}_{i3}\hat{\mathbf{R}}_{i3} + \mathbf{R}_{i4}\hat{\mathbf{R}}_{i4}$. The geodesic error, θ, can be calculated using the inverse cosine function:

$$\theta = 2\arccos\left(\left|\langle \mathbf{R}_i, \hat{\mathbf{R}}_i \rangle\right|\right) \tag{2.4}$$

where $\langle \mathbf{R}_i, \hat{\mathbf{R}}_i \rangle$ denotes the dot product $| * |$ of $\mathbf{R}_i$ and $\hat{\mathbf{R}}_i$, and the absolute value ensures the principal value of the arccos function is taken, thus yielding the smallest rotation angle between the two quaternions.

In a similar way as the drift was calculated relative to the number of images in the sweep, we propose to calculate the integrated angular error over a sweep so we additionally evaluate the average geodesic rate (**AGR**) as the mean of the cumulative angle error over all frames divided by the expected angle. The Maximum Geodesic rate error (**MGR**) is the maximum rotation angle divided by the relative rotation. Finally, the sum of geodesic errors (**SG**) is the sum of the accumulated angle across all frames. During our experiments, we will evaluate drift metrics and rotation metrics per sweep in the test dataset: **ADR**, **MD**, **SD**, **HD**, **AGR**, **MGR**, **SG**.

2.3 Experiments

Next, we describe the experimental setup and several experiments performed to analyse and evaluate the proposed modifications presented in section 2.3.3. While some of the experiments will present quantitative scores, other experiments will be discussed in the light of exemplary qualitative results.

2.3.1 Experimental setup

For all the experiments, we split our datasets patient-wise, with 29 participants for training, 5 for validation, and 10 for testing. In terms of sweeps, this split corresponds to [116,20,40] and [147, 27, 56] sweeps for the simulated dataset and the real sweeps dataset, respectively. We train the evaluated architectures from scratch and apply min-max histogram normalisation to the images before feeding them to a network. We used Adam optimizer with a learning rate of 0.001 and a StepLR scheduler, with a learning rate decay frequency set at intervals of 10 epochs, coupled with a decay factor of 0.5.

2.3.2 Speckle correlation patterns

The goal of this experiment is to calculate the decorrelation curve for different types of movement to evaluate the possibility of predict the tracking with non-deep-learning methods. Speckle between pairs of images within a sweep under different types of probe motion (sweep types) can provide useful information about the motion. We proceed to calculate the correlation curves of sweeps using Equation 2.5.

$$\rho\left(P_1, P_2\right) = \frac{\text{covariance}\left(P_1, P_2\right)}{\text{std}\left(P_1\right) * \text{std}\left(P_2\right)} \tag{2.5}$$

If the speckle pattern of a patch is found in patches of continuous frames (P_1 and P_2), we can compute the displacement and relate it to the probe motion. In the case of an in-plane probe motion, this relationship is simple. However, what one observes for the typical linear sweep in the out-of-plane direction is a progressive degradation of the correlation score as the probe moves forward. Some examples of correlation curves computed on our simulated dataset are presented in Figure 2.7. First, as expected, the speckle correlation is higher between nearby frames and drops as the probe moves away from the reference frame. Ideally, for the out-of-plane motion sweep, when the correlation curve of 4 patches in the same row of an image is calculated. According to Gee *et al.* [22] original paper, it is only the case for patches on the same row; if the patches are on the same column, differences are expected. However, as presented in Figure 2.2 and Figure 2.7, the correlation of patches affected by the same motion can also degrade differently according to the imaged tissue, even when belonging to the same participant.

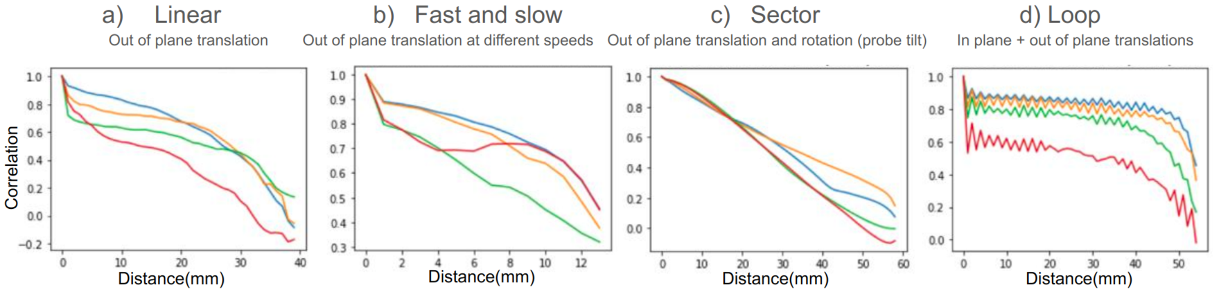


Figure 2.7 – Speckle correlation curves from our simulated dataset: Linear, fast and slow, sector, loop movements. Different colours stand for different patches.

Even when observing the results for the linear and fast-and-slow motion, where only out-of-plane motion is present for all four patches, we observe no clear overlap between the curves of the four patches. In the case of sector motion, we observed partial overlap at the beginning of the curves but not towards the end. Thereby, we argue that speckle correlation methods could be useful for recovering the probe motion for small angle movement with no translation of the top of the image. Conversely, for the loop sweeps, the correlation curves decrease slowly as the probe follows both out-of-plane motion and in-plane motion, which leads to oscillations and makes it difficult to decorrelate the given patches. Tissue dependency is illustrated by the curves in Figure 2.7. We conclude that speckle correlation is not, in general, sufficient enough to predict the probe motion of our data.

2.3.3 Analysis of the rigid motion enforcement proposition

Initial experiments using the DCL-Net on our real data lead to convergence issues. In order to understand if the problem of convergence came from tracking noise, we revert to the simulated dataset. Different values of α, β, and γ, controlling the influence of loss components in Equation 2.2 were evaluated to search for the best configuration enforcing the displacement field to be rigid. Just the best three configurations for DCL-Net+DF are presented in Table 2.2, and errors are calculated per type of sweep.

Experiment	Type	MSE [mm] ↓			MSE[°] ↓			Drift[mm]	Geodesic[°]
		$\hat{t_x}$	$\hat{t_y}$	$\hat{t_z}$	$\sphericalangle x$	$\sphericalangle y$	$\sphericalangle z$	MD	MG
DCL-Net [137]	Sector	**0,023**	**0,01**	**0,043**	**0,173**	**0,008**	**0,007**	**2,229**	**0,105**
$\alpha = 0.3,\ \beta = 0,\ \gamma = 0$	lineal	**0,051**	0,015	**0,286**	0,122	**0,0118**	**0,014**	**11,18**	0,095
	Loop	**5,981**	0,013	**0,152**	0,1308	0,013	0,0113	**10,87**	0,109
	Fast&slow	**0,043**	0,016	**0,4409**	0,1343	0,014	**0,014**	**8,101**	0,048
DCL-Net+DF	Sector	0,067	0,024	0,418	0,325	0,029	0,025	16,48	0,2
$\alpha = 0,\ \beta = 0,\ \gamma = 1$	Lineal	0,073	0,025	1,032	0,146	0,031	0,021	47,84	0,11
	Loop	5,98	0,025	0,179	0,156	0,029	0,024	13,77	0,13
	Fast&slow	0,066	0,027	1,42	0,143	0,021	0,024	31,25	0,05
DCL-Net+DF	loop	0,09	0,06	0,65	0,29	0,05	0,05	25,64	0,19
$\alpha = 0.2,\ \beta = 0.1,\ \gamma = 0.0$	linear	0,11	0,07	0,75	0,16	0,06	0,07	35,4	0,13
	zigzag	5,98	0,06	0,17	0,15	0,06	0,06	13,27	0,13
	fast and slow	0,1	0,06	1,17	0,17	0,06	0,06	24,7	0,071
DCL-Net+DF	Sector	0,09	0,014	0,447	0,37	0,011	0,017	17,64	0,24
$\alpha = 0,\ \beta = 0.1,\ \gamma = 1$	Lineal	0,07	**0,0129**	1,011	**0,1**	0,013	0,02	46,75	**0,081**
	Loop	5,98	**0,013**	0,1884	**0,12**	**0,013**	**0,017**	13,91	**0,1**
	Fast&slow	0,08	**0,011**	1,388	**0,09**	**0,011**	0,017	30,48	**0,034**

Table 2.2 – Error metrics for the DCL Net architecture on the Simulated dataset, before and after including the Displacement Field module and loss. Lower errors per sweep type are presented in bold.

As we observe, the original DCL-Net architecture performs best overall compared to the modifications. The rigidity loss improves results by a few points in the t_y translation and the Geodesic error in linear, loop, and fast and slow sweeps, but DCL-net continues to outperform in t_y of the sector scan and on the other metrics including the drift for all the types of sweep. Figure 2.8 presents the qualitative best results of our rigidity constraint for the four types of simulated sweeps where the ground truth image planes are represented in blue with their corresponding predictions in red. We also display the displacement field in 3D and its 2D projection onto the ZX plane. A red circle highlights the outliers. Despite the encouraging visual results, we conclude that the intermediate displacement field representation was not effective enough to improve the sweep predictions systematically across the dataset.

2.3.4 Addition of a CLSTM module

This experiment evaluates 2 architectures, DCL-Net [137] and CNN-Prevost *et al.*. [14], using the real-sweeps dataset. With the loss presented on Equation 2.6 and with α equal to zero for the CNN architecture.

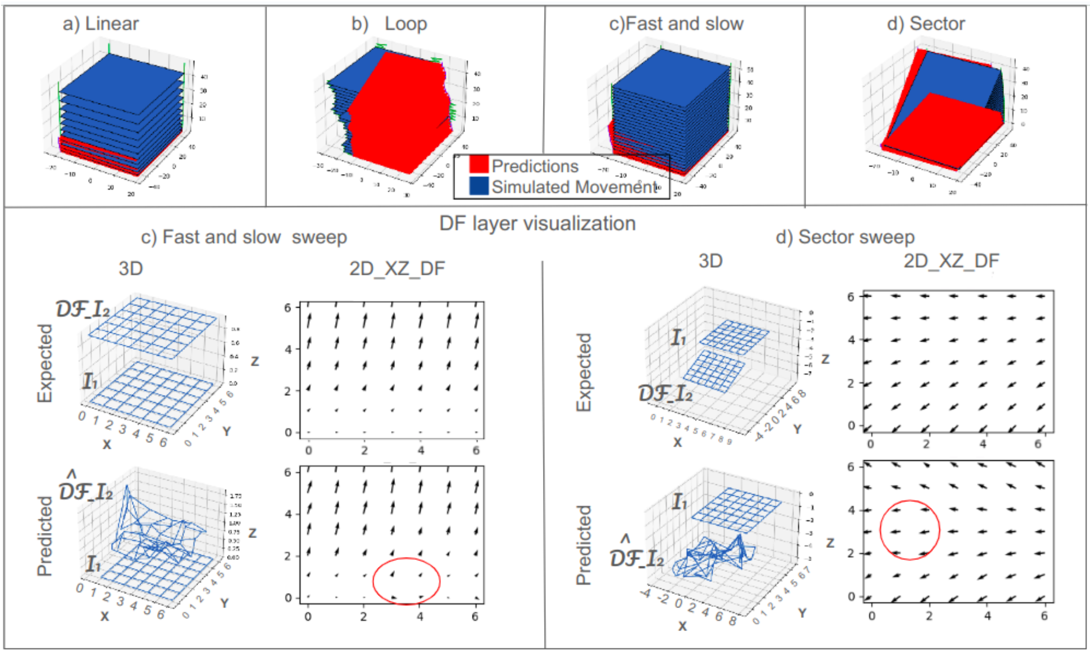


Figure 2.8 – Results of displacement field modification of the DCL-Net architecture on four test sweeps of the simulated dataset: $\alpha = 0$, $\beta = 0.1$, $\gamma = 1$

$$Loss = (1 - \alpha) \times \mathcal{L}_{mse} + \alpha \times \mathcal{L}_{corr} \tag{2.6}$$

.

Different values of α were evaluated for DCL-Net, but only the best two results are presented in the second and third lines of Table 2.3. When $\alpha = 0.3$, the sweeps present an average accomulated drift error of 75 millimetres and a maximum geodesic error of 0.15 degrees.

MODEL	MAE[mm]						Drift[mm]				Geodesic[°]		
	$\hat{t_x}$	$\hat{t_y}$	$\hat{t_z}$	$\sphericalangle\hat{x}$	$\sphericalangle\hat{y}$	$\sphericalangle\hat{z}$	MD	SD	FDR	ADR	MG	AG	AGR
CNN [14]	0.470	0.385	0.593	0.0848	0.076	0.968	132.280	27552	0.457	18295	0.283	0.146	7.49
DCL-Net (0.5) [137]	0.440	0.329	0.472	**0.0762**	0.067	0.0873	**71.40**	15271	**0.232**	9404.6	0.159	0.86	3.75
DCL-Net *(0.3)* [137]	**0.439**	**0.3267**	**0.4741**	0.0763	**0.067**	**0.0858**	75.45	**14604**	0.2384	**7753.4**	**0.1311**	**0.07459**	**3.18**
CNN+CLSTM	0.460	0.372	0.522	0.0786	0.0682	0.08673	83.7	17092	0.243	7751.5	0.152	0.0858	3.69
CNN+CLSTM *(SW)*	0.465	0.356	0.537	0.0818	0.0721	0.0905	1022	20337	0.315	10503	0152	0.239	566

Table 2.3 – Error metrics for CNN architectures, containing our proposed inclusion of a CLSTM module with and without sliding window inference (SW).

However, as we can observe in the fourth and fifth lines of Table 2.3, the addition of a CLSTM module did not significantly improve any of the metrics, probably because the memory module should be added earlier in the architecture where the dimension reduction is not that strong. The way of performing the inference, providing images in overlapped instead of non-overlapped batches (i.e. with a sliding window approach (SW)), does not have impact either.

When comparing the Table 2.3 with the graphs of the qualitative relative trans-

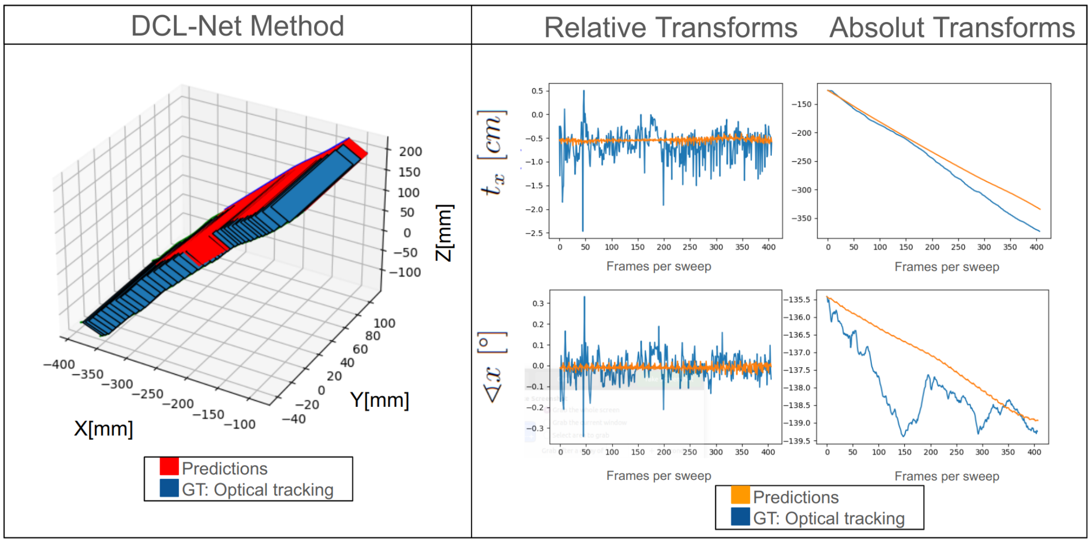


Figure 2.9 – Qualitative predictions of DCL-Net method in the real lower limb dataset. Ground truth is in blue, and predictions are in red/orange.

forms for the methods DCL-Net (Figure 2.9) and DCL-Net+CLSTM with sliding window (SW) (Figure 2.10), we observe that the proposed modification better reflects the oscillations instead of predicting the average displacement as the DCL-Net, which converges to $t_x = -0.68mm$ and $\sphericalangle x = -0.009°$, the mean values per axis across the dataset are presented in Table 2.1.

The horizontal axis in the graphs represents the number of images in the sweep, while the vertical axis represents the displacement between images in mm or degrees, respectively. The 3D plot presents a weak overlapping of the predictions over the ground truth image positions (Red over blue). The qualitative results without the slicing window were omitted as they were very close to those in Figure 2.10. Although the relative predictions and the 3D overlap of the sequence are not too far off, the final accumulated error for both the DCL-Net and our variant with a CLSTM module is large, around 50mm in the t_x and 7° in $\sphericalangle x$. Both errors are still too important to use in position prediction for reconstructing a 3D volume and for clinical use.

2.3.5 Data augmentation techniques

In order to add more variability to the dataset and prevent the network from converging to the mean of the DOFs, we implemented some data augmentation techniques explained in section 2.2.3. Figure 2.11 shows that after augmentation, the final absolute angular error $\sphericalangle y$ seems smaller (3°), but it is because it converges to 0, ignoring the rotation completely, which is not close to reality. However, the network improves the prediction of the translations by better following the ground truth relative motion, t_y, especially in the first frames (Red circle). This result encourages us to think that learning out-of-plane

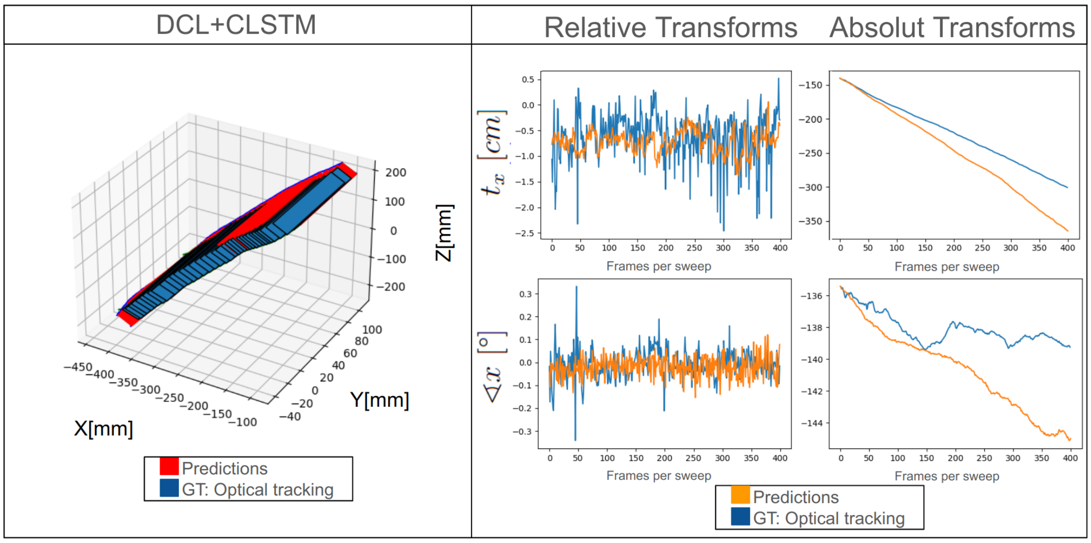


Figure 2.10 – Qualitative predictions of DCL-Net+CLSTM with sliding window in the real lower limb dataset. Ground truth is in blue, and predictions are in red/orange.

translation could be feasible and, in any case, easier than learning rotations.

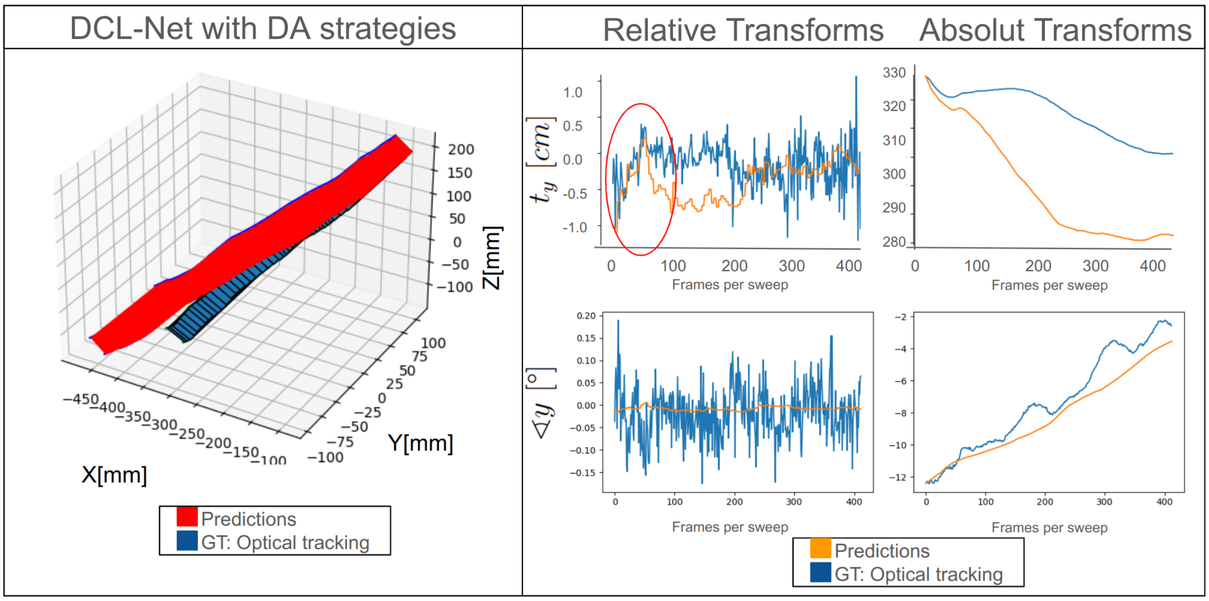


Figure 2.11 – Qualitative predictions of DCL-Net method with data augmentation. Ground truth is in blue, and predictions are in red/orange.

2.3.6 2 DCL-Net architectures results

Towards improving the rotation predictions, we independently train two DCL-Net architectures: One for predicting the rotations and one for predicting the translations. Comparing independent learning in Figure 2.12 with joint learning in Figure 2.11 suggests separate learning of the translation and rotation hinders the translation predictions, which has a larger influence on the drift error. However, for rotations, there is no major improvement nor degradation with this modification. Similar behavior was observed for other test sweeps.

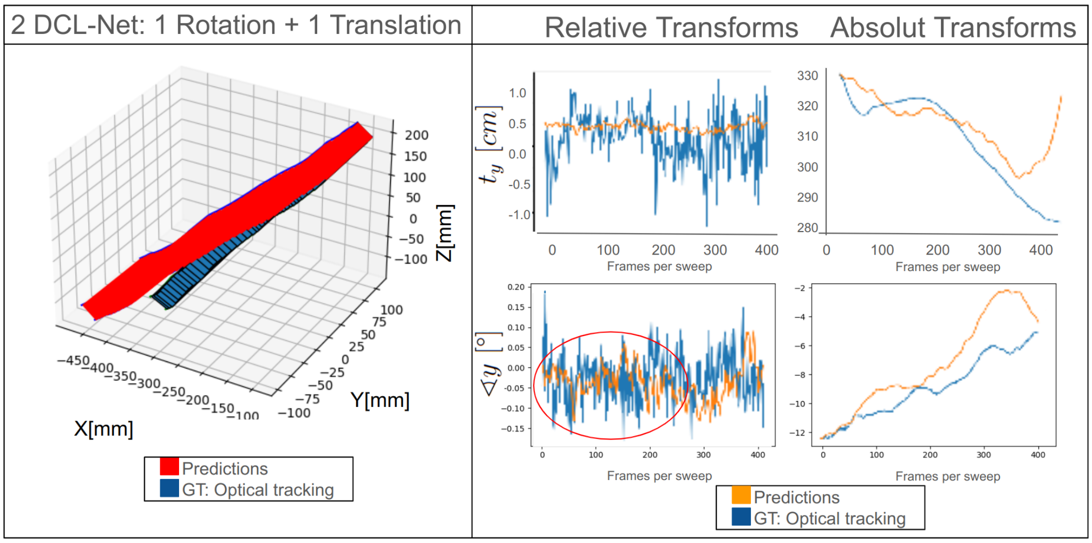


Figure 2.12 – 2 DCL-Net: Training independently rotations from translations

2.3.7 Angle representation results

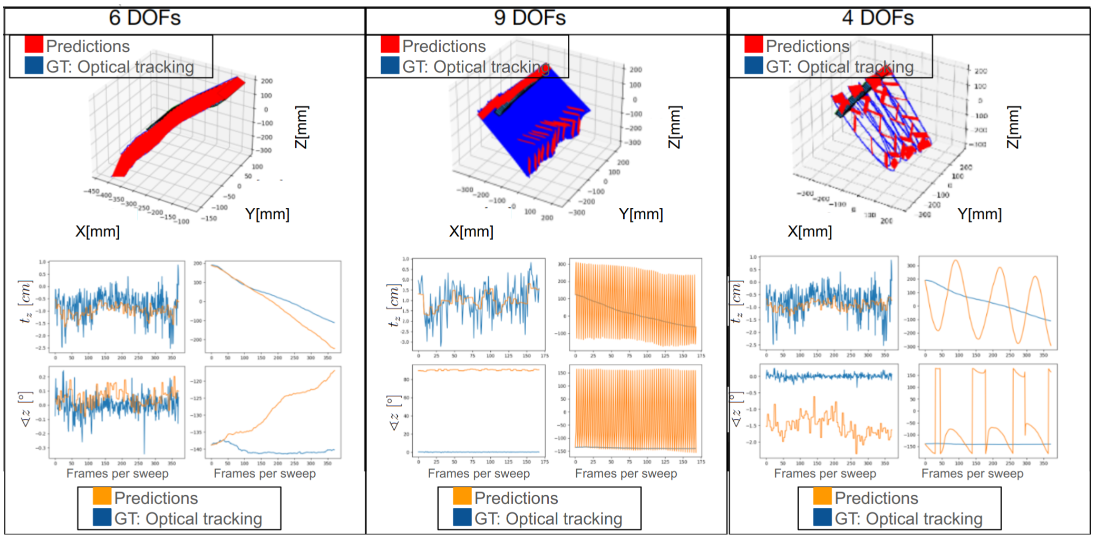


Figure 2.13 – Rotation representation results of the DCL-Net on a test sweep when building the rotation matrix from 4,6,9 predicted values. The 3D sweep is plotted at the top and the relative and absolute displacement of the t_z axis and the $\sphericalangle z$ angle is presented at the bottom.

Figure 2.13 illustrates how the network struggles to learn R_4-Quaternion representations and $\mathbf{R_9}$-complete rotation matrices. Only the $\mathbf{R_6}$ representation provides similar, but not better results than the original $\mathbf{R_3}$-Euler representation.

2.4 Discussion

Our primary goal was to reduce the tracking requirements for building volumes suitable for the segmentation of ultrasound images. To analyze the feasibility of a sensorless approach, we simplified our imaging approach from five overlapping sweeps to single sweeps, reducing volume complexity. Additionally, we compiled a control dataset featuring four distinct movements: linear, varying speeds (fast and slow), loop, and sector. The dataset sizes were extensive, with 314 sweeps from 44 participants for the real dataset and 176 sweeps for the simulated dataset.

In our experimental phase, we initially tested speckle correlation methods. However, these methods proved ineffective for the types of motion characterising our sweeps. Integrating a correlation loss into the DCL-Net [137] architecture, as suggested by Guo *et al.*(2020), seemed to align the architecture more closely with the dataset's average but did not surpass its performance. We employed data augmentation techniques to render the network's predictions independent of the speed of acquisition, which improved translation accuracy but not rotational precision. It became evident that the prediction of both translations and rotations needed to be learned concurrently.

A significant portion of our research concentrated on exploring different angle representations. Beyond the traditional Euler angles (a three-value representation), we experimented with quaternions, full rotation matrix predictions, six-value predictions, and the displacement field intermediate representation. While representations with nine and four degrees of freedom (DOFs) overwhelmed the network, the six-value representation did not enhance positioning accuracy beyond the original $\mathbf{R_3}$-Euler representation. The displacement field intermediate representation with varied loss functions did not surpass DCL-Net's error metrics either. Discussions with the Prevost author of [14] made us realise that excessive speckle filtering in our images hindered the method's effectiveness.

Also, our data presented complexities beyond mere single sweep prediction, and the methods we explored faced challenges even with individual sweeps. It seems that, for the moment, the best-performing learned-based sensorless tracking methods are not yet developed enough to handle real acquisition conditions with multiple sweeps of ultrafast images and slowly varying anatomies. We will, therefore, continue to rely on tracking-based compounding for the following chapters.

Part II

Deep Neural networks for 3D ultrasound segmentation

Abstract

This part presents the research done to automatically segment ultrasound images with high accuracy using neural networks. We will make emphasis on 3D ultrasound segmentation methods. Our goal is to analyze and design a learning-based method to segment with high-accuracy 3D volumes. in particular, for our low limb dataset, our aim is to perform with high accuracy segmentation, with similar volumetric error to the inter-operative volumetric error of 5%. Our proposed methods exploit the advantages of memory modules, custom losses, and weakly supervised annotations. We propose 2 models to handle the specific challenges of ultrasound datasets: First, we study the influence of negative labels to provide additional information when datasets are partially labelled. Moreover, we consider memory modules and an additional auxiliary classification task. Gathering these conditions, we propose the UNet-C-S-R architecture for the segmentation of muscles in 3D ultrasound volumes. Second, we propose the IFSS-net architecture, a network that segments with high certainty the muscles of the lower limb in 3D ultrasounds high-quality volumes.

Clinical motivation and introduction

Quantifying muscle volume is crucial in monitoring neuromuscular diseases and assessing sports performance [4]. This typically involves segmenting lower limb muscles [7,8], while these measurements can be performed in Magnetic Resonance Images (MRI), 3D Ultrasound (US) offers a cost-effective and portable alternative. Nevertheless, manual segmentation in both cases is time-consuming [153, 154] and depends heavily on the operator. Additionally, 3D Ultrasound segmentation is a challenging modality for both radiologists and neural networks. Among the challenges, we fund anatomical variability, indistinct contrast or texture among muscles, and ultrasound-specific issues like incomplete boundaries or uneven intensity distribution [44]. Additional complexities arise when propagating a sub-volume mask across the entire volume due to changes in muscle position, shape, and appearance caused by ultrasound beam physics or probe movement [7].

Recent advancements in deep learning have significantly improved the analysis of ultrasound images and videos in various medical fields, including cartilage tracking segmentation in knees [7], cervical muscle segmentation [19], fetal localisation [16], and lesion classification in breast and liver [17, 18]. Nevertheless, the effectiveness of these supervised methods depends on the availability of extensive datasets annotated by clinical experts. Since 2019, there has been a growing interest in learning from limited annotated data, following, for instance, few-shot learning [155] or self-supervision [156] strategies. Few-shot learning [157], relies on extensive training data first before addressing limited annotated data. Self-learning methods typically use auxiliary tasks like image reconstruction [158] or context restoration [159] for learning from unlabelled data. Self-supervision can also involve pseudo-labelling, where certain unannotated data predictions are considered as labels for further refinement. In this direction, we propose two architectures that handle some specific challenges of ultrasound modality:

- For the low-resolution dataset, our first architecture contains an additional reconstruction task and makes use of negative labels.
- For the high-resolution dataset, our second architecture uses label propagation methods and balances sensitivity and specificity during training.

Part II of the thesis aims to facilitate the segmentation and volume calculation of lower limb muscles from 3D freehand ultrasound volumes, with potential applications in other clinical areas requiring 3D ultrasound organ segmentation, as well as in other.

Chapter 1 introduces the UNet-C-S-R architecture that utilises negative labels to handle partially labelled data. The research explores an approach to incorporate negative information within the labels. Such additional information is anticipated to reduce false positives and improve overall segmentation accuracy, towards better informed clinical decisions and patient outcomes in the context of ultrasound imaging of muscular structures.

Chapter 2, on the other hand, introduces the IFSS-net architecture. A network strategically designed to achieve highly accurate segmentation of the muscles within the lower limb, particularly performing when applied to 3D ultrasound volumes characterised by high image quality and detail. This network incorporates the best of the previous studies, handling the 3D ultrasound volume as a sequence of high-resolution images, using a loss for true positive and true negative balance, relying on the previous segmentation masks to ensure smoothness, and using memory modules to enforce image correlation.

Chapter 1

Muscle segmentation on low-resolution US images with negative label priors

This chapter specifically targets binary segmentation of the lower limb's Gastrocnemius Medialis (GM) muscle of the lower limb dataset in low-resolution US images. Our proposed **UNet-S-R-CLSTM** automatic segmentation method for 3-D ultrasound images aims to support Duchenne muscular dystrophy (DMD) assessment. It uses image sequence patterns (3D slices) to enhance incomplete data. The network design uses encoder-decoder structure, separable depth-wise convolutions [160], and spatio-temporal data to improve missing annotations and boundary detection. The loss is inspired by Petit *et al.* [161] that uses negative masks to constraint the area of prediction. Additionally, our network extends the information propagation across the sequence through the integration of a Convolutional Long Short Term Memory (CLSTM) mechanism [162], strategically positioned within the constriction point of an encoder-decoder framework. The CLSTM adeptly captures potential muscular distortions spanning both short and prolonged intervals, simultaneously averting the dissemination of erroneous or disorderly data via its gated mechanism learning. In pursuit of augmented network convergence and the prevention of excessive adaptation. Lastly, to retain the structural coherence of boundaries, which significantly amplifies pixel-level predictions, we induce the encoding pathway to assimilate a representation that conserves the geometric attributes of the input sequence. This is achieved through the integration of an auxiliary reconstruction decoder trained in an unsupervised manner. A more detailed explanation can be found in this chapter.

1.1 Low-resolution limb muscles dataset

As described in Section : "3D datasets used in this thesis, we use the low-resolution compounding of the 2D ultrasound sweeps of 59 recordings from 44 volunteers. Low resolution is due to the compounded using the Stradwin software [22], from where only low-resolution cross-sectional images could be extracted. Images suffer from artefacts like intersection lines. The data for each patient consists of around 300 images of size 227×544 pixels.

