Automatisation de la segmentation sémantique de structures cardiaques en imagerie ultrasonore par apprentissage supervisé

Sarah Marie-Solveig Leclerc

N°d'ordre NNT : 2019LYSEI121

THESE de DOCTORAT DE L'UNIVERSITE DE LYON
opérée au sein de
l'INSA de Lyon

Ecole Doctorale N° 160
Electronique, Electrotechnique, Automatique (EEA)

Spécialité/ discipline de doctorat :
Traitement du Signal et de l'Image

Soutenue publiquement le 11/12/2019, par :
Sarah Marie-Solveig Leclerc

Automatisation de la segmentation sémantique de structures cardiaques en imagerie ultrasonore par apprentissage supervisé

Devant le jury composé de :

Rueckert, Daniel	Professor	Imperial College of London	Rapporteur
Garreau, Mireille	Professeur	Université de Rennes 1	Rapporteure
Thiran, Jean-Philippe	Professeur	EPFL	Examinateur
Noble, Alison	Professor	Oxford University	Examinatrice
Lartizien, Carole	Directrice de recherche	CNRS	Directrice de thèse
Bernard, Olivier	Maître de conférence HDR	INSA de Lyon	Co-directeur
Grenier, Thomas	Maître de conférence	INSA de Lyon	Examinateur
Jodoin, Pierre-Marc	Professeur	Université de Sherbrooke	Examinateur

Département FEDORA – INSA Lyon - Ecoles Doctorales – Quinquennal 2016-2020

SIGLE	ECOLE DOCTORALE	NOM ET COORDONNEES DU RESPONSABLE
CHIMIE	**CHIMIE DE LYON** **http://www.edchimie-lyon.fr** Sec. : Renée EL MELHEM Bât. Blaise PASCAL, 3e étage **secretariat@edchimie-lyon.fr** INSA : R. GOURDON	**M. Stéphane DANIELE** Institut de recherches sur la catalyse et l'environnement de Lyon IRCELYON-UMR 5256 Équipe CDFA 2 Avenue Albert EINSTEIN 69 626 Villeurbanne CEDEX **directeur@edchimie-lyon.fr**
E.E.A.	**ÉLECTRONIQUE, ÉLECTROTECHNIQUE, AUTOMATIQUE** **http://edeea.ec-lyon.fr** Sec. : M.C. HAVGOUDOUKIAN **ecole-doctorale.eea@ec-lyon.fr**	**M. Gérard SCORLETTI** École Centrale de Lyon 36 Avenue Guy DE COLLONGUE 69 134 Écully Tél : 04.72.18.60.97 Fax 04.78.43.37.17 **gerard.scorletti@ec-lyon.fr**
E2M2	**ÉVOLUTION, ÉCOSYSTÈME, MICROBIOLOGIE, MODÉLISATION** **http://e2m2.universite-lyon.fr** Sec. : Sylvie ROBERJOT Bât. Atrium, UCB Lyon 1 Tél : 04.72.44.83.62 INSA : H. CHARLES **secretariat.e2m2@univ-lyon1.fr**	**M. Philippe NORMAND** UMR 5557 Lab. d'Ecologie Microbienne Université Claude Bernard Lyon 1 Bâtiment Mendel 43, boulevard du 11 Novembre 1918 69 622 Villeurbanne CEDEX **philippe.normand@univ-lyon1.fr**
EDISS	**INTERDISCIPLINAIRE SCIENCES-SANTÉ** **http://www.ediss-lyon.fr** Sec. : Sylvie ROBERJOT Bât. Atrium, UCB Lyon 1 Tél : 04.72.44.83.62 INSA : M. LAGARDE **secretariat.ediss@univ-lyon1.fr**	**Mme Sylvie RICARD-BLUM** Institut de Chimie et Biochimie Moléculaires et Supramoléculaires (ICBMS) - UMR 5246 CNRS - Université Lyon 1 Bâtiment Curien - 3ème étage Nord 43 Boulevard du 11 novembre 1918 69622 Villeurbanne Cedex Tel : +33(0)4 72 44 82 32 **sylvie.ricard-blum@univ-lyon1.fr**
INFOMATHS	**INFORMATIQUE ET MATHÉMATIQUES** **http://edinfomaths.universite-lyon.fr** Sec. : Renée EL MELHEM Bât. Blaise PASCAL, 3e étage Tél : 04.72.43.80.46 **infomaths@univ-lyon1.fr**	**M. Hamamache KHEDDOUCI** Bât. Nautibus 43, Boulevard du 11 novembre 1918 69 622 Villeurbanne Cedex France Tel : 04.72.44.83.69 **hamamache.kheddouci@univ-lyon1.fr**
Matériaux	**MATÉRIAUX DE LYON** **http://ed34.universite-lyon.fr** Sec. : Stéphanie CAUVIN Tél : 04.72.43.71.70 Bât. Direction **ed.materiaux@insa-lyon.fr**	**M. Jean-Yves BUFFIÈRE** INSA de Lyon MATEIS - Bât. Saint-Exupéry 7 Avenue Jean CAPELLE 69 621 Villeurbanne CEDEX Tél : 04.72.43.71.70 Fax : 04.72.43.85.28 **jean-yves.buffiere@insa-lyon.fr**
MEGA	**MÉCANIQUE, ÉNERGÉTIQUE, GÉNIE CIVIL, ACOUSTIQUE** **http://edmega.universite-lyon.fr** Sec. : Stéphanie CAUVIN Tél : 04.72.43.71.70 Bât. Direction **mega@insa-lyon.fr**	**M. Jocelyn BONJOUR** INSA de Lyon Laboratoire CETHIL Bâtiment Sadi-Carnot 9, rue de la Physique 69 621 Villeurbanne CEDEX **jocelyn.bonjour@insa-lyon.fr**
ScSo	**ScSo*** **http://ed483.univ-lyon2.fr** Sec. : Véronique GUICHARD INSA : J.Y. TOUSSAINT Tél : 04.78.69.72.76 **veronique.cervantes@univ-lyon2.fr**	**M. Christian MONTES** Université Lyon 2 86 Rue Pasteur 69 365 Lyon CEDEX 07 **christian.montes@univ-lyon2.fr**

*ScSo : Histoire, Géographie, Aménagement, Urbanisme, Archéologie, Science politique, Sociologie, Anthropologie

“
Penser, ce n’est pas unifier, rendre familière l’apparence sous le visage d’un grand principe. Penser, c’est réapprendre à voir, diriger sa conscience, faire de chaque image un lieu privilégié.

Thinking is learning all over again how to see, direct one’s consciousness, make of every image a privileged place.
”

Le mythe de Sisyphe, Albert Camus, 1942

Foreword

This thesis entitled "Supervised machine learning for the automatic semantic segmentation of anatomical structures in cardiac ultrasound imaging" is about the automation of the segmentation task in echocardiographic images. The investigation of this on-going problem is primordial in order to alleviate the workload of cardiologists and improve the reliability of the estimation of the clinical indices used to establish medical diagnosis.

As supervised learning have shown an astounding potential in replicating human performance on complex tasks, including in medical imaging, we decided to investigate the clinical potential of such methods for echocardiographic semantic segmentation. The main objective is to reach human expert performance on this task. This manuscript reports the experiments and results obtained throughout my PhD, and was written to serve as a stepping stone for further study on the subject.

Several of the implementations detailed here are to be attributed to my collaborators, these contributions being as stated:

1. Olivier Bernard was in charge of the implementation of the CAMUS dataset, on which we beneficiated from the help of the three cardiologists Florian Espinosa (from the university hospital of Saint-Etienne, Saint-Etienne, France), and Erik Andreas Rye Berg and Torvald Espeland (from the Saint Olavs' hospital and the centre for innovative ultrasound solutions of the NTNU university, Trondheim, Norway);

2. Erik Smistad (from NTNU, Trondheim, Norway), Joao Pedrosa (at that time at KU Leuven, Leuven, Belgium), and Ferriel Khellaf (at that time at Erasmus MC, Rotterdam, Netherlands) respectively implemented the U-Net 1 model presented in Chapter 7, the BEASM algorithm in Section 7.3.4, and the Active Shape Models in Section 6.4.2.1;

3. Erik Smistad provided the code used to evaluate the clinical indices on the CAMUS dataset, as well as the back-bone code of a common evaluation platform involving cross-validation, plot and data loading functions;

4. Ferriel Khellaf computed the scores on the CETUS dataset of the pipeline described in Section 6.4.2.1;

5. Frederic Cervenansky set up the CAMUS online evaluation platform along with the corresponding web site for the CAMUS challenge [1], with the help of Olivier Bernard.

A few details on all the collaborators of the project are given in Appendix A.

INSA DE LYON

Abstract

Université de Lyon
Ecole EEA de Lyon: thématique Traitement du signal et de limage

PhD / Doctorat

Supervised machine learning for the automatic semantic segmentation of anatomical structures in cardiac ultrasound imaging

by Sarah Leclerc

The analysis of medical images plays a critical role in cardiology. Ultrasound imaging, as a real-time, low cost and bed side applicable modality, is nowadays the most commonly used image modality to monitor patient status and perform clinical cardiac diagnosis. However, the semantic segmentation (i.e the accurate delineation and identification) of heart structures is a difficult task due to the low quality of ultrasound images, characterized in particular by the lack of clear boundaries.

To compensate for missing information, the best performing methods before this thesis relied on the integration of prior information on cardiac shape or motion, which in turns reduced the adaptability of the corresponding methods. Furthermore, such approaches require manual identifications of key points to be adapted to a given image, which makes the full process difficult to reproduce. In this thesis, we propose several original fully-automatic algorithms for the semantic segmentation of echocardiographic images based on supervised learning approaches, where the resolution of the problem is automatically set up using data previously analyzed by trained cardiologists.

From the design of a dedicated dataset and evaluation platform, we prove in this project the clinical applicability of fully-automatic supervised learning methods, in particular deep learning methods, as well as the possibility to improve the robustness by incorporating in the full process the prior automatic detection of regions of interest.

Acknowledgements

Contents

List of Figures

List of Tables

List of Abbreviations

Chapter 2

SVM	**S**upport **V**ector **M**achine
RF	**R**andom **F**orest(s)
CAMUS	**C**ardiac **A**cquisitions for **M**ulti-structure **U**ltrasound **S**egmentation

Chapter 3

SNR	**S**ignal to **N**oise **R**atio
LV	**L**eft **V**entricle
RV	**R**ight **V**entricle
RA	**R**ight **A**trium
LA	**L**eft **A**trium
ED	**E**nd **D**iastole
ES	**E**nd **S**ystole
$\mathbf{LV_{EDV}}$	**L**eft **V**entricle **E**nd **D**iastole **V**olume
$\mathbf{LV_{ESV}}$	**L**eft **V**entricle **E**nd **S**ystole **V**olume
EF	**L**eft **V**entricle **E**jection **F**raction
$\mathbf{LV_{EF}}$	**E**jection **F**raction
A4CH	**A**pical **4** **C**hamber **V**iew
A2CH	**A**pical **2** **C**hamber **V**iew

Chapter 4

TP	**T**rue **P**ositive
FP	**F**alse **P**ositive
TN	**T**rue **N**egative
FN	**F**alse **N**egative
D	**D**ice
JAC	**Jac**card index
MAD	**M**ean **A**bsolute **D**istance
HD	**H**ausdorff **D**istance
MRI	**M**agnetic **R**esonance **I**maging
CT	**C**omputed **T**omography
CNN	**C**onvolutional **N**eural **N**etwork
ACDC	**A**utomated **C**ardiac **D**iagnosis **C**hallenge
CETUS	**C**hallenge on **E**ndocardial **T**hree-dimensional **U**ltrasound **S**egmentation
ASM	**A**ctive **S**hape **M**odel
AAM	**A**ctive **A**ppearance **M**odel
BEAS(M)	**B**-spline **E**xplicit **A**ctive **S**urface (**M**odel)
SRF	**S**tructured **R**andom **F**orest
ANN	**A**rtificial **N**eural **N**etwork

DL	**D**eep **L**earning
CNN	**C**onvolutional **N**eural **Network**
ACNN	**A**natomically **C**onstrained **N**eural **N**etwork
corr	**corr**elation
LOA	**L**imit **O**f **Agreement**

Chapter 5

IQ	**I**mage **Q**uality
$\mathbf{LV_{endo}}$	**L**eft **V**entricle **E**ndocardium
$\mathbf{LV_{epi}}$	**L**eft **V**entricle **E**picardium
$\mathbf{O_1}$	**O**bserver **1**
$\mathbf{O_2}$	**O**bserver **2**
$\mathbf{O_3}$	**O**bserver **3**

Chapter 6

PCA	**P**rincipal **C**omponent **A**nalysis
HOG	**H**istogram **O**f **G**radients
LAX	**L**ong **AX**is view

Chapter 7

CPU	**C**entral **P**rocessing **U**nit
GPU	**G**raphics **P**rocessing **U**nit
EDN	**E**ncoder **D**ecoder **N**etwork

Chapter 8

DS	**D**eep **S**upervision
SHG	**S**tacked **H**our**G**lasses

Chapter 9

BB	**B**ounding **B**ox
ROI	**R**egion **O**f **I**nterest
RU-Net	**R**efining **U-Net**
LU-Net	**L**ocalization **U-Net**

À Ginette et Jacqueline, mes grands-mères adorées,

To Ginette and Jacqueline, my beloved grandmothers,

Part I

Presentation

Chapter 1

Résumé en Français (French Summary)

L'étude présentée dans ce manuscript porte sur l'**Automatisation de la segmentation sémantique de structures cardiaques en imagerie ultrasonore par apprentissage supervisé**. Ce premier chapitre est dédié à la reprise en français des points clés abordés dans le manuscript complet écrit en anglais.

À ce titre, il contient dans cet ordre :

- un abstract, synthétisant en quelques mots la thématique et les contributions;
- une traduction du Chapitre d'introduction (2);
- une sélection du Chapitre présentant l'échocardiographie (3);
- une sélection du Chapitre sur l'état de l'art de la segmentation en échocardiographie (4);
- un résumé du Chapitre décrivant la base de données construite pour le projet (5);
- un résumé du Chapitre sur l'adaptation de la méthode des forêts aléatoires structurées pour la segmentation d'images échocardiographiques (6);
- un résumé du Chapitre sur l'évaluation du potentiel des méthodes d'apprentissage profond pour la segmentation automatique d'images échocardiographiques (7);
- un résumé du Chapitre sur le perfectionnement des modèles d'apprentissage profond et de leurs métriques d'évaluation dans le cadre de la segmentation d'images échocardiographiques (8);
- un résumé du Chapitre sur l'amélioration de la robustesse de la segmentation produite par des modèles d'apprentissage profond grâce à l'ajout de mécanismes d'attention, appliqué à la segmentation d'images échocardiographiques (9);

1.1 Abstract

L'analyse d'images médicales joue un rôle essentiel en cardiologie pour la réalisation du diagnostique cardiaque clinique et le suivi de l'état du patient. Parmis les modalités d'imagerie utilisées, l'imagerie par ultrasons, temps réelle, moins coûteuse et portable au chevet du patient, est de nos jours la plus courante.

Malheureusement, l'étape nécessaire de segmentation sémantique (soit l'identification et la délimitation précise) des structures cardiaques est difficile en échocardiographie à cause de la faible qualité des images ultrasonores, caractérisées en particulier par l'absence d'interfaces nettes entre les différents tissus.

Pour combler le manque d'information, les méthodes les plus performante, avant ces travaux, reposaient sur l'intégration d'informations a priori sur la forme ou le mouvement du cœur, ce qui en échange réduisait leur adaptabilité au cas par cas. De plus, de telles approches nécessitent pour être efficaces l'identification manuelle de plusieurs repères dans l'image, ce qui rend le processus de segmentation difficilement reproductible.

Dans cette thèse, nous proposons plusieurs algorithmes originaux et entièrement automatiques pour la segmentation sémantique d'images échocardiographiques. Ces méthodes génériques sont adaptées à la segmentation échocardiographique par apprentissage supervisé, c'est-à-dire que la résolution du problème est construite automatiquement à partir de données pré-analysées par des cardiologues entraînés.

Grâce au développement d'une base de données et d'une plateforme d'évaluation dédiées au projet, nous montrons le fort potentiel clinique des méthodes automatiques d'apprentissage supervisé, et en particulier d'apprentissage profond, ainsi que la possibilité d'améliorer leur robustesse en intégrant une étape de détection automatique des régions d'intérêt dans l'image.

1.2 Introduction

Ce chapitre expose les motivations de ce travail, et les objectifs correspondants. Nous y présentons ensuite la méthodologie appliquée tout au long de la thèse et terminons par l'organisation du manuscrit.

1.2.1 Motivation

Le contexte scientifique dans lequel la présente étude a été menée comprend d'une part des défis cliniques et d'autre part des défis méthodologiques. L'identification de ces barrières nous a incité à définir des objectifs clairs pour chacun de ces aspects, et à élaborer des stratégies spécifiques pour les atteindre.

1.2.1.1 Contexte scientifique

Contexte clinique Les maladies cardiovasculaires figurent parmi les principales causes de mortalité et de morbidité dans le monde. Alors qu'elles représentaient 30 % du total des décès en 2008, le nombre de victimes ne cesse d'augmenter, en partie à cause du vieillissement de la population. D'ici 2030, on estime que plus de 20 millions de personnes mourront chaque année des suites de telles maladies, ce qui encourage le développement de nouvelles pratiques en clinique afin de favoriser leur diagnostic précoce [2]. Comme la plupart des pathologies cardiaques affectent la forme et le comportement des structures cardiaques, l'imagerie médicale non invasive est une solution de choix car elle permet d'établir le diagnostic à partir de la visualisation des différentes structures cardiaques et de l'évaluation de leur fonctionnement.

L' imagerie par ultrasons est actuellement la modalité la plus utilisée pour l'imagerie cardiaque [3], car elle permet une bonne résolution temporelle à un coût relativement faible comparé aux autres modalités. Puisque l'usage de l'échocardiographie 3D est encore nouveau en routine clinique [4], l'imagerie ultrasonore 2D reste la modalité la plus répandue pour réaliser l'estimation d'indices cliniques tels que les volumes ventriculaires et les fractions d'éjection. Ces mesures reposent alors sur l'approximation de volumes à partir de surfaces, et dépendent fortement du processus d'acquisition.

En échocardiographie, le tissu myocardique apparaît dans l'image en forte intensité (blanc) et peut être différencié des cavités remplies de sang qui sont elles associées à des intensités basses (noir). La séparation et l'identification des différentes structures à partir d'une délimitation précise, tâche d'analyse appelée segmentation sémantique, est la première étape pour effectuer des mesures de surfaces et de volumes.

Cependant, la segmentation en échocardiographie est une tâche particulièrement difficile en raison de l'absence de frontières nettes, du faible rapport signal/bruit et de la texture chatoyante propre aux images échographiques appelée speckle, auxquels s'ajoutent de nombreux et complexes artéfacts d'image. La conséquence directe de ces attributs est que les logiciels réalisant de manière entièrement automatique la segmentation d'images échocardiographiques fonctionnent assez mal, ce qui oblige les cliniciens à tracer les différents contours à l'aide d'outils semi-automatiques [5].

Contexte algorithmique Les annotations réalisées manuellement ou semi-automatiquement ne sont pas reproductibles et sont sujettes, en plus d'être coûteuses en terme de temps, à des différences inter- et intra- observateurs. Afin d'améliorer le déroulement de l'analyse cardiaque en routine clinique, l'automatisation de la segmentation du cœur a donc fait l'objet

d'intenses recherches au cours des dernières décennies, avec un effort particulier sur la segmentation du ventricule gauche [3].

La segmentation en échocardiographie a été historiquement abordée par des méthodes de traitement d'image morphomathématique reposant sur lexploitation d'information pixellique et donc uniquement locale, puis par des algorithmes à base de contour actifs exploitant des contraintes pré-établies pour régulariser globalement les résultats de segmentation. En particulier, l'utilisation d'informations a priori sur la forme recherchée s'est avérée particulièrement efficace pour inférer à partir du contexte les interfaces localement manquantes dans l'image. D'autre part, l'incorporation de lissage temporel des résultats de segmentation a permis d'encourager une cohérence de la segmentation au long du cycle cardiaque.

Les méthodes d'apprentissage supervisé rassemblent des algorithmes effectuant un mappage entre l'espace image et l'espace de solution dont les paramètres sont déduits de cas résolus (paires image/solution). Ces méthodes sont devenues populaires en vision par ordinateur au cours des années 90 grâce au développement des modèles de forme actifs [6], des machines à vecteurs de support (SVM) et des forêts aléatoires (RF) [7]. Cependant, la difficultés d'obtenir des données annotées par des experts a limité leur application en imagerie médicale, et en particulier en échocardiographie.

1.2.1.2 Verrous techniques

L'établissement d'ensembles de données adaptés Le premier verrou technique pour l'automatisation de la segmentation échocardiographique concerne le développement de larges ensembles de données annotées. Ces ensembles de données sont non seulement nécessaires pour l'entraînement des méthodes d'apprentissage supervisé, mais aussi pour l'évaluation précise de la performance des algorithmes.

Afin de répondre aux besoins cliniques, un ensemble de données annotées de segmentation en imagerie médicale doit contenir et être représentatif de la variabilité à laquelle les cliniciens sont confrontés dans leur pratique quotidienne (pathologies, qualités d'image, matériels et paramètres d'acquisition...), c'est-à-dire ne pas être limité à des simulations, même réalistes, ou comporter uniquement des cas réels avec une qualité d'image élevée [3].

De plus, les annotations doivent être établies par des experts pour toutes les vues et tous les instants pertinents suivant un protocole fixé et consensuel, ce idéalement dans un contexte favorisant la précision (hors exercice et avec un logiciel de traçage approprié). Les mêmes structures d'intérêt doivent être annotées sur toutes les images, et tous les renseignements cliniques pertinents pour le patient doivent être renseignés avec les images de l'examen. Enfin, pour être entièrement validé et bénéfique pour la communauté, un tel ensemble de données devrait être publique d'accès, tout du moins aux chercheurs et aux cliniciens.

La construction d'algorithmes robustes et entièrement automatiques Le second défi technique consiste à mettre en place des algorithmes de segmentation robustes et entièrement automatiques comme socles de logiciels d'analyse fiables, assistant les cardiologues lors des examens (analyse d'image, calcul d'indices cliniques, guide d'acquisition...). En vue d'un diagnostic précis et reproductible, la robustesse des algorithmes devient l'un des principaux critères de qualité, en tant que mesure de fiabilité.

Afin d'encourager leur adoption en pratique clinique [3], la comparaison des méthodes doit être effectuée sur un ensemble de données large et validé, ce qui permet en plus de sélectionner les meilleurs algorithmes de mieux analyser les limites des différentes approches. En ce qui concerne les méthodes d'apprentissage supervisé, l'analyse des erreurs réalisées par de multiples algorithmes sur un ensemble de données validé par la communauté peut également aider à évaluer les cas manquants dans la base pour une meilleure robustesse des modèles d'apprentissage.

Enfin, les solutions développées devraient idéalement être optimisées sur le plan de la vitesse dans le but d'être intégrées dans des logiciels d'analyse temps réel qui seront embarqués dans de nouvelles générations d'échographes.

1.2.2 Méthodologie

1.2.2.1 Objectifs

En raison de l'absence d'ensembles de données appropriés, la performance réelle des méthodes de l'état de l'art en analyse d'images échocardiographiques n'est pas établie. Afin de contribuer au développement d'une solution de segmentation automatique des structures cardiaques en imagerie ultrasonore qui soit adaptée aux besoins cliniques, ce travail a pour objectif d'apporter des réponses aux questions suivantes :

1. Quel est le potentiel des méthodes de segmentation sémantique par apprentissage supervisé en échocardiographie 2D ?
2. Sommes-nous proches d'une complète automatisation de l'analyse cardiaque en imagerie ultrasonore ?

1.2.2.2 Méthode

L'évaluation complète du potentiel des méthodes d'apprentissage supervisé pour l'analyse d'images échocardiographiques ne peut être effectuée que par un étalonnage des méthodes les plus performantes sur un même ensemble de données. Cela nécessite :

1. d'utiliser un large ensemble de données annotées ;
2. d'établir un ensemble approprié de métriques géométriques et cliniques ;
3. de mettre en œuvre et d'adapter les techniques de l'état de l'art à notre problèmatique;
4. d'évaluer les méthodes sur une même plate-forme d'évaluation.

Ainsi, le potentiel clinique des algorithmes de segmentation automatique pourrait être établi à partir de :

1. la comparaison entre la meilleure méthode et les scores inter- et intra-experts ;
2. l'analyse des cas aberrants pour évaluer la robustesse ;
3. l'analyse d'erreurs afin d'établir des pistes d'amélioration prometteuses;
4. l'élaboration de nouvelles méthodes dédiées à l'amélioration de la robustesse et de la précision de la segmentation.

La rigueur et la qualité de l'évaluation ont été notre principale priorité, ce qui nous a amenés à contribuer non seulement sur les aspects algorithmiques de la thématique, mais aussi sur les métriques utilisées pour mesurer la qualité du résultat.

1.2.3 Organisation du manuscript

Le manuscrit est découpé en quatre parties principales, toutes composées de chapitres indépendants qui abordent progressivement les aspects méthodologiques suivants :

1. Présentation
    - Chapitre 1 : le résumé de la thèse, en français à la demande de l'école doctorale EEA, qui couvre toutes les discussions abordées dans le manuscrit en se concentrant sur les points clés ;
    - Chapitre 2 : l'introduction, dans laquelle les motivations et la stratégie méthodologique de la thèse sont présentées, ainsi que l'organisation détaillée du manuscrit.
2. Contexte
    - Chapitre 3 : les bases de l'échocardiographie, où nous décrivons les aspects les plus pertinents de l'imagerie échographique cardiaque par rapport à cette étude;
    - Chapitre 4 : la revue de l'état de l'art, qui détaille les ensembles de données, méthodes et métriques éxistants en segmentation d'images échocardiographiques.
3. Contributions
    - Chapitre 5 : la création de la base de données CAMUS, à ce jour l'ensemble de données de référence libre d'accès le plus complet en échocardiographie 2D ;
    - Chapitre 6 : l'adaptation des forêts aléatoires structurées (méthode d'apprentissage automatique) à la segmentation échocardiographique 2D et 3D;
    - Chapitre 7 : l'étude et l'évaluation de méthodes d'apprentissage profond à travers l'architecture U-Net ;
    - Chapitre 8 : le perfectionnement des modèles encodeur-décodeurs et de leur évaluation grâce à la conception de métriques anatomiques;
    - Chapitre 9 : le développement de modèles intégrant des mécanismes d'attention, destinés à améliorer la robustesse de la segmentation.
4. Epilogue
    - Chapter 10 : la conclusion, avec un retour sur les principales réalisations et le détail des perspectives de nos travaux.

Le manuscrit est complété par une série d'annexes :

A la présentation des collaborateurs de ce projet ;

B la liste des publications ;

C la description des ordinateurs utilisés dans l'étude ;

D des informations supplémentaires sur le Chapitre 5 ;

E des informations supplémentaires sur le Chapitre 7 ;

F des informations supplémentaires sur le Chapitre 8 ;

G des informations supplémentaires sur le Chapitre 9 ;

Pour finir, toutes les références sont listées dans la section bibliographique qui clôt le manuscrit.

1.3 Échocardiographie

L'échocardiographie est une modalité d'imagerie dédiée au diagnostique cardiaque qui repose sur l'insonification du corps au moyen d'ultrasons afin d'en visualiser les structures internes, le cœur dans ce cas précis. L'échographie est la principale modalité utilisée en imagerie cardiaque car c'est la modalité non invasive la moins coûteuse et la plus rapide. Le principal inconvénient à l'utilisation des ultrasons réside dans la basse qualité des images reconstruites, qui rend difficile leur interprétation et leur analyse automatique. Dans cette section, nous décrivons en guise d'introduction à l'échocardiographie :

1. la génération, la propagation et la réception des ondes ultrasonores;
2. la formation, les caractéristiques, et les différents modes d'images en échocardiographie;
3. l'estimation d' indices cliniques à partir d'images échocardiographiques.

1.3.1 Formation des images ultrasonores

L'imagerie par ultrasons est basée sur la transmission d'impulsions ultrasonores aux tissus mous tels que les muscles et les vaisseaux sanguins. Une image des tissus est reconstruite à partir des échos renvoyés alors que l'onde est rétrodiffusée et réfléchie. La section suivante donne une explication basique du processus de formation des images échocardiographiques.

1.3.1.1 Emission et réception de l'onde

La partie principale des sondes à ultrasons cardiaques correspond à un réseau phasé 1D ou 2D de transducteurs piézoélectriques permettant respectivement des acquisitions 2D ou 3D. Les cristaux piézoélectriques servent ici à transformer l'énergie électrique en pression acoustique, et inversement.

Les pulses d'ondes ultrasonores sont générés par de petites vibrations des éléments piézoélectriques de la sonde, induites par l'application d'un signal électrique sinusoïdal d'amplitude et de fréquence appropriées. Après la phase d'émission, la sonde reçoit les échos de l'onde émise et produit en réponse un signal électronique caractéristique des interfaces entre milieux d'impédances acoustiques différentes rencontrées.

Pour les applications médicales, les fréquences des ultrasons utilisées se situent généralement entre 1 et 15 MHz, soit bien au-delà du seuil ultrasonore situé à ($f_u > 20$ kHz) qui définit la limite de l'audition humaine [8]. La longueur d'onde associée, $\lambda = \frac{c}{f}$, joue un rôle important dans l'interaction entre l'onde et les tissus biologiques.

1.3.1.2 Propagation et réflection de l'onde

Les ultrasons sont des ondes mécaniques longitudinales, aussi appelées ondes de compression, où l'oscillation des particules du milieu se fait dans la direction de propagation des ondes. L'équation de l'évolution de la pression acoustique **??** peut être considérée vraie sous l'hypothèse qu'aucune énergie n'est perdue au cours du trajet dans le milieu :

$$\frac{\partial^2 p}{\partial t^2} = c_m^2 \times \frac{\partial^2 p}{\partial x^2} \tag{1.1}$$

avec c_m la vitesse de propagation de l'onde dans le milieu homogène et p la pression acoustique à la position x et au temps t. c_m dépend des caractéristiques du tissu, plus précisément de sa compressibilité κ et de sa densité ρ :

$$c_m = \frac{1}{\sqrt{\kappa \times \rho}} \tag{1.2}$$

Étant donné que la vitesse moyenne à l'intérieur des tissus mous ne varie que légèrement selon le tissu, les échographes sont calibrés en supposant que le son voyage à l'intérieur du corps à une vitesse constante de $c_s = 1540$ m/s.

En raison du changement d'impédance, une réflexion spéculaire se produit à l'interface entre deux milieux si celle-ci est de taille supérieure à la longueur d'onde de l'onde ultrasonore. La quantité d'échos réfléchis, à laquelle l'amplitude du signal électrique est proportionnelle, dépend de la nature des structures (échogénicité) et de l'angle entre la sonde et l'interface.

En supposant que la vitesse soit constante, le temps de vol de l'onde permet de localiser les interfaces. Cependant, il est important de garder à l'esprit que :

- les échos des éléments sont réfléchis et réfractés selon la loi de Snell-Descartes, d'où le besoin de focaliser les faisceaux afin d'éviter de mal localiser les réflecteurs ;
- les tissus mous ne sont pas homogènes, donc la vitesse à l'intérieur d'un tissu donné n'est pas constante en vérité. De plus, les inhomogénéités provoquent des effets de diffusion s'additionnant à la réflection. Ces petits diffuseurs agissent comme des sources secondaires d'ondes sphériques, produisant une texture d'image caractéristique appelée speckle;
- l'onde ultrasonore est atténuée au cours de sa propagation selon un coefficient qui augmente avec la fréquence et varie avec le milieu. Par conséquent, l'amplitude de l'écho rétrodiffusé doit être pondérée en fonction de la profondeur pour une bonne reconstruction de l'image.

1.3.1.3 Adaptation à la profondeur

La profondeur maximale est d'environ 5 à 10 cm pour l'échographie cardiaque pédiatrique et de 10 à 20 cm pour l'imagerie cardiaque chez l'adulte [9]. Pour compenser l'atténuation à l'intérieur du corps, on applique un gain proportionnel au temps de vol des échos, soit à la profondeur d'image. Cette pondération est nécessaire pour garantir la même visibilité sur tout le support insonifié. Elle est traditionnellement appliquée au niveau de l'acquisition, à la réception des échos, plutôt qu'en post-traitement de l'image formée. Le gain est défini manuellement ou automatiquement selon l'échographe.

1.3.1.4 Formation de voie

La technique de formation de voie la plus courante s'appelle "Delay And Sum" (DAS) soit "délai et sommation". Afin de créer des faisceaux focalisés orientés, le front d'onde de l'onde ultrasonore est orienté par l'application de temporisations sur les différents éléments de la sonde. Ces délais sont également appliqués sur les échos reçus avant sommation pour compenser les différences de distance entre les différents éléments piézo-électriques émetteurs et le point de focalisation.

Traditionnellement, pour reconstruire un visuel du cœur, plusieurs faisceaux focalisés sont émis successivement à traver les côtes pour concentrer localement l'énergie délivrée [9], en balayant le champ de vue. L'image est ensuite reconstruite en utilisant l'enveloppe des signaux formés par le faisceau afin de récupérer les intensités (B-mode), et une compression logarithmique est utilisée pour améliorer le contraste.

1.3.2 Caractéristiques des images ultrasonores

Les images échographiques présentent des caractéristiques et des niveaux de qualité de reconstruction variables, en partie liés au processus d'acquisition. Nous abordons brièvement dans cette partie les notions de résolution d'image, de contraste et d'artéfacts en imagerie échocardiographique.

1.3.2.1 Résolutions spatiales

En raison de l'atténuation dans les tissus, l'imagerie ultrasonore impose un compromis entre la profondeur accessible et la résolution spatiale, contrôlé par la fréquence de l'onde.

Résolution axiale La distance minimale permettant de différencier deux réflecteurs adjacents le long de l'axe de propagation des ondes est proportionnelle à la longueur d'onde et à la durée de l'impulsion [10] :

$$d_{a_{min}} = \tau \times \frac{c_s}{2} = \frac{n}{f} \times \frac{\lambda \times f}{2} = \frac{n \times \lambda}{2} \tag{1.3}$$

où τ correspond à la durée d'une impulsion et n au nombre de cycles de l'impulsion avant de revenir à l'équilibre.

Résolution latérale La résolution latérale, soit l'épaisseur apparente des diffuseurs le long de la direction perpendiculaire au faisceau est [9] :

$$d_{l_{min}} = \lambda \times \frac{d_F}{L_p} \tag{1.4}$$

où d_F correspond à la profondeur de champ et L_p à la longueur de la sonde. La meilleure résolution latérale est donc obtenue pour de hautes fréquences et de larges sondes. Cependant, l'étroitesse de l'ouverture entre les côtes impose une petite longueur de sonde cardiaque, de l'ordre de 3 cm.

1.3.2.2 Résolution temporelle

Une nouvelle impulsion peut être émise une fois que les échos sont revenus après avoir atteint la profondeur maximale souhaitée : $t_r = \frac{2 \times d_{max}}{c_s}$. Ainsi, la profondeur et le nombre de lignes conditionnent la fréquence d'images atteignable [9] :

$$F_r = \frac{1}{t_r \times n_l} \tag{1.5}$$

avec t_r le temps de retour défini ci-dessus, et n_l le nombre de lignes. En pratique clinique, F_r est habituellement d'environ 50 images par seconde, soit le double de la fréquence d'images en vidéo classique.

1.3.2.3 Contraste

Le contraste fait référence à la capacité de distinguer les zones sombres des zones éclairées, ainsi que de détecter les variations d'amplitude [10]. Il est étroitement lié au rapport signal sur bruit (SNR). Le contraste entre deux structures peut être formulé comme :

$$C = \frac{|S_A - S_B|}{S_A + S_B} \tag{1.6}$$

où S_A, S_B correspondent aux intensités des structures à différencier. Le contraste peut être amélioré à l'aide d'agents de contraste, comme l'injection de bulles d'air microscopiques, ou par un post-traitement spécifique, comme la normalisation de l'histogramme.

1.3.2.4 Artéfacts

Les artéfacts en imagerie ultrasonore apparaissent sous la forme de structures dupliquées, manquantes, mal situées ou déformées. Ce sont des conséquences directes du processus d'acquisition et de formation de l'image qui peuvent entraver l'évaluation et le diagnostic [11]. Parmi les artéfacts existants, on peut citer :

1. les artéfacts de réverbération, les zones d'ombres et les artéfacts en miroir, qui résultent de réflexions multiples;
2. les artéfacts de réfraction, dû au comportement de lentille de certains réflecteurs ;
3. les artéfacts de lobes latéraux et de champ proche, liés à l'équipement.

1.3.3 Modes d'imagerie

Trois principaux modes d'imagerie échographique sont utilisés dans l'échocardiographie clinique : le mode B, le mode M (ou mode mouvement) et le mode Doppler. Nous décrivons dans ce résumé uniquement le mode B.

Le mode B (pour "brightness", luminosité) est le mode le plus courant en routine clinique, permettant d'obtenir des visuels comme dans la Fig 1.1. Il consiste à scanner une partie du cœur à l'aide d'une succession de faisceaux ultrasonores de différentes orientations. La position d'un réflecteur sur l'image est alors estimée à partir de :

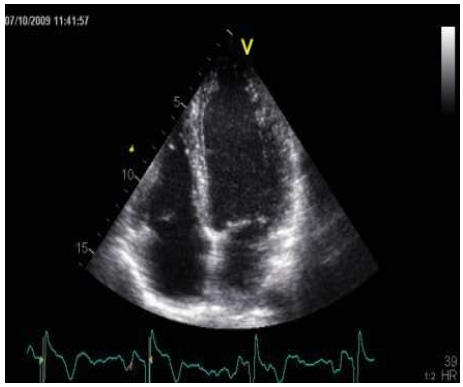


FIGURE 1.1: image B-mode du cœur.

1. la profondeur, déduite du temps de voyage ;
2. la position latérale, déduite à partir de l'orientation du faisceau.

Les images en mode B sont affichées en niveaux de gris : les réflecteurs puissants, tels que les valves ou les muscles, apparaissent brillants tandis que les structures moins échogènes, telles que les cavités remplies de sang, apparaissent sombres. Dans notre étude, nous travaillons exclusivement avec des images en mode B.

1.3.4 Analyse de la fonction cardiaque

L'imagerie médicale permet l'évaluation non invasive d'un ensemble d'indices complémentaires de la fonction cardiaque. L'analyse de ces indices détermine le risque de maladie et mène à la prise en charge et au traitement du patient [12]. Cette section du résumé présente les indices cliniques utilisés en échocardiographie et étudiés dans cette thèse.

1.3.4.1 Anatomie et cycle du cœur

Le cœur agit comme une pompe pour le système cardiovasculaire chargé de l'approvisionnement en oxygène et nutriments de l'organisme. Il est symmétriquement divisé par une paroi, appelée septum ventriculaire, qui forme une partie du myocarde, le muscle cardiaque. La couche interne du myocarde est appelée endocarde tandis que la couche externe est l'épicarde. Quatre chambres équipées de valves contrôlent le flux sanguin :

- les ventricules droit et gauche, respectivement responsables de l'envoi du sang pauvre en oxygène aux poumons et du sang oxygéné à l'ensemble du corps;
- les oreillettes droite et gauche, qui récupèrent le sang à transférer aux ventricules, respectivement venant de tout le corps et des poumons.

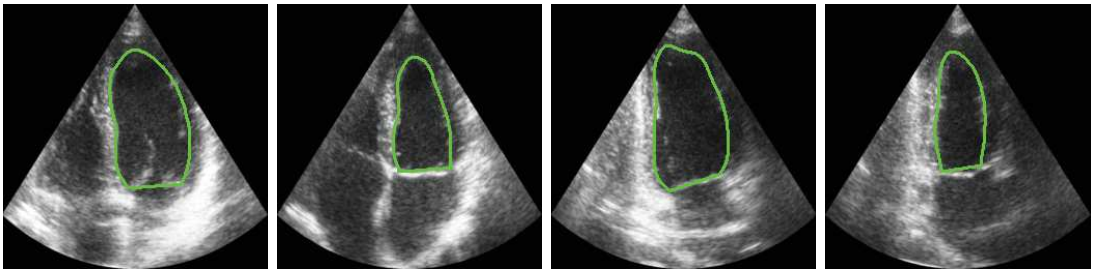
Le cycle cardiaque est composé de deux phases principales : une phase de contraction permettant d'envoyer le sang ventriculaire via les artères (systole), et une phase de relaxation dédiée au remplissage des ventricules (diastole). Ces deux événements sont déclenchés par des impulsions électriques envoyés par le cerveau.

1.3.4.2 Indices globaux

La détermination des volumes ventriculaires en fin diastole (ED) et en fin systole (ES) permet de calculer les fractions d'éjection des ventricules, soit le pourcentage du volume sanguin éjecté. Bien que l'imagerie par résonance magnétique cardiaque (CMR) demeure la référence en matière d'évaluation précise de ces indices, l'échocardiographie est la modalité la plus utilisée en raison de son faible coût, de son applicabilité au chevet du patient, et de son excellente résolution temporelle (temps réel).

L'échocardiographie 3D permet une visualisation de l'ensemble du cœur, cependant les résolutions spatiale et temporelle, inférieures à celle de l'échocardiographie 2D, limitent encore son utilisation en routine clinique. En échocardiographie 2D, l'estimation des volumes ventriculaires repose sur deux vues apicales orthogonales, la vue 4 chambre (A4C) et la vue 2 chambre (A2C), sur lesquelles la délimitation précise des structures est requise à ED et ES. Un exemple d'annotation d'expert sur les quatre trames correspondantes est donné en Fig 1.2.

Les contours endocardiques sont utilisés pour estimer les volumes du ventricule gauche avec la formule biplan de Simpson [13]. Cette méthode estime le volume total en additionnant les



(a) A4C-ED (b) A4C-ES (c) A2C-ED (d) A2C-ES

FIGURE 1.2: Vues apicales utilisées pour estimer la fraction d'éjection avec la méthode Biplane de Simpson, illustrées sur le patient 206 de l'ensemble de données CAMUS.

surfaces A_i de disques elliptiques de hauteur h dont la largeur et la hauteur sont estimées à partir des contours A2C et A4C.

$$V = \sum_{i=0}^{n} A_i \times h \tag{1.7}$$

La fraction d'éjection du ventricule gauche est directement obtenue à partir des volumes LV_{EDV} et LV_{ESV} :

$$LV_{EF}(\%) = \frac{LV_{EDV} - LV_{ESV}}{LV_{EDV}} \times 100 \tag{1.8}$$

1.3.4.3 Pratique et besoins cliniques

En raison du manque de précision et de reproductibilité des méthodes de segmentation entièrement automatiques, l'annotation semi-automatique ou manuelle des images échocardiographiques 2D pour réaliser l'estimation des indices cardiaques demeure un travail quotidien. Cela conduit à des actions cliniques chronophages sujettes à des variabilités intra- et inter-observateurs [14]. Les difficultés inhérentes à la segmentation des images échocardiographiques sont les suivantes :

1. le faible contraste entre les tissus cardiaques et le sang, ainsi que les inhomogénéités d'intensité ;
2. les variations du speckle le long du myocarde, dues à l'orientation de la sonde cardiaque par rapport au tissu et à la présence des trabécules et des muscles papillaires, dont l'intensité est similaire à celle du myocarde ;
3. les différences d'échogénicité, forme, intensité et mouvement des tissus selon les patients et les pathologies ;
4. les sorties de plan des structures d'intérêt.

Toute méthode de segmentation entièrement ou partiellement automatique doit pouvoir surmonter ces obstacles. Afin de situer les contributions de ce travail, nous passons en revue dans la partie suivante les méthodes proposées pour la segmentation échocardiographique, les ensembles de données disponibles et les métriques utilisées dans la littérature.

1.4 État de l'art

L'état de l'art correspondant à cette étude englobe les différentes méthodes de segmentation appliquées en échocardiographie. Leur performance est traditionnellement évaluée à travers un ensemble de métriques géométriques bien établies, mais sur des ensembles de données distincts, ce qui rend la comparaison difficile. Récemment, le développement de "challenges" internationaux a permis de comparer les algorithmes participants sur des bases de données publiques, et donc de réaliser une comparaison objective de la performance des méthodes de segmentation. Dans cette section, nous passons en revue la littérature en relevant :

1. les métriques géométriques communes utilisées pour l'évaluation de la segmentation en imagerie médicale ;
2. les ensembles de données publiques existants en échocardiographie (dans ce résumé) ;
3. les méthodes les plus modernes de segmentation d' images échocardiographiques.

1.4.1 Métriques de segmentation en imagerie médicale

L'évaluation en imagerie médicale repose sur l'établissement d'ensembles de critères de qualité spécifiques à la tâche à effectuer afin de fournir des scores sur le degré d'accord entre les résultats produits et le résultat attendu, dit "vérité terrain". En apprentissage supervisé, cela correspond aux annotations fournies par des experts [15]. Nous présentons ici les métriques classiques utilisées dans le cas d'une segmentation binaire, soit la séparation d'une structure (étiquetée 1) du fond de l'image (étiqueté 0).

1.4.1.1 Chevauchement de régions

Les métriques de chevauchement donnent des scores sur la qualité globale de la segmentation de l'image. L'indice Dice et l'indice Jaccard JAC (aussi appelé intersection sur union IOU) sont fréquemment utilisés dans la littérature :

$$D(A,B) = 2 * \frac{A \cap B}{|A| + |B|} = 2 \times \frac{JAC}{1 + JAC} \tag{1.9}$$

où A est la structure segmentée par la méthode et B la vérité terrain.

Dans notre étude, nous utilisons le Dice (meilleur score de 1) ou le Dice inversé $1 - D$ (meilleur score de 0) pour représenter la performance globale des méthodes de segmentation.

1.4.1.2 Distances spatiales entre les contours

Les métriques de distance spatiale permettent une meilleure évaluation de la précision du contour. La distance absolue moyenne (MAD) représente l'erreur moyenne de segmentation, tandis que la distance de Hausdorff (HD) indique l'erreur maximale [15]. Soit $d_{C_1}(C_2)$ l'ensemble des distances euclidiennes obtenues par projection perpendiculaire des points du contour C_2 sur le contour C_1. Pour assurer un comportement symétrique, nous utilisons les formulations correspondantes des distances MAD et HD :

$$MAD(C_A, C_B) = \frac{\overline{d_{C_A}(C_B)} + \overline{d_{C_B}(C_A)}}{2} \tag{1.10}$$

$$HD(C_A, C_B) = \max\left(\max d_{C_A}(C_B), \max d_{C_B}(C_A)\right) \tag{1.11}$$

où C_A et C_B sont les ensembles de points de chaque objet, d la distance euclidienne, et $\bar{\bullet}$ l'opérateur moyen.

L'association du Dice et de la distance de Hausdorff permet d'évaluer les précisions globale et locale. Dans cette étude, nous indiquons également la distance absolue moyenne pour représenter la proximité globale des contours.

1.4.2 Bases de données échocardiographiques de référence en libre accès

Plusieurs bases de données de segmentation cardiaque annotées par des experts ont servi de support d'étude. La plupart des bases de données publiques d'accès ont été mises en ligne dans le cadre de compétitions internationales et permis aux organisateurs de comparer les méthodes de l'état de l'art.

Afin de souligner le manque de données publiques d'accès en imagerie échocardiographique, nous présentons dans le manuscrit complet les différentes initiatives entreprises en imagerie par résonance magnétique (IRM) et en tomodensitométrie (CT), mais nous concentrons ici sur les ensembles de données échocardiographiques.

1.4.2.1 Échocardiographie 3D

A notre connaissance, il n'existe qu'un seul ensemble de données échocardiographiques publique, publié en 2014 [16] et utilisé lors du challenge CETUS (Challenge on Endocardial Three-dimensional Ultrasound Segmentation). Cette étude se concentrait sur une population de 45 sujets répartis dans trois centres différents. Pour chaque individu, des acquisitions 3D ont été effectuées sur l'ensemble du cycle cardiaque et le ventricule gauche a été annoté à ED et ES.

Plus récemment, (Dong et al., 2018) [17] ont construit une base de 60 patients avec le ventricule gauche segmenté à ED et ES, mais n'ont pas donné accès aux données.

1.4.2.2 Échocardiographie 2D

Avant notre initiative, il n'existait aucun ensemble de données publiques de segmentation échocardiographique 2D. Parmi les études réalisées, les plus proches de notre initiative sont :

1. le travail de (Carneiro et al., 2012) [18], qui ont utilisé un ensemble de données de 12 patients (400 images avec annotations manuelles pour le LV) pour entraîner un réseau de neurone, en gardant 2 patients (40 images) pour l'évaluation ;

2. les expérimentations dans (Smistad et al., 2017) [19], dans lesquelles les auteurs ont construit un ensemble de données de 100 000 images à partir des vues apicales de 100 patients. Cependant les annotations de référence ont été obtenues avec un algorithme automatique;

3. les travaux récents de (Azarmehr et al., 2019) [20], qui ont étudié la performance de U-Net sur un ensemble de données de 61 patients (992 images A4CH) avec le LV annoté par deux experts.

Aucun de ces trois ensembles de données n'a été rendu public, ce qui signifie que l'analyse de la segmentation en échocardiographie 2D n'a jamais fait l'objet d'une étude approfondie au moyen d'un large ensemble de données publique d'accès.

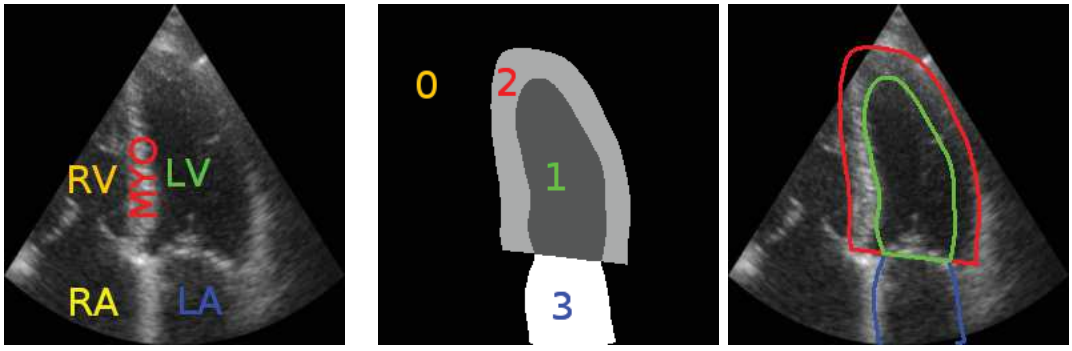
1.4.3 La segmentation sémantique en échocardiographie

La segmentation est une étape nécessaire à l'estimation des indices cliniques en échocardiographie. Comme mentionné précédemment, les images échographiques ont des caractéristiques inhérentes (contraste, texture, artéfacts...) qui entravent le processus de segmentation et expliquent la prévalence actuelle des méthodes semi-automatiques en routine clinique.

1.4.3.1 Définition de la segmentation sémantique

La segmentation d'image est la tâche consistant à partitionner une image en objets d'intérêt, soit en traçant des contours, comme sur la Fig. 4.3 c), soit en classifiant chaque pixel, comme dans la Fig. 4.3 b). La segmentation sémantique implique d'identifier les objets par leur nature au moyen d'une étiquette numérique, aussi appelée classe. En cas d'occurrences multiples d'un même type d'objet, l'étiquetage peut être adapté pour séparer les éléments distincts d'un même type (segmentation d'instances multiples).

Dans cette thèse, l'identification et la délimitation des structures cardiaques sont exprimées comme une tâche de segmentation sémantique, car nous assignons des étiquettes uniques aux différentes cavités et au myocarde (Tab. 1.1). La segmentation peut être vue comme un mappage $X \rightarrow Y$, où X est l'image ultrasonore comme dans la Fig. 4.3 a) et Y un masque de segmentation comme dans la Fig. 4.3 b). Comme nous ne recherchons pas nécessairement une annotation détaillée de toutes les structures visibles, tout pixel qui n'appartient pas à une structure d'intérêt (ici le ventricule gauche LV, le myocarde myo et l'oreillette gauche LA) est par défaut associé à la classe "fond".



(a) Image échocardiographique 2D avec structures indiquées (b) Masque de la vérité terrain (c) Contours de la vérité terrain

FIGURE 1.3: La segmentation multi-structure, considérée comme un problème de classification multi-classes. Les algorithmes d'apprentissage supervisé sont entraînés à prédire b) à partir de a). En guise de visuel, on affiche traditionnellement les contours sur l'image comme en c).

TABLE 1.1: Étiquettes identifiant les structures cardiaques dans ce projet

Structure	ventricule gauche	myocarde	oreillette gauche	autre (fond)
Étiquette	1	2	3	0

1.4.3.2 Vue d'ensemble des méthodes de segmentation en échocardiographie

De nombreuses revues listent les méthodes de segmentation existantes en échocardiographie 2D [18], en échocardiographie 3D [21]. [16] ou les deux [3] [22]. La plupart des travaux portent exclusivement sur la détection de l'endocarde. Comme expliquée dans l'étude de (Noble et al., 2006) [3] et confirmé dans (Kong et al., 2012) [22], la faible qualité d'image de l'imagerie ultrasonore comparée aux autres modalités a incité la communauté à proposer des méthodes spécifiques à ce type d'images.

Les rapports décrivent six grandes catégories de techniques de segmentation en imagerie ultrasonore. Leurs caractéristiques, énumérées dans la section 1.2, comprennent :

1. le formalisme : la segmentation est soit basée sur la détection de transitions d'intensité dans l'image associées à des frontières anatomiques (approche contour), soit sur le regroupement de distributions spécifiques de pixels/voxels basé sur des critères de similarité (approche région), soit un mix des deux ;
2. l'utilisation de connaissances a priori : des informations pré-établies sur la forme, l'emplacement ou la texture des régions anatomiques sont ajoutées dans le but de contraindre la segmentation ;
3. la cohérence temporelle : une évolution cohérente des contours à travers les séquences temporelles est obtenue par du suivi, des contraintes spatio-temporelles, ou un lissage en post-traitement ;
4. l'apprentissage supervisé : l'optimisation des paramètres d'un modèle pré-établi est guidée par un ensemble de cas résolus.

1.4.3.3 Méthodes non supervisées

Les méthodes non supervisées ne nécessitent pas de données d'entraînement et incorporent plutôt des connaissances a priori sous forme d'initialisation, de contraintes de forme ou de pré et post-traitement pour guider la segmentation. Parmis elles, les modèles déformables (serpents, level sets, canevas déformables) ont été prédominants en échocardiographie par rapport à d'autres techniques telles que les méthodes de regroupement (k-means, lignes de

Table 1.2: Caractérisation des différents types de méthodes de segmentation en échocardiographie, inspirée de la revue par (Carneiro et al., 2012) [18]

Méthode	Contour	Région	Information a priori	Cohérence temporelle	Super-vision
Bas en haut	✓	✓		✻	
Contours actifs	✓	✻	✻	✻	
Level Sets	✓	✓	✻	✻	
Modèles déformables	✓	✓	✓	✻	
Modèles de forme	✓	✻	✻	✻	✓
Apprentissage automatique / profond	✻	✓	✻	✻	✓

✓: *Propriété inhérente au formalisme originel*
✻: *Propriété ajoutée au formalisme originel*

partage des eaux) ou d'accroissement de région, et les approches probabilistes (coupes de graphe, champs aléatoires markoviens). Plus de détails sur l'application de méthodes non supervisées à l'échocardiographie sont donnés dans le manuscript anglais.

1.4.3.4 Modèles supervisés

Les modèles appris par apprentissage supervisées sont entraînés à reproduire une forme d'expertise en optimisant leurs paramètres sur des cas résolus. Des modèles de forme actifs ont été utilisés pour la segmentation échocardiographique depuis les années 1990, et intégrés au sein de pipelines semi-automatiques, c'est-à-dire nécessitant une saisie manuelle et/ou une adaptation à chaque image. Pour les solutions entièrement automatiques, la tendance actuelle favorise les algorithmes d'apprentissage automatique et d'apprentissage profond.

Dans ce résumé, nous proposons la description des différentes méthodes. Le détail des performances et des applications en échocardiographie est donné dans le manuscript complet.

Modèles de forme actifs Les modèles de forme actifs (ASM) [6] définissent un espace d'évolution pour les points de contour de la segmentation. Le prototype de forme et les déformations autorisées sont établies à partir de cas résolus. Tout d'abord, toutes les formes sont recalées entre elles à partir de repères annotés manuellement. L'espace de forme est ensuite construit autour de la forme moyenne comme un modèle de variations statistiques, dans lequel les déformations suivent des distributions gaussiennes.

Les principaux modes de variation peuvent être déduits en appliquant une analyse en composantes principales (PCA), et des limites peuvent être posées sur l'ampleur des déformations autorisées depuis la forme moyenne. Pour une nouvelle image, une fois recalée dans l'espace de forme, les caractéristiques de l'image sont utilisées pour produire une segmentation plausible par rapport à la base d'apprentissage.

Modèles d'apparence actif Afin de guider l'ajustement itératif de la forme à l'image, [23] a proposé dans les modèles d'apparence actifs (AAM) d'associer un modèle d'intensité au modèle de forme. La distribution des intensités est également définie selon des modes de variation gaussiens. Cette approche s'est avérée particulièrement efficace en imagerie ultrasonore où les distributions d'intensité spécifiques ont pu être modélisés pour améliorer la précision de la segmentation [3].

Forêts aléatoires Les forêts de décision aléatoires (RF) consistent en un ensemble d'arbres de décision entraînés sur des sous-ensembles aléatoires de données annotées afin d'éviter le sur-apprentissage. Les arbres sont construits comme des successions de décisions heuristiques binaires sur des caractéristiques pré-établies par le développeur. A chaque intersection, ou nœud, les données sont dirigées vers une des deux branches sortantes de manière à optimiser le gain d'information.

Le modèle des forêts aléatoires est flexible et générique : les extrémités, appelées feuilles, peuvent stocker tout type et toute quantité d'informations sur la tâche à accomplir, souvent une tâche de classification ou de régression. De plus, la phase de test est très rapide par calcul parallèle car chaque arbre est indépendant et n'effectue qu'une poignée d'opérations de seuillage afin de fournir une solution. Étant donné que la segmentation multi-structure peut être vue comme un problème de classification multi-classes, les RF sont une solution intéressante pour la segmentation d'images médicales [24].

Réseaux de neurones Les réseaux de neurones artificiels réalisent un mappage au moyen d'un ensemble de couches performant chacune une projection dans un espace intermédaire. Ces couches sont habituellement plus de 2 (apprentissage profond). Les réseaux multi-couches sont entraînés itérativement par rétro-propagation de l'erreur [25]. En multipliant les couches, les réseaux de neurones sont capables de capturer des mappages complexes et directement construits à partir de l'image entière, ou d'une partie. Nous distinguons ici les réseaux de neurones entièrement connectés des réseaux de neurones convolutifs.

Dans les perceptrons multi-couches (MLP), également appelés réseaux de neurones entièrement connectés, chaque couche est composée de plusieurs perceptrons, soit d'une unité dont la sortie est le résultat de l'application d'une fonction non linéaire (activation) à la somme pondérée de toutes les entrées.

Les réseaux de neurones convolutionnels (CNN) se sont établis comme des méthodes de pointe en traitement d'images, y compris en imagerie médicale [26]. Les modèles sont composés de couches convolutives appliquant un filtrage local de l' entrée, et comportent souvent des couches supplémentaires de régularisation et de normalisation. Un de leurs avantages réside dans la possibilité de ne stocker que les paramètres des filtres, ce qui réduit l'utilisation de la mémoire et permet d'augmenter la taille de l'entrée par rapport aux MLP.

1.4.4 Challenge CETUS

La compétition internationale CETUS sur la segmentation du ventricule gauche en 3D à partir d'images ultrasonores a montré que des méthodes entièrement automatiques pouvaient donner de meilleurs résultats que des approaches semi-automatiques. Cependant, aucune méthode n'a atteint la variabilité inter-observateur, particulièrement faible en raison du consensus entre les experts de référence. La communauté continue de proposer et d'appliquer de nouvelles méthodes sur CETUS, y compris des méthodes par apprentissage.

L'une des principales limites de CETUS concerne la validation des méthodes. Premièrement, aucune validation croisée n'est effectuée dans le cadre d'une compétition, alors que cela pourrait limiter le biais entraîné par le petit nombre de patients concerné. C'est particulièrement problématique lorsque les méthodes montrent des performances très proches. De plus, concernant les métriques cliniques, l'interprétation des performances à partir des biais et des limites d'accord est difficile en l'absence d'information sur l'erreur moyenne. Enfin, aucune métrique n'a évalué la plausibilité anatomique des formes prédites du ventricule. Ces observations ont fortement guidé notre propre méthodologie.

1.4.5 Conclusion

L'état de l'art de la segmentation échocardiographique n'est pas clair en raison de l'absence d'un ensemble de données publique en 2D. Cependant, la compétition CETUS a vu l'application avec succès de méthodes génériques d'apprentissage supervisé, telles que les forêts aléatoires et les réseaux de neurones convolutifs, en échocardiographie 3D.

Au cours de cette thèse, nous avons proposé des innovations sur i) les ensembles des données, puisqu'il n'existait pas d'ensemble de données publiques en 2D; ii) les modèles supervisés, puisqu'aucune méthode n'avait démontré la fiabilité géométrique et clinique de ses prédictions par rapport à la performance humaine; iii) les métriques, car trop peu d'étude impliquent l'évaluation de la plausibilité des formes générées en imagerie médicale.

1.5 La base de données CAMUS

Comme décrit dans la section 4.3, le problème de segmentation en échocardiographie 2D n'a jamais été étudié dans la littérature à partir d'un large ensemble de données publique. Dans ce contexte, nous décrivons ici le plus grand ensemble de données annotées libre d'accès pour l'évaluation de la segmentation échocardiographique 2D. Le but de cette base de données clinique est de :

- permettre d'entraîner correctement des modèles d'apprentissage supervisé ;
- permettre une comparaison directe entre les méthodes de pointe ;
- évaluer jusqu'où les méthodes d'apprentissage supervisé peuvent aller dans l'analyse d' images échocardiographiques 2D, soit la segmentation des structures cardiaques ainsi que l'estimation des indices cliniques.

CAMUS signifie "Cardiac Acquisitions for Multi-structure Ultrasound Segmentation", et est également le nom d'un écrivain et révolutionnaire français du 20ème siècle. Cet ensemble de données est disponible au téléchargement sur [1], où se trouvent également une plate-forme commune d'évaluation.

1.5.1 Propriétés

1.5.1.1 Population

L'ensemble de données proposé se compose des séquences apicales deux chambre (A2C) et quatre chambre (A4C) d'au moins un cycle cardiaque complet pour 500 patients. Leur inclusion dans l'étude a été conforme au règlement établi par le comité d'éthique local de l'Hôpital Universitaire de Saint-Étienne (France). Pour renforcer le réalisme clinique, aucune sélection des données n'a été effectuée selon aucune condition préalable. Il en résulte un ensemble de données très hétérogène, tant en termes de qualité d'image que de pathologie, ce qui correspond à la situation classique en routine clinique.

Étant donné que la fraction d'éjection pour des patients en bonne santé se situe normalement autour de 64 $\pm 6,5$ % [27], la plupart des patients de la base de données CAMUS ($> 80\%$) sont susceptibles de présenter une condition cardiaque pathologique au regard de cet indice. De plus amples renseignements sur la population de l'étude se trouvent dans l'annexe D.

1.5.1.2 Acquisition

Les acquisitions ont été réalisées au CHU de Saint-Étienne au moyen d'échographes GE Vivid E95 équipés de sondes GE M5S et du logiciel d'analyse EchoPAC. Les examens cliniques ont été optimisés pour effectuer la mesure de la fraction d'éjection du ventricule gauche (LV_{EF}) à partir des vues apicales deux et quatre chambre (A2C, A4C), comme décrit en section 3.4. La base de données comporte une grande variabilité concernant les paramètres d'acquisition (positionnement et orientation de la sonde), en particulier :

- pour certains patients, des parties du myocarde n'étaient pas visibles sur les images (hors secteur) ;
- dans certains cas, la recommandation sur l'orientation de la sonde pour acquérir une vue quatre chambre était tout simplement impossible à suivre et une vue cinq chambre a été acquise.

Table 1.3: Principales caractéristiques de la base de données CAMUS (500 patients)

Ensemble	**Qualité d'image** %			LV_{EF} %		
	Bonne	Intermédiaire	Faible	$\leq 45\%$	$\geq 55\%$	Entre
Full	35	46	19	49	19	32
pli 1	34	48	18	48	20	32
pli 2	34	46	20	50	18	32
pli 3	34	46	20	48	20	32
pli 4	34	46	20	50	20	30
pli 5	34	46	20	48	20	32
pli 6	36	46	18	50	20	30
pli 7	36	46	18	50	20	30
pli 8	36	46	18	50	18	32
pli 9	36	46	18	48	20	32
pli 10	36	46	18	50	18	32

1.5.1.3 Qualité d'image

Les séquences exportées du système GE correspondent à des séquences d'images mode B exprimées en coordonnées polaires. La même procédure d'interpolation a été utilisée pour exprimer toutes les images en coordonnées cartésiennes avec une résolution de grille unique, soit $\lambda/2 = 0.3$ mm le long de l'axe x (parallèle à la sonde) et $\lambda/4 = 0.15$ mm sur l'axe z (perpendiculaire à la sonde), où λ correspond à la longueur d'ondes du capteur ultrason.

La qualité de l'image a été évaluée subjectivement par l'observateur O_1, révélant 19% de patients avec au moins une séquence de mauvaise qualité d'image (Tab. 1.3). Pour ces patients, la localisation de l'endocarde (LV_{Endo}) et de l'épicarde (LV_{Epi}) et les indices cliniques dérivés ne sont pas considérés précis.

En analyse classique, les images de mauvaise qualité sont écartées en raison de leur inutilité en clinique. Dans notre étude, nous avons conservé toutes les images et adapté l'évaluation en fonction de la qualité de l'image.

1.5.1.4 Partitionnement

Les principales informations qui caractérisent la base CAMUS sont présentées dans le tableau 1.3. En s'appuyant sur la qualité d'image et la fraction d'éjection, l'ensemble des données a été divisé en 10 plis équilibrés.

Chaque pli contient 50 patients et les mêmes distributions en termes de qualité d'image et LV_{EF} que l'ensemble. Ce partitionnement a été utilisé pour construire des ensembles d'entraînement et de tests pour les méthodes d'apprentissage supervisé, en conservant 8 plis pour l'entraînement, 1 pour la validation (pour les méthodes d'apprentissage profond), et le dernier pour le test. Les 10 rotations possibles ont été utilisées pour effectuer une validation croisée classique des méthodes évaluées.

1.5.2 Annotations

Les propriétés de généralisation des algorithmes supervisés sont liées à la reproductibilité des annotations des experts, notamment à la cohérence et à la précision des contours établis manuellement. L'établissement d'un protocole de segmentation bien défini pour la vérité terrain était donc de la plus haute importance pour ce travail.

1.5.2.1 Sélection des trames ED et ES

O_1 a sélectionné une trame de fin diastole (ED) et une trame de fin systole (ES) pour toutes les séquences, selon le protocole et les recommandations dans (Lang et al., 2015) [27]. Selon les recommandations de l'American Society of Echocardiography et de l'European Association of Cardiovascular Imaging, la trame ED est ainsi définie de préférence soit comme la première image après la fermeture de la valve mitrale, soit comme l'image dans laquelle la surface associée au ventricule gauche est la plus grande. De même, ES est définie comme l'image après la fermeture de la valve aortique, ou comme l'image dans laquelle la surface du ventricule est minimale.

Dans ce travail, les trames ED et ES ont été identifiées comme les images où la taille du ventricule gauche était la plus grande / la plus petite en raison du manque d'ECG fiable pour plusieurs patients. Comme cette approche simpliste n'est pas la plus précise, surtout en présence d'anomalies, les valeurs des indices cliniques que nous rapportons doivent être interprétées avec cet aspect à l'esprit.

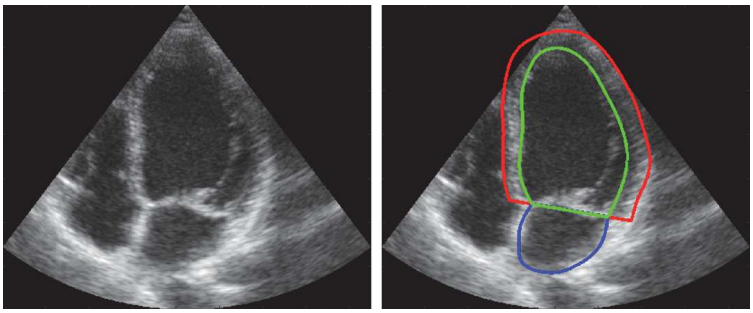
1.5.2.2 Protocole de segmentation

Les acquisitions de la base de données CAMUS sont optimisées pour le calcul de la fraction d'éjection et se concentrent donc sur une bonne visibilité du ventricule gauche, mais le myocarde et l'oreillette gauche sont également au moins partiellement visibles (Fig. 1.4).

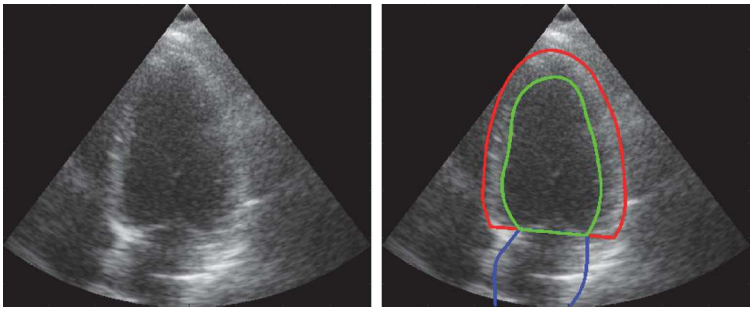
Ces structures ont également été annotées, principalement pour étudier l'influence de la contextualisation sur la performance des techniques d'apprentissage supervisé, et ont été contourées comme étant adjacentes au ventricule gauche. Elles commencent ainsi à partir des mêmes points de jonction, situés entre les valves et le muscle (Fig. 1.4).

Le protocole suivant a été mis en place :

- LV_{Endo} : le protocole décrit dans (Lang et al., 2015) [27] a été appliqué pour la paroi interne du LV, le plan de la valve mitrale, les trabeculations, les muscles papillaires et le sommet. Cela impliquait en particulier de :
  1. terminer les contours dans le plan de la valve mitrale ;
  2. exclure partiellement l'écoulement ventriculaire gauche en reliant l'articulation de la valve mitrale septale à la paroi septale de façon à créer une forme lisse ;
  3. inclure les trabécules et les muscles papillaires ;
  4. maintenir la cohérence temporelle des tissus entre les instants ED et ES.
- LV_{Epi} : Il n'y a pas de recommandation pour délimiter l'épicarde. Nous avons donc esquissé l'épicarde pour qu' :
  1. il soit dessiné comme l'interface entre le péricarde et le myocarde pour les segments antérieur (sur A2C), antérolatéral (sur A4C) et inférieur, et pour la frontière entre la cavité ventriculaire droite et le septum pour le segment inferoseptal ;



(a) Image de bonne qualité



(b) Image de moyenne qualité

(c) Image de faible qualité

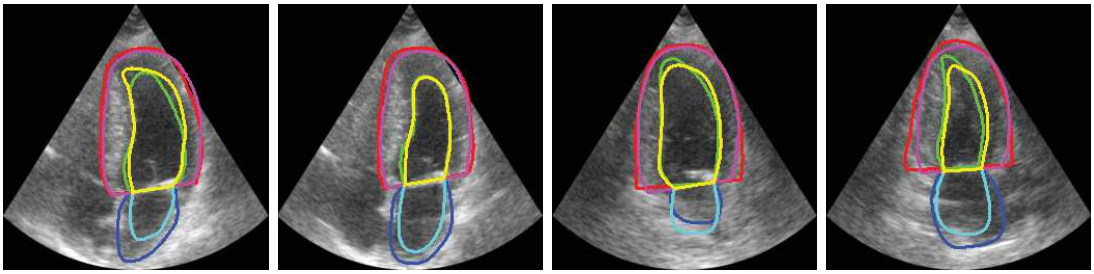
FIGURE 1.4: Images tirées de la base de données CAMUS. L'endocarde, l'épicarde et l'oreillette gauche sont respectivement représentés en vert, rouge et bleu. (Gauche) images d'entrée, (Droite) annotations manuelles.

2. en cas d'information localement manquante, les contours soient dessinés en gardant la forme et l'épaisseur continus;
3. la base s'aligne avec la base du ventricule gauche, en suivant une ligne droite.

- oreillette gauche (LA) : il est recommandé pour la segmentation de l'oreillette d'utiliser des acquisitions dédiées. Étant donné que nous avons utilisé des acquisitions axées sur le ventricule gauche, une partie de la base ne couvre pas toute la surface de la LA et n'est donc pas adaptée pour effectuer des mesures sur celle-ci. Nous avons donc décidé d'utiliser le protocole de contourage suivant :

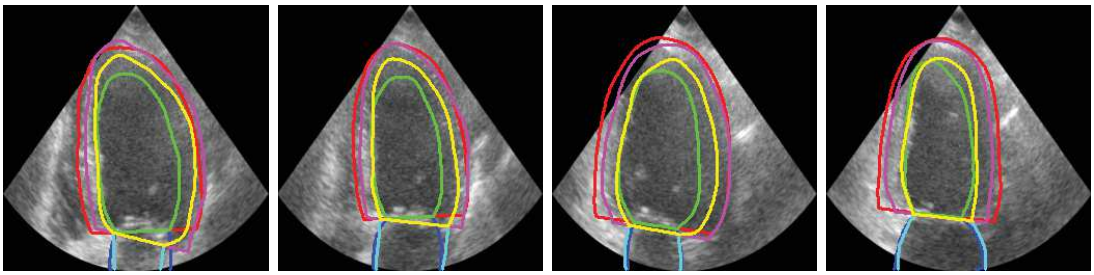
  1. commencer le contour de l'oreillette à partir des extrémités du contour du ventricule gauche, aux points d'articulation des valves ;
  2. le contour passe ensuite par la paroi intérieure de l'oreillette;
  3. le contour s'arrête si nécessaire sur le bord de l'image (Fig. 5.1 b.).

P-252: $O1a$ (rgb) VS $O1b$ (myc)



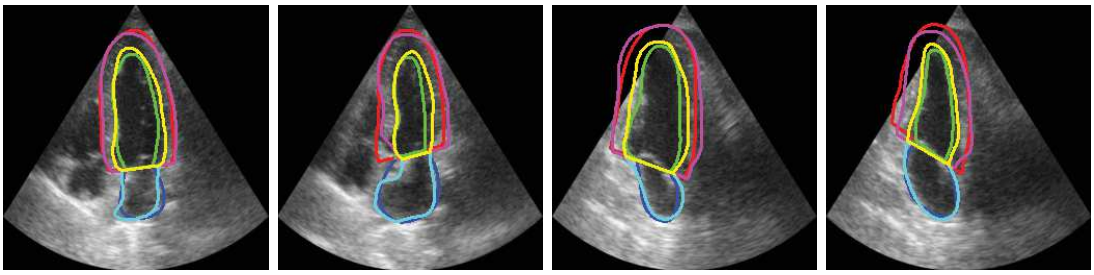
(a) A4CH-ED (b) A4CH-ES (c) A2CH-ED (d) A2CH-ES

P-266: $O1a$ (rgb) VS $O2$ (myc)



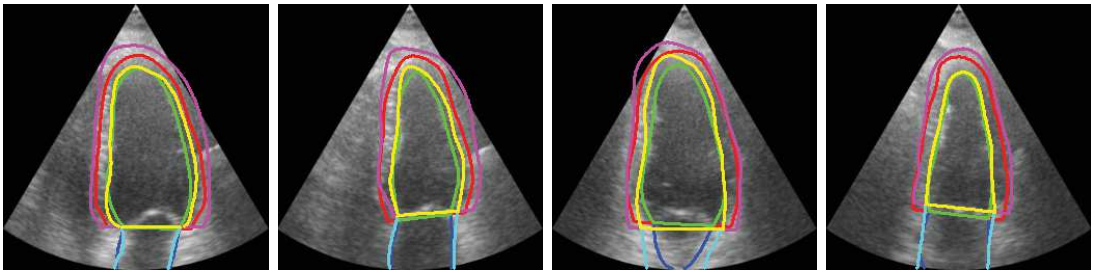
(e) A4CH-ED (f) A4CH-ES (g) A2CH-ED (h) A2CH-ES

P-242: $O1a$ VS $O3$ (myc)



(i) A4CH-ED (j) A4CH-ES (k) A2CH-ED (l) A2CH-ES

P-260: $O2$ (rgb) VS $O3$ (myc)



(m) A4CH-ED (n) ACH-ES (o) A2CH-ED (p) A2CH-ES

FIGURE 1.5: Exemples de CAMUS illustrant la variabilité moyenne entre les experts pour la distance absolue moyenne (MAD). Une série d'annotations est en RGB, et l'autre en MYC. L'image intitulée en rouge représente pour chaque paire d'annotations la variabilité moyenne sur l'endocarde.
Première rangée : intra-variabilité. Trois dernières lignes : inter variabilité.

1.5.2.3 Inter- and intra- variability

Trois cardiologues ($O1$, $O2$ et $O3$) ont participé à l'annotation de l'ensemble de données. Des efforts considérables ont été déployés pour définir avec O_1 un protocole de segmentation manuelle cohérent. Nous avons ensuite demandé à $O1$ d'effectuer l'annotation manuelle pour l'ensemble des données en dehors du cadre de l'examen cardiaque, tandis que les deux autres cardiologues ont contourné le pli 5 (50 patients). $O1$ a annoté ce pli une seconde fois sept mois plus tard (ses annotations sont appelées $O1a$ et $O1b$, respectivement). Ce pli est donc utilisé dans cette étude pour mesurer à la fois la variabilité inter-observateurs et la variabilité intra-observateur.

Des exemples sont présentés dans la Fig. 5.2. Les scores géométriques et cliniques peuvent être trouvés dans la section 7.3.

1.5.3 Conclusion

Nous avons construit un ensemble de données spécifique à l'entraînement, l'évaluation et la comparaison des méthodes de segmentation en échocardiographie 2D. La compétition ouverte que nous proposons au travers de cet ensemble de données publique est de concevoir un algorithme capable de produire des résultats de segmentation plus précis que les variabilité inter- et intra- experts.

A l'exception de l'étude sur l'échocardiographie 3D présentée dans la section 6.4.2.1, toutes nos contributions algorithmiques ont été conçues et testées sur la base de données CAMUS.

1.6 Adapter le formalisme des forêts aléatoires structurées pour la segmentation sémantique d'images échocardiographiques

La section suivante décrit l'algorithme des forêts aléatoires structurées -Structured Random Forests (SRF)- à travers ses propriétés et son potentiel pour la segmentation automatique du cœur en imagerie ultrasonore. Elle répond aux questions suivantes :

1. Comment le formalisme des SRF peut-il être adapté aussi bien pour prédire des probabilités de bord ou des étiquettes associées aux structures, et être intégré dans des pipelines de segmentation ?

2. Comment pouvons-nous intégrer de l' information contextuelle dans l'espace de caractéristiques des SRF ?

3. Peut-on encourager de la plausiblité anatomique concernant les prédictions des SRF ?

1.6.1 Forêts aléatoires

Les forêts aléatoires -"Random Forests" (RF)- impliquent l'apprentissage d'arbres de décision qui performent une succession d'heuristiques sur des caractéristiques pré-établies. Les RF peuvent en particulier être entraînées à prédire à quelle structure ou à quel objet appartient un pixel (classification). Le mappage correspondant peut s'écrire $P \rightarrow X \rightarrow Y$, où P est l'ensemble de pixels à classer, à partir duquel on extrait un ensemble de caractéristiques discriminantes X. La forêt apprend ensuite le mapping $X \rightarrow Y$ de manière à affecter automatiquement à chaque pixel $p \in P$ une étiquette $y \in Y$ en fonction de X.

1.6.1.1 Phase d'entraînement

Chaque arbre est entraîné sur un sous-ensemble aléatoire des données $p \in P$, selon une stratégie appelée "bootstrap", et a accès à un nombre limité de caractéristiques, aussi choisies au hasard. Chaque arbre pousse en séparant à chaque nœud le jeu de données d'entrée en deux selon le gain d'information résultant, jusqu'à ce que tous les chemins soient terminés par une feuille. Une feuille est créée lorsqu'un critère d'arrêt est rempli et contient les informations qui caractérisent les données qui l'ont atteinte (ex : histogramme de classes).

1.6.1.2 Phase de prédiction

Chaque arbre peut effectuer une prédiction sur une nouvelle donnée selon $\hat{y_t} = \text{argmax}_{y \in Y}\, p_t(y/x)$, mais conserver une sortie probabiliste est souvent préférée afin de l'intégrer dans un modèle d'ensemble et bénéficier de la régularisation et du débruitage induits. Comme chaque pixel est considéré indépendamment des autres, la classification des RF a tendance à produire des segmentations bruitées, d'où la nécessité d'apprendre à prédire des résultats structurés (ex: patch de masque de segmentation) comme dans les forêts aléatoires structurées.

1.6.2 Forêts aléatoires structurées

(Kontschieder et al., 2011) [28] ont proposé de changer le mappage de $p \rightarrow X \rightarrow Y$ à $P \rightarrow X \rightarrow L$, où L correspond au masque de segmentation du patch image P, qui comprend une structure cohérente comme observable dans toute image naturelle [29].

1.6.2.1 Phase d'entraînement

Pour entraîner les fonctions de séparation des SRF, les étiquettes discrètes étant remplacées par des patchs de segmentation, il est nécessaire de définir la notion de distance entre les patchs de segmentation, pour d'une part regrouper ceux qui sont similaires et d'autre part séparer ceux qui sont différents. En d'autres termes, une correspondance intermédiaire $P \rightarrow X \rightarrow T \rightarrow L$ est nécessaire, avec T un espace de classe discret similaire à Y. A chaque feuille, l'information stockée reflète la structure cohérente des patchs qui l'ont atteinte, par exemple au travers d'un patch médoïd.

