Deep learning motion estimation in ultrasound imaging : Application to myocardial deformation quantification

Ewan Evain

N°d'ordre NNT : 2022LYSEI037

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

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

Spécialité de doctorat :
Traitement du signal et des images

Soutenue publiquement le 29/04/2022, par :
Ewan EVAIN

Deep learning motion estimation in ultrasound imaging: application to myocardial deformation quantification

---

Apprentissage profond pour l'estimation de mouvement en imagerie ultrasonore: application à la quantification des déformations du myocarde

---

Devant le jury composé de :

Petitjean, Caroline	Professeure	Université de Rouen	Rapporteure
Thome, Nicolas	Professeur	CNAM	Rapporteur
Bloch, Isabelle	Professeure	Telecom Paris	Examinatrice
De Craene, Mathieu	Ingénieur de recherche	Philips	Examinateur
Van Sloun, Ruud	Associated Professor	Université d'Eindhoven	Examinateur
Bernard, Olivier	Professeur	INSA de Lyon	Directeur de thèse

Département FEDORA – INSA Lyon - Ecoles Doctorales

SIGLE	ECOLE DOCTORALE	NOM ET COORDONNEES DU RESPONSABLE
CHIMIE	**CHIMIE DE LYON** **https://www.edchimie-lyon.fr** Sec. : Renée EL MELHEM Bât. Blaise PASCAL, 3e étage secretariat@edchimie-lyon.fr	**M. Stéphane DANIELE** C2P2-CPE LYON-UMR 5265 Bâtiment F308, BP 2077 43 Boulevard du 11 novembre 1918 69616 Villeurbanne **directeur@edchimie-lyon.fr**
E.E.A.	**ÉLECTRONIQUE, ÉLECTROTECHNIQUE, AUTOMATIQUE** **https://edeea.universite-lyon.fr** Sec. : Stéphanie CAUVIN Bâtiment Direction INSA Lyon Tél : 04.72.43.71.70 secretariat.edeea@insa-lyon.fr	**M. Philippe DELACHARTRE** INSA LYON Laboratoire CREATIS Bâtiment Blaise Pascal, 7 avenue Jean Capelle 69621 Villeurbanne CEDEX Tél : 04.72.43.88.63 **philippe.delachartre@insa-lyon.fr**
E2M2	**ÉVOLUTION, ÉCOSYSTÈME, MICROBIOLOGIE, MODÉLISATION** **http://e2m2.universite-lyon.fr** Sec. : Sylvie ROBERJOT Bât. Atrium, UCB Lyon 1 Tél : 04.72.44.83.62 secretariat.e2m2@univ-lyon1.fr	**M. Philippe NORMAND** Université Claude Bernard Lyon 1 UMR 5557 Lab. d'Ecologie Microbienne Bâtiment Mendel 43, boulevard du 11 Novembre 1918 69 622 Villeurbanne CEDEX **philippe.normand@univ-lyon1.fr**
EDISS	**INTERDISCIPLINAIRE SCIENCES-SANTÉ** **http://ediss.universite-lyon.fr** Sec. : Sylvie ROBERJOT Bât. Atrium, UCB Lyon 1 Tél : 04.72.44.83.62 secretariat.ediss@univ-lyon1.fr	**Mme Sylvie RICARD-BLUM** Institut de Chimie et Biochimie Moléculaires et Supramoléculaires (ICBMS) - UMR 5246 CNRS - Université Lyon 1 Bâtiment Raulin - 2ème étage Nord 43 Boulevard du 11 novembre 1918 69622 Villeurbanne Cedex Tél : +33(0)4 72 44 82 32 **sylvie.ricard-blum@univ-lyon1.fr**
INFOMATHS	**INFORMATIQUE ET MATHÉMATIQUES** **http://edinfomaths.universite-lyon.fr** Sec. : Renée EL MELHEM Bât. Blaise PASCAL, 3e étage Tél : 04.72.43.80.46 infomaths@univ-lyon1.fr	**M. Hamamache KHEDDOUCI** Université Claude Bernard Lyon 1 Bât. Nautibus 43, Boulevard du 11 novembre 1918 69 622 Villeurbanne Cedex France Tél : 04.72.44.83.69 **hamamache.kheddouci@univ-lyon1.fr**
Matériaux	**MATÉRIAUX DE LYON** **http://ed34.universite-lyon.fr** Sec. : Yann DE ORDENANA Tél : 04.72.18.62.44 yann.de-ordenana@ec-lyon.fr	**M. Stéphane BENAYOUN** Ecole Centrale de Lyon Laboratoire LTDS 36 avenue Guy de Collongue 69134 Ecully CEDEX Tél : 04.72.18.64.37 **stephane.benayoun@ec-lyon.fr**
MEGA	**MÉCANIQUE, ÉNERGÉTIQUE, GÉNIE CIVIL, ACOUSTIQUE** **http://edmega.universite-lyon.fr** Sec. : Stéphanie CAUVIN Tél : 04.72.43.71.70 Bâtiment Direction INSA Lyon mega@insa-lyon.fr	**M. Jocelyn BONJOUR** INSA Lyon Laboratoire CETHIL Bâtiment Sadi-Carnot 9, rue de la Physique 69621 Villeurbanne CEDEX **jocelyn.bonjour@insa-lyon.fr**
ScSo	**ScSo*** **https://edsciencessociales.universite-lyon.fr** Sec. : Mélina FAVETON INSA : J.Y. TOUSSAINT Tél : 04.78.69.77.79 melina.faveton@univ-lyon2.fr	**M. Christian MONTES** Université Lumière Lyon 2 86 Rue Pasteur 69365 Lyon CEDEX 07 **christian.montes@univ-lyon2.fr**

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

Remerciements

Résumé

La modalité d'imagerie la plus utilisée en pratique clinique est actuellement l'imagerie ultrasonore car celle-ci est peu couteuse, rapide et non-invasive. En échocardiographie, plusieurs indices caractérisant la fonction cardiaque peuvent être extraits de ces acquisitions, parmi lesquels la déformation globale longitudinale joue un rôle important dans l'établissement d'un diagnostic. Cependant l'estimation de cet indice souffre d'un manque de reproductibilité du fait des caractéristiques inhérentes à l'imagerie ultrasonore. En effet, les méthodes traditionnelles comme le flux optique ou la correspondance de blocs ne permettent pas de gérer des artéfacts types tels que la décorrélation de texture ultrasonore. Récemment, les approches d'apprentissage profond ont battu les méthodes de l'état de l'art en estimation de mouvement, portées par les applications en robotique ou sur les voitures autonomes. Dans la première partie de cette thèse, nous présentons une étude pilote afin d'évaluer la capacité des méthodes d'apprentissage profond à estimer le mouvement en imagerie ultrasonore malgré les nombreux artéfacts sous-jacents. Pour ce faire, nous avons créé une base de données composée d'images ultrasonores simulées et in-vitro incluant un disque tournant avec des vitesses variables. Dans la seconde partie de cette thèse, nous détaillons le réseau de neurones pyramidal que nous avons développé afin d'estimer la déformation du muscle myocardique et qui améliore de façon significative les performances des méthodes de l'état de l'art. Pour entrainer et évaluer notre méthode d'apprentissage, nous avons également implémenté un pipeline de simulations permettant de générer des séquences d'images échocardiographiques réalistes avec un champ dense du mouvement du muscle myocardique de référence et présentant une grande variabilité anatomique et fonctionnelle.

Abstract

Ultrasound is the most widely used imaging modality in clinical practice because it is fast, non-invasive and less expensive than other modalities. In echocardiography, several metrics characterizing the cardiac function can be extracted from these acquisitions, among which the global longitudinal strain (GLS) plays an important role in establishing a diagnosis. However, the estimation of this index suffers from a lack of reproducibility due to the specific ultrasound's characteristics. Indeed, traditional methods such as optical flow or block matching do not handle typical artifacts such as ultrasound texture decorrelation. Recently, deep learning approaches have beaten state-of-the-art methods in motion estimation, driven by applications to robotics or autonomous cars. In the first part of this thesis, we present a pilot study to evaluate the ability of deep learning methods to estimate motion in ultrasound imaging despite the many underlying artifacts. To do so, we created a database composed of simulated and in-vitro ultrasound images including a rotating disk with varying speeds. In the second part of this thesis, we detail the pyramidal neural network that we have developed to estimate the deformation of the myocardial muscle and that significantly improves the performances of the state-of-the-art methods. To train and evaluate our learning method, we also implemented a simulation pipeline to generate realistic echocardiographic image sequences with a dense reference field and with high anatomical and functional variability.

Contents

List of Figures

List of Tables

List of Abbreviations

2D Two Dimension. 5

3D Three Dimension. 41

A2C Apical Two Chamber. 16

A3C Apical Three Chamber. 16

A4C Apical Four Chamber. 16

A5C Apical Five Chamber. 16

AE Angular Error. 36

AHA American Heart Association. 11

AS Aortic Stenosis. 83

BI-RADS Breast Imaging Reporting and Data System. 132

CNN Convolutional Neural Network. 31

CNSN Convolutional Neural Sub-Network. 79

DL Deep Learning. 4

EM Electromechanical. 45

EPE Endpoint Error. 28

GAN Generative Adversarial Networks. 47

GLS Global Longitudinal Strain. 4

GRU Gated Recurrent Unit. 33

HCM Hypertrophic Cardiomyopathy. 83

HF Heart Failure. 83

IQ In-phase and Quadrature. 12

LVEF Left Ventricle Ejection Fraction. 18

MAD Median Absolute Deviation. 64

MAPSE Mitral Annular Plane Systolic Excursion. 18

MRI Magnetic Resonance Imaging. 46

MUST Matlab Ultrasound Toolbox. 47

NCC Normalized Cross Correlation. v, 24

PIV Particle Image Velocimetry. 24

PLAX Parasternal long-axis. 15

PRF Pulse Repetition Frequency. 15

PSAX Parasternal short-axis. 15

R-CNN Region-based Convolutional Neural Network. 132

RF Radio Frequency. 12

TGC Time Gain Compensation. 13

TTE Transthoracic Echocardiography. 15

Part I

Introduction

Chapter 1

Introduction

Contents

1.1 Motivations

Cardiovascular diseases (CVDs) are the leading cause of death worldwide, representing 32% of deaths in 2019 [1]. The most common risk factors for these diseases (hypercholesterolemia, hypertension and diabetes) are often the result of poor lifestyle habits and are accentuated by them (smoking, sedentary lifestyle, overweight) [2]. Identification of these risk factors allows the organization of prevention policies and regular monitoring of cardiac function to avoid future complications. This monitoring can be done in several ways, including cardiac imaging. For instance, several cardiac diseases impact the myocardial deformation and can be detected or quantified thanks to different imaging techniques in clinical routine.

Ultrasound imaging is the most widely used imaging modality in cardiology because it is fast, non-invasive and less expensive than other modalities [3]. Echocardiography enables the analysis of cardiac function by the extraction of relevant clinical indices such as volumes and myocardial deformation. In this context, one typical index recommended in clinical practice guidelines for assessing myocardial deformation is the global longitudinal deformation (GLS) [4]. GLS is computed from B-mode images acquired in any standard apical view as the percentage of myocardial shortening between the end-diastolic and end-systolic instants [5]. Current solutions of myocardial deformation estimation suffer from measurement reproducibility problems due to the specific ultrasound's characteristics (e.g., artifacts, lack of information and speckle decorrelation). These methods are mainly based on conventional block-matching and optical flow techniques.

Recently, deep learning (DL) methods have become the most efficient approaches in computer vision in most of the domains. Motion estimation is no exception to the rule, with DL solutions obtaining the best results on natural image databases. Among these methods, supervised learning techniques have shown better performances than unsupervised approaches. Thus, it appears that simulated databases are fundamental both for evaluating the performance of the methods and for training neural networks for motion estimation tasks. Indeed, only simulations allow to have a reference motion per pixel to train motion estimators in a supervised way.

1.2 Objectives

Because of the recent improvement of motion estimation performances on natural images, the use of DL methods to solve the problem of cardiac structure tracking and myocardial deformation estimation seems to be a solution of choice to explore. In this context, the main objectives of my thesis are the following:

- Demonstrate the ability of deep neural networks to estimate the motion in ultrasound imaging.
- Develop a robust DL method to estimate the myocardial motion and the corresponding deformation (also named strain) in ultrasound imaging.
- Share with the research community the synthetic open-access databases generated to train DL methods and evaluate their accuracy on ultrasound images.

1.3 Thesis organization

This thesis is divided into four main parts composed of different chapters as follows.

Part 1,

- Chapter 1 introduces the methods, the objectives and the organization of the manuscript.

Part 2 describes the state-of-the-art methods in the main areas covered by this thesis.

- Chapter 2 discusses the basis of echocardiography: the cardiac function and anatomy, the ultrasound imaging formation and type, the classical views in clinical practice and the most used indices to assess the cardiac function in clinical practice.
- Chapter 3 details the motion estimation methods traditionally used in ultrasound imaging and the DL approaches that recently challenged them. A description of the metrics used to evaluate the accuracy of the estimated motion field is also provided.
- Chapter 4 presents the synthetic open-access two-dimensional databases used for training and evaluation of DL methods. Open access databases composed of natural data and cardiac images are both described.

Part 3 outlines the contributions of this thesis.

- Chapter 5 presents the methodology and results of our pilot study investigating the ability of CNNs to estimate displacement in 2D ultrasound imaging.
- Chapter 6 describes the design and the performance of our deep neural network to estimate myocardial motion from B-mode 2D images.

Part 4 concludes the manuscript.

- Chapter 7 summarizes the contributions and discusses the future work.

A series of appendices completes the manuscript in which we find:

- A summary of the thesis in French as requested by the EEA doctoral school.
- A contribution realized in parallel with the subject of my thesis and valued by a publication. This article describes the method we developed to win the JFR 2021 (French Days of Radiology) challenge. This challenge organized by the French Society of Radiology concerned the classification of breast nodules in 2D ultrasound imaging.
- A list of publications produced during my thesis.

Part II

Background

Chapter 2

Echocardiography

Contents

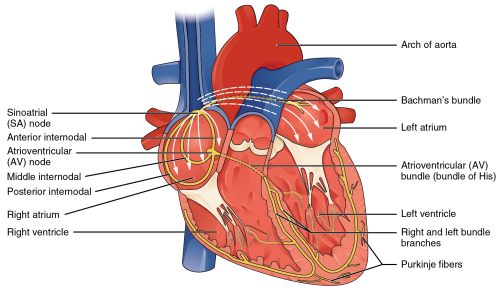
2.1 Introduction

Ultrasound imaging is currently the most widely used imaging modality in clinical practice due to its rapid, non-invasive, and low-cost nature compared to other modalities. These advantages also make it the most widely used modality in cardiology for both flow and tissue analysis.

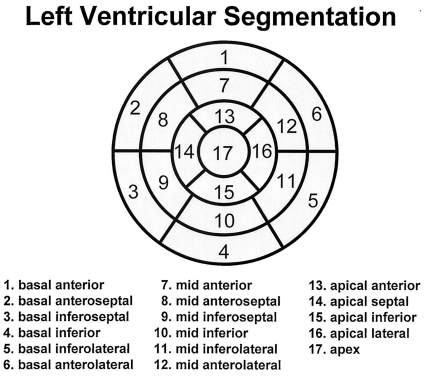
In this chapter, we will first introduce the anatomy and function of the heart. We will then describe the principle of image formation in ultrasound, and the different types of acquisition that are performed during cardiac examinations. Finally, we will introduce the main indices computed from the 2D B-mode acquisitions in clinical practice.

2.2 Cardiac function and anatomy

The heart is a muscle that pumps blood to the rest of the human body through the circulatory system. It comprises two parts separated by the septum: the right and left side of the heart (see Figure 2.1). These two parts include an atrium towards the base and a ventricle towards the apex separated by valves, mitral for the left heart, and tricuspid for the right heart. The right side of the heart collects deoxygenated blood and sends it to the lungs for reoxygenation through the pulmonary artery, while the left side of the heart collects oxygenated blood and sends it through the aorta to the general circulation to perfuse all organs. To avoid backflow, two valves are located at the outlet of the ventricles: the pulmonary valve and the aortic valve, between the right ventricle and the pulmonary artery, and between the left ventricle and the aorta, respectively. The heart is supplied by the right and left coronary arteries, the left coronary artery dividing itself into a circumflex branch and an anterior interventricular branch.



(a) Anatomy of the heart from the anterior view of the frontal section. Modified from [6].



(b) Nomenclature of the 17 myocardial segments of the left ventricle in bull-eye view [7].

Figure 2.1 – Anatomy of the heart (2.1a) and nomenclature of the different segments (2.1b)

The heart muscle is called the myocardium. It is surrounded by two layers: the endocardium to the chambers layer and the epicardium to the pericardium layer. It contracts to push blood out from the ventricles during systole and fills during diastole.

These movements are in synergy with the opening and closing of the valves in order to increase the pressure in the cavities and thus push blood out more efficiently.

The rhythm of muscular contractions and relaxations of the cardiac cycle, called sinus rhythm, is imposed by the sinoatrial node, which generates electrical pulses that are responsible for the myocardial contraction. These electrical signals travels to the atrioventricular node, the bundle of His, and finally to the Purkinje fibers, which transmit the electrical impulse to the myocardium. Most of the physiological variations during the cardiac cycle are illustrated in Figure 2.2.

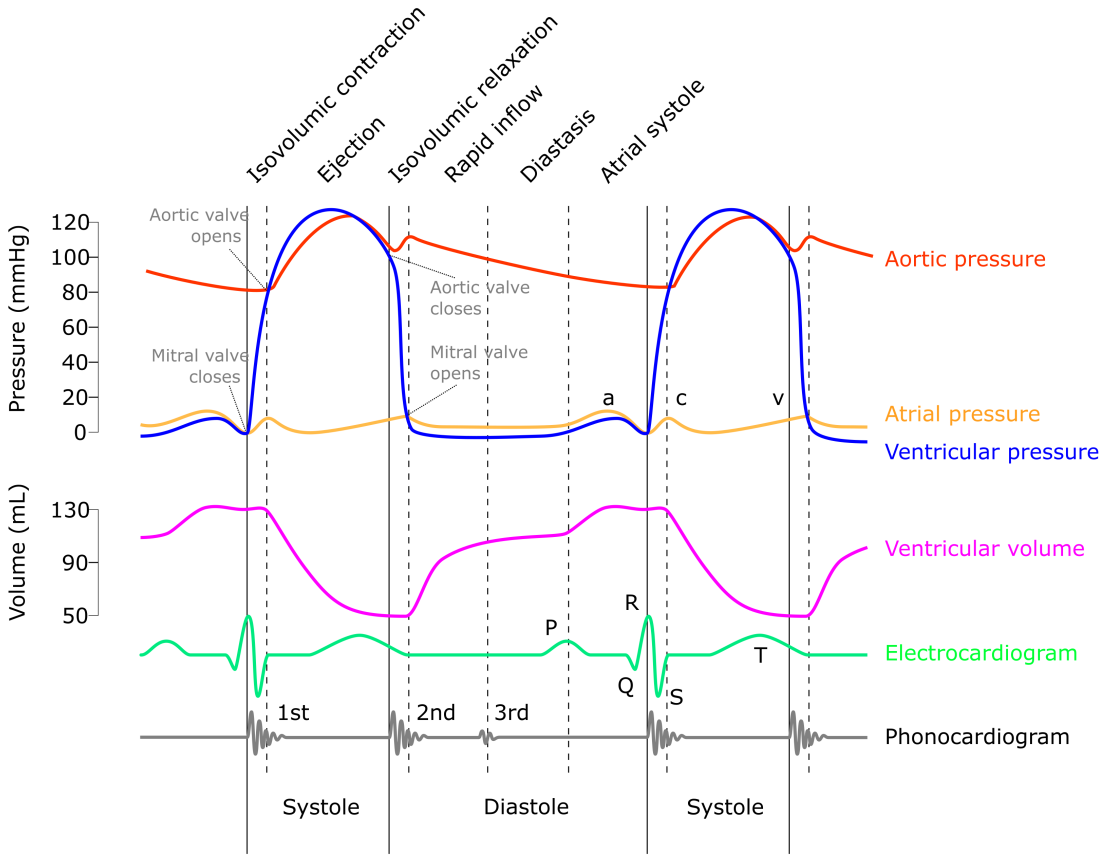


Figure 2.2 – Wiggers diagram showing the variations of physiological values during the cardiac cycle [8].

The myocardium in the left ventricle was mapped as segments classified according to their location. A total of 17 segments were defined by the American Heart Association (AHA) [7] according to their positioning on the long axis of the ventricle (basal, mid, and apical) and on the circumferential location (see Figure 2.1b). This nomenclature standardizes the description of the myocardium and assigns individual segments to specific coronary artery territories.

2.3 Ultrasound Imaging

2.3.1 Ultrasound image formation

2.3.1.1 From wave to image

Ultrasound waves are mechanical waves between 16 kHz and 1 GHz. In cardiac imaging, the ultrasounds used have a frequency between 2.5 MHz and 5 MHz. Ultrasound waves are generated at the level of the probe by piezoelectric elements that transform an electrical signal into mechanical waves and vice versa.

In transmission, the piezoelectric crystals are activated by a sinusoidal electrical signal and generate ultrasonic waves. An apodization step is applied in transmission/reception and corresponds to the multiplication of the electrical signals by some coefficients to reduce the side lobe artifacts. Delays are applied on the piezoelectric crystals according to the targeted location to focus the incident waves on a precise point. In reception, the piezoelectric crystals convert the backpropagated echoes into electrical signals (Figure 2.3).

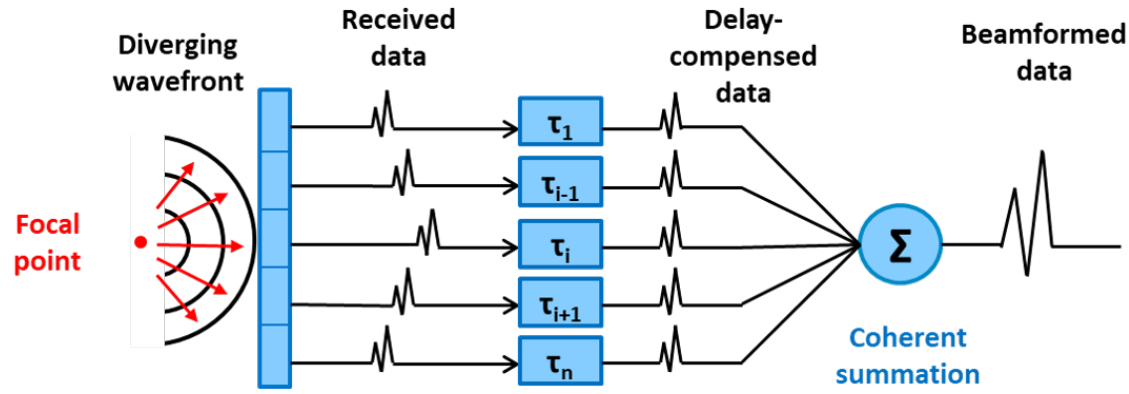


Figure 2.3 – Schematic representation of beamforming in ultrasound imaging.

Standard delay and sum techniques are finally applied to create beamformed radiofrequency (RF) signals. These signals are demodulated for storage purposes and are referred to as In-phase/Quadrature (IQ) signals. The corresponding values are complex. To compensate from a large dynamic, the envelope of IQ signals is finally log-compressed to create the intensity of the final image named as B-mode image. B-mode images are acquired in polar coordinate system due to the acquisition process. A scan conversion is thus applied to express the images into the cartesian coordinate system for the final display on the equipment screens.

2.3.1.2 Wave interactions in the tissues

During the propagation of the ultrasound waves in the insonified tissues, different phenomena happened depending on the properties of the media (Figure 2.4). Indeed, the speed of wave propagation, the acoustic impedance and the absorption coefficients of the media can vary greatly and cause physical phenomena that have a significant impact on the beamformed image. The values of the speed of sound and the acoustic impedance of different media are provided as examples in Table 2.1.

Medium	Speed ($m.s^{-1}$)	Acoustic impedance ($Mrayls$)
Air	330	0.0004
Blood	1500	1.50
Bone	3500	7.80
Fat	1450	1.38
Muscle	1580	1.71

Table 2.1 – Speed of sound and acoustic impedance in several media

In addition, during propagation in soft tissue, ultrasound waves undergo attenuation proportional to the distance traveled. In ultrasound imaging, attenuation is compensated by an increasing amplification of the received signal with depth called temporal gain compensation (TGC).

When passing from one medium to another, part of the wave is transmitted, and the other part is reflected, generating a backscattered echo. The amount of reflected wave is proportional to the reflection coefficient which depends on the difference in acoustic impedance between the two media. The impedance is a function of the speed of wave propagation in the medium and the density of the medium. If the size of the elements causing the reflection of the wave are smaller than the wavelength of the emitted wave, then a scattering situation happens and the elements are named as scatterers.

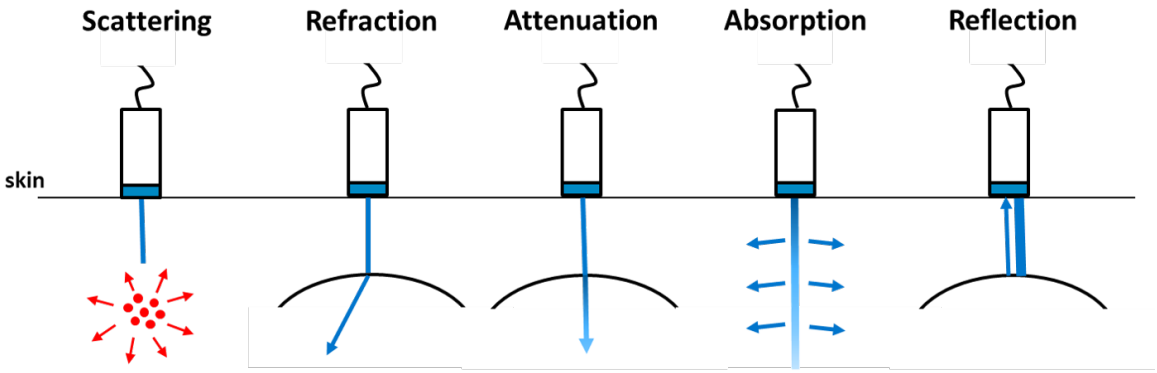
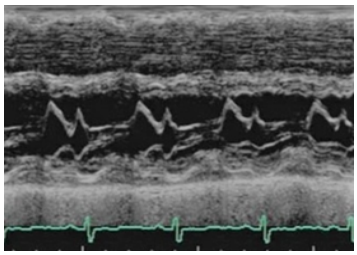


Figure 2.4 – Illustration of the different types of ultrasound interactions in tissues.

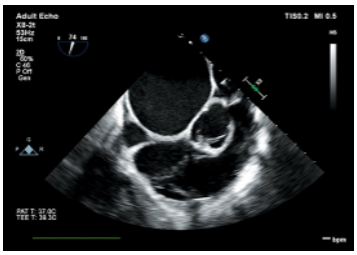
The combination of scattering phenomenon mixed with random constructive and destructive interferences creates speckle pattern. The size and shape of the speckle are directly linked to the number and relative positions of the scatterers per resolution cell. The resulting pattern can be seen as a local signature of the tissue and its shape is tracked over time by traditional algorithms to perform motion estimation.

2.3.2 Different types of ultrasound imaging

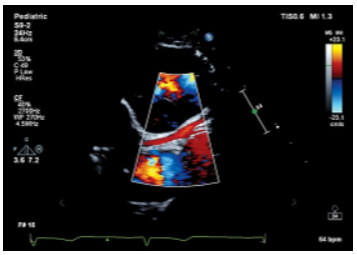
The study of the left ventricle in clinical practice relies mainly on three different types of ultrasound imaging which are: the Motion mode (M-mode), the Brightness Mode (B-mode), and the Doppler.



(a) M-Mode imaging of a mitral valve [9].



(b) B-mode imaging of the heart [10].



(c) Color Doppler imaging [10].

Figure 2.5 – Main types of ultrasound imaging in echocardiography.

2.3.2.1 M-Mode

This mode is the primary form of ultrasound imaging. M-mode images record the movement of structures located on a single acquisition line over time. Because of its high temporal and axial resolution, this mode is mainly used to measure high-velocity structures such as valves. It can also be used to measure some dimensions of cardiac cavities and thickness. Finally, the M-mode acquisition line can be superimposed on the current B-mode image to position it correctly. An example of M-mode image is given in Figure 2.5-a.

2.3.2.2 B-Mode

B-mode is the most commonly used imaging mode in ultrasound. The pixel intensity is represented in grayscale with a value proportional to the amplitude of the returned echoes. Several acoustic waves are transmitted to reconstruct a sectorial image and allows the visualization of the heart anatomy. This real-time imaging can be used to measure the myocardial deformation and is the one we will input to our deep learning solution to estimate the myocardial motion (Chapters 5 and 6). An illustration of a B-mode image is given in Figure 2.5-b.

2.3.2.3 Doppler

Doppler imaging enables blood flow and cardiac tissue displacement characterization. This modality is based on the Doppler effect, *i.e.* the frequency difference of a wave when a relative displacement occurs between its transmitter and its receiver.

In such situation, the reflected frequency (f_r) is different from that emitted (f_o) and increases if the source or the observer gets closer and decreases otherwise. The same principle is used in ultrasound where the difference between the frequency of the received wave and that of the transmitted wave is known as the Doppler shift frequency (Δf) and is defined as follows:

$$\Delta f = f_r - f_o = \frac{2 \times f_o \times V \times cos(\theta)}{C}, \tag{2.1}$$

with V the blood velocity, θ the angle between the flow and the propagation direction of the ultrasonic wave, and C the speed of sound in soft tissue. This imaging technique allows the visualization and quantification of the flows by the reflection of the transmitted waves on the moving red blood cells. Doppler frequency shift can also

be used to study myocardial wall motion in tissue Doppler. There are two types of acquisition in Doppler imaging:

- Continuous-wave Doppler: Ultrasonic waves are continuously emitted by a crystal while another crystal collects the reflected waves. This mode has no measured velocity limit and an excellent velocity resolution. However, it does not allow the spatial localization of the measured velocities along the acquisition direction due to the continuous emission of ultrasonic waves.

- Pulsed wave Doppler: The crystals of the probe alternate between wave emission and reception of the backscattered signals. Pulses are emitted in transmission and the time between two pulses corresponds to the pulse repetition frequency (PRF). This mode has a good spatial resolution but suffers from aliasing when the PRF is less than twice the maximum Doppler frequency that can be measured.

Pulsed wave Doppler mode allows the visualization of the projection of the actual blood flow velocity along the direction of the transmitted wave on a large scan sector. This mode is named color Doppler. The section plane is scanned in the same way as in B-mode imaging. On each of the acquisition lines, the estimated Doppler velocity is represented by a color coded value, which is red if the measured motion is towards the probe, and blue if the motion is away from the probe. An illustration of a color Doppler image is given in Figure 2.5-c.

2.3.3 Views

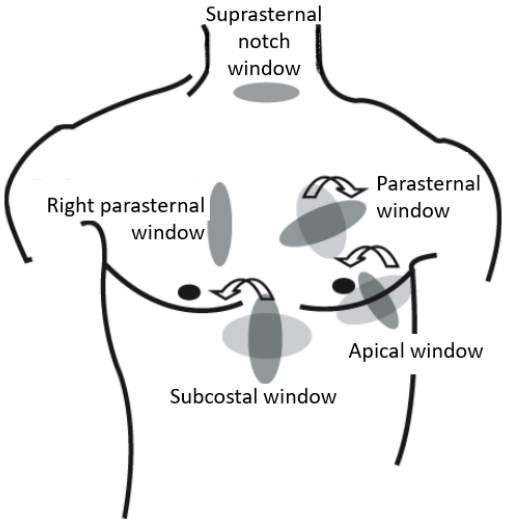
Because the clinicians observe different heart structures depending on the position and the orientation of the probe, the acquisitions in echocardiography have been standardized for the analysis of cardiac function in clinical practice, as shown in Figure 2.6. In this section, I will describe the main views used in echocardiography, especially the parasternal and apical views in transthoracic echocardiography (TTE) [11].

2.3.3.1 Parasternal

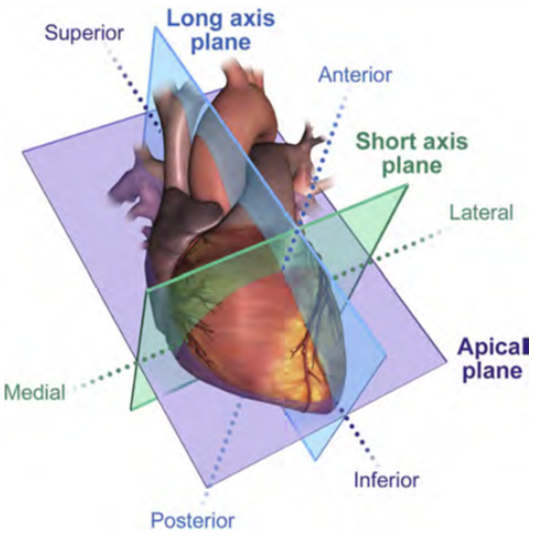
A shown in Figure 2.6b, the long axis and short-axis views are the two standards parasternal acquisitions.

- Parasternal long-axis view (PLAX): This view allows the assessment of the left ventricle (anterosternal and inferolateral walls), the left atrium (anterior and posterior walls), the right ventricle, the ascending aorta, and three valves: mitral, tricuspid, and pulmonary.

- Parasternal short-axis view (PSAX): There are several PSAX views corresponding to different plane positions from the apex to base. This allows the visualization of different structures, among which the basal and middle segments of all walls of the left ventricle, the lateral and septal walls of the left atrium, the right ventricle, the posterior and lateral walls of the rights atrium. The pulmonary artery and the aorta are also visible according to the section, as well as the septa and the tricuspid, mitral and pulmonary valves.



(a) Image acquisition windows in echocardiography.



(b) Scanning planes of the heart [12].

Figure 2.6 – Standard acquisition views in echocardiography.

2.3.3.2 Apical

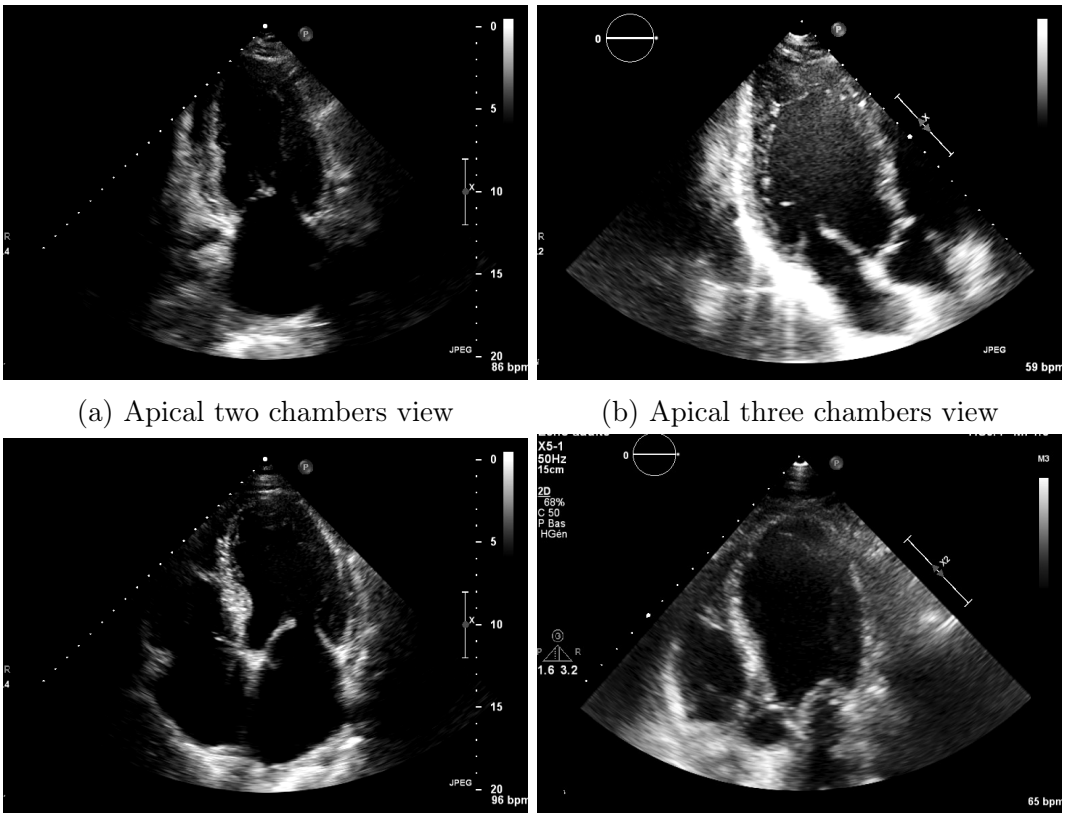
There exists four apical views that are used both in B-mode and Doppler. They allows the assessment of the heart anatomy and function, such as the analysis of the myocardial strain, the measurement of the volume of the left ventricle as well as the regurgitation of the mitral valve.

- Apical 2 Chambers (A2C): This view is centered on the left heart so that we can observe the apex, the anterior and inferior walls of the left ventricle, the anterior and posterior walls of the left atrium and the mitral valve between them (see Figure 2.7a).

- Apical 3 Chambers (A3C): In this view, the probe's position is close to that used in A2C, which explains why the same structures are seen with the addition of the aorta (see Figure 2.7b).

- Apical 4 Chambers (A4C): In this view, the four chambers of the heart are seen (see Figure 2.7c). More particularly, the apex, anterolateral and anteroseptal walls of the left ventricle, the lateral and septal walls of the left atrium, the anterior wall of the right ventricle, and the lateral walls are viewed. In addition, the interventricular and interatrial septa and the mitral and tricuspid valves are visible.

- Apical 5 Chambers (A5C): This view shows the four chambers of the heart and the aorta (see Figure 2.7d). The lateral and septal walls as well as the apex of the left ventricle are visible. The lateral and septal walls of the left atrium can also be seen. For the right heart, only the atrial wall of the right ventricle and the lateral and septal walls of the right atrium can be viewed.



(a) Apical two chambers view (b) Apical three chambers view

(c) Apical four chambers view (d) Apical five chambers view

Figure 2.7 – Examples of B-mode images acquired in the four apical views.

2.3.4 Clinical Indices

Several clinical indices extracted from the images are used in echocardiography. We will focus in this section on some important indices for the study of the left ventricle from B-mode ultrasound imaging, namely the ejection fraction and the myocardial deformation.

2.3.4.1 Ejection fraction of the left ventricle

The volume of the left ventricle and the resulting ejection fraction are important information estimated by the cardiologist during clinical exams. The guideline recommends to compute the volume of the cavity by the modified Simpson's method [13]. This calculation must be made from the two orthogonal A2C and A4C views, from which the endocardium is segmented. The ventricular cavity is divided along the long axis into a series of 20 equally sized disks (see Figure 2.8). The sum of the volumes of these disks corresponds to the estimated volume of the left ventricle. Equation 2.2 is the final formula used to compute the volume according to the two views with a corresponding to the diameter of disk i in A4C, b to the diameter of disk i in A2C, and L to the length of the long axis of the ventricle.

$$V = \frac{\pi}{4} \sum_{i=1}^{20} a_i \times b_i \times \frac{L}{20} \tag{2.2}$$

This computation must be done at the diastole (EDV) and systole (ESV) instants.

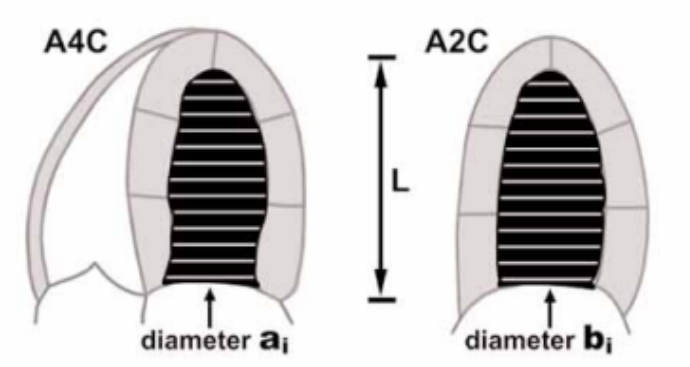


Figure 2.8 – Modified Simpson's method applied to the left ventricular volume estimation in 2D B-mode ultrasound imaging [14].

The difference between the two corresponding volumes is then used to compute the ejection fraction of the left ventricle (LVEF) from the following equation:

$$LVEF(\%) = \frac{EDV - ESV}{EDV} \times 100 \tag{2.3}$$

This value is used in the classification and treatment of heart failures [15]. If LVEF is greater than 50%, the heart is functioning normally. Severe heart failure is diagnosed if the LVEF is less than 40%. Heart failure is referred to as moderate when the corresponding LVEF is between 40% and 50%.

2.3.4.2 Myocardial strain

Several clinical metrics in clinical practice are derived from left ventricular strain computed using the B-mode image, among them the global longitudinal strain (GLS) and the regional strain are the most well known [16]. In the GLS metric, the longitudinal ventricular length L is extracted at each timestep t either from a contour in the middle of the myocardium or from the endocardial contour. The different lengths $L(t)$ are used to calculate the Lagrangian strain [17], defined as:

$$S_L = \frac{L(t) - L(t_0)}{L(t_0)}, \tag{2.4}$$

where t_0 corresponds to end-diastole time. When the GLS is computed, the full endocardial/mid-myocardial contour is taken into account to estimate $L(t)$. In the regional strain estimation, the contours are divided into several regions (see Figure 2.9) and a local strain per region is finally computed.

Regional strain can be interesting to detect local contraction troubles on the different heart segments (Figure 2.1b) but suffers from the lack of reproducibility of the measurements compared to GLS [4], [18]. Thanks to its relative reproducibility, GLS is now part of the recommended metrics to perform during clinical exams. It is worth noting that the GLS computed from the endocardial contours takes into account part of the radial motion in addition to the longitudinal strain [19].

Another metric was therefore developed to take into consideration only the strictly longitudinal deformation: the Mitral Annular Plane Systolic Excursion (MAPSE) [20]. This index is computed in the same way as the GLS, with the difference that L denotes

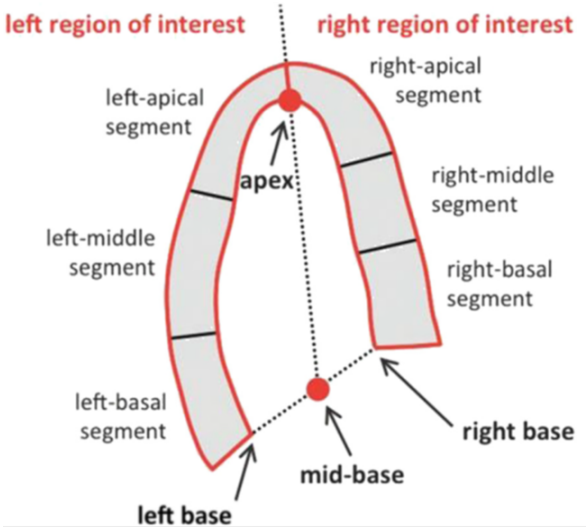


Figure 2.9 – Representation of the myocardium in apical view [17].

the distance between the apex and the mid-basal point of the endocardial border (Figure 2.9).

2.4 Conclusions

We have seen in this chapter that the anatomy and function of the heart are monitored in clinical practice by using ultrasound imaging with different modes. In B-mode imaging, several indices are used to analyze and quantify cardiac function. Even if some metrics such as the ejection fraction or the GLS are considered robust enough to be part of the clinical guidelines, most of them still suffer from a lack of reproducibility due to the noisy nature of ultrasound images and the presence of numerous artifacts. In this thesis, we intend to take a step forward in the estimation of the GLS by developing a dedicated deep learning solution that will robustly and efficiently estimate tissue motion from B-mode sequences. These displacement fields will then be use to propagate the endocardial contour at end-diastole to the rest of the sequence, allowing the computation of the GLS from Equation 2.4.

Chapter 3

Motion estimation

Contents

3.1 Introduction

As discussed in the previous chapter, the estimation of myocardial motion is an important component in the characterization of cardiac function. Indeed, several relevant clinical indices can be computed from the estimated motion, among which the myocardial global longitudinal strain is part of the recommendations during clinical examinations while the regional strain is currently not used because of a lack of reproducibility. In echocardiography, the motion is classically estimated through speckle tracking algorithms. These methods are based on the hypothesis that the local tissue properties that result in speckle patterns should be consistent between two consecutive frames (see Section 2.3.1.2 for more details on the origin of the speckle).

These methods were first developed from other applications in the literature and then adapted to ultrasound modality. Among the most known traditional methods, we can cite the block matching and optical flow techniques. Deep learning approaches have recently been studied in this area, with significant success for natural scene images, as illustrated in Figure 3.2.

In this chapter, I will first describe the main traditional methods developed for motion estimation in echocardiography. I will then provide an in-depth overview of deep learning architectures developed for motion estimation, as well as an explanation of their evolution over the past few years. Finally, I will analyze the application of these techniques in 2D echocardiographic imaging.

3.2 Non-DL approaches

Several approaches have been developed to estimate the motion between two images or over a sequence. In this section, the two main non-DL approaches that have been successfully applied in echocardiography are reviewed: the optical flow and block-matching techniques.

3.2.1 Optical flow approaches

Optical flow is one of the most used methods for motion estimation in computer vision. There are mainly two types of optical flow techniques, one based on the pixel intensities and the other on the phase information. Both methods generally iterate over multiple scales, where coarse displacements are first estimated and then refined by successive iterations to the full resolution of the input images.

3.2.1.1 Intensity-based optical flow techniques

Intensity-based optical flow is built on the assumption that the brightness (I) remains constant over time. In these conditions, the following equation can be obtained:

$$I(x, y, t) = I(x + \Delta x, y + \Delta y, t + \Delta t) \tag{3.1}$$

Considering that the motion between two consecutive instants is small enough, the Taylor expansion applied to the previous equation allows the derivation of the following relation:

$$\frac{\partial I}{\partial t} + \boldsymbol{v} \cdot \boldsymbol{\nabla} I = 0 \tag{3.2}$$

This equation has two unknowns and is therefore under-determined. This situation is referred to as *the aperture problem*. Making the problem well-posed requires the addition of constraints for computing $\boldsymbol{v}$. These constraints are then expressed either globally or locally.

Global methods impose overall constraints on the estimated field, such as regularity or smoothness [21]. The displacement field is then considered as the solution of the minimization process applied on an energy functional. The constraints are usually expressed through a regularization term which is added to a data attachment term, as illustrated in the following equation [21]:

$$\iint \left(\frac{\partial I}{\partial t} + \boldsymbol{v} \cdot \boldsymbol{\nabla} I \right)^2 dx_1 dx_2 + \lambda \iint \|\Delta \boldsymbol{v}\|^2 dx_1 dx_2, \tag{3.3}$$

where λ corresponds to a positive parameter which balances the influence of the two terms.

The local methods aim to reinforce the consistency of the estimated movement at the level of a few pixels. One of the historical method corresponds to the Lukas Kanade algorithm. The underlying strategy resides in the assumption that the neighboring pixels have a common displacement field. This notion of neighborhood is then introduced through the definition of a window. The use of a window of size $n \times n$ pixels allows to solve equation (3.2) through an approximation problem with a system of n^2 equations with 2 unknowns. This problem can be efficiently formulated using a matrix notation of the form $WAv = B$ where W corresponds to a diagonal pixel weighting matrix, A is a matrix containing the spatial derivatives of the intensities and B a vector of the temporal derivatives. These matrices are expressed as:

$$W = \begin{bmatrix} W_{x_1,y_1} & & \\ & \ddots & \\ & & W_{x_n,y_n} \end{bmatrix} \quad B = \begin{bmatrix} I_t(x_1,y_1) \\ \vdots \\ I_t(x_n,y_n) \end{bmatrix} \quad A = \begin{bmatrix} I_x(x_1,y_1) & I_y(x_1,y_1) \\ \vdots & \vdots \\ I_x(x_n,y_n) & I_y(x_n,y_n) \end{bmatrix}$$

Generally, W involves weighting coefficients computed from a Gaussian function. This allows to give more importance to the pixels close to the pixel of interest. Finally, the use of the standard linear least square method allows the derivation of the following optimal solution:

$$\mathbf{V} = (\mathbf{A}^\intercal \mathbf{W} \mathbf{A})^{-\mathbf{1}} \cdot \mathbf{A}^\intercal \mathbf{W} \mathbf{B} \tag{3.4}$$

One of the most powerful methods based on optical flow in echocardiographic imaging is the Farneback algortihm [22]. This method uses second order polynomials to efficiently define the neighborhood of each pixel. A multi-scale analysis is then performed, in which large motions are first estimated at the lowest resolution. These motions are then iteratively refined to the resolution of the input images.