Muscle annotations were done sparsely on the B-mode images, for some selected frames, approximately every 10 slides. In each annotated 2D B-mode image, the visible parts of the Gastronemius medialis (GM), Gastronemius lateralis (GL), and Solius (SOL) muscles were meticulously identified. Annotations for 3D muscles were compounded from 2D labels employing surface fitting techniques. Two experts performed annotations on a subset of volumes, ensuring 3% of volumetric error, effectively validating the accuracy of the manual segmentation procedure.

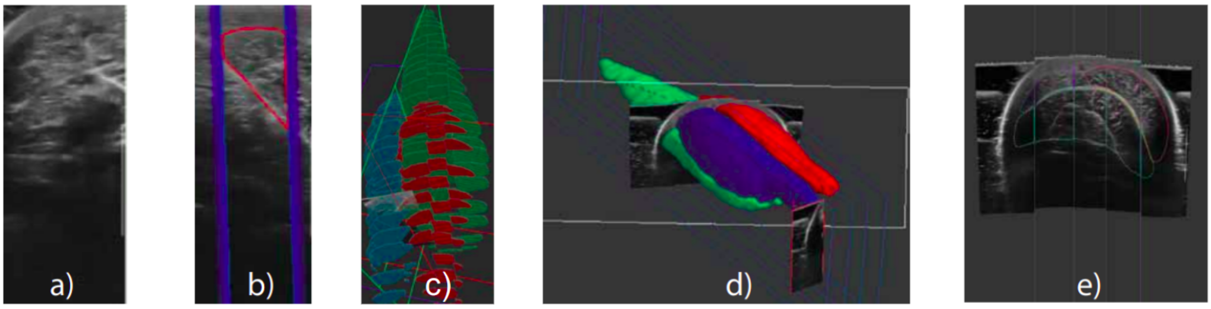


Figure 1.1 – Low-resolution dataset description: a) B-mode image. b) manual annotation over B-mode. c) Sparse annotations on 3D. d) 3D muscle label interpolation. e) Cross-section ultrasound images with GM (Blue), GL(Red), and SOL(Green) annotations overlaid.

1.2 Related work: memory modules and weakly supervised training.

As a baseline, we study the work of Azad *et al.* [163], who in 2019 proposed a learning-based approach suitable for segmenting retinal blood vessels, skin lesions, and lung nodules. All three tasks were performed on natural and CT images. The proposes BCDU-Net architecture, extended from the well-known UNet [48] incorporates a bi-directional Convolutional Long Short-Term-Memory (BI-CLSTM) for handling memory information. It also makes use of densely connected convolutions to enhance feature propagation in the encoding path. In other methods had integrated memory modules for segmentation, for instance, in 2015, Stollenga *et al.* [164] proposed to use multi-dimensional recurrent NNs neural networks to segment brain structures. Focusing on arranging computations pyramidally to enhance GPU parallelisation. In 2016, Chen *et al.* [165] combined a fully convolutional network (FCN) and a recurrent neural network (RNN) to exploit intra-slice and inter-slice contexts for neural structure segmentation and fungus segmentation on natural images. In 2019, Arbelle *et al.* [166] proposed U-Net-CLSTM for cell segmentation in pathology data.

All these works rely on the use of memory modules like RNNs or CLSTMs, which enhance spatio-temporal understanding and provide discriminative features. Nevertheless, these works are not suited for handling incomplete annotation masks. Weakly-supervised segmentation methods aim to reduce full image annotation costs using different types of low-cost labels, such as incomplete masks. In 2015, Dai *et al.* [167] explored bounding box annotations for supervising an FCN recovering borders in a coarse-to-fine fashion. In 2016, Kolesnikov *et al.* [168] proposed a deep neural network with a loss function using weak localisation seeds expanding seeds towards the border's objects based on class information. In a similar direction, Li *et al.* [169] used region proposal networks for approximate mask generation. However, it was not until 2016 and 2018 that Lu *et al.* [170] and Petit *et al.* [161] proposed to use background labels to compensate for missing organ annotations. Their idea was that background pixels belonging to other classes with incomplete masks could provide additional information to the network, helping it to predict the structure of interest with better accuracy.

We model the 3D segmentation problem as segmenting sub-volumes along the volume's axial direction, which we sometimes refer to as the temporal direction. Our proposed method fuses spatio-temporal feature propagation and prior weak information to boost the network performance. The method generates a true negative mask for the background using other organ annotations. The architecture relies on a CLSTM for feature propagation using a true positive annotated mask. More details are presented in the next section.

1.3 UNet-S-R-CLSTM method

With the objective of aiding quantitative measurements in DMD patient follow-ups, we tackle the challenge of segmenting muscles within sequences of 2-D images. Specifically focusing on 3-D freehand ultrasound (US) images from healthy subjects, our aim is to extract segmentation masks for three muscles: GM, GL, and SOL, in lower limb images. Given the complexities of manually annotating such sequences, our approach utilises an FCN model and training strategy that relies on incomplete 2-D annotations, allowing only some slices to be annotated and where each slice does not necessarily encompass all muscle masks.

To successfully train a deep learning model under these constraints, we have developed a training strategy capable of handling partial annotations while maximising the use of available information. Leveraging the knowledge of the locations of other muscles, which aids in limiting foreground predictions, our proposed approach involves a spatio-temporal multi-task strategy encompassing two tasks: 1) segmentation and 2) image reconstruction. The segmentation task employs a spatio-temporal U-Net featuring a CLSTM within its bottleneck to ensure information propagation between slices. Additionally, we consider two competing masks: a foreground mask for the muscle of interest and a background mask containing negative evidence from other annotated muscles. The auxiliary reconstruction task compresses and stores spatio-temporal data in a compact representation.

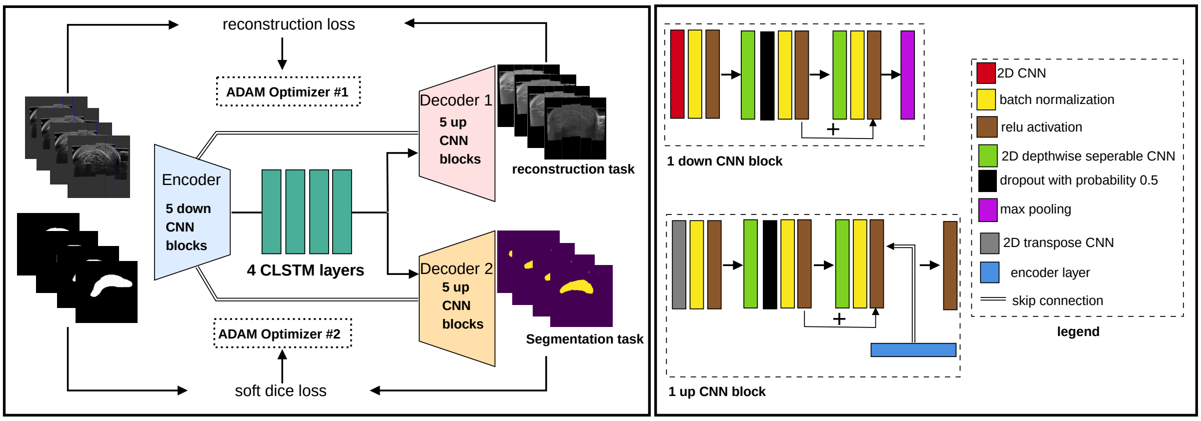


Figure 1.2 – Schematic representation of UNet-S-R-CLSTM.

Architecture: The structure of our FCN model consists on the union of two decoders that share an identical encoding path (as illustrated in Fig. 1.2). This encoding path extracts compact low-resolution features using convolutional blocks. The feature maps from the final encoder layer are fed into a CLSTM module to capture the temporal correlations between slices and effectively address the absence of complete annotations. The output of the CLSTM module is then directed to two decoders: the first is dedicated to reconstructing the original image, while the second is focused on the segmentation task. The last layer of the reconstruction decoder is transformed into a single channel, whereas the segmentation decoder is projected into C maps, with C representing the number of classes. Subsequently, the output feature maps of dimension C undergo a pixel-wise softmax transformation to generate probabilities for the predicted masks. Our model comprises an encoder featuring a configuration of 5 residual depthwise-separable convolutional blocks interspersed with max-pooling operations. Thereby, the encoder gradually transforms the single gray-scale channel into 16, 32, 64, 128, and 256 feature maps at each respective layer. The bottleneck houses a CLSTM composed of 4 sequential cells, each containing 256 feature maps. The decoder structure mirrors arrangement.

Multi-task Learning: Our method jointly trains the network's weights for ultrasound image reconstruction and muscle segmentation. In this way, the encoder learns to extract a compact representation that benefits not only the segmentation task but also the reconstruction task. By conditioning the encoder to retain the spatio-temporal data essential for image sequence reconstruction, it encourages a more accurate geometric representation at the bottleneck. This refined representation subsequently contributes to improved and smoother segmentation masks, particularly along the boundaries.

Typically, multi-task learning involves a weighted combination of two criteria. However, finding suitable weights for different tasks can be intricate due to potential conflicts. Alternatively, we adopt a multi-objective approach using as reference the work of Sener *et al.* [171]. In this manner, we optimize each objective function independently through two ADAM optimizers [96]. Given that both networks, denoted as $f_\theta(.)$ and $g_\omega(.)$, share a common encoder path, we alternately update the network parameters θ and ω.

Loss functions. The volumes are provided to the network as a series of 2D slices with T frames each. Hereafter, we denote one such sequence as $\mathbf{X} = \{\mathbf{x}_1, \ldots, \mathbf{x}_T\}$.

The *reconstruction loss* of the network $f_\theta(.)$ parameterized by θ, the weights in the segmentation branch of the network, is the average Mean Square Error (st-MSE) between an input sequence ($\mathbf{X}$) and the corresponding output reconstructions ($\hat{\mathbf{X}}$), expressed by equation 1.1 where $\mathbf{x}_t$ is the t-th 2-D ultrasound slice, and $f_\theta(\mathbf{x}_t)$ denotes the corresponding output of the reconstruction branch.:

$$\text{st-MSE}(\mathbf{X}, \hat{\mathbf{X}}, \theta) = \sum_{t=1}^{T} (\mathbf{x}_t - f_\theta(\mathbf{x}_t)) \tag{1.1}$$

We name this loss st-MSE, where st stands for spatio-temporal. The *segmentation loss* of the network $f_\omega(.)$ parameterized by ω, the weights in the segmentation branch of the network, is a dice modification that uses both positive and negative labels. The conventional dice loss for segmentation can be expressed as follows :

$$\text{dice}(\mathbf{y}_t, g_\omega(\mathbf{x}_t)) = \left(\frac{2 \sum_{\text{pixels}} \mathbf{y}_t \, g_\omega(\mathbf{x}_t)}{\sum \mathbf{y}_t^2 + \sum g_\omega(\mathbf{x}_t)^2} \right) \tag{1.2}$$

Our loss called the st-SDC loss (slices soft dice loss), is formally described in equation 1.3, where $\mathbf{Y}_t$ and $\hat{\mathbf{Y}}_t$ correspond to the 2D ultrasound ground truth and corresponding predicted masks for a chosen structure at time t of the input sequence $\mathbf{X}$.

$$\text{st-SDC}(\mathbf{Y}, \hat{\mathbf{Y}}, \delta) = \min_{\delta} \sum_{batch} \frac{1}{T} \sum_{t=1}^{T} (\text{SDC}(\mathbf{y}_t, \hat{\mathbf{y}}_t, \delta)) \tag{1.3}$$

The chosen mask $\mathbf{y}_t$ is selected depending on the labels available at time t, as shown in equation (1.4). δ are the weights for the segmentation architecture. This involves either an annotated slice $\mathbf{y}^a$ with its accurate positive foreground or an unannotated slice $\mathbf{y}^n$ with its negative prior. We define SDC as the adaptive Dice score, taking into account the label availability for each slice. Formally:

$$\mathrm{SDC}(\mathbf{y}_t, \hat{\mathbf{y}}_t, \boldsymbol{\omega}) = \begin{cases} 1 - \mathrm{dice}(\mathbf{y}_t^a, g_\delta(\mathbf{x}_t)), & \text{if } \mathbf{y}_t^a \text{ is annotated} \\ \mathrm{dice}(\mathbf{y}_t^n, g_\delta(\mathbf{x}_t)) - 1, & \text{if } \ \mathbf{y}_t^n \text{ is the negative prior} \end{cases} \tag{1.4}$$

1.4 Experiments and results

In order to assess the advancements introduced by our novel model components in comparison to a standard U-Net, we conducted two studies. First, a comparative study using a fully supervised training dataset. Second, we examine the model's capacity to effectively learn from fewer annotations while utilising true negative insights from other muscle masks.

1.4.1 Model Ablation Study

This comparative study uses a fully supervised training dataset focusing on soleus muscle segmentation, which is recognised as one of the most challenging muscles to segment. Initially, we evaluated the conventional *U-Net* model. Subsequently, we replaced the fully convolutional operators with separable depth-wise convolutions, resulting in the *U-Net-S* model. Next, we integrated a secondary decoder focused on reconstructing 2-D image slices sequentially alongside the primary segmentation decoder, forming the *U-Net-S-R* model. Finally, our comprehensive model, *U-Net-S-R-CLSTM*, encompassing the *CLSTM* module atop the *U-Net-S-R*, was evaluated. Table 1.1 highlights the efficacy of employing separable depthwise convolutions within the U-Net architecture, which is particularly interesting for datasets of moderate size as it reduces the number of weights to train. Furthermore, our findings indicate that the inclusion of the supplementary reconstruction decoder enhances both the mean Intersection over Union (mIoU) and Dice Similarity Coefficient (DSC) metrics. Notably, the integration of CLSTM leads to a substantial performance improvement, as it capitalizes on the complete spatio-temporal structure of the input data.

Ablation studies	DSC (%)		mIoU (%)	
Ablation studies	mean	std	mean	std
U-Net	78	8.3	76	11.7
U-Net-S	82	7.2	79	8.9
U-Net-S-R	87	5.4	85	5.2
U-Net-S-R-CLSTM (ours)	**91**	5.2	**89**	5.0

Table 1.1 – Comparison of the baseline (U-Net) for segmenting the soleous muscle versus our model and its different variants.

Figure 1.3 visually showcases the disparities in the predicted mask across various models, demonstrating that our network's preservation of geometrical structure is notably superior, attributed to the presence of the reconstruction decoder.

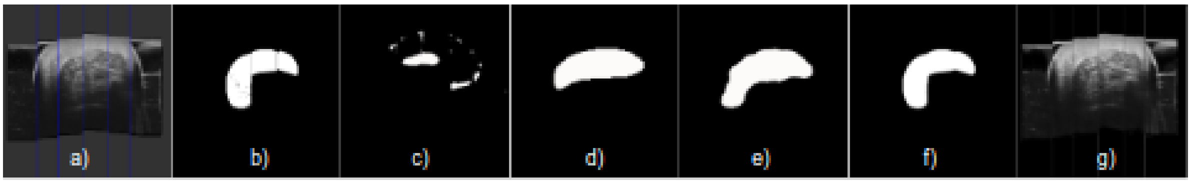


Figure 1.3 – Qualitative results of solius muscle predictions with different methods: **Input:** a) 2D US input slice at t_1, b) annotated mask. **Prediction:** c) U-Net, d) U-Net with separable depthwise convolution, e) U-Net with separable depthwise convolution and reconstruction decoder, f) *Our proposed U-Net-S-R-CLSTM Model.* g) Reconstructed slice.

1.4.2 Negative Priors' influence

We examine the model's capacity to effectively learn from fewer annotations while utilising true negative insights from other muscle masks. We evaluate the GM muscle's performance using as metrics the DSC, the mIoU, and the HD, for a limited number of annotated slices. The true positive masks and the true negative masks are presented in Figure 1.4 e)-f).

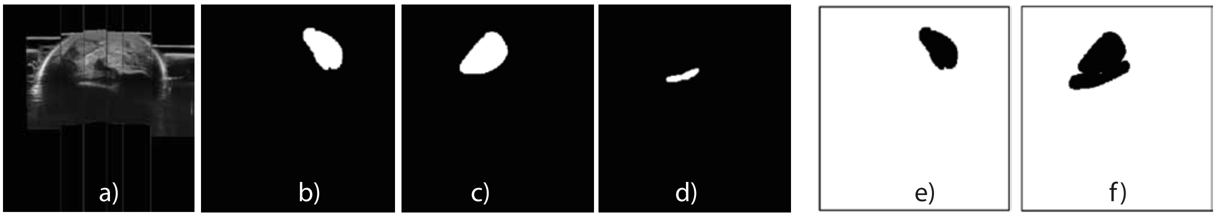


Figure 1.4 – Visualisation of negative priors: a) 2-D US image: its Manual segmentation masks for three muscles: (b) GM, c) GL, d) SOL); e) The background of an annotated GM muscle. f) Generated true negative background: For a slice where the GM was not annotated, the negative prior is built using information from the available annotations, e.g. here those of the GL and SOL muscles.

To gauge the impact of negative priors, we compare our results with a version of the model trained with fully annotated ground truth masks for the GM muscle without utilising prior knowledge from other muscles. The results are presented in Table 1.2. The second column shows the dice/DSC/Dice score for a decreasing number of annotated slices using only the available percentage of relevant muscle annotations. The third column displays model performance with varying percentages of negative evidence. The third column displays model performance when considering, in addition, the negative evidence from other muscles. The fourth column measures the gain w.r.t. the fully supervised Dice score.

annotation percentage [%]	DSC	DSC	mIoU	HDE	performance gain w.r.t DSC score
100	—	94.5	91	2.42	—
	without negative prior	with negative prior			ratio over 100
90	**91.4**	90.1	88	2.85	-1.44
70	**88.8**	83.4	76	4.20	-6.47
50	66.2	**70.8**	61	4.87	+6.49
30	43.1	**50.6**	42	6.50	+14.82

Table 1.2 – Performance of the proposed model under different percentages of annotations using supervised and weakly-supervised settings.

The influence of prior negative knowledge becomes apparent when the un-annotated data is above 50%. In such cases, performance improvement gradually increases from 6.49% to 14.82%. However, with 90% and 70% of annotations available, the performance improvement diminishes to -1.44% and -6.47%, respectively. When sufficient annotations are accessible for a fully supervised model, negative priors can hinder generalisation performance by introducing bias. Conversely, negative priors prove valuable when limited labelled data is available, facilitating gradient flow and loss updates through negative prior information.

1.5 Discussion & perspectives

In this work, we proposed a deep-learning approach to segment muscles in 3D freehand low-resolution ultrasound data. Our model benefits from the spatio-temporal structure of the data at the feature level as well as from an auxiliary reconstruction task. We also presented a multi-objective training strategy that avoids the need to find loss weights. We explore different means to transfer prior knowledge from complementary masks and study the behaviour of the different components under fewer annotations. Experimental results show that with good amounts of supervision, the spatio-temporal consistency enforced through the CLSTM, as well as the addition of a parallel reconstruction decoder, are effective tools to improve the segmentation results. The use of complementary negative masks is the most useful when the amount of the annotated ground truth is relatively small (up to 1000 images). It would be of interest to evaluate the time reduction achieved when using the proposed method as mask initialisation. Some perspectives include training and evaluating our model on other 3D ultrasound datasets, especially public ones.

Chapter 2

Muscle segmentation on High-resolution large US volumes with few annotations and mask propagation

This chapter's focus is on the segmentation of lower limb muscles in 3D high-resolution ultrasound datasets. Among the contributions, we include a novel segmentation and propagation method, a sequential pseudo-labelling strategy for weak supervision, a bidirectional spatio-temporal model, a decremental learning strategy, and an adaptive loss function. Deeper details can be found in the methodology section. In simple words, this study introduces a novel deep learning method for segmenting and propagating muscle masks in 3D US data, called IFSS-Net. The Siamese network is designed to establish a common feature representation between ultrasound and mask sub-volumes, which is further enhanced by a global feature-matching module. A Bidirectional Long Short-Term Memory (Bi-CLSTM) addresses shape and structure changes across the volume. The model weights are trained with a decremental learning strategy, reducing the reliance on expert annotations over time. Such an approach requires a small number of expert-annotated slices per 3D volume and leverages unannotated sub-volumes using sequential pseudo-labelling. To handle class imbalance, a modified Tversky loss adapts weights to penalise false positives and false negatives.

Comparative evaluations, presented in the experiments and results sections, demonstrate the efficacy of the proposed approach. The method shows promising results for segmenting and quantifying muscle volumes accurately, even with limited annotations.

Validation is conducted on the segmentation, label propagation, and volume computation of lower limb muscles using a dataset of 44 subjects and 3D ultrasound images. These findings advance the automation of 3D ultrasound image analysis toward improving both clinical research and medical diagnostics.

2.1 High-resolution limb muscles dataset

The creation of this high-resolution dataset is presented in Section 1.1: "Datasets used in this thesis". To summarize, the dataset contains 59 3D ultrasound Volumes from 44 participants of size 564x632x1443 (± 49x38x207), with an average isotropic voxel spacing of 0.276993 mm. Volumes were reshaped to size 512x512x1400 to be given to the network in a subset of 2D slices. Volumes were reconstructed from 5 parallel overlapping sweeps, from the knee to the ankle, using their optical tracking. Muscles 3D segmentation was obtained using the **ZOI** method [69] from sparse 2D annotations performed on B-mode images as described in Chapter I. For this work, the network was trained 3 times, one per muscle, using only binary masks. The split was done patient-wise, with 29 participants for training, 10 for testing, and 5 for validation. As we provide the cross-sectional images as input to the network, the split leads to a total of 40600 images for training, 14000 for testing, and 7000 for validation. Figure 2.1 presents an overview of the dataset.

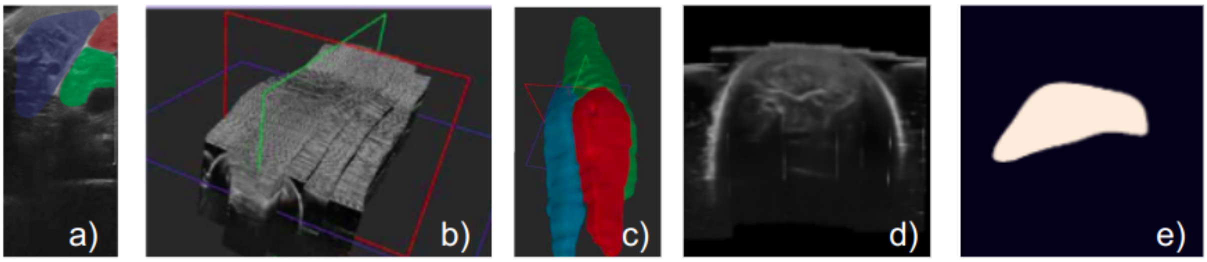


Figure 2.1 – High-resolution ultrasound 3D low limb dataset: a) 2D ultrasound B-mode images with labels. b) 3D ultrasound volume. c) 3D muscle-labels. d)Cross-sectional image given to the network e) Example of binary mask of the soleus muscle.

2.2 Sequential architectures and methods for handling large datasets

Typical volumetric segmentation architectures are extensions of the UNet [48] architecture to the 3D domain. Three main CNN architecture methods of interest are the 3D UNet, the V-net [112] and the Daf3d [2] architectures. These models are characterised by the use of 3D convolutional layers or spatial and channel-wise attention mechanisms to adaptively fuse information from different convolutional pathways.

In order to handle hardware memory constraints, which is a typical challenge for volumetric data, Hesamian *et al.* [172] conducted an exhaustive analysis of deep learning methodologies for medical image segmentation. Roth *et al.* [173] augmented the 3D-UNet's skip connections by replacing concatenation layers with summation layers and harnessing multi-GPU computation for pancreas volume segmentation. Aligning with this approach, we embrace summation layers for parameter reduction. Despite this reduction, our volumes are still too large for common middle-scale GPUs. Therefore, we propose to process them sequentially in sub-volumes, thus enabling efficient large-volume segmentation with a single GPU.

To mitigate potential discontinuities originating from sequential processing, we adopt a bidirectional spatio-temporal module (BiCLSTM) [162]. In a similar vein, Novikov *et al.* [174] approached volume segmentation via sequential 2D slice analysis utilizing a 2D UNet and two BiCLSTMs. While our framework also embraces BiCLSTMs, our emphasis lies in preserving the inherent 3D characteristics by applying 3D convolutions to sub-volumes [175], ensuring local contextual coherence. The sequential treatment of sub-volumes leverages a BiCLSTM across the depth dimension, capturing intricate textures and broader-scale deformations. Notably, in our BiCLSTM architecture, we introduce weight learning for the forward CLSTM, subsequently reused for the backward CLSTM, thereby reducing unnecessary parameter proliferation. In contrast, our approach diverges from Wang *et al.*'s intricate DAF3D [44] architecture for ultrasound image segmentation, opting for a lightweight model that conceives the challenge as a synergistic optimisation of segmentation and propagation tasks through the prism of a Siamese Network [118], instead of concurrent loss functions.

2.3 Interactive Few Shot-Siamese Network

Figure 2.2 provides an overview of the IFSS-Net architecture. Instead of taking the entire ultrasound volume as input, IFSS-Net divides each volume into sub-volumes, which are distinct 3-dimensional tensors, where the number of images in the sub-volume determines its depth. These sub-volumes are processed sequentially through the network, wherein the output of one sub-volume aids the prediction of the subsequent one. This recurrent mechanism is achieved by merging the new input with the predicted mask from the previous sub-volume. To handle this dual input scenario, the IFSS-Net incorporates two encoders, structured using convolutional layers with an identical structure and sharing their weights.

Post-encoding, the inputs are fused into a single tensor via the feature fusion module. This combined output is directed to a bidirectional long-short-term memory module (BiCLSTM). This memory module leverages dependencies across images within the same sub-volume to refine predictions. The decoder employs skip-connections to the correspond-

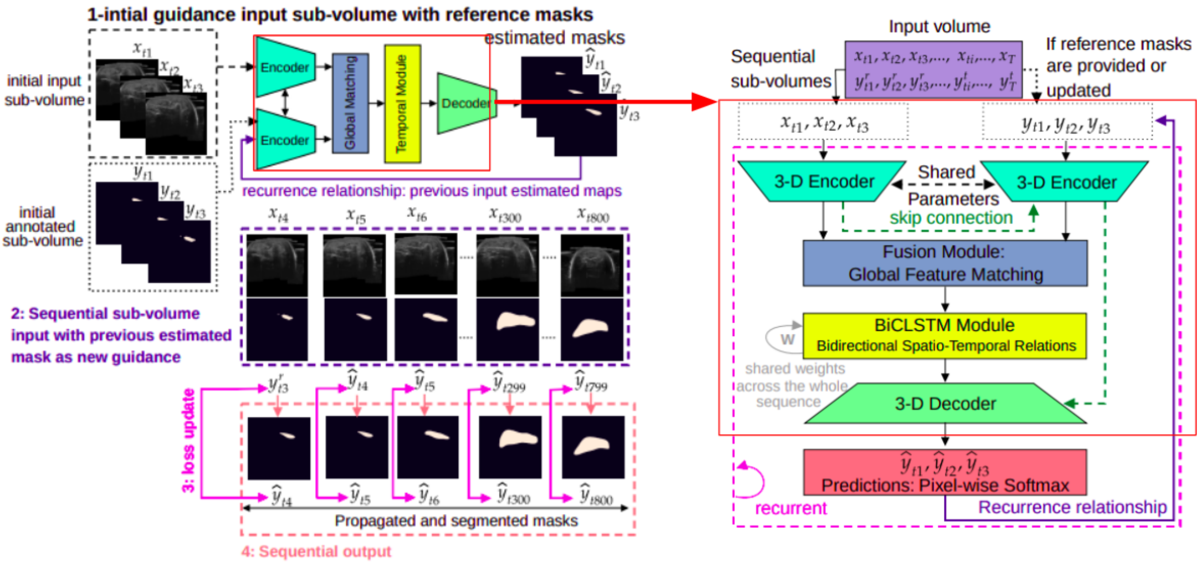

1-intial guidance input sub-volume with reference masks
initial input sub-volume
x_{t1} x_{t2} x_{t3}
initial annotated sub-volume
y_{t1} y_{t2} y_{t3}
Encoder
Encoder
Global Matching
Temporal Module
Decoder
estimated masks
$\hat{y}_{t1}$ $\hat{y}_{t2}$ $\hat{y}_{t3}$
recurrence relationship: previous input estimated maps
x_{t4} x_{t5} x_{t6} x_{t300} x_{t800}
2: Sequential sub-volume input with previous estimated mask as new guidance
3: loss update
y^r_{t3} $\hat{y}_{t4}$ $\hat{y}_{t5}$ $\hat{y}_{t299}$ $\hat{y}_{t799}$
$\hat{y}_{t4}$ $\hat{y}_{t5}$ $\hat{y}_{t6}$ $\hat{y}_{t300}$ $\hat{y}_{t800}$
Propagated and segmented masks
4: Sequential output
Input volume
Sequential sub-volumes
$x_{t1}, x_{t2}, x_{t3},..., x_{ti},..., x_T$
$y^r_{t1}, y^r_{t2}, y^r_{t3},..., y^t_{ti},..., y^r_T$
If reference masks are provided or updated
x_{t1}, x_{t2}, x_{t3}
y_{t1}, y_{t2}, y_{t3}
3-D Encoder
Shared Parameters
3-D Encoder
skip connection
Fusion Module: Global Feature Matching
BiCLSTM Module
Bidirectional Spatio-Temporal Relations
W
shared weights across the whole sequence
3-D Decoder
$\hat{y}_{t1}, \hat{y}_{t2}, \hat{y}_{t3}$
Predictions: Pixel-wise Softmax
recurrent
Recurrence relationship

Figure 2.2 – Schematic of the IFSS-Net architecture.

ing encoder layer, akin to a UNet-style architecture [48]. The predicted mask from the decoder serves as input for predicting the subsequent sub-volume, thereby perpetuating the sequential coherence of the predictions.

The IFSS-Net accurate performance relies on 3 main principles: **Sub-volume processing, Mask propagation, and Pseudo-labelling**.

Mask propagation is a common problem in computer vision. The objective is to track a target's shape and position across frames by modelling its spatio-temporal coherence. Siamese networks in Computer Vision [118, 176] aid this task by projecting images into a shared feature space, learning similarities across frames. Such networks, used in medical imaging, have successfully tracked landmarks in ultrasound sequences and knee cartilage in ultrasound images [7]. Our study adopts a similar joint approach of segmentation and propagation, avoiding manual priors towards tracking and propagating either a reference mask or, in its absence, the network predictions from previous frames. To this end, we process our volume data as sequential **sub-volumes**. Let the dataset used to train the network be defined as $D = \{\mathcal{V}_p, \mathcal{Y}_p\}_{p=1}^{P}$ where each pair $\mathcal{V}_i, \mathcal{Y}_i$is a couple of 3D ultrasound volume and its corresponding binary segmentation mask for the muscle of interest for a single patient, in the set of P patients. $\mathcal{V}_P$ and $\mathcal{Y}_P$ can be expressed as an ordered sequence of T stacked 2D gray-scale US slices and their corresponding binary masks, i.e. $\mathcal{V}_P = \{x_1, \ldots, x_t, \ldots x_T\} \in \mathbb{R}^{T\times512\times512\times1}$ and $\mathcal{Y}\mathcal{Y}_P = \{y_1, \ldots, y_t, \ldots, y_T\}$ for a patient P, with 512 being the image size and T being variable among different patients and muscles. Next, we denote $\hat{\mathcal{Y}}$ the prediction of the network for image x_i. Instead of directly feeding sequences or images, we split each volume $\mathcal{V}_P$ into overlapping sub-volumes $\mathcal{V}_\rangle \in \mathcal{V}_p$ as shown in Figure 2.3 to better model the 3D nature of our data. In order to feed back the current prediction of the network, we also create the corresponding prediction sub-volumes $\hat{\mathcal{Y}}_i$. For a sliding window of size w and a step size of 1, this splitting procedure reorganises our dataset into a new set of $T - w + 1$ overlapped sub-volumes $\{\mathcal{V}_i, \mathcal{Y}_i\}$ given as input to our network.

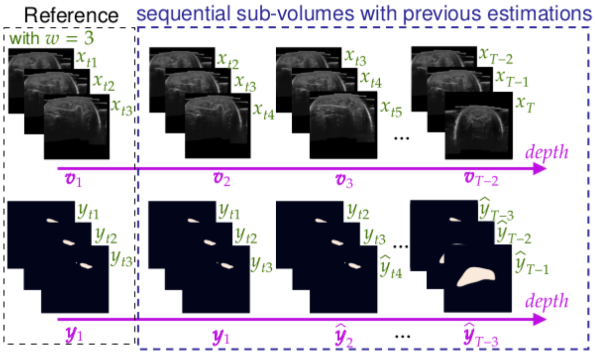


Figure 2.3 – Sequential input of sub-volumes with their corresponding previous time-step estimated masks. This way of handling the data can be seen as a 3D+time processing.

Mask propagation is a well-known task for Siamese architectures, like the PG-Net by Wug et.al [118] in the context of object detection in videos. While PG-Net generates sharp masks, it lacks smooth temporal transitions, which is especially important for our sequential ultrasound data. To enhance mask consistency over time, we establish a recurrence relationship that connects predictions back to the input, akin to Hu et.al. [177] and Perazzi et.al [178]. We incorporate a Bidirectional Convolutional Long-Short-Term Memory (BiCLSTM) [179] to model muscle pixels across past and future slices, ensuring temporal coherence. Furthermore, we introduce Atrous Separable Convolutions (ASC) [180] into our model to reinforce learning local deformation patterns in image space and time within each sub-volume. While ASC captures contextual information at different scales, we mitigate potential boundary issues by employing a series of 3D ASC in a recurrent manner, complemented by bidirectional contextual propagation using the BiCLSTM.

