Automatic analysis of cardiac function with artificial intelligence: multimodal approach for portable echocardiographic devices

Yingyu Yang

UNIVERSITÉ CÔTE D'AZUR

ÉCOLE DOCTORALE
SCIENCES ET TECHNOLOGIES DE L'INFORMATION ET DE LA COMMUNICATION

THÈSE DE DOCTORAT

Analyse automatique de la fonction cardiaque par intelligence artificielle: approche multimodale pour un dispositif d'échocardiographie portable

YingyuYANG

Centre Inria d'Université Côte d'Azur, Équipe Epione

Thèse dirigée par Maxime Sermesant et co-dirigée par Pamela Moceri

Soutenue le 19 Decembre 2023

Présentée en vue de l'obtention du grade de Docteur en Automatique, Traitement du Signal et des Images d'Université Côte d'Azur.

Devant le jury composé de :

Olivier Bernard	INSA Lyon	Président
Andrew King	King's College London	Rapporteur
Julia Schnabel	Technical University of Munich	Rapporteur
Maria A. Zuluaga	EURECOM	Examinateur
Mehdi Benchoufi	echOpen factory	Examinateur
Pamela Moceri	Centre Hospitalier Universitaire de Nice	Co-directrice de thèse
Maxime Sermesant	Centre Inria d'Université Côte d'Azur	Directeur de thèse

Inria

UNIVERSITÉ CÔTE D'AZUR

chu de Nice

Membres du Jury

Titre français
Analyse automatique de la fonction cardiaque par intelligence artificielle: approche multimodale pour un dispositif d'échocardiographie portable

Titre anglais
Automatic analysis of cardiac function with artificial intelligence: multimodal approach for portable echocardiographic devices

Devant le jury composé de :

Président du jury
Olivier BERNARD, Professeur, INSA Lyon

Rapporteurs

Andrew KING, Reader, King's College London
Julia SCHNABEL, Professeure, Technical University of Munich

Examinateurs

Maria A. ZULUAGA, Professeure associée, EURECOM
Mehdi BENCHOUFI, Médecin de santé Publique, Docteur, echOpen factory
Pamela MOCERI, Professeure des universités - praticienne hospitalière, Centre Hospitalier Universitaire de Nice
Maxime SERMESANT, Directeur de recherche - HDR, Centre Inria d'Université Côte d'Azur

Résumé

Selon le rapport annuel de la Fédération Mondiale du Cœur de 2023, les maladies cardiovasculaires (MCV) représentaient près d'un tiers de tous les décès mondiaux en 2021. Comparativement aux pays à revenu élevé, plus de 80% des décès par MCV surviennent dans les pays à revenu faible et intermédiaire. La répartition inéquitable des ressources de diagnostic et de traitement des MCV demeure toujours non résolue. Face à ce défi, les dispositifs abordables d'échographie de point de soins (POCUS) ont un potentiel significatif pour améliorer le diagnostic des MCV. Avec l'aide de l'intelligence artificielle (IA), le POCUS permet aux non-experts de contribuer, améliorant ainsi largement l'accès aux soins, en particulier dans les régions moins desservies.

L'objectif de cette thèse est de développer des algorithmes robustes et automatiques pour analyser la fonction cardiaque à l'aide de dispositifs POCUS, en mettant l'accent sur l'échocardiographie et l'électrocardiogramme. Notre premier objectif est d'obtenir des caractéristiques cardiaques explicables à partir de chaque modalité individuelle. Notre deuxième objectif est d'explorer une approche multimodale en combinant les données d'échocardiographie et d'électrocardiogramme.

Nous commençons par présenter deux nouvelles structures d'apprentissage profond (DL) pour la segmentation de l'échocardiographie et l'estimation du mouvement. En incorporant des connaissance a priori de forme et de mouvement dans les modèles DL, nous démontrons, grâce à des expériences approfondies, que de tels a priori contribuent à améliorer la précision et la généralisation sur différentes séries de données non vues. De plus, nous sommes en mesure d'extraire la fraction d'éjection du ventricule gauche (FEVG), la déformation longitudinale globale (GLS) et d'autres indices utiles pour la détection de l'infarctus du myocarde (IM).

Ensuite, nous proposons un modèle DL explicatif pour la décomposition non supervisée de l'électrocardiogramme. Ce modèle peut extraire des informations explicables liées aux différentes sous-ondes de l'ECG sans annotation manuelle. Nous appliquons ensuite ces paramètres à un classificateur linéaire pour la détection de l'infarctus du myocarde, qui montre une bonne généralisation sur différentes séries de données.

Enfin, nous combinons les données des deux modalités pour une classification multimodale fiable. Notre approche utilise une fusion au niveau de la décision intégrant de

l'incertitude, permettant l'entraînement avec des données multimodales non appariées. Nous évaluons ensuite le modèle entraîné à l'aide de données multimodales appariées, mettant en évidence le potentiel de la détection multimodale de l'IM surpassant celle d'une seule modalité.

Dans l'ensemble, nos algorithmes proposés robustes et généralisables pour l'analyse de l'échocardiographie et de l'ECG démontrent un potentiel significatif pour l'analyse de la fonction cardiaque portable. Nous anticipons que notre cadre pourrait être davantage validé à l'aide de dispositifs portables du monde réel.

Mots-clés: Analyse de la fonction cardiaque, Apprentissage profond, Segmentation de l'échocardiographie, Suivi du mouvement en échocardiographie, Décomposition de l'électrocardiogramme, Apprentissage multimodal, Apprentissage profond avec incertitude

Abstract

According to the 2023 annual report of the World Heart Federation, cardiovascular diseases (CVD) accounted for nearly one third of all global deaths in 2021. Compared to high-income countries, more than 80% of CVD deaths occurred in low and middle-income countries. The inequitable distribution of CVD diagnosis and treatment resources still remains unresolved. In the face of this challenge, affordable point-of-care ultrasound (POCUS) devices demonstrate significant potential to improve the diagnosis of CVDs. Furthermore, by taking advantage of artificial intelligence (AI)-based tools, POCUS enables non-experts to help, thus largely improving the access to care, especially in less-served regions.

The objective of this thesis is to develop robust and automatic algorithms to analyse cardiac function for POCUS devices, with a focus on echocardiography (ECHO) and electrocardiogram (ECG). Our first goal is to obtain explainable cardiac features from each single modality respectively. Our second goal is to explore a multi-modal approach by combining ECHO and ECG data.

We start by presenting two novel deep learning (DL) frameworks for echocardiography segmentation and motion estimation tasks, respectively. By incorporating shape prior and motion prior into DL models, we demonstrate through extensive experiments that such prior can help improve the accuracy and generalises well on different unseen datasets. Furthermore, we are able to extract left ventricle ejection fraction (LVEF), global longitudinal strain (GLS) and other useful indices for myocardial infarction (MI) detection.

Next, we propose an explainable DL model for unsupervised electrocardiogram decomposition. This model can extract interpretable information related to different ECG subwaves without manual annotation. We further apply those parameters to a linear classifier for myocardial infarction detection, which showed good generalisation across different datasets.

Finally, we combine data from both modalities together for trustworthy multi-modal classification. Our approach employs decision-level fusion with uncertainty, allowing training with unpaired multi-modal data. We further evaluate the trained model using paired multi-modal data, showcasing the potential of multi-modal MI detection to surpass that from a single modality.

Overall, our proposed robust and generalisable algorithms for ECHO and ECG analysis demonstrate significant potential for portable cardiac function analysis. We anticipate that our novel framework could be further validated using real-world portable devices. We envision that such advanced integrative tools may significantly contribute towards better identification of CVD patients.

Keywords: Cardiac function analysis, Deep learning, Echocardiography segmentation, Echocardiography motion tracking, Electrocardiogram decomposition, Multi-modal learning, Deep learning with uncertainty

Acknowledgements

First and foremost, I would like to express my gratitude to my supervisor, Maxime Sermesant. Since my first internship with you, your trust and encouragement have been invaluable in shaping my confidence and interest in research. Your positive attitude towards research has left a lasting impact. I sincerely appreciate your support and guidance for each of my research projects during my thesis.

I am also deeply thankful to my co-supervisor, Pamela Moceri. Despite your busy schedule as a cardiologist, you have always been there, providing valuable feedback on my projects and helping me prepare clinical data for evaluation (Thank you Marie for your help in multi-modal data construction!).

I extend my appreciation to each member of the jury for agreeing to evaluate my work. Special thanks to Prof. Julia Schnabel and Prof. Andrew King for your dedicated commitment to reviewing my thesis and offering insightful feedback. Thank you, Prof. Oliver Bernard, as the president of the jury, for your continuous attention and constructive comments, which have greatly contributed to the improvement of my work. Prof. Maria A. Zuluaga, your feedback for my defense has been inspiring. Dr. Mehdi Benchoufi, your initiative for affordable and accessible ultrasound has been crucial to the establishment of this thesis.

I would like to thank all the senior researchers and young colleagues in the Epione team. The collaborative and supportive work environment you have created is unparalleled, and I am grateful for the journey we've shared. Even as we may be apart, I look forward to continuing our collaboration in building a better world through medical imaging and modeling for improved diagnosis and therapy.

To my friends from the music band "At the moment," graduating together is a special milestone. I hope we can continue playing music together in the future.

Lastly, I want to express my gratitude to my family for their everlasting support in my decision to study in France, pursue a PhD degree, and work in research. A special thank you to my husband, Enlin, for your companionship and love throughout this challenging thesis journey. Though it hasn't been easy, I believe that with you, we can conquer even greater challenges in life.

Financial Support

The research leading to these results has been supported by the Inria PHD funding and the National Research Agency (ANR) 3IA Côte d'Azur (ANR-19-P3IA-0002). The project was also supported by the Inria Sophia Antipolis - Méditerranée, "NEF" computation cluster.

Contents

Clinical

CVD	Cardiovascular Diseases
POCUS	Point-of-care Ultrasound
AHA	American Heart Association
MI	Myocardial Infarction
AMI	Anterior Myocardial Infarction
IMI	Inferior Myocardial Infarction
LMI	Lateral Myocardial Infarction
ECG	Electrocardiogram
ECHO	Echocardiography
ENDO	Endocardium
EPI	Epicardium
MYO	Myocardium
EF	Ejection Fraction
LV	Left Ventricle
LVEF	Left Ventricular Ejection Fraction
LVEDV	Left Ventricular End-diastole Volume
LVESV	Left Ventricular End-systole Volume

Methodology

AI	Artificial Intelligence
ML	Machine Learning
DL	Deep Learning
CNN	Convolutional Neural network
CVAE	Conditional Variational Autoencoder
FCN	Fully Connnected Network
AUROC	Area Under Reveiver Operator Curve
HD	Hausdorff Distance
MSD	Mean Surface Distance
LR	Logistic Regression
SVM	Support Vector Machine

1 Introduction

Contents

This thesis explores how AI-based models could help automatic cardiac function analysis for non-expert practitioners, using portable modalities such as echocardiography and electrocardiogram.

1.1 Thesis context

1.1.1 Point-of-care Ultrasound (POCUS)

For the past decades, cardiovascular diseases (CVDs) have stood as a significant global disease burden, witnessing a doubling in prevalent cases from 1990 to 2019 [Roth, 2020]. Addressing the prevention and management of CVD patients requires effective diagnostic methods, with point-of-care ultrasound (POCUS) proving to enhance the diagnostic landscape [King, 2016]. However, the intricacies of echocardiography analysis demand years of medical training and experience from cardiologists in order to provide accurate

diagnostic. Unfortunately, the poorly balanced distribution of cardiovascular medical professionals also hinders the CVD diagnostics and management [Narang, 2016]. To tackle this challenge, automatic cardiac function analysis emerges as a potential solution. With the help of Artificial Intelligence (AI), POCUS could add more benefits for cardiac diagnosis and decision making. One study showed that medical students using a hand-held ultrasound device with the aid of AI, can achieve very good inter-rater reliability when evaluating ejection fraction from 4-chamber view echocardiography [Dadon, 2020]. Another study demonstrated that AI-enabled POCUS movement guidance can help novices to acquire high-quality echocardiography [Cheema, 2021] and improve trainees' confidence in acquisition [Waldman, 2022]. AI-enabled POCUS not only can improve the workflow of cardiologists but also enables trainees/nurses to help, facilitating the usage of POCUS for patient care at a large scale.

1.1.2 echOpen Project

There are already many market players who are trying to develop POCUS for efficient diagnosis . However, their high cost may be a burden for public use at large scale. echOpen, which aims at developing open-source and low-cost hand-held ultrasound device, will become a strong game changer in the landscape of POCUS. The echOpen device enhanced by AI will simplify and improve health-care in different scenarios, for example, at home, in health centers, elderly houses. It can also be used by non medical doctors especially in an under-served area.

The goal of echOpen is to achieve widespread availability of ultrasound imaging. This objective is pursued by harnessing cost-effective Point-of-Care Ultrasound (POCUS) technology in conjunction with AI tools.

1.1.3 Thesis motivation

This thesis was inspired from our collaboration with echOpen, which started in 2020. Our objective is to explore AI algorithms for portable cardiac function analysis that may be deployed with portable ultrasound devices such as those from echOpen project. In the meantime, we focus on exploring the possibility of combining portable ultrasound and wearable electrocardiogram devices for more robust diagnosis or cardiac function evaluation from a multi-modal point-of-view.

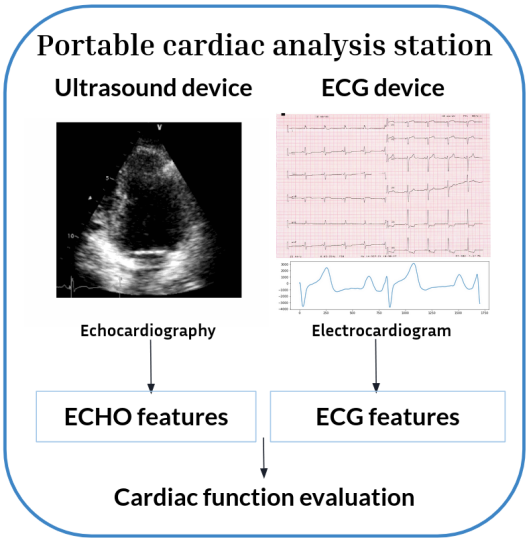


Fig. 1.1.: Thesis motivation and aim: explore AI algorithms for portable cardiac function analysis

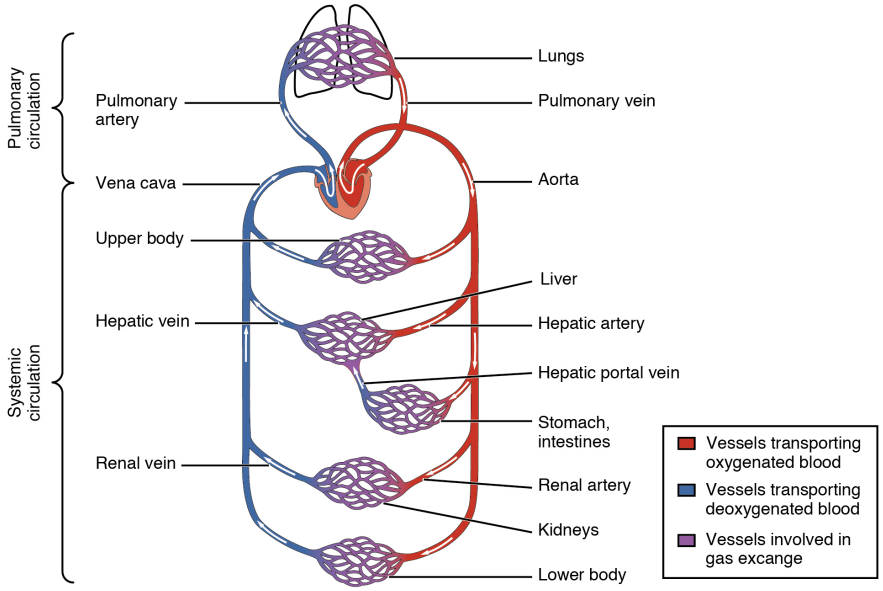


Fig. 1.2.: An overview of cardiovascular circulation system. (Illustration by Open Stax Anatomy & Physiology, CC-BY license.)

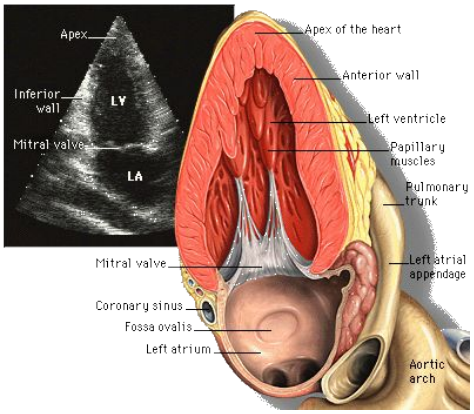
1.2 Clinical context

Why the heart is so important? The heart is the central element of the human's circulation system, the engine that supports body activity. It can be regarded functionally as two blood pumps, with the left heart and the right heart working jointly in order to maintain the systemic circulation and pulmonary circulation, respectively. In particular, the left ventricle (LV) ejects blood into the aorta (Ao), which transports oxygenated blood to all organs involved in the arterial system. Systemic veins receive deoxygenated blood to entry the right atrium (RA). The right ventricle (RV) then pumps blood into pulmonary

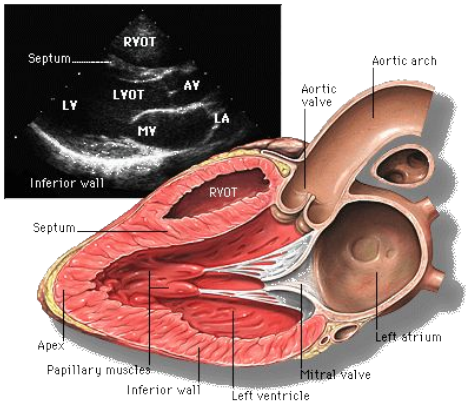
artery (PA). Oxygen and carbon dioxide get exchanged during pulmonary circulation. The oxygenated blood passes through the left atrium (LA) and arrives at the left ventricle for another systemic circulation.

1.2.1 Echocardiography

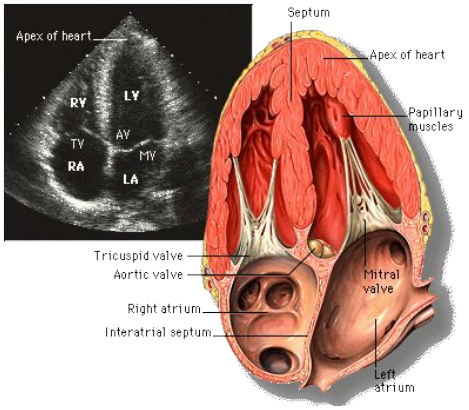
2D echocardiography (ECHO) is a very useful tool to view the heart in real-time. The standard machine to obtain 2D echocardiograhy is the phased array ultrasound transducer. At the front-end of the transducer, it contains thousands of piezoelectric crystals, which are able to convert electric currents to ultrasound waves and vice versa. Once the reflected waves are received, the vibration will be converted to electric currents and sent for image reconstruction. The standard examination by echocardiography includes



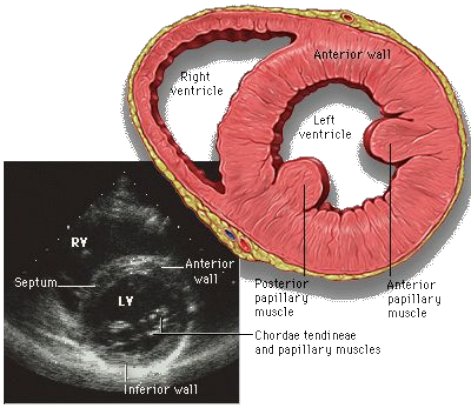
(a) Apical 2-chamber view



(b) Parasternal long-axis view



(c) Apical 4-chamber view



(d) Parasternal short-axis view (middle left ventricle)

Fig. 1.3.: Example views of echocardiography. (Illustration by Patrick J. Lynch and C. Carl Jaffe from Wikipedia commons, CC-BY license.)

different imaging windows and imaging views, whose choice are based on the targeted structure of the heart. For example, as the left ventricle plays an important role in blood bumping, echocardiography is often used to observe left ventricle wall motion and to detect abnormalities by conducting apical 4-chamber, 2-chamber view, long axis view, mid/basal short axis view etc. (Figure 1.3) The most commonly extracted parameter from echocardiography is the left ventricular ejection fraction (LVEF), an important clinical index. This index is computed from the measurement of the left ventricle volume from 4-chamber and 2-chamber apical view, jointly. Ejection fraction is the percentage of blood ejected by LV during one heart beat, which reflects the effectiveness of the LV for pumping blood into the systemic circulation. Using recent techniques (e.g. speckle tracking echocardiography) one can calculate several myocardial deformation related parameters, such as global longitudinal strain (GLS) which represents the percentage of longitudinal shortening at peak-systole. Studies have shown that GLS is able to detect more subtle changes in myocardial function compared to LVEF [Kalam, 2014; Morris, 2014].

1.2.2 Electrocardiogram

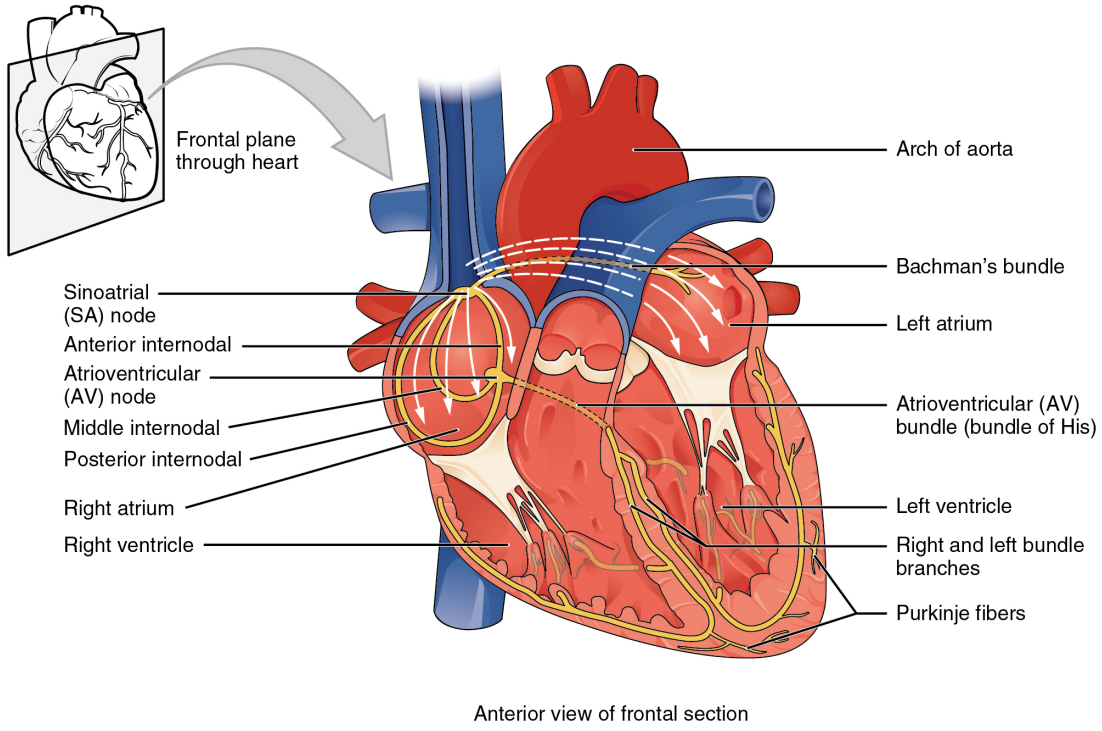


Fig. 1.4.: An overview of the conduction system of the heart. (Illustration by Open Stax Anatomy & Physiology, CC-BY license.)

A healthy heartbeat is controlled by a harmonized series of contractions of the four heart chambers, which depend on the electrical conduction within the heart. The sinoatrial (SA) node (i.e., the primary pacemaker site of the heart) generates an action potential wave that is rapidly propagated across the atria. Active potential waves then arrive at

the atrioventricular (AV) node, which is normally the only pathway for them to enter the ventricles. The propagation rate slows down at the AV node so that the atria has enough time for depolarization and contraction, letting atria blood entering the ventricles. Next, the action potentials propagate through the AV node to the ventricle base via the bundle of His, then through left and right bundle branches on the septum. These specialized fibers have high conduction speeds (2 m/s), and branch out into Purkinje fibers, which rapidly transmit impulses (4 m/s) throughout the ventricles, culminating in a direct connection with ventricular myocytes for final conduction at cellular level [Klabunde, 2011]. A well-functioning conduction system ensures rapid and near-synchronized depolarization and contraction of cardiac myocytes, which is essential for the normal ejection of ventricles.

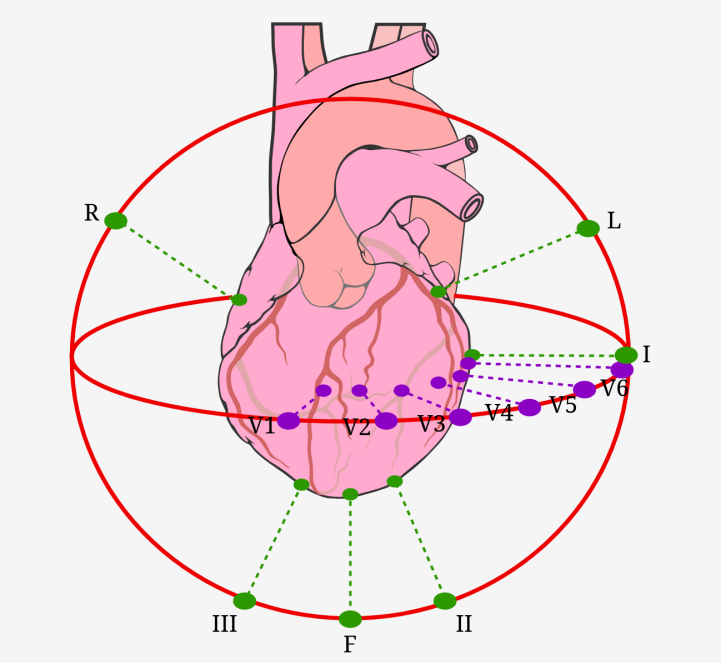


Fig. 1.5.: 12-lead electrocardiogram lead axes. Illustration by Wikipedia Commons, CC-BY-SA license.

The electrocardiogram (ECG) is an important tool to detect abnormalities in the electrical conduction within the heart. As the body tissues are able to conduct electrical currents that are generated from the heart, ECG can be recorded by electrodes located on the body's surface. The standard ECG used in clinics usually refers to the 12-lead ECG, and different leads record the cardiac electrical activity from different angles. In particular, the limb leads I, II, III, aVR, aVL and aVF (green points in Figure 1.5) record the electrical currents in the frontal plane of the heart. Precordial chest leads V1-V6 (purple leads in Figure 1.5) measure the electrical activity in the horizontal plane, which is perpendicular to the frontal plane. By analyzing the ECG signals, cardiologists are capable to identify abnormal conduction activity and select patients for further diagnostic or treatment. A single-lead ECG signal (for example those acquired from wearable ECG device), can easily detect common rhythm or conduction related disorders, but are less informative

for structural problems, such as myocardial infarction, ventricular hypertrophy and so on [Witvliet, 2021].

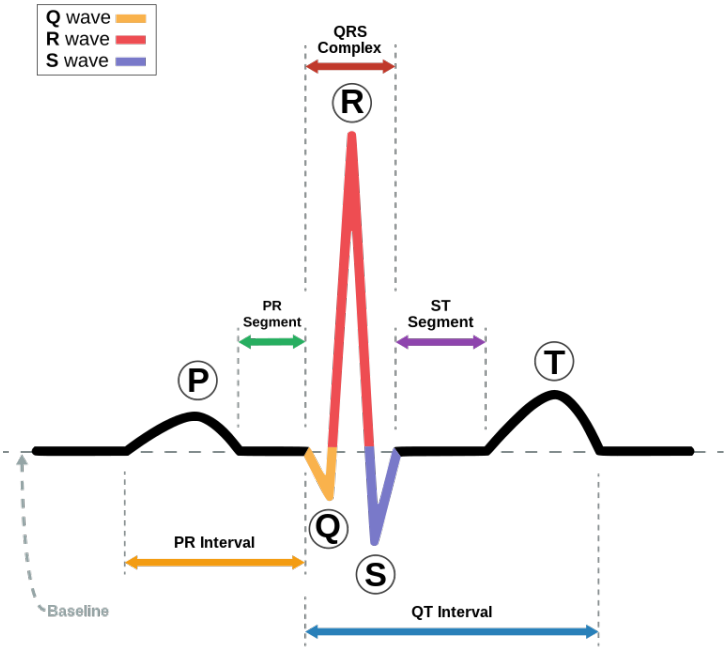


Fig. 1.6.: A typical example of ECG signal. Illustration by Agateller from Wikimedia commons, Public domain.

A typical ECG trace of one single heartbeat consists of: P wave; QRS complex; and, T wave (Figure 1.6). These sequences represent the atrial depolarization, ventricular depolarization and ventricular repolarization, respectively. The morphology of different ECG components plays an important role in diagnosing cardiac pathologies, such as myocardial ischemia, infarction etc.

1.2.3 Myocardial infarction

Coronary arteries are responsible for blood delivery to the heart muscle, which is crucial for a normal functioning of the heart. According to a national survey in the United States, coronary related heart disease accounts for near 4.9% in adults[1].Myocardial infarction (MI), also known as heart attack, is the most significant contributor to sudden cardiac death. It's reported that there are around 803,000 people encounter a heart attack in the United States every year [Tsao, 2023]. MI occurs due to a gradual accumulation of atherosclerotic plaque within a coronary artery. Eventually, this plaque may suddenly rupture, leading to the rapid formation of a clot (Fig.1.7) that completely blocks the artery. This blockage, in turn, disrupts the flow of blood to the heart muscle, resulting in severe tissue damage due to lack of oxygen supply.

[1] https://wwwn.cdc.gov/NHISDataQueryTool/SHS_adult/index.html

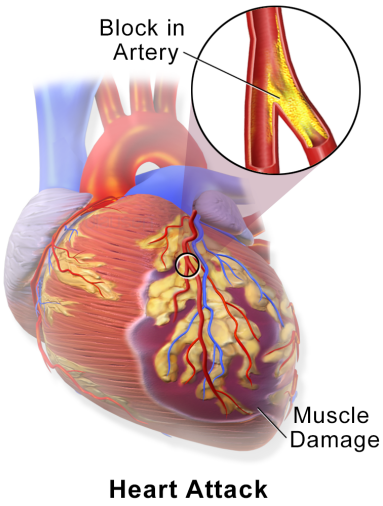


Fig. 1.7.: Myocardial infarction. Illustration by Blausen Medical Communications, Inc. from Wikimedia commons, CC-BY license.

A patient is diagnosed with MI if he/she presents with elevated cardiac troponin values and also falls into at least one of the following conditions: symptoms of myocardial ischaemia; new changes of ST-segment/T-wave in ECG; development of pathological Q waves in ECG; abnormal myocardium motion; and, the presence of coronary thrombus [Thygesen, 2018]. In other terms, ECG and echocardiography could serve as primary tools for detection of MI patients, thus facilitating the classification of MI patients at hospitals.

1.3 Methodological context

1.3.1 Automatic echocardiography analysis

2D Echocardiography serves as a widely adopted and economically advantageous method for diagnosing cardiac dysfunction. There is a significant demand for automated solutions that can proficiently and cost-effectively assess cardiac function during clinical evaluations. In this context, the techniques of segmentation and motion tracking play pivotal roles in extracting essential cardiac parameters like the left ventricle ejection fraction (LVEF) and global longitudinal strain (GLS). Nevertheless, these tasks are intricate due to challenges encountered in ultrasound images, such as low signal-to-noise ratios, unclear boundaries, and issues with visibility.

1.3.1.1 Cardiac segmentation

Recently, deep learning methods have significantly advanced in the medical image segmentation domain, including automatic segmentation in 2D echocardiography (i.e., identifying the structure of left ventricle blood pool, myocardium and left atrium). In particular, the UNet architecture and its variants have demonstrated exceptional performance [Leclerc, 2019a; Ling, 2022]. Different regularization techniques have been proposed to constrain the shape regularity for cardiac imaging segmentation [Oktay, 2017; Clough, 2020]. Segmentation output is also a valuable source for both systolic and diastolic function analysis [Puyol-Antón, 2022a]. One crucial application of cardiac segmentation is the estimation of LVEF, a fundamental cardiac measurement in assessing cardiac function [Folse, 1962]. The quantification of LVEF involves using the bi-plane or single-plane Simpson's method [Folland, 1979] for volume assessment, thus is sensitive to image quality and segmentation process. Moreover, automatic identifications of end-diastole (ED) and end-systole (ES) frames [Smistad, 2020; Leclerc, 2019a] are also necessary. An alternative approach for LVEF estimation involves analyzing entire echocardiography sequences. This includes methods employing recurrent neural networks or transformer architectures to predict LVEF [Kazemi Esfeh, 2020; Reynaud, 2021].

1.3.1.2 Cardiac motion tracking

Traditional motion estimation methods, such as block matching [Azarmehr, 2020; Boukerroui, 2003], optical-flow [Ahn, 2013] and non-rigid registration [Vercauteren, 2008; Chakraborty, 2016], usually demand heavy computations. Recent work using deep learning models for echocardiography motion estimation have achieved good trade-off between run time and accuracy [Ahn, 2020; Ta, 2020]. Unsupervised motion estimation relies solely on image pairs (or annotated masks) for training [Ahn, 2020; Ta, 2020; Balakrishnan, 2019; Wang, 2022]. In contrast, supervised motion estimation employs dense displacement fields from ground truth data in order to penalize predicted deformation fields [Østvik, 2021]. However, obtaining accurate ground-truth myocardial displacement is challenging, often necessitating the use of synthetic echocardiographic images with known motion [Alessandrini, 2018; Evain, 2022]. However, it remains unclear how synthetic data aligns with *in vivo* data due to potential domain shifts [Deng, 2022].

1.3.2 Automatic electrocardiography analysis

For automatic ECG analysis, the application of deep learning methods have shown great effectiveness in various tasks, such as beat and rhythm detection [Teplitzky, 2020], ECG delineation of P, QRS, T segments [Jimenez-Perez, 2021], CVD diagnosis [Jahmunah, 2021; Dai, 2021; Li, 2023], etc. The most common architecture for ECG analysis is the

convolutional neural network (CNN) due to its great non-linear modelling potential and recurrent neural network (RNN) owing to its strong processing ability for time series. 12-lead ECG signals provide more information about electrical activities, while researchers are also exploring using reduced lead signals, such as single lead, 2 leads etc, for more general and portable application scenarios [Nejedly, 2022; Xu, 2022; Yu, 2022].

1.3.3 Multi-modal learning

Multi-modal learning, which includes multiple modalities such as image, text, audio etc, have shown better performance than using only one modality in different aspects: representation, fusion, co-learning and so on [Rahate, 2022]. In the field of biomedical data mining, it has also shown great potential [Stahlschmidt, 2022]. For example, some researchers [Soto, 2022; Goto, 2022] fused features from echocardiography and electrocardiogram to distinguish hypertrophic cardiomyopathy patients from individuals with similar pathology. They have shown better diagnostic performance using two modalities than using only one. [Puyol-Antón, 2022b] searched for a shared representation space between cardiac magnetic resonance (CMR) and 2D echocardiography, improving the cardiac resynchronisation therapy (CRT) response prediction with joint modalities.

1.4 Manuscript Organisation and Contributions

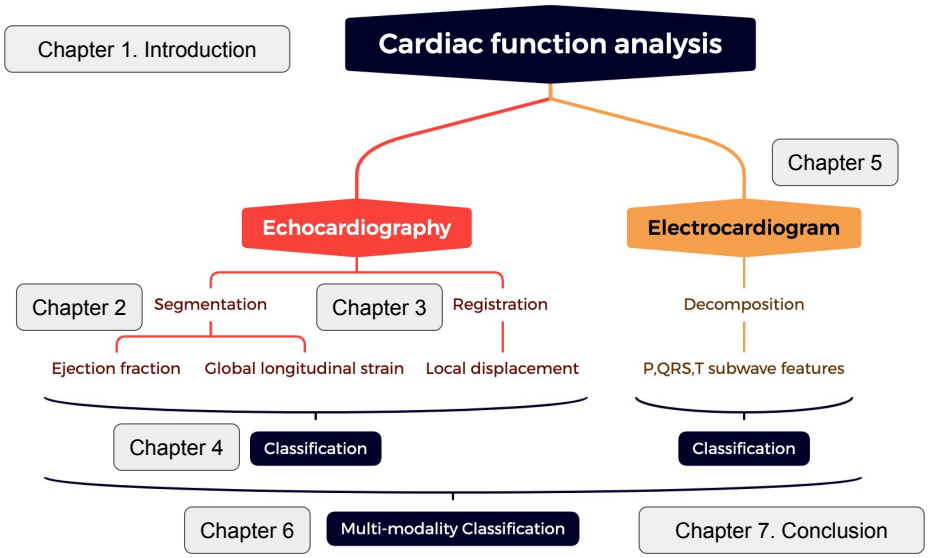


Fig. 1.8.: The road-map of manuscript organisation.

In the first chapter, we have presented the thesis background and motivation, as well as clinical and methodological context of this thesis. A more precise literature review will be presented afterwards, in the corresponding chapters.

Since two modalities are involved in this thesis, we design a road map for this objective (Figure 1.8). First, we treat echocardiography (Chapter 2, 3, 4) and electrocardiogram (Chapter 5) separately. Then, in Chapter 6, we explore multi-modal learning combining the two modalities together.

In Chapter 2, we describe three explorations for shape-aware segmentation in 2D echocardiography. In particular, we propose three models that incorporate shape information from global level (landmarks), regional level (AHA segments) and pixel level (structure contours). To improve generalisation, a stack of augmentation strategies tailored for ultrasound images is introduced. By conducting a thorough evaluation on different datasets, we demonstrate that with a pixel-level contour regularisation, we are capable of using a simple U-Net for robust segmentation with reduced anatomy outliers.

In Chapter 3, we focus on weakly-supervised motion tracking using real echocardiography sequences. Inspired from previous regional methods, we propose a novel motion estimation framework: the polyaffine motion model (PAM), which parameterise the dense deformation field by only 60 parameters. Compared to the state-of-art models, our proposed PAM model not only registers image frames with very good accuracy, but also demonstrates more regularity in terms of Jacobian determinant. By applying the trained model on unseen datasets, we obtain comparable results with models trained directly on the same dataset, supporting the conclusion that our PAM model is robust and has strong potential for clinical applications.

In addition, prior to this work, we have designed three neural network structures composing multi-layer perceptrons (MLP) and transformers for patch-based image-to-image registration. The improved regularity of displacement field has inspired us with respect to the potential of regional parameterisation, which has led to the work presented in Chapter 3. As the prior one was a collateral work with a colleague (Zihao Wang), we present the development of this work in Appendix A to keep the integrity of the main thesis.

In Chapter 4, a pipeline for robust echocardiograhy analysis is described. By using global indexes extracted from segmentation and motion tracking results, we achieve generalisable classification of myocardial infarction patients (MI) and non-MIs. This chapter serves as a conclusion for automatic echocardiography analysis.