3.2.1.2 Phase-based optical flow techniques

One of the main weaknesses of intensity-based optical flow techniques is related to the assumption of intensity conservation during motion. Indeed, intensity can be modified

according to various factors such as the change of luminosity. To solve this issue, other methods based on optical flow and exploiting phase information have been introduced [23]. Indeed, phase information is known to be more robust to intensity changes. These algorithms thus exploited the phase consistency to estimate motion. Different methods were used to extract the phase information from the input images, the most common being the Gabor filter and the techniques exploiting the monogenic signal [24]. Both approaches transformed the input images into magnitude and phase information which are then used to solve equation (3.2) or (3.4). Since the extraction of phase can be easily applied at different scales, most of the phase-based optical flow techniques are based on multi-scale motion analysis strategy. Most of these approaches are time consuming because of the computation of the phase information. GPU solutions therefore have been proposed for real time applications [25].

3.2.2 Block matching approaches

Block matching are also popular methods for motion estimation in ultrasound imaging. In these approaches, the region to be tracked in the source image is divided into patches named blocks. Each block is then searched in the target image inside a predefined window centered around the block of interest. Standard similarity metrics are used to find the optimal match between the source and target blocks. Once the displacement of all blocks is calculated, a filtering step is required to smooth the final displacement field attached to the region of interest.

One of the most powerful methods based on block matching in echocardiographic imaging is the Particle Image Velocimetry (PIV) algorithm. This method has recently won a challenge in ultrasound during the International Ultrasonics Symposium in 2018 [26]. In this technique, the cross-correlation computed through fast Fourier transform is used to efficiently measure the degree of similarity between 2 blocks. In particular, the peak of the cross-correlation is used to find the optimal displacement at the pixel level. Parabolic peak fitting can also be applied to the cross-correlation operator to estimate displacements at the sub-pixel level. PIV also allows multi-scale motion analysis by adapting the size of the search window which makes it possible to estimate large or small displacements. Finally, an extension of this method allows a global improvement of the accuracy by using n successive frames as input to estimate the motion on one single image from $n-1$ cross-correlation operators. Figure 3.1 gives an illustration of the PIV method when two images are used as input.

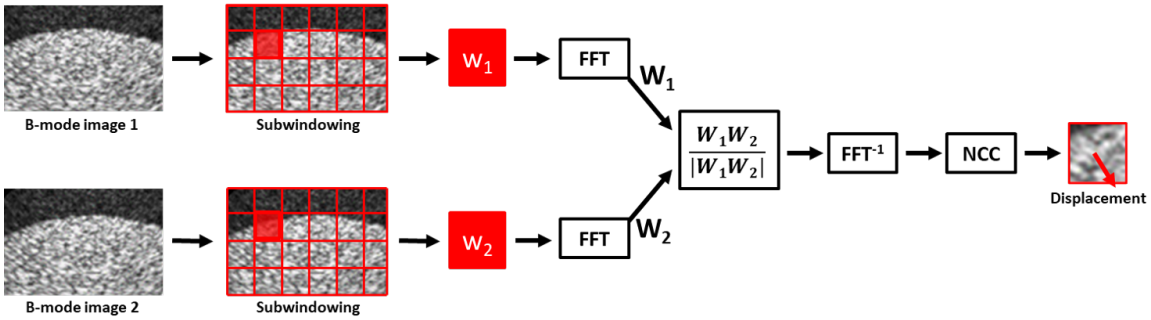


Figure 3.1 – PIV block-matching method as proposed in [26]. NCC stands for Normal Cross Correlation, while w_i and W_i corresponds to the input image blocks and the corresponding Fourier transforms, respectively.

3.3 State-of-the-art deep learning methods applied to motion estimation

DL-based motion estimation is a field in which a lot of research has been conducted in recent years, especially because of its many growing applications like robotics or autonomous cars. The application of deep learning methods in this field has contributed to a significant improvement in performance which has generated a strong interest with more than 110 methods developed in 6 years. An overview of the number of DL methods developed each year is given in Figure 3.2. Many advances made in other research areas have been used to further improve the performance of DL methods for motion estimation. For instance, several significant advances have been made through the use of the well-known U-Net architecture introduced for image segmentation [27] and more recently through the use of recurrent networks or transformers used in natural language processing. The best performing DL solutions for motion estimation correspond to supervised learning approaches, with currently achieve 40% higher accuracy than unsupervised techniques. Supervised methods are trained on dedicated databases with reference dense motion fields. A detailed description of such databases is provided in Chapter 4.

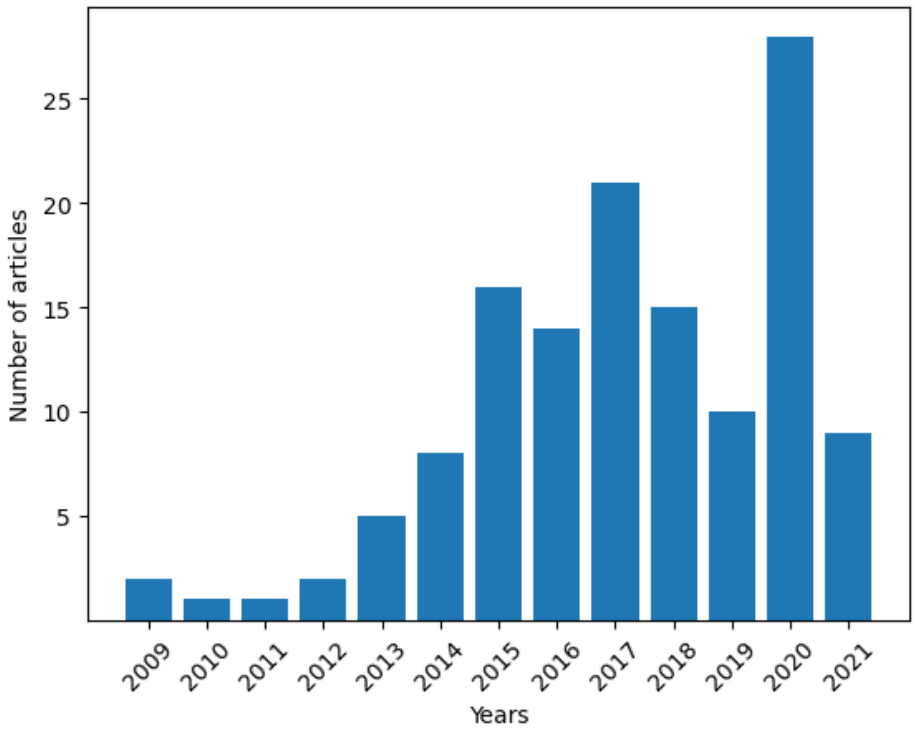


Figure 3.2 – Evolution of the number of methods benchmarked on MPI Sintel [28] per year (survey done in September 2021).

In this section, I will provide a detailed overview of the DL architectures developed to estimate motion from natural images. In particular, I will focus on the originality of architectures that gave a leap in performance. An overview of these DL architectures is given in Figure 3.3.

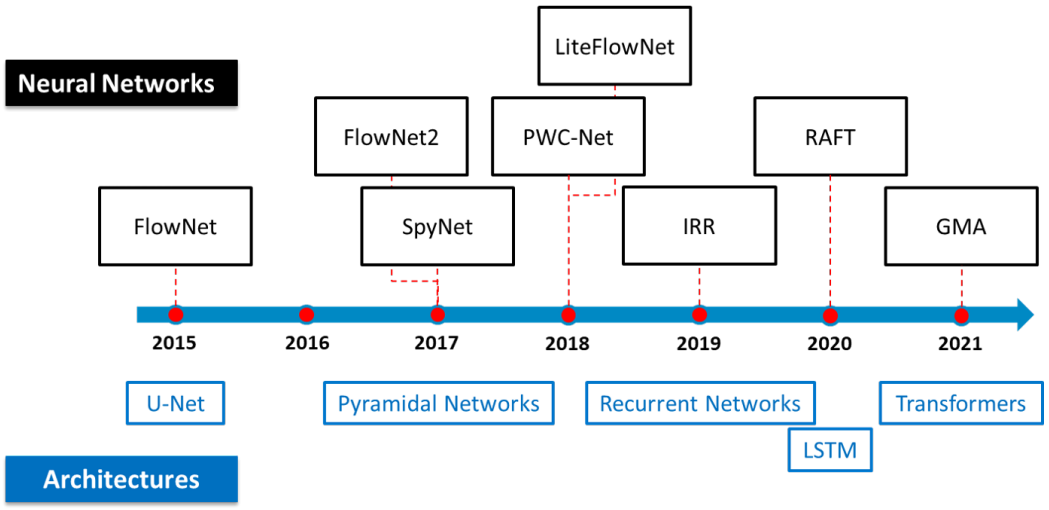


Figure 3.3 – Overview of the main DL methods in motion estimation on natural images.

3.3.1 FlowNet: a U-Net based network

3.3.1.1 Network architecture

FlowNet is the first end-to-end motion estimation network. Over the years, different versions of FlowNet have been proposed to enrich the network. The first FlowNet model was based on the U-Net [27] architecture, published the same year [29]. Its architecture is composed by three main components: an encoding part, a bottleneck and a decoding part. The conventional pooling layers used in the encoder were replaced by convolution layers with a stride of 2. The model produces a displacement map at a fourth of the original image resolution, followed by bilinear interpolation to generate the final deformation field at full resolution. The authors justify this choice by the small benefit of producing displacement maps up to the full resolution at the expense of a much larger number of parameters to optimize. In the first version of FlowNet, two architectures were introduced: FlowNet-C and FlowNet-S, where the two characters c and s stand for correlation and simple, respectively. A synthetic description of the two networks is provided below.

- FlowNet-S: implements a classical U-Net architecture for which the two input images are concatenated and the displacement field is estimated as output. The first three convolution layers of the network have a kernel size equal to 7, 5 and 5 pixels, respectively. A kernel size of 3 is then used for the other convolution layers involved in the architecture. This network is represented in blue in Figure 3.4

- FlowNet-C: This network is also based on the U-Net architecture but is also inspired by Siamese networks [30]. Indeed, the images at time t and $t+1$ are processed separately by convolution layers with shared weights before being correlated by a dedicated layer, named as correlation layer. The idea behind this architecture is to reproduce as closely as possible what has been done in traditional motion estimation techniques. As for FlowNet-S, the first three convolution layers have a kernel size of 7, 5 and 5 before decreasing to 3 for the remaining layers. This network is represented in red in Figure 3.4

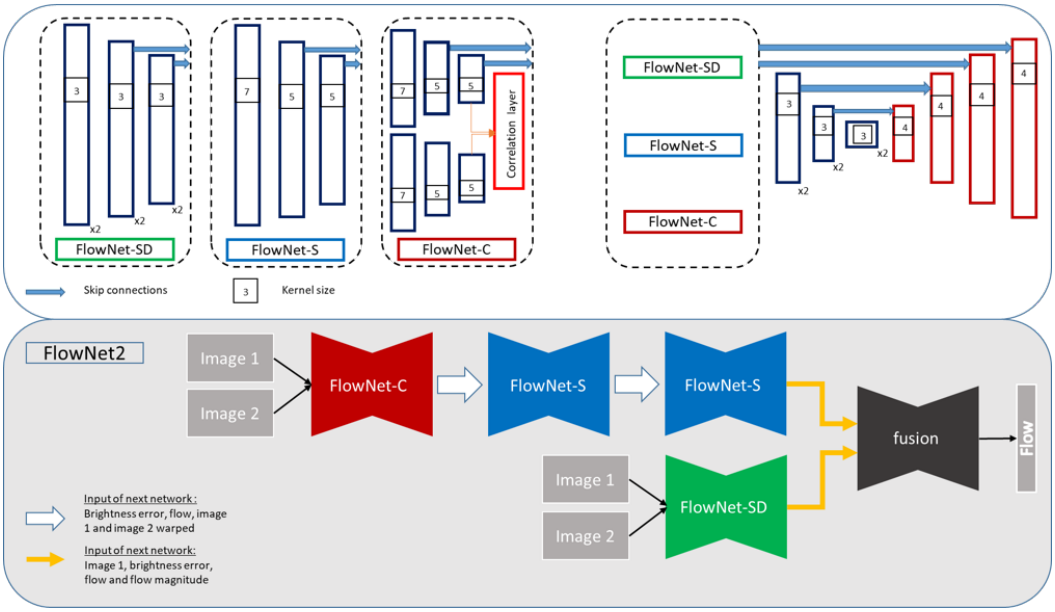


Figure 3.4 – Schematic view of the overall architecture of FlowNet2 (bottom) [31] and the architecture of all the sub-networks that compose it (top). Convolution layers are shown in blue and deconvolution layers in red

It is interesting to note that this first version was used as a sub-network of more complex models. For instance, SegFlow [32] used a FlowNet-S-like network coupled with a ResNet architecture to jointly predict a segmentation map and a displacement field, with the idea of improving results through a multi-tasks approach [33].

FlowNet2 was later proposed in [31] as an extension of FlowNet by the same authors. This network is based on two branches: one specialized for the estimation of large displacements thanks to the stacking of FlowNet-C and two FlowNet-S sub-networks and one dedicated to small displacements. In the branch specialized for large displacements (on the top of the FlowNet2 diagram given in Figure 3.4), the three sub-networks have convolution kernels of greater size (i.e. 7, 5 and 5) suitable for the estimation of large range of movements. These sub-networks give as output an estimation of the flow. FlowNet-C exploits the source and target images as input while the two Flownet-S take as input a concatenation of *i)* the source image; *ii)* the flow estimated by the previous sub-network; *iii)* the target image warped by the estimated flow and *iv)* the brightness error (white arrows on the Figure 3.4). This corresponds to the norm per channel of the difference between the original image and the target image warped by the estimated field. The branch specializing in small displacements (on the bottom of the FlowNet2 diagram given in Figure 3.4) only involves a single network named FlowNet-SD (where SD stands for small displacements). This network has a similar architecture to FlowNet-S except that the convolution kernels of the first layers have been replaced by kernels of size three (network in green in Figure 3.4).

The estimated movements at the output of the two branches are finally merged by a dedicated network to produce the final prediction (this network is named Fusion in Figure 3.4). More particularly, this fusion network takes as inputs *i)* the original image; *ii)* the brightness error; *iii)* the estimated flow and *iv)* the flow magnitude (corresponding to the channel norm of the estimated flow) from each of the two branches. The network corresponds to a simple U-Net with two levels. The different sub-networks

were trained sequentially using Adam [34] optimizer and the End Point Error (EPE) loss function. This function is described in Section 3.5. An exception is the fusion network which is trained by freezing the weights of the two branches and by minimizing a L_{pq} loss with $p = 2$ and $q = 0.2$.

In parallel with the development of the architectures, the authors of FlowNet also created two key synthetic databases that have been crucial to the performance of their algorithm. Indeed, facing the lack of data to train neural networks, the authors first generated a large synthetic database named FlyingChairs. A second synthetic database called ChairsSDHom was also designed to assist in small displacement estimation through sub-pixel motion simulation. Details on these two databases are given in Section 4.2.4 and 4.2.5. These databases are now state-of-the-art and used by most deep learning methods in motion estimation.

FlowNet2 has significantly improved the results in terms of motion estimation compared to other neural networks and traditional methods. It therefore corresponds to a key solution in the domain. Nevertheless, this network has several weaknesses, among which we can mention the number of parameters to learn (more than 162M) and the complex training strategy involving several stages.

3.3.1.2 Key fundamental concepts

The success of FlowNet also resides in the introduction of key concepts which are used nowadays by most DL methods in motion estimation. In particular, a dedicated data augmentation strategy including geometric augmentations (translation, rotation and scaling) and image alterations (noise, brightness, contrast, gamma, color) was proposed to improve the generalization capacity of the network. The corresponding parameters are detailed in the original paper [29]. The performance of the network has also been improved by the introduction of the deep supervision technique. In particular, the loss function was optimized at several levels of the network and with larger coefficients for the layers with the lowest resolution. Finally, the use of learning rate decay was introduced in the context of motion estimation. The authors proposed to divide by two the initial learning rate every n iterations according to different schemes depending on the targeted networks. An illustration of this learning rate strategy is given in Figure 3.5.

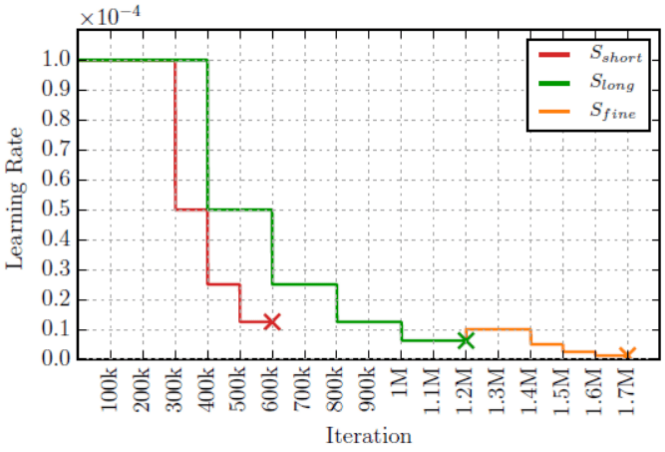


Figure 3.5 – Learning rate schedules as proposed in [31]

3.3.2 SpyNet: image warping based pyramidal networks

Contrary to FlowNet, the authors of SpyNet [35] developed a network closer to traditional methods by implementing a multi-scale strategy of which an illustration is provided in Figure 3.6. To this end, a pyramidal architecture was introduced to refine the motion estimation at different scales, from coarse to fine. Specifically, the network is composed of different levels, each of them having a dedicated sub-network with five convolutional layers. The size of the input images is divided by a factor two at each level. The sub-networks, named G_i in Figure 3.6, received as inputs the source image, the warped target image and the previously estimated flow expressed at the corresponding resolution level. They output a residual displacement which is added to the previously estimated flow to refine it for the next level. Each of the sub-network was trained separately at their resolution level based on the Adam optimizer. The learning rate was set to $1e^{-4}$ for the first 60 epochs and then decreased until the value of $1e^{-5}$ at convergence. Because of its strategy, SpyNet comes with fewer parameters than state-of-the-art methods and has the advantage of being faster for both learning and inference. Its performance are slightly lower than FlowNet but with 96% fewer parameters.

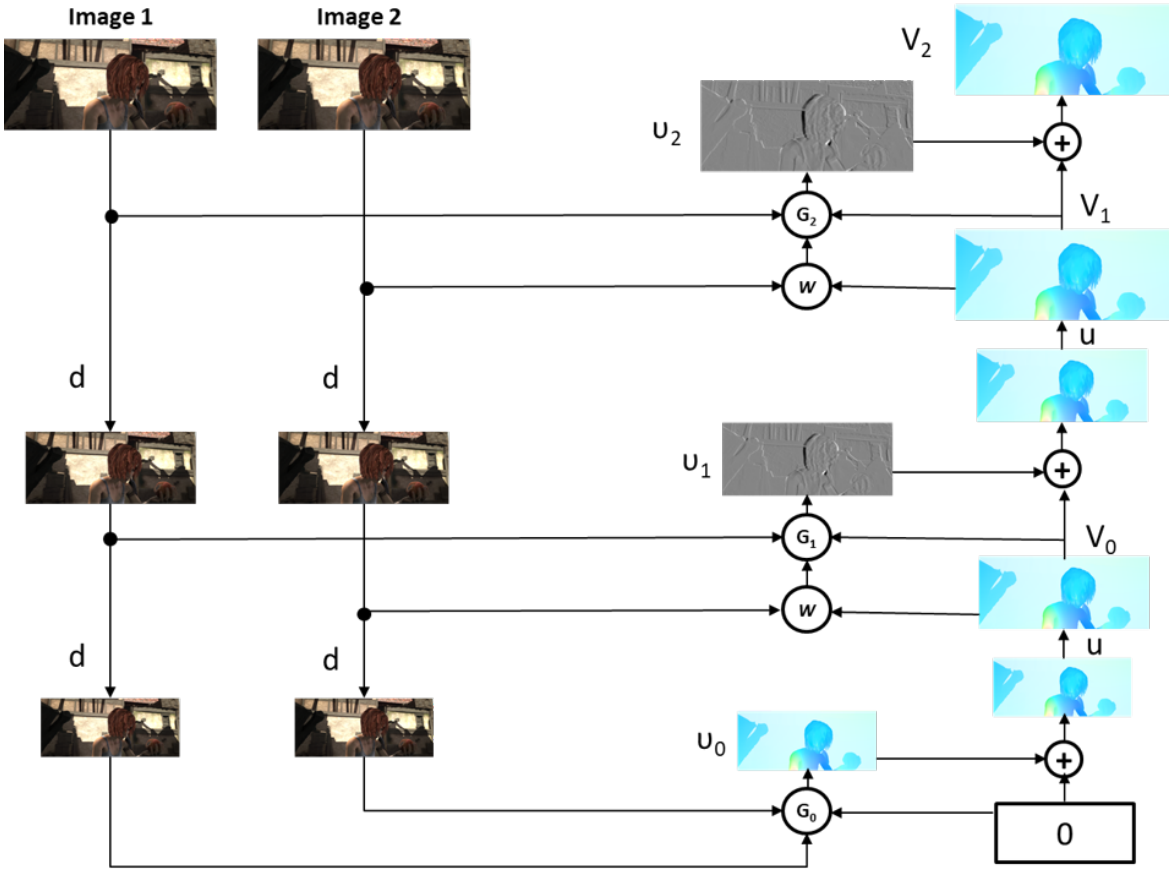


Figure 3.6 – Schematic representation of SpyNet [35] as 3-Level Pyramid Network

3.3.3 Feature warping based pyramidal networks

Following the same idea as SpyNet, many neural networks have been inspired by pyramid networks, which are lighter and faster at training and inference than FlowNet. In

this context, two close architectures were released simultaneously: PWC-Net [36] and LiteFlowNet [37]. Contrary to SpyNet, these two networks expressed the pyramidal structure into the feature space, where warping and correlation operations are used to efficiently estimate the motion. Correlation measurements are performed based on the creation of a cost volume which allows to capture motion information. This component is essential and has been the subject of intense research. For instance, Devon *et al.* introduced the concept of deformable cost volume with dilated rates to capture the motion of fast objects with small size [38]. More recently, Hui *et al.* proposed to make the cost volume more accurate by introducing weighting coefficients computed from the correlation among different displacement for each pixel [39].

3.3.3.1 LiteFlowNet

LiteFlowNet is a network composed of the following two key parts [37]:

- NetC is a multi-resolution feature extractor that generates 6-level pyramid features from the two input images. It uses shared weights to calculate the features attached to each image. The spatial resolution is divided by two between each level.

- NetE predicts the displacement field at each level of the pyramid. Specifically, a first displacement field is estimated by two modules from the lowest resolution feature maps computed from the input images. The first module (named M in Figure 3.7) corresponds to a descriptor matching unit which makes a first motion estimation at the pixel level. The second module (named S in Figure 3.7) refines the motion estimated by M to increase the accuracy at a sub-pixel level. The M module contains warped features and correlation layers to compute a cost volume. This cost volume is then used to estimate residual displacements which are then added to the motion field estimated at the previous level and scaled in terms of resolution and magnitude. The S module takes as input the features of the source image warped by the previously estimated motion and the features of the target image. It then refines the motion field by adding the residual sub-pixel displacements. The output of S in finally regularized to remove outliers by adapting the regularization kernel for each pixel through a feature-driven local convolution layer (module named R in Figure 3.7).

The final network is composed of 6 levels of pyramids. During training, the same data augmentation strategy, optimizer, loss function and parameters introduced in FlowNet2 were used. The displacement field is also estimated up to a quarter of the original resolution of the input images. The final estimated motion is then obtained by a bilinear interpolation operation. This network has recently been improved by modulating the cost volume and introducing a confidence map to estimate the movement on more reliable neighboring features [40].

3.3.3.2 PWC-Net

PWC-Net is based on the same key elements as LiteFlowNet, namely feature warping, cost volume and pyramid structure. An illustration of the overall architecture is provided in Figure 3.8. A pyramid of seven levels with shared weights is first used to

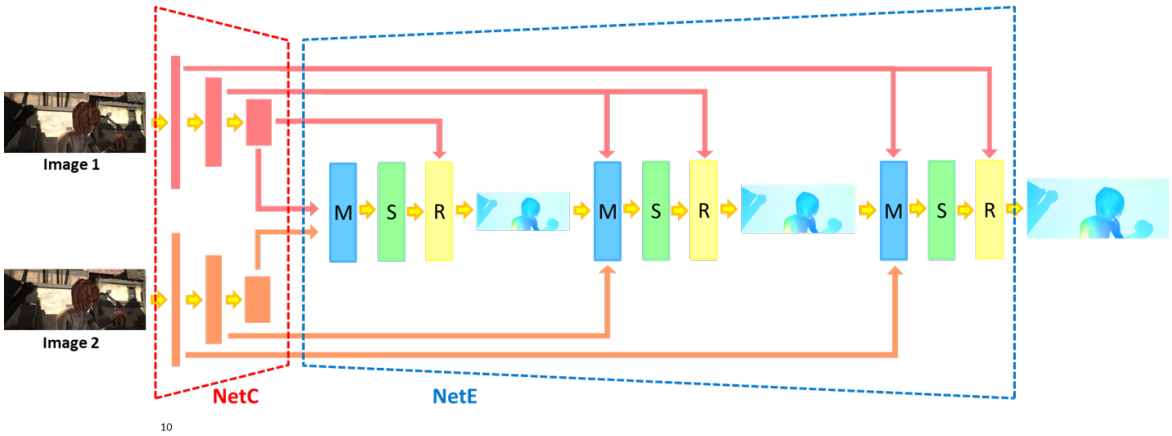


Figure 3.7 – Schematic representation of LiteFlowNet [37] as 3-Level Pyramid Network.

extract features maps from input images at different resolution. The feature dimensions are divided by two at each level. The cost volume is then calculated between the feature maps of the source image and those of the target image warped by the flow estimated at the previous level and upsampled. This volume is then concatenated with the feature maps of the source image and given as input to a convolutional neural network to estimate the dense displacement field at a given resolution level. These steps are iterated until the second to last level of the pyramid. At this point, the estimated motion is passed to a final CNN named context network. This network uses dilated convolutions to extract contextual information, increase the receptive field and thus refine the motion estimation. Finally, a bilinear interpolation upsamples the final flow to produce a displacement map of the same size as the input image. The archiectural elements of PWC-Net and LiteFlowNet are similar except for the context network of PWC-Net and the regularization module of LiteFlowNet. The training parameters of the network are also based on the elements introduced with FlowNet2 (whose details are given in Section 3.3.1.2). In particular the same loss management, data augmentation strategy, optimizer and learning rate schedule were used.

While PWC-Net involves almost the same number of parameters, it outperforms LiteFlowNet, making it one of the currently best performing networks for motion estimation. Another version of PWC-Net named PWC-Net+ has been recently released. The corresponding architecture remains the same, but several changes have been made to the way the network is trained [41]. Over the past three years, numerous studies have been conducted to further improve the results from the initial PWC-Net architecture. In [42], the authors proposed to jointly estimate the motion and the corresponding confidence map. More recently, Yang *et al.* developed an approach with a pyramidal feature extraction for the creation of a cost volume and a volumetric U-Net to process the cost volume and extract the displacement field [43]. Table 3.1 provides an overview of the performance of FlowNet-based and PWC-Net-based approaches obtained from the same open-access database.

3.3.3.3 IRR

IRR [44] was built on the PWC-Net or FlowNet-S architectures, as shown in Figure 3.9. The general principle of this method is to iteratively refine the motion estimation through the use of a simple block with shared weights. Furthermore, the method is

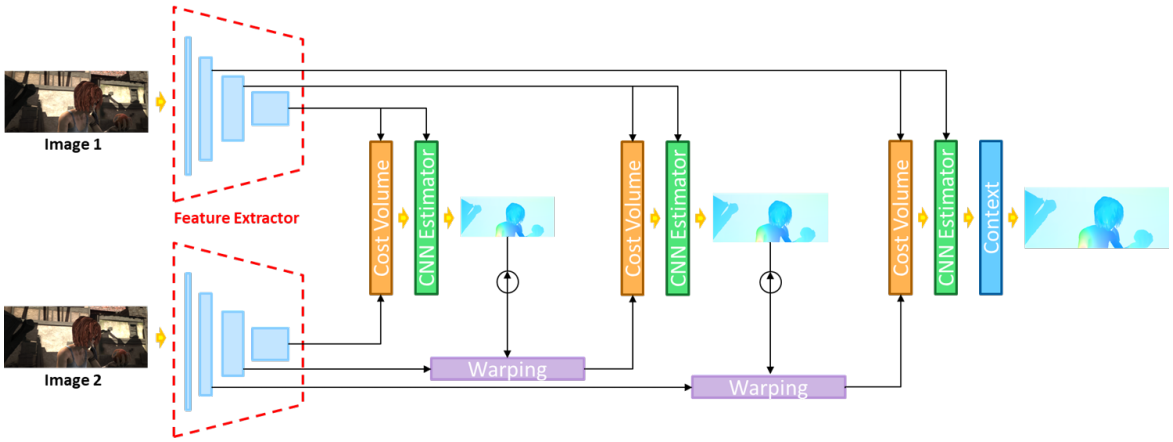


Figure 3.8 – Schematic representation of PWC-Net [36].

	Methods						
	FlowNet2 [31]	*FlowNet-S [29]*	*FlowNet-C [29]*	*SpyNet [35]*	*PWC-Net [36]*	*PWC-Net+ [41]*	*LiteFlowNet [37]*
Parameters (M)	160	38	39	1.2	8.75	8.75	5.37
EPE (px)	5.74	7.22	7.88	8.36	5.04	4.60	5.38
Runtime (s)	0.12	0.03	0.03	0.16	0.03	0.03	0.09

Table 3.1 – Model size, runtime and accuracy on MPI Sintel test final dataset [28] of the previously mentioned networks. The runtime is taken from the respective articles and is calculated on RTX 1080Ti.

based on a multi-task approach where the bidirectional motion field and occlusion maps are jointly estimated. An occlusion map is defined by the pixels in the source image that do not have a match in the target image. In addition, by reversing the order of the input images and duplicating the network, IRR estimates the displacement field in both directions simultaneously, improving the overall accuracy of the network. The network was trained by minimizing a loss function composed of the EPE term for motion estimation and a weighted binary cross entropy for occlusion map estimation. The same loss management, data augmentation strategy, optimizer and learning rate schedule as the ones introduced with FlowNet2 were used. IRR is often considered as one of the first recurrent networks because the estimated flow is used in an iterative way to warp the feature maps extracted from the images. However, several algorithms have gone further in this approach and are described in the next section.

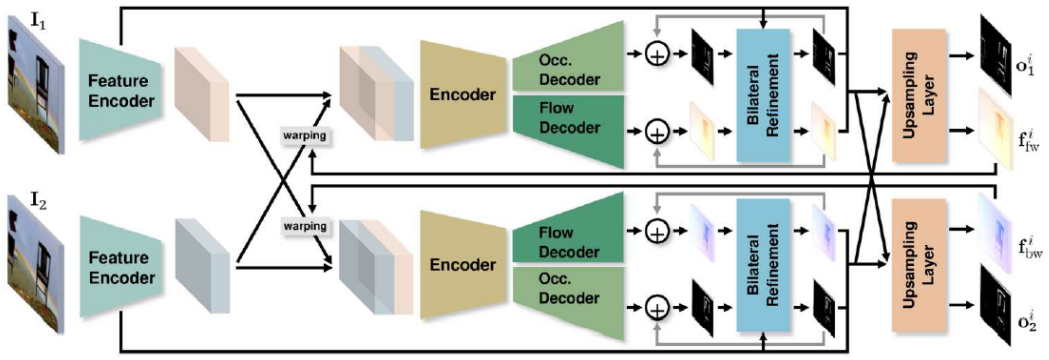


Figure 3.9 – Illustration of the IRR-FlowNet architecture [44]. This network produces both a bidirectional motion estimate and the corresponding occlusion map.

3.3.4 Recurrent Networks

As in many application areas, motion estimation has also benefited from innovations in other deep learning areas of research. In this section, I will describe some of the best performing networks in motion estimation based on recurrent networks and transformers.

3.3.4.1 RAFT: recurrent all-pairs field transforms network

RAFT is a recurrent displacement field estimation network [45]. This network was inspired by natural language processing and uses the Gated Recurrent Unit (GRU) as a basis. The corresponding architecture is provided in Figure 3.10.

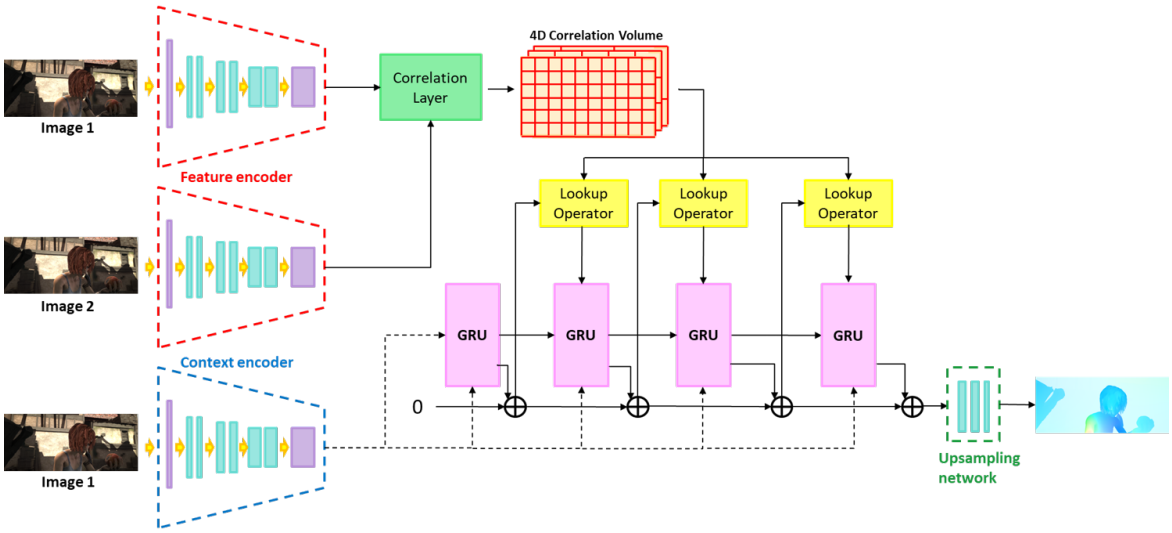


Figure 3.10 – Schematic representation of RAFT network [45].

The network is composed of four main parts:

- a feature encoder composed of 6 residual blocks producing the features of the source image and the target image separately at one eighth of the original resolution. A second encoder with the same architecture is also used to extract context information from the source image.
- a correlation layer that builds a correlation volume by computing the dot product between all pairs of feature vectors extracted by the first encoder. The last two dimensions of the 4D correlation volume are then pooled by kernels of 1, 2, 4 and 8 to integrate information at different resolution. It allows the extraction of relevant information of large and small displacements.
- a Lookup operator that generates a feature map by indexing the correlation volume. Indexing consists in defining a local grid around the propagated pixel according to the previous estimate limited by a constant radius for all resolutions.
- a GRU block [46] which recurrently updates the estimated displacement field.

The displacement field generated by the GRU block is one eighth of the resolution of the original image. This information is therefore input into an upsampling network

composed of two convolution layers and a softmax activation function to reach the resolution of the input image. RAFT is currently one of the best performing methods in motion estimation.

3.3.4.2 GMA: global motion aggregation network

Very recently, a new algorithm called GMA has been proposed to further improve the performance in motion estimation [47]. This network extended the RAFT architecture by integrating attention mechanisms through the use of transformers [48]. The idea of this network is to preserve the performance of RAFT while improving the results in situations with the presence of occlusions. Specifically, hidden motions are handled with a transformer module called GMA in Figure 3.11. In this module, the query and key vectors are computed from the features extracted by the context encoder, while the value vectors are derived from the correlation volume. The correlation volume is constructed in the same way as in the RAFT network. The GMA module then calculates the feature vector updates as a weighted attention sum of the motion features. The concatenation of the updated feature vectors with the motion features is then given as input to the GRU block of a standard RAFT network. As shown in Table 3.2, this network is currently the best performing solution for motion estimation on the MPI Sintel database [28].

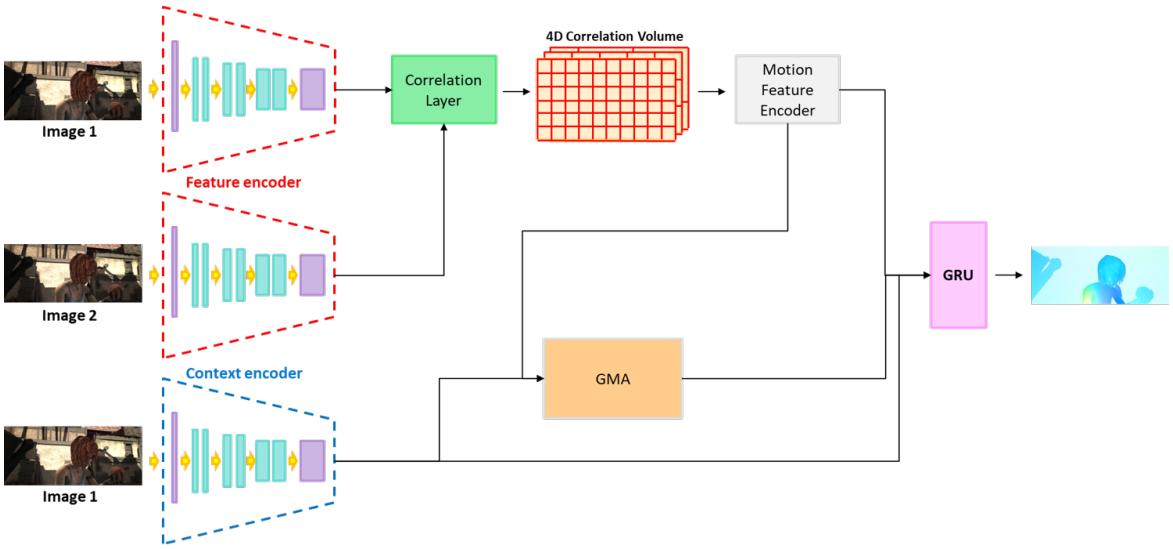


Figure 3.11 – Illustration of the GMA architecture as proposed in [47]

	Methods			
	PWC-Net+ [41]	*IRR-PWC [44]*	*RAFT [45]*	*GMA [47]*
Parameters (M)	8.75	6	5.3	5.9
EPE (px)	4.60	4.58	2.86	2.47
Runtime (s)	0.03	0.2	0.1	/

Table 3.2 – Model size, runtime and accuracy on MPI Sintel test final dataset [28] of the recurrent networks vs the best method mentioned in the table 3.1. The runtime is taken from the respective articles and is calculated on RTX 1080Ti.

3.4 Applications of DL methods to motion estimation in US imaging

Several deep learning motion estimation methods have been applied in ultrasound imaging, mainly since 2019. Most of them are based on architectures that have been previously validated on natural scene images. In terms of application, these methods have been mainly used for structural tracking and tissue deformation (one of the main objectives being elastography). It is important to note that at the beginning of my PhD, very few methods were already published. I thus realized my project on a very competitive context. This section will detail the deep learning methods developed for motion estimation in ultrasound imaging, both for the analysis of tracking and deformation. We will first focus on supervised learning methods and then on unsupervised and semi-supervised learning approaches.

3.4.1 Supervised Methods

As is the case for natural images, supervised approaches are the methods that currently give the best results in ultrasound imaging. In this context, the networks achieving the best performances on natural data have been re-used as basic architectures for ultrasound image analysis.

In [49], the authors estimated the carotid wall motion by mimicking the operation of a traditional block matching algorithm using Siamese networks to extract correlated feature maps before being post-processed by a linear Kalman filter.

Another hot topic is the analysis of the left ventricle in echocardiography. In several studies, this estimation occurs in a complete cardiac function analysis pipeline based on view recognition, myocardial segmentation, myocardial motion estimation and clinical indices computation. In a first paper, FlowNet2 was used to estimate myocardial motion without ultrasound-specific learning [50]. More recently, the same authors improved their results by developing a new motion estimator based on PWC-Net called EchoPWC-Net [51]. To adapt this network to ultrasound images, the authors removed the feature maps warping, propagated the first feature maps and added finer resolutions to the loss. While the results obtained on the synthetic dataset were much better than those produced by conventional methods (based on optical flow), the scores obtained on real images deteriorated, mainly because of the lack of variability in the training database.

The networks developed in elastography are also based on the most efficient deep learning architectures on natural images. The first networks applied in this field were based on FlowNet2 architectures [52], [53]. More recently, a comparison of FlowNet2, PWC-Net and LiteFlowNet trained on the same simulated elastography data was realized [54]. The results showed that the best performing network was the PWC-Net with a transfer learning strategy. At the same time, a network based on PWC-Net called MPWC-Net was developed. The originality of this network is that it takes as input a concatenation of the RF data with the corresponding envelope and B-modes images to improve its performance [55]. More recently, this network was still improved by integrating some concepts from the IRR-PWC architecture [56]. Finally, recurrent networks were also applied in elastography to extract relevant temporal correlations between frames to estimate shear modulus images [57].

3.4.2 Semi-supervised & Unsupervised Methods

Unsupervised and weakly supervised methods are usually developed to cope with the lack of annotated databases. These methods have been mainly applied in ultrasound imaging for the analysis of the liver and the left ventricle. The underlying architectures are mostly based on the classical U-Net and Siamese networks [30].

In [58], an unsupervised method named Cascaded One-shot Deformable Convolutional Neural Network (COSD-CNN) was proposed to track landmarks in ultrasound images. For training purposes, the final target images to track were generated by combining a source image and a random image of the sequence. A Siamese network [30] was first used to extract relevant features at different scales. A cross-correlation response map was then estimated from these features and used as initialization for a second branch. This branch was composed of another Siamese network with a one-shot deformable convolution (OSDC) module to adapt the receptive field to the deformation of the target to be followed.

In [59], a classical U-Net was trained in an unsupervised way to estimate the motion in ultrasound by minimizing the distance between the source image distorted by the estimated field and the target image. This approach does not allow a reliable estimation of the movements inside a region but is more appropriated for contour tracking. At the same time, a semi-supervised method of joint learning of segmentation and motion estimation of the left ventricle was proposed in [60]. Specifically, this network is composed of *i)* a supervised U-Net for left ventricular segmentation; *ii)* an unsupervised FlowNet-C network for motion estimation and *iii)* a recurrent network to impose temporal continuity.

In [61], The authors developed a method to jointly segment the lumen and thickness of the intima-media and estimate carotid wall motion. Their network received as input a whole sequence of 2D images and provided as output the motion of the carotid through a pyramidal Siamese sub-network and the segmentation of the different structures through a multi-task regression network. The displacement field was estimated in an unsupervised way and bidirectionally while the segmentation task learning was supervised.

In elastography, a semi-supervised method was implemented for fine-tuning purposes based on the LiteFlowNet architecture from which the weights were taken to initiate the transfer learning to the unannotated RF data [62]. From these weights, an unsupervised learning was performed by minimizing the Charbonnier penalty between the first image and the second image warped by the estimated forward flow to enhance the accuracy and by enforcing the smoothness on the axial and lateral displacements and derivatives.

3.5 Evaluation Metrics

Motion estimation methods are evaluated on benchmark databases for which the motion reference is known. The endpoint error (EPE), the angular error (AE) or the F1-all score are the metrics classically used for this task [63].

3.5.1 Endpoint Error

The endpoint error (EPE) is defined as the Euclidean distance between the reference displacement vector (u_{GT}, v_{GT}) and the estimated one (u, v) [64]. Its expression is given as:

$$EPE = \sqrt{(u - u_{GT})^2 + (v - v_{GT})^2} \tag{3.5}$$

3.5.2 Angular Error

The angular error (AE) corresponds to the angle between the reference and the estimated displacement vectors. This value is computed from the scalar product of the two vectors. To avoid division by the zero value, the angle is calculated between the 3D displacement fields whose z component is equal to 1:

$$AE = \arccos\left(\frac{1.0 + u \times u_{GT} + v \times v_{GT}}{\sqrt{1.0 + u^2 + v^2}\sqrt{1.0 + u_{GT}^2 + v_{GT}^2}}\right) \tag{3.6}$$

The disadvantage of this metric is that it penalizes small displacements more heavily.

3.5.3 F1-all score

This metric corresponds to the percentage of outliers among all the estimated values in the image. Its expression in given as:

$$F1 - all = \frac{B}{H \times W} \times 100 \tag{3.7}$$

where B corresponds to the number of outliers, H the height of the image and W its width.

3.6 Conclusions

Motion estimation is a problem that has been addressed for many years. Deep learning solutions have recently made a breakthrough in this field by outperforming all traditional methods. Different architectures have been successfully designed for this task, among which pyramidal structures such as PWC-Net and recurrent networks are among the most efficient.

At the beginning of my thesis, there were very few deep learning techniques applied to ultrasound imaging. The corresponding studies were limited to using pre-trained methods on natural scene images without specific learning on ultrasound data. One wonders why no further work was done at the time to improve motion estimation in ultrasound imaging. One of the main reasons was probably the difficulty of accessing referenced datasets for this modality, i.e. data whose motions are known and which can be used as a reference for learning algorithms. In this context, ultrasound simulations can play a key role in building such datasets.

Chapter 4

Open-access two-dimensional databases

Contents

4.1 Introduction

Benchmarking or training neural networks in a supervised way requires databases with references. For motion estimation, it is impossible to manually generate such databases, as it can be done in other tasks, such as segmentation. Indeed, estimating the displacement of each pixel of the image between two frames by hand would be tedious, if not impossible. One effective solution is to generate simulated images and motion fields, for which motion is define in a dense way for all or part of the images. In the DL community on motion estimation, the reference motion fields are usually displayed in an HSV color code by a test pattern describing the magnitude of the displacement field and its direction, as illustrated in Figure 4.1.

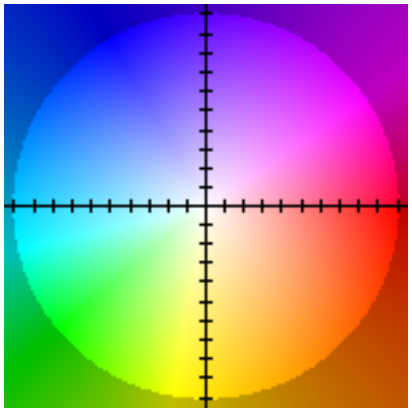


Figure 4.1 – Reference test pattern for color-coded representation of displacement fields.