Pseudo Labelling is a semi-supervised technique to address annotation challenges with large datasets. Typically involving two stages, pseudo-labelling first trains on labelled data and then utilises predictions from a deep network as pseudo-labels associated with unlabelled data to retrain the model. This approach has been employed in various domains, including classification, segmentation, and noisy label correction [181, 182]. In this study, our focus lies in minimising annotation costs for segmentation training, capitalising on pseudo-labelling to enhance our propagation model using abundant unannotated slices. Unlike previous methods [172–174], we introduce a novel continuous and sequential pseudo-labelling approach tailored for volumetric data, exploiting spatio-temporal smoothness between slices. This strategy initiates with an annotated 3D US sub-volume extracted from the full volume and sequentially propagating annotations to unlabelled sub-volumes. To this end, the objective function is adjusted to compare current and previous time-step predictions in the absence of annotated data. Sequential pseudo-labelling offers a consistent gradient flow during training, relying on the assumption of high similarity among contiguous slices. In practice, we use manual annotations whenever they are provided, and we relabel unannotated slices from predictions obtained from the last updated state of the network. The pseudo-labelling is done sequentially by adapting the training loss and enforcing label propagation smoothness.

Learnable Tversky Loss Common losses for image segmentation are the cross-entropy loss [48], the Dice score [112], or a combination of the two [34]. However, these choices are not adapted to handle a large imbalance between the background and foreground classes [172] in our case. To this end, Salehi et.al [183] proposed the Tversky loss, which achieves a trade-off between precision and recall by manually controlling the penalties for False positives (FP) and False negatives (FN) with two weights α and β. The Tversky loss is formally expressed as follows:

$$T_{\alpha,\beta}(y,\hat{y}) = 1 - \frac{TP}{TP + \alpha FP + \beta FN}$$

Here, TP denotes the count of true positives, FP signifies the count of false positives, and FN represents the count of false negatives. For instance, a larger β value gives more importance to false negatives, promoting larger predicted masks. Such adjustment could enhance recall while potentially compromising precision. Ordinarily, these parameter values would be predetermined before training, and tuning would be guided by the validation set results. However, IFSS-Net deviates from this norm by incorporating α and β as trainable parameters within the network architecture while imposing the constraint that the combined sum of both α and β must equate to 1. This approach enables the network to dynamically adjust the loss function's behaviour during training, fostering improved convergence and adaptability.

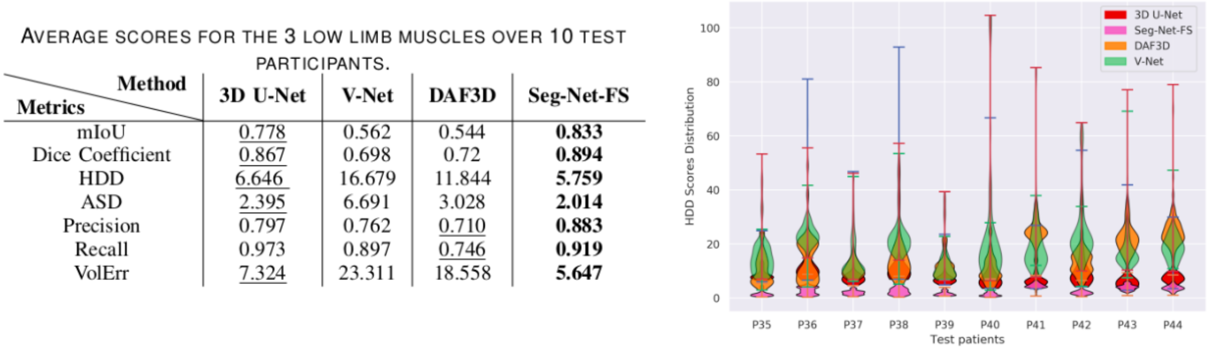
Decremental learning strategy is a method aimed at evaluating the impact of the dataset size for model training accuracy. This approach has been applied especially in medical tasks where datasets are small compared with the computer vision datasets [184]. It usually operates in two phases, the first of which involves training with a dataset fully annotated in every single frame in the volume to establish a strong initial model. The second phase then incrementally removes less critical data from the training process, allowing the model to maintain its performance with a reduced dataset. In our case, handling the volumes as a sequence of images, we reduce the amount of frames annotated. It reduces the model's reliance on extensive training datasets, providing a minimum amount of annotated data to achieve certain accuracy. This approach not only preserves model performance but also enhances computational efficiency. We propose a decremental weakly supervised training strategy to smoothly transition from a weakly-supervised training to a few-shot setting. In this decremental training approach, we progressively add volumes to the training set, each time requiring fewer annotations from the expert. With such gradual decay, very few shot annotations can be utilised efficiently at the end. For our experiments, we define two main trainings.

2.4 Experiments and results

For all the different experiments, it is worth clarifying that PGnet is the name given to the IFSS-Net architecture without the BiCLSTM module, and SegNet is the IFSS-Net architecture without the trainable Twersky loss. Several comparisons were also made when all labels were used under a fully supervised scheme (FS), using 100% of the annotations, and for a decreasing percentage of available annotated labels under a weakly supervised training strategy(WS), using 3% of the annotations.

2.4.1 State of art architectures performance

We trained the Seg-Net-FS in a fully supervised manner and compared it to V-Net, 3D U-Net, and DAF3D, also trained under the same training protocol.



AVERAGE SCORES FOR THE 3 LOW LIMB MUSCLES OVER 10 TEST PARTICIPANTS.

Metrics \ Method	3D U-Net	V-Net	DAF3D	Seg-Net-FS
mIoU	0.778	0.562	0.544	**0.833**
Dice Coefficient	0.867	0.698	0.72	**0.894**
HDD	6.646	16.679	11.844	**5.759**
ASD	2.395	6.691	3.028	**2.014**
Precision	0.797	0.762	0.710	**0.883**
Recall	0.973	0.897	0.746	**0.919**
VolErr	7.324	23.311	18.558	**5.647**

Figure 2.4 – Quantitative evaluation of state-of-the-art segmentation methods: In the left, the average score of 6 different segmentation metrics. On the right is the Hausdorff distance distribution over each of the test participants on 4 different supervised segmentation methods: Unet3D, V-net, DAF3D and Seg-Net-FS (ours).

Among the assessed methods, Seg-Net-FS emerges as the top performer across all the metrics reported. In the second position, we found 3D U-Net and the DAF3D, with the latter providing a better balance between precision and recall. However, the V-Net's qualitative performance is hindered by the utilisation of striding operations that result in the loss of boundary information. Quantitative results for this comparison are reported in the table embedded in Figure 2.4, where our segment performs better than other state-of-the-art approaches across different metrics. The violin plots also display the distribution of HDD scores among participants in the test set, offering insights into the methods' proficiency in representing muscle shape deformations and volume-depth variations. The HDD score distribution for both Seg-Net-FS and 3D U-Net shows a compact pattern, centred around mean values of 5.75 mm and 6.64 mm, respectively. In contrast, DAF3D's HDD scores vary among patients, with scores of certain individuals falling between the first and third quartiles, such as "P36, P40, and P42." This indicates that some slices were accurately segmented (first quartile), while other slices within the same patients suffered from inadequate segmentation (third quartile). On the other hand, V-Net exhibits a higher mean HDD of 16.65 mm and reflects higher variability across patients, particularly "P36, P38, P42, and P44."

2.4.2 Mask-propagation and few-shot inference

After studying our Seg-Net-FS under a 3D fully supervised training scheme, we move toward propagation methods for segmentation in fully and weakly supervised schemes.

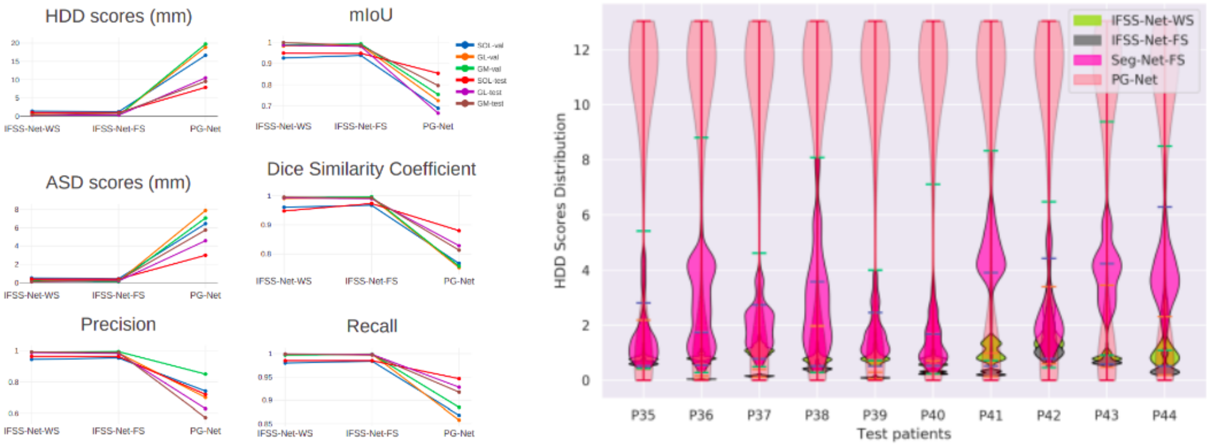


Figure 2.5 – Quantitative results for decremental learning strategy: IFSS-Net-Weakly-supervised(WS), IFSS-Net-Full supervised(FS), SEG-NET Full supervised (FS) and PG-NET Full supervised (FS)

In the proposed propagation method, we provide initial guidance with a first set of sub-volume annotations to find the target muscle in a test volume. In this sense, our method follows a few-shot influence strategy. Then, predictions are propagated progressively, covering larger parts of the volume. Quantitative results are presented at left in Figure 2.5. As we infer from our results, our recurrent loop makes the model aware of the muscle changes over time. Our experiments show that our propagation-based method IFSS-Net does better than Seg-Net-FS. We tested our new model in different ways, comparing it under weak supervision (WS) and full supervision (FS) schemes. In the WS case, we consider only 3% of the annotations (one mask every 10 frames), while in FS, we consider all frames in the volume annotated. We also compared our method with another propagation-based approach called PG-Net (IFSS-Net without the BICLSTM memory module). Different validation and test set metrics were computed for the 3 muscles.

The violin plots on the right of Figure 2.5 present the distribution of HDD scores across the propagation networks and compare it to the Seg-Net-FS. It is evident that transitioning from volumetric segmentation (Seg-Net-FS) to segmentation with propagation (IFSS-Net) results in a significant reduction in HDD scores. This shift leads to smoother predictions and enhances the accuracy of volume measurements. Examining the PG-Net results, we notice two distinct sets of quartiles. The first set is concentrated below 4 mm, corresponding to propagated masks that closely resemble the reference masks. The second set lies above 8 mm, corresponding to propagated masks that differ considerably from the reference masks. Consequently, PG-Net encounters challenges in effectively propagating the reference mask throughout the depth.

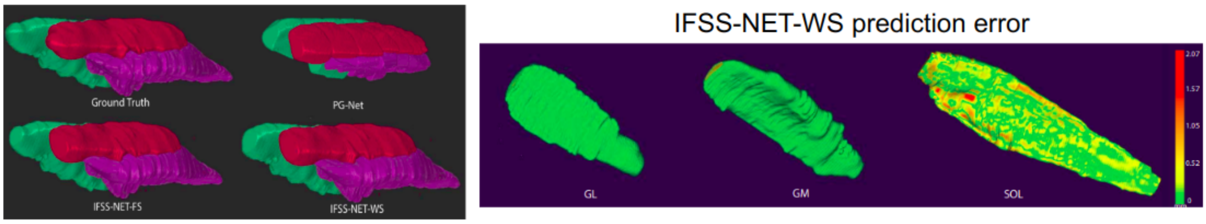


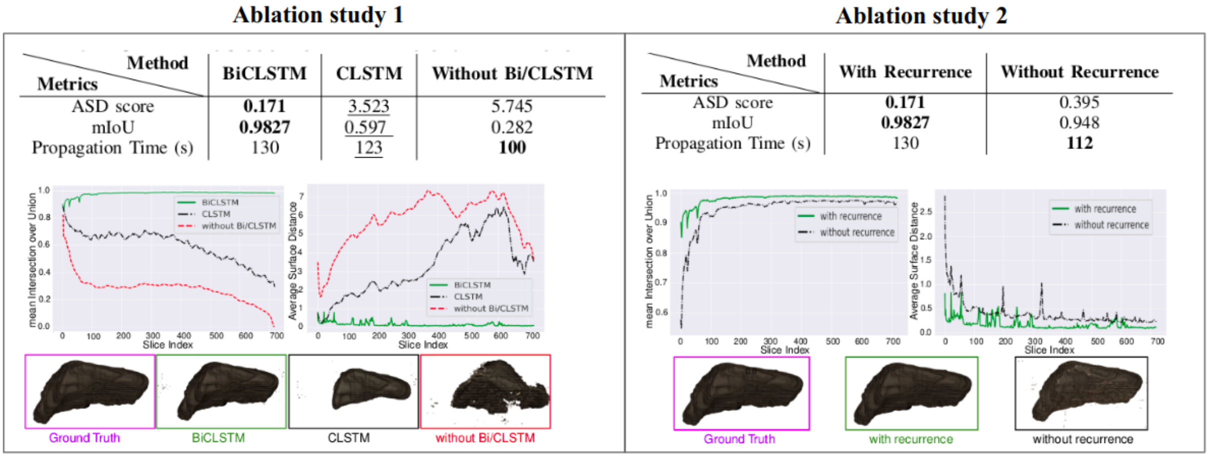
Figure 2.6 – Qualitative results of the Hausdorff surface error: (left) Qualitative results of different IFSS-Net configurations. (right) Color-coded surface distance error for each of the 3 muscles.

Qualitative results can be observed in Figure 2.6. On the left is an illustration of the predictions for one test participant. The PG-net inference resembles a zero-order interpolation, resulting in an uneven surface prediction for the volume. This outcome can be attributed to the architecture of PG-Net, which employs an RNN module. Indeed, RNN modules require transforming feature maps into vectors, causing a loss of spatial structure. At first glance, there is not much difference between the fully supervised and weakly supervised approaches. However, the weakly supervised approach uses only 3% of the annotations. The Figure 2.6-right on the right presents the surface error map across the GL, GM, and SOL muscles for IFSS-Net-WS. The colour-coded surface is used to represent the distance of the prediction from the ground truth for each muscle. As before, the SOL muscle emerges as the most challenging to segment. The highlighted regions in red primarily stem from sub-volumes where the US image reconstruction quality is notably poor. In the case of GM and GL muscles, the endpoints proved more difficult to segment, although the distance error remained within 0.52 mm. This difficulty might arise from our model struggling to accurately propagate the masks to cover the entire muscle extent, particularly in regions where the US sub-volume is affected by noise.

2.4.3 Ablation studies

Five different ablation studies were performed in order to evaluate the contribution of each module and parameter configuration.

Ablation studies 1 and 2, presented in Figure 2.7, focus on the influence of the memory modules (CLSTM, BiCLSTM) and the recurrence loop. As reflected by the ASD and IoU scores in Figure 2.7-Left, the BiCLSTM module stands out as a critical component of our approach .Substituting the BiCLSTM module with a CLSTM initially maintains the smoothness, but it later deteriorates along the sequence. The absence of BiCLSTM or CLSTM modules causes inconsistent propagation and leaking of the SOL mask predictions into neighbouring muscles (GL and GM). In the second ablation study (Figure 2.7-Right), the incorporation of current, past, and future annotations is shown to help the network learn SOL muscle variation patterns through depth. Moreover, the absence of a recurrent loop produces a noisy and unsmooth boundary. The recurrent loop allows the network to be updated constantly with new reference pseudo masks. Thereby, this module provides guidance to the network while the expert intervention is minimal.



Ablation study 1

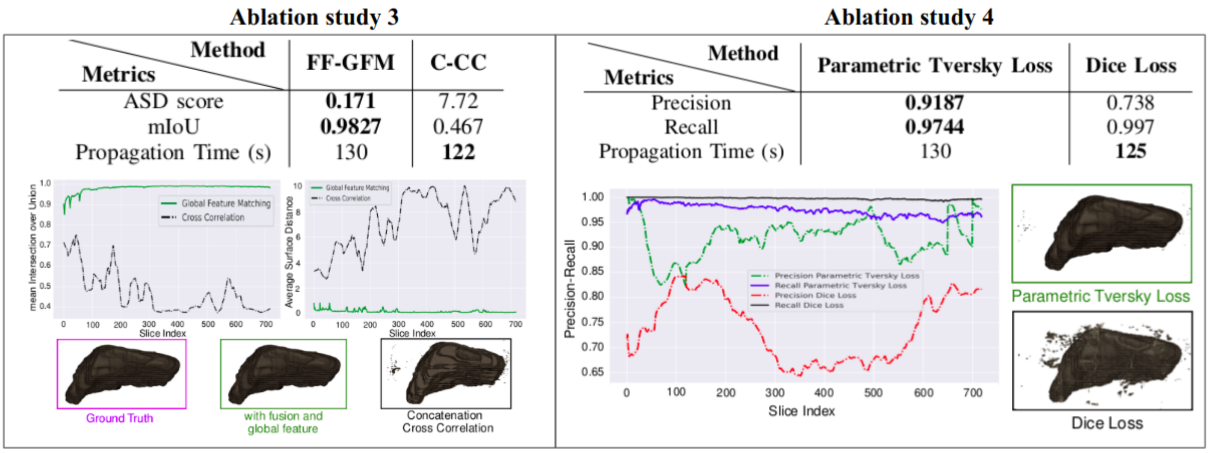
Metrics \ Method	BiCLSTM	CLSTM	Without Bi/CLSTM
ASD score	**0.171**	3.523	5.745
mIoU	**0.9827**	0.597	0.282
Propagation Time (s)	130	123	**100**

Ablation study 2

Metrics \ Method	With Recurrence	Without Recurrence
ASD score	**0.171**	0.395
mIoU	**0.9827**	0.948
Propagation Time (s)	130	**112**

Figure 2.7 – Memory module and recurrence ablation experiments on SOL muscle: **(Left)** Performance using different temporal modules. (Top) mIoU and ASD scores over depth (slice index) to assess the propagation for each slice. (Bottom) Resulting rendered volumes for one subject. **(Right)** Propagation performance with and without the recurrence relation. The recurrence relation module obtains a smoother and less noisy graph and volume prediction.

Ablation studies 3 and 4, presented in Figure 2.8, evaluate the influence of the feature fusion module and the Tversky loss. We contrasted a simplified approach by substituting feature fusion (FF) with basic concatenation (C) and replacing the Global Feature Match-



Ablation study 3

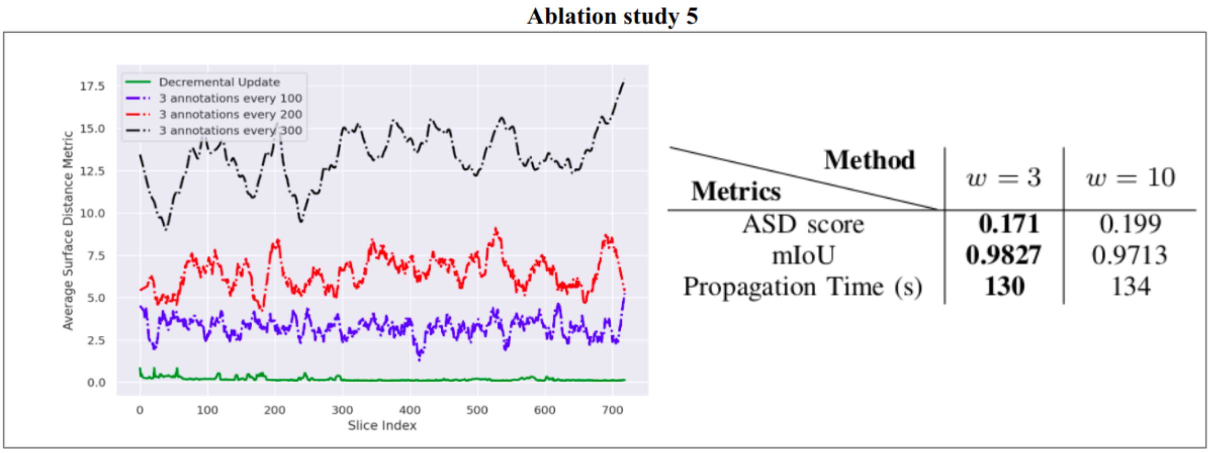
Metrics \ Method	FF-GFM	C-CC
ASD score	**0.171**	7.72
mIoU	**0.9827**	0.467
Propagation Time (s)	130	**122**

Ablation study 4

Metrics \ Method	Parametric Tversky Loss	Dice Loss
Precision	**0.9187**	0.738
Recall	**0.9744**	0.997
Propagation Time (s)	130	**125**

Figure 2.8 – Global feature fusion and parametric Tversky ablation experiments: **(Left** Propagation performance with and without global feature matching. **(Right)** Influence of the loss function in handling input and output voxel imbalance. With the Dice loss, high recall and low precision are reported, while with the parametric Tversky loss, a trade-off between the recall and the precision is maintained.

ing (GFM) [185] with a simple cross-correlation (CC) operator. Predictions appear noisier with a simple cross-correlation approach. The quantitative results in Figure 2.8-Left emphasise the significant role of the FF-GFM module in establishing a coherent feature space for images and guiding masks. In Figure 2.8-Right, we also compare the Tversky loss with a Dice loss. Dice tends to bias predictions towards background pixels, causing overfitting and giving FN pixels undue importance. The Tversky loss reduces overfitting by achieving the desired FNs-FPs balance. Learning the Tversky loss parameters is also vital as muscle size evolves along depth, starting with a small foreground muscle amid numerous background voxels, moving toward more balanced middle slices and going back to imbalance at the end of the muscle.



Ablation study 5

Metrics \ Method	$w = 3$	$w = 10$
ASD score	**0.171**	0.199
mIoU	**0.9827**	0.9713
Propagation Time (s)	**130**	134

Figure 2.9 – Decremental learning and sub-volume window size ablation experiments: **(Left)** Supervision strategy (decremental vs fixed). **(Right)**Exploring the temporal depth of sub-volumes, with $w = 3$ and $w = 10$.

Ablation studies 5 and 6, presented in Figure 2.8, evaluate the influence of the decremental learning strategy and window size parameters determining sub-volume depth. We compare our approach under different amounts of weak supervision, considering a fixed percentage of annotated images per volume or the decremental approach. In the fixed case, we include three annotation masks every 100, 200, and 300 slices, corresponding to a weak supervision with 3.5%, 2%, and 1% of annotations, respectively. In the decremental scheme, we stick to the 3.5% but we apply a decremental decay across volumes. For our dataset with 29 training volumes, and assuming the sequence length is 1400, this means the first volume requires 233 annotated slices, the second 116, and so on, i.e. [233; 116; 58; 56; 53; 51; ...], ensuring the total amount of annotations (3,5%) is preserved. Fixed updates with the same 3.5% supervision yield a higher 3.2 mm ASD score (blue graph). Further reducing the annotation percentage (2% and 1%) leads to noisier pseudo labels, causing less smooth predictions with higher ASD scores of 6.3 mm and 13.8 mm. The table embedded in Figure 2.9-Right reports the evaluation of the size of the sub-volume window denoted as w. With $w = 10$ instead of 3, the sequential pseudo-labelling method becomes slightly noisier. This leads to more unnecessary errors and longer computation times.

2.5 Discussion and Perspectives

The IFSS-Net is a novel approach that combines the advantages of expert involvement and deep learning methods for segmenting sequential or volumetric data. We integrated various strategies, such as Siamese networks with sub-volume recurrency, Bi-CLSTM, 3D ACS, and pseudo-labelling, to effectively leverage the spatio-temporal consistency in such data. The resulting IFSS-Net enables the propagation of a small number of reference annotations across the entire volume or sequence, reducing the expert effort required during training. We conducted a comprehensive evaluation of muscle segmentation and volume estimation tasks on ultrasound volumes.

One of the directions for future research in this field is to validate the performance of our IFSS-Net on 3D freehand ultrasound volumes from children with Duchenne Muscular Dystrophy. Given that muscles are progressively replaced by fatty tissues as the disease advances, adaptations will be necessary. Possible solutions include fine-tuning the model on a limited set of DMD patients or training it on synthetic fatty tissue. Another approach could involve adapting the established zero-shot learning approach from classification to segmentation tasks. Our proposed methodology holds potential not only for segmenting muscles with volume measurements but also for other anatomies and various medical image analysis tasks involving sequential data. Furthermore, two additional directions for exploration include testing the scalability of IFSS-Net across multiple anatomies simultaneously and assessing its ability to generalise across multiple imaging modalities.

Part III

Analysis of ultrasound segmentation architectures

Abstract

In this part, I propose two strategies to analyse the quality of the segmentations obtained with different architectures on 2D and 3D ultrasound consistency. We mainly focus on the quality of borders, where annotations suffer from variability due to the intrinsic principles of the modality. The first study evaluates the label variability of annotators in uncertain areas of the images, while the second study analyses the performance of ultrasound segmentation networks by differentiating between distinct and completed borders. Our goal is to highlight and quantitatively evaluate how specificities of ultrasound influence both human annotations and machine predictions for a better-informed interpretation of the results. Our studies make use of the Seg-Grad-Cam strategy to visualise where networks focus their attention when segmenting distinct versus completed borders, and to quantify the accuracy of such predictions. Finally, in order to boost the awareness of label variability in the networks, we propose to provide additional information from ultrasound confidence-maps to segmentation architectures, with the objective of teaching the network border uncertainty.

Motivation

Accurate delineation of borders on ultrasound images is challenging due to their inherent noise, attenuation, speckle, shadows, signal dropout, and low contrast between areas of interest [186]. Different from other modalities such as computed tomography (CT) or magnetic resonance imaging (MRI), which display homogeneous tissue distributions across similar organ structures, ultrasound relies on the differential absorption of acoustic waves and their reflective interactions at tissue interfaces characterised by varying acoustic impedance. Therefore, the delineation of anatomical boundaries can be particularly challenging for both physicians and networks in regions exhibiting low signal intensity or in the presence of artefacts.

In order to understand how ultrasound segmentation deep-neural networks behave on such a challenging task, we perform two experimental studies towards:

— Making deep-learning segmentation models aware of the specific uncertainties underlying ultrasound images, which also affect experts' annotations.

— Understanding the features networks rely on to complete borders on low cotrast regions.

To address the **labelling variability** among observers in medical image segmentation, supervised learning methods often rely on ground truth data generated by popular fusion techniques such as majority voting [187] or STAPLE [188]. Such techniques focus on obtaining a single label map given a set of annotations from different experts, but do not model annotation variability per se. Several strategies have been developed to incorporate uncertainty directly within deep learning segmentation methods. For instance, Baumgartner *et al.* [189] introduced a layered probabilistic model, and Jungo *et al.* [190] explored the uncertainty and calibration in segmenting brain tumours with UNet-like structures [48]. Monteiro *et al.* [191] proposed stochastic segmentation networks and Rousseau *et al.* [192] propose post-hoc network calibration methods. All these techniques involve training with single or multiple annotations, which is expensive. A drawback of the above methods is their need of architectural alterations and/or additional labels. Instead, in this part, we opt for retraining well-studied segmentation architectures and evaluate the possibility of providing a pre-computed ultrasound Confidence Map (CM), introduced by Kalamaris *et al.* [193], as an additional input channel to the network. The pre-calculated CMs have been used in ultrasound for improving reconstruction [194], registration [195], and non-deep-learning bone segmentation [196]. To the best of our knowledge, this is the first study using ultrasound confidence maps in the context of deep-neural neural networks for semantic segmentation and border variability analysis.

To improve the understanding on how **segmentation networks process anatomical borders in ultrasound images**, one can rely on methods developed for enhanced model interpretation [197, 198]. Initially introduced for classification in [199], Grad-Cam highlights regions of an input image that significantly influence a model's prediction, providing insights into why a model made a specific decision. More recently, some works have employed Grad-Cam to evaluate, in the context of medical images, the learned information for classification [199,200] or segmentation [201,202] tasks. The above methods have primarily centred on MRI or CT analysis, and only few have been evaluated for ultrasound segmentation [172] but looking at the full predictions. To the best of our knowledge, there are no qualitative Grad-Cam evaluations specific to ultrasound borders. Nevertheless, we argue the correct understanding of how new deep-neuronal networks perform border segmentation is crucial to enable reliable diagnoses and effective treatment planning [203], for tumour localisation [204], lesion detection [205], or quantitative volume evaluation [206–209]. Therefore, in chapter 2.4, we propose to use Seg-Grad-CAM to analyse the network's activation for borders, particularly differentiating between distinct/evident and completed/interpolated edges.

Chapter 1

Evaluation of labelling variability

Expert annotations on ultrasound images are influenced by several factors such as the anisotropy and the orientation-dependency arising from the physical acquisition principles behind this imaging modality, see section Scientific Background. As a consequence, annotations are susceptible to substantial variations along the borders and to significant discrepancies between different operators but also for the same operator repeating the task twice. Moreover, acoustic waves interact with surface layers, giving rise to regions of attenuation, shadowing, or indistinct boundaries. These artefacts increase the challenge of this time-consuming task, which is nonetheless indispensable for accurate volumetric computations in the diagnosis and follow-up of medical conditions like hyperthyroidism [6] or Duchenne muscular dystrophy [4].

Understanding the variability of annotations is important for clinicians and engineers designing segmentation architectures, as awareness of where challenging regions occur has the potential of impact their performance. Non-learning-based tools exist to estimate how much confidence can be attributed to ech pixel in an ultrasound images. They are know as confidence maps (CMs). In this work, we evaluate the influence of Confidence maps (CMs) [193] given as additional information to a neural networks, and relate it to label variability. CMs are image-based simplified approximations of wave propagation through the imaged mediums. They were first introduced by Kalamaris *et al.*in 2012, where they were used to estimate an uncertainty value for each pixel in the image. The problem is formulated as a label propagation on a graph solved with a random walker algorithm [210].

1.1 Related work

Various approaches have been devised to tackle the inter-observer variability in the context of medical image segmentation. Among these methodologies are fusion techniques such as majority voting [187] or STAPLE [188]. Fusion methods focus on aggregating multiple annotations to create a more reliable ground truth or reference segmentation for training or evaluation purposes. Majority voting [187] relies on annotations provided by multiple annotators for a particular image. Each pixel or voxel in the image is assigned a label that the majority of annotators agree upon. In contrast, STAPLE [188] (Simultaneous Truth and Performance Level Estimation) takes into account not only the majority agreement but also the performance level of individual annotators. As a result, STAPLE provides probabilistic segmentation label map, containing the likelihood of each label assignment for every pixel or voxel within the image. This probabilistic label map is conceived considering both concordance and discordance among annotators'. A primary drawback of the above fusion techniques is their demand for multiple annotations for a single image, which makes them time-consuming and expensive. Additionally, Jungo *et al.* [190] examine the impact of prevalent image label fusion techniques on the process of label uncertainty estimation. Their results highlight a negative effect when associating of fusion methods to deep-learning methods to obtain reliable estimates of segmentation uncertainty.

In light of these challenges, different approaches have surfaced, aiming to directly integrate segmentation uncertainty into the model's predictive capacity through the utilisation of annotations from either single or multiple experts. Baumgartner *et al.* [189] introduced a hierarchical probabilistic model to model segmentation across varying resolutions. By adopting this hierarchical paradigm, they incorporate an uncertainty representation into the segmentation process. Following a different direction, Monteiro *et al.*put forward the concept of Stochastic Segmentation Networks (SSNs) [191], modelling aleatoric uncertainty within image segmentation. The distinguishing feature of SSNs lies in their capability to capture probability distributions with spatial coherence from which is then possible to sample multiple credible hypotheses. In a third direction, Rousseau *et al.* [192] highlight that while neural networks are often trained to optimize segmentation accuracy, less emphasis has been placed on calibrating the confidence scores associated with their predictions. Well-calibrated confidence scores are crucial as they provide meaningful information to users about the reliability of the network's predictions. They explore various post-hoc calibration methods and find a correlation between loss functions and calibration performance. However, the above strategies necessitate alterations to architectures and detailed hyper-parameter optimisation, which are costly in terms of time and computational resources.

Other approaches exist to studying label variability without modifying the network's architecture. such as label smoothing [211–213], temperature scaling [214], annotator error disentangling [215], and non-parametric calibration [216]. Most of these techniques have been applied on MRI or CT. We align ourselves with these ideas, applying our method to ultrasound, a modality characterised by blurred edges, low signal-to-noise ratio, speckle noise, and other challenges.

Confidence maps (CMs) have been used until now for improving reconstruction [194], registration [195], and non-deep-learning bone segmentation [196] in ultrasound. To the best of our knowledge, we are the first to study ultrasound confidence maps in the context of neural networks for semantic segmentation and analyse its influence on border variability. We propose to incorporate the CM either as a second channel or in the loss function. We evaluate our approach on two volumetric datasets, and under 3 architectures and 2 different loss functions. Architectures include SOA methods such as UNet [48], UNet transformer [49] and Attention UNet [45]. We also qualitative demonstrate that variability of CM-based model closely reflects the variability of expert annotations.

1.2 Method: Using confidence maps in neural networks

Our methodology consists of two main steps computing the confidence maps, and integrating them as an additional input channel to the network, or as part of the loss function.

Pre-calculating the "Confidence Maps" (CM).
We follow the approach presented by Karamalis *et al.* [193]. The method seeks to assign uncertainty values to individual pixels within ultrasound (US) images by introducing a simplify model for wave propagation based on random walks on a graph. When pressure waves traverse tissue during an ultrasound examination, they undergo a series of intricate interactions, including transmission, reflection, absorption, refraction, dispersion, and diffraction. Consequently, the fidelity of recorded wave intensities diminishes as they propagate through the tissue medium. In the context of this model, an ultrasound image is graphically represented as a graph wherein the uppermost pixels serve as source nodes, and the lowermost pixels act as sink nodes. Edges connect neighbouring pixels and edge weights reflect the transmission likelihood according to the nodes geometrical placement and their geometrical similarity. Thereby, a graph represents the sound flow from the source pixels at the top towards the sink nodes (as depicted in Figure 1.1-b). The transmission is modelled as a random walk following a path extending from the upper to

the lower region, approximately orthogonal to the direction of the beam or scanline. Yet, slight deviations in the horizontal and diagonal directions remain possible. More specifically, graph edges are defined in accordance with an 8-neighbourhood rule, establishing connections between adjacent pixels. The weights assigned to these edges encompass diverse physical properties inherent to ultrasound:

- The vertical wave propagation involves an exponential attenuation described by the Beer-Lambert law, governed by a parameter α.
- An additional penalisation factor regulated by the parameter β is introduced to effectively capture the interplay between reflections and transmissions across tissue boundaries, particularly when neighbouring pixels exhibit distinct intensities.
- To account for beam shape effects, a penalty mechanism is employed for horizontal and diagonal propagation.

