الجـــــــــــــــــــــمهـوريـــــة الجـــــــــــــــزائـريـة الديمـــــــقـراطيـة الـشـــــــــعبيـة

Democratic and Popular Republic of Algeria

وزارة التـــــــــــــــــــــــــــــعـليـم العـــــــالي و البـــــــــــحث العـــــلـمـــي

Ministry of Higher Education and Scientific Research

جـــــــــــــــامعة أبي بكـــر بـلـقـايد – تـلمســـــــــــــان –

Aboubakr Belkaïd University – Tlemcen –
Faculty of TECHNOLOGY

THESIS

Presented to obtain the **degree** of **DOCTORATE (Third Cycle)**

In : Biomedical Engeneeing

Speciality : Medical Imaging

By : Hafida BELFILALI

Topic

Analysis of echocardiographic image sequences to study left ventricular performance.

Publicly defended, on 25 / 09 / 2023 , to the jury composed of :

M. DJEBBARI Abdelghani	Professor	Univ. Tlemcen	President
M. MESSADI Mahammed	Professor	Univ. Tlemcen	Advisor
M. BOUSEFSAF Frédéric	Associate Professor	Univ. Lorraine	Co-Advisor
M. MEGNAFI Hicham	Associate Professor	ESSA, Univ. Tlemcen	Examiner 1
M. DIB Nabil	Associate Professor	Univ. Tlemcen	Examiner 2
Mme. Maaoui Choubeila	Professor	Univ. Lorraine	Guest

I proudly dedicate this thesis to my hero, my FATHER
You will never be able to see this work
But you are on every page ...

Acknowledgement

First and foremost, I want to express my gratitude to my supervisor Mr. Mahammed MESSADI for his leadership, guidance, and support since the beginning. I am also grateful to my endless source of inspiration, my co-supervisor Mr. Frédéric BOUSEFSAF for his insightful advice, inspiring knowledge, and motivation behind carrying out this project.

I would like to thank the GBM laboratory, under the direction of Mr. Abdelghani DJEBBARI, for enabling me to pursue my research work. My special thanks also go to Mr. Imed KACEM for allowing me to join the LCOMS during the last year of my thesis.

There are so many people to thank and acknowledge. I express my thanks to my friends for their encouragement and love. The enthusiasm of my sisters and brother towards my research and their support were a vital source of motivation, especially in the difficult moments. I would like to express my gratitude to them.

Last but not least, I want to thank the woman who is so dedicated, passionate, ambitious, and courageous, my wonderful mother. She has always been my backbone in everything I do and has given me all the support I need. You have taught me so much (mama).

ملخص

أمراض القلب والأوعية الدموية هي أمراض تؤثر على القلب والأوعية الدموية. وفقا لمنظمة الصحة العالمية ، فهي تعتبر السبب الرئيسي للوفيات في جميع أنحاء العالم. يعد التشخيص المبكر لاضطرابات وظائف القلب أمرا بالغ الأهمية في تقليل معدل الوفيات. البطين الأيسر هو عنصر حيوي في الجهاز القلبي الوعائي ويلعب دورا مهما في الدورة الدموية. يمكن تقدير العديد من المعلمات السريرية من بنية البطين الأيسر أثناء فحوصات القلب والأوعية الدموية لضمان التشخيص الموثوق به ، بما في ذلك حجم البطين الأيسر والكسر القذفي.

تسمح طرق التصوير القلبي المختلفة بتصوير تجويف البطين الأيسر. تخطيط صدى القلب هو الأسلوب الأكثر استخداما من قبل أطباء القلب في الممارسة السريرية الروتينية نظرا لمزاياها العديدة. الطريقة الأساسية لتقدير المعلمات السريرية هي تجزئة سطح البطين الأيسر من متواليات صور تخطيط صدى القلب ثنائية الأبعاد. يعتمد التقييم الدقيق لوظيفة البطين الأيسر على جودة نتائج التجزئة. ومع ذلك ، فإن التحديد اليدوي للبطين الأيسر من قبل أطباء القلب أمر صعب ويستغرق وقتا طويلا وغير دقيق بسبب الجودة المنخفضة لصور تخطيط صدى القلب. لذلك ، هناك حاجة لتقسيم البطين الأيسر تلقائيا من تسلسل صور تخطيط صدى القلب للتغلب على هذه التحديات.

في هذه الأطروحة ، هدفنا هو تطوير إطار عمل تجزئة آلي بالكامل يعتمد على تقنيات التعلم العميق لتقييم أداء البطين الأيسر باستخدام صور تخطيط صدى القلب. اختبرنا فعالية الأساليب المقترحة من خلال مقارنة النتائج التي تم الحصول عليها مع بيانات الحقيقة الأساسية والأساليب الحديثة الموجودة في هذا المجال. كانت النتائج مرضية ، مما يؤكد الإمكانات الكبيرة للتقنيات الآلية لتحليل صور تخطيط صدى القلب لمساعدة أطباء القلب في ممارستهم السريرية اليومية.

الكلمات المفتاحية: البطين الايسر ؛ تخطيط صدى القلب ؛ تجزئة ؛ تحليل صور تخطيط القلب ؛ تعلم عميق ؛ تحليل صور تخطيط القلب ؛ خوارزمية U-Net ؛ آلية الانتباه ؛ نقل التعلم.

Résumé

Les maladies cardiovasculaires sont des pathologies qui affectent le cœur et les vaisseaux sanguins. Selon l'organisation mondiale de la santé, elles sont la première cause de mortalité dans le monde. Le diagnostic précoce des troubles de la fonction cardiaque est essentiel pour réduire le taux de mortalité. Le ventricule gauche (VG) est un élément vital du système cardiovasculaire et joue un rôle important dans la circulation sanguine. Plusieurs paramètres cliniques peuvent être estimés à partir de la structure du ventricule gauche lors des examens cardiovasculaires afin de garantir des diagnostics fiables, notamment les volumes ventriculaires gauches et la fraction d'éjection ventriculaire gauche.

Diverses modalités d'imagerie cardiaque permettent de visualiser la cavité ventriculaire gauche. L'échocardiographie est la technique la plus utilisée par les cardiologues dans la pratique clinique courante en raison de ses nombreux avantages. La principale méthode d'estimation des paramètres cliniques est la segmentation de la surface du ventricule gauche à partir de séquences d'images échocardiographiques 2D. L'évaluation précise de la fonction de la cavité ventriculaire gauche dépend de la qualité des résultats de la segmentation. Cependant, la délimitation manuelle du VG par les cardiologues est difficile, longue et imprécise en raison de la faible qualité des images échocardiographiques. Par conséquent, il est nécessaire de segmenter automatiquement le VG à partir de séquences d'images échocardiographiques afin de surmonter ces difficultés.

Dans cette thèse, notre objectif est de développer un système de segmentation entièrement automatique basé sur des techniques d'apprentissage profond pour évaluer la performance du VG à l'aide d'images échocardiographiques. Nous avons testé l'efficacité des approches proposées en comparant les résultats obtenus avec les données de vérité terrain et les méthodes existantes dans ce domaine. Les résultats sont satisfaisants, soulignant le potentiel significatif des techniques automatisées pour l'analyse des images échocardiographiques afin d'aider les cardiologues dans leur pratique clinique quotidienne.

Mots Clée : Ventricule gauche ; Échocardiographie ; Segmentation ; Analyse d'images échocardiographiques ; Apprentissage profond ; Architecture U-Net ; Mécanisme d'attention ; Apprentissage par transfert.

Abstract

Cardiovascular diseases are pathologies that affect the heart and blood vessels. According to the world health organization, they are the leading cause of mortality worldwide. Early diagnosis of cardiac function disorders is crucial in reducing the mortality rate. The Left Ventricle (LV) is a vital component of the cardiovascular system and plays a significant role in blood circulation. Several clinical parameters can be estimated from the LV structure during cardiovascular exams to ensure reliable diagnoses, including left ventricular volumes and ejection fraction.

Various cardiac imaging modalities allow visualization of the left ventricular cavity. Echocardiography is the most widely used technique by cardiologists in routine clinical practice due to its many advantages. The primary method for estimating clinical parameters is LV surface segmentation from 2D echocardiographic image sequences. The accurate evaluation of the LV chamber's function relies on the quality of the segmentation results. However, LV manual delineation by cardiologists is difficult, time-consuming, and imprecise due to the low quality of echocardiographic images. Therefore, there is a need to automatically segment the LV from echocardiographic image sequences to overcome these challenges.

In this thesis, our objective is to develop a fully automatic segmentation framework based on deep learning techniques to assess LV performance using echocardiographic images. We tested the effectiveness of the proposed approaches by comparing the obtained results with ground truth data and existing state-of-the-art methods in this field. The results are satisfactory, underlining the significant potential of automated techniques for echocardiographic image analysis to help cardiologists in their daily clinical practice.

Keywords: Left ventricle; Echocardiography; Segmentation; Echocardiographic image analysis; Deep learning; U-Net architecture; Attention mechanism; Transfer learning.