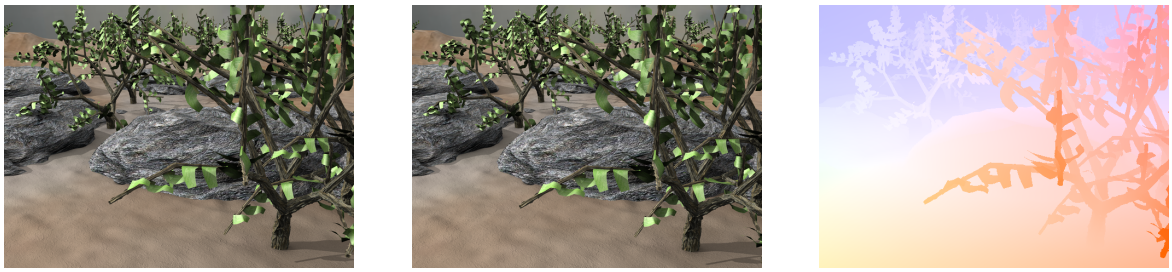
Since synthetic databases are a core part of the state of the art in DL tracking, I will describe in the next section established databases for both natural and 2D echocardiographic images.

4.2 Open-access 2D synthetic natural images

4.2.1 Middlebury

Faced with the impossibility of manually having a dense reference field to estimate the movement on the images, many synthetic databases have been created in order to benchmark tracking methods. The Middlebury database [63] is one example.

The initial database was composed of 150 image pairs divided into two sets, 72 image pairs for training and 78 for testing. However, reference motion fields were made available only for the training dataset. Estimated displacement fields on the test set must be sent to the platform to be benchmarked. The resolution of the images ranges from 316×252 pixels to 640×480 pixels. Unfortunately, the number of frames with references is too low for training neural networks. Also, the motion range is rather limited (less than 35 pixels).



(a) (b) (c)

Figure 4.2 – Example of an image pair (a, b) and the corresponding ground truth (c) from the Middlebury database [63].

4.2.2 KITTI

The recent interest in intelligent and autonomous cars triggered the development of databases specializing in car movements. KITTI2012 [65] and KITTI2015 [66] were built from real street stereo images acquired by cameras mounted on a car and used for tracking benchmarks.

In KITTI2012, the ground truths of the motion field were derived from the motion of the car and the lidar placed on the roof of the vehicle. The lidar captures static objects in the scene, but at a coarse resolution. The resulting ground truth is sparse and reduced to the static parts of the scenes. The labeling of KITTI2015 also includes the trajectories of moving objects in a sparse density. To add these trajectories, the authors generated 3D cars models and fit them to the points cloud.

KITTI2012 has 195 pairs of frames with sparse reference, and KITTI2015 has 200 training sequences. In both cases, the reference field has a density of about 50% of pixels in the images. The final resolution of the images is 1226×370 pixels for KITTI2012 and 1242×375 for KITTI2015. Motion fields are difficult to estimate from these databases because of the presence of many occlusions, illumination changes, and a large range of motion (average displacement of 8 and 9 pixels but a maximum displacement of 549 and 724 pixels on KITTI2012 and KITTI2015, respectively). The main weaknesses of these two databases are the limited number of images with sparse reference displacement fields.

4.2.3 MPI Sintel

Contrary to the previously mentioned databases, MPI Sintel is a fully synthetic database built from an open-source 3D short movie named Sintel [28]. Two versions of the database exist: one with the original images and one with the same images but with additional blur and atmosphere effects. The database is composed of 1041 pairs of images of 1024×436 pixels with dense ground truth motion fields. The average displacement is 13.5 pixels on the whole dataset and 445 pixels at the maximum. This database is one of the most commonly used to benchmark tracking methods but is unfortunately not large enough to train a convolutional neural network. Moreover, motion estimation is particularly difficult with this database because of the presence of large motions, scene illumination, specular reflections, numerous occlusions and the addition of the blurring and atmosphere effects.

(a) (b) (c)

Figure 4.3 – Example of an image pair (a, b) and the corresponding ground truth (c) from the MPI Sintel database [28].

4.2.4 FlyingChairs2D

To overcome the lack of training data, a synthetic database named FlyingChairs2D was generated [29]. This database consists of 2D images of flying chairs from different angles superimposed on various backgrounds. More than 800 chair models and 60 views per chair were generated to increase the database diversity. Synthetic motions were generated from affine transformations applied to the background and the chairs. The variety of displacement is identical to that of Sintel but limited to 150 pixels maximum. A total of 22872 pairs of images with a resolution of 512×384 pixels compose the database, which is important to train the DL methods.

(a) (b) (c)

Figure 4.4 – Example of an image pair (a, b) and the corresponding ground truth (c) from the FlyingChairs2D database [29].

4.2.5 ChairSDHom

The ChairsSDHom [31] database was created to compensate for the lack of small displacements in other databases. Indeed, it was observed that neural networks trained on the other datasets performed well on large displacements but not on small ones. Therefore, the focus was on generating pairs of images containing flying chairs like FlyingChairs2D, but with a large proportion of displacements with an amplitude smaller than 1 pixel. Contrary to other databases, this dataset does not involve moving backgrounds to match real sequences as closely as possible. This dataset is composed of 20966 pairs of images of resolution 512×384 pixels.

4.2.6 FlyingThings3D

This database was synthesized to further increase the diversity of samples and thus facilitate the training of neural network with more and more complex architecture [67]. Each scene is composed of a textured background with randomly selected cuboid and

(a) (b) (c)

Figure 4.5 – Example of an image pair (a, b) and the corresponding ground truth (c) from the ChairSDHom database [31].

cylindrical shapes on which 3D models from ShapeNet, a large-scale dataset of 3D shapes, were superimposed. Between 5 and 20 3D models were randomly distributed and textured before being translated and rotated along a smooth 3D modeled trajectory. This database contains a total of 21818 images of size 960×540 pixels with the corresponding motion reference.

(a) (b) (c)

Figure 4.6 – Example of an image pair (a, b) and the corresponding ground truth (c) from the FlyingThings3D database [67].

4.2.7 Monkaa

Monkaa database [67] was created to add nonrigid and softly articulated motions in a training database with other motion characteristics similar to the MPI Sintel database. Multiple versions of a scene were generated with different camera positions to increase the amount of data. This dataset contains 8591 pairs of images of size 960×540 pixels and with the corresponding ground truths.

(a) (b) (c)

Figure 4.7 – Example of an image pair (a, b) and the corresponding ground truth (c) from the Monkaa database [67].

4.2.8 Driving

A simulated database named Driving [67] was created to look like KITTI by depicting street scenes from the viewpoint of a driving car. In addition, 3D tree models and traffic light models were used to have a more realistic rendering. This results in 4392 frames of size 960×540 pixels with the corresponding ground truths.

(a) (b) (c)

Figure 4.8 – Example of an image pair (a, b) and the corresponding ground truth (c) from the Driving database [67].

4.2.9 CrowdFlow

Other databases have been developed for more specific applications. One example is the CrowdFlow database [68] which is specialized in crowd movement analysis. A specific engine, named Unreal Engine, was used to generate thousands individual movements of people in a simulated urban environment. Ten sequences between 300 and 450 frames of 1280×720 pixels size were generated for a total amount of 3200 frames.

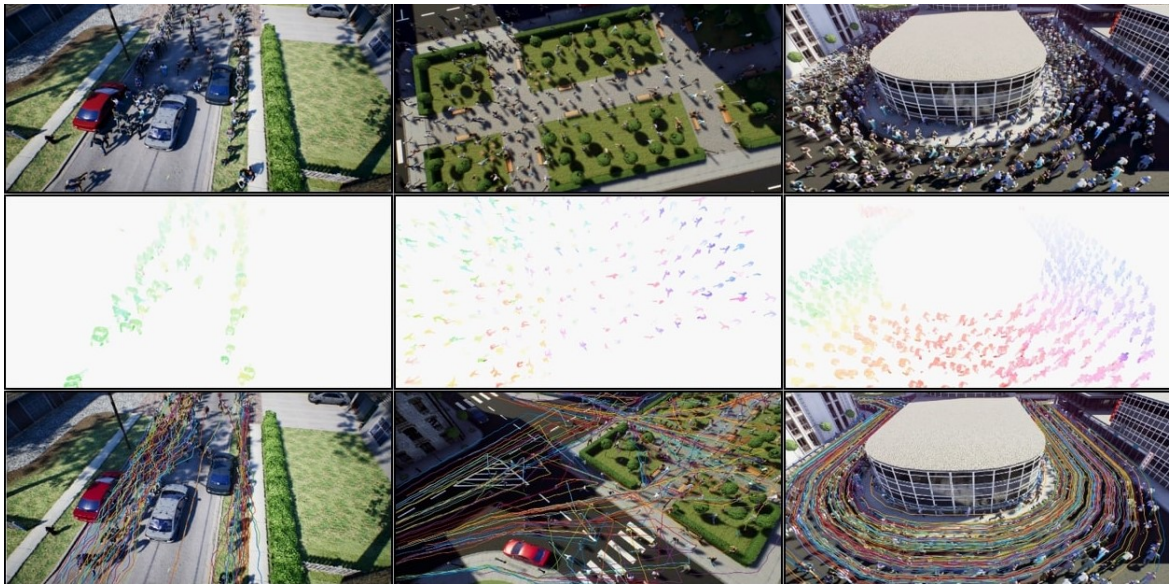


Figure 4.9 – Example of images from CrowdFlow database [68] with their corresponding ground truths.

4.2.10 CreativeFlow+

Recently, the largest open-access dataset has been released [69]. This database, named CreativeFlow+, was created to generalize the methods of real world computer vision to stylized content. Using the same motion magnitude distribution as Sintel, 124390 training frames and 10031 test frames divided into 3000 sequences were simulated with their corresponding ground truth.

4.3 Open-access 2D cardiac ultrasound images

4.3.1 Global context of echocardiographic simulation

In the field of cardiac ultrasound, several simulations of ultrasound sequences have been produced in recent years, mainly to evaluate segmentation or tracking algorithms. Despite the potential lack of realism, these have the advantage of known and controlled geometries and movements, qualities required for benchmarking or evaluating an algorithm. Simulation methods of echocardiographic images must generate a cardiac movement close to reality while synthesizing the particular texture of ultrasound images. Most existing approaches to echocardiographic image simulation have focused on the myocardium. They usually consist of at least two models: one for myocardial anatomy and one for myocardial motion. There exists mainly three different types of heart motion models in the literature:

- derived from the mathematical modeling of cardiac motion [70].
- derived from the cardiac motion observed in another modality such as in 3D cine-MR images. These models are named kinematic [71], [72].
- derived from the physics of the observed phenomena as well as from the modeling of the physiology of the heart. Among the most successful models of this type, we can mention the electromechanical model (EM) proposed in [73]. In this approach, cardiac structures are derived from segmented 3D cine-MRI images, while motion is generated by modeling both electrical activation and the induced myocardial contraction.

These models can be used to customize myocardial motion during simulations. This allows in particular to simulate pathological cases. Most of the proposed simulations were performed in 3D to better capture the complex movement of the heart [71], [74].

The first simulations mainly focused on generating realistic heart motion, without worrying too much about the visual quality of the synthetic data generated. The main weakness was too place to much emphasis on myocardial simulation and not enough on surrounding structures, such as valves or papillary muscles [75]. To solve this problem, a new family of image-based simulation methods has emerged. In [76], the authors proposed to use a real acquisition to *i)* directly compute a reference motion of the myocardium from an optical flow technique and *ii)* compute the reflection coefficients of the acoustic scatterers used as input data in a physical simulator to generate ultrasound sequences. Although this technique greatly improves the degree of realism of the simulated images in terms of ultrasonic texture, the movement of the myocardium suffers from a bias towards the method used and can be physiologically

incorrect. In [77], the authors used a warping strategy to deform a real sequence through the motion of an EM model. This allows the introduction of a controlled myocardial model while maintaining a certain degree of image realism. Unfortunately, this method creates several artifacts due to the successive warpings and unrealistic speckle motion due to the absence of speckle decorrelation.

Among all the advances of the last few years, one approach has emerged as a breakthrough by effectively combining key concepts from state-of-the-art solutions to generate realistic 3D ultrasound simulations [78]. A schematic diagram of this technique is given in Figure 4.10. The proposed pipeline consists of a real sequence to synthesize a realistic ultrasound texture, an EM model to generate controlled myocardial motion, and a physical simulator named COLE to guarantee acoustic properties. COLE is based on an efficient implementation of a convolution operation with a spatially varying point spread function to simulate 3D volumes in few minutes (while the other classical physical simulators like Field II need several hours.) [79]. By doing so, the authors managed to create realistic ultrasound image content (e.g. presence of valves, trabeculae, papillary muscles) with speckle decorrelation and without warping artifacts. Based on this technique, the authors provided the access to 8 simulated sequences, a first iteration towards open-access synthetic images to validate motion estimation in echocardiography.

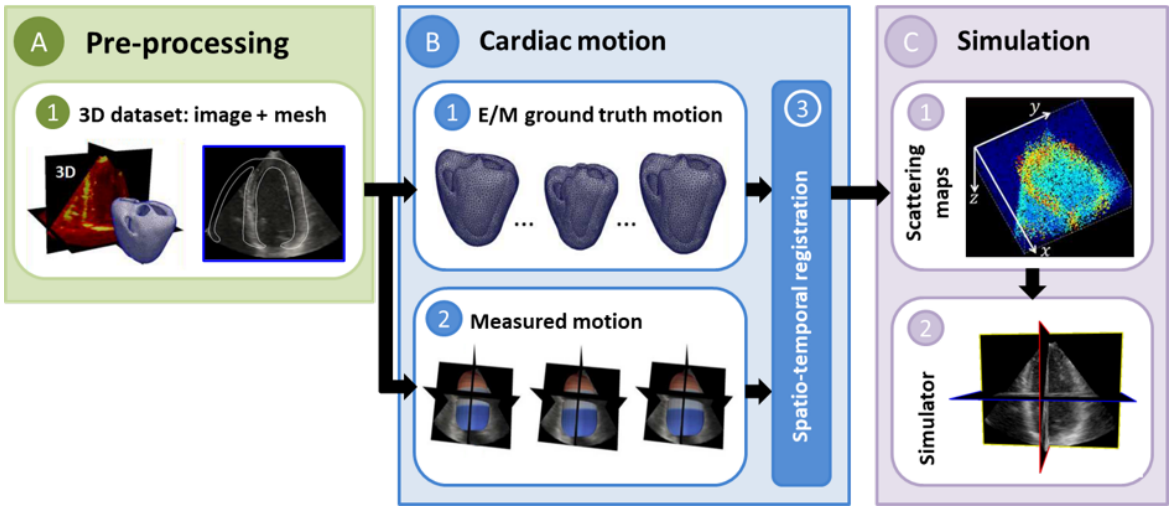


Figure 4.10 – Schematic view of the pipeline developed by [78] to generate realistic ultrasound simulations from real images.

The same simulation pipeline was recently extended to MRI with an improved transition between the myocardium and the surrounding structures to make the simulation even more realistic. A multi-modal open access database of 90 sequences of cardiac volumes from 18 virtual patients with ultrasound, cine MRI and tagged MRI sequences was made available [80].

Most echocardiographic image simulation methods have focused on 3D data. However, 2D echocardiography remains the most commonly used modality in clinical routine. It is therefore necessary to adapt the 3D simulator to be able to generate synthetic 2D sequences. Unfortunately, this step is not straightforward. Indeed, selecting 2D slices from simulated echocardiographic volumes does not allow to obtain realistic sequences because the image resolution thus obtained will be lower than the one of real 2D acquisitions. This is due to the use of ultrasound probes with different

characteristics between 3D and 2D.

In [81], 2D apical sequences based on a kinematic model with cardiac contraction were simulated. These simulations have a great diversity thanks to the use of four different contraction models, several noise levels and numerous scatterer positions. However, a simple strategy was used to simulate the speckle information, resulting in high unrealistic contrast between the myocardium and the surrounding structures. Moreover, the simulations do not include key structures such as the valves, the pericardium, or the papillary muscles. Finally, the kinematic models used are quite simple and do not represent the complexity of physiological contractions of the heart.

To solve these problems, an extension of the 3D ultrasound pipeline proposed in [78] was developed [82]. More specifically, two real sequences were used: one 3D acquisition to perform a registration with the EM model and one 2D acquisition to synthesize realistic 2D ultrasound texture. Moreover, some improvements on the management of the acoustic scatterers were made to better control the degree of speckle decorrelation during the myocardial motion. Based on this framework, the authors created a new open access database composed of 2D apical two-three-four chamber view sequences for seven vendors and five different motion patterns, including one healthy and four pathologies. This resulted in a dataset composed of 6,060 pairs of synthetic ultrasound images with the corresponding myocardial displacement fields. This was the only open-access database of 2D echocardiographic sequences that was existing at the beginning of this PhD. More details are given in Section 4.3.2.

Very recently, Sun *et al.* developed the first simulation pipeline for the simulation of color Doppler echocardiography. This approach includes reference blood flows derived from a computational fluid dynamics model proposed in [83]. The authors used the physical simulator named SIMUS [84] to generate a realistic B-mode texture and synthetic signals with the presence of typical artifacts such as velocity aliasing and clutter noise from wall motion.

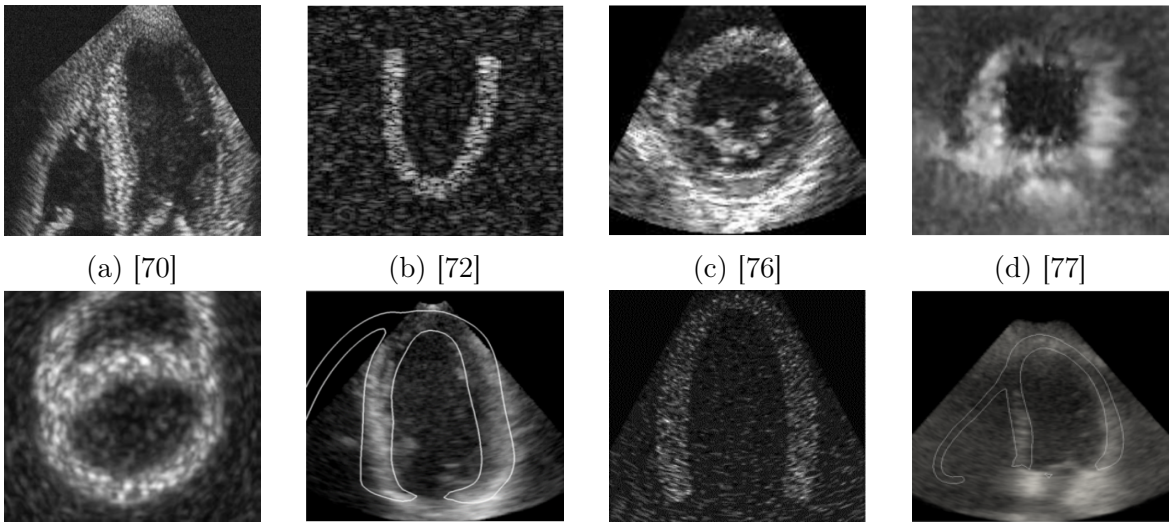
In parallel with the development of the ultrasound simulation pipeline, there has been increasing interest in designing new physical simulators with their own characteristics. For instance, Garcia recently developed the SIMUS simulator based on the MUST open-access toolbox [1] [85]. The interest of this software is to express the laws of ultrasound physics in the Fourier domain, which allows to efficiently integrate different key concepts. Perdios *et al.* developed an analytical version of the ultrasound principles expressed in the time domain through the B-Spline formalism. Based on an efficient GPU implementation, their solution allows a drastic reduction in simulation time. Their simulator was then used to generate a large amount of synthetic ultrasound images to train deep learning methods for high quality ultrafast image reconstruction (i.e. ultrasound images obtained from the transmission of a few plan-waves).

Finally, the deep learning community has also recently investigated the ability of generative adversarial networks (GANs) to generate 2D echocardiographic images in apical view from simple binary input masks [86], [87]. While the results obtained on single images are very encouraging, no study has yet addressed the problem of simulating a complete echocardiographic sequence with realistic motion references.

4.3.2 Existing open-access 2D echocardiographic database

As described in the previous section, there existed at the beginning of this thesis only

[1] `www.biomecardio.com/MUST`



(a) [70] (b) [72] (c) [76] (d) [77]

(e) [75] (f) [78] (g) [81] (h) [80]

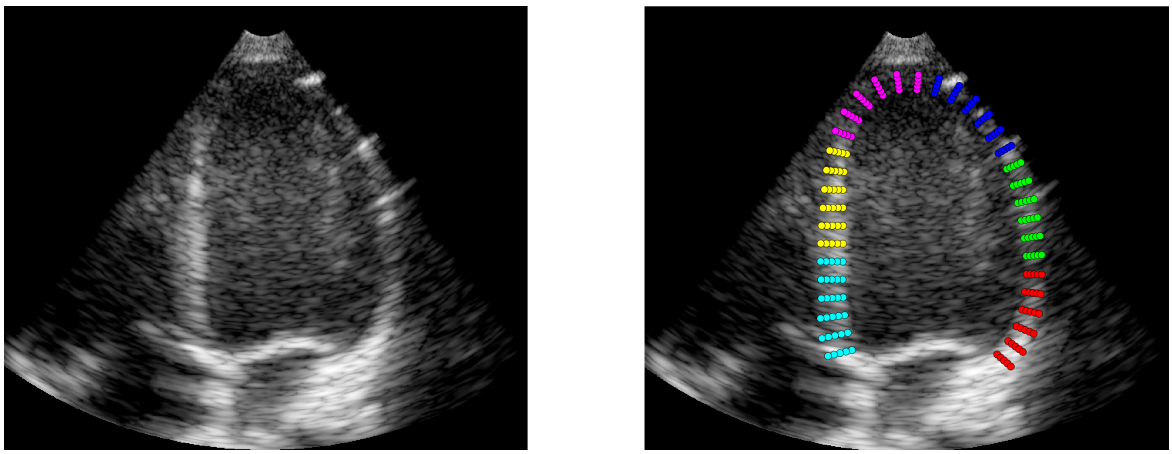
Figure 4.11 – Examples of simulated ultrasound data in several articles.

one open-access database with realistic 2D echocardiographic sequences [82]. This database consists of 105 simulated sequences and includes:

- ultrasound systems from seven different vendors: i.e. Esaote, General Electric, Philips, Siemens, Samsung, Toshiba, and Hitachi.
- three different echocardiographic views per vendor: apical 2-chambers, apical 3-chambers, and apical 4-chambers.
- five motion patterns: one healthy and four ischemic ones, i.e. occlusion of the right coronary artery, occlusion of the left circumflex, and distal and proximal occlusions of the left anterior descending artery.

As mentioned in the previous section, the pipeline used to simulate this database used a real acquisition to synthesise realistic ultrasound texture. In this context, a single healthy subject per vendor was acquired for the three different views. Pathological behaviors were then simulated thanks to the EM model. In particular, specific pathological motion patterns were applied on targeted cardiac segments defined from the standard model of the heart of the American Heart Association (AHA) [7]. A total of 6,060 pairs of simulated 2D echocardiographic images were generated, along with the associated myocardial reference fields. These fields can be derived from a sequence of meshes. Each mesh is composed of 36 intramoycardial segments that join the epicardial and endocardial contours. Each intramyocardial segment contained 5 distinct points, resulting in a final mesh with 180 points (36 longitudinal $\times$ 5 radial) and 280 triangle cells. These points were grouped into 6 subsets according to the AHA representation, as illustrated in Figure 4.12. This strategy allows the analysis of strain both locally and globally.

Despite a high degree of realism of the simulated images, this database suffers from a main limitation. Indeed, the pipeline relies on a personalization procedure that consists of adjusting the various parameters of the model so that its shape matches the anatomy present in the actual sequence. This personalization operation remains tedious, and currently limits the deployment of such scheme to small dataset (*i.e.*



(a) (b)

Figure 4.12 – Example of realistic simulated data from an Esaote sequence in A4C view with 4.12b and without mesh 4.12a.

number of patients lower than 10 with the same kind of heart motion) with synthetic myocardial deformations that remain low as compared to reported normality ranges (*e.g.* simulated peak systolic longitudinal strain lower than 10% instead of 20% in real cases for healthy subjects).

4.4 Conclusions

In this chapter, we have shown that many simulated databases with reference fields have been proposed in the literature to evaluate motion estimation algorithms. These databases are essential since it is impossible to manually annotate dense motion fields on real images. Thanks to a growing number of simulated cases, some of these databases can also be used to train deep learning algorithms, the most interesting being the FlyingChairs2D and FlyingThings3D databases.

Regarding echocardiographic data, there existed only one open access database of realistic synthetic 2D sequences at the beginning of this thesis. Despite the high degree of realism of the synthetic ultrasound images, this database suffers from the following drawbacks: *i)* there is a lack of diversity since the simulated cases were generated from only 7 real subjects; *ii)* because of the personalization operation, the full pipeline is difficult to deploy on many patients, restricting the number of the simulated virtual patients; *iii)* the actual personalization procedure also limits the range of the synthetic myocardial deformations which remain low as compared to real cases. This revealed the need for more complete synthetic databases in echocardiography with a wider variety of myocardial shapes and movements in order to be used for the training of deep learning algorithms.

Part III

Contributions

Chapter 5

Assessment of the ability of CNNs to estimate displacement in 2D ultrasound imaging

Contents

This chapter is an extension of our article "A pilot study on convolutional neural networks for motion estimation from ultrasound images" published in IEEE Transactions on Ultrasonics, Ferroelectrics, and Frequency Control [88].

5.1 Introduction

The rise of deep learning has been a breakthrough in computer vision and has improved performance in terms of accuracy and speed. As it has been discussed in Chapter 3, motion estimation is no exception. In this field, deep learning methods have also significantly improved the performance obtained compared to traditional methods. At the time of this first contribution, FlowNets [29], [31] were among the best networks for motion estimation on natural images. Thus, in the few papers using neural networks to estimate motion in ultrasound, only U-Net based networks were used. In [53], FlowNet2 was embedded with its pre-trained weights (as provided in [31]) to estimate coarse displacements in the context of elastography. Similarly, FlowNet2 was applied as such in [89] with a view classification and semantic partitioning of the mycoardium for quantifying longitudinal deformation. In [52], the authors used the first branch of the FlowNet2 architecture with transfer learning from a simulated dataset to retrieve displacements in ultrasound breast imaging for elastography. While this pioneer study reveals the value of adapting deep learning solutions for motion estimation in ultrasound, it is limited to the direct use of a branch of an existing network without any evaluation on the architecture for the targeted application. Moreover, this study focused on estimating strain from relatively small displacements, where the decorrelation of speckle is limited. Based on this literature review, we conducted a pilot study to answer the following three questions:

1. How do different CNN architectures compare on a given set of echo images in terms of motion accuracy?

2. How does it compare to non-DL algorithms that have been designed for ultrasound?

3. What is the gain brought by transfer learning, *i.e.* by re-training the weights of a network already pre-trained on natural (video) scene sequences.

For that purpose, we focused on synthetic and *in-vitro* datasets involving controlled motion fields (in our case rotations) on a simplified geometry - a disk. Assessing accuracy on these images before and after transfer learning on simulated data, for a number of FlowNet2-based networks, allowed for an accurate quantitative comparison of these networks. Moreover, we decided to focus on rigid rotations as these motions are important sources of speckle decorrelation in ultrasound, making it particularly difficult to estimate actual motion from apparent displacement (*i.e.* motion measured from the image itself). Finally, although our dataset involves simpler motion fields and less realistic images than in synthetic echocardiography of [82], our solution has the advantage of providing a reference displacement field over the entire image domain for training DL-based motion estimators.

5.2 Evaluated networks

In this study, we investigated the performance of FlowNet2 in estimating displacements in ultrasound imaging, and that of each individual CNN involved in this architecture, namely FlowNetC, FlowNetS and FlowNetSD. A detailed description of these architectures is provided in Section 3.3.1.1. Recently, Cai *et al.* used the FlowNet-SD architecture for the estimation of particle displacements in velocimetry imaging [90]. They showed that it was possible to improve the overall accuracy of this network by replacing the bilinear interpolation at the end of the expansion part with two additional upsampling layers, leading to a more classical U-Net-like architecture. This was justified by the fact that small displacements, especially sub-pixel displacements, would not be sufficiently addressed by bilinear interpolation. Inspired by this work, we also investigated the influence of replacing the bilinear interpolation with two upsampling layers for the FlowNetS and FlowNetSD architectures, leading to two modified networks referred to as FlowNetS* and FlowNetSD* in the following. We thus investigated the performance of 6 different networks for estimating motion in ultrasound imaging: FlowNet2, FlowNetC, FlowNetS, FlowNetS*, FlowNetSD and FlowNetSD*.

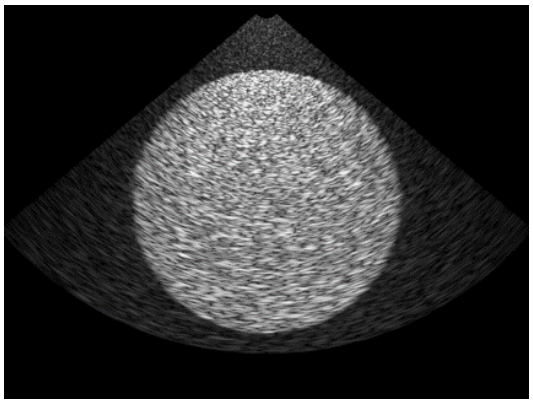
5.3 Simulated & In-Vitro dataset

All networks evaluated in this study were trained in a supervised manner, which required setting up an ultrasound dataset with reference motion fields. To this end, we created a dataset consisting of synthetic and *in-vitro* data with known motion. We used B-mode images after scan conversion and log-compression since, in practice, they are the only data that can be retrieved from clinical ultrasound scanners and they require low storage compared with RF data. In particular, we worked on a spinning disk scenario, where the amount of displacement is well-controlled, both in simulations and *in-vitro* experiments. This strategy allowed us to design simulated and real image sequences with similar displacements and image intensities. It also targets a well-known challenge in the field of ultrasound, as it may be difficult to recover accurate rotations due to the greater speckle decorrelation it induces compared with translations. Working with a combination of simulated and *in-vitro* data allowed us to assess accuracy and robustness. Regarding accuracy, the simulated data showed the value of transfer learning for specializing the different networks to ultrasound, as described in Section 5.4. Regarding robustness, the *in-vitro* data allowed us to evaluate how several networks that were trained on synthetic data performed on real ultrasound images. Examples of *in-vitro* and simulated images, with the corresponding reference motion fields are provided in Figure 5.1. The full database was made available for download at the following link[1].

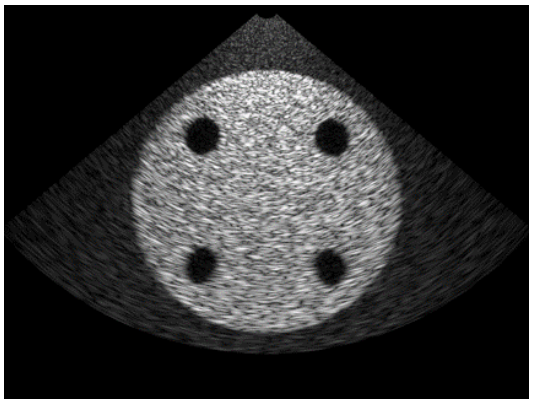
5.3.1 In-Vitro Data

We re-processed the *in-vitro* data described in [91]. These images were acquired with a Verasonics research scanner (V-1-128, Verasonics Inc., Redmond,WA) and a 2.5 MHz phased-array transducer (ATL P4–2, 64 elements) on a agar-based disk phantom with incremental angular velocities. 32 diverging waves with a triangle steering strategy

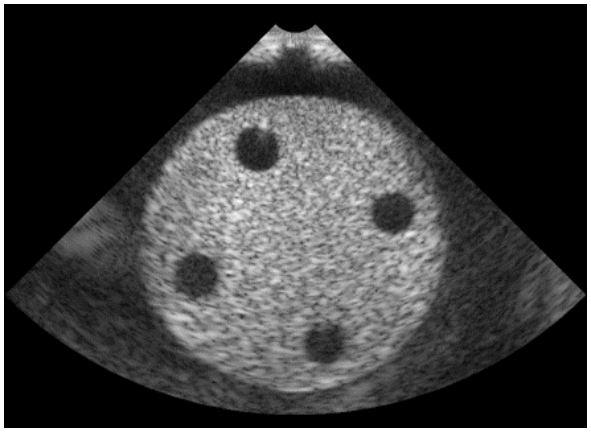
[1] `http://humanheart-project.creatis.insa-lyon.fr/revolus.html`



(a) Simulated disk without anechoic cysts.



(b) Simulated disk with anechoic cysts.



(c) In-vitro disk from [91].

Figure 5.1 – Example of simulated (a,b) and in-vitro (c) spinning disks.

were emitted to reconstruct one single image based on a dedicated delay and sum technique. The disk had four anechoic cysts positioned symmetrically with respect to its center, as shown in Figure 5.1c. Each sequence was composed of 293 frames with angular velocities from 1 to 5 rad/s. From the B-mode images delivered into a polar coordinate system, we reconstructed all the B-mode images into a Cartesian coordinate system on a uniform grid 441 $\times$ 321 with a pixel area of 0.45 mm^2. Each sequence was reconstructed with a frame rate of 312 Hz. In our experiments, we used pairs of images separated by 6 frames in order to work with a frame rate of 52 Hz. This temporal and spatial imaging resolutions are similar to the ones classically used in clinical echocardiography practice. Based on the range of the angular velocities, this meant estimating displacements between 0 (at the center of the disk) and 11 pixels (taking into account the spatial resolution of the grid) between two images.

5.3.2 Simulated Data

Synthetic images of a spinning disk were generated using the ultrasound simulator proposed in [92]. The same acquisition protocol as the one used to acquire the *in-vitro* data was simulated. Foreground scatterers were randomly positioned inside a disk and moved at angular velocities ranging from 1 to 5 rad/s. The full synthetic dataset was composed of 14300 pairs of images with the corresponding reference displacement

fields per pixel.

For each angular velocity, 10 sequences of 287 images were simulated: 5 sequences with a homogeneous disk at the center of the image, and 5 including four anechoic cysts placed symmetrically with respect to the center of the disk, as illustrated in Figure 5.1. For each sequence, background and foreground scatterers were randomly distributed on the first frame, leading to different speckle textures over the dataset. The spinning disk involved in the *in-vitro* experiment corresponds to an agar phantom immersed in water. There is therefore an intrinsic difference in terms of backscattering coefficient of the acoustic wave, and thus in the reconstructed B-mode images, between the background (water) and the object (disk). We experimentally fixed a ratio of 8 between the backscattered coefficients of the foreground and background scatterers involved in our simulations so to generate B-mode images with similar intensity histograms to the ones of the in-vitro dataset (see Figure 5.2). As for the *in-vitro* data, all the simulated B-mode images were reconstructed into a Cartesian coordinate system on a uniform grid 428×321 with a pixel area of 0.47 mm^2. The frame rate was set at 312 Hz and pairs of images separated by 6 frames were used for motion estimation. To further analyze the robustness of the resulting neural network, we also generated four sequences by angular velocity for each type of disk (with and without anechoic cysts) in which the disk is shifted by +/- 30 pixels in both lateral and axial directions, as illustrated in Figure 5.3.

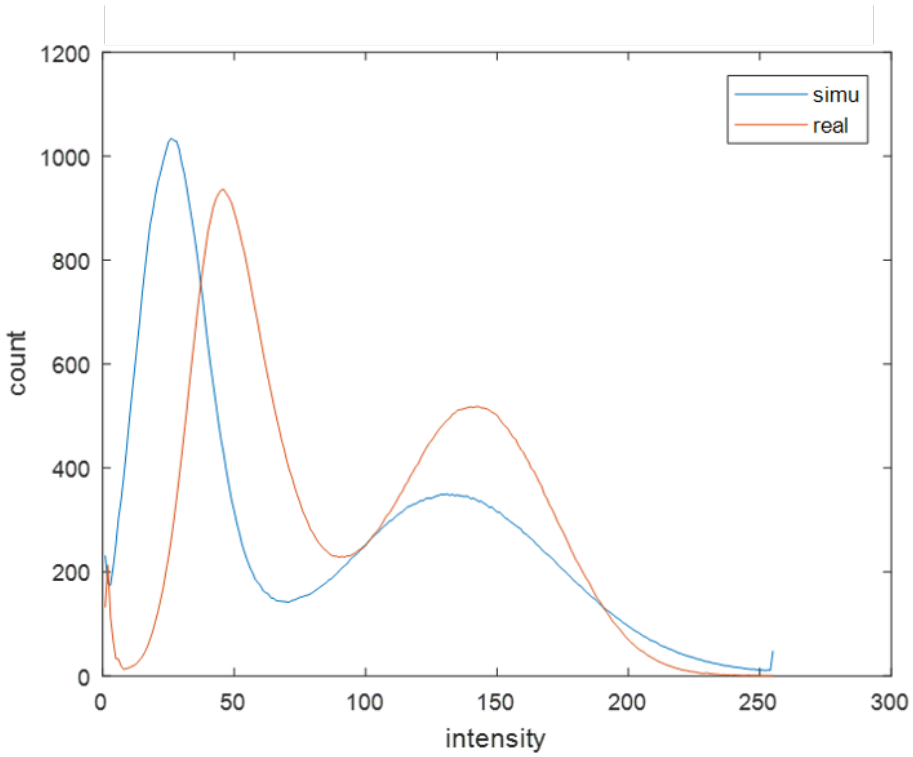
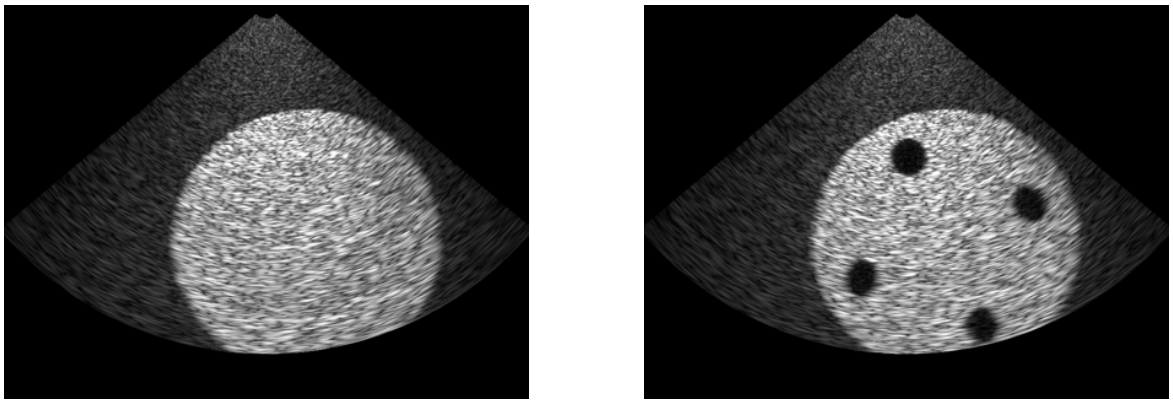


Figure 5.2 – Histogram of B-mode intensities of in-vitro (red) and simulated (blue) images.



(a) Shifted disk simulated without anechoic cysts.

(b) Shifted disk simulated with anechoic cysts.

Figure 5.3 – Example of simulations of shifted rotating disks with (b) and without (a) anechoic cysts.

5.4 Transfer Learning

This section provides details on the transfer learning strategy that we applied to specialize the networks described in Section 5.2 in ultrasound images. In particular, starting from the networks weights learned from several synthetic sequences provided by the FlowNet2 authors [31], we launched a new learning procedure independently for each network based exclusively on the 14300 synthetic ultrasound data described in Section 5.3.2.

5.4.1 Loss functions

To perform transfer learning on the FlowNet2 architecture, we followed the recommendations of [31]. From the network already trained from several synthetic sequences, we restricted the update procedure to the weights belonging to the merging branch during the learning phase (fusion network displayed in the bottom part of Figure 3.4). For this purpose, the L_{pq} loss function with $p = 2$ and $q = 0.2$ was used, as described in Section 3.3.1. To perform transfer learning on the remaining networks described in Section 5.2, we opted for the EndPoint Error (EPE) as loss (Equation 3.5). For FlowNetC, FlowNetS and FlowNetSD architectures, the actual loss function corresponded to the weighted sum of the EPE calculated at different resolution levels in the expansion part of the networks as in [31]. The values of the weights involved in this loss function were the same than those chosen experimentally in [31]. For FlowNetS* and FlowNetSD*, since two additional layers were inserted at the end of the expansion part of the network, it was necessary to adapt the loss function by adding two EPE terms computed from each new resolution. The corresponding weights were the same as those proposed in [90], thus adding more importance to the last layers.

5.4.2 Hyper-parameters

An initial learning rate of $\lambda = 1e - 4$ was set experimentally. Moreover, as recommended in the original paper [31] and described in Section 3.3.1.2, λ was then divided

by two after reaching 40% (and every 20%) of the total number of epochs. This procedure was performed to ease convergence of the optimization process. Indeed, this scheme allows for fast learning at the beginning of the training process, making gradually smaller updates over the course of the optimization to refine the weights. The Adam optimizer [34] was used in this study and a batch size of 4 pairs of images was chosen mainly for reasons of memory capacity.

5.4.3 Dataset split

The full set of the synthetic data was divided into three folds: 60% for the training set, 20% for the validation set and 20% for the test set. Moreover, each fold contained data for the entire angular velocity range (*i.e.* from 1 to 5 rad/s). The validation set was used to select the most efficient weights of the deep learning architectures during the training process, while the test set was used to produce all the results given in this paper. It is important to note that all algorithms were tested only once on the test set and that no optimization was performed on it in order to avoid overfitting. For each velocity, a balanced selection of sequences with and without anechoic cysts was realized. Data from a whole sequence were not mixed across the 3 folds. In particular, 6 full sequences were chosen for the training phase, 2 full sequences for the validation phase and 2 full sequences for the testing phase. This procedure ensured that each fold contained the same diversity of information without introducing any bias.

5.4.4 Data Augmentation

To avoid overfitting, data augmentation was realized following the procedure described in [29], [31], [93] detailed in Section 3.3.1.2. The deployed augmentation integrated some typical alterations that could happen in ultrasound, *i.e.* variations in brightness, saturation and contrast. We involved variability in terms of translation, Gaussian noise, black borders and cut-outs, as they are reported to improve the generalization of networks [94]. Concerning the additive noise, we used a Gaussian noise with small variance in order to remain close to the original images while increasing the network robustness against intensity fluctuations. We also added random cropping to increase the variability of the dataset and reduce the dimensions of the images to 384×320 pixels to respect the input image size of the different networks described in Section 5.2. Finally, to ensure that the networks did not overspecialize in a single direction of rotation, pairs of images were randomly flipped with a probability of 0.5. In this way, both clockwise and counterclockwise rotations were equally included during the training process.

5.5 Evaluation protocol

5.5.1 Metrics

Performance of all evaluated networks was assessed through: *i)* the EPE metric described in Section 5.4.1 and computed inside the disk only; *ii)* the error on the angular velocity, which was computed by dividing at each point inside the disk the estimated velocity magnitude by the distance to the disk center. We reported the distribution

of each of these metrics using the median and the median absolute deviation (MAD) defined as

$$MAD = median(|X - median(X)|)\,, \tag{5.1}$$

where X stands for the distribution of the metric values inside the disk. MAD was preferred to the standard deviation as it is more robust to outliers.

5.5.2 State-of-the-art method

We compared all CNN networks described in Section 5.2 with a state-of-the-art Particle Imaging Velocimetry (PIV) method. We decided to use this block-matching algorithm since it won the challenge on synthetic aperture - vector flow imaging organized during the International Ultrasonic Symposium in 2018 [26]. As detailed in Section 3.2.2, PIV is a block-matching algorithm that worked with ensembles of n consecutive images, under the assumption that the motion remained unchanged during that temporal window, to calculate the average of $n-1$ cross correlation matrices (ensemble correlation, see [95]). Peak detection of the averaged normalized cross correlation provided the displacements with a pixel precision. Subpixel precision of the displacement estimates was then obtained through parabolic peak fitting of the cross correlation. Taking into account the physical properties of the ultrasound images involved in our experiments, we applied a multiscale strategy to estimate motion by using 3 different sizes of search areas, namely 24×24, 16×16 and 12×12 pixels. These values span a search range from 3 to 6 times the speckle size and were experimentally tuned to obtain the best results. We also verified that adding a larger window did not improve the results and that the multiscale strategy returned the best results regardless of the frame rate.

Experimentally, we observed that PIV produced erroneous results when the displacement was either too low (average displacements lower than one pixel) or too large (average displacements higher than ten pixels) between two consecutive frames. For this reason, we applied the PIV algorithm under two different conditions:

- by using pairs of images corresponding to a frame rate of 52 Hz and used to train the different networks (referred to as PIV);
- by reducing the number of frames that separate two images of a pair in order to adapt the frame rate for each angular velocity and taking 16 images into account to obtain the best possible results (referred to as PIV-adapt).

In practice, the higher the angular velocity, the higher the optimal frame rate chosen, up to 312 Hz for a velocity of 5 rad/s. This way of proceeding gives natural bounds of PIV accuracy, as giving to PIV data at a higher temporal resolution than the evaluated CNNs gives an *upper bound* of PIV's accuracy.

5.6 Accuracy benchmarks

5.6.1 Network Selection

We first investigated the performance of the six selected networks on the test set of the simulated dataset. Table 5.1 summarized all the results obtained during this phase. For each network, the median EPE, the estimated angular velocity and the MAD

Table 5.1 – Median EPE, estimated angular velocity and MAD dispersion values computed inside the spinning disk on the synthetic dataset from the networks described in the Section 5.2. For each network, different learning strategies were assessed (✓ transfer learning, - pre-trained weights from natural scene images; ✗ random initialization). The best scores for each category are highlighted in bold while the overall best network is shaded.

Methods	**TL**	**1 rad/s**		**2 rad/s**		**3 rad/s**		**4 rad/s**		**5 rad/s**	
		EPE	*Velocity*	*EPE*	*Velocity*	*EPE*	*Velocity*	*EPE*	*Velocity*	*EPE*	*Velocity*
		px.	rad/s.	px.	rad/s.	px.	rad/s.	px.	rad/s.	px.	rad/s.
FlowNet2	✓	1.4 ± 0.4	0.0 ± 0.0	1.6 ± 0.8	0.2 ± 0.2	2.8 ± 1.1	0.6 ± 0.5	4.5 ± 1.5	0.4 ± 0.3	6.1 ± 1.9	0.2 ± 0.2
	-	0.2 ±0.1	0.9 ± 0.1	0.5 ± 0.2	1.7 ± 0.1	2.0 ± 0.6	1.6 ± 0.2	4.3 ± 1.3	1.0 ± 0.4	6.2 ± 2.0	0.6 ± 0.4
FlowNetC	✓	5.3 ±2.6	4.0 ± 2.4	6.5 ± 3.0	4.6 ± 2.7	7.8 ± 3.4	5.2 ± 3.1	8.7 ±3.5	5.3 ±3.1	9.6 ±3.5	5.3 ±3.1
	✗	1.5 ±0.5	0.1 ±0.1	2.9 ±0.8	0.1 ±0.1	4.4 ±1.1	0.1 ±0.1	5.9 ±1.5	0.1 ±0.1	7.3 ±1.9	0.1 ±0.1
	-	1.0 ±0.3	0.4 ±0.2	2.2 ±0.6	0.6 ±0.2	3.7 ±1.0	0.7 ±0.3	5.0 ±1.4	0.8 ±0.3	6.4 ±1.8	0.8 ±0.4
FlowNetS	✓	0.2 ±0.1	0.9 ±0.1	**0.3** ±0.1	**1.9** ±0.1	0.5 ±0.2	2.7 ±0.2	1.1 ±0.4	3.2 ±0.2	2.2 ±0.6	3.4 ±0.2
	✗	1.2 ±0.5	**1.0** ±0.5	1.5 ±0.7	1.3 ±0.5	2.6 ±1.1	1.5 ±0.5	3.8 ±1.5	1.7 ±0.6	4.9 ±1.8	1.9 ±0.6
	-	1.0 ±0.3	0.6 ±0.2	1.9 ±0.6	0.9 ±0.3	3.3 ±1.0	1.0 ±0.4	4.7 ±1.5	1.2 ±0.5	6.0 ±1.8	1.4 ±0.6
FlowNetS*	✓	2.0 ±1.0	1.2 ±0.8	3.2 ±1.3	1.1 ±0.8	4.6 ±1.6	1.2 ±0.8	5.9 ±1.9	1.2 ±0.8	7.3 ±2.2	1.2 ±0.8
	✗	1.9 ±0.7	1.1 ±0.7	3.1 ±1.0	1.1 ±0.7	4.5 ±1.3	1.1 ±0.7	5.9 ±1.6	1.2 ±0.7	7.4 ±1.9	1.4 ±0.8
FlowNetSD	✓	**0.1** ±0.0	1.1 ±0.0	0.4 ±0.1	2.2 ±0.0	**0.4** ±0.1	**3.2** ±0.1	**0.3** ±0.1	**4.2** ±0.1	**0.4** ±0.1	**4.9** ±0.1
	✗	1.6 ±0.4	0.3 ±0.2	3.3 ±0.7	0.6 ±0.3	4.5 ±1.0	0.6 ±0.3	5.8 ±1.4	0.7 ±0.4	7.3 ±1.7	0.8 ±0.4
	-	0.2 ±0.1	0.9 ±0.1	0.5 ±0.2	1.8 ±0.1	1.5 ±0.5	2.1 ±0.3	3.5 ±1.0	1.7 ±0.3	5.0 ±1.4	1.7 ±0.4
FlowNetSD*	✓	**0.1** ±0.0	1.1 ±0.0	0.5 ±0.1	2.3 ±0.0	0.6 ±0.2	3.4 ±0.0	0.6 ±0.2	4.4 ±0.0	0.4 ±0.1	5.2 ±0.1
	✗	1.4 ±0.4	0.1 ±0.1	2.8 ±0.7	0.1 ±0.1	4.3 ±1.1	0.1 ±0.1	5.8 ±1.4	0.1 ±0.1	7.2 ±1.8	0.1 ±0.1

dispersion were computed inside the spinning disk for the five angular velocities (from 1 to 5 rad/s). Increasing the rotation speed by 1 rad/s leads to an augmentation of the speckle decorrelation and an increment of the maximum motion amplitude by approximately two pixels per radian.