1.6.2.2 Phase de prédiction

Au moment de l'inférence, une nouvelle image de test est divisée en un ensemble de patchs se chevauchant et centrés sur les nœuds d'une grille régulière préalablement définie sur le support de l'image. Une solution simple pour le modèle d'ensemble consiste à établir pour tout pixel la classe qui a été la plus votée parmi toutes les prédictions des arbres.

1.6.3 Application des forêts aléatoires structurées à la segmentation en échocardiographie 2D

Ce travail a été publié lors de la Conférence IEEE IUS de 2017 [30].

1.6.3.1 Méthodologie

Nous avons décidé de revisiter le formalisme de (Dollár et al., 2015) afin d'effectuer ici une segmentation multi-classe sur des images ultrasonores. Pour ce faire, nous avons apporté trois contributions méthodologiques :

1. Nous avons caractérisé la notion de similitude entre des patches de segmentation multi-classes qui stockent des informations de région et non de bord ;
2. Nous avons adapté le choix du patch médoïd associé à chaque feuille, ainsi que l'intégration des prédictions des arbres dans un modèle d'ensemble ;
3. Nous avons réalisé une étude complète pour proposer d'extraire des caractéristiques à plusieurs échelles tout en limitant l'impact sur la mémoire.

1.6.3.2 Expérience

Nous possédions alors les annotations en fin diastole et fin systole des vues A4C de 250 patients. Nous avons utilisé 200 patients pour entraîner deux forêt de segmentation (ED et ES), et gardé les 50 autres patients comme base de test. Nous avons généré comme caractéristiques des histogrammes de gradient calculés selon 4 directions aux échelles 1, 3 et 5 de l'image, afin de fournir le contexte local, intermédiaire et global.

Au moment du test, nous extrayons les patchs avec une foulée de 2 et associons à chaque pixel la classe la plus votée en considérant les prédictions des arbres sur tous les patchs qui le contienne. La figure 1.6 illustre la solution à base de SRF que nous proposons.

1.6.3.3 Modèle d'apparence actif

Nous avons comparé notre solution entièrement automatique au modèle semi-automatique d'apparence actif (AAM). Cette méthode correspond à un modèle déformable dont les statistiques de formes et d'intensités le long du contour sont apprises à partir d'un ensemble de

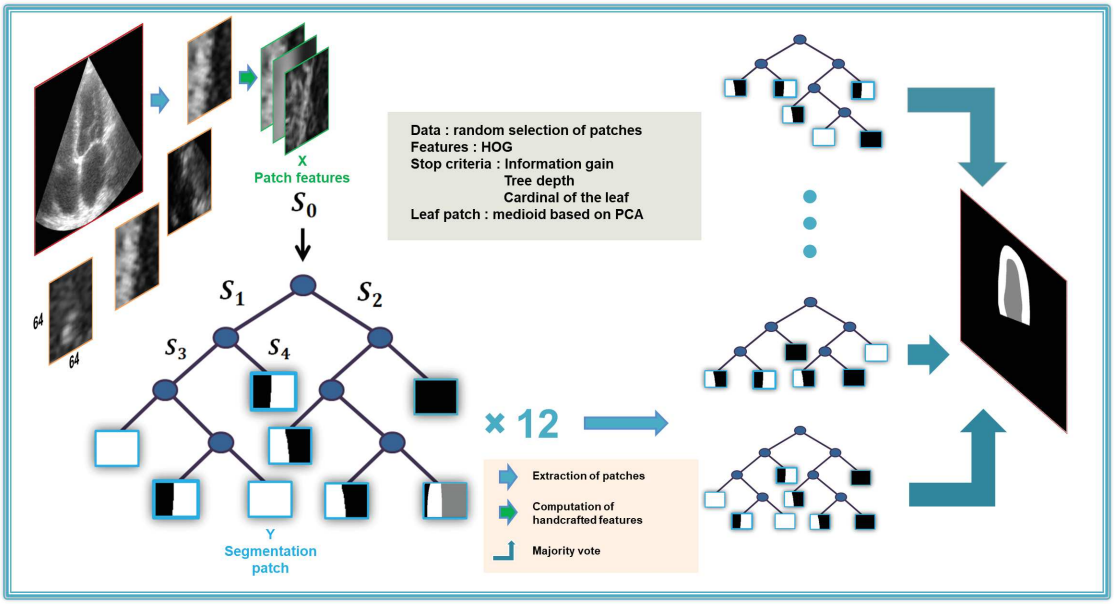


FIGURE 1.6: Notre méthode SRF [31]

données d'apprentissage.

Nous avons créé deux modèles indépendants : l'un dédié à la détection de l'endocarde, l'autre spécialisé dans la paroi épicardique. L'initialisation de cette méthode nécessite 5 points fournis manuellement par un utilisateur : 2 à la base, 1 à l'apex et 2 au septum. 4 points sont sur la bordure de l'endocarde et le dernier pour initialiser l'épaisseur du myocarde.

1.6.3.4 Résultats

A partir des résultats fournis dans les tableaux 6.3 et 6.4, nous pouvons faire les observations suivantes :

- Notre solution automatique obtient des résultats compétitifs par rapport à ceux de l'AAM semi-automatique ;
- Les SRF obtiennent cependant des résultats moins bons en terme de HD pour les deux structures;
- L'élimination des 6 valeurs aberrantes améliore significativement les scores SRFs. Cela suggère que la gestion des cas aberrants doit être une priorité.

1.6.3.5 Discussion

Afin d'améliorer la robustesse, nous pouvons soit ajouter plus de données d'entraînement dans l'espoir d'apprendre suffisamment de variabilité pour faire face à tous les cas possibles, soit ajouter des contraintes de forme en guise de régularisation.

1.6.4 Application à la segmentation d'images échocardiographiques 3D

Nous avons proposé une méthode combinant des SRFs spécialisées dans la détection des contours 3D du ventricule gauche et un modèle de forme actif (ASM) classique, que nous avons évaluée sur l'ensemble de données CETUS. Ce travail a été publié dans le cadre de la conférence IEEE SPIE 2018 [32].

1.6.4.1 Méthodologie

Nous avons adapté notre algorithme de SRF à la détection de contours dans des volumes échocardiographiques. La sortie est alors une carte de probabilité 3D où chaque valeur de voxel correspond à la probabilité d'avoir détecté la paroi endocardique. Un ASM exploite cette information tout en régularisant sa forme dans un espace pré-appris.

1.6.4.2 Pipeline

La figure 1.7 illustre le pipeline semi-automatique complet mis en place pour cette étude. Sur chaque image de test, les SRFs détectent automatiquement l'endocarde. L'ASM est ensuite initialisé à l'aide de trois repères indiqués manuellement et d'une détection automatique du plan de la valve mitrale à partir de la carte du bord. La carte de contours sert de terme d'attache aux données sur lequel l'ASM est adapté de manière itérative. Après convergence, la surface résultante est utilisée comme prédiction finale de l'endocarde.

1.6.4.3 Résultats

Scores géométriques Nous avons comparé notre méthode aux dix algorithmes du "challenge" CETUS. A partir des tableaux 1.4 et 1.5, nous pouvons établir que notre méthode a obtenu de très faibles erreurs moyennes, surpassant les BEAS et la méthode SRF de (Domingos et al., 2014) [33] sur toutes les métriques de distance.

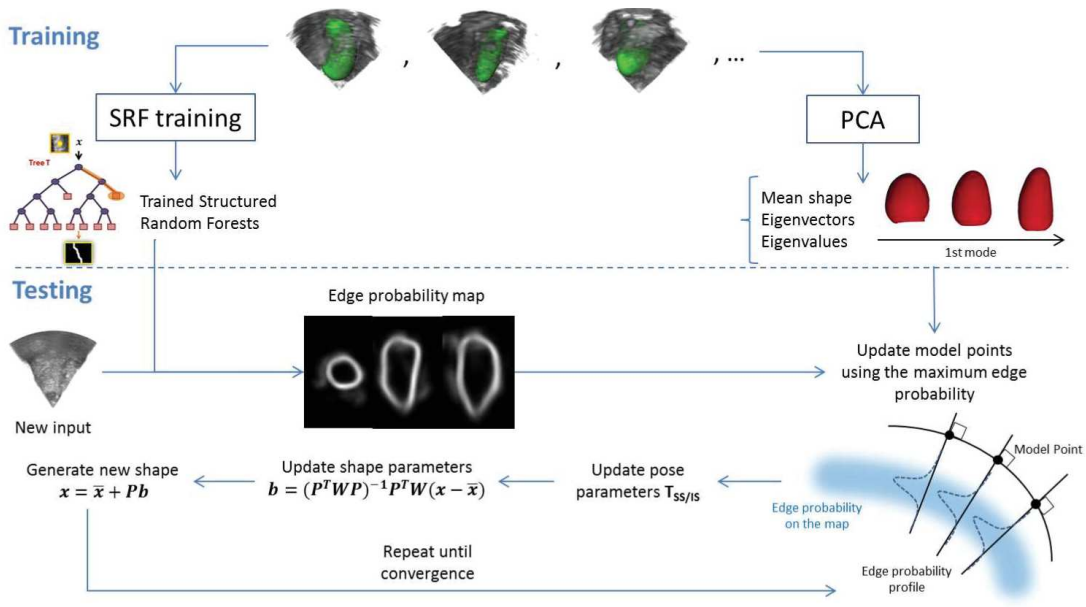


FIGURE 1.7: Vue d'ensemble de notre pipeline complet combinant SRF et ASM pour la segmentation du ventricule gauche en échocardiographie 3D.

TABLE 1.4: Scores géométriques en fin diastole

Méthodes	MAD	HD	1 - D
Inter-observateur	1.39 ± 0.40	4.70 ± 1.27	0.069 ± 0.021
SRF + ASM (30 cas d'entraînement)	$\mathbf{2.04 \pm 0.48}$	$\mathbf{6.67 \pm 1.98}$	$\mathbf{0.101 \pm 0.024}$
BEAS	2.26 ± 0.73	8.10 ± 2.66	0.106 ± 0.041
Domingos and al. [33] (15)	2.09 ± 0.68	9.31 ± 3.89	0.106 ± 0.038

TABLE 1.5: Scores géométriques en fin systole

Method	MAD	HD	1 - D
Inter-observateur	1.34 ± 0.35	4.70 ± 1.15	0.080 ± 0.021
SRF + ASM (30 cas d'entraînement)	$\mathbf{2.18 \pm 0.79}$	$\mathbf{7.76 \pm 2.47}$	$\mathbf{0.136 \pm 0.054}$
BEAS	2.43 ± 0.91	8.13 ± 3.08	0.144 ± 0.057
Domingos and al. [33] (15)	2.20 ± 0.72	8.35 ± 2.67	0.129 ± 0.050

Au final, nous sommes classés au premier rang pour le MAD à ES, et deuxième sur les autres. Les scores de distance à ED sont légèrement meilleurs qu' à ES, ce qui est cohérent avec les résultats des autres méthodes et peut s'expliquer, comme en 2D, par la plus grande complexité des formes à ES.

Scores cliniques Les scores cliniques correspondants sont donnés dans le tableau 6.8 et montrent des corrélations plus faibles que les BEAS sur tous les indices, avec des biais plus élevés sauf pour le volume à ES et la fraction d'éjection. Cependant, les résultats restent très compétitifs et notre mesure d'erreur globale calculée selon les lignes directrices du challenge CETUS $(0,627)$ est au final meilleure que celle des BEAS $(0,644)$.

Analyse visuelle La plupart des segmentations ratées proviennent d' erreurs locales sur les cartes de bord (surtout au niveau de l'apex, des parois septales et latérales et de la base). Ces erreurs sont très souvent corrélées avec des pertes de signal, fréquentes dans ces régions. Grâce à la régularisation, les prédictions de notre méthode semblent toujours plausibles.

1.6.4.4 Discussion

Les valeurs élevées de HD qui caractérisent les prédictions des SRF sont plus faibles dans cette étude, ce qui selon nous est dû au lissage de forme effectué par l'ASM. Les valeurs aberrantes correspondent souvent aux mêmes patients, ce qui confirme le manque de robustesse des SRF, qui échouent à généraliser le traitement à des cas particuliers.

1.6.5 Conclusion

Nous avons montré que les SRF peuvent être intégrées avec succès dans des pipelines entièrement automatiques et semi-automatiques. Un point clé de l'algorithme est l'extraction manuelle des caractéristiques de l'image, ce qui est aussi bien sa force que sa limite. Une extraction automatique des caractéristiques peut être réalisée par apprentissage profond, où le mappage complet de l'image à la segmentation est appris.

1.7 Évaluer le potentiel des méthodes d'apprentissage profond pour la segmentation automatique des images échographiques

Cette section couvre la description du réseau de neurones convolutionnel U-Net, en se concentrant sur ses propriétés et son potentiel pour la segmentation automatique du cœur. En particulier, nous souhaitons répondre aux questions suivantes :

- Combien de patients sont nécessaires pour entraîner un CNN et obtenir des résultats très précis en segmentation d'images échocardiographiques 2D vis-à-vis de l'expert de référence ?
- Où se situe la performance des architectures encodeur-décodeur (EDN) par rapport aux techniques de l'état de l'art sans apprentissage profond ?
- Quelle est la précision des indices cliniques estimés à partir des segmentation par CNN comparée aux variabilités inter et intra-experts ?

1.7.1 Réseaux convolutionnels

Parmis les techniques d'apprentissage profond, les réseaux de neurones convolutionnels (CNNs) consistent traditionnellement en des modèles pré-établis mappant la sortie directement depuis l'entrée par une succession de convolutions dont les paramètres sont appris par rétropropagation. Leur force vient de l'apprentissage automatique de toutes les transformations intermédiaires du mappage (y compris l'extraction de caractéristiques), et de la capacité robuste de la rétropropagation à converger vers un bon optima local.

1.7.2 U-Net

L'architecture U-Net [34] a été développée dans la communauté médicale. Ce réseau de segmentation est directement inspiré des auto-encodeurs convolutifs.

1.7.2.1 Architecture

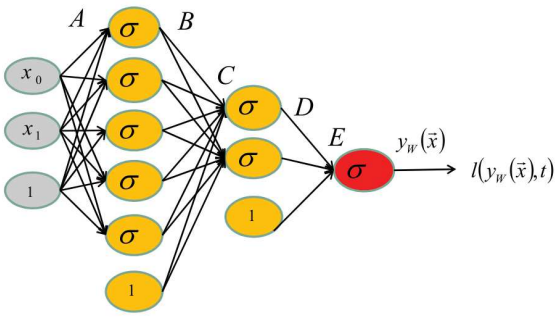
U-Net tire son nom de sa forme symétrique entre une phase de compression spatiale des caractéristiques de l'image (partie encodeur) et une phase de reconstruction de la segmentation (partie décodeur). U-Net inclut des connexions de saut entre l' encodeur et le décodeur afin de récupérer les informations bas niveau de l'image lors de la reconstruction. Au fur et à mesure que la taille des cartes de caractéristiques diminue, le nombre de filtres augmente.

1.7.2.2 Phase d'entraînement

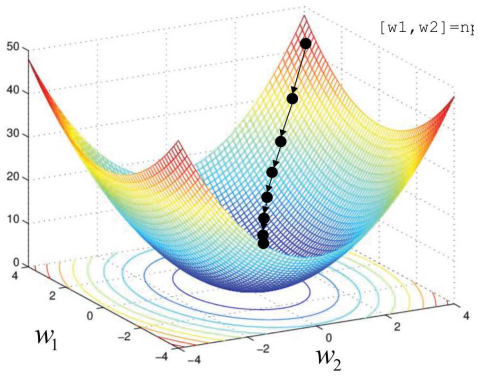
Fonction de coût Le modèle est adapté par une suite de mise à jour des paramètres des filtres de façon à ce que la sortie du réseau optimise une fonction objectif. Comme dans les problèmes d'optimisation classiques, nous optimisons généralement le réseau sur une fonction de coût plutôt que sur l'objectif, de sorte à chercher dans l'espace de solution un minimum plutôt qu'un maximum.

Pour la segmentation multi-classe, les fonctions de coût les plus fréquentes sont l'entropie croisée catégorique et le coût du Dice multi-classe [35]. Elles sont calculées entre les prédictions du modèle et les masques de vérité de terrain (apprentissage supervisé profond).

$$\frac{\partial(l(y_W(\vec{x}),t))}{\partial W^{[0]}} = \frac{\partial(l(y_W(\vec{x}),t))}{\partial y_W(\vec{x})} \frac{\partial(y_W(\vec{x}))}{\partial E} \frac{\partial(E)}{\partial D} \frac{\partial(D)}{\partial C} \frac{\partial(C)}{\partial B} \frac{\partial(B)}{\partial A} \frac{\partial(A)}{\partial W^{[0]}}$$



(a) Illustration de rétropropropagation sur un réseau neuronal entièrement connecté [36]



(b) Paysage d'optimisation : convergence itérative depuis une initialisation aléatoire [36]

FIGURE 1.8: Principales composantes de l'optimisation en apprentissage profond : a) Le gradient à appliquer à W pour corriger l'erreur sur l'objectif est obtenu en appliquant le théorème de dérivation des fonctions composées à toutes les couches intermédiaires. Chaque poids est mis à jour d'une fraction de son gradient (taux d'apprentissage). b) Progressivement, le modèle converge vers un minimum local de la fonction de coût calculée sur l'ensemble des données d'entraînement.

Rétropropagation Proposée dans (Rumelhart et al., 1986) [25], la rétropropagation a permis d'entraîner efficacement des réseaux de neurones à plusieurs couches. L'erreur sur la fonction de coût est propagée à chaque paramètre depuis la sortie à l'entrée selon le théorème de dérivation des fonctions composées [25], comme illustré par la figure 1.8.

1.7.2.3 Phase de test

Au moment de l'application sur une nouvelle donnée, la prédiction du modèle est obtenue en effectuant une passe en avant : l'image d'entrée est redimensionnée, puis traitée de couche en couche afin de construire le masque de segmentation.

1.7.3 Comparison aux SRF

Nous avons comparé U-Net aux SRF pour deux tailles d'ensembles de données d'entraînement, 50 patients annotés à ED et ES sur les vues A4C et A2C (200 images) et 400 patients (1600 images), ce qui est le maximum que nous pouvons utiliser afin de garder un pli de côté pour la validation et un autre de test.

Avec 50 cas (Tab. 7.3), les SRF ont systématiquement montré des performances en moyenne meilleures que le U-Net, mais inconstantes, comme le suggèrent les valeurs d' écart type. Avec 400 cas d'entraînement, U-Net a obtenu des scores plus élevés que les SRF avec une marge significative, comme le montre le tableau 7.4.

Tandis que les SRF n'ont que légèrement bénéficié d'une augmentation de 800% de la taille de l'ensemble d'entraînement et semblent atteindre un plateau, notre U-Net a presque doublé ses scores. Ce travail a été publié lors de la conférence IEEE IUS 2018 [31].

1.7.4 Potentiel de U-Net pour la segmentation cardiaque ultrasonore 2D

Ce travail a été publié dans le journal IEEE Transactions on Medical Imaging (TMI) [5].

1.7.4.1 Évaluation

Nous effectuons une validation croisée sur les dix plis de CAMUS, équilibrés en termes de qualité d'image et de valeur de fraction d'éjection. En plus des valeurs moyennes et des écarts-types sur chaque métrique géométrique, nous mesurons la robustesse en évaluant le nombre de cas où les contours prédits sont loin des annotations de la vérité terrain. Ces cas aberrants géométriques sont caractérisés par des valeurs MAD ou HD en dehors de la variabilité inter-observateurs moyenne.

1.7.4.2 Méthodes évaluées

U-Net 1 et U-Net 2 Nous avons décidé de comparer les performances de deux implémentations, *i)* U-Net 1, optimisé pour la vitesse et adapté depuis (Smistad et al., 2017) [19] et *ii)* U-Net 2, inspiré par notre propre étude (Leclerc et al., 2018) [31] et optimisé pour la précision.

B-spline explicit active surface model (BEASM) Nous comparons U-Net au vainqueur du challenge CETUS, l'algorithme "B-Spline Explicit Active Surface" (BEAS) [37]. Nous nous comparons à la version améliorée appelée BEASM [38], qui incorpore un modèle de forme actif.

La méthode BEASM est un modèle déformable, donc l'initialisation du contour joue un rôle crucial sur la qualité des résultats. Nous utilisons deux stratégies d'initialisation différentes :

1. BEASM-f, où le contour est automatiquement initialisé à partir de la méthode proposée dans (Barbosa et al., 2013) [37];
2. BEASM-s, où le contour est initialisé à partir de trois points (deux à la base et un au sommet de l'endocarde) extraits depuis les contours de référence.

1.7.4.3 Inter- et intra- variabilité entre experts sur la segmentation d'images en échocardiographie 2D

Les écarts inter et intra-observateurs ont été calculés à partir du pli 5 et limités aux patients ayant une bonne et moyenne qualité d'image (40 patients).

Inter-variabilité Nous fournissons la comparaison entre chaque paire de cardiologues. Les plus faibles écarts sont soulignés en cyan pour chaque métrique en haut du tableau 7.7. Les valeurs utilisées pour les critères aberrants sont surlignées en rouge.

Intra-variabilité L'intra-variabilité, représentée par O_{1b}, est significativement inférieure à l'inter-variabilité. Les valeurs suggèrent que O_1 a un contour spécifique, qui est reproduit indépendamment de l'instant.

1.7.4.4 Résultats géométriques et cliniques

Les tableaux 1.6 et 1.7 rapportent respectivement les scores géométriques et cliniques, où la supériorité des U-Nets est évidente.

TABLE 1.6: Précision de segmentation des méthodes évaluées sur les dix plis de CAMUS, restreinte aux patients ayant une bonne & moyenne qualité d'image.

Method	ED						ES					
	LV_{endo}			LV_{epi}			LV_{endo}			LV_{epi}		
	D	**MAD**	**HD**	**D**	**MAD**	**HD**	**D**	**MAD**	**HD**	**D**	**MAD**	**HD**
	val.	mm	mm	val.	mm	mm	val.	mm	mm	val.	mm	mm
O_{1a} vs O_2 (inter-obs)	0.919 ±0.033	2.2 ±0.9	6.0 ±2.0	0.913 ±0.037	3.5 ±1.7	8.0 ±2.9	0.873 ±0.060	2.7 ±1.2	6.6 ±2.4	0.890 ±0.047	3.9 ±1.8	8.6 ±3.3
O_{1a} vs O_3 (inter-obs)	0.886 ±0.050	3.3 ±1.5	8.2 ±2.5	0.943 ±0.018	2.3 ±0.8	6.5 ±2.6	0.823 ±0.091	4.0 ±2.0	8.8 ±3.5	0.931 ±0.025	2.4 ±1.0	6.4 ±2.4
O_2 vs O_3 (inter-obs)	0.921 ±0.037	2.3 ±1.2	6.3 ±2.5	0.922 ±0.036	3.0 ±1.5	7.4 ±3.0	0.888 ±0.058	2.6 ±1.3	6.9 ±2.9	0.885 ±0.054	3.9 ±1.9	8.4 ±2.8
O_{1a} vs O_{1b} (intra-obs)	0.945 ±0.019	1.4 ±0.5	4.6 ±1.8	0.957 ±0.019	1.7 ±0.9	5.0 ±2.3	0.930 ±0.031	1.3 ±0.5	4.5 ±1.8	0.951 ±0.021	1.7 ±0.8	5.0 ±2.1
SRF	0.895 ±0.074	2.8 ±3.6	11.2 ±10.2	0.914 ±0.057	3.2 ±2.0	13.0 ±9.1	0.848 ±0.137	3.6 ±7.8	11.6 ±13.6	0.901 ±0.078	3.5 ±4.7	13.0 ±11.1
BEASM-f	0.879 ±0.065	3.3 ±1.8	9.2 ±4.9	0.895 ±0.051	3.9 ±2.1	10.6 ±5.1	0.826 ±0.092	3.8 ±2.1	9.9 ±5.1	0.880 ±0.054	4.2 ±2.0	11.2 ±5.1
BEASM-s	0.920 ±0.039	2.2 ±1.2	6.0 ±2.4	0.917 ±0.038	3.2 ±1.6	8.2 ±3.0	0.861 ±0.070	3.1 ±1.6	7.7 ±3.2	0.900 ±0.042	3.5 ±1.7	9.2 ±3.4
U-Net 1	0.934 ±0.042	1.7 ±1.0	5.5 ±2.9	0.951 ±0.024	1.9 ±0.9	5.9 ±3.4	0.905 ±0.063	1.8 ±1.3	5.7 ±3.7	0.943 ±0.035	2.0 ±1.2	6.1 ±4.1
U-Net 2	**0.939** ±0.043	**1.6** ±1.3	**5.3** ±3.6	**0.954** ±0.023	**1.7** ±0.9	6.0 ±3.4	**0.916** ±0.061	**1.6** ±1.6	**5.5** ±3.8	**0.945** ±0.039	**1.9** ±1.2	6.1 ±4.6

TABLE 1.7: Paramètres cliniques des méthodes évaluées sur les dix plis de CAMUS, restreinte aux patients ayant une bonne & moyenne qualité d'image.

Method	LV_{EDV}			LV_{ESV}			LV_{EF}		
	corr	**bias**$\pm\sigma$	**mae**	**corr**	**bias**$\pm\sigma$	**mae**	**corr**	**bias**$\pm\sigma$	**mae**
	val.	ml	ml	val.	ml	ml	val.	%	%
O_{1a} vs O_2	0.940	18.7±12.9	18.7	0.956	18.9±9.3	18.9	0.801	-9.1±8.1	10.0
O_{1a} vs O_3	0.895	39.0±18.8	39.0	0.860	35.9±17.1	35.9	0.646	-12.6±10.0	13.4
O_2 vs O_3	0.926	-20.3±15.6	21.0	0.916	-17.0±13.5	17.7	0.569	3.5±11.0	8.5
O_{1a} vs O_{1b}	0.978	-2.8±7.1	6.2	0.981	-0.1±5.8	4.5	0.896	-2.3±5.7	4.5
SRF	0.755	-0.2±25.7	17.4	0.827	9.3±18.0	14.8	0.465	-11.5±15.4	12.8
BEASM-f	0.704	13.4±30.6	22.9	0.713	18.0±25.8	22.5	0.731	-9.8±8.3	10.7
BEASM-s	0.886	14.6±19.2	17.8	0.880	18.3±16.9	19.5	0.790	-9.4±7.2	10.0
U-Net 1	0.947	-8.3±12.6	10.9	0.955	-4.9±9.9	8.2	0.791	-0.5±7.7	5.6
U-Net 2	**0.954**	-6.9 ±11.8	**9.8**	**0.964**	-3.7± 9.0	**6.8**	**0.823**	-1.0±7.1	**5.3**

1.7.4.5 Discussion

De cette étude, il est apparu qu'un CNN bien conçu pouvait atteindre des scores de segmentation impressionnants en analyse d'images échocardiographiques. Les deux U-Nets surpassent les méthodes de l'état de l'art entièrement et semi-automatiques sans apprentissage profond.

Les résultats géométriques et cliniques sont également meilleurs que les scores inter-observateurs sur toutes les structures et métriques. Ils restent inférieurs aux scores intra-observateurs lorsque l'on considère l'ensemble des données.

Trois limites principales ressortent de notre analyse :

1. Bien que les résultats obtenus par les U-Nets soient inférieurs à la moyenne des scores inter-observateurs, ils comportent encore trop d'erreurs par rapport aux valeurs intra-observateurs ;
2. 18% des segmentations produites par les deux U-Nets peuvent être considérées comme des cas aberrants géométriques, tandis que le taux de l'intra-variabilité est de 13%. Cela suggère la nécessité d'une méthode plus robuste ;
3. Nous avons évalué visuellement qu'environ 2% des prédictions étaient manifestement invraisemblables d' un point de vue anatomique, ce qu'un expert ne produirait jamais. L' évaluation géométrique classique est donc insuffisante pour juger de la qualité des segmentations médicales.

1.7.5 Comportement du réseau de neurones convolutionnel U-Net

Nous avons étudié le comportement de U-Net pour des variations de qualité et de quantité des données. Les principales conclusions sont que le modèle :

- est toujours rapide à entraîner et à appliquer ;
- apprend un protocole spécifique ;
- est robuste par rapport à la qualité de l'image.

Au-delà d'un ensemble d'entraînement de 250 patients, la performance ne s'améliore que légèrement, ce qui indique qu'un plateau a été atteint en termes de performance.

1.7.6 Conclusion

Nous avons montré à travers une analyse approfondie du comportement et des performances de U-Net que les modèles de type encodeur-décodeur sont des candidats très prometteurs pour la segmentation automatique multi-structures en échocardiographie 2D. U-Net a battu les SRF et les BEASM avec une marge significative, et a également obtenu de meilleurs résultats géométriques et cliniques que l'inter-variabilité entre trois experts.

L'intra-variabilité moyenne n'est pas atteinte, ce qui suppose une marge d'amélioration pour l'apprentissage. En conclusion, aussi bien le modèle actuel que les critères d'évaluation sont insuffisants et doivent être affinés.

1.8 Dépasser la performance et l'évaluation conventionnelles des modèles d'apprentissage profond de segmentation

Cette section porte sur l'étude de modèles avancés d'apprentissage profond basés sur l'architecture U-Net, l'objectif étant de comparer les approches encodeur-décodeur les plus récentes sur la base de données CAMUS. Pour compléter l'évaluation et le classement de ces méthodes, nous proposons également des métriques de plausibilité anatomique. Les interrogations suivantes sont étudiées :

- Peut-on améliorer les scores obtenus par l'architecture U-Net par des propositions plus complexes récemment proposées dans la littérature ?
- Comment pouvons-nous construire un critère automatique et objectif de la validité des formes prédites ? Quel est l'impact sur le classement des méthodes de segmentation ?

1.8.1 Encoder-decoders de l'état de l'art en segmentation

1.8.1.1 Supervision profonde

La supervision profonde a été proposée dans la littérature comme un moyen d'entraîner plus efficacement les réseaux de neurones profonds grâce à l'ajout d'objectifs intermédiaires [39]. Nous avons considéré dans ce travail deux approches de supervision profonde de l'architecture U-Net : la supervision profonde imbriquée [40], et la supervision profonde en cascade [41].

U-Net ++ (Zhou et al., 2018) [40] ont proposé d'ajouter des couches de convolution le long des connexions de saut de U-Net, et d'utiliser la supervision profonde pour forcer les caractéristiques extraites de l'image à être sémantiquement proches en tout lieu du réseau via l'ajout de trois objectifs de segmentation intermédiaires. Lors de la phase de test, la moyenne des quatre sorties est utilisée comme segmentation finale, suivant une stratégie de modèle d'ensemble. Après adaptation, notre version de U-Net++ comprend 1.1M de paramètres, moins que l'original (9M), pour une meilleure performance sur CAMUS.

"Stacked hourglasses" Le modèle des sabliers empilés - "Stacked Hourglasses" (SHG) - [41], intègre une succession de plusieurs réseaux encodeur-décodeurs (généralement avec la même architecture) en un seul réseau. Les premiers sous-réseaux sont utilisés comme des blocs résiduels, c'est-à-dire que l'entrée d'un sous-réseau est le résultat de l'addition entre la segmentation et l'entrée précédentes. Chaque sortie de sous-réseau est associée à un objectif de segmentation intermédiaire selon une stratégie de supervision profonde qui, combinée aux connexions résiduelles, force les sous-réseaux à apprendre à affiner la segmentation précédente. Pour notre version, nous avons utilisé l'architecture U-Net 1 comme motif.

1.8.1.2 Réseaux de neurones avec contrainte anatomique

Le réseau de neurones avec contrainte anatomique - "Anatomically Constrained Neural Network" (ACNN) - proposé dans (Oktay et al. 2017) [42], encourage des résultats de segmentation anatomiquement plausibles par l'ajout d'une contrainte de forme implicite générée par un auto-encodeur. L'optimisation de la segmentation est réalisée sur la fonction de coût :

$$L = L_x + \lambda_1 \times L_{he} \tag{1.12}$$

avec L_x le coût de segmentation (ici l'entropie croisée catégorique), L_{he} le coût anatomique dérivé de l'auto-encodeur, et λ_1 un hyper-paramètre qui équilibre les deux fonctions.

L_{he} correspond à la somme des distances euclidiennes entre les coefficients des codes générés par l'auto-encodeur pour les masques de la vérité terrain et ceux générés à partir des segmentations prédites par le réseau de segmentation.

Nous avons utilisé U-Net 1 comme réseau de segmentation dans notre implémentation d'ACNN. Les modèles résultants comportaient 2,2 millions de paramètres.

1.8.1.3 Résultats

Résultats géométriques Les scores géométriques sur les images de bonne et moyenne qualité sont reportés dans le tableau 8.2. On peut y observer que les architectures encodeur-décodeur ont obtenu des résultats très proches, toujours inférieurs à l'inter-variabilité. Cependant, par rapport au modèle de référence U-Net 1, les scores obtenus sont :

- équivalents, bien qu'une très faible amélioration soit constatée avec le réseau SHG ;
- légèrement dégradés pour l'approche ACNN en ce qui concerne les métriques HD ;
- dégradés pour l'architecture U-Net++.

Résultats cliniques Le tableau 8.4 montre les scores cliniques obtenus pour les 3 méthodes encodeur-décodeur évaluées comparées à U-Net 1. D'après ce tableau, on peut observer que SHG et ACNN obtiennent des scores similaires à U-Net 1, tandis que les estimations des volumes à partir des prédictions de U-Net++ sont moins précises que celles des trois autres encodeurs-décodeurs.

1.8.1.4 Conclusion

L'étude menée montre que les trois réseaux encodeurs-décodeurs testés, tous impliquant une architecture plus complexe que U-Net 1, ne produisent pas de meilleurs résultats sur l'ensemble de la base de données CAMUS. Cette observation confirme l'idée que U-Net atteint un plateau de performance, supposé à la fin du chapitre 7.

1.8.2 Métriques de plausibilité de forme en imagerie cardiaque

Dans cette section, nous complétons l'évaluation avec des métriques d'appréciation des formes associées aux structures cardiaques, à partir desquelles nous construisons la notion de cas aberrant anatomique en échocardiographie 2D. Ce travail a été publié dans le cadre de la conférence MIDL 2019 [43].

1.8.2.1 Simplicité et convexité

Pour comparer automatiquement la segmentation de plusieurs structures S par différents annotateurs, les auteurs dans (Zhu et al., 2017) [44] ont utilisé deux critères géométriques, la convexité et la simplicité :

$$\text{Convexité} : Cx(S) = \frac{Aire(S)}{Aire(ConvHull(S))} \tag{1.13}$$

avec $ConvHull(S)$ l'enveloppe convexe de S.

$$\text{Simplicité} : Sp(S) = \frac{\sqrt{4\pi \times Aire(S)}}{\text{Périmètre}(S)} \tag{1.14}$$

Table 1.8: Critère d'aberrance anatomique

Structure	Cx	Sp
LV	< 0.741	< 0.529
Epi	< 0.960	< 0.694

Ces deux métriques ont des valeurs comprises entre 0 et 1, et sont maximisées pour un cercle. Ce qui nous intéresse est que la convexité et la simplicité donnent potentiellement des valeurs discriminantes pour toute forme convexe telles que les formes ovales des cavités cardiaques, et pour les formes en pont comme le myocarde.

1.8.2.2 Aberrance anatomique

Nous avons établi comme critère de cas anatomiquement aberrant tout relevé de valeurs de convexité et/ou de simplicité en deçà des valeurs observées sur les annotations des experts (pli 5 de CAMUS). Les valeurs limites sont données dans le tableau 1.8.

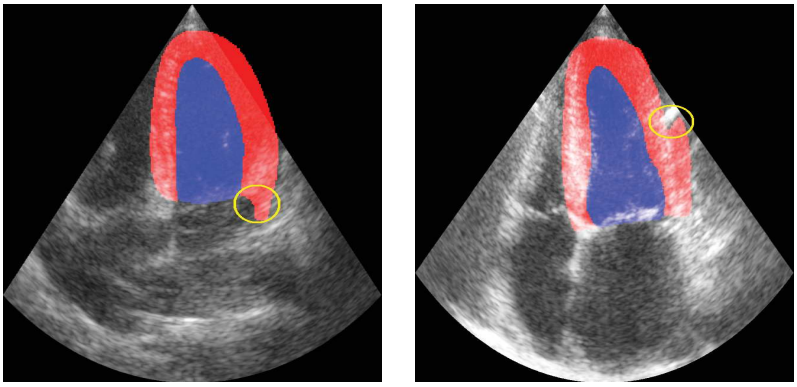
1.8.2.3 Impact sur le classement des méthodes de segmentation par apprentissage supervisé

A partir du tableau 8.7 sur les scores anatomiques des encodeurs-décodeurs testés, plusieurs observations peuvent être faites :

1. tous les modèles produisent en moyenne des formes moins convexes et plus complexes que les experts ;
2. U-Net 1 ne produit que 5% de cas aberrants anatomiques, ce qui soutient l'idée que le réseau a implicitement appris à reconstruire des masques de segmentation cohérents ;
3. Bien que U-Net 2 ait surclassé U-Net 1 sur toutes les métriques géométriques classiques (D, MAD, HD), il produit trois fois plus de formes anatomiquement invraisemblables. Cela peut être dû à son nombre de paramètres beaucoup plus élevé, et donc être un signe de sur-apprentissage ;
4. L'effet de raffinement de la segmentation dans SHG peut être observé à partir de la réduction significative du nombre de cas aberrants anatomiques (de 95 à 47) ;
5. la contrainte anatomique implicite des ACNNs a tendance à créer des aberrations anatomiques sur notre jeu de données;
6. Les cas aberrants anatomiques et géométriques sont souvent liés, car la plupart des cas aberrants anatomiques sont aussi des cas aberrants géométriques.

1.8.2.4 Discussion

Les critères anatomiques proposés sont sensibles aux déformations locales, comme le montre la figure 1.9. Cependant, ils ont été établis sur un ensemble d'annotations restreint, aussi bien sur le plan de la variété d'image que de la variété d'experts, ce qui implique qu'ils sont nécessairement imprécis.



(a) Cx= 0.96|0.94Sp= 0.67|0.71 (b) Cx= 0.93|0.95Sp= 0.65|0.70

FIGURE 1.9: Anatomical outliers from U-Net 2: a) is also a geometrical outlier but not b). Local shape irregularities are cercled in yellow.

De plus, la plausibilité anatomique n'implique pas nécessairement une meilleure précision, de sorte que ce critère de qualité doit être secondaire comparé aux valeurs de MAD et HD. Nos critères de plausibilité sont donc à considérer comme des indicateurs de risque d'échec de la segmentation, plutôt que des limites strictes.

1.8.2.5 Conclusion

L'introduction de métriques anatomiques permet de compléter l'évaluation des modèles d'apprentissage supervisé, en couplant la précision sur les contours et sur les indices cliniques avec la régularité et la justesse des formes observées.

Bien qu'imprécis, les critères de cas aberrants anatomiques que nous avons conçus ont permis d'observer à la fois le sur-apprentissage de U-Net 2 et l'effet de raffinement dans SHG. Nous avons ansi établi que U-Net 1 est la méthode d'apprentissage supervisé la plus prometteuse, à ce stade de cette étude, et qu'il est difficile d'améliorer les résultats obtenus.

1.9 Améliorer la robustesse de la segmentation par apprentissage profond par l'incorporation de mécanismes d'attention

Afin de créer un modèle capable de dépasser les résultats obtenus par U-Net 1, nous nous sommes intéressés aux mécanismes d'attention pouvant être utilisés pour concentrer le traitement dans une région d'intérêt. Cette section répond aux trois questions suivantes :

1. Est-il possible d'améliorer la précision des réseaux de neurones convolutionnels (CNN) pour la segmentation d'images échocardiographiques ?
2. Le nombre de cas aberrants peut-il être réduit de manière significative en modifiant l'architecture du réseau ?
3. Un CNN peut-il d'obtenir des résultats inférieurs à la variabilité intra-observateur, aussi bien pour la segmentation que pour l'estimation d'indices cliniques ?

Les résultats correspondants ont été publiés dans la conférence IEEE International Ultrasonic Symposium conference [45] et récemment soumis dans le journal Transactions on Ultrasonics, Ferroelectrics and Frequency Control (TUFFC) [46].

1.9.1 Mécanismes d'attention en échocardiographie

1.9.1.1 Définition de l'attention dans le cadre de l'apprentissage profond

Les mécanismes d'attention correspondent à l'intégration d'une procédure de contextualisation à l'intérieur d'un algorithme d'apprentissage supervisé afin d' améliorer sa performance. La contextualisation est généralement appliquée directement sur l'image ou sur un espace de caractéristiques, et consiste à apprendre à concentrer l'extraction d'information et le traitement sur des parties isolées de l'image.

1.9.1.2 Potentiel des mécanismes d'attention sur la base de données CAMUS

Nous avons conduit une expérience consistant à évaluer la performance de U-Net 1 en supposant une pré-localisation parfaite dans le but d' étudier l'intérêt d'effectuer une tâche de localisation avant la segmentation. Les résultats sont donnés dans la section 9.1.

En résumé, cette expérience a révélé que l'insertion d'une étape de localisation précise avant la segmentation d'un modèle U-Net 1 pourrait amener à des résultats remarquablement précis, avec des scores moyens toujours inférieurs à la variabilité intra-observateur et un nombre de cas aberrants inférieur à 8%.

1.9.2 Architectures d'apprentissage avec attention pour la segmentation échocardiographique 2D

Nous avons créé et comparé deux modèles à base de CNNs, les premiers réseaux avec attention appliqués dans notre contexte.

1.9.2.1 "Refining U-Net"

"Refining U-Net" (RU-Net) consiste en une succession de deux U-Nets, le second affinant le résultat de la segmentation du premier. Ce modèle est donc similaire à l'architecture

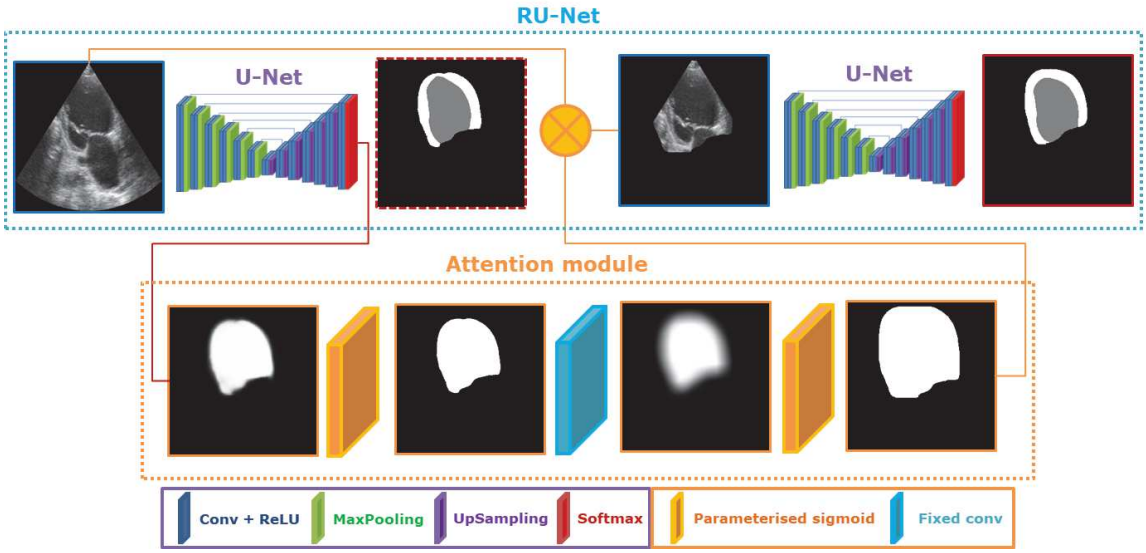


FIGURE 1.10: Illustration du modèle RU-Net. Les deux U-Nets sont indépendants (paramètres séparés).

SHG présentée dans le chapitre précédent, sauf que la première segmentation est ici utilisée pour concentrer le traitement dans une région d'intérêt de l'image ultrasonore d'entrée. L'architecture globale est fournie dans la Fig. 1.10, où le mécanisme d'attention est encadré en jaune. Le module d'attention est composé de deux fonctions sigmoïdes paramétrées qui sont respectivement appliquées avant et après une couche de dilatation réalisée par une convolution, ce qui permet de rendre le processus entier différenciable.

1.9.2.2 "Localization U-Net"

"Localization U-Net" (LU-Net) a pour objectif de localiser et de segmenter les parois endocardique et épicardique du ventricule gauche dans une procédure d'apprentissage de bout en bout. La différence avec RU-Net est que LU-Net incorpore une étape de localisation après la première segmentation, sous la forme de prédiction des coordonnées d'une boîte englobante. L'hypothèse sous-jacente de cette stratégie est que l'optimisation conjointe de la localisation et de la segmentation devrait conduire à une meilleure segmentation. L'architecture générale est illustrée dans la Fig. 1.11. Il est intéressant de noter que l'étape de segmentation intermédiaire a permis de réduire de moitié les erreurs de localisation.

1.9.3 Expériences

1.9.3.1 Méthodes de segmentation

Nous avons comparé nos modèles d'attention au réseau Attention U-Net (AG-U-Net), récemment proposé dans (Oktay et al., 2018) [47]. Pour RU-Net, nous avons utilisé une dilatation de 30 pixels et un seuil à 0,7, ce qui s'est avéré être la combinaison la plus performante (4M de paramètres). Pour LU-Net, nous avons utilisé l'architecture U-Loc2-multi comme réseau de proposition de région et U-Net 1 comme réseau de segmentation. Deux marges $m = 5\%$ et $m = 15\%$ de boîte englobante ont été évaluées (13M de paramètres).

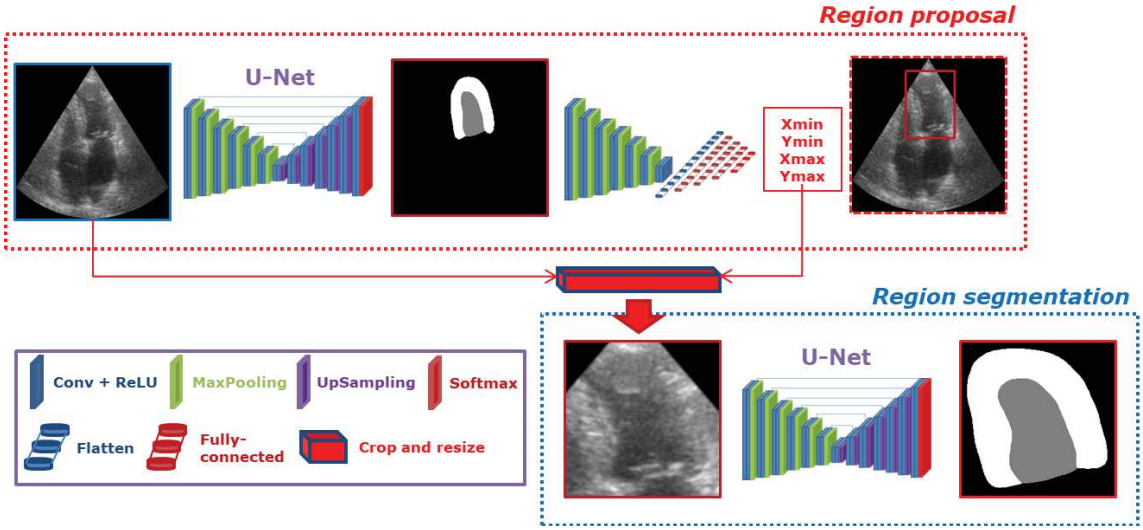


FIGURE 1.11: Illustration du modèle LU-Net avec le réseau de proposition de région U-Loc2-multi-région, décrit dans la section 9.5.1.
Les deux U-Nets sont indépendants.

1.9.3.2 Résultats géométriques

A partir du tableau 1.9, on peut tout d'abord observer que tous les réseaux incorporant un méchanisme d'attention ont produit des résultats soit similaires, soit meilleurs, que le réseau U-Net 1 d'origine. Le modèle RU-Net a obtenu des résultats similaires pour le LV_{endo} et une faible amélioration pour le LV_{epi} (surtout visible pour la métrique HD, avec une amélioration de $0,4$ mm), comparé à U-Net 1. Il a également obtenu une réduction de 2% des cas aberrants géométriques sur les données de bonne et moyenne qualité (5% pour la base entière).

Les modèles les plus performants ont été AG-U-Net et LU-Net. AG-U-Net a obtenu les meilleurs résultats pour la segmentation de la paroi LV_{endo}, conduisant à des scores de segmentation proches mais toujours supérieurs à la variabilité intra-observateur. L'approche LU-Net-m5 a obtenu les meilleurs résultats pour la segmentation de la paroi LV_{epi} et le moins de cas aberrants géométriques (11%). Il est intéressant de noter que ces scores sont soit équivalents, soit inférieurs à la variabilité intra-observateur pour ces deux aspects.

1.9.3.3 Résultats cliniques

Les modèles AG-U-Net et LU-Net-m5 ont obtenu les meilleurs scores pour tous les indices cliniques (Tab. 9.4). Cependant, même si les scores de LU-Net-m5 et AG-U-Net étaient légèrement meilleurs que ceux de U-Net 1, les erreurs restaient supérieures à l'intra-variabilité.

1.9.4 Discussion

1.9.4.1 Réseaux d'attention

Les résultats soulignent la capacité des réseaux basés sur le mécanisme d'attention à améliorer la segmentation et l'estimation des indices cliniques associée en échocardiographie 2D.

TABLE 1.9: Scores géométriques des 4 méthodes évaluées sur les patients de bonne et moyenne qualité d'image (406 au total).

Modèle	LV_{endo}			LV_{epi}			outl.
	D	***MAD***	***HD***	***D***	***MAD***	***HD***	**geo.**
	val.	mm	mm	val.	mm	mm	# %
intra-observateur	0.937 ±0.027	1.4 ±0.5	4.5 ±1.8	0.954 ±0.020	1.7 ±0.8	5.0 ±2.2	21 13
U-Net 1	0.920 ±0.056	1.7 ±1.2	5.6 ±3.3	0.947 ±0.030	1.9 ±1.1	6.2 ±3.7	282 17%
RU-Net	0.925 ±0.049	1.7 ±1.0	5.4 ±3.3	**0.950** ±0.030	1.8 ±1.1	5.8 ±3.9	240 15%
AG-U-Net [47]	0.930 ±0.049	**1.5** ±1.3	**5.3** ±3.4	**0.950** ±0.026	1.8 ±1.0	5.9 ±3.7	270 17%
LU-Net-m5	**0.953** ±0.026	1.7 ±0.9	5.5 ±3.6	0.932 ±0.043	**1.5** ±0.8	**5.1** ±3.3	**186** 11%
LU-Net-m15	0.952 ±0.029	1.7 ±1.1	5.6 ±4.0	0.931 ±0.049	**1.5** ±1.1	5.3 ±3.6	203 12%

1.9.4.2 Comparaison à la variabilité intra-observateur

LU-Net a atteint la variabilité intra-observateur moyenne pour la paroi LV_{epi}. Le nombre de cas aberrants géométriques produits par cette méthode (c.-à-d. 11%) est également inférieur au score intra-observateur. A notre connaissance, c'est la première fois qu'un tel résultat est obtenu dans le cadre de la segmentation d'images échocardiographiques 2D. Cependant, les scores obtenus par notre modèle restent insuffisants pour la paroi LV_{endo}.

1.9.4.3 Pistes d'amélioration

Nous avons identifié deux pistes d'amélioration potentielles. Tout d'abord, en se basant sur les tableaux 9.2 et 9.1, il semble que l'étape de localisation pourrait être optimisée plus avant afin d'améliorer les scores de LU-Net.

Deuxièmement, il semble incontournable d'introduire de la cohérence temporelle dans les architectures d'apprentissage profond. En effet, alors que la stratégie actuelle (où ED et ES sont traités séparément) fournit des résultats de corrélations élevés pour les indices LV_{EDV} et LV_{ESV} $(0,956)$, l'estimation de la fraction d'éjection descend à $0,829$. Cela révèle un manque de cohérence entre les résultats de segmentation de LU-Net à ED et ES.

1.9.5 Conclusion

Nous avons proposé deux méthodes basées sur l'incorporation d'attention dans l'image afin d'améliorer la robustesse de la segmentation de l'endocarde et de l'épicarde en échocardiographie 2D. Nous avons montré que l'optimisation conjointe des tâches de localisation et de segmentation du modèle LU-Net conduisait à de meilleurs résultats de segmentation.

Bien qu'il nous reste à atteindre l'intra-variabilité sur plusieurs métriques, ce travail a établi la localisation comme une piste de choix pour une analyse plus robuste des images ultrasonores par apprentissage profond.

1.10 Conclusion

Catte dernière section résume les principales contributions et conclusions tirées de l'analyse des résultats rassemblés dans le manuscrit. Des pistes d'amélioration et des perspectives à court et long termes sont ensuite fournies pour ouvrir la porte à d'autres investigations.

1.10.1 Principales contributions

Cette thèse a été le lieu de trois principales contributions :

1. Nous avons construit et rendu publique le plus grand ensemble de données cliniques dédié à l'analyse d'images échocardiographiques 2D, contenant des contours de référence pour trois structures (ventricule gauche, myocarde, et oreillette gauche). De plus, nous maintenons en parallèle de l'accès à la base de données CAMUS une plate-forme d'évaluation permettant une comparaison directe avec nos solutions [1] ;

2. Nous avons démontré que l'apprentissage supervisé, en particulier l'apprentissage profond, est actuellement l'approche la plus prometteuse en vue d'établir une segmentation entièrement automatique, précise et rapide en échocardiographie [5] ;

3. Nous avons montré le fort potentiel des mécanismes d' attention pour l'amélioration de la robustesse et de la précision des résultats de segmentation prédits par des modèles d'apprentissage profond [46].

1.10.2 Bilan

1.10.2.1 Aspects méthodologiques

Sur le plan méthodologique, une étude très complète a été menée dans le cadre de ce doctorat: i) l'adaptation de l'algorithme des forêts aléatoires structurées à l'échocardiographie 2D et 3D [30] [32] ; ii) l'analyse approfondie de l'application de l'architecture U-Net sur la base de données CAMUS [31] [5] ; iii) l'extension de l'évaluation à la validation anatomique des structures [48] ; iv) le développement d'architectures encodeur-décodeurs plus robustes et précises grâce à l'incorporation de mécanismes d'attention [45] [46].

Cela nous permet de répondre ici à notre premier objectif : l'évaluation du potentiel des méthodes d'apprentissage supervisé pour la segmentation sémantique en échocardiographie 2D. Nos résultats ont montré que les méthodes d'apprentissage profond, au travers d'architectures encodeur-décodeur, peuvent produire des résultats de segmentation très précis, se situant en-dessous de la variabilité inter- et approchant la variabilité intra- expert. De plus, elles sont robustes à la qualité d'image et également, dans une certaines mesure, aux paramètres d'acquisition.

1.10.2.2 Aspects cliniques

Nous avons établi que les techniques d'apprentissage profond obtiennent actuellement les meilleurs scores cliniques sur la base de données CAMUS, avec des estimations très précises des valeurs du volume ventriculaire gauche en fin diastole et systole (corrélations d'environ $0,95$ et erreurs moyennes d'environ 8 ml). Les résultats sont plus contrastés pour l'estimation de la fraction d'éjection, avec des corrélations autour de $0,80$ et une erreur moyenne moyenne autour de 5%. Cette observation souligne la nécessité d'introduire de la cohérence temporelle dans le cadre de la segmentation du cœur.

Nous pouvons alors apporter une réponse à notre deuxième objectif : évaluer si nous sommes sur le point d'automatiser complètement l'analyse cardiaque en imagerie ultrasonore. Nos résultats montrent que les plus performantes de nos méthodes, entièrement automatiques, obtiennent des résultats inférieurs à la variabilité inter-observateurs pour tous les indices, ce qui dénote le potentiel élevé des méthodes d'apprentissage profond pour l'analyse d'images échocardiographiques. De plus, les réseaux convolutionnels permettent une inférence rapide et demande une mémoire de stockage raisonnable, ce qui rend ces solutions encore plus attrayantes d'un point de vue clinique. Les résultats semblent cependant encore manquer de robustesse, ce qui est confirmé par les valeurs de l'intra-variabilité.

1.10.3 Perspectives

1.10.3.1 Perspectives à court terme

Perspectives algorithmiques Ma thèse a permis de fournir une preuve de concept soutenant que l'analyse d'images échocardiographiques peut être abordée avec succès par des méthodes d'apprentissage supervisé. Néanmoins, l'étude dans son ensemble souffre de plusieurs limitations qui devraient idéalement être l'objet d'investigations approfondies :

- l'amélioration de la robustesse de la segmentation multi-structures automatique : les résultats présentés dans le chapitre 9 montrent que la robustesse de la segmentation pourrait encore être augmentée aussi bien pour l'endocarde que pour l'épicarde. Nous pensons qu'une solution permettant d'atteindre l'intra-variabilité moyenne sur ces deux structures peut être construite autour de mécanismes d'attention ;

- l'évaluation du potentiel des techniques d'apprentissage profond en échocardiographie 3D : en raison de la petite taille du seul ensemble de données accessible au public, CETUS [16], la précision potentielle des méthodes d'apprentissage profond pour l'analyse d'images ultrasonores 3D ne peut être établi. A partir des conclusions de notre étude en 2D, il serait intéressant d'adapter l'algorithme appelé LU-Net [46] à la localisation et segmentation du cœur dans des volumes 3D. En fonction de la consommation de mémoire, nécessairement plus grande en 3D qu'en 2D, il sera peut-être préférable d'étudier un traitement multi-coupes ou des approches 2.5D [49] ;

- la régularisation de l'évolution de la segmentation au cours du cycle cardiaque a été établie comme une nécessité pour obtenir des estimations robustes de la fraction d'éjection. En outre, nous avons observé que certains cas difficiles à résoudre de par la présence d'artéfacts, pourraient bénéficier de contraintes temporelles. Il apparaît donc essentiel d'introduire ou d'encourager de la cohérence temporelle dans l'apprentissage;

- la plausibilité anatomique : la conception de métriques anatomiques basées sur l'analyse des formes géométriques produites nous a permis de montrer que les réseaux convolutionnels apprennent un mappage qui incorpore déjà une certaine validité des formes, mais sans aucune garantie. Il serait donc intéressant d'étudier des solutions de recalage multi-atlas [50] ou des méthodes d'apprentissage incluant une correction [51] pour assurer la prédiction de formes anatomiquement plausibles en toute situation;

- le développement de simulations réalistes : nous avons observé sur CAMUS que certaines prédictions aberrantes seraient très difficiles à résoudre en cela qu'elles sont le produit d'acquisitions atypiques (artéfacts, zoom, orientation, très faible contraste...) et correspondent à des cas très inhabituels pour le réseau. Une façon de résoudre cela consisterait à ajouter les variations manquantes dans notre ensemble de données sous

forme d'images virtuelles réalistes. La simulation devra reposer sur des paramètres contrôlés pour générer un large éventail de cas.

Perspectives cliniques La présente étude a révélé qu'avant d'appliquer des solutions d'apprentissage profond en routine clinique, les besoins suivants doivent être satisfaits :

1. la création d'un large ensemble de données publiques comportant des annotations consensuelles : la base de données CAMUS comprend les vues apicales 2 et 4 chambre de 500 patients, pour un total de 2000 images annotées. Différentes qualités d'image y sont représentées. Cela représente une partie intéressante de la variabilité notable, mais ne recouvre qu'une petite partie des populations, pathologies, fournisseurs d'échographes, et experts existants. Le nombre de vues et de structures impliquées est également limité. C'est pourquoi nous qualifions cette étude de preuve de concept, précurseuse, destinée à motiver la communauté à établir une base de données plus complète. Cette nouvelle base de taille et de variété suffisantes devrait idéalement comporter des annotations soigneusement validées par le corps médical, afin d'évaluer avec précision l'inter- et l'intra-variabilité sur l'analyse d'images échocardiographiques 2D, et d'exploiter pleinement le potentiel des méthodes d'apprentissage supervisées démontré ici;

2. la validation clinique des méthodes : nous avons tenté d'effectuer une évaluation approfondie des méthodes d'apprentissage supervisé, en combinant de la validation croisée et des métriques complémentaires (géométriques, cliniques, anatomiques). Cependant, la conformité des prédictions aux exigences cliniques n'est pas évaluée. En particulier, il apparaît primordial d'étendre l'étude à davantage de cas, et d'intégrer une notion plus fine de pathologie que la seule estimation de la fraction d'éjection du ventricule gauche;

3. le développement de logiciels guidant l'acquisition : nous avons montré le fort potentiel des méthodes d'apprentissage supervisé pour l'analyse automatique des images échocardiographiques, mais la nécessité d'automatiser la procédure d'acquisition subsiste. En effet, guider les praticiens dans le positionnement de la sonde est nécessaire pour garantir l'utilisation de visuels adéquats du cœur avant la segmentation [52]. Pour ce faire, la conception d'algorithmes qui apprennent à reconnaître si l'acquisition correspond à un contexte approprié, par exemple en exprimant une incertitude sur la validité de la vue, semble être une piste intéressante [53].

1.10.3.2 Perspectives à long terme

Perspectives cliniques Cette étude fait partie d'une révolution dans le traitement d'image médical, qui appelle des interactions plus étroites entre chercheurs, développeurs et cliniciens afin de potentiellement amener aux avancées suivantes :

1. l'utilisation quotidienne de méthodes d'apprentissage pour aider les médecins dans leurs analyses, ce qui améliorerait considérablement leur productivité;

2. un algorithme de segmentation automatique appris à partir d'un ensemble de données consensuelles permettrait non seulement de générer des résultats reproductibles, mais également de représenter la norme de la segmentation en échocardiographie. Un tel outil pourrait ainsi être utilisé pour former de nouvelle générations de médecins, et d'aider la transmission de connaissances des experts aux novices ;

3. des logiciels d'analyse d'image robustes permettraient d'évaluer le potentiel de modalités innovantes par rapport à celles établies, et d'ouvrir la porte à de nouvelles pratiques. Par

exemple, le couplage efficace de l'apprentissage supervisé avec l'imagerie échographique pourrait permettre de diagnostiquer un plus grand nombre de patients, y compris ceux qui se trouvent dans des endroits éloignés (télémédecine), et de favoriser un diagnostic précoce en permettant à des non-professionnels de faire des acquisitions.

Perspectives algorithmiques Les modèles d'apprentissage en profondeur peuvent être compressés pour consommer moins de ressource et être plus rapides lors de leur utilisation. Cependant, la phase d'entraînement reste une étape incontournable, et coûteuse en de multiples ressources (temps, électricité, processeurs, pollution...). Par exemple, pour cette seule étude, plusieurs centaines de modèles ont été entraînés en raison de changements d'ensembles de données (entraînement, test et validation), et de variations d' hyperparamètres. Par conséquent, il est possible que se développe bientôt un fort besoin de modèles capables de ne pas apprendre de zéro. L'apprentissage par transfert va dans ce sens, mais dans le but de compenser la petite taille des ensembles de données plutôt que de raccourcir la phase d'entraînement. De même, il y'a un besoin de modèles convergeant systématiquement vers une solution optimale fixe et unique (sans stochasticité). Ces deux éléments nécessitent des recherches plus approfondies sur l'optimisation des modèles d'apprentissage et la recherche de paramètres optimaux.

Dans un avenir proche, les ensembles de données médicales recueillis pour concevoir des solutions d'apprentissage supervisé pourraient croître de façon exponentielle et continue. Il est donc intéressant d'anticiper sur les besoins de stockage en développant des algorithmes de compression d'ensembles de données, adaptés aux applications médicales, et qui résumeraient une grande quantité de données en une fraction de cas pour une variabilité et une performance similaires. Cela pourrait être permis par des innovations théoriques en théorie de l'information de l'apprentissage profond.

Chapter 2

Introduction

This chapter introduces the motivations behind this dissertation work, and outlines the corresponding objectives. It then presents the methodology applied throughout the thesis and concludes with the organization of the manuscript.

2.1 Motivation

The scientific context in which the present study was conducted includes both clinical and methodological challenges. The underlying difficulties motivated us to define clear objectives and to build dedicated strategies to reach them.

2.1.1 Scientific context

2.1.1.1 Clinical context

Cardiovascular diseases (CVDs) rank among the top major causes of mortality and morbidity worldwide. While they accounted for 30 % of total deaths in 2008, the number of casualties is steadily growing, due in part to the aging of the population. By 2030, it is estimated that more than 20 million people will die from CVDs per year, which encourages the development of new clinical practices to improve early diagnosis [2]. As most developing cardiac pathologies affect the shape and behavior of the heart structures, non-invasive medical imaging is the primary solution to establish the diagnosis from the visualization of the heart structures and the evaluation of its contractile function.

Ultrasound imaging is currently the most widely used modality in cardiac imaging [3] as it allows a good temporal resolution at a relatively low cost. Since 3D echocardiography is still relatively new in clinical practice [4], 2D imaging remains the standard modality to measure clinical indexes such as ventricular volumes and ejection fractions from the image support. Such measurements rely on approximations of volumes from surfaces, and are highly dependent of the acquisition process. In cardiac ultrasound, the myocardial tissue appears in high intensities (white) and can be differentiated from blood-filled cavities associated to low intensities (black). The separation and identification of the different structures from accurate delineation, called semantic segmentation, is the first step to measure surfaces or volumes.

However, segmentation in echocardiography is a particularly difficult task due to the lack of clear boundaries, a low signal-to-noise ratio, the speckle texture specific to ultrasound images and the presence of numerous and complex image artifacts such as reverberations and signal dropout. The direct consequence is that embedded fully-automatic ultrasound cardiac segmentation softwares do not perform well, which forces clinicians to delineate the different contours using semi-automatic tools [5].

2.1.1.2 Algorithmic context

Manual and semi-automatic annotations are not reproducible and prone to inter- and intra-observer differences, in addition to being time consuming. In order to improve the clinical work flow, automatic cardiac segmentation has therefore been the subject of intense research over the past decades, with a strong focus on the left ventricle [3]. The segmentation in echocardiography has been traditionally addressed using either generic low-level image processing or active contour algorithms exploiting dedicated prior information to regularize the segmentation results. In this context, shape priors and constraints have proven to be especially efficient to infer missing boundaries from contextual information, while temporal smoothing is used to encourage coherent segmentation throughout the cardiac cycle.

Supervised learning methods correspond to a set of mapping algorithms whose parameters are inferred from solved cases. These methods were made popular in Computer Vision during the 90s through the development of active shape models [6], kernel Support-Vector Machines (SVMs) and Random forests (RF) [7]. However, the difficulties in obtaining medical data with expert annotations has limited their application in the medical community in general, and in echocardiography in particular.

2.1.2 Challenges

2.1.2.1 Appropriate datasets

The first challenge in echocardiography segmentation concerns the development of large annotated datasets, not only to develop supervised learning methods, but also for the evaluation of the performance of segmentation algorithms. To answer the clinical needs, the dataset has to include all the variability that clinicians face in their daily practice (pathology, image quality, acquisition material and settings...), i.e. not restricted to simulations or to real cases with good image quality [3].

Expert annotations have to be established on all relevant views and frames from a fixed and consensual protocol, ideally in a context favoring their accuracy (off-line, with an appropriate tracing software). The same structures of interest have to be annotated on all images, and clinically relevant patient information should be collected along with the exam images. Finally, to be thoroughly validated and beneficial to the community, the dataset should be publicly accessible, at least to researchers and clinicians.

2.1.2.2 Robust and fully-automatic algorithms

The second challenge consists in establishing robust and fully automatic segmentation algorithms as basis of reliable analysis tools to assist the cardiologists during exams (image analysis, clinical indices computation, acquisition guidance...). As accurate and reproducible diagnosis is required, the robustness is one of the main quality criteria, as a guarantee of reliability. The comparison of existing methods has to be performed on a large and standard dataset to encourage their adoption in clinical practice [3]. This further allows to better analyze the limitations of state-of-the-art methods, and to select the best algorithms.

Concerning supervised learning methods, error analysis from multiple algorithms on a dataset validated by the community would also help in assessing which cases are missing to improve the robustness of the learning models. The solutions should ideally be optimized for speed to be embedded inside ultrasound scanners as part of real time analysis softwares.

2.2 Methodology

2.2.1 Objectives

Due to the lack of standard datasets, the real performance of state-of-the-art methods in echocardiographic image analysis is unclear. In order to contribute to the development of a clinically suitable solution for the automatic segmentation of the heart structures in ultrasound imaging, this work aims at providing answers to the following queries:

1. What is the potential of supervised learning methods for 2D echocardiography semantic segmentation?
2. How close are we to fully automatize the cardiac analysis in ultrasound imaging?

2.2.2 Method

The complete study of the potential of supervised learning methods for echocardiographic image analysis can only be efficiently set up through a benchmarking which reports the current best performing methods on a same dataset. This requires to:

1. use a large annotated dataset;
2. establish a meaningful set of metrics, geometrical and clinical;
3. implement and adapt state-of-the-art technics;
4. evaluate them on a common evaluation platform.

The assessment of fully automatic segmentation could then be established from:

1. comparing the performance of the best method to inter- and intra-expert scores;
2. performing outlier analysis to evaluate the robustness;
3. performing error analysis to establish leads of improvement;
4. building new methods dedicated to improve the robustness and accuracy of the best segmentation methods.

The quality of the evaluation procedure was our main priority. This led us to contribute not only on algorithmic aspects, but also on the metrics used to assess the quality of the result.

2.3 Thesis organization

The manuscript is organized in four main parts composed of independent chapters that progressively address the following methodological aspects:

1. Presentation
    - Chapter 1: thesis summary, in French as requested by the doctoral school EEA, which covers every discussion of the manuscript, focusing on key points;
    - Chapter 2: introduction, in which the motivations and the methodological strategy of the thesis are presented before the detail of the organization of the manuscript.

2. Background
    - Chapter 3: echocardiography basics, which introduces the most relevant aspects of cardiac ultrasound imaging;
    - Chapter 4: state-of-the-art review, which details the segmentation dataset, methods and the evaluation in echocardiography.
3. Contributions
    - Chapter 5: CAMUS dataset properties, to our knowledge the most complete open access benchmark dataset in 2D echocardiography;
    - Chapter 6: structured random forest adaptation to 2D and 3D echocardiography segmentation (machine learning method);
    - Chapter 7: deep learning study and evaluation based on the U-Net architecture;
    - Chapter 8: enhancement of encoder-decoder models and of their evaluation through the design of anatomical metrics in 2D echocardiography;
    - Chapter 9: attention- based models design, dedicated to improve the robustness of the segmentation of 2D echocardiography images.
4. Epilogue
    - Chapter 10: conclusion, with the key achievements and the perspectives of the thesis.

The manuscript is further completed by a set of appendices:

A presentation of the collaborators of this project;

B list of publication;

C description of the computers that were used in the study;

D supplementary information on Chapter 5;

E supplementary information on Chapter 7;

F supplementary information on Chapter 8;

G supplementary information on Chapter 9;

Finally, all references are listed in the bibliography section that closes the manuscript.

Part II

Background

Chapter 3

Echocardiography

Echocardiography is a diagnostic imaging modality based on ultrasonic insonification to visualize the interior of the heart. Ultrasound imaging is the primary modality used in cardiac imaging as it is less invasive, cheapest, and fastest when compared to other non-invasive modalities. Its main drawback lies in its lower image quality, which leads to difficulties in the interpretation and processing. In this chapter, we provide general concepts for:

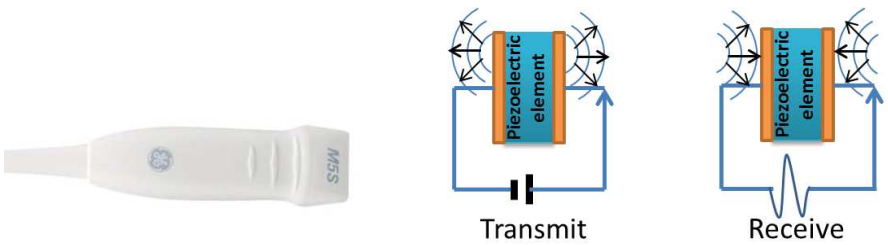
1. the generation, propagation, reception and beamforming of ultrasound waves
2. the characteristics of ultrasound images, and the different imaging modes of echocardiography
3. the estimation of clinical cardiac indices from ultrasound images

3.1 Ultrasound image formation

Ultrasound imaging is based on the transmission of ultrasound pulses to soft tissues, such as muscles and blood vessels. An image of the tissues is inferred from the returning echoes as the wave is back-scattered and reflected. The following section gives a basic explanation of the formation process of echocardiographic images.

3.1.1 Wave emission and reception

The main part of cardiac ultrasound probes (Fig. 3.1 a.) corresponds to 1D/2D phased-arrays composed of piezoelectric transducers, enabling respectively 2D/3D acquisitions. Piezoelectric crystals convert electrical energy into acoustic pressure, and the opposite.



(a) probe used for the CAMUS dataset acquisitions (b) inverse piezoelectric effect (c) direct piezoelectric effect

FIGURE 3.1: a: 2D cardiac ultrasound probe (GE M5S). b: The emission / c: reception are performed using piezoelectric materials [54].

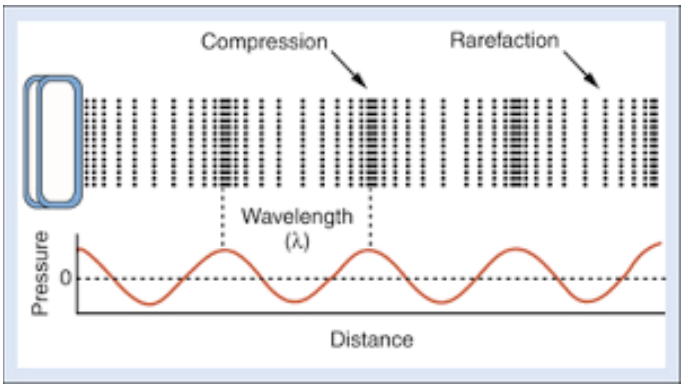


FIGURE 3.2: Longitudinal wave at a given time. As the wave propagates, particles oscillate between a compression or rarefaction state [8].

TABLE 3.1: Speed of sound in common media

Material	**Speed** (m/s)
Air	330
Fat	1450
Water	1520
Brain	1540
Blood	1570
Muscle	1580
Skin	1600
Bone	3500

Pulsed ultrasound waves are generated from small vibrations of the probe elements, induced by the application of a sinusoidal electric signal with appropriate amplitude and frequency, as shown in Figure 3.1 (c). After emission, the probe receives the reflected echoes (Fig. 3.1 b.), producing in response an electronical signal carrying information about the encountered interfaces between media of different acoustic impedance. For medical applications, frequencies typically range from 1 MHz to 15 MHz, a lot higher than the ultrasound threshold ($f_u > 20$ kHz) that defines the limit of human hearing [8]. The wavelength of the wave illustrated on (Fig. 3.2), computed as $\lambda = \frac{c}{f}$, plays an important role in the interaction between the wave and the biological tissues.

3.1.2 Wave propagation and reflection

Ultrasounds are longitudinal mechanical waves, i.e. compression waves, causing the medium particles to oscillate along the direction of the wave propagation (Fig. 3.2).
The following acoustic wave equation can be derived assuming no energy is lost during the travel in the medium:

$$\frac{\partial^2 p}{\partial t^2} = c_m^2 \times \frac{\partial^2 p}{\partial x^2} \tag{3.1}$$

where c_m is the wave propagation speed inside the homogeneous medium and p the acoustic pressure at position x and time t.

c_m depends on the tissue characteristics, i.e. its compressibility κ and density ρ:

$$c_m = \frac{1}{\sqrt{\kappa \times \rho}} \tag{3.2}$$

Table 3.1 lists the average speed of sound for some common biological tissues [55] [56]. As the average speed inside soft tissues changes only slightly with the structure, ultrasound scanners are calibrated assuming the sound travels inside the body at the constant speed $c_s = 1540$ m/s. Due to the change in impedance, specular reflection (Fig. 3.3) happens when the interface between two media is greater than the wavelength. The amount of reflected echoes, and hence the amplitude of the electric signal, depends on the nature of the structures (echogenecity) and on the angle between the probe and the interface. Assuming the speed is constant, the time travel of the wave enables to localize the interfaces. However:

- echoes are reflected and refracted following the Snell-Descartes' law, hence focused beams are used to avoid mislocating the reflectors;
- soft tissues are not homogeneous, so the speed inside a given tissue is not constant. Moreover, scattering effects (as depicted in Fig. 3.3) occur due to these inhomogeneities. Small scatterers act as spherical wave secondary sources, producing an image texture named speckle which characterizes ultrasound images;
- the ultrasound wave is attenuated as it propagates, with a coefficient that increases with the frequency and varies with the medium. Consequently, the amplitude of the back-scattered echo has to be weighted according to the depth.

3.1.3 Depth adaptation

The maximum depth ranges is about $5 - 10$ cm for pediatric cardiac sonography and $10 - 20$ cm for adult cardiac imaging [9]. To compensate for the attenuation inside the body, gain proportional to the time travel of the echoes / imaging depth is applied (Time Gain Compensation). This step is necessary to guarantee the same visibility along the insonified medium. It is traditionally performed at the acquisition level rather than the post-processing, and is either set manually or automatically.

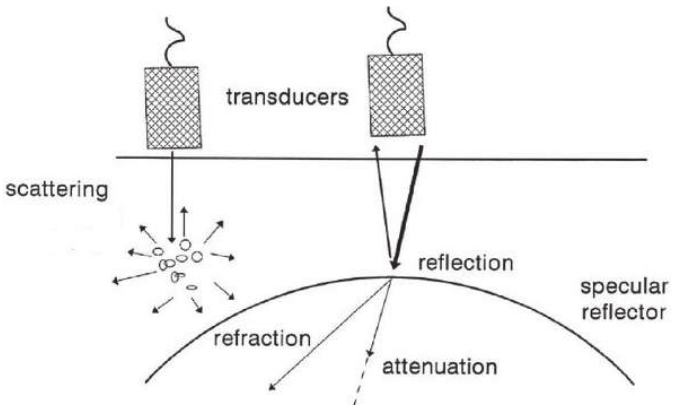
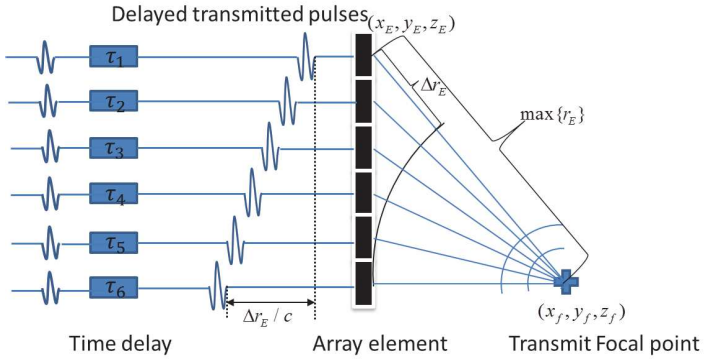
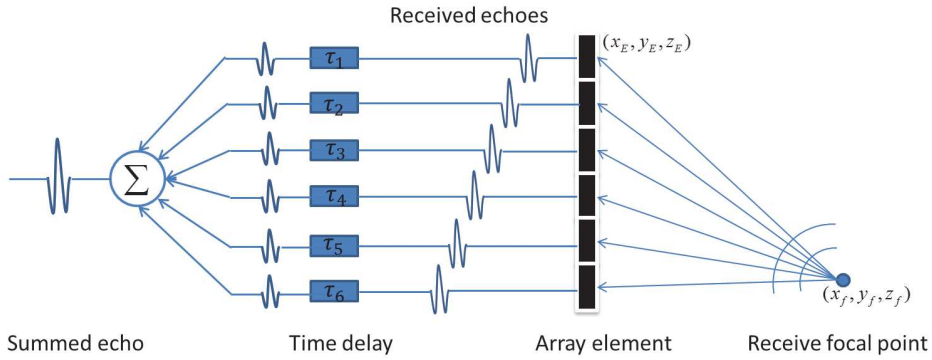


FIGURE 3.3: Interactions of the sound wave with soft tissues. Left: scattering effect. Right: Reflection, refraction, and attenuation [57].



(a) Delayed emission



(b) Delayed reception and summation

FIGURE 3.4: Focused beam onto a focal point. a: Each element receives a different delay according to the distance to the point. b: The same delays are applied on the received echoes before summation to create the radiofrequence (RF) echo signal [58].

3.1.4 Beamforming

The most common beamforming technic is called Delay And Sum (DAS). The ultrasound wavefront is oriented by the application of time delays on the different probe elements in order to create oriented focused beams (Fig 3.4 a.). Delays are also applied on the received echoes before summation (Fig 3.4 b.). The delays compensate the additional distance covered by the waves to reach each piezzo-electric element from the focused point.

To visualize the heart through the ribs in a B-mode image, the ultrasound beam is swept across a conic sector as seen in Fig 3.5, so that the ultrasound waves cover the expected field from a fixed probe position. Traditionally, several focused beams are emitted successively to locally concentrate the delivered energy (Single line acquisition) [9]. The image is reconstructed using the envelope of the beamformed signals to retrieve intensities. Logarithmic compression is used to enhance the contrast.

3.2 Image characteristics

Ultrasound images present different characteristics and levels of quality that are directly linked to the acquisition process. We briefly address in this part the notions of image resolution, contrast, and artifacts which are specific to cardiac ultrasound imaging.

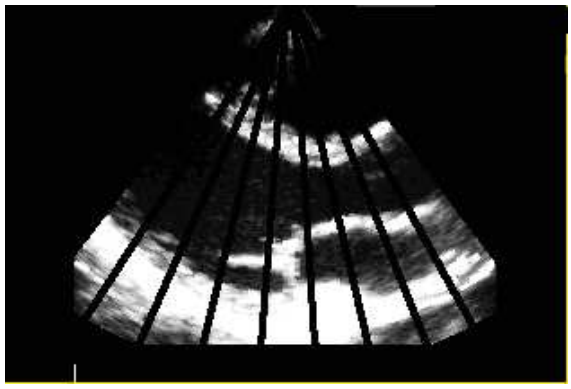


FIGURE 3.5: 2D Long axis (LAX) view of the heart. The triangular sector clearly apparent in the B-mode image is decomposed into several lines.