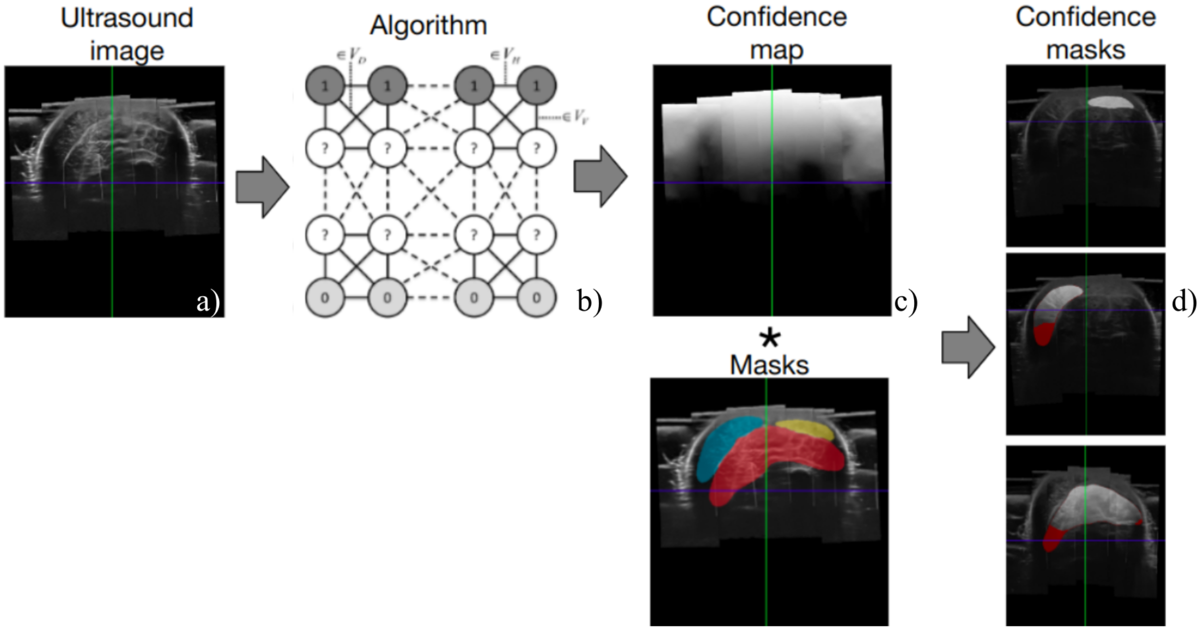


Figure 1.1 – From confidence maps to confidence masks. From left to right: a) US image, b) image graph representation, c) Confidence map and label masks, d) Confidence masks of the gastrocnemius medialis, lateralis and soleus muscles.

The computational process depends on the resolution of a linear system of equations. Once the graph has been defined, the confidence maps seek to estimate the random walk likelihood for all undefined pixels/nodes in the graph. These values are estimated through the resolution of a linear system. The estimated values are then interpreted as "confidence": values are high near the probe and progressively decrease when interfaces are found or through the effect of attenuation. An example of estimated confidence map values is presented in Figure 1.1-c.

Ultrasound images frequently manifest regions and islands characterised by diverse degrees of certainty. This variability has the potential to confound the network, as it might erroneously interpret all pixels as uniform representations of truth. To assimilate this information into the network's functioning, two distinct approaches are investigated:

1. Augmenting the input to the network by adding CM as an additional channel. Consequently, the input becomes $[\mathbf{X}|\mathbf{CM}]$, where $\cdot|\cdot$ represents concatenation.
2. Combining CMs with the labels to form a "Confidence Mask" $(Y \cdot CM)$ (see Figure 1.1(c,d)). The CE confidence loss is then defined over the m voxels of the image as follows

$$\text{CE}_{conf}(\mathbf{Y}, \hat{\mathbf{Y}}, \mathbf{CM}) = -\frac{1}{m}\sum_{i=1}^{m}(Y_i \cdot CM_i) \cdot \log\left(\hat{Y}_i\right) \tag{1.1}$$

   Respectively the DiceCE confidence loss is defined over the m voxels of the image with λ_{Dice} and λ_{CE} being the weights of each loss:

$$\text{DiceCE}_{conf}(\mathbf{Y}, \hat{\mathbf{Y}}, \mathbf{CM}) = \lambda_{Dice} * \text{Dice}(\mathbf{Y}, \hat{\mathbf{Y}}) + \lambda_{CE} * \text{CE}_{conf}(\mathbf{Y}, \hat{\mathbf{Y}}, \mathbf{CM}) \tag{1.2}$$

Guiding segmentation networks with Confidence Maps: At the core of our approach lies the principle of incorporating pre-calculated Confidence Maps (CMs) within the training of a segmentation deep neural network. Let $\mathbf{X} \in \mathbb{R}^{W \times H \times D}$ denote the input volume, while $\mathbf{Y}$ and $\hat{\mathbf{Y}} \in \mathbb{R}^{W \times H \times D \times C}$ respectively represent the labels encoded in a one-hot manner and the network's predictions across C classes. The initial step involves the computation of a CM from the input image, denoted as $\mathbf{CM} : \mathbf{X} \mapsto [0,1]^{W \times H \times D}$. In our first proposition, we advocate the utilisation of CMs as an additional input channel, such that $[\mathbf{X}|\mathbf{CM}]$, where $\cdot|\cdot$ signifies concatenation.

Our second proposition uses CMs as prior knowledge to weight the labels. To this end, we build a "confidence mask", resulting from the product $(Y \cdot CM)$, where the "$\cdot$" operator is the element-wise multiplication. The confidence mask sets the stage for formulating a Cross-Entropy Confidence Loss over the m voxels within to the image. Formally:

$$\text{CE}_{conf}(\mathbf{Y}, \hat{\mathbf{Y}}) = -\frac{1}{m}\sum_{i=1}^{m}(Y_i \cdot CM_i) \cdot \log\left(\hat{Y}_i\right) \tag{1.3}$$

1.3 Experiments and results

1.3.1 Understanding expert and network variability

We propose to study the variations inherent to expert annotations and network predictions relying on a Monte Carlo (MC) dropout approach. This process involves evaluating 100 annotation of one single expert over the same single image at different times. Then, we activate dropout layers of our two models during inference and obtain 100 predictions for each. Finally, we compute the pixel-wise entropy over 100 instances for each case. Original image and entropy results are presented in Figure 1.2. The entropy of expert annotations (1.2b) showcases anisotropy, particularly evident in challenging regions such as the convergence of three muscles and in deeper anatomical areas. When compounding, the entropy map of a simple UNet model (UNet-1ch-Dice) along side our model incorporating CMs as a second input channel (unet-2ch-Dice), see Figure1.2c and 1.2d, we observe the utilisation of CMs yields results closer to the entropy of experts heightening locations where experts face challenges. In this sense, MC visualisation of our model UNet-1ch-Dice show a potential to raise awareness about regions associated to high expert variability.

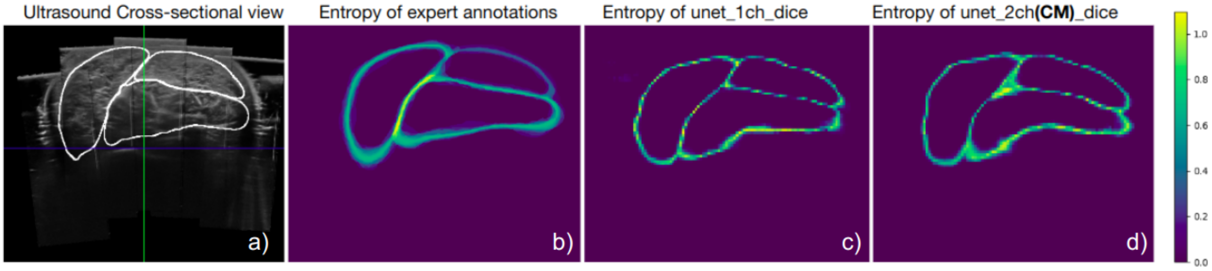

Ultrasound Cross-sectional view
Entropy of expert annotations
Entropy of unet_1ch_dice
Entropy of unet_2ch(CM)_dice
a)
b)
c)
d)
1.0
0.8
0.6
0.4
0.2
0.0

Figure 1.2 – Labelling variability analysis: Ultrasound image with the labels of the three lower limb muscles, b) Expert variability, c) MC dropout for a baseline UNet method. d) MC dropout of our method with CMs as the second channel.

1.3.2 Evaluating the influence of confidence maps across multiple network configurations

The following series of experiments offer a means to assess the impact of Confidence Maps (CMs) across ten different network configurations. Employing the UNet architecture as a baseline, CMs were either directly integrated into the network as a supplementary channel, denoted by the designation -2ch-, or embedded within the loss function with the structure $UNet - ch - (loss)$**conf**. Three distinct loss functions were considered: Dice, Cross-entropy (CE), and a composite DiceCE loss. The experiments were categorised into the following groups:

Method	**CM as Input**	**Loss: Dice**	**Loss: CE**	**CM in Loss**
unet_1ch_Dice		✓		
unet_1ch_CE			✓	
unet_1ch_DiceCE		✓	✓	
unet_2ch_Dice	✓	✓		
unet_2ch_CE	✓		✓	
unet_2ch_DiceCE	✓	✓	✓	
unet_1ch_CEconf			✓	✓
unet_1ch_DiceCEconf		✓	✓	✓
unet_2ch_CEconf	✓		✓	✓
unet_2ch_DiceCEconf	✓	✓	✓	✓

These models are then evaluated on two datasets, the thyroid 3D dataset and the Leg-3D dataset; refer to chapter 3D datasets (1.1).

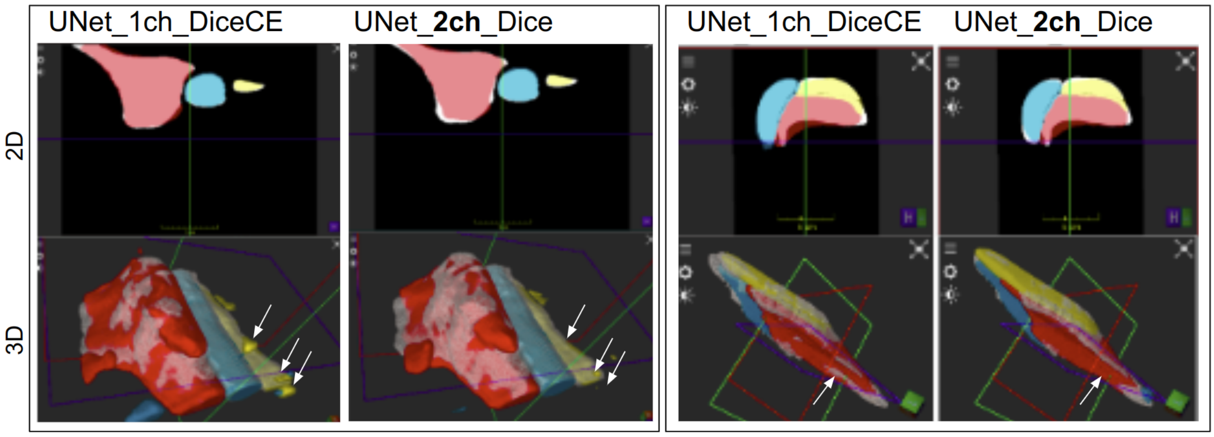


Figure 1.3 – Qualitative comparison of our UNet best CM configuration (*unet_2ch_Dice*) and the baseline (*unet_1ch_CE*) in terms of performance. We show results on two datasets (left) Thyroid, then right Low-limb. On the right is the Thyroid, and on the left, the Lower limb. See chapter 3D datasets 1.1 for more details.

Qualitative outcomes are shown in Figure 1.3, illustrating a comparison between baseline (*unet_1ch_DiceCE*) results and our best CM-based configuration (*unet_2ch_Dice*), on two datasets. In the case of the thyroid, there is an observable reduction in the occurrence of isolated regions (as illustrated in Figure 1.3-left). Conversely, predictions exhibit smoother transitions for the lower limb dataset (as showcased in Figure 1.3-right), suggesting improved interpolation capabilities.

We quantitatively evaluate the three configuration groups on different metrics (Dice, Intersection over union, Average surface distance, Hausdorff, miss rate, precision). Considering a trade-off between Dice score and Hausdorff score, the two most optimal configurations are *unet_2ch_Dice* and *unet_1ch_CEconf*. These outcomes are detailed in Figure 1.4, where it is marked with numbers of the most favourable tree configurations. To ease the comparison, figure 1.4 resumes the box plots of the best configurations of each group. To ensure the robustness of the results across different participant divisions, a 3-fold cross-validation was executed. Our findings indicate that the inclusion of Confidence Maps (CMs) generally reduces the standard deviation of DSC scores. Notably, while the HD metric of CM configurations either remains similar or experiences slight increments for the thyroid dataset, the positive impact of CMs on the muscles dataset is evident. This observed behaviour can be attributed to the smoothness of muscle shapes.

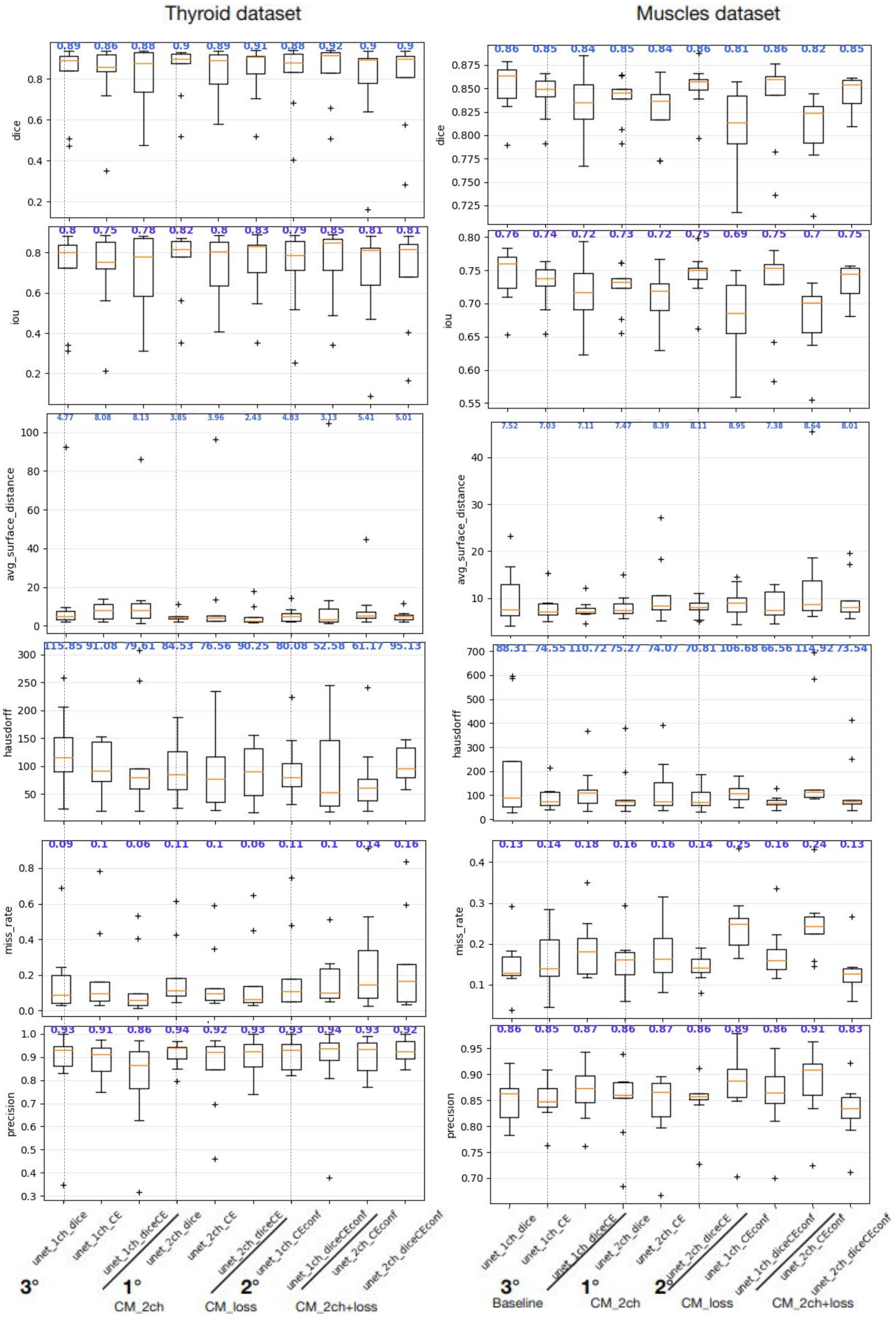


Figure 1.4 – Quantitative results of UNet configurations with confidence maps: First and second columns report the thyroid and muscle metrics results, respectively. The best four performing methods are ranked 1°, 2°, 3°.

1.3.3 Impact of confidence map hyperparameters

Confidence maps depend on two main hyperparameters, α and β, termed as the weighting ratio and adaptiveness, respectively. Confidence maps resulting from multiple variations of such parameters, are illustrated in Figure 1.5.

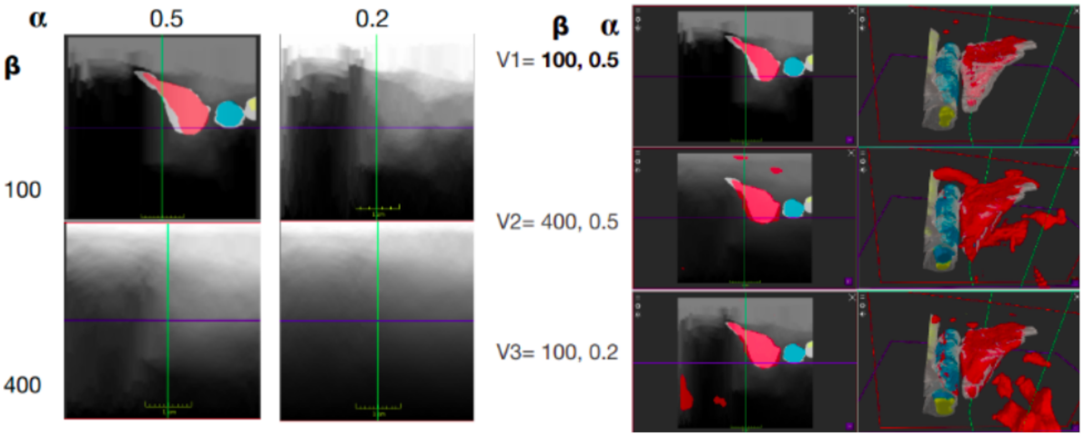


Figure 1.5 – Qualitative results of confidence maps hyperparameters: **(Left)** A qualitative view of the confidence maps with different hyperparameters α and β (resolution and adaptiveness). **(Right)** A 3D view of the inference of *unet_2ch_dice* using the confidence maps as the second channel, the configuration $\beta = 100$, $\alpha = 0.5$ decreases the number of islands like outliers.

The subsequent experiments study the impact of different CM hyper-parameters in the network's performance. Three sets of hyper-parameters were considered, each evaluating under two schemes: one integrating the confidence map as an extra channel, and the other embedded it within the cross-entropy loss function. Throughout all these experiments, the foundational framework employed was the UNet architecture.

α, β	Network	DSC	IoU	Precision	Miss rate	ASD[mm]
100, 0.5	2ch-Dice	0,84±0,03	0,75±0,04	0,91±0,04	0,20±0,03	4,41±5,13
400, 0.2	2ch-Dice	0,84±0,03	0,75±0,05	0,88±0,06	0,14±0,03	11,87±5,31
100, 0.2	2ch-Dice	0,83±0,02	0,73±0,03	0,84±0,04	0,15±0,03	14,84±1,15
100, 0.5	1ch-CEconf	0,83±0,04	0,73±0,06	0,87±0,05	0,17±0,04	15,57±2,68
400, 0.5	1ch-CEconf	0,82±0,03	0,72±0,05	0,90±0,06	0,20±0,02	10,15±2,55
100, 0.2	1ch-CEconf	0,82±0,04	0,72±0,07	0,86±0,06	0,16±0,03	13,58±4,67

Table 1.1 – Metrics comparison of the baseline (U-Net) network for two main uses of the confidence maps: as a second channel (2Ch) and in the loss (CEconf).

The results presented in Table 1.1 reveal that the method dependency on these parameters is relatively low. We observe that increasing adaptiveness (β) leads to a higher likelihood of diagonal and horizontal shifts, resulting in smoother images. Additionally, reducing resolution (α) leads to an aggregation of initial nodes at the image's uppermost region, causing the appearance of narrower columns. Quantitative results also indicate that employing a higher weighting ratio of resolution with $\alpha = 0.5$ and a lower adaptiveness with $\beta = 100$ improves results when incorporating confidence maps as an additional channel. This improvement can be attributed to the effective representation of layers within the confidence maps. Conversely, when incorporating confidence maps in the loss function, superior results are achieved with a higher weighting ratio resolution ($\alpha = 0.5$) and greater adaptiveness ($\beta = 400$). This enhancement is attributed to the fact that confidence maps using these parameters assign higher values to the confidence masks, leading to more effective penalisation of errors. For a visual representation of these findings, refer to Figure 1.5.

1.3.4 Evaluation of CMs on different Networks comparison

We assessed three additional architectures to explore potential enhancements in network architecture for the given task beyond the UNet: UNetr [49], Deep-Atlas [52] and Attention UNet [45]. The Friedman statistical significance test was employed to evaluate the null hypothesis that "all methods perform equally". Following this, pairwise post-hoc cross-validation was conducted between the baseline networks and the modified networks ($2ch$ and $DiceCEconf$). Methods that reject the null hypothesis, with $p < 0.05$, for the Hausdorff distance are denoted with an asterisk ($*$). The results of segmentation metrics for one of the subject of the dataset are presented in Table 1.2. We observe that the incorporation of CM slightly improves segmentation metrics.

Network	DSC ↑	precision ↑	miss rate ↓	ASD ↓
UNet 1ch Dice CE	0.84 ± 0.02	0.84 ± 0.06	0.16 ± 0.04	8.37 ± 2.19
UNet 2ch Dice∗	**0.85 ± 0.01**	0.85 ± 0.01	**0.16 ± 0.02**	**8.20 ± 1.81**
UNet 1ch DiceCEconf∗	0.81 ± 0.01	**0.86 ± 0.05**	0.18 ± 0.04	8.36 ± 0.78
AttUNet 1ch CE	0.57 ± 0.49	0.58 ± 0.50	0.44 ± 0.49	40.00 ± 15.01
AttUNet 2ch Dice ∗	**0.86 ± 0.00**	**0.85 ± 0.04**	**0.12 ± 0.04**	**7.03 ± 1.08**
AttUNet 1ch DiceCEconf	0.57 ± 0.50	0.59 ± 0.51	0.44 ± 0.48	40.00 ± 14.09
DeepAtlas Dice	0.82 ± 0.02	0.84 ± 0.05	0.18 ± 0.06	**11.22 ± 1.54**
DeepAtlas 2ch Dice ∗	0.84 ± 0.01	0.84 ± 0.06	0.15 ± 0.07	11.89 ± 6.00
DeepAtlas 1ch DiceCEconf ∗	**0.85 ± 0.01**	**0.84 ± 0.02**	**0.13 ± 0.02**	11.74 ± 5.50
UNETR 1ch CE	**0.66 ± 0.06**	0.72 ± 0.12	**0.37 ± 0.02**	23.52 ± 8.97
UNETR 2ch Dice∗	0.48 ± 0.12	0.77 ± 0.11	0.62 ± 0.12	23.74 ± 8.72
UNETR 1ch DiceCEconf∗	0.56 ± 0.07	**0.82 ± 0.13**	0.55 ± 0.09	**19.54 ± 9.65**

Table 1.2 – Metrics on the muscle dataset for the baselines and the modified versions of 4 different networks: UNet3D [48], attention Unet (AttUNet) [45], DeepAtlas [52] and UNet Transformer(UNETR) [49].

Figure 1.6 presents qualitative images from different trainings for the thyroid dataset.

Notably, UNetr exhibits lower accuracy, potentially due to the requirement for more data. Conversely, A-UNet achieves notably high accuracy by incorporating CM. This improvement can be attributed to the attention layers effectively utilising the additional informative cues provided by the CM.

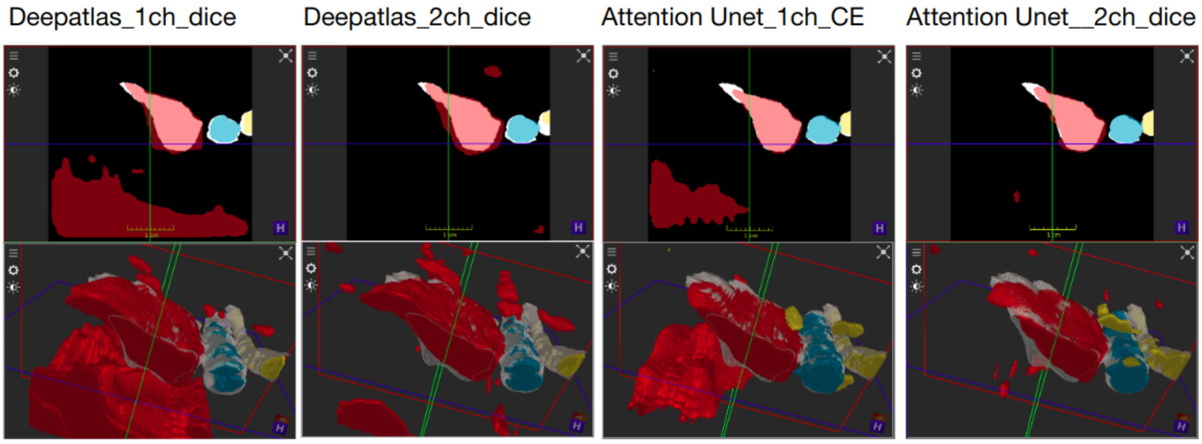


Figure 1.6 – Qualitative results of different networks with CM addition. From left to right: Deep-atlas with and without CMs and Attention UNet with and without CMs. Evaluation on the thyroid dataset.

1.4 Discussion & future work

In summary, this study introduced a novel strategy to enhance the label variability awareness for deep learning ultrasound segmentation methods approximating the inherent variability in expert annotations. We made use of pre-calculated ultrasound Confidence Maps (CMs). The method effectively estimates uncertainties that arise due to fundamental ultrasound wave propagation principles, influencing the annotators' decisions. The CMs are incorporated into the network either as an extra input channel or within the loss function, guiding the network to predict segmentation that faithfully captures expert-like variability in borders. The variability tends to produce predictions with low entropy borders for confident regions and high entropy borders for uncertain regions, thus offering multiple solutions for physician assessment. Notably, the CM loss method exhibits two benefits: it does not increase the number of parameters, indicating its versatility across architectures, even as a fine-tuning strategy post-transfer learning. Our experimental outcomes also underscore that both proposed approaches involving CMs do not penalise convergence during model training. Additionally, the pre-computation of CMs is simple, achieved by solving a linear system with a sparse matrix. Finally, our evaluation spanned two datasets—private and public—to ensure the robustness of our findings.

Moving forward, there are several promising avenues for future research that could enhance the application of confidence maps (CMs). Also, while our study focused on Dice loss and cross-entropy loss, there is room to investigate the impact of alternative loss functions or their combinations on the effectiveness of CMs. This exploration could further leverage the information provided by CMs. Additionally, future research could delve into the efficacy of CMs in scenarios with limited available data. Understanding how CMs perform in situations where data is scarce or imbalanced could provide valuable insights into their robustness and generalizability across different contexts. Another intriguing direction lies in refining the confidence maps themselves. Specifically, avoiding the assumption that fixes the bottom region with all its pixels set to zero.

In summary, the future trajectories of research involve automating the confidence maps extraction, exploring diverse loss functions, assessing their behaviour with limited data, and refining their design to maximise their contribution.

Chapter 2

Ultrasound segmentation analysis via distinct and completed anatomical borders

Medical professionals often address ultrasound segmentation difficulties by initially identifying key structures, then focusing on specific areas in the image, and finally differentiating the structures that define tissue boundaries [69]. However, in areas with reduced visibility a border completion is carried by the annotator relying on prior knowledge and expertise. Moreover, since ultrasound is a dynamic imaging technique, it necessitates an understanding of the anatomy and the relative probe position for a precise identification of the structures present in the images. In this chapter, we argue that these additional requirements should be considered when assessing ultrasound deep-learning segmentation techniques. Consequently, we recommend separately examining the capability of such methods to outline the distinct versus challenging boundary zones. To this end, we propose evaluating the performance of ultrasound segmentation networks based on attribute maps in boundary regions on top of boundary metrics evaluating the accuracy, Dice, Hausdorff, etc. Notably, we differentiate the performance in evident and completed borders, as similar to medical professionals. We distinguish between the performance of evident and completed borders, as we anticipate that segments with clear boundaries will be simpler for a network to identify compared to those requiring expert prior knowledge. While deep learning methods tend to succeed on distinct borders when sufficient data is available for training, we investigate in this paper how different architectures behave on ambiguous boundary zones. Despite the significance of this question, the majority of neural networks for ultrasound segmentation have been evaluated directly against full expert labels, with metrics such as the Dice score or the Hausdorff distance, which do not explicitly capture the network's behaviour on different types of boundaries. Especially for ultrasound imaging, where boundaries are known to be challenging to detect, a deeper understanding of how networks segment borders seems crucial for solving the task.

2.1 Related work

2.1.1 Evaluating learned information

In recent years, deep learning methods have become the prevailing approach for addressing image segmentation tasks [172]. However, a substantial drawback of these methods lies in their limited interpretability, which hinders the comprehension of their decision-making processes [217]. This deficiency becomes particularly critical in medicine, where there is a higher demand for model accuracy and explainability to ensure reliable diagnoses and effective treatment planning [203]. Among explainability techniques for deep convolutional neural networks, Gradient-weighted Class Activation Mapping (Grad-CAM) stands out as a simple and effective technique. Initially introduced by Selvaraju *et al.* [199] in 2016. It offers a heat map visualisation approach to reveal layer activation.

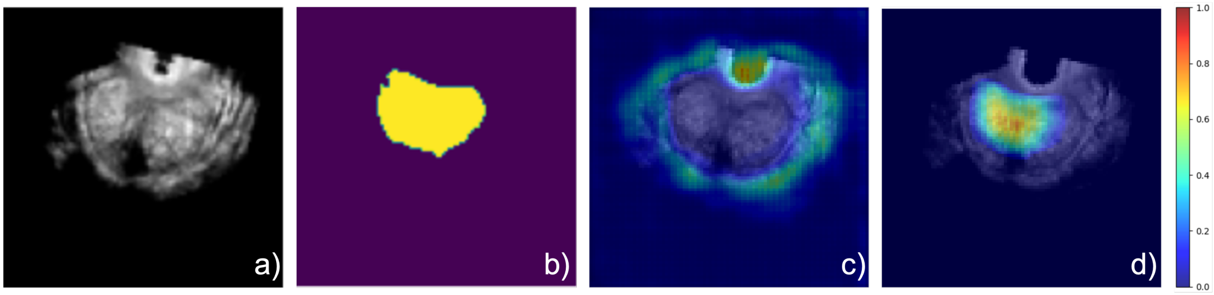


Figure 2.1 – GradCAM of the prostate UNet network: a) Ultrasound image, b) Network Prediction, c) Grad-CAM for target-label 0 (background) d) Grad-CAM for target-label 1 (prostate)

Several studies have adopted Grad-CAM to understand the information processed by classification networks [199, 200, 218]. In the context of medical image segmentation [172], the role of Grad-CAM is to reveal as a heat map the image regions that contribute the most to the decision-making. Yet, to our knowledge, no qualitative assessment of ultrasound boundary predictions using Grad-CAM has been conducted before. Some works have used Grad-CAM to ensure accurate diagnoses and efficient treatment strategies [203], or for pinpointing tumours [204] and identifying lesions [205] or diseases (e.g. Covid-19 [218]). Grad-CAM has also been used for CT and MRI image segmentation [206–209] in a qualitative fashion. This work searches instead to examine network activations relative to boundaries, especially distinguishing between evident and completed categories, using

Seg-Grad-CAM. Figure 2.1 presents an example of applying Seg-Grad-CAM to a prostate segmentation network when considering the full segmentation and not just the boundaries. In this work, we propose to detect the border regions based on a post-processing of the ground truth labels. Then, we adapt Seg-Grad-CAM to the separation of distinct and completed borders. Finally, we propose several border-aware metrics to compare the performance of 3 neural network architectures of 4 public and private datasets.

2.1.2 Public Ultrasound Datasets

Table 2.1 presents an overview of the open-source and private datasets US datasets used in this work. More details can be found in the Fundamentals chapter, Section 3D datasets 1.1. We briefly describe again the datasets below. Example images from the different datasets can be seen in 2.1.

	Name	Labels	Size of set	Resolution	Probe	Ref
2D	UTP Nerves	Nerves in 4 body points: sciatic, ulnar, median & femoral	1857 images	360 × 279	Linear 4-16MHz	[63]
3D	Thyroid	Thyroid, carotid artery & jugular vein	16 participants: 32 volumes	380 × 330 × 300	3D curvilinear probe 64 channels	[5]
3D	Prostate	Prostate	40 Volumes	230 × 230 × 70	Rectal 4-16Mhz	[H]
3D	Low-limb	Leg	44 Volumes	230 x 230 x 70	Linear 4-16Mhz	[50]

Table 2.1 – Overview of Ultrasound Datasets used in this chapter. Reference "H" denotes in-house data.

The thyroid dataset by Kronke *et al.* [5] offers 32 3D volumes from 16 individuals, with annotations of the thyroid gland, jugular vein, and carotid artery. Images have a resolution $380 \times 330 \times 300$ pixels with $0.12mm$ of voxel spacing. The acquisition was made using a 3D curvilinear probe with a magnetic tracking system.