In Chapter 5, we propose an unsupervised decomposition framework to analyze the morphology of a single heartbeat electrocardiogram. The decomposition network estimates the parameters of underlying subwaves, which provides explainable information of electrical events. Once again, the estimated parameters could serve for generalisable detection of MI patients using 12-lead ECG signals.

In Chapter 6, we adapt an uncertainty-based fusion strategy for multi-modal decision fusion using single modality predictions from echocardiography and electrocardiogram. Compared with conventional late fusion methods, uncertainty fusion leverages on the most trustworthy modality by measuring the single modality uncertainty, thus improving multi-modal performance by a margin of 7% compared with the single modality classification.

Finally, in Chapter 7, we summarize the main contributions of this thesis and discuss potential directions for future work.

1.5 Publications

The contributions during this thesis have resulted in the following peer-reviewed publications:

Journal articles

- [Yang, 2023a] **Yingyu Yang**, Rocher Marie, Moceri Pamela, and Maxime Sermesant. "Shape and Motion Priors for Generalisable Echocardiography Analysis using Deep Learning". 2023. *Paper in preparation for submission.*

Conference papers

- [Yang, 2021] **Yingyu Yang** and Maxime Sermesant. "Shape constraints in deep learning for robust 2D echocardiography analysis". *In: International Conference on Functional Imaging and Modeling of the Heart. Springer. 2021, pp. 22–34*

- [Yang, 2023b] **Yingyu Yang** and Maxime Sermesant. "Unsupervised PolyaffineTransformation Learning for Echocardiography Motion Estimation". *In: International Conference on Functional Imaging and Modeling of the Heart. Springer. 2023, pp. 384–393*

- [Yang, 2022] **Yingyu Yang**, Marie Rocher, Pamela Moceri, and Maxime Sermesant. "Explainable Electrocardiogram Analysis with Wave Decomposition: Application to Myocardial Infarction Detection". *In: InternationalWorkshop on Statistical Atlases and Computational Models of the Heart. Springer. 2022, pp. 221–232*

- [Wang, 2022] Zihao Wang[†], **Yingyu Yang**[†], Maxime Sermesant, and Hervé Delingette. "Unsupervised Echocardiography Registrationthrough Patch-based MLPs and Transformers". *In: International Workshop on Statistical Atlases and Computational Models of the Heart. Springer. 2022, pp. 168–178*

[†]These authors contributed equally to this work.

2 Automatic segmentation of echocardiography images

Contents

Abstract Automatic segmentation of cardiac structures in 2D echocardiography plays an important step for downstream analysis, such as volume quantification of the left ventricle, ejection fraction calculation etc. However, due to the noisy appearance of ultrasound images, it's not trivial to obtain robust segmentation results with the right anatomy.

In this chapter, we explore how to introduce shape constraints from global, regional and pixel level into a baseline U-Net model, for better segmentation and landmark tracking. Our experiments show that all three propositions perform similarly as a baseline model in terms of geometrical scores, while our pixel-level model (which uses a multi-class contour loss) reduces segmentation outliers and improves the tracking accuracy of 3 landmarks used for GLS computation. With appropriate augmentation techniques, our models also show a good generalisation performance

when testing on a larger unseen cohort. This chapter was published in the Proceedings of the 11th biennial International Conference on Functional Imaging and Modeling of the Heart (FIMH)[Yang, 2021]. We include extensive evaluation results that will be published in one journal submission [Yang, 2023a] in Appendix section of this chapter.

The main contributions of this chapter are summarized as follows:

- We propose three strategies for anatomy-aware segmentation using deep learning (Section 2.2).
- We introduce data augmentation techniques specifically designed for echocardiography to improve its generalisation (Section 2.3.2.2).
- We conduct extensive experiments on evaluating segmentation performance, EF and GLS values across datasets from different centers (Section 2.3.4 and Appendix 2.5).

2.1 Introduction

Echocardiography, a non-invasive and cost-efficient imaging technique, is widely used by cardiologists to evaluate the cardiac function. Segmentation and motion tracking are two essential tasks that can help cardiologists in clinical decision-making. Segmentation offers important information regarding the shape and volume, while motion tracking provides knowledge on myocardial deformation and function.

Methods leveraging deep learning have consistently demonstrated exceptional performance in the fields of medical segmentation and registration. As for segmentation, the U-Net architecture has proved its overwhelming power when used in large cohort echocardiography segmentation [Leclerc, 2019a]. With appropriate adaptation of U-Net model and data augmentation, the U-Net architecture also demonstrated good generalisation ability in segmenting cardiac magnetic resonance images (CMRI) [Chen, 2020]. To tackle the incorrect anatomy problem in cardiac segmentation, researchers have proposed regularization techniques and refinement models [Oktay, 2017; Leclerc, 2019b; Clough, 2020]. To incorporate temporal information into the segmentation process, Wei et al. [Wei, 2020] proposed two co-learning strategies of parallel segmentation and motion estimation from 2D echocardiography. Painchaud et al. [Painchaud, 2022] introduced a post-processing auto-encoder that corrects the temporal inconsistency of sequence segmentation outputs. Nonetheless, the segmentation of echocardiography remains challenging due to issues like out-of-view structures and suboptimal signal-to-

noise ratios, which pose significant hurdles, particularly in the context of myocardium segmentation.

In the field of cardiac motion tracking, unsupervised deep learning are very popular and these schemes reach similar performance or even outperform traditional registration methods. Krebs et al. proposed a conditional variational autoencoder which learned a diffeomorphic transformation from pairwise CMRI in an unsupervised way [Krebs, 2019]. Shawn et al. [Ahn, 2020] designed a U-Net like network for unsupervised pairwise echocardiography motion tracking. However, the displacement field can be unrealistic without a relevant regularisation.

Multiple research investigations have consistently demonstrated that global longitudinal strain (GLS) exhibits higher sensitivity as a metric for assessing systolic function when compared to left ventricular ejection fraction (LVEF). These findings suggest the potential utility of GLS in the clinical identification of left ventricular dysfunction, as reported in studies such as [Kraigher-Krainer, 2014][Hasselberg, 2014]. GLS can be approximated through the measurement of left ventricle length change, as indicated in the work by [Støylen, 2019]. Hence, there might be no imperative need to estimate a dense displacement.

U-Net like deep learning models depend largely on pixel-level classification, which can generate artefacts which are irregular for the organ shape. Researchers are seeking to combine shape constraints with deep learning methods [Bohlender, 2021]. Having the same intention to improve the segmentation consistency with anatomical shapes in 2D echocardiography, in our work, we explore the introduction of shape constraints from global, regional and pixel level into a baseline U-Net model. From the segmentation results, useful information such as EF and landmark based GLS can be extracted. The detailed model architecture will be explained in Section 2.2. We then present the implementation and experiment results for the segmentation and landmark detection steps in Section 2.3.

2.2 Methods

We used a U-Net model as our baseline model. Its encoder consisted of 5 down-sampling (MaxPool + Conv) blocks with ReLU activation after the 3x3 convolution. The corresponding decoder had 5 up-sampling (UpSample + Conv) blocks and is skip-connected with the encoder. Based on this model, we considered to incorporate shape constraints from three levels:

- **Global-level**: estimate a triangle like landmark map in parallel with segmentation (SEG-LM)

- **Regional-level**: add a poly-affine myocardium reconstruction network to constrain the shape of myocardium mask (SEG-AFFINE)

- **Pixel-level**: use a multi-class contour-loss to finely classify the boundary pixel (SEG-CONTOUR)

The three methods will be explained in detail in the following subsections.

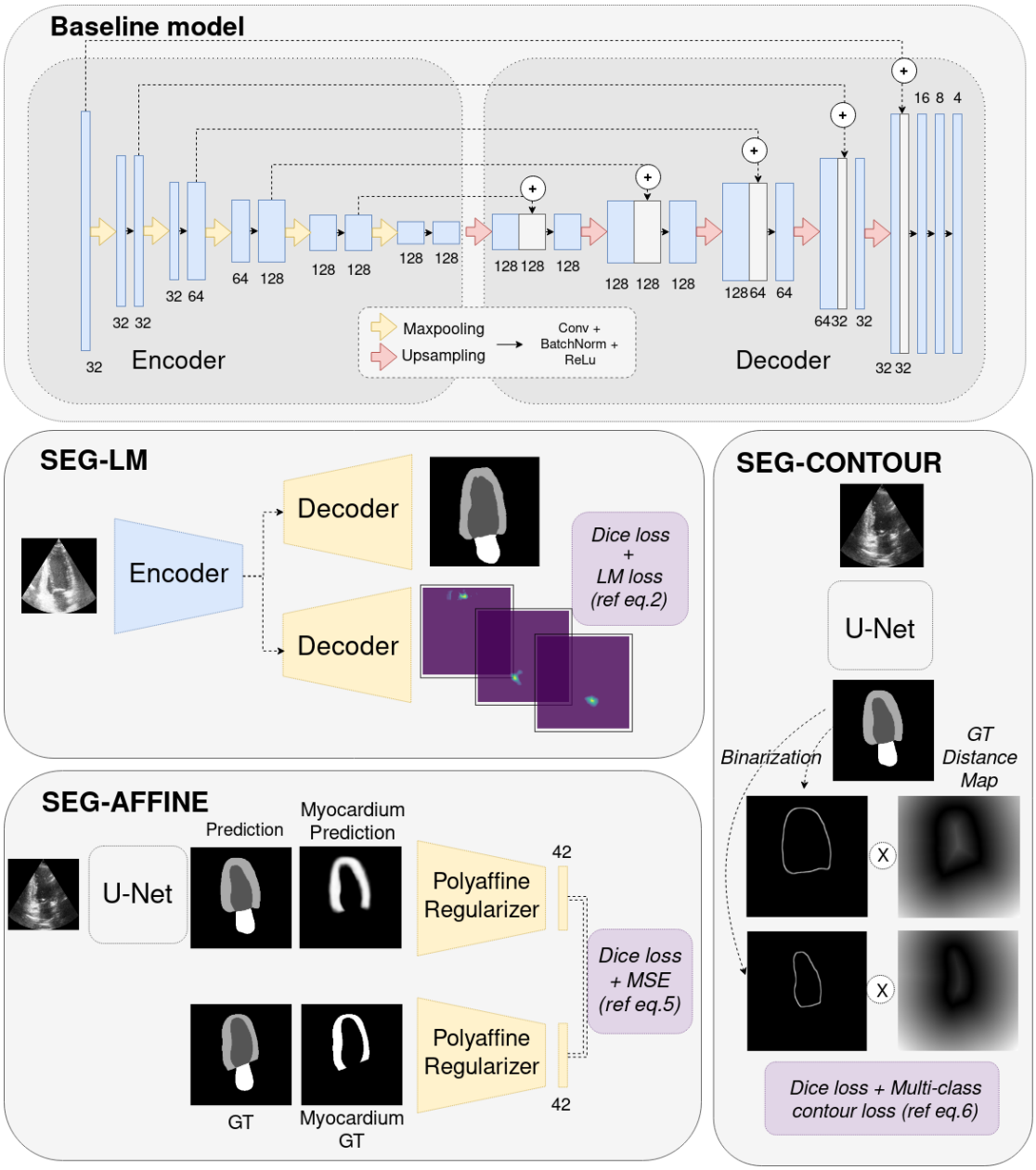


Fig. 2.1.: Detailed information of the 4 explored methods.

2.2.1 SEG-LM: Parallel segmentation and landmark detection

We adapted the baseline U-Net model for simultaneous segmentation and landmark prediction by adding a separate branch of decoder for landmark map estimation. The two decoders processed the encoder information in parallel. The final layer of segmentation branch and landmark detection were followed by SoftMax activation and Sigmoid activation, respectively. In particular, we considered the two end points of mitral valve (basal points) and the apex along the endocardial contour. The basal points were identified as the two end-points of adjacent boundary of both left ventricle and left atrium. The endo-apex point was then calculated as the furthest point to the mid-basal point along the endocardium. The output of landmark detection network was a heatmap of the corresponding target point. From the output heatmap, we extracted the landmark position by either finding the location of maximum or computing the centroid.

As we had different labels in the ground-truth data (myocardium, blood pool, atrium), a multi-class dice $\mathcal{L}_{dice}$ was used as the segmentation loss. As for landmark detection, we first penalised on the squared error of landmark heat-map (L2 loss: $\mathcal{L}_{l2}$). In order to avoid landmark overlapping on different output layers, we regularised the centre distance loss $\mathcal{L}_{CD}$ of different landmark heat-maps as proposed in [Wang, 2019]:

$$\mathcal{L}_{CD} = \frac{1}{2} \sum_{i=1, j \neq i}^{3} 1/(C(\mathcal{H}_i) - C(\mathcal{H}_j))^2 \tag{2.1}$$

with C the operation to obtain the centre position and $\mathcal{H}_i$ the landmark heat-map. Finally, a mean squared distance loss between the predicted heatmap centre and the ground truth point $\mathcal{L}_{point}$ was applied. Thus, the total loss $\mathcal{L}_{total}$ for optimisation was given by:

$$\mathcal{L}_{total} = \mathcal{L}_{dice} + \alpha \mathcal{L}_{l2} + \beta \mathcal{L}_{CD} + \gamma \mathcal{L}_{point} \tag{2.2}$$

2.2.2 SEG-AFFINE: Poly-affine Regulariser for Myocardium

With the intention to constrain the regularity of predicted myocardium mask, we proposed to model the myocardium mask as a combination of 6 AHA regions [Cerqueira, 2002]. We first chose a reference myocardium mask R from the training set. All of the N training myocardium masks $(M_i)_{i=1}^{N}$ were aligned to the reference mask by an affine transform $(T_i)_{i=1}^{N}$ estimated from the three landmarks (left basal, right basal and endo-apex). Then all aligned masks were averaged to a mean mask $\bar{R}$. The mean mask $\bar{R}$ is threshold-ed ($\bar{R}^f$) and splitted into 6 AHA regions $(\bar{I}_j)_{j=1}^{6}$. For every myocardium mask M_i, we aimed to first find 1 affine matrix A_g that globally transform the reference mask to $\hat{M_i}^g$. The corresponding reference regions became $(\bar{I}_j^g)_{j=1}^{6} = A_g \bar{I}_j$. We then found 6 affine matrices $(A_{ij})_{j=1}^{6}$, that transform the transformed (globally) mean AHA regions

$(\bar{I}_j^g)_{j=1}^6$ into $\hat{M}_i = \sum_{j=1}^6 A_{ij}\bar{I}_j^g$ that best reconstructed the target mask M_i, i.e. $\hat{M}_i \approx M_i$. For better fusion of the transformed 6 regions, we used the spatially weighted regions (multi-variate Gaussian) $\tilde{I}_j^w$ instead of $\bar{I}_j$, s.t. $\sum_{j=1}^6 \tilde{I}_j^w = \bar{R}^f$.

We used a CNN to estimate the affine parameters and reconstruct the given mask. The proposed network had two sub-networks. The first one aimed to estimate global affine parameters for global alignment which consisted of two hidden convolutional layers with down-sampling. The second sub-network was designed find the 6 regional affine matrix. It started with an encoder for high level feature extraction. The extracted features were passed through fully connected layers for affine matrix estimation $\hat{A}_{ij}$. By using the affine transform, we could reconstruct $\hat{M}^0$ from the 6 mean regions $\hat{M}^0 = \sum_{j=1}^6 \hat{A}_{ij}\tilde{I}_j^{wg}$. Two Conv layers were followed to refine the fusion mask $\hat{M}^0$, and thus we obtained the final output $\hat{M}^f$ (detailed information in fig.3).

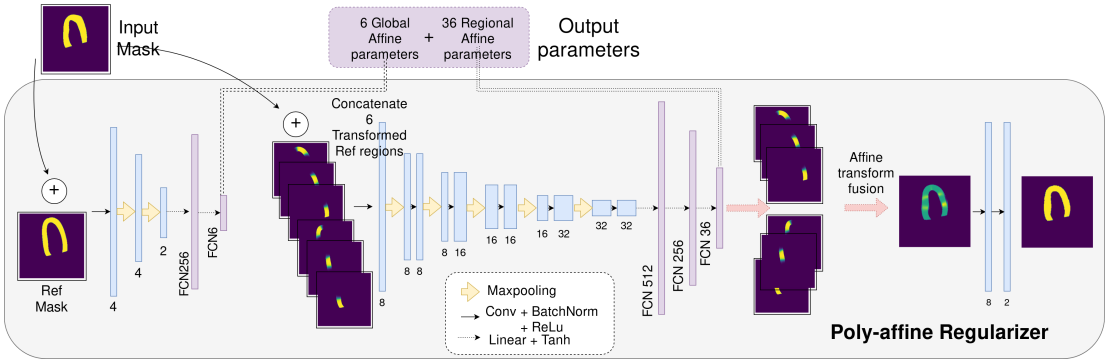


Fig. 2.2.: Architecture of Proposed Poly-affine Regulariser

In order to regularise the value of the affine parameters, we approached the affine parameter estimation problem using *Maximum A Posteriori* (MAP) with prior probabilities on the parameter values $P(A)$. The MAP aims to optimise: $\arg\max[P(A|M)] \propto P(M|A)P(A)$ (i.e. $\arg\min[-\log P(M|A) - \log P(A)]$). We used Gaussian distribution for conditional likelihood and priors. For $P(M|A) \propto \exp(-\frac{1}{2}(M - \hat{M}(A))^T\Sigma^{-1}(M - \hat{M}(A)))$, the variance is identity. For $P(A) \sim \mathcal{N}(\hat{\mu}, \hat{\Sigma})$, $\hat{\mu}$ and $\hat{\Sigma}$ are the maximum likelihood estimate (here we only consider diagonal covariance matrices) from the aligned transformation parameters $(T_i)_{i=1}^N$. The regularisation for affine parameters was $\mathcal{L}_{affine} = \alpha\mathcal{L}_{l2} + \beta\mathcal{L}_{prior}$, where $\mathcal{L}_{l2}$ was the mean square error between the reconstructed image and the input and

$$\mathcal{L}_{prior} = \sum_{k=1}^{K}\sum_{j=1}^{6} \delta_j \frac{(A_j^k - \hat{\mu}_j)^2}{\hat{\Sigma_{jj}}} \tag{2.3}$$

where $K = 1$ for global affine parameter and $K = 6$ for regional affine parameters.

Thus, the total loss for the poly-affine reconstruction network was:

$$\mathcal{L}_{total} = \underbrace{\mathcal{L}_{dice}(\hat{M}^g) + \beta^g \mathcal{L}^g_{prior}}_{\textit{global sub-net}} + \underbrace{\mathcal{L}_{dice}(\hat{M}^0) + \mathcal{L}_{dice}(\hat{M}^f) + \alpha\mathcal{L}_{l2} + \beta^r \mathcal{L}^r_{prior}}_{\textit{regional sub-net}} \tag{2.4}$$

Once the poly-affine regulariser network was trained, the 6+36 affine parameters served as an explicit hidden vector to regularise the shape of myocardium prediction. We trained a U-Net model (same as baseline model) which seeks for the best overlapping of mask as well as the minimum distance between the corresponding affine parameters of predicted myocardium and that of ground truth myocardium. The loss function for this method was

$$loss = dice + \alpha MSE(PA(P) - PA(M)) \tag{2.5}$$

where MSE represented the mean squared error, PA presented a poly-affine regulariser that outputs the 42 affine parameters.

2.2.3 SEG-CONTOUR: multi-class contour-loss

In order to increase the classification accuracy on the boundary, we chose to use an adapted multi-class contour loss[Jia, 2019]. Firstly, a distance map $D(M)$ was calculated from ground truth mask and it illustrated the shortest euclidean distance of each pixel to the closest border. Then the contour loss was calculated as

$$loss_{contour} = \sum(D(M) \circ contour(B(P))) \tag{2.6}$$

where $\circ$ performed element-wise multiplication. P represented the prediction output after SoftMax activation of U-Net for a certain class. $B(P)$ represented a differentiable thresholded Sigmoid for binarisation

$$B(P) = \frac{1}{1 + \exp^{-\gamma(P-T)}} \tag{2.7}$$

where $\gamma = 20$ and $T = 0.5$.

The contour of the binarised mask was obtained by applying a 2D Sobel filter

$$contour(P) = |G_x * P| + |G_y * P| \tag{2.8}$$

where $*$ denotesd 2D convolution and G_x, G_y were 2D Sobel kernel in x-,y- dimension:

$$G_x = \begin{bmatrix} 1 & 0 & -1 \\ 2 & 0 & -2 \\ 1 & 0 & -1 \end{bmatrix}, G_y = \begin{bmatrix} 1 & 2 & 1 \\ 0 & 0 & 0 \\ -1 & -2 & -1 \end{bmatrix}.$$

2.3 Experiments and Results

2.3.1 Datasets

In this work, we worked on two public data sets: CAMUS[1] and ECHONET[2]. CAMUS dataset consists of publicly accessible 2D echocardiographies and the corresponding annotations of 450 patients. For each patient, 2D apical 4-chambers (A4C) and 2-chambers (A2C) view sequences are available. Manual annotation of cardiac structures (left endocardium, left epicardium and left atrium) were acquired by expert cardiologists for each patient in each view, at end-diastole (ED) and end-systole (ES) [Leclerc, 2019a]. Along with the image and annotation data, the following information are also provided: image quality (good/medium/poor), left ventricle end-diastole volume (LVedv), left ventricle end-systole volume (LVesv) and left ventricle ejection fraction(LVEF).

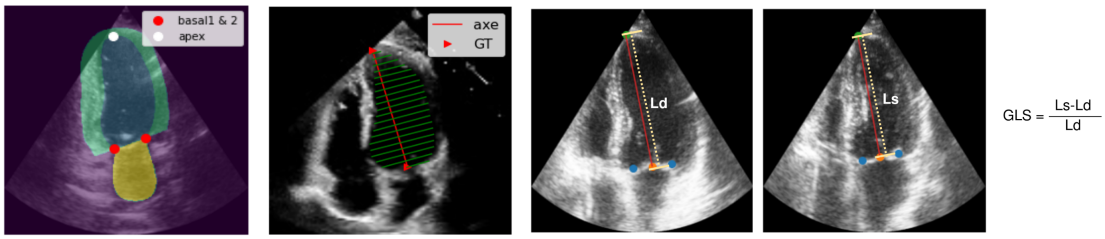


Fig. 2.3.: (a)An example of segmentation ground truth of CAMUS dataset. The two basal and apex landmarks were extracted following the procedures described in Section 2.2.1. (b) An example of annotation provided by ECHONET dataset. We generated the ground truth mask of LV by linearly connecting the border points. The grand axe (in red) was considered as the line connecting the apex and mid-basal point. The length of grand axe was considered as the LV length. (c-d) GLS calculation illustration (background is one echo image from CAMUS). We calculated the GLS from the LV length change by following the approximation method in [Støylen, 2019].

ECHONET dataset contains 10 030 apical 4-chambers echocardiography videos as part of routine clinical care at Stanford University Hospital [Ouyang, 2020]. Segmentation

[1] https://www.creatis.insa-lyon.fr/Challenge/camus/databases.html
[2] https://echonet.github.io/dynamic/

measurements (left endocardium) at end-diastole and end-systole are available for all videos. The corresponding LVedv, LVesv, LVEF are also provided for each video.

2.3.2 Experiments

We trained the baseline UNet, SEG-LM, SEG-AFFINE, SEG-CONTOUR models on the CAMUS dataset (both 2-chamber and 4 chamber ED/ES frames) using 10-fold cross validation. The 450 patients were randomly splitted into 10 folds. Each fold had a similar distribution of image quality and LVEF distribution. For every turn we used: 8 fold data for training; 1 fold for validation (for model selection); and, 1 fold for testing.

2.3.2.1 Data preprocessing

From the segmentation result, we first found the largest connected component for each class and applied a closing operation to fill the potential hole that could exist inside the predicted mask. We then extracted the basal and apex landmark points following protocol described in Section 2.2 for the four models except SEG-LM whose landmark is extracted from the landmark branch. The direction of mid-basal to apex was regarded as the grand axe of LV and used for LV volume estimation by using a modified Simpson's rule [Folland, 1979], and for the calculation of ejection fraction. The distance from mid-basal to apex formed the ventricle length and served for GLS calculation [Støylen, 2019]. The dataset didn't contain all the 3 apical views for the left ventricle; therefore, the GLS estimation was not be very accurate. However, we still computed the GLS from 2 views (CAMUS) and 1 view (ECHONET) as reference, in order to test the landmark detection accuracy.

2.3.2.2 Data augmentation

In order to avoid over-fitting and to improve generalisation performance, we applied random data augmentation at training phase for all the networks. Specifically, the implemented augmentation consisted of the following related changes designed for echocardiography (adapted from [Zhang, 2020]):

- **Rotation**: Rotation is usually caused by different probe orientation and we set the rotation range to [-15,15] degrees.

- **Crop and Scaling**: Scaling is due to the heart shape variability and myocardium motion. We set the scaling factor to [0.7,1.3] and then cropped (or pad) the resized image to be the same size.

- **Brightness adjustment**: Image intensity characteristic often varies among device vendors and depends on each image system settings. For brightness adjustment, we shifted the image intensity histogram by a factor of [0.9,1.1].

- **Contrast adjustment**: Contrast adjustment is a popular tool to improve the visualisation of cardiac structures. We applied a Sigmoid correction $I_{new} = \frac{1}{1+\exp C*(S-I)}$ on the input image, with a random range of cutoff S in [0.4,0.6] and a constant multiplier C from [4,10].

- **Sharpness adjustment**: To modify the sharpness of the image, we applied an unsharp mask $I_{new} = I + A * (I - I_\sigma)$, where A ranged from [1,2] and the σ of the Gaussian filter from [0.25,1.5].

- **Blurriness adjuestment**: For image blurring, we applied a Gaussian filter with σ in (0.25,1.5).

- **Speckle noise**: Due to the acoustic interference phenomenon, speckle is the main type of noise present in echocardiography. Speckle is usually modelled as a Rayleigh multiplicative noise. Since our training data is already log-compressed, we added speckle noise following an additive fashion $I_{new} = I + \sqrt{I} * G_\sigma$, where G is a multiplicative Gaussian noise with σ in the range [0.01,0.1].

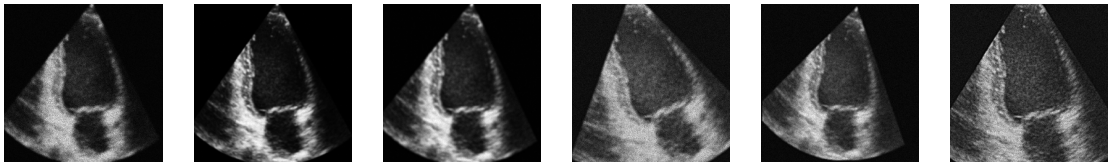


Fig. 2.4.: Randomly selected examples of augmented input images

The augmentation was applied in the following order: rotation; crop (pad) and scaling; brightness; contrast; sharpening; blurring; and, speckle noise. For each step, the application of the corresponding change was controlled by a Binominal random variable with $\rho = 0.5$.

2.3.2.3 Implementation

All segmentation models were implemented with Pytorch, with a batch size of 8 and input image resized to 256×256. CAMUS dataset does not contain the ground truth of desired landmarks. Thus, we generated the 'ground truth' landmark positions from ground truth segmentation masks. The Gaussian heatmap of ground-truth landmarks was computed with $\sigma = 4$.

- Baseline model: The baseline U-Net was trained with a multi-class dice loss. An Adam optimiser was applied with a learning rate of 1^{e-3}. The training was early stopped when the dice loss on validation data demonstrated no increase for more than 5 epochs.

- SEG-LM model: For SEG-LM model, we set $\alpha = 0.05, \beta = 0.5, \gamma = 0.5$ in Equation 2.2. An Adam optimiser was applied (lr$= 10^{-4}$). The output heatmap of landmarks was first processed to keep only one point cluster per layer and then the landmark location was extracted as the centroid.

- SEG-AFFINE model: The poly-affine regulariser network was first trained with the myocardium ground truth for reconstruction. We set $\alpha = 0.005, \beta^r = \beta^g = 0.01, \sigma_{i=1}^{6} = 1$ (Equation 2.4) and learning rate at 1e-4. We then trained a baseline U-Net with the polyaffine regulariser using loss function (Equation 2.5) with $\alpha = 10$. The parameter for global prior was calculated from the training-specific $(T_i)_{i=1}^{N}$. The parameter for regional prior was set as $\mu_{1...6} = [1, 0, 0, 0, 1, 0]$ and $[\Sigma_{i=1}^{6}] = 0.1$.

- SEG-CONTOUR model: As for SEG-CONTOUR model, contour loss was optimised along with the dice loss. The contour loss was easy to fall into a local minimum of 0, thus we set a weight of 100, 1e-4 for dice loss and contour loss respectively.

2.3.3 Evaluation metrics

For segmentation results, apart from the most used geometrical metrics: Dice coefficient, Hausdorff distance (HD) and Mean Surface Distance (MSD), we also used two anatomical metrics: Convexity(Cx) and Simplicity(Sp) [Leclerc, 2019b].

$$Convexity(Cx) = \frac{Area(P)}{Area(ConvexHull(P)} \tag{2.9}$$

$$Simplicity(Sp) = \frac{\sqrt{4\pi * Area(P)}}{Perimeter(P)} \tag{2.10}$$

Based on these metrics, we calculated the number of outliers for algorithm/model robustness evaluation. The outlier of segmentation prediction for CAMUS dataset was established from the inter-variability tests with the upper limit values for HD and MSD, and lower limit values for the simplicity and convexity [Leclerc, 2019b]. A prediction mask was considered as a geometrical outlier if its $HD > 3.5mm$ or $MSD > 8.2mm$ at ED, if $HD > 4mm$ or $MSD > 8.8mm$ at ES. The corresponding limit for anatomical outlier was as follows: if $Cx < 0.529$ or $Sp < 0.741$ for endocardium, if $Cx < 0.694$ or $Sp < 0.960$ for epicardium[Leclerc, 2019b].

2.3.4 Results

	Method	Baseline	SEG-LM	SEG-AFFINE	SEG-CONTOUR
Endo	Dice	**0.931** ± 0.040	0.928 ± 0.042	0.930 ± 0.042	**0.931** ± 0.041
	HD (mm)	5.04 ± 3.00	5.47 ± 3.04	5.14 ± 3.00	**4.99** ± 2.95
	MSD (mm)	1.51 ± 0.83	1.59 ± 0.84	1.53 ± 0.88	**1.50** ± 0.73
Epi	Dice	0.951 ± 0.025	0.950 ± 0.025	0.951 ± 0.027	**0.952** ± 0.026
	HD (mm)	5.75 ± 3.61	6.22 ± 3.74	5.84 ± 3.75	**5.63** ± 3.32
	MSD (mm)	1.71 ± 0.91	1.79 ± 0.93	1.72 ± 0.95	**1.67** ± 0.87
Outlier	Geo.	15%	20%	14.3%	**13.8%**
	Ana.	2.8%	7.6%	5.1%	**1.5%**
	Both	2.5%	5.3%	3.7%	**1.2%**

Tab. 2.1.: Segmentation Metric on CAMUS training data (450 patients, 10-fold cross validation) Endo.: endocardium, Epi: epicardium, HD: Hausdorff distance, MSD: mean surface distance, Geo.: Geometrical outlier, Ana.: Anatomical outlier. Values in bold represent the best score.

	Method	Baseline	SEG-LM	SEG-AFFINE	SEG-CONTOUR
Basal1	MAE (mm)	2.36	2.82	2.35	**2.25**
Basal2	MAE (mm)	3.06	3.45	2.97	**2.80**
Apex	MAE (mm)	4.06	4.48	4.22	**3.97**
GLS(%)	MAE (%)	3.97	4.51	4.03	**3.96**
	Corr.	0.74	0.70	0.74	**0.75**
	Bias(%) ± std	-0.73 ± 5.20	-1.74 ± 5.27	-0.9 ± 5.15	-0.36 ± 5.34
EF(%)	MAE (%)	**4.76**	5.26	4.96	5.06
	Corr.	**0.86**	0.84	0.85	0.84
	Bias(%) ± std	0.93 ± 7.07	2.51 ± 8.04	1.16 ± 7.31	0.7 ± 7.50

Tab. 2.2.: Landmark/GLS and EF Prediction on CAMUS training data (450 patients, 10-fold cross validation) Basal1: the left mitral valve end point, Basal2: the right mitral valve end point, EF: ejection fraction, GLS: global longitudinal strain. Values in bold represent the best score.

Table 2.1 and 2.2 include the results computed from 450 patients of CAMUS dataset using 10 fold cross-validation for the four methods detailed in Section 2.2. Compared with the

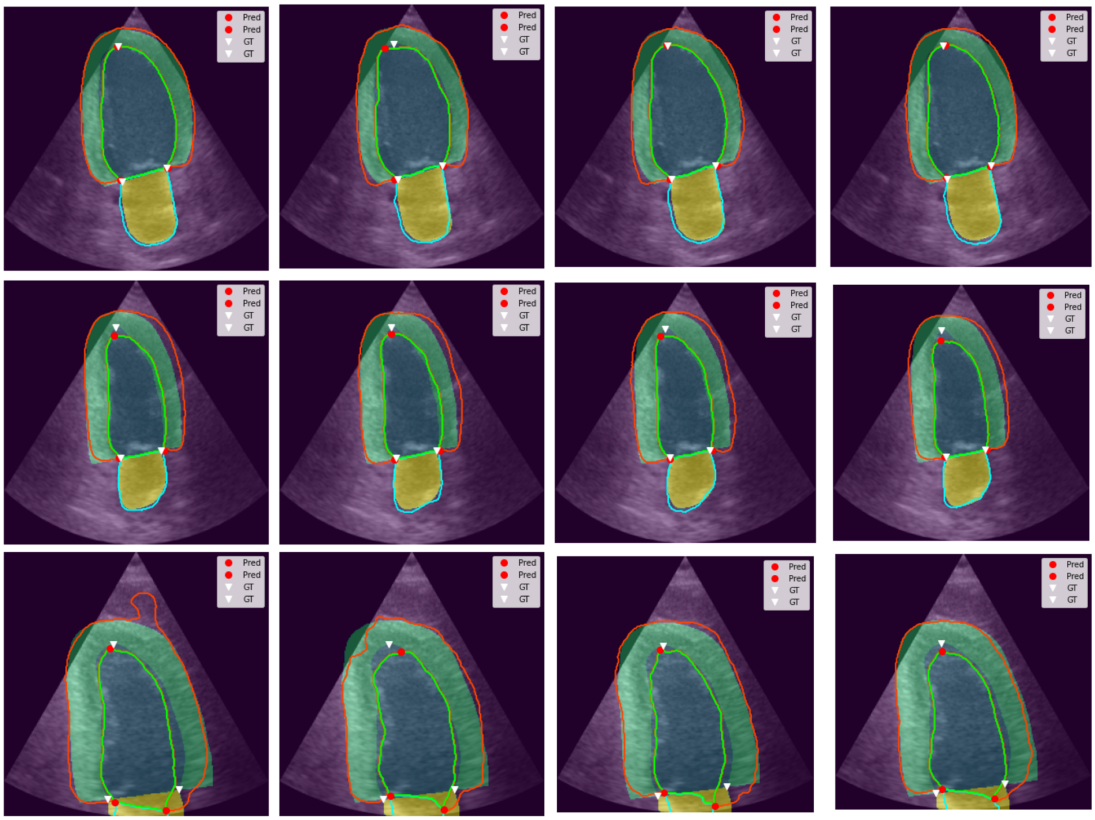


Fig. 2.5.: 3 CAMUS segmentation examples (good/medium/bad in terms of HD). The four columns represent: the baseline model, SEG-LM, SEG-AFFINE, SEG-CONTOUR respectively, from left to right. The red, green, cyan lines represent: the predicted segmentation contours of epicardium, endocardium and left atrium. The transparent green, blue and yellow regions are the ground truth masks of myocardium, left ventricle blood pool and left atrium, respectively.

baseline model, the SEG-LM, SEG-AFFINE models demonstrated only a slight decrease in terms of segmentation metric, ejection fraction (EF) prediction accuracy, and landmark detection accuracy. The SEG-CONTOUR model shows a similar performance with baseline model in Dice score (Table 2.1), but reduces greatly the number of geometrical and anatomical outliers (Table 2.1). It is reasonable that the SEG-CONTOUR reduced the classification error along the boundary area thus less outlier predictions were observed. This is consistent with the observations that a good Dice score does not always guarantee a good HD [Bernard, 2018], and, in our case, not always leads to anatomically-plausible segmentation. The SEG-CONTOUR model was also capable of tracking more precisely the boundary especially the landmarks (Table 2.2); thus resulting in a smaller bias of GLS prediction. In terms of EF calculation, the baseline U-Net model shows a smaller mean absolute error but a larger bias than the SEG-CONTOUR loss.

We show three CAMUS test examples (see Figure 2.5), where each row has a good/medium/bad performance in terms of HD score. Comparing with the other 3 models, SEG-CONTOUR has a more fluent border and similar to the ground truth annotation.

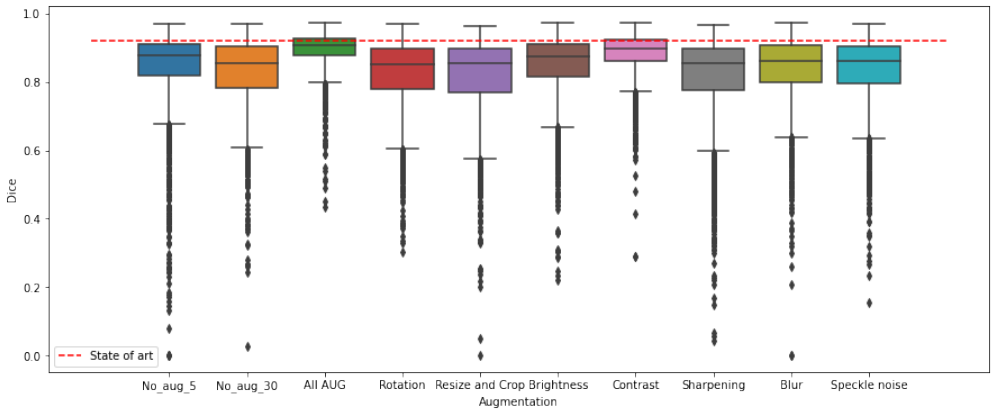


Fig. 2.6.: The baseline U-Net model was trained with only one of the mentioned augmentation methods on CAMUS dataset and then was evaluated on the same test fold of ECHONET segmentation model, whose dice coefficient was 0.92. Contrast adjustment contributed most to the improvement of generalisation result, while with all the techniques we obtained the best Dice score and less variation. No-aug-5: trained model of 5^{th} epoch without augmentation, No-aug-30: trained model of 30^{th} epoch without augmentation. At 30^{th} epoch, the model had already over-fitted the CAMUS data.