3.2.1 Spatial resolution

Because of the attenuation inside tissues, ultrasound imaging requires a trade-off between the reachable depth and the spatial resolution through the chosen frequency of the sound wave.

3.2.1.1 Axial resolution

The minimum distance allowing to differentiate two adjacent reflectors along the wave propagation axis is proportional to the wavelength and the pulse duration [10]:

$$d_{a_{min}} = \tau \times \frac{c_s}{2} = \frac{n}{f} \times \frac{\lambda \times f}{2} = \frac{n \times \lambda}{2} \tag{3.3}$$

where τ corresponds to the duration of a pulse and n to number of cycles of the pulse

It is therefore required to use the shortest pulse duration and the highest frequency available for a dedicated application. For instance, to image up to 20 cm, frequencies ranging from 2 to 5 MHz are used in echocardiography. With f= $2.5MHz$, for a single cycle duration pulse, the axial resolution is $d_{a_{min}} = 0.3$ mm.

3.2.1.2 Lateral resolution

The lateral resolution, corresponding to the apparent thickness of scatterers along the direction perpendicular to the beam, mainly depends on the beam width, which can be expressed as a function of the wavelength, the focal depth and the probe diameter [9], as given by:

$$d_{l_{min}} = \lambda \times \frac{d_F}{L_p} \tag{3.4}$$

where d_F corresponds to the focal depth, and L_p to the probe length.

The best lateral resolution is obtained for high frequencies and large probes. However the narrow aperture between the ribs imposes the cardiac probe length to be small, around 3 cm. For a focal point at 7.5 cm and f= $2.5MHz$, the optimal lateral resolution is: $d_{l_{min}} = 1.5$ mm. The lateral resolution worsens with depth, as the beam diverges after the focal point.

3.2.2 Temporal resolution

A new pulse can be emitted once the echoes have returned after reaching the maximum desired depth: $t_r = \frac{2 \times d_{max}}{c_s}$. Thus, the depth along with the number/density of lines condition the attainable frame rate [9]:

$$F_r = \frac{1}{t_r \times n_l} \tag{3.5}$$

with t_r the time-to-return defined above, and n_l the number of lines.

Recently, technics based on different strategies (i.e. line density reduction, multi-line acquisition, diverging waves) have been proposed in the literature to drastically increase the frame rate (up to a factor 5 with ultrafast technics), at the cost of lower image quality. In clinical practice, F_r is usually around 50 fps, twice the standard video frame rate.

3.2.3 Contrast resolution

The contrast resolution refers to the ability to distinguish dark and illuminated areas, as well as to detect variations of amplitude [10]. It is closely related to the signal to noise ratio (SNR). Among existing definitions, it can be mathematically formulated as:

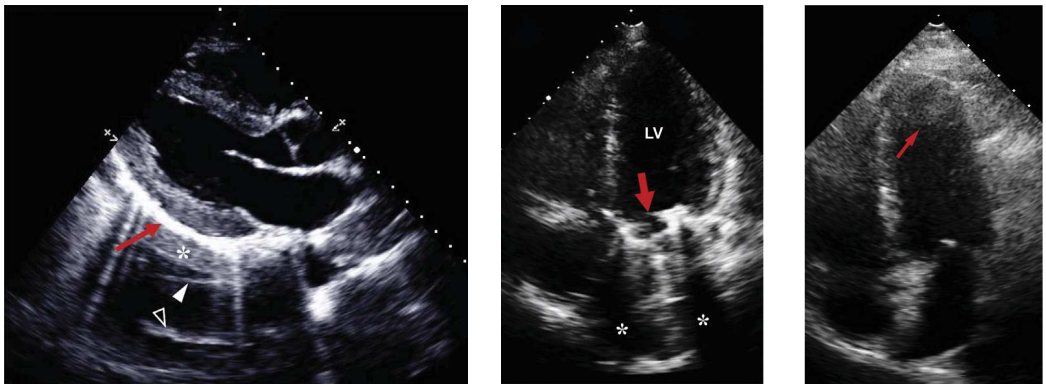
$$C = \frac{|S_A - S_B|}{S_A + S_B} \tag{3.6}$$

where S_A, S_B correspond to the intensities of the structures to differentiate.

The contrast is inherently linked to the object of interest, for instance to the echogenecity of the patient. Different contrast can be observed in Fig 3.6. Contrast can be enhanced using contrast agents, such as injections of microscopic air bubbles, or with dedicated post-processing, such as histogram normalization.

3.2.4 Artifacts

Artifacts in ultrasound imaging frequently appear in the form of duplicated, missing, badly located or distorted structures. Artifacts are direct consequences of the acquisition and image



(a) Mirror artifact below the pericardium (red arrow) and comet-tail reverberations due to the lung in a LAX image

(b) Reverberations (arrow) and acoustic shadows (asterisks) due to a mitral valve prosthesis

(c) Motionless near-field clutter created by the transducer oscillations

FIGURE 3.6: Common artifacts in echocardiography (B-mode images).

formation process and may hinder the evaluation and diagnosis [11]. Among the possible artifacts, one can cite:

1. reverberation, shadows and mirror artifacts, that result from multiple reflections;
2. refraction artifacts, due to the lens behavior of some reflectors;
3. side-lobe, beam-width and near-field artifacts, that are linked to the equipment.

Some typical artifacts that happen in cardiac ultrasound B-mode imaging are shown in Fig 3.6 (the arrows pointing out the artifacts are best seen in color).

3.3 Imaging modes

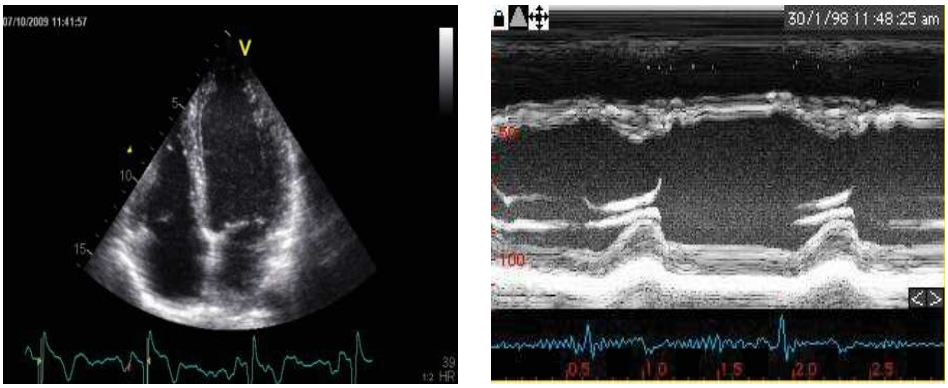
Three main ultrasound imaging modes are employed in clinical echocardiography: B-mode imaging, M-mode (or motion mode) imaging, and Doppler imaging.

3.3.1 B-mode imaging

Brightness-mode is the most common mode in clinical routine, rendering visuals as shown in Fig 3.6. As mentioned previously, it consists in scanning a section of the heart using ultrasound beams with different successive orientations. The position of a reflector on the image is estimated from:

1. the depth, deduced from the travel time;
2. the lateral position, inferred from the beam orientation.

B-mode images are displayed with a gray-scale colormap: strong reflectors, such as the valves or the muscles, appear bright while less echogeneous structures, such as the blood, appear dark. In our study, we work exclusively with B-mode images.



(a) B-mode image representing a heart section [9]

(b) M-mode representation of a line extracted from the B-mode image on the left [9]

FIGURE 3.7: B-mode and M-mode images.

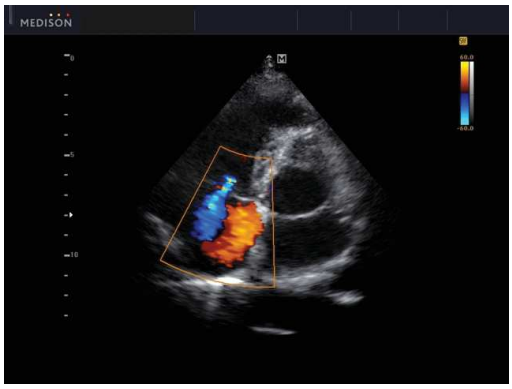


FIGURE 3.8: Example of mitral regurgitation observed with color flow Doppler imaging.

3.3.2 M-mode imaging

A M-mode image is obtained by displaying the variation of the echoes in a single line, which ensures a higher temporal sampling to observe tissues with very fast movements, such as the valves, or to perform accurate displacement measurements. However, it is difficult to align the probe perpendicularly to the structures of interest, so M-mode suffers from false measurements [13]. It is now possible to reconstruct the M-mode from the B-mode image (Anatomical M-mode) to freely position the line, but the time resolution improvement is lost. Figure 3.7 gives an example of the joint use of B- and M-mode images.

3.3.3 Doppler imaging

Doppler echocardiography is based on the Doppler principle. It allows hemodynamic evaluation of the heart[59] by revealing events like unusual obstructions or blood flows, such as valvular regurgitation (Fig. 3.8). It also allows to measure regular blood flows such as diastolic and systolic outputs and to estimate intracardiac pressures from blood velocity. The Doppler effect states that a frequency shift occurs when an ultrasound beam interacts with a moving target, such as the red blood cells [59]. The shift is related to the frequency of the transducer, the speed v of the moving object, and the angle between the ultrasound beam and the target's trajectory θ:

$$\Delta f = \frac{2 \times v \times cos(\theta)}{c_s} \times f \tag{3.7}$$

From this equation, it is easy to understand that:

1. the velocity cannot be measured if the transmitted beam and the direction of the moving target are perpendicular ($\Delta f = 0$);
2. at least two measurements are necessary to deduce v and θ from $v \times cos(\theta)$.

There are two types of Doppler imaging commonly used in clinical practice:

- continous (wave) Doppler: Mode used to display high velocities. A first crystal emits continuously while a second receives the echoes. Measured shifts relate to the entire beam (no depth localization).
- pulsed (wave) Doppler: Mode used to obtain local velocities. In a distinct region of interest, a single crystal emits several short ultrasound bursts and receives the echoes.

Aliasing occurs if the maximal shift is higher than $PRF/2$, with PRF the pulse repetition frequency that decrease with:

1. the imaging depth;
2. the width of the sample volume.

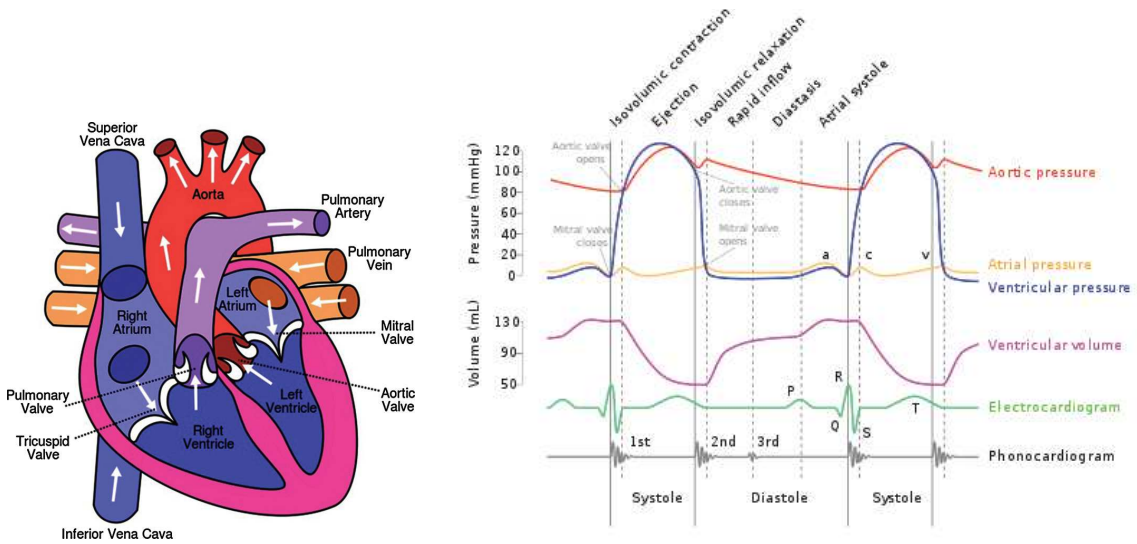
Therefore, continuous Doppler is used to measure maximal velocities, and pulsed Doppler to observe velocities in a particular sample. Color flow Doppler echocardiography displays the velocity of blood flows obtained from Pulse Wave Doppler with a color map that can be superimposed on the B-mode image. Conventionally, blood flowing away from the probe is shown in blue while blood flowing toward the probe is displayed in red. A third color, usually yellow or green, characterizes areas of accelerated or turbulent flow. Figure 2.4 gives an example of color Doppler imaging of a mitral regurgitation.

3.4 Cardiac function analysis

Medical imaging allows the non-invasive assessment of a set of complementary indices of the cardiac function, and is a routine task for cardiac diagnosis. Cardiac function analysis determines the risk of disease and leads to patient management and therapy [12]. After a short description of the heart functioning, this section introduces the main clinical indices used in echocardiography.

3.4.1 Cardiac anatomy and cycle

The heart acts as the pump of the cardiovascular system, which provides oxygen and nutrients to the whole body. It is divided into its right and left side by a wall called the ventricular septum, which is part of the myocardium, i.e. the heart muscle (in pink in Fig. 3.9 a.). The inner layer of the myocardium is called the endocardium while the outter layer is the epicardium. There are four chambers equipped with valves to control the blood flow:



(a) Schematic representation of the heart structures

(b) Wiggers diagram of a cardiac cycle. Different signals (ECG and phonocardiogram) variations are associated to cardiac events

FIGURE 3.9: Cardiac structures (a) and cycle (b).

- The right and left ventricles (RV, LV) , respectively responsible of sending deoxygenated blood to the lungs, and oxygenated blood to the whole body via the aorta;
- The right and left atria (RA, LA), which receive the blood to transfer to the ventricles respectively from the whole body and from the lungs.

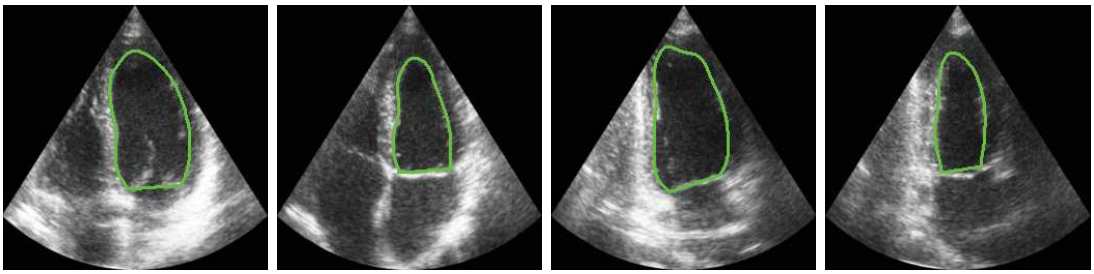
The heart cycle is composed of two main phases: a contraction, to send the ventricular blood through the arteries (systole), and a relaxation to refill the ventricles (diastole). Both events are triggered by electrical impulses, as seen on electrocardiograms (Fig. 3.9b.).

3.4.2 Global indices

The determination of ventricular volumes at end diastole (ED) and end systole (ES) (LV_{EDV} and LV_{ESV} for the left ventricle) allows to compute the stroke volumes, i.e. the volume of blood ejected by the ventricles at each beat (LV_{SV} for the LV), and the ejection fractions, i.e. the percentage of blood volume ejected (LV_{EF} for the LV). The left ventricular mass and the myocardium thickness are also measured. While Cardiac Magnetic Resonance Imaging (CMR) remains the gold standard for accurate assessment of global indices, echocardiography is the most employed modality in clinical routine because of its low cost, bedside applicability, excellent temporal resolution (real time), and absence of ionizing radiation. 3D echocardiography enables a visualization of the entire heart and would be the best choice to compute volumes. However its lower spatial and temporal resolutions compared to 2D echocardiography still limit its use in clinical practice.

In 2D echocardiography, the estimation of ventricular volumes is based on two orthogonal apical views, the 4 chamber view (A4C) and 2 chamber view (A2C), on which accurate delineation of the structures is required at both ED and ES. An example of expert annotation on the four corresponding frames is given in Fig 3.10. Endocardial contours are used to estimate the corresponding volumes with the biplane Simpson formula [13]. This method approximates the full volume as a summation of the areas A_i of elliptic disks of height h whose width and height are estimated from the A2C and A4C contours, as illustrated in Fig. 3.11.

$$V = \sum_{i=0}^{n} A_i \times h \tag{3.8}$$



(a) A4C-ED (b) A4C-ES (c) A2C-ED (d) A2C-ES

FIGURE 3.10: Apical views used to estimates the ejection fraction with the Biplane Simpson method, illustrated on Patient 206 of the CAMUS dataset (detailed in Chapter 5).

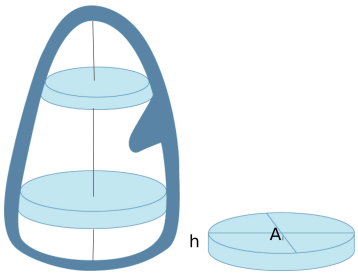


FIGURE 3.11: Monoplane Simpson decomposition of the heart's volume from A4C views. If we use the A2C, the disks are elliptic and better representative of the cavity [13].

Volume approximations are further error-prone due to the plan orientation that is manually selected by the cardiologist during the examination. The ejection fraction index is directly obtained from LV_{EDV} and LV_{ESV} as:

$$LV_{EF}(\%) = \frac{LV_{EDV} - LV_{ESV}}{LV_{EDV}} \times 100 \tag{3.9}$$

3.4.3 Local indices

Heart diseases such as myocardial ischemia and ventricular dyssynchrony may be identified and localized by the analysis of motion and deformation of the myocardium at different regions. The estimation of such indices requires not only accurate delineation of both the endocardium and epicardium during the full cardiac cycle, but also accurate tracking of the myocardial region over time. The speckle pattern inherent to ultrasound imaging can be used as natural markers as speckle remains locally stable for a few consecutive frames. Local myocardial strains curves can be further derived from the tracking field. Those curves can be used directly for cardiac function analysis, as illustrated in Fig 3.12 where the heart is divided into 17 segments as prescribed by the American Heat Association (AHA) model. Local strain indices have the potential to bring additional useful information for cardiac function analysis [60] but are still under investigation in order to ensure these measurements are reproducible in clinical context.

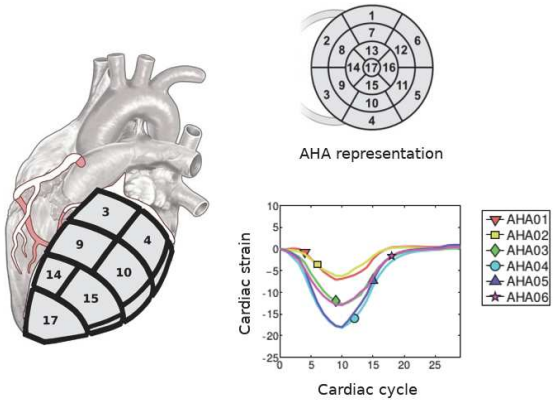


FIGURE 3.12: Cardiac function analysis through myocardial strain curves (on the right) computed for each AHA segment (on the left).

3.4.4 Daily practice and needs

Semi-automatic and manual annotation from real-time 2D echocardiography is still daily work in clinical routine for both global and local indices estimation, due to the lack of accuracy and reproducibility of fully-automatic cardiac segmentation methods. This leads to time consuming tasks prone to intra- and inter-observer variability [14]. The inherent difficulties for segmenting echocardiographic images are the following:

1. poor contrast between heart tissues and the blood pool along with brightness inhomogeneities;
2. variation in the speckle pattern along the myocardium due to the orientation of the cardiac probe with respect to tissue and presence of trabeculae and papillary muscles with intensities similar to the myocardium;
3. significant tissue echogenicity, shape, intensity and motion variability across patients and pathologies;
4. out-of-plane motion.

Any proposed fully or semi-automatic segmentation method has to be able to overcome these obstacles. In order to situate the contributions of this work, we review in the next chapter the methods that have been proposed for echocardiographic segmentation, the available datasets, and the common metrics used in the literature.

Chapter 4

State-of-the-art

The state-of-the-art in relation to our study corresponds to the segmentation methods applied in echocardiography, i.e. the identification and delineation of the different structures. The performance is usually assessed through well-established geometrical metrics but performed on distinct datasets, which makes the comparison difficult. Recently, international challenges have allowed to compare algorithms on public datasets, enabling an objective comparison of the performance of several segmentation methods. In this chapter, we review from the literature:

1. the common geometrical metrics used for segmentation evaluation in medical imaging;
2. the existing open-access datasets in cardiac segmentation;
3. the state-of-the-art methods for echocardiographic image segmentation.

4.1 Medical image segmentation metrics

Standard evaluation in medical imaging requires the establishment of task-specific sets of quality criteria designed to provide meaningful information over the degree of agreement between method predictions and expert annotations [15]. Segmentation metrics traditionally focus on the accuracy of contouring. We present here the classical metrics used in the case of binary segmentation which separates a single structure (labeled 1) from the background (labeled 0).

4.1.1 Region overlap

Overlap metrics give global scores on the quality of the image segmentation. All overall indices are derived from the four cardinalities of the confusion matrix:

1. True Positive (TP): number of elements (pixels / voxels) correctly assigned to 1
2. False Positive (FP): number of elements wrongly assigned to 1
3. True Negative (TN): number of elements correctly assigned to 0
4. False Negative (FN): number of elements wrongly assigned to 0

The sensitivity (true positive rate), specificity (true negative rate) and fallout (false positive rate) are often used to evaluate classification tasks. However, they are not truly appropriate for segmentation as they are very sensitive to objects size [61]. The overlap index between two segmentations A and B is called the Dice, and noted $D(A, B)$ thereafter. The Dice and

the Jaccard Index JAC (also called the intersection of union IOU) are less sensitive to the number of elements and more frequently used in the literature:

$$D(A,B) = 2 * \frac{A \cap B}{|A| + |B|} = 2 \times \frac{TP}{2TP + FP + FN} = 2 \times \frac{JAC}{1 + JAC} \tag{4.1}$$

In our study, we use the Dice (best score of 1) or the inverted Dice $1 - D$ (best score of 0) to represent the overall performance of segmentation methods.

4.1.2 Spatial distances between contours

Spatial distance based metrics allow a better assessment of the contouring accuracy. They are often restricted to the elements of the contours for computational efficiency. The Mean Absolute Distance (MAD) represents the average error in segmentation, while the Hausdorff Distance (HD) shows the maximum error [15]. Let $d_{C_1}(C_2)$ be the set of euclidean distances obtained from perpendicularly projecting the points of contour C_2 onto the contour C_1. To ensure a symmetric behavior, we use the corresponding formulations for MAD and HD:

$$MAD(C_A, C_B) = \frac{\overline{d_{C_A}(C_B)} + \overline{d_{C_B}(C_A)}}{2} \tag{4.2}$$

$$HD(C_A, C_B) = \max\left(\max d_{C_A}(C_B), \max d_{C_B}(C_A)\right) \tag{4.3}$$

where C_A and C_B are the sets of points of each object, d the euclidean distance, and $\overline{\bullet}$ the average operator.

The association of Dice and HD enables to encode both local and global accuracy through the measurement of the overlap and maximum error. In this study, we also indicate the MAD to represent the average closeness of the borders.

4.2 Open-access benchmark cardiac datasets

Several cardiac segmentation datasets annotated by experts have been studied in the community. Most publicly available datasets were released in conjunction with an international challenge, permitting the organizers to benchmark state-of-the-art methods. To underline the lack of public datasets in 2D ultrasound imaging, we report the different initiatives undertaken in magnetic resonance imaging (MRI) and computed tomography (CT).

4.2.1 MRI datasets

MRI imaging is characterized by a high capacity of discriminating types of tissues based on their magnetic response, enabling accurate structure delineation and diagnosis. During the last ten years, a few public datasets were built in cardiac MRI to assess the potential of state-of-the-art methods to estimate clinical indices from the segmentation, as shown in Fig 4.1. The outcome of the Kaggle challenge on volume estimation revealed that the top-performing methods relied on deep learning technics, in particular the U-Net architecture [34]. As a result, nine research groups proposed CNN-based solutions for the ACDC challenge, demonstrating that highly accurate segmentation results can be obtained in cardiac MRI with deep learning methods. Other machine learning methods, such as random forests [62], were successfully applied for pathology prediction from the derived clinical indices.

Other fully-annotated cardiac datasets have been released through different channels than during challenges:

CMRI datasets								
Name	Year	Nb Subjects		Ground truth				Active evaluation platform
		train	test	LV	RV	Myo	Pathology	
Sunnybrook	2009	45	—	✔	✘	✔	✔	✘
STACOM	2011	100	100	✔	✘	✔	✘	✘
MICCAI RV	2012	16	32	✘	✔	✘	✘	✘
Kaggle	2015	500	200	✘	✘	✘	✘	✘
ACDC	2017	100	50	✔	✔	✔	✔	✔

FIGURE 4.1: Summary and comparison of the existing cardiac MRI datasets which were released for challenges and are publicly available in 2017 [63].

- HVSMR 2016 [64];
- the Multi-Modality Whole Heart Segmentation dataset [65];
- a small dataset in [66];
- the left atrium multi-modality segmentation challenge [67];
- the UK biobank imaging study, the largest existing cardiac MRI dataset, which includes several vital organs exams and is still on-growing [68].

Although interesting, the first two datasets contain images that are clinically atypical, and the UK biobank is not free of charge and therefore has a restricted access.

4.2.2 CT datasets

CT imaging reconstructs visuals of the heart and coronary arteries using X-rays. Several public datasets of CT volumes of the human heart have been released to benchmark methods in the last ten years:

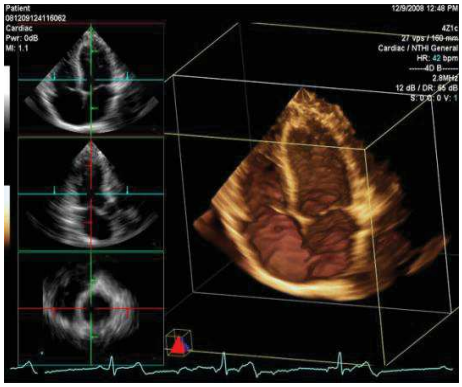
- the CVRG CT dataset composed of 25 scans [69];
- the left atrium multi-modality segmentation challenge mentioned above [67];
- the Cardiac Fat database [70] with 20 patients;
- the Multi-Modality Whole Heart Segmentation challenge dataset mentioned above, involving 60 patients (20 for training) [65].

Two of the listed datasets [67] [65] are dedicated to multi-modality studies allowing to investigate registration as a mean to improve segmentation. One dataset corresponds to a multi-structure study but involves rather few cases [65]. Though the number of patients remains quite low, there exists one multi-structure dataset. As in MRI, we observe a clear trend in building bigger datasets.

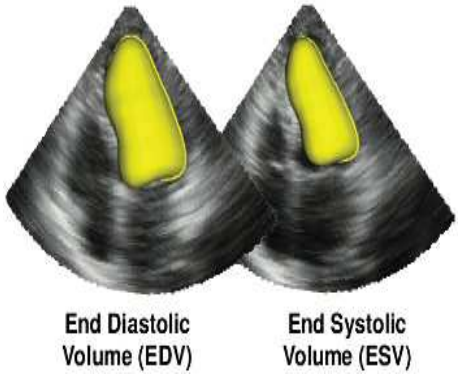
4.2.3 Ultrasound datasets

4.2.3.1 3D echocardiography

To our knowledge, there exists only one public echocardiographic dataset released in 2014 [16] and promoted during the CETUS challenge (Challenge on Endocardial Three-dimensional



(a) Example of 3D ultrasound



(b) 3D LV segmentations shown on an apical view [16]

FIGURE 4.2: Cardiac segmentation in 3D ultrasound.

Ultrasound Segmentation). This study focused on a population of 45 subjects examined in three different centers. For each individual, 3D echocardiographies were acquired over the full cardiac cycle and the LV was annotated at ED and ES (Fig. 4.2b). Among the subjects, 15 were healthy, 15 had dilated cardiomyopathy and the last third was examined at least three months after a myocardiac infarction. A balanced set of 15 patients was dedicated to a training dataset, while the rest was kept for evaluation. Two kinds of metrics were computed: geometrical scores and clinical indices.

More recently, (Dong et al., 2018) [17] built a 3D echocardiography dataset of 60 patients with the left ventricle segmented at ED and ES, but did not give public access to the data.

4.2.3.2 2D echocardiography

Prior to our initiative, there existed no public dataset for 2D echocardiographic segmentation. Among the conducted studies, the closest studies to our initiative are:

1. the work of (Carneiro et al., 2012) [18], who used a dataset of 12 patients (400 images with manual annotations for the LV) to train a deep belief network (DBN), with 2 patients (40 images) for the evaluation;

2. the experimentations in (Smistad et al., 2017) [19], in which the authors built up a dataset of 100 000 images from apical views of 100 patients annotated with an automatic algorithm;

3. the recent work of (Azarmehr et al., 2019) [20], who studied the performance of U-Net on a dataset of 61 patients (992 A4CH images) with the LV annotated by two experts.

None of these three datasets were made public, implying that the analysis of 2D echocardiography has never been thoroughly investigated via the use of a large scale public dataset.

4.3 Echocardiographic image semantic segmentation

Robust image segmentation is a necessary step to perform accurate estimation of clinical indices in echocardiography (see Section 3.4). As mentioned in Chapter 3, ultrasound images have inherent characteristics (contrast, texture, artifacts...) which hinder the segmentation

process and explain the current prevalence of semi-automatic methods in clinical routine.

In this section, we:

1. define the process of semantic segmentation and the notations used in this work;
2. present an overview of the different types of segmentation methods that have been applied in echocardiography;
3. review the unsupervised and supervised learning methods which have successfully been applied to cardiac segmentation in both 2D and 3D ultrasound imaging;
4. analyze the outcome of the CETUS challenge, which enabled a direct comparison of many state-of-the-art methods for 3D echocardiography segmentation.

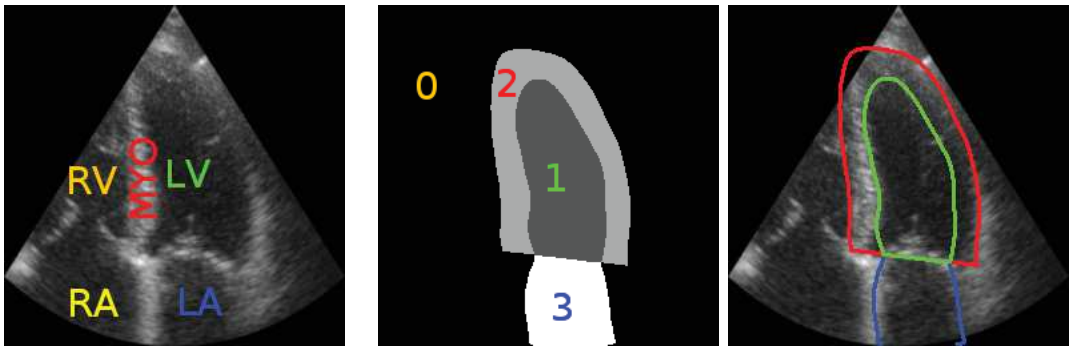
4.3.1 Semantic segmentation

Image segmentation is the task of partitioning an image into objects of interest, either by tracing contours as in Fig. 4.3 c) or by classifying each pixel as in Fig. 4.3 b). Semantic segmentation implies to identify objects by their nature, through a numeric label i.e. a class. In case of multiple occurrences of a same type of object, the labeling may be adapted to separate distinct items (instance segmentation).

In this work, the identification and delineation of the heart structures is expressed as a semantic segmentation task (values in Tab. 4.1), as we assign unique labels to the different cavities and the myocardium. Segmentation can be seen as the mapping $X \rightarrow Y$, with X is the ultrasound image as in Fig. 4.3 a) and Y a segmentation mask as in Fig. 4.3 b). As we do not necessarily search for the detailed annotation of all visible structures, any pixel that do not belong to a structure of interest (here the left ventricle LV, the myocardium myo and the left atrium LA) is associated to the background class.

4.3.2 Semantic segmentation methods overview

Many reviews list the existing segmentation methods in 2D echocardiography [18], 3D echocardiography [21] [16] or both [3] [22]. Most works focus exclusively on the endocardium



(a) 2D echocardiographic image with structures (b) Ground truth mask (c) Ground truth contours

FIGURE 4.3: Multi-structure segmentation seen as a multi-label classification problem. Supervised learning algorithms are trained to predict b) from a). In visuals, we traditionally display the contours over the image as in c).

TABLE 4.1: Labels associated to identify cardiac structures in this project

Structure	Left ventricle	Myocardium	Left atrium	Other (background)
Label	1	2	3	0

detection. As explained in the review by (Noble et al., 2006) [3] and confirmed in [22], the low image quality involved in ultrasound imaging compared to other modalities stimulated the community to come up with customized methods.

These surveys depict six main categories of ultrasound imaging segmentation technics, whose characteristics listed in Tab 4.2 include:

1. the formalism: the segmentation is based on the detection of image transitions associated to anatomical borders (boundary-driven), on the grouping of specific distributions of pixel/voxel based on similarity criteria (region-driven), or both;
2. the use of prior knowledge: pre-established information on the shape, location, or texture of anatomical regions is added to constrain the segmentation;
3. the temporal consistency: a coherent evolution of contours through time sequences is obtained with tracking, spatio-temporal constraints, or smoothing post-processing;
4. supervised learning: the optimization of the parameters of a pre-established model is guided by a dataset of solved cases.

This table reveals that:

- Boundary- and region- driven approaches are equally popular, leading to the development of hybrid approaches;
- Priors are frequently incorporated into frameworks, proving the interest of such schemes for ultrasound applications;
- Improvement of baseline methods is often obtained by encouraging coherent segmentation over the cardiac cycle;

TABLE 4.2: Characteristics of segmentation methods in echocardiography, inspired from the review of (Carneiro et al., 2012) [18]

Method	Boundary	Region	Prior	Time consistency	Super-visation
Bottom-up	✓	✓		✻	
Active Contours	✓	✻	✻	✻	
Level Sets	✓	✓	✻	✻	
Deformable templates	✓	✓	✓	✻	
Shape models	✓	✻	✻	✻	✓
Machine / Deep learning	✻	✓	✻	✻	✓

✓: *Property inherent to the original formalism*
✻: *Property added to the original formalism*

- There is a trend to derive models from expert annotated datasets.

A brief description of each method category is given in the following. We chose to cluster them according to their non-supervised or supervised nature.

4.3.3 Non-supervised learning methods

Unsupervised learning methods do not require training data. They rather involve prior knowledge in the form of initialization, shape constraints or pre- and post-processing to guide the segmentation. Deformable models (snakes, level-sets, deformable templates) have been predominant in echocardiography compared to other technics such as clustering (k-means, watershed), region growing and probabilistic methods (graph cuts, Markov Random fields).

4.3.3.1 Bottom-up technics

Bottom-up algorithms rely solely on low-level image information to detect anatomical borders. These technics involve combinations of classic morphological region operators (dilation, erosion), edge detectors (gradient filters, Hough transforms), and simple mathematical models (circle, ellipse). They are computationally efficient but greatly lack in robustness with regards to varying imaging conditions [18].

(Zhang et al., 1984) [71] proposed to partially tackle this issue by using pre-processing (thresholding) and temporal smoothing to segment the LV in 2D sequences, while (Klingler et al., 1988) [72] used morphological operators on a denoised time-averaged cycle to identify the LV. The semi-automatic segmentations of the LV were highly correlated (r=0.93) to the manual segmentation provided by one expert on seven canine hearts.

4.3.3.2 Active contours

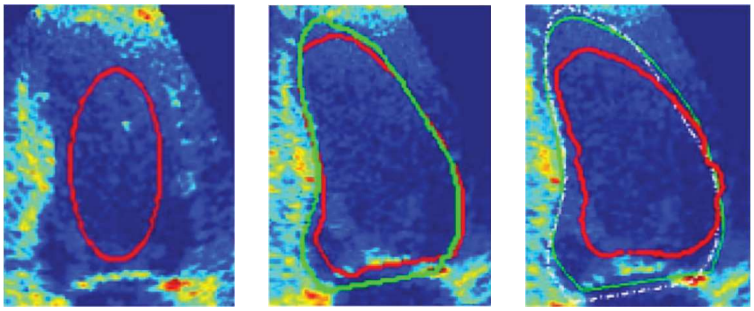
Active contours, also called snakes, are variational methods that correspond to the iterative fitting of an elastic contour to minimize a set of energy terms E_{snake} [73]:

$$E_{snake} = E_{int} + E_{ext} \tag{4.4}$$

with E_{int} an internal energy term and E_{ext} an external energy term.

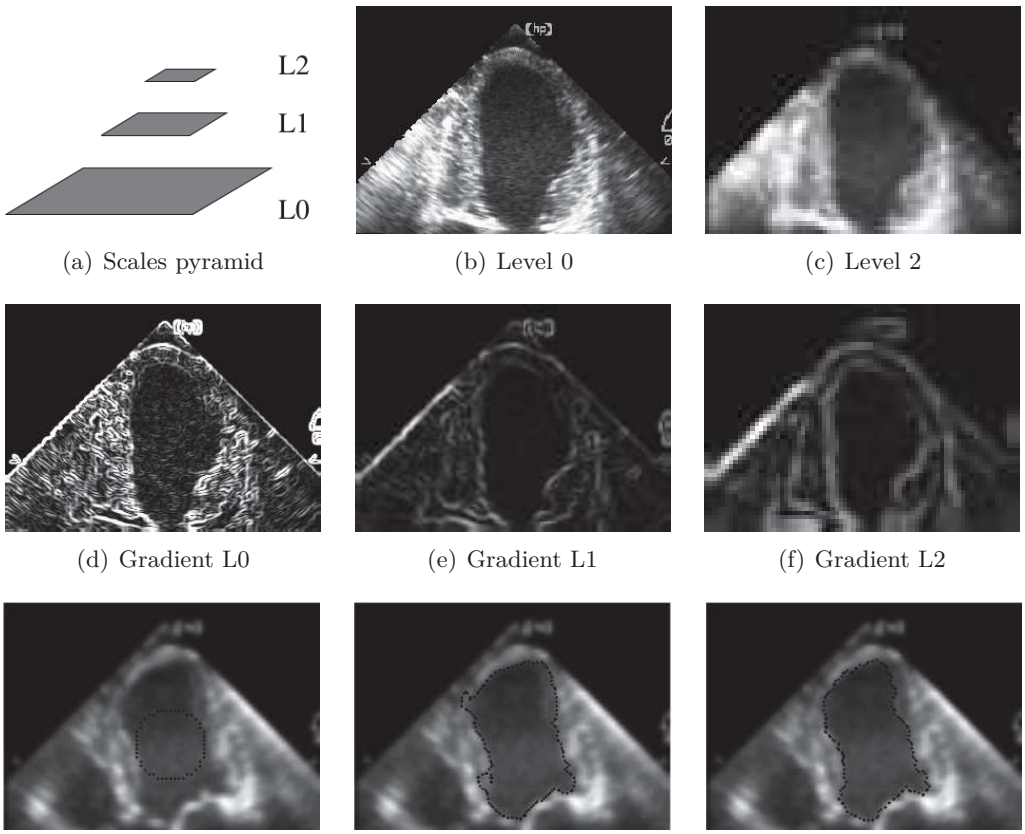
The internal energy imposes on the model a smoothness constraint (regularization term), while the external energy term pushes the contour towards edges or other image features (data-fidelity term) [74]. The design of specific features, contour models, and energy terms dedicated to ultrasound imaging has made active contours a popular research topic for echocardiographic segmentation [3]. The main limitation of active contours is their strong dependence on initialization.

In (Mishra et al., 2003) [75], the optimization was carried by a genetic algorithm from a bottom-up initialization. Their fully-automatic method produced an area correlation of 0.92. Shape and intensity priors were added in (Chen et al., 2007) [76] as energy terms. The contour evolved from an elliptic initialization and its convergence was strongly sensitive to the hyper-parameter balancing the external energy and the shape prior, as shown in Fig. 4.4.



(a) Initialization (b) Enhanced data-term (c) Enhanced shape constraint

FIGURE 4.4: Illustrations from the active contour method in Chen et al. [76], with the prediction in red and the ground truth in green. From a same initialization (a), enhancing the data-term (b) allows for a much better result than an enhanced shape constraint, which tends to shrink the LV (c).



(a) Scales pyramid (b) Level 0 (c) Level 2

(d) Gradient L0 (e) Gradient L1 (f) Gradient L2

(g) Initialization (h) Without region homogeneity constraint (i) With region homogeneity constraint

FIGURE 4.5: Illustrations from the level-set approach from (Ning et al., 2002) [77]. First rows: Multi-scale data-terms obtained by applying edge detection on the blurred downsampled image. Last row: Examples of segmentation results.

4.3.3.3 Level Sets

Level Sets are similar to active contours except they rely on an implicit representation of the contour Ω. C_{LS} is expressed as the zero-level of a distance function $\Phi : \Omega \rightarrow \mathbb{R}$:

$$C_{LS} = \{x \in \Omega \,|\, \Phi(x) = 0\} \tag{4.5}$$

The driving forces are functions of Φ, which inherently allows automatic topological changes such as splitting and merging, and reduces the sensitivity to initial conditions [22]. Concerning their application to echocardiographic segmentation, the incorporation of priors and the reduction of search dimensionality have been the most investigated research axes [18].

A multi-region level-set relying on statistics of the radiofrequency signal and two manual initializations has been proposed for echocardiographic structures in (Bernard et al., 2007) [78]. (Cremers et al., 2006) [79] investigated the incorporation of a statistical shape prior invariant to translation and scale. A combinative level-set was applied successfully to segment the LV in 3D echocardiography in (Ning et al., 2002) [77], showing the benefit of combining region and edge constraints as well as multiple scale processing in ultrasound imaging (Fig. 4.5). The automatic initialization is derived from gaussian-blurred L2 scales (Fig. 4.5 f.). The fully-automatic algorithm provided MAD scores ($MAD = 1.64 \pm 0.50$ mm) below the inter-variability of 3 experts on 24 apical images ($MAD = 1.86 \pm 0.67$ mm).

Active Geometric Functions, as presented in (Duan et al., 2010) [80], allow to reduce the search space while keeping an implicit formulation, leading to real-time applicability. On 35 4D canine sequences containing 425 frames, the authors reported a MAD value of 4.00 ± 3.23 mm while the two experts' inter-variability was assessed to be 4.23 ± 3.26 mm.

4.3.3.4 Spatio-temporal analysis

Nevertheless, echocardiographic segmentation is often approached as a spatio-temporal problem to ensure the robustness of the estimation of clinical indices (2D+t, 3D+t) [3].

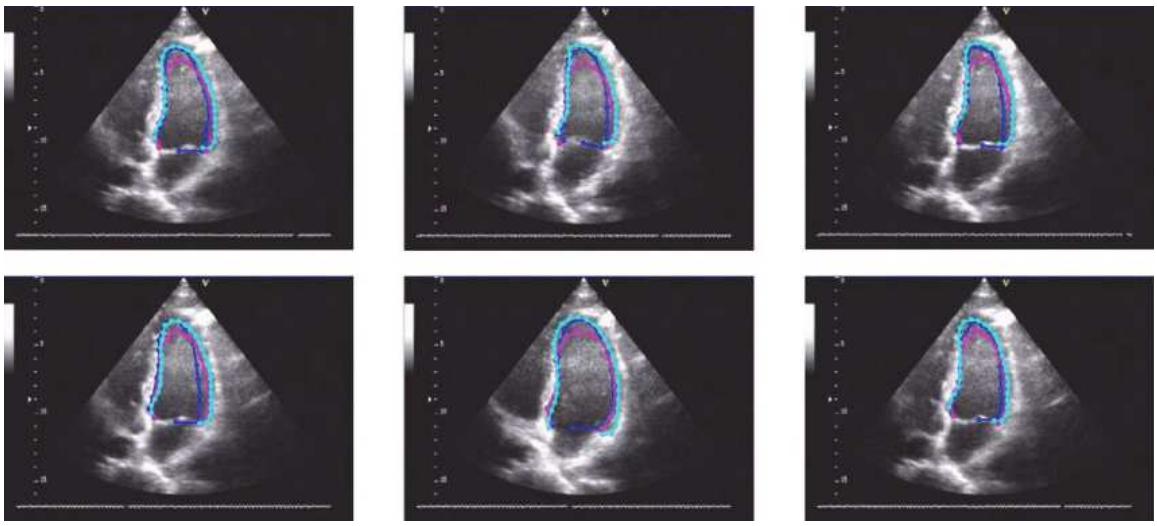


FIGURE 4.6: Illustrations of the deformable model from [81]. The groundtruth is in blue while the proposed tracker's prediction is in cyan and the comparative tracker in magenta.

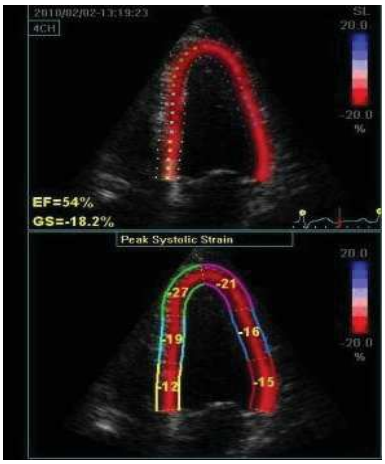


FIGURE 4.7: Myocardium segmentation for strain estimation with speckle tracking [89].

Markov Random fields [82], optical flow [83], and motion tracking from local phase features [84] have been investigated to provide segmentations of the LV coherent along time sequences. Deformable templates associated to tracking algorithms proposed a strong alternative to level-sets by restricting non-affine transformations of a pre-defined prototype shape. Using probability maps as data-fidelity terms, the deformable template in (Mignotte et al., 2001) [85] obtained TP rates on the LV slightly below inter- and intra-observer scores on 50 2D short-axis cycles (78.9% VS 84.5% and 88.3%, respectively). The solution was constructed from random initializations, making for a fully-automatic pipeline. In (Nascimento et al., 2008) [81], a deformable model accounting for outlier points was associated to a Kalman filter. Visuals on several frames are shown in Fig. 4.6 for a good case. Hausdorff distance values remained high (above 17mm), especially on the lateral wall of A4C views.

Endocardial and epicardial borders were simultaneously tracked in (Jacob et al., 2002) [86]. Recently, speckle tracking of the myocardium has showed great potential in myocardial segmentation for strain estimation [87] [88].

4.3.4 Supervised learning models

Supervised learning methods are trained to replicate expertise by optimizing model parameters on solved cases. The derived models should be able to encode the relevant information from the training dataset and to improve with further cases. Active Shape Models have been integrated into echocardiographic segmentation pipelines since the 1990s in semi-automatic pipelines, i.e. needing manual input and/or refinement. For fully automatic solutions however, the current trend favors machine and deep learning algorithms.

4.3.4.1 Active Shape models

Similar to deformable templates, active shape models (ASM) [6] define a space of evolution for the segmentation contour points. The main difference is that the prototype and the allowed deformations are computed from solved cases. First, all shapes are registered, based for instance on manually annotated stable landmarks. The shape space is then built around the mean shape as a statistical model in which deformations follow gaussian distributions.

The main modes of variations can therefore be inferred by applying Principal Component Analysis (PCA), and limitations be placed on the magnitude of deformations from the mean shape. After registration to the shape space, image information is used to fit a plausible shape with regard to the training set.

In (Paragios, 2015) [90], the authors proposed to build two ASM to segment the left ventricle with dedicated shape spaces at ED and ES. Interestingly, 5 modes of variations were sufficient in (Hamarneh et al, 2000) [91] to explain 95 % of the LV shape in 2D echocardiography, for a dataset of 105 images. A Kalman filter was associated to a shape model for the segmentation of both the endocardium and epicardium in (Jacob et al., 2002) [92], relying on a manual initialization. More recently, in the B-spline explicit active surface model (BEASM) method of (Pedrosa, 2017) [38], a statistical shape model established on MRI data was used to regularize the B-spline coefficients of an active contour method in 3D echocardiography. The segmentation obtained at ED was then propagated to ES using localized anatomically constrained affine optical flow, resulting in MAD values of 1.81 ± 0.59 mm and 1.98 ± 0.56 mm respectively at ED and ES.

4.3.4.2 Active Appearance models

In order to guide the iterative fitting of the controlled shape, [23] proposed to associate an intensity prior to the shape prior, resulting in active appearance models (AAM). The intensity distribution prior is defined similarly as gaussian modes of variation. This approach proved especially useful in ultrasound imaging where the specific intensity patterns were modeled to increase the accuracy of the segmentation [3].

In (Bosch et al., 2002) [93], the authors built AAM on the normalized 2D sequences of 65 patients to generate time continuous segmentation over the full cardiac cycle on A4C views,

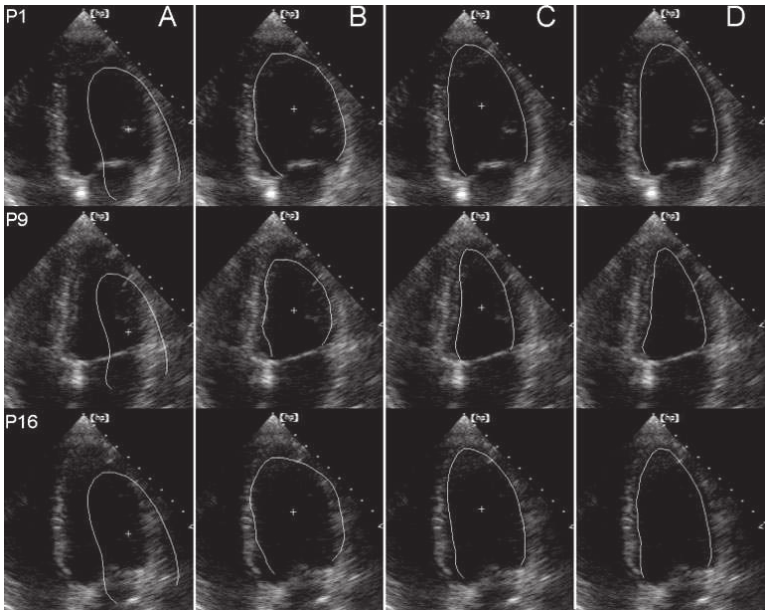


FIGURE 4.8: Visuals from the AAM of (Bosch et al., 2002) [93] on 3 frames of a single sequence. A: initialization, B: AAM position after 5 iterations, C: AAM position after 20 iterations, D: ground truth.

which proved more accurate than a classical AAM. Especially, the average accuracy on 592 points forming the contours was reported to be of 3.35 ± 1.22 mm, below the average inter-observer variability of 3.82 ± 1.44 mm defined from 2 experts. (Van Stralen et al., 2015) [94] used time-instant dedicated AAM trained by extending the CETUS dataset with 25 additional cases annotated by a different expert, still showing improvement over the baseline.

The greatest difficulty of such frameworks is to design an efficient mapping between the image and the shape space. Furthermore, as mentioned in (Noble et al., 2006) [3], ASM and AAM can only be as good as the dataset they are derived from. This holds especially true for medical applications where pathologies will be associated to irregular shapes.

4.3.4.3 Random forests

Random decision forests (RF) consist in an ensemble of decision trees trained on annotated data using bootstrap aggregating (bagging) to avoid over-fitting, i.e. each tree learns on a random subset of the available data. Trees are built as successions of heuristic binary decisions on hand-crafted features that route the data to one branch or the other, based on an information gain criteria. The RF framework is extremely flexible and generic, because the extremities, called leaves, can store any type and quantity of information over the task at hand, often classification or regression. Also, the testing phase is very fast with parallel computing as each independent tree only performs a handful of thresholding operations to provide a solution. As multi-structure segmentation can be approached as a multi-class classification problem, RF are an interesting solution for medical image segmentation [24].

In echocardiography, (Lempitsky et al., 2009) [95] used random forests to segment the myocardium on 14 3D echocardiographic systolic frames, reaching a 92% TP rate. At the CETUS challenge in 2014, three teams proposed RF frameworks: (Keraudren et al., 2014) [96] proposed a fully-automatic solution able to segment a volume in 90 seconds, based on a cascade of RF trained using auto-context, i.e. the features of the forests were augmented with the prediction of the previous one. In (Milletari et al., 2015) [97], Hough forests learnt to predict the position of the center of the LV in addition to the class (LV or background) of the voxels based on a patch approach. Once the center was established, voxels from the endocardial region that predicted a close LV position and stored close intensity patches were used to obtain a denoised segmentation from the associated segmentation patches. This fully automatic pipeline took 40 seconds to annotate one volume.

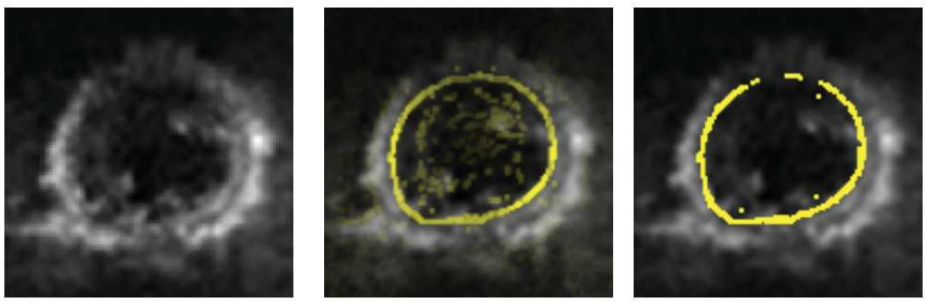


FIGURE 4.9: Edge maps from the SRF in (Domingos et al., 2014) [33]. The original slice is on the left, the SRF prediction in the midle, and the refined version using non maximum suppression on the right.

(Domingos et al., 2014) [98] used Structured Random Forests (SRF) to predict endocardium probability edge maps on short-axis slices. The probability maps were then used as the data-driven term of a continuous explicit surface model to regularize the predicted shape. As the model required a manual initialization, the full pipeline was semi-automatic. This method produced accurate results, with a MAD of 2.09 ± 0.68 mm and 2.20 ± 0.71 mm at ES. Edge maps derived from SRF were also successfully applied for myocardium segmentation in [33].

4.3.4.4 Neural networks

Artificial neural networks (ANN) are models that learn a mapping as a set of layers, usually more than 2 which is called Deep Learning (DL). Multi-layer networks are trained iteratively using back-propagation of the error [25]. By stacking layers, neural networks are able to capture complex mappings directly built from the image, or part of the image. We here distinguish fully-connected neural networks from convolutional neural networks.

Fully-connected neural networks In multi-layer perceptrons, also called fully-connected neural networks, each layer is composed of several perceptrons, i.e. unit whose output is the result of passing a weighted sum of input units through an activation non-linear function. In (Binder et al., 1999) [99], a 2 layers network was trained on patches extracted from 8 patients to differentiate tissue and blood pool regions. Spatial temporal contour linking was used on the boundary candidates elected from the segmentation to contour parasternal short axis frames at ED and ES. Interestingly, the authors balanced poor quality, medium quality and good quality images in the test set and reported good correlations even for poor quality images. (Carneiro et al., 2012) [18] used a combination of two multi-layer perceptrons (also called deep belief networks) for the segmentation of the endocardium on 2D A4C acquisitions. The first was trained to predict whether an image patch contained the full LV or not while the second performed the contour extraction. Their method was trained on 400 images from 12 different patient sequences with various pathologies and tested on 50 images from 2 healthy subject sequences. They obtained an average HD of 18 mm and an average MAD of 8 mm.

Convolutional neural networks Convolutional neural networks (CNNs) grew to be state-of-the-art in image processing on multiple tasks, including in medical imaging [26]. They are composed of convolutional layers that apply local filtering, often with the addition to regularization and normalization layers.

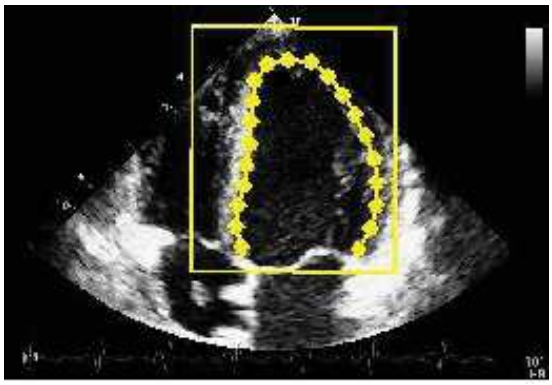
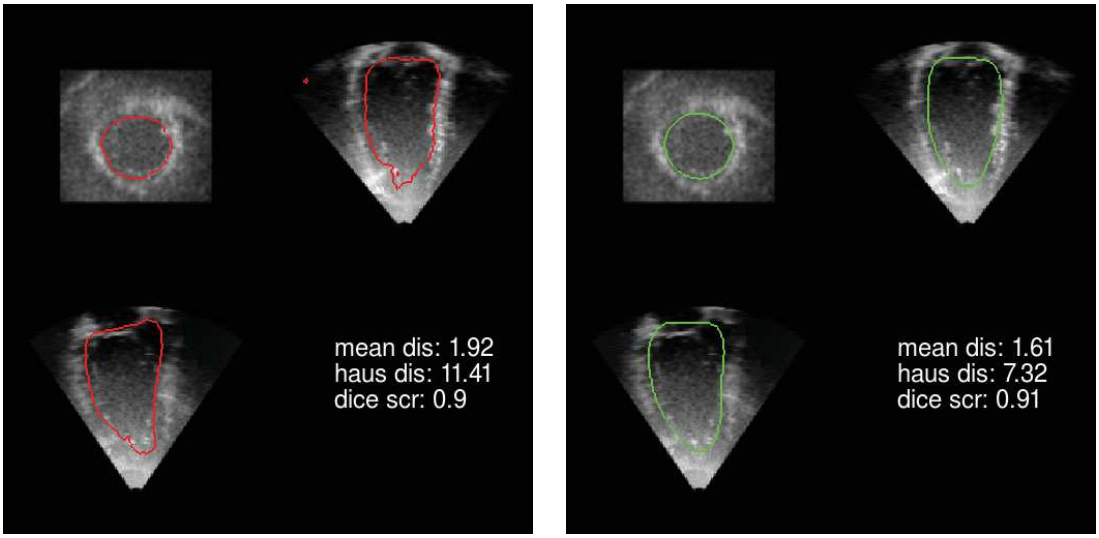


FIGURE 4.10: LV localization and segmentation from the deep learning approach in (Carneiro et al., 2012) [18].



(a) Original prediction (b) ACNN prediction

FIGURE 4.11: Regularized CNN segmentation in 3D echocardiography [42].

Conveniently, CNNs only need to store the filter parameters, which allowed to augment the size of the input when compared to MLP.

In (Smistad et al., 2017) [19], the authors showed that the U-Net architecture [34] could be trained to successfully segment the LV in 2D ultrasound imaging. However, due to lack of annotations, the network was trained with the output of a state-of-the-art deformable model segmentation method. On a manually segmented test set, the results revealed that the network and the deformable model obtained the same accuracy, with a Dice score of 0.87. The U-Net architecture was also very recently used in (Azarmehr et al., 2019) [20] on a manually annotated dataset of 61 patients, providing dice scores of 0.92 ± 0.05 mm and HD scores of 3.97 ± 0.82 mm, better than the reported 2 expert inter-variability of $D = 0.88 \pm 0.06$ and $\text{HD} = 4.50 \pm 0.87$ mm. In the Anatomically Constrained Neural Networks (ACNN) of (Oktay et al., 2017) [42], a shape constraint was derived implicitly from an auto-encoder and directly incorporated in a encoder-decoder framework for the segmentation of the CETUS dataset. The results obtained from a small dataset of 15 patients, with MAD values of 1.9 and 2.1 mm at ED and ES, reveal the strong potential of CNNs to analyze echocardiographic images.

4.3.4.5 Conclusion

In this part, we reviewed the main types of methods that have been applied to echocardiographic segmentation. It appears from this overview that the addition of priors (intensity distributions, shape constraints and temporal behavior models) in the frameworks of both unsupervised and supervised methods has been key to obtain good results in ultrasound imaging [3], probably to account with the lack of clear boundaries. It also clearly reveals that the lack of data, especially publicly available data, has for long limited the scope of evaluation. For instance, several studies [77] [85] [93] [20] reported results below the average expert-variability, but on different settings (dataset, expert, center, population, pathologies), which makes which makes difficult the direct comparison of their approaches.

4.3.5 State-of-the-art algorithms from the CETUS challenge

The CETUS challenge has been the only initiative to propose a public dataset and evaluation platform in order to fairly compare algorithms. The analysis of the results allows to establish the most promising methods for echocardiographic segmentation.

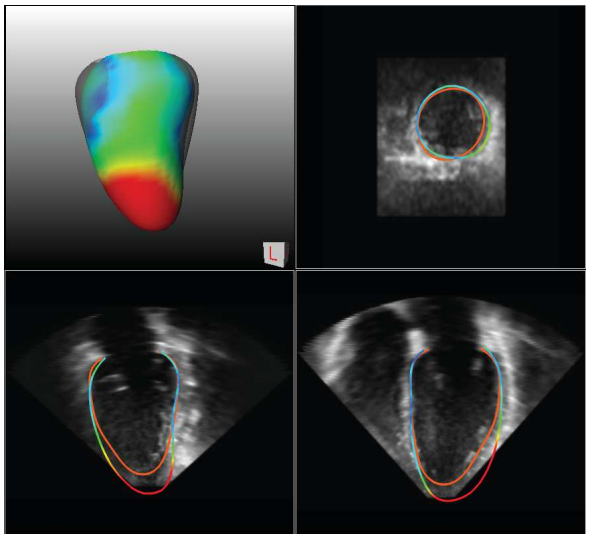
4.3.5.1 Algorithms

9 methods were evaluated during the CETUS challenge, 5 being fully-automatic:

- the B-Spline Explicit Active Surface (BEAS), which consists in an active contour method where the contour is decomposed as a set of continuous splines [37]. It was extended to 3D and integrated in the fully-automatic pipeline of (Barbosa et al., 2014) [100] to provide the segmentation of the ED volume. The ES frame was segmented based on a tracking solution using optical flow and local block matching;
- the auto-context forests of (Keraudren et al., 2014) [96] described in Section 4.3.4.3;
- the Hough forests of (Milletari et al., 2015) [97], described in Section 4.3.4.3;
- the kalman filter framework of (Smistad et al., 2014) [101], which defines the LV as a mesh model with 21 transformation parameters estimated using edge detection and an extended Kalman filter, allowing non-linear transitions over time;
- the AAM of (Van Stralen et al., 2015) [102] described in Section 4.3.4.1.

The 4 other methods were semi-automatic:

- the graph-cut method proposed in (Bernier et al., 2014) [103] used a manual initialization involving three landmarks to derive a T bar from the apex to the base. The



(a) AAM prediction

(b)

FIGURE 4.12: Average result from (Van Stralen et al., 2015) [102], color-coded in function of the distance to the ground truth in orange.

image was then projected onto a polar coordinate system to build a graph where each node was associated to a voxel and each edge to a gradient-based energy term. The LV was assumed to be shaped in U and delineated from the background with a graph-cut procedure that forced the separation to pass through the three annotated landmarks;

- the surface model using edge detection SRF in (Domingos et al., 2014) [98], described in Section 4.3.4.3;
- the multi-atlas approach of (Oktay et al., 2014) [104], rarely applied in ultrasound imaging as it requires an accurate registration between the target and the atlas image. To tackle this issue, the authors proposed to use as feature a shape representation built from mapping images patches to a segmentation patch dictionary using a sparse coding manifold. The method used a speckle reduction pre-processing step and three initialization points: the apex, the center of the LV and the mitral valve;
- the fast level-set approach in (Wang et al., 2014) [105] guided by multi-scale features computed with quadrature filters that perform edge and ridge detection.

The winner of the challenge was ultimately the fully-automated BEAS, which suggested the possibility to provide robust solutions for ultrasound segmentation without user input, and hence reproducible. Since the challenge, except our work presented in Section 6.4, four teams reported better results than the BEAS on CETUS:

- the BEASM [38], as described in Section 4.3.4.1;
- the association of AAM and tracking in [94], where the ES segmentation is obtained from the AAM segmentation at ED by tracking the LV;
- the tracking method in (Queirós et al., 2017) [106] based on localized anatomically constrained affine optical flow (LACAOF) from BEAS segmentation at ED;
- the ACNN of (Oktay et al., 2017) [42], as described in Section 4.3.4.4.

4.3.5.2 Geometrical results

Tab. 4.3 summarizes the results obtained on the geometrical metrics, i.e. the Dice, the MAD and the HD at ED and ES. The inter-expert variability was established from a consensus between three experts. Three main conclusions can be drawn from this table:

1. using dedicated pre-processing and relevant priors allows generic methods to perform well in ultrasound, as observed from the scores of (Oktay et. al, 2014) [104] and (Bernier et al., 2014) [103];
2. RF frameworks worked well even on a small training dataset (15 patients), especially in the pipeline of (Domingos et al., 2014) [98] which produced the overall best MAD scores;
3. it is possible to achieve high accuracies using fully-automatic methods to segment the LV, as the performances of all methods are quite close. Moreover, the winner of the challenge and the current best performing methods are both fully-automatic.

TABLE 4.3: Segmentation scores for the 14 evaluated methods on the test set of the CETUS dataset (30 patients). The inter-expert variability is given prior to fully-automatic methods, semi-automatic methods, and algorithms proposed after the challenge. The best scores in 2014 are given in blue while current ones are shown in bold.

Method	ED			ES		
	D	**MAD**	**HD**	**D**	**MAD**	**HD**
	val.	mm	mm	val.	mm	mm
inter-obs	0.931 ±0.021	1.4 ±0.4	4.7 ±1.27	0.920 ±0.021	1.3 ±0.3	4.7 ±1.15
BEAS [100]	0.894 ±0.041	2.3 ±0.7	8.1 ±2.7	0.856 ±0.057	2.4 ±0.9	8.1 ±3.1
Auto-context RF [96]	0.870 ±0.048	2.4 ±0.9	9.0 ±3.1	0.842 ±0.057	2.5 ±0.7	9.1 ±3.2
Hough forests [97]	0.893 ±0.031	2.1 ±0.7	8.2 ±3.9	0.838 ±0.062	2.9 ±1.0	8.5 ±2.3
Kalman filter [101]	0.885 ±0.038	2.6 ±0.9	8.3 ±3.0	0.844 ±0.050	2.9 ±0.9	9.0 ±3.0
AAM [102]	0.879 ±0.054	2.4 ±0.9	8.4 ±3.5	0.835 ±0.079	2.8 ±1.2	8.6 ±2.8
Graph-cut [103]	0.882 ±0.029	2.4 ±0.6	9.4 ±2.6	0.837 ±0.047	2.6 ±0.6	9.3 ±2.1
SRF [98]	0.894 ±0.038	2.1 ±0.7	9.3 ±3.9	0.871 ±0.050	2.2 ±0.7	8.3 ±2.7
Multi-atlas [104]	0.894 ±0.033	2.2 ±0.7	7.5 ±1.8	0.849 ±0.049	2.5 ±0.7	8.6 ±3.0
Level-set [105]	0.815 ±0.042	2.5 ±1.0	9.0 ±3.6	0.841 ±0.057	2.7 ±0.3	9.1 ±3.3
LACAOF [106]	0.894 ±0.040	2.3 ±0.7	8.1 ±2.6	0.861 ±0.054	2.4 ±0.8	8.2 ±3.0
Tracking AAM [94]	**0.910** -	1.9 ±0.8	6.7 -	0.862 -	2.5 ±0.8	7.4 -
ACNN [42]	0.904 ±0.020	1.9 ±0.5	7.0 ±2.0	0.874 ±0.040	2.2 ±0.6	7.5 ±2.2
BEASM [38]	0.909 ±0.034	**1.8** ±0.6	**6.3** ±1.7	**0.875** ±0.046	**2.0** ±0.7	**6.9** ±2.1

TABLE 4.4: Accuracy on clinical indices of the 14 evaluated methods on the test set of the CETUS dataset (30 patients). The inter-expert variability is given prior to fully-automatic methods, semi-automatic methods, and algorithms proposed after the challenge. The best scores in 2014 are given in blue, current ones in bold.

Method	LV_{EDV}		LV_{ESV}		LV_{EF}	
	corr	**LOA**	**corr**	**LOA**	**coor**	**LOA**
	val.	ml	val.	ml	val.	ml
inter-obs	0.985	-3.0 ± 11.1	0.993	-1.9 ± 6.5	0.952	-0.1 ± 3.3
BEAS [100]	0.965	-5.0 ± 17.7	0.967	-6.8 ± 13.9	0.889	2.9 ± 5.2
Auto-context RF [96]	0.921	15.9 ± 24.6	0.952	-6.2 ± 16.6	0.719	12.1 ± 10.6
Hough forests [97]	0.953	5.1 ± 19.0	0.960	-1.9 ± 6.5	0.745	15.2 ± 7.6
Kalman filter [101]	0.941	-10.1 ± 19.4	0.964	-11.3 ± 14.6	0.879	3.7 ± 5.2
AAM [94]	0.966	-15.4 ± 16.0	0.964	-13.2 ± 14.4	0.611	3.7 ± 8.8
Graph-cut [103]	**0.979**	2.7 ± 13.9	0.968	2.2 ± 13.7	0.811	0.1 ± 7.8
SRF [98]	0.917	8.7 ± 25.0	0.956	-5.2 ± 15.9	0.819	8.3 ± 7.2
Multi-atlas [104]	0.945	-6.0 ± 20.8	0.924	-0.4 ± 20.6	0.780	-1.5 ± 6.9
Level-set [105]	0.927	2.0 ± 23.8	0.956	-3.9 ± 16.1	0.881	3.5 ± 5.2
LACAOF [106]	0.965	-4.99 ± 17.7	0.971	5.83 ± 13.1	**0.927**	2.3 ± 4.20
Tracking AAM [94]	0.958	-4.9 ± 18.1	0.965	-15.4 ± 15.1	0.751	8.4 ± 7.7
ACNN [42]	0.961	-4.1 ± 17.3	**0.973**	-3.5 ± 13.6	0.892	0.48 ± 5.50
BEASM [38]	0.953	-3.3 ± 19.0	0.960	-4.8 ± 16.1	0.911	1.7 ± 5.2

4.3.5.3 Clinical indices

Tab. 4.4 summarizes the results obtained for all methods on clinical indices, i.e. the LV volume at ED and ES and the ejection fraction for which the correlation *corr* and limit of agreement *LOA* ($mean \pm 1.96 \times std$) are reported.

Two main conclusions can be drawn from this table:

1. the clinical indices obtained with RF frameworks were not so accurate, showing high biases and limits of agreement on all indices;

2. algorithms involving tracking [100] [38] [106] [101] showed to be particularly accurate for the estimation of the LV_{EF}, as well as a few methods that do not involve temporal coherency but do include shape constraints or priors [103] [42];

4.3.5.4 Outcome

A surprising outcome of the CETUS challenge was that fully-automatic state-of-the-art methods could perform on par and even better than semi-automatic pipelines. No method has reached the inter-observer variability, especially low due to consensual manual annotations.

The community is still proposing state-of-the-art approaches on CETUS, including high potential machine and deep learning methods.

One main limitation concerns the validation of the methods. First, no cross-validation is performed in a challenge and the small amount of patients involved biases the evaluation. It is especially problematic when methods show very close performances. In addition, the interpretation of the bias and LOA for clinical indices is not straightforward as it does not give information on the average error. Finally, no metric assessed anatomical shape plausibility of the LV predicted by the algorithms. These observations strongly guided our own methodology.

4.4 Conclusion

The state-of-the-art of cardiac ultrasound segmentation is unclear due to the absence of a public dataset in 2D. However, the CETUS challenge saw the successful application on 3D echocardiography of generic supervised learning methods, such as random forests and convolutional neural networks, which showed a high potential for fully-automatic segmentation. These methods are easily extended to multi-class segmentation and can be augmented with appropriate pre-processing, prior knowledge, and temporal smoothing.

We establish from our review three needs for improvement:

1. the data: no public annotated dataset existed in 2D ultrasound before our initiative, and the only existing dataset in 3D involved few patients, which limitates the potential of learning methods but also limits the significance of the evaluation;

2. the models: though the obtained results are tremendously encouraging, no solution proved robust enough to provide reliable segmentation results and clinical indices;

3. the metrics: too few studies evaluate the segmentation results under a large set of complementary geometrical and clinical metrics, yet evaluation is of primary importance for medical applications.

In this work, we proposed innovations on all three axes. These contributions are detailed in Chapters 5, 6, 7, 8 and 9.

Part III

Contributions

Chapter 5

The CAMUS dataset

As described in Section 4.3, the problem of 2D echocardiography segmentation has never been investigated in the literature through a public large scale dataset. A single echocardiographic dataset, composed of 45 3D sequences, has been made public and broadly validated by the community. However, as it includes only 15 patients for training, this dataset is not appropriate to study the behavior of machine learning approaches. In this context, we introduce in this chapter the largest publicly-available and fully annotated dataset for the purpose of 2D echocardiographic assessment. The purpose of this clinical dataset is to:

- enable to appropriately train machine / deep learning models;
- allow a meaningful comparison between state-of-the-art methods;
- evaluate how far supervised learning methods can go at assessing 2D echocardiographic images i.e. segmenting cardiac structures as well as estimating clinical indices.

CAMUS stands for Cardiac Acquisitions for Multi-structure Ultrasound Segmentation, and is also the name of a french writer and revolutionary of the 20th century. The dataset is available for download at [1], along with a Girder evaluation platform.

5.1 Properties

5.1.1 Population

The proposed dataset consists of apical two-chamber (A2C) and four-chamber (A4C) sequences of at least one full cardiac cycle for 500 patients. Their inclusion followed the regulation set by the local ethical committee of the University Hospital of Saint- Étienne (France). To enforce clinical realism, no data selection was performed on any prerequisite. This produced a highly heterogeneous dataset, both in terms of image quality and pathology, which is the classical situation in clinic practice. As the ejection fraction for healthy patients should normally be around $64 \pm 6,5$ % [27], most patients of the CAMUS dataset ($> 80\%$) are likely to present a pathological heart with regard to the ejection fraction.

More information on the study population is given in Appendix D.

5.1.2 Acquisition

The acquisitions were performed at the University Hospital of Saint- Étienne, with GE Vivid E95 ultrasound scanners equipped with GE M5S probe and the EchoPAC analysis software. The clinical exams were optimized to perform left ventricule ejection fraction (LV_{EF}) measurements from apical two and four chamber views (A2C, A4C), as explained in Section 3.4.

TABLE 5.1: The main characteristics of the CAMUS dataset (500 patients)

Dataset	Image Quality %			LV_{EF} %		
	Good	Medium	Poor	$\leq 45\%$	$\geq 55\%$	In between
Full	35	46	19	49	19	32
fold 1	34	48	18	48	20	32
fold 2	34	46	20	50	18	32
fold 3	34	46	20	48	20	32
fold 4	34	46	20	50	20	30
fold 5	34	46	20	48	20	32
fold 6	36	46	18	50	20	30
fold 7	36	46	18	50	20	30
fold 8	36	46	18	50	18	32
fold 9	36	46	18	48	20	32
fold 10	36	46	18	50	18	32

The dataset involves a wide variability of acquisition settings (probe placement and orientation), in particular:

- for some patients, parts of the wall were not visible in the images;
- for some cases, the recommendation on the probe orientation to acquire a rigorous four-chambers view was simply impossible to follow and a five-chambers view was acquired.

5.1.3 Image quality

The sequences exported from the GE system correspond to sets of B-mode images expressed in polar coordinates. The same interpolation procedure was used to express all images in Cartesian coordinates with a unique grid resolution, i.e. $\lambda/2 = 0.3$ mm along the x-axis (parallel to the probe) and $\lambda/4 = 0.15$ mm along the z-axis (perpendicular to the probe), where λ corresponds to the wavelength of the ultrasound probe. The image quality (IQ) was evaluated subjectively by O_1, revealing 19% of patients with at least one poor image quality sequence (Tab. 5.1). For these patients, the localization of the endocardium (LV_{Endo}) and epicardium (LV_{Epi}) and the derived clinical indices are not considered clinically accurate.

In classical analysis, poor quality images are removed because of their clinical uselessness. In our study, we keep all the images and ponder the evaluation according to the image quality.

5.1.4 Partition

The main information which characterizes the collected dataset are presented in Tab 5.1. Relying on image quality and ejection fraction, the dataset was divided into 10 balanced folds. Each fold contains 50 patients with the same distributions in terms of image quality and LV_{EF} as the full dataset. This partition was used to derive training and test sets for the supervised learning methods, keeping 8 folds for training, 1 for validation (only for deep learning methods, added to the training set if not), and the last for testing. The 10 possible rotations were used to perform standard cross-validation for the evaluated methods.

5.2 Annotations

The generalization properties of the algorithm is linked to the reproducibility of the expert annotations, especially the coherence and accuracy of the manual contouring. Establishing a well-defined ground-truth segmentation was therefore of utmost importance for this work.

5.2.1 Selection of ED / ES frames

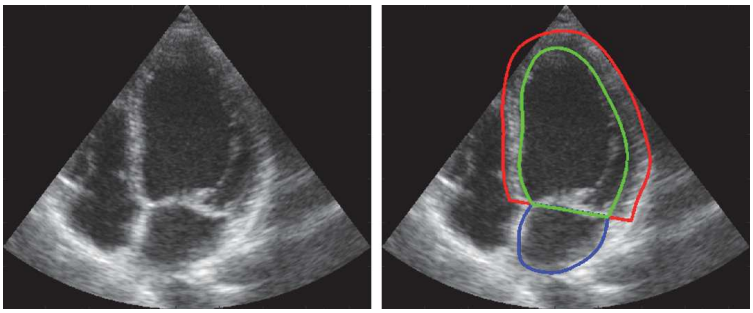
O_1 selected one End Diastole (ED) and one End Systole (ES) frame for all sequences, following the protocole and recommendations in (Lang et al., 2015) [27]. According to the recommendation of the American Society of Echocardiography and the European Association of Cardiovascular Imaging, ED is preferably defined either as the first frame after mitral valve closure, or as the frame in which the respective LV dimension or volume measurement is the largest. Similarly, ES is best defined as the frame after aortic valve closure, or as the frame in which the cardiac dimension or volume is the smallest.

In this work, ED and ES were selected as the frames where the LV dimension was at its largest / smallest due to the lack of reliable ECG. As this simpler approach is not the most accurate, especially in the presence of abnormalities, the values of the clinical indices we report have to be interpreted keeping this in mind.

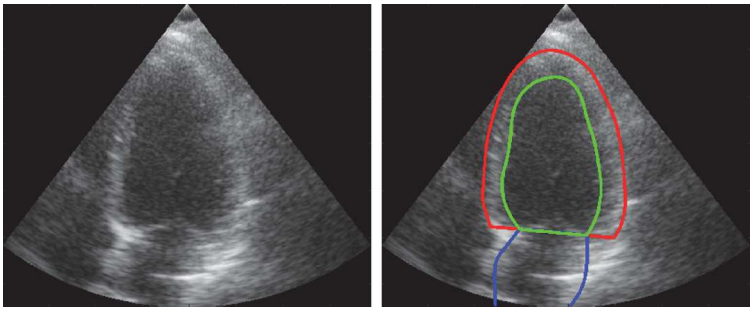
5.2.2 Segmentation

The acquisitions were optimized for the computation of the ejection fraction, and hence focus on having a good visual of the left ventricle, but the myocardium and the left atrium are also at least partially visible (Fig. 5.1). These structures were also annotated mostly to study the influence of contextualization on the performance of supervised learning techniques, and were drawn as adjacent, starting from the same points of junction located between the valves and the muscle (Fig. 5.1). The following protocol was set up:

- LV_{Endo}: the protocole described in (Lang et al., 2015) [27] was applied for the LV inner border, the mitral valve plane, the trabeculations, the papillary muscles and the apex. It implied in particular to:
    1. include trabeculae and papillary muscles in the LV cavity;
    2. terminate the contours in the mitral valve plane on the ventricular side of the bright ridge, at the points where the valve leaflets are hinging;
    3. partially exclude left ventricular outflow tract from the cavity by drawing from septal mitral valve hinge point to the septal wall to create a smooth shape;
    4. keep tissue consistency between ED and ES instants.
- LV_{Epi}: There is no recommendation for delineating the epicardium. We thus outlined the epicardium so that:
    1. it is drawn as the interface between the pericardium and the myocardium for the anterior (on A2C), anterolateral (on A4C) and inferior segments and the frontier between the right ventricle cavity and the septum for the inferoseptal segments;
    2. in case of missing local information, contours are drawn assuming shape and thickness continuity;
    3. its base aligns with the base of the left ventricle, drawing a straight line.



(a) Good image quality



(b) Medium image quality

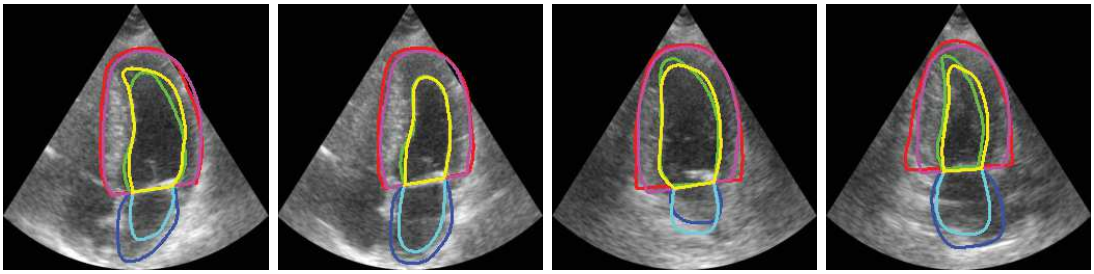
(c) Poor image quality

FIGURE 5.1: Typical images extracted from the CAMUS dataset. The endocardium, epicardium and left atrium wall are respectively shown in green, red and blue. (Left) input images, (Right) corresponding manual annotations.