Contents

List of Figures

List of Tables

List of Algorithms

Nomenclature

Acronymes / Abbreviations

A2C	Apical 2 Chamber view
A4C	Apical 4 Chamber view
AAMs	Active Appearance Models
Acc	Accuracy
Adam	Adaptive Moment Estimation
AI	Artificial intelligence
AMBE	Absolute Mean Brightness Error
ANN	Artificial Neural Network
ASMs	Active shape models
AUC	Area Under the ROC Curve
B-mode	Brightness-mode
CAMUS	Cardiac Acquisitions for Multi-structure Ultrasound Segmentation
CLAHE	Contrast Limited Adaptive Histogram Equalization
CNN	Convolutional Neural Network
CO	Cardiac Output
Conv2DTranspose	transpose convolutional layer
Corr	Pearson correlation coefficient
CT	Computed Tomography

DBNs	Deep Belief Networks
DenseNet	Densely Connected Convolutional Networks
DSC	Dice Similarity Coefficient
EF	Ejection Fraction
ED	End Diastole
LV_{Endo}	Left Ventricle Endocardium
ES	End Systole
LV_{Epi}	Left Ventricle Epicardium
FC	Fully Connected layer
FCM	fuzzy C-Means
TP	False Positive
GAN	Generative Adversarial Networks
HD	Hausdorff Distance
HR	Heart Rate
JC	Jaccard Coefficient
LA	Left Atrium
LV	Left Ventricle
LSTM	Long Short-Term Memory
LV_{EDV}	Left Ventricular End Diastolic Volume
LV_{EF}	Left Ventricular Ejection Fraction
LV_{ESV}	Left Ventricular End Systolic Volume
MAD	Mean Absolute Distance
MAE	Mean Absolute Error
ML	Machine learning

M-mode	Motion-mode
MRI	Magnetic Resonance Imaging
MSE	Mean Squared Error
LV_{Myo}	Left Ventricle Myocardium
PSNR	Peak Signal to Noise Ratio
RA	Right Atrium
ReLU	Rectified Linear Unit
ResNet	Deep Residual Learning Network
RMS-Prop	Root Mean Square Propagation
RNN	recurrent neural network
ROC	Receiver Operating Characteristic
RV	Right Ventricle
SGD	Stochastic gradient descent
SSIM	Structural Similarity Index
SV	Stroke Volume
tanh	hyperbolic tangent activation function
FN	False Negative
TN	True Negative
TP	True Positive
TTE	Transthoracic Echocardiogram
VGG-Net	Visual Geometry Group
ViTs	Vision Transformers

General Introduction

Problem statement

According to the world health organization, cardiovascular diseases caused 17.9 million deaths worldwide in 2019, accounting for 32% of all global fatalities [19]. Heart attacks and strokes were the primary causes. Hence, cardiovascular diseases are the leading cause of death globally. These diseases fall under the category of heart and blood vessel disorders. They have a significant impact on the vital functions of the body. One integral component of the cardiovascular system is the Left Ventricle (LV), which contracts to force oxygenated blood through the aortic valve, distributing it throughout the body [20]. The LV is particularly susceptible to most cardiovascular diseases. Any abnormality in its function can lead to pathological symptoms. Early detection of cardiac function anomalies is crucial for effective treatment and reduction in mortality rates. For that, it is necessary to develop innovative clinical procedures for early diagnosis.

Various cardiac imaging modalities allow for the visualization of the structure and function of the heart, enabling medical professionals to diagnose a range of cardiac anomalies and guide therapeutic interventions and invasive procedures for cardiovascular diseases. Among these modalities, echocardiography is the most commonly used technique in clinical practice. An echocardiogram can directly visualize the size of the heart chambers, ventricular wall thickness, and any structural anomalies [21]. It also enables the evaluation of contractility and assessment of the left ventricular ejection function.

Echocardiography utilizes safe sound waves to obtain cardiac images. This imaging modality doesn't present any known danger to the body. Echocardiography offers several advantages, including wide availability, affordability, and high temporal resolution. There are several echocardiography techniques, such as 2D or 3D modalities. The 2D modality is the most widely used technique for measuring Ejection Fraction (EF). This parameter is calculated based on the segmentation and measurement of left ventricular volumes from the surface. Therefore, we must segment the 2D echocardiographic images to obtain the left ventricular area.

Image segmentation in echocardiography plays a crucial role in cardiac image processing. Accurate LV segmentation is essential for comprehensive analysis and assessment of its function. Precise segmentation in the End Diastole (ED) and End Systole (ES) phases enables the quantification and evaluation of LV chamber function. Usually, cardiologists perform LV delineation manually during the interpretation of echocardiograms. However, manual segmentation of the LV in echocardiography presents several challenges.

Ultrasound images have low contrast and a low signal-to-noise ratio. The presence of speckle noise and other artifacts contributes to poor image quality, making the boundaries of the cardiac chambers unclear. Additionally, manual segmentation of the LV is a time-consuming and laborious task that requires skilled clinicians. This process may also introduce variability in LV delimitations between multiple users and even by the same user.

Accurate automatic segmentation of the LV from echocardiographic image sequences is essential to overcome the complications and issues mentioned earlier. It serves as a reliable solution for automatically measuring cardiac morphology and function. The development of fully-automatic ultrasound cardiac segmentation software holds great potential in aiding cardiologists in the early detection and diagnosis of cardiovascular diseases.

Motivation

Medical image processing plays a relevant role in extracting critical and pertinent information from medical images automatically or semi-automatically. It has a significant impact on clinical procedures, particularly in the field of cardiac analysis of echocardiographic images. As a result, extensive research has been conducted in the past few decades, focusing on LV segmentation. Semi-automatic techniques have been used traditionally to segment the LV in echocardiographic images. These methods require the observer to outline the region of interest, after which the algorithm determines the best-fitting contour of the LV. However, automatic image segmentation offers several advantages over semi-automatic techniques, making it more approved. Automatic segmentation algorithms can process large datasets much faster than semi-automatic methods. They can be easily scaled to handle a wide range of image sizes and complexities without a significant increase in time and effort. Moreover, they requires minimal or no user interaction, reducing the time and effort required to annotate or refine segmentation masks manually.

In recent years, there has been significant progress in computer-aided diagnosis systems that leverage medical image processing and artificial intelligence. These systems play a crucial role in assisting clinicians in the early and accurate identification of diseases. They

have attracted attention as valuable tools for making timely clinical decisions and have been the subject of numerous studies. Researchers have been highly motivated to design fully automatic systems for medical image analysis, particularly leveraging deep learning techniques. Convolutional Neural Networks (CNNs) have gained prominence due to their excellent performance in computer vision tasks. CNNs are a type of Artificial Neural Network (ANN) that incorporate one or more convolutional layers. The architecture and configuration of CNNs greatly influence the performance of deep learning models.

Training a deep learning model requires a large amount of data. However, in many application domains, including medical imaging, obtaining a sufficient quantity of labeled data for training neural networks can be challenging and resource-intensive [22]. The parameters of artificial intelligence algorithms are estimated through supervised learning using annotated data cases. The use of such models in the medical community, particularly in echocardiography, has been limited by the difficulty of obtaining expert-annotated medical data. Echocardiographic data stored in medical archives are rarely annotated or labeled by experts. Consequently, the scarcity of labeled echocardiographic images has significantly restricted the availability of publicly accessible datasets.

These challenges and limitations have motivated our research to address the assessment of left ventricular function in echocardiographic images. The identified difficulties have guided us in defining precise aims and objectives, leading us to develop dedicated approaches for rapid and accurate LV chamber assessment.

Aims and contributions

This Ph.D. work aims to develop the framework described earlier to assess LV performance. The primary objective is to automate the process of left ventricular segmentation and accurately estimate the end-diastolic and end-systolic volumes and the EF from echocardiographic image sequences. For that, it is necessary to develop robust, fully automatic segmentation algorithms that strongly support cardiologists in their clinical routine and help reduce inter- and intra-observer variability.

There are numerous intelligent algorithms applied to various tasks in medical image interpretation. CNNs, such as U-Net architecture [9], have been extensively utilized in medical image segmentation. U-Net architecture has demonstrated fast and accurate segmentation capabilities for medical images. It has been successfully adapted in many research works for segmenting medical ultrasound images [23]. In particular, it has shown exceptional performance for echocardiographic image segmentation.

The primary aim of this thesis is to address the following question: Can an efficient

technique be developed using the U-Net architecture to enhance the assessment of LV performance from echocardiographic image sequences? The proposed framework should effectively address the scarcity of available data and exhibit generalizability to unseen data.

Based on the comprehensive analysis of the challenges identified in the literature and the careful considerations accumulated through years of research, we formulated the following goals to address the research question:

- Conduct a comprehensive survey of the current research to review recent progress and the state of knowledge on this topic.
- Develop an automatic algorithm that accurately assesses the LV performance from echocardiographic image sequences.
- Evaluate the robustness of the proposed framework using a variety of metrics. Establish geometric parameters for LV segmentation evaluation and estimation of clinical parameters.
- Assess the performance of the investigated algorithms using a large publicly available dataset.
- Evaluate the generalizability of the proposed framework by testing it on an external and independent echocardiography dataset.

We designed these goals to address the research question effectively and contribute to advancing knowledge in the context of LV performance assessment from echocardiographic image sequences.

The main contributions of this work can be summarized as follows:

- Development of two frameworks based on attention mechanism and transfer learning concepts. We designed these approaches to address the research goals and improve the assessment of LV performance from echocardiographic image sequences.
- Analysis of the performance of the developed algorithms using the CAMUS dataset, a large public dataset proposed by Leclerc et al. [18]. We evaluated the algorithms using geometrical and clinical parameters, demonstrating their effectiveness compared to existing works in the field.
- Achieving excellent performance results which surpass the performance of previous approaches. These findings highlight the effectiveness of the proposed algorithms and their potential for enhancing clinical practice in the field of echocardiography.

- Collection of a private echocardiography dataset to further evaluate the clinical generalizability. This additional dataset allowed for a more comprehensive assessment and validation of the algorithms' performance.

- Demonstration of the generalizability of the proposed framework based on transfer learning. By leveraging knowledge and representations learned from the public dataset, the algorithms showcased their capability to adapt and perform well on the private dataset, indicating their potential applicability to diverse clinical scenarios.