The evaluation results of applying the trained model on a totally different dataset ECHONET (the same test fold of 1277 patients as in [Ouyang, 2020]), showed the same trend of performance for all our four methods. The SEG-CONTOUR method demonstrated a good performance on tasks related to boundary information, for example, lower HD and MSD, better GLS estimation. It is less accurate on area based task (i.e., volume estimation thus ejection fraction prediction). It is noticeable that all of the four models demonstrated acceptable values for Dice coefficient on this different dataset (the Dice coefficient of models trained on ECHONET data was 0.92% [Ouyang, 2020]), which proves the importance of appropriate image augmentation techniques.

2.4 Conclusion

In this chapter, we explored methods to introduce shape constraints into 2D Echocardiography segmentation models from three levels: global-level (SEG-LM), regional-level (SEG-AFFINE), pixel-level (SEG-CONTOUR). From the evaluation results on CAMUS dataset and its generalisation result on a unseen dataset ECHONET, we suggest that it is more efficient to introduce pixel-level shape constraint than global or regional level constraints for U-Net based models. With a multi-class contour loss, SEG-CONTOUR model achieved better classification on the boundary pixels with a reduced a Hausdorff distance and more accurate landmark detection result. Overall, our experiments showed the great potential of SEG-CONTOUR for robust segmentation and deformation analysis.

	Method	Baseline	SEG-LM	SEG-AFFINE	SEG-CONTOUR
LV	Dice	0.892 ± 0.069	0.886 ± 0.063	0.887 ± 0.068	**0.895** ± 0.057
	HD (pxls)	12.99 ± 7.96	13.54 ± 6.86	13.12 ± 7.25	**12.56** ± 6.34
	MSD (pxls)	3.74 ± 3.42	3.87 ± 2.63	3.87 ± 2.97	**3.58** ± 2.16
EF	MAE (%)	**7.97**	8.79	9.35	8.39
	Corr.	0.69	**0.72**	0.67	0.69
	Bias(%) ± std	4.46 ± 12.80	6.21 ± 13.39	7.11 ± 14.41	5.03 ± 13.21
GLS	MAE (%)	4.96	5.61	5.88	**4.39**
	Corr.	0.38	0.36	0.29	**0.51**
	Bias(%) ± std	0.36 ± 7.94	-1.9 ± 7.07	-0.75 ± 9.29	0.59 ± 6.52

Tab. 2.3.: ECHONET Prediction

2.5 Appendix

2.5.1 Evaluation on CAMUS test data

ED	Endo			Epi		
	Dice	HD	MSD	Dice	HD	MSD
U-Net[Leclerc, 2019a]	0.939	5.3	1.6	0.954	6.0	1.7
ACNNs[Oktay, 2017]	0.936	5.6	1.7	0.953	5.9	1.9
CLAS[Wei, 2020]	0.947	4.6	1.4	0.961	4.8	1.5
nnUNet[Ling, 2022]	0.952	4.3	1.4	0.963	4.6	1.5
Ours	0.949	4.8	1.4	0.961	5.0	1.6
ES	**Endo**			**Epi**		
	Dice	HD	MSD	Dice	HD	MSD
U-Net[Leclerc, 2019a]	0.916	5.5	1.6	0.945	6.1	1.9
ACNNs[Oktay, 2017]	0.913	5.6	1.7	0.945	5.9	2.0
CLAS[Wei, 2020]	0.929	4.6	1.4	0.955	4.9	1.6
nnUNet[Ling, 2022]	0.935	4.2	1.3	0.959	4.4	1.5
Ours	0.930	4.7	1.4	0.955	5.0	1.6
	EDV		ESV		EF	
	Corr.	MAE	Corr.	MAE	Corr.	MAE
U-Net[Leclerc, 2019a]	0.926	11.2	0.960	7.5	0.845	5.0
ACNNs[Oktay, 2017]	0.928	9.7	0.954	6.9	0.807	5.5
CLAS[Wei, 2020]	0.958	7.7	0.979	4.4	0.926	4.0
nnUNet[Ling, 2022]	0.977	5.9	0.987	4.0	0.857	4.7
Ours	0.965	6.4	0.983	4.6	0.912	4.4

Tab. 2.4.: Segmentation evaluation on CAMUS test split data (50 patients).

Method	Geometrical	Anatomical
R-Unet [Leclerc, 2019b]	16%	1.2%
U-Net [Leclerc, 2019a]	18%	3.75%
Ours	**15.4%**	**1.15%**

Tab. 2.5.: CAMUS Segmentation shape outliers evaluated on 500 patients (2,000 images in total) with geometrical and anatomical outliers.

We further evaluated the trained SEG-CONTOUR model on the unseen test split of CAMUS dataset and compared the performance with other state-of-art segmentation methods using the same training set. Our proposed method achieved comparable performance with the best performing method either on cardiac structure segmentation or segmentation-based EF estimation. In terms of correct anatomy, comparing with the

previous work of [Leclerc, 2019b] which added a refinement network after the U-Net work to improve the shape regularity of predicted LV structures, our U-Net model with a contour prior (SEG-Contour) can directly generate less geometrical and anatomical outliers for the same dataset.

2.5.2 Evaluation on local private dataset

A second external validation dataset contains 2D echocardiography sequences of 76 patients collected from Nice CHU hospital approved by the local ethics guidelines. The final diagnostic as well as EF were given by an expert cardiologist. This dataset served as an evaluation set for real-world monitoring. Since no segmentation ground truth was available for our private dataset, we compared the EF estimated from segmentation model with the value provided by one expert cardiologist. We achieved a MAE of 8.3% and correlation coefficient of 0.74.

2.5.3 Discussion

From segmentation prediction, the most important cardiac index we can extract is EF. The three datasets involved in the segmentation evaluation phase demonstrate different annotation schemes, thus influencing the estimation of EF. For example, CAMUS dataset provided a detailed annotation of LV for ED/ES. EchoNet-Dynamic only provided endocardial traces that were used for volume estimation. The ground truth EF of our private data was obtained by one expert cardiologist's visual estimation. Despite of the diverse discrepancy between different datasets, our model still predict very satisfactory EF results, within the inter-observer's variability (MAE: 10%, Corr. 0.8) reported in [Leclerc, 2019a].

3 Automatic motion estimation from echocardiography images

Contents

Abstract The second important task in echocardiography analysis is to track the movement of left ventricle (LV), in order to reveal the motion pattern of the left ventricle wall. However, motion estimation in echocardiography is a challenging task since the ultrasound images suffer largely from low signal-to-noise ratio and out-of-view problems. Current deep learning-based models for cardiac motion estimation in the literature estimate the dense motion field with spatial regularization. However, the underlying spatial regularization can only cover a very small region in the neighboring areas, which is not enough to derive a smooth and realistic motion field for the myocardium in echocardiography.

In this chapter, we introduce a novel framework for cardiac motion tracking. Our method applies polyaffine transformation for motion estimation, which intrinsically regularizes the myocardium motion to be smooth. In order to constrain the motion

model to focus on the LV wall, we introduced a motion prior that can provide guidance for locating the myocardium. The main advantage of our model is that it provides a compact, smooth and realistic parameterisation of the LV deformation. This chapter is based on our preliminary results published in the Proceedings of the 12th biennial International Conference on Functional Imaging and Modeling of the Heart (FIMH) [Yang, 2023b] and its further improvement included in a journal submission [Yang, 2023a].

Our main contributions in this chapter include:

- the development of a novel polyaffine motion model (PAM) for unsupervised cardiac motion estimation (Section 3.2);
- the design of a comprehensive pipeline for using PAM in a weakly-supervised manner on a large public echocardiography dataset (Section 3.3);
- comprehensive experiments on various real-world datasets, showing that PAM outperforms other popular deep learning-based motion estimation methods and thereby highlighting its potential for clinical applications (Section 3.4).

3.1 Introduction

Echocardiography is one of the most widely used modalities for cardiac dysfunction diagnosis. It is a non-invasive radiation-free and real-time imaging modality, very suitable for portable analysis, such as myocardial motion evaluation. However, ultrasound images generally have poorer quality than other cardiac imaging modalities (e.g. MRI and CT), due to their low signal-to-noise ratio. Additionally, limitations in acquisitions and patient variability may cause the myocardium to occasionally be outside of the imaging field-of-view. These critical issues make myocardial motion estimation in echocardiography a particularly challenging task.

Traditional methods for motion estimation in echocardiography, such as block matching [Azarmehr, 2020; Alessandrini, 2016] or optical flow [Farneb, 2003; Barbosa, 2013; Alessandrini, 2016], are typically time-consuming and not suitable for real-time or portable analysis. However, recent advances in deep learning (DL) have improved both the time efficiency and tracking performance of motion estimation algorithms. DL-based models can generally be classified into two types: optical flow-based models and registration-based models, repsectively. Optical flow-based DL models estimate dense displacement fields between consecutive image frames and typically require ground truth displacement for supervised training [Østvik, 2021; Evain, 2022]. Since obtaining

ground truth displacement from real-world echocardiography images can be difficult, synthetic echocardiography datasets are often used for supervised training of these models [Alessandrini, 2018; Evain, 2022]. Registration-based DL models typically register all other frames to end-diastole either in an unsupervised manner or with weak supervision using segmentation masks, enabling them to work on real-world datasets [Ahn, 2020; Ta, 2020]. U-net-like architectures are often used as the core design of such methods [Ahn, 2020; Ta, 2020]. Many other studies focused on cardiac MRI motion estimation have also utilized the registration approach with different temporal smoothness strategies [Qin, 2018; Krebs, 2020]. Recently, researchers have also incorporated bio-mechanical modeling knowledge into deep learning networks with the aim of improving the generalizability of motion estimation performance [Qin, 2020; Zhang, 2022].

The motion of the myocardium is not arbitrary, and various spatial regularization techniques have been explored in the literature, such as: the divergence penalty [Mansi, 2011] for enforcing incompressibility; the rigidity penalty for smoothness [Staring, 2007]; and, the elastic strain energy for mechanical correctness [Papademetris, 2001], among others. Many deep learning models have incorporated these regularization techniques into their work [Ahn, 2020; De Vos, 2019; Zhang, 2022]. However, these regularization techniques are usually based on first- or second-order derivatives of the displacement vector field, which are applied at the pixel level in discrete implementation and are insufficient for studying the myocardium in echocardiography. The high noise-to-signal ratio in echocardiography degrades the quality of the estimated motion field, despite the pixel-wise constraints in deep learning methods. Lastly, other methods tackle the smoothness issue at the transformation level, for instance using regionally affine deformations [McLeod, 2015].

In this work, we introduced a deep learning-based motion model leveraging such approach and tailored for 2D echocardiography motion estimation, called the PolyAffine Motion model (PAM). In particular, based on an encoder-decoder backbone [Siarohin, 2019], we proposed the following modelling and regularisation strategies aiming to improve motion estimation in a global manner.

- Smoothness ++ : We parameterised the deformation field as polyaffine transformation, which intrinsically incorporated global-level smoothness regularization.

- Right Position ++ : We designed a motion prior and integrate it into loss functions, which helped the model to concentrate on myocardium region.

- Incompressiblity ++ : We proposed an incompressiblity loss term by regularisation through regional transformation, which helped us reduce large unrealistic volume change.

3.2 Methodology

3.2.1 Polyaffine motion model

Given a source image I_S and target image I_T, we aimed to estimate a dense motion field $\mathcal{T}_{S \leftarrow T} : \mathrm{R}^2 \rightarrow \mathrm{R}^2$ from target image I_T to source image I_S such that:

$$I_S(\mathcal{T}_{S \leftarrow T}(\cdot)) = I_T(\cdot). \tag{3.1}$$

Inspired by the polyaffine motion fusion framework proposed by Arsigny et al. [Arsigny, 2009], we adapted the motion estimation module from [Siarohin, 2019] to develop our proposed method for cardiac motion estimation in echocardiography.

The Polyaffine motion model consisted of two steps. First, the motion of the left ventricle was approximated through the sparse motion of several key points. An encoder-decoder network was then used to output the location of key points along with their local affine mapping for both I_S and I_T separately. Second, from the sparse motion, we obtained the final dense motion field through polyaffine motion fusion. The proposed method is illustrated in Figure 3.1.

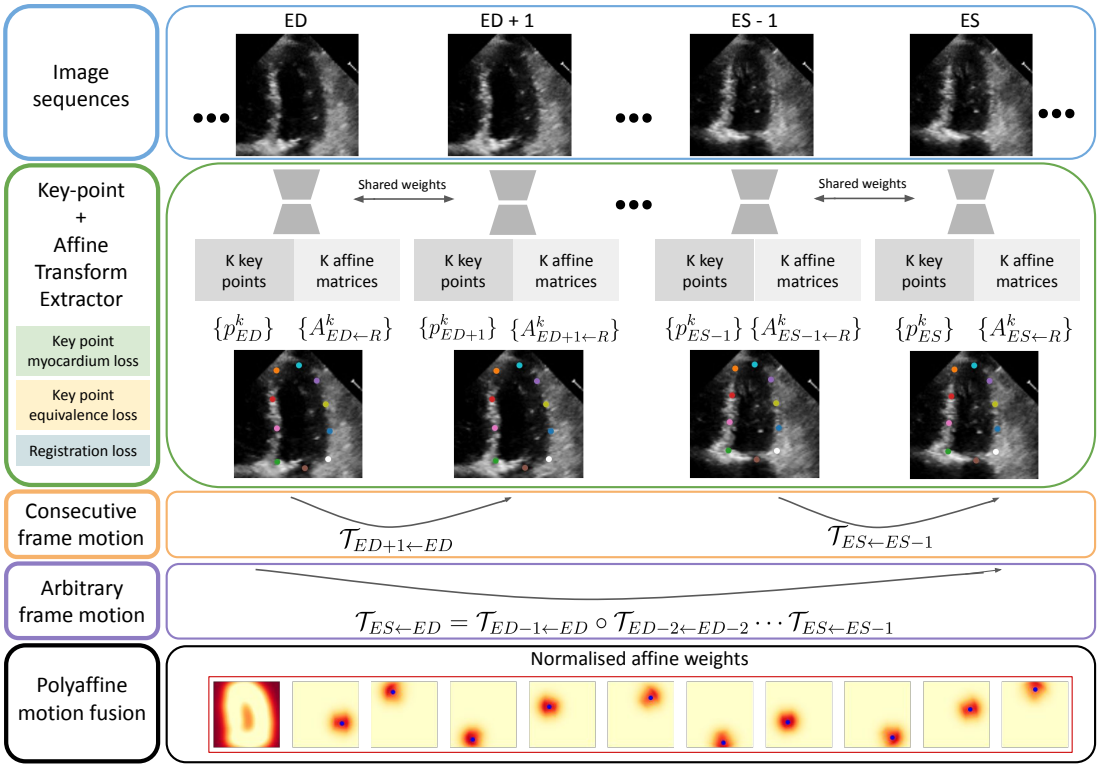


Fig. 3.1.: Method overview.

3.2.1.1 Key point and affine transformation estimation

We adopted the encoder-decoder architecture for key point extraction as presented in [Siarohin, 2019] and provided a brief review of the method from the point of view of affine transformation. In order to process each image separately, we assumed that there exists an abstract reference frame R. Given an image X, the encoder-decoder network generated the estimated key points $p_X^k, k = 1, 2, ...K$ as well as the corresponding linear mappings $A_{X\leftarrow R}^k \in \mathcal{R}^{2\times 2}, k = 1, 2, ...K$. The local affine transformation from target image I_T to source image I_S was then computed using the following equation:

$$\mathcal{T}_{S\leftarrow T}^k(z) = \underbrace{\bar{A}_{S\leftarrow T}^k \cdot z}_{\text{Linear mapping}} + \underbrace{(p_S^k - \bar{A}_{S\leftarrow T}^k \cdot p_T^k)}_{\text{Translation}}, \tag{3.2}$$

where $z \in \mathrm{R}^2$ represents the coordinate in the target image, and $\bar{A}_{S\leftarrow T}^k = A_{S\leftarrow R}^k (A_{T\leftarrow R}^k)^{-1}$. In order to capture motion around the myocardium, we introduced a myocardium-related key-point prior (Section 3.3.2) and an associated loss functions (Section 3.2.2.1). This encouraged the network to output key points close to the myocardium, in contrast to [Siarohin, 2019], which employed a self-supervised approach for learning key-point positions.

3.2.1.2 Polyaffine motion fusion

Once we obtained the local affine transformation, the dense motion field was computed through direct polyaffine motion fusion. For each local affine motion, a spatial weight $W_k(p_T^k, \sigma^2)$ controlled its influenced region; we achieved this by applying a 2D isotropic Gaussian distribution centered at key point p_T^k with variance σ^2. Here, W_0 represented the weight for the background region and the left ventricle cavity area, and was computed as follows:

$$W_0 = \textbf{RELU}(1 - \sum_{k=1}^{K} W_k(p_T^k, \sigma^2)). \tag{3.3}$$

We normalised all weights into $\bar{W}_k = \frac{W_k}{\sum_{k=0}^{K} W_k}$, where $k = 0, 1, 2...K$. An example is shown in Figure 3.1. Next, we applied these to compute the polyaffine dense motion

$$\mathcal{T}_{S\leftarrow T}(z) = \bar{W}_0 z + \sum_{k=1}^{K} \bar{W}_k(p_T^k)\mathcal{T}_{S\leftarrow T}^k(z). \tag{3.4}$$

The original first-order motion model (FOMM) [Siarohin, 2019] utilised a second encoder-decoder network to estimate the normalised weights for each local affine transformation, which can not guarantee the center of weights close to the corresponding key point, thereby being unstable for myocardial motion estimation.

3.2.1.3 Sequence motion estimation

Considering a sequence of image frames $I_1, I_2, ...I_N$, the assumption of abstract reference frame enabled a fast way to obtain the dense motion field between any arbitrary pair of frames in the sequence. Without loss of generality, we assumed that I_{ED} was the frame at end-diastole and I_{ES} was the frame at end-systole, and that I_{ED} was ahead of I_{ES} in time. The local affine transformation from I_{ED} to I_{ES} could be calculated by composition:

$$\begin{aligned} \mathcal{T}^k_{ES\leftarrow ED}(z) &= \mathcal{T}^k_{ED+1\leftarrow ED} \circ \mathcal{T}^k_{ED+2\leftarrow ED+1} \cdots \mathcal{T}^k_{ES\leftarrow ES-1} \\ &= \bar{A}^k_{ES\leftarrow ED} \cdot z + (p^k_{ES} - \bar{A}^k_{ES\leftarrow ED} \cdot p^k_{ED}), \end{aligned} \tag{3.5}$$

where $\bar{A}^k_{ES\leftarrow ED} = A^k_{ES\leftarrow R}(A^k_{ED\leftarrow R})^{-1}$. The final dense motion was computed by combining the local motion, as described in Equation 3.4.

3.2.2 Loss functions

The model was trained end-to-end using a combination of loss functions, which can be grouped into the following subsets for sequence motion estimation.

3.2.2.1 Keypoint myocardium losses

We proposed two loss functions to enforce the position of key points near the myocardium region. The first loss, denoted by $\mathcal{L}_{\text{kp\_prior}}$, penalized the L_2 norm between the estimated key-point position and a prior position. We obtained the prior key-point position from the available training masks, as described in Section 3.3.2. The second loss, denoted by $\mathcal{L}_{\text{kp\_myo\_ED}}$, constrained key points at end-diastole (ED) to reside within the myocardium region by penalizing the distance between each key point and the myocardium.

$$\mathcal{L}_{\text{kp\_myo\_ED}} = -\sum_{k=1}^{K} H(p^k, \sigma_H^2) * (\text{Mask}_{\text{myo\_ED}} - 0.5), \tag{3.6}$$

where $H(p^k, \sigma_H^2)$ was the isotropic Gaussian heatmap centered at p^k with variance σ_H^2 and $\text{Mask}_{\text{myo\_ED}}$ the binary mask of myocardium at ED.

3.2.2.2 Keypoint equivalence losses

Another two losses $\mathcal{L}_{\text{equi\_kp}}$ and $\mathcal{L}_{\text{equi\_affine}}$ imposed an equivalence constraint to the detected key points [Siarohin, 2019]. These forced the model to predict consistent key points and local linear mappings when applying a known transformation to the input image.

3.2.2.3 Registration losses

The last four losses regularized the final dense motion field through image similarity and key-point similarity. First, for each pair of consecutive frames in the sequence, we constrained

$$\mathcal{L}_{\text{seq\_im}} = \sum_{seq} NCC(I_T, I_S(\mathcal{T}_{I_S \leftarrow I_T})), \tag{3.7}$$

$$\mathcal{L}_{\text{seq\_kp}} = \sum_{seq} \frac{1}{K} \sum_{k=1}^{K} |H(p_T^k, \sigma_H^2) - H(p_S^k, \sigma_H^2)(\mathcal{T}_{I_S \leftarrow I_T})|, \tag{3.8}$$

where NCC represented the normalised cross correlation. Second, in order to force the model to learn temporal motion, we applied a second pair of similarity loss between the image at end-diastole and all frames after the end-diastole frame, denoted as $\mathcal{L}_{\text{ED2any\_im}}$ and $\mathcal{L}_{\text{ED2any\_kp}}$ for image similarity and key-point similarity respectively.

3.2.2.4 Incompressibility penalisation

In order to constrain the compressibility of the myocardium within physiological ranges, we applied an incompressibility penalisation to the deformation field. Different from former work that applied divergence-free regularisation to the dense motion field [Mansi, 2011; Ahn, 2020], here we can efficiently constrain the Jacobian determinant of each local affine transformation to be close to 1:

$$\mathcal{L}_{\text{incomp}} = \sum_{k=1}^{K} (|\det(A_{X \leftarrow R}^k)| - 1)^2 \tag{3.9}$$

The total loss for the PAM model was

$$\begin{aligned} \mathcal{L}_{\text{total}} = & \lambda_1 \mathcal{L}_{\text{seq\_im}} + \lambda_2 \mathcal{L}_{\text{seq\_kp}} \\ & + \lambda_3 \mathcal{L}_{\text{equi\_kp}} + \lambda_4 \mathcal{L}_{\text{equi\_affine}} \\ & + \lambda_5 \mathcal{L}_{\text{kp\_prior}} + \lambda_6 \mathcal{L}_{\text{kp\_myo}} \\ & + \lambda_7 \mathcal{L}_{\text{ED2any\_im}} + \lambda_8 \mathcal{L}_{\text{ED2any\_kp}} \\ & + \lambda_9 \mathcal{L}_{\text{incomp}} \end{aligned} \tag{3.10}$$

3.3 Experiments

3.3.1 Datasets

Three public datasets of echocardiography were included in our study.

- EchoNet[1] [Ouyang, 2020] contained 10030 echocardiography videos of long-axis 4 chamber view. Left-ventricle tracings at end-diastole and end-systole were available (see Figure 3.2(a)).

- CAMUS dataset[2] [Leclerc, 2019a] consisted of apical 4-chamber view videos from 450 patients whose left heart segmentation masks were publicly accessible (see Figure 3.2(e)).

- HMC-QU dataset[3] [Degerli, 2021b] contained 109 4-chamber view echocardiography recordings with the segmentation of the left myocardium along one cardiac cycle (see Figure 3.2(d)).

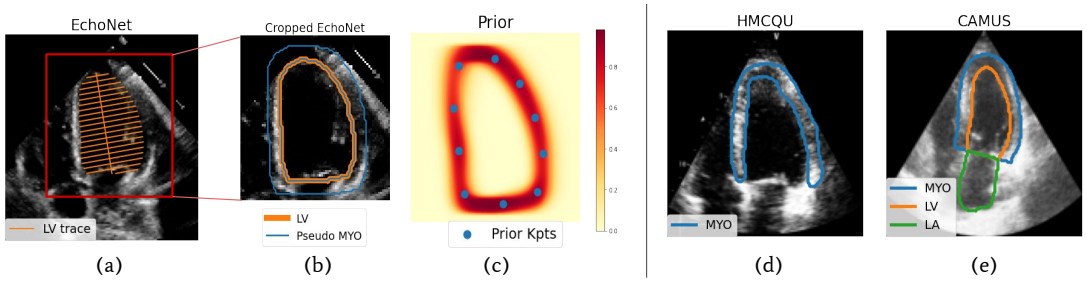


Fig. 3.2.: Dataset overview. (a) EchoNet example and the given annotations of LV tracing. (b) Left ventricle cropping with the generated pseudo myocardium contour (blue). (c) The mean mask from the EchoNet training set and the 10 prior key points. (d) HMC-QU example and the given annotation of the myocardium. (e) CAMUS example and the given annotation of different cardiac structures. (MYO: myocardium, LV: left ventricle, LA: left atrium)

All image sequences were cropped around the LV center according to the provided segmentation/tracings (see Figure 3.2(b)). In this study, we followed the given data division of the EchoNet dataset, with 7465 samples for training, 1288 samples for validation and 1277 samples for testing. CAMUS and HMC-QU datasets were used for evaluation during test phase.

[1] https://echonet.github.io/dynamic/
[2] https://www.creatis.insa-lyon.fr/Challenge/camus/
[3] https://www.kaggle.com/datasets/aysendegerli/hmcqu-dataset

3.3.2 Dataset preprocessing

3.3.2.1 Pseudo myocardium mask

There was no ground truth of myocardium mask in EchoNet dataset. To provide guidance for key points, we generated a pseudo myocardium mask for end-diastole (ED) and end-systole (ES) by applying a dilation operation using 13x13 and 17x17 structure element to the left ventricle mask at ED and ES frame respectively. The difference between the dilated mask and the original one was regarded as the pseudo myocardium mask (see Figure 3.2(b)).

3.3.2.2 Key-point prior at end-diastole

A mean myocardium mask was computed using all the pseudo myocardium masks at ED from the training set. Then, all the pixels of the mean mask were clustered into 10 groups using the KMeans function from the scikit-learn package [Pedregosa, 2011], where the center of each group was considered as one key-point prior (see Figure 3.2(c)).

3.3.3 Implementation

We compared our proposed model with the conditional variational autoencoder (CVAE) method proposed in [Krebs, 2020], which demonstrated effectiveness in sequential motion modelling of cardiac images. Its former registration version [Krebs, 2019] had shown more regular motion field than other deep learning methods, including VoxelMorph [Dalca, 2018]. We implemented the CVAE method following its description in [Krebs, 2020] and used the same training hyper-parameters. However, we added a Dice loss between the transformed end-systolic (ES) mask and end-diastolic (ED) mask during CVAE training to keep it consistent with our PAM model, which used segmentation mask information during training (see Section 3.2.2.1). We also integrated a divergence-free term into the final loss function for incompressibility regularisation.

We trained the PAM model using an Adam optimizer with a learning rate of $1e-4$. To determine the optimal hyperparameters, we conducted experiments on a randomly selected subset of 1000 training examples from the EchoNet dataset. The hyperparameters for the loss function were set to $\lambda_1 = \lambda_2 = 100, \lambda_3 = \lambda_4 = 50, \lambda_5 = 20, \lambda_6 = 0.1, \lambda_7 = \lambda_8 = 100, \lambda_9 = 1$ in Equation 3.10. The variance of the Gaussian heatmap was set to $\sigma^2 = 0.05$ for affine transformation weights (see Equation 3.3), and $\sigma_H^2 = 0.005$ for all other scenarios. We trained the model for a maximum of 100 epochs, and applied early stopping if the validation loss did not improve over 10 epochs.

3.3.4 Alation study

The backbone FOMM model [Siarohin, 2019] optimized the first four terms of Equation 3.10. New strategies tailered for echocardiography proposed in this work are denoted as follows:

- +*Prior*: integration of prior-related losses (5th and 6th loss terms in Equation 3.10);
- +*Polyaffine*: replacing the encoder-decoder module for weight estimation by designed polyaffine weights;
- +*ED*: incorporation of registration losses between ED frame and others (7th and 8th terms in Equation.3.10); and,
- PAM(all): the model that optimized the total loss function of Equation 3.10.

3.4 Results

3.4.1 Registration accuracy

We first evaluated the motion estimation accuracy by assessing the registration performance using the available segmentation masks from all test datasets. For the EchoNet test split, we compared the ground truth of the LV mask for end-diastole (ED) with that transformed from end-systole (ES). For the CAMUS dataset, the same evaluation was applied to the myocardium mask for ED/ES. For the HMC-QU dataset, the myocardium masks of one cardiac cycle were all transformed to ED using the estimated motion field. To enable group-level statistical analysis along the cardiac cycle, the temporal metric was interpolated to the same length.

3.4.1.1 EchoNet

We compared our proposed PAM model with a state-of-art model CVAE and our backbone model FOMM. The main difference between PAM and FOMM was our proposition of motion prior and explicit polyaffine motion fusion. From Table 3.1 we observed that the introduction of motion prior and polyaffine motion framework improved largely the performance of registration on EchoNet-Dynamic test split. We also achieved a performance comparable performance with that of the state-of-art CVAE model.

Method	Endo		
	Dice	HD(*pixels*)	MSD(*pixels*)
CVAE [Krebs, 2020]	0.91	7.97	2.30
FOMM [Siarohin, 2019]	0.75	22.99	6.07
FOMM+*Polyaffine*	0.82	11.50	4.61
FOMM+*Prior*	0.77	18.36	5.46
FOMM+*Prior*+*Polyaffine*	0.91	7.53	2.36
FOMM+*Prior*+*Polyaffine*+*ED*	0.92	7.34	2.23
PAM(all)	**0.92**	**7.17**	**2.14**

Tab. 3.1.: Registration evaluation on EchoNet-Dynamic test split. *Endo: endocardium. HD: Hausdorff distance. MSD: mean surface distance.*

Method	Epi		Endo	
	Dice	MSD (mm)	Dice	MSD (mm)
OTA+OTS [Wei, 2020]	0.933	2.0	0.898	1.8
CVAE [Krebs, 2020]	0.917	3.0	0.859	3.3
FOMM [Siarohin, 2019]	0.883	4.3	0.778	5.2
FOMM+*Polyaffine*	0.881	4.3	0.796	4.8
FOMM+*Prior*	0.893	3.8	0.804	4.3
FOMM+*Prior*+*Polyaffine*	0.932	2.5	0.885	2.6
FOMM+*Prior*+*Polyaffine*+*ED*	**0.933**	**2.4**	**0.894**	**2.4**
PAM(all)	0.927	2.6	0.883	2.6

Tab. 3.2.: Registration evaluation on CAMUS training set (450 patients with 2CH and 4CH samples). *Endo: endocardium. Epi: Epicardium. MSD: mean surface distance.*

3.4.1.2 CAMUS

Without alternative training or fine-tuning, we applied directly the trained motion model on unseen dataset CAMUS training set. Table 3.2 summarizes the generalisation performance of our proposed PAM model and CVAE model. Our PAM model not only generalized well on CAMUS dataset when compared with the backbone FOMM and the state-of-art CVAE, but also achieved the closest registration performance with the best tracking model (OTA+OTS) [Wei, 2020] that was trained on CAMUS training set using 10-fold cross-validation.

3.4.1.3 HMC-QU

HMC-QU dataset contained 109 4-chamber view samples with myocardium annotated automatically for one cycle [Degerli, 2021a]. The proposed PAM model achieved satisfac-

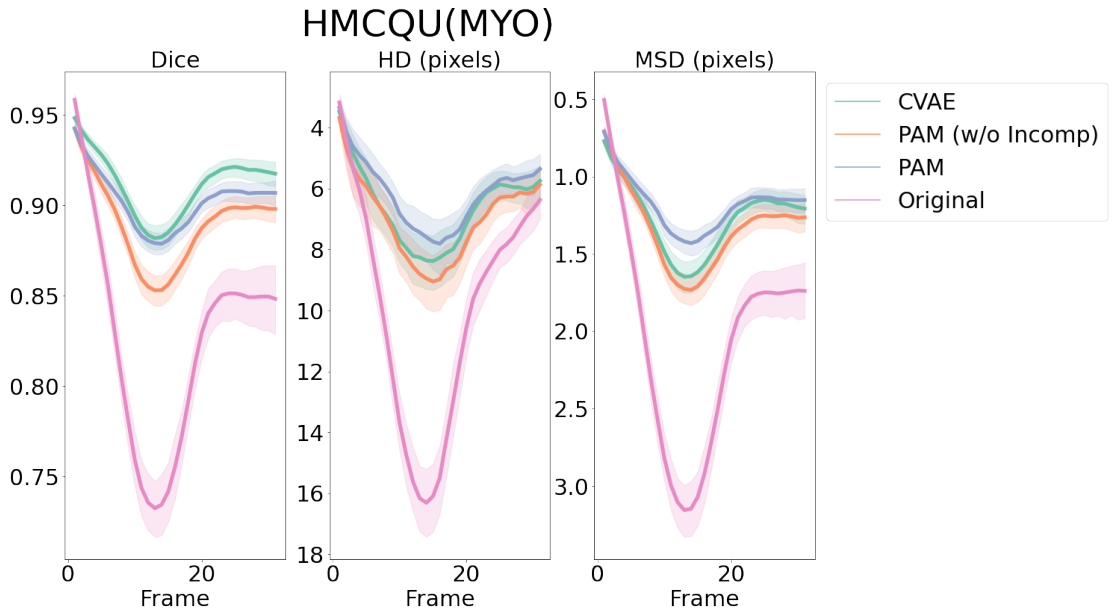


Fig. 3.3.: Registration results on HMC-QU dataset (109 A4C samples) using frame-wise myocardium masks. Original: Comparison between the ground truth masks along one cardiac cycle and that of end-diastole. Curves of all samples were interpolated to the same length. *MYO: myocardium. HD: Hausdorff distance. MSD: mean surface distance.*

tory tracking performance (Dice score 0.90, HD 6.1 pixels, MSD 1.2 pixels) compared with CVAE model (Dice score 0.91, HD 6.6 pixels, MSD 1.3 pixels). The introduction of the incompressibility penalisation improved the tracking performance, especially in terms of Hausdorff distance and mean surface distance (Figure 3.3).

3.4.2 Deformation regularity

Furthermore, our PAM model excelled in producing motion fields with enhanced regularity as evidenced by Jacobian Determinant closer to 1 and smoother volume change as evidenced by the gradient (see Figure 3.4). The smooth displacement field was advantageous for computing dense strain tensor, which was typically constructed using the first-order derivative of the displacement field. We show examples of Jacobian Determinant of motion field from myocardial infarction patient (MI) and non-MI patient in Figure 3.5.

Besides of the good tracking performance from different evaluation metrics, the proposed PAM model demonstrated good potential for abnormal wall motion detection. We show an example of myocardial infarction (MI) from HMC-QU dataset (Figure 3.6) to visualize how it works. Our prediction shows that the patient has a small radial strain in SEG14 region, which is in accordance with the fact that the patient is diagnosed with MI and in particular the SEG14 is affected.

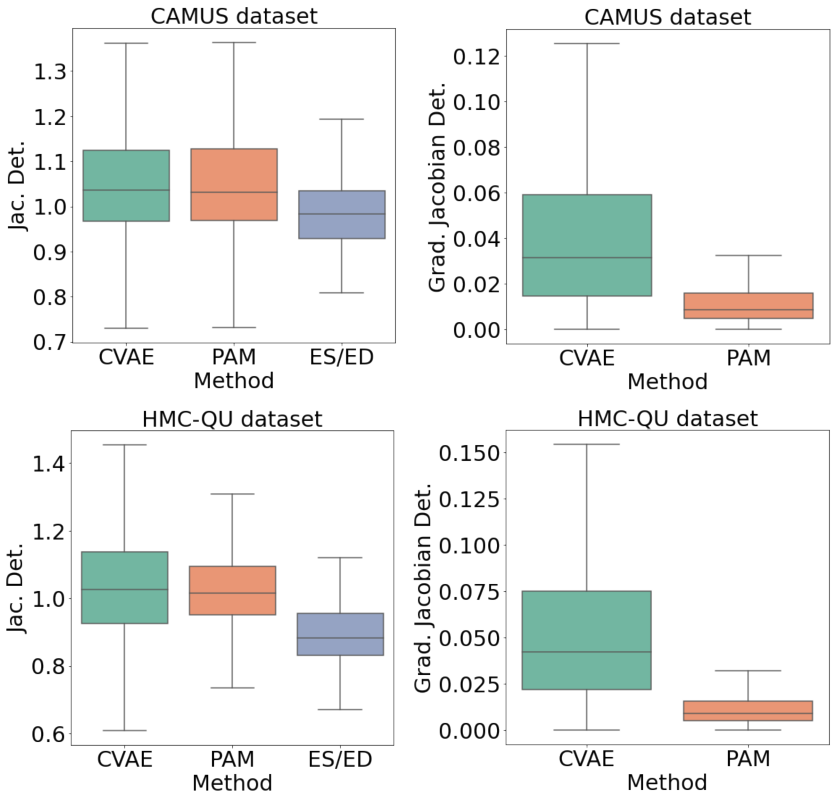


Fig. 3.4.: Evaluation results of Jacobian Determinant (Jac. Det.) and its gradient in the myocardium regionon on the CAMUS dataset and on the HMC-QU dataset. ES/ED represents the area ratio of the myocardium between end-systole and end-diastole obtained using ground truth masks.

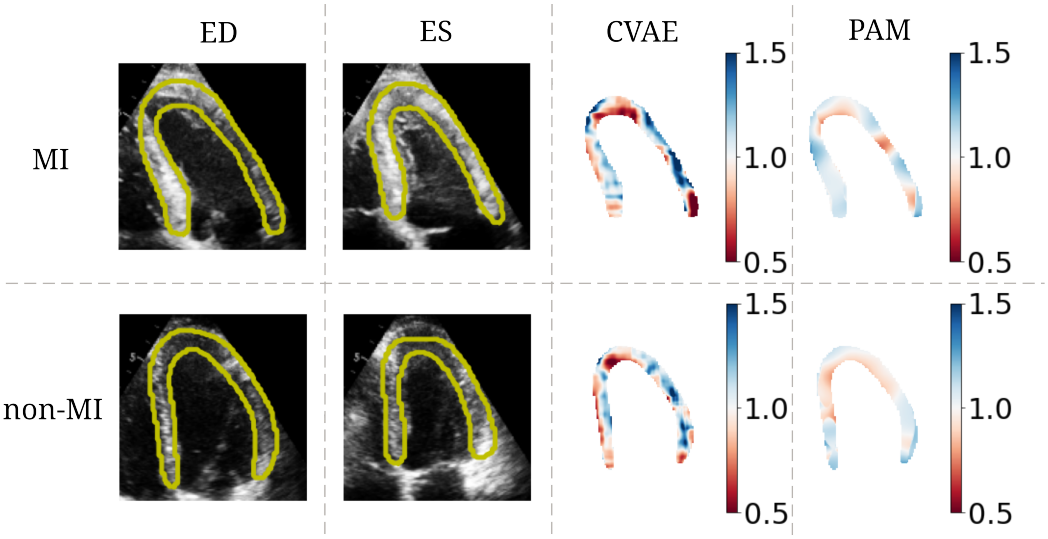


Fig. 3.5.: Examples of Jacobian determinant map in the myocardium region from HMC-QU dataset using CVAE and PAM methods.

3.5 Discussion and conclusion

In this chapter, we proposed a polyaffine motion model (PAM) for echocardiography motion estimation. The PAM model demonstrated excellent motion estimation performance

on real-world echocardiography datasets and showed good generalization to unseen datasets from other centers. Our explicit design of fusion weights enabled efficient learning of local affine transformation, and the intrinsic polyaffine structure improved the smoothness of the motion field, showing potential for abnormal wall motion detection. In the future, we will focus on integrating temporal regularization for the PAM model and conducting evaluations on synthetic datasets with known ground-truth displacement.