- Left atrium (LA): it is recommended for the LA segmentation to assess the full LA area from dedicated recordings. Since we have used acquisitions focusing on the LV, part of the dataset does not cover the full LA surface and is therefore not suited to perform such measurement. We used the following contouring protocol:

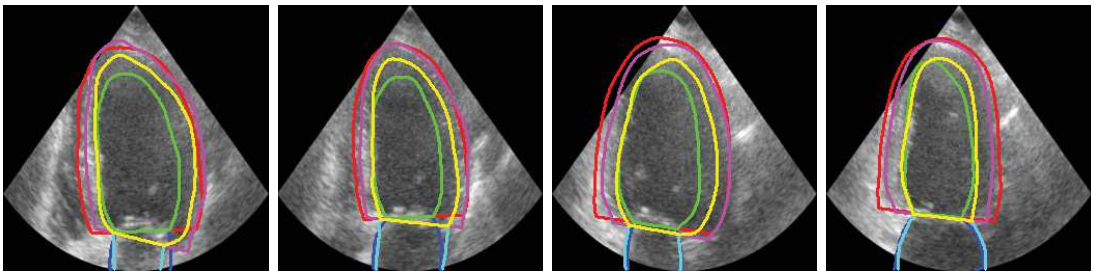
  1. start the LA contour from the extremities of the LV_{Endo} contour, at the points where the valve leaflets are hinging;
  2. have the contour pass by the LA inner border;
  3. stop at the image edges when necessary, as seen in Fig 5.1 b).

P-252: $O1a$ (rgb) VS $O1b$ (myc)



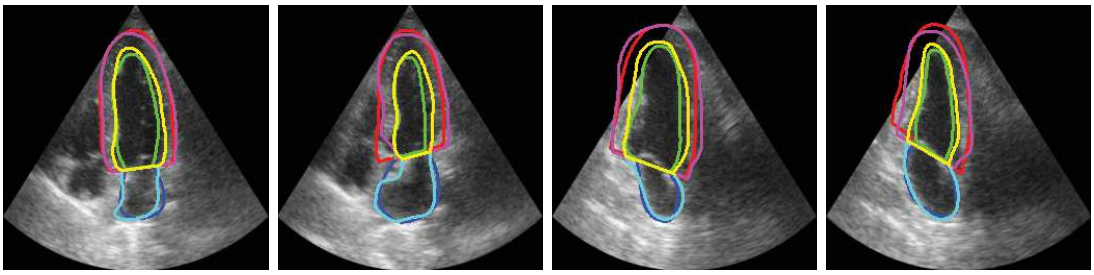
(a) A4CH-ED (b) A4CH-ES (c) A2CH-ED (d) A2CH-ES

P-266: $O1a$ (rgb) VS $O2$ (myc)



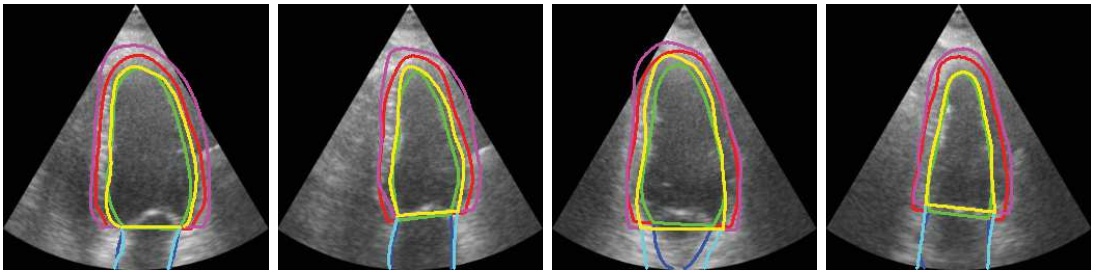
(e) A4CH-ED (f) A4CH-ES (g) A2CH-ED (h) A2CH-ES

P-242: $O1a$ VS $O3$ (myc)



(i) A4CH-ED (j) A4CH-ES (k) A2CH-ED (l) A2CH-ES

P-260: $O2$ (rgb) VS $O3$ (myc)



(m) A4CH-ED (n) ACH-ES (o) A2CH-ED (p) A2CH-ES

FIGURE 5.2: CAMUS dataset samples to show the average expert variability with respect to the mean absolute distance (MAD). One set of annotations is in RGB, and the other in MYC. On the selected cases, the frame entitled in red depicts for each pair the average variability on the endocardium. First row: intra-variability. Last three rows: inter- variability.

5.2.3 Inter- and intra- variability

Three cardiologists ($O1$, $O2$ and $O3$) participated in the annotation of the dataset. Considerable effort was spent to define with O_1 a consistent off-line manual segmentation protocol. We then asked $O1$ to perform the manual annotation for the entire dataset, while the other two contoured the 5^{th} fold (50 patients). O1 annotated this fold a second time seven months after (the annotations are refered to as $O1a$ and $O1b$, respectively). This fold is therefore used in this study to measure both the inter- and intra-observer variability. Examples are shown in Fig. 5.2. Geometrical and clinical scores may be found in Section 7.3.

5.3 Conclusion

We built a specific dataset to properly train, evaluate and compare segmentation methods in 2D echocardiography. The open challenge set by this public dataset is to design an algorithm able to produce segmentation results more accurate than the inter- and intra- expert variability. Except the investigation on 3D echocardiography presented in Section 6.4.2.1, all our algorithmic contributions were designed for and tested on the CAMUS dataset.

Chapter 6

Revisiting the formalism of Structured Random Forests for semantic segmentation in echocardiography

The following chapter is a discussion on the Structured Random Forest (SRF) algorithm, its properties, and its potential for the automatic segmentation of the heart in ultrasound imaging. It answers the following queries:

1. How can the SRF formalism be adapted to predict either edge probabilities or structure labels, and accordingly be integrated into segmentation pipelines?
2. How can we integrate contextual information into the feature space of SRF ?
3. Can anatomically plausible shapes be encouraged in a SRF framework ?

6.1 Introduction

6.1.1 Motivations

During the CETUS challenge of 2014 [16], several supervised machine learning technics were successfully applied to the segmentation of the left ventricle (LV) on 3D echocardiographic images. One team, Domingos et al., integrated the Structured Random Forest (SRF) algorithm into a semi-automatic pipeline that ranked 1st on average geometrical scores.

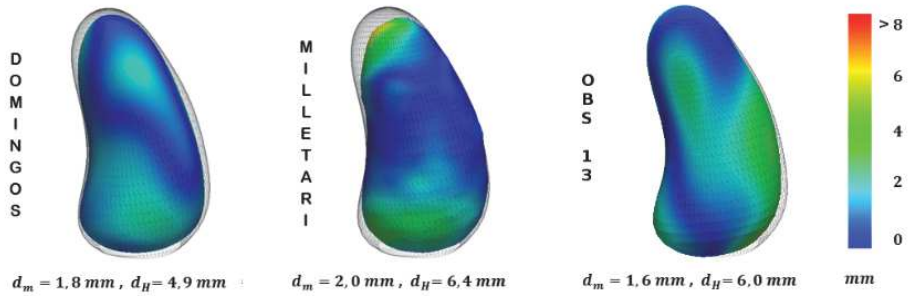


FIGURE 6.1: Error maps of the method from Domingos and al. [16], compared to the second best machine learning method and an expert. The expert consensus is shown in grey.

Interestingly, this machine learning method showed its capacity to produce accurate segmentation results even on:

- a small training dataset (15 patients);
- a 3D dimensional problem;
- a large variety of pathological cases and patient morphologies;
- heterogeneous, noisy, and multi-centric ultrasound data.

An illustration of the smooth error maps of the method is given in Fig 6.1. Based on these observations, we decided to investigate SRF as a primary candidate for building a robust and fully-automatic segmentation pipeline for both 2D and 3D echocardiographic images.

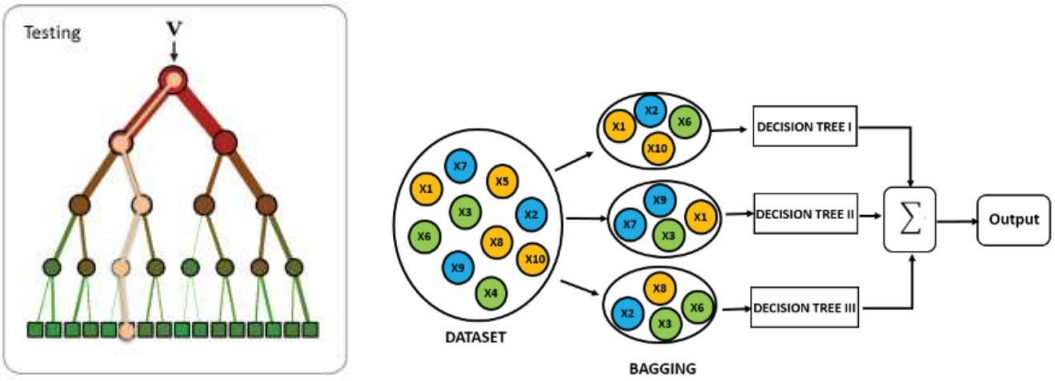
6.1.2 Random forests

6.1.2.1 Principle

The Structured Random Forests were introduced in Computer Vision in [28] as an extension of the popular Random Forest (RF) algorithm. RFs gained momentum in the nineties, often replacing Markov and Conditional Random Field (CRF) or Support Vector Machines (SVMs). Random Forests involve the learning of decision trees that perform successive heuristics (greedy decisions) on manually-crafted features. They have been applied to several mapping problems $X \rightarrow Y$ in image processing such as classification and regression, density estimation and manifold learning [62].

As segmentation can be approached as a pixel labeling task, RFs can be trained to predict to which structure or object a pixel belongs to. The corresponding mapping may be written as $P \rightarrow X \rightarrow Y$, where P is the set of pixels to classify from which we manually extract a set of discriminant features X to automatically assign to each pixel $p \in P$ a label $y \in Y$.

In the random forest paradigm, a forest F is a consortium of decision(-making) trees, wherein each tree $t \in [1, T]$ is able to make a prediction on and of its own (Fig. 6.2 a.).



(a) Binary decision tree making a decision from a new data [62]

(b) Bagging: using one subset of the data per tree [107]

FIGURE 6.2: Core elements of the RF algorithm: binary decision tree (a) allied to randomizing strategies (such as b). Decision trees are a set of nodes routing the data to terminal nodes - called leaves - where the answer is stored.

The trees are built to be decorrelated so that the forest's average prediction can correct the potential bias of a single tree brought on by over-fitting the training dataset [62]. In order to do so, each tree is trained on a random subset of the data $x \in X$, following a bootstrap strategy (Fig. 6.2 b.), and has access to a limited amount of features (often also randomly selected). The boot-strapping aggregation (bagging) consists in virtually creating several small data subsets of from one larger set through random sampling [7]. It allows to prevent over-fitting while still capturing the diversity of the training dataset by building distinct classifiers and taking their average prediction.

6.1.2.2 Training phase of a tree

We will here follow but simplify the notations and equations from (Criminisi et al., 2012) [62], restraining their use to classification for segmentation. During the training phase, each tree is built recursively as a succession of nodes where the data S_i (composed by a set of pixels) reaching node i is split into two subsets S_{i_1} and S_{i_2} based on a split function h_i (also called weak learner) of parameters k_i and τ_i having the following usual expression:

$$h_i(x_i, k_i, \tau_i) = [x_i(k_i) < \tau_i] \tag{6.1}$$

where x_i is a random subset of the feature space x computed from the set S_i, k_i is an index to select a single feature from x and τ_i corresponds to a threshold value.

This formulation of the split function is called a stump, as it performs a linear separation in the feature space (Fig. 6.3). From this model, the function h_i splits the data S_i into two subsets S_{i_1} and S_{i_2}, where S_{i_1} corresponds to the subset of pixels having their feature value $x_i(k_i)$ lower than τ_i, S_{i2} being the rest of the data. The parameters k_i and τ_i for each node i are searched in order to optimize their corresponding splitting function h_i by maximizing the entropy-based information gain expression I_g given below:

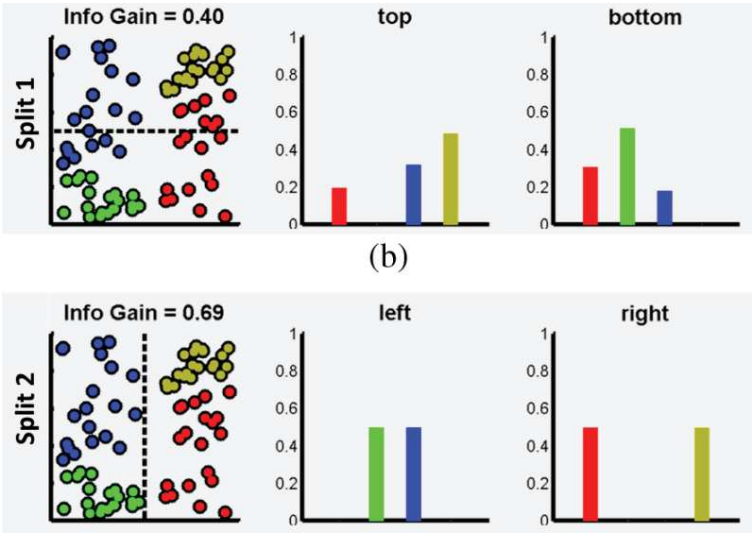


FIGURE 6.3: Stump illustration and resulting information gain [62]. The stump function is dotted. RFs perform linear cuts in the representation space.

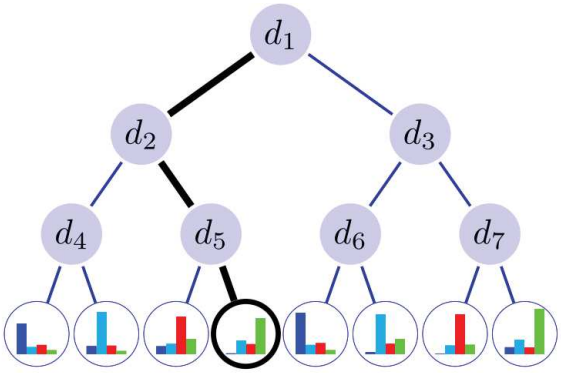


FIGURE 6.4: Tree representation: The information stored at the leaves after the training (here the histogram of labels) can be used at test time to assign a new class to the new data routed down to the leaf [108].

$$I_g(k_i, \tau_i) = H(S_i) - \sum_{j \in \{1,2\}} \frac{|S_{i_j}(k_i, \tau_i)|}{|S_i|} \times H(S_{i_j}(k_i, \tau_i)) \tag{6.2}$$

where the subsets S_{i_j} are subjected to the split function $h_i(x, k_i, \tau_i)$ and $H(S)$ is the discrete Shannon entropy formula given by:

$$H(S) = -\sum_{y \in Y} p_y \times log_2(p_y) \tag{6.3}$$

Each tree is trained until all paths are terminated by a leaf node. A leaf is created when a stopping criteria is met (e.g. maximal depth, minimal information gain, or minimal number of data at a node), and stores information that characterizes the data that reached it. For instance, leaves can store the maximum a posteriori label or the normalized histogram of classes to keep a sense of uncertainty, as illustrated in Fig. 6.4.

6.1.2.3 Testing phase of the forest

At inference time, all pixels of a given image are fed to a set of trees T_p from F. The data is submitted to the learnt split functions of each tree, starting from the root node, and gradually routed to one of the leaves (Fig. 6.2 a., 6.4). The information of the reached leaves are then brought together to assign an output value to the processed pixel (ensemble model). Each tree can make a prediction on its own $\hat{y}_t = \text{argmax}_{y \in Y}\, p_t(y/x)$, but exploiting the probabilistic output from each leaf is often used to benefit from the ensemble model theory. As the trees are trained independently, this step usually corresponds to a simple averaging procedure which has convenient denoising effect:

$$p(y/x) = \sum_{t=1}^{|T_p|} \sum_{y \in Y} p_t(y/x) \tag{6.4}$$

In a classification context, such probability can be directly used to assign as output the most voted class through the following equation:

$$\hat{y} = \underset{y \in Y}{\text{argmax}}\, p(y/x) \tag{6.5}$$

6.1.2.4 Summary of properties

Here is a summary of the main strong and weak points associated to random forests:

Strong points

- good generalization properties, even on small datasets, thanks to the bagging strategy and to the nonlinear splitting of the feature space;
- fast inference: as most trees do not reach depths greater than 50 successive tests, the test is manageable in real time, an appealing property for commercialized products;
- parallelization: as trees are unrelated to the others, they can be trained and applied separately, and hence benefit from parallel processing (multithread CPUs / GPUs);
- multiclass / multitask capacity: Contrary to SVMs, a single decision tree is directly expandable to multiclass application. Furthermore, it can also be trained to perform several tasks simultaneously (e.g multiple classifications, or regression and classification as in Hough forests [109]);
- uncertainty: by keeping the probabilistic formulation of the trees' predictions, a forest can be used to express confidence in a result, as well as several guesses.

Weak points

- feature engineering is required to encode the relevant information from high dimensional data (such as images) on complex tasks (such as segmentation);
- low representation power: each split function makes a greedy decision based on a small set of features. However, this aspect has been shown to help preventing over-fitting;
- potentially slow training: RFs require to learn several classifiers instead of a single one;
- potentially noisy predictions: for segmentation tasks, as each pixel is classified independently from the others, the resulting segmentations are usually very noisy (Fig. 6.5). This can be improved by incorporating contextual features (mid-level/or high-level) [110], mutualizing neighbouring labels, or directly predicting structured outputs as in SRF (detailed in the next section).

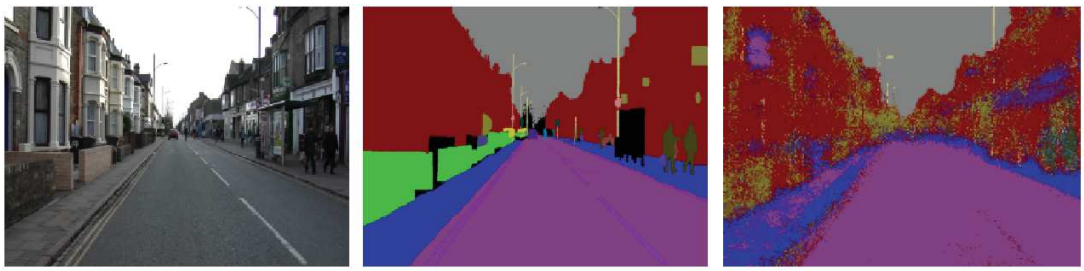


FIGURE 6.5: (left) Street view, (middle) its corresponding segmentation masks, (right) the prediction obtained from using RFs. [28].

Both / Neither Two traits of RFs are subjected to circumstances:

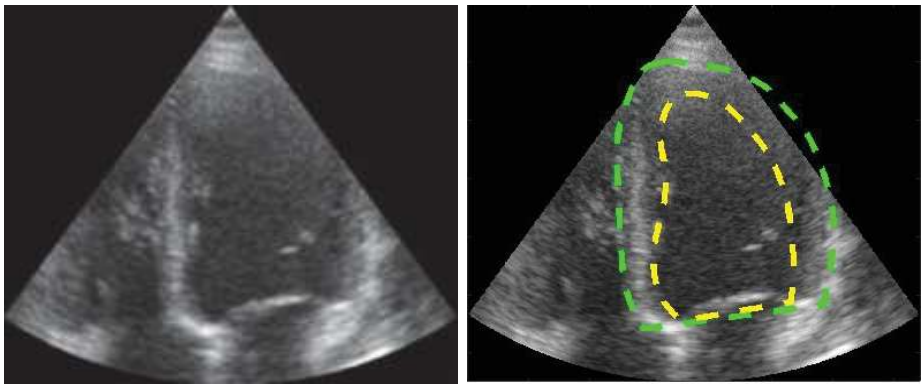
- interpretability: it is easy to trace why a decision was made based on the features that were used, provided that the features themselves are easily interpretable;
- tuning: the choices of features, weak learner function, energy function and stopping criteria can be intuitive when they are directly linked to the requirements (what information is discriminant for such task on such image modality, what defines the quality of a split, what is our computational budgets, etc...). However, the number of hyperparameters in RFs is quite large (as can be seen in Tab. 6.1), and we rarely have direct links between them and the specifications, which implies time consuming experiments.

6.1.3 Structured Random Forests

6.1.3.1 Principle

The segmentation maps obtained by classical Random Forests can be noisy due to the voting scheme applied independently to each pixel (Fig.6.5). However, natural objects have characteristic topologies, showing continuous and full structures with clear boundaries. For instance, in the particular case of ultrasound imaging, the edges of anatomical structures appear fuzzy with possibly missing information due to low signal to noise ratio in some parts of the image, as illustrated in Fig. 6.6.

Thus, there is a need to adapt the RF formalism to improve the coherency in segmentation tasks. In order to smooth the segmentation mask obtained from RFs, (Kontschieder et al., 2011) [28] proposed to change the mapping from $p \rightarrow X \rightarrow Y$ to $P \rightarrow X \rightarrow L$, where L corresponds to the segmentation mask of the image patch P, encompassing the coherent structures present in natural images (e.g. clean boundaries between full regions, as illustrated in Fig. 6.5). Enforcing such mapping should therefore result in a better capture of the topology present in the image [29].



(a) Ultrasound image

(b) Corresponding segmentation of the left ventricle and myocardium

FIGURE 6.6: Ultrasound image of the heart and its corresponding segmentation superimposed. The drawn contours are smoothed though the image information shows no clear boundary.

6.1.3.2 Training phase

One key aspect of the theory of Random Forests is the possibility to split the data according to similarity criteria, usually computed directly from the label space Y and based on a fixed number of classes. To train the split functions of SRF, as discrete labels are replaced by segmentation patches, it is necessary to define the notion of distance between segmented patches to group those that are closely similar, and separate those that are strongly different. In other words, an intermediary mapping $P \rightarrow X \rightarrow T \rightarrow L$ is required, with T a discrete label space like Y.

In (Dollár et al., 2015) [111], the authors proposed to model each patch by an edge vector constructed by unrolling the corresponding patch and to replace the value of each pixel by a binary value reflecting the presence or not of a contour at its position. At each node, the vectors associated to the input data S_i are randomly under-sampled to a fixed dimension of size 256 before a Principal Component Analysis (PCA) is applied. The signs of the first components are used to assign a virtual label (or tag) to each patch.

The tags cluster similar patches, which allows to utilize the classical RF theory to split S_i into two subsets S_{i1} and S_{i2}. At each leaf, the stored information reflects the coherent structure of the patches that reached it, for instance the mean patch with either probabilistic or binary values. Otherwise, as in (Kontschieder et. al, 2014) [29], a representative patch which characterizes the best the mode of the joint distributions of the patches can be chosen among the leaf patches i.e. the node patch that has the more pixel labels in common with the others.

6.1.3.3 Testing phase

At inference time, a new test image is divided into a set of overlapping patches centered on the nodes of a regular grid previously defined on the image support. Each image patch is fed to a set of trees. The information stored at the reached leaves is then used to create a single segmentation patch added to the output image at the same position as the image patch. A straightforward solution for the ensemble model consists in establishing for all pixels the class that was the most voted for among all tree predictions.

As patches are overlapping, the same strategy is applied to mutualize the overlapping predictions (fusion model). Such a scheme creates as output image a probability map of the presence of a region or the presence of an outline, depending on the relevant information.

6.1.3.4 Implementations

We provide in table 6.1 the main aspects of two distinct implementations of SRF, proposed respectively in (Kontschieder et al., 2014) [29] and (Dollár et al., 2015) [111] for computer vision applications. From this table at line "patch cluster", one can observe that both frameworks involve a dynamic classification of the patches: the tag characterizing a label patch is changed at each node to either focus on a different local information [29] or to better reflect the node data distribution [111].

In (Dollár et al., 2015) [111], the authors observe that to capture patch similarities in 16×16 patches, at least 64 pixels should be taken into account in multi-variate joint distributions. Instead, they propose to use PCA on 1D edge vectors to map high dimensional patches to a low dimensional label space.

TABLE 6.1: Differences between Kontschieder's and Dollár's approaches of the structured random forests

Method	Kontschieder et al. [29]	Dollár et al. [111]
Application	Street scenes (CamVid, MSRCv2)	Natural images (Berkeley dataset)
Problem	Multi-class segmentation	Edge detection
Features	$I + I' + I''$ + HOG pixel lookups	$CIE-LUV + \|I'\|$ + HOG at 2 scales lookups + pairwise differences
Number of trees	10	8 trained, ≤ 4 at test time
Patch size	24×24	16×16
Stop criteria	$D_t > 500$ $\|l\| < 5$	D, Ig, $\|l\|$ tuning on the training data
Patches selection	Overlapping	Overlapping
Patches similarity	Joint probability (labels considered independent)	K-means / PCA on the edge vector: $nSamples$ random pairs $[y(i) = y(j)]_{i \neq j}$
Patch cluster	Dynamic classification center + 1 random labels: $t \in \|Y\|^2$	Dynamic classification signs of the first 5 PCA components
Split function	random stump function type	classical stumps
Split evaluation	Information gain	Information gain
Leaf content	From the leaf set Maximizes the joint probability	medoid patch closer to the mean
Ensemble model	From the set of predictions Maximizes the joint probability	averaging
Fusion model	Most voted class Refined iteratively	averaging
Observation	even a small patch size > 5 is helpful	Few trees are necessary due to the fusion model
Limit	Patch tagging: only one transition is taken into account	The edge vector is binary not multiclass

I: image intensity, I': gradient of I, HOG: Histogram of oriented gradients
D_t: tree depth, l: leaf patches, I_g: Information gain

Thus, Dollár's method enables the tag to carry more and more discriminating information about the patches present at a given node as the tree grows deeper. On the contrary, Kontschieders method uses a simpler scheme involving little information in order to rapidly perform the separation at the nodes (two-label joint distribution, or a single pixel class), at the cost of representation power. Piotr Dollár gave public access to his matlab code [112], enabling a fast application to medical imaging [33].

6.2 Methodology

Based on the analysis performed in table 6.1, we decided to revisit the formalism of (Dollár et al., 2015) to perform multi-class segmentation of echocardiographic images. To this end, we made three methodological contributions:

1. we characterized the similarity between multi-class segmentation patches storing region information instead of edge content in the 1D vector;
2. we appropriately adapted the choice of the medoid patch associated to each leaf as well as the integration of tree predictions in the ensemble model;
3. we realized a comprehensive study to extract the most suited features among a given set in a RF framework.

Through this study, we showed in particular the importance of mixing multiple scales of features to encode contextual information in echocardiography. In parallel, we developed several tree visualization tools to investigate the pertinence and refinement of the data separation in SRF.

6.2.1 From edge to multi-region formalism

6.2.1.1 Multi-class patch separation

To generalize the methodology of (Dollár et al., 2015) [111] to region-based segmentation, it is necessary to derive a mapping function from the label space L to the tag space T $L \rightarrow T$ that takes into account a multi-class distribution of labels (as illustrated in Fig.6.7). To this aim, we first propose to replace the $nSamples$ pair-wise comparisons (Patches Similarity line of Tab. 6) used to compute the edge vector by the label values of $nSamples$ pixels selected by the same random sub-sampling scheme. This procedure allows us to derive a vector, called the patch code in the aftermath, consisting in a sub-dimensional representation of the patch. From such vectors, we then considered several approaches for the separation of patches based on their similarity:

1. scheme 1: compute PCA directly on the corresponding patch codes (regardless of the fact that distances between classes are not equal). Then, as proposed by Dollár et al., the signs of the first n_{PCA} components determine the tag value. This scheme thus creates at most 2^{nPCA} different tags at each node;
2. scheme 2: assign a tag value according to which classes are present in each patch code. There are at most $2^{|Y|}$ tags with this method, with $|Y|$ the number of labels;
3. scheme 3: take both schemes 1 and 2 into account independently, leading to the creation of $2^{(nPCA+|Y|)}$ possible tags;
4. scheme 4: generate $|Y-1|$ binary codes from each patch code, one for each class other than the background, by assigning 1 to pixels belonging to this given class, and 0 to other labels. Scheme 1 is applied to every set of binary codes. This approach allows to consider the classes as equidistant, but will create up to $2^{(nPCA\times|Y-1|)}$ different tags at each node.

We empirically assessed the performance of the tag information generated by these four multi-class schemes on a subset of echocardiographic images. Thanks to these experiments, the following conclusions were drawn:

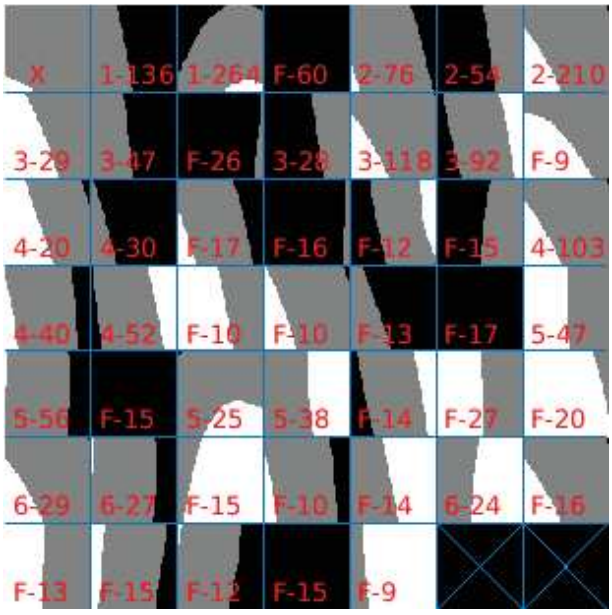
- scheme 1 produces the overall best results. This illustrates the capacity of the PCA procedure to capture the structural information in the patch, both in terms of spatial and labeling information;
- scheme 2 performed significantly worse than the others solutions. This proves that characterizing patches only from the class information in the patch is not sufficient. This result also highlights the ability of the PCA approach to extract useful structural information to assess the similarity of segmented patches;
- schemes 3 and 4 significantly slowed down the training process due to the high number of tags without any improvement compared to the results obtained with scheme 1.

We thus retained **scheme 1** in our SRF formalism to efficiently compute similarities between segmented patches with multi-class information.

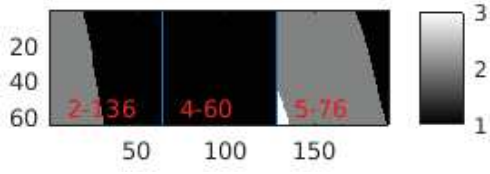
6.2.1.2 Multi-class leaf content

The **medoid** associated to each leaf can either be picked among the set of patches that reached it during the training process or computed from them. Considering that SRFs aim at reducing the noise in label distributions, we keep the idea of choosing an existing patch as leaf information. Since scheme 1 is a direct extension of Dollaŕ's patch similarity function, we can also directly apply the patch selection they proposed of setting the representative patch to be the medoid, obtained by identifying the patch whose code is the closest to the mean code.

To visualize the quality of split, and the diversity of medoids, we can display the medoids of each node, along with the number of patches it represents. In Fig. 6.7a), we display the nodes of a small tree (learnt on 400 patches). The root node is identified by 'X', leaves as 'F', and others as 'node number' - 'number of patches'. From left to right and top to bottom we display the nodes in order of construction, hence from the top of the tree to the bottom.



(a) The final nodes "F" are leaves. Most of the time, it is above 8, the minimal number of patch at a node.



(b) Split visual: node 2 splits its data between nodes 4 and 5. Medoids are refined.

FIGURE 6.7: Visualization of the nodes of a small tree (400 patches learnt).

On Fig. 6.7 b), the result of a split allows to observe the evolution of medoids after the information gain-based separation. We can observe that:

- medoids are refined properly with the similarity function described above: the even split in Fig. 6.7 b. result in two very different medoids both equally similar to the medoid of the parent node;
- as the tree gets deeper, more complex segmentation patches are present as medoids, which suggest both a good separation of patches and a good choice of medoids.

6.2.1.3 Multi-class patch fusion

Once the tag strategy is defined, we have to determine how to assign a label at a given position. We propose, in order to beneficiate from all predictions to associate each pixel to the class most voted over:

1. all the predictions of a set of trees. In practice we use all the trained trees;
2. all the overlapping patches. At inference, patches are sampled from the image with a stride of 2.

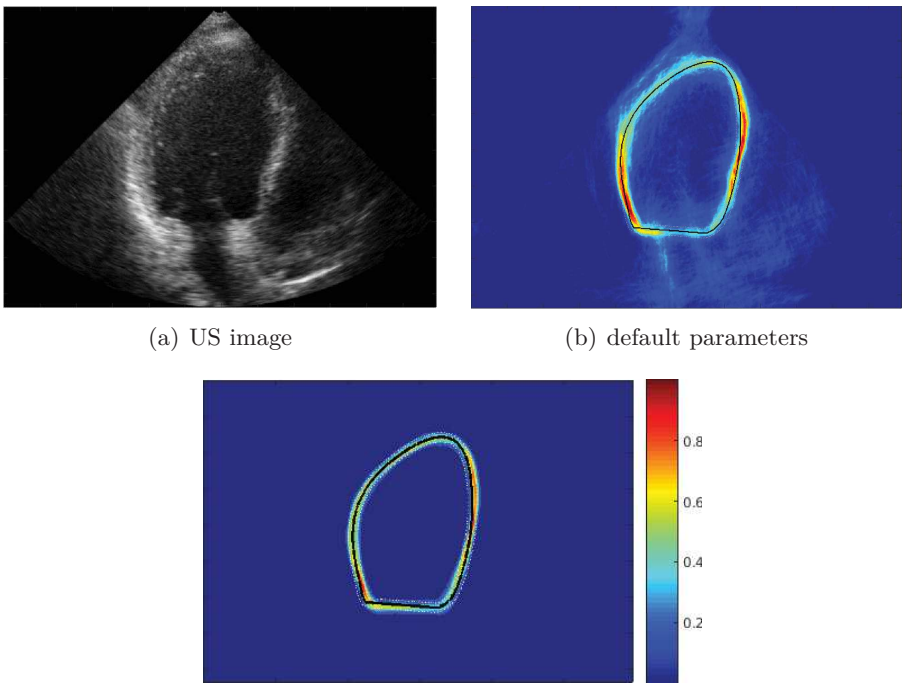
6.2.2 Multi-level features

Inspired by [111], we compute **gradient-type features** from every image. A patch P is associated to the concatenation of the features of all its pixels, but also to a few pair-wise differences of features to take into account spatial evolution of edges across the patch. The feature space includes the intensity at the initial resolution as well as the magnitude of gradient and histogram of gradient (HOG) computed at several scales to provide complementary contextual information, from local to global.

Each scale s of the image is obtained by downsampling the image by a factor $2 \times s$ and applying a smoothing triangular filter. A small filter size is used for "regular" feature channels, whose values are stored directly (Fig. 6.10), and a larger for "similarity" channels, used to compute pairwise feature differences (Fig. 6.11). The HOG space is divided into 4, resulting in a total of 5 features maps per scale (magnitude + HOG on 4 directions). All are brought back to the initial resolution by bilinear interpolation.

Since contextual information can bring meaningful information in echocardiographic images, we decided to investigate its influence on a dedicated subset of echocardiographic images composed of 50 images acquired from an A4C view at ED (Fig. 6.8). In particular, features were computed for scales s = 1 to 7. To investigate which scales are the most discriminant, we trained several trees with features of all scales and computed how often they were used by the trees to split the data.

Fig. 6.9 provides the results we obtained. Thanks to this figure, one can empirically observe that the higher scales (from 4 to 7) are very often used to separate the label patches. Thus, in order to make a compromise between the memory usage and the contextual information brought at different scales, we decided to involve **scales 1, 3 and 5** for the computation of the feature space in our experiments. From the statistics shown from the small experiment in Fig. 6.9, this choice might appear surprising but it was motivated by: i) the fact that scale 7 could not be computed for the smallest images of the dataset ; ii) when looking at performance, scale 5 and 6 were inter-changeable. Fig. 6.10 and 6.11 provide examples of features computed at several scales.



(a) US image (b) default parameters

(c) tuned

FIGURE 6.8: Detection of the endocardium using default or tuned SRF (multi-scale features + larger patch size + all trees used at test time).

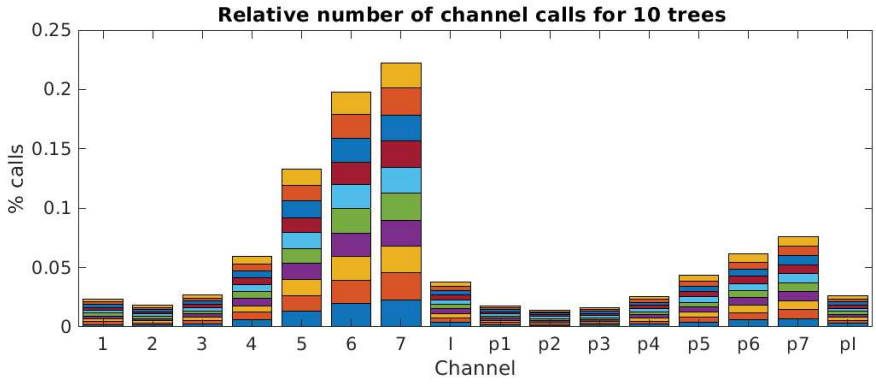
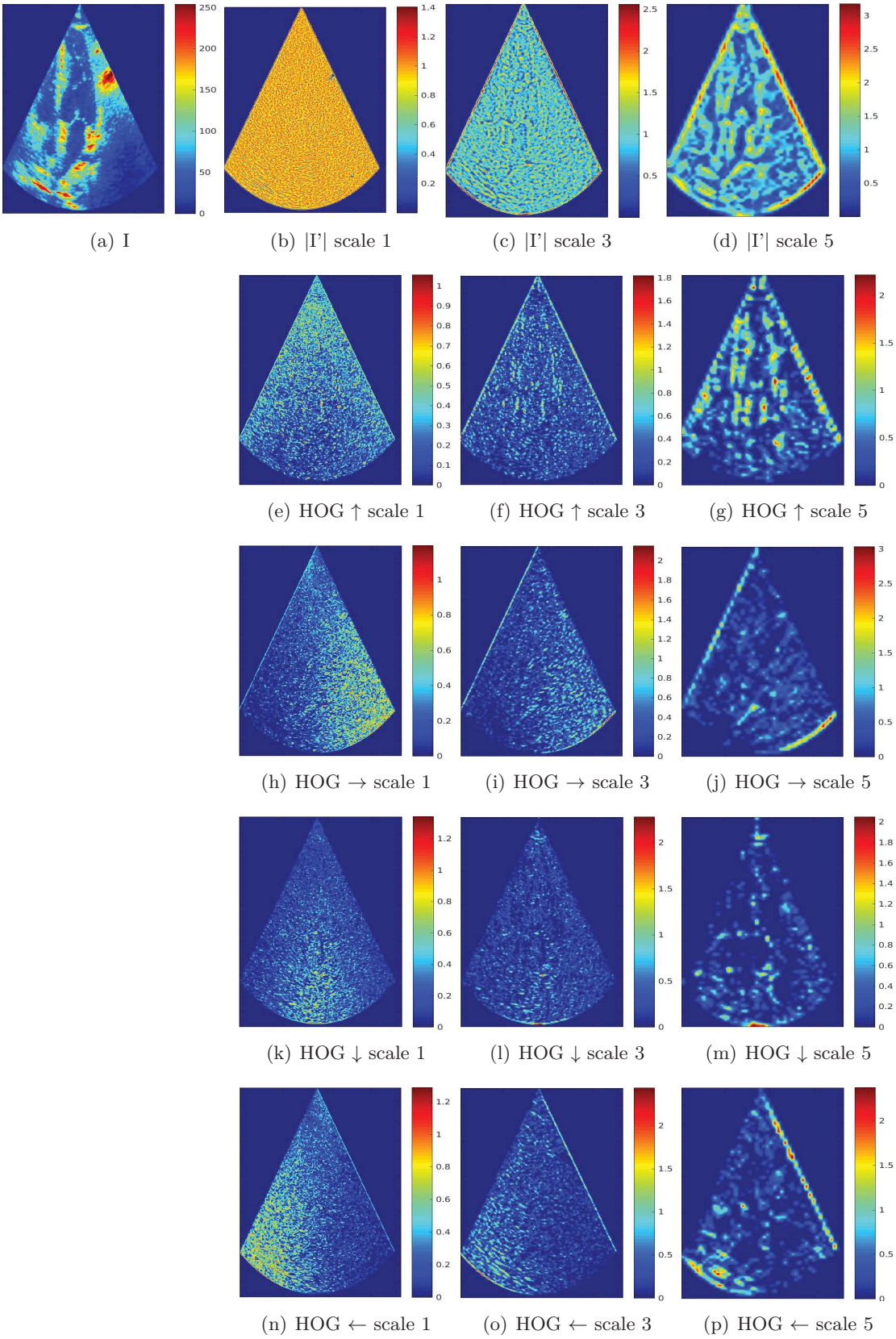


FIGURE 6.9: Relative number of calls for 7 scales of features on ten mid-size trees (5×10^4 patches). Regular features on the left, pairwise on the right.



(a) I (b) |I'| scale 1 (c) |I'| scale 3 (d) |I'| scale 5

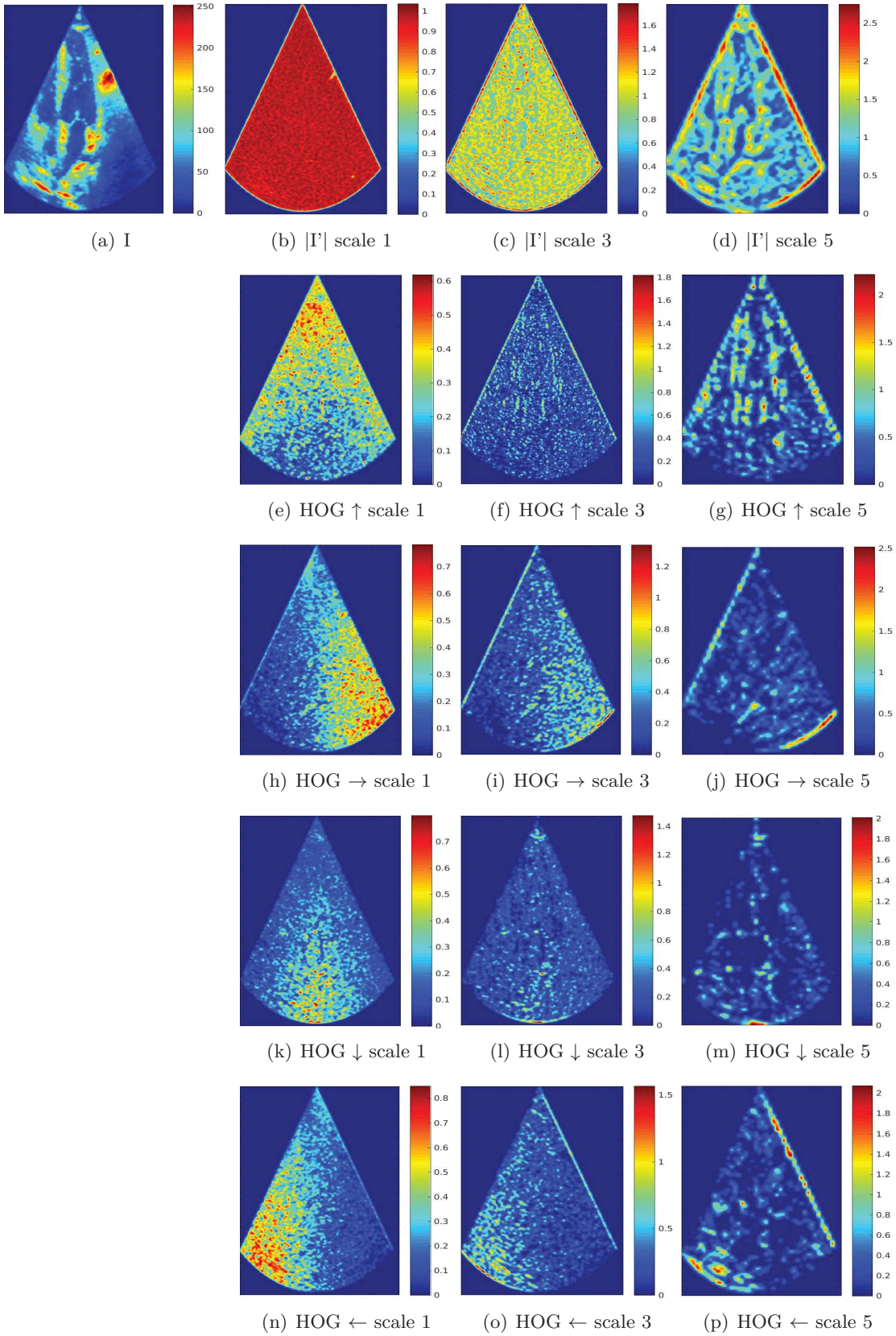
(e) HOG ↑ scale 1 (f) HOG ↑ scale 3 (g) HOG ↑ scale 5

(h) HOG → scale 1 (i) HOG → scale 3 (j) HOG → scale 5

(k) HOG ↓ scale 1 (l) HOG ↓ scale 3 (m) HOG ↓ scale 5

(n) HOG ← scale 1 (o) HOG ← scale 3 (p) HOG ← scale 5

FIGURE 6.10: Regular feature maps of a given image, the HOG arrows represent the direction on which the gradient is projected



(a) I (b) |I'| scale 1 (c) |I'| scale 3 (d) |I'| scale 5

(e) HOG ↑ scale 1 (f) HOG ↑ scale 3 (g) HOG ↑ scale 5

(h) HOG → scale 1 (i) HOG → scale 3 (j) HOG → scale 5

(k) HOG ↓ scale 1 (l) HOG ↓ scale 3 (m) HOG ↓ scale 5

(n) HOG ← scale 1 (o) HOG ← scale 3 (p) HOG ← scale 5

FIGURE 6.11: Similarity feature maps of a given image.

6.3 Application to 2D echocardiography segmentation

This work has been published in the IEEE IUS Conference of 2017 [30].

6.3.1 Dataset

At the time of the study, the CAMUS dataset was not completed. We worked on a version of 250 A4C view acquisition. For each sequence, we compared the performance of our SRF approach with the annotation of one expert cardiologist at end-diastole (ED) and end-systole (ES), i.e. respectively the end of the dilation and compression phases. For both ED and ES, we used 200 patients as train subjects to build a multi-class segmentation forest and the remaining 50 as test subjects.

6.3.2 Hyper-parameters

6.3.2.1 Optimization study

Our proposed method depends on several hyper-parameters, the most important for the accuracy being:

- the size and number of the patches;
- the number of PCA components used to compute the tag information;
- the number of trees.

Moreover, many other hyper-parameters in SRF have to be optimized according to these three main hyper-parameters. For instance:

- the number of trees from which performance begins to reach a plateau depends on the quantity of patches and features used to train each tree;
- the patch size affects the number of patches that should be sampled from each image to cover it, and also has an impact on the choice of the patch code dimension described in section 6.2.1.1.

All hyper-parameters were fixed from a cross-validation study performed on a different dataset from a previous study, which acted as a validation set. The best configuration is summarized in table 6.2. It is interesting to note that a higher patch size than the one used in Dollár et al. [111] was most beneficial to produce smooth and coherent shapes. We also observed that the greater the number of trees used, the better the results. We thus used the 16 trained trees during the testing procedure. However, there was little improvement in beyond 8 trees, possibly because of the redundancy in training sets introduced from this level.

6.3.2.2 Stopping criteria

The following stopping criteria were used during the training phase:

- a maximum tree depth of 32;
- a minimum number of patches at a leaf of 8;
- a minimum entropy gain of 1×10^{-10}.

We observed that the last two criteria prevailed since most trees did not reach a depth of 32.

6.3.2.3 Training patches

The original bagging strategy was kept, implying the number of positive patches (patches containing at least two labels) and negative patches (patches containing only one unique label) was balanced in the training set of each tree. There are on average about 2×10^4 possible positive patches in a CAMUS image, and 1×10^5 negative ones, so a tree learnt approximately $\frac{1}{10}$ of available positive patches and $\frac{1}{50}$ of negative ones. These numbers include patches just translated from one pixel, so with these proportions a tree learns a good representation of the full dataset. The patches are chosen at random, which improved the performance compared with forcing the learned patches to be on a regular grid.

6.3.2.4 Summary

Here is a summary of the settings of our study. For both ED and ES, we used 200 patients as train subjects to build a multi-class segmentation forest, and the remaining 50 as test subjects. We build 8 trees with the first 100 train patients and 8 others on the other 100 patients. The resulting 16 trees have each learnt from 400 000 2D patches equally distributed and are all able to predict the segmentation of a new test image. We compute features at scales 1, 3 and 5 as previously described in Section 6.2.2 in order to provide local, mid-level and global context. A 2D patch of size 64×64 is summarized by 64 randomly selected pixels and we use the 3 first PCA components to establish labels. Tree maximum depth is settled at 32. At test time, we extract patches with a stride of 2 and associate to each pixel the most voted class by considering the tree predictions on all patches that contain it. Figure 6.12 provides an illustration of the overall strategy of our proposed SRF solution.

6.3.3 Evaluation

6.3.3.1 Metrics

We computed three geometric metrics to account for the segmentation quality of our algorithm and its robustness: The Dice, the mean absolute distance (MAD) and the Hausdorff distance (HD), presented thoroughly in Section 4.1. The Dice and the MAD give global information (overlap and closeness), while the HD highlights local errors. Both distance metrics were computed in mm and relate to the borders: the endocardium for the left ventricle (LV_{endo}) and the epicardium (LV_{epi}) for the myocardium. For this particular study, we chose not to take into account contours outside of the ultrasound sector in our evaluation, because they are not based on any image information.

Table 6.2: Main differences between Dollár and al.'s and our hyperparameters

Hyperparameter	Dollár and al. [111]	2D US [30]
Patch size / number per tree	16×16 / $10e^6$	64×64 / 4×10^5
1D vector length (used for PCA)	256	64
nb PCA components / clusters used for labeling	2 / 4	3 / 8
Fraction of features used to train each tree	0.25	0.5
pair-wise features grid size	5	8
nTreesEval	50%: 4	all: 16 (8 per 100 patients)

6.3.3.2 Active Appearance Model

We compared our fully automatic solution to the semi-automatic Active Appearance Model (AAM) previously presented in the state-of-the-art review (Chapter 4) in Section 4.3.4.1. AAM were first introduced in [23] and successfully applied to ultrasound segmentation in [93]. This method corresponds to a deformable model whose statistics of shapes and intensities along the moving contour are learned from a training dataset. The shape statistics are obtained by applying a classical PCA on the training dataset so that the evolving contour is modeled as the addition of a mean shape with possible deformations as given by the following equation [23]:

$$S = \bar{s} + P_s \times b_s \tag{6.6}$$

where $\bar{s}$ corresponds to the mean shape, P_s is a projection matrix which encodes the modes of shape variation, and b_s corresponds to a scalar vector which determines the amount of variation along each mode (each component of this vector is bounded by a fixed value computed from the training set).

In parallel to the shape model, an appearance model is used to encode the possible evolution of the pixel intensities along the evolving contour by applying another PCA through the following equation:

$$A = \bar{a} + P_a \times b_a \tag{6.7}$$

where $\bar{a}$ corresponds to the mean intensity values computed for each point of the evolving contour, P_a is a projection matrix which encodes the modes of intensity variation and b_a a scalar vector which determines the amount of variation along each mode (each component of this vector are bounded by fixed values computed during the training phase).

The shape and appearance representations are united in a single model which forces a joint optimization of both shape and appearance simultaneously. We created two independent

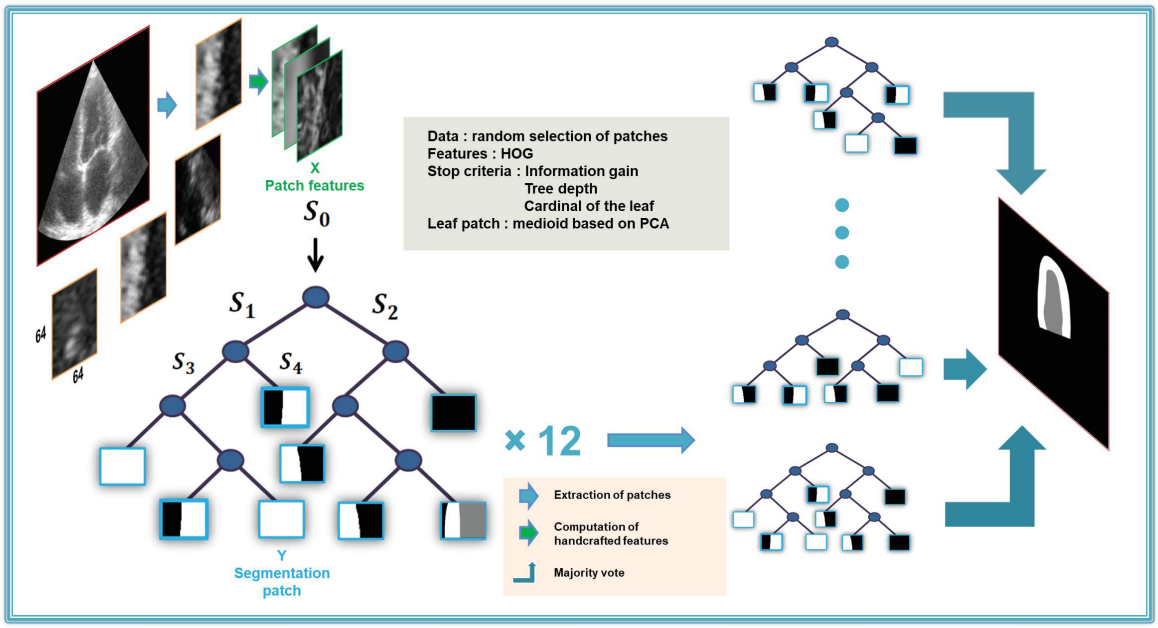


FIGURE 6.12: Our SRF summary [31]

models: one to detect the endocardium border, the other specialized in the epicardium border. The initialization requires 5 points manually provided by a user: 2 at the basis, 1 at the apex and 2 at the septum. 4 points are on the endocardium border and the last one initiates the myocardium thickness. To make this initialization phase accurate and reproducible, we automatically selected those points from the annotations of the expert used as reference. This biases somehow our study, but allows us to compare our approach with a semi-automatic method that has been favored.

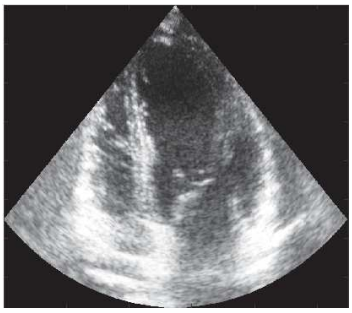
6.3.4 Pre- and Post-processing

For all the experiments, the data was pre-processed so that pixel intensities within the sectorial region were normalized following the proposition in [93]. In particular, the following two-steps were applied on each image:

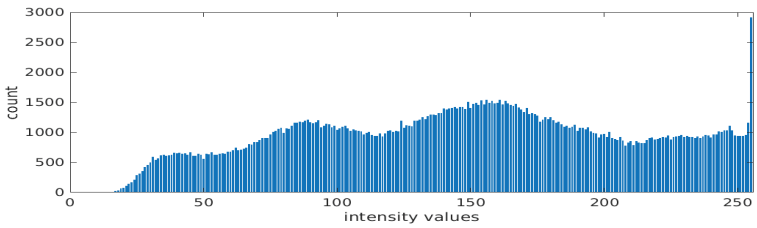
- computation of the 0.1 percentiles f_{pmin} and f_{pmax} from the cumulative probability density function;
- normalization of intensities by applying the following operation: $I_N = 255 \times \frac{I - fpmin}{fpmax - fpmin}$

Fig. 6.13 illustrates the application of such normalization on an example.

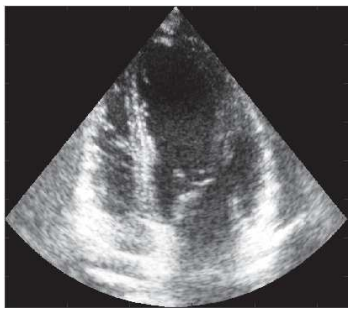
The same post-processing was applied for both approaches to the segmentation results obtained for each evaluated method, which consists in removing so-called false positives by keeping only the biggest predicted structure for each class. This procedure is later called "blob selection".



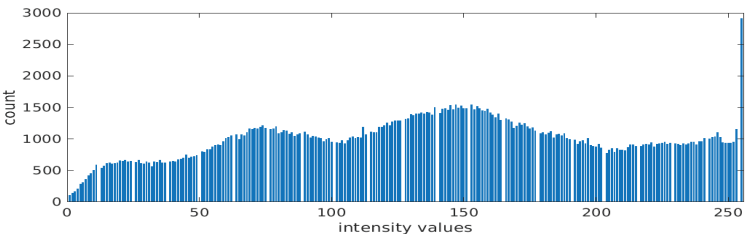
(a) US image



(b) Histogram without the zero values



(c) Normalized image



(d) Normalized histogram

FIGURE 6.13: Image and histogram before (top) and after (bottom) the pre-processing histogram normalization. A better contrast can be seen on the pre-processed data.

TABLE 6.3: SRF VS AAM segmentation results at ED

Struct	Algorithm	Dice	HD	MAD
LV_{endo}	AAM	0.90±0.04	**7.24±2.77**	2.84±1.20
	MS-SRF	**0.92±0.03**	8.20±4.77	**2.13±0.91**
	MS-SRF-r	0.92±0.03	7.41±3.91	2.04±0.86
LV_{epi}	AAM	**0.92±0.03**	**7.51±2.37**	2.91±1.10
	MS-SRF	0.88±0.08	8.51±6.9	**2.42±1.95**
	MS-SRF-r	0.90±0.05	6.73±3.89	2.01±0.85

AAM: Active Shape Model, MS-SRF: Multi-Scale SRF
MS-SRF-r: Multi-scale SRF with the 6 worst cases removed

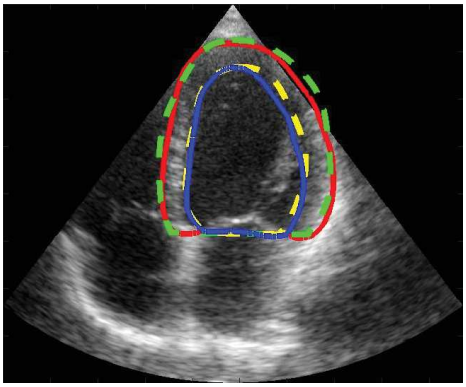
TABLE 6.4: SRF VS AAM segmentation results at ES

Struct	Algorithm	Dice	HD	MAD
LV_{endo}	AAM	0.89±0.06	**6.9±3.04**	**2.25±1.29**
	MS-SRF	**0.93±0.04**	10.23±5.44	2.88±1.44
	MS-SRF-r	0.93±0.03	9.04±4.24	2.57±1.17
LV_{epi}	AAM	**0.93±0.03**	**6.64±2.16**	**2.43±0.9**
	MS-SRF	0.90±0.08	12.71±13.14	3.33±2.53
	MS-SRF-r	0.92±0.04	8.77±4.50	2.53±1.16

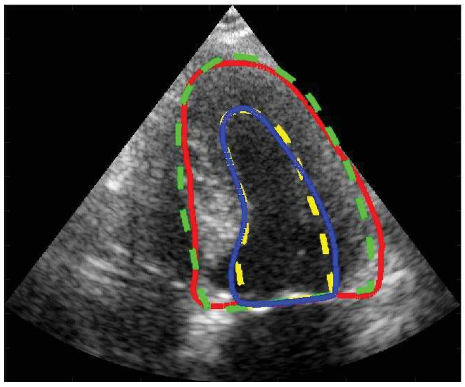
6.3.5 Results

Tables 6.3 and 6.4 summarize the results obtained separately on both ED and ES. We used the results on the full test set in our comparison analysis between Multi-structural SRF (MS-SRF) and AAM. Though our automatic solution performs generally well, there are 6 cases on which it performs significantly worse. This happens on both ED and ES, which indicates it struggles on specific patients. The most likely reasons for these failures are unusual intensity patterns and contrast. It is possible that with more training cases, the algorithm could learn to solve them. We also display the statistical results that were obtained without these particular cases (MS-SRF-r), as an indicator to the potential of this method. From the results provided in tables 6.3 and 6.4, we can make the following observations:

- our automatic solution obtained competitive results compared to those of the semi-automatic AAM, especially at ED where MAD scores are better for both structures;
- the SRFs obtained worse results in terms of HD for both ED and ES and both structures. This suggests that our method tends to produce inaccurate local segmentation;
- in terms of MAD and HD, the SRFs performed better at ED than at ES. This result is usual in echocardiography since the heart involves more complex shapes at ES. It is interesting to note that the AAM performed better at ES, which implies that this model is able to efficiently encode the cardiac shape variability at this instant;
- the LV_{endo} appears easier to segment than the LV_{epi} for both methods, possibly because of a better contrast between the LV and the myocardium;
- removing the 6 outliers significantly improves the SRFs scores, especially for the HD values (e.g. the HD score goes from 12.7 mm to 8.8 mm for the myocardium). At ED,



(a) Segmentation at ED for a test patient with good contrast.



(b) Segmentation at ES for a test patient with mid-quality contrast.

FIGURE 6.14: Visuals for solved cases. Ground truth contours are dotted, while the algorithm prediction is displayed in full line.

the SRFs would outperform the AAM on average and close the gap at ES, where the standard deviation for the HD goes from 13.14 mm to 4.50 mm. This clearly suggests that coping with outliers should be a priority.

To summarize, the SRFs provide a very interesting fully automatic solution for multi-structure cardiac segmentation, as its performance is comparable to that of the semi-automatic AAM. They solve a large variety of configurations as seen on Fig. 6.14. Thanks to the averaging of local segmentation patches, the predicted structures are coherent, full and closed.

6.3.6 Discussion

The weakness of our SRF lies in its lack of robustness to unusual cases, of which an example is provided in Fig. 6.15. One way to improve robustness would be to add more training data, hoping to learn enough variability to cope with all possible cases. Others leads include adding shape priors as in the following study, or extracting more discriminant features from ultrasound images. However, feature extraction is a tedious task without the insurance of a good generalization. This aspect motivated us to study and compare the capacity of deep learning solutions to perform the segmentation of echocardiographic images in Chapter 7.

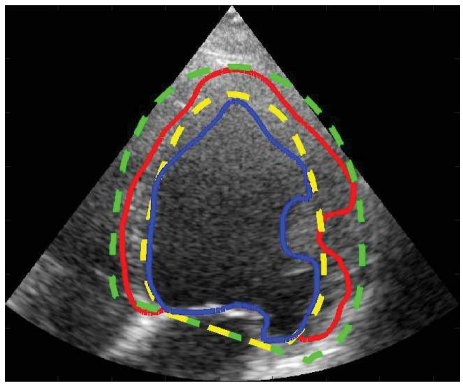


FIGURE 6.15: Visual for an unsolved case: Low contrast, unusual intensity patterns and shape configuration are in our opinion responsible.

6.4 Application to 3D echocardiography segmentation

This work has been published in the IEEE SPIE conference 2018 [32].

6.4.1 Motivations

Based on the observation from the previous study that adding shape constraints would lead to better results for our SRF segmentation framework. We proposed a novel framework combining SRFs specialized in 3D edge detection with a conventional Active Shape Model (ASM), which we evaluated on the CETUS dataset (see Section 4.2.3.1). Segmentation was performed and evaluated for the end-diastolic (ED) and end-systolic (ES) phases and compared to the results of state-of-the-art algorithms.

6.4.2 Methodology

In order to easily introduce shape constraints into our formalism, we decided to adapt our SRF to 3D edge detection. The output of this updated version of the algorithm produces a 3D probability map where each voxel value amounts to the probability of having detected the endocardial boundary. An ASM is then applied on the derived data-term to optimally position the 3D evolving surface on the local maxima of the map while regularizing its shape in a pre-learned space.

6.4.2.1 Structured Random Forests for the creation of 3D edge probability maps

In (Domingos et al., 2014) [33], the authors also used SRF to create edge probability map but from 2D slices. We proposed in this study to directly extend the SRF algorithm for the detection of 3D edge probability maps.

Extension to 3D We kept the edge SRF formalism described in table 6.1 while retaining the idea of using multiple scales of features. The main issue was the extension of the formalism from 2D to 3D. Indeed, this forced us to make a few choices for memory cost reasons, whose main aspects are listed below:

- use of a smaller patch size than in 2D (from 64 to 16);
- reduce the number of scales used to compute the feature space (from 3 to 2);
- increase in the number of HOG orientations to take into the supplementary dimension (from 4 to 8);
- increase of the dimension of the 1D edge vector for a better representation of the edges in the 3D patches (from 256 to 512).

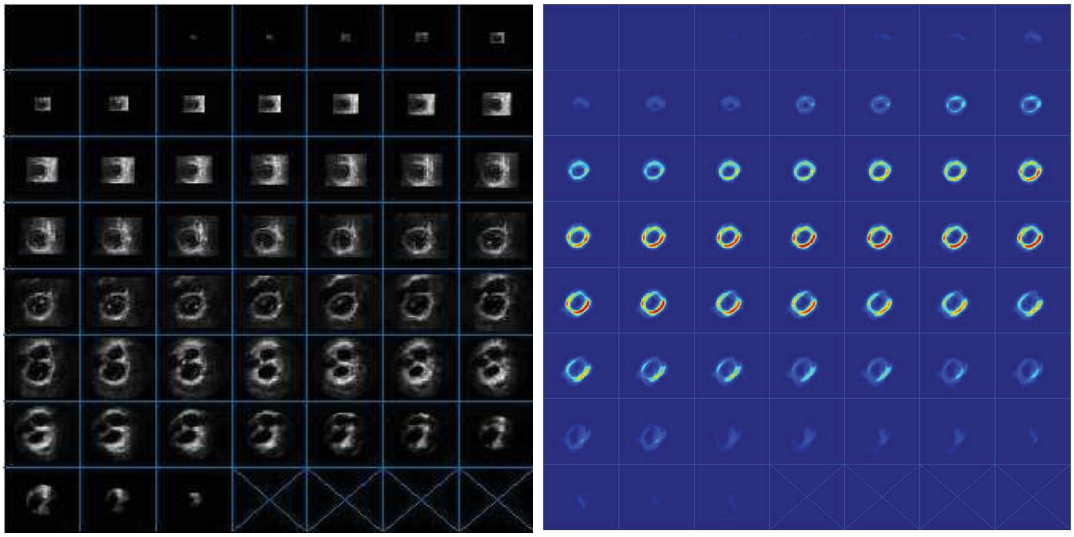
When applying our 3D SRF to a new image, 3D patches are extracted from the volume with a stride of 2 to accelerate the inference phase. Moreover, the probabilities across the image are normalized and a threshold is applied to discard low probability edges. A refining pass is also applied to obtain a more accurate prediction by keeping exclusively the predictions for the voxels previously associated to the endocardium.

TABLE 6.5: 3D SRF main hyperparameters

nb of trees	12	patch size	16	patches per tree	4×10^5
pair-wise features grid size	5	nb PCA components	7	1D vector length	512
HOG orientations	8	Feature scales	3 , 5	smoothing radius	2
max tree depth	32	min information gain	10^{-10}	min leaf size	8
fraction of features per tree	0.5	stride	2	probability threshold	20%

Hyper-parameters Table 6.5 provides the main hyper-parameters of our study. The most determinant (i.e. patch size, feature scales, number of patches per tree) were optimized heuristically by cross validation on a subset of 3D echocardiographic volumes. Memory consumption and training time were highly decisive in the choice of these values, as training a forest took several days using our hardware system described in Appendix C, Section C.2 (about six days were necessary for the last version).

Training and testing We trained trees using the ED manual ground truth references for a first forest and the ES segmentations for a second. For each phase, we randomly divided the entire CETUS dataset into 3 folds of 15 patients and learned for each test fold on the remaining 30 images. One should keep in mind that we had access to the manual references for the 45 ultrasound volumes of the CETUS dataset, while other participants could only work with 15 manual ground truth references. In summary, for each fold and phase, 12 trees were built, each learning from 400 000 3D patches of size 16 voxels, and using half of the available features, selected randomly. At test time, patches are extracted with a stride of 2 voxels and we apply a 0.20 threshold on the normalized probability edge maps.



(a) Slices of the 3D images

(b) Corresponding SRF edge probability map of the endocardium

FIGURE 6.16: 3D volumes slices and the corresponding edge map slices from the apex to the base.

6.4.2.2 Active Shape Model for 3D echocardiography

Model We used the 30 manually annotated volumes of the test set of the CETUS challenge to build two models dedicated to each time instant, i.e. one for the ED and one for the ES. The ASM we implemented is based on the model proposed in (Haak et al., 2015) [113]. In particular, from eq. 6.6 the vector of shape parameters b_s can be formulated as:

$$b_S = (P_s^T W P_s)^{-1} P_s^T W (s - \bar{s}) \tag{6.8}$$

where W is a diagonal matrix that weighs the deviation of each point of the evolving surface and P_s the covariance matrix resulting from PCA. The weight associated to each point s_i is:

$$w_i = w_{\text{US}}(s_i)^2 \times s_{\text{EPM}}(s_i) \tag{6.9}$$

where $s_{\text{EPM}}(s_i)$ is the probability that the point s_i belongs to an edge given by the Structured Random Forests. $w_{\text{US}}(s_i)$ is a ponderation applied to give less importance to points close to the ultrasound cone. The addition of this term is justified by the fact that our SRF implementation tends to provide false positives near the cone. Distance weights are obtained by convolving the binary ultrasound cone mask with a Gaussian kernel.

As in (Haak et al., 2015) [113], the shape parameters values are forced to remain inside a hyper-ellipsoid defined in the shape eigen vector space to avoid sudden shape variations. In addition, they are constrained at each iteration along with the pose parameters $T_{\text{SS/IS}}$ (translation, rotation, scaling), defining the transformation from the shape space to the image space. Then, an edge map outlier detection and correction is performed on the edge candidate points. In particular, points associated to strong shape deviations (more than 4 times the mean variation) along at least one eigenvector are considered to be outliers [114]. Their new position is inferred by projecting them on the alignment of the rest of the edge candidates, after rigidly registering them to the mean shape.

Initialization To initiate the ASM, three landmarks need to be manually selected on the B-mode image to guide an initial rigid transform $T_{\text{SS/IS}}$ between the image space (IS) and the shape space (SS): one at the apex, one for the mitral valve, and one for the aortic valve. $T_{\text{SS/IS}}$ involved translation, rotation and scale parameters. As the ASM tends not to follow the mitral valve plan only partially indicated on the edge maps, an adaptation of (Van Stralen et al., 2008) [115] is used to automatically detect it from the edge map.

The mitral valve plane is estimated using a spherical projection of the LV on a plane through the long axis of the prediction (LAX). The LAX is detected by considering 2D slices perpendicular to the acquisition axis. The center of the 2D edge maps (Fig. 6.16) is localized applying a circle Hough transform on each of them. The LAX is then inferred from these points using dynamic programming.

Iterative fitting The model is gradually fitted to the edge information by the following procedure. Using the edge probability map, we define new candidate points and update both the transformation parameters in $T_{\text{SS/IS}}$ and the shape parameters b_S. For each point of the shape, we seek a new candidate along the line normal to the boundary. The probability profile is weighted so as to decrease with the distance to the actual point and the maximum edge probability along this direction is used to update the point. Model points identified as belonging to the mitral valve ring are projected on the previously detected plane instead of the edge map points.

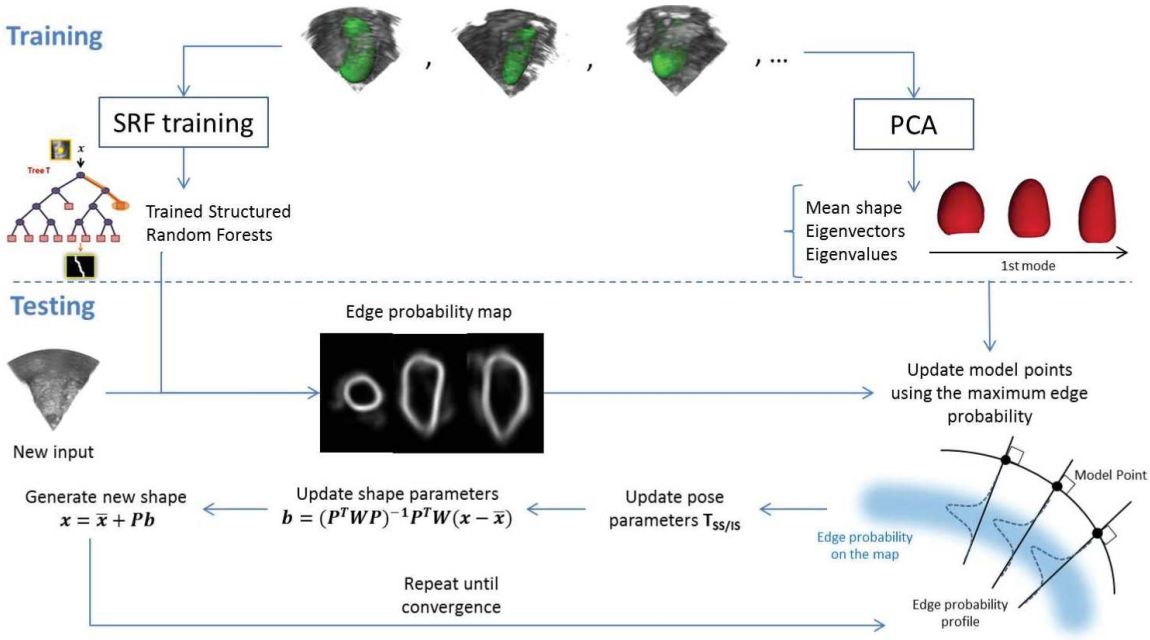


FIGURE 6.17: Overview of our complete pipeline combining SRF and ASM for 3D echocardiography segmentation.

The shape points and update points are used to refine $T_{\mathrm{SS/IS}}$ and the b_S vector is updated according to Eq. 6.8. The variation constraints are strengthened progressively so that shape variations become smaller as a local optimum is approached, which is ultimately detected by monitoring the difference between two successive shapes.

6.4.2.3 Segmentation pipeline

Figure 6.17 illustrates the full semi-automatic pipeline that was set up for this study. The shape model is built from a set of annotated data. On each test image, the SRFs automatically detect the endocardium. The ASM is then initialized using three manually indicated landmarks, and an automatic detection of the mitral valve plan from the edge map. The edge map further acts as the data-fidelity term on which the ASM is iteratively fitted. After convergence, the resulting surface is used as the final prediction of the endocardium.

6.4.3 Evaluation

The geometrical metrics used to evaluate the quality of the segmentation are the same as those described in 4.1 and used in our 2D study. We also assessed the capacity of our pipeline to estimate the ejection fraction LV_{EF} from the segmentation of the volumes at ED and ES (LV_{EDV} and LV_{ESV}).

To compare to other challengers, we ran our hybrid solution on the 30 testing ultrasound volumes of the CETUS dataset. All metrics (geometrical and clinical) were computed on the CETUS evaluation platform [116].

6.4.4 Results

Results include geometrical and clinical scores, as well as a visual analysis.

6.4.4.1 Geometrical scores

We compared our method to all ten algorithms of the challenge. In (Khellaf et al., 2018) [32], we provided the comparison of the results we obtained and those of the BEAS, the winner of the challenge. For comparison purposes, we also indicated the experts' inter-variability and the results of (Domingos et al., 2014) [33], whose approach is also based on a semi-automatic pipeline involving SRFs (see Section 4.3 for more details).

From tables 6.6 and 6.7, one can see our method achieved very low errors in terms of distance scores, outperforming the BEAS and the pipeline of (Domingos et al., 2014) [33] on all distance metrics. Compared to all 10 algorithms listed in the CETUS challenge, we ranked first for the MAD at ES, and obtained the 2nd lowest mean errors elsewhere. The distance scores for ED are slightly better than for ES, which is consistent with the results of other methods and can be explained, as in 2D, by the higher shape complexity at ES.

TABLE 6.6: Segmentation distance scores for end-diastole

Method	MAD	HD	1 - D
Inter-observer	1.39 ± 0.40	4.70 ± 1.27	0.069 ± 0.021
ours (30 training cases)	$\mathbf{2.04 \pm 0.48}$	$\mathbf{6.67 \pm 1.98}$	$\mathbf{0.101 \pm 0.024}$
BEAS	2.26 ± 0.73	8.10 ± 2.66	0.106 ± 0.041
Domingos and al. [33] (15)	2.09 ± 0.68	9.31 ± 3.89	0.106 ± 0.038

TABLE 6.7: Segmentation distance scores for end-systole

Method	MAD	HD	1 - D
Inter-observer	1.34 ± 0.35	4.70 ± 1.15	0.080 ± 0.021
ours (30 training cases)	$\mathbf{2.18 \pm 0.79}$	$\mathbf{7.76 \pm 2.47}$	$\mathbf{0.136 \pm 0.054}$
BEAS	2.43 ± 0.91	8.13 ± 3.08	0.144 ± 0.057
Domingos and al. [33] (15)	2.20 ± 0.72	8.35 ± 2.67	0.129 ± 0.050

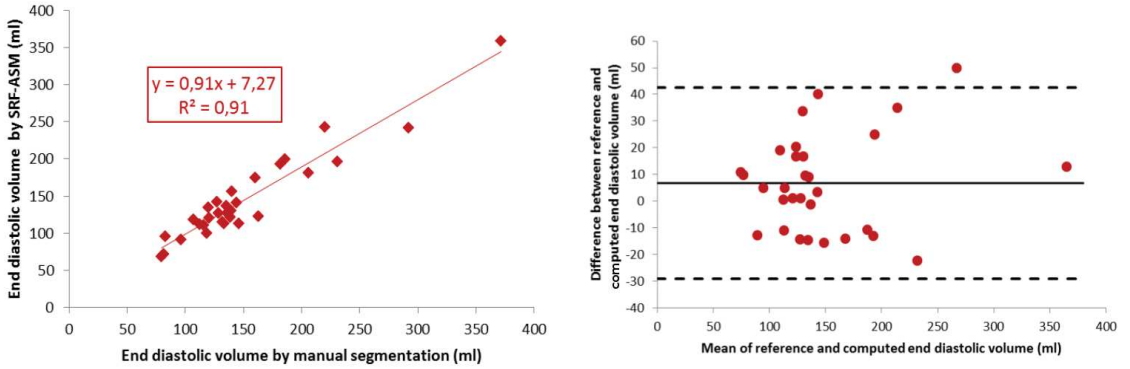
TABLE 6.8: Clinical indices scores for end-diastole and end-systole

	LV_{EDV} (ml)			LV_{ESV} (ml)			LV_{EF} (%)		
Algorithm	*corr**	*bias*	*std*	*corr**	*bias*	*std*	*corr**	*bias*	*std*
ours	0.044	6.66	18.26	0.040	**6.36**	15.05	0.127	**1.80**	5.64
BEAS	**0.035**	**−4.99**	**17.66**	**0.033**	−6.78	**13.86**	**0.111**	2.88	**5.24**

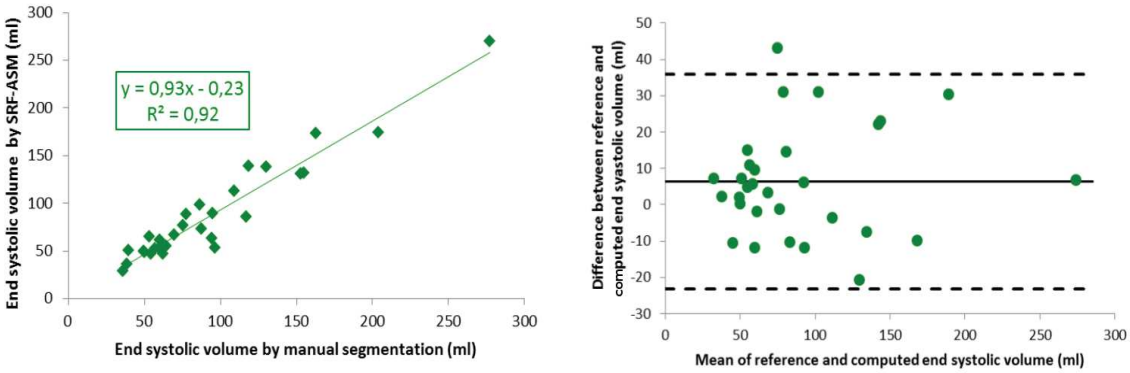
corr = 1 − r with r the correlation coefficient*
bias is the mean bias between ground truth values and calculated values.
std is the standard deviation

6.4.4.2 Clinical Scores

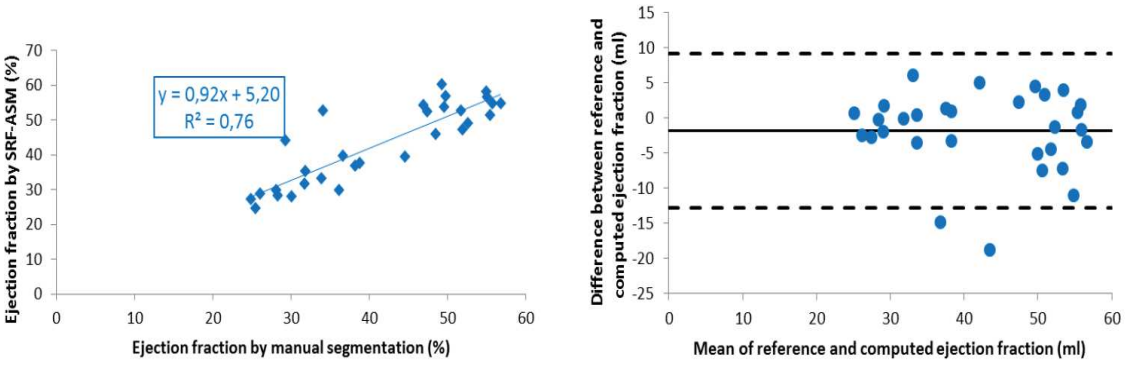
The corresponding clinical scores in Tab. 6.8 show lower correlations than the BEAS on all indices, and higher biases except for the volume at ES and the ejection fraction, but remain competitive. Ultimately, our global error measure computed according to the CETUS challenge guidelines, 0.627, is better than the BEAS' (0.644). Concerning the robustness of our algorithm, the linear regressions provided in Fig. 6.18 show that errors do not get higher for large volumes, while the Bland-Altman plots reveal the absence of consistent bias between the ground truth volumes and our predictions.