The Nerve-UTP-2D dataset, presented by Jimenez *et al.* [63], consists of 691 2D ultrasound images sourced from a SONOSITE Nano-Max device with annotations by a certified anesthesiologist. Covering nerve types such as sciatic, ulnar, median, and femoral, these images of 360×279 pixels, aid in peripheral nerve studies and are valuable for training models in nerve identification.

The prostate 3D dataset is an in-house dataset containing 40 3D US volumes (230 × 230 × 70 pixels, 0.27mm voxel spacing) of the prostate for examination of cancer, obtained using a rectal ultrasound probe. This dataset correlates ultrasound and MRI data to ensure accurate prostate contours.

Low-limb muscles 3D dataset comprises 44 ultrasound volumes of legs, focused on muscles such as the gastrocnemius and soleus, using a unique freehand ultrasound tracking method. This method accommodates comprehensive imaging from knee to ankle, filling a voxel grid (564 × 632 × 1443) with an average voxel spacing of $0.277mm^3$.

Our focus in this paper is on evaluating the quality of border predictions. To this end, our experiments are based on existing and well-studied architectures, namely:

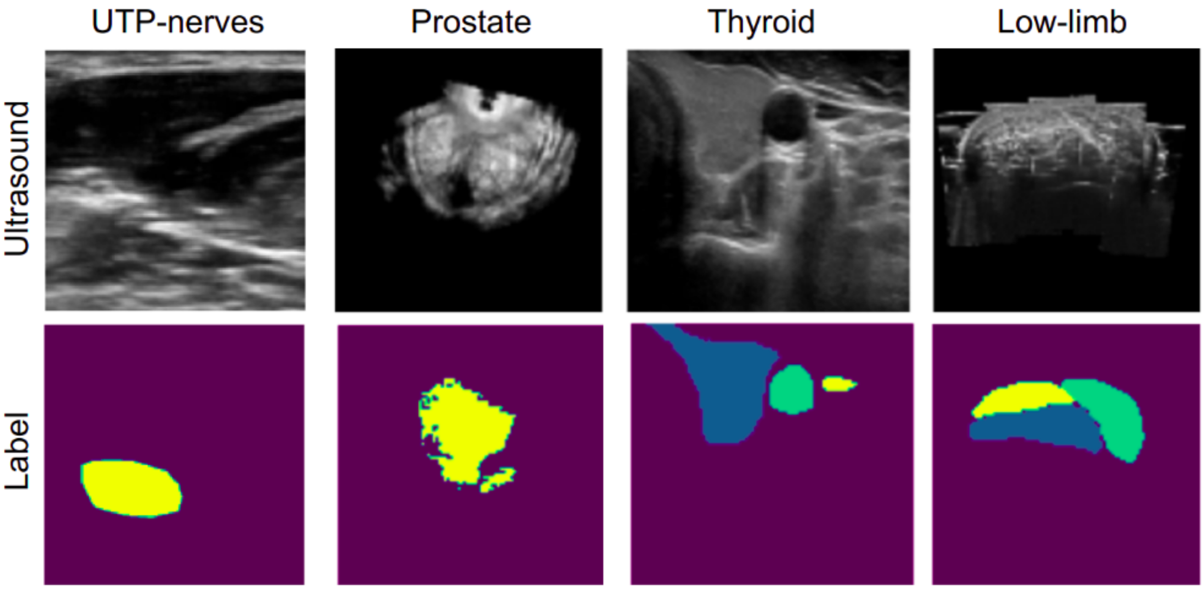


Figure 2.2 – 4 Ultrasound datasets: The first 2 datasets contain binary segmentation of the nerves and the prostate. While the last 2 multi-label datasets contain 3 labels each, corresponding to the thyroid(blue)-carotid(green)-jugular(yellow) and the Solius(blue)-gastrocnemius Medialis(green)- and gastrocnemius lateral (yellow).

- a classical UNet [48].
- an attention UNet [45], incorporating attention modules in the skip-connections.
- the recent UNeTR [49], which combines transformer blocks in the encoder with a convolutional decoder.

For the experiment in Section 2.3, we also consider a multi-task Y-Net [219], an encoder-decoder architecture mixing a segmentation with a classification task at the bottleneck. This configuration has given promising outcomes for breast cancer tumors [219] and Chest x-rays [220].

2.2 Proposed clinician-inspired border evaluation

The essence of our approach lies in using the Seg-Grad-Cam to assess the quality of segmentation networks in defining completed and distinct borders. This distinction is crucial for deciphering the nuances of expert annotations. Unlike CT or MRI scans, where similar intensity regions are more discernible, ultrasound imaging presents non-uniform intensity patterns due to the varying acoustic interactions within tissues. This complexity often increases the challenge of structure delineation. Consequently, practitioners rely on their knowledge of anatomy and previous empirical observations, with a particular emphasis on identifying tissue interfaces that exhibit pronounced reflections. Evident borders are identified as the shiny areas in the ultrasound images, while the concept of completed borders refers to a process which is common in clinical practice, physicians interpolating and connecting discontinuities to segment anatomically accurate and smooth structures, often by evaluating adjacent frames for context.

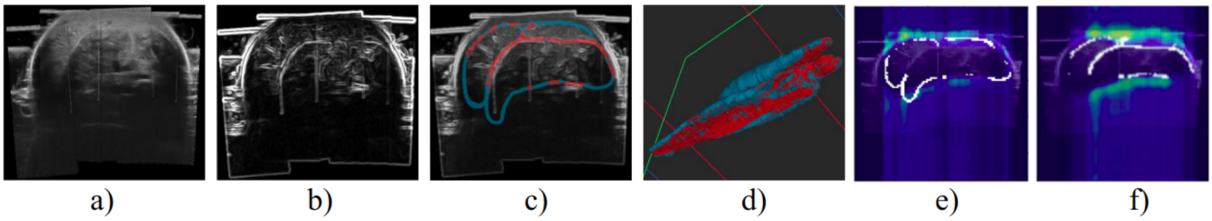


Figure 2.3 – Distinct (red) and completed (blue) borders. (a) Ultrasound images, (b) Edges on the image, (c) 2D cross-sectional view with borders, and (d) 3D-view of the borders. e)Activations for completed Borders, f)Activation for Distinct Borders.

We propose to compute a map of the distinct borders as the Hadamard product between the border label (B) and an edged-smoothed ultrasound image. The latter is created by applying smoothing and gradient filters convolving with kernels K_{smooth} and K_{sobel} respectively to the ultrasound image. Formally:

$$B_{distinct} = \textbf{Thres}(I * K_{\text{smooth}} * K_{\text{sobel}}) \odot B, \tag{2.1}$$

Completed borders are then obtained by calculating the complementary of evident borders. Being the border label the union of evident borders and completed borders. Figure 2.3 presents challenging ultrasound images for the lower limb in a cross-sectional view (a) and the extracted edges (b). The obtained distinct borders are in red and the completed in blue (c). We can observe in the 3D view (d) the high probability of getting completed borders in deeper regions where the ultrasound signal is low. The creation of images (e) and (f) is explained below.

Pre-evaluated Seg-Grad-Cam: The main objective of our method is to evaluate the ability of the network to delineate the borders, as physicians do. To this end, we visualise the Seg-Grad-Cam of the completed and distinct borders. We observe the activation areas for specific pixels in order to understand the decisions taken by the networks. The core of the Seg-Grad-Cam is the gradient computations. Seg-Grad-Cam heatmaps, hereafter denoted as $L^c_{\text{Seg-Grad-Cam}}$ are computed by first obtaining the gradient of the loss function with respect to a specific pixel's score class Y^c in the output segmentation map (for $Y^c \in \{B_{distinct}, B_{completed}\}$), then computing the global average pooling of these gradients at a given layer, and finally combining yhe pooled gradients with the activation maps and applying a ReLU activation. This process is represented by:

$$L^c_{\text{Seg-Grad-Cam}} = \text{ReLU}\left(\sum_k \left(\frac{1}{w \times h}\sum_i \sum_j \frac{\partial Y^c_{i,j}}{\partial A^f_{i,j}}\right) A^f\right), \tag{2.2}$$

where A^f is the activation map of the f^{th} feature map from the last convolutional layer

L of the segmentation network when processing image X, being $\frac{1}{w \times h}$ the global average pooling operator, and (i, j) the spatial coordinates in the feature map. Figure 2.3 presents the Seg-Grad-Cam for the pixels of $B_{completed}$ (e) and $B_{distinct}$ (f) borders.

2.3 Experiments and results

To assess the precision with which various networks delineate distinct and completed borders in ultrasound images, we conducted a series of experiments involving three distinct encoder-decoder network architectures. Each architecture was trained across four varied datasets, undergoing three separate cross-validation iterations. In our study design, patients' data were portioned into 70%, 20%, and 10% segments for training, validation, and testing phases, respectively. We implemented data augmentation strategies, including image flipping and normalisation. Networks were trained to the point of convergence using a batch size of two. The optimisation was carried out using the Adam optimiser with an initial learning rate of 0.001, and we employed a StepLR scheduler to adjust the learning rate by a factor of 0.5 every 10 epochs. The effectiveness of the models was measured by calculating the mean and standard deviation of the performance metrics on the test image sets.

2.3.1 Common border segmentation metrics evaluation

Dataset		LEG-3D-US			Nerve-UTP-2D	
Metrics	Dice	HD95	NSD	Dice	HD95	NSD
UNet	**0.789 ± 0.029**	**5.61 ± 1.62**	**0.962 ± 0.023**	**0.739 ± 0.283**	**11.28 ± 15.62**	**0.815 ± 0.319**
A-UNet	0.587 ± 0.048	8.29 ± 1.58	0.866 ± 0.036	0.73 ± 0.298	11.72 ± 11.07	0.791 ± 0.316
UNeTR	0.519 ± 0.161	14.75 ± 4.12	0.757 ± 0.107	0.661 ± 0.249	18.06 ± 18.56	0.719 ± 0.258
Dataset		Thyroid			Prostate	
Metrics	Dice	HD95	NSD	Dice	HD95	NSD
UNet	0.888 ± 0.053	**5.12 ± 2.71**	0.966 ± 0.018	**0.816 ± 0.075**	6.84 ± 4.02	**0.952 ± 0.048**
A-UNet	**0.892 ± 0.033**	6.81 ± 1.67	**0.972 ± 0.014**	0.815 ± 0.092	**6.71 ± 4.05**	0.947 ± 0.065
UNeTR	0.636 ± 0.105	27.20 ± 8.12	0.766 ± 0.034	0.523 ± 0.179	10.66 ± 6.28	0.77 ± 0.217

Table 2.2 – Evaluated metrics with mean and standard deviation of the test datasets for 2D and 3D architectures.

Our assessment of ultrasound segmentation models focused on a range of border metrics. The detailed findings of this evaluation are reported in Table 2.2. Our analysis indicates that the UNet Transformer architecture did not perform as well as its counterparts in all examined datasets. This under-performance is potentially linked to the heavy data demands of transformer architectures, resulting in a performance gap of 6% in 2D assessments that widened to 27% in 3D analyses. Choosing the more effective model between UNet and A-UNet proved challenging due to their similar performance metrics, especially noted in studies involving the thyroid and prostate datasets.

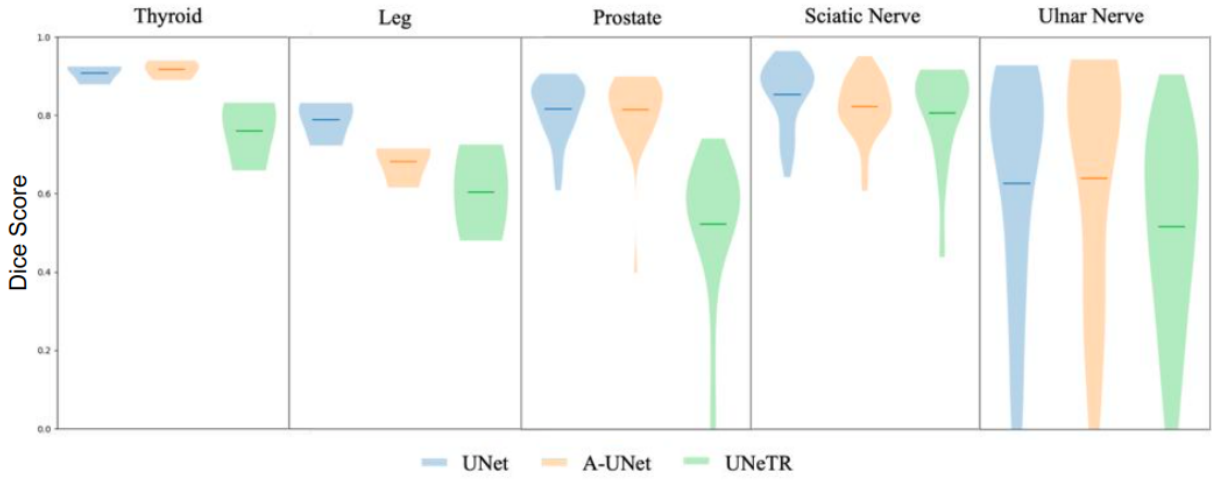


Figure 2.4 – Violin plots of the Dice score evaluated in the test datasets.

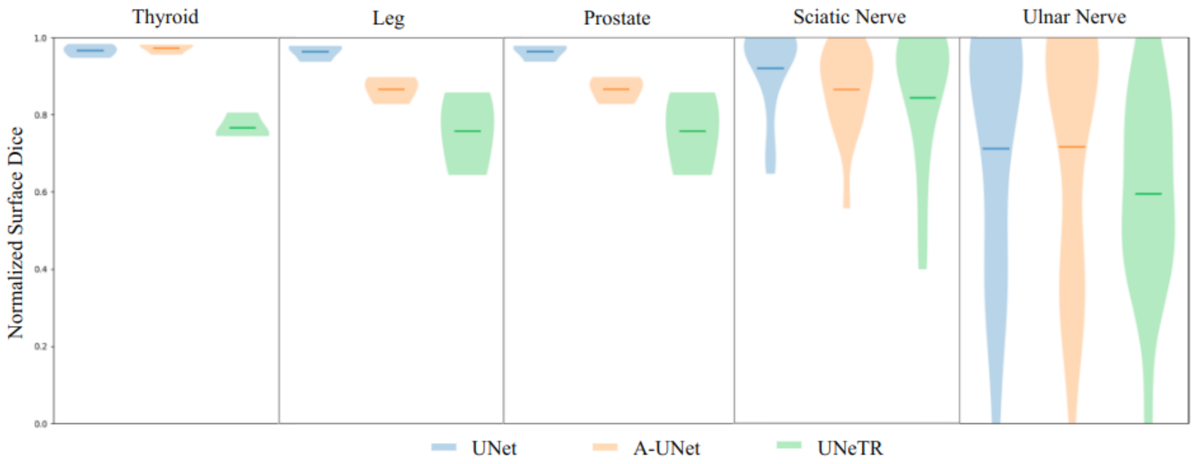


Figure 2.5 – Violin plots of the normalised surface Dice evaluation metric in the test datasets.

We undertook a deeper examination of metric consistency across individual slices using violin plots to visualise the data spread (see Figure 2.4 and Figure 2.5). This additional scrutiny, however, did not yield a conclusive ranking of architectures.

A Grad-Cam analysis of the segmentation architectures was used in order to determine potential differences between the UNet and A-UNet architectures, which proved challenging due to their comparable performance metrics, particularly within the thyroid and prostate datasets. We analysed the Seg-Grad-Cam for the background label to discern how each network considers local and contextual information of the foreground structures, as depicted in Figure 2.6. In exploring the gradient activation mechanisms of the respective architectures, we observed that both networks predominantly concentrate their attention on high-reflection zones, such as the interfaces between tissues. The focus of the background attention varies slightly, either becoming more scattered or more concentrated depending on the network. Nonetheless, such gradient activation maps do not target specific anatomical landmarks, thus offering limited insights for a thorough comparative analysis of the network architectures. Similar results were observed in maps of the foreground (the labels).

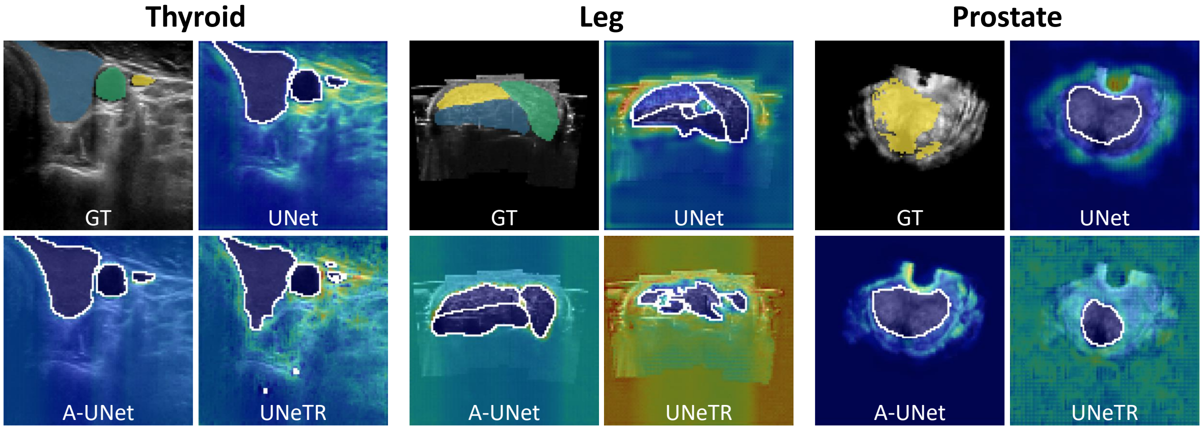


Figure 2.6 – 3D Seg-Grad-Cam evaluation of the background label with predictions border overlapped in white.

2.3.2 Evaluation of distinct and completed prediction

	LEG-3D-US			Thyroid		
	B_{TP}	B_{Distinct}	$B_{\text{Completed}}$	B_{TP}	B_{Distinct}	$B_{\text{Completed}}$
$\text{Reference}_{Border}[\%]$	100	54.8 ± 7.6	45.2 ± 4.3	100	86.7 ± 4.8	13.3 ± 4.8
$\text{UNet}_{Predictions}$	82.1 ± 3.3	42.2 ± 1.9	39.9 ± 2.1	86.3 ± 1.5	80.4 ± 2.7	5.9 ± 1.4
$\text{A-UNet}_{Predictions}$	75.5 ± 7.4	50.4 ± 2.4	35.1 ± 2.8	88.4 ± 2.3	83.0 ± 1.2	5.1 ± 1.3
$\text{UNetR}_{Predictions}$	37.2 ± 3.6	23.6 ± 2.5	13.6 ± 1.3	72.8 ± 7.9	70.5 ± 3.3	4.3 ± 2.1

Table 2.3 – True positive percentage in the test dataset predicted with respect to the complete border. Reference expresses the percentage of distinct and completed borders with respect to the total border.

Next, we focus specifically on the border's performance. Table 2.3 next, examines the lower limb and thyroid ultrasound datasets on the true positive rate metric of the borders, making a differentiation in evident and completed border accuracy. The methodology, detailed in Section 2.2, involves isolating the evident and completed borders, both summing 100% of the border. The initial row of the table reports the percentage of the reference border occupied by distinct and completed edges. The evaluation is done on the predictions from three segmentation models: UNet, A-UNet, and UNetR, against the complete reference border. The results indicate a notably higher prediction accuracy for distinct borders across all models, underscoring their enhanced performance in this specific area and highlighting the need for methods to perform more accurate border completion.

A qualitative evaluation highlights the image areas activated using Seg-Grad-Cam for distinct and completed borders (see Figure 2.7). The focal points of network gradient activations vary with the architecture employed. Typically, distinct border activations occur at interfaces that exhibit strong reflections. For instance, in segmenting lower limb muscles, there is heightened attention on the fatty layer that demarcates the muscle's top boundary, enhancing the definition of the border. When it comes to completed borders, the activation is influenced by the network's own attention method. For example, UNetR's approach of segmenting the volume into patches tends to impede the completion of certain areas. Conversely, UNet manages multiple feature scales via its encoder, which enables the integration of distinct borders in the prediction of completed ones. Moreover, the inclusion of additional anatomical structures—like bones or the trachea—provides contextual information that enhances the delineation of completed borders, mirroring the process by which medical practitioners identify landmark points to delineate missing or ambigous boundaries.

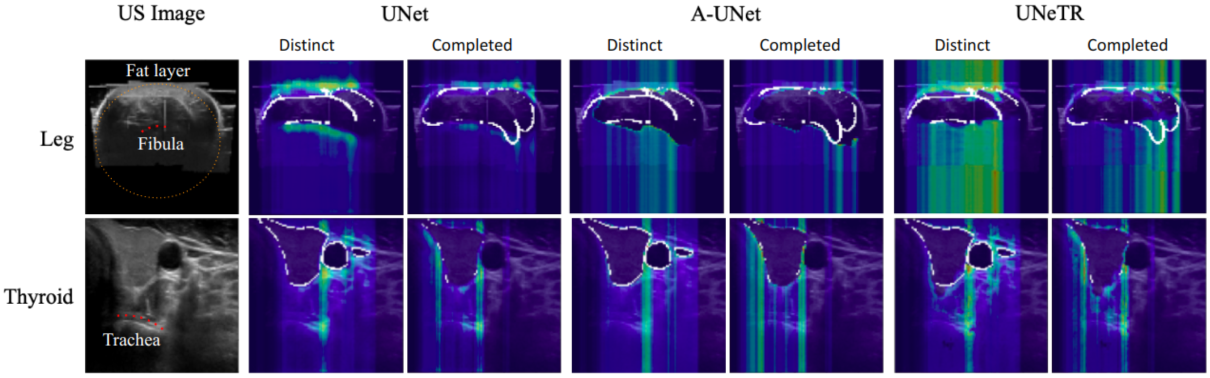


Figure 2.7 – Anatomical networks observations using Seg-Grad-Cam on borders: On the thyroid, we can identify the Trachea, while on the low limb, other structures like the Fat layer and the Fibula are also activated.

2.3.3 Y-net for multi-task learning paradigm

The training of UNet, Attention UNet, and UNet Transformer in a 2D multi-task learning setting faced similar challenges as with 3D datasets. The 2D image set includes 169 and 228 ulnar and sciatic images, respectively. Comparative analysis of their performance on sciatic and ulnar nerves is presented in Table 2.4.

For the first time, UnetR presents relatively better performances especially when segmenting ultrasound images for the sciatic nerve. For all the architectures, the Ulnar dataset seems to be the most challenging dataset. Our observations also indicate that considering a classification task in a multitask setting boosts segmentation capabilities.

	Sciatic Nerve			Ulnar Nerve		
Architecture	Dice ↑	HD95 ↓	NSD ↑	Dice ↑	HD95 ↓	NSD ↑
UNet – single	0.849 ± 0.083	**6.04 ± 15.62**	0.918 ± 0.110	0.627 ± 0.283	**11.77 ± 9.79**	0.711 ± 0.319
UNet – joined	0.817 ± 0.089	13.07 ± 14.31	0.884 ± 0.119	0.632 ± 0.232	20.18 ± 17.14	0.679 ± 0.261
YNet-U	**0.854 ± 0.093**	6.44 ± 4.47	**0.927 ± 0.122**	**0.676 ± 0.269**	12.16 ± 11.60	**0.777 ± 0.273**
A-UNet – single	0.839 ± 0.075	9.97 ± 10.43	0.864 ± 0.116	0.639 ± 0.298	11.78 ± 11.07	0.717 ± 0.316
A-UNet – joined	0.846 ± 0.069	9.91 ± 12.35	0.922 ± 0.091	0.681 ± 0.211	12.90 ± 11.42	0.756 ± 0.240
YNet-A	**0.859 ± 0.058**	**6.59 ± 4.35**	**0.922 ± 0.086**	**0.771 ± 0.175**	**9.34 ± 9.11**	**0.776 ± 0.210**
UNeTR – single	**0.817 ± 0.124**	**13.35 ± 11.91**	**0.843 ± 0.180**	0.516 ± 0.249	23.77 ± 18.56	0.595 ± 0.258
UNeTR – joined	0.799 ± 0.118	19.27 ± 14.63	0.813 ± 0.163	0.588 ± 0.284	**18.76 ± 17.19**	**0.678 ± 0.268**
YNet-TR	0.798 ± 0.115	17.15 ± 14.26	0.824 ± 0.167	**0.59 ± 0.273**	25.10 ± 19.32	0.655 ± 0.260

Table 2.4 – Quantitative comparison of multiple trainings of UNet and Attention UNet. Training on datasets encompassing one single nerve is denoted with "single", and training on datasets encompassing both nerves is denoted with "joined". YNet training is only performed on the joined datasets.

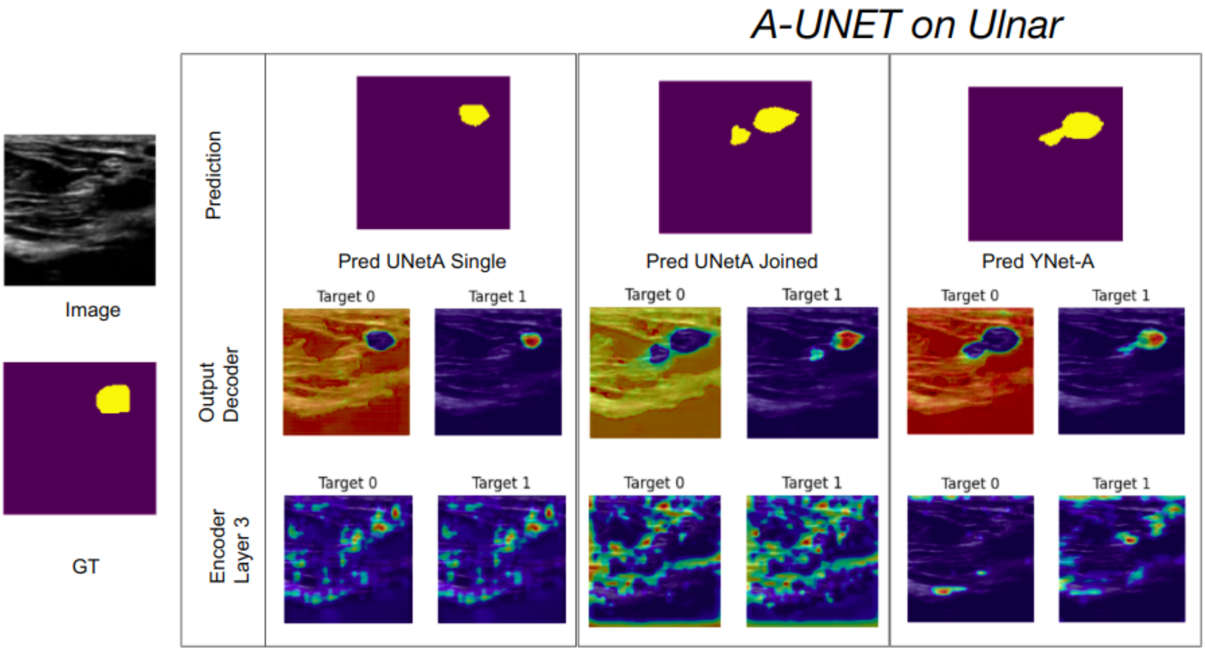


Figure 2.8 – Qualitative visualisation of the last and the 3rd encoder layer of the Attention UNet architecture when trained on the ulnar dataset only (first column), ulnar and sciatic datasets simultaneously (second column) and on both datasets with an auxiliary classification task included (third column).

We proceeded to use Seg-Grad-CAM to evaluate how the encoder behaviour changed when the additional classification task was added. Figure 2.8 presents a better focus of the encoder layer gradient activations around the important regions when the classification task is included. We observe some activation in anatomical key points such as bones, ligaments, veins or arteries. This could be due to the need for structures other than nerves to do the classification of anatomical points. Deeper research must be done in order to correlate them with the physician's key points used for segmentation.

Ablation study: Frequency of Classification-segmentation join training: Table 2.5 presents an ablation study of the YNet-UNet when training with two common methods to boost multi-task performance: The first three lines apply the classification loss in epochs multiples of the frequency factor. The second, applied for the two bottom lines, corresponds to initialising the network with pre-trained segmentation weights. Interestingly, variations perform worse than the initial configuration across all metrics for both nerves. Consequently, the initial training configuration, with no adjustments to the classification loss frequency or pretraining, stands out as the optimal choice.

2*Classification loss frequency	2*Pretraining	Sciatic nerve		Ulnar nerve	
		Dice	HD95	Dice	HD95
1	–	0.854	6.44	0.676	12.16
15	–	**0.860**	8.72	0.707	10.41
110	–	0.851	9.15	0.693	16.68
1	14 epochs	0.858	**5.97**	**0.724**	**10.25**
1	29 epochs	0.846	6.50	0.672	15.65
15	14 epochs	0.849	7.77	0.685	15.62

Table 2.5 – Quantitative comparison of different YNet-UNet training processes. The first row is equivalent to the initial YNet-U training.

2.4 Discussion and perspectives

This study concentrated on evaluating the performance of deep-learning ultrasound segmentation techniques in segmenting borders. The task was shown to be challenging even for human experts, given that ultrasound scans often show fluctuating intensities within the same tissue type while actual organ borders may be absent or ambiguous. We propose a new evaluation approach, acknowledging the fact that experts adopt different strategies when it comes to segmenting borders that are either clearly visible or ambiguous. Utilising the Seg-Grad-Cam, we analyse where the network concentrates its activations when completing anatomical borders. Our research uncovers that these networks mimic the adaptive focusing of medical experts, attending to distinct borders. They take into account not only the sharply defined edges but also significant anatomical landmarks, all while incorporating an inherently learned model of shape.

For a quantitative evaluation, we independently measure the network's ability to identify evident versus completed borders. This differentiation in metrics, which could be applied to other measures like Intersection over Union (IoU) or normalised surface distance, provides insights into how the network's activation varies with the complexity of the task at hand. Looking ahead, there is potential to investigate how networks internalise anatomical shape models and to enhance their effectiveness by directly incorporating the difference between distinct and completed boundaries into the loss function.

Finally, we contribute a publicly available LEG-3D-US dataset to the research community. This open-access resource is particularly valuable for enhancing medical diagnostic processes in biomedical research. The lower limb dataset captures the complex anatomy of the muscles, aiding in the development of better treatments, especially for muscle diseases like Duchenne muscular dystrophy. The dataset also has applications in sports medicine, helping to investigate the connection between athletic performance and muscle dynamics, thus contributing to the customisation of training programs.

Conclusions and perspectives

THE main objective of this thesis was the development and analysis of deep learning methods for the creation and segmentation of ultrasound volumes, towards improving the volume quality itself, but also the quantitative measurements extracted from it. This objective was to be achieved while considering certain constraints:

- A relatively challenging ultrasound dataset from 2D ultrasound images with noisy tracking.
- The difficulty of annotating large volumes leading often to sparse and incomplete labels.
- The physical properties of ultrasound images resulting in low contrast borders and high labelling variability

To overcome the above challenges, we made several propositions divided into three groups, aiming respectively to build datasets suitable for learning, designing automatic 3D ultrasound muscle segmentation, deep learning models and analysing the network's segmentation performance.

Part I: Building 3D ultrasound annotated datasets focused on the creation of a high-resolution 3D ultrasound multi-label dataset of the lower limb with reliable 3D annotations. To this end, we proposed two different studies focusing on:

1. Addressing the challenge of creating 3D labels from sparse 2D annotations with probe tracking information.
2. Reducing the need of probe tracking towards improving the portability of ultrasound volumes with high quality.

The contributions made in these works are as follows:

- To the best of our knowledge, the LEG-3D-Ultrasound dataset is the largest available multi-label ultrasound database. We are preparing its open-source release in the months to come.
- We studied different image-based and interpolation-based methods to create 3D labels from sparse 2D annotations in ultrasound volumes acquired at different frequencies. We proposed the "ZOI" semi-automatic method to create 3D annotations in 10% of the time needed compared to full annotations done by an expert.
- We explored different learning-based methods for sensorless freehand ultrasound. To this end, we adapted to the specificities of our data, several existing methods that predict the DOF position of the probe from image sequences. Although experimentally inconclusive with our data, we also proposed changes in terms of the architecture and loss towards forcing a rigid displacement field between slices or in terms of the rational representation.

In conclusion of this part, we provide an open-source dataset created and a simple and cost-effective method, primarily relying on open-source software, for producing a high-quality, fully annotated ultrasound 3D dataset. This proposed approach significantly reduces the need for manual intervention while offering reasonably accurate data for segmentation and volume computation. We are aware that a better volume compounding could improve segmentation, especially in artifacts such as the vertical lines at image borders. These lines caused by probe motion or stitching errors, disrupt uniformity and degrade segmentation performance. Future works could focus on techniques like interpolation, overlapping, or edge blending to reduce such artifacts and improve volume accuracy and segmentation reliability. Regarding the fully sensorless freehand tracking approach, we believe its limited performance can be explained by several reasons linked to our data, as the tiny motion between parts of images in ultrafast ultrasound acquisitions such as ours, which are hard to predict. Also, the nature of the muscle structure changes very smoothly along the acquisition direction, making the out-of-plane estimation highly ill-posed.

Finally, and also related to the previous remark, the speckle correlation across muscle with similar tissue texture seems to be challenging for both speckle decorrelation and learning-based approaches. Nevertheless, the created dataset will be used by the team in future research on sensorless ultrasound algorithms for multiple sweeps. Perspectives for this part are adding low-cost sensors and focusing on reducing tracking errors, similar to [221,222]. A second direction could be the joint modelling of the volume compounding and the segmentation tasks [223,224].

Part II: Deep neural networks for 3D ultrasound segmentation focused on the design of two deep learning models for muscle segmentation, towards accurately delimiting their contours muscles in 3D volumes in low and high-resolution datasets. We designed the two models with the following objectives in mind:

1. Dealing with limited and sparse annotated data.
2. Addressing the challenge of deep neural networks with very large volumes.
3. Improving the prediction accuracy until being close to the inter-expert volumetric segmentation error of 5%.

The main contributions from this part are:

— We proposed an automatic segmentation method called "UNet-S-R-CLSTM" that employs an auxiliary reconstruction task alongside a multi-objective training strategy with a loss that dynamically adapts to the available annotations in each part of the sequence, making also use of negative labels from neighbouring muscles. This approach is particularly effective when dealing with limited annotated data.