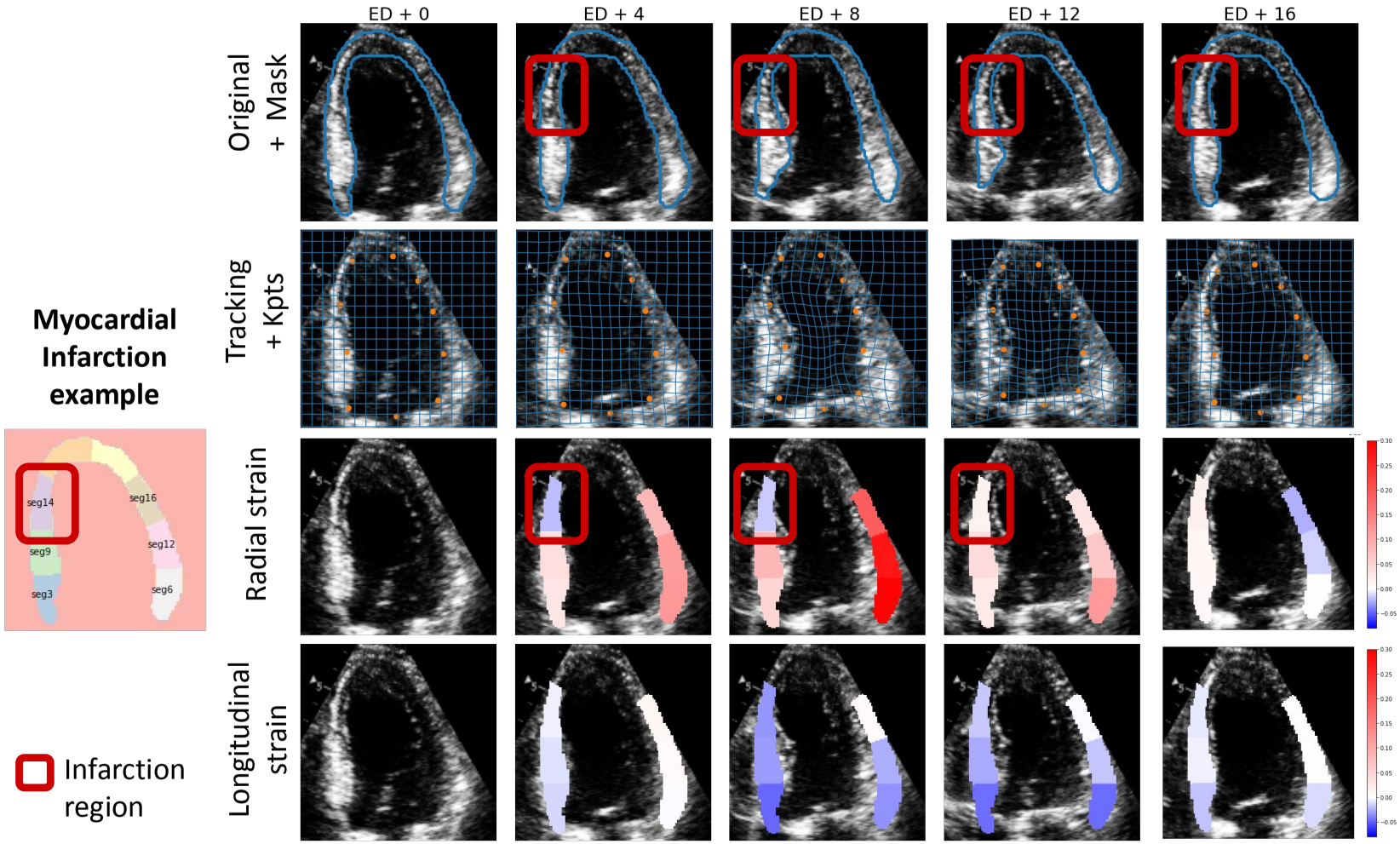


Fig. 3.6.: Visual results of a myocardial infarction sequence from HMC-QU dataset where SEG14 is annotated as infarcted region.

3.6 Appendix

In this appendix, we demonstrate other advantages of using PAM for echocardiography analysis.

3.6.1 Cardiac motion transfer between sequences

One important application is to generate new synthetic datasets that could be used for data augmentation or representation learning and others. PAM model facilitates motion transfer from one sequence to another while preserving the appearance along the total sequence. This is different from other deep learning methods that deform one single frame (such as end-diastole frame) to obtain new sequences [Krebs, 2020].

Given two sequences $\{X_i, i \in N\}$ and $\{Y_i, i \in N\}$, our objective was to transfer motion of $\{X_i\}$ to $\{Y_i\}$. We denoted the new sequence of Y as G. The motion transfer process contained the following steps:

1. Motion estimation: By applying PAM model to input sequences, we obtained a sequence of key points $\{p^k_{X_i}\}$, $\{p^k_{Y_i}\}$ and local affine transform from abstract reference frames $\{A^k_{X_i \leftarrow R}\}$, $\{A^k_{Y_i \leftarrow R}\}$ respectively for $\{X_i\}$ and $\{Y_i\}$.

2. Align ED and ES frames: We aligned the key point sequences and local affine transform sequences temporally through linear interpolation such that ED and ES frames of sequences $\{X_i\}$ were aligned to $\{Y_i\}$ (i.e. $t^Y_{ED} = t^X_{ED}, t^Y_{ES} = t^X_{ES}, t \in N$). We denoted the interpolated key point and local affine transform as $\{\bar{p}^k_{X_i}\}$ and $\{\bar{A}^k_{X_i \leftarrow R}\}$.The ED and ES frames of each image sequence were identified either by detection of volume change extremes or from synchronized ECG signal.

3. Relative motion transfer:

- 1. We deformed all image frames of $\{Y_i\}$ and its corresponding key points $\{p^k_{Y_i}\}$ to ED frame through

$$\begin{aligned} \hat{Y}_i &= Y_i(\mathcal{T}_{i \leftarrow ED}(\cdot)), \\ \hat{p}^k_{Y_i} &= \arg\max H(p^k_{Y_i}, \sigma^2_H)(\mathcal{T}_{i \leftarrow ED}(\cdot)) \end{aligned} \tag{3.11}$$

  where $\mathcal{T}_{i \leftarrow ED}$ was from the polyaffine dense motion according to Equation 3.4 and Equation 3.2 using its key points and local affine transform matrices. $\hat{Y}_i$ represented a sequence of pseudo ED frames of $\{Y_i\}$.

- 2. We transferred relative motion between X_{ED} and other frames X_i to $\hat{Y}$ frame-wisely, to obtain the objective sequence G. Specifically, we applied $\bar{A}^k_{X_{ED}\leftarrow X_i}$ to the neighbourhood of $\hat{p}^k_{Y_i}$

$$\mathcal{T}^k_{\hat{Y}_i\leftarrow G_i}(z) = \bar{A}^k_{X_{ED}\leftarrow X_i}\cdot z + (\hat{p}^k_{Y_i} + \bar{A}^k_{X_{ED}\leftarrow X_i}(p^k_{X_i} - p^k_{X_{ED}} + \hat{p}^k_{Y_i})) \tag{3.12}$$

where $\bar{A}^k_{X_{ED}\leftarrow X_i} = A^k_{X_{ED}\leftarrow R}(A^k_{X_i\leftarrow R})^{-1}$. We obtained the final dense motion through polyaffine motion fusion

$$\mathcal{T}_{\hat{Y}_i\leftarrow G_i}(z) = \bar{W}_0 z + \sum_{k=1}^{K}\bar{W}_k(p^k_{X_i})\mathcal{T}^k_{\hat{Y}_i\leftarrow G_i}(z) \tag{3.13}$$

where we applied directly the key point neighbour weight from sequence X_i.

We present two different examples of transferring motion between patients with different ejection fraction (EF) in Figure 3.7 and Figure 3.8.

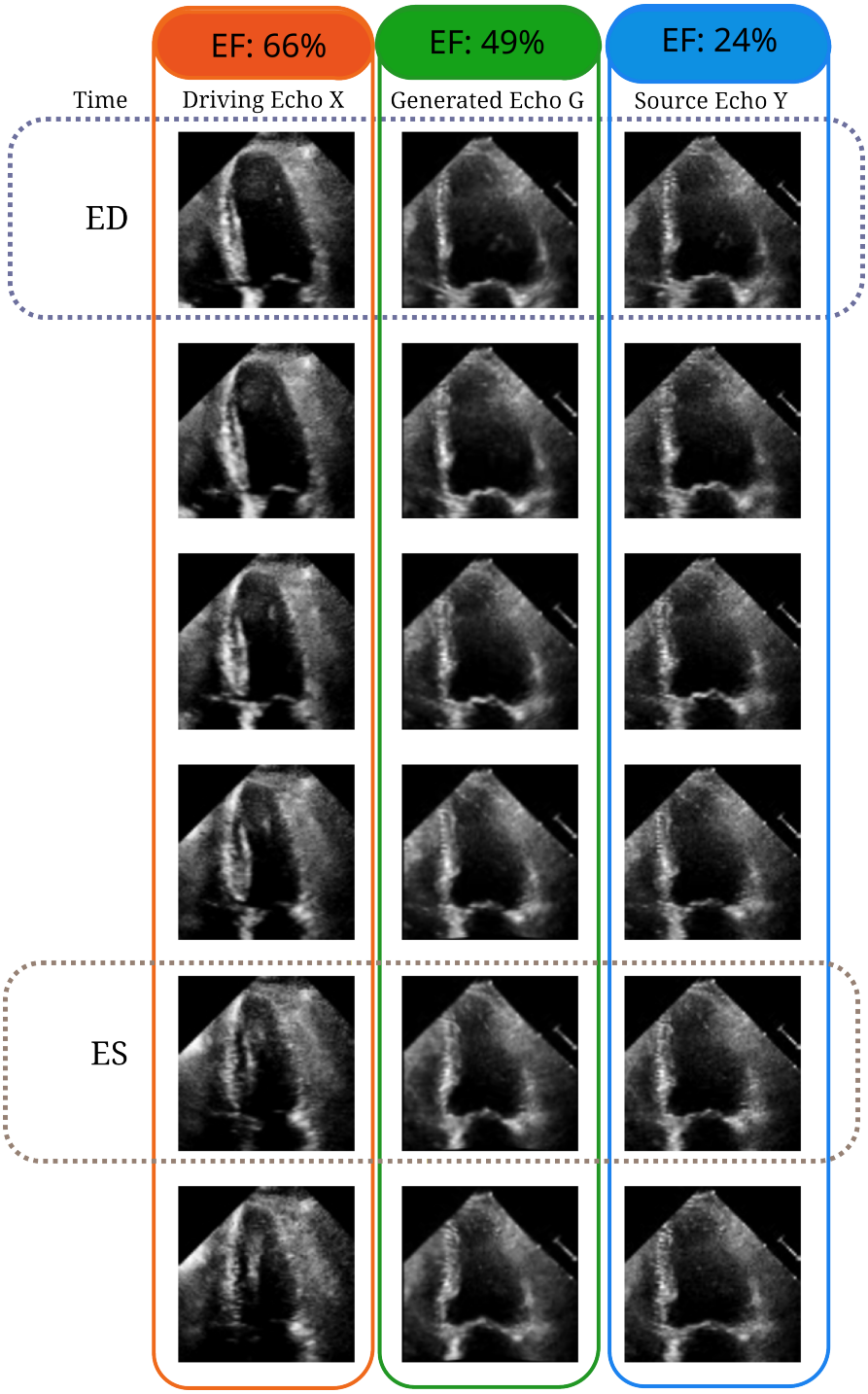


Fig. 3.7.: An example of transferring motion from sequences with large EF to sequences with small EF.

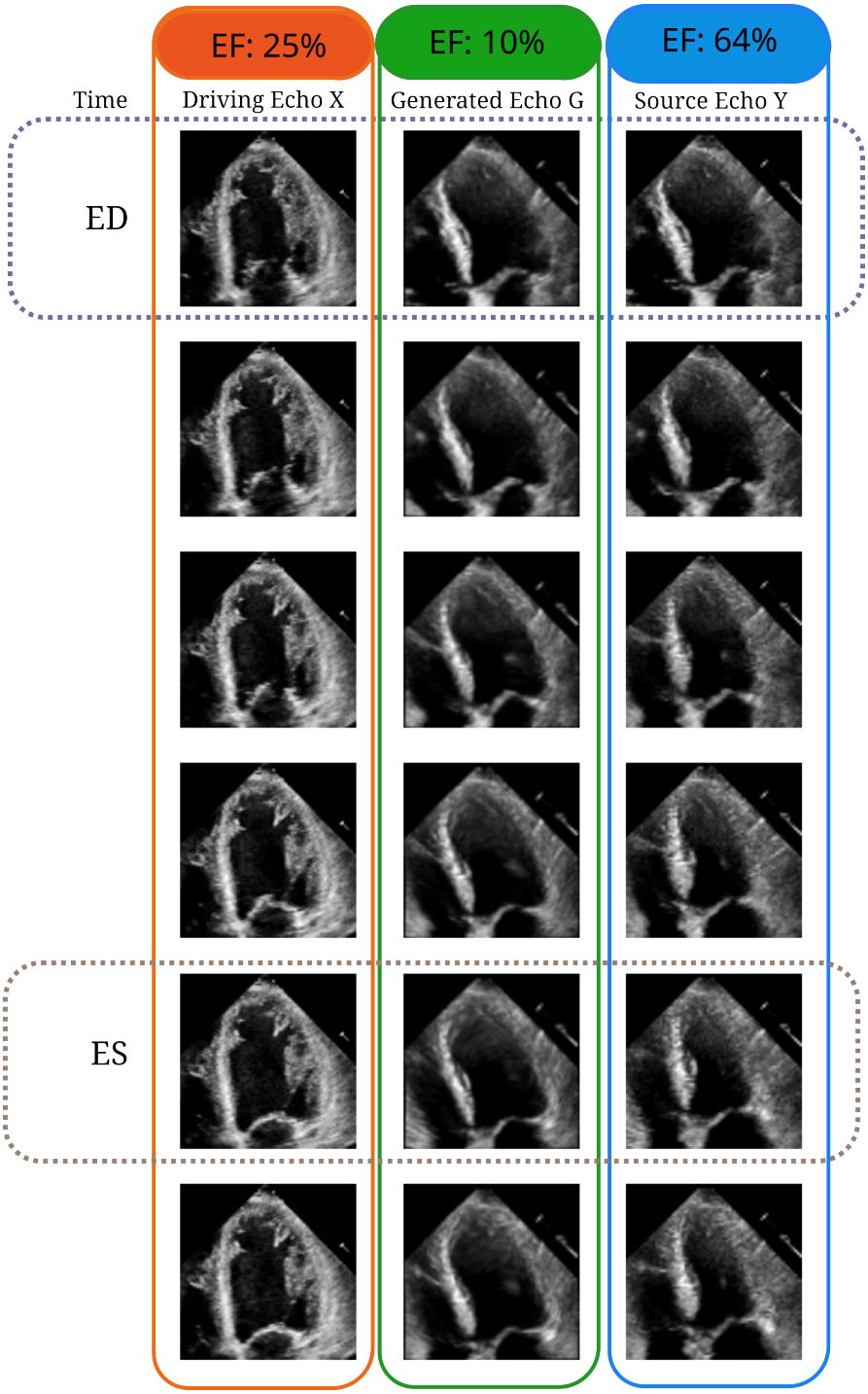


Fig. 3.8.: An example of transferring motion from sequences with small EF to sequences with large EF.

4 Echocardiography analysis pipeline

Contents

In chapter 2 and chapter 3, we have proposed generalisable solutions for cardiac segmentation and cardiac motion tracking, respectively. It is unknown how these models perform in a down-stream task, such as cardiac pathology detection.

In this chapter, we propose a robust and generalisable pipeline for echocardiography analysis, notably with application in myocardial infarction detection. This chapter is partially included in a journal submission [Yang, 2023a].

The main contributions of this chapter are listed as follows:

- We introduce a robust pipeline for echocardiography analysis using deep learning that make use of shape prior and motion prior (Section 4.2).
- We provide evaluation results across different datasets and compared them with a benchmark study (Section 4.4).

4.1 Introduction

As stated in former chapters, automatic algorithms play an important role for cardiac function analysis. Segmentation and motion tracking are two most used approaches to fulfill such objective.

From segmentation output, it is not hard to obtain left ventricle ejection fraction (LVEF) using Simpson's method. Although there exists numerous works that deploy neural networks in order to directly estimate LVEF, we rely on the segmentation mask, this prioritizing its interpretability in our approach.

From deformation field, we can obtain another important index of cardiac function: the global longitudinal strain (GLS). GLS reveals cardiac contractility and is reported to be more sensitive in detecting early cardiac diseases than LVEF [Cikes, 2015]. Computation of GLS can stem from cardiac landmarks [Smistad, 2020; Østvik, 2021] or dense deformation fields [Morales, 2021]. The strain tensor's sensitivity to the smoothness of the deformation field necessitates a regularization in deep learning models. Common strategies include Gaussian smoothing layers [Krebs, 2020] and rigidity penalties [Staring, 2007; De Vos, 2019]. However, these techniques often apply within limited pixel neighborhoods in the image domain and might not consistently adhere to anatomical constraints. In our approach, we addressed this issue through a poly-affine motion framework, which serves as a well-established motion approximation for myocardial movement [McLeod, 2015]. By integrating a motion prior, we can achieve the capability to track myocardial motion with a small number of parameters.

4.1.1 Myocardial infarction detection in echocardiography

Deep learning and machine learning-based classification methods have demonstrated good performance in myocardial infarction (MI) detection using 2D echocardiography. Convolutional neural networks are commonly employed to extract deep features from 2D echocardiography images for classification purposes [Kusunose, 2020; Huang, 2020; Omar, 2018]. Additionally, handcrafted features, including frequency-related attributes, have been extracted from 2D echocardiography images to aid in MI detection [Liu, 2023; Raghavendra, 2018]. However, much of the existing research relies on training models with local large datasets, posing limitations with respect to external validation by other researchers. A notable exception is the work of Degerli et al. [Degerli, 2024], who introduced a publicly available dataset for MI detection and assessed its performance using segment displacement features. Given the relatively small size of this dataset (130 patients), we opt to extract interpretable features from segmentation and motion

estimation outputs. Subsequently, we employ machine learning techniques for MI detection.

4.2 Method

The proposed pipeline for 2D echocardiography analysis consisted of three steps. First, a segmentation module estimated the detailed structure of left ventricle, from which the value of ejection fraction was extracted. Second, a motion module predicted the regional myocardium displacement, where strain values can be easily computed. Third, a classification module took the predicted indexes as input and detected patients with myocardial infarction in a multi-view setting.

4.2.1 Shape and motion priors for echocardiography analysis

We used the segmentation model SEG-CONTOUR as introduced in Section 2.2.3, and the motion tracking model PAM as introduced in Section 3.2. The segmentation model incorporated a shape prior through a contour loss. The motion tracking model integrated a motion prior through weakly-supervised myocardium key point position. A summary of the proposed method is illustrated in Figure 4.1.

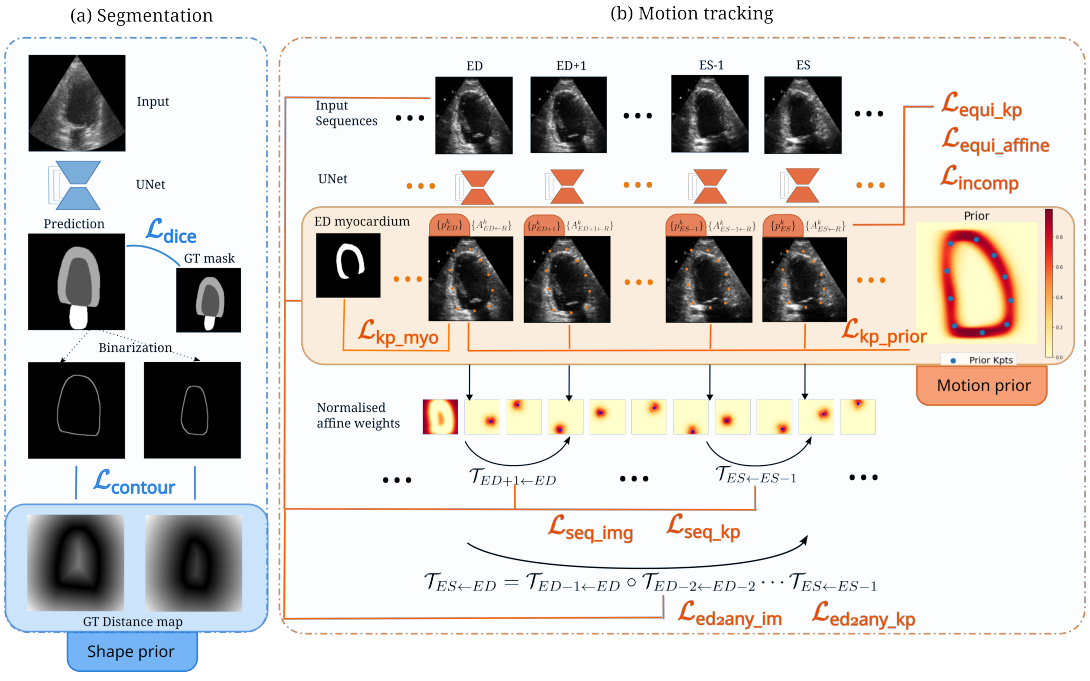


Fig. 4.1.: Shape and motion priors for echocardiography analysis.

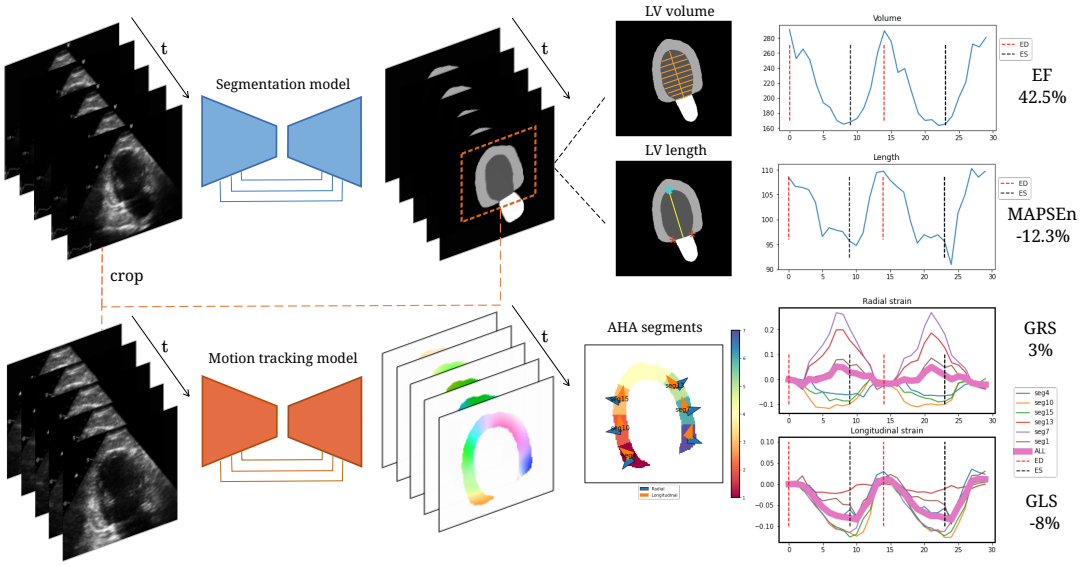


Fig. 4.2.: Pipeline for interpretable cardiac features extraction.

4.2.2 Multi-view myocardial infarction detection

In our study, we chose to formulate a 8-dimension feature vector for multi-view myocardial infarction classification. This vector contained volume and deformation related features, that could be extracted from either segmentation output and motion estimation output. The 8-dimension feature was a concatenation of two 4-dimension vectors from 2-chamber view and 4-chamber view, respectively. Each 4-dimension feature consisted of the following: ejection fraction (EF); global longitudinal strain (GLS) at end-systole; global radial strain (GRS) at end-systole; and, normalised mitral annular plane systolic excursion (MAPSEn). The pipeline is depicted in Figure 4.2.

We choose global features for downstream classification for two primary reasons. Firstly, global features show lower sensitivity to noise and domain shift issues compared to local features. Local features depend on precise segmentation and tracking results, introducing more noise into downstream training when segmentation or tracking is not perfect. In contrast, global features induce less noise due to the denoised average. Secondly, global features provide greater standardization across various views and patients, while local features are contingent on segment split, which may lack consistency across different patients.

4.2.2.1 Ejection fraction

We applied Simpson's rule [Folse, 1962] to the estimated left ventricle (LV) segmentation output to obtain approximated volume curve along cardiac sequences. End-diastole and

end-systole frames were identified as when the volume curve is at its largest and smallest point. EF was calculated following this formula EF $= 1 - \frac{V_{ES}}{V_{ED}}$.

4.2.2.2 Normalised Mitral annular plane systolic excursion (MAPSEn)

MAPSE measures the displacement of mitral annular plane at systole, which is a helpful cardiac index for the evaluation of LV systolic function. MAPSEn normalizes MAPSE for LV length [Støylen, 2018].

$$\text{MAPSEn} = \frac{\text{MAPSE}}{L_{ED}} - 1 = \frac{L_{ES} - L_{ED}}{L_{ED}} \tag{4.1}$$

where L_t refers to the LV length at time t. We computed the LV length by connecting the apex and mid-basal points from segmentation outputs.

4.2.2.3 Myocardial strain

We took the end-diastole myocardium segmentation output as the region of interest (ROI) and obtained the pixel displacement within this ROI from motion estimation output across the whole cardiac sequences. For all myocardial pixels at end-diastole, we obtained the 2D Lagrangian finite strain tensor $\mathbf{E}$ using

$$\mathbf{E} = \frac{1}{2}(\mathbf{F}^T\mathbf{F} - \mathbf{I}), \tag{4.2}$$

where $\mathbf{I}$ was the identity matrix and $\mathbf{F}$ was the deformation gradient tensor. Radial direction was computed as the tangent field of the correspondence trajectories between endocardial contour and epicardial contour [Yezzi, 2003]. Longitudinal direction was taken perpendicular to radial direction.

We next identified 8 segments by clustering all the myocardium pixels into 8 clustering using the KMeans method, where the segments 4 and 5 were considered to form together an apex segment. For each segment, one single radial direction and one longitudinal direction were kept, representing the mean of all pixel directions within the chosen segment. Segment strain was the averaged value of pixel strain projected into its segment-wise radial and longitudinal direction. The average strain of all segments were regarded as the global strain. Strains at the end-systolic phase were extracted as features for myocardial infarction detection.

4.3 Experiments

4.3.1 Datasets

Three public datasets (i.e., CAMUS, EchoNet, HMCQU) and one private dataset (i.e., CHU) with 2D echocardiography data were included in our study.

Tab. 4.1.: Datasets inncluded in this study

Dataset	Training Task	Pathology	View		
			A2C	A4C	2+4
CAMUS	Segmentation	–	500	500	500
EchoNet	motion estimation	–	–	10030	–
HMC-QU	Classification	MI/Non-MI	130	162	130
CHU	–	MI/Non-MI	66	76	64

Tab. 4.2.: Datasets with pathology diagnostics

Dataset	MI	Non-MI	Total
HMC-QU	88	42	130
CHU	42	22	64

CAMUS dataset [Leclerc, 2019a] contains apical 4-chamber (A4C) view and apical 2-chamber (A2C) view of 500 patients. The following annotations are also publicly available: segmentation, volume information and ejection fraction of the left heart at end-diastole (ED) and at end-systole (ES). The dataset contains an official train-test split of 450 patients vs. 50 patients.

EchoNet-Dynamic [Ouyang, 2020] is a large open dataset of 10030 A4C echocardiography videos. Left ventricle traces at ED and ES as well as EF values are provided for all videos. The official split of train vs. validation vs. test is 7465/1288/1277.

HMC-QU [Degerli, 2024] is the first public dataset for MI detection. It contains 162 A4C videos, 130 A2C videos, where A2C and A4C views of 130 patients are both available. Notably, this dataset features annotated infarct wall segments across all videos. Moreover, for 109 of the A4C videos, the dataset supplies segmentation masks for the myocardium, generated through a semi-automatic approach [Degerli, 2021a].

CHU dataset contains 2D echocardiography sequences of 76 patients collected retrospectively from Nice CHU hospital. This study protocol was in compliance with the declaration of Helsinki and was approved by the local research Ethics committee. These examinations were conducted and interpreted by four different cardiologists, with only one cardiologist per examination. This clinical dataset contains challenging echocardiography images, with various image quality, coming from different caring scenarios and device vendors. This dataset serves as an evaluation set for the real-world patient monitoring.

A detailed summary of all the datasets are presented in Table.4.1 and Table.4.2.

4.3.2 Experiments and implementation

The segmentation model and the motion tracking model were implemented as described in Chapter 2 and Chapter 3 respectively.

4.3.2.1 Myocardial infarction detection

We trained the segmentation model and the motion tracking model using CAMUS training data and EchoNet training data respectively (detailed evaluation was presented in Chapter 2 and Chapter 3). In order to perform MI detection, we extracted cardiac features from segmentation and motion estimation outputs in the HMC-QU dataset and CHU dataset, respectively. In particular, we performed 5-fold cross-validation on HMC-QU dataset using multi-view features (130 patients) and evaluated the classifier directly without retraining or finetuning on CHU dataset (64 patients).

Implementation: We applied two popular machine learning methods from sklearn package [Pedregosa, 2011] for classification: random forest (RF) and support vector machine (SVM), respectively. A resampling strategy was conducted before classification training for the unbalanced HMC-QU dataset.

The best performing method on HMC-QU dataset used normalised segment displacement as features [Degerli, 2024].

$$f_k = \frac{\max D_k(t)}{\min I_{(k,p)}(t)} \tag{4.3}$$

where $D_k(t)$ referred to the mean displacement of segment k at time t, and $I_{(k,p)}(t)$ represented the averaged Manhattan distance between segment k and its opposite segment p. Authors in [Degerli, 2024] used active polynomials to extract displacement information. In our case, we obtained the normalised segment displacement using the predicted segmentation and motion estimation results, which resulted in a 6-dimension vector for each apical view. We denoted the method using the proposed global features as 'global', and the method using normalised displacement as 'local'.

4.4 Results and Discussion

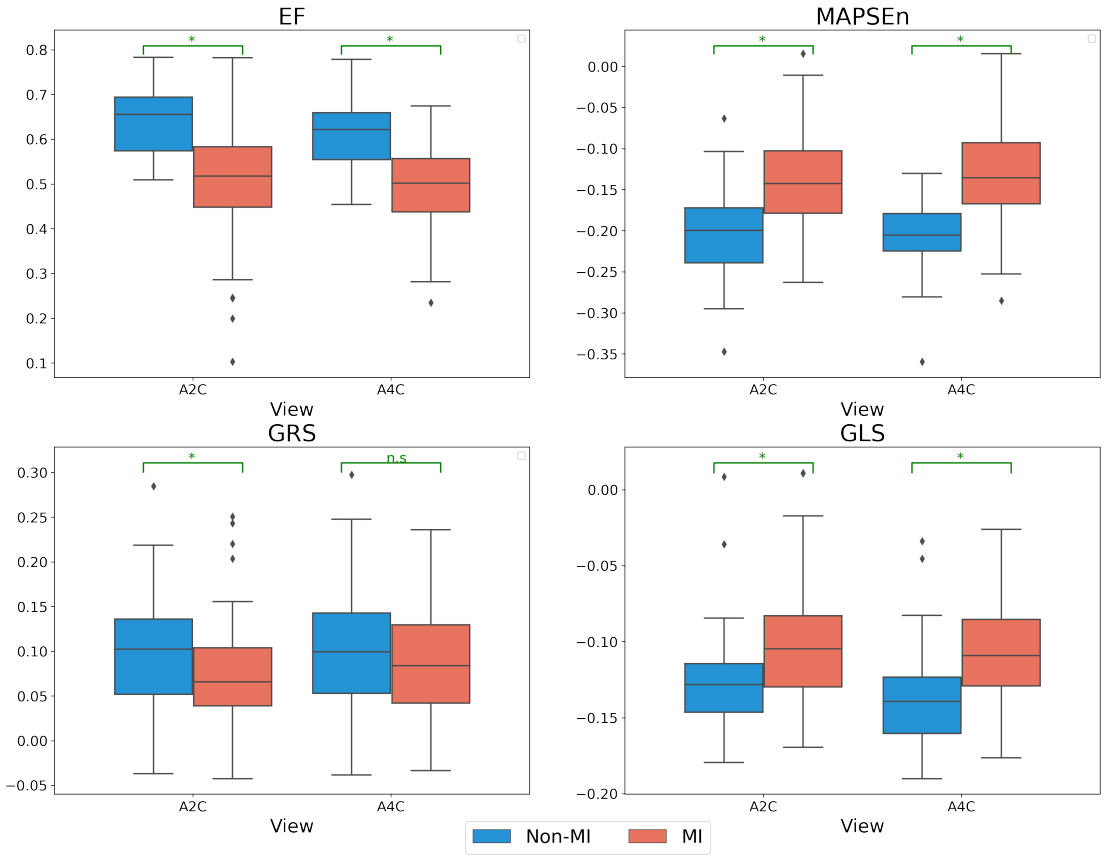


Fig. 4.3.: Distribution of global features between healthy (Non-MI) and MI patients from HMC-QU dataset. (Green $*$: *p-value* < 0.05 under the one-sided assumption using Wilcoxon rank-sum test.)

Tab. 4.3.: Myocardial infarction detection results

Method	Feature	Sen.	Spe.	Prec.	F1.	Accuracy
HMC-QU (*5 fold cross-validation*)						
RF	[Degerli, 2024]	0.875	0.619	0.828	**0.951**	0.792
SVM	[Degerli, 2024]	**0.910**	0.429	0.769	0.877	0.754
RF	Local	0.773	0.714	0.850	0.810	0.754
SVM	Local	0.761	0.786	0.881	0.817	0.769
RF	Global	0.841	0.786	0.891	0.865	**0.823**
SVM	Global	0.784	**0.833**	**0.908**	0.841	0.800
CHU (*Evaluation*)						
RF	Local	**1.000**	0.091	0.677	0.808	0.688
SVM	Local	0.952	0.273	0.714	0.816	0.719
RF	Global	0.952	0.409	0.754	**0.842**	**0.766**
SVM	Global	0.881	**0.545**	**0.787**	0.831	**0.766**

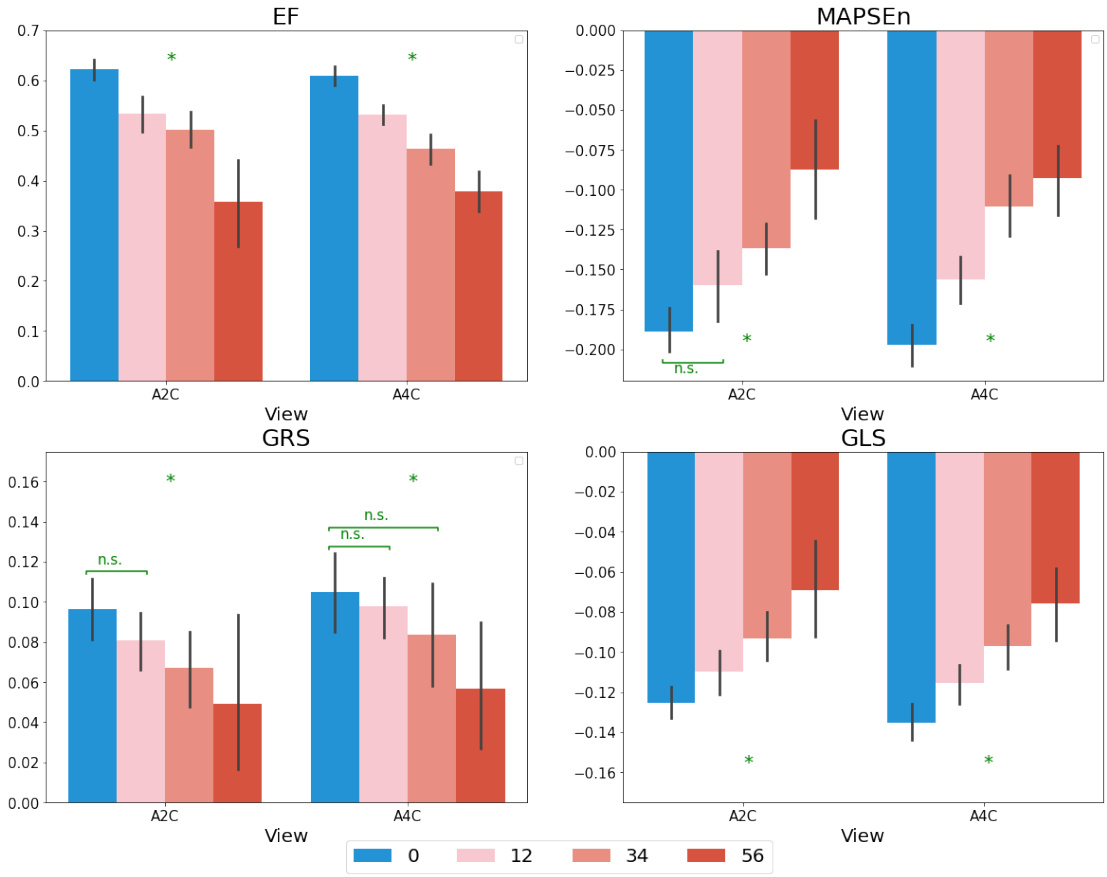


Fig. 4.4.: Distribution of global features between patients with different number of infarct segments from HMC-QU dataset. Legend 0: non-MI, 12: MI patients with 1-2 infarct segments, 34: MI with 3-4 segments, 56: MI with 5-6 segments. Wilcoxon rank-sum test is performed between non-MI and others. (Green $*$: *p-value* < 0.05)

4.4.1 HMC-QU

First, a significant group difference was observed between MI patients and non-MI patients in the four global parameters on the HMC-QU dataset, as illustrated in Figure 4.3, except for the GRS feature for the apical 4-chamber view.

Second, the results presented in Table 4.3 demonstrate that our method, utilizing global features, outperformed the use of local features, including the state-of-the-art result reported in [Degerli, 2024]. This showcases the effectiveness and generalizability of the proposed pipeline for myocardial infarction detection. Notably, local features extracted using a trained segmentation model and motion tracking model exhibited similar classification accuracy compared to those obtained through the time-consuming active polynomial method.