(a) LV volume at ED



(b) LV volume at ES



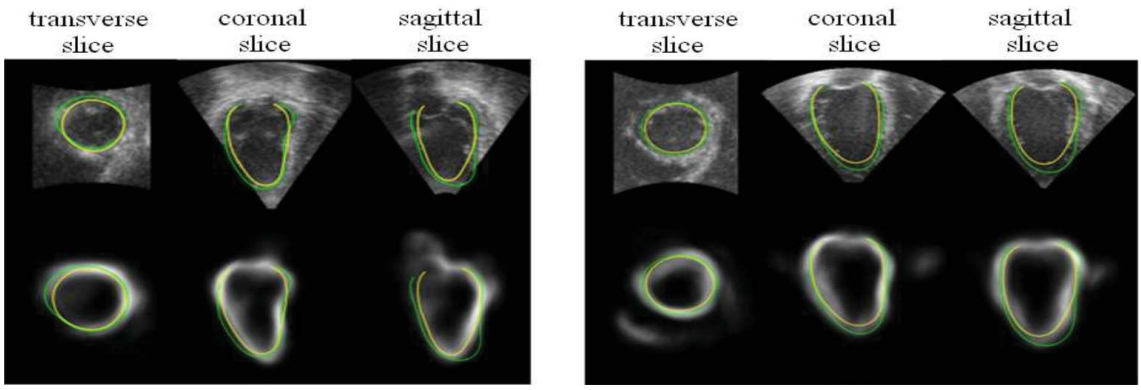
(c) EF

FIGURE 6.18: Correlation (left) and Bland Altman plots of all clinical indices.

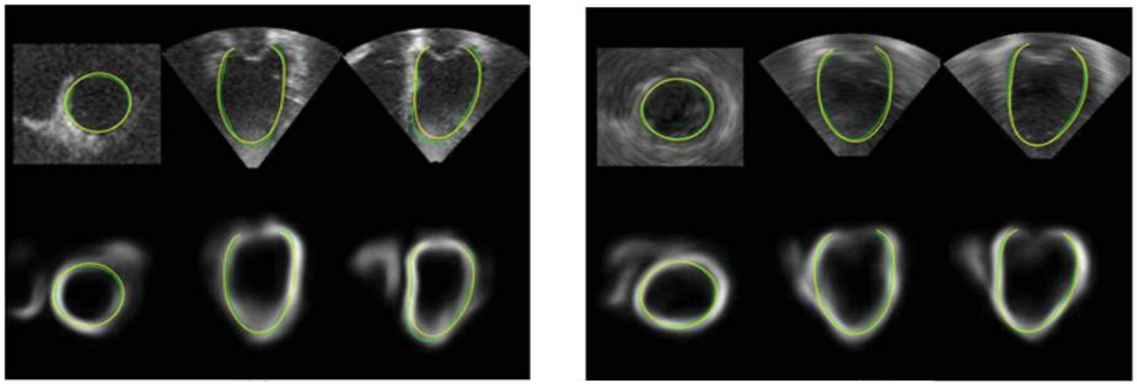
6.4.4.3 Visual analysis

We provide in Fig. 6.19 visuals of the two best and worst cases, to hint at our algorithm's behavior. On the worst segmentation (HD $\approx$ 12 mm), we can observe the SRFs mislocated the apex region. It should be noted that the apex is hard to detect even for a trained eye: during the creation of the reference delineations, the experts contours showed large differences near this area, and had to be reevaluated to reach consensus [16]. Edge map errors often correlated with signal dropouts, and most poor segmentations originated from local errors on the edge maps (apex, septal and lateral walls and basis). Since our SRFs use gradient-based features, they struggle on unusual patterns (e.g. signal dropouts, artifacts) that do not exist in the training set, and if the suggested shape is sufficiently plausible, it will not be corrected by the shape model. The two worst segmentations of Fig 6.19 a) presented large volume underestimations ($\approx$ 40 ml), and greatly contributed to the overall bias.

Even so, the predictions from our method always appear plausible. On Fig. 6.19 b), the two best cases show that our algorithm can be robust even when LV borders are not entirely visible. On the short axis view on the left, the edge of one of the LV walls appears to be partly missing, but our method still manages to get an accurate segmentation. As seen below, it is the combination of the edge map and the shape model that allowed this accurate detection: even though the edge prediction is blurry, the true edges are still associated with a weak probability that guided the shape model towards the true shape.



(a) Worst cases



(b) Best cases

FIGURE 6.19: Comparison between the prediction (yellow) and the groundtruth contours (green) of the endocardium on both the B-mode image and the SRF edge probability map.

6.4.5 Discussion

It is interesting to notice that the high HD values which characterizes the SRF predictions are lower in this study, which we believe is due to the ASM avoiding local errors by smoothing the LV shape. We noted that the outliers in the plots of Fig. 6.18 correspond to the same patients, which indicate as we observed in 2D that results could be greatly improved by enhancing the segmentation of only a small number of patients.

The lack of robustness of the SRFs, though reduced by the shape constraints, affected the whole pipeline. Still, with our combination of SRF and ASM, we could produce segmentation results on the CETUS dataset that outperform a few years later the winner of the challenge. Our method also performed better than the other pipeline involving SRFs. However, it should be noted that it:

1. is not fully automatic as the ASM requires inputs for the rigid registration initialization;
2. exploited training data for the SRFs that was not available to the participants, preventing a fair comparison. This shows however that even a slightly larger training set (30 instead of the 15) can make a machine learning pipeline outperform more classical image processing approaches.

6.5 Conclusion

In this chapter, we showed that:

- SRFs provide an interesting solution for multi-structure cardiac segmentation, as well as edge prediction in the context of echocardiographic imaging;
- the multi-scale strategy to compute low, mid and high-level features is key to obtain good segmentation results and accurate edge probability maps in echocardiography;
- SRFs can be successfully integrated into both fully automatic and semi-automatic pipelines;
- SRF results can be enhanced in a hybrid solution with active shape models.

The results that we obtained in both 2D and 3D echocardiography reveal that SRFs are a candidate of choice to solve the problem of segmentation in this modality. A key point of the algorithm is the manual feature extraction, which we believe is both a strength and a weakness of the approach. Well chosen, features enable a good generalization even on small datasets. Poorly chosen, they degrade the performance of the algorithm, and even if particular attention was paid to improving the creation of the feature space, it would be difficult to generalize sufficiently on a large set of heterogeneous echocardiographic images.

One alternative is to perform automatic extraction of features via deep learning, and learn the entire mapping from the image to the segmentation. In image processing, the current most efficient algorithms to do so correspond to convolutional neural networks. This motivated us to conduct a dedicated study presented in the next chapters.

Chapter 7

Assessing the potential of Deep Learning methods for the automatic segmentation of ultrasound images

This chapter covers the description of one of the most promising deep learning algorithm for the segmentation of medical images, the convolutional neural network U-Net, focusing on its properties and its potential for the automatic segmentation of the heart. For the purpose of obtaining insights on the best leads of improvement, we conduct experiments analyzing the behavior of the network. Especially, we want to answer the following queries:

- How many patients are needed to train a CNN to get highly accurate results in 2D echocardiographic image segmentation compared to the expert of reference ?
- How well do encoder-decoder (EDN) architectures perform compared to non-deep learning state-of-the-art techniques ?
- How accurate are the clinical indices estimated from the segmentation of CNNs compared to the inter and intra-expert variability ?

7.1 Introduction

7.1.1 Motivations

Since the breakthroughs in (LeCun et al., 1999) [118] and (Krizhevsky et al., 2012) [119], convolutional networks (CNNs) have gotten increasingly popular in image processing, becoming state-of-the-art in numerous image processing applications such as classification and segmentation, but also inpainting, generation, reconstruction, compression, text analysis, style transfer...

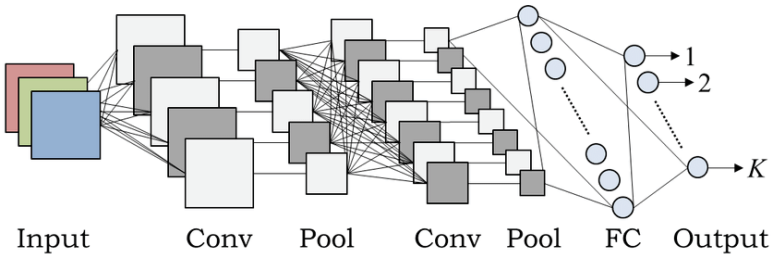


FIGURE 7.1: Generic CNN architecture [117]

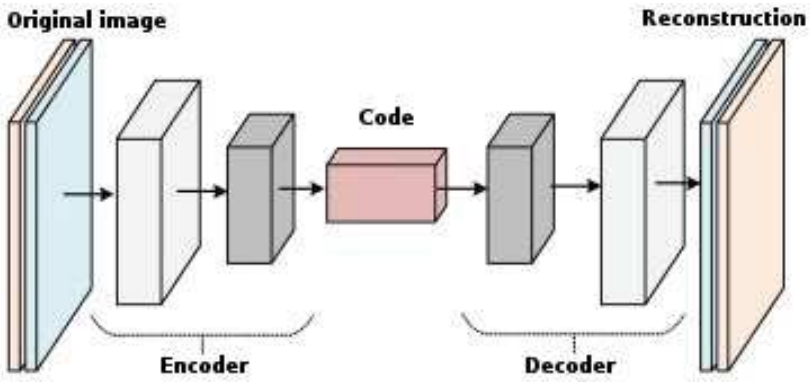


FIGURE 7.2: Auto-encoder architecture.

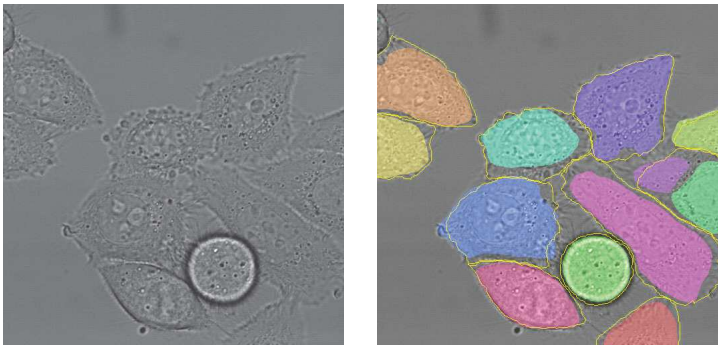
As seen in Section 4.3, the research in medical fields also established these technics as very promising for the future [26], though much concern is expressed concerning their clinical application and validation.

7.1.2 Convolutional neural networks

As part of deep learning technics, CNNs traditionally consist in pre-fixed models mapping the output directly from the input (Fig. 7.1) whose parameters are learnt by backpropagation. Their strength comes from the automatic learning of all intermediary transformations of the mapping (including feature extraction), and the surprisingly robust ability of backpropagation to converge to good local optima. Since their introduction, CNNs have gotten more and more stable and reliable by the addition of regularizing technics and optimizers [120].

7.1.3 U-Net

The U-Net architecture [34] was developed in the medical community as a segmentation network inspired from convolutional auto-encoders (Fig. 7.2). Since then, it established itself as state-of-the-art in numerous applications [121]. An example from the original paper is shown in Fig. **??**.



(a) original image

(b) segmentation, with the grountruth in yellow

FIGURE 7.3: Example of cell segmentation with U-Net in microscopy images (DIC-HeLa data set) [34].

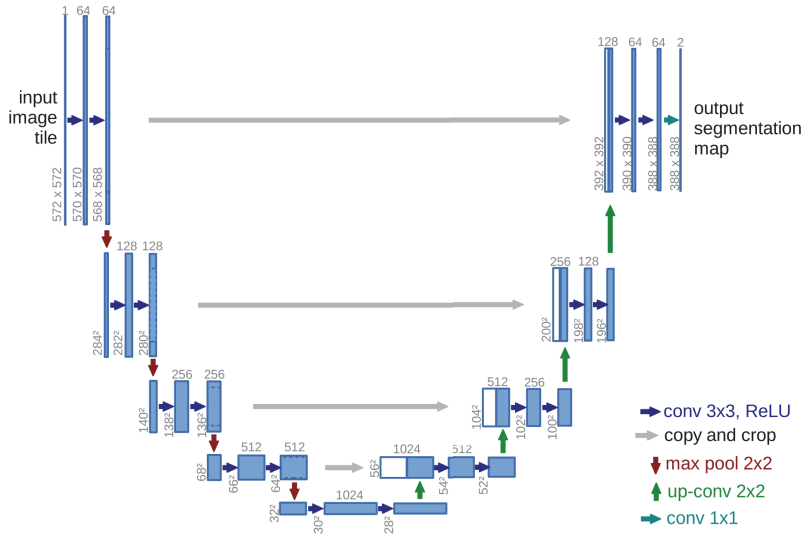


FIGURE 7.4: Representation of the layers and feature kernels of the U-Net [34]

TABLE 7.1: Original U-Net architecture (28M parameters)

Level	Layer	Kernel / Pool size	Activation	Connection
D1	Conv Conv MaxPooling	64 (3,3) 64 (3,3) (2×2)	ReLU ReLU	*
D2	Conv Conv MaxPooling	128 (3,3) 128 (3,3) (2×2)	ReLU ReLU	**
D3	Conv Conv MaxPooling	256 (3,3) 256 (3,3) (2×2)	ReLU ReLU	***
D4	Conv Conv MaxPooling	512 (3,3) 512 (3,3) (2×2)	ReLU ReLU	****
D5	Conv Conv	1024 (3,3) 1024 (3,3)	ReLU ReLU	
U1	UpConv Conv Conv	(2,2) 512 (3,3) 512 (3,3)	 ReLU ReLU	****
U2	UpConv Conv Conv	(2,2) 256 (3,3) 256 (3,3)	 ReLU ReLU	***
U3	UpConv Conv Conv	(2,2) 128 (3,3) 128 (3,3)	 ReLU ReLU	**
U4	UpConv Conv Conv	(2,2) 64 (3,3) 64 (3,3)	 ReLU ReLU	*
Seg	Conv	2 (1,1)	Softmax	

7.1.3.1 Architecture

U-Net is named after its symmetric shape, from the spatial compression of the encoding part to its reconstruction by the decoding part (Fig. 7.4). This sequential model is composed of blocks that each perform information filtering and spatial down/up -sampling. We call these blocks "levels", in a reminder of feature levels (Tab. 7.1). One particularity of U-Net is that as the spatial dimensions are reduced, the number of feature kernels increases to account for the loss of spatial information.

7.1.3.2 Differences between U-Net and auto-encoders

Similar to auto-encoders that learn to reconstruct an image from a latent sub-dimensional representation, U-Net is trained end-to-end and corresponds to a mapping $X \rightarrow E \rightarrow Y$, where E is a low dimensional space. However, it differs on three crucial points:

- The reconstruction space is different from the image space, hence why U-Net is called an encoder-decoder;
- The low dimensional code corresponds to a spatial compression and not a single vector;
- U-Net presents skip connections between the encoding and decoding parts that enable to extract features at several resolutions and re-use them for the reconstruction. This implies that in the case of U-Net, the decoder cannot work without the encoder.

7.1.3.3 Layers

The architecture contains many classical elements of CNNs, which we described here after, focusing on what is necessary to understand how encoder-decoder models such as U-Net work.

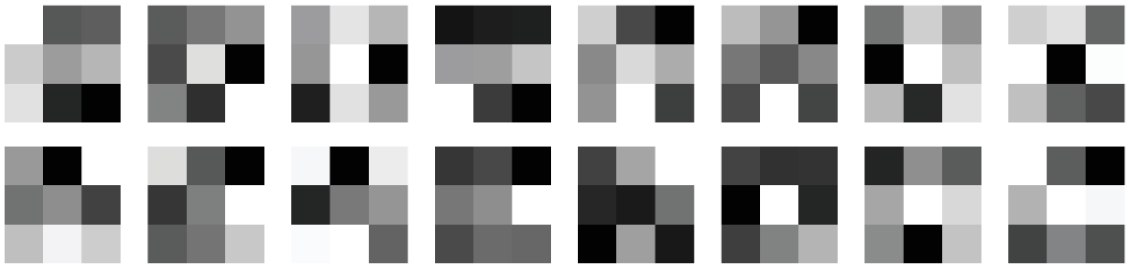
Convolutional layers All filters of a given layer function independently, in parallel with the others. Traditionally, filters mix information from all the input channels, so the number of output channels corresponds to the number of filters. The addition of filters therefore enlarges the network, while adding layers deepens the network.

To illustrate this in 2D and describe our notations in Tab. 7.1, let $width_{input} \times height_{input} \times nb_{channels}^{input}$ be the size of the input of a layer. A kernel $64(3,3)$ indicates that 64 filters of size $3 \times 3 \times nb_{channels}$ are applied to the input to compute the output. Hence, the output consists in 64 feature maps, each filter being the result of locally applying and translating a 9 pixels square filter that mixes the information of all the input image channels, as follows:

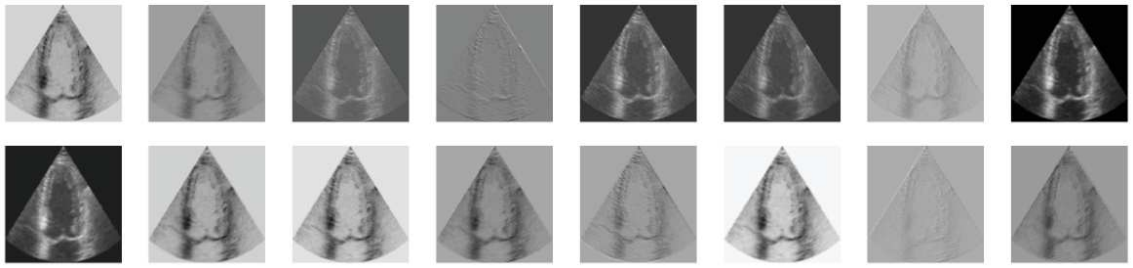
$$output[i,j] = \sum_{u=-1}^{1} \sum_{v=-1}^{1} \sum_{c=1}^{nb_{channels}^{input}} w[u,v,c] \times input[i-u, j-v, c] \tag{7.1}$$

with w the weights of the filters and $[i,j] \in [1, width_{input} - 1] \times [1, height_{input} - 1]$.

Except for the first layer, feature maps are difficult to interpret since they result from several operations and store complementary information. In Fig. 7.5, we display the filters of the first layer with the corresponding feature maps for a U-Net model trained on a subset of CAMUS. As expected for a segmentation task, the learned filters perform operations close to edge detection, inversion and blurring, although none corresponds to perfectly symmetrical hand-crafted filters [120]. Some feature maps appear redundant, but as the corresponding filters are not identical, they can server different purposes when combined.



(a) First layer filters ($nb_{channels}^{input} = 16$)



(b) Corresponding feature maps

FIGURE 7.5: Visualization of the first layer filters learned by a reduced U-Net, and corresponding feature maps on a random validation image [122].

Convolutions are not applied on border pixels since they require surrounding information. Consequently, the feature maps spatially decrease in size if no image padding is performed (i.e adding pixels around the edges of the image), or if the convolution is applied with a stride (i.e sampling values with a fixed pitch). As convolutions unfold, the receptive field of the network (i.e the processed zone in the image) is gradually enlarged as illustrated in Fig. 7.6. Instead of using large convolution filters, it is therefore more efficient to stack small filters since it creates a similar receptive field with fewer parameters [120]. Another possibility is to use dilated convolutions, i.e filter parameters are spaced apart to cover a wider receptive field.

Ultimately, each filter locally performs a linear transformation of the input (Fig. 7.7 a.). Similar in spirit to SVM kernels, a learnable bias is added to the filter parameters to enable the transformation not to preserve the center. The number of parameters learnt at each convolutional layer is therefore:

$$|w| = nb_{channels}^{output} \times (|w[c]|^2 \times nb_{channels}^{input} + 1) \tag{7.2}$$

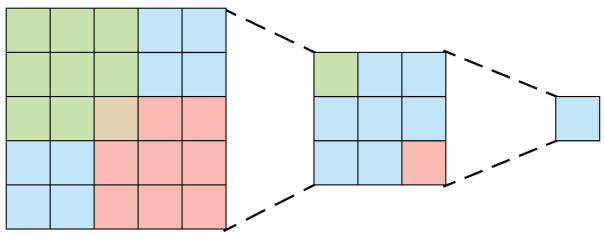


FIGURE 7.6: Receptive fields [123]: After 2 3×3 convolutions, the pixel on the right contains global information about the 5×5 region on the left. As CNNs get deeper, the receptive field increases, allowing to extract high-level features.

Activation layers A succession of linear operators (multiplication and sum) always amounts to a single linear operator. Nonlinearities are therefore applied after each convolutional layer to obtain a nonlinear mapping between the input and the output. If the activations were stumps, activated convolutions would perform decision as SVMs (Fig. 7.7 a.). In practice, neural network activations contain one or two nonlinear parts and a linear part (Fig. 7.7 b.). Such functions are traditionally called activations because they originally projected values onto 0 / 1. The sigmoid (Eq. 7.3) and ReLU (Eq. 7.4) are the two most popular.

$$sigmoid(x) = \frac{1}{1 + \exp(-x)} \tag{7.3}$$

$$ReLU(x) = max(0, x) = \{x < 0 : 0, x \geq 0 : x\} \tag{7.4}$$

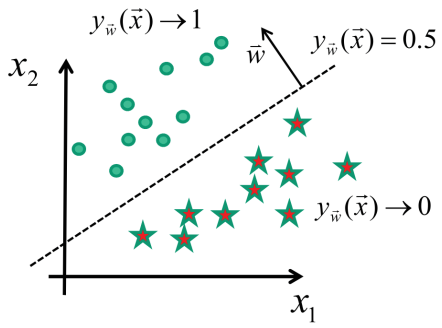
The sigmoid function is close to a stump between 0 and 1 while being fully differentiable. However, stacking sigmoids can cause downstream gradients to become very small [120]. To avoid the vanishing gradient effect, the ReLU activation is volontarily unbounded for positive values, however it may instead foster exploding gradients and the "death" of neurons due to absence of gradients for negative values. There exists several modifications to avoid this particular point (Fig. 7.7 b.) but with the Adam optimizer [124] (described here under) to scale gradients, ReLU is usually very efficient.

The softmax activation returns the normalized exponential of a set of values x (Eq. 7.5):

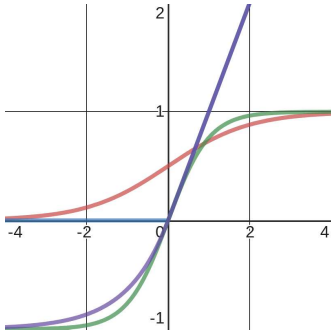
$$softmax(x_i) = \frac{\exp(x_i)}{\sum_{j=1}^{n} \exp(x_j)} \tag{7.5}$$

where n is the number of values in x.

This transformation maps a set of values between 0 and 1 so their sum equals 1. Its output can be interpreted as a probability. For segmentation, we use the softmax to derive class probabilities for each pixel by setting the number of last feature maps to be the number of classes in the image (including the background), and applying the activation along the channel



(a) Linear separation. The sigmoid has a similar behavior, except values close to 0 are linearly shifted while points far from the decision frontier tend towards 0 or 1 rather than being set to either. [36]



(b) Activation functions. Sigmoid in red, ReLU in blue, hyperbolic tangent in green and ELU in purple

FIGURE 7.7: Behavior of activated convolutions: while the convolution filters map values to another representation, the activation sets saturation values.

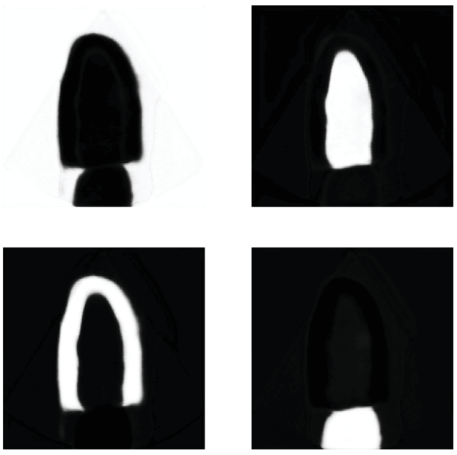


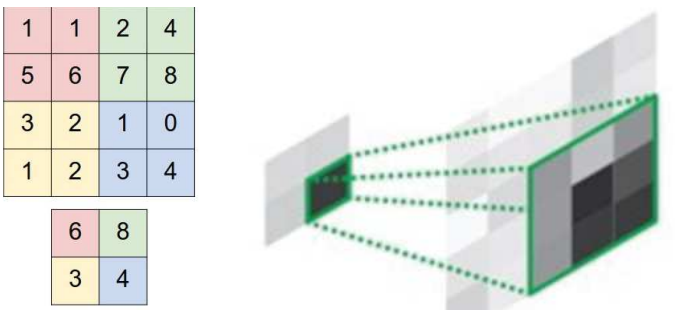
FIGURE 7.8: Softmax output for the image in Fig. 7.5. Here the U-Net is used to segment the LV (top right), the myocardium (bottom left) and the LA (bottom right).

axis (Fig. 7.8). Contrarily to other activations that are used in between convolutional layers, the softmax is therefore traditionally used only once, to conveniently format the output.

Spatial down and up - sampling The spatial downsampling is usually performed with pooling layers. Pooling consists in keeping a single value for a region. The role of pooling is two-fold [125]:

1. Reduce the spatial resolution (and hence the memory consumption);
2. Provide (partial) invariance to rotation.

Average Pooling (i.e taking the mean value) and Max Pooling as done in U-Net and illustrated in (Fig. 7.9 a.), are the most frequent. For a downsampling of 2, we consider 2×2 regions. The upsampling consists in increasing the resolution of the input feature map. It can be done either using nearest neighbor (upsampling layers) or by learning parameters to weigh the influence of surrounding pixels (Upconvolutions, also called deconvolutions and transposed convolutions), as shown in Fig. 7.9 b.).



(a) 2×2 Max Pooling [120]

(b) Transposed convolution [126]

FIGURE 7.9: Down and Up-sampling in CNNs examples.

7.1.3.4 Training phase

Loss function U-Net is trained by updating its parameters so the output fits an objective function. As in classical optimization problems, we usually optimize the network on a cost / loss function rather than the objective, so that we iteratively search the solution space for a minimum instead of a maximum (Fig. 7.10 b.). U-Net originally was trained with a binary cross-entropy loss. For multi-class segmentation, the most frequent cost functions are the categorical cross-entropy (Eq. 7.6) and the multi-class dice loss (Eq. 7.7) introduced in [35]. They are computed between the model predictions p and the ground truth masks gt:

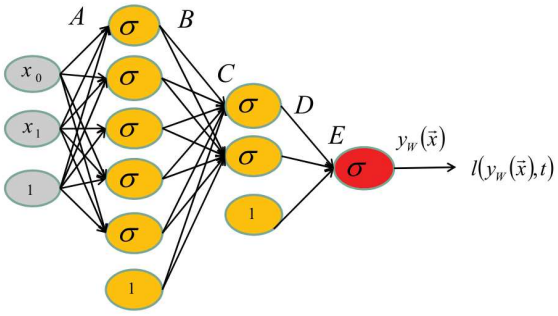
$$CCE(p, gt) = -\frac{1}{N \times (|C|)} \times \sum_{c=0}^{|Y|-1} \sum_{i=1}^{N} gt(i=c) \log(p(i,c)) \tag{7.6}$$

$$DL(p, gt) = 1 - \frac{1}{|C| \times N} \times \sum_{c=0}^{|Y|-1} \sum_{i=1}^{N} \frac{2 \times p(i,c) \times gt(i=c)}{p(i,c) + gt(i=c)} \tag{7.7}$$

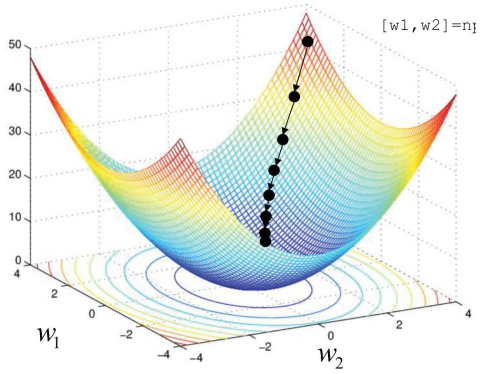
with $p(i,c)$ the probability for the sample i to belong to class $c \in Y$, and ϵ a small number added to avoid divisions by zero. The values $gt(i=c)$ are constants and binary:
$gt(i=c) = 1$ when $i = c$, 0 otherwise.

Back-propagation First introduced in (Rumelhart et al., 1986) [25], backpropagation has enabled to efficiently train multi-layers networks. The error on the loss function is propagated to each parameter [25] according to the chain rule, from the output to the input, as seen in Fig. 7.10. This implies that each layer is updated to correct the error independently from the others, which allows to express the current error as a function of each weight individually. A small step in the inverse gradient direction is then taken for each parameter w.

$$\frac{\partial(l(y_W(\vec{x}),t))}{\partial W^{[0]}} = \frac{\partial(l(y_W(\vec{x}),t))}{\partial y_W(\vec{x})} \frac{\partial(y_W(\vec{x}))}{\partial E} \frac{\partial(E)}{\partial D} \frac{\partial(D)}{\partial C} \frac{\partial(C)}{\partial B} \frac{\partial(B)}{\partial A} \frac{\partial(A)}{\partial W^{[0]}}$$



(a) Backpropagation illustration on a fully connected neural network [36]



(b) Optimization landscape: iterative convergence from a random initialization [36]

FIGURE 7.10: Main components of the optimization in DL: a) The gradient to apply to W to correct the error on the objective is obtained by applying the chain rule to all intermediary layers. Each weight is updated by a fraction of their gradient (learning rate). b) Progressively, the model converges to a local minima of the loss function computed on the training dataset.

Optimizer The training dataset is usually too large to be entirely used for gradient descent, so we use iterative algorithms to update the weights, called optimizers. The Stochastic Gradient Descent used in the original U-Net is the vanilla update [120]:

$$w = w - \lambda \times dw \tag{7.8}$$

with λ the learning rate and dw the backpropagated gradients computed on a single random example or on a random subset of the training data (referred to as mini-batch).

Working with mini-batches denoises the gradients while accelerating the training phase compared to using a single element, and lowering the memory consumption compared to using the full dataset. It participates in making the training process stochastic. Mini-batches are usually sampled without replacement so that each image is seen the exact same number of time, once per training epoch. At the end of each epoch, we reshuffle the data and create a new set of mini-batches.

Including momentum to the update process allows to take into account the landscape depicted by the loss function and the training set by adapting the learning rate to the slope and integrating the speed v of the last updates:

$$\begin{aligned} \hat{v} &= -\lambda \times dw + \beta \times v \\ \hat{w} &= w + \hat{v} \end{aligned} \tag{7.9}$$

This requires the introduction of the hyper-parameter β to balance the influence of previous speed and directions and the newly suggested step. There exists several other optimizers that either work on improving the gradient momentum as Nezterov momentum [120], or on adapting the learning rate to the magnitude of gradients, as adagrad [127] and rmsprop [128]. The most popular at the moment, Adam [129], intends to do both:

$$\begin{aligned} \hat{v} &= \beta_1 \times v + (1 - \beta_1) \times dw \\ v' &= \frac{\hat{v}}{1 - \beta_1^t} \\ \hat{a} &= \beta_2 \times a + (1 - \beta_2) \times (dw^2) \\ a' &= \frac{\hat{a}}{1 - \beta_2^t} \\ w &= -\frac{\lambda \times v'}{\sqrt{a'} + \epsilon} \end{aligned} \tag{7.10}$$

with ϵ a non null number (usually very small) to avoid divisions by zero, and t the number of iterations since the beginning of the training. The original version of Adam used $\hat{v}$ and $\hat{a}$ instead of v' and a'. The hyper-parameters β_1 and β_2 are usually slightly lower than 1 (0.90, 0.99 ...) in order to slowly decrease the influence of past steps. Since then, several augmentations of Adam [124] have been proposed (adamax, nadam, amsgrad...).

Regularizers Regularizers are added to the network to reduce the risks / amount of overfitting by hindering the training process on purpose through:

- setting penalties, such as weight regularization that forces the weight values to be small;
- reducing the model capacity (dropout).

Dropout [130] is used at the end of the contractile path of the U-Net. This technique consists in randomly shutting off a certain amount of neurons at a given layer. In CNNs, we usually

shut entire filters off. This will encourage the model to rely on multiple filters to perform a decision, while possibly favoring redundancy of the most important features. It is presumed that dropout has a denoising effect similar to ensemble models (Fig. 7.11), as we could consider that several small CNNs are trained in parallel and participate to the prediction. The notion of ensemble models is more thoroughly presented in Chapter 6 through the example of the Random Forest algorithm.

Initialization Initialization is a crucial step for a good convergence. Weights need an asymmetrical initialization because otherwise, if the weights all had the same value, they would participate similarly to the error, be associated to the same gradient, and maintain identical values. We usually pick small random values sampled from a centered gaussian distribution with a normalized variance [120], so the outputs of $\sum_{i=1}^{N} w_i \times x_i$ have an expectation of 0 and a variance of 1.

(Glorot et al., 2010) [131] proposed a popular variance normalization based on the number of input and output units, used either with normal or uniform distributions. Ideally, it is the activation outputs we should be normalizing, so (He et al., 2015) [132] also proposed a normalization appropriate for ReLU activations. In the original U-Net, the authors used He's initialization, and the weights were thus sampled from a centered gaussian distribution of variance $\sqrt{\frac{2}{N}}$, with N the number of weights of the layer.

Randomness in the training process of neural networks The random initialization and the stochastic optimization are the theoretical reasons why a same model does not converge twice to identical weights and performance. In practice, the hardware (GPUs, Multi-thread CPUs) also adds randomness to the training [133] [134].

Pre-processing As an encoder-decoder, U-Net takes the whole image as input. In the original U-Net, images were resized to a fixed size of [572, 572] to be coherent with the decrease of the image size. If no padding is used, the image size is indeed reduced to $[width_{output}, height_{output}] = ([width_{input} - 2, height_{input} - 4])/2$ at each level (two convolutions $3 * 3$ and a downsampling of factor 2 by the maxpooling operation). Another option is to use image sizes several times multiples of 2, and a conservative padding.

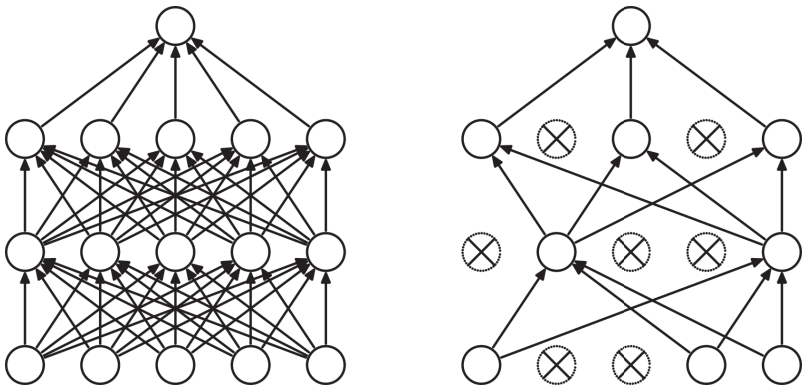


FIGURE 7.11: Dropout illustration [130]. Left: base network, right: network with a dropout rate around 40%.

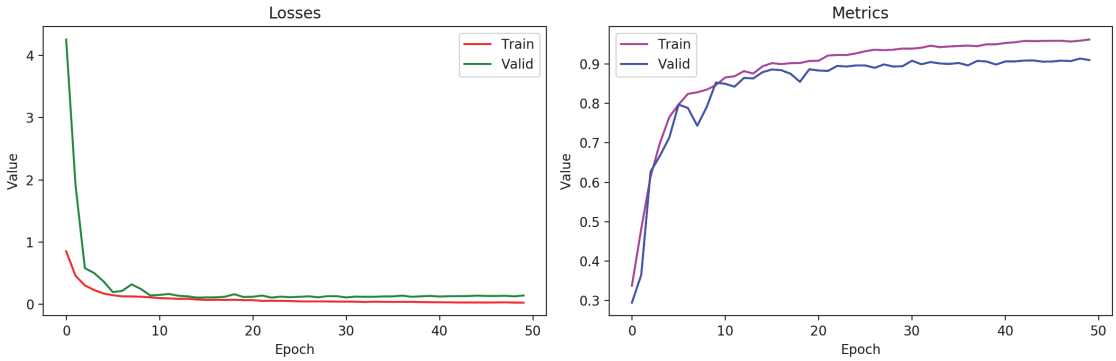


FIGURE 7.12: Training curves of a reduced U-Net trained on 200 patients and validated on 100 with a categorical cross-entropy loss and the Adam optimizer with a learning rate of 2×10^{-3}. Dice results are given on the right. [122].

Few training images (20, 30 and 35) were available for the experiments in the original paper of (Ronneberger et al., 2015) [34], so the authors used data augmentation to improve the robustness of their model. Data augmentation is traditionally the creation of virtual examples from real ones through the application of classical transformations. These transformations can impact the geometry of the image (translation, rotation, scaling, shearing, ...) or the histogram (dataset or sample normalization, brightness shift, ...) [135]. Appropriate data augmentation allows to complete a given dataset with plausible variability. For U-Net, the data augmentation consisted in small and smooth random deformations.

Training curves To assess whether the model was rightfully tuned and properly trained, we can display training curves, i.e the performance on both the training and validation sets (Fig 7.12) for the losses and metrics that were used. From these curves, we observe how the model performed on the training and validation sets at each epoch. Especially, we can assess:

1. whether the training loss converges to a low value (low bias), and if it does so smoothly, which hints at good optimization choices
2. whether the inflexion points of the two losses are close in time and value, which supposes a good balance of the data sets in terms of variety
3. whether the validation loss converges to a value that is close to the training one, revealing a good model choice (low variance)
4. how many epochs are necessary to properly train the model on this dataset by detecting the start of overfitting as the moment the validation loss stops improving
5. whether the metrics follow the trend of the loss (or the inverse if they are to be maximized), which gives us information on the loss' suitability to the problem at hand

On Fig. 7.12 for instance, the curves indicate that beyond 30 training epochs, the network started over-fitting the dataset. Therefore, we should either lower the number of epochs or drop the learning rate to get a smoother convergence. The training phase shows the model has low variance, however the close values between valid and train losses may suggest the model has high bias with regard to the training set, i.e has not enough parameters to solve the training set. Therefore, a wider or deeper network might get better performance. Indeed, a model of sufficient capacity should be able to over-fit the training dataset when given

a sufficient number of epochs. The validation set allows to detect when the model starts learning specificities of the training set rather than general characteristics of the problem, so in practice we keep the model configuration which worked best on the validation set.

7.1.3.5 Testing phase

At test time, the model prediction for a new image is obtained by performing a forward pass: the input image is resized, and processed from layer to layer. Instead of computing the loss function between the prediction and the output to derive the error to backpropagate as we would in the training phase, the obtained mask is resized to the original image size and evaluated on an appropriate set of metrics.

No post-processing is performed in the original version, except averaging the output of 7 rotations of the image. The output of the softmax is used as an edge probability map in the ISBI 2012 challenge [34] and thresholded to compute the intersection over union (IOU) on the ISBI cell tracking challenges of 2014 and 2015. For all these challenges, U-Net outperformed the past results, even doubling the best score for the third challenge.

7.2 Potential of U-Net for 2D ultrasound segmentation

7.2.1 Optimizing hyper-parameters

We implemented our version of U-Net in the vitalab Python library [136], using the Keras API [135] with tensorflow as back-end. As hyper-parameters are interdependent and not straightforward to interpret, it is common to conduct a random search (exploration) over the hyper-parameter space, and a grid search around the best values found. However, this optimization scheme does not provide insights on the hyper-parameters' influence.

Similar to how we optimized the hyper-parameters of SRF, we decided to tune each individual hyper-parameter on a subset of CAMUS (300 training patients, A4C chamber views, 30 for validation and 50 for test) by trying a few values adapted to train the network on a 8 GB GPU. We observed on CAMUS that the optimal parameters were dependent of the chosen training, validation and test sets, and that most had very little impact. Therefore, a careful tuning was unnecessary and potentially misleading. Tab. 7.2 summarizes the final hyper-parameter values we converged to.

7.2.2 Comparison to SRF in 2D echocardiography

This work has been published in the IEEE IUS conference 2018 [31].

7.2.2.1 Motivations

Our first inquiry was to determine whether the features automatically learned by U-Net, and the U-Net model in general, would outperform the SRF with our handcrafted features. Especially, we wanted to confirm our intuition that:

- SRF make the most of small datasets;
- CNNs would generalize well on our application when given enough training data, which is related to the task complexity.

TABLE 7.2: Differences between the original U-Net and our first architecture. Unmentioned parameters are unchanged.

Hyper-parameter	Original	ours	observation
width	64	32	better than 16 or 48;
batch size	1	10	non-linear impact;
initialization	He	Glorot	slightly better results
padding	none	same (size)	practical to
skip connection	copy and crop	concatenation	avoid cropping
loss function	Binary CE	Categorical CE	
optimizer	momentum	Adam	better (dynamic λ);
λ	$10e^{-2}$	10^{-4}	smoother convergence;
regularizer	dropout on the bottom layer	spatial dropout 20% + batchNorm	rate has to be small; global improvement
pre-processing	resize $[572, 572]$	resize $[256, 256]$	$\downarrow$ memory usage;
data augmentation	small deformations	none	data aug. unhelpful;
post-processing	averaging	blob selection	remove small false pos;
validation set	-	60 images	stabilized the val. loss

The data augmentation involved affine transformations

7.2.2.2 Evaluation

Dataset By the time of this study, the complete CAMUS dataset had been annotated by O_1, as presented in Chapter 5. We compared the algorithms on two training dataset sizes, 50 patients annotated at ED and ES on A4C and A2C views (200 images) and 400 patients (1600 images), which is the most we can use to keep one fold aside for validation and one for test. The validation loss was tracked to select the best U-Net model, i.e the one that generalized the best with regard to the validation set. As we preferred to keep the exact same training dataset for both algorithm, we did not add the validation set to the training set of the SRFs. Unlike the experiment in Section 6.3, we gathered the annotated images at ED and ES, A4C and A2C into a single training set to assess the potential of both methods in coping with instant and view variability.

SRF To account for the increase in training set size, we trained 12 random trees per subset of 50 patients (200 images) by extracting 2×10^5 patches, half of which include an anatomical border of interest (i.e the endocardium or the epicardium). The features included the intensity and the HOG at scales 1, 3 and 5.

Metrics Both algorithms were trained to segment the LV and myocardium simultaneously. We compared the algorithms on the three geometrical metrics introduced in Section 4.1, using a same evaluation platform based on the medpy library [137].

7.2.2.3 Results

Tables 7.3 and 7.4 summarize the results for the segmentation of the endocardial region $\boldsymbol{LV_{endo}}$ and epicardial regions $\boldsymbol{LV_{epi}}$. With only 50 cases (Tab. 7.3), the SRF systematically showed better but inconsistent mean performances as suggested by the standard deviation values. With 400 training cases, the U-Net obtained higher scores with a significant margin, as seen in Tab. 7.4. Visuals of average results for the two algorithms are provided in Fig. 7.13.

TABLE 7.3: Comparison between U-Net and SRF
Training size = 50 patients / 200 images

Model	LV_{endo}			LV_{epi}		
	D	**MAD**	**HD**	**D**	**MAD**	**HD**
U-Net	0.795 **±0.089**	4.1 **±1.9**	15.7 **±7.6**	0.838 ±0.083	5.3 ±2.8	18.3 **±8.5**
SRF	**0.843** ±0.095	**3.7** ±2.1	**13.9** ±9.5	**0.894** **±0.058**	**3.7** **±1.9**	**14.6** ±9.0

TABLE 7.4: Comparison between U-Net and SRF
Training size = 400 patients / 1600 images

Model	LV_{endo}			LV_{epi}		
	D	**MAD**	**HD**	**D**	**MAD**	**HD**
U-net	**0.896** **±0.047**	**2.3** **±1.0**	**7.3** **±3.2**	**0.931** **±0.028**	**2.6** **±1.1**	**8.1** **±4.7**
SRF	0.859 ±0.092	3.3 ±2.1	12.7 ±9.6	0.896 ±0.06	3.7 ±2.0	14.2 ±9.0

7.2.2.4 Discussion

We make three main observations from this comparison:

1. The SRF performed slightly worse than previously observed in Section 6.3. This is most likely due to the increase in task complexity, as we learn to segment without prior knowledge regarding the instant (ED or ES) and view (A4C or A2C) of the frame. The enlargement of the training set imposed stricter restrictions on memory and it is possible that the patches learned did not provide a sufficient data overlap between trees. (Another explanation is that we simply do not use the same training and test sets.)

2. More troubling, the SRF only slightly benefited from a $8x$ increase in the training set size. The standard deviation remained about the same, suggesting they did not get better at dealing with the variety, as we would hope. The SRF appear to reach a plateau not linked to the training dataset.

3. Our U-Net almost doubles its scores from Tab. 7.3 to Tab. 7.4 in both average performance and standard deviation. The variance is therefore divided by 4 by the addition of more cases. This confirms that more training data was necessary to train the U-Net, but also that it has the potential to further learn from new cases.

As mentioned previously, it is our belief that using handcrafted features simplifies the problem by focusing on relevant information, enabling good results on small datasets. However, doing so also limits the access to the information present in the ultrasound image, which is revealed by the plateau SRF reach.

Oppositely, the U-Net neural network which learns features automatically showed better performance as more training data was added because the feature extraction and interpretation became more robust. We therefore wondered whether the improvement was linear with the training set size, and if U-Net had reached its best performance on our application.

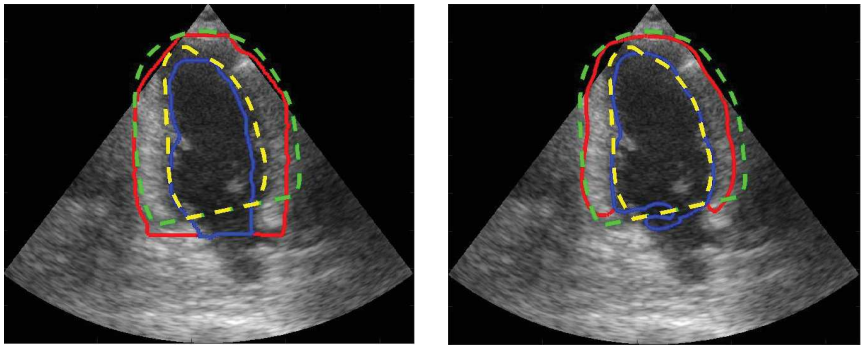
7.2.3 Influence of the training dataset size

7.2.3.1 Motivations

Though convolutional Neural Networks showed their potential in reproducing experts' actions on several medical applications, the amount of training data needed for a CNNs to achieve competitive results is unclear. We instinctively suppose that:

- it is task dependent (complexity, evaluation);
- quality of data is more relevant than the data quantity as the scope of a task rather comes down to data variety. In theory, we could extract from a very large training set a smaller set with the same heterogeneity (active learning [138]).

Much effort has been made in assessing the importance of size effect when comparing statistical results [139], but to our knowledge, there is no established solution to predict beforehand the minimum amount of training data needed for a CNN to solve a given image processing task. Therefore, we designed an extremely simple but efficient experiment to assess whether our training dataset was of sufficient size to allow proper generalization, focusing on 2D ultrasound segmentation of the heart.



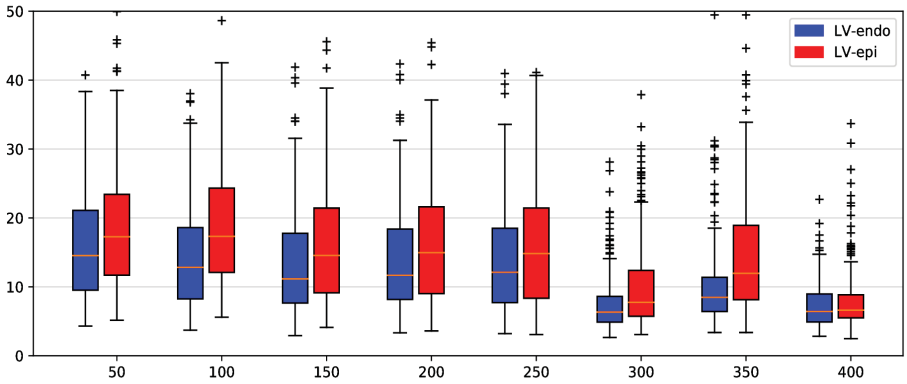
(a) D: 0.845 | 0.890 . MAD: 4.2 | 4.4 mm . HD: 15.1 | 19 mm.

(b) D: 0.922 | 0.911 . MAD: 1.9 | 3.6 mm . HD: 7.9 | 7.9 mm.

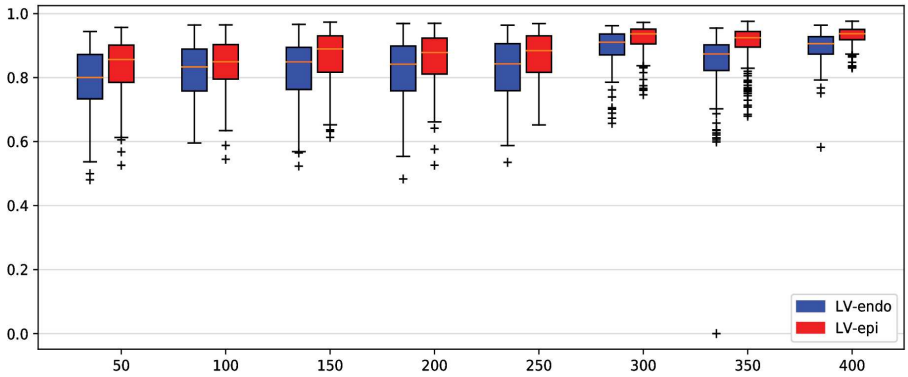
(c) D: 0.852 | 0.925 . MAD: 2.4 | 2.5 mm . HD: 8.3 | 5.7 mm.

(d) D: 0.796 | 0.899 . MAD: 4.4 | 3.5 mm . HD: 12.3 | 9.9 mm.

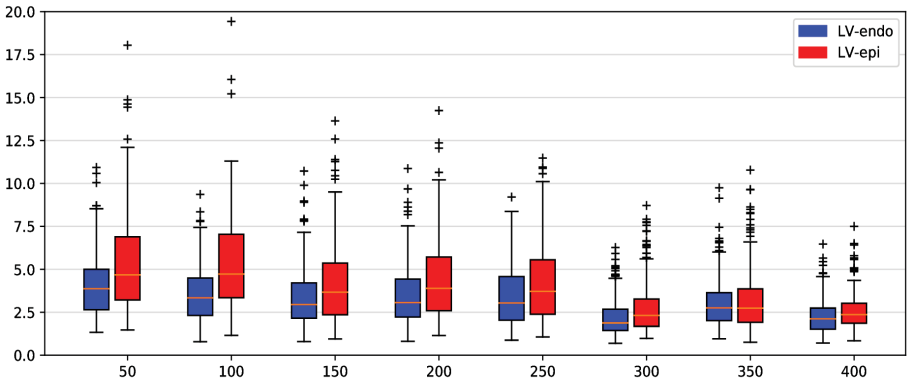
FIGURE 7.13: Average peformance representations. Left: U-Net, right: SRF. Top: 50 patients, bottom: 400. "metric: LV_{endo} | LV_{epi}.



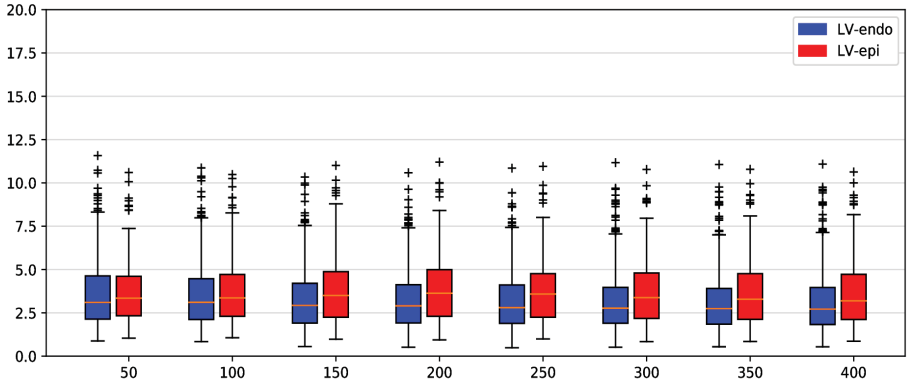
(a) Hausdorff distance (mm) evolution. The image is truncated at 50 mm for visibility.



(b) Dice evolution



(c) MAD distance (mm) evolution



(d) MAD distance (mm) evolution (SRF)

FIGURE 7.14: Evolution of the three geometrical metrics for an increasing training set size, from 50 to 400 patients (i.e 200 to 1600 images).

7.2.3.2 Evaluation

We kept the same test and validation set, and increasingly added other folds to the training set. The number of epochs was adapted so that the models had the same number of weight updates, with all models having converged with regard to the validation loss. Other hyper-parameters (architecture, learning rate, batch size etc...) were constant. Statistical results on the three geometrical metrics (Dice, MAD, HD) are illustrated with Tukey boxplots. They summarize distributions by showing the median, lower and upper quartiles. The whiskers are set at $1.5 \times IQR$, with IQR the inter-quartile range.

7.2.3.3 Results

Boxplots are provided in Fig. 7.14. For comparison, we also provide the MAD boxplot of the SRF. U-Net showed distinct improvement on the three metrics on both structures when increasing the training size. Interestingly, this improvement is not linear nor consistent:

- the evolution from 250 to 300 shows a much higher gain than for any other subset;
- the evolution from 300 to 350 corresponds to a decrease in performance.

The noisy evolution is likely due to the evaluation of a single combination of training-validation-test sets. A denoised curve could be obtained by performing cross-validation on the test set [140]. The overall trend on this particular setting is sufficient to suggest that:

- the robustness of the U-Net gets much better as we add training cases, as expressed by the reduction of the number and acuteness of the prediction;
- U-Net could still beneficiate from additional training data, but is slowly converging.

7.2.3.4 Discussion

This training size experiment is an interesting step for any CNN study, as it helps to evaluate whether the model and the dataset are a good match [122]. It also reflects the task complexity, and can reveal whether folds are slightly or heavily unbalanced. Ideally, it should involve more data points to observe a smooth convergence in performance, which allows to fit a predictive model as in Fig. 7.15 to estimate the potential gain of adding data [138]. On our dataset, U-Net gave competitive results on cardiac ultrasound segmentation with only 50 patients, and outperformed the SRF algorithm from 300 patients on on all metrics (200 for the MAD). We therefore decided to continue investigating this method.

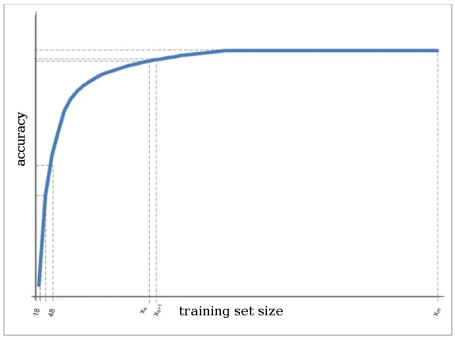


FIGURE 7.15: Learning curve model [138]: $y = (1 - a) - b \times x^c$ with $0 < a \ll 1$ and $c \in [-1, 0]$.

7.3 Clinical potential in 2D echocardiography

This work has been published in the IEEE TMI journal [5].

7.3.1 Motivations

As underlined from the previous experiments, U-Net is a promising automatic solution for 2D echocardiographic multi-class segmentation. To better evaluate its clinical potential, we decided to investigate:

1. the geometrical performance on the full CAMUS dataset;
2. the corresponding accuracy on clinical indices;
3. the comparison of U-net with another state-of-the-art non-machine learning method and with experts O_2 and O_3

7.3.2 Evaluation

7.3.2.1 Dataset

We indicated in Chapter 5 that 19% of the patients of the CAMUS dataset present at least one poor quality sequence, based on the opinion of one expert O_1. For these patients, the localization of the endocardium and epicardium, as well as the estimation of the derived clinical indices, are not considered reliable. To evaluate the clinical potential of U-Net, these cases were therefore excluded during the computation of the different metrics. We however studied their influence as part of the train and validation sets.

As we observed that results were skewed by particular combinations of train and test sets, we decided to perform cross-validation on ten folds of CAMUS, that we balanced in terms of image quality and ejection fraction value (see Chapter 5 for more details).

7.3.2.2 Metrics

Geometric metrics We assessed the performance of the algorithms on the geometric metrics described in Section 4.1, for the segmentation of the endocardium and the epicardium. In addition to the average and standard deviations values, we evaluate the robustness by assessing the number of cases where the predicted contours were far from the ground truth annotations. The criteria we designed are the following (Tab. 7.5): geometrical outliers are characterized at each instant by MAD or HD values outside the higher average inter-observer variability observed on the LV_{endo} and LV_{epi}. This criteria is voluntarily stricter than the usual $mean + 3 \times std$ [93], as our goal is to bring each case under the inter- expert variability.

Table 7.5: Geometrical outliers criteria

Instant	MAD (mm)	HD (mm)
ED	> 3.5	> 8.2
ES	> 4.0	> 8.8

Clinical metrics To evaluate the clinical potential, we evaluated the accuracy of the estimations of three indices: the left ventricle volumes at ED and ES (LV_{EDV} and LV_{ESV}) and the ejection fraction LV_{EF}, the computation of which is detailed in Section 3.4. We computed four metrics: the correlation (*corr*), the bias and the standard deviation (*std*) values, as well as the mean absolute error (*mae*). The *mae* is used to define the best solution rather than the bias since a lower bias alone does not necessarily mean a better performing method. We compared the algorithm scores to the corresponding expert inter- and intra-variability.

7.3.2.3 Statistical analysis

To comment on the statistical differences between the results of the algorithms, we ran paired Wilcoxon signed rank tests [141]. The Wilcoxon signed rank test consists in comparing two distributions D and E to assess whether the differences between them are statistically significant. It relies on the theory that under a sufficient number of observations, differences between the samples of a given distribution are equally distributed around zero (H0).

Contrary to the Student T-test, the data is not assumed to be normally distributed. We use the Wilcoxon signed rank test rather than the MannWhitney rank-sum test [142] because the samples are not independent. Indeed, we voluntarily pair the scores obtained on each image by the two algorithms. The procedure is the following: the N pairs are ranked (sorted) according to their non-null absolute difference $|d_i - e_i| > 0 \rightarrow r_i \in [1, N_d]$, with N_d the number of non-null differences. Then the sum W of the signed ranks is computed:

$$W = \sum_{i=1}^{N_d} r_i \times sgn(d_i - e_i) \tag{7.11}$$

Its distribution has an expected value of 0 and a variance of $\frac{N_d \times (N_d+1) \times (2 \times N_d+1)}{6}$ if H0 holds true. The size is therefore a factor for the establishment of critical values W_{N_d} . If $W > W_{N_d}$, the differences between the two samples are deemed statistically significant and H0 is rejected because the difference between the median of the two distributions, with regard to the sample size, is too high. If $W < W_{N_d}$, there is no ground to consider that the two medians are significantly different, and the two distributions may amount to the same [143].

In practice, we use the p-value of the test, which represents the probability that ($W < W_{N_d}$ when $H0$ is true). Small p-values therefore lead to a rejection of the null hypothesis. P-values are usually compared to an arbitrarily small threshold because p-values tend to get smaller when the number of samples is high, implying the test gets stricter when considering large amount of data points. Common threshold values are 0.05, 0.01, 0.005, or 0.001 [144]. We used the R language in this study to perform the statistical tests.

7.3.3 U-Net 1 and U-Net 2

7.3.3.1 Architectures

There is a wide range of possible U-Net designs, so we decided to compare the performance of two implementations, *i)* U-Net 1 optimized for speed adapted from (Smistad et al., 2017) [19] and *ii)* U-Net 2, inspired by our own study (Leclerc et al., 2018) [31] and optimized for accuracy. The two networks were trained to simultaneously segment the LV, the myocardium, and the left atrium. The comparison between the two architectures enabled us to investigate the impact of hyper-parameter and architecture choices on the quality of the results.
Tab. 7.6 summarize the differences between the two models, described in more details in Appendix E. For each implementation, we successively indicate the number of output feature

TABLE 7.6: Main characteristics of U-Net 1 and U-Net 2

U-Net	#Feat maps	Lowest Res	Up-sampling	Norma-lization	Batch Size	λ	Loss	# Train Prms
1	$32 \downarrow 128 \uparrow 16$	8×8	2×2 NN	-	32	10^{-3}	Dice	2M
2	$32 \downarrow 512 \uparrow 32$	16×16	Deconv	Batch Norm	10	10^{-4}	CCE + L2(W)	18M

NN: nearest neighbor
Res: resolution

maps for the first, the bottom (where the spatial information is the most compressed), and the last convolution layers.

U-Net 1 is deeper with less filters. In the U-Net 2 design, the number of filters per convolutional layer increases and decreases linearly from an initial kernel size of 48, which makes for a wider net. Spatial dropout was entirely removed, as it tended to lower the performance. Batch normalization (BN) was added before each activation, after we observed this technique significantly boosted the performance. Batch normalization is used to shift the input distribution so that each layer is less dependent of the previous ones. First, the inputs x_i are normalized to follow a standard normal distribution on each dimension k:

$$\hat{x}_i^{(k)} = \frac{x_i^{(k)} - \bar{\mu}^{(k)}}{\sigma^{(k)}} \tag{7.12}$$

Then, an affine transformation is performed, whose parameters are learned:

$$y_i^{(k)} = \gamma^{(k)} \times \hat{x}_i^{(k)} + \beta^{(k)} \tag{7.13}$$

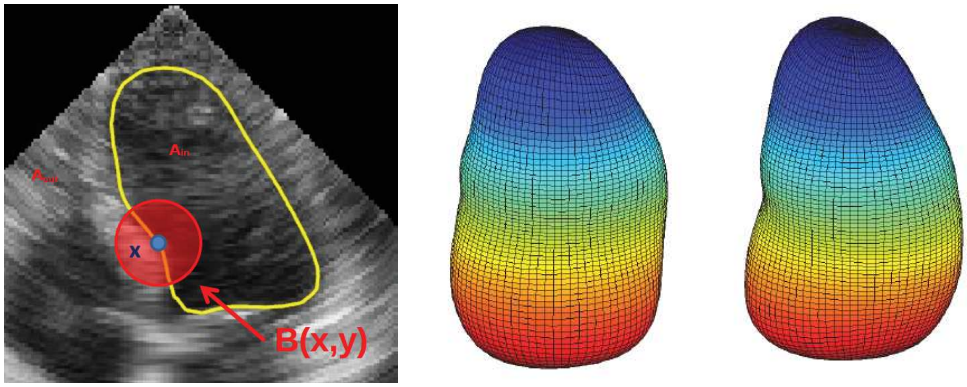
Batch normalization was originally proposed as a way to reduce the internal covariate shift in deep networks [145]. However, it was recently suggested that it instead smooths the overall objective function [146], improving the robustness to high learning rates and to initialization, which we confirmed in practice.

7.3.3.2 Implementation details

For a fair comparison, the same data pre- and post-processing were applied. In particular:

- images were resized to 256x256 pixels before performing density normalization;
- no data augmentation strategy was involved;
- principal simply connected objects (blob) selection was performed to remove small false positive detections.

For both U-Nets, the Adam optimizer was used, with a Glorot initialization. Convolutions were zero-padded to preserve the size of the feature maps. When learning on the full dataset, the number of epochs was set to 30 (U-Net 1) and 50 (U-Net 2) to allow the convergence of the validation loss. The model that performed best on the validation set was saved.



(a) BEAS: Contours are expressed as explicit functions of coordinates [58]

(b) Shape model added in the BEASM [38]: mean shape

FIGURE 7.16: Key components of the BEASM: an explicit formulation of contours incorporating shape constraints from an ASM.

7.3.4 B-spline explicit active surface model (BEASM)

7.3.4.1 Introduction

We compare U-Net to the winner of the CETUS challenge, the B-Spline Explicit Active Surface (BEAS) [37], which is not a machine learning method. The version of the BEAS we compare to is the enhanced BEASM [38], which were adapted to work in 2D.

7.3.4.2 Algorithm

The key concept of the BEAS method is to consider the boundary of a deformable interface as an explicit function (Fig. 7.16 a.), where one of the coordinates of points within the surface is expressed explicitly as a function ψ of the remaining coordinates [58]. ψ is defined as a linear combination of B-spline functions whose controlled knots are located on a regular rectangular grid defined on the polar space:

$$r = \psi(\theta) = \sum_{\theta \in [0,2\pi]} c[k] \times \beta_h(\theta - \theta_k) \tag{7.14}$$

where c[k] are the coefficients of the uniform spline modelised by β_h.

The evolution of the deformable surface is governed by the minimization of an energy function E_i that locally maximizes the differences between the mean intensity values on each side of the closed contour γ, with regard to the B-spline coefficients [37]:

$$\frac{\partial E_I}{\partial c[k]} = \int_{x \in \gamma} \left(\frac{\bar{I}(x) - I_{in}}{A_{in}} + \frac{\bar{I}(x) - I_{out}}{A_{out}} \right) \times \beta_h(\theta - k) dx \tag{7.15}$$

where A_{in} and A_{out} are the areas inside and outside the LV for a small circular neighborhood of x, as illustrated on Fig. 7.16.

The coefficients are then updated iteratively:

$$c[k]^{t+1} = c[k]^t - \lambda \times \frac{\partial E_I}{\partial c[k]^t} \tag{7.16}$$

The framework was extended to the coupled segmentation of the endocardium and epicardium in (Pedrosa et al., 2016) [147], and further improved with the integration of a statistical shape model built from 289 ciné-MRI volumes [38], namely the BEASM method. Similar to the shape models in Chapter 6, regularized shape coefficients encode plausible variations of the mean shape of the training dataset:

$$c_s[k] = \overline{c[k]} + P_s \times b_s \tag{7.17}$$

where the weights in b_s are limited to $\pm 2.5 \times \sigma(c[k])$.

An energy term is then derived to encourage B-spline coefficients to be close to the closest regularized ones:

$$E_S = \int_{x \in \gamma} \frac{1}{2} \times (c[k] - c_s[k])^2 \, dx \tag{7.18}$$

whose derivative with regard to the B-spline coefficients is:

$$\frac{\partial E_S}{\partial c[k]} = \int_{x \in \gamma} (c[k] - c_s[k]) \, dx \tag{7.19}$$

Both the local texture (represented by E_I) and the shape (represented by E_S) are now taken into account in the update:

$$c[k]^{t+1} = c[k]^t - \lambda \times \frac{\partial(\alpha E_I + \beta E_S)}{\partial c[k]^t} \tag{7.20}$$

where α and β are hyper-parameters balancing the influence of the two terms.

7.3.4.3 Initialization

The BEASM method amounts to a deformable model, so the initialization of the contour plays a crucial role on the quality of the results. We tried two different strategies:

1. BEASM-f, where the evolving contour is automatically initialized from the method proposed in (Barbosa et al., 2013) [37], where an ellipse is fitted using edge detection, here to detect the base and the septum;

2. BEASM-s, where the evolving contour of the LV is initialized from three points (two at the base and one at the apex of the endocardium) extracted from the reference contours.

7.3.5 Inter- and intra- variability between experts regarding image segmentation in 2D echocardiography

The inter and intra-observer measurements were computed from fold 5 and restricted to patients having good and medium image quality (40 patients).

7.3.5.1 Inter-variability

Rather than the average inter-variability between our three experts, we provide the comparison between each pair of cardiologists in annotating the 50 patients (200 images) of fold 5. We highlight in cyan the best scores for each metric at the top of Tab. 7.7. One can see there is a strong agreement between O_{1a} and O_3 on the epicardium, but they tend to disagree on

the placement of the endocardium. When we compare O_2 and O_3 to O_1 on clinical indices (Tab. 7.11), we can see that the *maes* of the volumes are the same as the biases, hinting that there exists a consistent bias in the volume estimations between these cardiologists. The scores on the LV_{endo} are similar when comparing O_1a and O_2, and O_3 and O_2, so O_2's contours may correspond to the strongest agreement between the three cardiologists of the study. The values used for the outlier criteria are highlighted in red.

7.3.5.2 Intra-variability

The intra-variability, represented by O_{1b}, is significantly lower than the inter-variability. The values suggest that:

- O_1 has a specific way of contouring, reproduced independently of the time instant;
- the epicardium contours show a higher variability than for the endocardium;
- the difference between O_{1a} and O_{1b} is on average below 2mm for both structures, and locally increases up to 5 mm.

7.3.5.3 Comparison to the literature

Compared to the values reported on other studies, the inter-variability we report appears to stand in-between, though a direct comparison is impossible due to difference of protocols, metrics, number of patients, pathologies, material and image properties.

In (Bosch et al., 2012) [93], the average distance between landmarks over the full cardiac cycle of 19 patients (supposedly higher than the MAD) for two independent expert annotations of the LV was of 3.82 ± 1.44 mm. The intra-variability estimated with a six month interval between the two contours was of 2.32 ± 0.75 mm. In (Azarmehr et al., 2019) [20], the inter-expert variability measured for the segmentation of the LV on 992 images (from 61 patients) annotated by two experts was $D = 0.88 \pm 0.06$ mm and $HD = 4.50 \pm 0.87$ mm.

7.3.6 Geometrical results

We present the geometrical scores achieved by all methods in function of the image quality. For a direct comparison to the inter- and intra- variability, we also provide the scores restricted to the fold 5.

7.3.6.1 On good and medium quality images

Tab. 7.7 lists the geometrical scores of all methods, with the best values in bold. One can see that the U-Nets got overall the best segmentation scores on all metrics, for both ED and ES. Interestingly, while the U-Net methods are fully-automatic, they still produced better segmentation results than the semi-automatic BEASM. BEASM-f obtained on average better HD scores, while the SRF got better Dice and MAD scores. However, the large standard deviation values for the SRF illustrate the difficulties of this method in obtaining consistent segmentations over the entire dataset. As for the BEASM-s, one can see that the manual initialization had a strong impact on the quality of the results, with an improvement of the mean value of 0.8 mm and 2.4 mm for the MAD and HD metrics respectively.

U-Net 1 and 2 achieved equivalent results for all metrics, with average results in between the inter-observer and intra-observer scores for all metrics. This shows the robustness of this

TABLE 7.7: Segmentation accuracy of the 5 evaluated methods on the ten test folds, restricted to patients having good & medium image quality (406 patients in total).

Method	ED						ES					
	LV_{endo}			LV_{epi}			LV_{endo}			LV_{epi}		
	D	MAD	HD	D	MAD	HD	D	MAD	HD	D	MAD	HD
	val.	mm	mm	val.	mm	mm	val.	mm	mm	val.	mm	mm
O_{1a} vs O_2 (inter-obs)	0.919 ±0.033	2.2 ±0.9	6.0 ±2.0	0.913 ±0.037	3.5 ±1.7	8.0 ±2.9	0.873 ±0.060	2.7 ±1.2	6.6 ±2.4	0.890 ±0.047	3.9 ±1.8	8.6 ±3.3
O_{1a} vs O_3 (inter-obs)	0.886 ±0.050	3.3 ±1.5	8.2 ±2.5	0.943 ±0.018	2.3 ±0.8	6.5 ±2.6	0.823 ±0.091	4.0 ±2.0	8.8 ±3.5	0.931 ±0.025	2.4 ±1.0	6.4 ±2.4
O_2 vs O_3 (inter-obs)	0.921 ±0.037	2.3 ±1.2	6.3 ±2.5	0.922 ±0.036	3.0 ±1.5	7.4 ±3.0	0.888 ±0.058	2.6 ±1.3	6.9 ±2.9	0.885 ±0.054	3.9 ±1.9	8.4 ±2.8
O_{1a} vs O_{1b} (intra-obs)	0.945 ±0.019	1.4 ±0.5	4.6 ±1.8	0.957 ±0.019	1.7 ±0.9	5.0 ±2.3	0.930 ±0.031	1.3 ±0.5	4.5 ±1.8	0.951 ±0.021	1.7 ±0.8	5.0 ±2.1
SRF	0.895 ±0.074	2.8 ±3.6	11.2 ±10.2	0.914 ±0.057	3.2 ±2.0	13.0 ±9.1	0.848 ±0.137	3.6 ±7.8	11.6 ±13.6	0.901 ±0.078	3.5 ±4.7	13.0 ±11.1
BEASM-f	0.879 ±0.065	3.3 ±1.8	9.2 ±4.9	0.895 ±0.051	3.9 ±2.1	10.6 ±5.1	0.826 ±0.092	3.8 ±2.1	9.9 ±5.1	0.880 ±0.054	4.2 ±2.0	11.2 ±5.1
BEASM-s	0.920 ±0.039	2.2 ±1.2	6.0 ±2.4	0.917 ±0.038	3.2 ±1.6	8.2 ±3.0	0.861 ±0.070	3.1 ±1.6	7.7 ±3.2	0.900 ±0.042	3.5 ±1.7	9.2 ±3.4
U-Net 1	0.934 ±0.042	1.7 ±1.0	5.5 ±2.9	0.951 ±0.024	1.9 ±0.9	5.9 ±3.4	0.905 ±0.063	1.8 ±1.3	5.7 ±3.7	0.943 ±0.035	2.0 ±1.2	6.1 ±4.1
U-Net 2	**0.939** ±0.043	**1.6** ±1.3	**5.3** ±3.6	**0.954** ±0.023	**1.7** ±0.9	6.0 ±3.4	**0.916** ±0.061	**1.6** ±1.6	**5.5** ±3.8	**0.945** ±0.039	**1.9** ±1.2	6.1 ±4.6

LV_{endo}: Endocardial contour of the left ventricle;
LV_{epi}: Epicardial contour of the left ventricle;
ED: End diastole; ES: End systole;
D: Dice index; MAD: mean absolute distance; HD: Hausdorff distance

architecture in obtaining accurate segmentation results. According to the Wilcoxon signed rank test, U-Net 1 and 2 produced scores that are statistically different (p-value < 0.05) for most metrics at ED and ES, apart for the LV_{epi} HD. However, the overlapping distributions of the results of the two networks show multiple similarities as seen in Fig. 7.17. For instance, both distributions have chi-square aspects, with very good mean performances and overall good robustness in spite of the production of a few outliers.

It is well known that the left ventricle shape is more difficult to segment at ES, leading to slightly worse performance for classical algorithms on this time instant. This property is also confirmed in this study since all the evaluated methods produced better results at ED on every metric. As complement, we provide in Tab. 7.8 the outliers rates produced by each method (using cardiologist $O1_a$ as reference) on the full dataset and by the experts on fold 5 (respectively on 406 and 40 patients).
From this table, we can observe that the U-Nets are the only methods that produced less

Table 7.8: Outliers rates

Method	SRF	BEASM-f	BEASM-s	U-Net 1	U-Net 2	O_2	O_3	$O_2 \vert O_3$	O_{1b}
# Outliers	69%	79%	59%	18%	18%	56%	58%	46%	13%

outliers than the rates observed between experts, but still more than observed in the intra-variability.

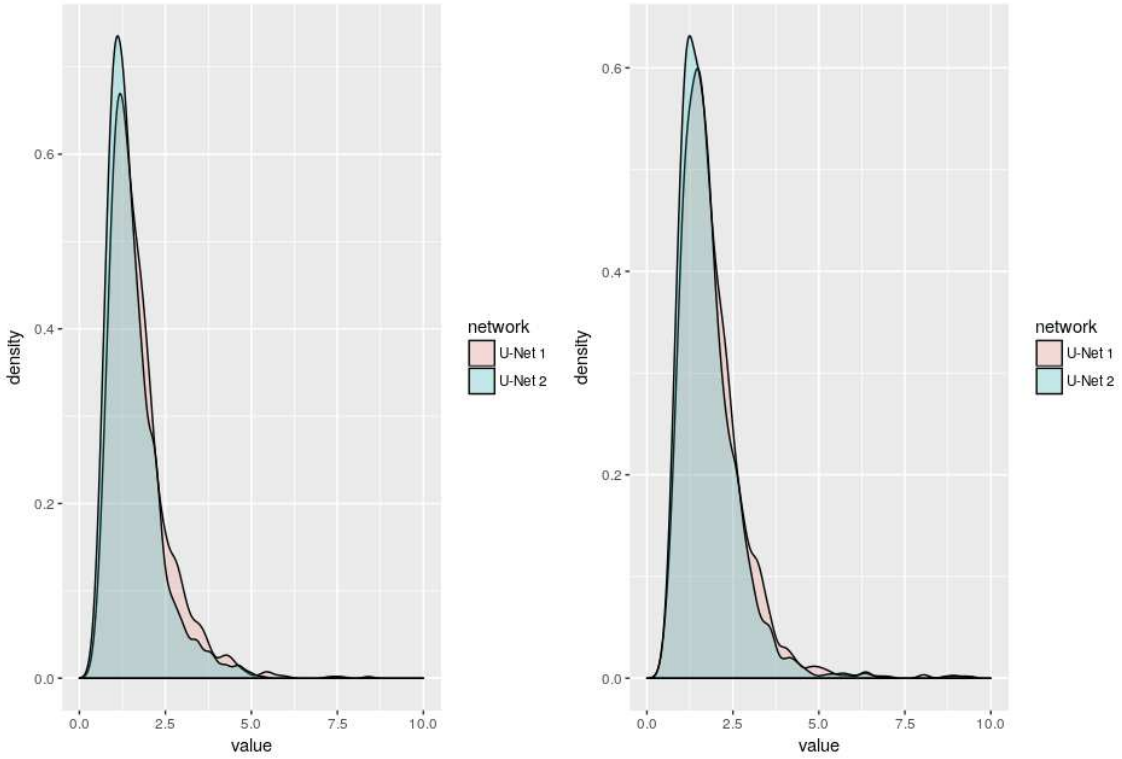
7.3.6.2 On fold 5

For a better comparison to the inter- and intra- observer variability, Tab. 7.9 summarizes the results obtained for all the methods on good and medium quality images of the fold 5. On this particular fold, all methods provided better scores than on the whole dataset, hinting it is overall composed of rather easy cases. The U-Nets remained the best performing methods, and U-Net 2 reached the intra-variability on the MAD LV_{endo} metric for both instants. The superiority of U-Net 2 over U-Net 1 is not regular, and as 10 out of 12 p-values are above the threshold, the differences between the results of the two models are here deemed not statistically significant, possibly because of a lower number of samples.

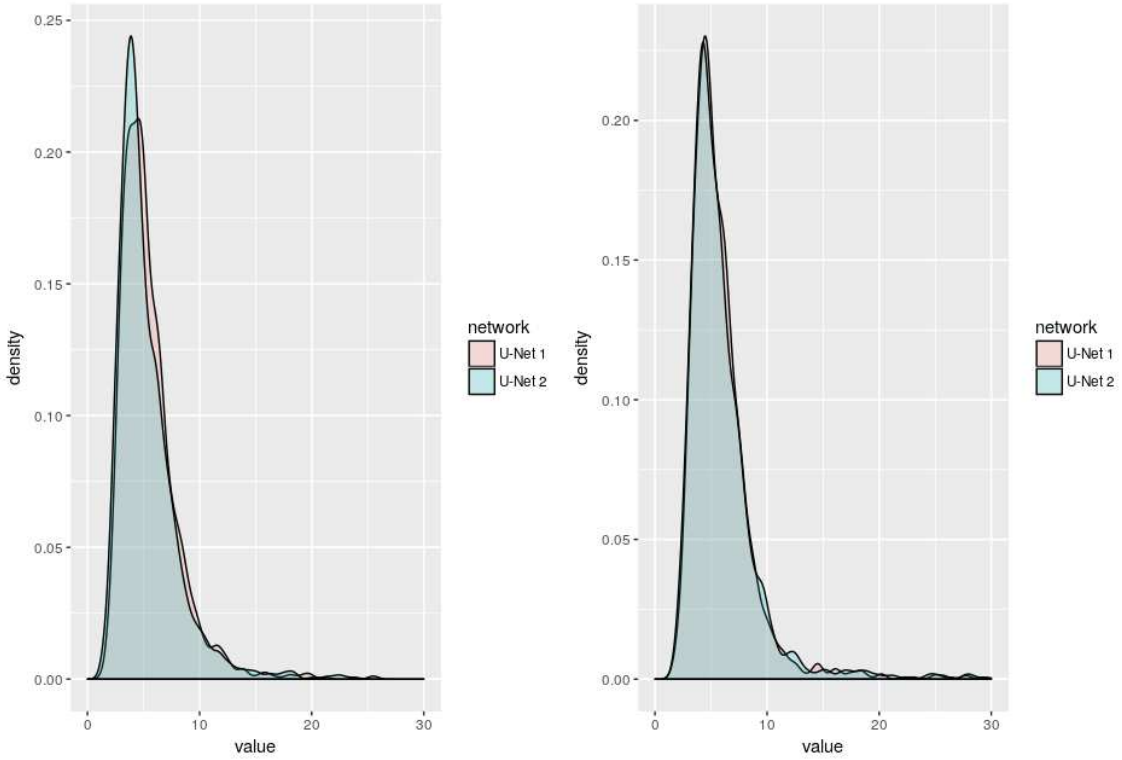
7.3.6.3 On poor quality images

To evaluate how methods coped with image quality, we report in Tab. 7.10 the results obtained on the poor quality images. As expected, all methods showed a drop in performance. However, possibly because poor quality images were added to the training set, the drop was limited. The U-Nets remained the best performing methods with MAD and Dice values below the inter-variability. The p-values indicate that the differences were not statistically significant, possibly because of a higher dispersion of the results. The HD values of the U-Nets were significantly worse than on good and medium quality images, and the semi-automatic BEASM obtained a better score with regard to the HD at ED for the LV_{endo}.