— We proposed the "IFSSnet" architecture for high-resolution muscle segmentation that benefits from spatio-temporal modelling of the segmentation task, where we treat volumes as sequential data to leverage the computational burden. Apart from performing fully automatic segmentation, our sequential modelling can also be used in interactive mode as a label propagation strategy, alleviating the need for extensive annotated datasets. We showed that providing just 3% of the annotations as initial masks is sufficient to obtain high-quality segmentation in a few-shot set-up. Moreover, we propose to rely on a learnable parametric Tversky loss to balance precision and recall, and thus adapt to the high imbalance in our dataset.

In conclusion, we proposed two different deep-learning architectures and losses for muscle volumetric segmentation with dedicated modules and training strategies to address the challenges of sparse and incomplete annotations and the computational burden associated with large volumes. Furthermore, both architectures represent a good compromise between performance and cost (computation, data annotation, etc.) compared to fully automatic segmentation methods. However, segmentation performance depends on evaluation metrics and ground truth label variability. Therefore, it seemed important for us to investigate the robustness of the networks trained to segment ultrasound volumes, as we did in part II. Other perspectives for this part could be incorporating priors from statistical muscle shape models [225, 226]. A final research direction for this part includes establishing a parallel between our sequential models, capable of learning long-dependencies across the volumes, to those present in vision transformers [227], towards further reducing the number of experiment inputs in the iterative case.

Part III: Analysis of ultrasound segmentation architectures focused on the experimental validation and understanding of existing segmentation methods. We designed the studies with the following objectives in mind:

1. Evaluate how to take into account the way clinical experts annotate ultrasound images and improve the network's robustness to expert annotation variability.
2. Provide a new evaluation procedure specific to segmented borders in ultrasound images.

The main contributions from this part are:

— We investigated the challenges of ultrasound image segmentation due to its inherent characteristics, such as noise and low contrast. We studied the advantages of using pre-computed Confidence Maps as additional information for the network. We evaluate the impact when CMs are added as a second channel or in the loss function. Our experiments probe an improvement in the awareness of the networks to expert label variability.

— We finally evaluate different architectures (3) on different ultrasound datasets (5), concluding that ultrasound segmentation metrics could be categorised into metrics for evident and completed borders. This division is necessitated by the anisotropic nature of ultrasound imaging, which results in images with no blurred borders, different from MRI and CT images. We propose a mathematical way to extract the evident borders and separate them from the completed ones. Additionally, we used Seg-Grad-CAM to study the attention points for completed borders, concluding that similar to physicians, the networks use evident borders and reference key points to interpolate and fill the gaps.

In conclusion, part III provides valuable insights into the decision-making processes of deep neural networks for ultrasound segmentation after a deeper understanding of the complexities inherent in ultrasound imaging. We make use of Confidence Maps and Seg-Grad-CAM to enhance and understand network performance. We propose a specific evaluation of ultrasound borders, dividing the border metrics in evident and completed. Future perspectives could focus on automating the CM extraction and exploring alternative loss functions to maximise the effectiveness of CMs in various training scenarios. Additionally, research directions for future works could be oriented towards improving architectures to explicitly take into account the border variability, for instance, with spatially variant convolution kernels [228] or designing ultrasound-specific attribution explanation heat-maps [229].

Finally, the most significant and impactful work that can be done lies in translating some of our work into the context, for instance, neurological (Duchenne) or geriatrical (Sarcopenia) applications.

Articles as first Author

Gonzalez Duque, V., *Al Chanti, D., Crouzier, M., Nordez, A., Lacourpaille, L., Mateus, D. (2020). Spatio-temporal consistency and negative label transfer for 3D freehand US segmentation. In MICCAI 2020: 23rd International Conference, Lima, Peru, October 4–8, 2020, Proceedings, Part I 23 (pp. 710-720). Springer International Publishing.*

Duque, V. G., *Alchanti, D., Crouzier, M., Nordez, A., Lacourpaille, L., Mateus, D. (2020, November). Low-limb muscle segmentation in 3D freehand ultrasound using non-learning methods and label transfer. In 16th International Symposium on Medical Information Processing and Analysis (Vol. 11583, pp. 154-163). SPIE.*

Al Chanti, D., ***Duque, V. G.***, *Crouzier, M., Nordez, A., Lacourpaille, L., Mateus, D. (2021). IFSS-Net: for faster muscle segmentation and propagation in volumetric ultra-*

sound. IEEE transactions on medical imaging, 40(10), 2615-2628.

***Duque, V. G.**, Zirus, L., Velikova, Y., Navab, N., Mateus, D. (2023). Can ultrasound Confidence Maps predict expert labels' variability? In ASMUS workshop at MICCAI 2023: 26th International Conference, Vancouver, Canada, October 8–14, 2023, Proceedings pp. 100-120). Springer International Publishing.*

***Duque, V. G.**, Marquardt, A., Velikova, Y., Lacourpaille, L., Nordez, A., Crouzier, M., Lee H.J., Mateus, D., Navab N., (2023). Ultrasound Segmentation Analysis via Distinct and Completed Anatomical Borders. 15th International Conference on Information Processing in Computer-Assisted Interventions(IPCAI). International Journal of Computer Assisted Radiology and Surgery (IJCARS).*

Articles as Co-author

*L. Piecuch, **V.G. Duque**, A. Sarcher, A. Nordez, G. Rabita, G. Guilhem, and D. Mateus. Muscle volume quantification: guiding transformers with anatomical priors. Workshop on Shape in Medical Imaging (ShapeMI) at MICCAI 2023: 26th International Conference, Vancouver, Canada, October 8–14, 2023. Springer International Publishing.*

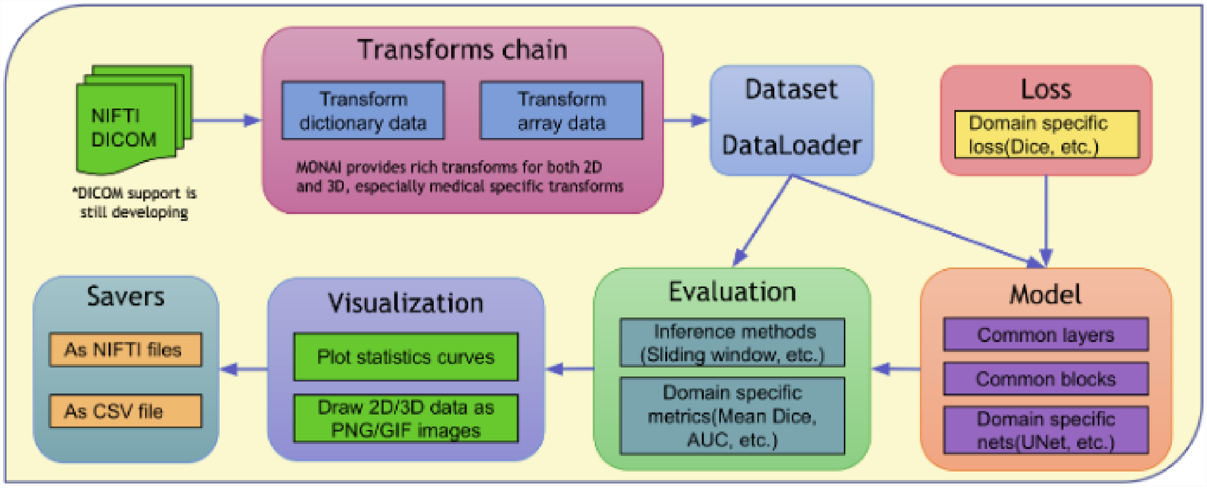
*Velikova, Y., Azampour, M. F., Simson, W., **Gonzalez Duque, V.**, & Navab, N. (2023, October). LOTUS: Learning to Optimize Task-based US Representations. In International Conference on Medical Image Computing and Computer-Assisted Intervention (pp. 435-445). Cham: Springer Nature Switzerland.*

Appendix

Appendix A

Technical contributions

MONAI is an open-source framework that specializes in deep learning for healthcare and medical imaging. Before MONAI, there existed different frameworks among which we found Tethano (Before 2012), Caffe (2013), CNTK from microsoft (2014), Tensorflow (2015), Keras(2015), Mxnet(2015), Chainer(2015), Caffe2(2016), Gluon(2016), Pytorch(2017).



NIFTI
DICOM
*DICOM support is still developing
Transforms chain
Transform dictionary data
Transform array data
MONAI provides rich transforms for both 2D and 3D, especially medical specific transforms
Dataset
DataLoader
Loss
Domain specific loss(Dice, etc.)
Savers
As NIFTI files
As CSV file
Visualization
Plot statistics curves
Draw 2D/3D data as PNG/GIF images
Evaluation
Inference methods (Sliding window, etc.)
Domain specific metrics(Mean Dice, AUC, etc.)
Model
Common layers
Common blocks
Domain specific nets(UNet, etc.)

Figure A.1 – End-to-end workflow in medical deep learning area. Taken from: https://docs.monai.io/en/0.5.0/highlights.html

The workflow of medical image analysis using Artificial intelligence (AI) can be summarized with the Figure A.6. Medical image analysis using AI begins with data acquisition, where various medical images, such as X-rays, MRIs, or CT scans, are collected. Once acquired, these images undergo pre-processing, which includes steps like image enhancement to improve clarity and normalization to standardize image values. Subsequently, data augmentation techniques might be applied to artificially expand the dataset, ensuring the AI model is robust and generalizes well. These processed images are then fed into machine learning or deep learning models for training, where the model learns to recognize patterns, anomalies, or specific features. It use specific losses to find the best parameters of the architecture. After training, the model is validated and tested on unseen data to evaluate its performance. Once satisfactory accuracy is achieved, the AI model can be deployed in a clinical setting to assist healthcare professionals in diagnosing diseases, identifying anomalies, or planning treatments based on the insights derived from the analyzed images.

MONAI appears in 2020, as a joint effort from researchers from various institutions to create a collection of tools and best practices to facilitate the development and validation of deep learning models in this domain. MONAI Key features and aspects include:

1. Modularity and Flexibility: MONAI is designed to be adaptable, enabling researchers and developers to use individual components with other libraries or frameworks if desired.
2. Transforms: MONAI provides a comprehensive set of transformations for both image preprocessing (like normalization, cropping, and resampling) and augmentation (like rotation, scaling, and elastic deformation).
3. Network Architectures: The framework incorporates various state-of-the-art neural network architectures optimized for medical imaging tasks.
4. Evaluation Metrics: MONAI includes metrics commonly used in medical imaging challenges to assess the performance of models.
5. Interoperability: MONAI is designed to be compatible with the PyTorch ecosystem, allowing seamless integration with other PyTorch libraries and tools.
6. Research Reproducibility: MONAI places an emphasis on reproducibility, providing consistent implementations and environments for research.
7. Community-Driven: The development of MONAI is driven by the community, with contributions from researchers, clinicians, and developers in the medical imaging domain.

A.1 Ultrasound confidence maps as a MONAI transform

Course: Project Management and Software Development for Medical Applications-2023. *Student:* Bugra.

Supervisor: **Vanessa Gonzalez**

Ultrasound imaging plays a crucial role in medical diagnostics by providing immediate visualization of internal body structures. However, its effectiveness can be hindered by artifacts such as shadow effects and reverberations. To combat these issues, the approach of attributing confidence levels to different areas of the ultrasound image has been introduced, aiding clinicians in distinguishing between trustworthy and less reliable regions. This concept was first put forward by Karamalis et al. in their pioneering study [193], where they developed a technique for creating these confidence maps. Addressing Key feature 2, our research aimed to re-implement in python the Matlab code. Initially, we adapted the method using the Science Python Library (SciPy) [230], resulting in a functional yet slower performance. By employing stochastic techniques, we were able to markedly enhance the algorithm's speed. A key accomplishment of our study was the successful incorporation of this improved method into the MONAI framework [231], a leading medical imaging platform based on PyTorch. This integration significantly extends its accessibility to a wider range of users. Processing time is presented in Figure A.2-a). Figure A.2-b) provide an evaluation of the error of both implementations for one image example, several evaluations were done in order to verify whether the implementation was correct.

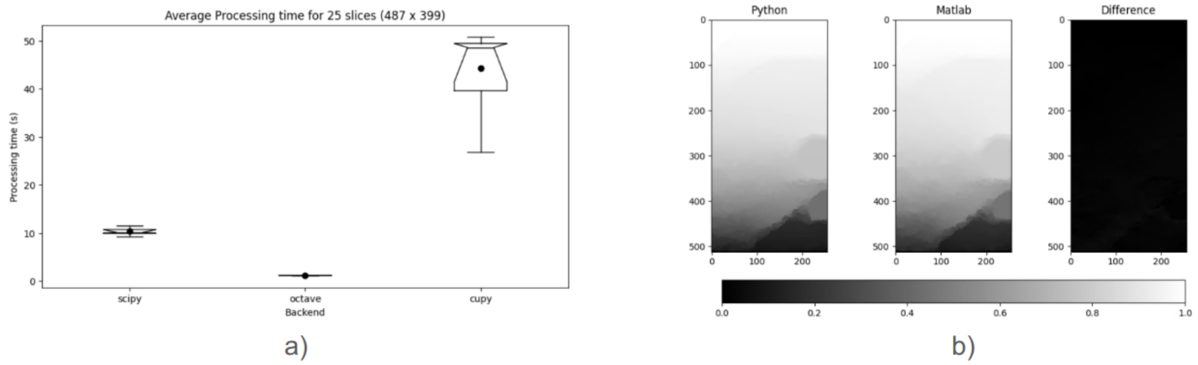


Figure A.2 – Confidence maps Monai implementation: a) Running time b) Re-implementation discrepancy.

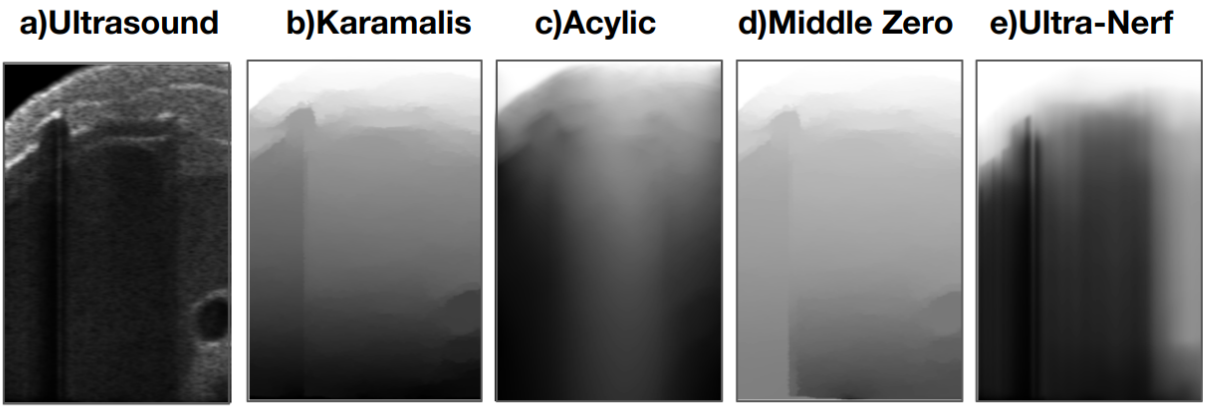


Figure A.3 – Qualitative visualisation of confidence maps versions: a) Kalamaris et al. [193], b) Cylic c) Bottom Zero d) Ultra-Nerf [232]

Different Confidence Maps representations can be found in the Figure A.3. Image b) obtained with Hung et al. [233] method is calculated raw by raw in a recursive manner. Thinking to provide a more reliable representation, we would like first to force the

Confidence Maps to have similar values in overlapping images, and second obtain values different to zero in inferior areas in the images when there is sound reaching the bottom. In this order, we proceed to calculate to proposals. First, we set just the middle center of the image as Zero for the Kalamaris' method, and second, we proceed to calculate CM using the Ultra-Nerf method of Wysocki et al. [232] The work is still in progress. Qualitative visualizations are presented in Figure A.3

A.2 Re-implementation of Quicknat and Daf3D networks

Course: Clinical Application Project-2023. *Student*: Alexandra Manquart & Carlotta Holze. *Supervisor:* **Vanessa Gonzalez**

Addressing the Key features 3 and 6, we proceed to re-implement in MONAI the networks Quicknat and Daf3D.

Quicknat is a 2.5D network that match the features of 3 different 2D views QuickNAT is a deep learning architecture tailored for the rapid and accurate segmentation of neuroanatomical structures in brain MRIs. Utilizing a modified U-Net structure, a type of Fully Convolutional Neural Network (FCNN), QuickNAT captures both local and global features of the brain. Designed with efficiency in mind, it boasts fewer parameters than many deep learning models, ensuring faster inference times. Additionally, it employs a multi-stage training strategy, initially leveraging a patch-based approach followed by fine-tuning on entire MRI volumes, and can be extended to provide Bayesian segmentation, offering uncertainty measures in its predictions

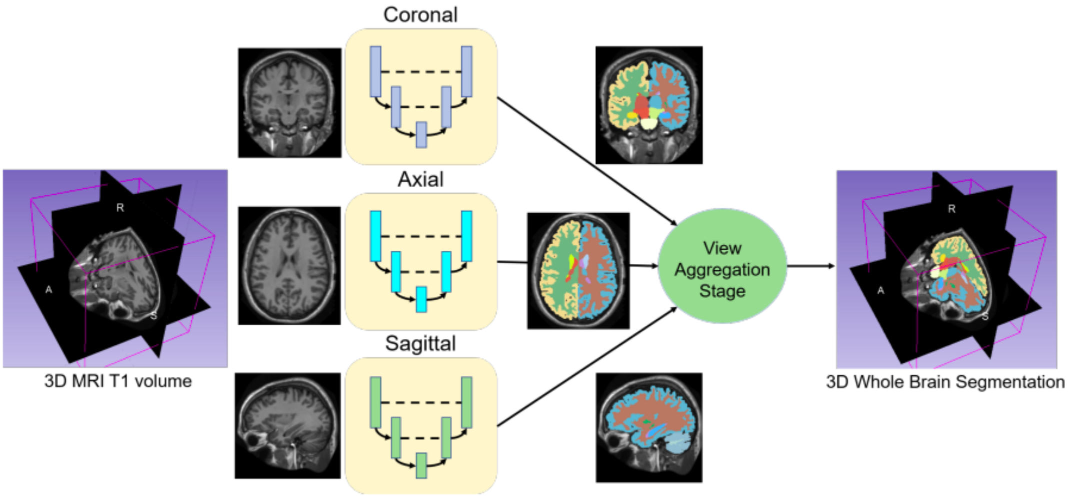


Figure A.4 – Quicknat architecture: THe multi-view aggregation step that combines segmentations from models trained on 2D slices along three principal axes: coronal, sagittal, and axial.

DAF3D is a 3D ultrasound binary segmentation architecture for the prostate. Deep Attentive Features for Prostate Segmentation in 3D Transrectal Ultrasound by Yi Wang proposes a deep learning architecture tailored for the precise segmentation of the prostate in 3D transrectal ultrasound images. Leveraging attention mechanisms, the model selectively focuses on critical regions within the ultrasound data, enhancing the differentiation between the prostate and surrounding tissues. This approach aims to address challenges such as speckle noise, shadowing, and varying prostate appearances by harnessing the model's capacity to prioritize relevant features, thereby improving the accuracy and robustness of prostate segmentation in clinical ultrasound scans.

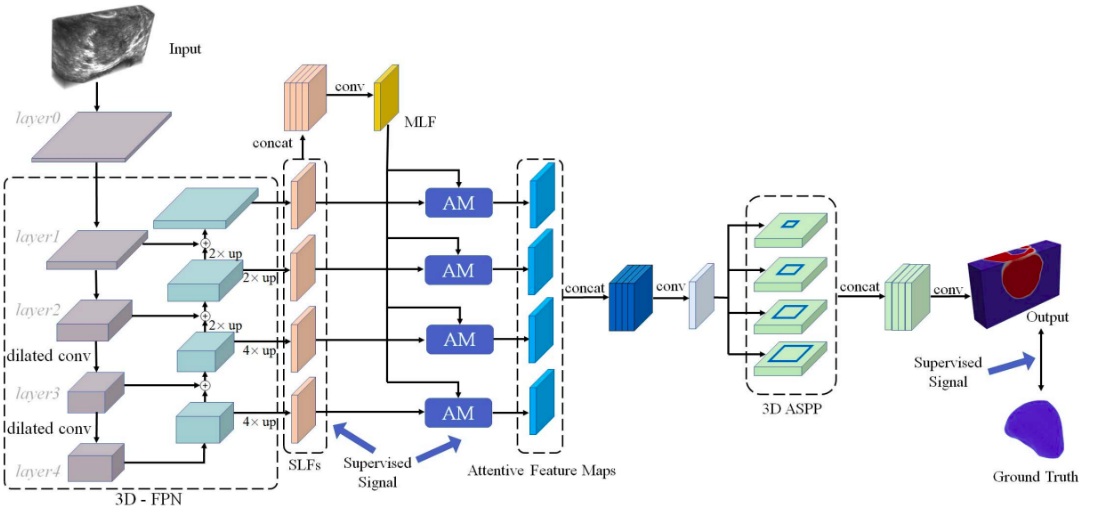
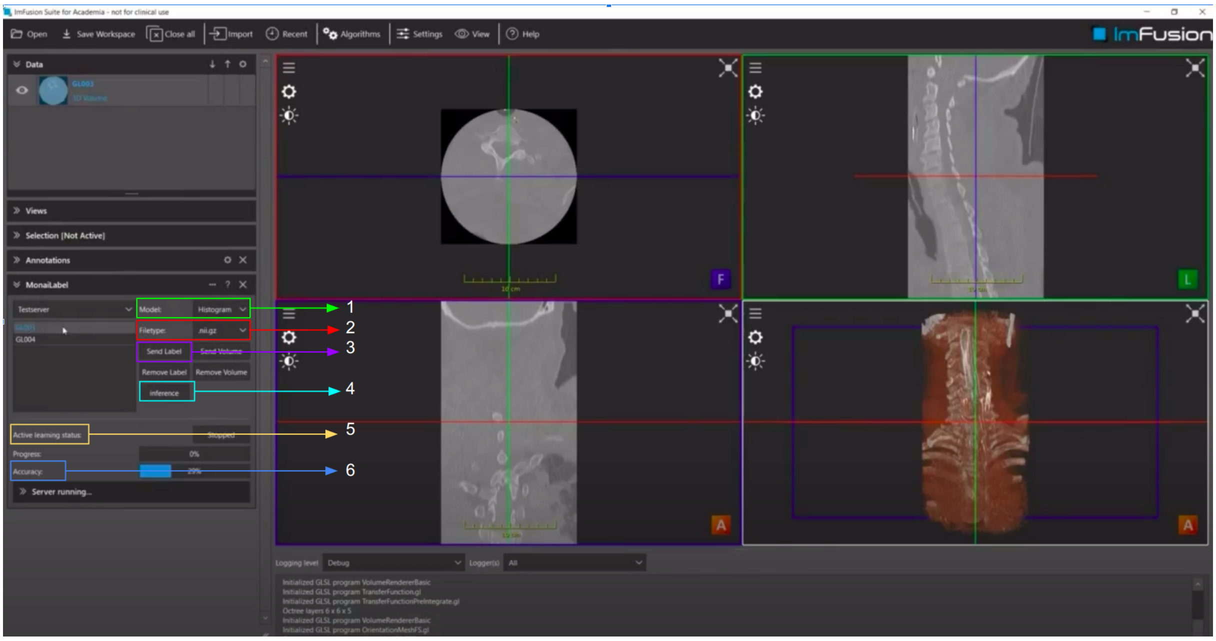


Figure A.5 – The schematic illustration of our prostate segmentation network equipped with attention modules. FPN: feature pyramid network; SLF: single-layer features; MLF: multi-layer features; AM: attention module; ASPP: atrous spatial pyramid pooling. Taken from [34]

A.3 Monai-label plugin in Imfusion software

Course: Clinical Application Project-2022. *Student:* Maximilian Bauregger. *Supervisor:* **Vanessa Gonzalez**

For addressing the Key feature 7, we create a plugin program in Imfusion software that allow the use of MONAI-Label for training on deep-learning algorithms. MONAI-Label is a component of the MONAI ecosystem designed to streamline the annotation of medical images. It offers interactive tools for annotating images and employs AI models to assist in this process by providing initial annotation suggestions, which human annotators can refine. This fusion of human expertise and AI accelerates the creation of high-quality annotated datasets. Furthermore, its extensibility allows for easy integration of new models or algorithms, and its position within the MONAI framework ensures seamless access to a vast array of medical image analysis tools. The platform also promotes collaboration among researchers and clinicians, fostering a community-driven approach.



Model: Histogram

Filetype: .nii.gz

Active learning status: Stopped

Progress: 0%

Accuracy: 29%

Logging level: Debug

Logger(s): All

```
Initialized GLSL program VolumeRendererBasic
Initialized GLSL program TransferFunction.gl
Initialized GLSL program TransferFunctionPreIntegrate.gl
Octree layers 6 x 6 x 5
Initialized GLSL program VolumeRendererBasic
Initialized GLSL program OrientationMeshFS.gl
```

Figure A.6 – Plugin features: 1)AI Model selection, 2)Search files with file-type selected, 3)Send a Label or a Volume in the dataset folder we are segmenting, 4)Create an inference label using the pre-trained model, 5),Retran the model with the new annotated data, 6)Calculate the accuracy of the model in one validation participant.

A.4 Conclusion

Our contributions to the MONAI framework mark significant advancements in the field of medical imaging and education. Firstly, the development of a transformation technique for confidence maps in ultrasound imaging greatly accelerates the process of generating new image representations. This enhancement is crucial for rapid diagnosis and effective patient care. Secondly, the implementation of the Quicknat and DAF3D networks, along with an instructional tutorial, is a major step forward in educational resources. It empowers students and professionals alike to delve into advanced neural network models with greater ease and understanding. Lastly, the integration of the MONAI Label plugin software with the Imfusion suite is a groundbreaking addition. It allows for the continuous retraining of networks with new data, ensuring that the models evolve and improve over time. This continuous learning aspect is vital in keeping pace with the ever-changing landscape of medical imaging, making our contributions not only innovative but also indispensable for future advancements in the field.

Appendix B

Summary of articles contributions

B.1 Articles as first author

***Gonzalez Duque, V.**, Al Chanti, D., Crouzier, M., Nordez, A., Lacourpaille, L., Mateus, D. (2020). Spatio-temporal consistency and negative label transfer for 3D freehand US segmentation. In MICCAI 2020: 23rd International Conference, Lima, Peru, October 4–8, 2020, Proceedings, Part I 23 (pp. 710-720). Springer International Publishing.*

Summary: We propose a one-encoder and two-decoder architecture with a CLSTM in the bottleneck that performs binary segmentation and image reconstruction using the multitask learning principle, that leads to a better geometrical estimation of the mask shape. Our 4 label dataset was compose of sparse 2d slices annotations, reason why we opted for a weak-label learning approach. In summary, when annotation was available, we perform a normal dice loss. But for such slices where the label of interest was not available, we penalize the predictions on the positions where the true negative of the other labels were annotated.

*Al Chanti, D., **Duque, V. G.**, Crouzier, M., Nordez, A., Lacourpaille, L., Mateus, D. (2021). IFSS-Net: for faster muscle segmentation and propagation in volumetric ultrasound. IEEE transactions on medical imaging, 40(10), 2615-2628.*

Summary: We present a two encoders - one decoder architecture with a Bidirectional Convolutional long-short term memory (BICLSTM) in the bottleneck that performs binary muscle segmentation of 3D ultrasound volumes. The main contribution is the small volumetric error of the predictions equivalent to an intra-operative error of 4%. The volume is passed sequentially in batches of 3 slices and the predictions are sent as input to the second decoder for the next sequence, since the muscles are smooth, the network learns to interpolate the previous annotations. We introduce a decremental update of the objective function to guide the model convergence in the absence of large amounts of annotated data. And to handle the class-imbalance between foreground and background muscle pixels, we propose a parametric Tversky loss function that learns to adaptively penalize false positives and false negatives.

***Duque, V. G.**, Zirus, L., Velikova, Y., Navab, N., Mateus, D. (2023). Can ultrasound Confidence Maps predict expert labels' variability?*

Summary: We propose, for the first time, the integration of [193] Confidence Map (CM) in the neuronal networks, to provide important prior information about where to find the uncertain image regions in order to boost segmentation. We propose two uses of CM: First as second channel and second in the confidence loss (Masking the CM with the label ground truth). Our main contributions are networks with predictions of high uncertainty in areas where the inherent physical principles governing the acquisition can be a source of uncertainty, particularly at farther distances from the transducer. Both tasks imply a minimal computational overhead and no changing in the architectures.

***Duque, V. G.**, Alchanti, D., Crouzier, M., Nordez, A., Lacourpaille, L., Mateus, D. (2020, November). Low-limb muscles segmentation in 3D freehand ultrasound using non-learning methods and label transfer. In 16th International Symposium on Medical Information Processing and Analysis (Vol. 11583, pp. 154-163). SPIE.*

Summary: This was the first approach to obtain 3D muscle labels from partial annotations done in B-mode images. We compared different algorithms without artificial intelligence, all available in Slicer 3D: Fill between slices, grow from seeds and watershed. Our proposal, the ZOI method, was a zero-order interpolation with smoothness. 15 of the 44 volumes were manually corrected by an expert for the use of the dataset in other projects, as they reached only a mean dice score of 0.89±0.03 and a mean volumetric measure error of 4.18%.

B.2 Articles as co-author

*L. Piecuch, **V.G. Duque**, A. Sarcher, A. Nordez, G. Rabita, G. Guilhem, and D. Mateus. Muscle volume quantification: guiding transformers with anatomical priors. ShapeMi 2023.*

The relationship with the thesis lies in the objective of automating muscle segmentation for more efficient and accurate morphometric analysis. Both endeavors recognize the limitations of manual segmentation and the potential of neural networks in transforming this process. This article specifically introduces a hybrid architecture that combines convolutional and visual transformer blocks, designed to capture the intricate details and long-range relations of muscles on CT 3D images. This is particularly relevant to the thesis as the consistent anatomical configuration of leg muscles in athletes would benefit from such an approach. Furthermore, the utilization of an adjacency matrix for muscle neighborhood estimation resonates with the thesis's goal of precise segmentation in 3D US volumes. Both converge on the need for automated, advanced neural network solutions for muscle segmentation in the realm of sports and medical imaging.

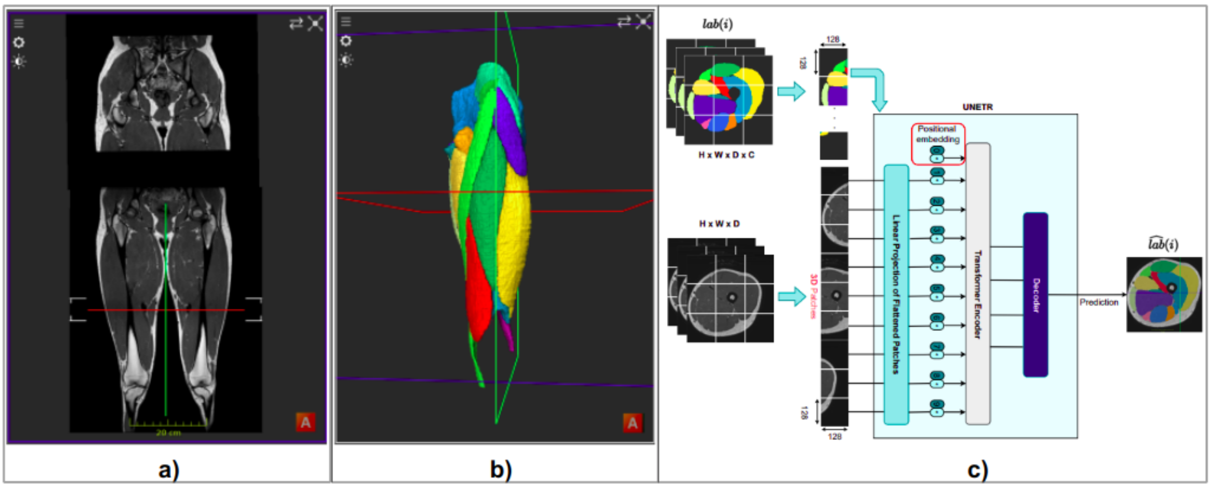


Figure B.1 – Overview of UNetr + adjacency loss method: a)Input MRI, b) Labelmap c) Architecture

Abstract: Muscle volume is a useful quantitative biomarker in sports, but also for the follow-up of degenerative musculo-skeletal diseases. In addition to volume, other shape biomarkers can be extracted by segmenting the muscles of interest from medical images. Manual segmentation is still today the gold standard for such measurements despite being very time-consuming. We propose a method for automatic segmentation of 18 muscles of the lower limb on 3D Magnetic Resonance Images to assist such morphometric analysis. By their nature, the tissue of different muscles is undistinguishable when observed in MR Images. Thus, muscle segmentation algorithms cannot rely on appearance but only on contour cues. However, such contours are hard to detect and their thickness varies across subjects. To cope with the above challenges, we propose a segmentation approach based on a hybrid architecture, combining convolutional and visual transformer blocks. We investigate for the first time the behaviour of such hybrid architectures in the context of muscle segmentation for shape analysis. Considering the consistent anatomical muscle configuration, we rely on transformer blocks to capture the long-range relations between the muscles. To further exploit the anatomical priors, a second contribution of this work consists in adding a regularisation loss based on an adjacency matrix of plausible muscle neighbourhoods estimated from the training data. Our experimental results on a unique database of elite athletes show it is possible to train complex hybrid models from a relatively small database of large volumes, while the anatomical prior regularisation favours

better predictions.

*Velikova, Y., Azampour, M. F., Simson, W., **Gonzalez Duque, V.**, & Navab, N. (2023, October). LOTUS: Learning to Optimize Task-based US representations. In International Conference on Medical Image Computing and Computer-Assisted Intervention (pp. 435-445). Cham: Springer Nature Switzerland.*

The connection with the thesis lies in the overarching theme of automating and refining the segmentation process in ultrasound imaging. While our main goal is ultrasound segmentation, Lotus uses network segmentation as an evaluation task. Lotus tries to solve the problem of small annotated datasets for networks that need a lot of data. It focuses on the generation of simulated ultrasound training data from annotated CT scans. Its emphasis lies on using a fully differentiable ultrasound simulator to optimize parameters for generating ultrasound images, combined with an end-to-end training setting for simultaneous image synthesis and segmentation. Both aim to harness the power of deep learning and neural networks to alleviate the manual and expertise-dependent nature of ultrasound image segmentation, albeit targeting different anatomical regions and applications.