However, the use of global features for MI detection had its limitations. Further analysis, as illustrated in Figure 4.5, revealed that our proposed method using global features

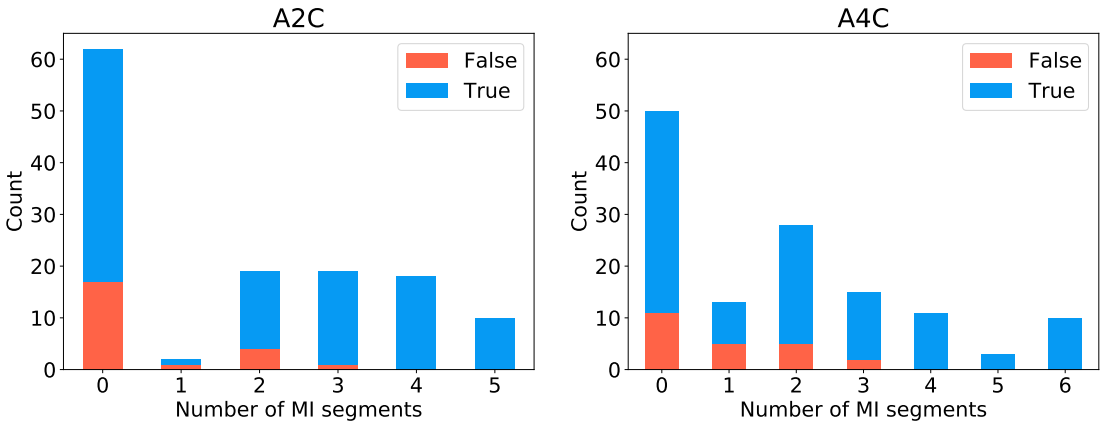


Fig. 4.5.: 5-fold cross validation result of MI detection using global features and SVM classifier. False: wrong classification, True: correct classification.

faced challenges in differentiating healthy samples from MI patients with fewer infarct segments. This observation aligns with the linear relation observed between the number of infarct segments and global features, as shown in Figure 4.4.

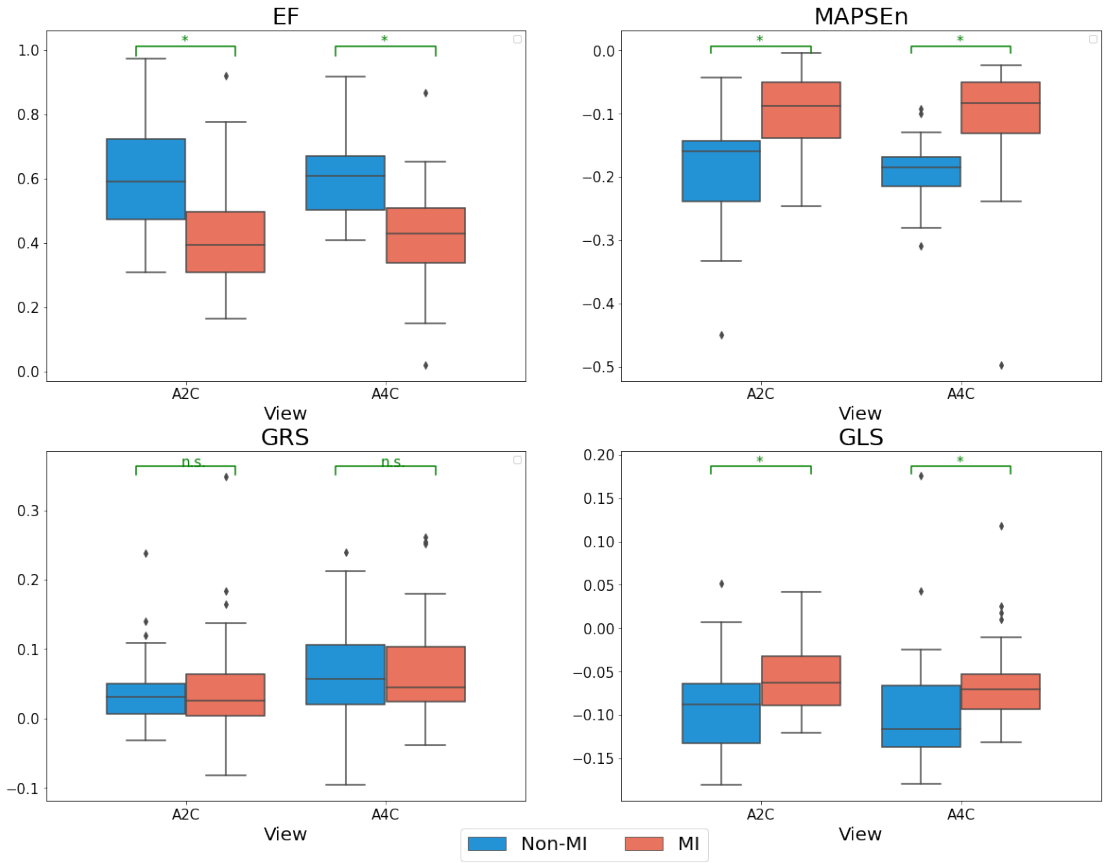


Fig. 4.6.: Distribution of global features between healthy (Non-MI) and MI patients from CHU dataset. (Green $*$: *p-value* < 0.05 under the one-sided assumption using Wilcoxon rank-sum test.)

4.4.2 CHU

As shown in the lower part of Table 4.3, classifiers that utilized global features demonstrated better classification performance on the CHU dataset. This is evident from comparable sensitivity, higher specificity, higher precision, higher F1-score, and higher accuracy. Local features tend to be more sensitive to noisy output from segmentation and motion tracking model due to the domain shift between the test dataset (CHU) and the training dataset (HMC-QU). In contrast, global features, often derived from averaged measurements, exhibit greater stability compared to local features, resulting in superior performance.

Additionally, a significant group difference between MI patients and non-MI patients on the CHU data was observed (Figure 4.6). The evaluation results on the CHU dataset strongly reinforce the robustness and good generalisability of our proposed pipeline for echocardiography analysis.

4.4.3 Discussion

First, all features of the HMC-QU dataset and CHU dataset were extracted from segmentation and motion estimation outputs, whose models were trained on other datasets. We observed significant differences in specific measurements between healthy and MI patients on these two unseen datasets, highlighting the effectiveness and robustness of our segmentation and motion estimation methods in distinguishing between the two groups.

Second, the application of the trained classifier on a real-world unseen dataset (CHU dataset) further supports the robustness of using global features for MI detection. However, we acknowledge that using global features for MI detection faces challenges in identifying patients with only one or two segments containing infarcted myocardium.

Future work will focus on the construction of effective local features and the integration of a second modality, such as an echocardiogram. ECG, revealing the electrical morphology of infarction patients from another perspective, will be explored with echocardiography together in Chapter 6.

4.5 Conclusion

In this chapter, we proposed a novel and generalizable pipeline for interpretable 2D echocardiography analysis, incorporating a segmentation model with a shape prior, a

motion estimation model with a motion prior, and a classification module using clinical features. Our experiments and comprehensive evaluation demonstrated the robustness of our proposed method compared with other competing methods at different stages. The validation conducted on diverse public and private datasets, including HMC-QU dataset, serves as benchmark results, highlighting the generalizability of our approach in echocardiography analysis.

5 Explainable analysis of electrocardiogram

Contents

Abstract Automatic analysis of electrocardiograms with adequate explainability is a challenging task. Many deep learning based methods have been proposed for automatic classification of electrocardiograms. However, very few of them provide detailed explainable classification evidence.

In this chapter, we explore explainable ECG classification through explicit decomposition of single-beat (median-beat) ECG signal. In particular, every single-beat ECG sample is decomposed into five subwaves and each subwave is parameterised by a Frequency Modulated Moebius. Those parameters have explicit meanings for ECG interpretation. In stead of solving the optimisation problem iteratively which is time-consuming, we design a Cascaded CNN network to estimate the parameters for each single-beat ECG signal. Our preliminary results show that with appropriate position regularisation strategy, our neural network is able to estimate the subwave for P, Q, R, S, T events and maintain a good reconstruction accuracy (with R2 score 0.94 on test dataset of PTB-XL) in a unsupervised manner. Using the estimated parameters, we achieved very good classification and generalisation performance on myocardial infarction detection on four different datasets. Features of high importance are in accordance with clinical interpretations.

This chapter was published in the Proceedings of the International Workshop on Statistical Atlases and Computational Models of the Heart (STACOM) [Yang, 2022]. The main contributions of this chapter are listed as follows:

- We propose a time-efficient and automatic pipeline for ECG decomposition and reconstruction by passing through a deep learning model: Cascaded FMMnet, which is capable to reconstruct single beat signals with high quality. The training is unsupervised and make it accessible for all kinds of ECG datasets, with or without annotations. The estimated parameters have explainable meanings for each subwave, such as the amplitude, the position etc. (Section 5.2)

- We have conducted comprehensive experiments using the estimated parameters as features to classify normal and myocardial infarction patients across different datasets. The important features identified are in accordance with the clinical interpretation indexes, such as T wave and Q wave change. (Section 5.3)

5.1 Introduction

Myocardial infarction (MI) is a kind of pathology where myocardial cells are necrotic due to the prolonged lack of oxygen (ischaemia). A patient is diagnosed as MI if he/she has elevated cardiac troponin values, and falls into at least one of the following conditions: symptoms of myocardial ischaemia; new changes of ST-segment/T-wave in electrocardiogram (ECG); development of pathological Q waves in ECG; abnormal myocardium motion; and, presence of coronary thrombus [Thygesen, 2018]. Among all diagnostic approaches, ECG is very easy and fast to perform on patients, even by non-experts. However, accurately diagnosing MI patients using ECG recordings is a challenging task. For example, one study showed that experienced cardiologists only identified 82% of the real ST-segment elevated MI patients [Mixon, 2012]. Computer-assisted ECG analysis could help cardiologists, non-experts better interpret ECG recordings for MI detection.

There exist many research works exploring automatic MI detection using ECG signals and they can be categorised into feature-based methods and neural network-based methods, respectively. The feature-based methods usually contains three stages: ECG delineation (segmentation); feature extraction; and. classification. Different kinds of features were explored: morphological features (such as ST-elevation value, QRS duration, T wave amplitude, Q wave amplitude, etc.) [Lu, 2000; Arif, 2012; Jaleel, 2016]; wavelet transform related features (coefficients) [Pereira, 2016; Jayachandran, 2010; Bhaskar, 2015]; empirical mode decomposition features [Acharya, 2017a], and so on. Compared with other features, morphological features are explainable but very sensitive to ECG delineation results. ECG delineation methods [Graja, 2005; Peimankar, 2021; Jimenez-

Perez, 2021], usually depend on annotations for all of the events (P,Q,R,S,T) to train models. They are constrained by the size and type of pathology in the annotated dataset. The neural network based methods for MI detection have overwhelmed in recent years. For example, single-beat ECG signals are either directly classified using a 1D convolutional neural network (CNN) [Acharya, 2017b] or are transformed into 12-lead 2D image for input into a 2D CNN [Fang, 2022]. For more detailed reviews of MI detection methods, interested readers are invited to read reviews [Ansari, 2017; Xiong, 2022].

In our work, we tackle this MI detection problem using the feature-based approach by ECG parameterisation. Actually, ECG parameterisation/modelling is not a new idea. For example, Liu et al. [Liu, 2015] proposed to fit a 20th order polynomial function to a given ECG signal and used the fitted coefficients as ECG features. Second-order ODEs were applied to model the 12-lead ECGs and the estimated time-varying coefficients were used as features [Zewdie, 2014]. Both methods achieved good accuracy for MI detection; however, the features did not have an explainable meaning for clinicians. In our case, we adapted the explainable Frequency Modulated Moebius (FMM) model as our parameterisation model [Rueda, 2021] for single-beat ECG.

5.2 Methods

The pipeline of our explainable ECG analysis consisted of three stages (Figure 5.1(a)). First, the 12-lead ECG recordings were filtered and segmented to obtain single-beat median ECG signals. Second, each lead-wise median signal was passed through an encoder network (Cascaded FMMnet) for subwave decomposition. The encoder network generated 21 estimated parameters crucial for signal reconstruction and interpretation. Third, for each sample (12-lead median ECG signal), myocardial infarction classification was conducted based on the 264(21x12+12) estimated parameters, where 12 additional parameters came from estimated ST-segment voltage value. To explain the classification result, we used the additive weighted features for linear classification models and SHAP value [Lundberg, 2017] for non-linear classification models. Finally, decomposed waveform and feature importance from classifier together gave visual and quantitative explainability of our prediction result.

5.2.1 Data Preprocessing

The preprocessing included 5 steps: resampling; filtering; R-peak detection; ECG segmentation; and, median signal generation.

The original 12-lead ECG recording was resampled to 500Hz if its original sampling rate was not 500Hz. A butterworth high-pass filter with cutoff frequency at 0.5Hz was

then applied to remove baseline wander. The R-peaks of Lead II were automatically detected and used as reference for all the other leads. For every lead ECG, each single beat segment was set from 35% heart beat duration (s) before the R-peak to 50% heart beat duration (s) after the R-peak. One 1.2-second median beat signal was calculated by aligning the R-peaks of all the single beats (at 0.5-second position) and padded by neighbouring values at the two ends if the medial signal was shorter than 1.2s. Neurokit2 package [Makowski, 2021] was used for filtering, R-peak detection and single beat segment calculation.

5.2.2 Cascaded FMMnet

In order to reinforce explainability in automatic ECG analysis, we utilised the decomposition model proposed by [Rueda, 2021]. The idea was to approximate the single-beat ECG signal by composition of five subwaves (P,Q,R,S,T), each of which was parameterised by a Frequency Modulated Moebius.

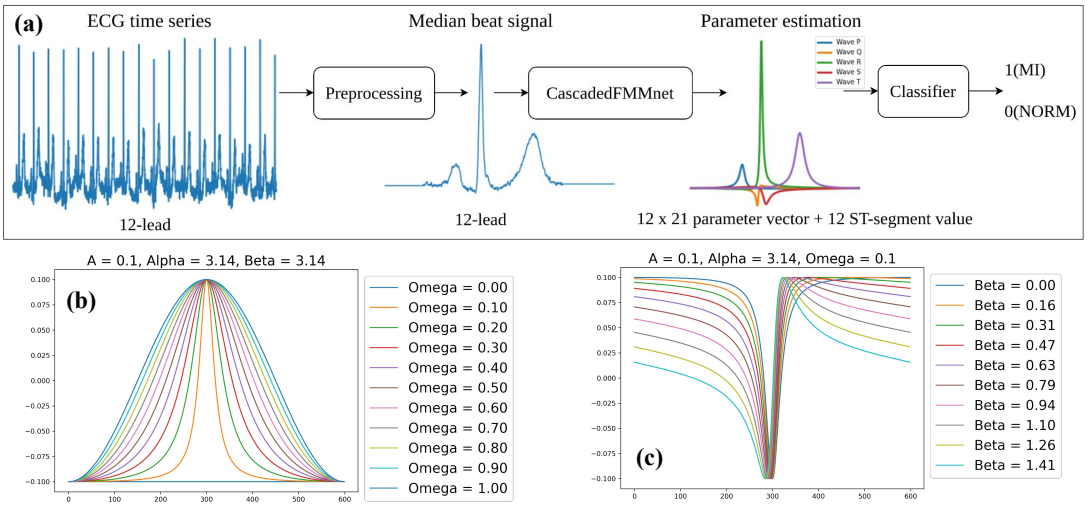


Fig. 5.1.: (a) Pipeline of our proposed explainable ECG classification. (b)ω controls the kurtosis of the wave signal. (d) β controls the skewness of the wave signal.

Assuming $X(t_i)$, $t_i \in [0, 2\pi]$, the original signal of a single-beat ECG record, could be decomposed into five subwaves W_s, $s \in \{P, Q, R, S, T\}$. Each subwave was described by a four-dimensional parameter $p_s = \{A_s, \alpha_s, \beta_s, \omega_s\}$ respectively,

$$W_s(t, p_s) = A_s \cos(\beta_s + 2 \arctan(\omega_s \tan(\frac{t - \alpha_s}{2}))) \tag{5.1}$$

where A, α, β, ω controlled the absolute amplitude, the position, the skewness and the kurtosis of the waveform (see Figure 5.1 (c)(d)). The approximation of original signal

$\hat{X}(t)$ was defined as the addition of the five subwaves W_s and an additional baseline shift M,

$$\hat{X}(t, \theta, M) = M + \sum_{s \in \{P,Q,R,S,T\}} W_s(t, p_s) \tag{5.2}$$

where $\theta = (M, p_S, p_Q, p_R, p_S, p_T)$ and they verified the following ranges

- 1. $M \in \mathcal{R}$
- 2. $p_s \in \mathcal{R}^+ \times [0, 2\pi] \times [0, 2\pi] \times [0, 1], s \in \{P, Q, R, S, T\}$
- 3. $\alpha_P \leq \alpha_Q \leq \alpha_R \leq \alpha_S \leq \alpha_T$

The aim of decomposition was to estimate the optimal 21 parameters $\hat{\theta}$ that best fit $\hat{\theta} = argmin_\theta \sum_{i=1}^{n} [X(t_i) - \hat{X}(t_i)]^2$.

Instead of using the computationally intensive iterative optimisation [Rueda, 2021], we estimated the 21 parameters through a data-driven deep learning model: namely the Cascaded FMMnet. This network consisted of 5 identical cascaded sub-network, each of which was responsible for estimating 5 parameters $(M_i, A_i, \alpha_i, \beta_i, \omega_i)$ of one subwave S_i, where $S_i(t) = M_i + A_i \cos(\beta_i + 2 \arctan(\omega_i \tan(\frac{t-\alpha_i}{2})))$. Assuming the original signal $X(t)$, the input of the ith sub-network $X_i(t)$ was the residual of the original signal subtracting former subwaves, i.e. $X_i(t) = X_{i-1}(t) - S_{i-1}(t)$, where $i \in [1, 5], X_0(t) = X(t), S_0(t) = 0$.

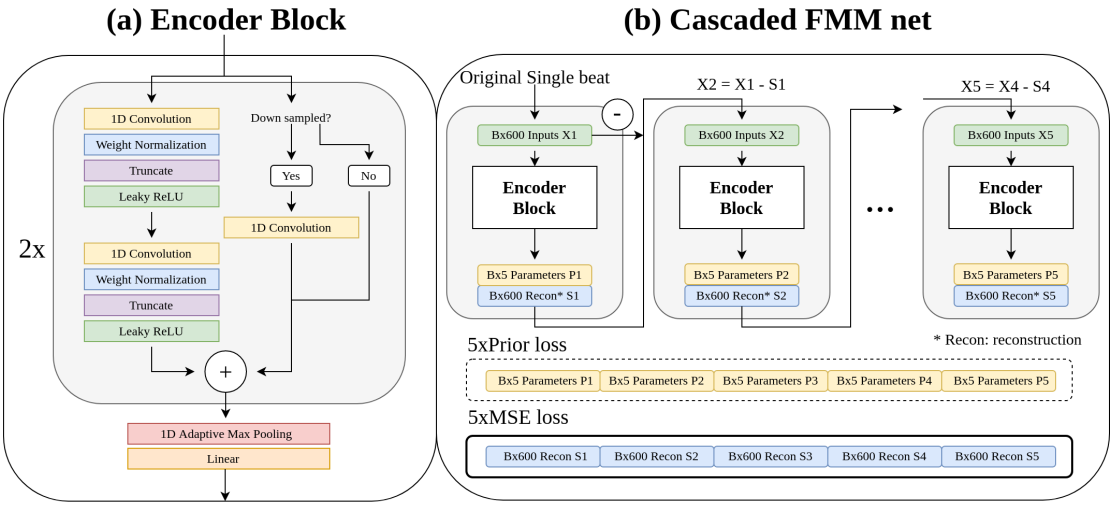


Fig. 5.2.: The detailed architecture of Cascaded FMMnet.

The encoder block in our network was comprised of: 2 stacks of causal convolution with down-sampled skip-connection; 1 max-pooling layer; and, 2 linear layers. It took an input of 1x600 dimension and outputted a 21-dimension vector which was the estimation of the parameters. The input median ECG signal was resized to be within the range of

$[-1, 1]$ and the last linear layer had a Sigmoid activation for $(A_i, \alpha_i, \beta_i, \omega_i)$ and a Tanh activation for M_i parameter. The final M was the sum of all the $M_i, i \in [1, 5]$. The final estimation of θ were obtained by multiplying M, A_i with the resize factor and by multiplying α_i, β_i with 2π.

We penalised the network by minimising the mean square error of reconstructed signal $S_i(t)$ and the input signal $X_i(t)$. In order to force each subnet to capture a fixed subwave, a regulariser called prior loss was added in the loss function. We randomly chose 100 median ECG samples and estimated the 21 hidden parameters (5 hidden waves) using the FMM R package [Rueda, 2019]. The mean μ_α and variance σ_α^2 of parameter α were computed and are used to constrain the subwaves' position. We let the Cascaded FMMnet estimate the T,R,S,P,Q subwaves sequentially by regularising the position α to be close to its corresponding pre-calculated distribution. The total loss function is:

$$loss = loss_{mse} + \gamma loss_{prior} \tag{5.3}$$

$$= \sum_{i=1}^{5} |X_i(t) - S_i(t)|^2 + \gamma \sum_{i=1}^{5} \frac{(\alpha_i - \mu_\alpha)^2}{\sigma_\alpha^2} \tag{5.4}$$

where γ controlled the balance of signal fitting and parameter distribution.

In order to ease the estimation for all leads, we assumed that the P,Q,R,S,T were sequentially positioned in all single beat ECG, which may be different with the conventional names of ECG wave peaks in some leads. For example, in Figure 5.7(a), the decomposed Q wave represented the conventional R wave for lead V1.

5.3 Experiments and Results

5.3.1 Datasets

Dataset	NORM	MI	AMI	IMI	LMI	Frequency(Hz)	Folds
PTB-XL [Wagner, 2020]	7185	2955	1937	1447	70	500	10
PTB [Bousseljot, 1995]	80	368	181	175	130	1000	5
CPSC(+Extra) [Liu, 2018]	922	370	–	–	–	500	5
CHU	22	37	–	–	–	Paper scan	-

Tab. 5.1.: The detailed information of the 4 datasets used in our study. AMI/IMI/LMI refer to anterior/inferior/lateral myocardial infarction.

We included three public ECG datasets and one private dataset in our study. All four datasets contained standard 12-lead ECG recordings and were (re)sampled to 500Hz. The PTB-XL dataset [Wagner, 2020] contained 21837 12-lead ECG recordings and covered

multiple ECG diagnostics and morphologies. Our Cascaded FMMnet was trained on this large dataset without considering specific pathology. For classification, as here we were concerned about detecting myocardial infarction (MI) patients from normal (NORM) cases, we included the detailed information of MI and NORM across different datasets in Table 5.1. The private dataset (CHU) was collected from Nice University Hospital in a scanned PDF format. Specific preprocessing was conducted to digitalize the ECG signals [Fortune, 2022].

5.3.2 Reconstruction

5.3.2.1 Experiment

The PTB-XL dataset was used to train the Cascaded FMMnet. As this was provided with 10 pre-defined folds, we randomly chose 8 folds as training set, 1 fold as validation set and the rest fold as test set. 100 random samples from training set were picked to compute the mean and variance used in regulariser (equation 4). The encoder network was implemented in Pytorch and trained with batch size of 192, learning rate of 0.0001. γ was initialised from 1 and it was updated to $\gamma = 0.1\gamma$ if $loss_{mse} \leq \gamma loss_{prior}$.

5.3.2.2 Results

We evaluated the reconstruction performance by R2 score and mean absolute error ($MAE = \frac{1}{600}\sum_{i=1}^{600}|X(t_i) - \hat{X}(t_i)|$). The proposed Cascaded FMMnet demonstrated very good reconstruction results and generalised well on three unseen datasets from different centers. First, the FMMnet was trained on 80% of the whole PTB-XL dataset and it demonstrated similar reconstruction result on the train/validation/test set of PTB-XL: they all presented a mean MAE of $0.016mV$ and a mean R2 score of 0.94. Second, all four datasets showed consistent reconstruction error on NORM/MI patients (Figure 5.3(a-b)) and lead-wisely (Figure 5.3(c-d)). Our Cascaded FMMnet took 0.09s/12 leads on a Dell laptop (Intel© Core™ i7-8650U CPU @ 1.90GHz × 4) while the original FMM optimisation [Rueda, 2019] takes more than 10s/1 lead on the same machine. We show an example of 12-lead ECG decomposition in Figure 5.7(a).

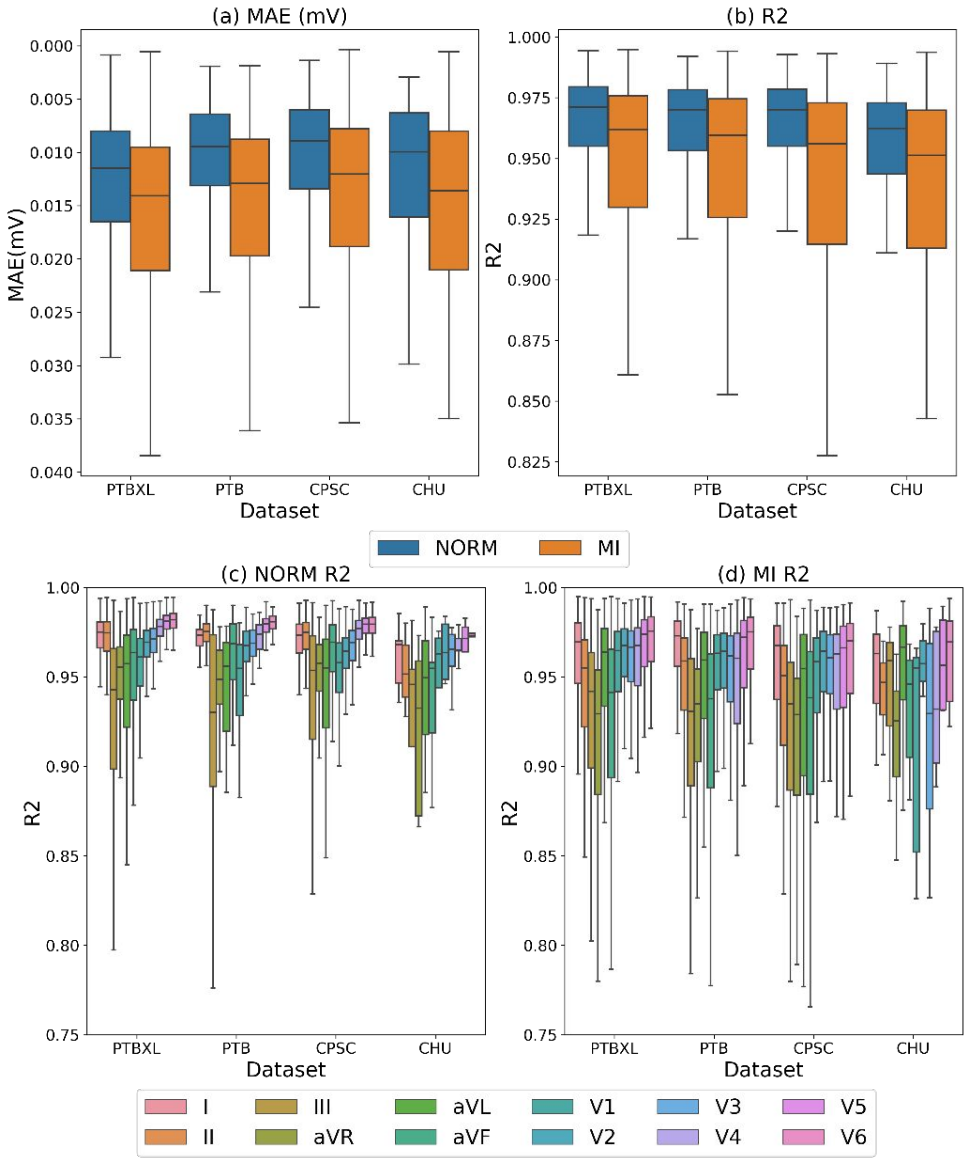


Fig. 5.3.: (a-b) The reconstruction metrics of different evaluation datasets on HC/MI separately. (c-d) Lead-wise R2 score of signal reconstruction of NORM patients and MI patients.

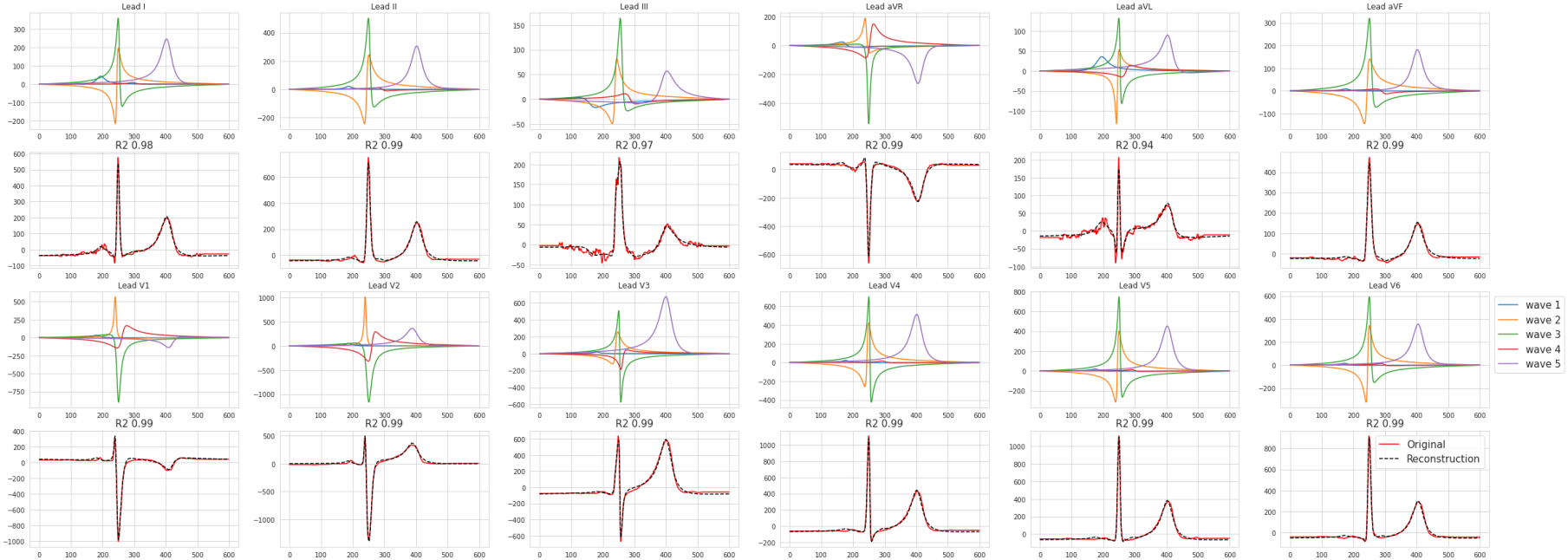


Fig. 5.4.: An example of signal decomposition for a 12-lead ECG and the corresponding reconstruction signals.

5.3.3 Classification

5.3.3.1 Experiment

For each 12-lead ECG sample (median beat), we obtained a 252-dimension feature vector from the Cascaded FMMnet. In addition, we included ST-segment voltage feature for each lead to form a 264-dimension vector. The ST-segment was identified as the short flat platform between peak S and peak T from each lead with the help of positional parameter α_T and α_S.

Explainable classification:

We explored two approaches to provide explainable classification. The first method began with a partial least squares regression that projects the 264-dimension vector to 3-dimension. A support vector machine (SVM) with linear kernel was applied then to find the best hyperplane that separated the MI/NORM patients. The additive nature of weighted features help to explain the classification results. The second method was to use SHAP value to explain a Logistic regression (LR) model for MI/NORM classification. We trained separate classifiers on PTB-XL, PTB and CPSC(+extra) datasets using 10-fold, 5-fold, 5-fold cross-validation, respectively. Two more specific classifiers for detecting AMI/IMI (vs. non AMI/IMI) were established on PTB-XL using 10-fold cross-validation.

Generalisable classification:

We tested the classifiers trained on PTB-XL to other three datasets: PTB, CPSC(+extra) and our private dataset (CHU), in order to evaluate the generalisation of our proposed classification pipeline.

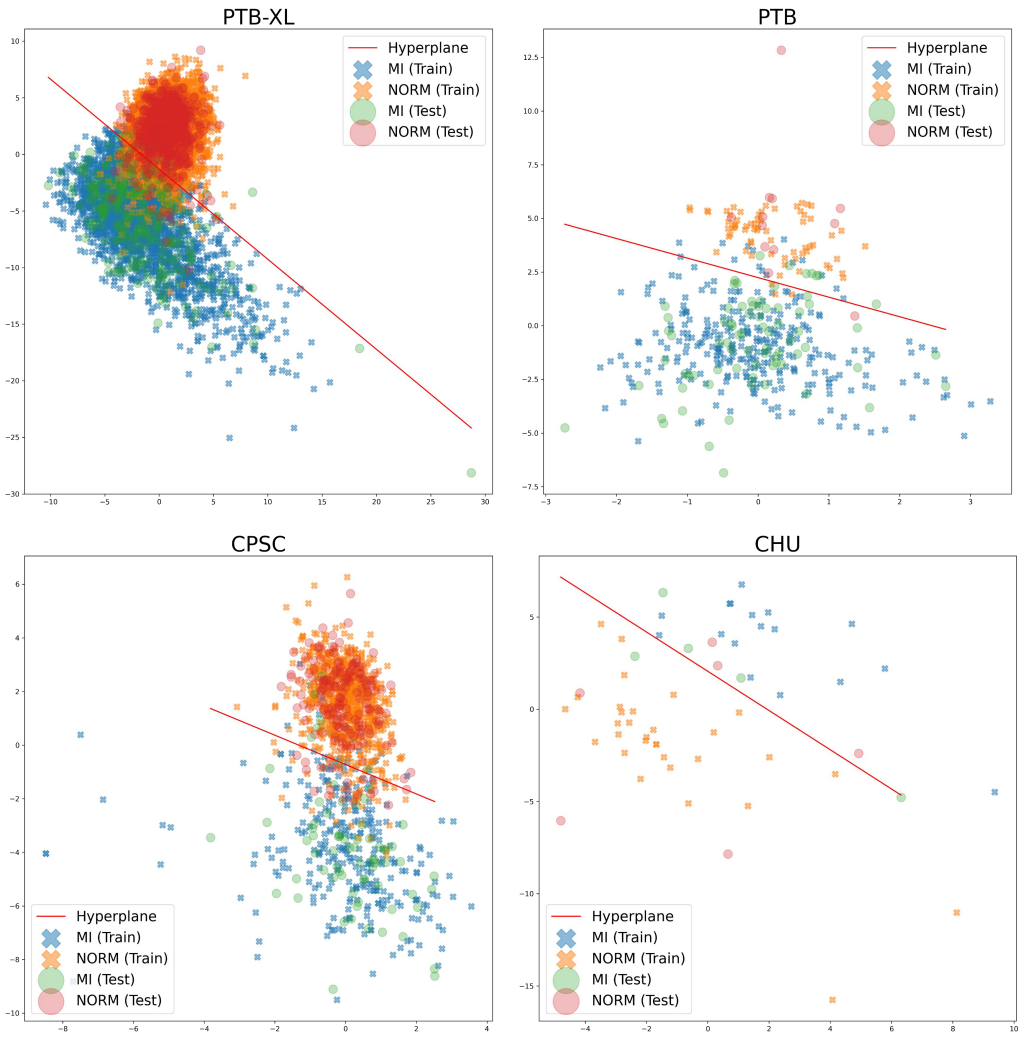


Fig. 5.5.: The classification boundary (hyperplane) of trained linear SVM classifiers and data points (3-dim) projected on one of the 2D plane orthogonal to the corresponding hyperplane.

Model	Dataset (evaluation)	Dataset (train)	Class	CV	AUROC	Accuracy	Sensitivity	Specificity
VGG [Fang, 2022]	PTB-XL	PTB-XL	MI	10-fold	1.00	0.97	0.96	0.98
Ours (LR)	PTB-XL	PTB-XL	MI	10-fold	0.99	0.96	0.93	0.96
Ours (SVM)	PTB-XL	PTB-XL	MI	10-fold	0.98	0.94	0.92	0.95
Ours (LR)	PTB-XL	PTB-XL	AMI	10-fold	0.98	0.93	0.92	0.93
Ours (LR)	PTB-XL	PTB-XL	IMI	10-fold	0.97	0.91	0.92	0.91
Ours (LR)	PTB-XL	PTB-XL	LMI	10-fold	0.91	0.85	0.87	0.85
VGG [Fang, 2022]	PTB	PTB	MI	5-fold	0.98	0.96	0.97	0.91
Ours (LR)	PTB	PTB	MI	5-fold	0.95	0.91	0.92	0.88
Ours (SVM)	PTB	PTB	MI	5-fold	0.95	0.90	0.92	0.81
Ours (LR)	PTB	PTB-XL	MI	–	0.95	0.84	0.83	0.99
Ours (SVM)	PTB	PTB-XL	MI	–	0.94	0.82	0.79	0.99
Ours (LR)	CPSC	CPSC	MI	5-fold	0.97	0.93	0.88	0.96
Ours (SVM)	CPSC	CPSC	MI	5-fold	0.97	0.92	0.88	0.93
Ours (LR)	CPSC	PTB-XL	MI	–	0.97	0.94	0.86	0.97
Ours (SVM)	CPSC	PTB-XL	MI	–	0.95	0.91	0.83	0.95
Ours (LR)	CHU	CHU	MI	5-fold	0.76	0.73	0.70	0.77
Ours (SVM)	CHU	CHU	MI	5-fold	0.77	0.68	0.59	0.73
Ours (LR)	CHU	PTB-XL	MI	–	0.77	0.76	0.92	0.50
Ours (SVM)	CHU	PTB-XL	MI	–	0.76	0.73	0.89	0.45

Tab. 5.2.: Classification results obtained on different datasets using Cascaded FMM net. *CV: cross validation.*

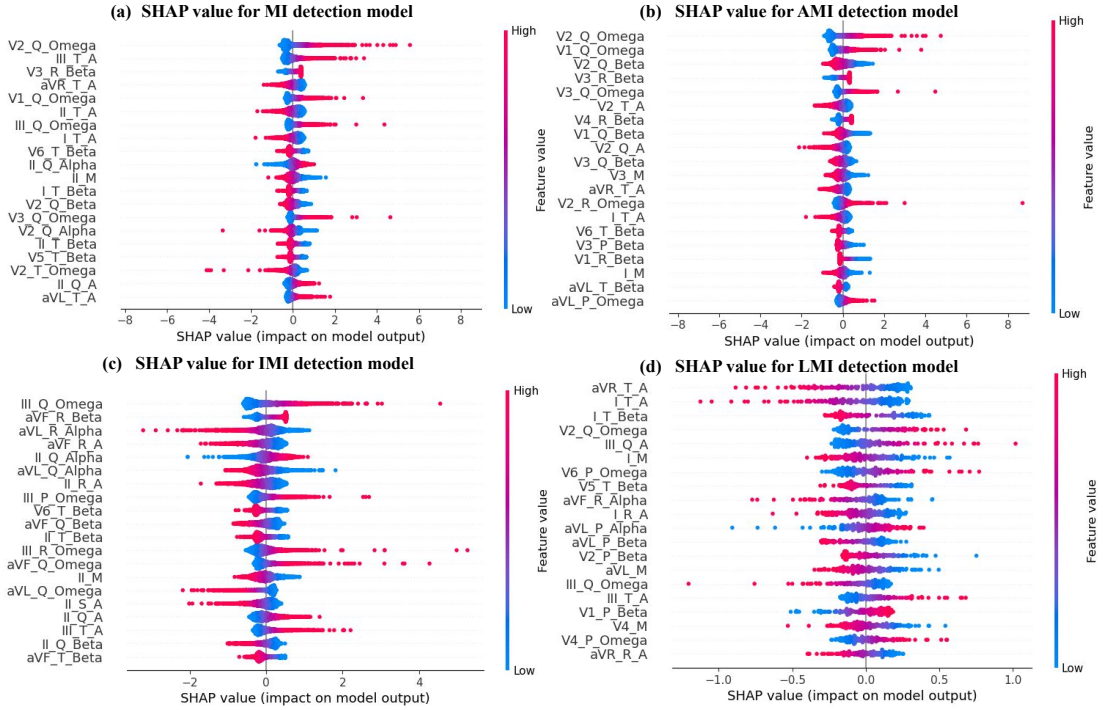
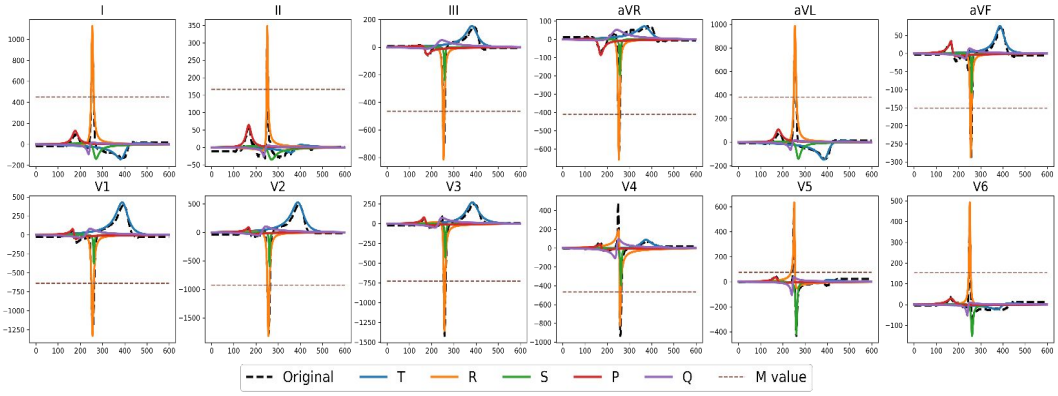


Fig. 5.6.: Explanations of feature importance for Myocardial Infarction (MI), Anterior Myocardial Infarction (AMI), Inferior Myocardial Infarction (IMI) and Lateral Myocardial Infarction (LMI) classification respectively using SHAP value on models trained on PTB-XL dataset. Higher shap value helps to augment the chances of detecting positive classes, in our cases, the MI/AMI/IMI/LMI classes.