On these poor quality images, the ground truth is considered unreliable, so we cannot conclude that image quality is not a performance factor. However, our results suggest that a U-Net trained on various image qualities can predict segmentations close to what an expert would do, overcoming the inherent difficulties of low quality images (signal dropout, poor contrast, artifacts...). The fact that results are not so dependent on image quality suggests that there are other predominant factors of difficulties.



(a) MAD LV_{endo} / LV_{epi}



(b) HD LV_{endo}/ LV_{epi}

FIGURE 7.17: Overlapping distributions of prediction results from U-Net 1 and U-Net 2.

TABLE 7.9: Segmentation accuracy of the 5 evaluated methods on fold 5 restricted to good & medium image quality (40 patients).

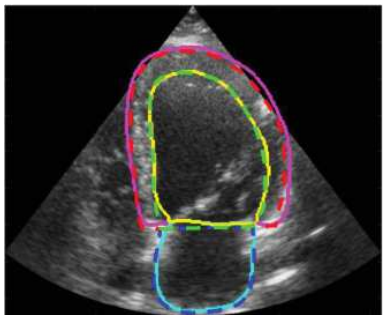
Method	ED						ES					
	LV_{endo}			LV_{epi}			LV_{endo}			LV_{epi}		
	D	MAD	HD	D	MAD	HD	D	MAD	HD	D	MAD	HD
	val.	mm	mm	val.	mm	mm	val.	mm	mm	val.	mm	mm
O_{1a} vs O_2 (inter-obs)	0.919 ±0.033	2.2 ±0.9	6.0 ±2.0	0.913 ±0.037	3.5 ±1.7	8.0 ±2.9	0.873 ±0.060	2.7 ±1.2	6.6 ±2.4	0.890 ±0.047	3.9 ±1.8	8.6 ±3.3
O_{1a} vs O_3 (inter-obs)	0.886 ±0.050	3.3 ±1.5	8.2 ±2.5	0.943 ±0.018	2.3 ±0.8	6.5 ±2.6	0.823 ±0.091	4.0 ±2.0	8.8 ±3.5	0.931 ±0.025	2.4 ±1.0	6.4 ±2.4
O_2 vs O_3 (inter-obs)	0.921 ±0.037	2.3 ±1.2	6.3 ±2.5	0.922 ±0.036	3.0 ±1.5	7.4 ±3.0	0.888 ±0.058	2.6 ±1.3	6.9 ±2.9	0.885 ±0.054	3.9 ±1.9	8.4 ±2.8
O_{1a} vs O_{1b} (intra-obs)	0.945 ±0.019	1.4 ±0.5	4.6 ±1.8	0.957 ±0.019	1.7 ±0.9	5.0 ±2.3	0.930 ±0.031	1.3 ±0.5	4.5 ±1.8	0.951 ±0.021	1.7 ±0.8	5.0 ±2.1
SRF	0.905 ±0.053	2.4 ±1.3	9.5 ±6.3	0.920 ±0.053	3.0 ±1.8	10.8 ±7.0	0.854 ±0.099	3.0 ±1.8	10.7 ±8.6	0.900 ±0.062	3.3 ±1.8	13.4 ±9.4
BEASM-f	0.882 ±0.065	3.2 ±2.1	8.5 ±5.0	0.897 ±0.052	3.9 ±2.3	9.9 ±5.0	0.832 ±0.089	3.6 ±2.2	9.3 ±5.4	0.883 ±0.051	4.1 ±2.2	10.4 ±4.9
BEASM-s	0.922 ±0.031	2.1 ±0.8	5.6 ±2.0	0.916 ±0.040	3.2 ±1.6	7.8 ±2.8	0.868 ±0.057	2.8 ±1.2	7.1 ±3.1	0.904 ±0.036	3.3 ±1.4	8.7 ±3.0
U-Net 1	0.941 ±0.021	1.5 ±0.5	5.0 ±1.4	**0.954** ±0.018	**1.7** ±0.7	**5.2** ±1.9	0.917 ±0.037	1.6 ±0.6	5.0 ±2.1	**0.947** ±0.020	**1.8** ±0.7	**5.4** ±2.1
U-Net 2	**0.945** ±0.021	**1.4** ±0.5	**4.9** ±1.9	**0.954** ±0.017	1.8 ±0.7	5.7 ±3.1	**0.927** ±0.039	**1.3** ±0.6	**4.9** ±2.5	0.946 ±0.023	1.9 ±0.8	6.0 ±3.4

TABLE 7.10: Segmentation accuracy of the 5 evaluated methods on the ten test datasets restricted to patients having poor image quality (94 patients).

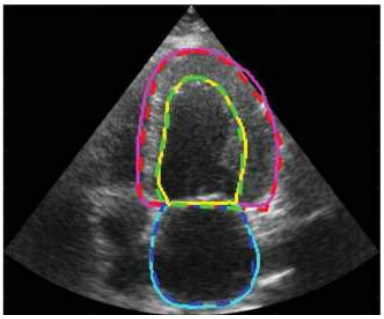
Method	ED						ES					
	LV_{endo}			LV_{epi}			LV_{endo}			LV_{epi}		
	D	MAD	HD	D	MAD	HD	D	MAD	HD	D	MAD	HD
	val.	mm	mm	val.	mm	mm	val.	mm	mm	val.	mm	mm
SRF	0.869 ±0.062	3.6 ±2.0	14.3 ±9.4	0.891 ±0.063	4.2 ±2.4	15.9 ±8.5	0.801 ±0.123	4.6 ±3.5	17.0 ±13.2	0.852 ±0.112	4.9 ±3.1	18.0 ±12.1
BEASM-f	0.857 ±0.083	4.1 ±2.6	10.5 ±6.3	0.888 ±0.058	4.5 ±2.6	11.9 ±6.2	0.801 ±0.102	4.7 ±2.7	12.3 ±6.6	0.873 ±0.063	4.7 ±2.6	12.4 ±6.2
BEASM-s	0.915 ±0.039	2.4 ±1.2	**6.4** ±2.7	0.914 ±0.035	3.4 ±1.5	8.5 ±2.9	0.859 ±0.063	3.3 ±1.6	8.3 ±3.6	0.900 ±0.039	3.6 ±1.6	9.5 ±3.4
U-Net 1	**0.921** ±0.037	**2.1** ±1.0	6.5 ±3.0	0.945 ±0.021	2.2 ±1.0	**6.8** ±3.0	0.893 ±0.059	2.2 ±1.2	6.8 ±4.2	0.935 ±0.031	2.4 ±1.3	**7.2** ±4.7
U-Net 2	**0.921** ±0.037	**2.1** ±1.0	6.8 ±3.3	**0.947** ±0.023	**2.1** ±1.0	7.2 ±3.8	**0.898** ±0.057	**2.1** ±1.2	**6.6** ±3.6	**0.937** ±0.032	**2.2** ±1.2	7.6 ±4.8

(a) 2CH-ED: MAD = 1.4 | 1.5
HD = 4.4 | 4.8 mm.

(b) 2CH-ES: MAD = 1.6 | 2.4
HD = 5.1 | 7.7 mm.

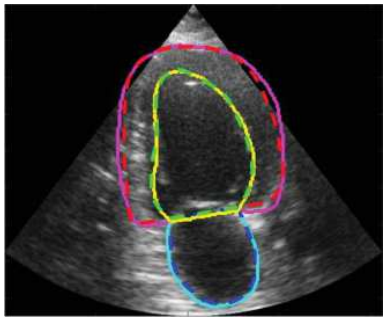


(c) 4CH-ED: MAD = 1.0 | 2.2
HD = 5.9 | 4.5 mm.



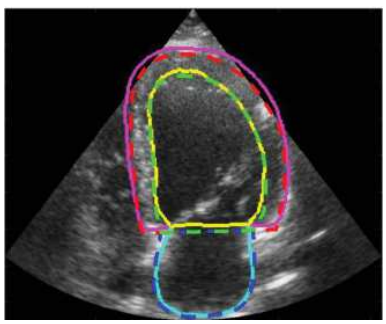
(d) 4CH-ES: MAD = 1.0 | 1.7
HD = 3.8 | 4.5 mm.

FIGURE 7.18: Segmentation results obtained by U-Net 1

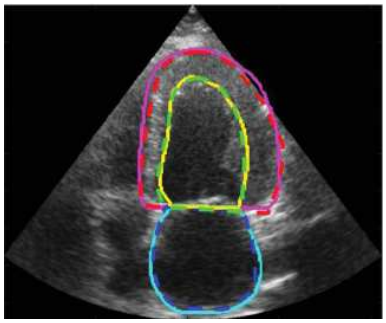


(a) 2CH-ED: MAD = 1.1 | 1.4
HD = 4.0 | 4.0 mm.

(b) 2CH-ES: MAD = 1.4 | 1.2
HD = 4.4 | 4.3 mm.

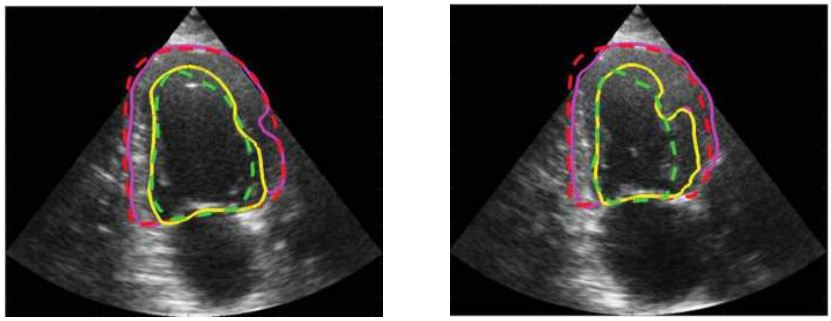


(c) 4CH-ED: MAD = 1.6 | 2.0
HD = 6.8 | 4.6 mm.



(d) 4CH-ES: MAD = 1.6 | 1.5
HD = 4.7 | 4.3 mm.

FIGURE 7.19: Segmentation results obtained by U-Net 2.



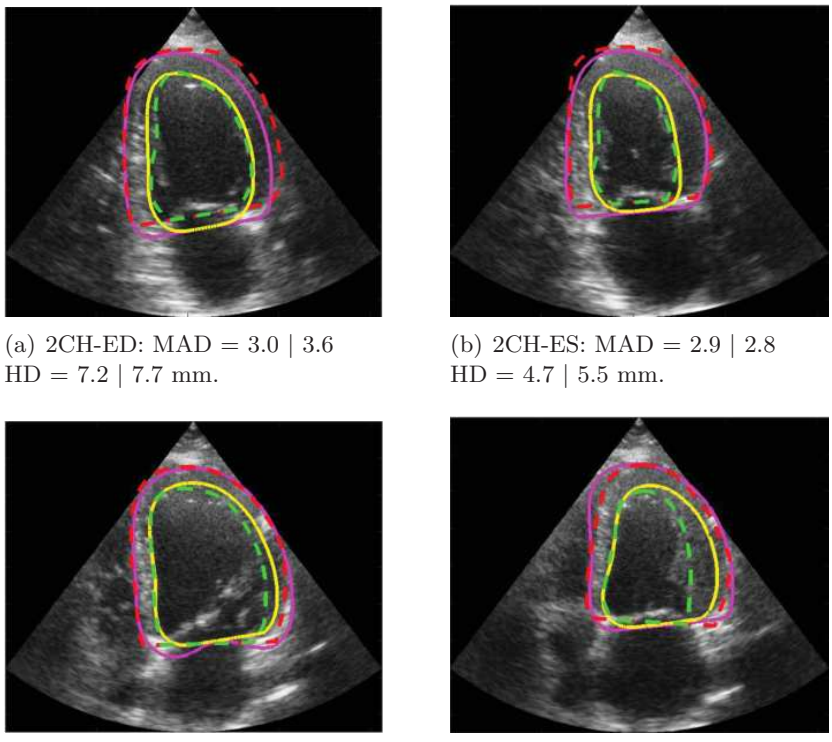
(a) 2CH-ED: MAD = 2.7 | 1.8
HD = 8.5 | 8.1 mm.

(b) 2CH-ES: MAD = 3.8 | 2.6
HD = 14.0 | 7.0 mm.

(c) 4CH-ED: MAD = 2.3 | 2.3
HD = 11.1 | 14.2 mm.

(d) 4CH-ES: MAD = 3.2 | 1.7
HD = 6.8 | 4.8 mm.

FIGURE 7.20: Segmentation results obtained by SRF.



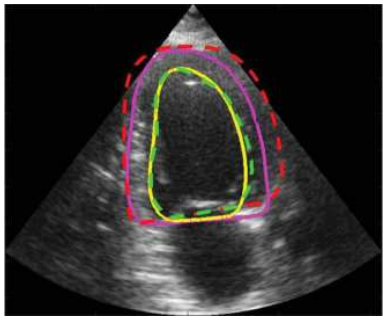
(a) 2CH-ED: MAD = 3.0 | 3.6
HD = 7.2 | 7.7 mm.

(b) 2CH-ES: MAD = 2.9 | 2.8
HD = 4.7 | 5.5 mm.

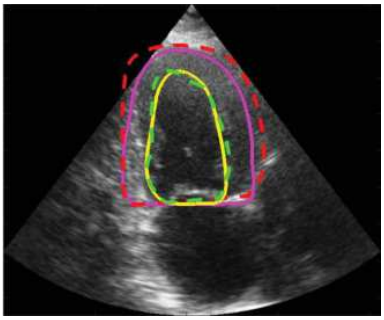
(c) 4CH-ED: MAD = 2.4 | 1.7
HD = 6.2 | 6.6 mm.

(d) 4CH-ES: MAD = 5.0 | 2.3
HD = 14.2 | 6.5 mm.

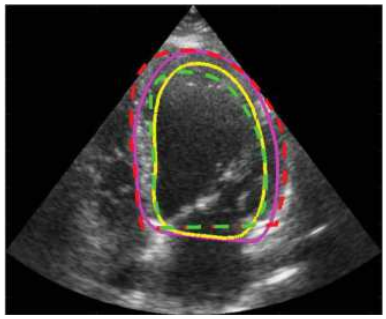
FIGURE 7.21: Segmentation results obtained by the BEASM-f.



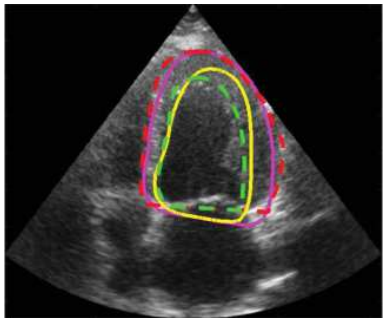
(a) 2CH-ED: MAD = 1.7 | 4.9
HD = 4.1 | 9.6 mm.



(b) 2CH-ES: MAD = 1.5 | 3.8
HD = 8.5 | 8.5 mm.



(c) 4CH-ED: MAD = 2.1 | 3.2
HD = 6.2 | 8.2 mm.

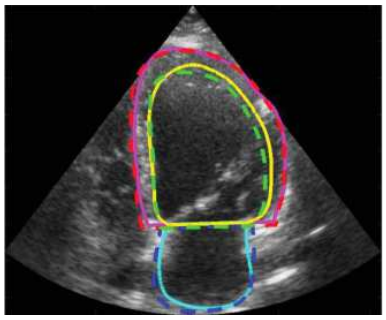


(d) 4CH-ES: MAD = 3.4 | 2.8
HD = 3.5 | 7.7 mm.

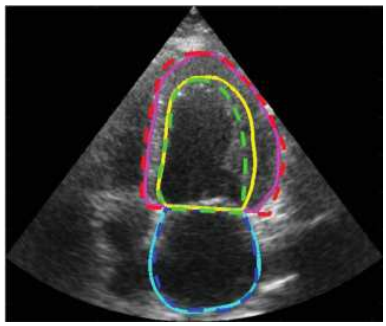
FIGURE 7.22: Segmentation results obtained by the BEASM-s.

(a) 2CH-ED: MAD = 3.1 | 4.2
HD = 5.9 | 9.0 mm.

(b) 2CH-ES: MAD = 3.3 | 5.5
HD = 8.1 | 9.5 mm.

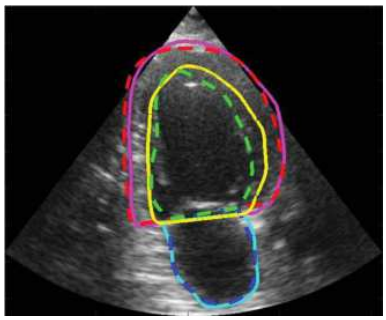


(c) 4CH-ED: MAD = 2.0 | 1.4
HD = 4.8 | 3.4 mm.



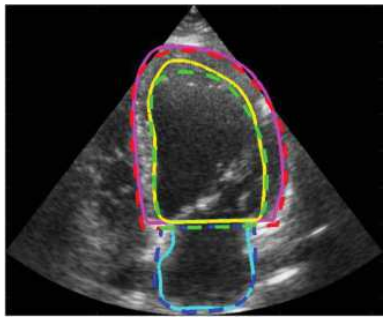
(d) 4CH-ES: MAD = 2.4 | 1.9
HD = 6.1 | 9.0 mm.

FIGURE 7.23: Segmentation results obtained by O_2.



(a) 2CH-ED: MAD = 4.5 | 4.1
HD = 9.7 | 4.1 mm.

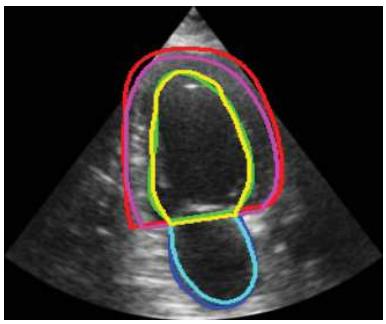
(b) 2CH-ES: MAD = 5.9 | 4.5
HD = 13.2 | 4.5 mm.



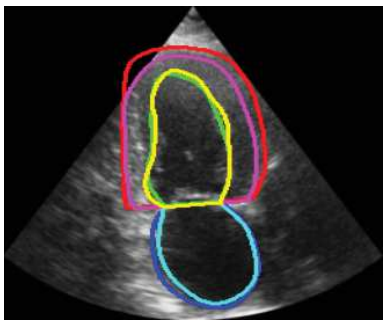
(c) 4CH-ED: MAD = 1.7 | 4.0
HD = 6.8 | 4.0 mm.

(d) 4CH-ES: MAD = 2.7 | 4.4
HD = 4.9 | 4.4 mm.

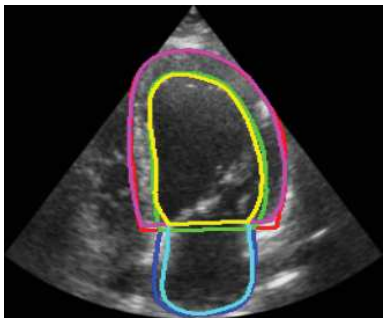
FIGURE 7.24: Segmentation results obtained by O_3.



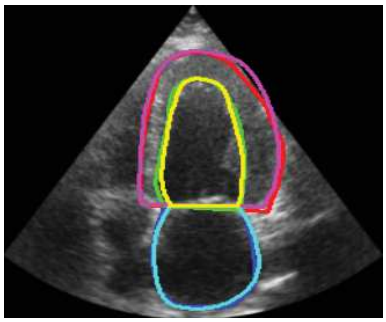
(a) 2CH-ED: MAD = 1.3 | 3.0
HD = 3.0 | 6.7 mm.



(b) 2CH-ES: MAD = 1.2 | 3.4
HD = 4.3 | 8.0 mm.



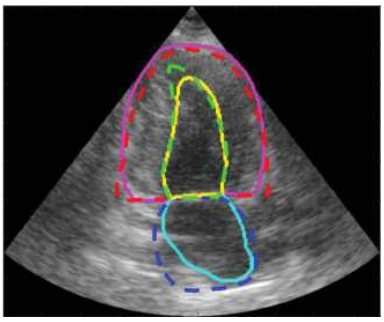
(c) 4CH-ED: MAD = 1.9 | 1.1
HD = 5.7 | 4.7 mm.



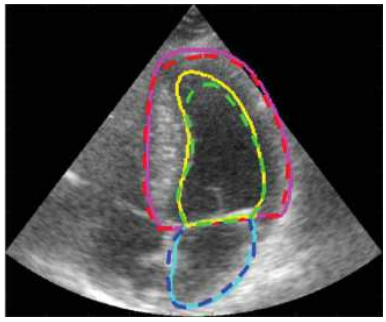
(d) 4CH-ES: MAD = 1.3 | 1.5
HD = 3.4 |5.1 mm.

FIGURE 7.25: Segmentation results obtained by O_1b.

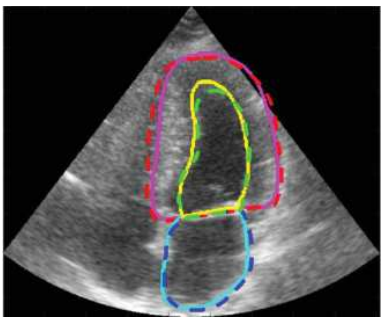
(a) 2CH-ED: MAD = 1.0 | 2.2
HD = 4.2 | 7.1 mm.



(b) 2CH-ES: MAD= 1.3 | 2.8
HD = 7.8 | 8.3 mm.

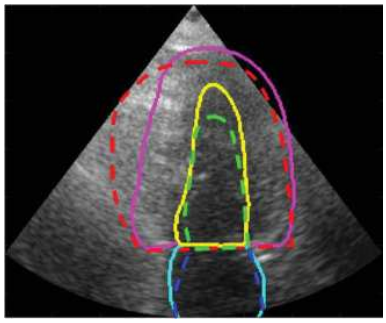


(c) 4CH-ED: MAD = 1.7 | 1.7
HD = 6.3 | 4.0 mm.

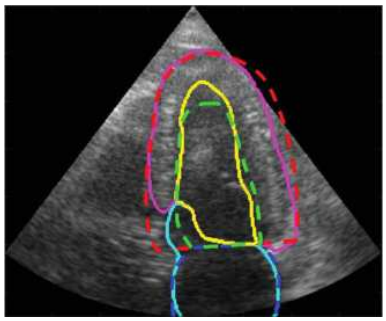


(d) 4CH-ES: MAD = 1.5 | 1.5
HD = 5.9 | 3.7 mm.

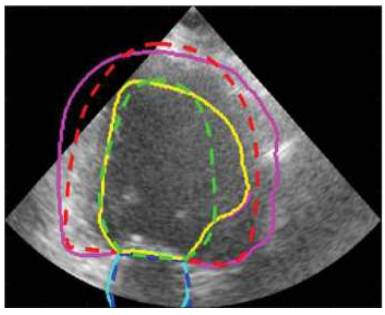
FIGURE 7.26: Segmentation results of U-Net 1 on Patient 252 (Medium IQ).



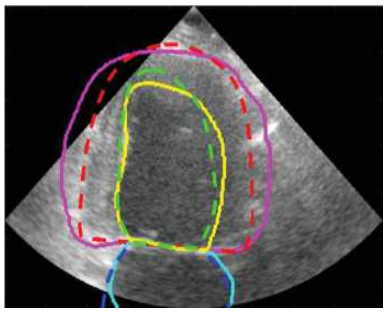
(a) 2CH-ES: MAD = 2.5 | 3.9
HD = 12.6 | 13.3 mm.



(b) 4CH-ES: MAD = 3.0 | 2.4
HD = 15.7 | 10.1 mm.



(c) 2CH-ED: MAD = 7.3 | 6.1
HD = 23.2 | 16.5 mm.



(d) 2CH-ES: MAD = 3.8 | 5.7
HD = 11.1 | 12.0 mm.

FIGURE 7.27: Anatomical outliers (up), one at ED only (down).

7.3.7 Visual analysis

To assess the behavior of the different methods, we provide visuals for all of them, including a direct comparison on a good quality case and of the U-Net on a medium quality case. The evaluation is completed by error analysis.

7.3.7.1 Comparison of all methods and experts on a given case

To allow visual assessment of the segmentation performance of the different methods implemented, we show in Fig. 7.18 to 7.24 the segmentation results obtained by each of the presented methods and the cardiologists on a given patient with a good image quality. Ground-truth contours are dotted and prediction contours are drawn in full line.

7.3.7.2 Medium image quality

Fig. 7.26 corresponds to the segmentation results obtained from U-Net 1 on a patient with medium quality. This case was been chosen because it represents on at least one image the median scores obtained by the U-Net 1 algorithm on the full dataset for the endocardium and epicardium: MAD $LV_{endo} = 1.6$ mm, MAD $LV_{epi} = 1.7$ mm.

7.3.7.3 Unsolved cases and limitations

Based on the outlier criteria of Tab. 7.11, we measured that 18% of the predictions of U-Net 1 on good and medium quality images were outliers. Among them, most (about 90%) could be associated to anatomically plausible LV and myocardium shapes, but for the last 10% (1.8% of the full dataset), the provided segmentation can not be assimilated to a heart shape (Figure 7.27 b.). This observation hints at the necessity to encourage shape plausibility as a quality criteria for echocardiographic segmentation.

Performing error analysis on outlier images reveal that the network is misled by peculiar context of acquisitions such as:

- non-frequent zoom and probe tilt;
- artifacts (shadowed zones, reverberation, fuzzy textures) as can be observed for the epicardium on Fig. 7.27 c) d);
- locally/globally extremely low contrast, which is patient-dependent (endocardium in Fig. 7.27 c) d).

This suggests that artificial data augmentation simulating the various acquisition conditions in which the network might be applied may be a good lead to ensure a good robustness to context variations. Another solution could be the addition of temporal constraints, as shown in Fig. 7.27 d) where the accurate segmentation at ES could help correct the error at ED.

7.3.8 Clinical results

7.3.8.1 On good and medium quality images

Clinical results on the CAMUS dataset restricted to good and medium image qualities are reported in Tab. 7.11. Here also, the U-Nets obtained the best scores on all metrics. The LV_{EF} estimations are close to unbiased, which is very encouraging, but the high standard deviations illustrate the need for temporal coherency. Volumes and ejection fraction obtained with U-Net 2 were statistically different with p-values < 0.05 compared to all the tested methods, including U-Net 1. The average scores of both U-Nets are all below the inter-observer scores, which proves the clinical interest of such approaches. They outperform with fine margins the other methods, including the BEASM with manual initialization. However, none of the method allowed to reach the intra-variability for any clinical indice.

7.3.8.2 On fold 5

In Tab. 7.12, we show the results of the 5 methods restricted to the fold 5. On this particular fold, U-Net 2 produced average scores below the average intra-variability for the correlation, the bias, and the mean absolute error for the LV_{EF}. It should be noted that the inter and intra-variability biases we report for the estimation of the LV_{EF} are higher than the ones measured in (Bosch et al., 2012) [93], which was of 0.88 ± 3.15, as well as the intra-variability (-1.71 ± 2.84). We assume that the variabilities we computed could be greatly reduced with a consensus. Therefore, it is more accurate to say that our results show the capacity of a properly trained U-Net to closely reproduce the specific protocol represented by O_1.

7.3.8.3 Bland Altman plots

To better visualize the distributions of the estimated ejection fractions, we provide in Fig. 7.28 the Bland Altman plots of the LV_{EF} measurements between the three cardiologists and the

TABLE 7.11: Clinical metrics of the 5 evaluated methods on the ten test folds restricted to patients having good & medium image quality (406 patients).

Method	$\mathbf{LV_{EDV}}$			$\mathbf{LV_{ESV}}$			$\mathbf{LV_{EF}}$		
	corr	**bias**$\pm\sigma$	**mae**	**corr**	**bias**$\pm\sigma$	**mae**	**corr**	**bias**$\pm\sigma$	**mae**
	val.	ml	ml	val.	ml	ml	val.	%	%
O_{1a} vs O_2	0.940	18.7±12.9	18.7	0.956	18.9±9.3	18.9	0.801	-9.1±8.1	10.0
O_{1a} vs O_3	0.895	39.0±18.8	39.0	0.860	35.9±17.1	35.9	0.646	-12.6±10.0	13.4
O_2 vs O_3	0.926	-20.3±15.6	21.0	0.916	-17.0±13.5	17.7	0.569	3.5±11.0	8.5
O_{1a} vs O_{1b}	0.978	-2.8±7.1	6.2	0.981	-0.1±5.8	4.5	0.896	-2.3±5.7	4.5
SRF	0.755	-0.2±25.7	17.4	0.827	9.3±18.0	14.8	0.465	-11.5±15.4	12.8
BEASM-f	0.704	13.4±30.6	22.9	0.713	18.0±25.8	22.5	0.731	-9.8±8.3	10.7
BEASM-s	0.886	14.6±19.2	17.8	0.880	18.3±16.9	19.5	0.790	-9.4±7.2	10.0
U-Net 1	0.947	-8.3±12.6	10.9	0.955	-4.9±9.9	8.2	0.791	-0.5±7.7	5.6
U-Net 2	**0.954**	-6.9 ±11.8	**9.8**	**0.964**	-3.7± 9.0	**6.8**	**0.823**	-1.0±7.1	**5.3**

* LV_{EDV}: End diastolic left ventricular volume;
* LV_{ESV}: End systolic left ventricular volume;
corr: Pearson correlation coefficient; mae: mean absolute error.

TABLE 7.12: Clinical metrics of the 5 evaluated methods on fold 5 restricted to patients having good & medium image quality (40 patients in total)

Method	$\mathbf{LV_{EDV}}$			$\mathbf{LV_{ESV}}$			$\mathbf{LV_{EF}}$		
	corr	**bias**$\pm\sigma$	**mae**	**corr**	**bias**$\pm\sigma$	**mae**	**corr**	**bias**$\pm\sigma$	**mae**
	val.	ml	ml	val.	ml	ml	val.	%	%
O_{1a} vs O_2	0.940	18.7±12.9	18.7	0.956	18.9±9.3	18.9	0.801	-9.1±8.1	10.0
O_{1a} vs O_3	0.895	39.0±18.8	39.0	0.860	35.9±17.1	35.9	0.646	-12.6±10.0	13.4
O_2 vs O_3	0.926	-20.3±15.6	21.0	0.916	-17.0±13.5	17.7	0.569	3.5±11.0	8.5
O_{1a} vs O_{1b}	0.978	-2.8±7.1	6.2	0.981	-0.1±5.8	4.5	0.896	-2.3±5.7	4.5
SRF	0.843	5.3±18.6	13.9	0.845	12.4±15.6	15.9	0.603	-11.9±10.8	13.1
BEASM-f	0.809	19.1±23.3	21.3	0.791	23.2±20.6	24.0	0.776	-12.1±8.2	12.6
BEASM-s	0.913	12.0±15.4	15.3	0.875	16.3±15.1	18.0	0.853	-10.2±6.7	10.5
U-Net 1	0.973	-7.7±8.3	8.7	0.945	-3.1±10.1	7.6	0.820	-2.6±7.5	5.9
U-Net 2	**0.977**	-4.0 ±7.3	6.6	**0.976**	-0.9± 7.0	**5.2**	**0.928**	-2.3±4.8	**4.0**

two sets of annotations from the first cardiologist, as well as the 5 evaluated methods. The X axis EF value corresponds to the average between the prediction and the value established from the contours of O_{1a}, and the Y axis shows the difference between the ground truth and the automatic algorithm estimations. Mean differences and 95% confidence intervals are represented with dotted horizontal lines. For a good visualization and an easy comparison, results are shown for errors between -40 and 40.

7.3.9 Discussion

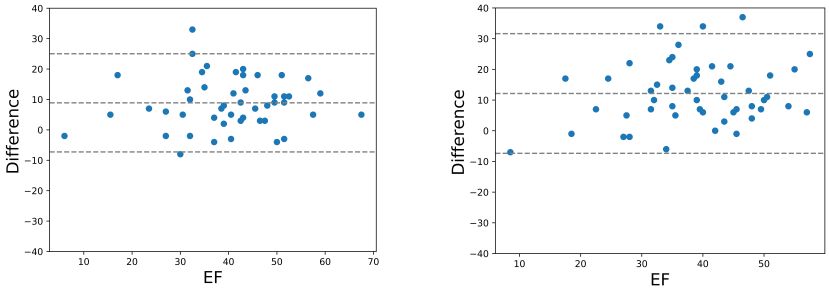
From this study, it appeared that a well-designed CNN could reach impressive segmentation scores in echocardiography. The two U-Nets clearly outperform the state-of-the-art fully and semi-automatic non-deep-learning methods. Interestingly, their geometric and clinical results are better than the inter-observer scores on all structures and metrics. They however remained lower than the intra-observer scores when considering the full dataset.

Three main limitations are revealed by our analysis:

1. 18% of the segmentations produced by both U-Nets could be seen as geometrical outliers, while the outlier rate from the intra-variability is 13%, hinting at the need for a more robust method;

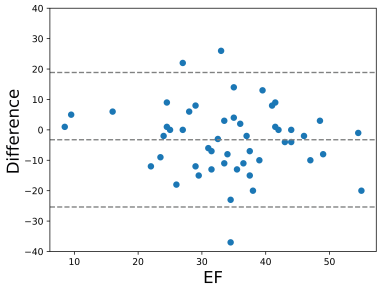
2. we visually assessed that about 2% of the predictions were obviously implausible in the anatomical sense, while an expert would never produce anatomically implausible results. Classical geometrical evaluation is therefore insufficient to judge the quality of medical segmentations;

3. Though the LV_{EF} results obtained by the U-Nets are below the average inter-observer scores with regards to the *mae* and above for the *corr*, they are still too low in comparison to the intra-observer value (*corr* of 0.82 vs 0.90). This confirms the need to further improve deep learning solutions to produce time coherent segmentation results.

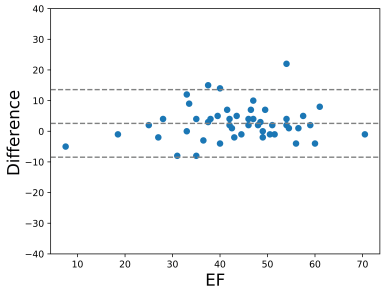


(a) O_2 vs O_{1a}

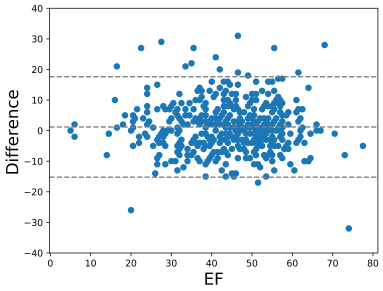
(b) O_3 vs O_{1a}



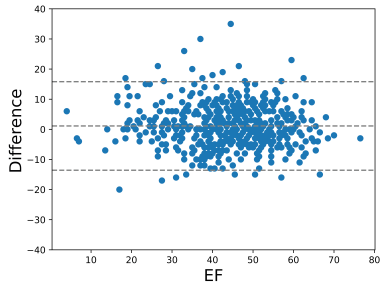
(c) O_3 vs O_2



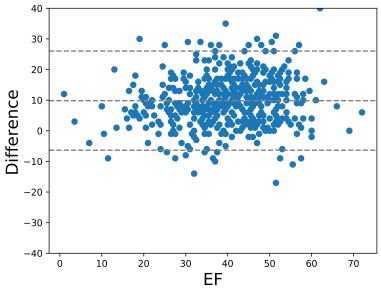
(d) O_{1b} vs O_{1a}



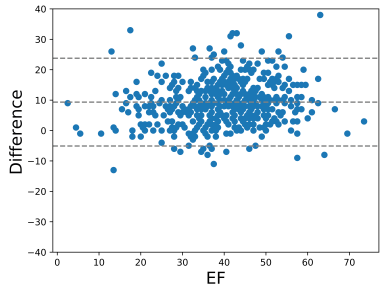
(e) U-Net 1 vs O_{1a}



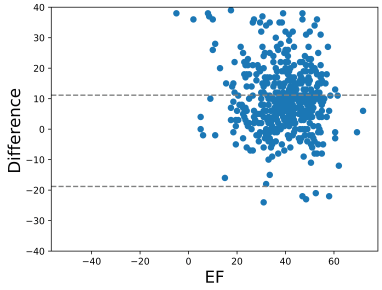
(f) U-Net 2 vs O_{1a}



(g) BEASM-auto vs O_{1a}



(h) BEASM-semi vs O_{1a}



(i) SRF vs O_{1a}

FIGURE 7.28: Bland Altman plots of the experts on fold 5 and of the algorithms on the full dataset.

7.4 Analysis of U-Net's behavior

With room for improvement, the obtained results tend to show that deep learning techniques are able to reproduce manual annotations with high fidelity. In order to define leads to improve the results, we designed and conducted a few experiments to better understand how U-Net behaves with changes in terms of models and training data.

7.4.1 Influence of the model

7.4.1.1 Stochasticity

We provide in Tab. 7.13 the results produced by our two U-Net architectures for two different trainings (U-Net 1, U-Net 1 bis, U-Net 2, U-Net 2 bis) to study the influence of the randomness, i.e without any changes in architecture, hyper-parameter nor dataset. This experiment showed that at worst the MAD and HD varied of 0.1 and 0.2 mm respectively, and that U-Net 2 were consistently better than U-Net 1. Below these variations, differences between encoder-decoders models should therefore be considered possibly due to randomness.

Table 7.13: Segmentation accuracy from U-Net 1 to U-Net 2. Bold values indicate a superior value to U-Net 1.

Methods *	LV_{endo}			LV_{epi}		
	D	**MAD**	**HD**	**D**	**MAD**	**HD**
	val.	mm	mm	val.	mm	mm
U-Net 1	0.920 ±0.056	1.7 ±1.2	5.6 ±3.3	0.947 ±0.030	1.9 ±1.1	6.0 ±3.8
U-Net 1 bis	**0.921** ±0.052	1.8 ±1.5	**5.5** ±3.1	**0.949** ±0.028	1.9 ±1.0	6.0 ±3.5
U-Net 1 Kernel 48	**0.923** ±0.047	1.9 ±1.0	**5.5** ±3.2	**0.948** ±0.026	1.9 ±1.2	6.0 ±3.8
U-Net 1 Cross-Entropy loss	0.916 ±0.047	1.9 ±1.3	5.7 ±3.5	**0.948** ±0.030	1.9 ±1.0	6.0 ±3.5
U-Net 1 Conv2DTranspose	**0.921** ±0.052	1.7 ±1.3	5.6 ±3.7	**0.948** ±0.027	1.9 ±1.0	6.1 ±4.5
U-Net 1 BatchSize 10	**0.923** ±0.055	1.7 ±1.0	5.6 ±4.3	**0.949** ±0.031	**1.8** ±1.1	6.4 ±5.2
U-Net 1 LR(1e-4)	0.913 ±0.055	1.9 ±1.1	6.1 ±3.9	0.943 ±0.034	2.1 ±1.2	6.7 ±4.4
U-Net 1 BatchNorm	**0.926** ±0.053	**1.6** ±1.4	**5.3** ±3.5	**0.948** ±0.030	1.9 ±1.0	6.0 ±3.5
U-Net 2	**0.928** ±0.054	**1.6** ±1.5	**5.4** ±3.7	**0.950** ±0.032	**1.8** ±1.0	6.1 ±4.1
U-Net 2 bis	**0.930** ±0.049	**1.5** ±0.9	**5.4** ±3.4	**0.951** ±0.024	**1.8** ±0.9	**5.9** ±3.3

7.4.1.2 Layers

We evaluate in Tab. 7.13 how the segmentation accuracy is affected by a single change of hyper-parameters. One can see that Batch Normalization is the only modification associated to an improvement consistently superior to what may be brought by pure randomness. Surprisingly, neither the loss, the batch size nor the upsampling scheme caused constant differences when changed. Augmenting the capacity by adding more filters per layer was not meaningfully beneficial, which suggests U-Net 1 already has a sufficient number of parameters.

Because of its larger number of trainable parameters, the training and running times of U-Net 2 were higher tha for U-Net 1. On our hardware (see Appendix C.1), training U-Net 1 and U-Net 2 on 400 patients respectively took 24 ± 5 min and 73 ± 1 min, while the inference lasted 0.09 ± 0.03 s and 0.14 ± 0.06 s.

In conclusion, U-Net 1 is a more efficient architecture when considering a trade-off between robustness and capacity, so we decided to focus our efforts on this model.

7.4.2 Influence of the data variety

7.4.2.1 Image quality

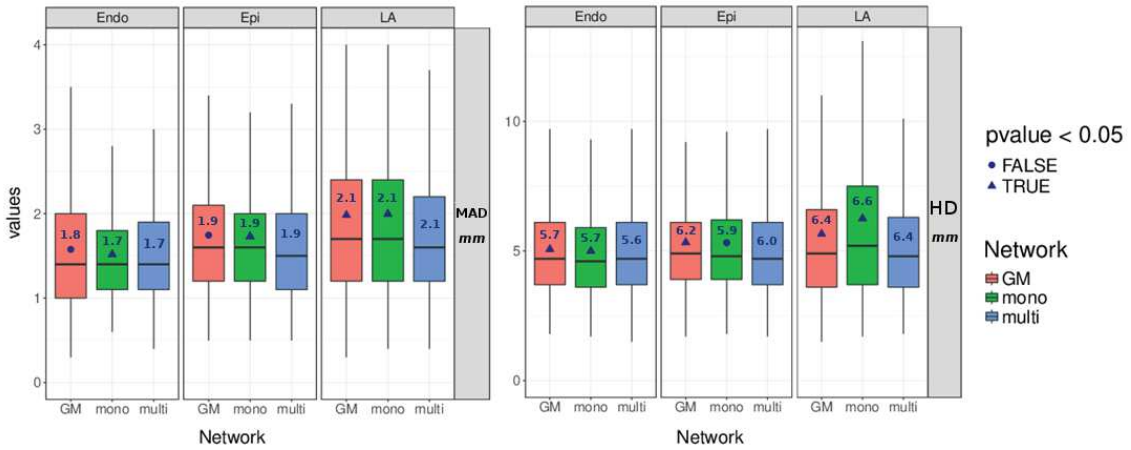
We previously observed that results obtained on poor quality images remained close to those obtained on good and medium quality images when poor quality images were present in the training set. We decided to investigate the influence of involving poor quality images in the training set for better quality images. Two U-Nets 1 were trained, one using the full training dataset (plotted in blue and referred as *multi*) and one using the training dataset restricted to patients having good and medium image quality (plotted in red and referred as *GM*).

Results are shown in Fig. 7.29. The numbers in blue correspond to the mean values of each set of measurements. All p-values were based on the Wilcoxon signed-rank test computed with the *multi* strategy as reference.

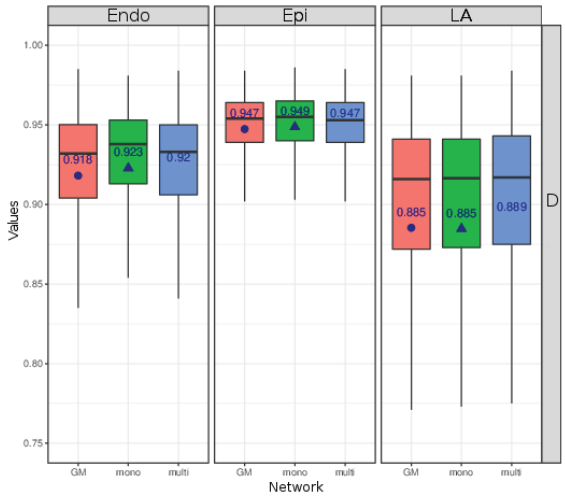
Except for the HD values, the differences were most of the time considered not statistically significant by the Wilcoxon test. The results suggest that when considering the full dataset, the 19% (94 patients) of poor quality images did not bring much additional information, supporting our intuition that the performance of the method is weakly linked to image quality. However, as including them did not decrease the performance either, it appears that learning to cope with poor image quality does not complicate the segmentation task. Therefore, U-Net is capable of handling all the variability in image quality found in echocardiography.

7.4.2.2 Number of structures

As explained in Chapter 5, the acquisitions were optimized to perform LV_{EF} measurements and the left atrium (LA) was often not fully visible. Knowing the segmentation of the LA would be very challenging on the CAMUS dataset due to the acquisition conditions, we investigated in the strategy called *multi* the capacity of U-Net to segment the LA in addition to the left ventricle and myocardium. The *mono* strategy corresponds to 3 separate U-Nets 1 trained on the full dataset with different ground truths, i.e one network trained to predict only the LV_{endo}, a second the myocardium, a third the LA, and the last all structures simultaneously. Results are plotted in green and blue in Fig. 7.29.



(a) Distance metrics



(b) Dice metric

FIGURE 7.29: Tukey box plots of the geometrical results a) MAD; b) HD; c) Dice of the U-Net 1 architecture for three different schemes.

Our findings are that U-Net 1 manages to get results on the LA close to what was obtained on the other structures, both in terms of mean absolute distance (mean values of 1.7, 1.9 and 2.1 mm for the endocardium, epicardium and LA, respectively) and average Hausdorff distance (mean values of 5.6, 6.0 and 6.4 mm for the endocardium, epicardium and LA).

Furthermore, one can see that the mono- and multi-structure approaches produced similar results, even if the corresponding differences were statistically significant. Therefore, with the proposed implementation, learning the segmentation of one structure in the context of the others does not improve the results compared to learning the segmentation of the structure alone. This hints at designing dedicated architectures and/or loss functions to better exploit the contextual information provided in the segmentation masks.

7.4.2.3 Influence of the size of the training dataset

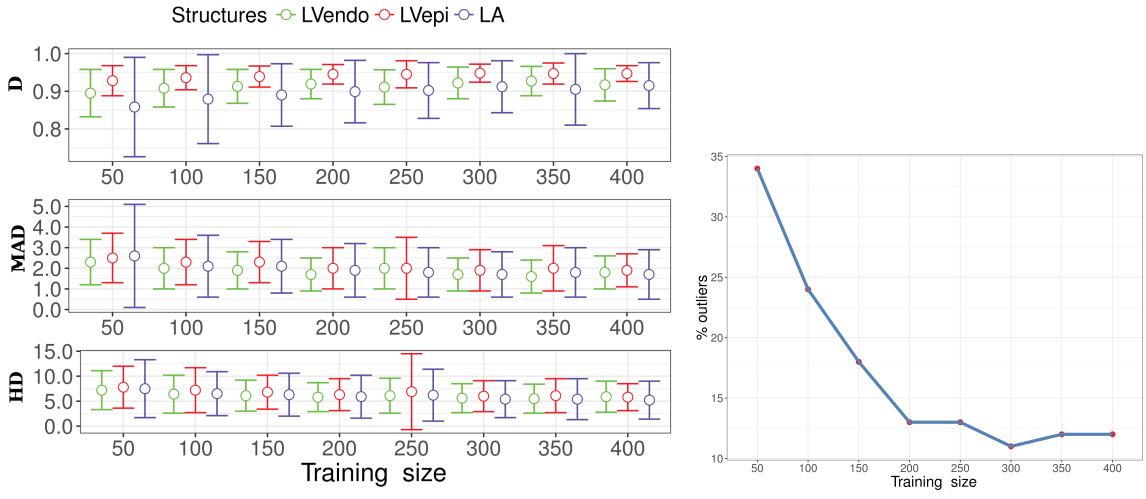
Similar to the experiment in 7.2.3, we studied in Fig. 7.30 the influence of the size of the training dataset on the quality of the segmentation of all structures, only this time we:

- used post-processing;
- used a different architecture (U-Net 1);
- inquired about all three structures;
- observed the impact on outliers.

The same folds were kept as test and validation sets. For the training sets, starting from 50 patients, we repeatedly added 50 other patients until 400 patients was reached, and trained distinct models on each training set size. The number of training epochs was proportionally lowered to ensure that each network went through the same number of iterations. As before, we observe an overall improvement of all metrics for the three cardiac structures with the increasing number of patients in the training set.

Interestingly, the improvement was stronger for the LA, possibly because of the higher variability in acquisition. For all three structures, while the improvement between 50 to 200 patients was quite pronounced, we again noted a change in the evolution of the performance from 250 patients on. The outlier rate stabilized around 12-13%, which confirms the convergence in performance.

Using post-processing appeared especially useful for small training sets, as the HD values for 50 patients, compared to the experiment in 7.2.3, was decreased by a factor 2. However, for a larger training set, it does not make as much a difference, meaning the number of false positive in the images greatly decreases as the training set size increases. Using post-processing, U-Net 1 only requires a training dataset of 50 patients to outperform both SRF and BEASM at their full capacity. However, from 250 patients on, the model appears to reach a plateau in performance.



(a) Metrics (b) Outliers G/M

FIGURE 7.30: Evolution of the segmentation scores of U-Net 1 computed on fold 5 according to the number of patients in the training dataset

7.4.2.4 Influence of the expert of reference

We investigated the influence of the expert annotations during the training phase, as shown in Fig. 7.31. To this aim, we trained three models on fold 5 with the same U-Net 1 architecture, each learning from the manual contours of a different annotator. The validation fold was kept the same for each experiment to avoid any validation bias, and models were evaluated on the remaining 400 patients (annotated by cardiologist O_{1a}).

From this figure, it is obvious that the best scores for the three structures are obtained for the model trained on the annotations of cardiologist O_{1a}, who also performed the manual contouring of the test and validation sets. Moreover, the performance of specific models is correlated with the inter-variability. This observation confirms that cardiologists have consistent differences in their way of contouring images and that an encoder-decoder architecture has the capacity to learn a specific way of segmenting echocardiographic images.

7.5 Conclusion

We showed in this chapter through a thorough analysis of the behavior and performance of U-Net that encoder-decoder models are very promising candidates for automatic multi-structure segmentation in 2D echocardiography. This algorithm outperformed the SRF and the BEASM with a significant margin, obtained better geometric and clinical results than the inter-variability between three experts, and got close to the intra-variability. We further observed that it:

- learns a specific protocol;
- is resilient to image quality;
- is fast to train and apply.

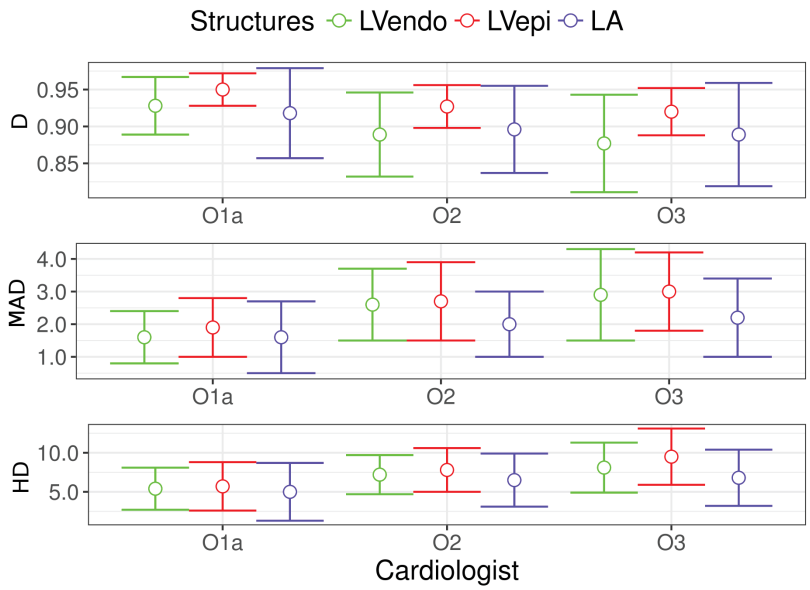


FIGURE 7.31: Geometric scores of the cardiologist-specific models.

These achievements show that U-Net is able to generalize rather properly on the task of cardiac segmentation of ultrasound images, however it suffers from four main limitations:

1. predictions are uneven, and the produced segmentation is still far from the ground truth for a significant amount of cases (18%);
2. beyond a training set size of 250 patients, the performance is only slightly improving;
3. the network does not necessarily predict a plausible shape (at least 2% of anatomical outliers, assessed visually);
4. the segmentation results are not necessarily coherent over time.

To address these limits, both the current model and the evaluation criteria are insufficient and need to be refined, as proposed in Chapter 8.

Chapter 8

Advanced deep learning models and evaluation

The following chapter concerns the study of advanced deep learning models based on the U-Net architecture, the goal being to benchmark state-of-the-art encoder-decoder approaches on the CAMUS dataset. To complete the evaluation and the ranking of the methods, we introduce anatomical metrics to refine the analysis of deep learning models. The following inquiries are therefore investigated:

- Can we improve the scores obtained by the U-Net architecture through more complex architectures recently proposed in the literature?
- How can we build an automatic and objective criteria for shape validity? What is the impact on the ranking of segmentation methods?

8.1 Deep supervision

Deep supervision (DS) was proposed in the literature as a way to efficiently train deep neural networks by setting intermediary objectives at early stages of the network [39]. DS limits the convergence problems linked to exploding and vanishing gradients. Indeed, depending on the activation function that is used, very large or small gradients can appear during the first optimization steps due to the random initialization and the multiple operations necessary to backpropagate gradients from the end to the beginning of the network [148].

We considered in this work two approaches to introduce deep supervision inside the U-Net architecture: nested deep supervision [40], and cascaded deep supervision [41].

8.1.1 U-Net++ for 2D echocardiography segmentation

8.1.1.1 U-Net ++

The authors in (Zhou et al., 2018) [40] made two observations from the analysis of U-Net and ancillary methods involving dense skip connections [149] [39] or grid connections [150]:

1. as the accurate segmentation of small regions is potentially very important in medical imaging, the segmentation network should include a progressive enhancement of the prediction. This particular attribute can be achieved with classical deep supervision;
2. encouraging semantic similarity in the skip connections, i.e. between the feature maps of the decoder and the encoder, would facilitate the learning process.

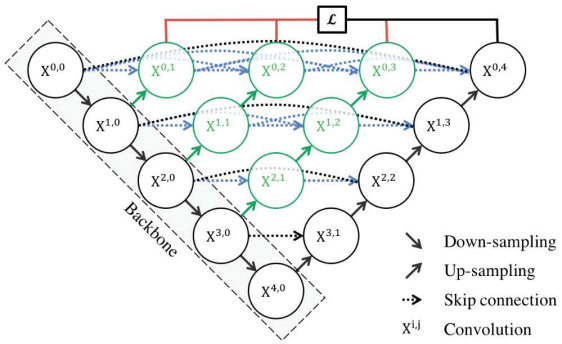


FIGURE 8.1: Original architecture of U-Net++. Deep supervision in red shows that a same loss L sends gradients to early reconstructions of the final segmentation in $X^{0,4}$. Additional convolutional layers on the skip connections are shown in green.

To this end, they proposed to decouple the original encoder feature maps and the skip-connected features by adding convolutional layers along the skip connections, and to use deep supervision to force features to be semantically close by assigning intermediary segmentation losses to their upsampled version. The full architecture is shown in Fig. 8.1.

At test time, the four outputs are averaged in an ensemble model strategy. U-Net++ was evaluated on several medical segmentation tasks including nodule segmentation in low-dose CT scans of chest, nuclei segmentation in microscopy images, liver segmentation in abdominal CT scans and polyp segmentation in colonoscopy videos.

8.1.1.2 Optimization on the CAMUS dataset

We adapted the U-Net++ architecture, starting from the official online version of the code [151], to obtain the best results possible on our CAMUS dataset, while keeping the central propositions of the paper. Table 8.1 summarizes the main results from our investigation.

TABLE 8.1: Segmentation accuracy of U-Net++ for different implementations, on the ten test sets restricted to patients having good or medium image quality (406 patients)

Methods	LV_{endo}			LV_{epi}		
	D	d_m	d_H	D	d_m	d_H
	val.	mm	mm	val.	mm	mm
U-Net++ Mono	0.909 ±0.063	2.0 ±1.6	6.6 ±4.8	0.941 ±0.033	2.2 ±1.2	**7.1** ±5.0
U-net++ Last 2	0.905 ±0.070	2.1 ±4.3	7.3 ±7.3	0.936 ±0.040	2.3 ±1.4	8.6 ±6.5
U-net++ 1	**0.915** ±0.055	**1.8** ±1.0	**6.4** ±4.0	**0.942** ±0.030	**2.1** ±1.1	**7.1** ±4.7

The original U-Net++ trained using deep supervision applied without averaging the outputs is called *U-Net++ Mono*. The results obtained by averaging the last two segmentation maps are associated to the model named *U-Net++ Last 2*. Finally, the performance obtained when changing the core with U-net 1 and not perform averaging of the outputs corresponds to the scores of *U-Net++ 1*. We removed the dropout as it was always detrimental.

From Table 8.1, one can observe that averaging the feature maps of the intermediate output worsened the results (especially visible on the standard deviations), showing that it is better to keep the last refined segmentation output without averaging in our application. We noted that the results progressively worsened as we added earlier outputs into the averaging, hinting that earlier segmentations are not refined enough to bring improvement in an ensemble strategy.

A significant gain was brought by replacing the original layers with the design of the U-Net 1 architecture. The resulting model was kept as the U-Net++ model in the sequel. In conclusion, the following changes were made compared to the original work:

- dropout was removed;
- averaging of the last feature maps of the intermediate outputs was removed;
- the original design of layers was adapted according to the choices we made to optimized U-Net 1 architecture while keeping the original network depth;
- the batch size however was set to 20 to train the network on a 8GB GPU.

Our final U-Net++ method comprises 1.1M parameters, significantly less than the original (9M), for a better performance on the CAMUS dataset.

8.1.2 Stacked hourglasses for 2D echocardiography segmentation

8.1.2.1 Stacked hourglasses

In the work of (Newell et al., 2016) [41], the authors use cascaded encoder-decoders to improve the performance of their human pose estimation network. A similar strategy was applied in (Vigneault et al., 2017) [152], in which a network was designed to automatically transform cardiac MR images into a standard orientation before the segmentation . Stacked Hourglasses (SHG) integrate successive encoder-decoder networks (usually of the same architecture) into a single network, as illustrated in Fig. 8.2.

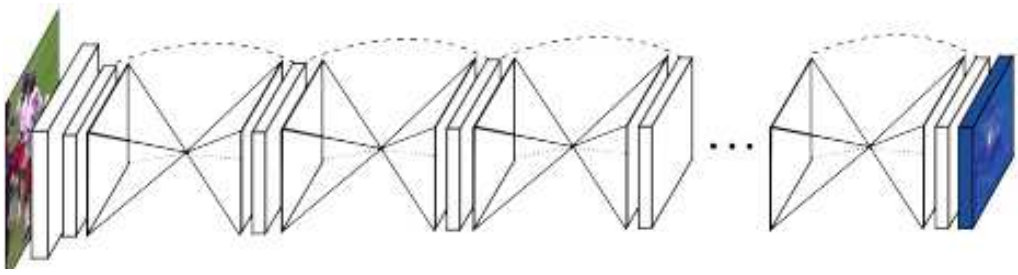


FIGURE 8.2: Original architecture of SHG. The input of cascaded networks is the concatenation of the original image and the previous segmentation result.

The first networks are used as residual blocks, i.e. the input of the next network is the result of the addition between the previous segmentation and the previous input. The image is transformed by a convolution layer to fit the size of the segmentation output (in our case 4 channels) before addition. Each output of the encoder-decoder networks is associated to an intermediate segmentation loss in a deep supervision strategy which, combined to the residual connections, implies that the cascaded networks learn to refine the previous segmentation. Lastly, the output of the third network is used as the final segmentation result.

8.1.2.2 Optimization on the CAMUS dataset

As the original network was trained to produce heat maps for the detection of several landmarks, the task is quite similar to semantic segmentation and the pipeline was kept the same. However, for comparison purposes, we used the U-Net 1 architecture as the key encoder-decoder network in our SHG implementation. We observed in particular that:

1. residual connections (summing of feature maps) improved results compared to skip connections (concatenation);
2. very little improvement was brought by the refining networks, and no improvement was obtained from three stacked networks on. We therefore used three stacked U-Net 1s;
3. we obtained slightly better results with a 4.5M parameters version than with 6M (stacking three light U-Net 1). This also allowed to keep the same batch size as U-Net 1 when training the network on a 8 GB GPU.

8.2 Encouraging shape validity in 2D echocardiography segmentation

One observation made in Chapter 3 was that U-Net does not always produce anatomically plausible shapes. Global shape requirements can be directly reinforced by their addition to the objective function of neural networks. To be optimized with gradient descent, the corresponding function must be differentiable according to the model parameters.

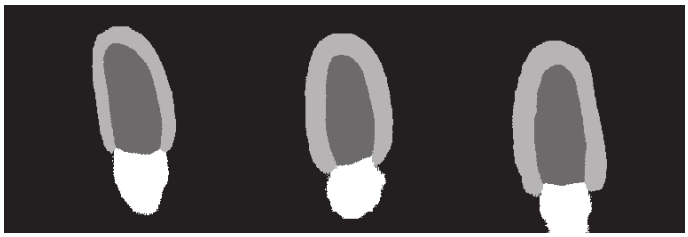
We study in the part the Anatomically Constrained Neural Network (ACNN) proposed in (Oktay et al. 2017) [42], and briefly described in Section 4.3.4.4. This network encourages segmentation results with plausible anatomical shapes through the design of an implicit shape constraint created from an auto-encoder.

8.2.1 Auto-encoder for 2D echocardiography reconstruction

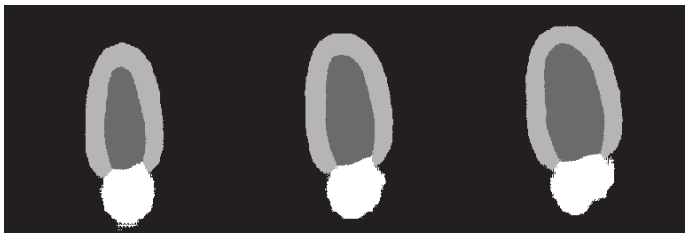
8.2.1.1 Auto-encoder

The shape constraint involved in the ACNN is built using an auto-encoder, i.e. a network that learns to reconstruct images from their projection onto a sub-dimensional representation space [120], usually a 1D vector of small size c (referred to as code). In other words, auto-encoders learn a mapping $X \to Z \to X$, with $Z \in \mathbb{R}^c$ and X a given image space.

Auto-encoders have been widely used in compression, inpainting and denoising tasks. Interestingly, auto-encoders have recently become a popular solution for image generation, especially through the variational auto-encoder and the conditional auto-encoder approaches which allow to structure the underlying latent space.



(a) Reconstruction of the first mode



(b) Reconstruction of the second mode

FIGURE 8.3: First two modes of an auto-encoder trained on fold 5 with a 94% average accuracy, i.e. quantity of pixels rightly classified. Left: $z = z_{mean} - 4 \times \lambda_p \times v_p$, middle: $z = z_{mean}$, right: $z = z_{mean} + 4 \times \lambda_p \times v_p$.

Variational auto-encoders force the latent space to be more compact and therefore more continuous by associating each representation to a gaussian distribution instead of a single point. Conditional auto-encoders include a pre-definite code coefficient that is a secondary input of the network and may represent a class [120].

8.2.1.2 Application on the CAMUS dataset

Details on the auto-encoder architecture, tuning strategy and results on the CAMUS dataset can be found on Appendix F. In particular, we could observe that:

- auto-encoders were less stable to train and to apply than a U-Net (divergence, strongly missed reconstructions because of high code coefficients), which led us to investigate activation functions and layer regularization to ensure a stable and smooth training for all ten models of the cross-validation;
- a code of 32 was sufficient to encode the variations of our segmentation masks;

Applying eigen decomposition to the covariance matrix of the latent space representations allows to observe its principal modes of variations through the corresponding deformations of the mean shape:

1. we sort the eigen values λ_p associated to the eigen vectors v_p according to $|\lambda_p|$ (PCA);
2. we reconstruct the images of $z = z_{mean} \pm 4 \times \lambda_p \times v_p$ using the decoder. A factor of 4 was chosen instead of the traditional value 3 to observe the variations more clearly.

We show visuals for the first two modes computed on one fold of the CAMUS dataset in Fig 8.3. Modes are not straightforward to interpret, as they are not bound to a single morphological deformation. For instance, the first mode seems to encode orientation- and position-related information, while the second one appears to control the left ventricle (LV) and the left atrium (LA) sizes, i.e. time instant-related information.

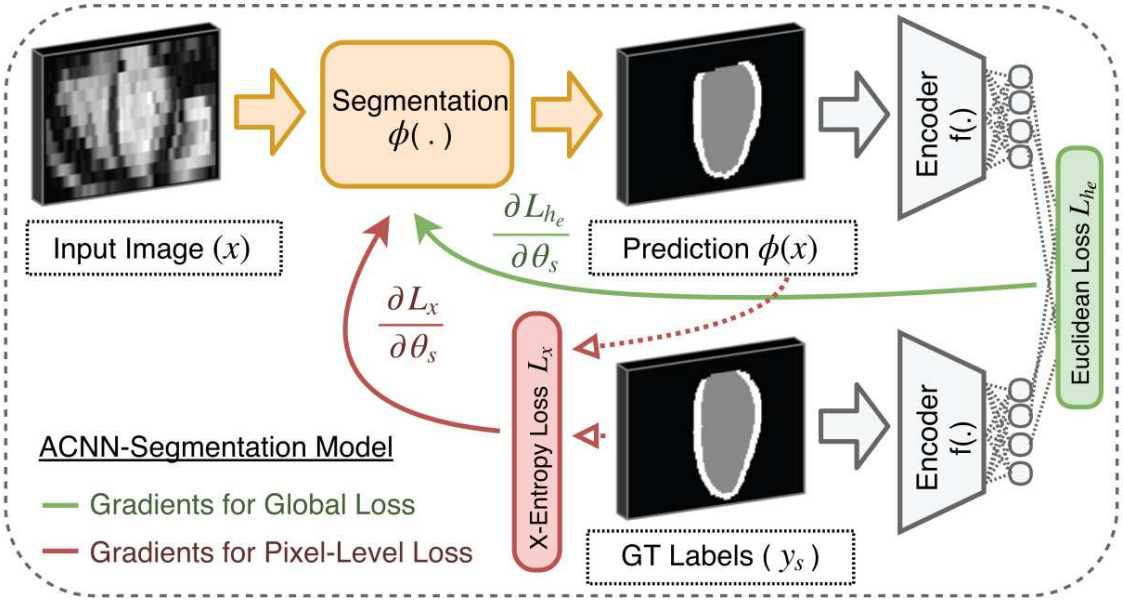


FIGURE 8.4: Pipeline for segmentation tasks of the ACNN model, introduced in (Oktay et al., 2017) [42].

8.2.2 Anatomically constrained neural network for 2D echocardiography segmentation

8.2.2.1 Anatomically constrained neural network

The ACNN pipeline for segmentation tasks in shown in Fig. 8.4. In particular, the segmentation optimization is performed through the following loss function:

$$L = L_x + \lambda_1 \times L_{he} \tag{8.1}$$

with L_x the segmentation loss (here the categorical cross-entropy), L_{he} the loss derived from the auto-encoder and λ_1 an hyper-parameter balancing the two losses.

L_{he} corresponds to the euclidean loss between the ground truth codes and the prediction ones obtained from the frozen auto-encoder, i.e. whose weights do not evolve through the training. This shape regularization loss aims at bringing the sub-dimensional representation of the ground truth and the prediction closer. Therefore, it implicitly encourages the overall segmentation result to be similar to the ground truth at the level of accuracy encoded in the latent space of the auto-encoder.

In the original paper of (Oktay et al., 2017) [42], the segmentation network was an encoder-decoder similar to a 3D U-Net whose architecture could be adapted to perform super-resolution instead of segmentation, i.e. to predict a high resolution segmentation mask from a low resolution image. As detailed in Section 4.3.5, segmentation ACNNs currently hold among the best performing methods on the CETUS dataset.

8.2.2.2 Optimization on the CAMUS dataset

For a fair comparison, we used U-Net 1 as the segmentation network in our ACNN implementation. The following changes were made to obtain the best results on our dataset:

1. a code of 32 coefficients was set for the auto-encoder network (which still allowed an average reconstruction accuracy of 97% on the training set);
2. the hyper-parameter balancing the segmentation and shape regularization losses was set so the two losses had close initialization values.

The resulting ACNN models had 2.2M parameters. Details on the behavior of ACNNs can be found in Appendix F. In particular, we investigated the influence of the loss balance and the training set size. The main observations were that:

- it was possible to train the segmentation network with L_{he} only, which ensured that the loss gradients were safely propagated from the auto-encoder to the segmentation network;
- the shape regularization could hinder results when weighted too strongly;
- the shape regularization improved results when a small training dataset was available, but this regularization benefit quickly faded as more training data was added.

8.3 Evaluation of advanced encoder-decoder models for 2D echocardiography image analysis

We evaluate in this section the performance of the three encoder-decoder based structures described above and compare their results with the ones obtained by the baseline approach, i.e. the U-Net 1 model. This work has been published in the IEEE TMI journal [5].

8.3.0.1 Evaluation

For a fair comparison with U-Net 1, the evaluation was kept the same as the one described in Section 7.3, involving the same geometrical and clinical metrics. We used the same pre- and post processing strategies, data formatting and selection. In addition, the same values of training hyper-parameters (otherwise mentioned in the description of the methods) were used to ensure that the comparison allowed to highlight the benefits or disadvantages of the extended architectures described above.

8.3.0.2 Geometrical results

The geometrical scores obtained on good and medium quality images are reported in Tab. 8.2. From this table, one can see all encoder-decoder architectures achieved close results, all below the inter-variability. In particular, compared to the U-Net 1 baseline model, the obtained scores were observed to be:

- equivalent, although marginal improvement can be noticed with the SHG network;
- slightly degraded for the ACNN approach for the HD metrics;
- degraded for the U-Net++ architecture;

Concerning ACNN, the similar scores may be explained by the simple shapes encountered in 2D echocardiography. Indeed, the reference contours drawn by the experts involve truncated ellipse-like shapes whose information seems to be easily learned by the different encoder-decoders when looking at visuals (Fig. 8.7 to 8.5).

TABLE 8.2: Segmentation accuracy of the 4 methods on the ten test folds, restricted to patients having good & medium image quality (406 patients).

Method	ED						ES					
	LV_{endo}			LV_{epi}			LV_{endo}			LV_{epi}		
	D	**MAD**	**HD**	**D**	**MAD**	**HD**	**D**	**MAD**	**HD**	**D**	**MAD**	**HD**
	val.	mm	mm	val.	mm	mm	val.	mm	mm	val.	mm	mm
O_{1a} vs O_2 (inter-obs)	0.919 ±0.033	2.2 ±0.9	6.0 ±2.0	0.913 ±0.037	3.5 ±1.7	8.0 ±2.9	0.873 ±0.060	2.7 ±1.2	6.6 ±2.4	0.890 ±0.047	3.9 ±1.8	8.6 ±3.3
O_{1a} vs O_3 (inter-obs)	0.886 ±0.050	3.3 ±1.5	8.2 ±2.5	0.943 ±0.018	2.3 ±0.8	6.5 ±2.6	0.823 ±0.091	4.0 ±2.0	8.8 ±3.5	0.931 ±0.025	2.4 ±1.0	6.4 ±2.4
O_2 vs O_3 (inter-obs)	0.921 ±0.037	2.3 ±1.2	6.3 ±2.5	0.922 ±0.036	3.0 ±1.5	7.4 ±3.0	0.888 ±0.058	2.6 ±1.3	6.9 ±2.9	0.885 ±0.054	3.9 ±1.9	8.4 ±2.8
O_{1a} vs O_{1b} (intra-obs)	0.945 ±0.019	1.4 ±0.5	4.6 ±1.8	0.957 ±0.019	1.7 ±0.9	5.0 ±2.3	0.930 ±0.031	1.3 ±0.5	4.5 ±1.8	0.951 ±0.021	1.7 ±0.8	5.0 ±2.1
U-Net 1	**0.934** ±0.042	**1.7** ±1.0	**5.5** ±2.9	**0.951** ±0.024	**1.9** ±0.9	5.9 ±3.4	0.905 ±0.063	**1.8** ±1.3	**5.7** ±3.7	0.943 ±0.035	**2.0** ±1.2	6.1 ±4.1
U-Net ++	0.927 ±0.046	1.8 ±1.1	6.5 ±3.9	0.945 ±0.026	2.1 ±1.0	7.2 ±4.5	0.904 ±0.060	**1.8** ±1.0	6.3 ±4.2	0.939 ±0.034	2.1 ±1.1	7.1 ±5.1
SHG	**0.934** ±0.034	**1.7** ±0.9	5.6 ±2.8	**0.951** ±0.023	**1.9** ±1.0	**5.7** ±3.3	**0.906** ±0.057	**1.8** ±1.1	5.8 ±3.8	**0.944** ±0.034	**2.0** ±1.2	**6.0** ±4.3
ACNN	0.932 ±0.034	**1.7** ±0.9	5.8 ±3.1	0.950 ±0.026	**1.9** ±1.1	6.4 ±4.1	0.903 ±0.059	1.9 ±1.1	6.0 ±3.9	0.942 ±0.034	**2.0** ±1.2	6.3 ±4.2

LV_{endo}: Endocardial contour of the left ventricle; LV_{epi}: Epicardial contour;
ED: End diastole; ES: End systole;

Concerning SHG, the similar scores may be explained by the plateau in performance observed in Chapter 7 when training U-Net 1 on more than 250 patients. This suggests that the capacity of a U-Net 1 is sufficient to generalize well on the CAMUS dataset, and explains why a more complex architecture like SHG did not bring any improvement.

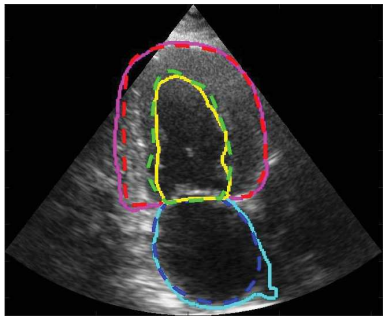
Outlier rates are given in Tab. 8.3. One can see from this table that U-Net and SHG achieved the same outlier rates, which reveals that the more complex architecture involved in the SHG model did not improve the robustness of the segmentation with respect to geometrical scores. ACNN increased the number of outliers, certainly due to higher HD scores.

Finally, the outlier rate of U-Net++ is of 30%, which, coupled with the geometrical scores, confirms it is less efficient than the others. This loss of performance may be due to forcing the feature maps of the network to correspond to segmentation masks too early.

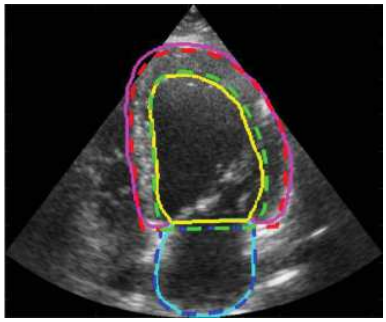
TABLE 8.3: Geometrical outlier rates

Method	U-Net 1	U-net ++	SHG	ACNN
# Outliers	18%	30%	18%	22%

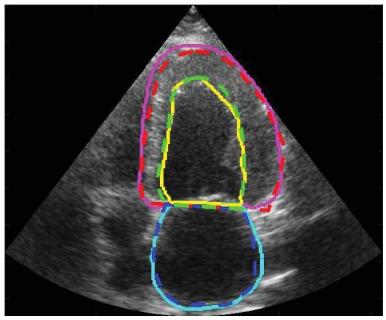
(a) 2CH-ED: MAD = 1.3 | 1.7
HD = 3.7 | 4.6 mm.



(b) 2CH-ES: MAD = 1.8 | 1.7
HD = 4.9 | 5.2 mm.

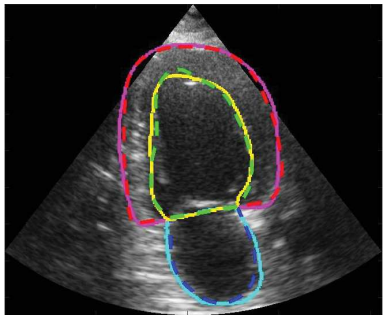


(c) 4CH-ED: MAD = 2.5 | 2.2
HD = 6.3 | 4.7 mm.

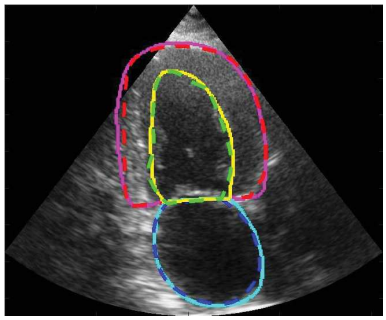


(d) 4CH-ES: MAD = 1.4 | 1.9
HD = 4.1 | 4.8 mm.

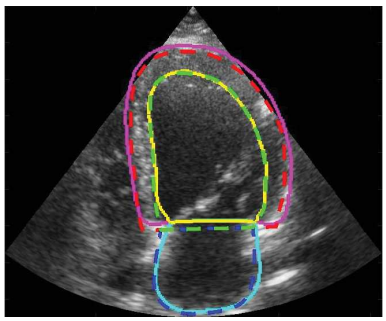
FIGURE 8.5: Segmentation results obtained by U-Net 1 ++.



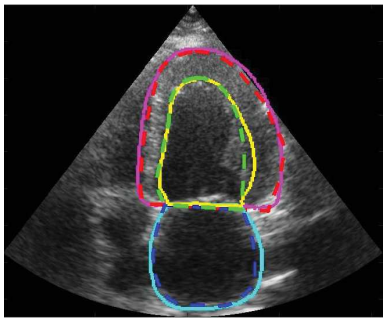
(a) 2CH-ED: MAD = 1.1 | 1.1
HD = 4.0 | 3.1 mm.



(b) 2CH-ES: MAD = 1.2 | 1.4
HD = 4.1 | 4.7 mm.

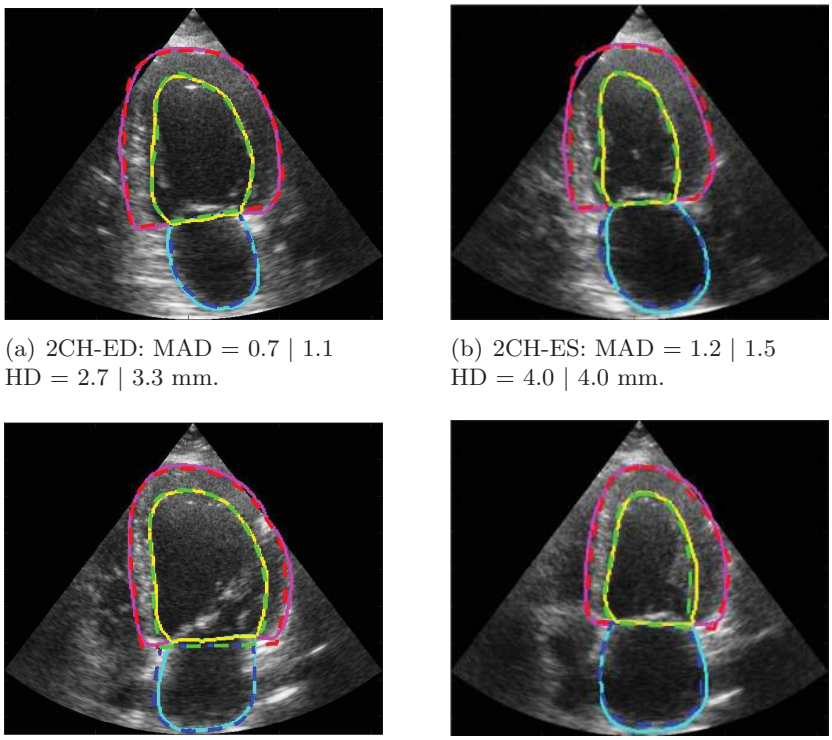


(c) 4CH-ED: MAD = 1.1 | 2.7
HD = 6.3 | 5.1 mm.



(d) 4CH-ES: MAD = 1.5 | 1.6
HD = 5.5 | 4.1 mm.

FIGURE 8.6: Segmentation results obtained by SHG.



(a) 2CH-ED: MAD = 0.7 | 1.1
HD = 2.7 | 3.3 mm.

(b) 2CH-ES: MAD = 1.2 | 1.5
HD = 4.0 | 4.0 mm.

(c) 4CH-ED: MAD = 1.2 | 1.7
HD = 5.9 | 5.1 mm.

(d) 4CH-ES: MAD = 1.3 | 1.3
HD = 3.7 | 3.7 mm.

FIGURE 8.7: Segmentation results obtained by ACNN.

8.3.0.3 Visual analysis

To allow visual assessment of the segmentation of quality of the results, we provide from Fig. 8.7 to 8.5 the segmentation results obtained by the different encoder-decoder methods tested on the same patient than the one used to illustrate the performance of the U-Net methods in Chapter 7. Ground-truth contours are dotted and prediction contours are drawn in full line.

From these visuals, it can be observed that SHG and ACNN predict smoother shapes compared to U-Net++ and U-Net 1 (see visuals of U-Net 1 in Fig. 7.18) on this patient. A trend is however hard to assess without an automatic criteria to assess the shape viability.

8.3.0.4 Clinical results

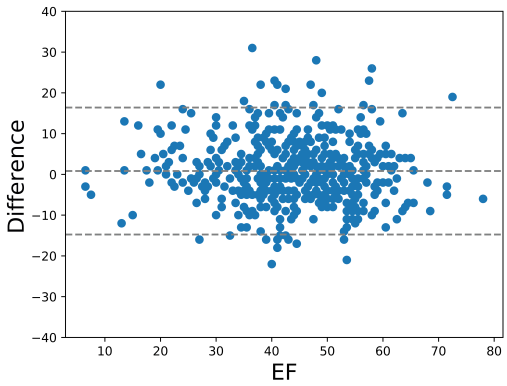
Tab. 8.4 provides the clinical scores obtained by the 4 evaluated encoder-decoder methods. From this table, it can be observed that SHG and ACNN obtain scores similar to U-Net 1. The LV volume estimations from U-Net++ predictions are worse than the ones of the three other encoder-decoders, however the ejection fraction estimation is similar to the score of U-Net 1.

Bland altman plots given in Fig. 8.8 confirm the lack of robustness of U-Net++ when compared to other methods. Still, an interesting result is that all encoder-decoders perform better than the inter-variability computed on the CAMUS dataset, reinforcing the high potential of such approaches.