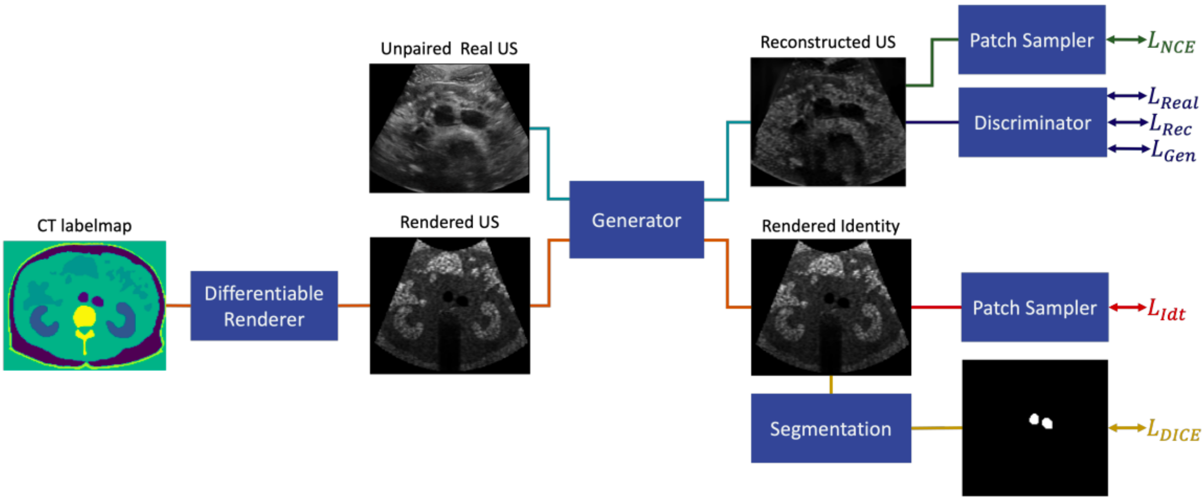


Figure B.2 – Overview of the proposed framework. During training, we render online US simulation images from CT label maps and use them as input to a segmentation network. Our ultrasound renderer is fully differentiable and learns to optimize the parameters based on the downstream segmentation task. At the same time, we train an unpaired and unsupervised image style transfer network between real and rendered images to achieve simultaneous image synthesis as well as automatic segmentation on US images in an end-to-end training setting

Abstract: Anatomical segmentation of organs in ultrasound images is essential to many clinical applications, particularly for diagnosis and monitoring. Existing deep neural networks require a large amount of labelled data for training in order to achieve clinically acceptable performance. Yet, in ultrasound, due to characteristic properties such as speckle and clutter, it is challenging to obtain accurate segmentation boundaries, and precise pixel-wise labelling of images is highly dependent on the expertise of physicians. In contrast, CT scans have higher resolution and improved contrast, easing organ identification. In this paper, we propose a novel approach for learning to optimize task-based ultrasound image representations. Given annotated CT segmentation maps as a simulation medium, we model acoustic propagation through tissue via ray-casting to generate ultrasound training data. Our ultrasound simulator is fully differentiable and learns to optimize the parameters for generating physics-based ultrasound images guided by the downstream segmentation task. In addition, we train an image adaptation network between real and simulated images to achieve simultaneous image synthesis and automatic segmentation on US images in an end-to-end training setting. The proposed method is evaluated on aorta and vessel segmentation tasks and shows promising quantitative results. Furthermore, we also conduct qualitative results of optimized image representations on other organs.

Bibliography

[1] Moore, C. L. and Copel, J. A. (2011) Point-of-care ultrasonography. *New England Journal of Medicine,* **364**(8), 749–757.

[2] Wang, Z. (2020) Deep learning in medical ultrasound image segmentation: A review. *arXiv preprint arXiv:2002.07703,*.

[3] Akkus, Z., Cai, J., Boonrod, A., Zeinoddini, A., Weston, A. D., Philbrick, K. A., and Erickson, B. J. (2019) A survey of deep-learning applications in ultrasound: Artificial intelligence–powered ultrasound for improving clinical workflow. *Journal of the American College of Radiology,* **16**(9), 1318–1328.

[4] Pichiecchio, A., Alessandrino, F., Bortolotto, C., Cerica, A., Rosti, C., Raciti, M. V., Rossi, M., Berardinelli, A., Baranello, G., Bastianello, S., et al. (2018) Muscle ultrasound elastography and MRI in preschool children with Duchenne muscular dystrophy. *Neuromuscular Disorders,* **28**(6), 476–483.

[5] Krönke, M., Eilers, C., Dimova, D., Köhler, M., Buschner, G., Schweiger, L., Konstantinidou, L., Makowski, M., Nagarajah, J., Navab, N., et al. (2022) Tracked 3D ultrasound and deep neural network-based thyroid segmentation reduce interobserver variability in thyroid volumetry. *Plos one,* **17**(7), e0268550.

[6] Nguyen, D. T., Choi, J., and Park, K. R. (2022) Thyroid Nodule Segmentation in Ultrasound Image Based on Information Fusion of Suggestion and Enhancement Networks. *Mathematics,* **10**(19), 3484.

[7] Dunnhofer, M., Antico, M., Sasazawa, F., Takeda, Y., Camps, S., Martinel, N., Micheloni, C., Carneiro, G., and Fontanarosa, D. (2020) Siam-U-Net: encoder-decoder siamese network for knee cartilage tracking in ultrasound images. *Medical Image Analysis,* **60**, 101631.

[8] Gomariz, A., Li, W., Ozkan, E., Tanner, C., and Goksel, O. (2019) Siamese networks with location prior for landmark tracking in liver ultrasound sequences. In *2019 IEEE 16th International Symposium on Biomedical Imaging (ISBI 2019)* IEEE pp. 1757–1760.

[9] WHO Leprosy. (2023).

[10] Oh, Y. S., Early, D. S., and Azar, R. R. (2005) Clinical applications of endoscopic ultrasound to oncology. *Oncology,* **68**(4-6), 526–537.

[11] Huang, Q., Zeng, Z., et al. (2017) A review on real-time 3D ultrasound imaging technology. *BioMed research international,* **2017**.

[12] Rohling, R., Gee, A., and Berman, L. (1997) Three-dimensional spatial compounding of ultrasound images. *Medical Image Analysis,* **1**(3), 177–193.

[13] Prager, R. W., Gee, A., and Berman, L. (1998) Stradx: real-time acquisition and visualisation of freehand 3D ultrasound.

[14] Prevost, R., Salehi, M., Sprung, J., Ladikos, A., Bauer, R., and Wein, W. (2017) Deep learning for sensorless 3D freehand ultrasound imaging. In *International conference on medical image computing and computer-assisted intervention* Springer pp. 628–636.

[15] Luo, M., Yang, X., Yan, Z., Li, J., Zhang, Y., Chen, J., Hu, X., Qian, J., Cheng, J., and Ni, D. (2023) Multi-IMU with Online Self-consistency for Freehand 3D Ultrasound Reconstruction. In *International Conference on Medical Image Computing and Computer-Assisted Intervention* Springer pp. 342–351.

[16] Chen, H., Ni, D., Qin, J., Li, S., Yang, X., Wang, T., and Heng, P. A. (2015) Standard plane localization in fetal ultrasound via domain transferred deep neural networks. *IEEE journal of biomedical and health informatics,* **19**(5), 1627–1636.

[17] Han, S., Kang, H.-K., Jeong, J.-Y., Park, M.-H., Kim, W., Bang, W.-C., and Seong, Y.-K. (2017) A deep learning framework for supporting the classification of breast lesions in ultrasound images. *Physics in Medicine & Biology,* **62**(19), 7714.

[18] Schmauch, B., Herent, P., Jehanno, P., Dehaene, O., Saillard, C., Aubé, C., Luciani, A., Lassau, N., and Jégou, S. (2019) Diagnosis of focal liver lesions from ultrasound using deep learning. *Diagnostic and interventional imaging,* **100**(4), 227–233.

[19] Cunningham, R. J., Harding, P. J., and Loram, I. D. (2016) Real-time ultrasound segmentation, analysis and visualisation of deep cervical muscle structure. *IEEE transactions on medical imaging,* **36**(2), 653–665.

[20] Cerrolaza, J. J., Sinclair, M., Li, Y., Gomez, A., Ferrante, E., Matthew, J., Gupta, C., Knight, C. L., and Rueckert, D. (2018) Deep learning with ultrasound physics for fetal skull segmentation. In *2018 IEEE 15th International Symposium on Biomedical Imaging (ISBI 2018)* IEEE pp. 564–567.

[21] Prabhu, S. J., Kanal, K., Bhargava, P., Vaidya, S., and Dighe, M. K. (2014) Ultrasound artifacts: classification, applied physics with illustrations, and imaging appearances. *Ultrasound quarterly,* **30**(2), 145–157.

[22] Gee, A., Prager, R., Treece, G., Cash, C., and Berman, L. (2004) Processing and visualizing three-dimensional ultrasound data. *The British journal of radiology,* **77**(suppl_2), S186–S193.

[23] Pieper, S., Halle, M., and Kikinis, R. (2004) 3D Slicer. In *2004 2nd IEEE international symposium on biomedical imaging: nano to macro (IEEE Cat No. 04EX821)* IEEE pp. 632–635.

[24] Liu, S., Wang, Y., Yang, X., Lei, B., Liu, L., Li, S. X., Ni, D., and Wang, T. (2019) Deep learning in medical ultrasound analysis: a review. *Engineering,* **5**(2), 261–275.

[25] Chen, H., Zheng, Y., Park, J.-H., Heng, P.-A., and Zhou, S. K. (2016) Iterative multidomain regularized deep learning for anatomical structure detection and segmentation from ultrasound images. In *Medical Image Computing and Computer-Assisted Intervention–MICCAI 2016: 19th International Conference, Athens, Greece, October 17-21, 2016, Proceedings, Part II 19* Springer pp. 487–495.

[26] Ravishankar, H., Prabhu, S. M., Vaidya, V., and Singhal, N. (2016) Hybrid approach for automatic segmentation of fetal abdomen from ultrasound images using deep learning. In *2016 IEEE 13th international symposium on biomedical imaging (ISBI)* IEEE pp. 779–782.

[27] Qi, H., Collins, S., and Noble, A. (2017) Weakly supervised learning of placental ultrasound images with residual networks. In *Medical Image Understanding and Analysis: 21st Annual Conference, MIUA 2017, Edinburgh, UK, July 11–13, 2017, Proceedings 21* Springer pp. 98–108.

[28] Gao, Y. and Alison Noble, J. (2017) Detection and characterization of the fetal heartbeat in free-hand ultrasound sweeps with weakly-supervised two-streams convolutional networks. In *Medical Image Computing and Computer-Assisted Intervention- MICCAI 2017: 20th International Conference, Quebec City, QC, Canada, September 11-13, 2017, Proceedings, Part II 20* Springer pp. 305–313.

[29] Huang, W., Bridge, C. P., Noble, J. A., and Zisserman, A. (2017) Temporal Heart-Net: towards human-level automatic analysis of fetal cardiac screening video. In *International Conference on Medical Image Computing and Computer-Assisted Intervention* Springer pp. 341–349.

[30] Ghesu, F. C., Georgescu, B., Zheng, Y., Hornegger, J., and Comaniciu, D. (2015) Marginal space deep learning: Efficient architecture for detection in volumetric image data. In *Medical Image Computing and Computer-Assisted Intervention–MICCAI 2015: 18th International Conference, Munich, Germany, October 5-9, 2015, Proceedings, Part I 18* Springer pp. 710–718.

[31] Pereira, F., Bueno, A., Rodriguez, A., Perrin, D., Marx, G., Cardinale, M., Salgo, I., and Del Nido, P. (2017) Automated detection of coarctation of aorta in neonates from two-dimensional echocardiograms. *Journal of Medical Imaging,* **4**(1), 014502–014502.

[32] Hiramatsu, Y., Muramatsu, C., Kobayashi, H., Hara, T., and Fujita, H. (2017) Automated detection of masses on whole breast volume ultrasound scanner: false positive reduction using deep convolutional neural network. In *Medical imaging 2017: Computer-aided diagnosis* Spie Vol. 10134, pp. 717–722.

[33] Bian, C., Lee, R., Chou, Y.-H., and Cheng, J.-Z. (2017) Boundary regularized convolutional neural network for layer parsing of breast anatomy in automated whole breast ultrasound. In *Medical Image Computing and Computer Assisted*

Intervention- MICCAI 2017: 20th International Conference, Quebec City, QC, Canada, September 11-13, 2017, Proceedings, Part III 20 Springer pp. 259–266.

[34] Wang, P., Cuccolo, N. G., Tyagi, R., Hacihaliloglu, I., and Patel, V. M. (2018) Automatic real-time CNN-based neonatal brain ventricles segmentation. In *2018 IEEE 15th International Symposium on Biomedical Imaging (ISBI 2018)* IEEE pp. 716–719.

[35] Yang, H., Shan, C., Kolen, A. F., and de With, P. H. (2019) Improving catheter segmentation & localization in 3d cardiac ultrasound using direction-fused fcn. In *2019 IEEE 16th International Symposium on Biomedical Imaging (ISBI 2019)* IEEE pp. 1122–1126.

[36] Li, X., Pang, S., Zhang, R., Zhu, J., Fu, X., Tian, Y., and Gao, J. (2023) ATTransUNet: An enhanced hybrid transformer architecture for ultrasound and histopathology image segmentation. *Computers in Biology and Medicine,* **152**, 106365.

[37] Wu, L., Xin, Y., Li, S., Wang, T., Heng, P.-A., and Ni, D. (2017) Cascaded fully convolutional networks for automatic prenatal ultrasound image segmentation. In *2017 IEEE 14th international symposium on biomedical imaging (ISBI 2017)* IEEE pp. 663–666.

[38] Cunningham, R., Sánchez, M. B., and Loram, I. D. (2019) Ultrasound segmentation of cervical muscle during head motion: A dataset and a benchmark using deconvolutional neural networks.

[39] Valanarasu, J. M. J., Oza, P., Hacihaliloglu, I., and Patel, V. M. (2021) Medical transformer: Gated axial-attention for medical image segmentation. In *Medical Image Computing and Computer Assisted Intervention–MICCAI 2021: 24th International Conference, Strasbourg, France, September 27–October 1, 2021, Proceedings, Part I 24* Springer pp. 36–46.

[40] Mishra, D., Chaudhury, S., Sarkar, M., and Soin, A. S. (2018) Ultrasound image segmentation: a deeply supervised network with attention to boundaries. *IEEE Transactions on Biomedical Engineering,* **66**(6), 1637–1648.

[41] Kim, B. S., Yu, M., Kim, S., Yoon, J. S., and Baek, S. (2022) Scale-attentional U-Net for the segmentation of the median nerve in ultrasound images. *Ultrasonography,* **41**(4), 706–717.

[42] Valanarasu, J. M. J., Yasarla, R., Wang, P., Hacihaliloglu, I., and Patel, V. M. (2020) Learning to segment brain anatomy from 2D ultrasound with less data. *IEEE Journal of Selected Topics in Signal Processing,* **14**(6), 1221–1234.

[43] Kumar, V., Webb, J. M., Gregory, A., Denis, M., Meixner, D. D., Bayat, M., Whaley, D. H., Fatemi, M., and Alizad, A. (2018) Automated and real-time segmentation of suspicious breast masses using convolutional neural network. *PloS one,* **13**(5), e0195816.

[44] Wang, Y., Dou, H., Hu, X., Zhu, L., Yang, X., Xu, M., Qin, J., Heng, P.-A., Wang, T., and Ni, D. (2019) Deep attentive features for prostate segmentation in 3D transrectal ultrasound. *IEEE transactions on medical imaging,* **38**(12), 2768–2778.

[45] Oktay, O., Schlemper, J., Folgoc, L. L., Lee, M., Heinrich, M., Misawa, K., Mori, K., McDonagh, S., Hammerla, N. Y., Kainz, B., et al. (2018) Attention u-net: Learning where to look for the pancreas. *IMIDL Conference (2018),*.

[46] Al Chanti, D., Duque, V. G., Crouzier, M., Nordez, A., Lacourpaille, L., and Mateus, D. (2021) IFSS-Net: Interactive few-shot siamese network for faster muscle segmentation and propagation in volumetric ultrasound. *IEEE transactions on medical imaging,* **40**(10), 2615–2628.

[47] Wang, P., Zhang, C., Qi, F., Liu, S., Zhang, X., Lyu, P., Han, J., Liu, J., Ding, E., and Shi, G. (2021) Pgnet: Real-time arbitrarily-shaped text spotting with point gathering network. In *Proceedings of the AAAI Conference on Artificial Intelligence* Vol. 35, pp. 2782–2790.

[48] Ronneberger, O., Fischer, P., and Brox, T. (2015) U-net: Convolutional networks for biomedical image segmentation. In *Medical Image Computing and Computer-Assisted Intervention–MICCAI 2015: 18th International Conference, Munich, Germany, October 5-9, 2015, Proceedings, Part III 18* Springer pp. 234–241.

[49] Hatamizadeh, A., Tang, Y., Nath, V., Yang, D., Myronenko, A., Landman, B., Roth, H. R., and Xu, D. (2022) Unetr: Transformers for 3d medical image segmentation. In *Proceedings of the IEEE/CVF winter conference on applications of computer vision* pp. 574–584.

[50] Vanessa Gonzalez Duque, Alexandra Marquardt, Y. V. L. L. A. N. M. C. H. J. L. D. M. N. N. (2024) Ultrasound Segmentation Analysis via Distinct and Completed Anatomical Borders.. *IJCARS, IPCAI,*.

[51] Crouzier, M., Lacourpaille, L., Nordez, A., Tucker, K., and Hug, F. (2018) Neuromechanical coupling within the human triceps surae and its consequence on individual force-sharing strategies. *Journal of Experimental Biology,* **221**(21).

[52] Xu, Z. and Niethammer, M. (2019) DeepAtlas: Joint semi-supervised learning of image registration and segmentation. In *Medical Image Computing and Computer Assisted Intervention–MICCAI 2019: 22nd International Conference, Shenzhen, China, October 13–17, 2019, Proceedings, Part II 22* Springer pp. 420–429.

[53] Patil, P. and Dasgupta, B. (2012) Role of diagnostic ultrasound in the assessment of musculoskeletal diseases. *Therapeutic advances in musculoskeletal disease,* **4**(5), 341–355.

[54] Jahanshir, A., Moghari, S. M., Ahmadi, A., Moghadam, P. Z., and Bahreini, M. (2020) Value of point-of-care ultrasonography compared with computed tomogra-

phy scan in detecting potential life-threatening conditions in blunt chest trauma patients. *The Ultrasound Journal,* **12**, 1–10.

[55] Yassa, M., Mutlu, M. A., Birol, P., Kuzan, T. Y., Kalafat, E., Usta, C., Yavuz, E., Keskin, I., and Tug, N. (2020) Lung ultrasonography in pregnant women during the COVID-19 pandemic: an interobserver agreement study among obstetricians. *Ultrasonography,* **39**(4), 340.

[56] Balocco, S., Gatta, C., Ciompi, F., Wahle, A., Radeva, P., Carlier, S., Unal, G., Sanidas, E., Mauri, J., Carillo, X., et al. (2014) Standardized evaluation methodology and reference database for evaluating IVUS image segmentation. *Computerized medical imaging and graphics,* **38**(2), 70–90.

[57] Donald, I., Macvicar, J., and Brown, T. G. (1958) Investigation of abdominal masses by pulsed ultrasound. *The Lancet,* **271**(7032), 1188–1195.

[58] Matrone, G., Savoia, A. S., Caliano, G., and Magenes, G. (2014) The delay multiply and sum beamforming algorithm in ultrasound B-mode medical imaging. *IEEE transactions on medical imaging,* **34**(4), 940–949.

[59] Khan, S., Huh, J., and Ye, J. C. (2021) Variational formulation of unsupervised deep learning for ultrasound image artifact removal. *IEEE Transactions on Ultrasonics, Ferroelectrics, and Frequency Control,* **68**(6), 2086–2100.

[60] Stasi, G. and Ruoti, E. M. (2015) A critical evaluation in the delivery of the ultrasound practice: the point of view of the radiologist. *Ital J Med,* **9**(1), 5.

[61] Kim, Y. H. (2021) Artificial intelligence in medical ultrasonography: driving on an unpaved road. *Ultrasonography,* **40**(3), 313.

[62] Toma, T. P. and Volpicelli, G. (2020) Essential image acquisition protocols for thoracic ultrasonography. *Respiration,* **99**(3), 231–238.

[63] Jimenez-Castaño, C. A., Álvarez-Meza, A. M., Aguirre-Ospina, O. D., Cárdenas-Peña, D. A., and Orozco-Gutiérrez, Á. A. (2021) Random fourier features-based deep learning improvement with class activation interpretability for nerve structure segmentation. *Sensors,* **21**(22), 7741.

[64] Ungi, T., Greer, H., Sunderland, K. R., Wu, V., Baum, Z. M., Schlenger, C., Oetgen, M., Cleary, K., Aylward, S. R., and Fichtinger, G. (2020) Automatic spine ultrasound segmentation for scoliosis visualization and measurement. *IEEE Transactions on Biomedical Engineering,* **67**(11), 3234–3241.

[65] Fenster, A., Downey, D. B., and Cardinal, H. N. (2001) Three-dimensional ultrasound imaging. *Physics in medicine & biology,* **46**(5), R67.

[66] Welch, J. N., Johnson, J. A., Bax, M. R., Badr, R., and Shahidi, R. (2000) A real-time freehand 3D ultrasound system for image-guided surgery. In *2000*

IEEE Ultrasonics Symposium. Proceedings. An International Symposium (Cat. No. 00CH37121) IEEE Vol. 2, pp. 1601–1604.

[67] Dai, Y., Tian, J., Xue, J., and Liu, J. (2006) A qualitative and quantitative interaction technique for freehand 3D ultrasound imaging. In *2006 International Conference of the IEEE Engineering in Medicine and Biology Society* IEEE pp. 2750–2753.

[68] Gao, H., Huang, Q., Xu, X., and Li, X. (2016) Wireless and sensorless 3D ultrasound imaging. *Neurocomputing,* **195**, 159–171.

[69] Duque, V. G., Alchanti, D., Crouzier, M., Nordez, A., Lacourpaille, L., and Mateus, D. (2020) Low-limb muscles segmentation in 3D freehand ultrasound using non-learning methods and label transfer. In *16th International Symposium on Medical Information Processing and Analysis* SPIE Vol. 11583, pp. 154–163.

[70] Terzakis, G., Lourakis, M., and Ait-Boudaoud, D. (2018) Modified Rodrigues parameters: an efficient representation of orientation in 3D vision and graphics. *Journal of Mathematical Imaging and Vision,* **60**(3), 422–442.

[71] Trobaugh, J. W., Trobaugh, D. J., and Richard, W. D. (1994) Three-dimensional imaging with stereotactic ultrasonography. *Computerized Medical Imaging and Graphics,* **18**(5), 315–323.

[72] Coupé, P., Hellier, P., Azzabou, N., and Barillot, C. (2005) 3D freehand ultrasound reconstruction based on probe trajectory. In *Medical Image Computing and Computer-Assisted Intervention–MICCAI 2005: 8th International Conference, Palm Springs, CA, USA, October 26-29, 2005, Proceedings, Part I 8* Springer pp. 597–604.

[73] Nelson, T. R. and Pretorius, D. H. (1997) Interactive acquisition, analysis, and visualization of sonographic volume data. *International Journal of Imaging Systems and Technology,* **8**(1), 26–37.

[74] Gobbi, D. G. and Peters, T. M. (2002) Interactive intra-operative 3D ultrasound reconstruction and visualization. In *International conference on medical image computing and computer-assisted intervention* Springer pp. 156–163.

[75] Huang, Q., Zheng, Y., Lu, M., and Chi, Z. (2005) Development of a portable 3D ultrasound imaging system for musculoskeletal tissues. *Ultrasonics,* **43**(3), 153–163.

[76] Huang, Q.-H. and Zheng, Y.-P. (2006) An adaptive squared-distance-weighted interpolation for volume reconstruction in 3D freehand ultrasound. *Ultrasonics,* **44**, e73–e77.

[77] Huang, Q., Zheng, Y., Lu, M., Wang, T., and Chen, S. (2009) A new adaptive interpolation algorithm for 3D ultrasound imaging with speckle reduction and edge preservation. *Computerized Medical Imaging and Graphics,* **33**(2), 100–110.

[78] Ohbuchi, R., Chen, D., and Fuchs, H. (1992) Incremental volume reconstruction and rendering for 3-D ultrasound imaging. In *Visualization in Biomedical Computing'92* SPIE Vol. 1808, pp. 312–323.

[79] Estépar, R. S. J., Martín-Fernández, M., Alberola-López, C., Ellsmere, J., Kikinis, R., and Westin, C.-F. (2003) Freehand ultrasound reconstruction based on roi prior modeling and normalized convolution. In *Medical Image Computing and Computer-Assisted Intervention-MICCAI 2003: 6th International Conference, Montréal, Canada, November 15-18, 2003. Proceedings 6* Springer pp. 382–390.

[80] Rohling, R., Gee, A., and Berman, L. (1999) A comparison of freehand three-dimensional ultrasound reconstruction techniques. *Medical image analysis,* **3**(4), 339–359.

[81] Sanches, J. M. and Marques, J. S. (2000) A Rayleigh reconstruction/interpolation algorithm for 3D ultrasound. *Pattern recognition letters,* **21**(10), 917–926.

[82] Oakden-Rayner, L. The rebirth of CAD: how is modern AI different from the CAD we know?. (2019).

[83] Sendak, M., Vidal, D., Trujillo, S., Singh, K., Liu, X., and Balu, S. (2023) Surfacing best practices for AI software development and integration in healthcare. *Frontiers in Digital Health,* **5**, 1150875.

[84] Khunte, M., Chae, A., Wang, R., Jain, R., Sun, Y., Sollee, J., Jiao, Z., and Bai, H. (2023) Trends in clinical validation and usage of US Food and Drug Administration-cleared artificial intelligence algorithms for medical imaging. *Clinical Radiology,* **78**(2), 123–129.

[85] Matsoukas, S., Chennareddy, S., Kalagara, R., Scaggiante, J., Smith, C. J., Bazil, M. J., Reford, E., Liu, K., Delman, B. N., Selim, M. H., et al. (2022) Pilot deployment of viz–intracranial hemorrhage for intracranial hemorrhage detection: real-world performance in a stroke code cohort. *Stroke,* **53**(9), e418–e419.

[86] Brattain, L. J., Ozturk, A., Telfer, B. A., Dhyani, M., Grajo, J. R., and Samir, A. E. (2020) Image processing pipeline for liver fibrosis classification using ultrasound shear wave elastography. *Ultrasound in medicine & biology,* **46**(10), 2667–2676.

[87] Gatos, I., Tsantis, S., Spiliopoulos, S., Karnabatidis, D., Theotokas, I., Zoumpoulis, P., Loupas, T., Hazle, J. D., and Kagadis, G. C. (2017) A machine-learning algorithm toward color analysis for chronic liver disease classification, employing ultrasound shear wave elastography. *Ultrasound in medicine & biology,* **43**(9), 1797–1810.

[88] Liang, X., Cao, Q., Huang, R., and Lin, L. (2014) Recognizing focal liver lesions in contrast-enhanced ultrasound with discriminatively trained spatio-temporal model. In *2014 IEEE 11th International Symposium on Biomedical Imaging (ISBI)* IEEE pp. 1184–1187.

[89] Hetherington, J., Lessoway, V., Gunka, V., Abolmaesumi, P., and Rohling, R. (2017) SLIDE: automatic spine level identification system using a deep convolutional neural network. *International journal of computer assisted radiology and surgery,* **12**, 1189–1198.

[90] Antico, M., Sasazawa, F., Dunnhofer, M., Camps, S., Jaiprakash, A., Pandey, A., Crawford, R., Carneiro, G., and Fontanarosa, D. (2020) Deep learning-based femoral cartilage automatic segmentation in ultrasound imaging for guidance in robotic knee arthroscopy. *Ultrasound in medicine & biology,* **46**(2), 422–435.

[91] Tardy, M., Scheffer, B., and Mateus, D. (2019) Uncertainty measurements for the reliable classification of mammograms. In *International Conference on Medical Image Computing and Computer-Assisted Intervention* Springer pp. 495–503.

[92] Lekadir, K., Galimzianova, A., Betriu, A., del Mar Vila, M., Igual, L., Rubin, D. L., Fernández, E., Radeva, P., and Napel, S. (2016) A convolutional neural network for automatic characterization of plaque composition in carotid ultrasound. *IEEE journal of biomedical and health informatics,* **21**(1), 48–55.

[93] Cheng, P. M. and Malhi, H. S. (2017) Transfer learning with convolutional neural networks for classification of abdominal ultrasound images. *Journal of digital imaging,* **30**, 234–243.

[94] Wu, L., Cheng, J.-Z., Li, S., Lei, B., Wang, T., and Ni, D. (2017) FUIQA: fetal ultrasound image quality assessment with deep convolutional networks. *IEEE transactions on cybernetics,* **47**(5), 1336–1349.

[95] Franco-Barranco, D., Muñoz-Barrutia, A., and Arganda-Carreras, I. (2022) Stable deep neural network architectures for mitochondria segmentation on electron microscopy volumes. *Neuroinformatics,* **20**(2), 437–450.

[96] Kingma, D. P. and Ba, J. (2014) Adam: A method for stochastic optimization. In *International Conference on Learning Representations.*

[97] Amari, S.-i. (1993) Backpropagation and stochastic gradient descent method. *Neurocomputing,* **5**(4-5), 185–196.

[98] Khirirat, S., Feyzmahdavian, H. R., and Johansson, M. (2017) Mini-batch gradient descent: Faster convergence under data sparsity. In *2017 IEEE 56th Annual Conference on Decision and Control (CDC)* IEEE pp. 2880–2887.

[99] Zhang, Y., Ying, M. T., Yang, L., Ahuja, A. T., and Chen, D. Z. (2016) Coarse-to-fine stacked fully convolutional nets for lymph node segmentation in ultrasound images. In *2016 IEEE International Conference on Bioinformatics and Biomedicine (BIBM)* IEEE pp. 443–448.

[100] Badrinarayanan, V., Kendall, A., and Cipolla, R. (2017) Segnet: A deep convolutional encoder-decoder architecture for image segmentation. *IEEE transactions on pattern analysis and machine intelligence,* **39**(12), 2481–2495.

[101] Isola, P., Zhu, J.-Y., Zhou, T., and Efros, A. A. (2017) Image-to-image translation with conditional adversarial networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition* pp. 1125–1134.

[102] Zhou, R., Fenster, A., Xia, Y., Spence, J. D., and Ding, M. (2019) Deep learning-based carotid media-adventitia and lumen-intima boundary segmentation from three-dimensional ultrasound images. *Medical physics,* **46**(7), 3180–3193.

[103] Pourtaherian, A., Ghazvinian Zanjani, F., Zinger, S., Mihajlovic, N., Ng, G., Korsten, H., and de With, P. (2017) Improving needle detection in 3D ultrasound using orthogonal-plane convolutional networks. In *Medical Image Computing and Computer-Assisted Intervention- MICCAI 2017: 20th International Conference, Quebec City, QC, Canada, September 11-13, 2017, Proceedings, Part II 20* Springer pp. 610–618.

[104] Roy, A. G., Conjeti, S., Navab, N., Wachinger, C., Initiative, A. D. N., et al. (2019) QuickNAT: A fully convolutional network for quick and accurate segmentation of neuroanatomy. *NeuroImage,* **186**, 713–727.

[105] Yang, X., Yu, L., Wu, L., Wang, Y., Ni, D., Qin, J., and Heng, P.-A. (2017) Fine-grained recurrent neural networks for automatic prostate segmentation in ultrasound images. In *Proceedings of the AAAI Conference on Artificial Intelligence* Vol. 31, .

[106] Simonyan, K. and Zisserman, A. (2014) Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556,*.

[107] Deng, J., Dong, W., Socher, R., Li, L.-J., Li, K., and Fei-Fei, L. (2009) Imagenet: A large-scale hierarchical image database. In *2009 IEEE conference on computer vision and pattern recognition* Ieee pp. 248–255.

[108] Yang, X., Yu, L., Li, S., Wang, X., Wang, N., Qin, J., Ni, D., and Heng, P.-A. (2017) Towards automatic semantic segmentation in volumetric ultrasound. In *Medical Image Computing and Computer Assisted Intervention- MICCAI 2017: 20th International Conference, Quebec City, QC, Canada, September 11-13, 2017, Proceedings, Part I 20* Springer pp. 711–719.

[109] Tu, Z. (2008) Auto-context and its application to high-level vision tasks. In *2008 IEEE Conference on Computer Vision and Pattern Recognition* IEEE pp. 1–8.

[110] Dai, W., Dong, N., Wang, Z., Liang, X., Zhang, H., and Xing, E. P. (2018) Scan: Structure correcting adversarial network for organ segmentation in chest x-rays. In *Deep Learning in Medical Image Analysis and Multimodal Learning for Clinical Decision Support: 4th International Workshop, DLMIA 2018, and 8th International Workshop, ML-CDS 2018, Held in Conjunction with MICCAI 2018, Granada, Spain, September 20, 2018, Proceedings 4* Springer pp. 263–273.

[111] Lei, Y., Tian, S., He, X., Wang, T., Wang, B., Patel, P., Jani, A. B., Mao, H., Curran, W. J., Liu, T., et al. (2019) Ultrasound prostate segmentation based on multidirectional deeply supervised V-Net. *Medical physics,* **46**(7), 3194–3206.

[112] Milletari, F., Navab, N., and Ahmadi, S.-A. (2016) V-net: Fully convolutional neural networks for volumetric medical image segmentation. In *2016 fourth international conference on 3D vision (3DV)* Ieee pp. 565–571.