It can be noted that the network architecture has a major influence on the ability to address this tracking problem. For instance, FlowNet-C showed a structural inability to accurately estimate motion in ultrasound imaging. Indeed, for all speeds and for all training strategies, we noticed a large error of estimation of the displacement field which causes a large underestimation (see FlowNet-C used with pre-trained weights or trained from scratch) or overestimation (see FlowNet-C with transfer learning) of the angular velocity.

When the speed is lower than 2 rad/s, many architectures showed their ability to estimate the motion, especially those inspired by a classical U-Net like FlowNet-S and FlowNet-SD or more complex architectures like FlowNet2. At these speeds, the magnitude of the displacement and therefore the speckle decorrelation were low, which allowed the networks trained on natural images to adapt to ultrasound images.

With an appropriate architecture, we saw that the best results were obtained by transfer learning, especially when the angular velocity exceeded 2 rad/s. In this range, only the FlowNet-SD and FlowNet-SD* networks trained on pre-trained weights offered the best results both in terms of EPE and estimated spinning speed. This ability to use the transfer learning to adapt to the ultrasound image characteristics, especially to the large speckle decorrelation generated by the rotational motion and the imposed speed, and to perform on it is illustrated by these performances.

The addition of the two extra layers realized on FlowNetS* and FlowNetSD* did not improve the results for FlowNetSD* and even degraded them for FlowNetS*. This revealed the uselessness of adding these two layers in the context of ultrasound motion estimation and corroborated the choice made by the authors of FlowNet2 to upsample the network outputs by a bilinear interpolation.

Finally, FlowNet-SD trained with transfer learning obtained the best results on the simulated data in terms of EPE for all rotation speeds, except for 2 rad/s for which the difference with the best estimate was 0.1 pixel. This low error on the estimated displacement field results in an accurate estimate of the rotation speed with a maximum error of 0.2 rad/s. For these two metrics, it is important to note that the corresponding MAD is also very low with a maximum at 0.1 pixels and 0.1 rad/s. An example of tracking with this method can be seen in Figure 5.4. This network has been used in the following comparisons with the non-DL methods of the state-of-the-art.

5.6.2 Comparison with non-DL methods

Once the network with the best performance was selected, the comparison with the non-DL methods of the state-of-the-art was conducted. Table 5.2 contained the results of the different methods namely FlowNetSD with transfer learning (FlowNetSD-TL) and the two versions of PIV described in Section 5.5.2.

5.6.2.1 In-silico results

In Table 5.2, PIV's velocity estimation results were consistent with the actual true values up to 2 rad/s, along with EPE errors ≤ 0.2 px and MAD values ≤ 0.1 px. At 3 rad/s, PIV slightly underestimated the velocity with a value of 2.5 rad/s but with higher median EPE (from 0.2 px to 0.9 px) and MAD (from 0.1 px to 0.6 px) values. For higher angular velocities, PIV estimates deteriorated with EPE errors over 6.4 px and an underestimation of the velocity, revealing the limitations of this algorithm for angular velocities higher than 2 rad/s at a frame rate of 52Hz.

Nonetheless, when adapting the frame rate for each angular velocity (thus assuming the true displacement range was known for every input sequence and providing PIV with images at a higher temporal resolution than FlowNetSD), PIV-adapt results became consistent and accurate. Indeed, all estimated angular velocity values were accurate and EPE errors were found to be constant and around 0.2 px. It is interesting to note that FlowNetSD produced slightly worse results but close to the PIV-adapt method, with EPE errors ≤ 0.4 px and angular velocity errors ≤ 0.2 rad/s. However,

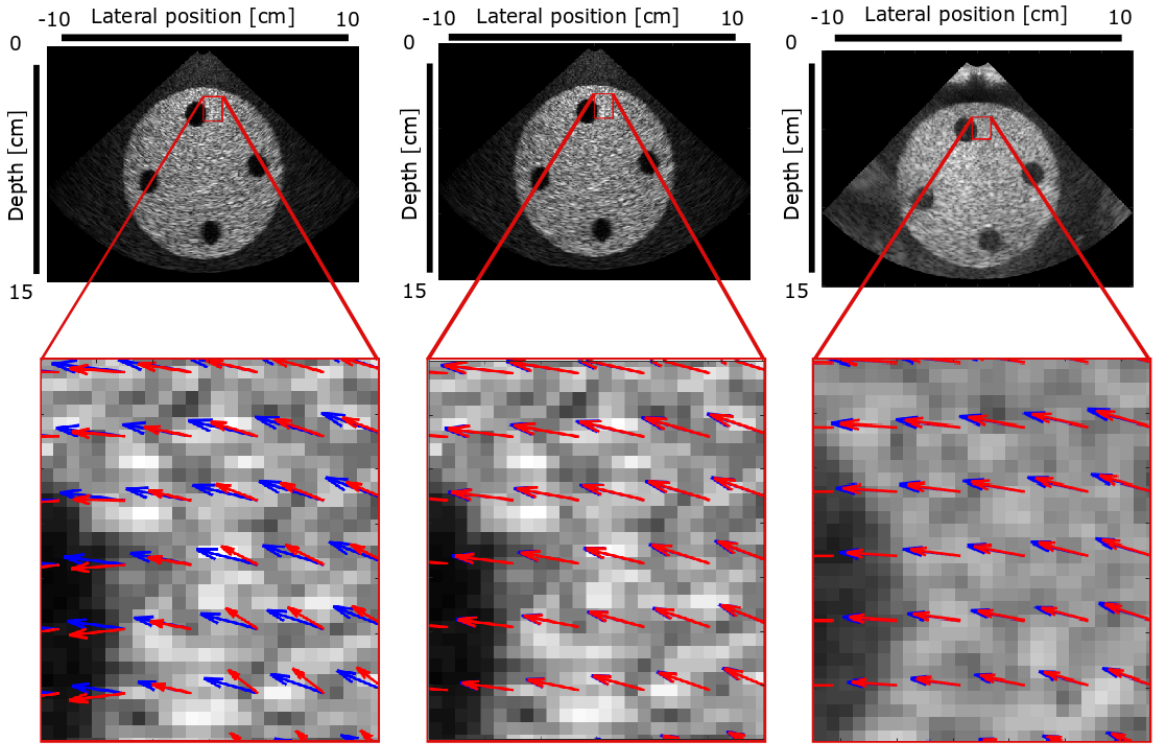


Figure 5.4 – Reference (blue) and estimated (red) motion fields using FlowNetSD on (left) a synthetic image without transfer learning; (middle) the same synthetic image but with transfer learning; (right) an in-vitro image with transfer learning. The displayed data were all extracted from a sequence with a spinning disk rotating at 1 rad/s.

the strong advantage of FlowNetSD is that all the results were obtained with data at the same temporal resolution, *i.e.* at 52 Hz, for all angular velocities. For qualitatively assessing the distribution of errors produced by FlowNetSD, we displayed in Figure 5.5 the spatial distribution of EPE and angular velocity errors obtained for an angular velocity of 1 rad/s. From this figure, it can be seen that the error was uniformly distributed over the entire disk, except at the center and on the edges of the disk, where higher values punctually appeared. Moreover, Table 5.3 showed that the error was uniformly distributed between the axial and lateral components without specific bias.

Regarding the computation time, FlowNetSD had a training time of about 10 hours on an *NVidia* 980Ti GPU and the inference time on a pair of images was 130 ms. PIV algorithms used the CPU only and had different computation times per pair of images: 1.12 s for PIV (n=2, where n is defined in Section 5.5.2) and 5.02 s for PIV-adapt (n=16) with a CPU at 2.3 GHz.

5.6.2.2 In-vitro results

The performance of FlowNetSD-TL was then assessed on the in-vitro dataset described in Section 5.3.1. Results were reported in Table 5.2. First of all, one can observe an increase of the EPE errors with respect to the values obtained on the simulated dataset for all angular velocities. This illustrates the challenge of processing real data

Table 5.2 – Median EPE, estimated angular velocity accuracy and MAD dispersion values computed inside the centered spinning disk on the synthetic dataset (first three rows) and on the in-vitro dataset (last three rows) for five different angular velocities. For this experiment, FlowNetSD with transfer learning (FlowNetSD-TL) was compared with the two versions of the non-deep learning state-of-the-art PIV technique described in Section 5.5.2.

	Methods	**1 rad/s**		**2 rad/s**		**3 rad/s**		**4 rad/s**		**5 rad/s**	
		EPE	*Velocity*	*EPE*	*Velocity*	*EPE*	*Velocity*	*EPE*	*Velocity*	*EPE*	*Velocity*
		px.	rad/s.	px.	rad/s.	px.	rad/s.	px.	rad/s.	px.	rad/s.
In-silico	FlowNetSD-TL	0.1 ±0.0	1.1 ± 0.0	0.4 ± 0.1	2.2 ± 0.0	0.4 ± 0.1	3.2 ± 0.1	0.3 ±0.1	4.2 ±0.1	0.4 ±0.1	4.9 ±0.1
	PIV	0.2 ±0.1	1.0 ± 0.1	0.2 ± 0.1	2.0 ± 0.1	0.9 ± 0.6	2.5 ± 0.6	3.7 ± 2.1	1.4 ± 1.1	6.4 ± 2.5	0.6 ± 1.0
	PIV-adapt	0.2 ± 0.1	1.0 ± 0.1	0.2 ± 0.1	2.0 ± 0.1	0.2 ± 0.1	2.9 ± 0.2	0.2 ± 0.1	3.9 ± 0.2	0.1 ± 0.1	4.9 ± 0.2
In-vitro	FlowNetSD-TL	0.7 ±0.2	1.3 ± 0.1	0.8 ± 0.3	2.3 ± 0.2	1.1 ± 0.4	3.4 ± 0.3	1.0 ±0.4	4.3 ±0.3	0.8 ±0.3	5.0 ±0.4
	PIV	0.4 ± 0.1	1.2 ± 0.1	0.8 ± 0.3	2.3 ± 0.4	2.3 ± 1.5	1.6 ± 1.2	4.8 ± 2.3	0.8 ± 1.1	6.5 ± 2.6	0.5 ± 1.0
	PIV-adapt	0.2 ± 0.1	1.2 ± 0.1	0.3 ± 0.1	2.4 ± 0.2	0.3 ± 0.1	3.6 ± 0.3	0.2 ± 0.1	4.5 ± 0.5	0.2 ± 0.1	5.6 ± 0.6

compared to simulated ones. However, EPE errors remained around or under 1.1 px with MAD values $\leq$ 0.4 px for all angular velocities. This shows a good generalization of our network to real data, despite the fact of being trained solely on simulated data. FlowNetSD-TL also returned angular velocities close to the true values, with a maximum error of 0.4 rad/s for the 3 rad/s velocity. Moreover, MAD values were $\leq$ 0.4 rad/s for all angular velocities.

PIV also produced results that were worse on in-vitro data, with an increase of the EPE error of 0.2 and 0.6 px for the 1 and 2 rad/s velocities, respectively. In line with the simulated case, PIV results degenerated for angular velocities higher than 3 rad/s, with an EPE error of 6.5 px and an angular velocity error of 4.5 rad/s at 5 rad/s. It can thus be observed that FlowNetSD produced more accurate and stable results (in terms of EPE, MAD values and velocity errors) than the PIV method on in-vitro data across the different angular velocities at a frame rate of 52 Hz. Regarding PIV-adapt, results for the velocity estimation were consistent with the true values up to 4 rad/s, along with EPE errors $\leq$ 0.3 px and MAD values $\leq$ 0.1 px. At 5 rad/s, PIV-adapt slightly overestimated the velocity with a value of 5.6 rad/s. Interestingly, even if FlowNetSD works at a lower frame rate of 52 Hz (vs. 312 Hz for PIV-adapt), it produced very close velocity estimates to those of PIV-adapt (mean difference of 0.1 rad/s) for angular velocities $\leq$ 4 rad/s and a better velocity estimate at 5 rad/s. This demonstrates the strong potential FlowNetSD in robustly assessing motion from ultrasound images, even at a lower temporal resolution closer to typical echocardiography values.

5.6.3 Ablation Studies

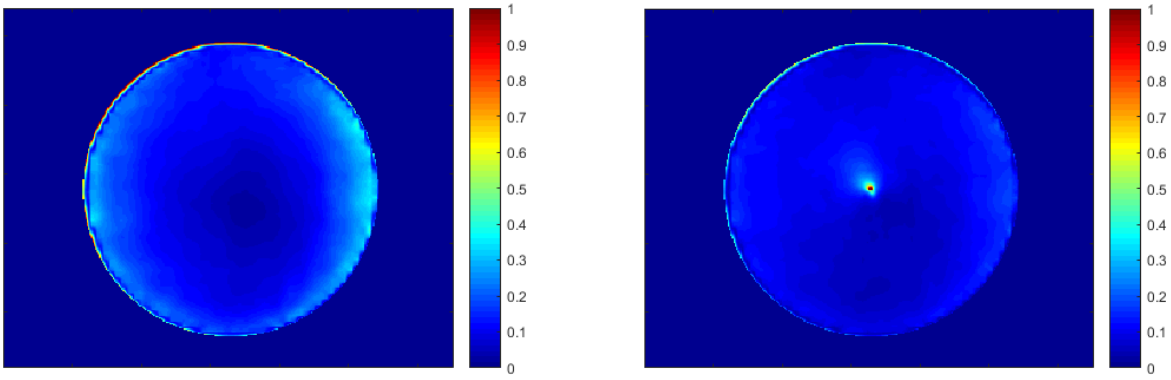
In order to optimize our training, to be able to determine elements allowing a better adaptation of neural networks to ultrasound imaging and to evaluate the robustness of our methods, we have conducted several ablation studies. The first one concerned

	Methods	**1 rad/s**	**2 rad/s**	**3 rad/s**	**4 rad/s**	**5 rad/s**
		rad/s.	rad/s.	rad/s.	rad/s.	rad/s.
In-silico	FlowNetSD-TL	1.1 ±0.0	2.2 ± 0.0	3.2 ± 0.1	4.2 ± 0.1	4.9 ± 0.1
	PIV	1.0 ±0.1	2.0 ± 0.1	2.5 ± 0.6	1.4 ± 1.1	0.6 ± 1.0
	PIV-adapt	1.0 ± 0.1	2.0 ± 0.1	2.9 ± 0.2	3.9 ± 0.2	4.9 ± 0.2
In-vitro	FlowNetSD-TL	1.3 ±0.1	2.3 ± 0.2	3.4 ± 0.3	4.3 ± 0.3	5.0 ± 0.4
	PIV	1.2 ± 0.1	2.3 ± 0.4	1.6 ± 1.2	0.8 ± 1.1	0.5 ± 1.0
	PIV-adapt	1.2 ± 0.1	2.4 ± 0.2	3.6 ± 0.3	4.5 ± 0.5	5.6 ± 0.6

	Methods	**1 rad/s**	**2 rad/s**	**3 rad/s**	**4 rad/s**	**5 rad/s**
		px.	px.	px.	px.	px.
In-silico	FlowNetSD-TL	0.1 ±0.0	0.4 ± 0.1	0.4 ± 0.1	0.3 ± 0.1	0.4 ± 0.1
	PIV	0.2 ±0.1	0.2 ± 0.1	0.9 ± 0.6	3.7 ± 2.1	6.4 ± 2.5
	PIV-adapt	0.2 ± 0.1	0.2 ± 0.1	0.2 ± 0.1	0.2 ± 0.1	0.1 ± 0.1
In-vitro	FlowNetSD-TL	0.7 ±0.2	0.8 ± 0.3	1.1 ± 0.4	1.0 ± 0.4	0.8 ± 0.3
	PIV	0.4 ± 0.1	0.8 ± 0.3	2.3 ± 1.5	4.8 ± 2.3	6.5 ± 2.6
	PIV-adapt	0.2 ± 0.1	0.3 ± 0.1	0.3 ± 0.1	0.2 ± 0.1	0.2 ± 0.1

Table 5.3 – The absolute value of the error between estimated angular velocity and corresponding ground truth analyzed in axial and lateral components on the synthetic dataset (first three rows) and on the in-vitro dataset (last three rows) for five different angular velocities. For this experiment, FlowNetSD with transfer learning (FlowNetSD-TL) was compared with the two versions of the non-deep learning state-of-the-art PIV technique described in Section 5.5.2.

	Methods	1 rad/s		2 rad/s		3 rad/s		4 rad/s		5 rad/s	
		Axial	*Lateral*	*Axial*	*Lateral*	*Axial*	*Lateral*	*Axial*	*Lateral*	*Axial*	*Lateral*
		px.	px.	px.	px.	px.	px.	px.	px.	px.	px.
In-silico	FlowNetSD-TL	0.1	0.1	0.2	0.2	0.2	0.2	0.2	0.2	0.2	0.2
	PIV	0.2	0.1	0.2	0.1	0.6	0.4	2.0	1.8	3.5	3.4
	PIV-adapt	0.2	0.1	0.1	0.1	0.2	0.1	0.2	0.1	0.1	0.1
In-vitro	FlowNetSD-TL	0.4	0.5	0.6	0.6	0.8	0.6	0.7	0.6	0.5	0.5
	PIV	0.2	0.2	0.5	0.4	1.3	1.2	2.6	2.5	3.5	3.5
	PIV-adapt	0.1	0.1	0.2	0.2	0.2	0.2	0.1	0.1	0.1	0.2



(a) EPE error map (Ours). (b) Angular velocity error map.

Figure 5.5 – Mean error maps of FlowNetSD with transfer learning computed (a) from the EPE (expressed in px) and (b) the angular velocity metrics (expressed in rad/s) over the synthetic dataset at 1 rad/s.

the influence of the type of noise to be used for data augmentation during training on ultrasound images. The second one was about the robustness of our method to a major change in the database: the displacement of the center of the rotating disk.

5.6.3.1 Noise addition during training on ultrasound imaging

We ran different experiments with different noise models to examine the impact of this marginal change in the data augmentation procedure on the final tracking accuracy. Two noises have been studied, the Gaussian noise and the Rayleigh noise. The Gaussian noise was already used during the FlowNet initial training (Section 3.3.1.2) and the Rayleigh noise is common in ultrasound imaging before logarithmic compression.

We evaluated the impact of noise on learning in general and whether differences existed between the two types of noise compared, and potentially determined whether one type of noise was more suitable for ultrasound imaging. Table 5.4 summarized the results of the different configurations (Gaussian, Rayleigh and without noise) on both databases. The choice of the model seemed to have a limited impact on the results

Table 5.4 – Ablation study on the influence of noise type in data augmentation applied to ultrasound imaging. Median EPE, estimated angular velocity accuracy and MAD dispersion values computed inside the centered spinning disk on the synthetic dataset (first three rows) and on the in-vitro dataset (last three rows) for five different angular velocities. For this experiment, FlowNetSD with transfer learning (FlowNetSD-TL) was compared with the two versions of the non-deep learning state-of-the-art PIV technique described in Section 5.5.2.

	Methods	**1 rad/s**		**2 rad/s**		**3 rad/s**		**4 rad/s**		**5 rad/s**	
		EPE	*Velocity*	*EPE*	*Velocity*	*EPE*	*Velocity*	*EPE*	*Velocity*	*EPE*	*Velocity*
		px.	rad/s.	px.	rad/s.	px.	rad/s.	px.	rad/s.	px.	rad/s.
In-silico	Gaussian Noise	0.1 ±0.0	1.1 ± 0.0	0.4 ± 0.1	2.2 ± 0.0	0.4 ± 0.1	3.2 ± 0.1	0.3 ±0.1	4.2 ±0.1	0.4 ±0.1	4.9 ±0.1
	Rayleigh Noise	0.1 ±0.0	1.1 ± 0.0	0.3 ± 0.1	2.2 ± 0.1	0.4 ± 0.1	3.3 ± 0.1	0.4 ± 0.1	4.3 ± 0.1	0.3 ± 0.1	5.0 ± 0.1
	No Noise	0.1 ±0.0	1.1 ± 0.0	0.3 ± 0.1	2.2 ± 0.1	0.4 ± 0.1	3.3 ± 0.1	0.4 ± 0.1	4.3 ± 0.1	0.3 ± 0.1	5.0 ± 0.1
In-vitro	Gaussian Noise	0.7 ± 0.2	1.3 ± 0.1	0.8 ± 0.3	2.3 ± 0.2	1.1 ± 0.4	3.4 ± 0.3	1.0 ± 0.4	4.3 ± 0.3	0.8 ± 0.3	5.0 ± 0.4
	Rayleigh Noise	0.5 ± 0.2	1.3 ± 0.1	0.8 ± 0.2	2.4 ± 0.1	1.1 ± 0.3	3.4 ± 0.2	1.1 ± 0.4	4.4 ± 0.2	0.9 ± 0.3	5.1 ± 0.2
	No Noise	0.5 ± 0.1	1.2 ± 0.1	0.8 ± 0.2	2.4 ± 0.1	1.1 ± 0.4	3.5 ± 0.2	1.1 ± 0.4	4.5 ± 0.2	0.9 ± 0.3	5.2 ± 0.3

compared to the case where no noise model was used.

On the simulated data, the three methods always remained at the maximum within 0.1 of each other both in terms of EPE and estimated speed. For each estimate, the MAD was the same between the methods. For example, for a velocity of 4 rad/s, the estimated velocity was 4.2 rad/s with a MAD of 0.1 rad/s for the Gaussian noise and 4.3 rad/s with the same MAD for the two other configurations.

The impact of these changes during training was more easily observed during the transfer to in-vitro data. We noticed that the Gaussian noise took a slight ascendancy over the Rayleigh noise. These two configurations obtained better results than the one without noise during training. Since results are marginally better using the Gaussian noise model, we decided to keep it in the data augmentation procedure.

5.6.3.2 Robustness to non-centered disks

We investigated the ability of our FlowNetSD-TL network to estimate angular velocity for a spinning disk that was *not* centered with respect to the acquisition field of view. This allowed us to verify whether using only a centered dataset during the training phase introduced a bias. This experiment was carried out for a shift of the disk of +/- 30 pixels in both lateral and axial directions for all 5 angular velocities. The obtained results are reported in Table 5.5. Concerning FlowNetSD-TL, we observed that even if the EPE scores were not as good as those obtained from the centered spinning disks, they remained close with a median error between 0.3 and 1.1 pixels and MAD values ≤ 0.3 pixels, showing similar performance as on the in-vitro dataset. In terms of angular velocity estimates, results remained accurate between 1 and 4 rad/s (error ≤ 0.2 rad/s), and with a maximal error of 0.4 rad/s for the 5 rad/s velocity. For all estimated velocities, MAD values remained lower than 0.3 rad/s, revealing a good

Table 5.5 – Median EPE, estimated angular velocity accuracy and MAD dispersion values computed inside the non-centered spinning disk on the synthetic dataset for five different angular velocities. For this experiment, FlowNetSD-TL was compared with the two versions of the non-deep learning state-of-the-art PIV technique described in Section 5.5.2.

Methods	**1 rad/s**		**2 rad/s**		**3 rad/s**		**4 rad/s**		**5 rad/s**	
	EPE	*Velocity*	*EPE*	*Velocity*	*EPE*	*Velocity*	*EPE*	*Velocity*	*EPE*	*Velocity*
	px.	rad/s.	px.	rad/s.	px.	rad/s.	px.	rad/s.	px.	rad/s.
FlowNetSD-TL	0.3 ±0.1	1.0 ± 0.1	0.6 ± 0.2	2.2 ± 0.1	0.8 ± 0.2	3.2 ± 0.1	0.8 ±0.2	4.0 ±0.2	1.1 ±0.3	4.6 ±0.2
PIV	0.3 ± 0.1	1.0 ± 0.1	0.5 ± 0.2	1.9 ± 0.3	1.9 ± 1.2	1.7 ± 0.9	4.7 ± 2.0	0.7 ± 0.9	6.6 ± 2.2	0.4 ± 0.9
PIV-adapt	0.3 ± 0.1	1.0 ± 0.1	0.3 ± 0.2	2.0 ± 0.2	0.3 ± 0.1	3.0 ± 0.3	0.3 ± 0.2	3.9 ± 0.4	0.3 ± 0.1	4.9 ± 0.4

consistency of these measurements.

Regarding PIV, like in the case of the centered images, results for the velocity estimation were consistent with the actual true value up to 2 rad/s, along with EPE errors $\leq$ 0.5 px and MAD values $\leq$ 0.2 px. However, from 3 rad/s, PIV estimates deteriorated with EPE errors over 6.6 px and a systematic underestimation of the velocity that worsens with speed, which confirms the limitations of this algorithm for angular velocities higher than 2 rad/s at a frame rate of 52Hz. As long as PIV-adapt is concerned, results stay close to those obtained on the centered images, both in terms of EPE and estimated velocity. The median error over all speeds remained constant at 0.3 px with MAD values between 0.1 and 0.2 pixels, while the error on the estimated velocity remained lower than 0.1 rad/s. Finally, it is interesting to note that in the case of non-centered images, PIV-adapt produced even better results compared to those obtained with FlowNetSD-TL, revealing the need for adding non-centered scenario cases during data augmentation to make FlowNetSD-TL more robust to such situations.

5.7 Discussion

In this chapter, we introduced an evaluation framework for quantitative comparison of different deep learning architectures to quantify motion from ultrasound images. To the best of our knowledge, this was the first time a study evaluated on ultrasound images: *i)* how different CNN architectures compare in terms of motion accuracy; *ii)* what is the impact of transfer learning when specializing networks trained on generic video sequences to ultrasound images; *iii)* how networks trained on simulated data perform on real ultrasound images (*i.e.* generalization ability to real data); *iv)* how a DL-based tracking solution compares with standard state-of-the-art tracking methods tuned for processing ultrasound images.

In terms of comparing different network architectures on our ultrasound database, it appeared that the FlowNetC network performed poorly both before and after transfer learning on the simulated images. This could be explained by the inability of the correlation layer to cope with speckle decorrelation induced by the rotations in our images. When comparing networks with similar architectures (FlowNetS and

FlowNetSD), we observed different performances in terms of accuracy. Interestingly, the main difference between the two networks mainly lies in the size of the convolution kernels. This suggests that this parameter impacts a lot the network accuracy and that it needs to be tuned carefully with respect to the acquisition system. In particular, if we assume a speckle size around 2 to 4 times the wavelength of the system, this amounts to 1.2 to 2.2 mm which corresponds to 3 to 5 pixels in our experiments. Regarding the best performing method (FlowNetSD), the size of the underlying convolution kernels was equal to 3 pixels, leading to a receptive field for the first two convolution layers of 3 and 5 pixels, which corresponds to the speckle size involved in the images. This tends to suggest that the extraction of features at the scale of the speckle size represents a good choice for the first layers of a CNN for capturing motion between ultrasound images. Finally, adding convolutional layers to reach the full image resolution at the output of the network did not improve the accuracy on our simulated dataset.

We compared FlowNetSD with the non-deep learning state-of-the-art PIV method, which showed excellent performance in a recent challenge [26]. At a sampling frequency of 52 Hz, PIV obtained accurate results for angular velocities ≤ 2 rad/s and failed to estimate velocities for values higher than 3 rad/s. To cope with large displacements, this method needs data at higher sampling frequencies, up to 312 Hz at 5 rad/s. In these conditions, PIV-adapt produced lower EPE errors and MAD values than FlowNetSD, especially for in-vitro images. Despite a motion estimate performed at 52 Hz, FlowNetSD yielded results with EPE errors ≤ 1 px for the full range of the tested velocities and with an accuracy on angular velocity estimates of less than 0.2 rad/s for simulated data and 0.4 rad/s for in-vitro data (with a better estimate of the angular velocity at 5 rad/s compared to PIV-adapt). Achieving a similar tracking accuracy over a large range of displacement amplitudes is remarkable, as standard registration algorithms' performance usually tends to deteriorate with larger motion.

One limitation of this study was the focus on a single motion pattern (rotations). Such focus allowed us to complete a study on a fully controlled motion, known to be one of the most important sources of speckle decorrelation. Results obtained from this pilot study revealed that deep learning solutions can be robust and accurate for the estimation of displacement fields in ultrasound, despite high speckle decorrelation. It is therefore important that future works extend this study to more complex and realistic motion patterns. As out-of-plane motion is known to occur for many imaging scenarios and to bring an additional source of speckle decorrelation, it should also be addressed in future work. This would allow to address this challenge during the training phase, and to evaluate the impact of that specific motion artifact on the tracking accuracy. In particular, 3D numerical simulations could introduce such decorrelation during the synthetic generation of 2D images, in the prospect of incorporating physical distortions in the data augmentation strategy. This would certainly be beneficial to improving the robustness of the tracking algorithm with respect to ultrasound-specific motion artifacts. Finally, the geometry involved in the simulated training dataset (*i.e.* spinning disk centered on the image) was the same as the one in the in-vitro dataset, leading to ideal conditions that could improve the performance of the evaluated network. This aspect has been evaluated in Section 5.6.2.2. From this experiment, it can be seen that velocity estimation remained accurate but the EPE increased with the angular speed, revealing the importance of designing simulations with as much variability as in real cases.

5.8 Conclusions

In this chapter, we benchmarked different CNN architectures on simulated and in-vitro data for tracking rotations between 1 and 5 rad/s from pairs of ultrasound images. Different networks, all derived from the FlowNet2 architecture, were compared with the PIV method, a state-of-the-art block matching algorithm tailored to ultrasound. Our quantitative evaluation on both simulated and in-vitro images revealed that the FlowNetSD network, after adapting its weights to ultrasound using transfer learning on simulated data, produced accurate motion estimation on in-vitro data for the full range of angular velocities and at a single frame rate of 52 Hz. Interestingly, FlowNetSD obtained angular velocity estimates comparable with the PIV-adapt method when the latter required adapting the acquisition frequency up to 312 Hz. This pilot study therefore reveals that deep learning solutions represent a potentially powerful alternative to standard tracking algorithms that can prove to be both robust and accurate for retrieving displacement fields from ultrasound images, including for large displacements and rotations, despite speckle decorrelation. For open science purposes, the databases used during these experiments were made available for download [2]. Based on the findings of this study, we investigated in the next chapter the ability of DL methods to outperform state-of-the-art methods for myocardial deformation estimation in echocardiography.

[2] `http://humanheart-project.creatis.insa-lyon.fr/revolus.html`

Chapter 6

Application of CNNs to the estimation of myocardial motion in 2D ultrasound

Contents

This chapter is an extension of our article "Motion estimation by deep learning in 2D echocardiography: synthetic dataset and validation" published in IEEE Transactions on Medical Imaging [96].

6.1 Introduction

As we previously saw in Chapter 2, ultrasound imaging is a widely used imaging modality in cardiology because it is inexpensive, fast and non-invasive. Echocardiography enables the extraction of clinical indices relevant to study the cardiac function and anatomy such as volumes and myocardial deformation [3]. Deformation indices are usually estimated by conventional motion estimation techniques which suffer from difficulties inherent to ultrasound images, such as artifacts (shadow, reverberation), lack of information or speckle decorrelation. The latter corresponds to the fact that the speckle pattern which is tracked from B-mode sequences can change over time. This phenomenon depends on the type of movement of the tissues (rotation being one of the worst) and is all the more true as the movements are fast, as demonstrated in the previous chapter. This results in a lack of accuracy and reproducibility in current embedded solutions. Therefore, improvements in motion estimation are crucial in echocardiography to obtain reproducible indices.

One index that has attracted considerable attention is the global longitudinal strain. As described in Section 2.3.4, GLS is defined as the percentage of myocardial longitudinal shortening between the end-diastolic and end-systolic instants [17]. It is a global value that proved to be robust enough to be part of the recommendations during clinical exams [4]. GLS is computed from B-mode images acquired in any standard apical view and by tracking a myocardial contour using the conventional block-matching [26] or optical flow techniques [22]. Tissue Doppler techniques can also be used to estimate GLS without the use of speckle tracking.

As detailed in Chapter 3, deep learning approaches have recently outperformed standard tracking methods on natural images. In particular, the benchmark on the Sintel dataset [1] shows that the top-ranked algorithms are all based on DL approaches and that the first non-DL method (FlowFields [97]) is currently ranked above 100. We thus hypothesized that DL can significantly improve tracking accuracy and robustness over traditional methods in echocardiography. Instead of relying only on the intensity or the phase information in the image to evaluate the motion, DL networks can learn to estimate complex tissue motion with the associated speckle decorrelation. Moreover, the addition of typical ultrasound artifacts during training should provide greater robustness of motion estimation and better adaptation to ultrasound images.

At the time this study began, most DL methods in ultrasound were applied to elastography [53], [56], among which some are based on extensions of key architectures such as PWC-Net [54], [55]. Some studies have also been conducted to estimate myocardial motion in echocardiographic imaging. In [59], an unsupervised approach based on the U-Net architecture was used to estimate the canine myocardial motion from a short-axis view. Evaluated on the same data, another network with an architecture derived from FlowNet-C was developed and trained in a semi-supervised way [98]. Short axis view essentially provided information on radial and circumferential strain. To track longitudinal motion, a pipeline was implemented to automate the GLS computation

[1] http://sintel.is.tue.mpg.de/results

using view classification, segmentation, motion estimation and Kalman filters on apical four chambers views [50]. The motion estimation part was based on FlowNet2 with the original network weights learned from natural synthetic images. Recently, another pipeline with a modified version of PWC-Net named EchoPWC-Net was introduced [51]. To adapt this network to ultrasound images, the authors removed the feature maps warping, propagated the first feature maps and added finer resolutions to the loss. This network was trained on a realistic simulated ultrasound dataset [82] in a supervised way and evaluated on the same *in-silico* dataset and on 30 *in-vivo* patients. Despite all the architectural modifications, the clinical measurements obtained on the real data were only slightly better than those obtained by a state-of-the-art method. Based on these results, the authors highlighted the importance of simulated data and pointed out the lack in quantity and diversity of training data currently available.

In this context, we decided to make contributions regarding the PWC-Net architecture, synthetic training data for capturing motion in ultrasound, a thorough investigation of different temporal strategies for improving results, and the first study on the generalization of this type of network in echocardiography:

- To overcome the problem of limited synthetic data in number and diversity, we created a new pipeline to generate large-scale synthetic ultrasound sequences with a wide range of cardiac deformations. Two types of synthetic data were thus generated, with and without reverberation artifacts.
- In contrast to [51], we showed that the PWC-Net architecture has the potential to produce relevant results on ultrasound images thanks to an adapted transfer learning procedure. This allows a better generalization of the network and a significant improvement of the results on clinical data.
- We further improve the performance of this network on ultrasound data by modifying its architecture to enhance its multi-scale analysis capability.
- We performed a thorough study of several temporal strategies that can be used to improve results during both the training and inference phases.
- We conducted the first study on the generalization of deep learning algorithms for motion estimation in echocardiography using a multi-center, multi-vendor and multi-disease dataset of real patients.

6.2 Methods

6.2.1 Synthetic dataset for relevant transfer learning

Two recent studies have shown that supervised DL techniques can learn from synthetic ultrasound sequences to improve motion estimation on *in-vitro* [88] and *in-vivo* data [51]. In this context, the realism of synthetic image sequences is key for improving the performance of DL models. In both studies, a physical simulator was used to generate synthetic data and special care was taken to define a realistic medium from acoustic scatterers. Besides the realism of the ultrasound image, the motion must also be realistic. In [78], [82], the motion field was generated through a bio-mechanical personalized simulation. The personalization operation remains tedious, and currently

limits the deployment of such scheme to small dataset (*i.e.* number of patients lower than 10 with the same kind of heart motion) with synthetic myocardial deformations that remain low as compared to reported normality ranges (*e.g.* simulated peak systolic longitudinal strain lower than 10% instead of 20% in real cases). In this study, we proposed a dedicated simulation strategy to tackle this issue, and augment the database with diverse ranges on motions, cardiac geometries and image quality.

6.2.1.1 Overall strategy

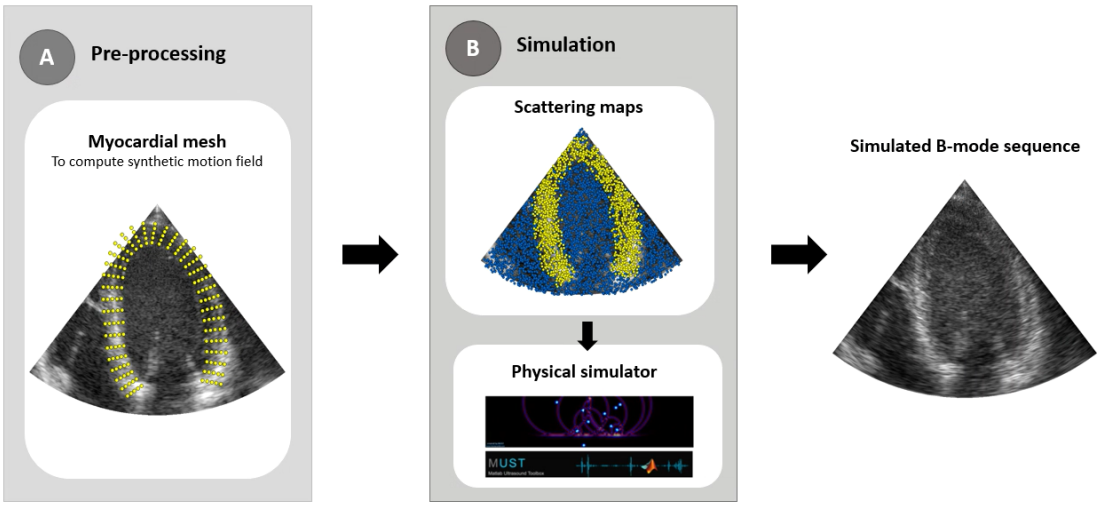


Figure 6.1 – Proposed pipeline for the simulation of B-mode sequences. (A) A clinical recording works as a template for speckle texture, anatomy definition, and myocardial motion estimation; (B) An ultrasound simulation environment merging information from the template image sequence and the myocardial meshes accounts for the image formation process. In the simulated sequence, the myocardial motion is fully controlled by the sequence of meshes while the visual appearance is very similar to the one of a real acquisition.

Our overall strategy builds upon the same core concepts as in the previous state-of-the-art papers [78], [82]. A schematic figure showing the workflow of the simulated pipeline is given in Figure 6.1. Clinical apical four-chamber B-mode recordings (called as template in the sequel) were used to simulate sequences with realistic tissue texture. For each frame of the template sequence, a scatter map was computed and fed to a physical simulator to produce the corresponding synthetic B-mode image. The scatterer maps were composed of two types of elements: the background and the myocardial scatterers. The full scatterers were distributed within the sector of the first frame according to a uniform random distribution. A density of 10 per square wavelength was chosen to ensure realistic speckle statistics. To avoid flickering effects, the background scatterers were kept motionless. To mimic the local echogenicity of the recorded model, the local intensities I_m of the actual B-mode images were used to calculate the reflection coefficients RC_m of the scatterers, *i.e.* $RC_m = (I_m/255)^{(1/\gamma)} \cdot \mathcal{N}(0, 1)$, where $\mathcal{N}(\cdot)$ is the normal distribution, and γ is a constant for gamma compression. The myocardial scatterers were selected on the first simulated frame using manual annotations. The positions of these scatterers were then computed for each B-mode frame of the simulated sequence using the strategy described at the end of this sec-

tion. The reflection coefficients of these scatterers were kept constant to maintain the speckle texture throughout the cardiac cycle. The final scatterers were obtained by combining the background and myocardial scatterers using the same scheme as in [82]. This strategy allows a smooth transition at the myocardial borders. Finally, a homemade open-source software called SIMUS from the MUST Matlab ultrasound toolbox[2] [99] was used to generate the synthetic ultrasound data. Each B-mode frame was generated by transmitting 128 focused beams, regardless of the acquired sector width (ranging from 60 to 90 degrees). In addition, the focal point was automatically chosen for each patient to be equal to half of the total acquired depth (ranging from 11 to 20 cm). The synthetic signals generated by SIMUS were demodulated to obtain IQ signals. The IQ signals were beamformed using a delay-and-sum technique to obtain B-mode images [100].

6.2.1.2 Template image sequences

The template cine loops used in our simulation pipeline come from the CAMUS open access dataset which consists of exams from 500 patients acquired in clinical routine from the University Hospital of St-Etienne (France) and using a GE system [101]. This dataset was built without any specific image quality or patient selection criteria to match the heterogeneity of texture, shape and cardiac motions seen in clinical routine. We selected a subset of 100 apical four-chamber sequences, where the ultrasound machine settings were adjusted to scan the myocardium (Figure 6.2). The same probe settings used to acquire the CAMUS dataset were simulated: a 2.5 MHz 64-elements cardiac phased array.

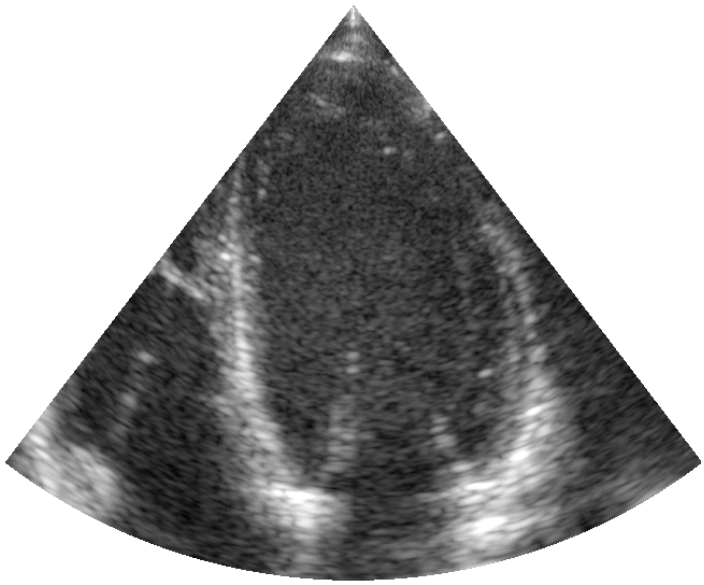


Figure 6.2 – Example of an apical four chamber echocardiography from the CAMUS database.

[2] `www.biomecardio.com/MUST`

6.2.1.3 Synthetic myocardial motion field

Endocardial and epicardial borders were delineated manually on the template sequences to obtain myocardial ROIs over the entire cardiac cycle. Time-varying surface meshes were generated for each of these ROIs following the resampling scheme given in Figure 6.3. Specifically, the base of the left ventricle was defined by the segment linking the two extreme endocardial points. The apex was defined as the furthest point from the base in the epicardial contour. 36 points were then evenly distributed over the epicardial contour: 18 on the septum, and 18 on the lateral wall. Intramyocardial perpendicular segments were then drawn from these epicardial points to join the epicardial and endocardial contours. Each intramyocardial segment contained 5 evenly distributed points. This resampling scheme meshed the myocardium with 180 points (36 longitudinal $\times$ 5 radial) and 280 triangle cells.

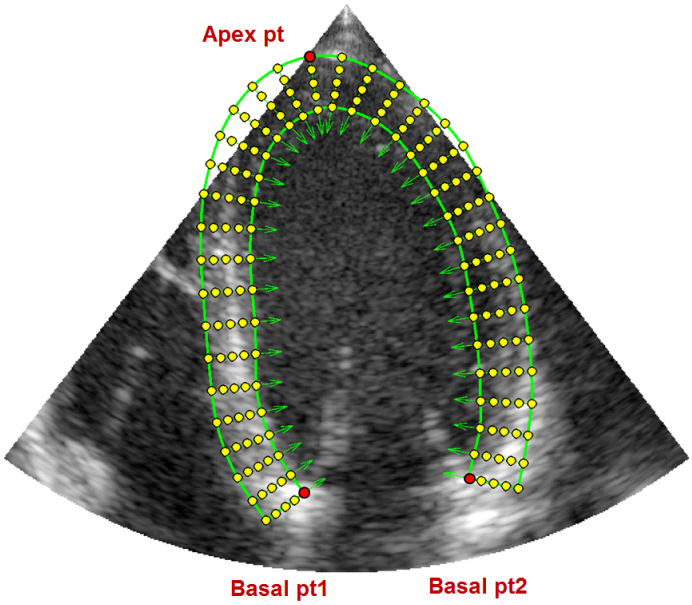


Figure 6.3 – Illustration of the resampling scheme used to generate a myocardial mesh (yellow nodes) from the corresponding segmentation mask (green lines).

For each simulation, a set of points was randomly distributed over the myocardial mesh at end-diastole. Each of these points was then propagated over the full sequence by interpolating the displacements of the corresponding cell. This simple procedure allowed us to compute the temporal trajectory of any point belonging to the myocardium. The resulting synthetic myocardial motion field does not correspond to the actual motion field, which is not the purpose here. The interest of this procedure is to efficiently generate a wide variety of cardiac motions/deformations that are realistic enough to serve as a relevant data augmentation for DL methods.