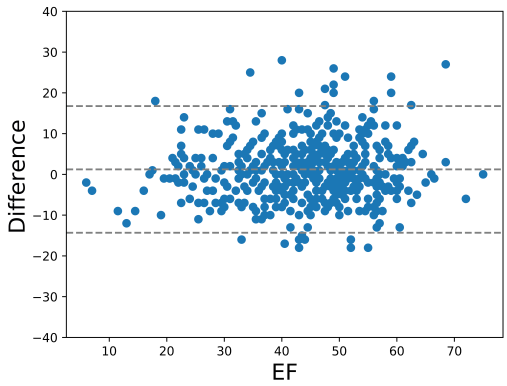
TABLE 8.4: Clinical metrics of the 4 evaluated methods on the ten test folds restricted to patients having good & medium image quality (406 patients).

Method	$\mathbf{LV_{EDV}}$			$\mathbf{LV_{ESV}}$			$\mathbf{LV_{EF}}$		
	corr	**bias**$\pm\sigma$	**mae**	**corr**	**bias**$\pm\sigma$	**mae**	**corr**	**bias**$\pm\sigma$	**mae**
	val.	ml	ml	val.	ml	ml	val.	%	%
O_{1a} vs O_2	0.940	18.7±12.9	18.7	0.956	18.9±9.3	18.9	0.801	-9.1±8.1	10.0
O_{1a} vs O_3	0.895	39.0±18.8	39.0	0.860	35.9±17.1	35.9	0.646	-12.6±10.0	13.4
O_2 vs O_3	0.926	-20.3±15.6	21.0	0.916	-17.0±13.5	17.7	0.569	3.5±11.0	8.5
O_{1a} vs O_{1b}	0.978	-2.8±7.1	6.2	0.981	-0.1±5.8	4.5	0.896	-2.3±5.7	4.5
U-Net 1	**0.947**	-8.3±12.6	10.9	0.955	-4.9±9.9	**8.2**	0.791	-0.5±7.7	**5.6**
U-Net ++	0.946	-11.4± 12.9	13.2	0.952	-5.7± 10.7	8.6	0.789	-1.8± 7.7	**5.6**
SHG	0.943	6.4±12.8	**10.5**	0.938	-3.2±11.3	**8.2**	0.770	-1.4±7.8	5.7
ACNN	0.945	-6.7± 12.9	10.8	0.947	-4.0± 10.8	8.3	**0.799**	-0.8±7.5	5.7

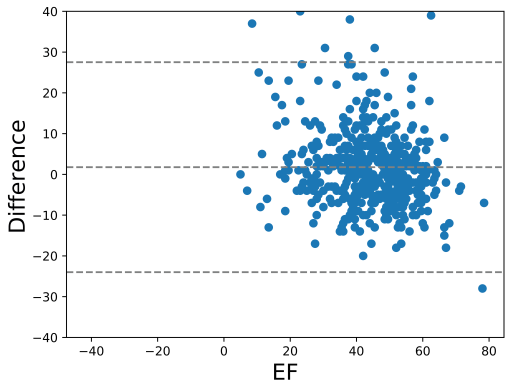
corr: Pearson correlation coefficient; mae: mean absolute error.



(a) ACNN vs O_{1a}



(b) SHG vs O_{1a}



(c) U-Net++ vs O_{1a}

FIGURE 8.8: Bland Altman of the three EDNs on the full dataset.

8.3.0.5 Conclusion

The study conducted in this first part of the chapter shows that the tested state-of-the-art encoder-decoder networks, involving more complex architecture than the U-Net 1, do not produce better results on the CAMUS dataset. This observation can be partly explained by the fact that U-Net already reaches a plateau in performance as shown in Chapter 7. However, to better assess the quality of results produced by the different methods (i.e. avoiding misguiding interpretations of isolated cases), it appears interesting to use additional metrics derived from the inter-observer variability. This aspect is described in the second part of this chapter.

8.4 Designing cardiac shape plausibility metrics

So far, we evaluated the methods of our study on geometrical and clinical metrics. We further completed the robustness assessment through the notion of geometrical outliers, i.e established from the geometrical scores as cases showing high MAD and HD values. In this section, we enhance the evaluation by adding complementary shape assessment metrics, from which we build the notion of anatomical outlier for 2D echocardiography.

This work has been published in the MIDL conference 2019 [43].

8.4.1 Motivations

The interest in forcing anatomically viable shapes for automatic algorithm predictions is not obvious, as shape plausibility does not necessarily correlate with contour and indice accuracy. However, shape plausibility remains a given for manual annotations, as no expert would produce contours such as the outliers we presented in the SRF study (Chapter 6) and the U-Net study (Chapter 7). Therefore, it appears that learning to reproduce expert contours implies for supervised learning methods to product shapes of expected appearances.

Shape plausibility assessment is rarely performed in the evaluation of automatic segmentation algorithms. When it is, the assessment is often performed visually. In the experiment conducted in Chapter 7 for instance, we estimated the number of anatomical outliers, i.e. segmentation masks containing at least one implausible shape variation, to be about 2%. However, visual assessment has two limitations:

1. it requires a human eye and attention, which does not scale with large datasets and numerous experiments;
2. it is subjective and therefore not reproducible.

Local distances are sometimes used to ponder the plausibility of a segmentation (structure diameter or area [153] [51], however such information does not encode local validation of the shape plausibility.

It is therefore necessary to establish a way to characterize geometrical shapes through objective, automatic and differentiating measurements.

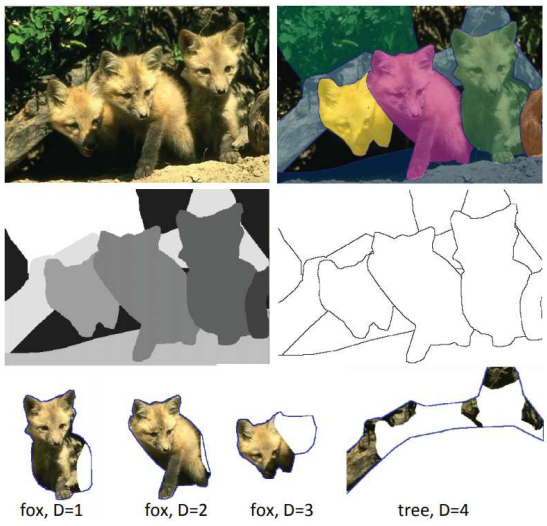


FIGURE 8.9: Semantic amodal segmentation as introduced in (Zhu et al. 2017) [44]. Instance segmentation masks are shown in the upper right, next to the image. Semantic masks are shown below the image next to corresponding contours. Finally, amodal contours are shown at the bottom, displaying continuity and simplicity to infer the hidden information.

8.4.2 Cardiac shape characterization in 2D echocardiography

8.4.2.1 Simplicity and convexity

In the work of (Yan et al., 2017) [44], the authors invented a new semantic segmentation problem in computer vision, that they called amodal segmentation. Amodal segmentation consists in predicting contours including hidden parts of objects, as illustrated in Fig. 8.9. In order to assess whether this problem is well-posed or not, the authors investigated the characteristics of the shapes predicted by human operators to show that a human would naturally use symmetry, object knowledge and continuity to infer hidden shapes, as well as opt for simplicity (straight lines) in case of missing information.

To automatically compare the segmentation of several structures S by different annotators, the authors used two geometrical criteria, the convexity and the simplicity:

$$Convexity : Cx(S) = \frac{Area(S)}{Area(ConvHull(S))} \tag{8.2}$$

with $ConvHull(S)$ the convex envelope of S.

$$Simplicity : Sp(S) = \frac{\sqrt{4\pi \times Area(S)}}{Perimeter(S)} \tag{8.3}$$

These two metrics hold values between 0 and 1, and are maximized for a circle. Convexity and simplicity give discriminative values for any convex shapes, such as oval shapes like heart cavities, and bridge-like shapes like the myocardium.

8.4.2.2 Cardiac shape validity assessment on the CAMUS dataset

In echocardiography, the predicted contours for the left ventricle, the myocardium, and the left atrium show characteristic shapes, as shown on Fig 8.10. Especially, due to locally missing

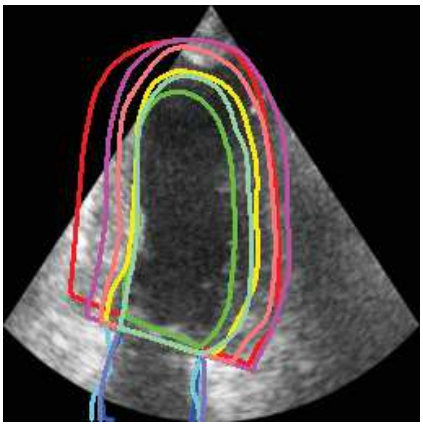


FIGURE 8.10: Contours drawn by the two cardiologists of our study on the same case. Despite the high distance between the annotations, the predicted shapes are similar in appearance.

information, the predicted shapes are characterized by:

- pre-definite forms (circular, oval, elongated...);

TABLE 8.5: Simplicity and convexity values computed from the three experts' annotations on 50 patients (200 images). Values in red correspond to the minimal value used for the outlier criteria.

Method	LV_{endo}		LV_{epi}	
	Cx	**Sp**	**Cx**	**Sp**
All experts	0.975 ±0.022 >0.741	0.722 ±0.040 >0.529	0.992 ±0.004 >0.960	0.794 ±0.022 >0.694
O_{1a}	0.970 ±0.027 >0.741	0.707 ±0.046 >0.529	0.993 ±0.002 >0.983	0.801 ±0.023 >0.721
O_{1b}	0.973 ±0.024 >0.862	0.714 ±0.045 >0.560	0.993 ±0.001 >0.988	0.806 ±0.020 >0.744
O_2	0.984 ±0.014 >0.871	0.728 ±0.037 >0.629	0.990 ±0.005 >0.960	0.773 ±0.025 >0.694
O_3	0.973 ±0.020 >0.832	0.738 ±0.031 >0.638	0.989 ±0.006 >0.964	0.794 ±0.021 >0.717

TABLE 8.6: Anatomical outlier criteria

Structure	Cx	Sp
LV	< 0.741	< 0.529
Epi	< 0.960	< 0.694

- smooth contours, resulting from manual annotations.

Tab. 8.5 summarizes the simplicity and convexity scores computed from all four sets of manual annotations on fold 5 (O_{1a}, O_{1b}, O_2 and O_3) for the endocardial and epicardial borders (as the LA is not fully present in all images). From this table it can be noted that:

- the endocardium and the epicardium are, as expected, both characterized by a high convexity and simplicity;
- the endocardium and the epicardium show distinct values of simplicity and convexity. Especially, the LV_{epi} shows on average more convex and more simple shapes than the LV_{endo}, which is also supported by the smaller standard deviations and minimum values;
- cardiologists learn to produce contours based on pre-established shape priors (close average values for all metrics for all experts), however their way of contouring is also reflected on the simplicity and the convexity of the shapes they draw (0_{1b} scores close to the ones of O_{1a}).

From these observations, we established that an anatomical outlier criteria would be characterized by unusually low values of convexity and simplicity, as summarized in Tab. 8.6.

8.4.3 Impact on the ranking of supervised learning segmentation methods

We provide in Tab. 8.7 the scores obtained by four encoder-decoders on the full CAMUS dataset:

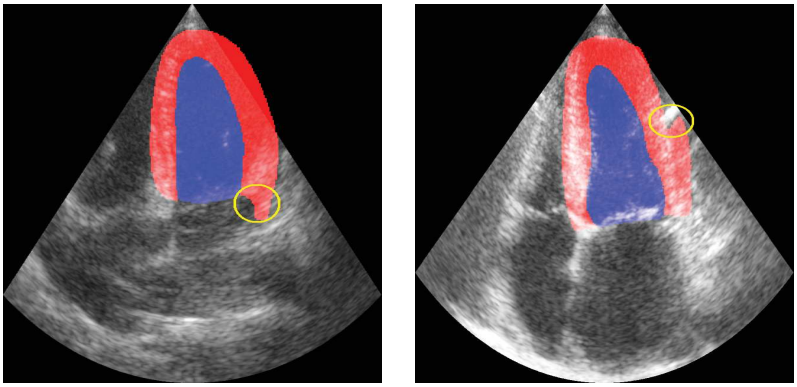
- U-Net 1 and U-Net 2, the two models from Chapter 7;
- SHG, which supposedly refines the segmentation result;
- ACNN, which supposedly constrains the produced shapes.

From these results, several observations can be made:

1. all models produced on average less convex and more complex shapes than experts;
2. U-Net 1 only produced 5% of anatomical outliers, which support the idea that the network implicitly learned to reconstruct coherent segmentation masks;
3. though U-Net 2 outperformed U-Net 1 on all classical geometrical metrics (D, MAD, HD), it produced three times more anatomically implausible shapes. This may be due to its higher number of parameters, and therefore a sign of over-fitting;
4. the refinement effect of SHG can be observed based on the significant reduction of anatomical outliers (from 95 to 47);

TABLE 8.7: Anatomic scores and outlier rates computed on the full dataset (500 patients).
ana: anatomical , geo: geometrical

Method	LV-endo		LV-epi		Outliers		
	Cx	**Sp**	**Cx**	**Sp**	**geo**	**ana**	**geo ∩ ana**
U-Net 1	0.958 ±0.022	**0.665** ±0.037	0.976 ±0.012	**0.743** ±0.022	423 21%	95 5%	71 4%
U-Net 2	0.952 ±0.030	0.658 ±0.045	0.970 ±0.028	0.732 ±0.045	519 26%	318 16%	231 12%
SHG	**0.960** ±0.018	0.664 ±0.036	**0.979** ±0.008	0.742 ±0.021	425 21%	**47** 2.4%	**29** 1.5%
ACNN	0.956 ±0.019	0.663 ±0.036	0.974 ±0.014	0.740 ±0.026	**417** 21%	147 7%	98 5%



(a) Cx= 0.96|0.94Sp= 0.67|0.71 (b) Cx= 0.93|0.95Sp= 0.65|0.70

FIGURE 8.11: Anatomical outliers from U-Net 2: a) is also a geometrical outlier but not b). Local shape irregularities are cercled in yellow.

5. the anatomical constraints of ACNNs creates more anatomical outliers on our dataset;
6. anatomical and geometrical outliers are often linked, as most anatomical outliers are also geometrical outliers.

Therefore, the design of anatomical metrics has allowed to confirm with an objective criteria that the U-Net 1 architecture is a better candidate than the U-Net 2 architecture to reproduce the hand of a cardiologist. It also allowed to confirm that the cascaded architecture of SHG performed a shape refinement, which did not show either on classical geometrical metrics. Finally, we could also show that on our specific dataset and application and with the present auto-encoder and segmentation models, the shape constraint of ACNN was detrimental.

8.4.4 Discussion

The proposed anatomical criteria has two main strengths:

1. it is sensitive to local deformations, as seen on Fig. 8.11;

2. it does not require the ground truth contour to be computed, and can be used to detect potentially missed cases since most anatomical outliers are also geometrical outliers.

However, it also suffers from three main limitations:

1. it was established on a small set of annotations, image-wise and expert-wise, implying our outlier criteria is necessarily imprecise;

2. anatomical plausibility does not imply a better accuracy, so this criteria can only be second to MAD and HAD values;

3. we did not take into account the fact that the shapes, and especially the LV_{endo}, should be characterized by different limits of convexity and simplicity at ED and ES;

4. we only considered lower limits for the shape plausibility, and not intervals.

These limitations would have to be addressed for a better shape assessment. However, the definition we propose is sufficient to compare the predictions from encoder-decoders.

8.4.5 Conclusion

The introduction of anatomical metrics allows to complete the evaluation of supervised learning models, coupling the accuracy on contours and estimated clinical indices with the shape smoothness and regularity observed on manual contours.

Though imprecise, the anatomical outlier criteria we design has allowed to both observe the over-fitting of U-Net 2 and the refinement effect in SHG. U-Net 1 was confirmed to be the most promising supervised learning method of this study.

8.5 Conclusion

According to classical evaluation metrics, the enhancement of U-Net 1 through classical additions (SHG, U-Net++, ACNN) did not yield a significant improvement, which supports the hypothesis that a plateau has been reached in performance. To better compare models, we completed the evaluation metrics of our study with anatomical metrics based on geometric shape assessment. These metrics show distinctive values for the cardiac structures extracted from echocardiography images, and can be used as a detector for potentially failed cases. Furthermore, anatomical metrics revealed that multiple successive segmentations could improve results, though marginally.

In the following and last contribution chapter, we investigate the design of multi-task pipelines to produce more robust segmentation results, and exceed the limit in performance observed on the CAMUS dataset.

Chapter 9

Attention-learning models to improve the robustness of deep learning segmentation in 2D echocardiography

The study presented in Chapter 7 revealed that segmentation methods based on encoder-decoder networks (U-Net [34]) could outperform state-of-the-art methods with a fine margin and produce accurate results on average lower than the inter-observer variability with regards to distance metrics. However, the intra-variability remained out of reach, and the investigation conducted in Chapter 8 to improve results further using model augmentations such as deep supervision and anatomical constraints was unsuccessful.

In this context, this chapter aims at providing answers to the following three questions:

1. Is it possible to further improve the accuracy of convolutional neural networks (CNNs) for segmentation in echocardiographic imaging?
2. Can the number of outliers be significantly reduced by changing the architecture?
3. Do CNNs allow the achievement of results below intra-observer variability, both in terms of segmentation and clinical index estimation?

The corresponding results were published in the IEEE International Ultrasonic Symposium conference [45] and recently submitted in the IEEE Transactions on Ultrasonics, Ferroelectrics and Frequency Control (TUFFC) journal [46].

9.1 Motivations

The experiments carried out in Chapters 7 and 8 highlighted two interesting outputs:

- the scores produced by U-Net models were little sensitive to hyper-parameter choices (see Section 7.4.1.2), which highlights the stability of this architecture, but also limits the improvement to seek through hyper-parameter tuning;
- the use of more sophisticated encoder-decoder architectures (i.e. U-Net ++ [40], stacked hourglasses network (SHG) [41] and anatomically constrained neural network (ACNN) [42]) did not produce better results than the baseline on our dataset.

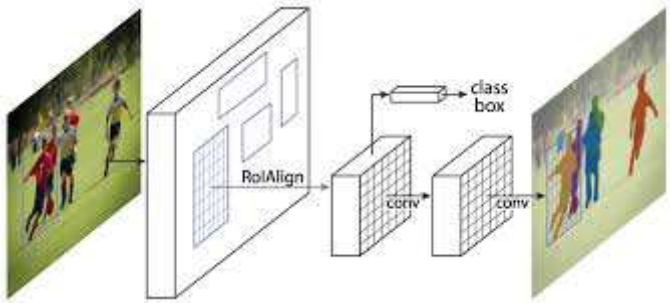


FIGURE 9.1: Mask R-CNN architecture of (He et al., 2017) [157]. A region proposal network isolates regions of interest that two parallel branches classify and segment.

Therefore, while U-Net appears as a good choice for the segmentation of echocardiographic images, the improvement of its performance through the extension of its architecture is not straightforward. In parallel to our experiments on deep supervision and implicit anatomical constraints described in Chapter 8, we observed an increasing interest in the computer vision community for methods based on attention mechanisms to improve accuracy on a wide range of image analysis tasks: classification [154], localization [155] [156] and segmentation [157].

9.2 Introduction

We first define the principle of attention in deep learning models, then present a few attention architectures that have been proposed for segmentation in medical imaging.

9.2.1 Definition of attention in deep learning frameworks

Attention mechanisms correspond to the integration of a contextualization procedure inside a supervised learning pipeline to improve its overall performance. Contextualization is usually applied either on the image itself or on a derived feature space, and consists in learning to focus both the information extraction and the processing on isolated parts of the image.

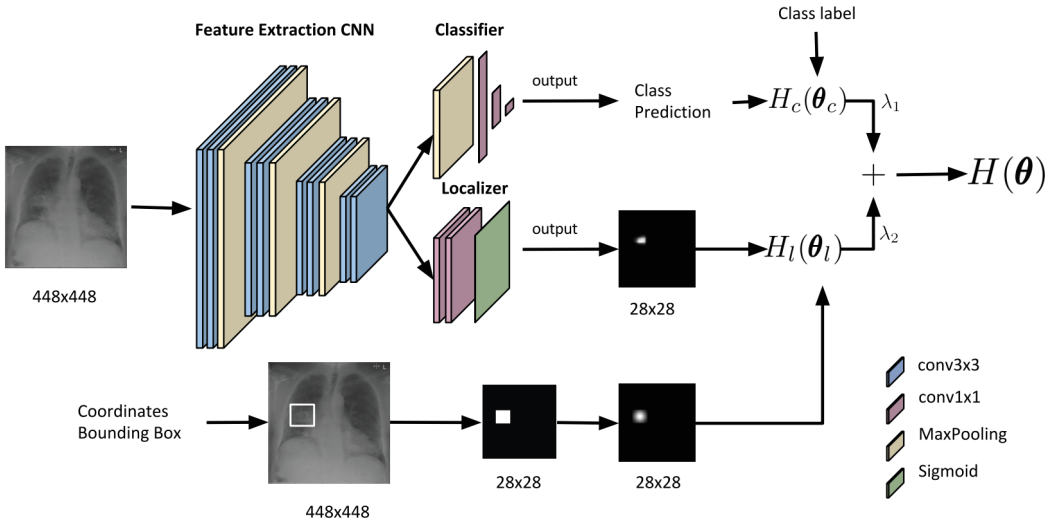
One of the best performing approach is the Mask R-CNN method that was recently proposed in (He et al., 2017) [157], and which provides the best current results in all three tracks of the COCO suite of challenges. This network is composed of three stages, as shown in Fig 9.1:

1. a region proposal network (RPN) which scans boxes distributed over the image area and finds the ones that contain objects;

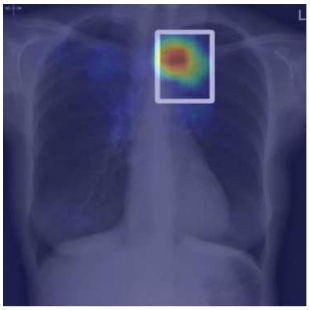
2. a classification network that scans each of the regions of interest (ROIs) proposed by the RPN and assigns them to different classes while refining the location and size of the bounding box to encapsulate the object;

3. a convolutional network that takes the regions selected by the ROIs classifier and generates masks for them.

The regularization effect brought by performing and jointly optimizing multiple tasks (localization, classification and segmentation) with a single network appeared a good lead to refine results obtained on the CAMUS dataset.



(a) Network architecture



(b) Produced localization

FIGURE 9.2: Pipeline and localization of lesion of (Pesce et al., 2019) [164].

9.2.2 Attention mechanisms in medical imaging

Attention-based approaches have been successfully applied in medical imaging throughout the last two years [158] [159] [160] [161].

In (Vigneault et al., 2018) [162], the authors proposed a dedicated CNN architecture for simultaneous localization and segmentation in cardiac magnetic resonance imaging. Their model was built around three stages: *i)* an initial segmentation is performed on the input image; *ii)* the features learned at the bottom layers are used to predict the parameters of a spatial transformer network [163] that transforms the input image into a canonical orientation; *iii)* a final segmentation is performed on the transformed image.

In (Pesce et al., 2019) [164], two attention networks were developed and compared for the detection of chest radiographs containing pulmonary lesions. The first solution involved the extraction of saliency maps from high level layers of the network, and compared the predicted position of a lesion with the true position as shown in Fig 9.2, while the second approach consisted in using a recurrent attention model which learns to process a short sequence of smaller image portions. The first approach involving heat map predictions (Fig 9.2) outperformed the later.

Recently, a generic attention model was proposed to automatically learn to focus on target structures in medical image analysis in (Schlemper et al., 2019 [165]). Based on attention gate modules that can be integrated in any existing CNN architecture introduced in (Oktay et al., 2018) [47], the proposed formalism intrinsically promotes the suppression of irrelevant regions in an input image while highlighting salient features useful for a specific task. This approach was validated for 2D fetal ultrasound image classification and 3D-CT abdominal image segmentation.

To our knowledge, no attention study prior to ours had been conducted for echocardiographic image segmentation. Therefore, we decided to investigate the capacity of attention mechanisms to improve the current best segmentation scores on the CAMUS dataset.

9.2.3 Potential of attention mechanisms on the CAMUS dataset

To investigate the interest of performing a localization task prior to the application of a U-Net segmentation model, we set up the following simulation:

1. we manually selected regions of interest (ROIs) around the reference segmentation mask for the endocardial region. Each ROI corresponds to the ideal bounding box (BB) with an additional margin m of 5, 15 or 30% of its size along the axes;

2. the ultrasound images and corresponding segmentation masks of the CAMUS dataset were cropped according to the bounding boxes to create new datasets on which we trained three U-Net 1 architectures, as described in Chapter 7, one for each margin.

The models are in the later referred to as BB-m5, BB-m15 and BB-m30, respectively. The scores are reported in Tab. 9.1 and reflect the segmentation results that a U-Net 1 would obtain with a perfect localization as prior.

From Tab. 9.1 in which best values are shown in bold, we can observe:

Table 9.1: Segmentation accuracy of U-Net 1 for a perfect prior localization of the endocardial region, restricted to patients having good and medium image quality (406 in total). m indicates the margin value.

Model	LV_{endo}			LV_{epi}			outliers
	D	***MAD***	***HD***	***D***	***MAD***	***HD***	**geo.**
	val.	mm	mm	val.	mm	mm	# %
intra-observer	0.937 ±0.027	1.4 ±0.5	4.5 ±1.8	0.954 ±0.020	1.7 ±0.8	5.0 ±2.2	21 13
U-Net 1	0.920 ±0.056	1.7 ±1.2	5.6 ±3.3	0.947 ±0.030	1.9 ±1.1	6.2 ±3.7	282 17%
BB-m5	**0.941** ±0.034	**1.3** ±0.6	**4.3** ±1.9	**0.971** ±0.011	**1.0** ±0.4	**4.1** ±1.8	**89** 5.5
BB-m15	0.940 ±0.034	**1.3** ±0.6	4.4 ±1.9	0.969 ±0.011	1.1 ±0.4	4.3 ±2.0	106 6.5
BB-m30	0.937 ±0.035	1.4 ±0.6	4.7 ±2.1	0.966 ±0.013	1.2 ±0.5	4.6 ±2.2	124 7.6

* LV_{endo}: *Endocardial contour of the left ventricle;*
* LV_{epi}: *Epicardial contour of the left ventricle*

- the strong contribution of the cropping stage, leading to a significant improvement of the baseline U-Net 1 results with average scores below the intra-observer variability (except for BB-m30 with the HD metric), and a number of outliers lower than 8%;
- the improvement is less pronounced for larger margins but still significant even for 30%, hinting that the network learns to rely on the localization and adapts to the margin;
- the MAD scores are much more improved for the epicardium (that is closer to the edges of the BB) than for the endocardium. HD values are similar.

This experiment revealed that the insertion of an accurate localization step prior to the segmentation of a U-Net 1 model could yield remarkable results.

9.3 Attention-learning architectures for 2D echocardiographic segmentation

Based on the motivations and the literature review on attention presented respectively in Section 9.1 and Section 9.2.3, we developed and compared two different approaches:

1. an attention-based method named RU-Net, for Refining U-Net;
2. a multi-task network named LU-Net, for Localization U-Net.

Since the U-Net 1 model produced high-performance segmentation results in echocardiography [5], we decided to use this architecture as back-bone for our two solutions. Moreover, we focused in this chapter on the joint segmentation of the left ventricle and the myocardium, since including the segmentation of the left atrium provided no significant improvement in overall results for the segmentation of the LV_{endo} and LV_{epi}. One other reason is that it allowed to study a smaller ROI to better observe the impact of attention in image processing.

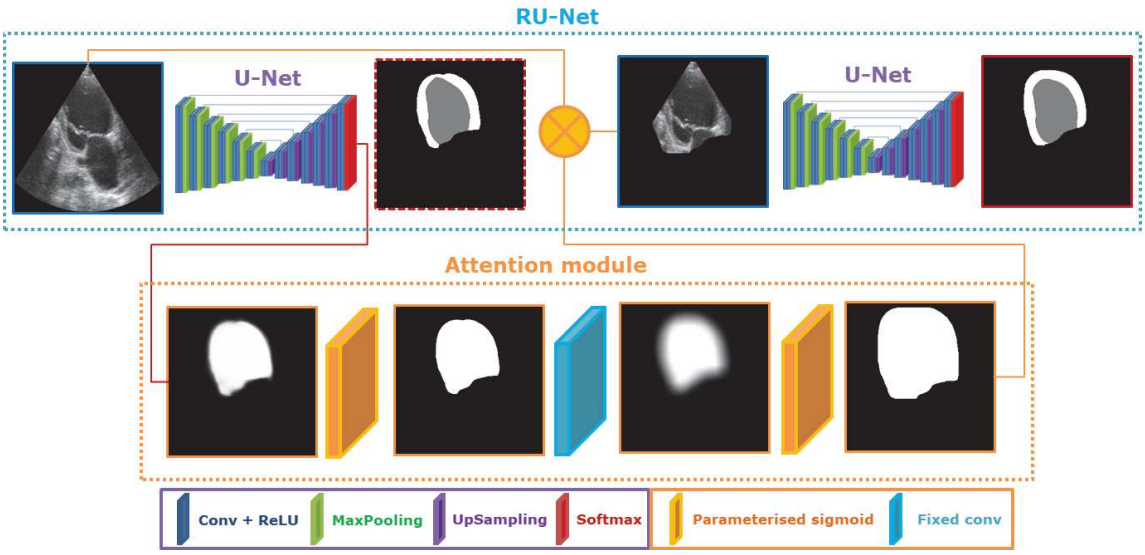


FIGURE 9.3: Illustration of the RU-Net pipeline. The two U-Nets are independent (different parameters).

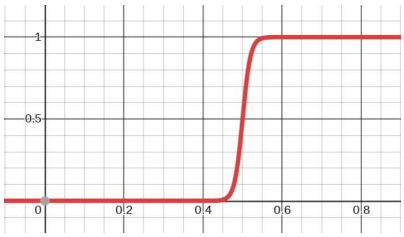


FIGURE 9.4: Parameterized sigmoid with a $slope = 100$ and $shift = 0.5$.

9.3.1 Refining U-Net

Refining U-Net (RU-Net) consists in two stacked U-Nets, the second refining the segmentation result of the first. In spirit, it is therefore similar to the SHG architecture introduced in Section 8.1.2, except that the first segmentation is used to perform attention in the ultrasound input image. The overall architecture is provided in Fig. 9.3, with the attention mechanism in yellow.

The union of the left ventricle and the myocardium segmentation provided by the first network is passed to an attention module that creates a dilated mask encompassing the structures of interest. The result of the attention module is then multiplied with the input image to create a contextualized image which is passed to a second U-Net 1.

The attention module is composed of two parameterized sigmoid functions applied before and after a convolution-based dilation layer, making the full process differentiable. The parameterized sigmoid functions have the following expression:

$$sig(x) = \frac{1}{1 + \exp\left(-slope \times (x - shift)\right)} \tag{9.1}$$

where the *slope* coefficient was set to 100 to have a quasi-instant transition (Fig. 9.4). The $shift$ parameter was left as a hyper-parameter for the first sigmoid while it was set to a fixed value of 0.01 for the second sigmoid to retrieve the energy lost through the dilation layer. RU-Net therefore has 2 hyper-parameters in addition to those involved in U-Net:

1. the dilation rate involved in the dilation layer, referred to as *dil* in the sequel;
2. the shift parameter of the first parameterized sigmoid function.

Since the attention module is differentiable, the full network is naturally trainable end-to-end. The overall loss of RU-Net corresponds to the sum of two identical multi-class dice functions computed at the output of each U-Net network. More details on the behavior of RU-Net can be found in Appendix G. RU-Net includes a total of 4M parameters ($2 \times 2M$).

9.3.2 Localization U-Net

The Localization U-Net (LU-Net) architecture aims at locating and segmenting the endocardial and the epicardial borders of the left ventricle in an end-to-end learning procedure. The difference with RU-Net is that LU-Net incorporates a standard localization step after the segmentation, with the prediction of the coordinates of a bounding box.

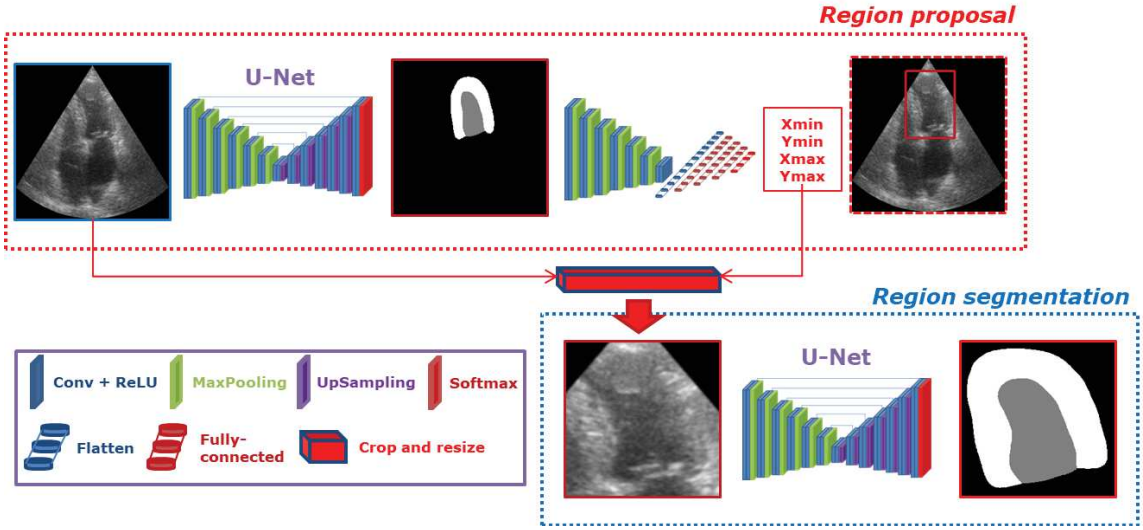


FIGURE 9.5: Illustration of the LU-Net pipeline with the U-Loc2-multi region proposal network, described in Section 9.5.1.
The two U-Nets are independent.

The underlying assumption of this strategy is that the joint optimization of localization and segmentation should lead to better segmentation results. LU-Net is therefore decomposed in two connected networks:

1. a region proposal network, performing localization;
2. a segmentation network, predicting the segmentation mask of the proposed ROI.

The overall architecture is illustrated in Fig. 9.5.

9.3.2.1 Region proposal

The region proposal network performs a mapping between the input ultrasound image and four coordinates that define a bounding box (BB) around the structure of interest, namely the union of the left ventricle and myocardium. The reference BB is defined as the minimal bounding box in contact with the epicardium border. The target coordinates are computed with an additional margin m as:

$$\begin{aligned} x^m_{min} &= x_{min} - m \times h \quad \text{and} \quad x^m_{max} = x_{max} + m \times h, \\ y^m_{min} &= y_{min} - m \times w \quad \text{and} \quad y^m_{max} = y_{max} + m \times w. \end{aligned} \tag{9.2}$$

where $(x_{min}, x_{max}, y_{min}, y_{max})$ are the relative coordinates of the reference BB (i.e. between 0 and 1) and (w, h) its width and height.

The motivation for adding a margin was to provide some context around the targeted structures for the segmentation task. Fig 9.5 shows the best performing localization method, that involved a first segmentation with a U-Net, an encoder and fully-connected regression layers. More details on the tested localization models are given in Section 9.4.

9.3.2.2 ROI segmentation

The output of the region proposal network is used as an attention mechanism to crop and resize the input ultrasound image. The resulting image is then fed to a segmentation network,

as illustrated in Fig 9.5. We use the U-Net 1 model described in Chapter 7 to learn and perform the segmentation, as it is the most efficient model evaluated on the CAMUS dataset considering a trade-off between accuracy, speed and size.

9.3.2.3 End-to-end approach

In order to make the full network trainable end-to-end, the crop and resize step was implemented using a bilinear differentiable sampling. In addition, the segmentation loss involved in the second U-Net was modified to evolve dynamically over the training phase with respect to the varying ROI. Please note the two U-Nets are independent networks, that have the same architecture but distinct parameters.

At inference time, we use the localization outputs to return the segmented ROI to the original coordinate system of the input image. The LU-Net architecture includes 13M parameters (11M for localization and 2M for segmentation).

9.4 Experiments

Since LU-Net performs a joint localization and segmentation of the structures of interest, we decided to compare the performance of state-of-the-art methods for these two tasks.

9.4.1 Evaluation metrics

We evaluated different methods on:

1. the localization task, to sort out the best region proposal network for LU-Net;
2. the segmentation task for attention methods;
3. the estimation of clinical indices.

9.4.1.1 Localization metrics

The performance of the localization procedure was assessed through the Intersection Over Union (IOU) metric (Section 4.1) and the euclidean distance errors between the predicted and reference BB coordinates (i.e. its central position (x_c, y_c), its height h and width w). In addition, we provided the "BB out" metric which corresponds to the number of cases where the predicted BB does not completely encompass the reference mask.

9.4.1.2 Segmentation metrics

To measure the accuracy of the segmentation output for the LV_{endo} and LV_{epi}, the same metrics as the ones used in Chapter 7 were used, i.e. the Dice similarity index, the mean absolute distance (MAD) and the Hausdorff Distance (HD). In addition, we assessed the quality of the segmentation with regards to cardiologists' annotations through the notions of outliers described in Chapter 7 and 8 (Section 8.4):

- geometric outlier: an image segmentation is seen as a geometric outlier if at least one of its eight corresponding distance scores (i.e. MAD and HD values) is out of the corresponding bounds defined from the inter-observer variability at ED and ES [5];

- anatomical outlier: an image segmentation is seen as an anatomical outlier if the simplicity and convexity [43] of the corresponding segmented contours are lower than the lowest values computed from expert annotations on 50 patients for at least one of the anatomical structures.

9.4.1.3 Clinical metrics

As for the experiments in Chapter 7 and 8, we evaluated the performance of the proposed methods with 3 clinical indices: *i)* the ED volume (LV_{EDV}, in ml); *ii)* the ES volume (LV_{ESV}, in ml) and *iii)* the ejection fraction (LV_{EF}, as a percentage), for which we computed two metrics: the correlation ($corr$) and the limit of agreement (LOA).

9.4.2 Methods

This study involved both localization and segmentation methods, that are presented here after. The learning strategy is briefly described in case of mono- and multi- tasking.

9.4.2.1 Localization methods

We implemented and assessed the performance of four different convolutional networks dedicated to the prediction of bounding boxes, i.e. predicting $(x^m_{min}, x^m_{max}, y^m_{min}, y^m_{max})$ values as defined in Section 9.3.2.1:

1. an AlexNet-like network [119], composed of a succession of convolutional layers of varying filter size and max pooling. Our version ends with three fully-connected layers of size 4096, 4096 and 4. Except for the additional last layer, this architecture is therefore the same as the original, without any dropout and data augmentation strategy, and includes 71M parameters;

2. a VGG19-like network [166], composed of 20 layers that alternate between convolutions and max pooling. The last fully-connected layers are made respectively of 4096, 4096 and 4 units, for a total of 70M parameters;

3. U-Loc1, based on a U-Net model performing the segmentation of the left ventricle and the myocardium. The bottom layer of this U-Net was used to carry out the localization procedure with a branch composed of four fully connected layers of 1024, 256, 32, and 4 units in parallel to the upsampling segmentation branch. This model was inspired by the work of (Vigneault et al., 2018) [162]. The network includes 9M parameters;

4. U-Loc2, also based on a U-Net model performing the segmentation of the left ventricle and the myocardium. The output of this U-Net was then connected to a downsampling branch similar to the encoder except it ended with four fully-connected layers of 1024, 256, 32 and 4 units. This multi-task approach and architecture is novel and corresponds to one of the innovations proposed in this study. We evaluated two versions of this network, one optimizing only the localization loss (referred to as U-Loc2-mono) and one optimizing both the localization and the segmentation losses (referred to as U-Loc2-multi). The network includes 11M parameters.

9.4.2.2 Segmentation methods

The performance of the joint segmentation of the endocardial and epicardial borders was assessed through the following four networks:

1. U-Net 1, corresponding to the best performing network on the CAMUS dataset when considering a trade-off between compacity and efficiency [5] (2M parameters);

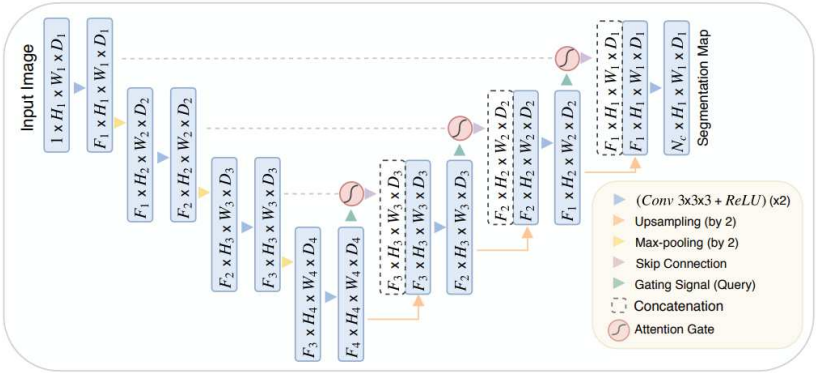
2. Attention-gated U-Net (AG-U-Net), recently proposed in (Oktay et al., 2018) [47], in which attention layers are used at each skip connection to locally weigh the concatenated features with coefficients established from the previous layer (Fig. 9.6). It includes batch normalization before each activation, and deep supervision by aggregating the feature maps produced after each attention layer at the last level of U-Net 1, i.e. before the last convolution and the softmax (2M parameters);

3. RU-Net, as introduced in Section 9.3, with a dilation of 30 pixels and a threshold at 0.7, which proved to be the best performing combination (4M parameters);

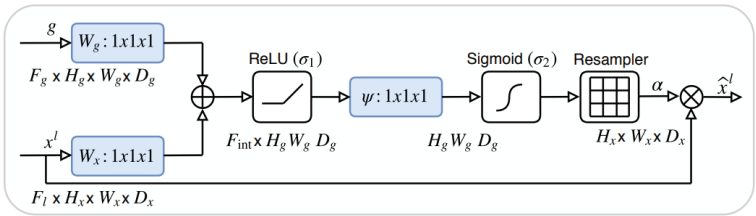
4. LU-Net, as introduced in Section 9.3.2, built using U-Loc2-multi as the region proposal network and U-Net 1 as the segmentation network. Two margins of $m = 5\%$ and $m = 15\%$ were evaluated (13M parameters).

9.4.2.3 Learning strategy

Optimizer All methods were optimized using Adam's optimizer with a learning rate (10^{-3} or to 10^{-4}) and a number of epochs (controlled using early stopping with a patience of 20) that experimentally allowed to observe a smooth convergence of the training and validation losses. The best model on the validation loss was selected after each training phase.



(a) AG-U-Net



(b) soft attention layer

FIGURE 9.6: Attention-gated U-Net (a) and soft attention layer (b) introduced in (Oktay et al., 2018). Attention layers are used in the upsampling branch to focus the skip connected features on regions of interest through the multiplication with attention maps built from the previous layer and the features to concatenate [47].

Loss Localization networks were optimized using a L1 loss clipped at 0.99 that sums the errors on the four relative BB values (i.e. $BB_{corr} = (x^m_{min}, x^m_{max}, y^m_{min}, y^m_{max})$):

$$L_1^{localization} = \frac{1}{4} \times \sum_{i=1}^{4} |gt_{corr}[i] - pred_{corr}[i]| \tag{9.3}$$

with gt_{corr} the targeted coordinates and $pred_{corr}$ the coordinates predicted by the model.

Segmentation networks were optimized using the multi-class Dice loss, taking into account the LV and myocardium predictions. For multi-task prediction, a weighting of 10 was applied to the localization term in order to balance the localization and the segmentation objectives.

9.5 Results

In order to easily compare the results with those provided in Chapter 7 and 8 on the CAMUS dataset, we followed the same strategy by training a single model for each deep learning method on the annotated images of both apical two and four-chamber views, regardless of the time instant.

9.5.1 Localization results

9.5.1.1 Influence of the architecture

Tab. 9.2 shows the localization accuracy computed on the full dataset (500 patients) for the algorithms described in section 9.4. Mean and standard deviation values for each metric were obtained from cross-validation on the 10 folds of the dataset. The values in bold correspond to the best scores for each metric. Comparing the methods using a margin of 5%, the proposed U-Loc2-multi got the overall best localization scores on all metrics (mean errors of 1.6, 1.9, 3.3 and 3.6 mm for x_c, y_c, h and w, respectively), except for the error on y_c with a difference of 0.2 mm with the best method.

The results highlight the interest of perfoming both segmentation and localization to improve the performance of the U-Loc2 method, with an average gain of 2.5 mm over the BB centre estimate and 3.6 mm over the BB dimension estimate. This significant improvement demonstrates that forcing segmentation as an intermediate step to localization is beneficial. Indeed, by performing multi-tasking, the U-Net architecture can produce localization results that outperform more established localization architectures (AlexNet, VGG) with a lot less parameters (11M instead of 70M).

9.5.1.2 Influence of the bounding box margin

We also investigated the influence of the choice of the margin value m on the accuracy of the localization results produced by the U-Loc2 method. Results were contrasted. Indeed, while the use of a lower margin (i.e. 5%) produced slightly better results with regards to the estimation of the BB position, the use of a higher margin (i.e. 15%) considerably reduced the number of cases where the BB did not encompass the reference mask (from 36% to 2%). Based on this experiment, it is clear that the U-Loc2-multi model produced the best localization results, but unclear which margin would be better in the end-to-end segmentation pipeline.

We therefore decided to use this network as the region proposal part of the LU-Net architecture, as illustrated in Fig. 9.3.2, and experiment with both margins, i.e. 5% and 15%.

TABLE 9.2: Localization accuracy on 4 evaluated methods on the full dataset (500 patients). The m information contained in each method name indicates the margin value defined in Section 9.3.2.1

Model	IOU	Error (mm)				BB out
		x_c	y_c	h	w	
AlexNet-m5	0.880 ±0.062	2.2 ±2.4	1.9 ±1.8	4.2 ±4.1	4.1 ±4.1	866 43%
VGG-m5	0.888 ±0.060	1.9 ±2.4	**1.7** ±1.7	4.0 ±3.9	4.0 ±3.7	903 45%
U-Loc1-m5	0.849 ±0.072	3.1 ±2.9	2.7 ±2.4	5.3 ±4.5	4.9 ±4.3	1094 55%
U-Loc2-mono-m5	0.791 ±0.138	4.2 ±4.7	4.4 ±6.0	7.1 ±6.4	6.9 ±6.7	1393 70%
U-Loc2-multi-m5	0.898 ±0.053	**1.6** ±1.8	1.9 ±1.9	**3.2** ±3.1	**3.6** ±3.2	712 36%
U-Loc2-multi-m15	**0.907** ±0.054	**1.6** ±2.0	**1.7** ±1.7	3.7 ±4.0	4.3 ±4.3	**31** 2%

9.5.2 Segmentation results

Tab. 9.3 displays the segmentation accuracy computed on the full CAMUS dataset for patients having good and medium image quality (406 patients) for the 4 algorithms described in Section 9.4.2.2. Mean and standard deviation values for each metric were obtained from cross-validation on the 10 folds of the dataset. The values in bold correspond to the best scores for each metric.

From these results, one can first see that all the attention-based networks produced either the same, or better results than the baseline U-Net 1. RU-Net model obtained similar results on the LV_{endo} and little improvement on the LV_{epi} (especially for the HD metric with an improvement of 0.4 mm) compared to U-Net 1. It also allowed a 2% reduction of the geometric outliers (5% on the full dataset). The limitation of RU-Net's improvement appears due to the second network strongly relying on the first segmentation and replicating errors. We did observe a true refinement effect in RU-Net however, the details being given in Appendix G.

The best performing methods were AG-U-Net and LU-Net. Indeed, AG-U-Net obtained the overall best results for the segmentation of the LV_{endo} border (MAD values of 1.5 mm and HD value of 5.3 mm), leading to segmentation scores close but still higher than the intra-observer variability for this structure. The LU-Net-m5 approach obtained the best results for the segmentation of the LV_{epi} border (MAD value of 1.5 mm and HD value of 5.1 mm) and the lowest number of geometric outliers (11%). Interestingly, these scores were either equivalent or lower than the intra-observer variability for this structure. It is also worth noting the robustness of the LU-Net model with respect to the choice of margin parameter, as margins of $m = 5\%$ and $m = 15\%$ produced almost the same segmentation scores for all metrics. An illustration of the segmentation performance of the LU-Net-m5 network compared to the baseline U-Net 1 model on three different cases is provided in Fig. 9.7.

TABLE 9.3: Segmentation accuracy on the 4 evaluated methods restricted to patients having good and medium image quality (406 in total).

Model	LV_{endo}			LV_{epi}			outl.
	D	***MAD***	***HD***	***D***	***MAD***	***HD***	**geo.**
	val.	mm	mm	val.	mm	mm	# %
intra-observer	0.937 ±0.027	1.4 ±0.5	4.5 ±1.8	0.954 ±0.020	1.7 ±0.8	5.0 ±2.2	21 13
U-Net 1	0.920 ±0.056	1.7 ±1.2	5.6 ±3.3	0.947 ±0.030	1.9 ±1.1	6.2 ±3.7	282 17%
RU-Net	0.925 ±0.049	1.7 ±1.0	5.4 ±3.3	**0.950** ±0.030	1.8 ±1.1	5.8 ±3.9	240 15%
AG-U-Net [47]	0.930 ±0.049	**1.5** ±1.3	**5.3** ±3.4	**0.950** ±0.026	1.8 ±1.0	5.9 ±3.7	270 17%
LU-Net-m5	**0.953** ±0.026	1.7 ±0.9	5.5 ±3.6	0.932 ±0.043	**1.5** ±0.8	**5.1** ±3.3	**186** 11%
LU-Net-m15	0.952 ±0.029	1.7 ±1.1	5.6 ±4.0	0.931 ±0.049	**1.5** ±1.1	5.3 ±3.6	203 12%

9.5.3 Clinical scores

Tab. 9.4 contains the clinical metrics computed on the full dataset for patients having good and medium image quality (406 patients) and for the 4 methods described in Section 9.4.2.2. The values in bold represent the best scores. The AG-U-Net and LU-Net-m5 models obtained the best scores for all clinical metrics (bias was not taken into account since the lowest bias value in itself does not necessarily mean the best performing method).

Regarding the estimation of the LV_{EDV}, the two methods produced high correlation scores (0.956), small biases (±1.4 ml) and reasonable limits of agreements (around 22 ml) and mean absolute errors (around 8 ml). The AG-U-Net produced the best LV_{ESV} results with a correlation of 0.962, while the LU-Net-m5 model produced the best LV_{EF} scores with a correlation of 0.829. However, even if the scores of LU-Net-m5 and AG-U-Net were slightly better than the ones of the baseline U-Net 1, they were still higher that the intra-observer results. This revealed there is still room for improvement, as discussed in Section 9.6.

9.5.4 LU-Net behavior

From the results given in Tab. 9.3 and Tab. 9.4, it appears that the LU-Net method outperforms the baseline U-Net 1 model both in terms of segmentation and clinical indice estimation. Furthermore, it is one of the most effective model, even compared to other attention-based networks. In order to complete the analysis of LU-Net, we applied this network to the full dataset (including poor image quality) and studied the generated outliers.

The corresponding results obtained with a margin of $m = 5\%$ are provided in Tab. 9.5. The results of the model named LU-Net-m5-o1 corresponds to the scores derived from the output of the first U-Net involved in the region proposal network, while the scores of the model named LU-Net-m5-o2 corresponds to the scores derived from the final output of the network (i.e. the one provided by the second U-Net).

TABLE 9.4: Clinical metrics of the 4 evaluated methods restricted to patients having good and medium image quality (406 in total)

Model	LV_{EDV}			LV_{ESV}			LV_{EF}		
	corr	***LOA***	***mae***	***corr***	***LOA***	***mae***	***corr***	***LOA***	***mae***
	val.	ml	ml	val.	ml	ml	val.	%	%
intra-observer	0.978	-2.8±14.3	6.2	0.981	-0.1±11.4	4.5	0.896	-2.3±11.2	4.5
U-Net 1	0.947	-8.3±24.7	10.9	0.955	-4.9±19.4	8.2	0.791	-0.5±15.1	5.6
RU-Net	0.946	-1.2±23.9	8.9	0.949	0.3±19.6	7.3	0.704	-2.1±14.3	6.0
AG-U-Net	**0.956**	-1.4±21.9	**8.1**	**0.962**	0.6±17.0	**6.2**	0.798	-2.2±15.1	5.5
LU-Net-m5	**0.956**	1.4 ±21.8	8.3	0.956	1.6± 18.0	7.0	**0.829**	-1.5±13.5	**5.0**
LU-Net-m15	0.952	2.4 ±22.9	**8.1**	**0.962**	1.8±16.7	6.5	0.821	-1.2±13.7	**5.0**

corr: Pearson correlation coefficient; LOA: limit of agreement; mae: mean absolute error. The values in bold refer to the best performance for each measure.

9.5.4.1 Stability of the results

From this table, one can see that LU-Net outperformed the U-Net 1 architecture for all the metrics for both LV_{endo} and LV_{epi} borders, this when considering all quality of images. The segmentation results produced by LU-Net also appeared to be remarkably stable when integrating poor image quality images, with a mean difference only of 0.1 mm for MAD, 0.2 mm for HD and 1% for the geometric outliers.

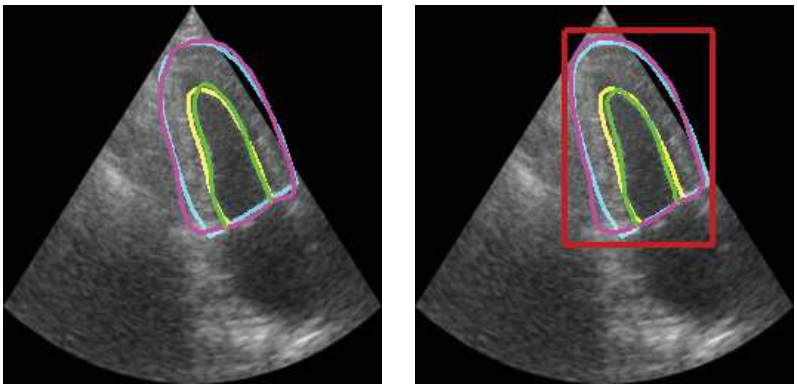
9.5.4.2 Localization

Concerning the localization scores, the LU-Net-m5 model obtained consistent results with respect to the U-Loc2-multi best performing method reported in Tab. 9.2 with an IOU of 0.906 and BB errors on (x_c, y_c, h, w) of $(1.5, 1.5, 3.3, 3.5)$ mm, respectively. Coupling this result with the last two lines of Tab. 9.5 shows that the first segmentation was less accurate than a single U-Net 1.

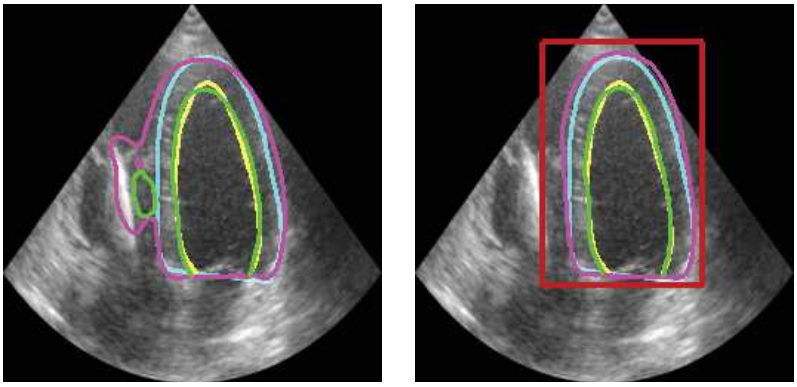
Therefore, it appears that the segmentation result produced in the region proposal part of the LU-Net was slightly degraded when optimizing the localization procedure, which in turn allowed for a significant improvement of the final segmentation results. Interestingly, we obtained the best scores when the three losses were balanced, rather than giving a stronger weight to either the localization or the last segmentation, which supposes that the robustness is improved through balanced multi-tasking rather than privileging the final objective.

9.5.4.3 Segmentation refinement

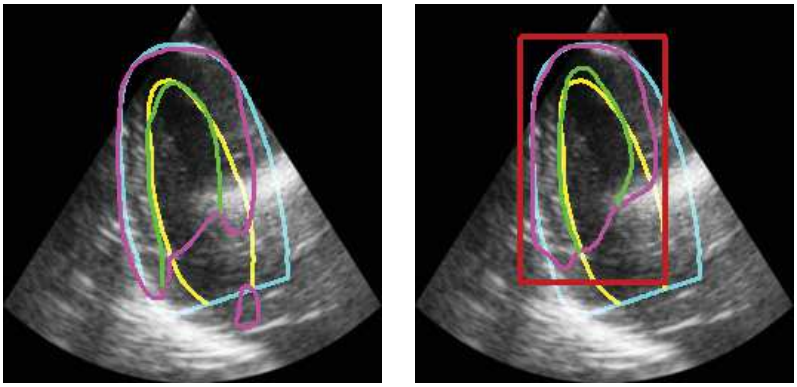
Concerning the segmentation scores, LU-Net-m5 produced 12% of geometric outliers, 2% of anatomical outliers and 1% of both, showing that half of the anatomical outliers were also geometric. Moreover, the geometric outlier rate was lower than the intra-observer variability one computed from a subset of 40 patients with good and medium image quality, which further highlights the quality of the results achieved by this method.



(a) case with similar results



(b) case in which the intermediate localization helps



(c) case in which a strong artifact hinders any improvement

FIGURE 9.7: Comparison of the segmentation performance of the baseline U-Net 1 (left column) and the proposed LU-Net architecture (right column). In each image, the prediction is in green and purple while the ground-truth is in yellow and cyan. The BB estimated is displayed in red.

9.6 Discussion

9.6.1 Attention-based networks

Tab. 9.3 and Tab. 9.4 underline the ability of attention-based networks to improve the segmentation and the estimation of clinical indices in 2D echocardiography. These results are even more interesting given that we had failed in improving the scores of the baseline U-Net 1 model through more sophisticated architectures, as detailed in Chapter 8. Although

TABLE 9.5: Segmentation accuracy and outliers on the full dataset (500 patients) including those with poor image quality

Model	LV_{endo}		LV_{epi}		outliers		
	MAD	***HD***	***MAD***	***HD***	**geo.**	**ana.**	**both**
	mm	mm	mm	mm		# %	
U-Net 1	2.0 ±1.2	6.1 ±3.9	2.0 ±1.1	6.5 ±4.5	423 21%	95 5%	71 4%
LU-Net-m5-o1	2.1 ±1.1	7.0 ±4.7	1.9 ±1.0	6.2 ±3.4	483 24%	201 10%	138 7%
LU-Net-m5-o2	**1.8** ±1.0	**5.7** ±3.6	**1.6** ±0.9	**5.3** ±3.3	**240** 12%	**31** 2%	**20** 1%

AG-U-Net produced the best scores on the LV_{endo} and the estimation of the LV_{ESV}, LU-Net provided the best trade-off between the achieved improvements on all metrics and the decrease of the number of geometric outliers.

9.6.2 Comparison with intra-observer variability

As for the segmentation scores, LU-Net managed to reach the intra-observer variability for the LV_{epi} border (MAD and almost HD metrics). The number of geometric outliers produced by this method (i.e. 11%) was also below the intra-observer score. To the best of our knowledge, this is the first time that such result is obtained in the context of 2D echocardiographic image segmentation. Unfortunately, the scores obtained by our proposed model remained higher than the intra-observer variability for the LV_{endo} border.

Concerning the estimation of clinical metrics, although LU-Net improved the results compared to the baseline U-Net, its scores remained slightly higher than the intra-observer variability. This reveals that while attention-based networks clearly enhanced the results produced by the baseline U-Net 1 model, there still exists room for improvement to faithfully reproduce the manual annotations of one expert.

9.6.3 Areas for improvement

We identified two leads of potential improvement to allow competitive results with respect to the intra-observer variability. First, based on Tab. 9.2, it appears that the localization step can be further optimized to improve the LU-Net scores, as suggested by the results on ideal cases provided in Tab. 9.1.

Secondly, there is a need to introduce temporal coherency into deep learning architectures. Indeed, while the current strategy (i.e. ED and ES are treated independently) provides high correlations for the estimation of the LV_{EDV} and LV_{ESV} (0.956 for both indices), the estimation of the LV_{EF} is degraded to 0.829 . This reveals a lack of temporal consistency of the LU-Net segmentation results between ED and ES.

9.7 Conclusion

In this chapter, we introduced two novel attention-based methods to improve the robustness of the segmentation of the endocardium and epicardium in 2D echocardiography. We showed that the joint optimization of the localization and the segmentation tasks of the LU-Net model led to better segmentation results at the end of the process. This method:

1. outperformed U-Net 1, the current best performing deep learning solution on the CAMUS dataset;
2. produced among the best results from the tested attention-based networks;
3. produced overall segmentation scores less than or equal to the intra-observer variability for the epicardial border with 11% of outliers;
4. closely reproduced the expert analysis for the end-diastolic and end-systolic left ventricular volumes, with a mean correlation of 0.96;
5. improved the estimation of the ejection fraction of the left ventricle, with scores that remained slightly worse than the intra-observer's ones.

Though the intra-variability remains to be reached for a set of metrics, this work established localization as a lead for more robust 2D echocardiographic image analysis with a deep learning approach.

Part IV

Epilogue

Chapter 10

Conclusion

This last chapter summarizes the key contributions and conclusions drawn from the analysis of the main results gathered in this manuscript. Leads of improvement and perspectives are then provided to open the door to further investigation.

10.1 Key contributions

This thesis supported three key contributions for the community:

1. We built and provided the largest clinical dataset for 2D echocardiography image analysis, with annotations of three structures (left ventricle, myocardium, left atrium) along with a public evaluation platform [1];

2. We demonstrated that supervised learning, in particular deep learning methods, are currently the most promising approaches to provide fully-automatic, accurate and fast segmentation solutions in 2D echocardiography [5];

3. We showed the potential of attention-learning mechanisms to improve the robustness and accuracy of segmentation results predicted by deep learning models [46].

10.2 Conclusions

10.2.1 Methodological aspects

A complete study was conducted in this PhD in terms of methodology: i) an adaptation of the structured random forest algorithm to 2D and 3D echocardiography was investigated [30] [32]; ii) a thorough analysis on the application of the U-Net architecture to the CAMUS dataset was realized [31] [5]; iii) the evaluation was extended to anatomical validation [48]; iv) more robust and accurate encoder-decoder architectures based on attention-learning mechanisms were developed [45] [46].

This allows us to answer our first objective: the assessment of supervised learning methods for the semantic segmentation in 2D echocardiography. Our results showed that deep learning methods, through encoder-decoder architectures, produced highly accurate results in between the inter- and intra- variability and were robust to image quality and, to a certain extent, to acquisition settings.

10.2.2 Clinical aspects

We established that deep learning technics obtained the best clinical scores on the CAMUS dataset with accurate estimations of the LV value at ED and ES (correlations around 0.95

and mean average error around 8 ml). Results are more contrasted with the estimation of the LV_{EF} with correlations around 0.80 and mean average error around 5%, which emphasis the need to introduce temporal coherency into the segmentation framework.

This provides answers to our second objective: evaluate how close we are to fully automatize cardiac analysis in ultrasound imaging. Our results show that our best performing fully-automatic methods obtained results below the inter-observer variability for all metrics, which indicates the high potential of deep learning methods for the analysis of 2D echocardiography images. In addition, CNNs allow fast inference while demanding reasonable storage memory, making these solutions even more attractive for clinical applications.

10.3 Perspectives

10.3.1 Short-term perspectives

10.3.1.1 Algorithm perspectives

My thesis provided a proof of concept that ultrasound image analysis could be successfully addressed with supervised learning methods. Still, the overall study suffers from some limitations that would ideally have to be investigated in follow-up studies:

- improvement of the robustness of multi-structure segmentation: the results presented in Chapter 9 show that the robustness of the segmentation could still be increased for both the endocardium and the epicardium. It is our belief that a solution reaching the average intra-variability on both structures could be reached by investigating attention learning mechanisms further;

- assessment of the potential of deep learning technics in 3D echocardiography: due to the small size of the only publicly available CETUS dataset [16], the potential of deep learning methods to analyze 3D images could not be thoroughly assessed. From the conclusions of our study in 2D, it would be interesting to adapt the LU-Net algorithm [46] to 3D volumes. Depending on the memory consumption, investigating multi-slice processing or 2.5D approaches [49] might be preferable;

- time-consistent predictions: the regularization of the evolution of the segmentation through the cardiac cycle has been established as a requirement to obtain robust ejection fraction estimations. Furthermore, some difficult to solve cases may benefit from temporal constraints. It is thus essential to introduce time coherency into our deep learning solutions;

- anatomically plausible shapes guarantee: by designing anatomical metrics based on geometrical shapes, we showed that CNNs learn a mapping that already encourages a certain validity of the shapes, however without any guarantee. It would thus be of interest to study multi-atlas registation [50] or latent space learning and correction [51] to ensure the prediction of anatomically plausible shapes in all situation and at all time.

- realistic simulations: we observed on the CAMUS dataset that some outliers would be very hard to solve as they correspond to very unusual cases produced by atypical acquisitions (artifacts, zoom, orientation, very low contrast...). One way to solve this problem consists in adding these missing variations to our dataset through virtual realistic images with controlled parameters to generate a wide range of cases;

10.3.1.2 Clinical perspectives

The present study also revealed that in order to apply deep learning solutions in clinical routine, the following needs have to be addressed:

1. a consensual large scale public dataset: the CAMUS dataset already comprises apical chamber views from 500 patients for a total of 1000 sequences with 2000 annotated images of various image quality. However, it still covers a small part of existing populations, pathologies, vendors, and experts variabilities. The number of views and structures involved is also limited. This is why we refer to this study as a proof of concept which should motivate the community to establish a more complete annotated dataset of sufficient size and variety, and with annotations that have been carefully validated. This would allow to precisely evaluate the inter- and intra-variability in 2D echocardiography image analysis, and to exploit the proven potential of supervised learning methods in reproducing expert contouring;

2. clinical validation: while we tried to perform thorough evaluation of supervised learning methods through the combination of cross-validation and complementary metrics, the clinical validation remains limited and needs to be further extended to more cases with targeted pathologies;

3. guidance softwares: we showed the strong potential of supervised learning methods for the automatic analysis of echocardiographic images, but there remains the need to automatize the acquisition procedure. Indeed, guiding the practicians in adequately positioning the probe appears to be necessary to obtain the right visuals of the heart for the segmentation algorithm to produce accurate results [52]. To do so, the design of algorithms that learn to recognize whether the acquisition corresponds to a suitable context, for instance by computing uncertainty, appears to be a solution of choice [53].

10.3.2 Long-term perspectives

10.3.2.1 Clinical perspectives

This study is part of a revolution in medical imaging processing, which calls for closer interactions between researchers, developers, and clinicians to potentially bring the following advancements:

1. learning methods to assist doctors in their daily examinations would strongly improve their workflow;

2. automatic segmentation learnt from a consensual dataset would not only allow to generate reproducible results, but would also represent the standard of the right segmentation, and possibly be used to teach new generations of doctors;

3. robust image analysis softwares would enable to measure the potential of new modalities compared to previous ones, and open the door to new practices. For instance, coupling robust supervised learning with ultrasound imaging would enable to treat more patients, including those in remote places (tele-medicine), and favor early diagnosis by rendering acquisitions by non-professionals.

10.3.2.2 Algorithm perspectives

However, while deep learning methods have been shown to easily be compressed and fastened, the training phase remains a strong limitation, costly in multiple resources (time, electricity, processors, pollution...). For this study alone, several hundreds of models have been trained, because of changes in training datasets and hyper-parameters values. Therefore, there soon may be a strong need for models able not to learn from scratch (transfer learning follows this direction but for the purpose of compensating small datasets rather than shortening the training phase), as well as a need for models that always converge to a fixed optimal solution (no stochasticity). Both items require further research in model optimization and in parameter search.

In a close future, medical datasets collected to design supervised learning solutions may grow exponentially and continually. It is thus of interest to anticipate on the storage needs by developing dataset compression algorithms adapted to medical applications, that would sum up a large amount of data in a fraction of cases for similar variability and performance of learning algorithms. These two leads might be solved through theoretical innivations in information theory of deep learning.

Part V

Appendices

Appendix A

Collaborators

This section briefly introduces the many collaborators of this study, which I was very lucky to work with.

A.1 Creatis - INSA Lyon

(a) Sarah Leclerc - the author (b) Carole Lartizien - PhD director (c) Olivier Bernard - co-director (d) Thomas Grenier - co-supervisor (e) Frédéric Cervenansky - research engineer

FIGURE A.1: Collaborators from the Creatis laboratory: myself (left), my supervisors (center), and the research engineer for the CAMUS platform (right).

Sarah Leclerc I graduated in 2016 from INSA Lyon with an electrical engineering degree and an image processing master degree. During my thesis, from 2016 to 2019, I worked on the application of supervised learning methods for the analysis of echocardiographic images, and taught computer science at the electrical engineering department of INSA Lyon. My research interests focus on medical imaging-oriented machine learning applications.

Carole Lartizien Carole is a Research Director of the CNRS specialized in the identification of major health issues that can be addressed by imaging, and of theoretical barriers in biomedical imaging related to signal and image processing, modelization and numerical simulation. Her research interests include machine learning methods for classification problems and prototyping of computer aided diagnostic system for cancer and neuro-imaging.

Olivier Bernard Olivier is associate professor at INSA Lyon and the new director of the team MYRIAD of the Creatis laboratory. His current research interests include medical image analysis with a particular attention to cardiac imaging. He has a strong interest in machine learning, image segmentation, motion analysis, statistical modeling, sampling theories and image reconstruction and simulation.

Thomas Grenier Thomas Grenier is associate professor at INSA Lyon, strongly involved in the Electric Engineering department. His research focuses on longitudinal analysis of medical data to study evolution as multiple sclerosis lesions and functional activity (muscle and hydrocephaly), which involves segmentation tasks, dedicated pre- and post- processing steps, clustering, semi-supervised or fully supervised schemes with specific constraints.

Frederic Cervenansky Frederic is a research engineer at the Creatis laboratory of INSA Lyon, in which he is strongly involved in computer development and project support. He is in charge of the computer development department and one of the main investigator of the Virtual Imaging Platform, a web portal for medical simulation and image data analysis.

A.2 VITAL - Sherbrooke University

FIGURE A.2: Pierre-Marc Jodoin - co-supervisor

Pierre-Marc Jodoin Pierre-Marc is a full professor at the computer science department of the University of Sherbrooke. He is the head of the VITAL laboratory which focuses on researching state of the art algorithms in image processing, computer vision, video analytics, medical imaging and statistical models, by investigating novel machine learning solutions (both supervised and unsupervised) including deep learning models.

A.3 CIUS - NTNU

(a) Erik Smistad - post doc (b) Andreas Østvik - PhD student (c) Lasse Lovstakken - researcher (d) Torvald Espeland - cardiologist (e) Erik Andreas Rye Berg - cardiologist

FIGURE A.3: Collaborators from NTNU on the CAMUS study and attention-learning study: scientific researchers (left) and clinical cardiologists (right).

Erik Smistad Erik is a research scientist at SINTEF Medical Technology and has a post doctoral position at the Norwegian University of Science and Technology. His research is focused on the design of programs to automatically and quickly locate anatomical structures in medical images in order to help physicians interpret images and navigate inside the body during surgery (segmentation, tracking, object detection and classification, GPU processing).

Andreas Østvik Andreas is a PhD candidate at the Department of Circulation and Medical Imaging at NTNU. His current research projects involve automatic methods based on machine learning for analyzing medical images. The main application is cardiac ultrasound imaging, where the goal is to develop good diagnostic tools and improve workflow in the clinic. He is also interested in robotics and machine vision for medical intervention.

Lasse Lovstakken Lasse is professor at the Department of Circulation and Medical Imaging of the Norwegian University of Science and Technology. His research interests are targeted towards medical ultrasound imaging, including segmentation, reconstruction, flow analysis, cluttering (...) He is currently strongly involved in the development and evaluation of 3D ultrasound imaging of blood flow in the hearts in pediatric and adult cardiology.

Torvald Espeland Torvald is a PhD student in clinical cardiology at the Department of Circulation and Medical Imaging of the Norwegian University of Science and Technology who practices at St. Olavs Hospital. His research focuses on the investigation of aortic stenosis, valvular heart disease and myocardial fibrosis with the use of echocardiography, including doppler echocardiography, and machine learning.

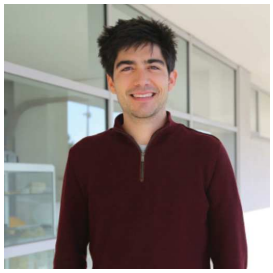
Erik Andreas Rye Berg Erik Andreas is a PhD student in clinical cardiology at the Department of Circulation and Medical Imaging of the Norwegian University of Science and Technology who practices at St. Olavs Hospital. A strong component of his research is 3D doppler echocardiography, with angle correction and flow estimation for aortic stenosis severity assessment.

A.4 CHU Saint-Etienne

FIGURE A.4: Florian Espinosa - cardiologist

Florian Espinosa Florian is a PhD student at the Department of Biology of Saint-Etienne university, and a cardiologist specialized in cardio-vascular diseases in adult cardiology. After practicing at the University Hospital of Saint-Etienne, he now practices in the cardiological intensive care unit at the Loire Private Hospital and the specialist clinic of Cardio Europa.

A.5 K.U Leuven



(a) João Pedrosa
- PhD student

(b) Jan D'hooge
- researcher

FIGURE A.5: Collaborators from KU Leuven on the CAMUS study.

João Pedrosa João recently finished his PhD in Biomedical Sciences at KU Leuven on 3D echocardiography segmentation. He now has a postdoctoral position at INESC TEC about lung nodule detection, segmentation and characterization on chest CT. His research interests are medical imaging acquisition and processing, machine learning and applied research for improved patient care.

Jan D'hooge Jan is associate professor at the Cardiovascular Imaging and Dynamics of the Department of Cardiovascular Sciences at KU Leuven. Has has worked on several medical imaging problems, such as elastic registration, segmentation, shape analysis, and data acquisition. His current research interests include myocardial tissue characterization, deformation imaging, and cardiac pathophysiology.

A.6 Erasmus

(a) Feriel Khellaf
- master student

(b) Jason Voorneveld
- PhD student

(c) Johan G. Bosch
- researcher

FIGURE A.6: Collaborators from Erasmus MC on the SRF study.

Feriel Khellaf Feriel received her biomedical engineering degree from Polytech Lyon in 2017, as well as a master in medical imaging signals and systems. After a master internship at Erasmus MC, she is now doing a PhD at the Creatis laboratory. Her work focuses on Proton Computed Tomography and is aimed at improving tumor targeting.

Jason Voorneveld Jason recently received his PhD from the Biomedical Engineering department of Erasmus MC, which aimed at developing and validating software and acquisition protocols suitable for quantifying blood flow in the left ventricle. He is now doing a post doctoral study at Erasmus MC. His research interests are ultrasound signal and image processing, and flow quantification and visualization.

Johan G. Bosch Hans is associate professor at the department of biomedical engineering at Erasmus Medical Center. His current research interests include 2-D and 3-D echocardiographic image processing and analysis as well as transducer development, with cardiac imaging one of the primary fields of investigation.

Appendix B

List of publications

Journal paper

1. S. Leclerc, E. Smistad, A. Østvik, F. Cervenansky, F. Espinosa, T. Espeland, E. A. R. Berg, T. Grenier, C. Lartizien, P.-M. Jodoin, L. Lovstakken, and O. Bernard, "Lu-net: A multi-task network to improve the robustness of deep learning segmentation in 2d echocardiography", [Submitted to TUFFC, special issue on Deep learning in medical ultrasound from image formation to image analysis]

2. E. Smistad, A. Østvik, I. Salte, D. Melichova, T. M. Nguyen, H. Brunvand, T. Edvardsen, S. Leclerc, O. Bernard, B. Grenne, and L. Lovstakken, "Real-time automatic ejection fraction and foreshortening detection using deep learning", [Submitted to TUFFC, special issue on Deep learning in medical ultrasound from image formation to image analysis]

3. S. Leclerc, E. Smistad, J. Pedrosa, A. Ostvik, F. Cervenansky, F. Espinosa, T. Espeland, E. A. R. Berg, P. Jodoin, T. Grenier, C. Lartizien, J. Dhooge, L. Lovstakken, and O. Bernard, "Deep learning for segmentation using an open large-scale dataset in 2d echocardiography", *IEEE Transactions on Medical Imaging*, pp. 1–12, 2019, ISSN: 0278-0062. DOI: doi:10.1109/TMI.2019.2900516

International Conference Paper

1. S. Leclerc, E. Smistad, T. Grenier, C. Lartizien, A. Østvik, F. Cervenansky, F. Espinosa, T. Espeland, E. A. R. Berg, P. Jodoin, L. Lovstakken, and O. Bernard, "Ru-net: A refining segmentation network for 2d echocardiography", in *2019 IEEE International Ultrasonics Symposium (IUS)*, 2019, pp. 1–4

2. S. Leclerc, E. Smistad, A. Ostvik, F. Cervenansky, F. Espinosa, T. Espeland, E. A. R. Berg, P.-M. Jodoin, T. Grenier, C. Lartizien, L. Lovstakken, and O. Bernard, "Deep learning segmentation in 2d echocardiography using the camus dataset : Automatic assessment of the anatomical shape validity", in *International Conference on Medical Imaging with Deep Learning – Extended Abstract Track*, London, United Kingdom, 2019

3. H.-T. Nguyen, P. Croisille, M. Viallon, S. Leclerc, S. Grange, R. Grange, O. Bernard, and T. Grenier, "Robustly segmenting quadriceps muscles of ultra-endurance athletes with weakly supervised u-net", in *International Conference on Medical Imaging with Deep Learning – Extended Abstract Track*, London, United Kingdom, 2019

4. S. Leclerc, E. Smistad, T. Grenier, C. Lartizien, A. Ostvik, F. Espinosa, P. Jodoin, L. Lovstakken, and O. Bernard, "Deep learning applied to multi-structure segmentation

in 2d echocardiography: A preliminary investigation of the required database size", in *2018 IEEE International Ultrasonics Symposium (IUS)*, 2018, pp. 1–4. DOI: doi:10.1109/ULTSYM.2018.8580136

5. A. MeidellFiorito, A. Østvik, E. Smistad, S. Leclerc, O. Bernard, and L. Lovstakken, "Detection of cardiac events in echocardiography using 3d convolutional recurrent neural networks", in *2018 IEEE International Ultrasonics Symposium (IUS)*, 2018, pp. 1–4. DOI: 10.1109/ULTSYM.2018.8580137

6. E. Smistad, A. Østvik, I. M. Salte, S. Leclerc, O. Bernard, and L. Lovstakken, "Fully automatic real-time ejection fraction and mapse measurements in 2d echocardiography using deep neural networks", *2018 IEEE International Ultrasonics Symposium (IUS)*, pp. 1–4, 2018

7. F. Khellaf, S. Leclerc, J. D. Voorneveld, R. S. Bandaru, J. G. Bosch, and O. Bernard, *Left ventricle segmentation in 3d ultrasound by combining structured random forests with active shape models*, 2018. DOI: doi:10.1117/12.2293544

8. S. Leclerc, T. Grenier, F. Espinosa, and O. Bernard, "A fully automatic and multi-structural segmentation of the left ventricle and the myocardium on highly heterogeneous 2d echocardiographic data", in *2017 IEEE International Ultrasonics Symposium (IUS)*, 2017, pp. 1–4. DOI: doi:10.1109/ULTSYM.2017.8092632

Appendix C

Material

We here give a few information on the hardware that we used to perform our experiments, focusing on details relevant to reproduce the results.

C.1 Saki

Saki is the DELL machine that was used to train and evaluate all the neural networks whose results are reported in this study. It was first equipped with 2 GPUs NVidia Tesla M60, each with a RAM of 8 GB. These GPUs were used to train and evaluate the models in Chapters 7 and 8. Saki was later equipped with 2 GPUs GTX 1080, each with a RAM of 16 GB, which allowed in particular to train the more complex models presented in Chapter 9. Saki also includes 2 CPUs Xeon E5, each of 10 cores with 2 threads to access a RAM of 512 GB. CPUs were used for the deep learning studies to compute the scores from the results of the test phase.

FIGURE C.1: Saki, indicated with the red arrow.

C.2 Houlock

Prior to Saki, we used a machine called Houlock that possesses 2 CPUs Xeon E5 of 12 cores (1 thread) each. Houlock was used to produce the results presented in Chapter 6. The 512 GB RAM allowed to train multiple random forests in parallel, and to deal with the increase in memory usage brought from working with 3D patches instead of 2D.