These contributions advance the assessment of LV performance from echocardiographic image sequences by introducing novel algorithms, validating their performance on public and private datasets, and highlighting their potential for improving clinical decision-making and patient care.

During this doctoral study, we published the following research papers to disseminate the findings and contributions of the research.

- Hafida Belfilali, Mohammed Messadi, Abdelhafid Bessaid, and Amine Abbou. *Analysis of ultrasound image sequences for the assessment of left ventricle performance*. In JD-GBM'2019, Tlemcen, Algeria, 13 june 2019.

- Hafida Belfilali, Frédéric Bousefsaf, and Mahammed Messadi. *Impact of attention mechanism on u-net architecture for the left ventricle segmentation*. In 2022 International Conference on Technology Innovations for Healthcare (ICTIH), pages 01–04, 2022.

- Hafida Belfilali, Frédéric Bousefsaf, and Mahammed Messadi. *Left ventricle analysis in echocardiographic images using transfer learning*. Physical and Engineering Sciences in Medicine, pages 1–16, 2022.

Thesis structure

The manuscript comprises five chapters. We give an overview of each chapter below:

- **Chapter 1:** provides the necessary clinical background of this thesis. It covers the anatomy and physiology of the heart, emphasizing the role of the LV cavity. This chapter also explains the ultrasound modality, describing the formation of echocardiographic images and their main properties. Additionally, it clarifies the assessment of cardiac function.

- **Chapter 2:** focuses on the technical background of the methods studied in the thesis. It discusses conventional techniques used for image preprocessing and provides an overview of segmentation, including its definition, types, and evaluation metrics. The chapter introduces the concept of ANN and CNN, followed by an overview of popular CNN segmentation models.

- **Chapter 3:** presents a comprehensive review of the existing methods in the literature for LV segmentation and function assessment in echocardiographic images. It categorizes the previous work into three main categories: conventional methods, shallow learning-based methods, and deep learning-based methods. The chapter highlights the state-of-the-art approaches in the field.

- **Chapter 4:** focuses on the segmentation of the LV cavity in 2D echocardiography. It introduces a segmentation method based on the attention mechanism. The chapter describes the experimental setup and presents the results of image preprocessing and segmentation. The proposed technique is evaluated and discussed to demonstrate its effectiveness.

- **Chapter 5:** forms the core of the work, proposing a framework based on transfer learning for echocardiographic image analysis. The chapter also discusses the generalizability of the suggested framework on a different dataset presented in detail. The experimental setup, findings, and discussion of the results are all given in this chapter.

Globally, these chapters provide a comprehensive overview of the clinical and technical background, literature review, and the methods proposed for LV segmentation and assessment. They demonstrate the contribution and novelty of the research conducted in the thesis.

The conclusion section of the manuscript serves as a concise summary of the works conducted in the thesis. It highlights the findings and conclusions derived from the research. Additionally, it offers multiple perspectives on the study, suggesting potential directions for future research. The conclusion section acts as a wrap-up, providing a comprehensive overview of the thesis and leaving the reader with a clear understanding of the contributions and the significance of the conducted research.

Research Consortium

The thesis is co-directed by Prof. Mahammed MESSADI and Assoc. Prof. Frédéric BOUSEFSAF. This collaborative supervision is the result of a partnership between the

GBM laboratory (Laboratoire de Génie Biomedical/Tlemcen University/Algeria)[1] directed by Prof. Abdelghani DJEBBARI and the LCOMS laboratory (Laboratoire de Conception, Optimisation et Modélisation des Systèmes/Lorraine University/France)[2] directed by Prof. Imed KACEM. The research focuses on the development of an intelligent system for the analysis of echocardiographic image sequences and the assessment of left ventricular cavity performance. This collaborative effort combines the expertise and resources of the two laboratories. It is worth mentioning that this work has received support from the Eiffel Excellence Scholarship provided by the French Government during the final year of the thesis.

[1] https://gbm.univ-tlemcen.dz
[2] https://lcoms.univ-lorraine.fr

Chapter 1

Clinical Background

1.1 Introduction

Cardiac imaging is a widely used technique for diagnosing and assessing the prognosis of patients. It can provide an established method for analyzing cardiac function, aiding cardiologists in diagnosing cardiac diseases and determining appropriate treatment. Numerous medical imaging modalities are available to evaluate heart function. With advancing technology, new techniques have been developed in the literature to assess cardiac conditions. These include cardiac Magnetic Resonance Imaging (MRI), cardiac Computed Tomography (CT), and ultrasound (Echocardiography).

Echocardiography offers several advantages over other diagnostic imaging modalities. It is a cost-effective option with superior temporal resolution, eliminating the need for ionizing radiation exposure in patients. Moreover, it is a portable technique. In addition to these benefits, echocardiography generates real-time images suitable for dynamic testing. Consequently, cardiologists rely on 2D echocardiographic images to evaluate cardiac function by estimating essential clinical parameters. However, ultrasound images may contain artifacts and suffer from lower quality. Hence, automated methods of processing images can address these challenges, minimize manual annotation, and enhance intra- and inter-observer variability.

This chapter will begin with a brief overview of cardiology, followed by a description of the ultrasound technique in cardiology. We will also explain how the echocardiography technique forms the images. Next, we will present the characteristics and types of echocardiographic images. Lastly, we will explain the assessment of cardiac function.

1.2 Overview of cardiology

1.2.1 Anatomy of the heart

The heart is an organ located between the lungs, in the middle of the chest. Figure 1.1 displays the anatomy of the heart. It is composed of the endocardium (LV_{Endo}), myocardium (LV_{Myo}), and epicardium (LV_{Epi}), three tissue layers. The inner contour of the heart (endocardium) connects to the heart muscles (myocardium). The epicardium is a connective tissue that provides a protection layer. The heart has two parts: left and right. Each of these parts comprises two chambers. The upper and lower chambers are the atria and ventricles, respectively. A muscular wall known as the septum separates the left and right ventricles and atria.The heart has four cavities: Left Ventricle (LV), Right Ventricle (RV), Left Atrium (LA) and Right Atrium (RA). The mitral valve separates the LV and LA, whereas the tricuspid valve the RV and RA. The largest and most powerful chamber in the heart is the LV. The walls of this chamber are thicker than those comprising the RV, which makes the function of the LV powerful as a pump [24].

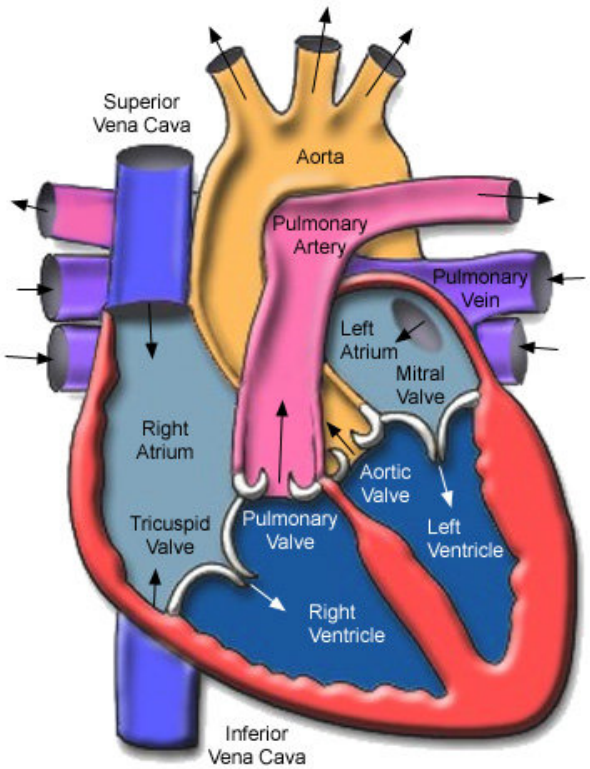


Figure 1.1: An illustration of the anatomy of the heart [1].

1.2.2 Cardiac cycle

The cardiac cycle is a rhythmic process wherein the atria and ventricles undergo alternating periods (contraction and relaxation). This process can effectively pump blood throughout the body. The RV and LV play distinct roles in this process. The RV pumps deoxygenated blood to the lungs through the pulmonary artery, while the LV pumps oxygenated blood to the entire body via the aorta. The RA and LA receive blood from the body and lungs before transmitting it to the ventricles. The cardiac cycle consists of two fundamental phases triggered by electrical impulses: systole and diastole. Systole represents the contraction period during which the ventricles contract forcefully to eject blood into the arteries. Diastole, on the other hand, is the relaxation phase when the ventricles refill with blood.

Figure 1.2 illustrates these two events initiated by electrical impulses. Various signals, including ventricular volume, electrocardiogram, and phonocardiogram, exhibit variations corresponding to different cardiac events. The cardiac cycle encompasses a series of electrical and mechanical events occurring with each heartbeat. It describes the coordinated sequence of events involving the contraction and relaxation of the heart. For an average heart rate of 75 beats per minute, an entire cardiac cycle duration is approximately 0.8 seconds [25].