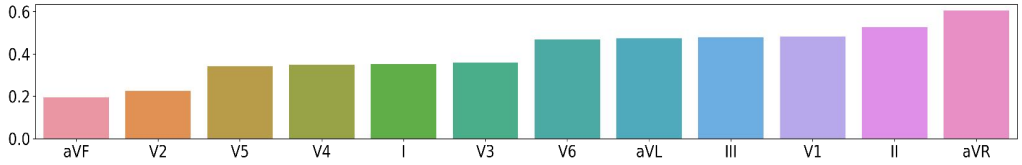
5.3.3.2 Results

We present the detailed classification evaluation in Table 5.2. Firstly, it's remarkable that using our classification pipeline, we were able to obtain satisfactory classification performance on different datasets (PTB-XL/PTB/CPSC) compared with other methods. As shown in Figure 5.5, a linear classifier (SVM with linear kernel) was already capable to obtain good separation of MI/NORM patients on both training and test data for PTB-XL, PTB and CPSC datasets, respectively. Using SHAP values, our models (Logistic regression classifiers) were capable to identify important infarction related features such as T wave amplitude change (T_A), T inversion (T_β) etc. They were also able to distinguish the influenced leads for infarction. For example, as shown in Figure 5.6, the model distinguished the V1,V2,V3 leads for AMI (Figure 5.6b), and the II,III,avF leads for IMI (Figure 5.6c). Since the MI label combined MI of different localisations, the important features for MI/NORM classification were spread widely across different leads (Figure 5.6a).

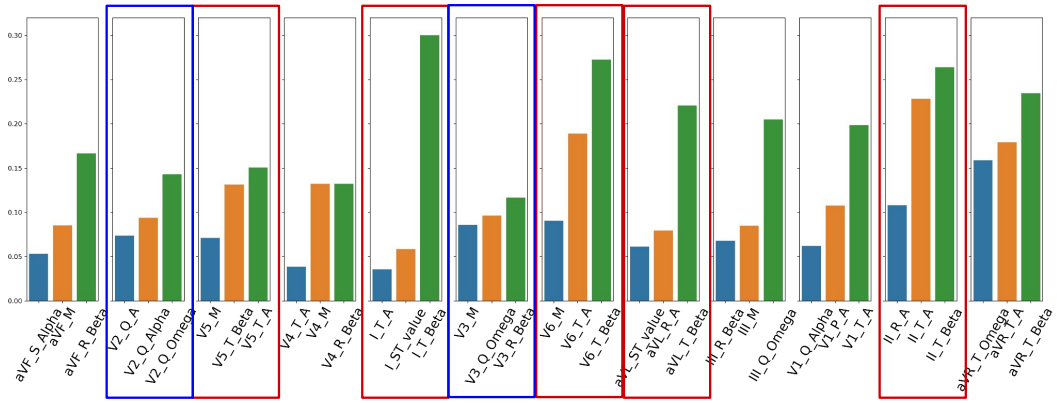
(a)12-lead median signal decomposition by Cascaded FMMnet



(b)Lead-wise weighted feature partition (linear SVM classifier)



(c)Lead-wise top-3 highly- weighted features (linear SVM classifier)



(d) SHAP value of features (Logistic regression classifier)

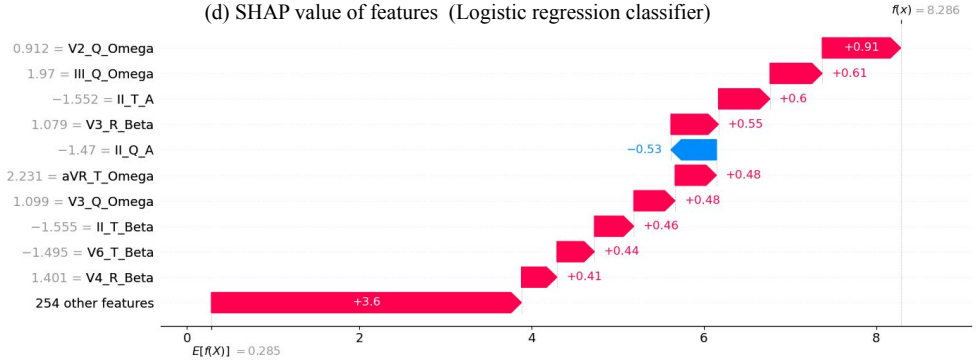


Fig. 5.7.: An example from PTB-XL test dataset classified as MI by the linear SVM classifier/Logistic regression classifier. According to the PTB-XL description, the patient was diagnosed with old anteroseptal infarction (tiny R waves present in V2,V3 leads), anterolateral ischaemia (inverted T waves and depressed ST-segments in I, avL, flat T waves in V5,V6 leads) and inferior infarction (flat T wave in II lead). (a) Decomposed 12-lead median ECG. (b)(c) Explainability provided by linear SVM model. Red rectangles mark the T-ST change related leads (acute infarction/ischaemia related) and blue rectangles mark the R wave change related leads (old infarction related and are represented here by our parameters prefixed by Q). (d) Explainability provided by Logistic regression model using SHAP value.

Second, models trained on PTB-XL generalised well on other datasets, without much drop of accuracy. However, we observed a drop in the specificity on our small private dataset. Since the signals were extracted from scanned ECG papers, the domain gap could be enlarged. In addition, the manner of annotation may be different. The label of NORM/MI in our private dataset is the final diagnosis decision, given by an experienced cardiologist after an overall examination of the patient, including echocardiography, ECG and sometimes coronary angiography. This may be different from the direct diagnosis based on the ECG alone from PTB-XL dataset.

We further include an example from PTB-XL test data that was correctly detected as MI (MI/NORM classifier) in Figure 5.7. The influenced ECG leads of this patient are I, II, avL, V2, V3, V5 and V6. First, our Cascaded FMMnet successfully identified the subwaves for most of the leads (Figure 5.7a). Second, we grouped the weighted features lead-wisely using the linear SVM model and observed that all leads helped us identify this patient as MI case (that is because they are all positive, as shown in Figure 5.7b). In addition, when examining the top-3 highest weighted features of every lead (Figure 5.7c), we found that most of the important features are in agreement with the clinical diagnostic (see red and blue rectangles in Figure 5.7c). We also observed a similar trend of important features explained by SHAP value from the Logistic regression classifier (Figure 5.7d).

5.4 Discussion and Conclusion

In this work, we proposed an automatic decomposition model, Cascaded FMMnet for ECG analysis which facilitated the downstream tasks, such as classification (in our case, the classification of normal vs. myocardial infarction patients). Overall, the Cascaded FMMnet was able to generalise reconstruction across datasets from different clinical centers and the estimated parameters could be used for MI classification with good explainability. It should be noticed that the Cascaded FMMnet was trained on PTB-XL dataset, which contains not only healthy samples but also samples with different pathologies (myocardial infarction, conduction disturbance, hypertrophy and so on). We believe that our novel network is capable to play an important role in other pathology classification. In the future, we will explore to improve the decomposition of different ECG leads with more specification. Our current Cascaded FMMnet assumes P,Q,R,S,T present in all leads; however, in some cases some subwaves could be absent. For example, in Figure 5.7a, we can observe the small S wave (in green) does not represent a meaningful peak and it's superposed onto the large R wave in lead avF, V1-4. This phenomenon is supported by the superior reconstruction accuracy on the leads close to lead II(I, II, V5, V6) than the other leads (Figure 5.3c-d). The original FMM paper [Rueda, 2021] also presented their results on lead II only. Some specific care for leads like V1-V4 should be considered.

5.5 Appendix

5.5.1 Paper ECG digitization

Our paper ECG digitization is based on a open-source software [Fortune, 2022] with adequate adaptation. A general pipeline of the ECG digitization is depicted in Figure 5.8. We only show one simple example here, for pipeline description and explanation. Since the CHU dataset contains several types of ECG paper formats (Figure 5.9), the automatic pipeline may fail and need manual interaction, such as signal masking and key point assignment.

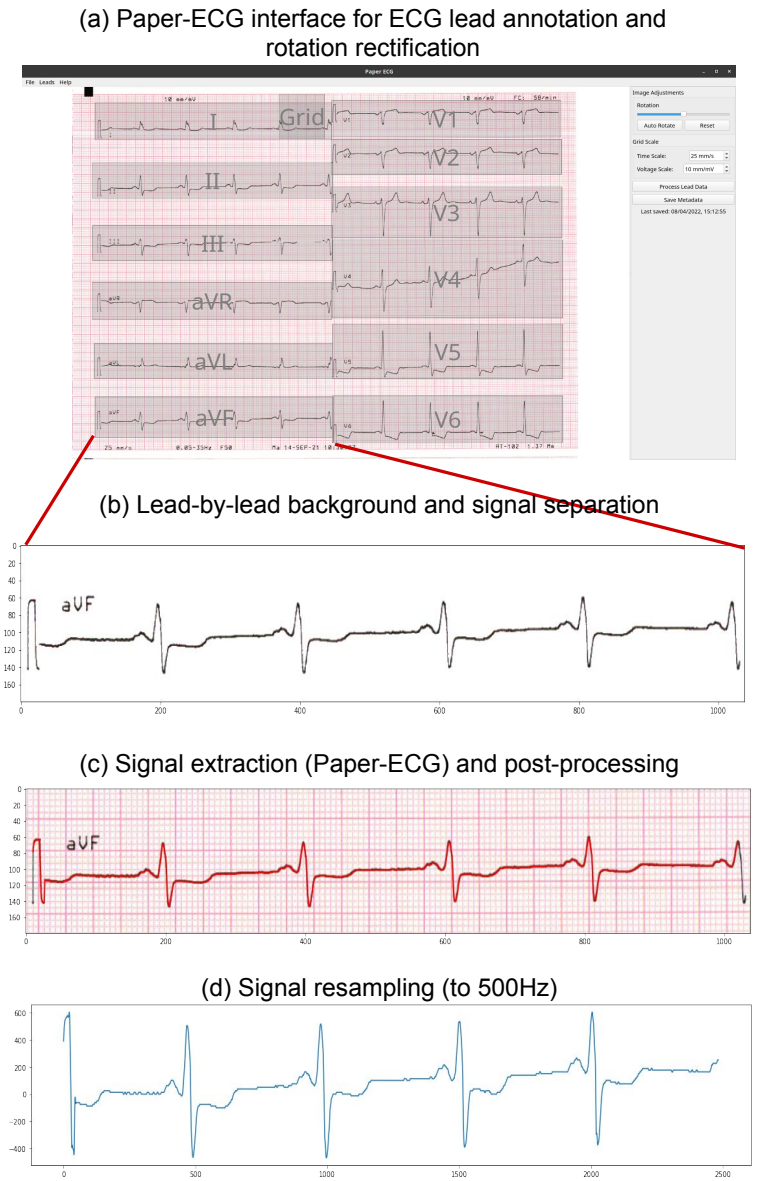


Fig. 5.8.: The general pipeline for ECG digitization.

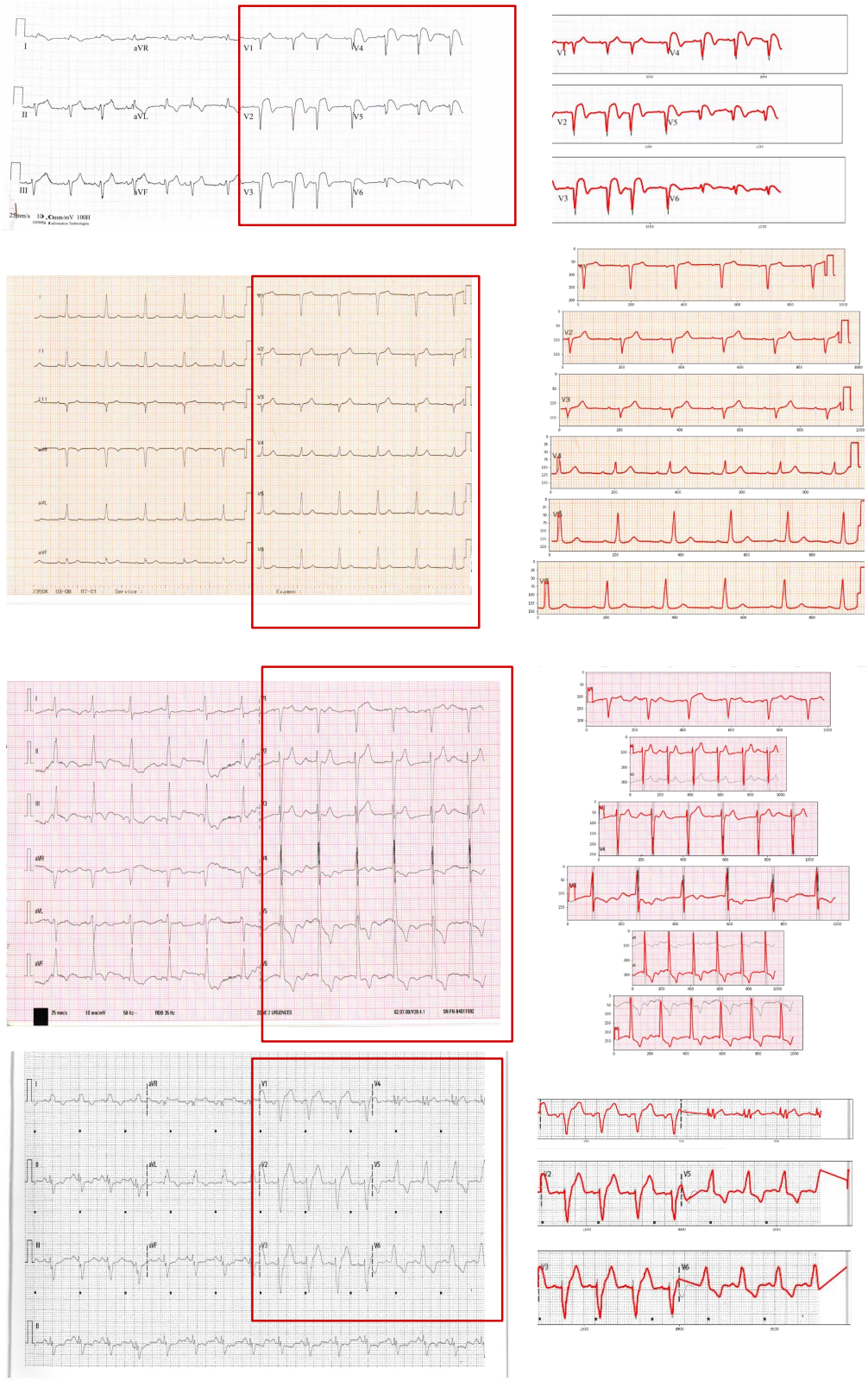


Fig. 5.9.: Different formats of ECG papers and some signal extraction examples.

6 Multi-modal detection of myocardial infarction: uncertainty-based fusion using echocardiography and eletrocardiogram

Contents

Abstract Medical devices used in diagnoses capture only one aspect of cardiac function. For example, 2D B-mode echocardiography reveals the anatomy and mechanical change of the heart, while electrocardiogram (ECG) captures electrical activities of the heart from different observation positions. As a routine for cardiac disease diagnosis, those examinations are not conducted synchronously, but sequentially and provide complementary proof for final decision. While most of the current AI-based models focus on a single modality, the combination of multi-modal information in AI research for healthcare has started to gain popularity and to become a hot topic. However, the scarcity of public multi-modal data for cardiac disease diagnosis make this task rather difficult. In this study, we adapted an uncertainty-based deep learning framework using unpaired single modality data, for better diagnosis of MI using echocardiography and eletrocardiogram data jointly. Specifically, we trained two single modality models using public datasets and evaluated the multi-modal classification performance on a small paired dataset collected from a local hospital. Our

experiment demonstrated that uncertainty-based multi-modal decision fusion outperforms popular fusion strategies and single modality models. Thus, uncertainty-based models could be a sustainable solution for unpaired multi-modal training.

The contribution of this chapter are listed as follows:

- We proposed to adapt a multi-modal decision fusion strategy with uncertainty using unpaired public single modality datasets (Section 6.2).
- We showed that uncertainty based multi-modal fusion improves the accuracy of detecting MI patients by 7% comparing to the single modality model, while the performance of other traditional fusion models are limited by the best-performing modality (Section 6.3).

6.1 Introduction

In real-world, clinicians combine information from different examinations and measurements to make clinical decisions. Most current AI research for healthcare simply consider one single modality, which does not profit from the complex and heterogeneous information that one can observe from patients using different imaging modalities, sensoring devices, biochemical tests etc. Multi-modal machine learning, which seeks to model the interactions between different modalties, brings opportunities for improving the prevention, diagnosis and therapy in AI-enabled healthcare [Acosta, 2022].

One challenge in biomedical multi-modal learning is to determine how to fuse information from different medical modalities for downstream tasks. Depending on when the fusion occurs, one can distinguish: early fusion and late fusion, respectively. Early fusion combines the raw modality or extracted features at the input level according to certain fusion approaches, such as concatenation, multiplicative interaction [Jayakumar, 2020], polynomial fusion [Kefalas, 2020], tensor fusion [Zadeh, 2017; Hou, 2019] etc. Late fusion aggregates the prediction outputs of different modalities at the decision level (e.g. using majority voting, weighted voting etc.) to generate a final decision. Early fusion usually demands paired multi-modality data to explore detailed interaction strategies, while late fusion only need single modality outputs, thus being less demanding for paired data.

In our study we were limited by the number of paired multi-modal data (only 56 patients). Therefore, we chose to do late fusion, which can leverage the usage of large single modality datasets. In addition, we adapted the uncertainty based fusion [Han,

2021], which generated the final decision based on the most trusted modality according to the single modality uncertainty.

6.2 Method

6.2.1 Single modality evidential deep learning

Uncertainty and the Theory of Evidence

Evidential deep learning (EDL) quantifies the class probabilities and overall uncertainty in a unified theoretical framework [Sensoy, 2018]. Based on the Dempster-Shafer Theory of Evidence [Yager, 2008], evidential deep learning adapts the idea from Sujective Logic (SL). SL associates the belief of possible class label assignments with the parameters of Dirichlet Distribution [Jsang, 2018], including the belief that the truth label is equally likely (i.e., "*I don not know*" for uncertainty quantification). Specifically, let us consider a K classification problem, where SL assigns K belief masses b_k to each class and an overall uncertainty mass u to the whole frame using the evidence $e_k \geq 0$. Evidence e_k represents a measure of the amount of support for kth class category collected from data input. The belief and uncertainty mass values are computed as follows:

$$b_k = \frac{e_k}{S} \text{ and } u = \frac{K}{S}, S = \sum_{i=1}^{K}(e_i + 1) \tag{6.1}$$

The sum of the $K+1$ mass values is one, i.e. $u + \sum_{k=1}^{K} b_k = 1$.

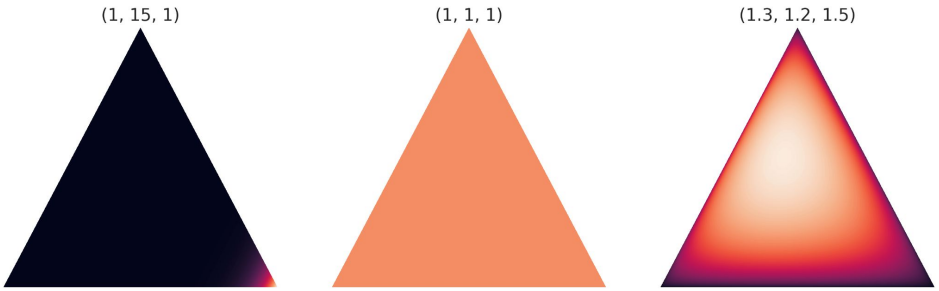


Fig. 6.1.: Examples of Dirichlet distribution (K=3). The values of α are listed above each figure.

The Dirichlet distribution is parameterized by K parameters $\alpha = [\alpha_1, ..., \alpha_K]$. Its probability density function (pdf) is given by

$$D(\mathbf{p}|\alpha) = \begin{cases} \frac{1}{B(\alpha)} \prod_{i=1}^{K} p_i^{\alpha_i - 1} & \text{for } \mathbf{p} \in \mathcal{S}_K, \\ 0 & \text{otherwise,} \end{cases} \tag{6.2}$$

where $\mathcal{S}_K$ represents the K-dimensional unit simplex $\mathcal{S}_K = \{\mathbf{p}|\sum_{i=1}^{K} p_i = 1 \text{ and } 0 \leq p_1, ..., p_k \leq 1\}$, and $B(\alpha)$ is the K-dimensional multinomial beta function. According to SL, the class belief assignment is connected to a Dirichlet distribution with $\alpha_k = e_k + 1$. The relationship between evidence, belief, uncertainty and Dirichlet parameters is summarized below :

$$b_k = \frac{e_k}{S} = \frac{\alpha_k - 1}{S} \text{ and } u = \frac{K}{S}, S = \sum_{i=1}^{K}(e_i + 1) = \sum_{i=1}^{K} \alpha_i. \tag{6.3}$$

The above relationship reveals that the higher the evidence e_k for k_{th} class is observed, the greater the class belief b_k and the corresponding Dirichlet parameter α_k will be (Figure 6.1 (a)). Similarly, when the total evidence observed from input data is small, i.e. $\sum e_k$ closer to 0 and $\alpha_k, k = 1, ..K$ closer to 1, the uncertainty of prediction becomes higher (Figure 6.1 (b-c)).

Learning to form opinions

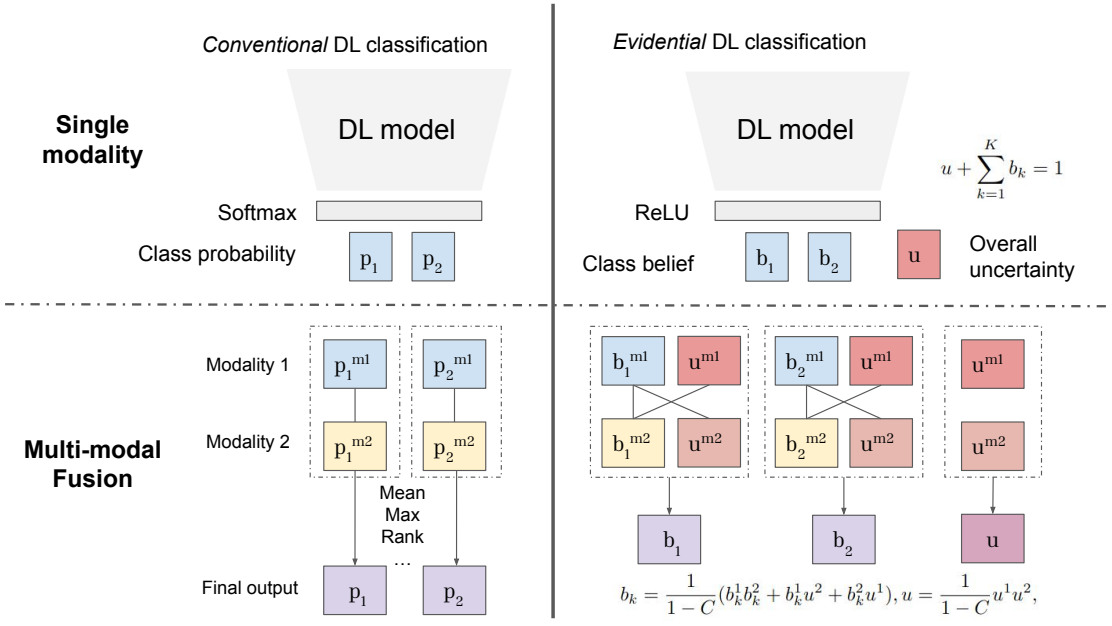


Fig. 6.2.: Comparison of conventional fusion strategies and uncertainty based fusion.

Evidential deep learning replaces the last *softmax* activation in neural network classifiers with non-negative activation, such as $ReLU$. The output of final activation layer is taken as the evidence vector. It forms class belief masses and constitutes the parameters for the estimated Dirichlet distribution (illustrated in the upper right part of Figure 6.2).

We assume that y_i is a one-hot vector of ground truth classification label for input data x_i. The cross-entropy loss is usually used in conventional neural network classifiers:

$$\mathcal{L}^{CE} = -\sum_{i=1}^{N}\sum_{j=1}^{K} y_{ij}\log(p_{ij}), \tag{6.4}$$

where p_{ij} is the predicted probability for sample x_i belonging to class j. Under the theory of evidence and Dirichlet distribution assumption, we can compute the Bayes risk of cross-entropy loss function as

$$\mathcal{L}_i^{UC} = \int [\sum_{j=1}^{K} -y_{ij}\log(p_{ij})]\frac{1}{B(\alpha_i)}\prod_{j=1}^{K} p_{ij}^{\alpha_{ij}-1} d\mathbf{p}_i = \sum_{j=1}^{K} y_{ij}(\psi(S_i) - \psi(\alpha_{ij})), \tag{6.5}$$

where $\psi(\cdot)$ represents the *digamma* function.

The minimization of the above loss does not guarantee that less evidence will be generated when the model predicts incorrect labels. To guide the network learn zero total evidence for uncertain samples, a regularization term is introduced. This term deploys a Kullback-Leibler divergence term to penalize the predictive Dirichlet distribution to be close to $D(\mathbf{p}|\mathbf{1})$.

$$KL[D(\mathbf{p}_i|\tilde{\alpha}_i)||D(\mathbf{p}_i|\mathbf{1})] = \log(\frac{\Gamma(\sum_{k=1}^{K}\tilde{\alpha}_{ik})}{\Gamma(K)\prod_{k=1}^{K}\Gamma(\tilde{\alpha}_{ik})}) + \sum_{k=1}^{K}(\tilde{\alpha}_{ik}-1)[\psi(\tilde{\alpha}_{ik}) - \psi(\sum_{k=1}^{K}\tilde{\alpha}_{ij})], \tag{6.6}$$

where $\Gamma(\cdot)$ represents the *gamma* function and $\mathbf{1}$ refers to a K-dim vector of all ones.

Thus, the final loss function for evidential deep learning neural networks reads:

$$\mathcal{L} = \sum_{i=1}^{N}\mathcal{L}_i^{UC} + \lambda_t\sum_{i=1}^{N} KL[D(\mathbf{p}_i|\tilde{\alpha}_i)||D(\mathbf{p}_i|\mathbf{1})], \tag{6.7}$$

where $\lambda_t = \min(1, t/10) \in [0, 1]$ is a balancing coefficient for regularization.

6.2.2 Multi-modal fusion with uncertainty

We adapt the multi-modal combination of belief and uncertainty masses presented in [Han, 2021] in an off-line manner (as illustrated in the lower right part of Figure 6.2).

Considering two independent sets of belief mass values $\mathcal{M}^1 = \{\{b_k^1\}_{k=1}^K, u^1\}$ and $\mathcal{M}^2 = \{\{b_k^2\}_{k=1}^K, u^2\}$ that are outputs of two independent single modality neural networks, the fusion of belief and uncertainty mass values is designed as follows:

$$b_k = \frac{1}{1-C}(b_k^1 b_k^2 + b_k^1 u^2 + b_k^2 u^1), u = \frac{1}{1-C} u^1 u^2, \tag{6.8}$$

where $C = \sum_{i \neq j} b_i^1 b_j^2$ measures the amount of conflict and $\frac{1}{1-C}$ is applied for normalisation of belief and uncertainty mass values.

6.3 Experiments and results

6.3.1 Datasets

In this study, we focused on MI detection using echocardiography and eletrocardiogram data. First, two independant datasets of ECHO and ECG respectively were involved:

- HMC-QU dataset: contains 130 long-axis 2-chamber view sequences (68 with myocardial infarction segments) and 162 long-axis 4-chamber view sequences (93 with MI segments).

- PTB-XL dataset: contains 12-lead ECG data (with 7185 samples of healthy controls and 2955 samples with 100%- certain MI).

In addition, a small number of paired ECHO and ECG data were collected retrospectively from CHU hospital. This dataset contains data from 56 patients, with 56 paired data of ECG and 4-chamber view ECHO, along with 50 paired data of ECG and 2-chamber view ECHO. We trained the models using HMC-QU and PTB-XL for single modalities separately, and evaluated the multi-modal classification performance on CHU dataset.

Dataset	Modality	MI	non-MI	Total
HMC-QU	ECHO 2ch	68	62	130
HMC-QU	ECHO 4ch	93	69	162
PTB-XL	ECG 12-lead	2955	7185	10140
CHU	ECHO(2ch) + ECG	33	17	50
CHU	ECHO(4ch) + ECG	36	20	56

Tab. 6.1.: Dataset statistics. *2ch: 2 chamber view, 4ch: 4 chamber view.*

6.3.2 Experiments

We first obtained input features for ECHO and ECG using the models described in Chapter 3 and Chapter 5 respectively. For ECHO data, we constructed a 40-dimension vector from temporal movement of the myocardium. This vector was composed of mean and standard deviation of each key point movement along the x- and y- axis, respectively, that were estimated from motion estimation deep learning model. For ECG data, we constituted a $21 * 12$ dimension vector from ECG decomposition model.

For all experiments, we used a 3-layer fully connected network (FCN) for single modality classification. We trained single modality models with 10-fold cross validation for ECG data and 5-fold cross validation for ECHO data. We then compared the multi-modal classification with uncertainty to several baseline fusion strategies.

We assumed that the output of FCN after Sigmoid function was p_c^k and the prediction class was $\bar{y^k}$, where c referred to class and k referred to modality, while the MI class was set to label 1. The following fusion strategies were included in our study :

- Max fusion: $p_c = \max\{p_c^k, k = 1, ..., K\}$, $\bar{y} = \arg\max_c p_c$;
- Mean fusion: $p_c = \text{mean}\{p_c^k, k = 1, ..., K\}$, $\bar{y} = \arg\max_c p_c$;
- Rank fusion: $\bar{y} = (\sum_k \bar{y^k}) \geq 1$;
- Multiply fusion: $p_c = \prod_k p_c^k$, $\bar{y} = \arg\max_c p_c$.

Finally, the baseline single modality model was trained using cross-entropy loss (Equation 6.4) and the uncertainty model with Equation 6.7. The multi-modal fusion with uncertainty was performed according to Equation 6.8.

6.3.3 Implementation

The uncertainty model for ECHO and ECG were trained using the following hyper-parameters:

- ECG: learning rate 0.01, batch size 512, number of epochs 1000, $\lambda_t = 10$;
- ECHO: learning rate 0.0001, batch size 8, number of epochs 500, $\lambda_t = 50$.

We chose the model with best validation loss during training.

6.3.4 Results

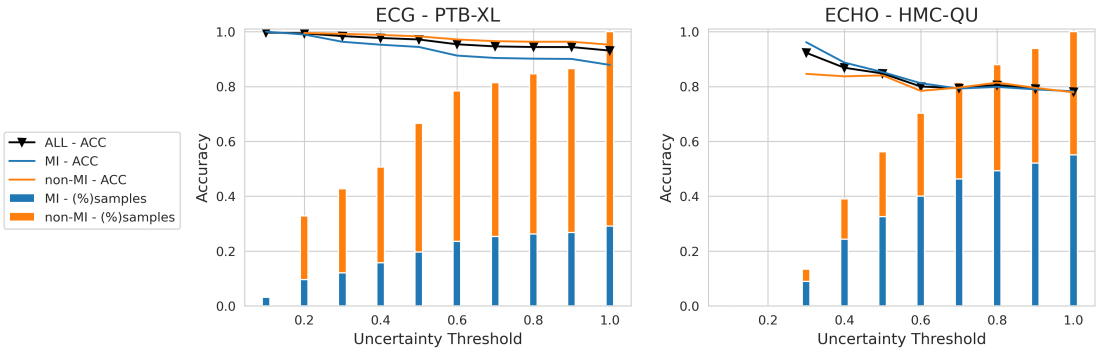


Fig. 6.3.: The change of prediction accuracy with respect to uncertainty threshold on PTB-XL ECG dataset and HMC-QU ECHO dataset (2CH/4CH mixed). Bar plots represent the percentage of samples kept under varying uncertainty thresholds.

First, we present the cross validation results on HMC-QU and PTB-XL dataset in Table 6.3 and Table 6.2. Although evidential deep learning (EDL) model demonstrated reduced performance compared with model trained using standard cross-entropy loss, its performance was comparable when using mixed 2-chamber and 4-chamber (2CH/4CH mixed) view together (292 samples in total). We made a similar observation on ECG classification using uncertainty-based loss (Table 6.3). Figure 6.3 shows how the test accuracy changes when EDL only keeps predictions under varying uncertainty thresholds. Notably, on both datasets, the accuracy decreased as the uncertainty threshold increased.

Method	View	Accuracy	Sensitivity	Specificity
[Degerli, 2024]	2CH	0.75	0.72	0.77
Ours (w/o UC)	2CH	**0.78**	**0.74**	0.82
Ours (w UC)	2CH	0.72	0.59	**0.85**
[Degerli, 2024]	4CH	**0.86**	**0.84**	**0.85**
Ours (w/o UC)	4CH	0.81	0.82	0.80
Ours (w UC)	4CH	0.82	0.83	0.81
Ours (w/o UC)	2CH + 4CH (mixed)	0.78	0.78	0.79
Ours (w UC)	2CH + 4CH (mixed)	0.78	0.78	0.78

Tab. 6.2.: ECHO classification: 5-fold CV results on HMC-QU dataset. *w/o UC: without uncertainty, w UC: with uncertainty.*

Method	Lead	Accuracy	Sensitivity	Specificity
[Yang, 2022]	12-lead	0.96	0.93	0.96
Ours (w/o UC)	12-lead	0.95	0.89	0.97
Ours (w UC)	12-lead	0.93	0.88	0.95

Tab. 6.3.: ECG classification: 10-fold CV results on PTB-XL dataset. *w/o UC: without uncertainty, w UC: with uncertainty.*

Method	Modality	Accuracy	Sensitivity	Specificity
Ours (w/o UC)	ECG	0.69	0.84	0.43
Ours (w/o UC)	ECHO	**0.75**	0.75	**0.76**
Max Fusion	ECG + ECHO	0.73	0.81	0.57
Mean fusion	ECG + ECHO	**0.75**	0.86	0.54
Rank fusion	ECG + ECHO	0.73	**0.96**	0.30
Multiply fusion	ECG + ECHO	0.68	0.87	0.32
Ours (w UC)	ECG	0.72	**0.86**	0.48
Ours (w UC)	ECHO	0.71	0.68	**0.76**
Uncertain fusion	ECG + ECHO	**0.79**	0.83	0.73

Tab. 6.4.: Evaluation on CHU dataset (with 2-chamber view and 4-chamber view mixed together, in total 106 paired samples). *w/o UC: without uncertainty, w UC: with uncertainty.*

We present the off-line multi-modal fusion evaluation on CHU dataset in Table 6.4. The performance of multi-modal conventional fusion (upper part) was limited by the best performing modality, in our case, the ECHO modality. The mean fusion strategy outperformed the other conventional methods, with a slight improvement in sensitivity but significant reduction in specificity due to the erroneous output of ECG prediction. In the lower part of Table 6.4, we can observe that uncertainty-based fusion improved largely the prediction accuracy compared to single modalities. In addition, this novel approach well combined the advantages of each modality: higher sensitivity than single ECHO output and higher specificity than single ECG output, with only a slight decrease compared with the best value of single modality outputs. Uncertainty-based fusion generated multi-modal prediction according to the most trustworthy modality, therefore improving the final prediction of diagnosis.

6.4 Conclusion

In this study, we explored different multi-modal late fusion strategies, demonstrating that uncertainty-based fusion outperformed other conventional fusion methods by improving the single modality accuracy by 7%. The uncertainty-based fusion was based on single modality evidential deep learning and examined the uncertainty of each modality prediction to balance the most trustworthy prediction for final prediction. In addition, the integration of uncertainty did not need any sampling steps and was easy to implement with deep learning methods. The off-line fusion setting made the most use of large public single modality dataset and kept the valuable paired multi-modal data for evaluation.

7 Conclusion

In this thesis, we proposed a framework for multi-modal cardiac function analysis using portable medical modalities: echocardiography (ECHO) and electrocardiogram (ECG). Specifically, we described in detail our solutions of analysing single modalities and multi-modal fusion with a focus on robustness, generalization, explainability and uncertainty. In the following, we summarize the main contributions of this thesis and discuss limitations of current work and potential directions of future research.

7.1 Main Contributions

Single-frame segmentation of echocardiography with shape prior
In Chapter 2, we introduced shape information from three-levels for 2D echocardiography segmentation and conducted thorough experiments across different datasets. In particular, the strategy from pixel level introduced shape prior in terms of contour loss and outperformed other strategies. With appropriate augmentation techniques, the trained model generalized well on segmentation prediction and left ventricle ejection fraction (LVEF) estimation on different test datasets. LVEF is an important clinical cardiac feature that cardiologists usually measure during examination. AI-enabled robust estimation of LVEF has significant potential to serve as a reference for experts and non-experts.