Appendix D

Supplementary information on the CAMUS dataset

D.1 Population characteristics

In the metadata of the dataset, the precise pathology is not known, but other statistics (age, sex, height, ejection fraction, image quality ...) allow the assessment of the population variety (Tab. D.1). For the age, height and weight, we chose thresholds based on the range and the distributions of values to represent three folds with at least two balanced:

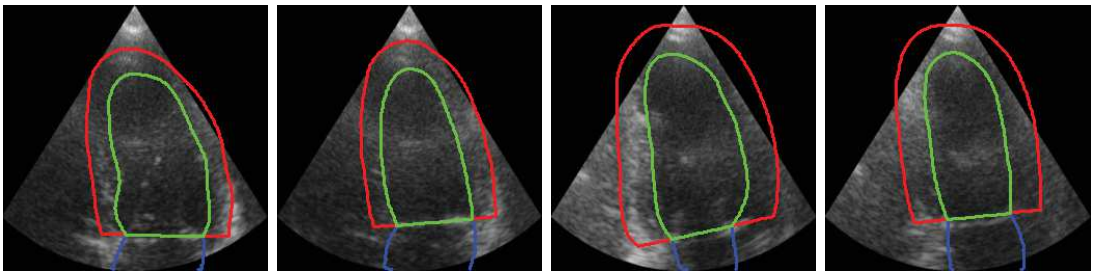
TABLE D.1: Population traits of the dataset

Information	**Value**	%
Sex	M	66
	F	34
Image quality	Good	35
	Medium	46
	Poor	19
Ejection Fraction	≤ 30	12
	$]30, 45[$	34
	$]45, 55[$	35
	≥ 55	19
Age	$[18, 40[$	5
	$[40, 70[$	50
	$[70, 93[$	45
Height (cm)	$[150, 165[$	40
	$[165, 175[$	38
	$[175, 187[$	22
Weight (kg)	$[42, 60[$	24
	$[60, 80[$	52
	$[80, 103[$	24

D.2 Visuals depicting the image variability

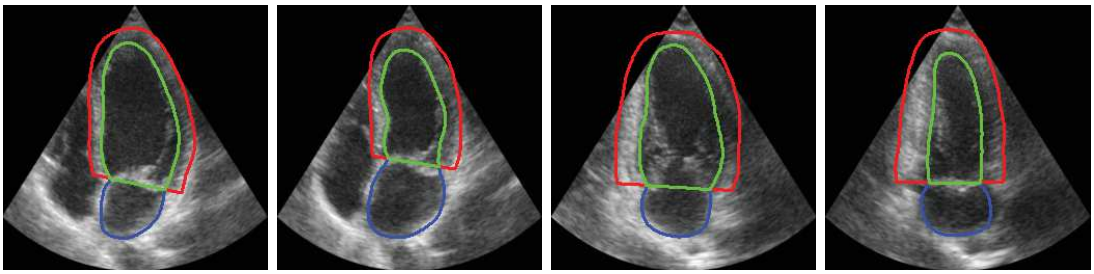
Fig. D.1 shows several illustrations of the expert O_1 annotations on the dataset. The reader may notice the variability in shapes and sizes of the structures which, apart from the LV, are not always fully visible. The epicardium may be extrapolated outside of the sector, and the left atrium appears cut at the bottom of the image.

P-273: EF=27 | IQ=G



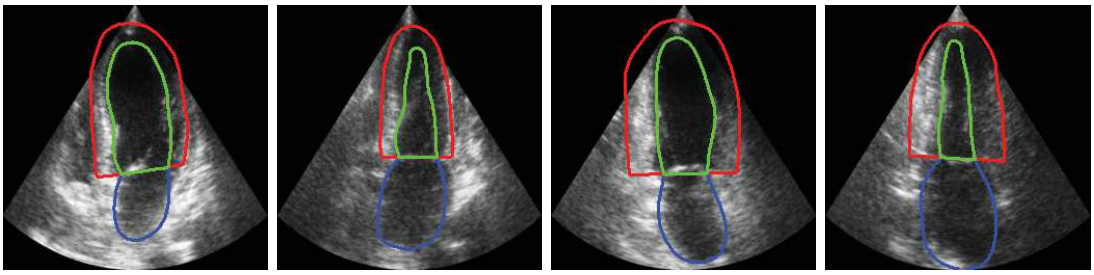
(a) A4C-ED (b) A4C-ES (c) A2C-ED (d) A2C-ES

P-208: EF=45 | IQ=G



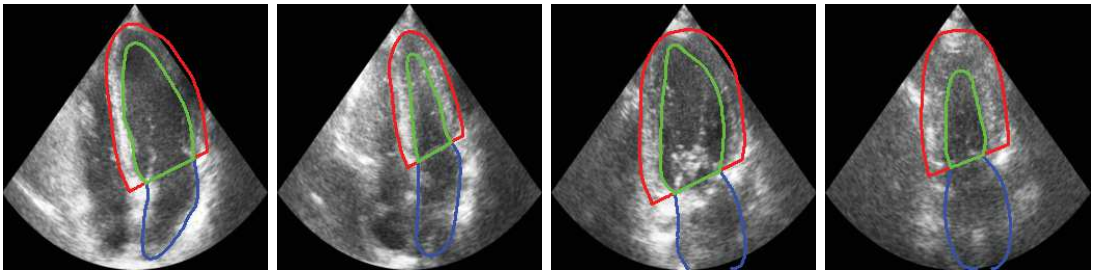
(e) A4C-ED (f) A4C-ES (g) A2C-ED (h) A2C-ES

P-189: EF=70 | IQ=G



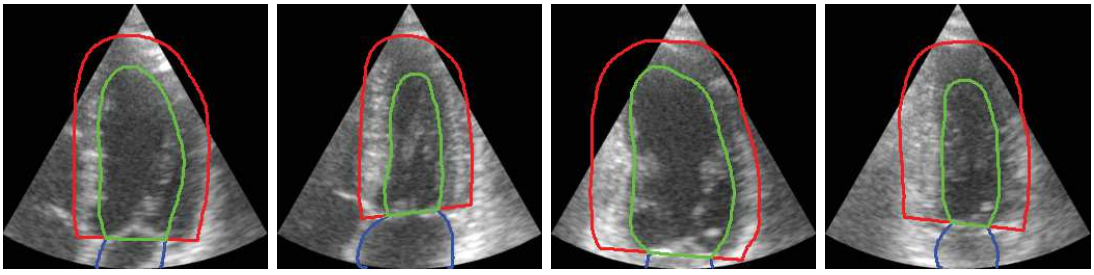
(i) A4C-ED (j) A4C-ES (k) A2C-ED (l) A2C-ES

P-23: EF=65 | IQ=M



(m) A4C-ED (n) A4C-ES (o) A2C-ED (p) A2C-ES

P-344: EF=61 | IQ=P



(q) A4C-ED (r) A4C-ES (s) A2C-ED (t) A2C-ES

FIGURE D.1: Camus dataset samples to show EF and IQ variability. First three rows: IQ = G(ood), EF < 45, [45, 55], > 55.
Last three rows: EF > 55, IQ = G(ood), M(oderate), P(oor).

D.3 Online challenge

The online site to the CAMUS challenge can be found at: `https://camus.creatis.insa-lyon.fr/challenge/`. The welcome page is divided into two redirections. The above leads to an overview of the challenge, where one can expect to find:

- a description of the challenge (context, scientific interest, organizers);
- the original paper [5], also publicly available on ArXiv;
- details on how to participate (registration, result submission);
- information on the CAMUS dataset (population, annotation, acquisition, separation into 450 patients for train and 50 for test);
- the geometrical and clinical metrics used for evaluation;
- contact information: `challengeCAMUS@creatis.insa-lyon.fr`.

The second item directs to the two phases of the evaluation platform, training and test. Each test can only be accessed once the participant has been registered, and enables to download the corresponding dataset and to submit results. Only the test phase is used to rank methods. All participants are allowed three submissions, and can ask to integrate the leader-board by contacting the organizers. This restriction is to limit over-fitting on the challenge, since only one test set is used for evaluation. However, as three submissions are allowed, the platform can be used to directly compare results of new methods to the one we published.
In order to accelerate the processing of the submitted results and the computation of geometrical and clinical indices, the evaluation code of our study was re-written in C++ for the online challenge. We noted differences between the scores computed from the python code and the C++ code, such as:

1. differences between MAD and HD metrics for both structures, due to the change of libraries (in particular different interpolation tactics);
2. differences for the computation of volumes and ejection fraction, due to a change of base detection algorithm;
3. differences in the ranking of methods, due to the limitation to one dataset, which biases the results. A direct comparison between the Python and C++ evaluation platform can be done by comparing the leaderboard in Fig. D.2 and table 7.9.

LEADERBOARD

User	Mean DICE ED	Mean DICE ES	Mean Hausdorff ED	Mean Hausdorff ES	Mean MAD ED	Mean MAS ES	EF correlation	EF bias	EF standard deviation(std)	Volume ED correlation	Volume ED bias	Volume ED std	Volume ES correlation	Volume ES bias	Volume ES std
Sarah Leclerc															
U-net-1(instance)															
endo	0.94	0.91	5.26	5.48	1.66	1.71	0.85	0.10	7.30	0.93	7.15	15.61	0.96	4.40	10.16
epi	0.96	0.95	5.23	5.74	1.74	1.90									
la	0.89	0.92	5.73	5.34	2.21	1.98									
U-net-2(instance)															
endo	0.92	0.90	5.70	5.29	1.62	1.66	0.79	2.60	8.51	0.96	-2.40	11.14	0.97	-3.01	7.62
epi	0.93	0.92	6.44	6.44	1.95	2.08									
la	0.85	0.89	6.92	6.19	2.58	2.13									
ACNNs(instance)															
endo	0.94	0.91	5.55	5.59	1.70	1.73	0.81	0.33	8.26	0.93	2.77	15.51	0.95	2.04	10.11
epi	0.95	0.94	5.86	5.87	1.86	1.98									
la	0.88	0.91	6.00	5.84	2.31	2.16									

FIGURE D.2: Leaderboard with the methods from [5]

Appendix E

Architectures of U-Net 1 and 2

For a better comparison with the original U-Net (Tab. 7.1), we provide the detailed model description of U-Net 1 and U-Net 2 respectively in Tab. E.1 and E.2.

Table E.1: U-Net 1 Architecture

Level	Layer	Kernel / Pool size	Activation	Connection
D1	Conv	32 (3,3)	ReLU	*
	Conv	32 (3,3)	ReLU	
	MaxPooling	(2*2)		
D2	Conv	32 (3,3)	ReLU	**
	Conv	32 (3,3)	ReLU	
	MaxPooling	(2*2)		
D3	Conv	64 (3,3)	ReLU	***
	Conv	64 (3,3)	ReLU	
	MaxPooling	(2*2)		
D4	Conv	128 (3,3)	ReLU	****
	Conv	128 (3,3)	ReLU	
	MaxPooling	(2*2)		
D5	Conv	128 (3,3)	ReLU	*****
	Conv	128 (3,3)	ReLU	
	MaxPooling	(2*2)		
D6	Conv	128 (3,3)	ReLU	
	Conv	128 (3,3)	ReLU	
U1	UpSampling	(2,2)		*****
	Conv	128 (3,3)	ReLU	
	Conv	128 (3,3)	ReLU	
U2	UpSampling	(2,2)		****
	Conv	128 (3,3)	ReLU	
	Conv	128 (3,3)	ReLU	
U3	UpSampling	(2,2)		***
	Conv	64 (3,3)	ReLU	
	Conv	64 (3,3)	ReLU	
U4	UpSampling	(2,2)		**
	Conv	32 (3,3)	ReLU	
	Conv	32 (3,3)	ReLU	
U5	UpSampling	(2,2)		*
	Conv	16 (3,3)	ReLU	
	Conv	16 (3,3)	ReLU	
Seg	Conv	4 (1,1)	Softmax	

TABLE E.2: U-Net 2 Architecture

Level	Layer	Kernel / Pool size	Activation	Connection
D1	Conv	48 (3,3)		
	BatchNorm		ReLU	
	Conv	48 (3,3)		
	BatchNorm		ReLU	*
	MaxPooling	(2*2)		
D2	Conv	96 (3,3)		
	BatchNorm		ReLU	
	Conv	96 (3,3)		
	BatchNorm		ReLU	**
	MaxPooling	(2*2)		
D3	Conv	192 (3,3)		
	BatchNorm		ReLU	
	Conv	192 (3,3)		
	BatchNorm		ReLU	***
	MaxPooling	(2*2)		
D4	Conv	384 (3,3)		
	BatchNorm		ReLU	
	Conv	384 (3,3)		
	BatchNorm		ReLU	****
	MaxPooling	(2*2)		
D5	Conv	768 (3,3)		
	BatchNorm		ReLU	
	Conv	768 (3,3)		
	BatchNorm		ReLU	
U1	UpConv	384 (2,2) - s(2,2)		
	BatchNorm		ReLU	
	Conv	384 (3,3)		****
	BatchNorm		ReLU	
	Conv	384 (3,3)		
	BatchNorm		ReLU	
U2	UpConv	192 (2,2) - s(2,2)		
	BatchNorm		ReLU	
	Conv	192 (3,3)		***
	BatchNorm		ReLU	
	Conv	192 (3,3)		
	BatchNorm		ReLU	
U3	UpConv	96 (2,2) - s(2,2)		
	BatchNorm		ReLU	
	Conv	96 (3,3)		**
	BatchNorm		ReLU	
	Conv	96 (3,3)		
	BatchNorm		ReLU	
U4	UpConv	48 (2,2) - s(2,2)		
	BatchNorm		ReLU	
	Conv	48 (3,3)		*
	BatchNorm		ReLU	
	Conv	48 (3,3)		
	BatchNorm		ReLU	
Seg	Conv	4 (1,1)	Softmax	

s: height and width strides

Appendix F

Supplementary experiments on ACNN

This appendix contains supplementary results on the ACNN architecture and the associated auto-encoder. In particular, we investigated:

F.1 Auto-encoder implementation, training and visuals

The auto-encoders of our study learn to reconstruct all segmentation masks of the training dataset from a low dimensional vector of fixed size with 32 values instead of 64 as in the original paper to limit the accuracy of the reconstruction on our 2D problem. The detailed architecture is given in Tab. F.1. Convolutions are not followed by any padding so the image

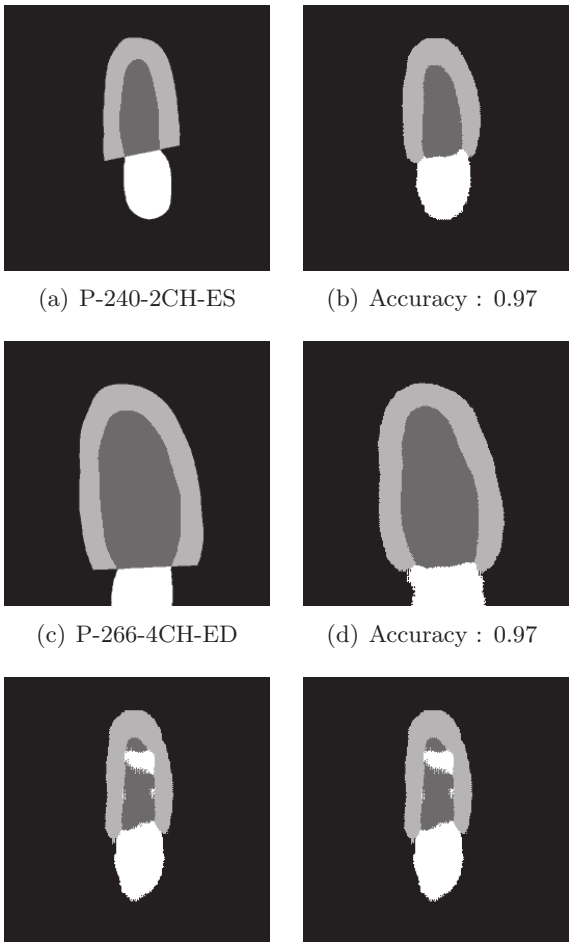
Table F.1: Auto-encoder Architecture

Level	Layer	Kernel size	Activation
Encoder	Conv	16 (3,3)	eLU
	Conv	16 (3,3)	eLU
	Conv	32 (3,3)	eLU
	Conv	32 (3,3)	eLU
	Conv	64 (3,3)	eLU
	Conv	64 (3,3)	eLU
Code	Conv+Flatten	1 (3,3)	eLU
	Dense	32	
Decoder	Dense	169	eLU
	Reshape	1 (13,13)	
	ConvTranspose	64 (4,4)	eLU
	ConvTranspose	64 (4,4)	eLU
	Conv	64 (3,3)	eLU
	ConvTranspose	32 (4,4)	eLU
	Conv	32 (3,3)	eLU
	ConvTranspose	16 (4,4)	eLU
	Conv	16 (3,3)	eLU
	ConvTranspose	16 (4,4)	eLU
Seg	Conv	4 (3,3)	Softmax

size is gradually reduced in the encoding part and enlarged in the decoding part. L1 regularization is applied on the code so that code coefficients remain small and of similar magnitude for all auto-encoders, which improved the stability of the training and the robustness of the test results. ELU activations are used as they showed to lead to a better convergence.

We train the networks during 50 epochs (long after convergence) with a batch size of 50. Cross-Entropy is used as the loss to optimize with the Adam optimizer and a learning rate of 1e-4. Train, validation and test sets are the same as the corresponding segmentation network and all ten auto-encoders achieve an accuracy superior to 92%.

Fig. F.1 shows examples of reconstructions from auto-encoders trained on the CAMUS dataset on images from the test set 5. As can be assessed visually, the reconstruction most often encodes correctly the position and the size of the structures while not being accurate on details. Some cases show errors of reconstruction due to the instability of the latent space, i.e a small change of coefficient can induce local errors in the reconstruction.



(a) P-240-2CH-ES (b) Accuracy : 0.97

(c) P-266-4CH-ED (d) Accuracy : 0.97

(e) P-242-4CH-ES (f) Accuracy : 0.96

FIGURE F.1: Reconstruction examples from the auto-encoders of our study. Left: ground truth; Right: reconstruction. The first two rows illustrate the average accuracy on the full CAMUS dataset. The last row shows an anatomically implausible reconstruction.

F.2 Impact of the regularization loss

The regularization loss corresponds to the euclidean distance of the code coefficients. As its values do not scale with the ones provided by the segmentation loss, a multiplying factor is applied. In our study, it is set at 10^4 so that the two losses have close initial values. In Table F.2, we compare the segmentation accuracy reached by our ACNN implementation for two different values of the regularization weight : the chosen one, and one 100 times stronger. From this table, it can be seen that the increase of the regularization weight affects the quality of the results, especially for the HD metric.

TABLE F.2: Segmentation accuracy for ACNN architecture with different shape regularization strengths

Methods *	LV_{endo}			LV_{epi}		
	D	*MAD*	*HD*	*D*	*MAD*	*HD*
	val.	mm	mm	val.	mm	mm
ACNN - $\lambda = 10^4$	**0.918** ±0.050	**1.8** ±1.0	**5.9** ±3.5	**0.946** ±0.030	**1.9** ±1.1	**6.4** ±4.2
ACNN - $\lambda = 10^6$	0.912 ±0.054	1.9 ±1.1	6.2 ±4.0	0.947 ±0.030	**1.9** ±1.1	6.9 ±5.8

F.3 Influence of the training set size

To study the influence of the training set of the auto-encoder and segmentation network of ACNNs, we first trained two auto-encoders using the fold 6 as validation set and fold 5 as test set :

- SR400, auto-encoder trained on 400 patients;
- SR15, auto-encoder trained on 15 patients from the same training set;

We then trained and evaluated five ACNN networks (using U-Net 1 for the segmentation part of the network) with the same validation and test set:

1. 50*p_no_SR*, which involved no shape regularization and was trained on 50 patients (200 images);
2. 5*p_no_SR*, which was trained on a smaller set of only 5 patients (20 images);
3. 50*p_with_SR*15, trained on the same training set as 50*p_no_SR* while optimizing the ACNN shape regularization auxiliary loss returned from *SR*15;
4. 50*p_with_SR*400, similar to 50*p_with_SR*15 but with the *SR*400 auto-encoder;
5. 5*p_SR*400, trained on the same training set as 5*p_no_SR*, with the shape regularization from *SR*400.

Fig. F.2 shows the geometrical results of all five models. It can be observed that :

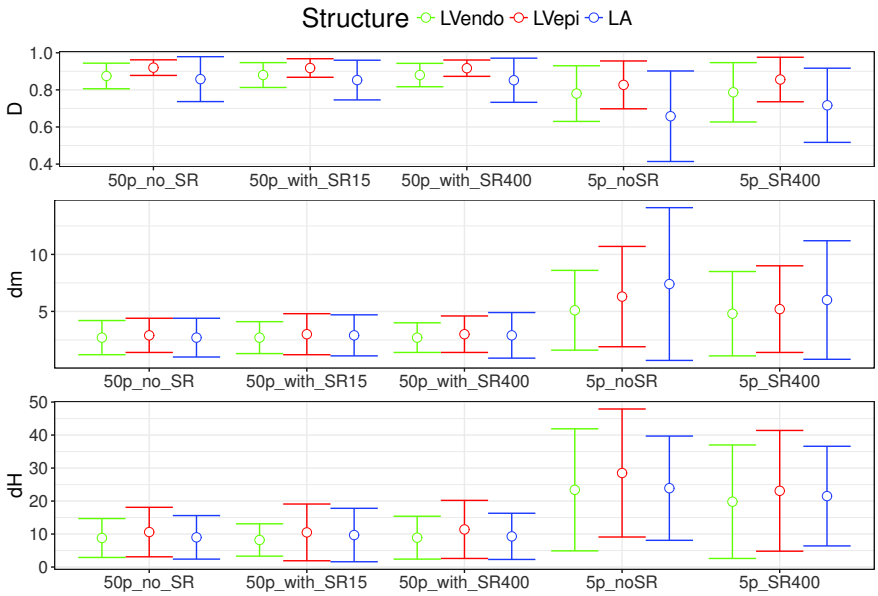


FIGURE F.2: Geometric performance of ACNNs illustrated by standard error bars around the mean values for the 5 segmentation networks.

- Shape regularization does not improve results for an ACNN trained on 50 patients but it does significantly improve the results for a smaller training set of 5 patients;
- Using for the ACNN an auto-encoder learnt from 15 patients (accuracy of 84%) produced close results compared to the one learnt on 400 patients.

These results support the idea that shape regularization can be helpful on datasets for which the training dataset size is not sufficient to learn the annotated shapes' complexity. For our specific task of 2D echocardiography, it appears that the necessary number of cases is low, inferior to 50 patients, probably because the shape variability is also low, allowing an auto-encoder to roughly infer it from a few number of cases. In the experiments in Chapter 8, we had access to 400 patients, which explains why the ACNN did not show any significant improvement compared to U-Net 1.

Appendix G

Supplementary experiments on RU-Net

This appendix contains supplementary results on the RU-Net architecture, focusing on the refining effect and the impact of hyper-parameters. To do so, Tab. G.1 shows the results of the two outputs of RU-Net in a balanced loss setting for $dil = 30$ and $shift = 0.7$, while Tab. G.2 shows the comparison of RU-Net, U-Net 1 and SHG for $dil = 11$ and $shift = 0.5$.

TABLE G.1: Refinement effect on RU-Net with $dil = 30$ and $shift = 0.7$
Cross validation on 10 subfolds of 200 images

	LV_{endo}		LV_{epi}		**Outliers # %**	
	MAD	**HD**	**MAD**	**HD**	**geo**	**ana**
RU-net- o1	1.8 ±1.2	5.9 ±3.6	2.0 ±1.1	6.5 ±3.9	403 20%	116 5.8%
RU-net - o2	**1.7** ±1.0	**5.6** ±3.3	**1.9** ±1.1	**6.0** ±4.1	**306** 15%	**33** 1.7%

TABLE G.2: Geometrical performance and outliers for RU-Net : $dil = 11, shift = 0.5$
Cross validation on 10 subfolds of 200 images

Model	**Prms**	LV_{endo}		LV_{epi}		**Outliers # %**	
	#	**MAD**	**HD**	**MAD**	**HD**	**geo**	**ana**
U-net 1	2.0M	**1.8** ±1.2	5.8 ±3.4	2.0 ±1.1	6.2 ±3.8	396 20%	95 4.8%
SHG	4.5M	**1.8** ±1.1	5.9 ±3.5	2.0 ±1.1	6.2 ±3.9	425 21%	47 2.4%
RU-net	3.9M	**1.8** ±1.1	**5.7** ±3.6	**1.9** ±1.2	**6.1** ±4.4	**320** 16%	**23** 1.2%

G.1 Comparison to other methods

From Tab. G.2, one can see that RU-Net outperforms the other two networks in terms of geometrical (-4%) and anatomical outliers (down to 1.2%), while maintaining high geometrical accuracy. Such two-step architecture therefore improves the robustness of the segmentation. On classic geometrical metrics, the SHG architecture does not bring improvement compared to the scores of U-Net 1, however, a decrease of the anatomical outlier rate can be noted.

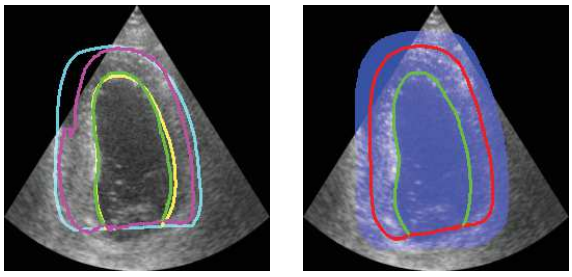
RU-Net in turns increases the geometrical performance consistently, though marginally, and induces a significant drop in anatomical outliers, which ultimately represent 1% of the dataset. Fig. G.1 shows an example of refined result. We can see that the performance is a lot smoother. However, the contour no longer is farther from the ground truth, which means the refining is double-edged, as the second network follows the guidance of the first segmentation.

G.2 Refining effect

Tab. G.1 shows the comparison between the two outputs of RU-Net. We can notice a small but consistent improvement over all segmentation metrics, with significant drops for the HD metrics. However the corresponding scores were higher than those observed for U-Net 1. Still, the scores of the first output are very close to what is obtained by U-Net 1, especially the number of outliers and the MAD values, which explicitly shows that the refinement occurring along RU-Net is not performed at the cost of the quality of the first output. We observe again on this table the decrease in outliers, both anatomical and geometrical.

G.3 Impact of hyper-parameters

The comparison of the scores of RU-Net on Tab. G.1 and Tab. G.2 allow to observe the impact of hyper-parameter changes for two sensible values: ($dil = 30 \rightarrow 11$; $shift = 0.7 \rightarrow 0.5$). The scores produced by RU-Net are quite stable when changing the two hyper-parameters. In our tests, we observed that $dil = 30$ was a good compromise between a too small dilation rate, which led to the second network following closely the first one's prediction, and a large dilation rate, where the benefit from the ROI detection was lost. Concerning the shift, keeping a higher value reduced the number of potentially misleading false positives. We therefore concluded that the solution $dil = 30$ and $shift = 0.7$ was the best.



(a) U-Net result (b) RU-Net result

FIGURE G.1: Illustration of the refinement in place in RU-Net. The ROI is shown in blue. The ground truth is shown in yellow and cyan while the prediction is on green and red/magenta. On the U-Net epicardium contour (magenta), we see the epicardium being locally discontinuous while it is not the case for RU-Net (red).

Bibliography

[1] O. Bernard, *Camus challenge*, camus.creatis.insa-lyon.fr/challenge/, [Online], 2019.

[2] P. Peng, K. Lekadir, A. Gooya, L. Shao, S. E. Petersen, and A. F. Frangi, "A review of heart chamber segmentation for structural and functional analysis using cardiac magnetic resonance imaging", *Magma (New York, N.Y.)*, vol. 29, no. 2, pp. 155–195, 2016, ISSN: 1352-8661. DOI: 10.1007/s10334-015-0521-4. [Online]. Available: https://www.ncbi.nlm.nih.gov/pubmed/26811173.

[3] J. A. Noble and D. Boukerroui, "Ultrasound image segmentation: A survey", *IEEE Transactions on Medical Imaging*, vol. 25, no. 8, pp. 987–1010, 2006, ISSN: 0278-0062. DOI: 10.1109/TMI.2006.877092.

[4] S. A. Kleijn and O. Kamp, "Clinical application of three-dimensional echocardiography: Past, present and future", *Netherlands heart journal : Monthly journal of the Netherlands Society of Cardiology and the Netherlands Heart Foundation*, vol. 17, no. 1, pp. 18–24, 2009, 19148334[pmid], ISSN: 1568-5888. DOI: 10.1007/bf03086210. [Online]. Available: https://www.ncbi.nlm.nih.gov/pubmed/19148334.

[5] S. Leclerc, E. Smistad, J. Pedrosa, A. Ostvik, F. Cervenansky, F. Espinosa, T. Espeland, E. A. R. Berg, P. Jodoin, T. Grenier, C. Lartizien, J. Dhooge, L. Lovstakken, and O. Bernard, "Deep learning for segmentation using an open large-scale dataset in 2d echocardiography", *IEEE Transactions on Medical Imaging*, pp. 1–12, 2019, ISSN: 0278-0062. DOI: doi:10.1109/TMI.2019.2900516.

[6] T. F. Cootes, C. J. Taylor, D. H. Cooper, and J. Graham, "Active shape models mdash;their training and application", *Comput. Vis. Image Underst.*, vol. 61, no. 1, pp. 38–59, 1995, ISSN: 1077-3142. DOI: 10.1006/cviu.1995.1004. [Online]. Available: http://dx.doi.org/10.1006/cviu.1995.1004.

[7] L. Breiman, "Random forests", *Machine Learning*, vol. 45, no. 1, pp. 5–32, 2001, ISSN: 1573-0565. DOI: doi:10.1023/A:1010933404324. [Online]. Available: https://doi.org/10.1023/A:1010933404324.

[8] J. A. Zagzebski, *Physics and instrumentation in doppler and b-mode ultrasonography*, radiologykey.com/physics-and-instrumentation-in-doppler-and-b-mode-ultrasonography/, 2019.

[9] S. Asbjorn, *Basic ultrasound for clinicians*, folk.ntnu.no/stoylen/strainrate/Basic_ultrasound, [Online], 2016.

[10] A. Ng and J. Swanevelder, "Resolution in ultrasound imaging", *BJA Education*, vol. 11, no. 5, pp. 186–192, 2011, ISSN: 2058-5349. DOI: 10.1093/bjaceaccp/mkr030. eprint: http://oup.prod.sis.lan/bjaed/article-pdf/11/5/186/794418/mkr030.pdf. [Online]. Available: https://doi.org/10.1093/bjaceaccp/mkr030.

[11] P. B. Bertrand, R. A. Levine, E. M. Isselbacher, and P. M. Vandervoort, "Fact or artifact in two-dimensional echocardiography: Avoiding misdiagnosis and missed diagnosis", *Journal of the American Society of Echocardiography*, vol. 29, no. 5, pp. 381–391, 2016, ISSN: 0894-7317. DOI: 10.1016/j.echo.2016.01.009. [Online]. Available: https://doi.org/10.1016/j.echo.2016.01.009.

[12] P. M. Elliott, A. Anastasakis, M. A. Borger, M. Borggrefe, F. Cecchi, P. Charron, A. A. Hagege, A. Lafont, G. Limongelli, H. Mahrholdt, W. J. McKenna, J. Mogensen, P. Nihoyannopoulos, S. Nistri, P. G. Pieper, B. Pieske, C. Rapezzi, F. H. Rutten, C. Tillmanns, H. Watkins, A. Contributor, C. O'Mahony, E. C. for Practice Guidelines (CPG), J. L. Zamorano, S. Achenbach, H. Baumgartner, J. J. Bax, H. Bueno, V. Dean, C. Deaton, . Erol, R. Fagard, R. Ferrari, D. Hasdai, A. W. Hoes, P. Kirchhof, J. Knuuti, P. Kolh, P. Lancellotti, A. Linhart, P. Nihoyannopoulos, M. F. Piepoli, P. Ponikowski, P. A. Sirnes, J. L. Tamargo, M. Tendera, A. Torbicki, W. Wijns, S. Windecker, D. Reviewers, D. Hasdai, P. Ponikowski, S. Achenbach, F. Alfonso, C. Basso, N. M. Cardim, J. R. Gimeno, S. Heymans, P. J. Holm, A. Keren, P. Kirchhof, P. Kolh, C. Lionis, C. Muneretto, S. Priori, M. J. Salvador, C. Wolpert, J. L. Zamorano, M. Frick, F. Aliyev, S. Komissarova, G. Mairesse, E. Smaji, V. Velchev, L. Antoniades, A. Linhart, H. Bundgaard, T. Heliö, A. Leenhardt, H. A. Katus, G. Efthymiadis, R. Sepp, G. Thor Gunnarsson, S. Carasso, A. Kerimkulova, G. Kamzola, H. Skouri, G. Eldirsi, A. Kavoliuniene, T. Felice, M. Michels, K. Hermann Haugaa, R. Lenarczyk, D. Brito, E. Apetrei, L. Bokheria, D. Lovic, R. Hatala, P. Garcia Pavía, M. Eriksson, S. Noble, E. Srbinovska, M. Özdemir, E. Nesukay, and N. Sekhri, "2014 esc guidelines on diagnosis and management of hypertrophic cardiomyopathy: the task force for the diagnosis and management of hypertrophic cardiomyopathy of the european society of cardiology (esc)", *European Heart Journal*, vol. 35, no. 39, pp. 2733–2779, 2014, ISSN: 0195-668X. DOI: 10.1093/eurheartj/ehu284. eprint: http://oup.prod.sis.lan/eurheartj/article-pdf/35/39/2733/17898410/ehu284.pdf. [Online]. Available: https://doi.org/10.1093/eurheartj/ehu284.

[13] 123sonography, *E-book on echocardiography*, www.123sonography.com/book/, [Online], 2019.

[14] A. C. Armstrong, E. P. Ricketts, C. Cox, P. Adler, A. Arynchyn, K. Liu, E. Stengel, S. Sidney, C. E. Lewis, P. J. Schreiner, J. M. Shikany, K. Keck, J. Merlo, S. S. Gidding, and J. A. C. Lima, "Quality control and reproducibility in m-mode, two-dimensional, and speckle tracking echocardiography acquisition and analysis: The cardia study, year 25 examination experience", *Echocardiography (Mount Kisco, N.Y.)*, vol. 32, no. 8, pp. 1233–1240, 2015, ISSN: 1540-8175. DOI: 10.1111/echo.12832. [Online]. Available: https://www.ncbi.nlm.nih.gov/pubmed/25382818.

[15] A. Fenster and B. Chiu, "Evaluation of segmentation algorithms for medical imaging", in *2005 IEEE Engineering in Medicine and Biology 27th Annual Conference*, 2005, pp. 7186–7189. DOI: 10.1109/IEMBS.2005.1616166.

[16] O. Bernard, J. G. Bosch, B. Heyde, M. Alessandrini, D. Barbosa, S. Camarasu-Pop, F. Cervenansky, S. Valette, O. Mirea, *et al.*, "Standardized evaluation system for left ventricular segmentation algorithms in 3d echocardiography", *IEEE Transactions on Medical Imaging*, vol. 35, no. 4, pp. 967–977, 2016.

[17] S. Dong, G. Luo, K. Wang, S. Cao, A. Mercado, O. Shmuilovich, H. Zhang, and S. Li, "Voxelatlasgan: 3d left ventricle segmentation on echocardiography with atlas guided generation and voxel-to-voxel discrimination", in *MICCAI*, 2018.

[18] G. Carneiro, J. C. Nascimento, and A. Freitas, "The segmentation of the left ventricle of the heart from ultrasound data using deep learning architectures and derivative-based search methods", *IEEE Transactions on Image Processing*, vol. 21, no. 3, pp. 968–982, 2012.

[19] E. Smistad, A. Ostvik, B. O. Haugen, and L. Lovstakken, "2d left ventricle segmentation using deep learning", in *2017 IEEE International Ultrasonics Symposium (IUS)*, 2017. DOI: doi:10.1109/ULTSYM.2017.8092573.

[20] N. Azarmehr, X. Ye, F. Janan, J. P. Howard, D. P. Francis, and M. Zolgharni, "Automated segmentation of left ventricle in 2d echocardiography using deep learning", in *International Conference on Medical Imaging with Deep Learning – Extended Abstract Track*, 2019. [Online]. Available: openreview.net/forum?id=Sye8klvmcN.

[21] K. Y. E. Leung and J. G. Bosch, "Automated border detection in three-dimensional echocardiography: Principles and promises.", *European journal of echocardiography : The journal of the Working Group on Echocardiography of the European Society of Cardiology*, vol. 11 2, pp. 97–108, 2010.

[22] D. Kang, J. Woo, C. C. J. Kuo, P. J. Slomka, D. Dey, and G. Germano, "Heart chambers and whole heart segmentation techniques: Review", *Journal of Electronic Imaging*, vol. 21, no. 1, pp. 1 –17 –17, 2012. DOI: 10.1117/1.JEI.21.1.010901. [Online]. Available: https://doi.org/10.1117/1.JEI.21.1.010901.

[23] T. F. Cootes, G. J. Edwards, and C. J. Taylor, "Active appearance models", in *Computer Vision — ECCV'98*, H. Burkhardt and B. Neumann, Eds., Berlin, Heidelberg: Springer Berlin Heidelberg, 1998, pp. 484–498, ISBN: 978-3-540-69235-5.

[24] A. Criminisi and J. Shotton, *Decision Forests for Computer Vision and Medical Image Analysis*. Springer Publishing Company, Incorporated, 2013, ISBN: 1447149289, 9781447149286.

[25] D. E. Rumelhart, G. E. Hinton, and R. J. Williams, "Learning representations by back-propagating errors", *Nature*, vol. 323, no. 6088, pp. 533–536, 1986, ISSN: 1476-4687. DOI: doi:10.1038/323533a0. [Online]. Available: https://doi.org/10.1038/323533a0.

[26] G. Litjens, T. Kooi, B. E. Bejnordi, A. A. A. Setio, F. Ciompi, M. Ghafoorian, J. A. van der Laak, B. van Ginneken, and C. I. Sánchez, "A survey on deep learning in medical image analysis", *Medical Image Analysis*, vol. 42, pp. 60 –88, 2017, ISSN: 1361-8415. DOI: https://doi.org/10.1016/j.media.2017.07.005. [Online]. Available: http://www.sciencedirect.com/science/article/pii/S1361841517301135.

[27] R. M. Lang, L. P. Badano, V. Mor-Avi, J. Afilalo, *et al.*, "Recommendations for cardiac chamber quantification by echocardiography in adults: an update from the american society of echocardiography and the european association of cardiovascular imaging", *Circulation*, vol. 28, no. 1, pp. 1–39, 2015.

[28] P. Kontschieder, S. R. Bulò, H. Bischof, and M. Pelillo, "Structured class-labels in random forests for semantic image labelling", in *2011 International Conference on Computer Vision*, 2011, pp. 2190–2197. DOI: doi:10.1109/ICCV.2011.6126496.

[29] P. Kontschieder, S. R. Bulò, M. Pelillo, and H. Bischof, "Structured labels in random forests for semantic labelling and object detection", *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 36, no. 10, pp. 2104–2116, 2014, ISSN: 0162-8828. DOI: doi:10.1109/TPAMI.2014.2315814.

[30] S. Leclerc, T. Grenier, F. Espinosa, and O. Bernard, "A fully automatic and multi-structural segmentation of the left ventricle and the myocardium on highly heterogeneous 2d echocardiographic data", in *2017 IEEE International Ultrasonics Symposium (IUS)*, 2017, pp. 1–4. DOI: doi:10.1109/ULTSYM.2017.8092632.

[31] S. Leclerc, E. Smistad, T. Grenier, C. Lartizien, A. Ostvik, F. Espinosa, P. Jodoin, L. Lovstakken, and O. Bernard, "Deep learning applied to multi-structure segmentation in 2d echocardiography: A preliminary investigation of the required database size", in *2018 IEEE International Ultrasonics Symposium (IUS)*, 2018, pp. 1–4. DOI: doi:10.1109/ULTSYM.2018.8580136.

[32] F. Khellaf, S. Leclerc, J. D. Voorneveld, R. S. Bandaru, J. G. Bosch, and O. Bernard, *Left ventricle segmentation in 3d ultrasound by combining structured random forests with active shape models*, 2018. DOI: doi:10.1117/12.2293544.

[33] J. S. Domingos, R. V. Stebbing, P. Leeson, and J. A. Noble, "Structured random forests for myocardium delineation in 3d echocardiography", in *Machine Learning in Medical Imaging*, G. Wu, D. Zhang, and L. Zhou, Eds., Cham: Springer International Publishing, 2014, pp. 215–222.

[34] O. Ronneberger, P. Fischer, and T. Brox, "U-net: Convolutional networks for biomedical image segmentation", in *Medical Image Computing and Computer-Assisted Intervention – MICCAI 2015*, N. Navab, J. Hornegger, W. M. Wells, and A. F. Frangi, Eds., Cham: Springer International Publishing, 2015, pp. 234–241, ISBN: 978-3-319-24574-4.

[35] F. Milletari, N. Navab, and S. Ahmadi, "V-net: Fully convolutional neural networks for volumetric medical image segmentation", in *2016 Fourth International Conference on 3D Vision (3DV)*, 2016, pp. 565–571. DOI: doi:10.1109/3DV.2016.79.

[36] P.-M. Jodoin, *Basics of deep learning*, deepimaging2019.sciencesconf.org/, [Online], 2019.

[37] D. Barbosa, T. Dietenbeck, B. Heyde, H. Houle, D. Friboulet, J. Dhooge, and O. Bernard, "Fast and fully automatic 3-d echocardiographic segmentation using b-spline explicit active surfaces: feasibility study and validation in a clinical setting", *Ultrasound in Medicine and Biology*, vol. 39, no. 1, pp. 89 –101, 2013.

[38] J. Pedrosa, S. Queirós, O. Bernard, J. Engvall, T. Edvardsen, E. Nagel, and J. Dhooge, "Fast and fully automatic left ventricular segmentation and tracking in echocardiography using shape-based b-spline explicit active surfaces", *IEEE Transactions on Medical Imaging*, vol. 36, no. 11, pp. 2287–2296, 2017.

[39] M. Drozdzal, E. Vorontsov, G. Chartrand, S. Kadoury, and C. Pal, "The importance of skip connections in biomedical image segmentation", 2016. DOI: 10.1007/978-3-319-46976-8_19.

[40] Z. Zhou, M. M. Rahman Siddiquee, N. Tajbakhsh, and J. Liang, "Unet++: A nested u-net architecture for medical image segmentation", in *Deep Learning in Medical Image Analysis and Multimodal Learning for Clinical Decision Support*, D. Stoyanov, Z. Taylor, G. Carneiro, T. Syeda-Mahmood, A. Martel, L. Maier-Hein, J. M. R. Tavares, A. Bradley, J. P. Papa, V. Belagiannis, J. C. Nascimento, Z. Lu, S. Conjeti, M. Moradi, H. Greenspan, and A. Madabhushi, Eds., Cham: Springer International Publishing, 2018, pp. 3–11, ISBN: 978-3-030-00889-5.

[41] A. Newell, K. Yang, and J. Deng, "Stacked hourglass networks for human pose estimation", in *Computer Vision – ECCV 2016*, B. Leibe, J. Matas, N. Sebe, and M. Welling, Eds., Cham: Springer International Publishing, 2016, pp. 483–499, ISBN: 978-3-319-46484-8.

[42] O. Oktay, E. Ferrante, K. Kamnitsas, M. Heinrich, W. Bai, J. Caballero, R. Guerrero, S. Cook, A. de Marvao, T. Dawes, D. O'Regan, B. Kainz, B. Glocker, and D. Rueckert, "Anatomically constrained neural networks (acnn): Application to cardiac image enhancement and segmentation", *IEEE Transactions on Medical Imaging*, vol. PP, 2017. DOI: 10.1109/TMI.2017.2743464.

[43] S. Leclerc, E. Smistad, A. Ostvik, F. Cervenansky, F. Espinosa, T. Espeland, E. A. R. Berg, P.-M. Jodoin, T. Grenier, C. Lartizien, L. Lovstakken, and O. Bernard, "Deep learning segmentation in 2d echocardiography using the camus dataset : Automatic assessment of the anatomical shape validity", in *International Conference on Medical Imaging with Deep Learning – Extended Abstract Track*, London, United Kingdom, 2019.

[44] Y. Zhu, Y. Tian, D. Mexatas, and P. Dollar, "Semantic amodal segmentation", in *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2017.

[45] S. Leclerc, E. Smistad, T. Grenier, C. Lartizien, A. Østvik, F. Cervenansky, F. Espinosa, T. Espeland, E. A. R. Berg, P. Jodoin, L. Lovstakken, and O. Bernard, "Runet: A refining segmentation network for 2d echocardiography", in *2019 IEEE International Ultrasonics Symposium (IUS)*, 2019, pp. 1–4.

[46] S. Leclerc, E. Smistad, A. Østvik, F. Cervenansky, F. Espinosa, T. Espeland, E. A. R. Berg, T. Grenier, C. Lartizien, P.-M. Jodoin, L. Lovstakken, and O. Bernard, "Lu-net: A multi-task network to improve the robustness of deep learning segmentation in 2d echocardiography", [Submitted to TUFFC, special issue on Deep learning in medical ultrasound from image formation to image analysis].

[47] O. Oktay, J. Schlemper, L. L. Folgoc, M. Lee, M. Heinrich, K. Misawa, K. Mori, S. McDonagh, N. Y. Hammerla, B. Kainz, B Glocker, and D. Rueckert, "Attention u-net: learning where to look for the pancreas ", in *Medical Imaging with Deep Learning (MIDL'18)*, 2018.

[48] H.-T. Nguyen, P. Croisille, M. Viallon, S. Leclerc, S. Grange, R. Grange, O. Bernard, and T. Grenier, "Robustly segmenting quadriceps muscles of ultra-endurance athletes with weakly supervised u-net", in *International Conference on Medical Imaging with Deep Learning – Extended Abstract Track*, London, United Kingdom, 2019.

[49] C. Angermann, M. Haltmeier, R. Steiger, S. P. Jr., and E. R. Gizewski, "Projection-based 2.5d u-net architecture for fast volumetric segmentation", *CoRR*, vol. abs/1902.00347, 2019. arXiv: 1902.00347. [Online]. Available: http://arxiv.org/abs/1902.00347.

[50] J. Duan, G. Bello, J. Schlemper, W. Bai, T. J. W. Dawes, C. Biffi, A. de Marvao, G. Doumoud, D. P. ORegan, and D. Rueckert, "Automatic 3d bi-ventricular segmentation of cardiac images by a shape-refined multi-task deep learning approach", in *IEEE Transactions on Medical Imaging*, 2018.

[51] N. Painchaud, Y. Skandarani, T. Judge, O. Bernard, A. Lalande, and P.-M. Jodoin, *Cardiac mri segmentation with strong anatomical guarantees*, 2019. arXiv: 1907.02865 [eess.IV].

[52] E. Smistad, A. Østvik, I. Salte, D. Melichova, T. M. Nguyen, H. Brunvand, T. Edvardsen, S. Leclerc, O. Bernard, B. Grenne, and L. Lovstakken, "Real-time automatic ejection fraction and foreshortening detection using deep learning", [Submitted to TUFFC, special issue on Deep learning in medical ultrasound from image formation to image analysis].

[53] A. Østvik, E. Smistad, S. A. Aase, B. O. Haugen, and L. Lovstakken, "Real-time standard view classification in transthoracic echocardiography using convolutional neural networks", *Ultrasound in Medicine and Biology*, vol. 45, no. 2, pp. 374–384, 2019, ISSN: 0301-5629. DOI: 10.1016/j.ultrasmedbio.2018.07.024. [Online]. Available: https://doi.org/10.1016/j.ultrasmedbio.2018.07.024.

[54] NDK, *Basic principle of medical ultrasonic probes (transducer)*, www.ndk.com/tc/sensor/ultrasonic/basi 2019.

[55] M. Daffertshofer and M. Hennerici, "Sonothrombolysis: Experimental evidence", *Frontiers of neurology and neuroscience*, vol. 21, pp. 140–9, 2006. DOI: 10.1159/000092396.

[56] B. Kennedy, *Ultrasound lectures*, slideplayer.com/slide/13784365/, 2018.

[57] J. P. Lawrence, "Physics and instrumentation of ultrasound", *Critical Care Medicine*, vol. 35, no. 8, 2007, ISSN: 0090-3493.

[58] O. Bernard, "Cardiac segmentation : Towards a robust estimation of volumetric indices", Habilitation à diriger des recherches, INSA de Lyon, 2019.

[59] N. S. Anavekar and J. K. Oh, "Doppler echocardiography: A contemporary review", *Journal of Cardiology*, vol. 54, no. 3, pp. 347–358, 2009, ISSN: 0914-5087. DOI: 10.1016/j.jjcc.2009.10.001. [Online]. Available: https://doi.org/10.1016/j.jjcc.2009.10.001.

[60] O. A. Smiseth, H. Torp, A. Opdahl, K. H. Haugaa, and S. Urheim, "Myocardial strain imaging: how useful is it in clinical decision making?", *European Heart Journal*, vol. 37, no. 15, pp. 1196–1207, 2015, ISSN: 0195-668X. DOI: 10.1093/eurheartj/ehv529. eprint: http://oup.prod.sis.lan/eurheartj/article-pdf/37/15/1196/24121129/ehv529.pdf. [Online]. Available: https://doi.org/10.1093/eurheartj/ehv529.

[61] A. A. Taha and A. Hanbury, "Metrics for evaluating 3d medical image segmentation: Analysis, selection, and tool", *BMC medical imaging*, vol. 15, pp. 29–29, 2015, ISSN: 1471-2342. DOI: 10.1186/s12880-015-0068-x. [Online]. Available: https://www.ncbi.nlm.nih.gov/pubmed/26263899.

[62] A. Criminisi, J. Shotton, and E. Konukoglu, "Decision forests: A unified framework for classification, regression, density estimation, manifold learning and semi-supervised learning", in *Foundations and Trends in Computer Graphics and Vision*, Foundations and Trendső in Computer Graphics and Vision: Vol. 7: No 2-3, pp 81-227, 2-3. NOW Publishers, 2012, vol. 7, pp. 81–227. [Online]. Available: https://www.microsoft.com/en-us/research/publication/decision-forests-a-unified-framework-for-classification-regression-density-estimation-manifold-learning-and-semi-supervised-learning/.

[63] O. Bernard, A. Lalande, C. Zotti, F. Cervenansky, X. Yang, P. Heng, I. Cetin, K. Lekadir, O. Camara, M. A. Gonzalez Ballester, G. Sanroma, S. Napel, S. Petersen, G. Tziritas, E. Grinias, M. Khened, V. A. Kollerathu, G. Krishnamurthi, M. Rohé, X. Pennec, M. Sermesant, F. Isensee, P. Jäger, K. H. Maier-Hein, P. M. Full, I. Wolf, S. Engelhardt, C. F. Baumgartner, L. M. Koch, J. M. Wolterink, I. Igum, Y. Jang, Y. Hong, J. Patravali, S. Jain, O. Humbert, and P. Jodoin, "Deep learning techniques for automatic mri cardiac multi-structures segmentation and diagnosis: Is the problem solved?", *IEEE Transactions on Medical Imaging*, vol. 37, no. 11, pp. 2514–2525, 2018, ISSN: 0278-0062. DOI: doi:10.1109/TMI.2018.2837502.

[64] M. H. Moghari, *Miccai 2016 whole heart segmentation from cmr*, segchd.csail.mit.edu/, 2016.

[65] X. Zhuang, *Miccai 2017 multi-modality whole heart segmentation*, stacom2017.cardiacatlas.org/, 2017.

[66] A. Andreopoulos and J. K. Tsotsos, "Efficient and generalizable statistical models of shape and appearance for analysis of cardiac mri", *Medical Image Analysis*, vol. 12, no. 3, pp. 335 –357, 2008, ISSN: 1361-8415. DOI: https://doi.org/10.1016/j.media.2007.12.003. [Online]. Available: http://www.sciencedirect.com/science/article/pii/S1361841508000029.

[67] C. Toboz-Gomez, *Cardiac atlas challenges*, www.cardiacatlas.org/challenges/left-atrium-segmentation-challenge/, [Online], 2013.

[68] *Uk biobank*, www.ukbiobank.ac.uk, [Online], 2019.

[69] S. Ardekani, R. G. Weiss, A. C. Lardo, R. T. George, J. A. C. Lima, K. C. Wu, M. I. Miller, R. L. Winslow, and L. Younes, "Computational method for identifying and quantifying shape features of human left ventricular remodeling", *Annals of Biomedical Engineering*, vol. 37, no. 6, pp. 1043–1054, 2009, ISSN: 1573-9686. DOI: 10.1007/s10439-009-9677-2. [Online]. Available: https://doi.org/10.1007/s10439-009-9677-2.

[70] . Rodrigues, F. Morais, N. Morais, L. Conci, L. Neto, and A. Conci, "A novel approach for the automated segmentation and volume quantification of cardiac fats on computed tomography", *Computer Methods and Programs in Biomedicine*, vol. 123, pp. 109 – 128, 2016, ISSN: 0169-2607. DOI: https://doi.org/10.1016/j.cmpb.2015.09.017. [Online]. Available: http://www.sciencedirect.com/science/article/pii/S0169260715002448.

[71] L. Zhang and E. A. Geiser, "An effective algorithm for extracting serial endocardial borders from 2-dimensional echocardiograms", *IEEE Transactions on Biomedical Engineering*, vol. BME-31, no. 6, pp. 441–447, 1984, ISSN: 0018-9294.

[72] J. W. Klingler, C. L. Vaughan, T. D. Fraker, and L. T. Andrews, "Segmentation of echocardiographic images using mathematical morphology", *IEEE Transactions on Biomedical Engineering*, vol. 35, no. 11, pp. 925–934, 1988, ISSN: 0018-9294. DOI: 10.1109/10.8672.

[73] K. Saini and M. Rohit, *Ultrasound imaging and image segmentation in the area of ultrasound: A review*, 2010.

[74] M. Kass, A. Witkin, and D. Terzopoulos, "Snakes: Active contour models", *International Journal of Computer Vision*, vol. 1, no. 4, pp. 321–331, 1988, ISSN: 1573-1405. DOI: 10.1007/BF00133570. [Online]. Available: https://doi.org/10.1007/BF00133570.

[75] A. Mishra, P. K. Dutta, and M. K. Ghosh, "A ga based approach for boundary detection of left ventricle with echocardiographic image sequences", *Image Vision Comput.*, vol. 21, pp. 967–976, 2003.

[76] Y. Chen, F. Huang, H. D. Tagare, and M. Rao, "A coupled minimization problem for medical image segmentation with priors", *International Journal of Computer Vision*, vol. 71, no. 3, pp. 259–272, 2007, ISSN: 1573-1405. DOI: 10.1007/s11263-006-8524-2. [Online]. Available: https://doi.org/10.1007/s11263-006-8524-2.

[77] N. Lin, W. Yu, and J. S. Duncan, "Combinative multi-scale level set framework for echocardiographic image segmentation", in *Medical Image Computing and Computer-Assisted Intervention — MICCAI 2002*, T. Dohi and R. Kikinis, Eds., Berlin, Heidelberg: Springer Berlin Heidelberg, 2002, pp. 682–689, ISBN: 978-3-540-45786-2.

[78] O. Bernard, B. Touil, A. Gelas, R. Prost, and D. Friboulet, "A rbf-based multiphase level set method for segmentation in echocardiography using the statistics of the radiofrequency signal", in *2007 IEEE International Conference on Image Processing*, vol. 3, 2007, pp. III –157–III –160. DOI: 10.1109/ICIP.2007.4379270.

[79] D. Cremers, S. J. Osher, and S. Soatto, "Kernel density estimation and intrinsic alignment for shape priors in level set segmentation", *International Journal of Computer Vision*, vol. 69, no. 3, pp. 335–351, 2006, ISSN: 1573-1405. DOI: 10.1007/s11263-006-7533-5. [Online]. Available: https://doi.org/10.1007/s11263-006-7533-5.

[80] Q. Duan, E. D. Angelini, and A. F. Laine, "Real-time segmentation by active geometric functions", *Computer methods and programs in biomedicine*, vol. 98, no. 3, pp. 223–230, 2010, ISSN: 1872-7565. DOI: 10.1016/j.cmpb.2009.09.001. [Online]. Available: https://www.ncbi.nlm.nih.gov/pubmed/19800708.

[81] J. C. Nascimento and J. S. Marques, "Robust shape tracking with multiple models in ultrasound images", *IEEE Transactions on Image Processing*, vol. 17, no. 3, pp. 392–406, 2008, ISSN: 1057-7149. DOI: 10.1109/TIP.2007.915552.

[82] N. Friedland and D. Adam, "Automatic ventricular cavity boundary detection from sequential ultrasound images using simulated annealing", *IEEE Transactions on Medical Imaging*, vol. 8, no. 4, pp. 344–353, 1989, ISSN: 0278-0062. DOI: 10.1109/42.41487.

[83] I. Mikic, S. Krucinski, and J. D. Thomas, "Segmentation and tracking in echocardiographic sequences: Active contours guided by optical flow estimates", *IEEE Transactions on Medical Imaging*, vol. 17, no. 2, pp. 274–284, 1998, ISSN: 0278-0062. DOI: 10.1109/42.700739.

[84] M. Mulet-Parada and J. Noble, "2d+t acoustic boundary detection in echocardiography", *Medical Image Analysis*, vol. 4, no. 1, pp. 21 –30, 2000, ISSN: 1361-8415. DOI: https://doi.org/10.1016/S1361-8415(00)00006-2. [Online]. Available: http://www.sciencedirect.com/science/article/pii/S1361841500000062.

[85] M. Mignotte, J. Meunier, and J.-C. Tardif, "Endocardial boundary estimation and tracking in echocardiographic images using deformable template and markov random fields", *Pattern Analysis and Applications*, vol. 4, no. 4, pp. 256–271, 2001, ISSN: 1433-7541. DOI: 10.1007/PL00010988. [Online]. Available: https://doi.org/10.1007/PL00010988.

[86] G. Jacob, J. A. Noble, C. Behrenbruch, A. D. Kelion, and A. P. Banning, "A shape-space-based approach to tracking myocardial borders and quantifying regional left-ventricular function applied in echocardiography", *IEEE Transactions on Medical Imaging*, vol. 21, no. 3, pp. 226–238, 2002, ISSN: 0278-0062. DOI: 10.1109/42.996341.

[87] M. Bansal and R. R. Kasliwal, "How do i do it?: Speckle-tracking echocardiography", *Indian heart journal*, vol. 65, no. 1, pp. 117–123, 2013, ISSN: 0019-4832. DOI: 10.1016/j.ihj.2012.12.004. [Online]. Available: https://www.ncbi.nlm.nih.gov/pubmed/23438628.

[88] P. Collier, D. Phelan, and A. Klein, "A test in context: Myocardial strain measured by speckle-tracking echocardiography", *Journal of the American College of Cardiology*, vol. 69, no. 8, pp. 1043–1056, 2017, ISSN: 0735-1097. DOI: 10.1016/j.jacc.2016.12.012. eprint: http://www.onlinejacc.org/content/69/8/1043.full.pdf. [Online]. Available: http://www.onlinejacc.org/content/69/8/1043.

[89] S. Asbjorn, *Basic concepts in myocardial strain and strain rate*, folk.ntnu.no/stoylen/strainrate/Myocardial_strain, [Online], 2016.

[90] N. Paragios, M.-P. Jolly, M. Taron, and R. Ramaraj, "Active shape models and segmentation of the left ventricle in echocardiography", in *Scale Space and PDE Methods in Computer Vision*, R. Kimmel, N. A. Sochen, and J. Weickert, Eds., Berlin, Heidelberg: Springer Berlin Heidelberg, 2005, pp. 131–142, ISBN: 978-3-540-32012-8.

[91] G. Hamarneh and T. Gustavsson, "Combining snakes and active shape models for segmenting the human left ventricle in echocardiographic images", in *Computers in Cardiology 2000. Vol.27 (Cat. 00CH37163)*, 2000, pp. 115–118. DOI: 10.1109/CIC.2000.898469.

[92] G. Jacob, J. A. Noble, C. Behrenbruch, A. D. Kelion, and A. P. Banning, "A shape-space-based approach to tracking myocardial borders and quantifying regional left-ventricular function applied in echocardiography", *IEEE Transactions on Medical Imaging*, vol. 21, no. 3, pp. 226–238, 2002. DOI: 10.1109/42.996341.

[93] J. G. Bosch, S. C. Mitchell, B. P. F. Lelieveldt, F. Nijland, O. Kamp, M. Sonka, and J. H. C. Reiber, "Automatic segmentation of echocardiographic sequences by active appearance motion models", *IEEE Transactions on Medical Imaging*, vol. 21, no. 11, pp. 1374–1383, 2002, ISSN: 0278-0062. DOI: doi:10.1109/TMI.2002.806427.

[94] M. van Stralen, A. Haak, K. Y. E. Leung, G. van Burken, C. Bos, and J. G. Bosch, "Full-cycle left ventricular segmentation and tracking in 3d echocardiography using active appearance models", in *2015 IEEE International Ultrasonics Symposium (IUS)*, 2015, pp. 1–4. DOI: 10.1109/ULTSYM.2015.0389.

[95] V. Lempitsky, M. Verhoek, J. A. Noble, and A. Blake, "Random forest classification for automatic delineation of myocardium in real-time 3d echocardiography", in *Functional Imaging and Modeling of the Heart*, N. Ayache, H. Delingette, and M. Sermesant, Eds., Berlin, Heidelberg: Springer Berlin Heidelberg, 2009, pp. 447–456, ISBN: 978-3-642-01932-6.

[96] K. Keraudren, O. Oktay, W. Shi, J. Hajnal, and D. Rueckert, "Endocardial 3d ultrasound segmentation using autocontext random forests", Oct. 2014.

[97] F. Milletari, S.-A. Ahmadi, C. Kroll, C. Hennersperger, F. Tombari, A. Shah, A. Plate, K. Boetzel, and N. Navab, "Robust segmentation of various anatomies in 3d ultrasound using hough forests and learned data representations", in *Medical Image Computing and Computer-Assisted Intervention – MICCAI 2015*, N. Navab, J. Hornegger, W. M. Wells, and A. Frangi, Eds., Cham: Springer International Publishing, 2015, pp. 111–118, ISBN: 978-3-319-24571-3.

[98] J. Domingos, R. Stebbing, and A. Noble, "Endocardial segmentation using structured random forests in 3d echocardiography", Oct. 2014.

[99] T. Binder, M. Süssner, D. Moertl, T. Strohmer, H. Baumgartner, G. Maurer, and G. Porenta, "Artificial neural networks and spatial temporal contour linking for automated endocardial contour detection on echocardiograms: A novel approach to determine left ventricular contractile function", *Ultrasound in Medicine and Biology*, vol. 25, no. 7, pp. 1069 –1076, 1999, ISSN: 0301-5629. DOI: https://doi.org/10.1016/S0301-5629(99)00059-9. [Online]. Available: http://www.sciencedirect.com/science/article/pii/S0301562999000599.

[100] D. Barbosa, D. Friboulet, J. D'hooge, and O. Bernard, "Fast tracking of the left ventricle using global anatomical affine optical flow and local recursive block matching", 2014.

[101] E. Smistad and F. Lindseth, "Real-time tracking of the left ventricle in 3d ultrasound using kalman filter and mean value coordinates", 2014. DOI: 10.13140/2.1.1330.6888.

[102] M. V. Stralen, A. Haak, K. Leung, G. V. Burken, and J. Bosch, "Segmentation of multi-center 3d left ventricular echocardiograms by active appearance models", 2014.

[103] M. Bernier, P.-M. Jodoin, and A. Lalande, "Automatized evaluation of the left ventricular ejection fraction from echocardiographic images using graph cut", 2014.

[104] O. Oktay, W. Shi, K. Keraudren, J. Caballero, and D. Rueckert, "Learning shape representations for multi-atlas endocardium segmentation in 3d echo images", 2014.

[105] C. Wang and O. Smedby, "Model-based left ventricle segmentation in 3d ultrasound using phase image", 2014.

[106] S. Queirós, J. L. Vilaça, P. Morais, J. C. Fonseca, J. D'hooge, and D. Barbosa, "Fast left ventricle tracking using localized anatomical affine optical flow", *International Journal for Numerical Methods in Biomedical Engineering*, vol. 33, no. 11, e2871, 2017, e2871 cnm.2871. DOI: 10.1002/cnm.2871. eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1002/cnm.2871. [Online]. Available: https://onlinelibrary.wiley.com/doi/abs/10.1002/cnm.2871.

[107] J. D'Souza, *A trip to random forest*, medium.com/greyatom/a-trip-to-random-forest-5c30d8250d6a, [Online], 2018.

[108] P. Kontschieder, M. Fiterau, A. Criminisi, and S. Rota Bulo, "Deep neural decision forests", in *The IEEE International Conference on Computer Vision (ICCV)*, 2015.

[109] J. Gall, A. Yao, N. Razavi, L. Van Gool, and V. Lempitsky, "Hough forests for object detection, tracking, and action recognition", *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 33, no. 11, pp. 2188–2202, 2011, ISSN: 0162-8828. DOI: doi:10.1109/TPAMI.2011.70.

[110] P. Kontschieder, S. R. Bulò, M. Donoser, M. Pelillo, and H. Bischof, "Semantic image labelling as a label puzzle game", in *Proc. BMVC*, http://dx.doi.org/10.5244/C.25.111, 2011, pp. 111.1–111.12, ISBN: 1-901725-43-X.

[111] P. Dollár and C. L. Zitnick, "Fast edge detection using structured forests", *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 37, no. 8, pp. 1558–1570, 2015, ISSN: 0162-8828. DOI: doi:10.1109/TPAMI.2014.2377715.

[112] P. Dollár, *Piotr's computer vision matlab toolbox (pmt)*, github.com/pdollar/toolbox, 2015.

[113] A. Haak, G. Vegas-Sanchez-Ferrero, H. W. Mulder, B. Ren, H. A. Kirili, C. Metz, G. van Burken, M. van Stralen, J. P. W. Pluim, F. W. van der Steen, T. van Walsum, and J. G. Bosch, "Segmentation of multiple heart cavities in 3-d transesophageal ultrasound images", *IEEE Transactions on Ultrasonics, Ferroelectrics, and Frequency Control*, vol. 62, no. 6, pp. 1179–1189, 2015, ISSN: 0885-3010. DOI: doi:10.1109/TUFFC.2013.006228.

[114] N. Duta and M. Sonka, "Segmentation and interpretation of mr brain images. an improved active shape model", *IEEE Transactions on Medical Imaging*, vol. 17, no. 6, pp. 1049–1062, 1998, ISSN: 0278-0062. DOI: doi:10.1109/42.746716.

[115] M. van Stralen, K. Y. E. Leung, M. M. Voormolen, N. de Jong, A. F. W. van der Steen, J. H. C. Reiber, and J. G. Bosch, "Time continuous detection of the left ventricular long axis and the mitral valve plane in 3-d echocardiography", *Ultrasound in Medicine and Biology*, vol. 34, no. 2, pp. 196–207, 2008, ISSN: 0301-5629. DOI: doi:10.1016/j.ultrasmedbio.2007.07.016. [Online]. Available: https://doi.org/10.1016/j.ultrasmedbio.2007.07.016.

[116] O. Bernard, *Cetus challenge*, www.creatis.insa-lyon.fr/Challenge/CETUS/, [Online], 2014.

[117] A. Hidaka and T. Kurita, "Consecutive dimensionality reduction by canonical correlation analysis for visualization of convolutional neural networks", *Proceedings of the ISCIE International Symposium on Stochastic Systems Theory and its Applications*, vol. 2017, pp. 160–167, 2017. DOI: doi:10.5687/sss.2017.160.

[118] Y. LeCun, P. Haffner, L. Bottou, and Y. Bengio, "Object recognition with gradient-based learning", in *Shape, Contour and Grouping in Computer Vision*. Berlin, Heidelberg: Springer Berlin Heidelberg, 1999, pp. 319–345, ISBN: 978-3-540-46805-9. DOI: doi:10.1007/3-540-46805-6_19. [Online]. Available: https://doi.org/10.1007/3-540-46805-6_19.

[119] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "Imagenet classification with deep convolutional neural networks", *Commun. ACM*, vol. 60, no. 6, pp. 84–90, 2017, ISSN: 0001-0782. DOI: doi:10.1145/3065386. [Online]. Available: http://doi.acm.org/10.1145/3065386.

[120] F.-F. Li, *Convolutional neural networks for visual recognition*, cs231n.github.io/, [Online], 2019.

[121] N. Ibtehaz and M. S. Rahman, "Multiresunet : Rethinking the u-net architecture for multimodal biomedical image segmentation", *ArXiv*, vol. abs/1902.04049, 2019.

[122] S. Leclerc, *Hands on session 2*, deepimaging2019.sciencesconf.org/, [Online], 2019.

[123] A. Dertat, *Applied deep learning - part 4: Convolutional neural networks*, towardsdatascience.com/applied-deep-learning-part-4-convolutional-neural-networks-584bc134c1e2, [Online], 2017.

[124] S. Ruder, *An overview of gradient descent optimization algorithms*, ruder.io/optimizing-gradient-descent/, [Online], 2016.

[125] C. Desrosiers, *Basics of deep learning 2*, deepimaging2019.sciencesconf.org/, [Online], 2019.

[126] L. A. Santos, *Image segmentation*, leonardoaraujosantos.gitbooks.io/artificial-inteligence/content/image_segmentation.html, [Online], 2017.

[127] J. Duchi, E. Hazan, and Y. Singer, "Adaptive subgradient methods for online learning and stochastic optimization", *J. Mach. Learn. Res.*, vol. 12, pp. 2121–2159, 2011, ISSN: 1532-4435. [Online]. Available: http://dl.acm.org/citation.cfm?id=1953048.2021068.

[128] T. Tieleman and G. Hinton, *Lecture 6.5 rmsprop: divide the gradient by a running average of its recent magnitude*, COURSERA: Neural Networks for Machine Learning, 2012.

[129] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization", *CoRR*, vol. abs/1412.6980, 2014.

[130] N. Srivastava, G. E. Hinton, A. Krizhevsky, I. Sutskever, and R. Salakhutdinov, "Dropout: A simple way to prevent neural networks from overfitting", *J. Mach. Learn. Res.*, vol. 15, pp. 1929–1958, 2014.

[131] X. Glorot and Y. Bengio, "Understanding the difficulty of training deep feedforward neural networks", in *Proceedings of the Thirteenth International Conference on Artificial Intelligence and Statistics*, Y. W. Teh and M. Titterington, Eds., ser. Proceedings of Machine Learning Research, vol. 9, Chia Laguna Resort, Sardinia, Italy: PMLR, 2010, pp. 249–256. [Online]. Available: http://proceedings.mlr.press/v9/glorot10a.html.

[132] K. He, X. Zhang, S. Ren, and J. Sun, "Delving deep into rectifiers: Surpassing human-level performance on imagenet classification", in *2015 IEEE International Conference on Computer Vision (ICCV)*, 2015, pp. 1026–1034. DOI: doi:10.1109/ICCV.2015.123.

[133] J. Brownlee, *How to get reproducible results with keras*, machinelearningmastery.com/reproducible-results-neural-networks-keras/, 2019.

[134] S. Collange, D. Defour, S. Graillat, and R. Iakymchuk, "Numerical reproducibility for the parallel reduction on multi- and many-core architectures", *Parallel Computing*, vol. 49, pp. 83 –97, 2015, ISSN: 0167-8191. DOI: https://doi.org/10.1016/j.parco.2015.09.001. [Online]. Available: http://www.sciencedirect.com/science/article/pii/S0167819115001155.

[135] F. Chollet *et al.*, *Keras*, keras.io, 2015.

[136] P. Jodoin *et al.*, *Vitalab public git repo*, bitbucket.org/vitalab/vitalabai_public/src/master/, 2019.

[137] O. M.Loli *et al.*, *Medpy*, github.com/loli/medpy, 2019.

[138] R. L. Figueroa, Q. Zeng-Treitler, S. Kandula, and L. H. Ngo, "Predicting sample size required for classification performance", *BMC Medical Informatics and Decision Making*, vol. 12, no. 1, p. 8, 2012, ISSN: 1472-6947. DOI: doi:10.1186/1472-6947-12-8. [Online]. Available: https://doi.org/10.1186/1472-6947-12-8.

[139] J. A. Durlak, "How to select, calculate, and interpret effect sizes", *Journal of Pediatric Psychology*, vol. 34, no. 9, pp. 917–928, 2009, ISSN: 0146-8693. DOI: doi:10.1093/jpepsy/jsp004. eprint: http://oup.prod.sis.lan/jpepsy/article-pdf/34/9/917/2768854/jsp004.pdf. [Online]. Available: https://doi.org/10.1093/jpepsy/jsp004.

[140] J. Brownlee, *Impact of dataset size on deep learning model skill and performance estimates*, machinelearningmastery.com/impact-of-dataset-size-on-deep-learning-model-skill-and-performance-estimates/, [Online], 2019.

[141] R. F. Woolson, "Wilcoxon signed-rank test", in *Wiley Encyclopedia of Clinical Trials*. American Cancer Society, 2008, pp. 1–3, ISBN: 9780471462422. DOI: 10.1002/9780471462422.eoct979. eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1002/9780471462422.eoct979. [Online]. Available: https://onlinelibrary.wiley.com/doi/abs/10.1002/9780471462422.eoct979.

[142] Wikipedia, *Mann-whitney u test*, en.wikipedia.org/wiki/MannWhitney_U_test, [Online], 2019.

[143] G. S. Guide, *Interpreting the p value*, www.graphpad.com/guides/prism/7/statistics/stat_interpreting_results_wilcoxon_.htm?toc=0&printWindow, [Online], 2019.

[144] S. Greenland, S. J. Senn, K. J. Rothman, J. B. Carlin, C. Poole, S. N. Goodman, and D. G. Altman, "Statistical tests, p values, confidence intervals, and power: A guide to misinterpretations", *European Journal of Epidemiology*, vol. 31, no. 4, pp. 337–350, 2016, ISSN: 1573-7284. DOI: 10.1007/s10654-016-0149-3. [Online]. Available: https://doi.org/10.1007/s10654-016-0149-3.

[145] S. Ioffe and C. Szegedy, "Batch normalization: Accelerating deep network training by reducing internal covariate shift", *ArXiv*, vol. abs/1502.03167, 2015.

[146] S. Santurkar, D. Tsipras, A. Ilyas, and A. Madry, "How does batch normalization help optimization?(no, it is not about internal covariate shift)", *ArXiv preprint arXiv:1805.11604*, 2018. [Online]. Available: https://papers.nips.cc/paper/7515-how-does-batch-normalization-help-optimization.

[147] J. Pedrosa, D. Barbosa, B. Heyde, F. Schnell, A. Rösner, P. Claus, and J. Dhooge, "Left ventricular myocardial segmentation in 3-d ultrasound recordings: effect of different endocardial and epicardial coupling strategies", *IEEE Transactions on Ultrasonics, Ferroelectrics, and Frequency Control*, vol. 64, no. 3, pp. 525–536, 2017.

[148] L. Wang, C.-Y. Lee, Z. Tu, and S. Lazebnik, "Training deeper convolutional networks with deep supervision", *ArXiv*, vol. abs/1505.02496, 2015.

[149] X. Li, H. Chen, X. Qi, Q. Dou, C. Fu, and P. Heng, "H-denseunet: Hybrid densely connected unet for liver and tumor segmentation from ct volumes", *IEEE Transactions on Medical Imaging*, vol. 37, no. 12, pp. 2663–2674, 2018. DOI: 10.1109/TMI.2018.2845918.

[150] D. Fourure, R. Emonet, É. Fromont, D. Muselet, A. Trémeau, and C. Wolf, "Residual conv-deconv grid network for semantic segmentation", *ArXiv*, vol. abs/1707.07958, 2017.

[151] Z. Zhou, *Repository for the code of u-net ++*, github . com / MrGiovanni / UNetPlusPlus, [Online], 2019.

[152] D. M. Vigneault, W. Xie, C. Y. Ho, D. A. Bluemke, and J. A. Noble, "-net (omega-net): Fully automatic, multi-view cardiac mr detection, orientation, and segmentation with deep neural networks", *Medical image analysis*, vol. 48, pp. 95–106, 2017.

[153] S.-Y. Huang, J. M. Boone, K. Yang, N. J. Packard, S. E. McKenney, N. D. Prionas, K. K. Lindfors, and M. J. Yaffe, "The characterization of breast anatomical metrics using dedicated breast ct.", *Medical physics*, vol. 38 4, pp. 2180–91, 2011.

[154] F. Wang, M. Jiang, C. Qian, S. Yang, C. Li, H. Zhang, X. Wang, and X. Tang, "Residual attention network for image classification", in *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2017, pp. 6450–6458.

[155] S. Ren, K. He, R. Girshick, and J. Sun, "Faster r-cnn: Towards real-time object detection with region proposal networks", in *Advances in Neural Information Processing Systems 28*, 2015, pp. 91–99.

[156] J. Redmon, S. Divvala, R. Girshick, and A. Farhadi, "You only look once: Unified, real-time object detection", in *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016, pp. 779–788.

[157] K. He, G. Gkioxari, P. Dollár, and R. Girshick, "Mask r-cnn", in *2017 IEEE International Conference on Computer Vision (ICCV)*, 2017, pp. 2980–2988.

[158] C. Payer, D. Štern, H. Bischof, and M. Urschler, "Multi-label whole heart segmentation using cnns and anatomical label configurations", in *Statistical Atlases and Computational Models of the Heart. ACDC and MMWHS Challenges*, M. Pop, M. Sermesant, P.-M. Jodoin, A. Lalande, X. Zhuang, G. Yang, A. Young, and O. Bernard, Eds., 2018, pp. 190–198.

[159] Q. Guan and Y. Huang, "Multi-label chest x-ray image classification via category-wise residual attention learning", *Pattern Recognition Letters*, 2018.

[160] Y. Wang, Z. Deng, X. Hu, L. Zhu, X. Yang, X. Xu, P.-A. Heng, and D. Ni, "Deep attentional features for prostate segmentation in ultrasound", in *Medical Image Computing and Computer Assisted Intervention – MICCAI 2018*, A. F. Frangi, J. A. Schnabel, C. Davatzikos, C. Alberola-López, and G. Fichtinger, Eds., Cham: Springer International Publishing, 2018, pp. 523–530.

[161] C. Li, Q. Tong, X. Liao, W. Si, Y. Sun, Q. Wang, and P.-A. Heng, "Attention based hierarchical aggregation network for 3d left atrial segmentation", in *Statistical Atlases and Computational Models of the Heart. Atrial Segmentation and LV Quantification Challenges*, M. Pop, M. Sermesant, J. Zhao, S. Li, K. McLeod, A. Young, K. Rhode, and T. Mansi, Eds., Cham: Springer International Publishing, 2019, pp. 255–264.

[162] D. M. Vigneault, W. Xie, C. Y. Ho, D. A. Bluemke, and J. A. Noble, "-net (omega-net): Fully automatic, multi-view cardiac mr detection, orientation, and segmentation with deep neural networks", *Medical Image Analysis*, vol. 48, pp. 95 –106, 2018, ISSN: 1361-8415.

[163] M. Jaderberg, K. Simonyan, A. Zisserman, and k. kavukcuoglu, "Spatial transformer networks", in *Advances in Neural Information Processing Systems 28*, 2015, pp. 2017–2025.

[164] E. Pesce, S. J. Withey, P.-P. Ypsilantis, R. Bakewell, V. Goh, and G. Montana, "Learning to detect chest radiographs containing pulmonary lesions using visual attention networks", *Medical Image Analysis*, vol. 53, pp. 26 –38, 2019, ISSN: 1361-8415.

[165] J. Schlemper, O. Oktay, M. Schaap, M. Heinrich, B. Kainz, B. Glocker, and D. Rueckert, "Attention gated networks: Learning to leverage salient regions in medical images", *Medical Image Analysis*, vol. 53, pp. 197 –207, 2019, ISSN: 1361-8415.

[166] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition", *CoRR*, vol. abs/1409.1556, 2014.

[167] A. MeidellFiorito, A. Østvik, E. Smistad, S. Leclerc, O. Bernard, and L. Lovstakken, "Detection of cardiac events in echocardiography using 3d convolutional recurrent neural networks", in *2018 IEEE International Ultrasonics Symposium (IUS)*, 2018, pp. 1–4. DOI: 10.1109/ULTSYM.2018.8580137.

[168] E. Smistad, A. Østvik, I. M. Salte, S. Leclerc, O. Bernard, and L. Lovstakken, "Fully automatic real-time ejection fraction and mapse measurements in 2d echocardiography using deep neural networks", *2018 IEEE International Ultrasonics Symposium (IUS)*, pp. 1–4, 2018.

FOLIO ADMINISTRATIF

THESE DE L'UNIVERSITE DE LYON OPEREE AU SEIN DE L'INSA LYON

NOM : LECLERC

DATE de SOUTENANCE : 11/12/2019

Prénoms : Sarah Marie-Solveig

TITRE : Automatisation de la segmentation sémantique de structures cardiaques en imagerie ultrasonore par apprentissage supervisé

NATURE : Doctorat

Numéro d'ordre : 2019LYSEI121

Ecole doctorale : EDA 160 : Electronique, Electrotechnique, Automatique (EEA)

Spécialité: Traitement du Signal et de l'Image

RESUME : La segmentation des structures du cœur en imagerie ultrasonore est rendue difficile par la faible qualité des images, notamment le manque de frontières nettes dans les images échocardiographiques 2D et 3D. Cette modalité temps-réelle et bas coût est néanmoins la plus utilisée de nos jours pour contrôler l'état des patients et réaliser le diagnostic clinique. L'état de l'art révèle que pour pallier au manque d'information dans les images, les méthodes les plus performantes jusqu'alors recouraient à l'intégration d'informations a priori sur les formes recherchées, ce qui réduit le potentiel d'adaptation de l'algorithme, ou à l'identification manuelle de points clés, ce qui rend le processus non reproductible. Dans cette thèse, nous proposons plusieurs algorithmes originaux de segmentation d'images échocardiographiques à base d'apprentissage supervisé, où la résolution du problème est automatiquement construite à l'aide de données résolues par des cardiologues experts. Grâce à la construction d'une base de données et d'une plateforme d'évaluation dédiées au projet, nous prouvons le fort potentiel clinique des méthodes par apprentissage profond, ainsi que la possibilité de rendre ces méthodes plus robustes encore en incorporant la détection automatique de régions d'intérêt.

MOTS-CLÉS : Segmentation, échocardiographie, apprentissage supervisé, apprentissage machine, apprentissage profond, évaluation, métriques anatomiques, multi-structure, multitâche, mécanisme d'attention, base de donnée, science ouverte.

Laboratoire (s) de recherche : Creatis

Directeurs de thèse: Carole LARTIZIEN et Olivier BERNARD

Président de jury : A établir par le jury

Composition du jury : Jean-Philippe THIRAN, Mireille GARREAU, Daniel RUECKERT, Alison NOBLE, Carole LARTIZIEN, Olivier BERNARD, Pierre-Marc JODOIN, Thomas GRENIER