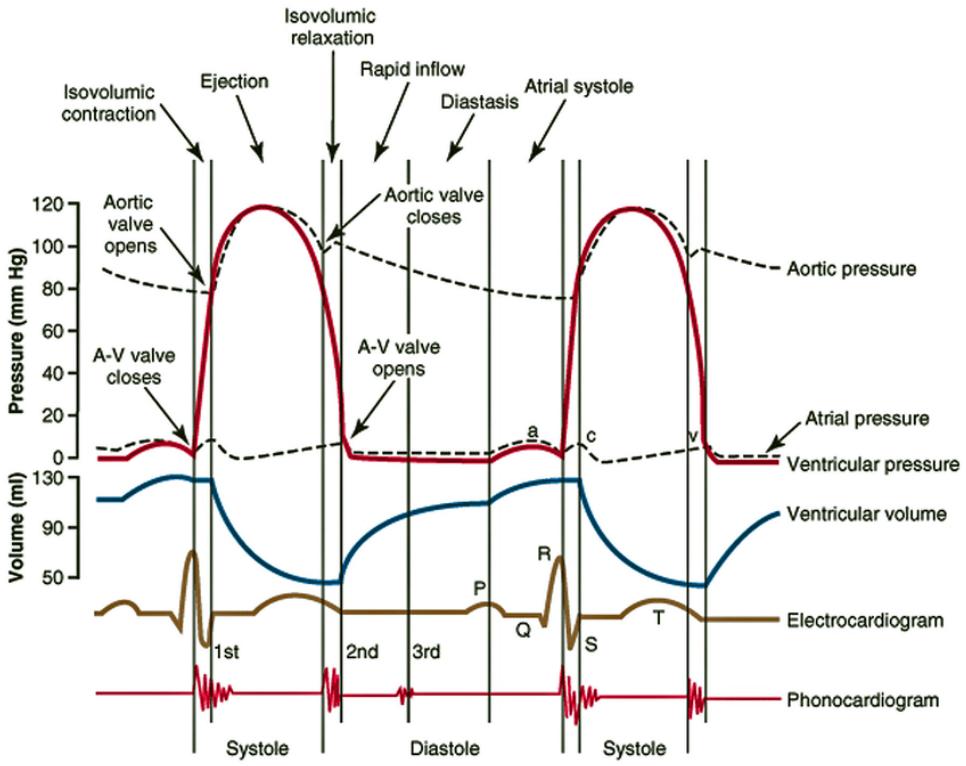


Figure 1.2: A Wiggers diagram illustrating events and details of the cardiac cycle [2].

1.2.3 Role of the left ventricle

The LV plays a vital role in the cardiovascular system. The thickness of the ventricular wall varies, being thickest near the base of the heart and gradually thinning to approximately 1-2 mm at the apex [26]. The concave shape of the LV is crucial in supporting its primary function, Which is ensuring adequate blood flow to other organ systems. The volume of blood pumped out by the heart per unit of time is the Cardiac Output (CO) calculated according to equation (1.1).

$$CO = HR \times SV \tag{1.1}$$

Heart Rate (HR) is the average number of heartbeats per minute, frequently measured in beats per minute (bpm). The amount of blood evacuated during a single ventricular contraction is known as the Stroke Volume (SV), calculated as the difference between the Left Ventricular End Diastolic Volume (LV_{EDV}) and the Left Ventricular End Systolic Volume (LV_{ESV}). The LV_{EDV} is the amount of blood just before the systole begins, and the LV_{ESV} is the quantity of blood remaining in the ventricle after the heart has contracted.

$$SV = LV_{EDV} – LV_{ESV} \tag{1.2}$$

In clinical practice, cardiologists don't directly measure the CO parameter. Therefore, it is common to use the EF to indicate the contractility of the heart. The EF is a metric that assesses the ability of the heart to pump oxygen-rich blood to the body, specifically, the Left Ventricular Ejection Fraction (LV_{EF}). LV_{EF} quantifies the percentage of blood ejected from the LV during systole. It is determined using equation (1.3).

$$LV_{EF} = \frac{SV}{LV_{EDV}} \tag{1.3}$$

For the proper management of patients with cardiovascular disease, a precise assessment of (LV_{EF}) is crucial. According to the American College of Cardiology, the following classification is utilized in clinical settings as follows [27]: normal (LV_{EF}) ranges are comprised between 50% to 70%, (LV_{EF}) above 70% denotes hyperdynamic, (LV_{EF}) from 40% to 49% indicates mild dysfunction, (LV_{EF}) from 30% to 39% presents moderate dysfunction, and (LV_{EF}) less than 30% designates severe dysfunction.

The LV connects nearly all organ systems by effectively pumping oxygenated blood throughout the body. Due to this close relationship, any decrease in left ventricular function can give rise to a wide range of potential issues. When left ventricular failure occurs, the heart exerts more effort to push oxygen-rich blood from the lungs to the LA, passing through the LV and subsequently throughout the body. This increased workload

places significant strain on the heart. Therefore, cardiologists pay particular attention to the function of the LV, as it is crucial for overall cardiac performance and the delivery of oxygenated blood to the body's organs and tissues.

1.3 Echocardiography: ultrasound in cardiology

Cardiac imaging encompasses various non-invasive imaging techniques to capture images of the heart and its surrounding structures. Cardiac imaging helps in diagnosing heart disease and assessing cardiac function. Some cardiac imaging methods include cardiac MRI, cardiac CT, and echocardiography. Cardiac MRI offers non-invasive imaging of biological tissues, including the heart, with excellent resolution and deep penetration. It can also manipulate the contrast of visualized structures using various mechanisms. However, it comes with a higher cost and longer imaging time. To obtain high-quality MRI images, patients must remain still for several seconds, often requiring breath-holding or apnea. Cardiac CT relies on X-ray technology and is particularly useful in diagnosing coronary artery disease. However, one significant drawback of cardiac CT is the radiation dose associated with X-ray exposure, which carries a potential risk of radiation-induced cancer [28]. Both cardiac MRI and cardiac CT imaging techniques are not real-time modalities, meaning they do not provide immediate, dynamic imaging. These limitations restrict their utility in some clinical settings.

Echocardiography, also known as an echocardiogram or cardiac echo, is the most widely used imaging modality for imaging the heart and assessing left ventricular function [29]. It is a diagnostic tool that does not involve ionizing radiation, making it safe for patients. Echocardiography offers several advantages, including being non-invasive, portable, cost-effective, easy to use, and real-time imaging. These multiple benefits make echocardiography the primary imaging modality recommended for measuring various clinical indices and effectively identifying heart dysfunction. However, it is necessary to note that echocardiographic images suffer from low image quality, making the interpretation and processing of echocardiographic images challenging. Nevertheless, despite its limitations, echocardiography remains a valuable and widely utilized imaging technique in cardiology due to its accessibility, real-time imaging capabilities, and ability to provide crucial diagnostic information without ionizing radiation.

1.3.1 echocardiographic exam

Cardiologists or specifically trained professionals known as echocardiographers experienced in conducting and interpreting echocardiographic exams often perform these tests.

The technique propagates ultrasound waves through biological tissues, including muscles and blood vessels. It enables the acquisition of real-time images that depict the acoustic properties of the examined structures.

Ultrasound waves used in echocardiography are pressure waves with frequencies above 20 kHz, making them higher than the audible range for humans. In echocardiography, frequencies ranging from 2 to 20 MHz can assess the location and severity of tissue damage. These frequencies are also suitable for evaluating the size, shape, motion, function, and performance of the heart and its valves. By utilizing this range of frequencies, echocardiography provides valuable information about the structure and function of the heart, assisting in the diagnosis and management of various cardiac conditions.

Cardiologists employ various echocardiogram types, including Transthoracic Echocardiogram (TTE), transesophageal echocardiogram, and stress echocardiogram. The choice of each echocardiogram depends on the specific evaluated cardiac condition. The TTE is the most commonly used form of echocardiogram in clinical practice. It is a non-invasive procedure performed externally on the chest. During a TTE, an echocardiographer applies a gel to the patient's chest and uses a handheld transducer (probe) to scan the heart. The transducer emits ultrasound waves and captures the echoes to create real-time images of the heart's structure and function. Figure 1.3 illustrates the procedure of echocardiography, specifically a TTE examination. The transesophageal echocardiogram involves inserting a specialized transducer into the esophagus to obtain detailed images of the heart structure that are not easily visible from the chest wall. The stress echocardiogram combines echocardiography with physical exercise or the administration of medications to evaluate the heart's function under stress conditions. The choice of the appropriate echocardiogram depends on the specific clinical indications and the information needed to assess and diagnose the cardiac condition.

The TTE relies on an ultrasound probe with transducers equipped with one or more piezoelectric elements. These elements generate ultrasound waves when excited by an electrical signal. During a TTE, the ultrasound probe directs the waves toward the heart and its surrounding structures. These waves interact with the different tissue structures, including the heart chambers, walls, and valves. The waves are scattered, reflected, and attenuated as they encounter the various interfaces within the tissues. A portion of the ultrasound wave is reflected towards the probe as echoes. The ultrasound probe receives these echoes and converts them into electrical signals. These signals are then processed and displayed on a monitor as moving images, allowing visualization in real-time of the heart's chambers, walls, and valves. It's important to note that the same transducer transmits the ultrasound waves and receives the reflected echoes. The transducer emits the ultrasound waves through small vibrations of the piezoelectric elements, and then it

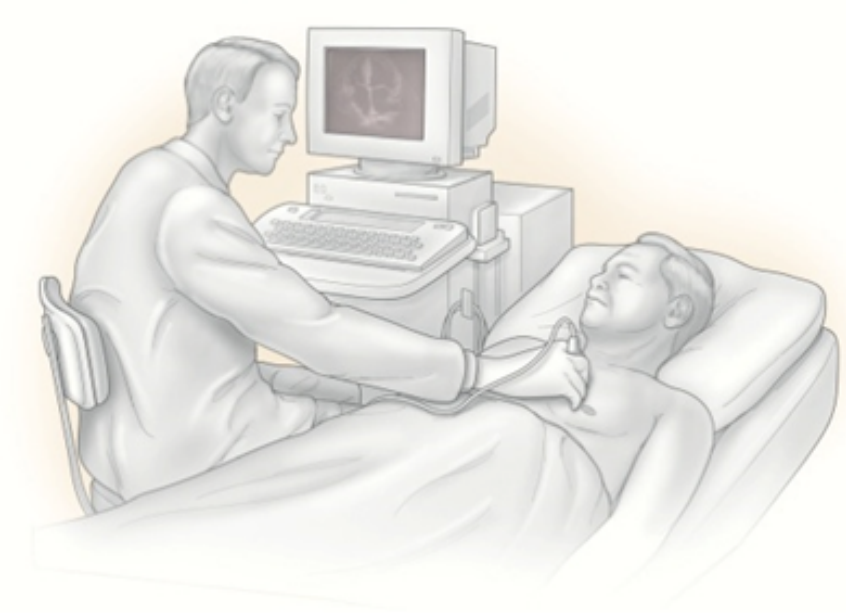


Figure 1.3: The process of echocardiography examination [3].

detects the echoes produced from the interfaces between tissues with different acoustic impedance. Figure 1.4 provides a visual representation of the emission of ultrasound waves from the probe's piezoelectric elements and the reception of the reflected echoes generated from the interfaces between tissues with different acoustic properties.