6.2.1.4 Reverberation artifacts

We incorporated reverberation artifacts into our synthetic dataset to challenge the network during training. Specifically, we placed scatterers near the mid-anterolateral wall with high reflection coefficients relative to their neighbors. The position and amplitude of these scatterers remained constant throughout the cardiac cycle. This simple strategy leads to stationary saturated areas in the simulated B-mode images

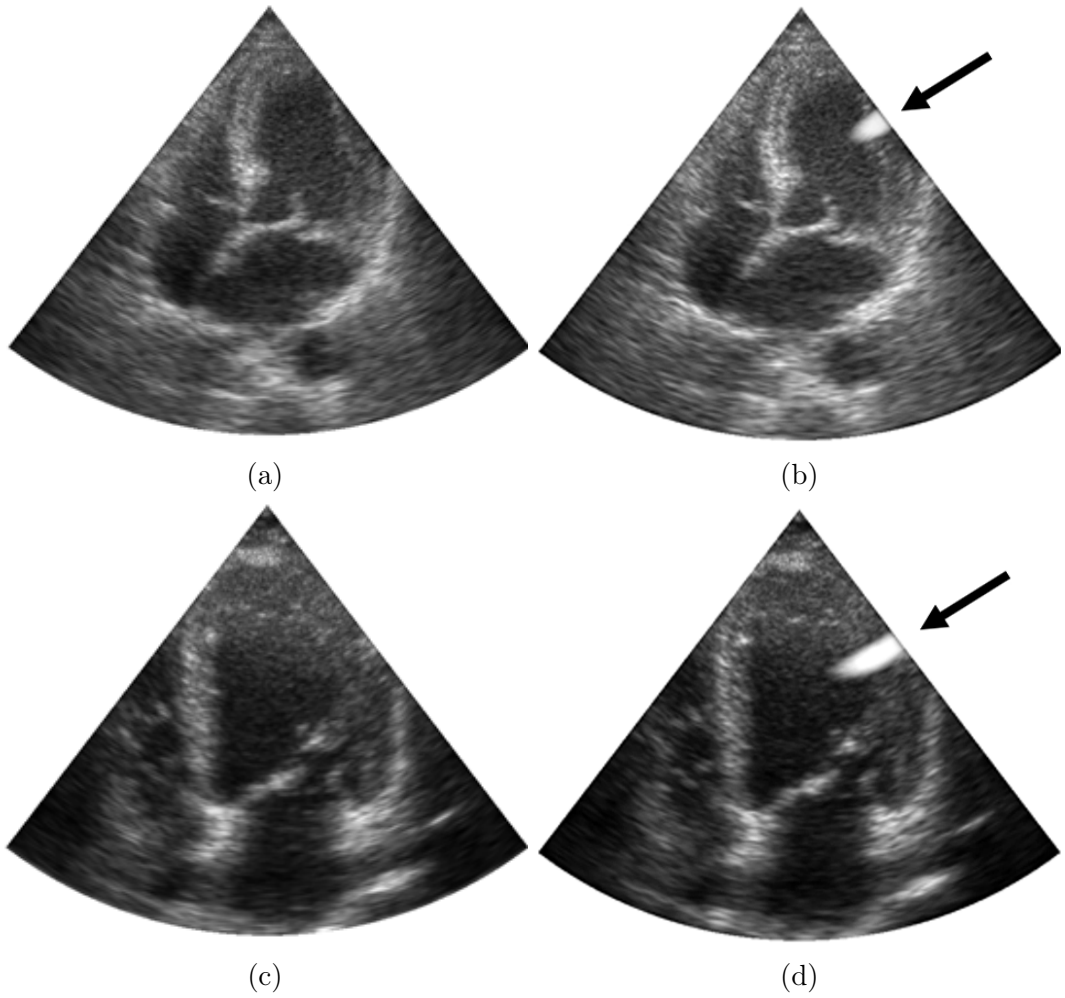


Figure 6.4 – Synthetic ultrasound images simulated from the proposed pipeline with (b, d) and without (a, c) reverberation artifacts for two different patients. The reverberation artifacts are identified by arrows.

to emulate reverberation artifacts that may come from the ribs, as shown in Figure 6.4. Real reverberation artifacts may have other characteristics such as multiple reverberation structures and clutter noise, but these are not taken into account in this simulation.

6.2.2 Optimization of PWC-Net for echocardiography

6.2.2.1 Overall architecture

PWC-Net is one of the most efficient DL networks for dense motion estimation between two frames [36]. This network borrows the concept of a multi-resolution pyramidal structure to standard image tracking algorithms. Motion is estimated from the coarsest to the most detailed spatial resolution. A pyramid of seven levels with shared weights downsamples successively the features maps by half. Input images are processed separately. A normalized cross-correlation between the feature map of the first image and the second image warped by the previous estimated flow is then computed. This operation named *cost volume* performs patch comparisons between two feature maps for a range of displacements. The cost volume, the feature map from the first

image, the upsampled estimated flow obtained at the previous level and the upsampled feature map are used as input in a Convolutional Neural Sub-Network (CNSN), referred to as *estimator*. This CNSN is in charge of predicting a dense displacement map. The steps previously described are iterated until obtaining a displacement field with a quarter of the size of the initial input image. This information is then provided as input to another CNSN, referred to as *context*, to improve the accuracy of the estimated flow. This is done by adding the previously estimated flow with the output of a branch involving dilated convolutions to reinforce the receptive field. Finally, a bilinear interpolation upsamples the final flow to output a displacement map of the same size as the input image. The parameters of this network are optimized through a multi-scale loss function. This function computes the distance between the intermediate estimated flows and the corresponding scaled ground truths. The detail of the full network configuration is described in the Section 3.3.3.2.

6.2.2.2 Proposed architecture

The overall architecture of our customized PWC-Net is given in Figures 6.5 and 6.6. The modifications we made from the original architecture are all displayed in green. Based on the observation that multi-scale analysis has proven to be efficient for motion estimation in ultrasound [24], we first added a contextual sub-network at each resolution level of the network (context blocks in Figure 6.5). In addition, the tracking of speckle patterns whose shapes can evolve between two consecutive frames make the motion estimation task particularly difficult in ultrasound. For this reason, we decided to reinforce the capacity of the network to extract relevant information by modifying each estimator sub-network as illustrated in Figure 6.6. These modifications correspond to skip connections concatenated to the output of each convolutional layer. The interest of these connections is twofold: *i)* since the PWC-Net architecture is deep, they limit the phenomenon of vanishing gradient; *ii)* the inputs of each convolutional layer are composed by the concatenation of the input and the outputs of the previous layer, leading to richer information sources. Similar to our intuition, Densenet connections were evaluated in [36], which improved the results by 5% but also increased the execution time up to 40%. Therefore, the authors leave the choice of using these connections according to the targeted objectives. The unlabeled blocks in Figure 6.5 represent the pyramidal feature extractors described in Section 6.2.2.1 and whose implementation details are described in Section 6.3.3. These feature extractors correspond to the ones proposed in the original PWC-Net implementation and do not involve any skip or residual connections.

6.2.2.3 Transfer learning strategy

In contrast to [51], we propose to keep the transfer learning strategy in order to strengthen the generalization capacity of the derived method, which should improve the results on clinical data. The specialization of the network from natural images to ultrasound was performed through different key steps. For adapting the proposed customized PWC-Net to gray level images, we first trained our network on a set of natural image pairs taken from the synthetic FlyingChairs2D and FlyingThings3D datasets [36]. Details on these datasets are given in Section 6.2.1. Once the network has been learned on this first dataset, two different transfer learning procedures were performed on simulated ultrasound images with the same weight regularization as the

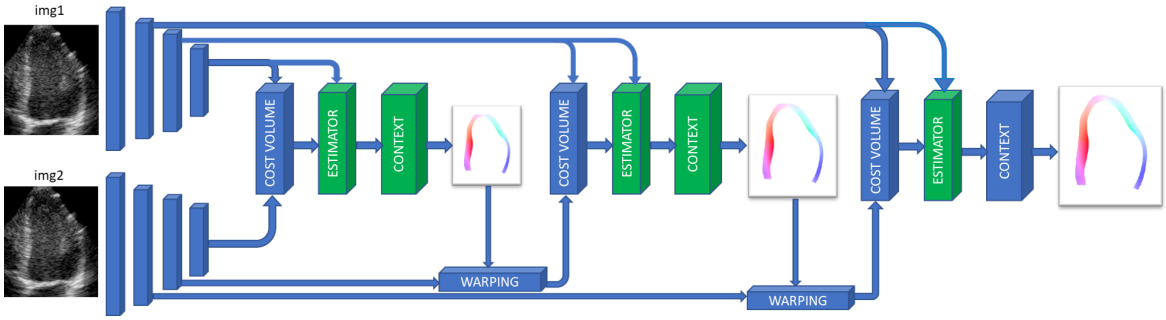


Figure 6.5 – Schematic view of our customized PWC-Net illustrated with a 4-level pyramid. The two input images are initially processed separately to extract the features, then the displacement fields are estimated in a coarse-to-fine manner (see Sections 3.3.3.2 and 6.2.2.1 for more details). The sub-networks modified as described in Section 6.2.2.2 are displayed in green.

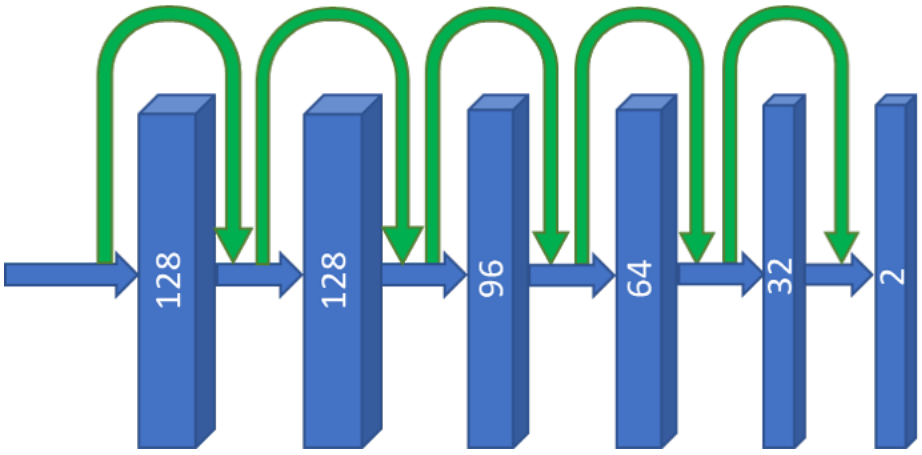


Figure 6.6 – Illustration of the estimator sub-network used in our customized PWC-Net with the added skip connections in green.

initial training and without freezing any layer. The first transfer used an open access dataset [82] for the purpose of comparison with [51]. A second transfer was made using the same open access dataset extended with a new simulated ultrasound dataset based on the pipeline described in Section 6.2.1. Both transfers used the same learning rate value ($\lambda = 1e^{-4}$) and the same progressive decay strategy as in the training on natural images, to ensure an efficient transfer to images of a different nature.

6.2.2.4 Temporal augmentation strategy

Different motion amplitudes between simulated and real data can worsen the performance of DL networks during inference. In addition, simulated data may be biased by certain types of motions and may not represent the variety of real movements, whether healthy or pathological. For addressing these issues, two temporal data augmentation strategies were combined during the training phase. First, to double the dataset with realistic movements of the ultrasound speckle, the pairs of forward frames with reference field ($t \rightarrow t+1$) were also presented to the network in the backward direction ($t+1 \rightarrow t$). In addition, rather than using only consecutive frames, we also provided image pairs separated by several frames to increase the amplitude of motion and the

levels of speckle decorrelation seen by the network during training.

6.2.2.5 Composition inference strategy

The speckle motion pattern was assumed to be consistent for an image pair $I_1 I_2$ in the forward ($I_1 \rightarrow I_2$) and backward ($I_2 \rightarrow I_1$) directions. This forward-backward composition consistency was exploited during inference to still improve the performance of the network. In particular, each motion estimation between two consecutive B-mode frames was performed as follows. The forward motion between I_1 and I_2 (F_f) was first computed and used to propagate the myocardial points. The backward motion field (F_b) was then computed at these coordinates. In an ideal case, the composition of these two displacement fields should return the identity transformation. To respect that constraint, we averaged the forward F_f and backward $-F_b$ displacement fields to compute the final motion estimation.

6.3 Experiments

6.3.1 Datasets

6.3.1.1 Synthetic natural datasets used for training

As described in Section 6.2.2.3, our network was first trained to model motion estimation from natural synthetic images. To this aim, we used two public datasets consisting of image pairs with the corresponding dense displacement field on the entire image and described in Section 4.2. The FlyingChairs dataset is composed of scenery images over which chairs with random orientations are over-imposed [29]. Random affine transformations were applied to the background and the chairs in the foreground. The FlyingThings3D dataset consists of images created from a mix of randomly textured 3D flying objects on a textured background [67]. The objects were randomly positioned in the image and then modified by geometrical transformations. Their motions correspond to randomized displacements along 3D linear trajectories. More details on these datasets are given in Chapter 4. We removed image pairs with displacement amplitudes that were too large for what can be expected in echocardiography. Each image was expressed in grayscale format. This resulted to a synthetic dataset composed of 42512 image pairs with the corresponding displacement fields.

6.3.1.2 Synthetic ultrasound datasets used for training and testing

We used the open access dataset of realistic 2D ultrasound sequences described in Section 4.3.2 during the different transfer learning procedures [82]. This dataset is based on an electromechanical model of the heart which was combined with template cine loop recordings to simulate realistic ultrasound sequences. This approach relies on a personalization procedure which currently limits the heterogeneity of the dataset. This open access dataset is composed of 2D apical two-three-four chamber view sequences for seven vendors and five different motion patterns, including one healthy and four pathologies. This resulted in a dataset composed of 6060 pairs of synthetic ultrasound images with the corresponding myocardial displacement fields.

To enrich this dataset, we developed a new simulation pipeline as described in Section 6.2.1.1. Based on this new approach, we simulated 2D apical four chamber

view sequences for 100 virtual patients from the CAMUS dataset. From Figure 6.7, one can appreciate the rich variety and the realistic nature of the cardiac deformations present in the simulated dataset. The same template cine loops were used to simulate new sequences with reverberation artifacts. This increased the diversity of speckle patterns and can make the network less sensitive to this type of artifact. The resulting synthetic dataset is composed of 8866 image pairs with the corresponding myocardial displacement fields. The full dataset is made available to the research community at `http://humanheart-project.creatis.insa-lyon.fr/medicaid.html`.

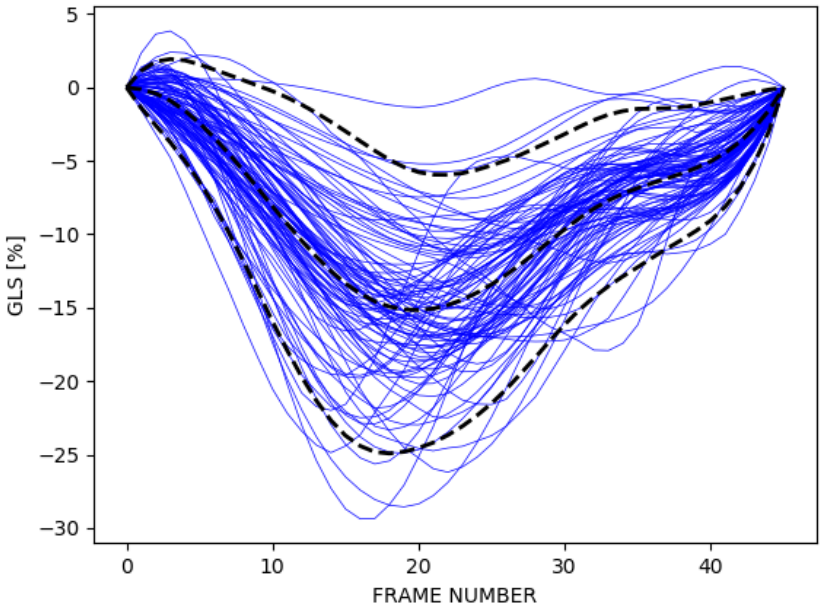


Figure 6.7 – Evolution of GLS as a function of time for all simulations. The mean and the limits of agreement are represented in black.

6.3.1.3 Clinical datasets used for testing

The open access CAMUS dataset described in Section 6.2.1.2 was first used to create a 30-patient clinical dataset consisting of apical four chamber view sequences acquired from a GE system. These sequences were selected according to their quality (only good and medium image quality were included), also ensuring the whole left ventricular myocardium was included in the field of view. The original CAMUS dataset is provided with annotations only at end-diastole and end-systole. We therefore asked an expert cardiologist to extend these annotations to the entire cardiac cycle. This work was done through an in-house web annotation platform based on the Desk library[3][102]. This resulted to clinical dataset composed of 1443 image pairs with reference contours for both endocardial and epicardial borders.

In order to assess the generalization capacity of our approach, an auxiliary dataset composed of 2D apical four chamber view sequences from 30 new patients was collected at the University Hospital of Caen (France) within the regulation set by the local ethical committee. The images were acquired using Philips scanners. The same

[3] `www.creatis.insa-lyon.fr/~valette/public/project/desk/`

protocol as the one used for the CAMUS dataset was used to manually annotate the entire cardiac cycle. To ensure a representative range of left ventricle pathologies, five different groups equally distributed were selected, resulting in six patients from each group. The groups were defined by a diagnosis of aortic stenosis (AS), hypertrophic cardiomyopathy (HCM), ischemic heart failure (HF), non-ischemic HF and no disease. This resulted to an auxiliary clinical dataset composed of 1536 image pairs with reference contours for both endocardial and epicardial borders. An image of this database is provided as an example in Figure 6.8.

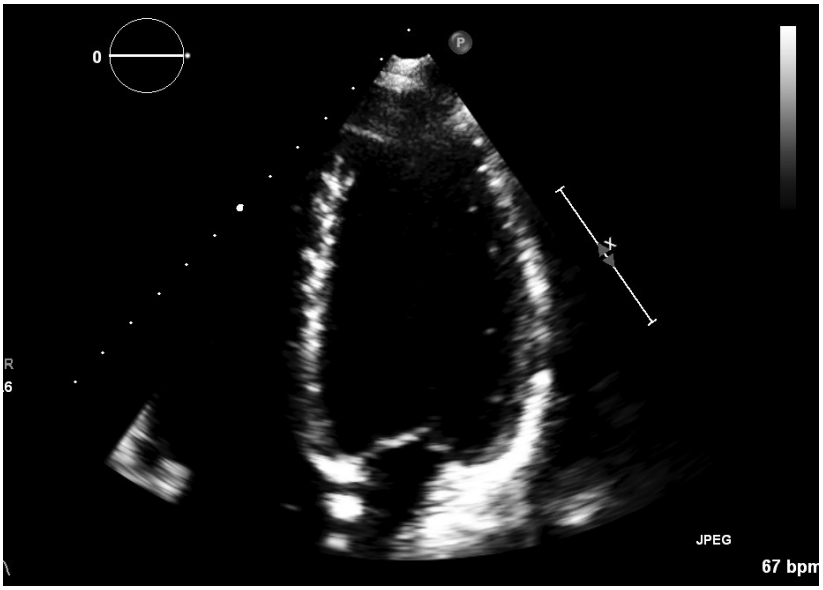


Figure 6.8 – Example of an apical four chamber echocardiography of a patient in the disease-free group from the Philips auxiliary database

6.3.2 Evaluated methods

As commercial applications like Tomtec and Echopac cannot be easily configured to modify the input contour to be tracked, we decided to assess the performance of the proposed network with the PIV block matching method [26], the Farnebäck optical flow method [22] and two DL methods, namely the PWC-Net [36] and the EchoPWC-Net [51]. Results from EchoPWC-Net and Farnebäck methods were taken directly from [51] due to difficulties in reproducing them.

6.3.2.1 PIV

Blockwise speckle tracking was implemented using standard FFT-based cross-correlations (PIV) on the B-mode images, further detailed in Section 3.2.2. The images were divided into subwindows with a 75% overlap. We detected small and large displacements with a multi-scale approach: displacement estimates were iteratively refined by decreasing the size of the subwindows (32×32, 16×16, and then 8×8). In contrast to [26], we computed the normalized cross-correlations from only two successive images (*i.e.* , no ensemble correlation). To obtain sub-pixel displacements, we used a parabolic fitting around the correlation peak. The estimated displacements were smoothed with an unsupervised robust spline smoother between two consecutive scales [103].

6.3.2.2 Farnebäck

This method is a traditional dense optical flow algorithm which is based on a pyramid of images at different resolution levels to track image points (see Section 3.2.1). A detailed description of the method and its parameters are given in [51].

6.3.2.3 EchoPWC-Net

This network is based on the PWC-Net architecture with several modifications to improve results on ultrasound data. The intuition behind these changes was to preserve information brought by speckle patterns by optimizing local variations. The authors also proposed to reinforce the resemblance of their solution with traditional speckle tracking approaches by integrating block matching aspects during the computation of the cost volume. The reader is referred to [51] for more details on this method.

6.3.3 Implementation details

Our experiments were realized using Keras 2.3.1 and Tensorflow 1.15 libraries. The modelling and experiments were conducted on a workstation with Ubuntu 20.10 operating system. The hardware consisted of an AMD Ryzen 9 3900X processor and an NVIDIA GeForce RTX 2080Ti GPU with 12 GB of memory. The runtime achieved in inference was 20 ms to estimate the myocardial motion between two consecutive frames.

6.3.3.1 Architecture parameters

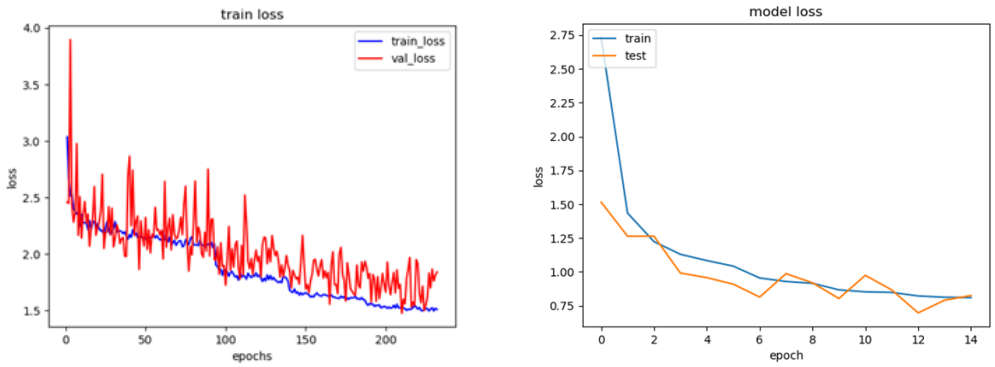
For the feature extractor of each level of the pyramid, we used three convolutional layers, including a convolution with a stride of two to downsample the final feature maps. The number of filters per level is constant and was set to 16, 32, 64, 96, 128 and 196 from top to bottom. The search range of the cost volume was fixed to 4 pixels. The same context network as the one proposed in [36] was used at each resolution level, composed by a succession of dilated convolutions with factors of 1, 2, 4, 8, 16, 1 and 1 respectively. The total number of parameters of our network was 14M.

6.3.3.2 Training procedures

Our network was trained with the Adam optimizer and a batch size of 4 for all the experiments. The initial learning rate was set to $1e^{-4}$, with a halving schedule every 20% of the training after 40% of the total number of iterations. 200 epochs (150 hours) and 15 epochs (15 hours) were used during the training on the natural and ultrasound datasets, respectively. An example of the evolution of the training and validation losses computed during two different learning sessions is provided in Figure 6.9. The weights were initialized following the He initialization scheme [104].

6.3.3.3 Data augmentation

Different data augmentation strategies were applied depending on the learning phase. During training on the synthetic dataset with natural images, the same data augmentation scheme as the one performed in [29] was carried out. A random crop of



(a) (b)

Figure 6.9 – Training and validation curves for two different training sessions. The curve on the left figure (a) corresponds to the training of c-PWC-Net on the natural gray images dataset described in Section 6.3.1.1, while the curve on the right figure (b) corresponds to the transfer learning of c-PWC-Net on our proposed synthetic dataset described in Section 6.3.1.2.

size 448×384 was used to select the images given as input to the network. In addition, geometric transformations (translations, scaling, flipping) and image alterations (brightness, Gaussian noise, contrast) were used. During the transfer learning phase, only image alterations (brightness and contrast) were used to respect the shapes and geometrical properties of the ultrasound images. Regarding rotation, we observed that its use during the augmentation procedure does not improve the results. This can be explained by the sectorial shapes present in these images. Indeed, all the sequences present in the *in-vivo* datasets do not have any tilde angle, which yields the same orientation of the sectorial shape. Using rotation as data augmentation thus leads to inappropriate rotated sectorial shapes.

6.3.3.4 Loss

We used a multi-scale loss function defined as

$$L(\Theta) = \sum_{l=l_0}^{L} \alpha_l \sum_{x} |w_\theta^l(x) - w_{GT}^l(x)|_2 + \gamma|\Theta|_2, \tag{6.1}$$

with Θ being the parameters, x the inputs, $|.|_2$ the $L2$ norm and w^l the flow field at the lth pyramid level. The values of the α_l parameters correspond to the one proposed in [29]. The regularization factor γ was set to $4e^{-4}$. During transfer learning, the optimization of this loss was restricted to the region where the reference dense displacement field was known, *i.e.* in the myocardial region.

6.3.4 Evaluation Metrics

6.3.4.1 Geometric Metrics

To measure the accuracy of the estimated motion and the tracked contours of a given method, the endpoint error (EPE) (Equation 3.5) computed only in the myocardial region, the mean absolute distance (d_m) and the Hausdorff distance (d_H) were used.

d_m corresponds to the average distance between two contours while d_H measures the local maximum distance between the two contours. In our experiments, d_m and d_H were used to assess tracking quality by comparing the reference endocardial contours with the contours obtained by propagating the reference contour at the first frame with the successive motion estimations obtained with the different techniques. The reported d_m and d_H scores correspond to the average values computed from all the contours obtained on the same sequence except the one at the first frame.

6.3.4.2 Clinical Metrics

We gauged the methods' performance with the Global Longitudinal Strain (GLS). There is currently no consensus on the way to compute it. Strain estimation from the epicardial border is generally avoided because of the proximity of the pericardium, which makes several areas difficult to segment. Because of a varying quality of the speckle pattern along the heart muscle, strain estimation from the myocardium is usually associated with a regularization term that affects the quality of the measurements. Finally, the strain estimation from the endocardium can be seen as the more robust to achieve thanks to a sufficient contrast between the blood pool and the myocardium. However, it is known that the computation of the GLS from the endocardium takes into account a part of the radial deformation [19]. In this study, we decided to use the endocardial contour to compute the longitudinal ventricular length L for each timestep t to estimate the Lagrangian strain (Equation 2.4). Moreover, to address the weakness of using the endocardial contour to calculate GLS, we use a complementary clinical index that provides information exclusively on longitudinal deformation, named Mitral Annular Plane Systolic Excursion (MAPSE) [20]. This index is computed in the same way as the GLS, with the difference that L denotes the distance between the apex and the mid-basal point of the endocardial border. Both GLS and MAPSE were computed for each time of the sequence.

The peak-systolic strain, defined as the minimum between end-diastole and end-systole strain values, was finally calculated for both indices, for which we computed three metrics: the bias, the standard deviation σ and the mean absolute error (mae). GLS and MAPSE ground truths were derived from the expert manual annotation of the endocardial contour. For fair comparisons, the same endocardial reference contours at end-diastole were used as the initial contours from which the different tracking methods were applied. All reference contours were manually delineated prior to the application of any method, thus avoiding any bias issues. It is important to note that the quality of the results of our tracking depends on the initial contour that is provided as input by an operator. Particular attention was therefore paid for accurately segmenting the myocardium.

6.4 Results

6.4.1 Simulations

6.4.1.1 Ablation Studies

We first conducted an ablation study to evaluate the influence of different design choices during the training and inference phases. We used for this purpose the open-

access dataset of synthetic ultrasound sequences described in Section 6.3.1.2. This dataset was divided into folds by vendor, of which five were used for training, one for validation and one for testing. During this experience, we trained our customized version of PWC-Net, referred as c-PWC-Net in the sequel, using a composition of the following temporal data augmentation strategies: *i)* using data in the forward direction only; *ii)* using data in the forward/backward directions and *iii)* using image pairs separated by one, two or three frames. During inference, the composition consistency described in Section 6.2.2.5 was also investigated. The obtained results are reported in Table 6.1. From these results, we can first observe the slight improvement (0.01 mm reduction of the EPE) brought by the composition consistency procedure. Regarding the training phase, the forward/backward and the spaced image pairs strategies further improved the EPE. The best configuration was obtained when combining the composition consistency in inference with the forward/backward data augmentation strategy during training using image pairs separated by one and two frames, the separation with more images leading to an unchanged average EPE of 0.04 ± 0.03 mm. This configuration was therefore employed for all the following experiments.

The importance of the modifications we brought to the original PWC-Net architecture was investigated through two complementary ablation studies. The first experiment was realized on the same dataset as for the ablation study given in Table 6.1. During this experiment, we trained c-PWC-Net using the same transfer learning strategy but for different versions of its architecture: *i)* without skip connection and contextual sub-networks; *ii)* with skip connection only; *iii)* with contextual sub-networks only and *iv)* with both of them (which corresponds to the proposed full architecture). From the results given in Table 6.2, one can clearly see the importance of the skip connections and the contextual sub-networks, the use of the contextual sub-networks resulting in an improvement of 0.2 mm while its combination with skip connections allowed to divide by 2 the average EPE compared to the original architecture of PWC-Net (*i.e.* c-PWC-Net without skip connection and contextual sub-network). Contrary to the two previous ablation studies, we used the proposed synthetic dataset to study the influence of the contextual sub-networks. Indeed, this dataset presents more variety in terms of strain deformations and therefore seems to be better suited. During this experiment, we trained c-PWC-Net using the same transfer learning strategy and with the following conditions: *i)* with a contextual sub-network for each resolution, except the two lowest ones; *ii)* with a contextual sub-network for each resolution, except the lowest one; *iii)* with a contextual sub-network for all resolutions (*i.e.* 6 sub-networks in total). Results given in Table 6.3 clearly show the interest of adding a contextual sub-network at each resolution level, even for the lowest ones. The best network architecture, subsequently adopted, was therefore the combination of skip connections with contextual sub-networks for all resolutions.

6.4.1.2 Open Access Synthetic US dataset

Table 6.4 shows the motion estimation performance of c-PWC-Net from the open access synthetic US dataset [82]. We used the same nomenclature as the one introduced in [51], by adding "-gray" in case of training performed on grayscale natural images, "-us" in case of training on the open access dataset and "-ft" in case of fine-tuning. For instance, PWC-Net-gray-usft refers to PWC-Net first trained on natural grayscale images, then on the synthetic US dataset of [82] through transfer learning. The Farnebäck, PWC-Net-gray-usft and EchoPWC-Net-us results were taken

Table 6.1 – Ablation study on the open access dataset [82] for the temporal data augmentation and inference strategies proposed in Section 6.2.2. The proposed customized PWC-Net (c-PWC-Net) was trained using forward (+), forward-backward (+/-) and image pairs separated by n frames (0 meaning two consecutive images). In inference, results were computed with and without the composition consistency procedure given in Section 6.2.2.5.

c-PWC-Net		**EPE**$\pm\sigma$
Training	Inference	mm.
+, 0	No composition	0.09 ± 0.07
+, 0	Composition	0.08 ± 0.07
+/-, 0	Composition	0.05 ± 0.04
+, 0, 1	Composition	0.05 ± 0.05
+/-, 0, 1	Composition	0.04 ± 0.04
+, 0, 1, 2	Composition	0.05 ± 0.03
+/-, 0, 1, 2	Composition	**0.04 ± 0.03**
+, 0, 1, 2, 3	Composition	**0.04 ± 0.03**
+/-, 0, 1, 2, 3	Composition	**0.04 ± 0.03**

Table 6.2 – Ablation study performed on the open access dataset [82] for the architectural modifications given in Section 6.2.2.2. The different networks were trained in the same conditions using the forward-backward strategy with image pairs separated by 0, 1, and 2 frames.

c-PWC-Net architecture		**EPE**$\pm\sigma$
Skip connections	Contextual sub-net.	mm.
✘	✘	0.08 ± 0.06
✔	✘	0.07 ± 0.06
✘	✔	0.06 ± 0.06
✔	✔	**0.04 ± 0.03**

Table 6.3 – Ablation study performed on the proposed synthetic dataset for the influence of the contextual sub-networks presented in Section 6.2.2.2. The different networks were trained under the same conditions as for the other ablation studies. The column labeled Positions provides information about the presence of a contextual sub-network relative to the pyramid level (1 stands for the highest resolution, 6 to the lowest).

c-PWC-Net architecture		**EPE**$\pm\sigma$
Number of sub-net.	Positions	mm.
4	1, 2, 3, 4	0.10 ± 0.08
5	1, 2, 3, 4, 5	0.09 ± 0.07
6	1, 2, 3, 4, 5, 6	**0.07 ± 0.06**

Table 6.4 – Results on the open access synthetic dataset [82]. The methods are compared on seven vendors in apical four chamber view. The metric used is the average endpoint error expressed in mm. The application of the Wilcoxon signed-rank test shows the statistical difference ($p<0.0001$) of c-PWC-Net-gray-usft with the methods for which we have the results for all the patients (referred by *).

Methods	*ESAOTE*	*GE*	*HITACHI*	*PHILIPS*	*SIEMENS*	*TOSHIBA*	*SAMSUNG*
	mm.	mm.	mm.	mm.	mm.	mm.	mm.
Farnebäck	0.08 (0.06)	0.09 (0.07)	**0.06 (0.04)**	0.08 (0.06)	0.06 (0.05)	0.07 (0.05)	0.07 (0.05)
PIV*	0.10 (0.08)	0.14 (0.11)	0.11 (0.08)	0.09 (0.07)	0.09 (0.08)	0.09 (0.07)	0.09 (0.07)
PWC-Net-gray-usft [51]	0.14 (0.10)	0.17 (0.12)	0.13 (0.09)	0.14 (0.10)	0.14 (0.10)	0.14 (0.11)	0.13 (0.09)
EchoPWC-Net-us [51]	**0.07 (0.06)**	**0.07 (0.06)**	**0.06 (0.04)**	**0.06 (0.05)**	0.06 (0.05)	**0.06 (0.04)**	**0.05 (0.04)**
PWC-Net-gray-usft (ours)*	0.08 (0.07)	0.10 (0.07)	0.07 (0.04)	0.09 (0.06)	0.08 (0.06)	0.07 (0.05)	0.08 (0.06)
c-PWC-Net-gray*	0.15 (0.13)	0.34 (0.72)	0.11 (0.08)	0.12 (0.10)	0.09 (0.08)	0.12 (0.10)	0.13 (0.08)
c-PWC-Net-gray-usft	**0.07 (0.06)**	0.08 (0.07)	**0.06 (0.04)**	0.07 (0.06)	**0.04 (0.03)**	**0.06 (0.04)**	0.07 (0.05)

from [51]. Concerning c-PWC-Net-gray-usft, the first training performed on the gray-scaled FlyingChairs and FlyingThings datasets was performed by splitting the full set of images into training and validation sets. On the validation set, the network achieved an average EPE of 1.53 ± 5.28 pixels. The comparison of the results obtained by the PWC-Net-gray-usft method using either the transfer learning strategy given in [51] or the one we proposed in Section 6.2.2.3 clearly illustrates the relevance of the choices we made. Indeed, using the same original PWC-Net architecture, our transfer learning strategy yields an overall improvement of 42% in the average EPE on the full dataset, from 0.14 mm to 0.08 mm. In addition, it appears that the architectural modifications we proposed further improve the results obtained by c-PWC-Net-gray-usft by reducing the average EPE by 0.02 mm. It is also worth noting that the two non-DL methods outperformed c-PWC-Net-gray, which gave notably bad results on the GE fold. This can be explained by the fact that c-PWC-Net-gray is a method trained only on simulated natural images. Its performance is naturally reduced on echocardiographic images. In the open access dataset of Alessandrini *et al.*, the simulated sequences from the GE vendor are found to be the most challenging in terms of image quality with the least defined speckle pattern, making the tracking task more difficult. This observation is confirmed by the fact that all evaluated methods score their worst on this fold. Finally, the best performing methods were obtained by the two DL techniques EchoPWC-Net-us and c-PWC-Net-gray-usft which reached the same average EPE of 0.06 ± 0.05 mm.

6.4.1.3 Proposed Synthetic US dataset

We conducted experiments to assess the added value of the proposed synthetic ultrasound dataset described in Section 6.3.1.2. The corresponding results are given in Tables 6.5 and 6.6. In these experiments, the training dataset was composed of the open access dataset proposed in [82] augmented with data from 60 virtual patients, while the validation and testing datasets were composed of the data from the 10 and 30 remaining virtual patients. Patients were randomly selected to create the different datasets. Several trainings were performed by varying the number of added virtual patients (from 20 to 60) and by including or not the data with reverberation artifacts for the same patients. In these tables, the different experiments are named according to the number of added patients, with an A appended if the simulated patients were

included with and without artefacts. For instance, c-PWC-Net-20A refers to c-PWC-Net trained using the open access dataset augmented with 20 virtual patients with and without reverberation artifacts. We compared the performance of the c-PWC-Net with the block-matching PIV approach. In these experiments, c-PWC-Net outperformed PIV for most of the metrics and training configurations.

In terms of geometric scores, it is worth noting the continuous improvement in overall scores with the increasing amount of synthetic data both in terms of mean and standard deviation, from 1.53 ± 1.12 mm to 0.73 ± 0.49 mm for the d_m metric and from 0.14 ± 0.11 mm to 0.07 ± 0.06 mm for the EPE computed on the testing dataset without artifact. The same trend can be observed with the computation of the cumulative EPE between ED and ES reported in Table 6.7, *i.e.* an improvement of c-PWC-Net from 2.66 ± 1.59 mm to 1.20 ± 0.67 mm when adding the 60 virtual patients. The same conclusions can be drawn on the data with artifacts, with an improvement from 1.64 ± 1.16 mm to 0.79 ± 0.53 mm for the d_m metric and from 0.16 ± 0.13 mm to 0.08 ± 0.06 mm for the EPE. The positive impact of including synthetic data with and without reverberation artifacts appeared systematically and for all metrics in Table 6.5.

For the clinical scores given in Table 6.6, the same trends as for the geometric metrics can be drawn, with an overall improvement of the different scores with the inclusion of more synthetic data. Concerning the bias and standard deviations for the two clinical indices, the main improvement occurred after adding the first 20 patients, from 2.95 ± 3.00 % to 0.57 ± 1.73 % for the GLS and from 4.72 ± 3.54 % to 1.45 ± 2.47 % for the MAPSE on the testing dataset without artifact. The addition of more data resulted in a stagnation of the bias, but a decrease in the standard deviation, leading to an overall improvement of both indices. The same happened for the testing dataset with reverberation artifacts. For both clinical indices, the addition of the 60 simulated patients (with or without artifacts) allowed a final improvement around 80% for the bias and 65% for the mae. A correlation plot between the estimated GLS and the ones from the ground truth is shown on Figure 6.10. A correlation coefficient of 0.95 was achieved with our method, demonstrating the great capacity of our method to estimate the strain on simulated dataset. We noticed that PIV obtained a lower correlation coefficient than our DL method (0.88), which corroborated the results obtained in the

Table 6.5 – Geometric results on the open access synthetic dataset [82] complemented with the proposed simulated database. The p-value computed on the EPE with the Mann-Whitney U test between PIV and c-PWC-Net-60A was equal to $6e^{-6}$, proving the statistical difference between these two methods.

Methods*	***Simulations***			***Artifacts***		
	EPE$\pm\sigma$	$d_m \pm \sigma$	$d_H \pm \sigma$	EPE$\pm\sigma$	$d_m \pm \sigma$	$d_H \pm \sigma$
	mm.	mm.	mm.	mm.	mm.	mm.
PIV	0.26 ± 0.18	1.63 ± 0.97	4.64 ± 1.88	0.27 ± 0.19	1.86 ± 1.17	5.02 ± 2.02
c-PWC-Net-0	0.14 ± 0.11	1.53 ± 1.12	3.76 ± 1.36	0.16 ± 0.13	1.64 ± 1.16	4.07 ± 1.51
c-PWC-Net-20	0.09 ± 0.07	0.87 ± 0.59	2.37 ± 0.68	0.10 ± 0.09	1.06 ± 0.79	2.68 ± 0.84
c-PWC-Net-20A	0.08 ± 0.06	0.84 ± 0.57	2.32 ± 0.66	0.09 ± 0.07	0.90 ± 0.61	2.30 ± 0.64
c-PWC-Net-40	0.08 ± 0.06	0.79 ± 0.53	2.12 ± 0.61	0.09 ± 0.08	0.96 ± 0.70	2.36 ± 0.75
c-PWC-Net-40A	0.08 ± 0.06	0.78 ± 0.52	2.05 ± 0.56	$\mathbf{0.08 \pm 0.06}$	0.84 ± 0.57	2.13 ± 0.60
c-PWC-Net-60	0.08 ± 0.06	$\mathbf{0.73 \pm 0.49}$	$\mathbf{1.98 \pm 0.54}$	0.09 ± 0.07	0.92 ± 0.70	2.34 ± 0.74
c-PWC-Net-60A	$\mathbf{0.07 \pm 0.06}$	$\mathbf{0.73 \pm 0.49}$	2.03 ± 0.56	$\mathbf{0.08 \pm 0.06}$	$\mathbf{0.79 \pm 0.53}$	$\mathbf{2.08 \pm 0.60}$

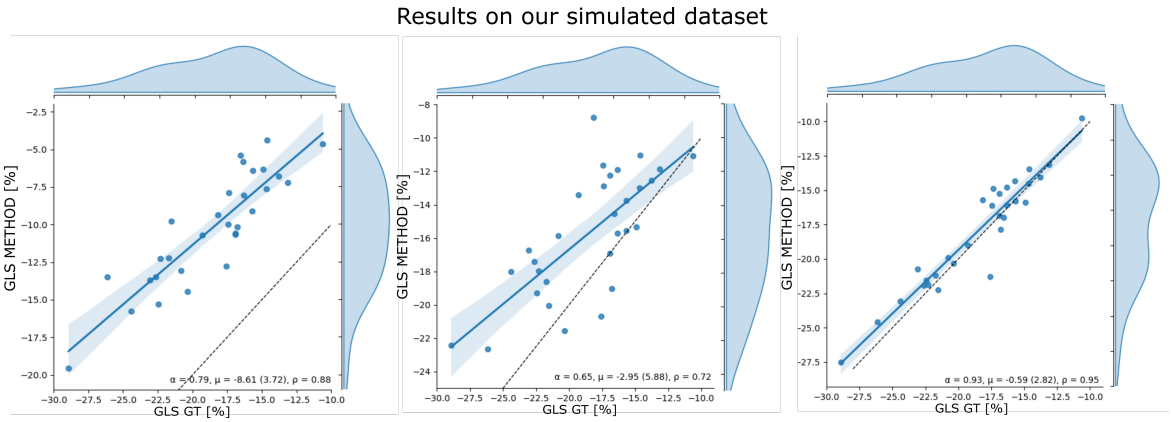


Figure 6.10 – Correlation plot of GLS computed on the endocardial contour between the PIV (left), c-PWC-Net (middle) and c-PWC-Net-60A (right) estimates and ground truths. Each point represents one of the 30 patients in the *testing simulated dataset*. Results are reported in the bottom right corner on the figure with α the slope of the regression line, μ the bias, the limits of agreement in parenthesis and ρ the correlation coefficient.

Table 6.6.

6.4.2 Clinical Data

6.4.2.1 Real patients from the CAMUS dataset

The performance of c-PWC-Net was assessed on the clinical data described in Section 6.3.1.3. PIV, c-PWC-Net and c-PWC-Net-60A, which performed best on the synthetic ultrasound dataset, were compared in this experiment. The comparison of our method with EchoPWC-Net would have been of interest. Unfortunately this is impossible since neither the trained model nor the test dataset were made publicly available by the authors. The geometric scores are given in Table 6.8. Since real motions are not known, only d_m and d_H metrics were computed from the tracked contours. c-PWC-Net outperformed the block-matching PIV method, with an improvement of 0.72 mm for the d_H metrics. Another interesting point is the improvement of our model brought by the proposed synthetic dataset, resulting in an overall performance of 1.86 ± 1.05 mm for d_m and 3.81 ± 1.18 mm for d_H. Table 6.9 lists the scores obtained

Table 6.6 – Clinical metrics on the open access synthetic dataset [82] complemented with the proposed simulated database.

Methods	*Simulations*				*Artifacts*			
	GLS		*MAPSE*		*GLS*		*MAPSE*	
	bias±σ	mae±σ	bias±σ	mae±σ	bias±σ	mae±σ	bias±σ	mae±σ
	%.	%.	%	%	%.	%.	%	%
PIV	4.89 ± 1.69	4.89 ± 1.69	2.77 ± 1.25	2.77 ± 1.25	6.01 ± 2.12	6.01 ± 2.12	3.20 ± 1.24	3.20 ± 1.24
c-PWC-Net-0	2.95 ± 3.00	3.40 ± 2.46	3.26 ± 2.13	3.30 ± 2.06	4.72 ± 3.54	4.94 ± 3.21	4.34 ± 2.45	4.39 ± 2.35
c-PWC-Net-20	0.57 ± 1.73	1.56 ± 0.90	0.56 ± 1.36	1.17 ± 0.88	1.45 ± 2.47	2.43 ± 1.47	0.98 ± 1.45	1.48 ± 0.91
c-PWC-Net-20A	0.48 ± 1.72	1.54 ± 0.88	0.43 ± 1.24	1.03 ± 0.79	0.70 ± 1.86	1.70 ± 0.99	0.54 ± 1.20	1.09 ± 0.72
c-PWC-Net-40	0.68 ± 1.47	1.35 ± 0.87	0.69 ± 1.13	1.08 ± 0.74	1.57 ± 2.11	2.29 ± 1.24	1.10 ± 1.33	1.47 ± 0.88
c-PWC-Net-40A	0.83 ± 1.30	1.26 ± 0.87	0.60 ± 1.00	0.93 ± 0.70	0.97 ± 1.60	1.57 ± 1.00	0.64 ± 1.03	1.02 ± 0.64
c-PWC-Net-60	**0.66 ± 1.31**	**1.20 ± 0.82**	0.60 ± 0.89	0.89 ± 0.60	1.50 ± 1.89	2.08 ± 1.20	1.00 ± 1.06	1.25 ± 0.73
c-PWC-Net-60A	0.59 ± 1.44	1.27 ± 0.87	**0.55 ± 0.90**	**0.85 ± 0.62**	**0.71 ± 1.55**	**1.42 ± 0.92**	**0.63 ± 0.92**	**0.95 ± 0.58**

Table 6.7 – EPE computed between ED and ES on the simulated dataset with and without artifacts. These two time instants are separated on average by 21 frames in our database.

Methods	***Simulations***	***Artifacts***
	EPE$\pm\sigma$	EPE$\pm\sigma$
	mm.	mm.
PIV	10.29 ± 2.40	7.86 ± 2.66
cPWC-Net	2.66 ± 1.59	2.7 ± 1.49
cPWC-Net-60	**1.20 ± 0.67**	**1.29 ± 0.70**

for the GLS and MAPSE clinical indices. The same trends as observed for the geometric metrics can also be drawn. Indeed, our DL solution outperformed PIV while the use of the synthetic dataset significantly improved the different clinical scores, with a mae from 4.29 ± 2.84 % to 2.55 ± 2.08 % for the GLS and from 4.19 ± 2.78 % to 2.62 ± 2.09 % for the MAPSE. As an illustration of our quality of tracking, we displayed on Figures 6.14 and 6.15 the evolution of the reference and the estimated GLS for all the test patients in the real CAMUS dataset. We also analyzed the results obtained on these 30 patients according to the quality of the sequences. From the Table 6.10, one can see that the tracking error is almost the same whatever the image quality. This can be explained by the fact that the two types of image quality were present in the synthetic dataset used during training. As a complement, a correlation plot between the estimated GLS and the ones from the ground truth is shown on Figure 6.11. A correlation coefficient of 0.77 was achieved, demonstrating the capacity of our method to reproduce manual annotations with good fidelity. It is also important to note that our correlation score was not as good as the one obtained in the recent study of [51]. This can be explained by the fact that the authors compare their method with the results obtained by commercial methods usually based on conventional speckle tracking algorithms. Using manual expert annotations as references can be more difficult because they are not based solely on image information and incorporate a variety of complex information that may be subjective. Finally, thanks to the values reported in Table 6.10, we observed that c-PWC-Net-60A has stable performances even for images of lower quality. This can be explained by the fact that the training dataset included a large range of image quality, and is encouraging as for the generalization capability of our algorithm.

Table 6.8 – Geometric results obtained on a subset of the CAMUS dataset composed of 30 real patients acquired with a GE system.

Methods	$d_m \pm \sigma$	$d_H \pm \sigma$
	mm.	mm.
PIV	2.27 ± 1.30	5.36 ± 2.07
c-PWC-Net	2.22 ± 1.34	4.64 ± 1.62
c-PWC-Net-60A	**1.86 ± 1.05**	**3.81 ± 1.18**

Table 6.9 – Clinical results obtained on a subset of the CAMUS dataset composed of 30 real patients acquired with a GE system.