Poly-affine motion estimation model for echocardiography with motion prior
In Chapter 3, we proposed a weakly-supervised motion estimation framework for 2D echocardiography sequences. Instead of estimating the dense motion field, we parameterized the movement of the left ventricle myocardium through a poly-affine motion framework, using only 60 parameters. Our design of motion prior helped significantly the model to predict key points around the myocardium and the corresponding local transformations. The compact modelling of the LV motion not only improved the registration accuracy but also the motion field regularity in terms of Jacobian determinant. Robust estimation of left ventricle motion help reveal global longitudinal strain (GLS), another sensitive feature for cardiac dysfunction detection. In addition, the poly-affine motion model was capable to transfer motion pattern from one sample to another while keeping the original appearance of ultrasound images. Effective motion transfer is able to enlarge the size of echocardiography dataset, for example for self-supervised learning.

Explainable echocardiography analysis pipeline
In Chapter 4, we introduced a pipeline for echocardiography analysis based on the method proposed in Chapter 2 and Chapter 3. Specifically, we extracted left ventricle ejection fraction (LVEF) and normalized mitral annular plane systolic excursion (MAPSEn) from segmentation prediction, and the global longitudinal strain (GLS) and global radial strain (GRS) from motion estimation prediction. On two test datasets, the extracted features showed significant group difference between healthy controls and myocardial infarction patients. The proposed pipeline demonstrated great potential for explainable cardiac function analysis using portable ultrasound devices.

Unsupervised decomposition of electrocardiogram for explainable analysis
In Chapter 5, we proposed an unsupervised decomposition model for electrocardiogram analysis. The model was able to take any single-beat lead-agnostic ECG signal as input and predict 21 explainable parameters that controlled P, Q, R, S, T subwaves. The ensemble parameters of 12-lead ECG demonstrated great classification performance when detecting myocardial infarction patients from healthy controls across various datasets. Besides, the visualization of ECG decomposition provided an explicit way to examine the analysis result qualitatively. The proposed model exhibits significant promise for application in the classification of other diseases due to its robust and adaptable framework.

Uncertainty guided multi-modal decision fusion
In Chapter 6, we adapted an uncertainty based classification framework for off-line multi-modal decision fusion and evaluated its performance on myocardial infarction detection using paired ECG and ECHO data. The proposed method generated final decision leveraging the most trustworthy modality and improved the multi-modal performance by 7% compared with single modality accuracy, while the performance of other conventional fusion strategies was limited by the best performing modality. Our findings illustrated that cardiac dysfunction detection can yield more reliable results compared to a single modality approach, thereby promoting the adoption of a multi-modal setup for real-world portable cardiac diagnosis.

7.2 Future research

In this thesis, we presented original contributions to the field of medical imaging and signal analysis, with a focus on employing 2D echocardiography, electrocardiogram and multi-modal classification for better detection of disease (e.g. MI). Although we have achieved promising results, there are several improvement needed which suggest several directions for future work.

Echocardiography analysis

We acknowledge that the segmentation model was applied on separate frames of echocardiography sequences, while temporal information of cardiac sequences were neglected by current model. Enforcing temporal consistency is a good way to reduce segmentation outliers and to engage unlabelled data for training [Wei, 2020; Painchaud, 2022]. In addition, for the motion tracking model, we imposed a strong regional weight for each key point without considering the various thickness of the LV wall across patients and across regions. Thus, it would be interesting to further combine temporal motion tracking model with temporal segmentation model, letting the predicted mask of the myocardium guide the region of effect for each local key point. Furthermore, motion tracking of LV regularizes vice versa the segmentation of cardiac structures for unlabelled temporal frames, accordingly. In addition, it's crucial to estimate the uncertainty of predicted segmentation and motion transformation for clinical applications. Methods proposed in [Grzech, 2022; Judge, 2023] are of great interest for the further improvement of our echocardiography analysis pipeline.

Eletrocardiogram analysis

Our proposed ECG decomposition method has two limitations. First of all, it can only analyze R-peak aligned single-beat signal. In the cases where it is difficult to detect the R-peak in the original ECG signal, the utility of the current model will be limited. Second, although the proposed model can process lead-agnostic single-beat ECG signal, the inter-lead relationship is neglected. We vise to analyze continuous time series instead of single-beat signal in the future. One possible approach is to learn contrastive representation of ECG by masking random leads as proposed in [Kiyasseh, 2021; Oh, 2022]. Specifically, we could integrate the information of physical lead position into the modelling process as in [Chen, 2021], who proposed an ECG synthesis model that takes into consideration the physical lead position encoding. In addition, we could further model the periodic characteristic of ECG time series, for instance, using the dynamic latent trajectory representation proposed in [Laumer, 2020; Ryser, 2022].

Multi-modal learning for portable analysis

In Chapter 6, we utilised the 12-lead ECG with 2D ECHO data features. While in portable examination, the 12-lead ECG is not always available. It would be more realistic to combine ECHO data with single-lead ECG data. For example, Lead I or Lead II could be easily measured using wearable ECG devices. However, such ambitious objective may be very challenging. That is because single-lead ECG is a useful diagnostic tool for rhythm-based and conduction disorder-related diseases; however, it is not very informative for MI detection [Witvliet, 2021]. Furthermore, it is crucial to model the interaction of full 12 leads and condition the representation on a desired single lead. For example, we could project each lead into a joint latent space using multi-channel variational autoencoder [Antelmi, 2019], or generate the 12-lead data using single-lead-based ECG synthesis [Lee, 2019; Golany, 2020; Beco, 2022]. Moreover, it demands effective modelling

between ECHO time series and single-lead ECG signal. Should synchronized ECHO and ECG is obtained, it would be entrancing to examine the temporal correlation between time series from ECHO and ECG. Another possible direction is to explore methods from multi-variate time series modelling, such as LSTM network [Karim, 2019], autoencoders [Audibert, 2020], transformer-based models [Zerveas, 2021] and graph neural networks [Duan, 2022].

Appendix

A

Unsupervised Echocardiography Registration through Patch-based MLPs and Transformers

Abstract Image registration is an essential but challenging task in medical image computing, especially for echocardiography, where the anatomical structures are relatively noisy compared to other imaging modalities. Traditional (non-learning) registration approaches rely on the iterative optimization of a similarity metric which is usually costly in time complexity. In recent years, convolutional neural network (CNN) based image registration methods have shown good effectiveness. In the meantime, recent studies show that the attention-based model (e.g., Transformer) can bring superior performance in pattern recognition tasks. In contrast, whether the superior performance of the Transformer comes from the long-winded architecture or is attributed to the use of patches for dividing the inputs is unclear yet. This work introduces three patch-based frameworks for image registration using MLPs and transformers. We provide experiments on 2D-echocardiography registration to answer the former question partially and provide a benchmark solution. Our results on a large public 2D-echocardiography dataset show that the patch-based MLP/Transformer model can be effectively used for unsupervised echocardiography registration. They demonstrate comparable and even better registration performance than a popular CNN registration model. In particular, patch-based models better preserve volume changes in terms of Jacobian determinants, thus generating robust registration fields with less unrealistic deformation. Our results demonstrate that patch-based learning methods, whether with attention or not, can perform high-performance unsupervised registration tasks with adequate time and space complexity. This chapter was a joint work with Zihao Wang and was published in the International Workshop on Statistical Atlases and Computational Models of the Heart (STACOM) [Wang, 2022].

A.1 Introduction

Image registration is essential for clinical usage; for example, the registration of cardiac images between end-diastole and end-systole is meaningful in myocardium deformation

analysis. Non-rigid echocardiography image registration is one of the most challenging image registration tasks, as finding the deformation field between noisy images is a highly nonlinear problem in the absence of ground truth deformation. Specifically, various image registration problems require the mapping between moving and fixed images to be folding free [Cao, 2005; Vercauteren, 2008; Arsigny, 2006]. Traditional non-learning approaches rely on the optimizing similarity metrics to measure the matching quality between image pairs [Oliveira, 2014; Davatzikos, 1997; Vercauteren, 2007]. With the rapid promotion of deep learning, various frameworks of convolutional neural networks (CNN) have been introduced in image registration and have shown impressive performance in many research works.

We consider a 2D non-rigid machine learning-based image registration task in this work. With two given images: $I_{fix}^N, N \in \mathbb{R}$ and $I_{mov}^N, N \in \mathbb{R}$, we want to learn a model $\mathcal{T}_\omega(I_{mov}, I_{fix}) \rightarrow \phi(\theta)$ that generates a constrained transformation $\phi(\theta)$ based on a similarity measurement $\mathcal{M}$ to warp the moving image by minimising the loss function:

$$L = \arg\min_{\theta} \mathcal{M}(I_{fix}, I_{mov} \circ \phi(\theta)) + \lambda \mathcal{C}(\phi(\theta)) \tag{A.1}$$

where the transformation ϕ is parameterised by the parameter θ and constrained by a regularisation term $C(\phi(\theta))$ to ensure ϕ to be a spatially smooth transformation. However, iterative optimization of Eq. A.1 is very time-consuming, whereas a well-trained CNN does not need any iterative minimization of the loss function at test time. This advantage drives researchers' attention to learning-based registration. Learning-based registration methods can be categorised into supervised and unsupervised registration approaches.

A.1.1 Supervised Registration

The supervised learning registration methods [Wu, 2016; Cao, 2017; Rohé, 2017] are primarily trained on a ground-truth training set for which simulated or registered displacement fields are available. The training dataset is usually generated with traditional registration frameworks or by generating artificial deformation fields as ground truth for warping the moving images to get the fixed images. [Sokooti, 2017; Yang, 2017]. One of the limitations of the supervised registration approaches is the registration quality, which is highly influenced by the nature of the training set of deformation map, although the requirement in terms of training set can be partially alleviated by using weakly-supervised learning [Hu, 2018b; Blendowski, 2021; Balakrishnan, 2019; Hu, 2018a; Ferrante, 2019].

A.1.2 Unsupervised Registration

In unsupervised registration [Krebs, 2019; Dalca, 2019; Balakrishnan, 2019; Hering, 2021; Mansilla, 2020], we rely on a similarity measure and regularisation to optimize the neural network for learning the transformations between the fixed and moving images. Usually, a CNN is used directly for warping the moving images, which is then compared to the fixed-image with the similarity loss. The displacement field can also be obtained from a generative adversarial neural network, which introduces a discriminator neural network for assessing the generated deformation field quality. [Tanner, 2018; Zheng, 2021; Mahapatra, 2020; Debayle, 2016].

A.1.3 Multi-layer Perceptron and Transformers

MLP is one of the most classical neural networks and consists of a stack of linear layers along with non-linear activation [Van Der Malsburg, 1986]. For several years, CNN has been widely used due to its performance on vision tasks and its computation efficiency [Krizhevsky, 2012]. Recently, several alternatives to the CNN have been proposed such as Vision Transformer (ViT) [Dosovitskiy, 2021] or MLP-Mixer [Tolstikhin, 2021] which demonstrated comparable or even better performance than CNN on classification or detection tasks. There are currently intense discussions in the community of whether patching, attention or simple MLP play the most important role in such good performance.

In this work, we propose three MLP/Transformer based models for echocardiography image registration and compare them with one representative CNN model in unsupervised echocardiography image registration. There are already works using transformers to register medical images, such as TransMorph [Chen, 2022] and Dual Transformer [Zhang, 2021], but they are mostly restricted to high signal-to-noise medical images, such as MRI images and CT images. While ultrasound images (2D) are actually the most popular imaging modality in the real world. Our inspiration not only comes from the trending debate over Transformer and MLP, but also stems from the intuition that patch-based learning methods share similar logic to traditional block matching method for cardiac tracking.

Our contributions are two-folded. First, we show the effectiveness of patch-based MLP/Transformer models in medical image registration compared with a CNN-based registration model. Second, we conduct a thorough ablation study of the influence of different structures (MLP, MLP-Mixer, Transformer) and different scales (single scale or multiple scales). Our results provide empirical support to the observation that the attention mechanism may not be the only key factor in the SOTA performances. [Liu, 2021a; Melas-Kyriazi, 2021], at least in the field of unsupervised image registration.

A.2 Methodology

A.2.1 Diffeomorphic Registration

We estimate a diffeomorphic transform between images, which preserves topology and is folding-free. Our model generates stationary velocity field $v(\theta)$ [Ashburner, 2000] instead of generating displacement maps, thanks to an integration layer applied to the velocity field leading to diffeomorphism $\phi(\theta)$. Formally, the diffeomorphic transformation ϕ is solution to a differential equation related to the predicted (stationary) velocity field V [Dalca, 2019]: $\frac{\partial \phi_t}{\partial t} = v(\phi_t); \phi_{t=0} = Id$. In a stationary velocity field setting, the transformation ϕ is defined as the exponential of the velocity field $\phi = \exp(v)$ [Arsigny, 2006]. The integration (exponential) layer applies the scaling and squaring method to approximate the diffeomorphic transform [Krebs, 2018]. The obtained transformation ϕ is then used by a spatial transform layer to deform the image.

A.2.2 Proposed frameworks

Given two images, I_{fix} and I_{mov}, we would to estimate the transformation $\phi(\theta)$ that transforms the moving image to the fixed image so that $I_{fix} \approx I_{mov} \circ \phi(\theta)$. We approximate the ideal $\phi(\theta)$ by the following proposed frameworks. The following three propositions are all based on patch-wise manipulations and share a similar general architecture. As shown in Fig.A.1, I_{fix} and I_{mov} are both processed by an identical feature extractor (green block) separately. The two feature maps are then passed through the cross feature block (blue block). After two linear layers, we obtain their corresponding velocity field. The velocity field passes through an integration layer and we obtain the final displacement field by upsampling it to the original image size.

A.2.2.1 Pure MLP registration framework

The same MLP block (Block I in Fig.A.1) are used for feature extractor and the cross feature block in this model. The outputs from two separate feature extractor (shared weights) are added together before feed into the cross feature block. We note this model **PureMLP** for abbreviation in the following paper.

A.2.2.2 MLP-Mixer registration framework

The MLP-Mixer registration framework is very similar to the former Pure MLP framework. The only difference is that the three MLP blocks used for separate feature extraction and cross feature processing are replaced by MLP-Mixer blocks [**mlpmixer**] (Block II in

Fig.A.1). The MLP-Mixer block has an identical structure as the MLP block, but with feature map transpose to obtain channel-wise feature fusion (the red cell of Block II in Fig.A.1). We note this model **MLPMixer** for abbreviation in the following paper.

A.2.2.3 Swin-Transformer registration framework

This model uses the MLP block (Block I in Fig.A.1) to first extract patch based features for both I_{fix} and I_{mov}. For cross feature block, we adapt the recent Swin Transformer [Liu, 2021b] to do the cross-patch attention locally (Swin Block in Fig.A.1). Our Swin block accepts feature input from both images (I_{fix} and I_{mov}), where key K and value V are normalised I_{fix} features while query Q comes from normalised I_{mov} features. Swin block calculates the cross-attention within a pre-defined window region. We perform normal window partition for I_{fix} features and one normal partition, one shifted partition for I_{mov} features. The cross-attention under the two types of window partition configurations are summed together before feeding to the final linear layers. Due to the page limit, we invite interested readers to refer to [Liu, 2021b] for detailed description of Swin transformer mechanism. We note this model **SwinTrans** for abbreviation in the following paper.

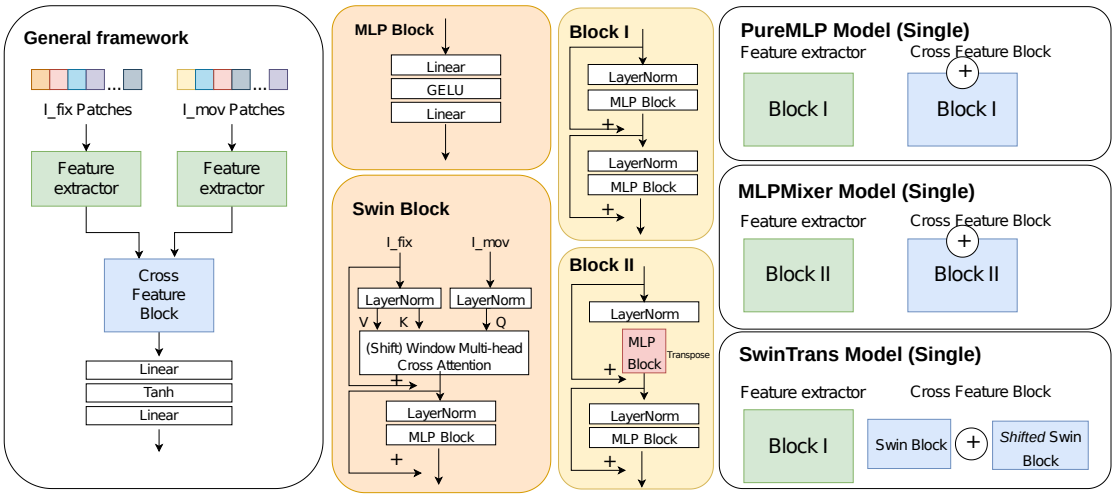


Fig. A.1.: The detailed composition of proposed three frameworks. Here we only show single-scale models. Please read Section.A.2.2 for more description.

A.2.3 Multi-scale features

In order to enforce different reception fields for patch-based models, we decide to combine multi-scale models together. This is accomplished by adopting models of different patch sizes together. It is quite similar to how CNN achieves this goal, by applying larger kernel or adding pooling layers. In particular, our multi-scale model consists of several parallel independent single-scale child models. The output of each

child model is upsampled and then combined together to form the final estimation of velocity field $v(\theta)$

$$v(\theta) = \sum_{C=1}^{N} \omega_C Out_C \tag{A.2}$$

where Out_C is the output of child model C. The final velocity field v is then passed to calculate the final transform ϕ as depicted in former subsections.

A.3 Experiments and Results

A.3.1 Dataset

To evaluate the effectiveness of our unsupervised registration models, we use a publicly accessible 2D echocardiography dataset CAMUS[1]. This dataset consists of 500 patients, each has 2D apical 4-chamber (A4C) and 2-chamber (A2C) view sequences. Manual annotation of cardiac structures (left endocardium, left epicardium and left atrium) were acquired by expert cardiologists for each patient in each view, at end-diastole (ED) and end-systole (ES) [Leclerc, 2019a]. The structure annotations of 450 patients are public available while that of the other 50 patients unreleased. In total we have 1000 pairs of ED/ES images and we randomly split (still considering age and image quality distribution) the 900 pairs (with annotations) into training (630), validation(90) and test data (180). The 100 pairs (without annotations) are included into the training set (730).

A.3.2 Implementation

We compare our proposed three models with a very popular CNN registration model VoxelMorph [Balakrishnan, 2019]. To be consistent with our setting, we make use of the diffeomorphic version of VoxelMorph model (we use the abbreviation **Vxm** in the following paper). We train all the models with input images resized to 128x128 pixels and use an Adam optimiser (learning rate = 0.0001). We set training epoch to be 500 and training is early stopped when there is no improvement on validation set over 30 epochs. Our codes are available (https://gitlab.inria.fr/epione/mlp_transformer_registration).

[1] https://www.creatis.insa-lyon.fr/Challenge/camus/

A.3.2.1 Loss function

In order to enforce the diffeomorphic property of our registration model, we apply a symmetric loss function for all the unsupervised models:

$$\arg\min L = L_{mse}(\hat{\phi}(I_{move}), I_{fix}) + L_{mse}(\hat{\phi}^{-1}(I_{fix}), I_{move}) + \lambda * L_{diff}(\hat{\phi}) \tag{A.3}$$

where $\hat{\phi}^{-1}$ is the inverse of $\hat{\phi}$ and L_{diff} is a diffusion regularizer for smoothness $L_{diff} = \int ||\nabla_x \phi + \nabla_y \phi||^2$ and set $\lambda = 0.01$ according to [Balakrishnan, 2019].

A.3.2.2 Data augmentation

In order to improve model generalisation and avoid outfitting, we apply the same random data augmentation tricks for each image pair during training phase. The following augmentation techniques: rotation, cropping and resizing, brightness adjustment, contrast change, sharpening, blurring and speckle noise addition are conducted with a probability of 0.5 separately. No augmentation is applied during validation nor test phase.

A.3.3 Experiments

A.3.3.1 Multi-scale models

(abbreviation: model name + _M) we apply three child models for PureMLP, MLPMixer and SwinTrans (with patch size of 4x4, 8x8 and 16x16 respectively). For child model of size 4x4, 8x8 and 16x16 in SwinTrans, we set the number of window size to be 8, 4 and 2, the number of heads to be 32, 16 and 8 respectively. The dimension of patch-embedding is set to be 128 for all patch-based methods. ω_C in Equation.A.2 is set to be 0.5, 0.3, 0.2 for child model with patch size of 4x4, 8x8, 16x16 separately.

A.3.3.2 Single-scale models

(abbreviation: model name + _S) we run single-scale models for PureMLP, MLPMixer and SwinTrans three proposed frameworks (using patch size of 4x4 pixels). The same configuration is set as for child model with patch-size 4x4 in multi-scale models.

A.3.4 Results

Since our SwinTrans model relies mostly on features of I_{fix} (with skip-connection of I_{fix} features), we report only the metrics related to the transformation $\phi(\theta)$ that

$I_{fix} \approx I_{move} \circ \phi(\theta)$. For CAMUS test dataset, we report the Dice score, Hausdorff distance (HD) and mean surface distance (MSD) between ground truth ED mask and transformed ES mask and the Jacobian determinant in the area of myocardium region.

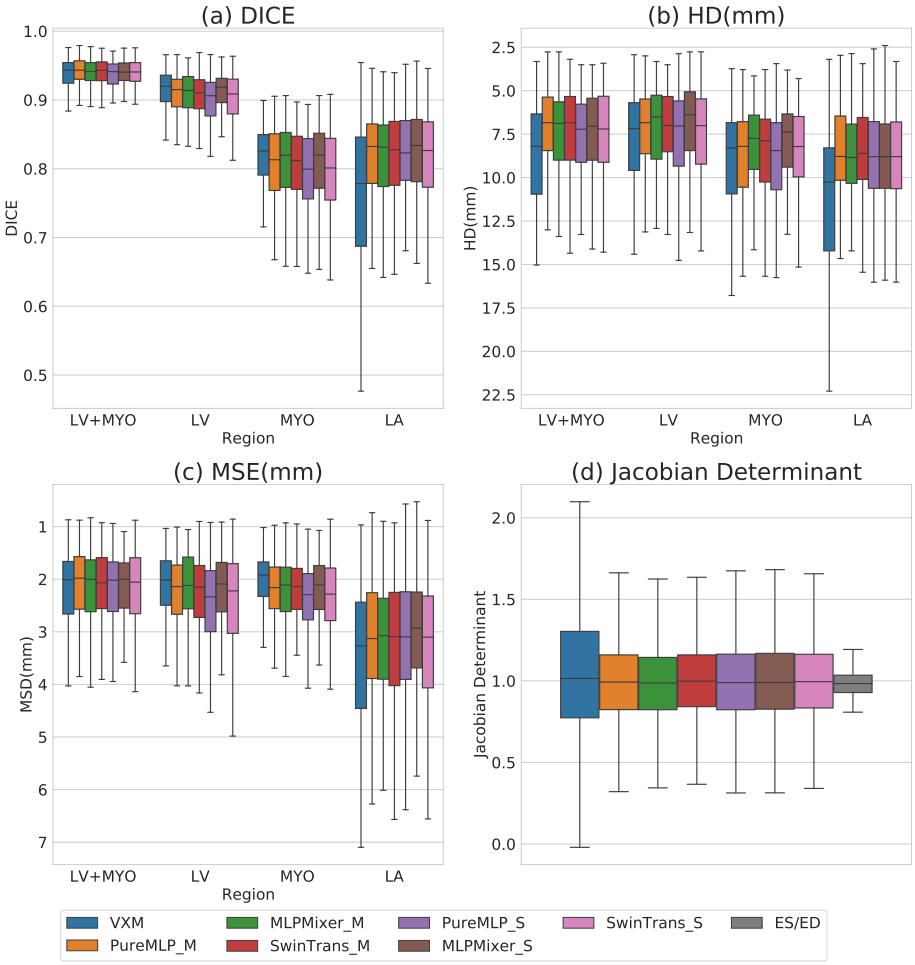


Fig. A.2.: Comparison of evaluation metrics (Dice score, Hausdorff distance (HD), mean surface distance (MSD) and Jacobin determinant) on test dataset of CAMUS. The Jacobin determinant is only computed in the myocardium region. Except Jacobin determinant figure, the higher the boxplot is in the figure, the better performance it will be.

A.3.4.1 Evaluation on CAMUS dataset

From Fig.A.2 we can observe that on CAMUS test dataset, almost all the proposed models, no matter it is multi-scale or single-scale, no matter what kind of sub-block it contains (MLP or Transformer or MLP-Mixer), have achieved comparable performance than the CNN model (Vxm), in particular for the whole left ventricle and left atrium registration. In addition, the distribution of Jacobin determinant shows that our patch based methods tend to generate more plausible transform, i.e. closer to real ES/ED myocardium area

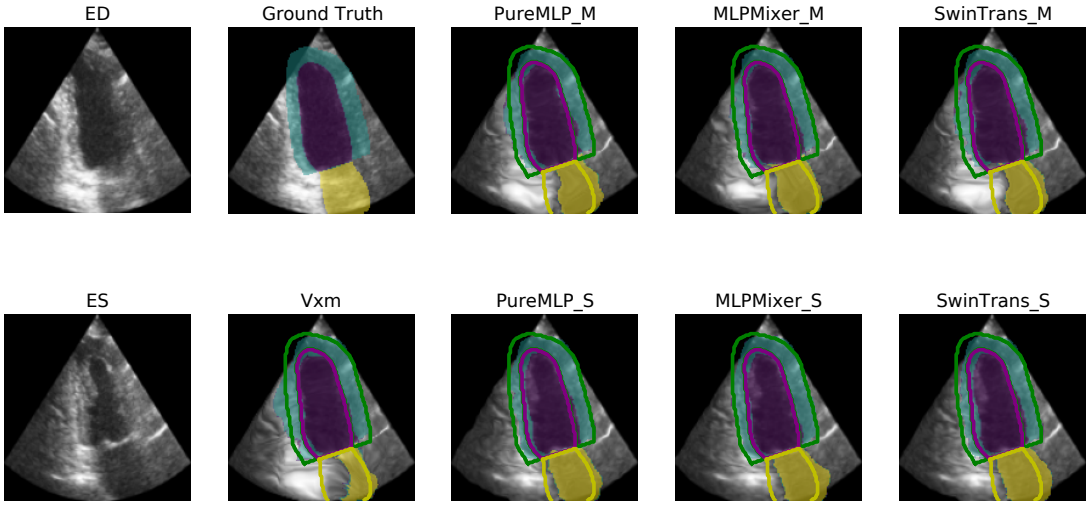


Fig. A.3.: The same registration example on CAMUS test data with transformed ES masks. Colourful patches are corresponding estimations while bold contours are the ground truth (Yellow: left atrium, Purple: left ventricle, Green: myocardium).

Tab. A.1.: Time and space complexity between different models (evaluated on a GTX 2080Ti)

Model	GPU Memory	Train time (s/pair)	Test time (s/pair)
Vxm	1365 MiB	0.020	0.0047
PureMLP_S	1411 MiB	0.020	0.0038
MLPMixer_S	1447 MiB	0.020	0.0038
SwinTrans_S	1479 MiB	0.029	0.0047

change. This is consistent with the Hausdorff distance results, which indicates that while preserving comparable registration performance, patch-based methods are more resistant to estimation of false large deformation (see the atrium and myocardium region of example in Fig.A.3). What's more, single-scale models and multi-scale models have similar performance. With single-sized patches, we are already capable to let feature information flow through the whole image area and estimate registration transform efficiently (see time and space complexity in Table.A.1).

A.4 Conclusion

In summary, we propose three novel patch-based registration architectures using only MLPs and Transformers. We show that our single and multi-scale models perform similarly and even better to CNN-based registration frameworks on a large echocardiography dataset. The three proposed models demonstrate similar performance among themselves. Our experiments show that patch-based models using MLP/Transformer can perform 2D medical image registration. We shared a similar conclusion with previous works [Liu, 2021a; Melas-Kyriazi, 2021] that the success of Transformer in vision tasks cannot

be simply attributed to the attention mechanism, at least in image registration task. Future works will concentrate on the application of MLP/Transformer in time-series motion tracking.

Bibliography

[Acharya, 2017a] U Rajendra Acharya, Hamido Fujita, Muhammad Adam, et al. "Automated characterization and classification of coronary artery disease and myocardial infarction by decomposition of ECG signals: A comparative study". In: *Information Sciences* 377 (2017), pp. 17–29 (cit. on p. 68).

[Acharya, 2017b] U Rajendra Acharya, Hamido Fujita, Shu Lih Oh, et al. "Application of deep convolutional neural network for automated detection of myocardial infarction using ECG signals". In: *Information Sciences* 415 (2017), pp. 190–198 (cit. on p. 69).

[Acosta, 2022] Julián N Acosta, Guido J Falcone, Pranav Rajpurkar, and Eric J Topol. "Multimodal biomedical AI". In: *Nature Medicine* 28.9 (2022), pp. 1773–1784 (cit. on p. 86).

[Ahn, 2013] Chi Young Ahn. "Robust myocardial motion tracking for echocardiography: Variational framework integrating local-to-global deformation". In: *Computational and Mathematical Methods in Medicine* 2013 (2013) (cit. on p. 11).

[Ahn, 2020] Shawn S Ahn, Kevinminh Ta, Allen Lu, et al. "Unsupervised motion tracking of left ventricle in echocardiography". In: *Medical imaging 2020: Ultrasonic imaging and tomography*. Vol. 11319. SPIE. 2020, pp. 196–202 (cit. on pp. 11, 19, 37, 41).

[Alessandrini, 2016] Martino Alessandrini, Brecht Heyde, et al. "Detailed evaluation of five 3D speckle tracking algorithms using synthetic echocardiographic recordings". In: *IEEE transactions on medical imaging* 35.8 (2016), pp. 1915–1926 (cit. on p. 36).

[Alessandrini, 2018] Martino Alessandrini, Bidisha Chakraborty, Brecht Heyde, et al. "Realistic Vendor-Specific Synthetic Ultrasound Data for Quality Assurance of 2-D Speckle Tracking Echocardiography: Simulation Pipeline and Open Access Database". In: *IEEE Transactions on Ultrasonics, Ferroelectrics, and Frequency Control* 65.3 (2018), pp. 411–422 (cit. on pp. 11, 37).

[Ansari, 2017] Sardar Ansari, Negar Farzaneh, Marlena Duda, et al. "A review of automated methods for detection of myocardial ischemia and infarction using electrocardiogram and electronic health records". In: *IEEE reviews in biomedical engineering* 10 (2017), pp. 264–298 (cit. on p. 69).

[Antelmi, 2019] Luigi Antelmi, Nicholas Ayache, Philippe Robert, and Marco Lorenzi. "Sparse multi-channel variational autoencoder for the joint analysis of heterogeneous data". In: *International Conference on Machine Learning*. PMLR. 2019, pp. 302–311 (cit. on p. 97).

[Arif, 2012] Muhammad Arif, Ijaz A Malagore, and Fayyaz A Afsar. "Detection and localization of myocardial infarction using k-nearest neighbor classifier". In: *Journal of medical systems* 36.1 (2012), pp. 279–289 (cit. on p. 68).

[Arsigny, 2006] Vincent Arsigny and et al. "A Log-Euclidean Framework for Statistics on Diffeomorphisms". In: *Medical Image Computing and Computer-Assisted Intervention – MICCAI 2006*. Berlin, Heidelberg: Springer Berlin Heidelberg, 2006, pp. 924–931 (cit. on pp. 102, 104).

[Arsigny, 2009] Vincent Arsigny, Olivier Commowick, Nicholas Ayache, and Xavier Pennec. "A fast and log-euclidean polyaffine framework for locally linear registration". In: *Journal of Mathematical Imaging and Vision* 33.2 (2009), pp. 222–238 (cit. on p. 38).

[Ashburner, 2000] John Ashburner and Karl J. Friston. "Voxel-Based Morphometry—The Methods". In: *NeuroImage* 11.6 (2000), pp. 805–821 (cit. on p. 104).

[Audibert, 2020] Julien Audibert, Pietro Michiardi, Frédéric Guyard, Sébastien Marti, and Maria A Zuluaga. "Usad: Unsupervised anomaly detection on multivariate time series". In: *Proceedings of the 26th ACM SIGKDD international conference on knowledge discovery & data mining*. 2020, pp. 3395–3404 (cit. on p. 98).

[Azarmehr, 2020] Neda Azarmehr, Xujiong Ye, et al. "An optimisation-based iterative approach for speckle tracking echocardiography". In: *Medical and Biological Engineering and Computing* 58.6 (2020), pp. 1309–1323 (cit. on pp. 11, 36).

[Balakrishnan, 2019] Guha Balakrishnan et al. "VoxelMorph: A Learning Framework for Deformable Medical Image Registration". In: *IEEE Transactions on Medical Imaging* 38.8 (2019), pp. 1788–1800 (cit. on pp. 11, 102, 103, 106, 107).

[Barbosa, 2013] Daniel Barbosa, Brecht Heyde, Thomas Dietenbeck, et al. "Fast Left Ventricle Tracking in 3D Echocardiographic Data Using Anatomical Affine Optical Flow". In: *Functional Imaging and Modeling of the Heart*. Ed. by Sébastien Ourselin, Daniel Rueckert, and Nicolas Smith. Berlin, Heidelberg: Springer Berlin Heidelberg, 2013, pp. 191–199 (cit. on p. 36).

[Beco, 2022] Sofia C Beco, João Ribeiro Pinto, and Jaime S Cardoso. "Electrocardiogram lead conversion from single-lead blindly-segmented signals". In: *BMC Medical Informatics and Decision Making* 22.1 (2022), pp. 1–15 (cit. on p. 97).

[Bernard, 2018] O. Bernard, A. Lalande, and et al. "Deep Learning Techniques for Automatic MRI Cardiac Multi-Structures Segmentation and Diagnosis: Is the Problem Solved?" In: *IEEE Transactions on Medical Imaging* 37.11 (2018), pp. 2514–2525 (cit. on p. 29).

[Bhaskar, 2015] Nitin Aji Bhaskar. “Performance analysis of support vector machine and neural networks in detection of myocardial infarction”. In: *Procedia Computer Science* 46 (2015), pp. 20–30 (cit. on p. 68).

[Blendowski, 2021] Max Blendowski and et al. “Weakly-supervised learning of multimodal features for regularised iterative descent in 3D image registration”. In: *Medical Image Analysis* 67 (2021), p. 101822 (cit. on p. 102).

[Bohlender, 2021] Simon Bohlender, Ilkay Oksuz, and Anirban Mukhopadhyay. “A survey on shape-constraint deep learning for medical image segmentation”. In: *IEEE Reviews in Biomedical Engineering* (2021) (cit. on p. 19).

[Boukerroui, 2003] Djamal Boukerroui et al. “Velocity estimation in ultrasound images: A block matching approach”. In: *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)* 2732 (2003), pp. 586–598 (cit. on p. 11).

[Bousseljot, 1995] R Bousseljot, D Kreiseler, and A Schnabel. “Nutzung der EKG-Signaldatenbank CARDIODAT der PTB über das Internet”. In: (1995) (cit. on p. 72).

[Cao, 2005] Yan Cao and et al. “Large deformation diffeomorphic metric mapping of vector fields”. In: *IEEE Transactions on Medical Imaging* 24.9 (2005), pp. 1216–1230 (cit. on p. 102).

[Cao, 2017] Xiaohuan Cao and et al. “Deformable Image Registration Based on Similarity-Steered CNN Regression”. In: *Medical Image Computing and Computer Assisted Intervention*. Cham: Springer International Publishing, 2017, pp. 300–308 (cit. on p. 102).

[Cerqueira, 2002] Manuel D. Cerqueira, Neil J. Weissman, and et al. “Standardized myocardial segmentation and nomenclature for tomographic imaging of the heart: A statement for healthcare professionals from the Cardiac Imaging Committee of the Council on Clinical Cardiology of the American Heart Association”. In: *Journal of Nuclear Cardiology* 9.2 (2002), pp. 240–245 (cit. on p. 21).

[Chakraborty, 2016] Bidisha Chakraborty et al. “Fast myocardial strain estimation from 3D ultrasound through elastic image registration with analytic regularization”. In: 9790 (2016). Ed. by Neb Duric and Brecht Heyde, pp. 48–54 (cit. on p. 11).

[Cheema, 2021] Baljash S Cheema, James Walter, Akhil Narang, and James D Thomas. “Artificial intelligence–enabled POCUS in the COVID-19 ICU: A new spin on cardiac ultrasound”. In: *Case Reports* 3.2 (2021), pp. 258–263 (cit. on p. 4).

[Chen, 2020] Chen Chen, Wenjia Bai, and et al. “Improving the Generalizability of Convolutional Neural Network-Based Segmentation on CMR Images”. In: *Frontiers in Cardiovascular Medicine* 7 (2020), p. 105 (cit. on p. 18).

[Chen, 2021] Jintai Chen, Xiangshang Zheng, Hongyun Yu, Danny Z Chen, and Jian Wu. “Electrocardio Panorama: Synthesizing New ECG views with Self-supervision”. In: *IJCAI*. 2021 (cit. on p. 97).

[Chen, 2022] Junyu Chen, Eric C Frey, Yufan He, et al. “Transmorph: Transformer for unsupervised medical image registration”. In: *Medical image analysis* 82 (2022), p. 102615 (cit. on p. 103).

[Cikes, 2015] Maja Cikes and Scott D. Solomon. “Beyond ejection fraction: an integrative approach for assessment of cardiac structure and function in heart failure”. In: *European Heart Journal* 37.21 (Sept. 2015), pp. 1642–1650 (cit. on p. 56).

[Clough, 2020] James R Clough, Nicholas Byrne, Ilkay Oksuz, et al. “A topological loss function for deep-learning based image segmentation using persistent homology”. In: *IEEE transactions on pattern analysis and machine intelligence* 44.12 (2020), pp. 8766–8778 (cit. on pp. 11, 18).

[Dadon, 2020] Ziv Dadon, David Rosenmann, Adi Butnaru, et al. “USE OF ARTIFICIAL INTELLIGENCE BY MEDICAL STUDENTS TO ENABLE ACCURATE POINT-OF-CARE ECHOCARDIOGRAPHIC ASSESSMENT OF EJECTION FRACTION”. In: *Journal of the American College of Cardiology* 75.11_Supplement_1 (2020), pp. 1568–1568 (cit. on p. 4).

[Dai, 2021] Hao Dai, Hsin-Ginn Hwang, and Vincent S Tseng. “Convolutional neural network based automatic screening tool for cardiovascular diseases using different intervals of ECG signals”. In: *Computer Methods and Programs in Biomedicine* 203 (2021), p. 106035 (cit. on p. 11).

[Dalca, 2018] Adrian V Dalca, Guha Balakrishnan, John Guttag, and Mert R Sabuncu. “Unsupervised learning for fast probabilistic diffeomorphic registration”. In: *MICCAI 2018: Granada, Spain, September 16-20, 2018, Proceedings, Part I*. Springer. 2018, pp. 729–738 (cit. on p. 43).