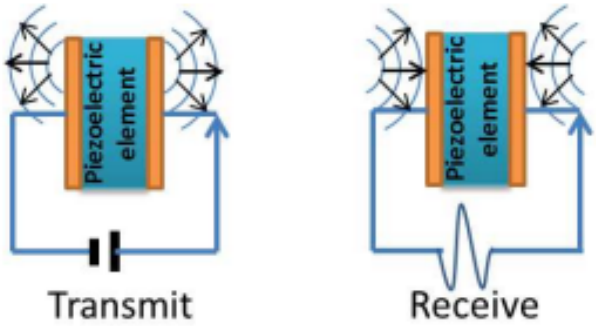
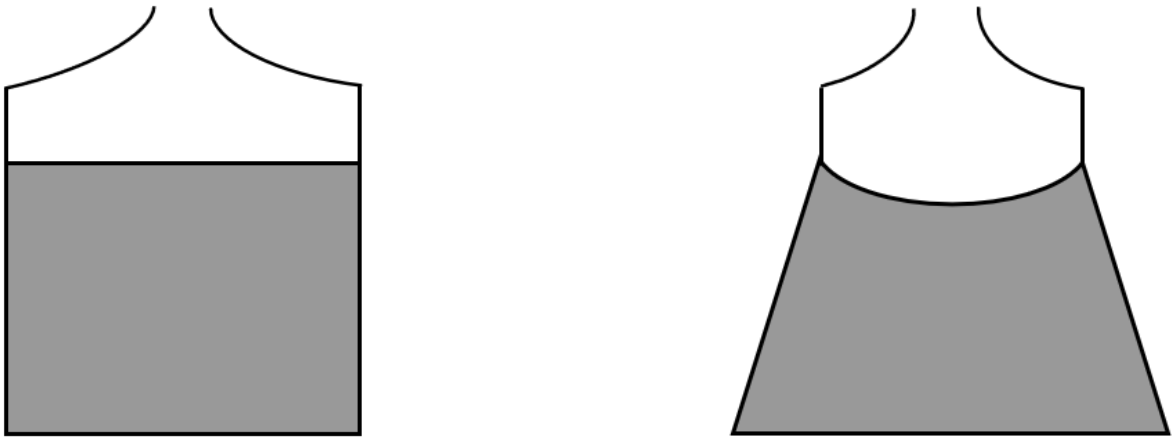


Figure 1.4: Illustration of the emission and reception performed using piezoelectric materials [4]

.

There are two main types of transducers: linear and sectorial (curvilinear). For instance, the linear transducer is used for thyroid or breast cancer detection or carotid artery imaging, while the sectorial transducer is utilized for cardiac or fetal imaging. Figure 1.5 illustrates the two types of probes.

In the echocardiographic exam, echocardiographers use curvilinear probes [30]. Access to the heart is difficult because of its form. The use of a curvilinear probe allows a sectorial



(a) **(b)**

Figure 1.5: Acquisition geometry for the ultrasound transducer. (a) linear. (b) sectorial

acquisition of the heart. Hence, it makes imaging of a large zone in-depth possible. The manipulation of the transducer by applying some movement such as rotation and tilting enables to obtain of multiple images in different views.

1.3.2 Interaction of the ultrasound wave with biological tissues

When an ultrasonic pulse propagates through soft tissues, such as muscles and blood vessels, it undergoes various phenomena that can alter the characteristics of the ultrasound waves. Figure 1.6 depicts these interactions. Here is a brief description of each interaction:

1.3.2.1 Reflection

Reflection is an interaction that occurs when an ultrasound beam encounters an interface between two tissues with different acoustic properties. When this happens, a portion of the ultrasound beam is reflected towards the ultrasound probe, while the remaining portion propagates through the tissue. The amount of reflection occurring at an interface depends on several factors, including the angle of incidence and the acoustic impedance of the involved tissues.

- Angle of incidence: the angle at which an incoming ultrasound beam strikes a tissue interface or boundary. It is the angle between the direction of the incident beam and a line perpendicular to the interface surface. This angle is a crucial factor in determining the behavior of the reflected ultrasound wave. It affects the amount of the reflected wave. Proper positioning of the transducer and careful control of the

angle of incidence help optimize the reflection of ultrasound waves and improve the accuracy and diagnostic value of the imaging results. When the incident ultrasound beam strikes the interface at a 90°angle (normal incidence), the reflected wave will be directed along the same path as the incident beam. This kind of incidence results in maximum reflection of the ultrasound wave. On the other hand, when the incident beam strikes the interface at an angle less than 90°(oblique incidence), the reflected wave is returned at an inclination equal to the angle of incidence [31]. This phenomenon is known as specular reflection.

- Acoustic impedance: the ultrasound wave travels through different mediums, and each medium has a specific acoustic impedance, given by the formula (1.4). Acoustic impedance is a parameter in ultrasound imaging that influences the behavior of ultrasound waves at tissue interfaces. When an ultrasound wave encounters a boundary between two tissues with different acoustic impedance, a portion of the wave is reflected while another is transmitted through the medium. The magnitude of reflection and transmission depends on the difference in acoustic impedance between the two tissues. The acoustic impedance allows us to predict and understand how ultrasound waves interact with and propagate through those tissues.

$$Z = \rho c \tag{1.4}$$

Where Z is the acoustic impedance, ρ is the density of the medium, and c is the speed of the wave.

The tissue's density affects the acoustic impedance. When the tissue density of the two mediums differs, their acoustic impedance will change significantly. Thus, the degree of reflection will depend on how much the acoustic impedance changes.

1.3.2.2 Scattering

Scattering is a phenomenon that occurs when an ultrasound wave interacts with small particles within a medium. When the size of the particles is smaller than the wavelength of the ultrasound wave, scattering can occur. These structures scatter the ultrasound waves in various directions, leading to a diffused pattern of reflected waves. The scattering contributes to the speckle pattern observed in ultrasound images.

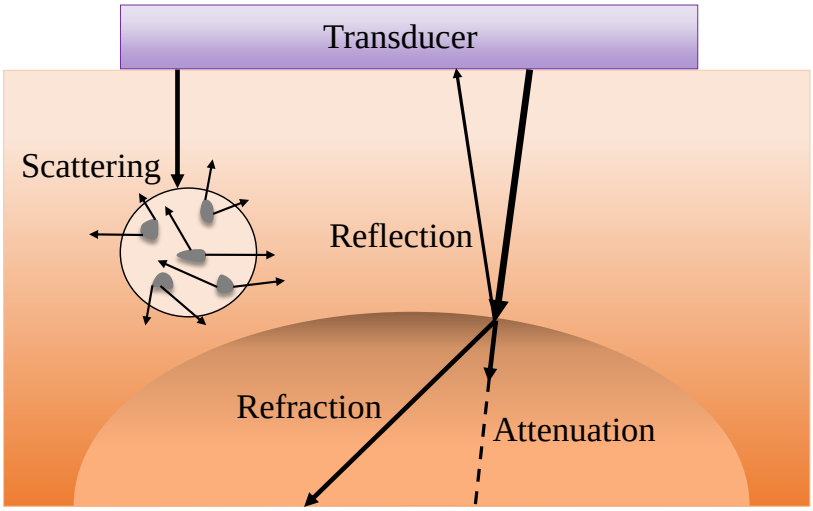


Figure 1.6: Interactions between the sound waves emitted by the transducer and soft tissues.

1.3.2.3 Refraction

Refraction happens when an ultrasound beam encounters a reflective surface with an oblique angle of incidence and interfaces with a different medium of various acoustic impedance. When the ultrasound beam reaches the interface, a portion of this beam is reflected toward the transducer. However, the remaining part of the ultrasound beam is transmitted through the interface into the second medium. The refraction phenomenon can lead to artifacts called refraction artifacts in ultrasound images. These artifacts can cause improper positioning and improper brightness of echoes displayed in clinical ultrasound [32]. Refraction artifacts occur due to the bending of the ultrasound beam as it passes through regions with different propagation speeds. This bending can cause distortions and misalignment of the imaged structures, affecting the accuracy and interpretation of the ultrasound image.

1.3.2.4 Attenuation

Attenuation in ultrasound refers to the reduction in the amplitude of ultrasound waves as they propagate through soft tissues [33]. As the ultrasound beam travels deeper into the tissue, its energy is progressively absorbed, resulting in a decrease in amplitude. Absorption is the primary mechanism contributing to attenuation, where the sound energy is converted into heat as it interacts with the tissue. Different tissues have varying absorption characteristics. The higher absorption occurs in tissues with higher density or higher attenuation coefficients. The attenuation of an ultrasound wave is proportional to its frequency. It increases with the increase of the ultrasound frequency. The attenuation phenomenon limits the depth of penetration of ultrasound waves and affects the overall

image quality.