Methods	***GLS***		***MAPSE***	
	bias$\pm\sigma$	**mae**$\pm\sigma$	**bias**$\pm\sigma$	**mae**$\pm\sigma$
	%.	%.	%.	%.
PIV	7.35 ± 3.42	7.35 ± 3.42	5.66 ± 3.43	5.66 ± 3.43
c-PWC-Net	3.96 ± 3.30	4.29 ± 2.84	3.90 ± 3.19	4.19 ± 2.78
c-PWC-Net-60A	**1.85 ± 2.73**	**2.55 ± 2.08**	**1.83 ± 2.83**	**2.62 ± 2.09**

Table 6.10 – Clinical results according to the image quality obtained on a subset of the CAMUS dataset composed of 30 real patients acquired with a GE system.

Methods	d_m	d_H	***GLS***		***MAPSE***	
	mean$\pm\sigma$	**mean**$\pm\sigma$	**bias**$\pm\sigma$	**mae**$\pm\sigma$	**bias**$\pm\sigma$	**mae**$\pm\sigma$
	mm.	mm.	%.	%.	%.	%.
c-PWC-Net-60A (#30)	1.86 ± 1.05	3.81 ± 1.18	1.85 ± 2.73	2.55 ± 2.08	1.83 ± 2.83	2.62 ± 2.09
Good Quality (#24)	1.84 ± 1.05	3.87 ± 1.21	1.71 ± 2.88	2.53 ± 2.16	1.77 ± 3.02	2.72 ± 2.15
Medium Quality (#6)	1.93 ± 1.04	3.56 ± 1.05	2.42 ± 2.18	2.62 ± 1.88	2.05 ± 2.15	2.22 ± 1.93

6.4.2.2 Real patients from the auxiliary dataset

The generalization capacity of c-PWC-Net-60A was assessed using the auxiliary dataset described in Section 6.3.1.3. This dataset contains echocardiographic sequences acquired exclusively from another hospital with a system from a different vendor than the one used to create the synthetic data. Moreover, this dataset was not used to generate new synthetic cases, so our algorithm never integrated this new type of data during its learning phase. From Table 6.11, it can first be observed that the geometric scores remain unchanged, with a mean value of 1.81 ± 1.11 mm for d_m and 3.45 ± 1.11 mm for d_H. It is also interesting to see that the quality of the tracking is homogeneous with respect to the type of pathology, with a variability of 0.58 mm for d_m and 0.88 for d_H.

Table 6.12 shows that the same trends are true for the clinical scores. Indeed, c-PWC-Net-60A obtained very similar results on the Philips dataset compared to the GE one, with a mae of 2.89 ± 2.08 % and 2.86 ± 1.88 % for the GLS and MAPSE, respectively. These results are also consistent between the different pathological groups. Figure 6.12 shows the correlation plot between the estimated GLS and the reference ones. Interestingly, a correlation coefficient of 0.93 was obtained, which can be explained by an overall better image quality in the Philips dataset. We display in Table 6.13 the reference and estimated mean GLS for each group from the dataset. One can clearly see that pathological patients lead to lower GLS values compared to healthy subjects and that our accuracy seems reasonable to detect altered GLS strain values.

During the manual contouring of the testing datasets, the points used to define the reference contours were selected independently from one frame to another by the expert

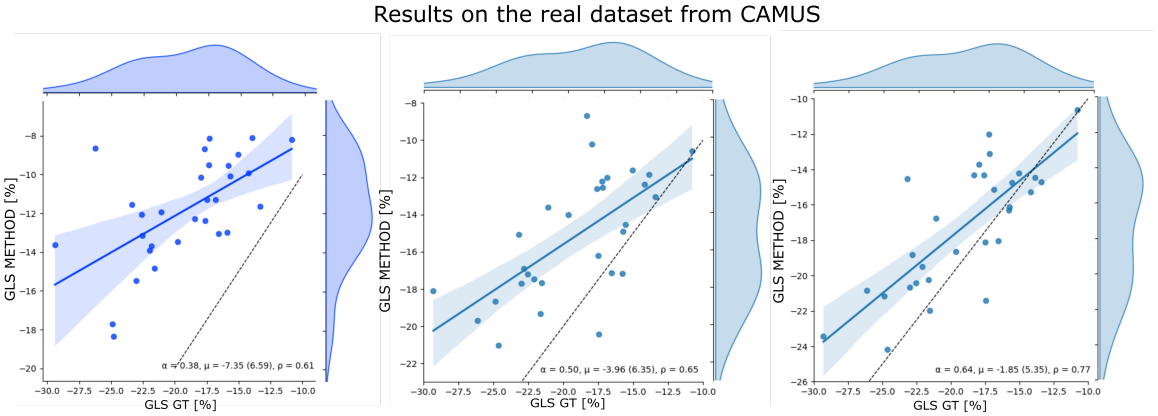


Figure 6.11 – Correlation plot of GLS computed on the endocardial contour between the PIV (left), c-PWC-Net (middle) and c-PWC-Net-60A (right) estimates and ground truths. Each point represents one of the 30 patients in the *real CAMUS dataset*. Results are reported in the bottom right corner on the figure with α the slope of the regression line, μ the bias, the limits of agreement in parenthesis and ρ the correlation coefficient.

Table 6.11 – Geometric results obtained with the c-PWC-Net-60A method on an auxiliary dataset composed of 30 real patients acquired with a Philips system.

Philips dataset	$d_m \pm \sigma$	$d_H \pm \sigma$
	mm.	mm.
Full dataset (#30)	1.81 ± 1.11	3.45 ± 1.11
Aortic Stenosis (#6)	1.72 ± 1.11	3.24 ± 1.02
Hypertrophic Cardiomyopathy (#6)	2.15 ± 1.26	3.91 ± 1.36
Ischemic (#6)	1.67 ± 1.08	3.38 ± 1.09
Non Ischemic (#6)	1.57 ± 0.95	3.03 ± 0.96
Normal (#6)	1.93 ± 1.14	3.69 ± 1.14

cardiologist. We therefore have no reference point tracked over the cardiac cycle, which prevents us from computing the reference regional strain. Nevertheless, we conducted an additional experience to assess the spatial distribution of the distance errors between the reference and the estimated contours. In particular, the orientation of the long axis was first extracted from the first frame of the processed sequence. For each contour of the sequence, apical points were defined as the intersection between the given contour and the line passing by the mid-basal point with the same orientation as the one computed on the first frame. A normalized parametrization for each contour was then computed, where 0 corresponds to the basal anterolateral point, 0.5 corresponds to the apex point and 1 corresponds to the basal inferoseptum point. Mean absolute distance between the reference and the estimated contours were finally computed along the parametric axis. Figure 6.13 displays the results obtained from the auxiliary dataset. Each curve corresponds to one of the 30 evaluated patients. On this figure, we can see that the errors are relatively homogeneous on each side of the myocardium, with slightly higher average values for the lateral side and at the apex (*i.e.* a mean error of 1.4 mm on the septal side and 1.9 mm on the lateral side).

Table 6.12 – Clinical results obtained with the c-PWC-Net-60A method on an auxiliary dataset composed of 30 real patients acquired with a Philips system.

Philips dataset	***GLS***		***MAPSE***	
	bias±σ	**mae±σ**	**bias±σ**	**mae±σ**
	%.	%.	%.	%.
Full dataset (#30)	2.85 ± 2.14	2.89 ± 2.08	2.74 ± 2.05	2.86 ± 1.88
Aortic Stenosis (#6)	2.85 ± 2.14	2.85 ± 2.14	2.22 ± 1.91	2.46 ± 1.52
Hypertrophic Cardiomyopathy (#6)	3.33 ± 2.26	3.33 ± 2.26	3.51 ± 2.27	3.51 ± 2.27
Ischemic (#6)	2.48 ± 1.59	2.50 ± 1.56	2.98 ± 1.29	2.98 ± 1.29
Non Ischemic (#6)	1.82 ± 1.92	2.01 ± 1.67	2.51 ± 1.77	2.51 ± 1.77
Normal (#6)	3.75 ± 2.84	3.75 ± 2.84	2.51 ± 3.08	2.85 ± 2.69

Table 6.13 – Reference and estimated mean GLS for each group of six patients auxiliary dataset composed of 30 real patients acquired with a Philips system.

Methods	**Full Database**	**AS**	**HCM**	**ISCH**	**NON ISCH**	**NORMAL**
	bias±σ	**bias±σ**	**bias±σ**	**bias±σ**	**bias±σ**	**bias±σ**
	%.	%.	%.	%.	%.	%.
Reference	-15.43 ± 5.48	-15.53 ± 4.00	-14.96 ± 5.85	-11.82 ± 6.44	-14.10 ± 4.04	-20.74 ± 3.77
Estimated	-12.58 ± 4.32	-12.68 ± 2.42	-11.63 ± 4.83	-9.34 ± 5.72	-12.29 ± 2.65	-16.99 ± 1.45

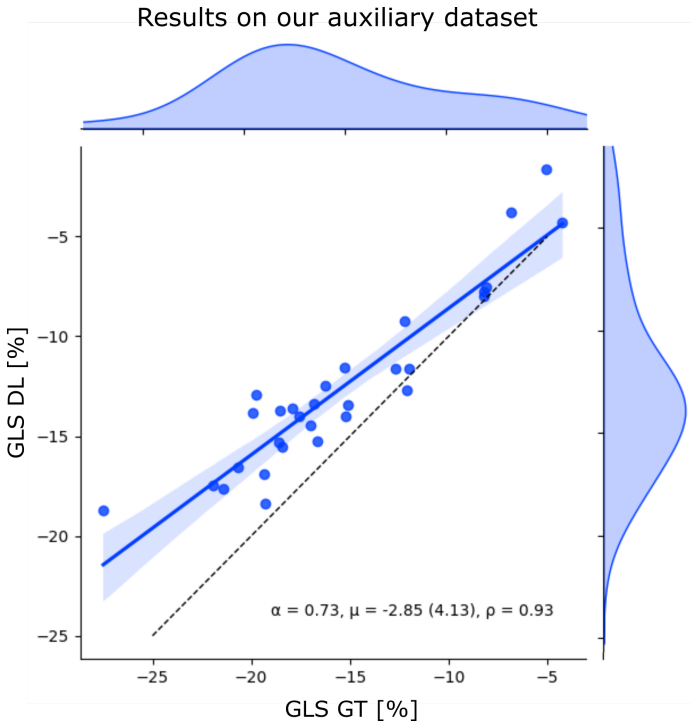


Figure 6.12 – Correlation plot of GLS computed on the endocardial contour between c-PWC-Net-60A estimates and ground truths. Each point represents one of the 30 patients in the *real auxiliary dataset*. Results are reported in the bottom right corner on the figure with α the slope of the regression line, μ the bias, the limits of agreement in parenthesis and ρ the correlation coefficient.

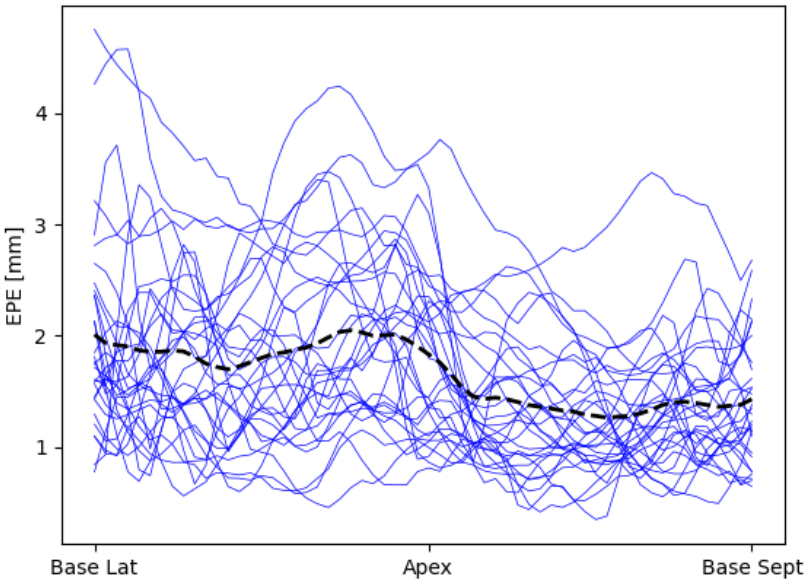


Figure 6.13 – Evolution of the distance error along the endocardium between the reference contours and the estimated ones using c-PWC-Net-60A. Each blue contour corresponds to the mean error computed over the cardiac cycle for one patient. The mean curve is represented in black.

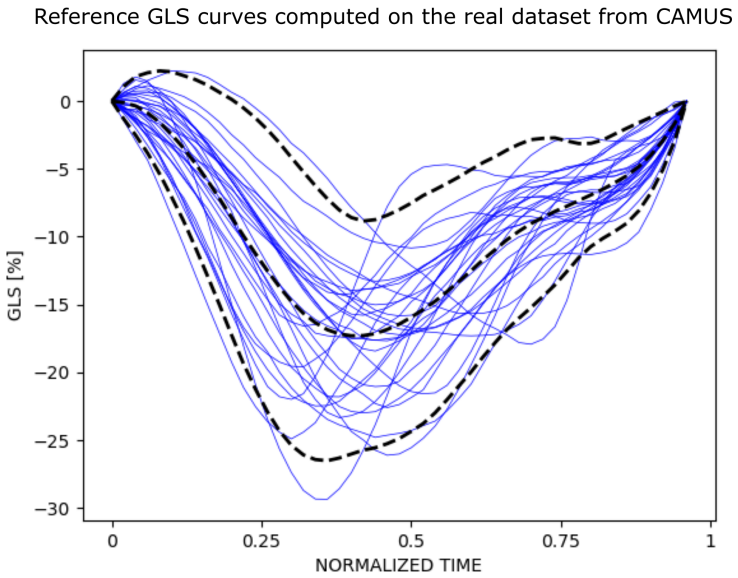


Figure 6.14 – Evolution of the *reference* GLS as a function of time for all the test patients in the *real CAMUS dataset*. The mean and the limits of agreement are represented by the black dashed lines. GLS curves were normalized along the time axis for display on a common graph.

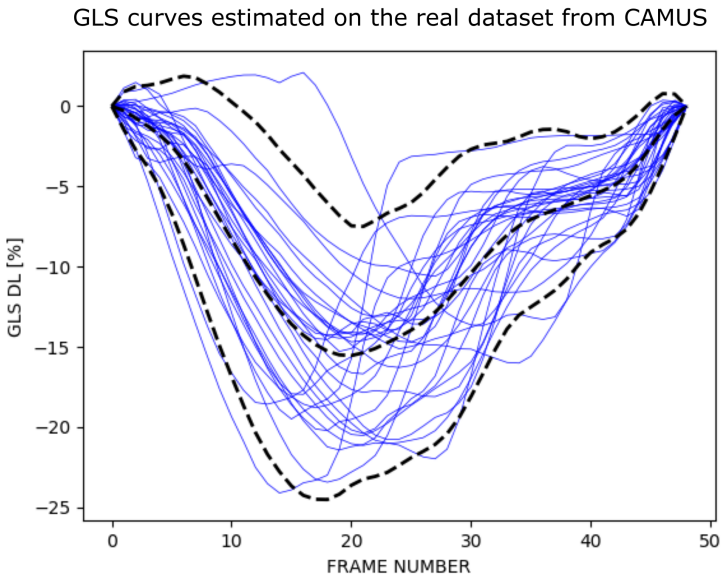


Figure 6.15 – Evolution of the *estimated* GLS as a function of time for all the test patients in the *real CAMUS dataset*. The mean and the limits of agreement are represented by the black dashed lines. GLS curves were normalized along the time axis for display on a common graph.

6.5 Discussion

6.5.1 A new open access simulated ultrasound dataset

The open access dataset of [82] was generated with synthetic deformation in lower ranges than normal strain values. Moreover, the complex personalization procedure in [82] limited the variability of geometries and motion types that can be simulated. This limits the relevance of using this dataset alone for DL training. Therefore, we designed a solely image-based pipeline, bypassing the need for an electromechanical model. This allowed to simulate many cases from B-mode template cine loops on which myocardial contours were manually annotated to generate synthetic motion fields. As illustrated in Figure 6.7, global deformation ranges with our simulation method match those of real sequences, although at a finer local scale disparities can occur. It is thus important to exploit this synthetic dataset with care. Indeed, although the corresponding global clinical indices seem relevant, the way in which the baseline myocardial motion was generated does not allow our dataset to be used to evaluate motion estimation algorithms. The accuracy of the myocardial motion pattern is necessarily limited and the proposed simulation pipeline should be viewed as a synthetic ultrasound sequence generator for data augmentation purposes only. With this in mind, we used this simulation pipeline to augment the existing open-access dataset of [82] and produced a more diverse synthetic dataset. Tables 6.5 and 6.6 showed the interest of our simulation strategy, with an improvement of the geometric and clinical scores when increasing the number of virtual patients in the training dataset. Specifically, the geometrical errors were mostly reduced by a factor of two thanks to the addition of the 60 simulated patients. Our pipeline can also simulate physical artifacts to improve the robustness of DL to these application-specific sources of noise. We focused in this

study on reverberation artifacts. As shown in Table 6.6, incorporating these artifacts in the training dataset significantly improved both the GLS and MAPSE scores, if the same type of artifact is present in the testing dataset. This validates the relevance of generating synthetic images with artifacts as a data augmentation procedure. The full set of the simulated data is made publicly available. We claim that the access to this synthetic dataset in addition to the one proposed in [82] will provide valuable and complementary tools for the research community. To better evaluate the quality of the simulated dataset, several videos of synthetic sequences are available at: `http://humanheart-project.creatis.insa-lyon.fr/medicaid.html`. Finally, it is worth noting that the manual annotation step involved in the current version of the simulation pipeline can be time consuming and is the main bottleneck to automatically deploy our solution on more than 100 patients regardless of the type of view during acquisition. This is the reason why we limited our study to 100 patients acquired on the apical four chamber view. In the near future, we plan to work on a fully automated and controlled deployment of our simulation pipeline to be able to generate larger and richer synthetic databases in terms of diversity of cases, pathologies and acquisition view.

6.5.2 Interest of the training/inference strategies

The use of data augmentation is key for DL-based methods. It thus appeared appealing to investigate dedicated strategies for tracking in ultrasound. Table 6.1 shows a reduction of the mean and standard deviation of the EPE by a factor of 2 when using forward/backward and spaced image pairs. This can be explained by the increase in motion diversity of the dataset while preserving the decorrelation of the associated speckle pattern. In parallel, we proposed a strategy for inference that incorporates temporal consistency. Although it provided an 11% improvement in the mean EPE, this procedure doubles the inference time, limiting its application to scenarios for which computation time is not a strong constraint.

6.5.3 Efficiency of the proposed transfer learning solution

We evaluated the relevance of our transfer learning solution on both simulated and clinical data. We first benchmarked our network against EchoPWC-Net [51] on the dataset of [82]. As illustrated in Table 6.4, the results obtained by the two networks are similar, despite different choices to extend the original PWC-Net architecture for ultrasound image processing. The authors of EchoPWC-Net obtained their best results by learning directly from simulated echocardiographic images, without any transfer learning. In contrast, we opted for a transfer learning approach starting first on natural images, before transferring to ultrasound data. Contrary to [51], we assumed that first confronting the network with a wide variety of images and motion types would enable a better generalization ability and avoid overfitting. In [51], transfer learning was evaluated but performed poorly. There may be several reasons for this, including the limited variability in geometry and motion types of the synthetic dataset used for training and the lack of learning on the ultrasound synthetic data due to a low learning rate. In our case, the learning rate was left unchanged between initial and transfer learning phases. The interest of our transfer learning approach was further validated by the results obtained on clinical data. Indeed, Tables 6.8 and 6.9 showed that

our method still significantly outperformed the state-of-the-art block-matching PIV method, with 18% and 29% improvements in the d_m and d_H scores, respectively. This improvement was even more significant for the clinical indices, with a reduction of the mae from 7.3% to 2.5% for the GLS and from 5.7% to 2.6% for the MAPSE. It would have been interesting to compare our method with EchoPWC-Net on clinical data. Unfortunately, this was impossible as neither the testing dataset nor the commercial software they used were accessible.

6.5.4 Capacity of Generalization

Finally, we realized the first study on the generalization of DL methods for motion estimation in echocardiography. To this aim, we used two complementary datasets, the composition of which allowed us to conduct a multi-center, multi-vendor and multi-disease study. Tables 6.11 to 6.12 illustrate the strong ability of our DL solution to provide accurate and consistent results for a wide range of situations, from different ultrasound machines to several pathologies with different motion patterns. This adaptability confirms the relevance of our transfer learning strategy as well as the quality of the synthetic data we have generated. To better assess the quality of the obtained results, several videos of tracking of the endocardial contours from both GE and Philips datasets are provided at: `http://humanheart-project.creatis.insa-lyon.fr/medicaid.html`.

6.5.5 Perspectives

Although our results on the clinical data are convincing, we can observe a decrease of performance between simulation and clinical data. There is therefore room for improvement. For a generalization point of view, it would be more interesting to have the simulated reverberation artifacts appear at random positions during training. Unfortunately, this implies the generation of new synthetic images when creating each new batch, which is currently not possible due to the computation time of the simulation (around 2 minutes per synthetic image). An intermediate possibility would be to simulate for each patient several sequences with reverberation artifacts at different positions and to draw the corresponding images randomly when creating the different batches. An alternative to enhance the generalization capability of DL methods would be to simulate a richer dataset by changing the motion or reflection of the scatterers to increase the variability of the image quality. It would also be interesting to integrate out-of-plane motions as they contribute to the deterioration of the motion estimation in 2D. Given the successful generalization of our method to different manufacturers, a next step would be to evaluate our method on echocardiographic data acquired from other views such as apical two chamber or three chamber. In parallel, another way to increase the accuracy of our network would be to optimize more tasks on the same cine loop. For instance, joint optimization of motion estimation and segmentation tasks, as recently proposed in [98], may be an interesting track to investigate.

6.6 Conclusions

In this chapter, we developed a deep learning method for motion estimation in echocardiography. We showed that the combination of a customized version of PWC-Net with

a new simulated synthetic dataset and a dedicated data augmentation strategy outperforms the current state-of-the-art methods, both for the tracking of endocardial borders and the estimation of the GLS and MAPSE indices. The genericity of our approach was also demonstrated from the first multi-center, multi-vendor and multi-disease study. The proposed synthetic dataset consists of 2D apical four chamber view sequences for 100 virtual patients with or without reverberation artifacts and with the corresponding myocardial displacement fields. For open science purposes, the full dataset can be directly accessed at `http://humanheart-project.creatis.insa-lyon.fr/medicaid.html`.

Part IV

Conclusions

Chapter 7

Conclusions and perspectives

Contents

7.1 Conclusions

Echocardiography is a very popular clinical modality to diagnose cardiac pathologies. The analysis of cardiac function is performed through the measurement of relevant clinical indices such as the ventricular volumes and myocardial deformation. One typical index recommended in clinical practice guidelines for assessing myocardial deformation is the global longitudinal strain. The techniques used to estimate this index from the B-mode images currently suffer from a lack of accuracy and reproducibility. In this context, we proposed the following contributions:

1. We conducted the first pilot study to evaluate the ability of convolutional neural networks to estimate motion in ultrasound images. In particular, we demonstrated that convolutional neural networks can be trained efficiently for the task of speckle tracking in ultrasound. To this end, we studied the ability of these methods to estimate rotational motion on simulated and *in-vitro* images. Such movements are among the most difficult to estimate in ultrasound due to the strong decorrelation of the related speckle. In order to train and evaluate our different methods, we created and shared with the community an open access database composed of simulated and real sequences including spinning disks with the corresponding ground truths. From the results obtained on these databases, we showed the interest of using transfer learning on ultrasound data and were able to empirically determine architectural features improving the performance of neural networks in ultrasound imaging. Finally, this pilot study demonstrated that deep learning methods outperform traditional speckle tracking methods. They can produce robust and accurate results despite speckle decorrelation and large displacements.

2. We designed a pipeline dedicated to the generation of realistic B-mode sequences to overcome the difficulty of establishing databases with reference motion fields. The interest of this pipeline resides in its ability in simulating B-mode sequences with a wide variety of cardiac deformations that are realistic enough to serve as a relevant data augmentation for deep learning methods. We thus generated an open access synthetic dataset which consists of 2D apical four chamber view sequences for 100 virtual patients with or without reverberation artifacts and with the corresponding myocardial displacement fields.

3. We developed a dedicated deep learning solution for motion estimation in echocardiography based on the PWC-Net architecture. In particular, we proved the relevance of deploying a transfer learning strategy and using temporal data augmentation during the learning phase. Together with our synthetic dataset, we were able to estimate myocardial deformation with less than 3% error, which is much better than the performance of the state-of-the-art methods. The generiticity of our network was finally demonstrated on a multi-center, multi-vendor and multi-disease study.

7.2 Perspectives

My thesis provided a proof of concept that global longitudinal strain could be successfully assessed with dedicated deep learning methods. Still, several methodological

improvements can be made to consolidate this study.

Strengthening of the methodological aspects

Although our DL method beat the state-of-the-art methods and obtained less than 3% error on the GLS estimation, several possibilities are to be considered in order to consolidate the robustness and the accuracy of the measurements.

- Integration of key concepts from the natural language processing domain: recurrent networks as well as transformers have made a remarkable entry in the field of motion estimation. Indeed, these structures have been successfully integrated into motion estimation networks, leading to the current best performing methods (RAFT and GMA) on open access dataset with natural images. We can therefore expect that the effective implementation of these structures in networks dedicated to echocardiographic imaging will further improve the performance of these algorithms.

- Implementation of a multi-task strategy: our method can be learned in conjunction with other tasks than pure motion estimation, such as segmentation, to increase its performance. Indeed, the joint optimization of several tasks can lead to a reinforcement of the temporal coherence of the obtained segmentations as well as to a better precision in the tracking of the myocardial contours. Moreover, this would make our method self-sufficient since the current version requires a segmentation of the left ventricle on the initial image so to compute the endocardial contour on the whole sequence by successive estimates of the displacement.

Strengthening of the simulation pipeline

A major contribution of my thesis concerns the design of a solely image-based pipeline, bypassing the need for an electromechanical model. This strategy allowed to simulate many cases from B-mode template cine loops on which myocardial contours were manually annotated to generate synthetic motion fields with great diversity. Even if the generated synthetic database allowed us to significantly improve our results, our simulation pipeline could benefit from the following contributions:

- Automation of the entire pipeline: the manual annotation step involved in the current version of the simulation pipeline is the main bottleneck to automatically deploy our solution on more than 100 patients regardless of the type of view during acquisition. It would be interesting in the near future to replace this step by a fully automated procedure with quality control using a deep learning formalism. This will allow us the deployment of our pipeline on large datasets ($>$1000) on all apical views (two, three, four and five-chamber view). Taking into account the consequent improvement of the performances of our method with the addition of only 60 simulated patients during the training, the study of a scaling up of the training would be interesting to carry out in order to determine to what extent we can improve our results thanks to the richness brought by simulated data.

- Improvement of the diversity of the properties of the simulated images: one of the main advantages of our simulation pipeline is that it is generic enough to simulate a wide variety of physical properties. So, it would be interesting to exploit this characteristic to reinforce the diversity in the simulated images, for instance from a data augmentation perspective. The main properties targeted should be the variation of the amount of speckle decorrelation over time and the insertion of artifacts other than reverberation.

Investigation of clinical indices

In this thesis, the interest of using deep learning methods in the estimation of the GLS has been demonstrated. Indeed, we were able to achieve less than 3% error in its estimation on real data from different manufacturers, outperforming all other state-of-the-art methods we evaluated. These results should encourage the community to continue in this direction to strengthen the measurement of clinical indices related to myocardial deformation.

- Estimation of the GLS on other apical views: the GLS estimated on clinical data showed a low bias against references generated by expert cardiologists. In this context, we have made an important validation effort on B-mode images containing different pathologies acquired with ultrasound scanners from different manufacturers. However, the evaluation includes only data acquired i, apical four-chamber view. To further demonstrate the robustness of our method, it would be interesting to evaluate it on different apical views in order to ensure its ability to obtain the same quality of GLS estimation whatever the type of views.

- Estimation of the regional longitudinal strain: In this thesis, we were mainly focused on the study of global myocardial deformation. This measurement is currently widely used in clinical practice although regional analysis of myocardial deformation has demonstrated its strong complementary potential in the analysis of cardiac function. Currently, the lack of reproducibility of this regional measure prevents it from being added to clinical guidelines. One reason for this drawback is the presence of many artifacts and a low signal-to-noise ratio in some areas of the myocardium during motion, problems that we also faced during the global deformation study. It would therefore be interesting, in light of the results obtained in this thesis, to extend our method to the study of the reliability of the measurement of regional deformations. This would also imply the generation of simulated databases for which the reference would contain the deformation of each myocardial segment.

Part V

Appendices

Chapter 8

Résumé en français

Chapitre 1 : Introduction

Les maladies cardiovasculaires sont la première cause de décès dans le monde, représentant 32% des décès en 2019 [1]. Les facteurs de risques les plus courants pour ces maladies (hypercholesterolemie, hypertension et diabète) résultent souvent de mauvaises habitudes de vie (tabagisme, mode de vie sédentaire, surpoids) [2]. L'identification de ces facteurs de risques permet l'anticipation et le suivi régulier de la fonction cardiaque dans le but de prévenir de futures complications. Ce suivi est effectué de diverses manières et notamment par l'imagerie cardiaque. En effet, la plupart de ces pathologies impactent la déformation du myocarde qui peut être détecté lors d'examens médicaux.

En cardiologie, l'imagerie ultrasonore est la modalité la plus largement utilisée car elle est rapide, non invasive et peu onéreuse [3]. L'échocardiographie permet l'analyse de la fonction cardiaque par l'extraction de métriques cliniques pertinentes comme la contraction du myocarde ou la fraction d'éjection. Le principal indice recommandé en clinique pour l'estimation de la déformation du myocarde est la déformation longitudinale globale (GLS) [4]. Le GLS est calculé à partir des images mode B acquises en vue apicale comme le pourcentage de raccourcissement du myocarde entre les instants fin de diastole et fin de systole [5]. Les solutions actuelles utilisées pour estimer la déformation du myocarde souffrent d'un manque de reproductibilité des mesures dû aux caractéristiques inhérentes à l'échographie (nombreux artefacts, manque d'information et décorrélation de speckle). Les méthodes d'estimation de mouvement en imagerie ultrasonore sont principalement fondées sur le flux optique et la correspondance de blocs.

Récemment, les méthodes d'apprentissage profond sont devenues les méthodes les plus performantes dans tous les domaines de vision par ordinateur. L'estimation de mouvement ne fait pas exception à la règle où les méthodes par apprentissage profond obtiennent les meilleurs résultats sur les bases de données d'images naturelles. Parmi ces méthodes, l'apprentissage supervisé a montré une plus grande efficacité que les méthodes non-supervisées. En estimation de mouvement, les bases de données simulées sont obligatoires pour l'évaluation et l'apprentissage de réseaux de neurones. En effet, seules les simulations permettent d'avoir la référence du mouvement pour chaque pixel de l'image sur une séquence entière.

Face aux excellentes performances obtenues sur les simulations d'images naturelles, le développement et l'amélioration de méthodes d'apprentissage profond pour estimer

la déformation du muscle myocardique en imagerie ultrasonore 2D constituent un axe de travail particulièrement intéressant à explorer. Dans ce contexte, les objectifs de cette thèse sont les suivants :

- Démontrer la capacité des réseaux de neurones convolutionnels à évaluer le mouvement en imagerie ultrasonore.
- Développer une méthode robuste d'apprentissage profond afin d'estimer le mouvement du myocarde et sa déformation en imagerie échocardiographique.
- Partager avec la communauté scientifique les bases de données synthétiques générées au cours de mes travaux dans le but d'entraîner des réseaux de neurones et d'évaluer la précision de ces méthodes.

Chapitre 2 : Echocardiographie

Anatomie et physiologie des structures cardiaques

Le coeur est un muscle qui pompe le sang vers le reste du corps humain à travers le système circulatoire. Il comprend 2 parties séparées par le septum : le cœur droit qui collecte le sang désoxygéné et l'envoie vers les poumons pour réoxygénation via l'artère pulmonaire et le cœur gauche qui collecte le sang oxygéné et l'envoie dans le reste du corps via l'aorte. Le muscle du cœur, nommé le myocarde, est entouré de deux couches : l'endocarde dans les chambres (oreillettes et ventricules) et l'épicarde vers le péricarde. Il se contracte sous l'impulsion de signaux électriques pour expulser le sang hors des ventricules en systole et se remplir durant la diastole.

Formation de l'image ultrasonore

Les ondes ultrasonores utilisées en imagerie cardiaque ont une fréquence comprise entre 2.5 MHz et 5 MHz. Ces ondes sont générées au niveau des sondes par les cristaux piezoelectriques qui transforment un signal électrique en ondes acoustiques et inversement. Une fois les ondes réfléchies réceptionnées, une technique basée sur la sommation des amplitudes des signaux reçus suivant des hyperboles est appliquée afin de créer un ensemble de signaux radio-fréquence (RF). Cette étape se nomme la formation de voies. Ces signaux sont ensuite démodulés afin de créer des signaux complexes dénommés signaux IQ (IQ signifie en anglais In-phase and Quadrature) qui possèdent le même contenu d'information mais à partir de moins d'échantillons que les signaux RF. L'enveloppe des signaux IQ est extraite puis compressée par une loi logarithmique afin de créer l'image mode B finale.

Lors de la propagation des ondes acoustiques dans les tissus, différents phénomènes liés aux propriétés du milieu ont lieu. Par exemple, lors de la propagation dans les tissus mous, les ondes ultrasonores subissent une atténuation proportionnelle à la distance parcourue. Cette atténuation est compensée en réception par l'application d'un gain des échos en réception en fonction du temps. Cette opération est nommée TGC en anglais pour Time Gain Compensation. Lors du passage d'un milieu à un autre, une partie de l'onde est transmise et une autre est réfléchie générant un écho renvoyé à la sonde. La quantité d'onde réfléchie est proportionnelle au coefficient de réflexion des milieux. Si la taille des éléments causant la réflection de l'onde est plus petite que la longueur d'onde de l'onde émise, la réflexion laisse place à la diffusion et les éléments sont nommés diffuseurs. La combinaison des phénomènes de diffusion avec des interférences constructives et destructives créée le motif de speckle. La taille et la forme du speckle sont liées au nombre et à la position des diffuseurs par rapport à la sonde par cellule de résolution. Le motif résultant peut être vu comme une signature locale du tissu pouvant être suivie par les algorithmes traditionnels pour en estimer le mouvement.

Etude de la déformation du myocarde

Plusieurs métriques cliniques peuvent être mesurées à partir des images en mode B acquises en échocardiographie. Parmi les métriques utilisées pour l'analyse fonctionnelle du myocarde, la déformation longitudinale globale (GLS) ainsi que la déformation lon-

gitudinale régionale sont les plus connues [16]. Lors du calcul du GLS, la longueur L du myocarde est extraite à chaque instant t. Cette longueur est calculée soit au centre du muscle, soit au niveau de la paroi endocardique. Le calcul de la déformation se fait de la manière suivante [17] :

$$S_L = \frac{L(t) - L(t_0)}{L(t_0)}, \tag{8.1}$$

où t_0 correspond à l'instant en fin de diastole.

Pour calculer la déformation longitudinale régionale, le myocarde est divisé en differentes régions appelées segment et une estimation de la déformation du muscle dans chaque région est alors réalisée. L'intérêt de cet indice est de permettre de détecter localement des problèmes de contraction du muscle cardiaque. Malheureusement cet indice souffre d'un manque de reproductibilité [18]. Contrairement à cet indice, le GLS fait désormais partie des mesures recommandées à effectuer lors des examens cliniques grâce à une meilleure reproductibilité [4]. Il est important de noter que le GLS calculé à partir des contours endocardiques prend en compte une partie du mouvement radial en plus de la déformation longitudinale [19]. Une autre métrique a donc été développée pour ne prendre en compte que la déformation strictement longitudinale : le MAPSE [20]. Cet indice est calculé de la même manière que le GLS, à la différence que L désigne la distance entre l'apex et le point médian de la base du ventricule gauche.

Conclusions

Nous avons vu dans ce chapitre que l'anatomie et la fonction du cœur sont contrôlées dans la pratique clinique en utilisant l'imagerie ultrasonore. Plusieurs indices sont utilisés pour analyser et quantifier la fonction cardiaque. Même si certaines mesures telles que la fraction d'éjection ou le GLS sont considérées comme suffisamment robustes pour faire partie des directives cliniques, la plupart d'entre elles souffrent encore d'un manque de reproductibilité dû à la nature bruitée des images échographiques et à la présence de nombreux artefacts. Dans cette thèse, nous avons l'intention d'améliorer l'estimation du GLS en développant une solution d'apprentissage profond dédiée qui estimera de manière robuste et efficace le mouvement des tissus à partir de séquences en mode B.

Chapitre 3 : Méthodes d'estimation de mouvement

Comme détaillé précédemment, l'estimation du mouvement du myocarde est un élément important pour caractériser la fonction cardiaque. En échocardiographie, le mouvement est traditionnellement estimé par des méthodes de suivi du motif de speckle fondées sur l'hypothèse que les propriétés locales des tissus doivent être conservées entre deux images consécutives. Nous présenterons dans cette partie les méthodes traditionnelles d'estimation de mouvement, puis les méthodes d'estimation de mouvement par apprentissage profond et pour finir l'application de ces réseaux de neurones convolutionnels en imagerie ultrasonore.

Approches traditionnelles

Il existe deux types d'approches principales en estimation de mouvement : le flux optique et la méthode de correspondance de blocs.

Le flux optique est une des méthodes les plus utilisées en estimation de mouvement en vision assistée par ordinateur. Deux types principaux de flux optiques sont utilisés en imagerie ultrasonore, un fondé sur la conservation de l'intensité des pixels et l'autre sur la conservation de la phase. Ces deux méthodes sont habituellement employées à plusieurs échelles à partir desquelles les déplacements grossiers sont estimés dans un premier temps puis raffinés par itérations successives jusqu'à la résolution initiale des images.

La méthode par correspondance de blocs est également une des méthodes les plus populaires d'estimation de mouvement en imagerie ultrasonore. La région à suivre dans l'image source est divisée en plusieurs blocs. Chaque bloc est recherché dans l'image cible dans une fenêtre centrée autour du pixel initial. Des métriques de similarités sont utilisées pour trouver la meilleure correspondance entre les blocs sources et cibles. Une fois le déplacement de tous les blocs calculé, une étape de filtrage est requise pour lisser le champ de déplacement final attaché à la région d'intérêt.

Approches par apprentissage profond sur des images naturelles

De nombreuses recherches ont été menées ces dernières années en estimation de mouvement par apprentissage profond, ce qui a permis une amélioration importante des performances. Certaines améliorations proviennent d'architectures introduites pour d'autres problèmes de traitement d'images. Par exemple, l'arrivée des réseaux U-Net initialement conçus pour la segmentation ou encore des réseaux récurrents et transformers utilisés en traitement du langage ont engendré des progrès significatifs. Les méthodes d'apprentissage profond se démarquent notamment avec des performances supérieures de près de 40% aux méthodes non supervisées. La Figure 8.1 fournit une vue d'ensemble des réseaux de neurones issus de l'état de l'art et parmi les plus performants.

Sur cette figure, on remarque que les premiers réseaux performants dans ce domaine (FlowNet [29] et FlowNet2 [31]) sont basés sur l'architecture U-Net [27]. Dans un second temps, des réseaux tels que SpyNet [35] ont introduit la notion d'architectures pyramidales permettant la mise en place de stratégie multi-échelles afin d'estimer le mouvement de plus en plus précisement. Ce réseau a l'avantage de posséder moins de paramètres et d'être plus rapide que les réseaux inspirés de U-Net [27], mais

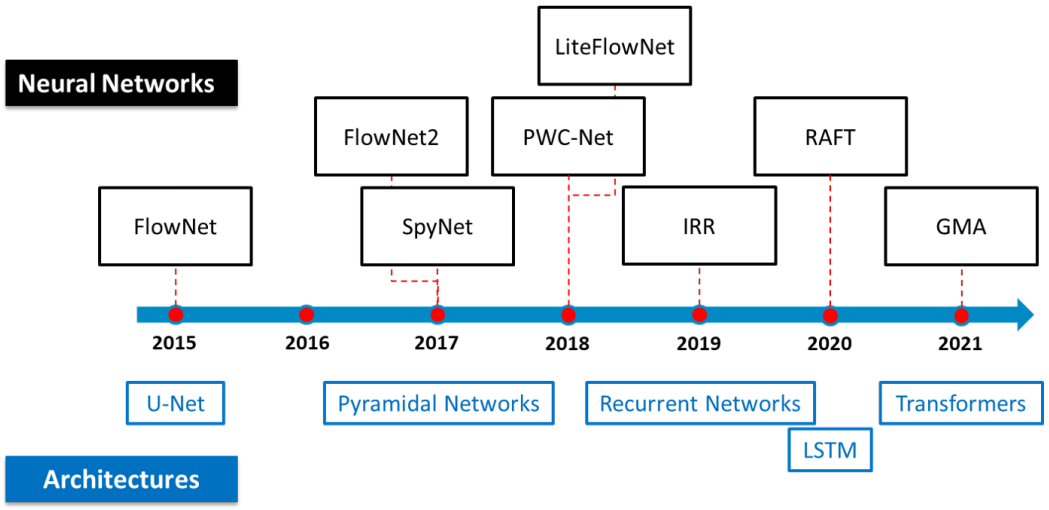


Figure 8.1 – Présentation des principales méthodes d'apprentissage profond en estimation du mouvement à partir d'images naturelles.

avec des performances moindres. Dans la lignée de SpyNet [35], PWC-Net [36] et LiteFlowNet [37] ont été développés simultanément avec l'idée d'implémenter cette structure pyramidale multi-échelle directement dans l'espace des caractéristiques apprises. Ces méthodes combinent les avantages de SpyNet [35], à savoir une plus grande légèreté du réseau ainsi qu'un faible temps d'exécution avec des performances parmi les meilleures de l'état de l'art encore actuellement. Plusieurs avancées en traitement du langage ont eu lieu ces dernières années et les concepts clés ont déja été appliqués en estimation de mouvement. Parmi les méthodes les plus performantes, nous pouvons citer la méthode RAFT [45] basée sur les réseaux récurrents et la méthode GMA[47] basée sur les transformers.

Applications des réseaux de neurones convolutionnels à l'estimation de mouvement en imagerie ultrasonore

De nombeuses méthodes en apprentissage profond ont été développées en imagerie ultrasonore. Dans cette partie nous allons décrire les méthodes supervisées car elles obtiennent de meilleures performances que les méthodes non ou semi-supervisées. Deux applications principales sont concernées : le suivi de tissus sur les images mode B et la déformation des tissus en élastographie. Les méthodes développées en estimation de mouvement en imagerie ultrasonore sont en général inspirées des techniques les plus performantes en imagerie naturelle. En début de thèse, seules deux méthodes avaient été publiées et celles-ci n'apprenaient pas spécifiquement sur les images ultrasonores. Les réseaux siamois [30] et FlowNet2 [31] étaient alors utilisés pour obtenir une première estimation grossière du mouvement avant un post-traitement [49], [50]. Dans un second temps, les méthodes ont été entraînées sur les images ultrasonores et, dans certains cas, leurs architectures ont été adaptées aux ultrasons.

Concernant l'étude de la déformation longitudinale du myocarde, EchoPWC-Net [50], un réseau dérivé de PWC-Net [36], a récemment été publié. Les auteurs ont procédé à des modifications architecturales par rapport au réseau initial telles que l'ajout de la contribution des résolutions les plus fines à la fonction de coût, la suppression des étapes de déformation des cartes de caractéristiques par un champ de déplace-

ment estimé et la propagation des premières cartes de caractéristiques dans le réseau. Cette méthode a été évaluée au travers de deux études impliquant respectivement 30 et 200 patients. En élastographie, PWC-Net [36] a également été utilisée en tant qu'architecture de base pour des méthodes dédiées expoitant soit des signaux RF B [55] soit des images mode B [56].

Conclusions

L'estimation de mouvement est un problème sur lequel de nombreuses recherches ont été effectuées depuis plusieurs années. Dans ce domaine, les méthodes d'apprentissage profond ont surpassé les méthodes traditionnelles utilisant le flux optique ou la correspondance de blocs. Dans ce contexte, mon étude bibliographique m'a permis d'évaluer quelles architectures étaient parmi les plus performantes, notamment en imagerie ultrasonores.

Chapitre 4 : Bases de données 2D en accès libre

L'évaluation ou l'entraînement de réseaux de neurones de manière supervisée requiert l'existence de bases de données avec les références correspondantes. Dans le cas de l'estimation de mouvement, il est impossible de générer manuellement de telles bases de données comme cela peut être fait dans d'autres domaines comme la segmentation. En effet, annoter à la main le déplacement de chaque pixel de l'image entre deux instants serait fastidieux, voire impossible. La solution est de générer des images et des champs de mouvement simulés, pour lesquels le mouvement est défini de manière dense sur toutes les images. Au vu de l'importance des bases de données synthétiques en apprentissage profond, il est important de présenter dans un premier temps les bases de données synthétiques d'images naturelles puis, dans un second temps, les bases de données échocardiographiques.

Bases de données 2D d'images naturelles

De nombreuses bases de données synthétiques dédiées à l'estimation de mouvements ont été publiées ces dernières années. Les premières bases de données ont été créées dans un but d'évaluation uniquement. En effet, ces bases ne possèdent pas suffisamment de données référencées pour pouvoir alimenter des méthodes par apprentissage profond. Parmi ces bases, les principales sont:

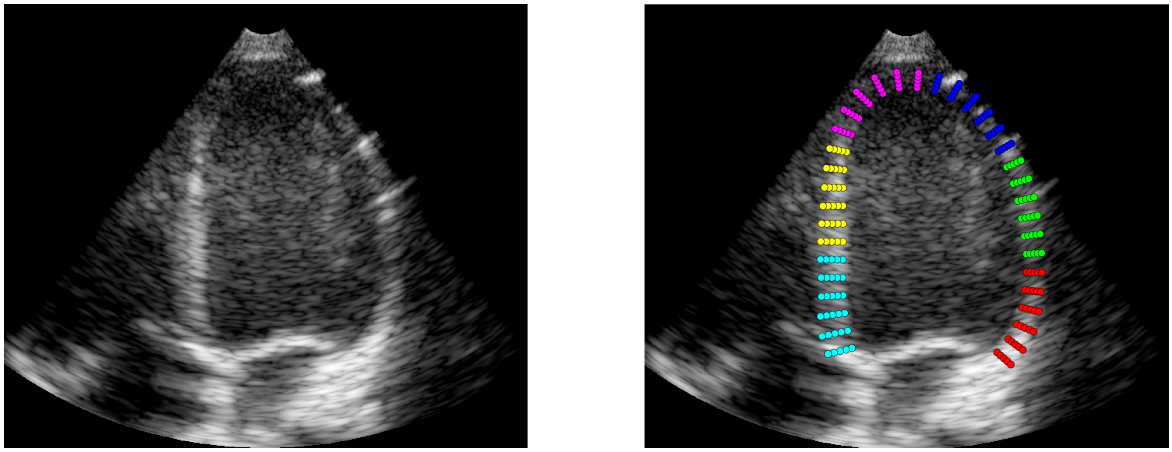
- KITTI, une base de données spécialisée dans les mouvements de voitures [65], [66]. Cette base a été développée dans la mouvance de la voiture autonome.
- MPI Sintel, une base de données synthétiques beaucoup plus riche, avec une grande variété de champs de déplacements, la présence d'occlusions ainsi que des images très bruitées (brouillard, effets atmosphériques, variation d'illumination) [28].