[113] Zhou, Z., Rahman Siddiquee, M. M., Tajbakhsh, N., and Liang, J. (2018) Unet++: A nested u-net architecture for medical image segmentation. In *Deep Learning in Medical Image Analysis and Multimodal Learning for Clinical Decision Support: 4th International Workshop, DLMIA 2018, and 8th International Workshop, ML-CDS 2018, Held in Conjunction with MICCAI 2018, Granada, Spain, September 20, 2018, Proceedings 4* Springer pp. 3–11.

[114] Ho, J., Kalchbrenner, N., Weissenborn, D., and Salimans, T. (2019) Axial attention in multidimensional transformers. *arXiv preprint arXiv:1912.12180,*.

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

[116] Gao, Y., Zhou, M., and Metaxas, D. N. (2021) UTNet: a hybrid transformer architecture for medical image segmentation. In *Medical Image Computing and Computer Assisted Intervention–MICCAI 2021: 24th International Conference, Strasbourg, France, September 27–October 1, 2021, Proceedings, Part III 24* Springer pp. 61–71.

[117] Cao, H., Wang, Y., Chen, J., Jiang, D., Zhang, X., Tian, Q., and Wang, M. (2023) Swin-unet: Unet-like pure transformer for medical image segmentation. In *Computer Vision–ECCV 2022 Workshops: Tel Aviv, Israel, October 23–27, 2022, Proceedings, Part III* Springer pp. 205–218.

[118] Oh, S. W., Lee, J.-Y., Sunkavalli, K., and Kim, S. J. (2018) Fast video object segmentation by reference-guided mask propagation. In *Proceedings of the IEEE conference on computer vision and pattern recognition* pp. 7376–7385.

[119] Reinke, A., Tizabi, M. D., Sudre, C. H., Eisenmann, M., Rädsch, T., Baumgartner, M., Acion, L., Antonelli, M., Arbel, T., Bakas, S., et al. (2021) Common limitations of image processing metrics: A picture story. *arXiv preprint arXiv:2104.05642,*.

[120] Visa, S., Ramsay, B., Ralescu, A. L., and Van Der Knaap, E. (2011) Confusion matrix-based feature selection.. *Maics,* **710**(1), 120–127.

[121] Lee, E. C., Fragala, M. S., Kavouras, S. A., Queen, R. M., Pryor, J. L., and Casa, D. J. (2017) Biomarkers in sports and exercise: tracking health, performance, and recovery in athletes. *The Journal of Strength & Conditioning Research,* **31**(10), 2920–2937.

[122] Lacourpaille, L., Gross, R., Hug, F., Guével, A., Péréon, Y., Magot, A., Hogrel, J.-Y., and Nordez, A. (2017) Effects of Duchenne muscular dystrophy on muscle stiffness and response to electrically-induced muscle contraction: a 12-month follow-up. *Neuromuscular Disorders,* **27**(3), 214–220.

[123] Zettinig, O., Salehi, M., Prevost, R., and Wein, W. (2018) Recent Advances in Point-of-Care Ultrasound Using the ImFusion Suite ImFusion Suite for Real-Time Image Analysis. In *Simulation, Image Processing, and Ultrasound Systems for Assisted Diagnosis and Navigation: International Workshops, POCUS 2018, BIVPCS 2018, CuRIOUS 2018, and CPM 2018, Held in Conjunction with MICCAI 2018, Granada, Spain, September 16–20, 2018, Proceedings* Springer pp. 47–55.

[124] Barber, L., Barrett, R., and Lichtwark, G. (2009) Validation of a freehand 3D ultrasound system for morphological measures of the medial gastrocnemius muscle. *Journal of biomechanics,* **42**(9), 1313–1319.

[125] Albu, A. B., Beugeling, T., and Laurendeau, D. (2008) A morphology-based approach for interslice interpolation of anatomical slices from volumetric images. *IEEE Transactions on Biomedical Engineering,* **55**(8), 2022–2038.

[126] Treece, G. M., Prager, R. W., Gee, A. H., and Berman, L. (2000) Surface interpolation from sparse cross sections using region correspondence. *IEEE transactions on medical imaging,* **19**(11), 1106–1114.

[127] Beare, R. and Lehmann, G. (2006) The watershed transform in ITK-discussion and new developments.

[128] Boykov, Y. Y. and Jolly, M.-P. (2001) Interactive graph cuts for optimal boundary & region segmentation of objects in ND images. In *Proceedings eighth IEEE international conference on computer vision. ICCV 2001* IEEE Vol. 1, pp. 105–112.

[129] Ihnatsenka, B. and Boezaart, A. P. (2010) Ultrasound: Basic understanding and learning the language. *International journal of shoulder surgery,* **4**(3), 55.

[130] Yoshizumi, N., Saito, S., Koyama, D., Nakamura, K., Ohya, A., and Akiyama, I. (2009) Multiple-frequency ultrasonic imaging by transmitting pulsed waves of two frequencies. *Journal of Medical Ultrasonics,* **36**, 53–60.

[131] Che, C., Mathai, T. S., and Galeotti, J. (2017) Ultrasound registration: A review. *Methods,* **115**, 128–143.

[132] Sun, S.-Y., Gilbertson, M., and Anthony, B. W. (2014) Probe localization for freehand 3D ultrasound by tracking skin features. In *International Conference on Medical Image Computing and Computer-Assisted Intervention* Springer pp. 365–372.

[133] Busam, B., Ruhkamp, P., Virga, S., Lentes, B., Rackerseder, J., Navab, N., and Hennersperger, C. (2018) Markerless inside-out tracking for interventional applications. *arXiv preprint arXiv:1804.01708,*.

[134] Cai, Q., Peng, C., Lu, J.-Y., Prieto, J. C., Rosenbaum, A. J., Stringer, J. S., and Jiang, X. (2021) Performance enhanced ultrasound probe tracking with a hemispherical marker rigid body. *IEEE Transactions on Ultrasonics, Ferroelectrics, and Frequency Control,* **68**(6), 2155–2163.

[135] Gee, A. H., Housden, R. J., Hassenpflug, P., Treece, G. M., and Prager, R. W. (2006) Sensorless freehand 3D ultrasound in real tissue: speckle decorrelation without fully developed speckle. *Medical image analysis,* **10**(2), 137–149.

[136] Xie, Y., Liao, H., Zhang, D., Zhou, L., and Chen, F. (2021) Image-based 3D ultrasound reconstruction with optical flow via pyramid warping network. In *2021 43rd Annual International Conference of the IEEE Engineering in Medicine & Biology Society (EMBC)* IEEE pp. 3539–3542.

[137] Guo, H., Xu, S., Wood, B., and Yan, P. (2020) Sensorless freehand 3D ultrasound reconstruction via deep contextual learning. In *International Conference on Medical Image Computing and Computer-Assisted Intervention* Springer pp. 463–472.

[138] Guo, H., Chao, H., Xu, S., Wood, B. J., Wang, J., and Yan, P. (2022) Ultrasound Volume Reconstruction From Freehand Scans Without Tracking. *IEEE Transactions on Biomedical Engineering,* **70**(3), 970–979.

[139] Miura, K., Ito, K., Aoki, T., Ohmiya, J., and Kondo, S. (2020) Localizing 2D Ultrasound Probe from Ultrasound Image Sequences Using Deep Learning for Volume Reconstruction. In *Medical Ultrasound, and Preterm, Perinatal and Paediatric Image Analysis* pp. 97–105 Springer.

[140] Miura, K., Ito, K., Aoki, T., Ohmiya, J., and Kondo, S. (2021) Probe localization from ultrasound image sequences using deep learning for volume reconstruction. In *International Forum on Medical Imaging in Asia 2021* SPIE Vol. 11792, pp. 133–138.

[141] Miura, K., Ito, K., Aoki, T., Ohmiya, J., and Kondo, S. (2021) Pose Estimation of 2D Ultrasound Probe from Ultrasound Image Sequences Using CNN and RNN. In *International Workshop on Advances in Simplifying Medical Ultrasound* Springer pp. 96–105.

[142] Li, Q., Shen, Z., Li, Q., Barratt, D. C., Dowrick, T., Clarkson, M. J., Vercauteren, T., and Hu, Y. (2023) Trackerless freehand ultrasound with sequence modelling and auxiliary transformation over past and future frames. In *2023 IEEE 20th International Symposium on Biomedical Imaging (ISBI)* IEEE pp. 1–5.

[143] Ning, G., Liang, H., Zhou, L., Zhang, X., and Liao, H. (2022) Spatial Position Estimation Method for 3D Ultrasound Reconstruction Based on Hybrid Transfomers. In *2022 IEEE 19th International Symposium on Biomedical Imaging (ISBI)* IEEE pp. 1–5.

[144] Luo, M., Yang, X., Huang, X., Huang, Y., Zou, Y., Hu, X., Ravikumar, N., Frangi, A. F., and Ni, D. (2021) Self Context and Shape Prior for Sensorless Freehand 3D Ultrasound Reconstruction. *arXiv preprint arXiv:2108.00274,*.

[145] Prevost, R., Salehi, M., Jagoda, S., Kumar, N., Sprung, J., Ladikos, A., Bauer, R., Zettinig, O., and Wein, W. (2018) 3D freehand ultrasound without external tracking using deep learning. *Medical image analysis,* **48**, 187–202.

[146] Mikaeili, M. and Bilge, H. Ş. (2022) Trajectory estimation of ultrasound images based on convolutional neural network. *Biomedical Signal Processing and Control,* **78**, 103965.

[147] Luo, M., Yang, X., Wang, H., Du, L., and Ni, D. (2022) Deep Motion Network for Freehand 3D Ultrasound Reconstruction. In *International Conference on Medical Image Computing and Computer-Assisted Intervention* Springer pp. 290–299.

[148] Mur-Artal, R., Montiel, J. M. M., and Tardos, J. D. (2015) ORB-SLAM: a versatile and accurate monocular SLAM system. *IEEE transactions on robotics,* **31**(5), 1147–1163.

[149] Housden, R. J., Gee, A. H., Treece, G. M., and Prager, R. W. (2006) Subsample interpolation strategies for sensorless freehand 3D ultrasound. *Ultrasound in medicine & biology,* **32**(12), 1897–1904.

[150] Tan, C., Feng, X., Long, J., and Geng, L. (2018) FORECAST-CLSTM: A new convolutional LSTM network for cloudage nowcasting. In *2018 IEEE Visual Communications and Image Processing (VCIP)* IEEE pp. 1–4.

[151] Zhou, Y., Barnes, C., Lu, J., Yang, J., and Li, H. (2019) On the continuity of rotation representations in neural networks. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition* pp. 5745–5753.

[152] Perrot, V. and Garcia, D. (2018) Back to basics in ultrasound velocimetry: tracking speckles by using a standard PIV algorithm. In *2018 IEEE International Ultrasonics Symposium (IUS)* IEEE pp. 206–212.

[153] Crawford, R. J., Cornwall, J., Abbott, R., and Elliott, J. M. (2017) Manually defining regions of interest when quantifying paravertebral muscles fatty infiltration from axial magnetic resonance imaging: a proposed method for the lumbar spine with anatomical cross-reference. *BMC musculoskeletal disorders,* **18**, 1–11.

[154] Morrow, J. M., Sinclair, C. D., Fischmann, A., Machado, P. M., Reilly, M. M., Yousry, T. A., Thornton, J. S., and Hanna, M. G. (2016) MRI biomarker assessment of neuromuscular disease progression: a prospective observational cohort study. *The Lancet Neurology,* **15**(1), 65–77.

[155] Zhao, A., Balakrishnan, G., Durand, F., Guttag, J. V., and Dalca, A. V. (2019) Data augmentation using learned transformations for one-shot medical image segmenta-

tion. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition* pp. 8543–8553.

[156] Zheng, X., Wang, Y., Wang, G., and Liu, J. (2018) Fast and robust segmentation of white blood cell images by self-supervised learning. *Micron,* **107**, 55–71.

[157] Kotia, J., Kotwal, A., Bharti, R., and Mangrulkar, R. (2021) Few shot learning for medical imaging. *Machine learning algorithms for industrial applications,* pp. 107–132.

[158] Hervella, Á. S., Rouco, J., Novo, J., and Ortega, M. (2018) Retinal image understanding emerges from self-supervised multimodal reconstruction. In *Medical Image Computing and Computer Assisted Intervention–MICCAI 2018: 21st International Conference, Granada, Spain, September 16-20, 2018, Proceedings, Part I* Springer pp. 321–328.

[159] Chen, L., Bentley, P., Mori, K., Misawa, K., Fujiwara, M., and Rueckert, D. (2019) Self-supervised learning for medical image analysis using image context restoration. *Medical image analysis,* **58**, 101539.

[160] Zhang, X., Zhou, X., Lin, M., and Sun, J. (2018) Shufflenet: An extremely efficient convolutional neural network for mobile devices. In *Proceedings of the IEEE conference on computer vision and pattern recognition* pp. 6848–6856.

[161] Petit, O., Thome, N., Charnoz, A., Hostettler, A., and Soler, L. (2018) Handling missing annotations for semantic segmentation with deep convnets. In *Deep Learning in Medical Image Analysis and Multimodal Learning for Clinical Decision Support: 4th International Workshop, DLMIA 2018, and 8th International Workshop, ML-CDS 2018, Held in Conjunction with MICCAI 2018, Granada, Spain, September 20, 2018, Proceedings 4* Springer pp. 20–28.

[162] Xingjian, S., Chen, Z., Wang, H., Yeung, D.-Y., Wong, W.-K., and Woo, W.-c. (2015) Convolutional LSTM network: A machine learning approach for precipitation nowcasting. In *Advances in neural information processing systems* pp. 802–810.

[163] Azad, R., Asadi-Aghbolaghi, M., Fathy, M., and Escalera, S. (2019) Bi-Directional ConvLSTM U-Net with Densley Connected Convolutions. In *Proceedings of the IEEE International Conference on Computer Vision Workshops* pp. 0–0.

[164] Stollenga, M. F., Byeon, W., Liwicki, M., and Schmidhuber, J. (2015) Parallel multidimensional lstm, with application to fast biomedical volumetric image segmentation. In *Advances in neural information processing systems* pp. 2998–3006.

[165] Chen, J., Yang, L., Zhang, Y., Alber, M., and Chen, D. Z. (2016) Combining fully convolutional and recurrent neural networks for 3d biomedical image segmentation. In *Advances in neural information processing systems* pp. 3036–3044.

[166] Arbelle, A. and Raviv, T. R. (2019) Microscopy cell segmentation via convolutional LSTM networks. In *2019 IEEE 16th International Symposium on Biomedical Imaging (ISBI 2019)* IEEE pp. 1008–1012.

[167] Dai, J., He, K., and Sun, J. (2015) Boxsup: Exploiting bounding boxes to supervise convolutional networks for semantic segmentation. In *Proceedings of the IEEE International Conference on Computer Vision* pp. 1635–1643.

[168] Kolesnikov, A. and Lampert, C. H. (2016) Seed, expand and constrain: Three principles for weakly-supervised image segmentation. In *European Conference on Computer Vision* Springer pp. 695–711.

[169] Li, B., Yan, J., Wu, W., Zhu, Z., and Hu, X. (2018) High performance visual tracking with siamese region proposal network. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition* pp. 8971–8980.

[170] Lu, Z., Fu, Z., Xiang, T., Han, P., Wang, L., and Gao, X. (2016) Learning from weak and noisy labels for semantic segmentation. *IEEE transactions on pattern analysis and machine intelligence,* **39**(3), 486–500.

[171] Sener, O. and Koltun, V. (2018) Multi-task learning as multi-objective optimization. In *Advances in Neural Information Processing Systems* pp. 527–538.

[172] Hesamian, M. H., Jia, W., He, X., and Kennedy, P. (2019) Deep learning techniques for medical image segmentation: achievements and challenges. *Journal of digital imaging,* **32**, 582–596.

[173] Roth, H., Oda, M., Shimizu, N., Oda, H., Hayashi, Y., Kitasaka, T., Fujiwara, M., Misawa, K., and Mori, K. (2018) Towards dense volumetric pancreas segmentation in CT using 3D fully convolutional networks. In *Medical imaging 2018: image processing* SPIE Vol. 10574, pp. 59–64.

[174] Novikov, A. A., Major, D., Wimmer, M., Lenis, D., and Bühler, K. (2018) Deep sequential segmentation of organs in volumetric medical scans. *IEEE transactions on medical imaging,* **38**(5), 1207–1215.

[175] Çiçek, Ö., Abdulkadir, A., Lienkamp, S. S., Brox, T., and Ronneberger, O. (2016) 3D U-Net: learning dense volumetric segmentation from sparse annotation. In *Medical Image Computing and Computer-Assisted Intervention–MICCAI 2016: 19th International Conference, Athens, Greece, October 17-21, 2016, Proceedings, Part II 19* Springer pp. 424–432.

[176] Bertinetto, L., Valmadre, J., Henriques, J. F., Vedaldi, A., and Torr, P. H. (2016) Fully-convolutional siamese networks for object tracking. In *Computer Vision–ECCV 2016 Workshops: Amsterdam, The Netherlands, October 8-10 and 15-16, 2016, Proceedings, Part II 14* Springer pp. 850–865.

[177] Hu, Y.-T., Huang, J.-B., and Schwing, A. (2017) Maskrnn: Instance level video object segmentation. *Advances in neural information processing systems,* **30**.

[178] Perazzi, F., Khoreva, A., Benenson, R., Schiele, B., and Sorkine-Hornung, A. (2017) Learning video object segmentation from static images. In *Proceedings of the IEEE conference on computer vision and pattern recognition* pp. 2663–2672.

[179] Liu, Q., Zhou, F., Hang, R., and Yuan, X. (2017) Bidirectional-convolutional LSTM based spectral-spatial feature learning for hyperspectral image classification. *Remote Sensing,* **9**(12), 1330.

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

[181] Liu, K., Ning, X., and Liu, S. (2022) Medical Image Classification Based on Semi-Supervised Generative Adversarial Network and Pseudo-Labelling. *Sensors,* **22**(24), 9967.

[182] Liu, F., Tian, Y., Chen, Y., Liu, Y., Belagiannis, V., and Carneiro, G. (2022) Acpl: Anti-curriculum pseudo-labelling for semi-supervised medical image classification. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition* pp. 20697–20706.

[183] Salehi, S. S. M., Erdogmus, D., and Gholipour, A. (2017) Tversky loss function for image segmentation using 3D fully convolutional deep networks. In *International workshop on machine learning in medical imaging* Springer pp. 379–387.

[184] Gao, J., Xu, L., and Wan, M. (2023) Incremental learning for an evolving stream of medical ultrasound images via counterfactual thinking. *Computerized Medical Imaging and Graphics,* **109**, 102290.

[185] Wang, D., Cui, C., and Wu, Z. (2006) Matching 3D models with global geometric feature map. In *2006 12th International Multi-Media Modelling Conference* IEEE pp. 4–pp.

[186] Noble, J. and Boukerroui, D. (2006) Ultrasound image segmentation: a survey. *IEEE Transactions on Medical Imaging,* **25**(8), 987–1010.

[187] Iglesias, J. E. and Sabuncu, M. R. (2015) Multi-atlas segmentation of biomedical images: a survey. *Medical image analysis,* **24**(1), 205–219.

[188] Warfield, S. K., Zou, K. H., and Wells, W. M. (2004) Simultaneous truth and performance level estimation (STAPLE): an algorithm for the validation of image segmentation. *IEEE transactions on medical imaging,* **23**(7), 903–921.

[189] Baumgartner, C. F., Tezcan, K. C., Chaitanya, K., Hötker, A. M., Muehlematter, U. J., Schawkat, K., Becker, A. S., Donati, O., and Konukoglu, E. (2019) Phiseg: Capturing uncertainty in medical image segmentation. In *Medical Image Computing and Computer Assisted Intervention–MICCAI 2019: 22nd International Conference, Shenzhen, China, October 13–17, 2019, Proceedings, Part II 22* Springer pp. 119–127.

[190] Jungo, A., Meier, R., Ermis, E., Blatti-Moreno, M., Herrmann, E., Wiest, R., and Reyes, M. (2018) On the effect of inter-observer variability for a reliable estimation of uncertainty of medical image segmentation. In *Medical Image Computing and Computer Assisted Intervention–MICCAI 2018: 21st International Conference, Granada, Spain, September 16-20, 2018, Proceedings, Part I* Springer pp. 682–690.

[191] Monteiro, M., Le Folgoc, L., Coelho de Castro, D., Pawlowski, N., Marques, B., Kamnitsas, K., van der Wilk, M., and Glocker, B. (2020) Stochastic segmentation networks: Modelling spatially correlated aleatoric uncertainty. *Advances in neural information processing systems,* **33**, 12756–12767.

[192] Rousseau, A.-J., Becker, T., Bertels, J., Blaschko, M. B., and Valkenborg, D. (2021) Post training uncertainty calibration of deep networks for medical image segmentation. In *2021 IEEE 18th International Symposium on Biomedical Imaging (ISBI)* IEEE pp. 1052–1056.

[193] Karamalis, A., Wein, W., Klein, T., and Navab, N. (2012) Ultrasound confidence maps using random walks. *Medical image analysis,* **16**(6), 1101–1112.

[194] Berge, C. S. z., Kapoor, A., and Navab, N. (2014) Orientation-driven ultrasound compounding using uncertainty information. In *Information Processing in Computer-Assisted Interventions: 5th International Conference, IPCAI 2014, Fukuoka, Japan, June 28, 2014. Proceedings 5* Springer pp. 236–245.

[195] Wein, W., Karamalis, A., Baumgartner, A., and Navab, N. (2015) Automatic bone detection and soft tissue aware ultrasound–CT registration for computer-aided orthopedic surgery. *International journal of computer assisted radiology and surgery,* **10**, 971–979.

[196] Beitzel, J., Ahmadi, S.-A., Karamalis, A., Wein, W., and Navab, N. (2012) Ultrasound bone detection using patient-specific CT prior. In *2012 Annual International Conference of the IEEE Engineering in Medicine and Biology Society* IEEE pp. 2664–2667.

[197] Draelos, R. L. and Carin, L. (2020) Use HiResCAM instead of Grad-CAM for faithful explanations of convolutional neural networks. *arXiv preprint arXiv:2011.08891,*.

[198] Jamil, M. S., Banik, S. P., Rahaman, G. A., and Saha, S. (2023) Advanced Grad-CAM++: Improved Visual Explanations of CNN Decisions in Diabetic Retinopathy. In *Computer Vision and Image Analysis for Industry 4.0* pp. 64–75 CRC Press 6000 Broken Sound Parkway NW, Suite 300, Boca Raton, FL 33487-2742.

[199] Selvaraju, R. R., Cogswell, M., Das, A., Vedantam, R., Parikh, D., and Batra, D. (2017) Grad-cam: Visual explanations from deep networks via gradient-based localization. In *Proceedings of the IEEE international conference on computer vision* pp. 618–626.

[200] Zhou, B., Khosla, A., Lapedriza, A., Oliva, A., and Torralba, A. (2016) Learning deep features for discriminative localization. In *Proceedings of the IEEE conference on computer vision and pattern recognition* pp. 2921–2929.

[201] Schulz, K., Sixt, L., Tombari, F., and Landgraf, T. (2020) Restricting the flow: Information bottlenecks for attribution. *arXiv preprint arXiv:2001.00396,*.

[202] Khakzar, A., Baselizadeh, S., Khanduja, S., Rupprecht, C., Kim, S. T., and Navab, N. (2021) Neural response interpretation through the lens of critical pathways. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition* pp. 13528–13538.

[203] Suara, S., Jha, A., Sinha, P., and Sekh, A. A. Is Grad-CAM Explainable in Medical Images?. (2023).

[204] Lee, S., Lee, J., Lee, J., Park, C.-K., and Yoon, S. Robust Tumor Localization with Pyramid Grad-CAM. (2018).

[205] Jiang, H., Xu, J., Shi, R., Yang, K., Zhang, D., Gao, M., Ma, H., and Qian, W. (2020) A multi-label deep learning model with interpretable grad-CAM for diabetic retinopathy classification. In *2020 42nd Annual International Conference of the IEEE Engineering in Medicine & Biology Society (EMBC)* IEEE pp. 1560–1563.

[206] Xiao, M., Zhang, L., Shi, W., Liu, J., He, W., and Jiang, Z. (2021) A visualization method based on the Grad-CAM for medical image segmentation model. In *2021 International Conference on Electronic Information Engineering and Computer Science (EIECS)* IEEE pp. 242–247.

[207] Vinogradova, K., Dibrov, A., and Myers, G. (2020) Towards interpretable semantic segmentation via gradient-weighted class activation mapping (student abstract). In *Proceedings of the AAAI conference on artificial intelligence* Vol. 34, pp. 13943–13944.

[208] Gildenblat, J. and contributors PyTorch library for CAM methods. https://github.com/jacobgil/pytorch-grad-cam (2021).

[209] Hasany, S. N., Petitjean, C., and Mériaudeau, F. (2023) Seg-XRes-CAM: Explaining Spatially Local Regions in Image Segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition* pp. 3732–3737.

[210] Xia, F., Liu, J., Nie, H., Fu, Y., Wan, L., and Kong, X. (2019) Random walks: A review of algorithms and applications. *IEEE Transactions on Emerging Topics in Computational Intelligence,* **4**(2), 95–107.

[211] Müller, R., Kornblith, S., and Hinton, G. E. (2019) When does label smoothing help?. *Advances in neural information processing systems,* **32**.

[212] Islam, M. and Glocker, B. (2021) Spatially varying label smoothing: Capturing uncertainty from expert annotations. In *Information Processing in Medical Imaging:*

27th International Conference, IPMI 2021, Virtual Event, June 28–June 30, 2021, Proceedings 27 Springer pp. 677–688.

[213] Lourenço-Silva, J. and Oliveira, A. L. (2021) Using soft labels to model uncertainty in medical image segmentation. In *International MICCAI Brainlesion Workshop* Springer pp. 585–596.

[214] Guo, C., Pleiss, G., Sun, Y., and Weinberger, K. Q. (2017) On calibration of modern neural networks. In *International conference on machine learning* PMLR pp. 1321–1330.

[215] Jacob, J., Ciccarelli, O., Barkhof, F., and Alexander, D. C. (2021) Disentangling Human Error from the Ground Truth in Segmentation of Medical Images. ACL.

[216] Wenger, J., Kjellström, H., and Triebel, R. (2020) Non-parametric calibration for classification. In *International Conference on Artificial Intelligence and Statistics* PMLR pp. 178–190.

[217] Xie, P., Zuo, K., Liu, J., Chen, M., Zhao, S., Kang, W., Li, F., et al. (2021) Interpretable diagnosis for whole-slide melanoma histology images using convolutional neural network. *Journal of healthcare engineering,* **2021**.

[218] Hamza, A., Attique Khan, M., Wang, S.-H., Alhaisoni, M., Alharbi, M., Hussein, H. S., Alshazly, H., Kim, Y. J., and Cha, J. (2022) COVID-19 classification using chest X-ray images based on fusion-assisted deep Bayesian optimization and Grad-CAM visualization. *Frontiers in Public Health,* **10**, 1046296.

[219] Mehta, S., Mercan, E., Bartlett, J., Weave, D., Elmore, J. G., and Shapiro, L. Y-Net: Joint Segmentation and Classification for Diagnosis of Breast Biopsy Images. (2018).

[220] McManigle, J. E., Bartz, R. R., and Carin, L. (2020) Y-Net for Chest X-Ray preprocessing: Simultaneous classification of geometry and segmentation of annotations. In *2020 42nd Annual International Conference of the IEEE Engineering in Medicine & Biology Society (EMBC)* IEEE pp. 1266–1269.

[221] Kim, T., Kang, D.-H., Shim, S., Im, M., Seo, B. K., Kim, H., and Lee, B. C. (2020) Versatile low-cost volumetric 3D ultrasound imaging using gimbal-assisted distance sensors and an inertial measurement unit. *Sensors,* **20**(22), 6613.

[222] Peng, C., Cai, Q., Chen, M., and Jiang, X. (2022) Recent advances in tracking devices for biomedical ultrasound imaging applications. *Micromachines,* **13**(11), 1855.

[223] Kinnari, J., Thomas, A., Lusk, P., Kondo, K., and How, J. P. (2024) SOS-SLAM: Segmentation for Open-Set SLAM in Unstructured Environments. *arXiv preprint arXiv:2401.04791,*.

[224] Tateno, K., Tombari, F., and Navab, N. (2015) Real-time and scalable incremental segmentation on dense slam. In *2015 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)* IEEE pp. 4465–4472.

[225] Marzola, A., Di Angelo, L., Di Stefano, P., and Volpe, Y. (2024) An enhanced statistical shape model for automatic feature segmentation of human vertebrae. *Biomedical Signal Processing and Control,* **91**, 105972.

[226] Eck, B. L., Yang, M., Elias, J. J., Winalski, C. S., Altahawi, F., Subhas, N., and Li, X. (2023) Quantitative MRI for evaluation of musculoskeletal disease: cartilage and muscle composition, joint inflammation, and biomechanics in osteoarthritis. *Investigative radiology,* **58**(1), 60–75.

[227] Parvaiz, A., Khalid, M. A., Zafar, R., Ameer, H., Ali, M., and Fraz, M. M. (2023) Vision transformers in medical computer vision—A contemplative retrospection. *Engineering Applications of Artificial Intelligence,* **122**, 106126.

[228] Dai, Y., Jin, T., Song, Y., Sun, S., and Wu, C. (2020) Convolutional neural network with spatial-variant convolution kernel. *Remote Sensing,* **12**(17), 2811.

[229] Achtibat, R., Dreyer, M., Eisenbraun, I., Bosse, S., Wiegand, T., Samek, W., and Lapuschkin, S. (2023) From attribution maps to human-understandable explanations through concept relevance propagation. *Nature Machine Intelligence,* **5**(9), 1006–1019.

[230] Virtanen, P., Gommers, R., Oliphant, T. E., Haberland, M., Reddy, T., Cournapeau, D., Burovski, E., Peterson, P., Weckesser, W., Bright, J., van der Walt, S. J., Brett, M., Wilson, J., Millman, K. J., Mayorov, N., Nelson, A. R. J., Jones, E., Kern, R., Larson, E., Carey, C. J., Polat, İ., Feng, Y., Moore, E. W., VanderPlas, J., Laxalde, D., Perktold, J., Cimrman, R., Henriksen, I., Quintero, E. A., Harris, C. R., Archibald, A. M., Ribeiro, A. H., Pedregosa, F., van Mulbregt, P., and SciPy 1.0 Contributors (2020) SciPy 1.0: Fundamental Algorithms for Scientific Computing in Python. *Nature Methods,* **17**, 261–272.

[231] Cardoso, M. J., Li, W., Brown, R., Ma, N., Kerfoot, E., Wang, Y., Murrey, B., Myronenko, A., Zhao, C., Yang, D., et al. (2022) Monai: An open-source framework for deep learning in healthcare. *arXiv preprint arXiv:2211.02701,*.

[232] Wysocki, M., Azampour, M. F., Eilers, C., Busam, B., Salehi, M., and Navab, N. (2023) Ultra-NeRF: Neural Radiance Fields for Ultrasound Imaging. *arXiv preprint arXiv:2301.10520,*.

[233] Hung, A. L. Y., Chen, W., and Galeotti, J. (2021) Ultrasound confidence maps of intensity and structure based on directed acyclic graphs and artifact models. In *2021 IEEE 18th International Symposium on Biomedical Imaging (ISBI)* IEEE pp. 697–701.

COLLEGE DOCTORAL PAYS DE LA LOIRE
SCIENCES DE L'INGENIERIE ET DES SYSTEMES

Titre : **Méthodes d'acquisition, de segmentation par apprentissage et d'analyse quantitative des volumes échographiques.**

Mots clés : Échographie 3D, interpolation des étiquettes, cartes de confiance, variabilité des annotations, annotations négatives, Grad-Cam, paradigme multi-tâches

Résumé : 'objectif de cette thèse est de faire progresser le domaine de la segmentation échographique 3D et de relever les défis liés à la variabilité des annotations, aux artefacts d'image et aux ensembles de données incomplets pour la délimitation des muscles de la jambe. Elle est divisée en trois parties principales. La première partie explore des méthodes pour générer des volumes échographiques 3D de haute fidélité et des annotations, notamment l'interpolation des étiquettes à partir d'entrées 2D clairsemées et des techniques de composition sans capteur. La deuxième partie présente deux modèles innovants pour la segmentation de volumes échographiques 3D : UNet-S-R-CLSTM, qui traite des sous-volumes tout en prenant en compte les annotations incomplètes et IFSSnet, qui utilise un cadre récurrent pour des prédictions aux contours lisses. La troisième partie se concentre sur l'analyse expérimentale des facteurs influençant les performances de segmentation, en mettant l'accent sur la qualité des contours et la variabilité des annotations à travers des cartes de confiance et des études comparatives. Ce travail contribue au domaine en 1) développant des méthodes pour la reconstruction de volumes 3D et la correction des erreurs, 2) proposant des architectures adaptées à la segmentation échographique 3D, et 3) réalisant des évaluations approfondies pour comprendre les limitations des performances et améliorer la fiabilité diagnostique.

Title : **Methods for the acquisition, learning-based segmentation, and quantitative analysis of ultrasound volumes.**

Keywords : 3D ultrasound, label interpolation, confidence maps, labeling variability, negative labels, Grad-Cam, multi-task paradigm

Abstract : The goal of this thesis is to advance the field of 3D ultrasound segmentation and address the challenges posed by annotation variability, image artifacts, and incomplete datasets for leg muscle delineation. It is divided into three main parts. The first part explores methods for generating high-fidelity 3D ultrasound volumes and annotations, including label interpolation from sparse 2D inputs and sensorless compounding techniques. The second part introduces two innovative models for segmenting 3D ultrasound volumes: UNet-S-R-CLSTM, which processes sub-volumes while accounting for incomplete annotations, and IFSSnet, which uses a recurrent framework for smooth border predictions. The third part focuses on experimental analysis of factors affecting segmentation performance, emphasizing border quality and annotation variability through confidence maps and comparative studies. This work contributes to the field by 1) developing methods for 3D volume reconstruction and error correction, 2) proposing architectures tailored for 3D ultrasound segmentation, and 3) conducting comprehensive evaluations to understand performance limitations and improve diagnostic reliability.