1.3.3 Ultrasound image formation

1.3.3.1 Beamforming

As previously mentioned, the transducers are equipped with several piezoelectric elements with transmission and reception modes to build an ultrasound image. The principle of a beamformer is based on the time-delay law to concentrate the energy emitted in a specific area of the medium. Applying time delays to the various probe elements creates aligned focused beams and orients the ultrasound wavefront, as illustrated in Figure 1.7. A beam is directed and concentrated toward a location of interest using a beamformer to transfer signals to piezoelectric elements with unique time delays. The same physical system of delays is also applied to the received echoes to convert a returned echo into an electrical signal (radio-frequency echo signal).

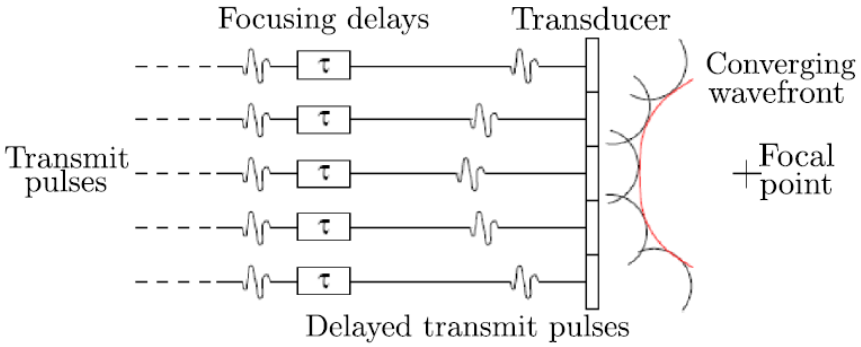


Figure 1.7: The principle of beamformer in transmission (the same principle in reception) [5].

The ultrasonic beam is swept through a conical sector to cover the intended field from a fixed probe location to form an image. In conventional echocardiography, an initial wave is focused along a line within the medium by adjusting the emission delay law. The returning signals reconstruct a line of the image through dynamic focusing during the reception. Then, a second wave focuses on the adjacent line. In the same way, the focus-receive operation has to be repeated on each line of the image. Scanning the beam over several lines allows for reconstructing the entire heart image. As depicted in Figure 1.8, the heart is imaged in real-time with approximately 50-100 frames per second.

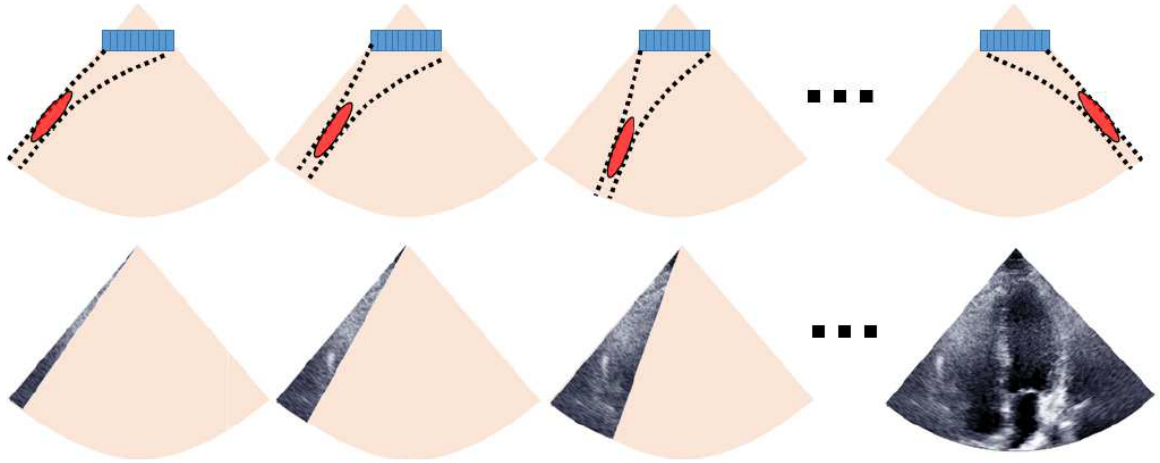


Figure 1.8: The basic idea behind image creation in traditional ultrasound imaging, as seen in echocardiography [6].

1.4 Echocardiographic images

1.4.1 Characteristics of echocardiographic images

The quality of ultrasound images varies based on the acquisition conditions. In cardiac ultrasound imaging, image quality primarily relies on resolution and the presence of artifacts in the visual representation. Below, we provide a brief description of each of these elements.

1.4.1.1 Spatial resolution

Spatial resolution is determined primarily by the transducer. It includes axial resolution and lateral resolution [34]. These resolutions of the acquisition system have a relevant role in the quality of an ultrasound system.

- Axial resolution: refers to the ability to distinguish between two successive echoes in the direction of propagation. The axial resolution of the 2D ultrasound image depends essentially on the ultrasound frequency. Higher frequencies result in shorter wavelengths, leading to improved axial resolution.

- Lateral resolution: refers to the minimal distance between two waves separated in a plane perpendicular to the ultrasound beam. The lateral resolution of the 2D ultrasound image depends on the size (thickness) of the ultrasound beam. Smaller beam thicknesses correspond to better lateral resolution. Large probes and higher frequencies typically yield a superior lateral resolution.

1.4.1.2 Temporal resolution

Temporal resolution is the time between the beginning of one frame and the beginning of the next. It refers to how well an ultrasound system can discriminate between successive image frames over time. The frame rate of the system is the primary determinant of temporal resolution. Overall, a higher frame rate means that it is easy to discriminate quick motions (such as the motion of heart chambers or valves during the cardiac cycle), providing better temporal resolution.

1.4.1.3 Contrast resolution

Contrast resolution indicates the capacity to distinguish differences between echo amplitudes of adjoining structures. It refers to the ability to discern between dark and light areas and spot amplitude differences. The contrast resolution and the signal-to-noise ratio are closely related. The contrast depends on some factors, for example, the echogenicity of the patient. It can be improved at various stages in the imaging process using contrast agents such as injections of microscopic air bubbles or specialized post-processing, such as histogram normalization.

1.4.1.4 Artifacts and speckle

Artifacts frequently appear in echocardiographic images. An artifact refers to information in an ultrasound image that results in an inaccurate representation of the proper anatomy [35]. In ultrasound imaging, artifacts usually have the shape of duplicated, missing, incorrectly placed, or warped structures. The incorrect interpretation of an artifact as a legitimate detection may result in unnecessary interventions, such as medical care and surgery. There are several types of artifacts, especially in TTE, and they are due to different reasons. For instance, multiple reflections can cause reverberation, shadows, and mirror artifacts. Furthermore, the behavior of some reflectors leads to refraction artifacts. Moreover, there are also other artifacts related to the equipment.

The speckle is a multiplicative noise corresponding to the granular aspect of the ultrasound image. For areas of homogeneous tissue, the speckle causes the signal to be inhomogeneous. It typically tends to lower image quality and contrast, thus impacting the diagnostic precision. The speckle results from the scattering interaction previously presented. It occurs when multiple emitted waves come from the surface of tiny structures within a specific tissue. This phenomenon depends on the resolution cell of the echocardiographic system.

1.4.2 Modes of transthoracic echocardiogram image display

Three modes of TTE acquisition exist in clinical echocardiography: M-mode, B-mode, and Doppler imaging. Figure 1.9 displays a typical example of each one.

1.4.2.1 M-mode imaging

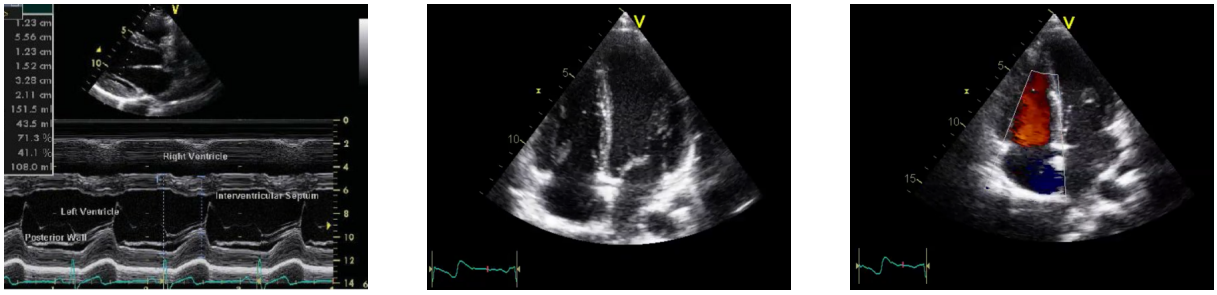
Motion-mode (M-mode) is frequently employed in echocardiography to observe the motions of the heart's walls and valves. It is a unidirectional examination that displays the variation of the position of the echoes in a single line according to time. M-mode imaging offers excellent temporal resolution, allowing for the efficient and convenient recording of multiple cardiac cycles [36]. However, a challenge in this mode is aligning the probe perpendicular to the object of interest, as improper alignment can lead to inaccurate measurements. Misalignment can lead to inexact measurements and interpretations of the recorded data.