[Dalca, 2019] Adrian V. Dalca and et al. “Unsupervised learning of probabilistic diffeomorphic registration for images and surfaces”. In: *Medical Image Analysis* 57 (2019), pp. 226–236 (cit. on pp. 103, 104).

[Davatzikos, 1997] Christos Davatzikos. “Spatial Transformation and Registration of Brain Images Using Elastically Deformable Models”. In: *Computer Vision and Image Understanding* 66.2 (1997), pp. 207–222 (cit. on p. 102).

[De Vos, 2019] Bob D De Vos, Floris F Berendsen, Max A Viergever, et al. “A deep learning framework for unsupervised affine and deformable image registration”. In: *Medical image analysis* 52 (2019), pp. 128–143 (cit. on pp. 37, 56).

[Debayle, 2016] Johan Debayle and Benoit Presles. “Rigid image registration by General Adaptive Neighborhood matching”. In: *Pattern Recognition* 55 (2016), pp. 45–57 (cit. on p. 103).

[Degerli, 2021a] Aysen Degerli et al. “Early detection of myocardial infarction in low-quality echocardiography”. In: *IEEE Access* 9 (2021), pp. 34442–34453 (cit. on pp. 45, 60).

[Degerli, 2021b] Aysen Degerli, Morteza Zabihi, Serkan Kiranyaz, et al. “Early detection of myocardial infarction in low-quality echocardiography”. In: *IEEE Access* 9 (2021), pp. 34442–34453 (cit. on p. 42).

[Degerli, 2024] Aysen Degerli, Serkan Kiranyaz, Tahir Hamid, Rashid Mazhar, and Moncef Gabbouj. “Early myocardial infarction detection over multi-view echocardiography”. In: *Biomedical Signal Processing and Control* 87 (2024), p. 105448 (cit. on pp. 56, 60–63, 92).

[Deng, 2022] Yinlong Deng et al. “Myocardial strain analysis of echocardiography based on deep learning”. In: *Frontiers in Cardiovascular Medicine* 9 (2022), p. 1067760 (cit. on p. 11).

[Dosovitskiy, 2021] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, et al. “An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale”. In: *International Conference on Learning Representations*. 2021 (cit. on p. 103).

[Duan, 2022] Ziheng Duan, Haoyan Xu, Yueyang Wang, et al. “Multivariate time-series classification with hierarchical variational graph pooling”. In: *Neural Networks* 154 (2022), pp. 481–490 (cit. on p. 98).

[Evain, 2022] Ewan Evain, Yunyun Sun, Khuram Faraz, Damien Garcia, et al. “Motion estimation by deep learning in 2D echocardiography: synthetic dataset and validation”. In: *IEEE transactions on medical imaging* (2022) (cit. on pp. 11, 36, 37).

[Fang, 2022] Rui Fang, Chih-Cheng Lu, Cheng-Ta Chuang, and Wen-Han Chang. “A visually interpretable detection method combines 3-D ECG with a multi-VGG neural network for myocardial infarction identification”. In: *Computer Methods and Programs in Biomedicine* 219 (2022), p. 106762 (cit. on pp. 69, 78).

[Farneb, 2003] Gunnar Farneb. “Two-Frame Motion Estimation Based on Polynomial Expansion”. In: *Scandinavian conference on Image analysis* 2749.1 (2003), pp. 363–370 (cit. on p. 36).

[Ferrante, 2019] Enzo Ferrante and et al. “Weakly Supervised Learning of Metric Aggregations for Deformable Image Registration”. In: *IEEE Journal of Biomedical and Health Informatics* 23.4 (2019), pp. 1374–1384 (cit. on p. 102).

[Folland, 1979] E. D. Folland, A. F. Parisi, and et al. “Assessment of left ventricular ejection fraction and volumes by real-time, two-dimensional echocardiography. A comparison of cineangiographic and radionuclide techniques”. In: *Circulation* 60.4 (1979), pp. 760–766 (cit. on pp. 11, 25).

[Folse, 1962] Roland Folse and Eugene Braunwald. “Determination of fraction of left ventricular volume ejected per beat and of ventricular end-diastolic and residual volumes. Experimental and clinical observations with a precordial dilution technic”. In: *Circulation* 25 (1962), pp. 674–685 (cit. on pp. 11, 58).

[Fortune, 2022] Julian D Fortune, Natalie E Coppa, Kazi T Haq, Hetal Patel, and Larisa G Tereshchenko. “Digitizing ECG image: a new method and open-source software code”. In: *Computer Methods and Programs in Biomedicine* (2022), p. 106890 (cit. on pp. 73, 82).

[Golany, 2020] Tomer Golany, Kira Radinsky, and Daniel Freedman. “SimGANs: Simulator-based generative adversarial networks for ECG synthesis to improve deep ECG classification”. In: *International Conference on Machine Learning*. PMLR. 2020, pp. 3597–3606 (cit. on p. 97).

[Goto, 2022] Shinichi Goto, Divyarajsinhji Solanki, Jenine E John, et al. “Multinational federated learning approach to train ECG and echocardiogram models for hypertrophic cardiomyopathy detection”. In: *Circulation* 146.10 (2022), pp. 755–769 (cit. on p. 12).

[Graja, 2005] Salim Graja and J-M Boucher. “Hidden Markov tree model applied to ECG delineation”. In: *IEEE Transactions on Instrumentation and Measurement* 54.6 (2005), pp. 2163–2168 (cit. on p. 68).

[Grzech, 2022] Daniel Grzech, Mohammad Farid Azampour, Ben Glocker, et al. “A variational Bayesian method for similarity learning in non-rigid image registration”. In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 2022, pp. 119–128 (cit. on p. 97).

[Han, 2021] Zongbo Han, Changqing Zhang, Huazhu Fu, and Joey Tianyi Zhou. “Trusted Multi-View Classification”. In: *International Conference on Learning Representations*. 2021 (cit. on pp. 86, 89).

[Hasselberg, 2014] Nina E. Hasselberg, Kristina H. Haugaa, and et al. “Left ventricular global longitudinal strain is associated with exercise capacity in failing hearts with preserved and reduced ejection fraction”. In: *European Heart Journal - Cardiovascular Imaging* 16.2 (Dec. 2014), pp. 217–224 (cit. on p. 19).

[Hering, 2021] Alessa Hering and et al. “CNN-based lung CT registration with multiple anatomical constraints”. In: *Medical Image Analysis* 72 (2021), p. 102139 (cit. on p. 103).

[Hou, 2019] Ming Hou, Jiajia Tang, Jianhai Zhang, Wanzeng Kong, and Qibin Zhao. “Deep Multimodal Multilinear Fusion with High-order Polynomial Pooling”. In: *Advances in Neural Information Processing Systems*. Ed. by H Wallach, H Larochelle, A Beygelzimer, et al. Vol. 32. Curran Associates, Inc., 2019 (cit. on p. 86).

[Hu, 2018a] Yipeng Hu and et al. “Label-driven weakly-supervised learning for multimodal deformable image registration”. In: *2018 IEEE 15th International Symposium on Biomedical Imaging (ISBI 2018)*. 2018, pp. 1070–1074 (cit. on p. 102).

[Hu, 2018b] Yipeng Hu and et al. “Weakly-supervised convolutional neural networks for multimodal image registration”. In: *Medical Image Analysis* 49 (2018), pp. 1–13 (cit. on p. 102).

[Huang, 2020] Mu-Shiang Huang et al. “Automated recognition of regional wall motion abnormalities through deep neural network interpretation of transthoracic echocardiography”. In: *Circulation* 142.16 (2020), pp. 1510–1520 (cit. on p. 56).

[Jahmunah, 2021] V Jahmunah, Eddie Yin Kwee Ng, Tan Ru San, and U Rajendra Acharya. “Automated detection of coronary artery disease, myocardial infarction and congestive heart failure using GaborCNN model with ECG signals”. In: *Computers in biology and medicine* 134 (2021), p. 104457 (cit. on p. 11).

[Jaleel, 2016] Abdul Jaleel, Reza Tafreshi, and Leyla Tafreshi. “An expert system for differential diagnosis of myocardial infarction”. In: *Journal of Dynamic Systems, Measurement, and Control* 138.11 (2016) (cit. on p. 68).

[Jayachandran, 2010] ES Jayachandran, Paul Joseph K, R Acharya U, et al. “Analysis of myocardial infarction using discrete wavelet transform”. In: *Journal of medical systems* 34.6 (2010), pp. 985–992 (cit. on p. 68).

[Jayakumar, 2020] Siddhant M Jayakumar, Wojciech M Czarnecki, Jacob Menick, et al. “Multiplicative interactions and where to find them”. In: (2020) (cit. on p. 86).

[Jia, 2019] Shuman Jia, Antoine Despinasse, and et al. “Automatically Segmenting the Left Atrium from Cardiac Images Using Successive 3D U-Nets and a Contour Loss”. In: *Lecture Notes in Computer Science* 11395 LNCS (2019), pp. 221–229 (cit. on p. 23).

[Jimenez-Perez, 2021] Guillermo Jimenez-Perez, Alejandro Alcaine, and Oscar Camara. “Delineation of the electrocardiogram with a mixed-quality-annotations dataset using convolutional neural networks”. In: *Scientific reports* 11.1 (2021), p. 863 (cit. on pp. 11, 68).

[Jsang, 2018] Audun Jsang. *Subjective Logic: A formalism for reasoning under uncertainty*. Springer Publishing Company, Incorporated, 2018 (cit. on p. 87).

[Judge, 2023] Thierry Judge, Olivier Bernard, Woo-Jin Cho Kim, et al. “Asymmetric Contour Uncertainty Estimation for Medical Image Segmentation”. In: *International Conference on Medical Image Computing and Computer-Assisted Intervention*. Springer. 2023, pp. 210–220 (cit. on p. 97).

[Kalam, 2014] Kashif Kalam, Petr Otahal, and Thomas H Marwick. “Prognostic implications of global LV dysfunction: a systematic review and meta-analysis of global longitudinal strain and ejection fraction”. In: *Heart* 100.21 (2014), pp. 1673–1680 (cit. on p. 7).

[Karim, 2019] Fazle Karim, Somshubra Majumdar, Houshang Darabi, and Samuel Harford. “Multivariate LSTM-FCNs for time series classification”. In: *Neural networks* 116 (2019), pp. 237–245 (cit. on p. 98).

[Kazemi Esfeh, 2020] Mohammad Mahdi Kazemi Esfeh et al. “A deep Bayesian video analysis framework: towards a more robust estimation of ejection fraction”. In: *International Conference on Medical Image Computing and Computer-Assisted Intervention*. Springer. 2020, pp. 582–590 (cit. on p. 11).

[Kefalas, 2020] Triantafyllos Kefalas, Konstantinos Vougioukas, Yannis Panagakis, et al. “Speech-driven facial animation using polynomial fusion of features”. In: *ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE. 2020, pp. 3487–3491 (cit. on p. 86).

[King, 2016] Kevin R. King et al. “Point-of-Care Technologies for Precision Cardiovascular Care and Clinical Research”. In: *JACC: Basic to Translational Science* 1.1 (2016), pp. 73–86 (cit. on p. 3).

[Kiyasseh, 2021] Dani Kiyasseh, Tingting Zhu, and David A Clifton. “Clocs: Contrastive learning of cardiac signals across space, time, and patients”. In: *International Conference on Machine Learning*. PMLR. 2021, pp. 5606–5615 (cit. on p. 97).

[Klabunde, 2011] Richard Klabunde. *Cardiovascular physiology concepts*. Lippincott Williams & Wilkins, 2011 (cit. on p. 8).

[Kraigher-Krainer, 2014] Elisabeth Kraigher-Krainer, Amil M. Shah, and et al. “Impaired Systolic Function by Strain Imaging in Heart Failure With Preserved Ejection Fraction”. In: *Journal of the American College of Cardiology* 63.5 (2014), pp. 447–456 (cit. on p. 19).

[Krebs, 2018] Julian Krebs and et al. “Unsupervised Probabilistic Deformation Modeling for Robust Diffeomorphic Registration”. In: *Deep Learning in Medical Image Analysis and Multimodal Learning for Clinical Decision Support*. Cham: Springer International Publishing, 2018, pp. 101–109 (cit. on p. 104).

[Krebs, 2019] J. Krebs and et al. “Learning a Probabilistic Model for Diffeomorphic Registration”. In: *IEEE Transactions on Medical Imaging* 38.9 (2019), pp. 2165–2176 (cit. on pp. 19, 43, 103).

[Krebs, 2020] Julian Krebs, Tommaso Mansi, Nicholas Ayache, and Hervé Delingette. “Probabilistic motion modeling from medical image sequences: application to cardiac cine-MRI”. In: *International Workshop on Statistical Atlases and Computational Models of the Heart*. Springer. 2020, pp. 176–185 (cit. on pp. 37, 43, 45, 50, 56).

[Krizhevsky, 2012] Alex Krizhevsky and et al. “ImageNet Classification with Deep Convolutional Neural Networks”. In: *Advances in Neural Information Processing Systems*. Ed. by F. Pereira, C. J. C. Burges, L. Bottou, and K. Q. Weinberger. Vol. 25. Curran Associates, Inc., 2012 (cit. on p. 103).

[Kusunose, 2020] Kenya Kusunose et al. “A deep learning approach for assessment of regional wall motion abnormality from echocardiographic images”. In: *Cardiovascular Imaging* 13.2_Part_1 (2020), pp. 374–381 (cit. on p. 56).

[Laumer, 2020] Fabian Laumer, Gabriel Fringeli, Alina Dubatovka, Laura Manduchi, and Joachim M Buhmann. “DeepHeartBeat: Latent trajectory learning of cardiac cycles using cardiac ultrasounds”. In: *Machine Learning for Health*. PMLR. 2020, pp. 194–212 (cit. on p. 97).

[Leclerc, 2019a] S. Leclerc, E. Smistad, and et al. “Deep Learning for Segmentation Using an Open Large-Scale Dataset in 2D Echocardiography”. In: *IEEE Transactions on Medical Imaging* 38.9 (2019), pp. 2198–2210 (cit. on pp. 11, 18, 24, 32, 33, 42, 60, 106).

[Leclerc, 2019b] S. Leclerc, E. Smistad, and et al. “RU-Net: A refining segmentation network for 2D echocardiography”. In: *2019 IEEE International Ultrasonics Symposium (IUS)*. 2019, pp. 1160–1163 (cit. on pp. 18, 27, 32, 33).

[Lee, 2019] JeeEun Lee, KyeongTaek Oh, Byeongnam Kim, and Sun K Yoo. “Synthesis of electrocardiogram V-lead signals from limb-lead measurement using R-peak aligned generative adversarial network”. In: *IEEE journal of biomedical and health informatics* 24.5 (2019), pp. 1265–1275 (cit. on p. 97).

[Li, 2023] Ya Li, Jing-hao Luo, Qing-yun Dai, et al. “A deep learning approach to cardiovascular disease classification using empirical mode decomposition for ECG feature extraction”. In: *Biomedical Signal Processing and Control* 79 (2023), p. 104188 (cit. on p. 11).

[Ling, 2022] Hang Jung Ling, Damien Garcia, and Olivier Bernard. “Reaching intra-observer variability in 2-D echocardiographic image segmentation with a simple U-Net architecture”. In: *IEEE International Ultrasonics Symposium (IUS)*. 2022 (cit. on pp. 11, 32).

[Liu, 2015] Bin Liu, Jikui Liu, Guoqing Wang, et al. “A novel electrocardiogram parameterization algorithm and its application in myocardial infarction detection”. In: *Computers in biology and medicine* 61 (2015), pp. 178–184 (cit. on p. 69).

[Liu, 2018] Feifei Liu, Chengyu Liu, Lina Zhao, et al. “An open access database for evaluating the algorithms of electrocardiogram rhythm and morphology abnormality detection”. In: *Journal of Medical Imaging and Health Informatics* 8.7 (2018), pp. 1368–1373 (cit. on p. 72).

[Liu, 2021a] Hanxiao Liu, Zihang Dai, David So, and Quoc V Le. “Pay attention to mlps”. In: *Advances in Neural Information Processing Systems* 34 (2021), pp. 9204–9215 (cit. on pp. 103, 109).

[Liu, 2021b] Ze Liu and et al. “Swin Transformer: Hierarchical Vision Transformer using Shifted Windows”. In: *International Conference on Computer Vision (ICCV)* (2021) (cit. on p. 105).

[Liu, 2023] Bohan Liu et al. “A deep learning framework assisted echocardiography with diagnosis, lesion localization, phenogrouping heterogeneous disease, and anomaly detection”. In: *Scientific Reports* 13.1 (2023), p. 3 (cit. on p. 56).

[Lu, 2000] HL Lu, K Ong, and P Chia. “An automated ECG classification system based on a neuro-fuzzy system”. In: *Computers in Cardiology 2000. Vol. 27 (Cat. 00CH37163)*. IEEE. 2000, pp. 387–390 (cit. on p. 68).

[Lundberg, 2017] Scott M Lundberg and Su-In Lee. “A Unified Approach to Interpreting Model Predictions”. In: *Advances in Neural Information Processing Systems*. Ed. by I. Guyon, U. Von Luxburg, S. Bengio, et al. Vol. 30. Curran Associates, Inc., 2017 (cit. on p. 69).

[Mahapatra, 2020] Dwarikanath Mahapatra and Zongyuan Ge. “Training data independent image registration using generative adversarial networks and domain adaptation”. In: *Pattern Recognition* 100 (2020), p. 107109 (cit. on p. 103).

[Makowski, 2021] Dominique Makowski, Tam Pham, Zen J. Lau, et al. “NeuroKit2: A Python toolbox for neurophysiological signal processing”. In: *Behavior Research Methods* 53.4 (Feb. 2021), pp. 1689–1696 (cit. on p. 70).

[Mansi, 2011] Tommaso Mansi, Xavier Pennec, Maxime Sermesant, Hervé Delingette, and Nicholas Ayache. “iLogDemons: A demons-based registration algorithm for tracking incompressible elastic biological tissues”. In: *International journal of computer vision* 92 (2011), pp. 92–111 (cit. on pp. 37, 41).

[Mansilla, 2020] Lucas Mansilla and et al. “Learning deformable registration of medical images with anatomical constraints”. In: *Neural Networks* 124 (2020), pp. 269–279 (cit. on p. 103).

[McLeod, 2015] Kristin McLeod, Maxime Sermesant, Philipp Beerbaum, and Xavier Pennec. “Spatio-temporal tensor decomposition of a polyaffine motion model for a better analysis of pathological left ventricular dynamics”. In: *IEEE transactions on medical imaging* 34.7 (2015), pp. 1562–1575 (cit. on pp. 37, 56).

[Melas-Kyriazi, 2021] Luke Melas-Kyriazi. *Do You Even Need Attention? A Stack of Feed-Forward Layers Does Surprisingly Well on ImageNet*. 2021. arXiv: 2105.02723 [cs.CV] (cit. on pp. 103, 109).

[Mixon, 2012] Timothy A Mixon, Eunice Suhr, Gerald Caldwell, et al. “Retrospective description and analysis of consecutive catheterization laboratory ST-segment elevation myocardial infarction activations with proposal, rationale, and use of a new classification scheme”. In: *Circulation: Cardiovascular Quality and Outcomes* 5.1 (2012), pp. 62–69 (cit. on p. 68).

[Morales, 2021] Manuel A. Morales et al. “DeepStrain: A Deep Learning Workflow for the Automated Characterization of Cardiac Mechanics”. In: *Frontiers in Cardiovascular Medicine* 8 (2021), p. 1041 (cit. on p. 56).

[Morris, 2014] Daniel A Morris, Kyoko Otani, Tarek Bekfani, et al. “Multidirectional global left ventricular systolic function in normal subjects and patients with hypertension: multicenter evaluation”. In: *Journal of the American Society of Echocardiography* 27.5 (2014), pp. 493–500 (cit. on p. 7).

[Narang, 2016] Akhil Narang et al. “The Supply and Demand of the Cardiovascular Workforce”. In: *Journal of the American College of Cardiology* 68.15 (2016), pp. 1680–1689 (cit. on p. 4).

[Nejedly, 2022] Petr Nejedly, Adam Ivora, Ivo Viscor, et al. “Classification of ECG using ensemble of Residual CNNs with or without attention mechanism”. In: *Physiological Measurement* 43.4 (2022), p. 044001 (cit. on p. 12).

[Oh, 2022] Jungwoo Oh, Hyunseung Chung, Joon-myoung Kwon, Dong-gyun Hong, and Edward Choi. “Lead-agnostic self-supervised learning for local and global representations of electrocardiogram”. In: *Conference on Health, Inference, and Learning*. PMLR. 2022, pp. 338–353 (cit. on p. 97).

[Oktay, 2017] Ozan Oktay et al. “Anatomically constrained neural networks (ACNNs): application to cardiac image enhancement and segmentation”. In: *IEEE transactions on medical imaging* 37.2 (2017), pp. 384–395 (cit. on pp. 11, 18, 32).

[Oliveira, 2014] Francisco P.M. Oliveira and João Manuel R.S. Tavares. “Medical image registration: a review”. In: *Computer Methods in Biomechanics and Biomedical Engineering* 17.2 (2014), pp. 73–93 (cit. on p. 102).

[Omar, 2018] Hasmila A Omar et al. “Automated myocardial wall motion classification using handcrafted features vs a deep cnn-based mapping”. In: *2018 40th Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC)*. IEEE. 2018, pp. 3140–3143 (cit. on p. 56).

[Østvik, 2021] Andreas Østvik, Ivar Mjåland Salte, Erik Smistad, Thuy Mi Nguyen, Daniela Melichova, et al. “Myocardial function imaging in echocardiography using deep learning”. In: *IEEE Transactions on Medical Imaging* 40.5 (2021), pp. 1340–1351 (cit. on pp. 11, 36, 56).

[Ouyang, 2020] David Ouyang, Bryan He, et al. “Video-based AI for beat-to-beat assessment of cardiac function”. In: *Nature* 580.7802 (2020), pp. 252–256 (cit. on pp. 24, 30, 42, 60).

[Painchaud, 2022] Nathan Painchaud et al. “Echocardiography segmentation with enforced temporal consistency”. In: *IEEE Transactions on Medical Imaging* 41.10 (2022), pp. 2867–2878 (cit. on pp. 18, 97).

[Papademetris, 2001] Xenophon Papademetris, Albert J Sinusas, Donald P Dione, and James S Duncan. “Estimation of 3D left ventricular deformation from echocardiography”. In: *Medical image analysis* 5.1 (2001), pp. 17–28 (cit. on p. 37).

[Pedregosa, 2011] F. Pedregosa, G. Varoquaux, A. Gramfort, et al. “Scikit-learn: Machine Learning in Python”. In: *Journal of Machine Learning Research* 12 (2011), pp. 2825–2830 (cit. on pp. 43, 61).

[Peimankar, 2021] Abdolrahman Peimankar and Sadasivan Puthusserypady. “DENS-ECG: A deep learning approach for ECG signal delineation”. In: *Expert systems with applications* 165 (2021), p. 113911 (cit. on p. 68).

[Pereira, 2016] Hope Pereira and Nivedita Daimiwal. “Analysis of features for myocardial infarction and healthy patients based on wavelet”. In: *2016 Conference on Advances in Signal Processing (CASP)*. IEEE. 2016, pp. 164–169 (cit. on p. 68).

[Puyol-Antón, 2022a] Esther Puyol-Antón, Bram Ruijsink, Baldeep S Sidhu, et al. “AI-Enabled Assessment of Cardiac Systolic and Diastolic Function from Echocardiography”. In: *International Workshop on Advances in Simplifying Medical Ultrasound*. Springer. 2022, pp. 75–85 (cit. on p. 11).

[Puyol-Antón, 2022b] Esther Puyol-Antón, Baldeep S Sidhu, Justin Gould, et al. “A multimodal deep learning model for cardiac resynchronisation therapy response prediction”. In: *Medical Image Analysis* 79 (2022), p. 102465 (cit. on p. 12).

[Qin, 2018] Chen Qin, Wenjia Bai, Jo Schlemper, et al. “Joint learning of motion estimation and segmentation for cardiac MR image sequences”. In: *International Conference on Medical Image Computing and Computer-Assisted Intervention*. Springer. 2018, pp. 472–480 (cit. on p. 37).

[Qin, 2020] Chen Qin, Shuo Wang, Chen Chen, et al. “Biomechanics-informed neural networks for myocardial motion tracking in MRI”. In: *International conference on medical image computing and computer-assisted intervention*. Springer. 2020, pp. 296–306 (cit. on p. 37).

[Raghavendra, 2018] U Raghavendra et al. “Automated technique for coronary artery disease characterization and classification using DD-DTDWT in ultrasound images”. In: *Biomedical Signal Processing and Control* 40 (2018), pp. 324–334 (cit. on p. 56).

[Rahate, 2022] Anil Rahate, Rahee Walambe, Sheela Ramanna, and Ketan Kotecha. “Multimodal co-learning: Challenges, applications with datasets, recent advances and future directions”. In: *Information Fusion* 81 (2022), pp. 203–239 (cit. on p. 12).

[Reynaud, 2021] Hadrien Reynaud et al. “Ultrasound video transformers for cardiac ejection fraction estimation”. In: *Medical Image Computing and Computer Assisted Intervention–MICCAI 2021: 24th International Conference, Strasbourg, France, September 27–October 1, 2021, Proceedings, Part VI 24*. Springer. 2021, pp. 495–505 (cit. on p. 11).

[Rohé, 2017] Marc-Michel Rohé and et al. “SVF-Net: Learning Deformable Image Registration Using Shape Matching”. In: *MICCAI 2017 - the 20th International Conference on Medical Image Computing and Computer Assisted Intervention*. Medical Image Computing and Computer Assisted Intervention – MICCAI 2017. Québec, Canada: Springer International Publishing, Sept. 2017, pp. 266–274 (cit. on p. 102).

[Roth, 2020] Gregory A Roth, George A Mensah, Catherine O Johnson, et al. “Global burden of cardiovascular diseases and risk factors, 1990–2019: update from the GBD 2019 study”. In: *Journal of the American College of Cardiology* 76.25 (2020), pp. 2982–3021 (cit. on p. 3).

[Rueda, 2019] Cristina Rueda, Yolanda Larriba, and Shyamal D Peddada. “Frequency modulated möbius model accurately predicts rhythmic signals in biological and physical sciences”. In: *Scientific Reports* 9.1 (2019), pp. 1–10 (cit. on pp. 72, 73).

[Rueda, 2021] Cristina Rueda, Yolanda Larriba, and Adrian Lamela. “The hidden waves in the ECG uncovered revealing a sound automated interpretation method”. In: *Scientific reports* 11.1 (2021), pp. 1–11 (cit. on pp. 69–71, 81).

[Ryser, 2022] Alain Ryser, Laura Manduchi, Fabian Laumer, et al. “Anomaly Detection in Echocardiograms with Dynamic Variational Trajectory Models”. In: *Machine Learning for Healthcare Conference*. PMLR. 2022, pp. 425–458 (cit. on p. 97).

[Sensoy, 2018] Murat Sensoy, Lance Kaplan, and Melih Kandemir. “Evidential deep learning to quantify classification uncertainty”. In: *Advances in neural information processing systems* 31 (2018) (cit. on p. 87).

[Siarohin, 2019] Aliaksandr Siarohin, Stéphane Lathuilière, et al. “First order motion model for image animation”. In: *Advances in Neural Information Processing Systems* 32 (2019) (cit. on pp. 37–41, 44, 45).

[Smistad, 2020] Erik Smistad et al. “Real-Time Automatic Ejection Fraction and Foreshortening Detection Using Deep Learning”. In: *IEEE Transactions on Ultrasonics, Ferroelectrics, and Frequency Control* 67.12 (2020), pp. 2595–2604 (cit. on pp. 11, 56).

[Sokooti, 2017] Hessam Sokooti and et al. “Nonrigid image registration using multi-scale 3D convolutional neural networks”. English. In: *Medical Image Computing and Computer Assisted Intervention - MICCAI 2017*. Springer, 2017, pp. 232–239 (cit. on p. 102).

[Soto, 2022] Jessica Torres Soto, J Weston Hughes, Pablo Amador Sanchez, et al. “Multimodal deep learning enhances diagnostic precision in left ventricular hypertrophy”. In: *European Heart Journal-Digital Health* 3.3 (2022), pp. 380–389 (cit. on p. 12).

[Stahlschmidt, 2022] Sören Richard Stahlschmidt, Benjamin Ulfenborg, and Jane Synnergren. “Multimodal deep learning for biomedical data fusion: a review”. In: *Briefings in Bioinformatics* 23.2 (2022), bbab569 (cit. on p. 12).

[Staring, 2007] Marius Staring, Stefan Klein, and Josien PW Pluim. “A rigidity penalty term for nonrigid registration”. In: *Medical physics* 34.11 (2007), pp. 4098–4108 (cit. on pp. 37, 56).

[Støylen, 2018] Asbjørn Støylen et al. “Relation between mitral annular plane systolic excursion and global longitudinal strain in normal subjects: the HUNT study”. In: *Echocardiography* 35.5 (2018), pp. 603–610 (cit. on p. 59).

[Støylen, 2019] Asbjørn Støylen, Harald Edvard Mølmen, and Håvard Dalen. "Left ventricular global strains by linear measurements in three dimensions: Interrelations and relations to age, gender and body size in the HUNT Study". In: *Open Heart* 6.2 (2019), pp. 1–9 (cit. on pp. 19, 24, 25).

[Ta, 2020] Kevinminh Ta, Shawn S Ahn, Allen Lu, et al. "A semi-supervised joint learning approach to left ventricular segmentation and motion tracking in echocardiography". In: *2020 IEEE 17th International Symposium on Biomedical Imaging (ISBI)*. IEEE. 2020, pp. 1734–1737 (cit. on pp. 11, 37).

[Tanner, 2018] Christine Tanner and et al. *Generative Adversarial Networks for MR-CT Deformable Image Registration*. 2018. arXiv: 1807.07349 [cs.CV] (cit. on p. 103).

[Teplitzky, 2020] Benjamin A Teplitzky, Michael McRoberts, and Hamid Ghanbari. "Deep learning for comprehensive ECG annotation". In: *Heart rhythm* 17.5 (2020), pp. 881–888 (cit. on p. 11).

[Thygesen, 2018] Kristian Thygesen, Joseph S Alpert, Allan S Jaffe, et al. "Fourth universal definition of myocardial infarction (2018)". In: *European Heart Journal* 40.3 (Aug. 2018), pp. 237–269 (cit. on pp. 10, 68).

[Tolstikhin, 2021] Ilya O Tolstikhin, Neil Houlsby, Alexander Kolesnikov, et al. "Mlp-mixer: An all-mlp architecture for vision". In: *Advances in neural information processing systems* 34 (2021), pp. 24261–24272 (cit. on p. 103).

[Tsao, 2023] Connie W Tsao, Aaron W Aday, Zaid I Almarzooq, et al. "Heart disease and stroke statistics—2023 update: a report from the American Heart Association". In: *Circulation* 147.8 (2023), e93–e621 (cit. on p. 9).

[Van Der Malsburg, 1986] C. Van Der Malsburg. "Frank Rosenblatt: Principles of Neurodynamics: Perceptrons and the Theory of Brain Mechanisms". In: *Brain Theory*. Ed. by Gunther Palm and Ad Aertsen. Berlin, Heidelberg: Springer Berlin Heidelberg, 1986, pp. 245–248 (cit. on p. 103).

[Vercauteren, 2007] Tom Vercauteren and et al. "Non-parametric Diffeomorphic Image Registration with the Demons Algorithm". In: *Medical Image Computing and Computer-Assisted Intervention – MICCAI 2007*. Berlin, Heidelberg: Springer Berlin Heidelberg, 2007, pp. 319–326 (cit. on p. 102).

[Vercauteren, 2008] Tom Vercauteren et al. "Symmetric Log-Domain Diffeomorphic Registration : A Demons-based Approach". In: *Medical Image Computing and Computer-Assisted Intervention – MICCAI 2008* 11 Pt 1 (2008), pp. 754–761 (cit. on pp. 11, 102).

[Wagner, 2020] Patrick Wagner, Nils Strodthoff, Ralf-Dieter Bousseljot, et al. "PTB-XL, a large publicly available electrocardiography dataset". In: *Scientific data* 7.1 (2020), pp. 1–15 (cit. on p. 72).

[Waldman, 2022] Carly Waldman, Robert Tickes, Megan Pelter, et al. “UTILITY OF A MACHINE LEARNING ALGORITHM TO INCREASE PHYSICIAN TRAINEE CONFIDENCE IN AND USAGE OF POINT-OF-CARE ECHOCARDIOGRAPHY”. In: *Journal of the American College of Cardiology* 79.9_Supplement (2022), pp. 1824–1824 (cit. on p. 4).

[Wang, 2019] X. Wang, X. Yang, and et al. “Joint Segmentation and Landmark Localization of Fetal Femur in Ultrasound Volumes”. In: *2019 IEEE EMBS International Conference on Biomedical Health Informatics (BHI)*. 2019, pp. 1–5 (cit. on p. 21).

[Wang, 2022] Zihao Wang, Yingyu Yang, Maxime Sermesant, and Hervé Delingette. “Unsupervised Echocardiography Registration through Patch-based MLPs and Transformers”. In: *International Workshop on Statistical Atlases and Computational Models of the Heart*. Springer. 2022, pp. 168–178 (cit. on pp. 11, 15, 101).

[Wei, 2020] Hongrong Wei et al. “Temporal-consistent segmentation of echocardiography with co-learning from appearance and shape”. In: *Medical Image Computing and Computer Assisted Intervention–MICCAI 2020: 23rd International Conference, Lima, Peru, October 4–8, 2020, Proceedings, Part II 23*. Springer. 2020, pp. 623–632 (cit. on pp. 18, 32, 45, 97).

[Witvliet, 2021] M Patrick Witvliet, Evert PM Karregat, Jelle CL Himmelreich, et al. “Usefulness, pitfalls and interpretation of handheld single-lead electrocardiograms”. In: *Journal of Electrocardiology* 66 (2021), pp. 33–37 (cit. on pp. 9, 97).

[Wu, 2016] G. Wu and et al. “Scalable High-Performance Image Registration Framework by Unsupervised Deep Feature Representations Learning”. In: *IEEE Transactions on Biomedical Engineering* 63.7 (2016), pp. 1505–1516 (cit. on p. 102).

[Xiong, 2022] Ping Xiong, Simon Ming-Yuen Lee, and Ging Chan. “Deep Learning for Detecting and Locating Myocardial Infarction by Electrocardiogram: A Literature Review”. In: *Frontiers in Cardiovascular Medicine* 9 (2022) (cit. on p. 69).

[Xu, 2022] Zhuoyang Xu, Yangming Guo, Tingting Zhao, et al. “Abnormality classification from electrocardiograms with various lead combinations”. In: *Physiological Measurement* 43.7 (2022), p. 074002 (cit. on p. 12).

[Yager, 2008] Ronald R Yager and Liping Liu. *Classic works of the Dempster-Shafer theory of belief functions*. Vol. 219. Springer, 2008 (cit. on p. 87).

[Yang, 2017] Xiao Yang, Roland Kwitt, Martin Styner, and Marc Niethammer. “Quicksilver: Fast predictive image registration – A deep learning approach”. In: *NeuroImage* 158 (2017), pp. 378–396 (cit. on p. 102).

[Yang, 2021] Yingyu Yang and Maxime Sermesant. “Shape constraints in deep learning for robust 2D echocardiography analysis”. In: *International Conference on Functional Imaging and Modeling of the Heart*. Springer. 2021, pp. 22–34 (cit. on pp. 14, 18).

[Yang, 2022] Yingyu Yang, Marie Rocher, Pamela Moceri, and Maxime Sermesant. "Explainable Electrocardiogram Analysis with Wave Decomposition: Application to Myocardial Infarction Detection". In: *International Workshop on Statistical Atlases and Computational Models of the Heart*. Springer. 2022, pp. 221–232 (cit. on pp. 14, 68, 92).

[Yang, 2023a] Yingyu Yang, Marie Rocher, Pamela Moceri, and Maxime Sermesant. "Shape and Motion Priors for Generalisable Echocardiography Analysis using Deep Learning". In: (2023) (cit. on pp. 14, 18, 36, 55).

[Yang, 2023b] Yingyu Yang and Maxime Sermesant. "Unsupervised Polyaffine Transformation Learning for Echocardiography Motion Estimation". In: *International Conference on Functional Imaging and Modeling of the Heart*. Springer. 2023, pp. 384–393 (cit. on pp. 14, 36).

[Yezzi, 2003] Anthony J Yezzi and Jerry L Prince. "An Eulerian PDE approach for computing tissue thickness". In: *IEEE transactions on medical imaging* 22.10 (2003), pp. 1332–1339 (cit. on p. 59).

[Yu, 2022] Zhaocheng Yu, Junxin Chen, Yu Liu, et al. "Ddcnn: A deep learning model for af detection from a single-lead short ECG signal". In: *IEEE journal of biomedical and health informatics* 26.10 (2022), pp. 4987–4995 (cit. on p. 12).

[Zadeh, 2017] Amir Zadeh, Minghai Chen, Soujanya Poria, Erik Cambria, and Louis-Philippe Morency. "Tensor Fusion Network for Multimodal Sentiment Analysis". In: *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*. Copenhagen, Denmark: Association for Computational Linguistics, Sept. 2017, pp. 1103–1114 (cit. on p. 86).

[Zerveas, 2021] George Zerveas, Srideepika Jayaraman, Dhaval Patel, Anuradha Bhamidipaty, and Carsten Eickhoff. "A transformer-based framework for multivariate time series representation learning". In: *Proceedings of the 27th ACM SIGKDD conference on knowledge discovery & data mining*. 2021, pp. 2114–2124 (cit. on p. 98).

[Zewdie, 2014] Getie Zewdie and Momiao Xiong. "Fully automated myocardial infarction classification using ordinary differential equations". In: *arXiv preprint arXiv:1410.6984* (2014) (cit. on p. 69).

[Zhang, 2020] Ling Zhang et al. "Generalizing Deep Learning for Medical Image Segmentation to Unseen Domains via Deep Stacked Transformation". In: *IEEE Transactions on Medical Imaging* 39.7 (2020), pp. 2531–2540 (cit. on p. 25).

[Zhang, 2021] Yungeng Zhang, Yuru Pei, and Hongbin Zha. "Learning dual transformer network for diffeomorphic registration". In: *International Conference on Medical Image Computing and Computer-Assisted Intervention*. Springer. 2021, pp. 129–138 (cit. on p. 103).

[Zhang, 2022] Xiaoran Zhang, Chenyu You, et al. "Learning Correspondences of Cardiac Motion from Images Using Biomechanics-Informed Modeling". In: *Statistical Atlases and Computational Models of the Heart*. Cham: Springer Nature Switzerland, 2022, pp. 13–25 (cit. on p. 37).

[Zheng, 2021] Yuanjie Zheng and et al. “SymReg-GAN: Symmetric Image Registration with Generative Adversarial Networks”. In: *IEEE Transactions on Pattern Analysis and Machine Intelligence* (2021), pp. 1–1 (cit. on p. 103).

List of Figures

List of Tables