Plus récemment, plusieurs bases de données ont été générées dans le but d'alimenter des méthodes d'apprentissage profond. En particulier, les bases de données FlyingChairs [29], FlyingThings [67] et ChairSDHom [31] sont constituées de plus de 20000 paires d'images de synthèse avec des chaises ou divers objets tournants. Ces trois bases de données se démarquent essentiellement par les amplitudes de mouvement qu'elles contiennent. D'autres bases de données ont également été créées afin de répondre à d'autres objectifs, tels que CrowdFlow spécialisée dans l'analyse des mouvements de foules [68] ou CreativeFlow+ dédiée aux objets stylisés [69].

Bases de données 2D échocardiographiques

Au début de ma thèse, il n'existait qu'une seule base de données de séquences d'images échocardiographiques réalistes [82]. Celle-ci compte 105 séquences et inclut :

- Des acquisitions issues de sept constructeurs différents : Esaote, General Electric, Philips, Siemens, Samsung, Toshiba et Hitachi.
- Trois vues apicales par constructeur : deux chambres, trois chambres et quatre chambres.



(a) (b)

Figure 8.2 – Exemple de données simulées réalistes en vue apicale quatre chambres avec (b) et sans (a) maillage superposé. Ces images ont été générées à partir d'une séquence réelle acquise par un échographe ESAOTE.

- Cinq types de mouvements du myocarde : un sain et quatre avec différentes ischémies.

Le pipeline développé pour simuler ces données utilise des acquisitions réelles pour synthétiser une texture échographique réaliste. Dans ce cadre, un seul sujet sain par constructeur a été acquis pour les trois vues différentes. Les cas pathologiques ont été simulés grâce à un modèle électromécanique.

6060 paires d'images ont été ainsi générées ainsi que des maillages de référence du myocarde dont les déplacements sont dérivés du modèle électromécanique. Les maillages sont définis par 180 points dont une illustration est fournie dans la Figure 8.2. Malgré un réalisme apparent des images, cette base de données souffre de plusieurs limitations. Le pipeline repose sur la procédure de personnalisation des données par le modèle électromécanique. Cette étape est fastidieuse et limite actuellement le déploiement de ce pipeline à un nombre restreint de données (moins de sept patients). De plus, l'amplitude maximale des déformations synthétiques du myocarde est limitée. En effet, le GLS correspondant est inférieure à 10% contre 20% dans les cas réels pour les sujets sains.

Conclusions

Nous avons vu que de nombreuses bases de données simulées avec un champ de déplacement de référence ont été proposées dans la littérature pour évaluer les algorithmes d'estimation de mouvement. Grâce à un nombre croissant de cas simulés, certaines de ces bases de données peuvent également être utilisées pour entraîner des algorithmes d'apprentissage profond. A ce titre, FlyingChairs2D et FlyingThings3D figurent parmi les plus intéressantes.

Concernant les données échocardiographiques, il n'existait au début de ma thèse qu'une seule base de données de séquences synthétiques d'images échocardiographiques 2D. Malgré le haut degré de réalisme des images générées, cette base de données souffre des inconvénients suivants : *i)* il y a un manque de diversité des données puisque les cas simulés ont été générés à partir de seulement 7 sujets réels, *ii)* en raison de

l'étape de personnalisation, le pipeline complet est difficile à déployer sur des patients, ce qui restreint le nombre de cas simulés, *iii)* la procédure de personnalisation proprement dite limite également l'amplitude des déformations synthétiques du myocarde qui restent plus faibles que celle des cas réels. Ces lacunes ont révélé le besoin de créer des bases de données synthétiques plus complètes en imagerie échocardiographique avec une plus grande variété de formes et de mouvements du myocarde afin d'alimenter des méthodes d'apprentissage profond.

Chapitre 5 : Evaluation de la capacité des réseaux de neurones convolutionnels à estimer les déplacements en imagerie ultrasonore 2D

Motivations

L'essor de l'apprentissage profond a constitué une percée dans le domaine de l'estimation de mouvement et a amélioré les performances en termes de précision et de vitesse. A l'époque de cette première contribution, les architectures de type FlowNets étaient parmi les meilleurs réseaux pour l'estimation du mouvement sur des images naturelles. Ainsi, dans les quelques articles utilisant des réseaux de neurones pour estimer le mouvement en imagerie ultrasonore, seuls les réseaux basés sur U-Net ont été utilisés. Les études pionnières ont ainsi révèlées l'intérêt d'adapter des solutions d'apprentissage profond pour l'estimation du mouvement en échographie. Cependant, elles se limitaient à l'utilisation directe de réseaux existants sans évaluer la pertinence de l'architecture pour l'application visée. Sur la base de cette analyse bibliographique, nous avons mené une étude pilote pour répondre aux trois questions suivantes

1. Comment différentes architectures de réseaux de neurones convolutionnels se comportent-elles sur un même ensemble de données d'images échocardiographiques en termes de précision de mouvement ?

2. Quelles sont les différences de performances vis à vis des algorithmes standards qui ne sont pas basés sur de l'apprentissage profond ?

3. Quel est le gain apporté par l'apprentissage par transfert, c'est-à-dire en ré-entraînant les poids d'un réseau déjà pré-entraîné sur des séquences d'images naturelles ?

Méthodologie

Dans cette étude, nous avons décidé de nous focaliser sur le cas particulier de l'estimation du mouvement de rotation d'un disque tournant, où la quantité de déplacement est bien contrôlée, à la fois dans des simulations et des expériences in vitro. Cette stratégie nous a permis de concevoir des séquences d'images simulées et réelles avec des déplacements et des intensités d'image similaires. Elle cible également un défi bien connu dans le domaine des ultrasons, car il peut être difficile d'estimer des rotations précises en raison de la plus grande décorrélation de speckle qu'elle induit par rapport à d'autres types de mouvements tels que la translation.

Nous avons ainsi divisé notre étude en deux phases principales. Une première étape dans laquelle nous comparons six réseaux issus des architectures FlowNet [29] et FlowNet2 [31] sur des données simulées que nous avons générées. Cette famille de réseaux de neurones fondée sur l'architecture U-Net [27] comptait parmi les plus performantes de l'état de l'art au moment de notre étude. Les architectures associées sont composées de plusieurs sous-réseaux comme on peut le voir sur la Figure 8.3. Une fois la comparaison réalisée sur la base de données simulées, le réseau ayant obtenu les meilleures performances est sélectionné et comparé à une méthode traditionnelle de correspondance de blocs issue de l'état de l'art et parmi les plus performantes en

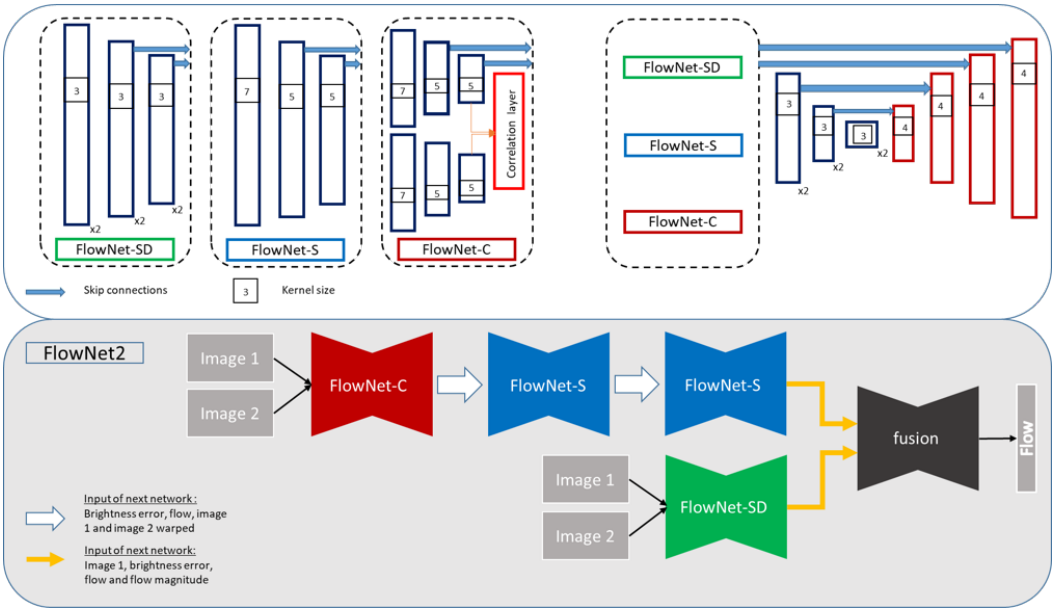


Figure 8.3 – Schéma de l'architecture globale de FlowNet2 [31] (partie basse de la figure) et de l'architecture des sous-réseaux qui le composent (partie haute de la figure). Les couches de convolution sont représentées en bleu et les couches de déconvolution en rouge.

estimation de mouvement en imagerie ultrasonore. Cette méthode est appelée PIV pour Particle Imaging Velocimetry.

Résultats

Dans la première phase de l'étude, à savoir la comparaison des différents réseaux de neurones sur les données simulées, FlowNet-SD avec apprentissage par transfert a obtenu les meilleurs résultats. En effet, de nombreuses architectures se sont révélées être efficaces à de faibles vitesses de déplacements ($<$3rad/s), ce qui s'explique par des amplitudes du déplacement réduites et donc de faibles décorrélations de speckle. Pour cette gamme de vitesses de rotation, des architectures dont les poids n'ont pas été appris sur des données ultrasonores produisent également de bons résultats. Cependant, avec une architecture appropriée (par exemple des noyaux de convolutions de petite taille) et une stratégie d'apprentissage par transfert, on note une adaptabilité de certains réseaux tel que FlowNet-SD à l'ensemble des vitesses de rotation testées.

La deuxième étape de l'étude correspond alors à la comparaison du réseau FlowNet-SD avec deux méthodes de correspondance de blocs issues de l'état de l'art : PIV et PIV-adapt. Nous remarquons que l'algorithme PIV est performant pour l'estimation des vitesses inférieures à 2 rad/s à la fois sur les données simulées et les données in vitro. Cependant, lorsque la vitesse de rotation augmente, PIV ne permet plus d'effectuer une bonne estimation des déplacements. Comme nous pouvons le voir dans le tableau 8.1, l'erreur sur les champs de déplacements (EPE) progresse alors de 0.2 px à plus de 0.9 px sur les données simulées et de 0.8 px à plus de 2.3 px sur les données in vitro. Ces erreurs sur l'EPE augmentent conjointement avec les erreurs sur la vitesse estimée. FlowNetSD-TL (FlowNetSD avec apprentissage par transfert) et PIV-adapt montrent, quant à eux, une grande précision en termes de vitesse estimée aussi bien sur les données simulées que in vitro. Ces résultats démontrent le potentiel de FlowNetSD-TL

Table 8.1 – Valeurs médianes de l'EPE, de la précision de la vitesse angulaire estimée et de la dispersion de la déviation médiane absolue (MAD) calculées à l'intérieur du disque tournant centré sur l'ensemble de données synthétiques (trois premières lignes) et sur l'ensemble de données in vitro (trois dernières lignes) pour cinq vitesses angulaires différentes. Pour cette expérience, FlowNetSD avec apprentissage par transfert (FlowNetSD-TL) a été comparé aux deux versions de la méthode de correspondance de blocs PIV et PIV-adapt.

	Methods	**1 rad/s**		**2 rad/s**		**3 rad/s**		**4 rad/s**		**5 rad/s**	
		EPE	*Velocity*	*EPE*	*Velocity*	*EPE*	*Velocity*	*EPE*	*Velocity*	*EPE*	*Velocity*
		px.	rad/s.	px.	rad/s.	px.	rad/s.	px.	rad/s.	px.	rad/s.
In-silico	FlowNetSD-TL	0.1 ±0.0	1.1 ± 0.0	0.4 ± 0.1	2.2 ± 0.0	0.4 ± 0.1	3.2 ± 0.1	0.3 ±0.1	4.2 ±0.1	0.4 ±0.1	4.9 ±0.1
	PIV	0.2 ±0.1	1.0 ± 0.1	0.2 ± 0.1	2.0 ± 0.1	0.9 ± 0.6	2.5 ± 0.6	3.7 ± 2.1	1.4 ± 1.1	6.4 ± 2.5	0.6 ± 1.0
	PIV-adapt	0.2 ± 0.1	1.0 ± 0.1	0.2 ± 0.1	2.0 ± 0.1	0.2 ± 0.1	2.9 ± 0.2	0.2 ± 0.1	3.9 ± 0.2	0.1 ± 0.1	4.9 ± 0.2
in vitro	FlowNetSD-TL	0.7 ±0.2	1.3 ± 0.1	0.8 ± 0.3	2.3 ± 0.2	1.1 ± 0.4	3.4 ± 0.3	1.0 ±0.4	4.3 ±0.3	0.8 ±0.3	5.0 ±0.4
	PIV	0.4 ± 0.1	1.2 ± 0.1	0.8 ± 0.3	2.3 ± 0.4	2.3 ± 1.5	1.6 ± 1.2	4.8 ± 2.3	0.8 ± 1.1	6.5 ± 2.6	0.5 ± 1.0
	PIV-adapt	0.2 ± 0.1	1.2 ± 0.1	0.3 ± 0.1	2.4 ± 0.2	0.3 ± 0.1	3.6 ± 0.3	0.2 ± 0.1	4.5 ± 0.5	0.2 ± 0.1	5.6 ± 0.6

et des réseaux de neurones en général, à estimer de manière robuste les mouvements en imagerie ultrasonore sans adaptation spécifique aux amplitudes des mouvements à estimer. En effet, dans notre expérience, FlowNetSD et PIV-adapt produisent des résultats sensiblement identiques. Cependant, FlowNetSD utilise uniquement en entrée une paire d'images à une fréquence fixe de 52 Hz tandis que PIV-adapt utilise jusqu'à 16 images en entrée à des fréquences comprises entre 52 Hz et 312 Hz.

Conclusions

Dans cette étude, nous avons évalué différentes architectures de réseaux de neurones convolutionnels sur des données simulées et in vitro pour le suivi de rotations entre 1 et 5 rad/s à partir de paires d'images échographiques. Différents réseaux, tous dérivés de l'architecture FlowNet2, ont été comparés à la méthode PIV, un des meilleurs algorithmes de mise en correspondance de blocs adapté aux ultrasons. Notre évaluation quantitative sur des images simulées et in vitro a révélé que le réseau FlowNetSD, après avoir adapté ses poids aux ultrasons en utilisant l'apprentissage par transfert sur des données simulées, a produit une estimation précise du mouvement sur des données in vitro pour toute la gamme des vitesses angulaires et à une fréquence d'images unique de 52 Hz. Il est intéressant de noter que FlowNetSD a obtenu des estimations de la vitesse angulaire comparables à celles de la méthode PIV-adapt lorsque cette dernière a nécessité une adaptation de la fréquence d'acquisition jusqu'à 312 Hz. Cette étude pilote révèle donc que les solutions d'apprentissage profond représentent une alternative potentiellement puissante aux algorithmes de suivi standard qui peut s'avérer à la fois robuste et précise pour récupérer les champs de déplacement à partir d'images échographiques, y compris pour de grands déplacements et rotations, malgré la décorrélation du speckle. Dans un souci d'ouverture scientifique, les bases de données

utilisées lors de ces expériences ont été mises à disposition pour être téléchargées [1]. Sur la base des résultats de cette étude, nous avons examiné, dans le chapitre suivant, la capacité des méthodes par apprentissage profond à surpasser les méthodes de pointe pour l'estimation de la déformation du myocarde en échocardiographie.

[1] http://humanheart-project.creatis.insa-lyon.fr/revolus.html

Chapitre 6 : Application des réseaux de neurones convolutionnels à l'estimation des mouvements du myocarde en imagerie ultrasonore 2D

Motivations

Après avoir démontré la capacité des réseaux de neurones à estimer le mouvement en imagerie ultrasonore, nous avons décidé de les appliquer à l'analyse de la déformation du myocarde. Le but de cette étude est d'améliorer la robustesse et la précision de la mesure de l'indice clinique de déformation longitudinale globale du muscle myocardique. Pour ce faire, nous avons décidé de partir de l'architecture PWC-Net [36] qui produit parmi les meilleurs résultats actuels en estimation de mouvements sur les images naturelles. Au moment de nos travaux, un réseau nommé EchoPWC-Net [51] également basé sur l'architecture PWC-Net a été publié pour l'estimation du mouvement du myocarde en imagerie échocardiographique. A l'inverse de notre démarche, ce réseau n'utilise pas de stratégie d'apprentissage par transfert mais est entraîné directement à partir de données synthétiques d'images échocardiographiques [82].

Méthodologie

Dans notre étude, nous avons décidé d'adapter l'architecture PWC-Net aux images ultrasonores dans le but d'obtenir les meilleurs résultats possibles à la fois sur des données simulées et sur des données cliniques. Pour ce faire, nous avons modifié la structure originale de PWC-Net en augmentant la capacité du réseau et en ajoutant des sous-réseaux de contexte à chaque niveau de résolution pour un meilleur suivi des motifs de speckle. Un schéma de l'architecture modifiée de PWC-Net que nous avons proposé est fourni dans la Figure 8.4. Du point de vue de l'entraînement du réseau, nous avons choisi d'utiliser un apprentissage par transfert au vu des résultats obtenus dans le chapitre précédent. L'entraînement se fait alors en deux étapes : *i)* un entraînement sur une base de données d'images naturelles composée d'images de FlyingChairs2D [29] et FlyingThings3D [67] en nuance de gris puis *ii)* un apprentissage par transfert sur des séquences d'images échocardiographiques simulées. Nous avons également proposé des schémas temporels d'augmentation de données à la fois en apprentissage et en inférence.

Comme nous l'avons évoqué dans le chapitre 8, une seule base de données synthétiques d'images échocardiographiques était disponible au moment de notre étude. De plus, cette base de données est mal adaptée à l'apprentissage par transfert du fait d'un nombre réduit de patients simulés et d'une faible variabilité des déformations simulées du muscle myocardique. Nous avons ainsi développé un pipeline de simulation permettant de générer un plus grand nombre de cas avec une plus grande diversité de mouvement du myocarde. Pour ce faire, notre pipeline prend en entrée des séquences échocardiographiques réelles et génère le jumeau numérique avec un champ de déplacement de référence du muscle myocardique calculé à partir d'annotations manuelles. Le schéma de notre pipeline est fourni dans la Figure 8.5. Notre pipeline inclus un simulateur physique d'ultrasons, permettant de générer des motifs de speckle qui se décorrélent au cours du mouvement. Nous pouvons également intégrer des artéfacts de réverberation au sein des données simulées. Nous avons simulé au total 8866

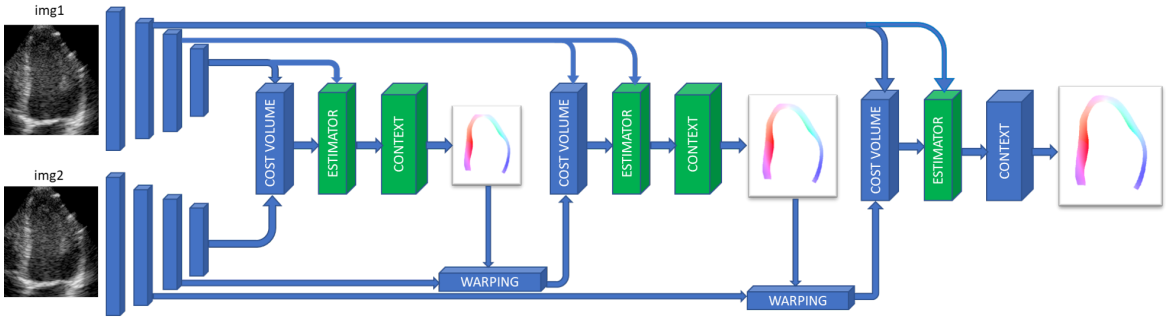


Figure 8.4 – Schéma de la version modifée de PWC-Net que nous avons développé, représentée ici à 4 niveaux. Les deux images en entrée sont d'abord traitées séparément pour extraire les cartes de caractéristiques, puis les champs de déplacements sont estimés de manière grossière et affinés itérativement. Les sous-réseaux que nous avons modifiés sont affichés en vert.

paires d'images échocardiographiques acquises en vue apical 4 chambres.

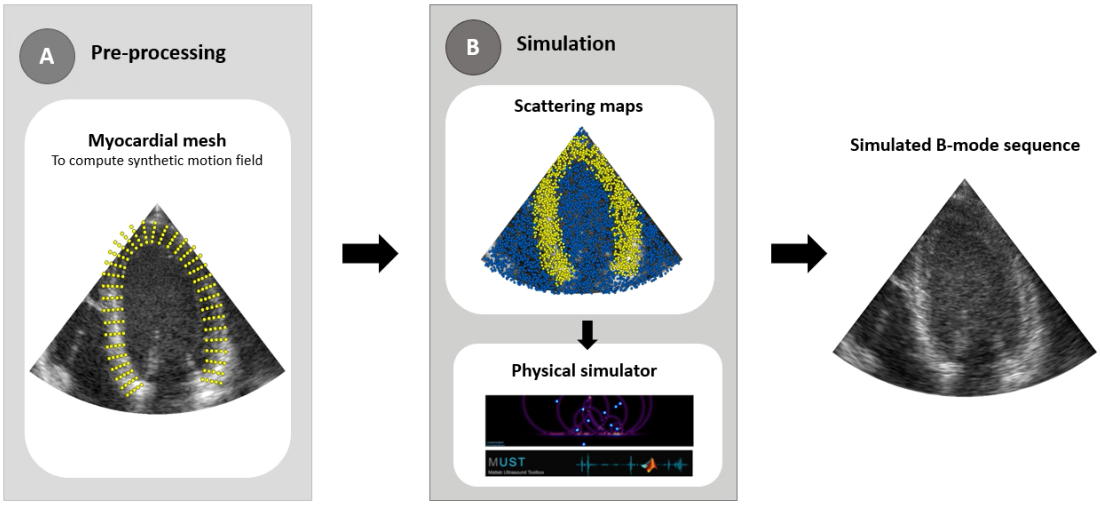


Figure 8.5 – Pipeline proposé pour la simulation de séquences en mode B. (A) Un enregistrement clinique sert de modèle pour la texture de speckle, la définition de l'anatomie et l'estimation du mouvement du myocarde ; (B) Un environnement de simulation échographique fusionnant les informations de la séquence d'images modèle et les maillages du myocarde contribuent à la formation de l'image. Dans la séquence simulée, le mouvement du myocarde est entièrement contrôlé par les variations des maillages tandis que l'aspect visuel est similaire à celui d'une acquisition réelle.

Nous avons évalué notre méthode à partir de deux apprentissges par transfert différents : *i)* soit en utilisant uniquement la base de données synthétiques en libre accès proposée dans [82]; *ii)* soit en combinant cette même base de données avec notre propre base de données synthétiques avec et sans artéfact de réverbération. Nous avons également comparé les performances de notre méthode avec la méthode de mise en correspondance de blocs PIV, la méthode de flux optique Farnebäck et le réseau EchoPWC-Net. Enfin, nous avons évalué la capacité de généralisation de notre algorithme au travers d'une étude multi-centrique, multi-vendeur et multi-pathologie.

Résultats

Après avoir démontré au travers de différentes études d'ablation l'intérêt de nos améliorations en termes de modifications architecturales du réseau, d'augmentation temporelle des données et de composition des champs de déplacements en inférence, nous avons évalué les performances de notre méthode obtenues sur les données synthétiques de référence [82]. A partir du tableau 8.2, nous pouvons voir que notre méthode (désignée sous le nom de c-PWC-Net-gray-usft) obtient avec EchoPWC-Net les meilleurs résultats en termes d'EPE devant les méthodes de l'état de l'art Farnebäck et PIV.

Fort de ces premiers résutlats, nous avons évalué notre méthode à la fois sur nos données simulées et des données réelles acquises via un imageur GE (30 patients acquis au CHU de St Etienne) et Philips (30 patients acquis au CHU de Caen). A partir des données simulées, nous avons observé que notre méthode obtenait de nouveau les meilleurs scores géométriques et cliniques. Nous avons également montré que l'ajout progressif de données synthétiques issues de notre pipeline de simulation lors de l'entrainement améliore constamment les résultats obtenus. En ce qui concerne les données réelles, notre méthode a surpassé les autres méthodes avec une amélioration de plus de 28% des erreurs sur les métriques géométriques et 65% sur les métriques cliniques.

Table 8.2 – Résultats obtenus sur l'ensemble des données synthétiques en libre accès [82]. Les méthodes sont comparées sur des données synthétiques simulées à partir de sept constructeurs différents en vue apicale quatre chambres. La métrique utilisée est l'EPE moyen exprimé en mm.

Methods	*ESAOTE*	*GE*	*HITACHI*	*PHILIPS*	*SIEMENS*	*TOSHIBA*	*SAMSUNG*
	mm.	mm.	mm.	mm.	mm.	mm.	mm.
Farnebäck	0.08 (0.06)	0.09 (0.07)	**0.06 (0.04)**	0.08 (0.06)	0.06 (0.05)	0.07 (0.05)	0.07 (0.05)
PIV	0.10 (0.08)	0.14 (0.11)	0.11 (0.08)	0.09 (0.07)	0.08 (0.07)	0.09 (0.07)	0.13 (0.10)
PWC-Net-gray-usft [51]	0.14 (0.10)	0.17 (0.12)	0.13 (0.09)	0.14 (0.10)	0.14 (0.10)	0.14 (0.11)	0.13 (0.09)
EchoPWC-Net-us [51]	**0.07 (0.06)**	**0.07 (0.06)**	**0.06 (0.04)**	**0.06 (0.05)**	0.06 (0.05)	**0.06 (0.04)**	**0.05 (0.04)**
PWC-Net-gray-usft (ours)	0.08 (0.07)	0.10 (0.07)	0.07 (0.04)	0.09 (0.06)	0.08 (0.06)	0.07 (0.05)	0.08 (0.06)
c-PWC-Net-gray	0.15 (0.13)	0.34 (0.72)	0.11 (0.08)	0.12 (0.10)	0.09 (0.08)	0.12 (0.10)	0.13 (0.08)
c-PWC-Net-gray-usft	**0.07 (0.06)**	0.08 (0.07)	**0.06 (0.04)**	0.07 (0.06)	**0.04 (0.03)**	**0.06 (0.04)**	0.07 (0.05)

Enfin, nous avons démontré la capacité de généralisation de notre réseau à d'autres données acquises sur d'autres matériels dans d'autres centres. En effet, nous avons collecté la base de données acquise sur des échographes Philips composée de 30 patients divisés en cinq groupes de pathologies différentes. Bien que notre méthode n'ait jamais appris sur des données réelles, elle a néanmoins démontré sa très bonne faculté d'adaptation aux données Philips sans réentraînement préalable. Sur cette base de données, les valeurs de métriques géométriques et cliniques obtenues par notre méthode sont proches de celles calculées sur les données CAMUS.

Conclusions

Dans ce chapitre, nous avons développé une méthode d'apprentissage profond pour l'estimation du mouvement en échocardiographie. Nous avons montré que la combinaison d'une version personnalisée de PWC-Net avec un nouvel ensemble de données

synthétiques simulées et une stratégie d'augmentation des données dédiée surpasse les méthodes actuelles de l'état de l'art, tant pour le suivi de la paroi endocardique que pour l'estimation des indices GLS et MAPSE. La généricité de notre approche a également été démontrée au travers de la première étude multi-centrique, multi-constructeur et multi-pathologique. L'ensemble de données synthétiques proposé consiste en des séquences d'images échocardiographiques 2D en vue apicale quatre chambres pour 100 patients virtuels avec ou sans artefacts de réverbération et avec les champs de déplacement du myocarde correspondants. Cette base de données est disponible en accès libre [2].

[2] `http://humanheart-project.creatis.insa-lyon.fr/medicaid.html`

Chapitre 7 : Conclusions et perspectives

Conclusions

L'échocardiographie est une modalité clinique très populaire pour diagnostiquer les pathologies cardiaques. L'analyse de la fonction cardiaque est réalisée par la mesure d'indices cliniques pertinents tels que les volumes ventriculaires et la déformation du myocarde. Dans ce contexte, la déformation longitudinale globale est l'un des indices recommandés dans les directives de pratique clinique pour évaluer la fonction cardiaque. Les techniques utilisées pour estimer la déformation cardiaque à partir des images mode B souffrent actuellement d'un manque de précision et de reproductibilité. Dans ce contexte, nous avons proposé les contributions suivantes :

1. Nous avons mené la première étude pilote visant à évaluer la capacité des réseaux de neurones à estimer le mouvement dans les images échographiques. En particulier, nous avons démontré que les réseaux de neurones convolutifs peuvent être entraînés efficacement pour la tâche de suivi du speckle en échographie. À cette fin, nous avons étudié la capacité de ces méthodes à estimer le mouvement de rotation sur des images simulées et *in vitro*. Ces mouvements sont parmi les plus difficiles à estimer en échographie en raison de la forte décorrélation du speckle associé. Afin d'entraîner et d'évaluer nos différentes méthodes, nous avons créé et partagé avec la communauté une base de données en libre accès composée de séquences simulées et réelles comprenant des disques en rotation avec les vérités terrains correspondantes. A partir des résultats obtenus sur ces bases de données, nous avons montré l'intérêt d'utiliser l'apprentissage par transfert sur des données ultrasonores et avons pu déterminer empiriquement des caractéristiques architecturales améliorant les performances des réseaux de neurones en imagerie ultrasonore. Enfin, cette étude pilote a démontré que les méthodes d'apprentissage profond sont plus performantes que les méthodes traditionnelles de suivi du speckle. Elles peuvent produire des résultats robustes et précis malgré la décorrélation du speckle et les grands déplacements.

2. Nous avons conçu un pipeline dédié à la génération de séquences mode B réalistes pour surmonter la difficulté d'établir des bases de données avec des champs de mouvement de référence. L'intérêt de ce pipeline réside dans sa capacité à simuler des séquences mode B avec une grande variété de déformations cardiaques suffisamment réalistes pour servir d'augmentation de données pertinente pour les méthodes d'apprentissage profond. Nous avons donc généré un ensemble de données synthétiques en libre accès qui consiste en des séquences 2D de vue apicale quatre chambres pour 100 patients virtuels avec ou sans artefacts de réverbération et avec les champs de déplacement du myocarde correspondants.

3. Nous avons développé une solution d'apprentissage profond dédiée à l'estimation de mouvement en échocardiographie basée sur l'architecture PWC-Net. En particulier, nous avons prouvé la pertinence de déployer une stratégie d'apprentissage par transfert et d'utiliser l'augmentation des données temporelles pendant la phase d'apprentissage. Avec notre jeu de données synthétique, nous avons pu estimer la déformation du myocarde avec moins de 3% d'erreur, ce qui est bien meilleur que les performances de l'état de l'art. La généricité de notre réseau

a finalement été démontrée sur une étude multi-centrique, multi-constructeur et multi-pathologie.

Perspectives

Ma thèse a fourni une preuve de concept que la déformation longitudinale globale pouvait être évaluée avec succès par des méthodes d'apprentissage profond dédiées. Pourtant, plusieurs améliorations méthodologiques peuvent être apportées afin de consolider cette étude.

Renforcement des aspects méthodologiques

Bien que notre méthode d'apprentissage profond ait battu les méthodes de pointe et obtenu moins de 3% d'erreur sur l'estimation du GLS, plusieurs possibilités doivent être envisagées afin de consolider la robustesse et la précision des mesures.

- Intégration de concepts clés du domaine du traitement du langage naturel : les réseaux récurrents ainsi que les transformateurs ont fait une entrée remarquée dans le domaine de l'estimation du mouvement. En effet, ces structures ont été intégrées avec succès dans des réseaux dédiés à l'estimation de mouvement, conduisant aux méthodes actuelles les plus performantes (RAFT et GMA) sur des jeux de données d'images naturelles en accès libre. On peut donc s'attendre à ce que l'implémentation effective de ces structures dans des réseaux dédiés à l'imagerie échocardiographique améliore encore les performances de ces algorithmes.

- Implémentation d'une stratégie multi-tâches : notre méthode peut être apprise en conjonction avec d'autres tâches que l'estimation pure du mouvement, comme la segmentation, afin d'augmenter ses performances. En effet, l'optimisation conjointe de plusieurs tâches peut conduire à un renforcement de la cohérence temporelle des segmentations obtenues ainsi qu'à une meilleure précision dans le suivi des contours du myocarde. De plus, cela rendrait notre méthode autosuffisante puisque la version actuelle nécessite une segmentation du ventricule gauche sur l'image initiale afin de calculer le contour endocardique sur toute la séquence par des estimations successives du déplacement.

Renforcement du pipeline de simulation

Une contribution majeure de ma thèse concerne la conception d'un pipeline basé uniquement sur l'image, évitant le besoin d'un modèle électromécanique. Cette stratégie a permis de simuler de nombreux cas à partir de séquences réelles d'images en mode B sur lesquelles les contours du myocarde ont été annotés manuellement afin de générer des champs de mouvement synthétiques d'une grande diversité. Même si la base de données synthétiques générée nous a permis d'améliorer significativement nos résultats, notre pipeline de simulation pourrait bénéficier des contributions suivantes :

- Automatisation de l'ensemble du pipeline : l'étape d'annotation manuelle impliquée dans la version actuelle du pipeline de simulation est le principal goulot d'étranglement pour déployer automatiquement notre solution sur plus de 100 patients quel que soit le type de vue lors de l'acquisition. Il serait intéressant

dans un futur proche de remplacer cette étape par une procédure entièrement automatisée avec un contrôle qualité utilisant un formalisme d'apprentissage profond. Cela nous permettra de déployer notre pipeline sur de grands ensembles de données (>1000) pour toutes les vues apicales (vues deux, trois, quatre et cinq cavités). Compte tenu de l'amélioration conséquente des performances de notre méthode avec l'ajout de seulement 60 patients simulés lors de l'entraînement, l'étude d'un passage à l'échelle de l'entraînement serait intéressante à réaliser afin de déterminer dans quelle mesure nous pouvons améliorer nos résultats grâce à la richesse apportée par les données simulées.

- Amélioration de la diversité des propriétés des images simulées : l'un des principaux avantages de notre pipeline de simulation est d'être suffisamment générique pour simuler une grande variété de propriétés physiques. Il serait donc intéressant d'exploiter cette caractéristique pour renforcer la diversité des images simulées, par exemple dans une perspective d'augmentation des données. Les principales propriétés visées devraient être la variation de la quantité de décorrélation du speckle au cours du temps et l'insertion d'artéfacts autres que la réverbération.

Autres mesures de déformation cardiaque

Dans cette thèse, l'intérêt d'utiliser des méthodes d'apprentissage profond pour l'estimation du GLS a été démontré. En effet, nous avons pu atteindre moins de 3% d'erreur dans son estimation sur des données réelles provenant de différents fabricants, surpassant toutes les autres méthodes de l'état de l'art que nous avons évaluées. Ces résultats doivent encourager la communauté à poursuivre dans cette voie pour renforcer la mesure des indices cliniques liés à la déformation du myocarde.

- Estimation du GLS sur d'autres vues apicales : le GLS estimé sur des données cliniques a montré un faible biais par rapport aux références générées par des cardiologues experts. Dans ce contexte, nous avons fait un effort important de validation sur des images en mode B contenant différentes pathologies acquises avec des échographes de différents fabricants. Cependant, l'évaluation ne comprend que les données acquises en vue apicale quatre cavités. Afin de démontrer davantage la robustesse de notre méthode, il serait intéressant de l'évaluer sur différentes vues apicales afin de s'assurer de sa capacité à obtenir la même qualité d'estimation du GLS quel que soit le type de vues.

- Estimation de la déformation longitudinale régionale : Dans cette thèse, nous nous sommes principalement intéressés à l'étude de la déformation globale du myocarde. Cette mesure est actuellement largement utilisée en pratique clinique bien que l'analyse régionale de la déformation myocardique ait démontré son fort potentiel complémentaire dans l'analyse de la fonction cardiaque. Actuellement, le manque de reproductibilité de cette mesure régionale l'empêche d'être ajoutée aux directives cliniques. Une des raisons de cet inconvénient est la présence de nombreux artéfacts et d'un faible rapport signal/bruit dans certaines zones du myocarde pendant le mouvement, problèmes auxquels nous avons également été confrontés pendant l'étude de la déformation globale. Il serait donc intéressant, à la lumière des résultats obtenus dans cette thèse, d'étendre notre méthode

à l'étude de la fiabilité de la mesure des déformations régionales. Ceci impliquerait également la génération de bases de données simulées dont la référence contiendrait la déformation de chaque segment myocardique.

Chapter 9

Classification of breast nodules in 2D ultrasound imaging

Contents

In parallel to this thesis, I participated with the Philips team to a challenge organized during the 2020 edition of the Journées Françaises de Radiologie by the French Society of Radiology. We won the "Breast Nodule Classification" challenge and published a corresponding article explaining our method. As this contribution does not concern my main field of research, I have attached its description in the following chapter.

9.1 Introduction

During the Journées Française de Radiologie (JFR), the French Society of Radiology (Société Française de Radiologie) organized many challenges on various topics in 2020. The competition was structured in 3 steps: *i)* first we received a first set of data to start designing solutions, *ii)* three weeks later additional data were received to further develop and evaluate our methods and select the best one and *iii)* finally, on the day of the competition, a test database was sent to us on which we had one hour to process all the data and send back a CSV file with our predictions. We therefore decided to participate in one of the proposed challenges concerning the classification of breast nodules in ultrasound imaging between categories hard to differentiate. The goal was to develop solutions to fight breast cancer and improve the patient's life by avoiding unnecessary biopsies.

Breast cancer accounts for a quarter of all new cancers in women worldwide and is the leading cause of death among women [105]. In many countries, national screening plans are organized and have shown that the risk of death from breast cancer among women invited for screening is reduced by 22% compared to women not invited [106]. In this clinical setting, breast ultrasound is routinely used in combination with mammography to distinguish malignant from benign nodules [107]. In order to standardize the patient monitoring, several standard criteria have been defined to characterize the potential malignancy of breast nodules from ultrasound images. These criteria are based on contours, height to width ratio, shape, echogenicity, and texture of the tumor [108]. Breast Imaging Reporting and Data System (BI-RADS) is a classification based on these criteria, which makes it possible to categorize the nodules and select the best approach [109], [110]. However, some breast nodules remain difficult to classify with imaging, in particular it is often challenging to differentiate BI-RADS 3 nodules (*i.e.*, probably benign nodules) from BI-RADS 4A nodules (*i.e.*, nodules with a low suspicion for malignancy). As a result, breast nodules assigned to BI-RADS 4 category require biopsy for further diagnosis when BI-RADS 3 do not. The development of a deep learning solution to better identify the nature of the breast nodule could lessen the need for breast biopsy in the presence of benign nodules. Regarding ultrasound, computer-aided breast nodules classifiers were proposed based on stacked auto encoder [111], deep polynomial network [112] and convolutional neural networks (CNN). Among CNNs, two major axes emerged. The first in which the network directly classifies and the second in which it performs a semantic segmentation. VGG [113]–[115] and GoogleNet[116]–[118] are two main architectures have been applied to directly label breast nodules using ultrasound data. On the other hand, the semantic segmentation approach simultaneously mixes segmentation and classification, which allows to explicitly extract information from the images. Mask region-based convolutional neural network (R-CNN) [119] is one of the most used architectures for this purpose in med-

ical imaging [120], in particular to distinguish malignant from benign nodules [121], [122]. The purpose of this study was to create a deep learning algorithm in order to infer the benign or malignant nature of breast nodules using two-dimensional (2D) BI-RADS 3 and 4 B-mode ultrasound data.

9.2 Materials and methods

9.2.1 Study population

The 2D B-mode ultrasound images were provided as part of the "Breast Nodule Challenge" organized during the 2020 edition of the Journées Françaises de Radiologie, which is the annual meeting of the French Society of Radiology (Société Française de Radiologie). The general description of the challenge, as well as details on the data provided and evaluation procedure, can be found in the related article [123]. Two training datasets were given, the first one containing 100 B-mode breast tumor images with balanced labels (50 benign and 50 malignant) and, three weeks later, the second one containing 360 images including 115 malignant nodules. These databases were provided with an indication of whether the image was malignant or benign, but without mention of the location of the tumor in the ultrasound scan. The malignant or benign nature of the breast nodules used as ground truth was previously obtained after biopsy and histopathological analysis. Finally, an unlabeled test dataset including 137 breast tumor images has been made available to evaluate the effectiveness and generalization of our model. Challenge participants had one hour to process this test set and submit their results. Breast tumor images, all labeled to BI-RADS 3 or 4. have been collected in several French institutions with various hardware.

9.2.2 Ground truth generation

The data initially received included only ultrasound images and their labels. In order to extract as much information as possible, the nodules were manually annotated using the software VGG Image Annotator [124] with the help of two expert radiologists. For the first dataset, all images were annotated directly by the radiologists, which can be time-consuming depending on the number of data to be processed. To speed up the process on the second dataset, a segmentation algorithm U-Net based [27] trained on the first dataset was used to generate tumor contours, which were then validated or corrected by the radiologists. These steps made it possible to generate accurate binary masks (Figure 9.1), providing explicit information on shape and ultrasound textures of the nodules. Nodules dimensions were not clearly defined because the pixel size information was not provided with the images.

9.3 Model

9.3.1 Pre-processing

All ultrasound images were available in the NifTi format and the corresponding references contained in JSON files. The ultrasound images and the binary masks extracted from them were then resized to 256×256 pixels.

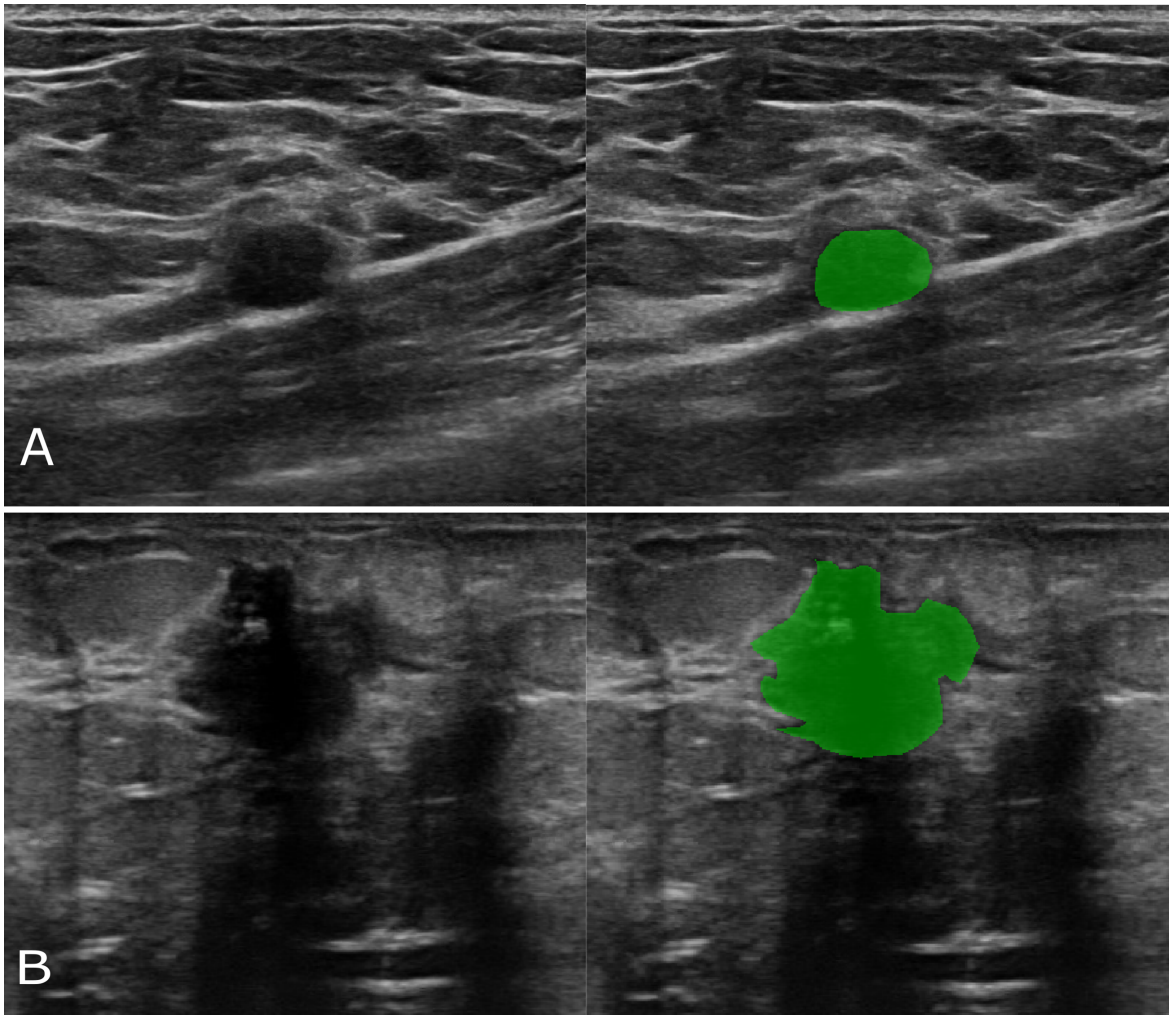


Figure 9.1 – Examples of masks obtained with the help of radiologists in order to train the Mask-RCNN. On the left, we have the B-mode images and on the right these overlaid images with corresponding ground truth binary masks. The regularity of the edges of the benign tumor (A) compared to the malignant tumor (B) can be seen here.

9.3.2 Model architecture

In order to extract as much information as possible from the B-mode images and thus obtain a better final classification of malignant and benign nodules, we opted for the mask region-based convolutional neural network (Mask-RCNN) architecture [119]. This architecture makes optimal use of all the information present in the input ultrasound image, extracting features relevant to both the segmentation mask output, such as tumor shapes and borders, and the classifying label output, such as texture. The network was built for the feature extraction part on a feature pyramid network (FPN) [125] and a ResNet-101 backbone [33]. Then, a region proposal network (RPN) [126] detected the regions of interest in the extracted features and these regions were explicitly aligned with the input features. Thereafter, the inference of classes and bounding boxes was separated from the estimation of the mask. The Mask-RCNN was trained using the ultrasound images of the nodules and the corresponding ground truth, which consists in the segmentation mask of the nodule and the label. As suggested in the related article [119], we used a multitask loss combining bounding box, segmentation and classification losses. During inference, the network takes the image as input and outputs probabilities of belonging to the benign and malignant classes, bounding boxes around the detected nodules and segmentation masks. Figure 9.3 shows several examples of the Mask-RCNN output.

9.3.3 Training

To avoid overfitting and increase the amount and diversity of training set, data augmentation was implemented. The deployed augmentation included random flips, noise addition and typical ultrasound alterations such as variations in brightness, saturation and contrast. For better results and convergence, the network was fine-tuned using weights learned from the Coco dataset [127]. For consistency with the suggested training strategy of Mask-RCNN [119] on the Coco dataset, the average pixel value per channel of the images in the Coco database was also subtracted from our images during the training. The training was done during 50 epochs, using Adam optimizer [34] with a learning rate and a batch size experimentally set at $1e^{-4}$ and 2 respectively.

9.3.4 Ensemble aggregation

To improve and better evaluate the performance of our network, we used a stratified five folds cross-validation strategy. Special attention was given to constructing all folds with a proportion of classes identical to the distribution of the provided training database. For each fold, a network was trained, with a common set of hyperparameters. The five resulting networks were then used for inference in an ensemble method (see below). The network produced a probability of a benign or malignant state of the tumor. If no tumor was detected in the input image, a probability of 0.5 was assigned. If more than one tumor was noticed, only the highest probability among the benign and malignant classes was kept to determine the final class. In the end, only the probability of malignancy was required, so if only a benign nodule was detected with a probability pb, the probability of malignancy pm was defined by the formula $pm = 1 - pb$. During inference on the final test set, the probabilities obtained by the five networks were averaged and the AUC was computed from the ROC curve (which is the true positive rate as a function of the false positive rate), in order to

Folds Index	AUC (CI)
0	0.72 (0.60-0.84)
1	0.69 (0.56-0.82)
2	0.54 (0.40-0.68)
3	0.77 (0.65-0.89)
4	0.75 (0.63-0.87)

Table 9.1 – Area Under Curve (AUC) and corresponding Confidence Interval (CI) obtained by our Mask-RCNN network on the different folds. We noticed that the results were disparate with AUCs between 0.54 and 0.77. This variability of results revealed the heterogeneity of the characteristics of the images composing the database. It also highlighted the difficulty of generalizing the performance obtained on one fold to the others.

measure the performance of the classification, and its ability to achieve high sensitivity and specificity.