1.4.2.2 B-mode imaging

Brightness-mode (B-mode) is the most frequently utilized mode in clinical practice. It involves employing successively oriented ultrasound beams to scan a section of the heart. It shows the ultrasound reflection as a gray-scale image composed of bright dots representing the ultrasound echoes. Less echogenic structures, like the blood, turn black, whereas powerful reflectors, such as the muscles or valves, appear bright. The time delay controls the vertical position of a point from pulse transmission to the returned echo. However, the horizontal position determines the location of the receiving transducer [37]. Due to recent developments in echocardiographic probes, B-mode can be either 2D or 3D to allow the imaging of depth, width (in 2D), and thickness (in 3D). Compared to the M-mode echocardiographic images, it is simpler to recognize the anatomic relationship between distinct structures. In this thesis, we use B-mode images.

1.4.2.3 Doppler imaging

Doppler imaging is an echocardiographic technique that allows the study of blood flow in real-time by the Doppler effect. It is the most widely used method to measure flow speed and direction and enables hemodynamic examination of the heart [38]. Additionally, it allows the estimation of intra-cardiac pressures from blood velocity; and the measurement of regular blood flows like diastolic and systolic outputs. When the ultrasound beam passes through the heart cavities or vessels, the blood's figurative elements (the transmitters) send back echoes. The reflected echo has a longer wavelength if it is farther



(a) **(b)** **(c)**

Figure 1.9: Some examples of different transthoracic echocardiogram modes. (a) M-Mode imaging. (b) B-Mode (2D imaging). (c) Doppler imaging.

away from the transducer and a shorter wavelength when it is nearer to the probe. An audio signal or a velocity curve are two possible exam outputs. The blood flows can also be color-coded using Doppler echocardiography. Typically, the positive flow far from the probe is shown in blue, while the positive flow near the transducer appears in blue.

1.4.3 Standard ultrasound views of the heart

In B-mode imaging, several preferred acquisition planes (views). A cardiac window is a region where the sonographer positions the transducer. The placement of the probe towards the precise cardiac window and its correct manipulation is critical to achieving the required view [38]. The most common standard views are:

- Parasternal long axis view (Figure 1.10a): This view is obtained by placing the transducer along the left parasternal border and angling it towards the heart. It provides a longitudinal section of the heart, allowing visualization of the LV, mitral valve, aortic valve, and a portion of the RV.
- Parasternal short axis view (Figure 1.10b): In this view, the transducer is placed at the left parasternal border and rotated 90 degrees to obtain a cross-sectional heart image. It provides information about the ventricular size, wall thickness, and the relationship between the ventricles and the valves.
- Apical 4 Chamber (A4C) (Figure 1.10c): this view is taken from the apex, where the four cavities are visible if a wide enough angle of view. The two ventricles areas are situated at the top of the image, while the atria are at the bottom, and the septum is in the central part of this view. This view is convenient for assessing chamber sizes, wall motion abnormalities, and valvular function.

- Apical 2 Chamber (A2C) view (Figure 1.10d): allows visualizing of two cavities with a smaller angle: the LV at the top of the image and the LA at the bottom. It provides a longitudinal section of these two chambers. The A2C views are obtained by tilting the transducer slightly from the apical position.

The A4C and A2C views are essential for evaluating left ventricular function and assessing clinical parameters such as LV_{EDV}, LV_{ESV}, and LV_{EF}. These measurements provide valuable information about the size, contractility, and pumping efficiency of the LV. In this study, focusing on the A4C and A2C views for assessing LV performance is appropriate, as these views provide relevant measurements for evaluating LV function and calculating the clinical parameters.

1.5 Cardiac function assessment

Medical imaging, a routine task for cardiac diagnostics, enables the non-invasive evaluation of many clinical indices. In cardiology, medical imaging aims to assess cardiac function accurately. Echocardiography is a widely available imaging technique for quantifying cardiac function. A single cardiac function analysis can give important diagnostic and prognostic information such as disease risk prediction, patient care, and therapy. B-mode imaging allows for determining heart function from cardiac volumes with the ability to visualize the heart. Specifically, it offers the opportunity to estimate local and global indices, i.e., LV volumes and EF.

2D echocardiography remains the most commonly used method for LV volume estimation in clinical practice. Accurate segmentation of the LV endocardium in 2D echocardiographic images at ED and ES frames is crucial for reliable volume calculations. Various image processing and segmentation techniques are employed to extract the LV endocardium from the 2D echocardiographic images, allowing for the accurate measurement of LV volumes and subsequent calculation of CO, SV, and LVEF. The most effective method for computing volumes is 3D echocardiography, which allows the visualization of the whole heart. However, The clinical applicability of 3D echocardiography is severely constrained by its low image quality [39].

There are different methods from which we can estimate the LV_{ED}, LV_{ES}, and LV_{EF} from the 2D echocardiography. Among them, we mention the following techniques:

- Modified Simpson's rule [40]: is also known as the biplane or disc summation method. The American Society of Echocardiography suggests using this technique to estimate the LV_{EF}. The area tracings of the LV cavity are necessary. According

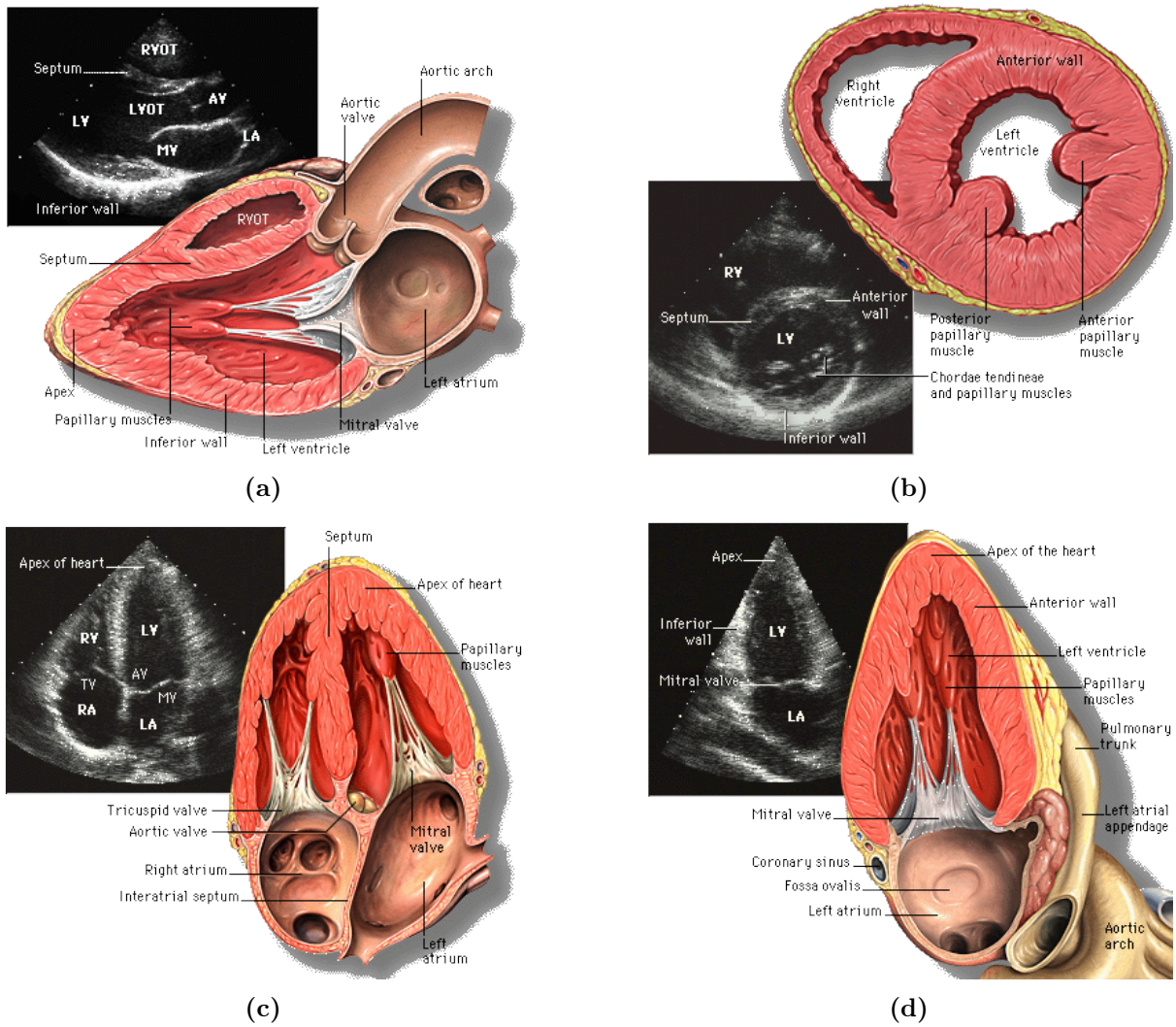


Figure 1.10: Standard ultrasound views of the heart. (a) Parasternal long axis view[a]. (b) Parasternal short axis view[b]. (c) Apical 4 Chamber view[c]. (d) Apical 2 Chamber view[d]

.

[a] https://commons.wikimedia.org/w/index.php?curid=21448310
[b] https://commons.wikimedia.org/wiki/File:LeftVentricleShortAxis.gif
[c] https://commons.wikimedia.org/wiki/File:Apical_4_chamber_view.png
[d] https://commons.wikimedia.org/wiki/File:Apical2Chamber.png

to this procedure, the LV_{Endo} must be traced in the A4C and A2C views in ED and ES. Eventually, these tracings separate the LV cavity into a specific number of disks (often 20). The final volume of the LV results from the summation of areas of the 20 cylinders or discs of equal heights.

- Teichholz method: is another technique used to estimate LV volumes and EF from 2D echocardiography. This method relies on the LV dimensions measurement in the parasternal long-axis view. This method calculates the LV_{EDV} and LV_{ESC} using the assumptions of a prolate ellipsoid shape for the LV and derives the LV_{EF} from these volume measurements. It relies on assumptions about LV shape and geometry that may not be accurate in certain cardiac conditions. Compared to Simpson's biplane method, it is considered less reliable for the LV volumes and EF estimation.