9.4 Results

Before obtaining the final test dataset, the results obtained on the five validation folds were first compared to the ground truths. Our network obtained different results with a mean AUC of 0.69 (0.57 – 0.82), a standard deviation of 0.08 and a maximum deviation of 0.23 depending on the folds on which it was trained and tested (Table 9.1 and Figure 9.2). Figure 9.3 shows an example of the results obtained on the validation data in terms of both segmentation and classification. We then conducted an initial attempt to aggregate the results as described in Section 9.3.4 and observed an average increase of 2 percent for all networks. On the test dataset received for the final evaluation, the AUC obtained by our aggregation method applied to the five-folds networks was 0.67.

9.5 Discussion

This study proposed a deep learning method to classify breast nodules on ultrasound images. This algorithm was based on the aggregation of results from an ensemble of Mask R-CNN neural networks. The final AUC obtained on the test set was 0.67, which is lower than our initial expectations and the results published in the literature. Best results from previous studies were 0.98 [114] and 0.95 AUC [118] using respectively VGG16 architecture and GoogleNet. The Mask-RCNN network reporting the best results in a study has an accuracy of 85% [122]. However, it could be noted that these results were obtained on databases from a single clinical center, with low variability in ultrasound protocols, a known pixel size and with similar proportion of malignant nodules in the test dataset as in the train dataset. The proportion of images belonging to each class between the training and test databases could also affect the results in a classification problem. Another notable difference between our study and these previous studies is that the images in the database belong to different BI-RADS classifications. Indeed, in our case, the objective is to differentiate the benign and

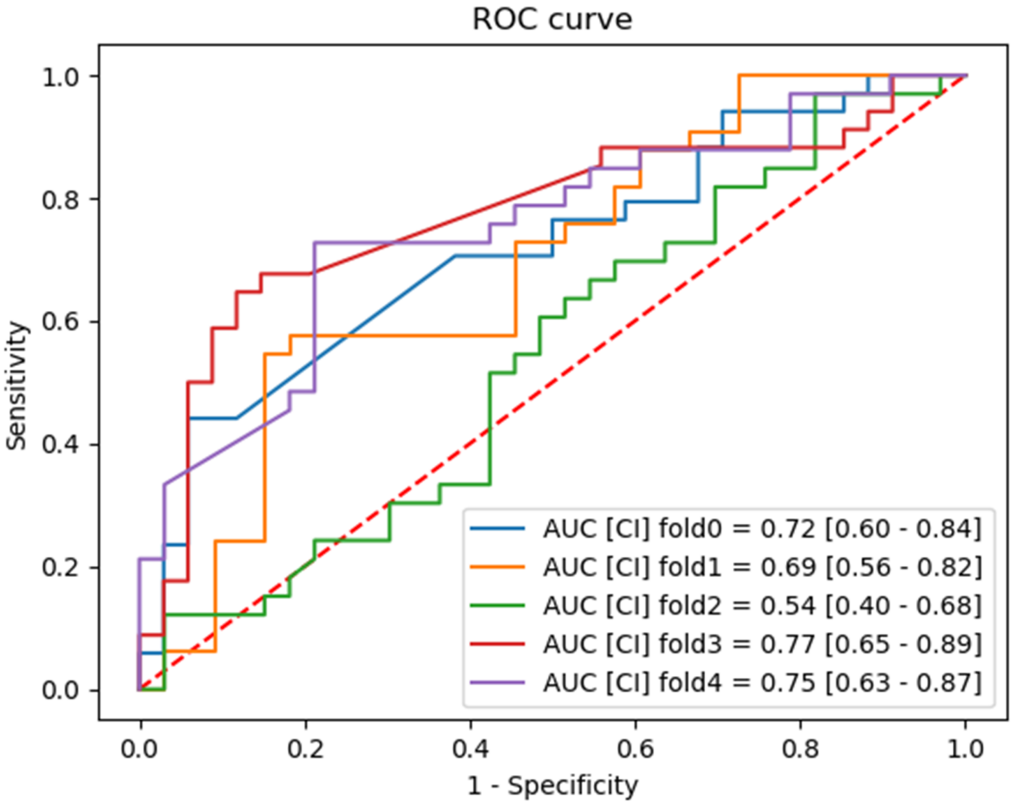


Figure 9.2 – The figure shows ROC curves obtained in cross-validation. We notice the important differences in performance according to the folds and in spite of the use of an identical network with the same hyperparameters. The area under the fold 2 curve has a difference of 0.19 with the average of the other folds. This underlines the high variability of the data.

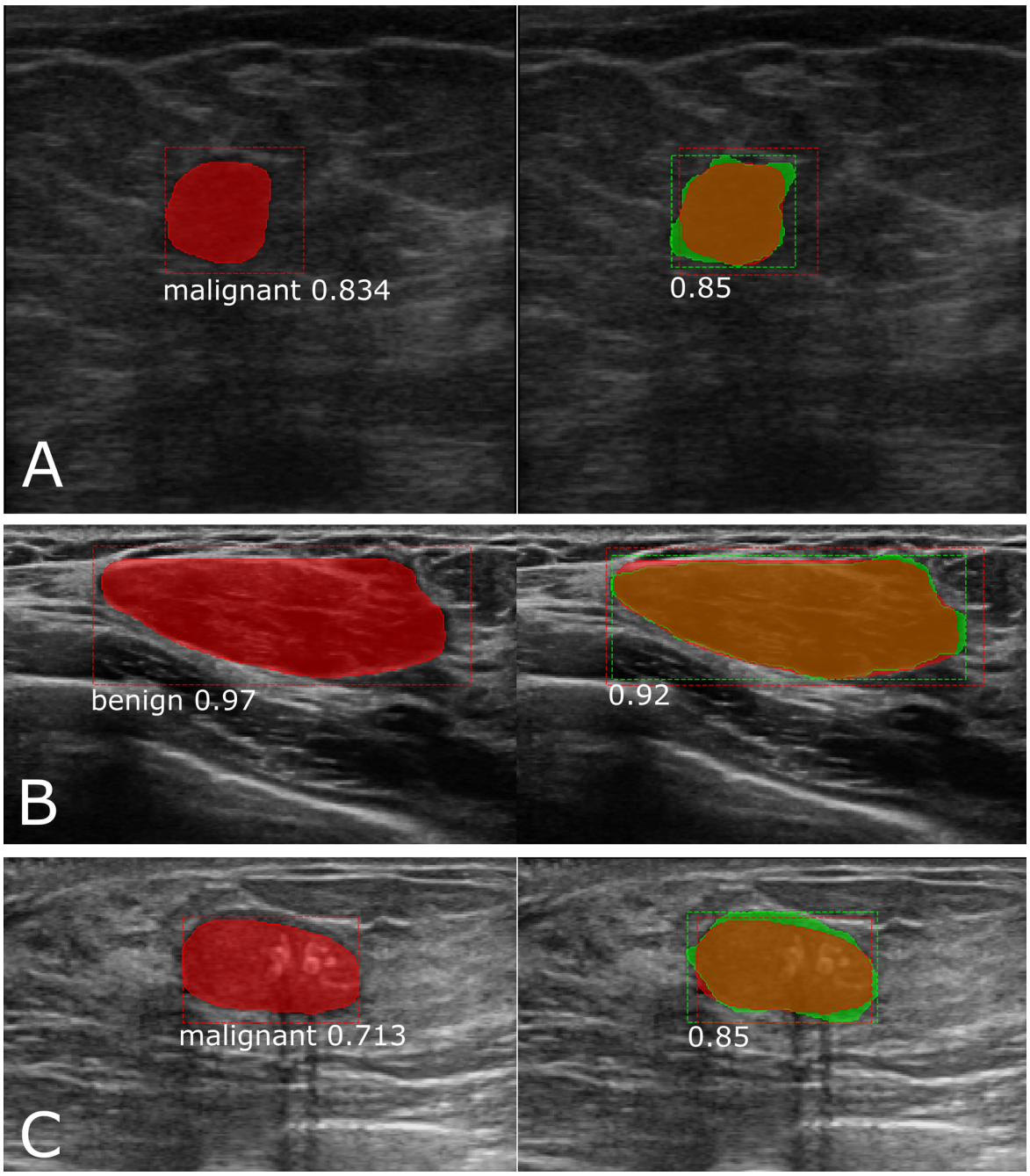


Figure 9.3 – Left part of the figure shows examples of network output predictions with each time the class found 'benign' or 'malignant', the estimated mask and the corresponding bounding box. Right part of the figure shows the overlaid ultrasound image with in green the reference mask and the corresponding bounding box and in red the predicted ones. The legend in white indicates the intersection over union overlap of the two masks. Here are examples of well-classified malignant tumor (A), well-classified benign tumor (B) as well as misclassified benign tumor (C).

malignant character of the borderline cases between BI-RADS 3 and 4 whereas in the cited articles, this classification is performed on all BI-RADS classes, which is an easier task and improves the performance of the classifiers. In order to improve our results, we could have used a training database whose proportions were closer to the test database, which was unknown here. Similarly, the knowledge of the physical size of the pixels according to the images could have been useful to make it possible to characterize the nodules using their size. Given the wide variety of data used (several clinical centers, several ultrasound machines, several sonographers), an increase in the number of training data could also have enhanced the final results. The disparate results of Table 9.1 and Figure 9.2 with AUCs between 0.54 and 0.77 revealed the heterogeneity of the characteristics of the images composing the database. Indeed, despite the identical ratio of benign to malignant nodules in each fold, this highlighted the difficulty of generalizing performance from one-fold to the others. Regarding the machine learning pipeline, there are two main directions for improving results. There is a detection threshold in the Mask-RCNN parameter below which outputs are not taken into account. After a quick study of this parameter, we set this threshold at 0.7 but a more in-depth study could allow a better overall performance. Another possibility would be to focus more precisely on the segmentation part of the Mask-RCNN to deduce the shape characteristics of the nodules.

9.6 Conclusion

In conclusion, the proposed deep learning solution helps with classification of benign and malignant breast nodules based solely on BI-RADS 3 and 4 classified 2D ultrasound images. However, the task remains complex and an increase in the number of images as well as a more in-depth study of the parameters and architecture of the Mask-RCNN is likely to improve the results.

Thanks to the developed method, we won this Challenge JFR 2020 and published an article on this subject in order to explain our approach.

Publications

Published Articles:

- E. Evain, K. Faraz, T. Grenier, D. Garcia, M. De Craene and O. Bernard, "A Pilot Study on Convolutional Neural Networks for Motion Estimation From Ultrasound Images," in IEEE Transactions on Ultrasonics, Ferroelectrics, and Frequency Control, vol. 67, no. 12, pp. 2565-2573, Dec. 2020, DOI:https://doi.org/10.1109/TUFFC.2020.2976809.

- E. Evain, C. Raynaud, C. Ciofolo-Veit, A. Popoff, T. Caramella, P. Kbaier, C. Balleyguier, S. Harguem-Zayani, H. Dapvril, L. Ceugnart, M. Monroc, F. Chamming's, I. Doutriaux-Dumoulin, I. Thomassin-Naggara, A. Haquin, M. Charlot, J. Orabona, T. Fourquet, I. Bousaid, N. Lassau, A. Olivier, "Breast nodule classification with two-dimensional ultrasound using Mask-RCNN ensemble aggregation", in Diagnostic and Interventional Imaging, vol. 102, no. 11, pp. 653-658, ISSN: 2211-5684, Sept. 2021, DOI:https://doi.org/10.1016/j.diii.2021.09.002.

- E. Evain, Y. Sun, K. Faraz, D. Garcia, E. Saloux, B. Gerber, M. De Craene and O. Bernard, "Motion estimation by deep learning in 2D echocardiography: synthetic dataset and validation," in IEEE Transactions on Medical Imaging, 2022, DOI:https://doi.org/10.1109/TMI.2022.3151606

Bibliography

[1] World Health Organization, *Cardiovascular diseases (cvds)*, https://www.who.int/en/news-room/fact-sheets/detail/cardiovascular-diseases-(cvds), 2021.

[2] V. Bittner, "The new 2019 aha/acc guideline on the primary prevention of cardiovascular disease", *Circulation*, vol. 142, no. 25, pp. 2402–2404, 2020.

[3] V. Mor-Avi, R. M. Lang, L. P. Badano, M. Belohlavek, N. M. Cardim, G. Derumeaux, M. Galderisi, T. Marwick, S. F. Nagueh, P. P. Sengupta, *et al.*, "Current and evolving echocardiographic techniques for the quantitative evaluation of cardiac mechanics: Ase/eae consensus statement on methodology and indications endorsed by the japanese society of echocardiography", *European Journal of Echocardiography*, vol. 12, no. 3, pp. 167–205, 2011.

[4] K. E. Farsalinos, A. M. Daraban, S. Ünlü, J. D. Thomas, L. P. Badano, and J.-U. Voigt, "Head-to-head comparison of global longitudinal strain measurements among nine different vendors: The eacvi/ase inter-vendor comparison study", *Journal of the American Society of Echocardiography*, vol. 28, no. 10, pp. 1171–1181, 2015.

[5] J.-U. Voigt, G. Pedrizzetti, P. Lysyansky, T. H. Marwick, H. Houle, R. Baumann, S. Pedri, Y. Ito, Y. Abe, S. Metz, *et al.*, "Definitions for a common standard for 2d speckle tracking echocardiography: Consensus document of the eacvi/ase/industry task force to standardize deformation imaging", *European Heart Journal-Cardiovascular Imaging*, vol. 16, no. 1, pp. 1–11, 2015.

[6] OpenStax College, *Conduction system of the heart — Wikipedia, the free encyclopedia*, https://commons.wikimedia.org/wiki/File:2018_Conduction_System_of_Heart.jpg, 2021.

[7] A. H. A. W. G. on Myocardial Segmentation, R. for Cardiac Imaging: M. D. Cerqueira, N. J. Weissman, V. Dilsizian, A. K. Jacobs, S. Kaul, W. K. Laskey, D. J. Pennell, J. A. Rumberger, T. Ryan, *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", *Circulation*, vol. 105, no. 4, pp. 539–542, 2002.

[8] Wikipedia contributors, *Wiggers diagram — Wikipedia, the free encyclopedia*, https://en.wikipedia.org/w/index.php?title=Wiggers_diagram&oldid=1029541282, 2021.

[9] Radiopaedia, *M-mode — radiopaedia.org*, https://radiopaedia.org/cases/, 2021.

[10] Philips Healthcare, *B-mode & doppler images — philips healthcare*, https://www.usa.philips.com/healthcare/, 2021.

[11] E. Abergel, A. Cohen, P. Gueret, and R. Roudaut, "Échocardiographie clinique de l'adulte", *Ed Estem DeBoeck*, pp. 1173–1175, 2003.

[12] C. Mitchell, P. S. Rahko, L. A. Blauwet, B. Canaday, J. A. Finstuen, M. C. Foster, K. Horton, K. O. Ogunyankin, R. A. Palma, and E. J. Velazquez, "Guidelines for performing a comprehensive transthoracic echocardiographic examination in adults: Recommendations from the american society of echocardiography", *Journal of the American Society of Echocardiography*, vol. 32, no. 1, pp. 1–64, 2019.

[13] R. M. Lang, M. Bierig, R. B. Devereux, F. A. Flachskampf, E. Foster, P. A. Pellikka, M. H. Picard, M. J. Roman, J. Seward, J. S. Shanewise, *et al.*, "Recommendations for chamber quantification: A report from the american society of echocardiography's guidelines and standards committee and the chamber quantification writing group, developed in conjunction with the european association of echocardiography, a branch of the european society of cardiology", *Journal of the American Society of Echocardiography*, vol. 18, no. 12, pp. 1440–1463, 2005.

[14] B. E. Bulwer, S. D. Solomon, and R. Janardhanan, "Echocardiographic assessment of ventricular systolic function", in *Essential Echocardiography*, Springer, 2007, pp. 89–117.

[15] P. Ponikowski, A. A. Voors, S. D. Anker, H. Bueno, J. G. F. Cleland, A. J. S. Coats, V. Falk, J. R. González-Juanatey, V.-P. Harjola, E. A. Jankowska, M. Jessup, C. Linde, P. Nihoyannopoulos, J. T. Parissis, B. Pieske, J. P. Riley, G. M. C. Rosano, L. M. Ruilope, F. Ruschitzka, F. H. Rutten, P. van der Meer, and E. S. D. Group, "2016 ESC Guidelines for the diagnosis and treatment of acute and chronic heart failure: The Task Force for the diagnosis and treatment of acute and chronic heart failure of the European Society of Cardiology (ESC)Developed with the special contribution of the Heart Failure Association (HFA) of the ESC", *European Heart Journal*, vol. 37, no. 27, pp. 2129–2200, May 2016, ISSN: 0195-668X. DOI: 10.1093/eurheartj/ehw128. eprint: https://academic.oup.com/eurheartj/article-pdf/37/27/2129/23748755/ehw128.pdf. [Online]. Available: https://doi.org/10.1093/eurheartj/ehw128.

[16] K. Takigiku, M. Takeuchi, C. Izumi, S. Yuda, K. Sakata, N. Ohte, K. Tanabe, S. Nakatani, J. investigators, *et al.*, "Normal range of left ventricular 2-dimensional strain", *Circulation Journal*, vol. 76, no. 11, pp. 2623–2632, 2012.

[17] J.-U. Voigt, G. Pedrizzetti, P. Lysyansky, T. H. Marwick, H. Houle, R. Baumann, S. Pedri, Y. Ito, Y. Abe, S. Metz, *et al.*, "Definitions for a common standard for 2d speckle tracking echocardiography: Consensus document of the eacvi/ase/industry task force to standardize deformation imaging", *European Heart Journal-Cardiovascular Imaging*, vol. 16, no. 1, pp. 1–11, 2015.

[18] K. Shiino, A. Yamada, M. Ischenko, B. K. Khandheria, M. Hudaverdi, V. Speranza, M. Harten, A. Benjamin, C. R. Hamilton-Craig, D. G. Platts, *et al.*, "Intervendor consistency and reproducibility of left ventricular 2d global and regional strain with two different high-end ultrasound systems", *European Heart Journal-Cardiovascular Imaging*, vol. 18, no. 6, pp. 707–716, 2017.

[19] J. E. Sanderson and A. G. Fraser, "Systolic dysfunction in heart failure with a normal ejection fraction: Echo-doppler measurements", *Progress in Cardiovascular Diseases*, vol. 49, no. 3, pp. 196–206, 2006.

[20] K. Hu, D. Liu, S. Herrmann, M. Niemann, P. D. Gaudron, W. Voelker, G. Ertl, B. Bijnens, and F. Weidemann, "Clinical implication of mitral annular plane systolic excursion for patients with cardiovascular disease", *European Heart Journal–Cardiovascular Imaging*, vol. 14, no. 3, pp. 205–212, 2013.

[21] B. K. Horn and B. G. Schunck, "Determining optical flow", *Artificial intelligence*, vol. 17, no. 1-3, pp. 185–203, 1981.

[22] G. Farnebäck, "Two-frame motion estimation based on polynomial expansion", in *Scandinavian conference on Image analysis*, Springer, 2003, pp. 363–370.

[23] D. J. Fleet and A. D. Jepson, "Computation of component image velocity from local phase information", *International journal of computer vision*, vol. 5, no. 1, pp. 77–104, 1990.

[24] M. Alessandrini, A. Basarab, H. Liebgott, and O. Bernard, "Myocardial motion estimation from medical images using the monogenic signal", *IEEE transactions on image processing*, vol. 22, no. 3, pp. 1084–1095, 2012.

[25] K. Pauwels and M. M. Van Hulle, "Realtime phase-based optical flow on the gpu", in *2008 IEEE Computer Society Conference on Computer Vision and Pattern Recognition Workshops*, IEEE, 2008, pp. 1–8.

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

[27] O. Ronneberger, P. Fischer, and T. Brox, "U-net: Convolutional networks for biomedical image segmentation", in *International Conference on Medical image computing and computer-assisted intervention*, Springer, 2015, pp. 234–241.

[28] D. J. Butler, J. Wulff, G. B. Stanley, and M. J. Black, "A naturalistic open source movie for optical flow evaluation", in *European Conf. on Computer Vision (ECCV)*, A. Fitzgibbon et al. (Eds.), Ed., ser. Part IV, LNCS 7577, Springer-Verlag, Oct. 2012, pp. 611–625.

[29] A. Dosovitskiy, P. Fischer, E. Ilg, P. Hausser, C. Hazirbas, V. Golkov, P. Van Der Smagt, D. Cremers, and T. Brox, "Flownet: Learning optical flow with convolutional networks", in *Proceedings of the IEEE international conference on computer vision*, 2015, pp. 2758–2766.

[30] S. Chopra, R. Hadsell, and Y. LeCun, "Learning a similarity metric discriminatively, with application to face verification", in *2005 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'05)*, IEEE, vol. 1, 2005, pp. 539–546.

[31] E. Ilg, N. Mayer, T. Saikia, M. Keuper, A. Dosovitskiy, and T. Brox, "Flownet 2.0: Evolution of optical flow estimation with deep networks", in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2017, pp. 2462–2470.

[32] J. Cheng, Y.-H. Tsai, S. Wang, and M.-H. Yang, "Segflow: Joint learning for video object segmentation and optical flow", in *Proceedings of the IEEE international conference on computer vision*, 2017, pp. 686–695.

[33] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition", in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2016, pp. 770–778.

[34] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization", *arXiv preprint arXiv:1412.6980*, 2014.

[35] A. Ranjan and M. J. Black, "Optical flow estimation using a spatial pyramid network", in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2017, pp. 4161–4170.

[36] D. Sun, X. Yang, M.-Y. Liu, and J. Kautz, "Pwc-net: Cnns for optical flow using pyramid, warping, and cost volume", in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2018, pp. 8934–8943.

[37] T.-W. Hui, X. Tang, and C. C. Loy, "Liteflownet: A lightweight convolutional neural network for optical flow estimation", in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2018, pp. 8981–8989.

[38] Y. Lu, J. Valmadre, H. Wang, J. Kannala, M. Harandi, and P. Torr, "Devon: Deformable volume network for learning optical flow", in *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, 2020, pp. 2705–2713.

[39] J. Wang, Y. Zhong, Y. Dai, K. Zhang, P. Ji, and H. Li, "Displacement-invariant matching cost learning for accurate optical flow estimation", *arXiv preprint arXiv:2010.14851*, 2020.

[40] T.-W. Hui and C. C. Loy, "LiteFlowNet3: Resolving Correspondence Ambiguity for More Accurate Optical Flow Estimation", 2020, pp. 169–184.

[41] D. Sun, X. Yang, M.-Y. Liu, and J. Kautz, "Models matter, so does training: An empirical study of cnns for optical flow estimation", *IEEE transactions on pattern analysis and machine intelligence*, vol. 42, no. 6, pp. 1408–1423, 2019.

[42] Z. Yin, T. Darrell, and F. Yu, "Hierarchical discrete distribution decomposition for match density estimation", in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2019, pp. 6044–6053.

[43] G. Yang and D. Ramanan, "Volumetric correspondence networks for optical flow", *Advances in neural information processing systems*, vol. 32, pp. 794–805, 2019.

[44] J. Hur and S. Roth, "Iterative residual refinement for joint optical flow and occlusion estimation", in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2019, pp. 5754–5763.

[45] Z. Teed and J. Deng, "Raft: Recurrent all-pairs field transforms for optical flow", in *European conference on computer vision*, Springer, 2020, pp. 402–419.

[46] H. Sak, A. W. Senior, and F. Beaufays, "Long short-term memory recurrent neural network architectures for large scale acoustic modeling", 2014.

[47] S. Jiang, D. Campbell, Y. Lu, H. Li, and R. Hartley, "Learning to estimate hidden motions with global motion aggregation", *arXiv preprint arXiv:2104.02409*, 2021.

[48] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Ł. Kaiser, and I. Polosukhin, "Attention is all you need", in *Advances in neural information processing systems*, 2017, pp. 5998–6008.

[49] S. Bharadwaj and M. Almekkawy, "Motion estimation for ultrasound image sequences using deep learning", *The Journal of the Acoustical Society of America*, vol. 148, no. 4, pp. 2487–2487, 2020.

[50] A. Østvik, E. Smistad, T. Espeland, E. A. R. Berg, and L. Lovstakken, "Automatic myocardial strain imaging in echocardiography using deep learning", in *Deep Learning in Medical Image Analysis and Multimodal Learning for Clinical Decision Support*, Springer, 2018, pp. 309–316.

[51] A. Østvik, I. M. Salte, E. Smistad, T. M. Nguyen, D. Melichova, H. Brunvand, K. Haugaa, T. Edvardsen, B. Grenne, and L. Lovstakken, "Myocardial function imaging in echocardiography using deep learning", *IEEE Transactions on Medical Imaging*, vol. 40, no. 5, pp. 1340–1351, 2021.

[52] B. Peng, Y. Xian, and J. Jiang, "A convolution neural network-based speckle tracking method for ultrasound elastography", in *2018 IEEE International Ultrasonics Symposium (IUS)*, IEEE, 2018, pp. 206–212.

[53] M. G. Kibria and H. Rivaz, "Gluenet: Ultrasound elastography using convolutional neural network", in *Simulation, Image Processing, and Ultrasound Systems for Assisted Diagnosis and Navigation*, Springer, 2018, pp. 21–28.

[54] B. Peng, Y. Xian, Q. Zhang, and J. Jiang, "Neural-network-based motion tracking for breast ultrasound strain elastography: An initial assessment of performance and feasibility", *Ultrasonic imaging*, vol. 42, no. 2, pp. 74–91, 2020.

[55] A. K. Tehrani and H. Rivaz, "Displacement estimation in ultrasound elastography using pyramidal convolutional neural network", *IEEE transactions on ultrasonics, ferroelectrics, and frequency control*, vol. 67, no. 12, pp. 2629–2639, 2020.

[56] ——, "Mpwc-net++: Evolution of optical flow pyramidal convolutional neural network for ultrasound elastography", in *Medical Imaging 2021: Ultrasonic Imaging and Tomography*, International Society for Optics and Photonics, vol. 11602, 2021, p. 1 160 206.

[57] T. Ahmed, M. Hasan, *et al.*, "Shear-net: An end-to-end deep learning approach for single push ultrasound shear wave elasticity imaging", *arXiv preprint arXiv:1902.04845*, 2019.

[58] F. Liu, D. Liu, J. Tian, X. Xie, X. Yang, and K. Wang, "Cascaded one-shot deformable convolutional neural networks: Developing a deep learning model for respiratory motion estimation in ultrasound sequences", *Medical Image Analysis*, vol. 65, p. 101 793, 2020.

[59] S. S. Ahn, K. Ta, A. Lu, J. C. Stendahl, A. J. Sinusas, and J. S. Duncan, "Unsupervised motion tracking of left ventricle in echocardiography", in *Medical Imaging 2020: Ultrasonic Imaging and Tomography*, International Society for Optics and Photonics, vol. 11319, 2020, 113190Z.

[60] K. Ta, S. S. Ahn, A. Lu, J. C. Stendahl, A. J. Sinusas, and J. S. Duncan, "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.

[61] C. Zhao, C. Feng, D. Li, and S. Li, "Of-msrn: Optical flow-auxiliary multi-task regression network for direct quantitative measurement, segmentation and motion estimation", in *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 34, 2020, pp. 1218–1225.

[62] A. K. Tehrani, M. Mirzaei, and H. Rivaz, "Semi-supervised training of optical flow convolutional neural networks in ultrasound elastography", *arXiv preprint arXiv:2007.01421*, 2020.

[63] S. Baker, D. Scharstein, J. Lewis, S. Roth, M. J. Black, and R. Szeliski, "A database and evaluation methodology for optical flow", *International journal of computer vision*, vol. 92, no. 1, pp. 1–31, 2011.

[64] M. Otte and H.-H. Nagel, "Optical flow estimation: Advances and comparisons", in *European conference on computer vision*, Springer, 1994, pp. 49–60.

[65] A. Geiger, P. Lenz, and R. Urtasun, "Are we ready for autonomous driving? the kitti vision benchmark suite", in *Conference on Computer Vision and Pattern Recognition (CVPR)*, 2012.

[66] M. Menze and A. Geiger, "Object scene flow for autonomous vehicles", in *Conference on Computer Vision and Pattern Recognition (CVPR)*, 2015.

[67] N. Mayer, E. Ilg, P. Hausser, P. Fischer, D. Cremers, A. Dosovitskiy, and T. Brox, "A large dataset to train convolutional networks for disparity, optical flow, and scene flow estimation", in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2016, pp. 4040–4048.

[68] G. Schröder, T. Senst, E. Bochinski, and T. Sikora, "Optical flow dataset and benchmark for visual crowd analysis", in *IEEE International Conference on Advanced Video and Signals-based Surveillance*, 2018.

[69] M. Shugrina, Z. Liang, A. Kar, J. Li, A. Singh, K. Singh, and S. Fidler, "Creative flow+ dataset", in *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, Jun. 2019.

[70] M. J. Ledesma-Carbayo, J. Kybic, M. Desco, A. Santos, M. Suhling, P. Hunziker, and M. Unser, "Spatio-temporal nonrigid registration for ultrasound cardiac motion estimation", *IEEE transactions on medical imaging*, vol. 24, no. 9, pp. 1113–1126, 2005.

[71] R. Haddad, P. Clarysse, M. Orkisz, P. Croisille, D. Revel, and I. E. Magnin, "A realistic anthropomorphic numerical model of the beating heart", in *International Workshop on Functional Imaging and Modeling of the Heart*, Springer, 2005, pp. 384–393.

[72] C. Butakoff, S. Balocco, and S. Ordas, "Simulated 3d ultrasound lv cardiac images for active shape model training", in *Medical Imaging 2007: Image Processing*, International Society for Optics and Photonics, vol. 6512, 2007, 65123U.

[73] M. Sermesant, H. Delingette, and N. Ayache, "An electromechanical model of the heart for image analysis and simulation", *IEEE transactions on medical imaging*, vol. 25, no. 5, pp. 612–625, 2006.

[74] O. Kutter, R. Shams, and N. Navab, "Visualization and gpu-accelerated simulation of medical ultrasound from ct images", *Computer methods and programs in biomedicine*, vol. 94, no. 3, pp. 250–266, 2009.

[75] M. De Craene, S. Marchesseau, B. Heyde, H. Gao, M. Alessandrini, O. Bernard, G. Piella, A. Porras, L. Tautz, A. Hennemuth, *et al.*, "3d strain assessment in ultrasound (straus): A synthetic comparison of five tracking methodologies", *IEEE transactions on medical imaging*, vol. 32, no. 9, pp. 1632–1646, 2013.

[76] M. Alessandrini, H. Liebgott, D. Friboulet, and O. Bernard, "Simulation of realistic echocardiographic sequences for ground-truth validation of motion estimation", in *2012 19th IEEE International Conference on Image Processing*, IEEE, 2012, pp. 2329–2332.

[77] A. Prakosa, M. Sermesant, H. Delingette, S. Marchesseau, E. Saloux, P. Allain, N. Villain, and N. Ayache, "Generation of synthetic but visually realistic time series of cardiac images combining a biophysical model and clinical images", *IEEE transactions on medical imaging*, vol. 32, no. 1, pp. 99–109, 2012.

[78] M. Alessandrini, M. De Craene, O. Bernard, S. Giffard-Roisin, P. Allain, I. Waechter-Stehle, J. Weese, E. Saloux, H. Delingette, M. Sermesant, *et al.*, "A pipeline for the generation of realistic 3d synthetic echocardiographic sequences: Methodology and open-access database", *IEEE transactions on medical imaging*, vol. 34, no. 7, pp. 1436–1451, 2015.

[79] H. Gao, H. F. Choi, P. Claus, S. Boonen, S. Jaecques, G. H. Van Lenthe, G. Van der Perre, W. Lauriks, and J. D'hooge, "A fast convolution-based methodology to simulate 2-dd/3-d cardiac ultrasound images", *IEEE transactions on ultrasonics, ferroelectrics, and frequency control*, vol. 56, no. 2, pp. 404–409, 2009.

[80] Y. Zhou, S. Giffard-Roisin, M. De Craene, S. Camarasu-Pop, J. D'Hooge, M. Alessandrini, D. Friboulet, M. Sermesant, and O. Bernard, "A framework for the generation of realistic synthetic cardiac ultrasound and magnetic resonance imaging sequences from the same virtual patients", *IEEE transactions on medical imaging*, vol. 37, no. 3, pp. 741–754, 2017.

[81] J. D'hooge, D. Barbosa, H. Gao, P. Claus, D. Prater, J. Hamilton, P. Lysyansky, Y. Abe, Y. Ito, H. Houle, *et al.*, "Two-dimensional speckle tracking echocardiography: Standardization efforts based on synthetic ultrasound data", *European Heart Journal–Cardiovascular Imaging*, vol. 17, no. 6, pp. 693–701, 2016.

[82] M. Alessandrini, B. Chakraborty, B. Heyde, O. Bernard, M. De Craene, M. Sermesant, and J. D'Hooge, "Realistic vendor-specific synthetic ultrasound data for quality assurance of 2-d speckle tracking echocardiography: Simulation pipeline and open access database", *IEEE transactions on ultrasonics, ferroelectrics, and frequency control*, vol. 65, no. 3, pp. 411–422, 2017.

[83] C. Chnafa, S. Mendez, and F. Nicoud, "Image-based large-eddy simulation in a realistic left heart", *Computers & Fluids*, vol. 94, pp. 173–187, 2014.

[84] D. Garcia, "Simus: An open-source simulator for ultrasound imaging. part i: Theory & examples", *arXiv preprint arXiv:2102.02738*, 2021.

[85] S. Shahriari and D. Garcia, "Meshfree simulations of ultrasound vector flow imaging using smoothed particle hydrodynamics", *Physics in Medicine & Biology*, vol. 63, no. 20, p. 205 011, 2018.

[86] A. H. Abdi, T. Tsang, and P. Abolmaesumi, "Gan-enhanced conditional echocardiogram generation", *arXiv preprint arXiv:1911.02121*, 2019.

[87] V. Zyuzin, J. Komleva, and S. Porshnev, "Generation of echocardiographic 2d images of the heart using cgan", in *Journal of Physics: Conference Series*, IOP Publishing, vol. 1727, 2021, p. 012 013.

[88] E. Evain, K. Faraz, T. Grenier, D. Garcia, M. De Craene, and O. Bernard, "A pilot study on convolutional neural networks for motion estimation from ultrasound images", *IEEE transactions on ultrasonics, ferroelectrics, and frequency control*, vol. 67, no. 12, pp. 2565–2573, 2020.

[89] A. Østvik, E. Smistad, S. A. Aase, B. O. Haugen, and L. Lovstakken, "Real-time standard view classification in transthoracic echocardiography using convolutional neural networks", *Ultrasound in Medicine and Biology*, vol. 45, no. 2, pp. 374–384, Feb. 2019, ISSN: 0301-5629.

[90] S. Cai, S. Zhou, C. Xu, and Q. Gao, "Dense motion estimation of particle images via a convolutional neural network", *Experiments in Fluids*, vol. 60, no. 4, p. 73, 2019.

[91] J. Porée, D. Posada, A. Hodzic, F. Tournoux, G. Cloutier, and D. Garcia, "High-frame-rate echocardiography using coherent compounding with doppler-based motion-compensation", *IEEE transactions on medical imaging*, vol. 35, no. 7, pp. 1647–1657, 2016.

[92] S. Shahriari and D. Garcia, "Meshfree simulations of ultrasound vector flow imaging using smoothed particle hydrodynamics", *Physics in Medicine & Biology*, vol. 63, no. 20, p. 205 011, Oct. 2018.

[93] P. Y. Simard, D. Steinkraus, and J. C. Platt, "Best practices for convolutional neural networks applied to visual document analysis", in *null*, IEEE, 2003, p. 958.

[94] T. DeVries and G. W. Taylor, "Improved regularization of convolutional neural networks with cutout", *arXiv preprint arXiv:1708.04552*, 2017.

[95] P. Joos, J. Porée, H. Liebgott, D. Vray, M. Baudet, J. Faurie, F. Tournoux, G. Cloutier, B. Nicolas, and D. Garcia, "High-frame-rate speckle-tracking echocardiography", *IEEE Transactions on Ultrasonics, Ferroelectrics, and Frequency Control*, vol. 65, no. 5, pp. 720–728, May 2018.

[96] E. Evain, Y. Sun, K. Faraz, D. Garcia, E. Saloux, B. L. Gerber, M. De Craene, and O. Bernard, "Motion estimation by deep learning in 2d echocardiography: Synthetic dataset and validation", *IEEE Transactions on Medical Imaging*, 2022.

[97] C. Bailer, B. Taetz, and D. Stricker, "Flow fields: Dense correspondence fields for highly accurate large displacement optical flow estimation", *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 41, no. 8, pp. 1879–1892, 2019.

[98] K. Ta, S. S. Ahn, A. Lu, J. C. Stendahl, A. J. Sinusas, and J. S. Duncan, "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)*, 2020, pp. 1734–1737. DOI: 10.1109/ISBI45749.2020.9098664.

[99] S. Shahriari and D. Garcia, "Meshfree simulations of ultrasound vector flow imaging using smoothed particle hydrodynamics", *Physics in Medicine & Biology*, vol. 63, no. 20, p. 205 011, Oct. 2018. DOI: 10.1088/1361-6560/aae3c3. [Online]. Available: https://doi.org/10.1088/1361-6560/aae3c3.

[100] V. Perrot, M. Polichetti, F. Varray, and D. Garcia, "So you think you can das? a viewpoint on delay-and-sum beamforming", *Ultrasonics*, vol. 111, p. 106 309, 2021.

[101] S. Leclerc, E. Smistad, J. Pedrosa, A. Østvik, F. Cervenansky, F. Espinosa, T. Espeland, E. A. R. Berg, P.-M. Jodoin, T. Grenier, *et al.*, "Deep learning for segmentation using an open large-scale dataset in 2d echocardiography", *IEEE transactions on medical imaging*, 2019.

[102] H. Jacinto, R. Kéchichian, M. Desvignes, R. Prost, and S. Valette, "A web interface for 3d visualization and interactive segmentation of medical images", in *Proceedings of the 17th International Conference on 3D Web Technology*, 2012, pp. 51–58.

[103] D. Garcia, "A fast all-in-one method for automated post-processing of piv data", *Experiments in Fluids*, vol. 50, no. 5, pp. 1247–1259, May 2011.

[104] K. He, X. Zhang, S. Ren, and J. Sun, "Delving deep into rectifiers: Surpassing human-level performance on imagenet classification", in *Proceedings of the IEEE international conference on computer vision*, 2015, pp. 1026–1034.

[105] C. Mattiuzzi and G. Lippi, "Current cancer epidemiology", *Journal of epidemiology and global health*, vol. 9, no. 4, p. 217, 2019.

[106] A. Dibden, J. Offman, S. W. Duffy, and R. Gabe, "Worldwide review and meta-analysis of cohort studies measuring the effect of mammography screening programmes on incidence-based breast cancer mortality", *Cancers*, vol. 12, no. 4, p. 976, 2020.

[107] C. M. Sehgal, S. P. Weinstein, P. H. Arger, and E. F. Conant, "A review of breast ultrasound", *Journal of mammary gland biology and neoplasia*, vol. 11, no. 2, pp. 113–123, 2006.

[108] A. T. Stavros, D. Thickman, C. L. Rapp, M. A. Dennis, S. H. Parker, and G. A. Sisney, "Solid breast nodules: Use of sonography to distinguish between benign and malignant lesions.", *Radiology*, vol. 196, no. 1, pp. 123–134, 1995.

[109] L. Liberman and J. H. Menell, "Breast imaging reporting and data system (bi-rads)", *Radiologic Clinics*, vol. 40, no. 3, pp. 409–430, 2002.

[110] E. B. Mendelson, W. A. Berg, and C. R. Merritt, "Toward a standardized breast ultrasound lexicon, bi-rads: Ultrasound", in *Seminars in roentgenology*, Elsevier, vol. 36, 2001, pp. 217–225.

[111] J.-Z. Cheng, D. Ni, Y.-H. Chou, J. Qin, C.-M. Tiu, Y.-C. Chang, C.-S. Huang, D. Shen, and C.-M. Chen, "Computer-aided diagnosis with deep learning architecture: Applications to breast lesions in us images and pulmonary nodules in ct scans", *Scientific reports*, vol. 6, no. 1, pp. 1–13, 2016.

[112] J. Shi, S. Zhou, X. Liu, Q. Zhang, M. Lu, and T. Wang, "Stacked deep polynomial network based representation learning for tumor classification with small ultrasound image dataset", *Neurocomputing*, vol. 194, pp. 87–94, 2016.

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

[114] A. Hijab, M. A. Rushdi, M. M. Gomaa, and A. Eldeib, "Breast cancer classification in ultrasound images using transfer learning", in *2019 Fifth International Conference on Advances in Biomedical Engineering (ICABME)*, IEEE, 2019, pp. 1–4.

[115] M. Byra, M. Galperin, H. Ojeda-Fournier, L. Olson, M. O'Boyle, C. Comstock, and M. Andre, "Breast mass classification in sonography with transfer learning using a deep convolutional neural network and color conversion", *Medical physics*, vol. 46, no. 2, pp. 746–755, 2019.

[116] C. Szegedy, W. Liu, Y. Jia, P. Sermanet, S. Reed, D. Anguelov, D. Erhan, V. Vanhoucke, and A. Rabinovich, "Going deeper with convolutions", in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2015, pp. 1–9.

[117] T. Fujioka, K. Kubota, M. Mori, Y. Kikuchi, L. Katsuta, M. Kasahara, G. Oda, T. Ishiba, T. Nakagawa, and U. Tateishi, "Distinction between benign and malignant breast masses at breast ultrasound using deep learning method with convolutional neural network", *Japanese journal of radiology*, vol. 37, no. 6, pp. 466–472, 2019.

[118] S. Han, H.-K. Kang, J.-Y. Jeong, M.-H. Park, W. Kim, W.-C. Bang, and Y.-K. Seong, "A deep learning framework for supporting the classification of breast lesions in ultrasound images", *Physics in Medicine & Biology*, vol. 62, no. 19, p. 7714, 2017.

[119] K. He, G. Gkioxari, P. Dollár, and R. Girshick, "Mask r-cnn", in *Proceedings of the IEEE international conference on computer vision*, 2017, pp. 2961–2969.

[120] V. Couteaux, S. Si-Mohamed, O. Nempont, T. Lefevre, A. Popoff, G. Pizaine, N. Villain, I. Bloch, A. Cotten, and L. Boussel, "Automatic knee meniscus tear detection and orientation classification with mask-rcnn", *Diagnostic and interventional imaging*, vol. 100, no. 4, pp. 235–242, 2019.

[121] Y. Liang, R. He, Y. Li, and Z. Wang, "Simultaneous segmentation and classification of breast lesions from ultrasound images using mask r-cnn", in *2019 IEEE International Ultrasonics Symposium (IUS)*, IEEE, 2019, pp. 1470–1472.

[122] J.-Y. Chiao, K.-Y. Chen, K. Y.-K. Liao, P.-H. Hsieh, G. Zhang, and T.-C. Huang, "Detection and classification the breast tumors using mask r-cnn on sonograms", *Medicine*, vol. 98, no. 19, 2019.

[123] N. Lassau, I. Bousaid, E. Chouzenoux, A. Verdon, C. Balleyguier, F. Bidault, E. Mousseaux, S. Harguem-Zayani, L. Gaillandre, Z. Bensalah, *et al.*, "Three artificial intelligence data challenges based on ct and ultrasound", *Diagnostic and Interventional Imaging*, vol. 102, no. 11, pp. 669–674, 2021.

[124] A. Dutta and A. Zisserman, "The via annotation software for images", *Audio and Video*, vol. 31, 2019.

[125] T.-Y. Lin, P. Dollár, R. Girshick, K. He, B. Hariharan, and S. Belongie, "Feature pyramid networks for object detection", in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2017, pp. 2117–2125.

[126] S. Ren, K. He, R. Girshick, and J. Sun, "Faster r-cnn: Towards real-time object detection with region proposal networks", *Advances in neural information processing systems*, vol. 28, pp. 91–99, 2015.

[127] T.-Y. Lin, M. Maire, S. Belongie, J. Hays, P. Perona, D. Ramanan, P. Dollár, and C. L. Zitnick, "Microsoft coco: Common objects in context", in *European conference on computer vision*, Springer, 2014, pp. 740–755.

FOLIO ADMINISTRATIF

THESE DE L'UNIVERSITE DE LYON OPEREE AU SEIN DE L'INSA LYON

NOM : EVAIN
(avec précision du nom de jeune fille, le cas échéant)

DATE de SOUTENANCE : 29/04/2022

Prénoms : Ewan

TITRE : Apprentissage profond pour l'estimation de mouvement en imagerie ultrasonore: application à la quantification des déformations du myocarde

NATURE : Doctorat

Numéro d'ordre : 2022LYSEI037

Ecole doctorale : EDA 160 : Electronique, Electrotechnique, Automatique (EEA)

Spécialité : Traitement du signal et des images

RESUME : La modalité d'imagerie la plus utilisée en pratique clinique est actuellement l'imagerie ultrasonore car celle-ci est peu couteuse, rapide et non-invasive. En échocardiographie, plusieurs indices caractérisant la fonction cardiaque peuvent être extraits de ces acquisitions, parmi lesquels la déformation globale longitudinale joue un rôle important dans l'établissement d'un diagnostic. Cependant l'estimation de cet indice souffre d'un manque de reproductibilité du fait des caractéristiques inhérentes à l'imagerie ultrasonore. En effet, les méthodes traditionnelles comme le flux optique ou la correspondance de blocs ne permettent pas de gérer des artéfacts types tels que la décorrélation de texture ultrasonore. Récemment, les approches d'apprentissage profond ont battu les méthodes l'état de l'art en estimation de mouvement, portées par les applications à la robotique ou les voitures autonomes. Dans la première partie de cette thèse, nous présentons une étude pilote afin d'évaluer la capacité des méthodes d'apprentissage profond à estimer le mouvement en imagerie ultrasonore malgré les nombreux artéfacts sous-jacents. Pour ce faire, nous avons créé une base de données composée d'images ultrasonores simulées et in-vitro incluant un disque tournant avec des vitesses variables. Dans la seconde partie de cette thèse, nous détaillons le réseau de neurones pyramidal que nous avons développé afin d'estimer la déformation du muscle myocardique et qui améliore de façon significative les performances des méthodes de l'état de l'art. Pour entrainer et évaluer notre méthode d'apprentissage, nous avons également implémenté un pipeline de simulations permettant de générer des séquences d'images échocardiographiques réalistes avec un champ dense de référence et présentant une grande variabilité anatomique et fonctionnelle.

MOTS-CLÉS : Echocardiographie, Tracking du myocarde, Déformation du myocarde, Apprentissage profond, Estimation de mouvement, Imagerie Ultrasonore, Simulation de données, Réseaux de neurones convolutionnels, Base de données

Laboratoire (s) de recherche : CREATIS

Directeur de thèse: BERNARD Olivier

Président de jury : BLOCH Isabelle

Composition du jury : PETITJEAN Caroline (Rapporteure), THOME Nicolas (Rapporteur), VAN SLOUN Ruud (Examinateur), DE CRAENE Mathieu (Examinateur), BLOCH Isabelle (Examinatrice, Présidente de Jury), BERNARD Olivier (Directeur de Thèse)