- Modified Quinones method: is based on linear measurements. This technique can be used in either M-mode or B-mode imaging. It uses a single quantification of the LV cavity in the mid-ventricle at both ED and ES. Because they depend on the assumption of a constant geometric LV form, such as a prolate ellipsoid, which does not apply in different cardiac diseases, volume calculations obtained from linear measurements may be unreliable. This method is no longer recommended for estimating LV volumes and EF in a clinical context.

- Automated speckle tracking echocardiography: This technique uses advanced image processing algorithms to track the movement of speckles (small acoustic markers) within the myocardium throughout the cardiac cycle. Analysis of the deformation of these speckles enables measurement of left ventricular deformation parameters necessary to estimate left ventricular volumes and EF.

1.6 Conclusion

This chapter provided a comprehensive overview of the clinical background. It first highlighted the clinical aspect by presenting the anatomy of the heart, an explanation of the cardiac cycle, and the role of the LV chamber. Moreover, it introduced the application of the ultrasound imaging modality in cardiology. This section focused on the echocardiographic examination and the basic principle of this technique. Then, it emphasized the echocardiographic images by outlining their characteristics, the modes of imaging, and the standard views of the heart. Finally, we described the process for assessing cardiac function.

Chapter 2

Technical Background

2.1 Introduction

Image processing involves modifying the characteristics of an image to enhance its information for human interpretation and make it more suitable for perception by autonomous machines. In the field of healthcare, digital image processing plays a crucial role. It is a powerful tool that facilitates the analysis of medical images, providing accurate anatomical information essential for diagnosis aid and early detection.

Artificial Intelligence (AI) has been successfully applied in various pattern recognition applications. Early research on deep learning for medical image analysis has demonstrated superior results compared to traditional techniques. Deep learning-based methods, in particular, enable autonomous prediction without relying on predefined features. Deep learning techniques are extensively employed in cardiovascular image analysis [8]. They assist in the evaluation, diagnosis, and prognosis of cardiovascular diseases. These networks, specifically CNNs, extract high-level and low-level information from input images, enabling the detection of complex image structures.

In this chapter, we will first introduce some conventional methods commonly used in the initial step of medical image processing, known as image preprocessing. Subsequently, we will describe the segmentation process, different types of segmentation techniques, and the evaluation metrics utilized to assess segmentation results. Following that, we will define the ANNs. Finally, we will present CNN and discuss the general CNN architectures used for segmentation.

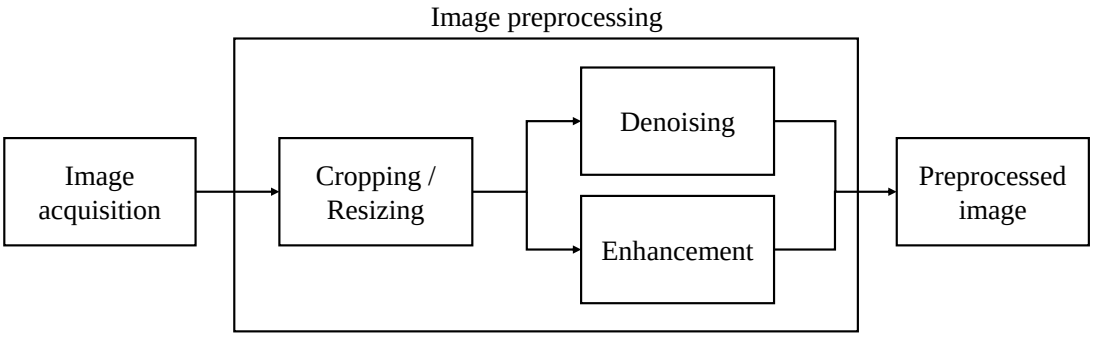


Figure 2.1: Flow diagram of image preprocessing.

2.2 Image preprocessing

Image preprocessing refers to a series of operations performed on image data to prepare and format it before being used in a specific process. In medical imaging, the main objectives of image preprocessing are to reduce acquisition artifacts and standardize images across a dataset. The specific preprocessing requirements depend on the imaging technique, the procedure used to acquire the data, and the intended workflow. Typical preprocessing steps include noise removal and image enhancement (as illustrated in Figure 2.1). These steps aim to improve the quality of the image and enhance specific features for better analysis and interpretation.

2.2.1 Image denoising

The purpose of denoising operations is to enhance image data by eliminating unwanted noise and distortions, such as speckle noise. Speckle artifacts manifest as a granular noise texture caused by the interference of wavefronts, and they degrade the quality of echocardiographic images. In many cases, reducing speckle noise can be accomplished through preprocessing filtering techniques. Some of these filters are listed below.

2.2.1.1 Gaussian filter

The Gaussian filter is a convolution operator. The following formula expresses the two-dimensional digital Gaussian filter:

$$G(x,y) = \frac{1}{2\pi\sigma^2} e^{-\left[\frac{x^2+y^2}{2\sigma^2}\right]} \tag{2.1}$$

Where σ^2 is the variance of the Gaussian filter and (x, y) are the pixel coordinates.

This filter replaces the value of a pixel with the mean value of the surrounding pixels, calculated based on a Gaussian distribution [41]. The filter works by convolving the image

with a Gaussian kernel. This filtering operation effectively blurs the image, reducing noise while preserving the overall structure and edges. It can also introduce distortions in an image that contain sharp variations in pixel brightness [42]. The extent of blurring or smoothing depends on the size of the Gaussian kernel or the standard deviation parameter of the Gaussian distribution.

2.2.1.2 Mean filter

The mean filter, also known as the average filter, enhances the pixel value in an image by replacing it with the average value of its grayscale neighborhood. It is a simple sliding-window filter that replaces the central pixel value of the kernel window with the average of all the pixel values within the window. The mean filter blurs and smooths images while reducing noise. However, it can also result in a loss of image detail by reducing variations in pixel intensities. Suppose that R_{xy} is the neighborhood window with size $m \times n$ and (x, y) is the point center. For the gray level of the pixel (x, y) in image K, the mean filter replaces the value $J(x, y)$ with the mean of the grayscale of the surrounding pixels.

$$J(x, y) = \frac{1}{m * n} \sum_{(u,v) \in R_{xy}} K(u, v) \tag{2.2}$$

2.2.1.3 Median filter

The median filter is a nonlinear filter commonly used for noise reduction in image processing. Unlike the mean filter, which uses the average value, the median filter replaces the gray level of each pixel with the median value of the gray levels within a specific neighborhood [43]. The principle of this filter is to arrange the pixel values within the neighborhood window in ascending order. The median value replaces the concerned pixel. Using the median instead of the average is less sensitive to extreme or outlier values. Moreover, it effectively reduces image noise and results in better preservation of edges and details of the image. The mathematical representation of the median filtered image $J(x, y)$ of the image $K(u, v)$ is as follows:

$$J(x, y) = \underset{(u,v) \in R_{xy}}{\text{median}} \{K(u, v)\} \tag{2.3}$$

2.2.1.4 Wiener filter

The Wiener filter is also known as the Least Mean Square Filter. The Wiener filter removes noise from each pixel in an image. It attempts to construct an image by applying a mean square error constraint between the denoised and the original image. Hence, the

mean square error is minimized as part of the Wiener filtering restoration process. The Wiener filter produces better results than linear filtering. However, This filter requires more computing time [44]. It analyzes data in the frequency domain but may fail to recover frequency components degraded by noise. Its mathematical formula is as follows:

$$f(u,v) = \left[\frac{H(u,v)^*}{H(u,v)^2 + \left[\frac{S_n(u,v)}{S_f(u,v)} \right]} \right] G(u,v) \tag{2.4}$$

Where $H(u,v)$: Degradation function, $H(u,v)^*$: Complex conjugate of $H(u,v)$, $G(u,v)$: Transform of the degraded image, $S_n(u,v)$: Power spectrum of the noise, and $S_f(u,v)$: Power spectrum of the original image.

2.2.2 Image enhancement

In image processing, image enhancement techniques play a crucial role in improving the quality of an image by emphasizing relevant information and suppressing irrelevant details [45]. Image enhancement can be used in specific applications by targeting different image features such as contrast, edges, and boundaries. The goal is to enhance the dynamic range of these selected features rather than increase the overall information richness of the image data. Image enhancement techniques often focus on highlighting details through increasing contrast and brightness. These techniques can be categorized into two main divisions: frequency domain and spatial domain. Frequency domain techniques operate on the frequency transform of the image, while spatial domain techniques operate directly on the pixel values. In recent years, neural network-based approaches have been applied for image enhancement. These approaches learn complex mappings between input and output images. Hence, they effectively enhance various aspects of an image.

2.2.2.1 Contrast stretching

Contrast stretching is a strategy used to enhance low-contrast images that may arise due to multiple factors, such as inadequate illumination, limited dynamic range in the imaging sensor, or incorrect lens aperture settings during image acquisition [46]. The goal of contrast stretching is to expand the pixel intensity values in the image to a desired range by applying a linear scaling function to the original pixel values. The contrast stretching process involves remapping or stretching the gray-level values of the image so that the histogram covers the entire range of possible values. This transformation effectively increases the difference in pixel intensities, resulting in an improved contrast. The equation below expresses the contrast stretching:

$$y = \begin{cases} \alpha x, & 0 \leq x < a \\ \beta(x-a) + y_a, & a \leq x < b \\ \gamma(x-b) + y_b, & b \leq x < L \end{cases} \tag{2.5}$$

Where x: is the input image, y: is the output stretched image, (α, β, and γ): the stretching constants, a and b: the lower and higher range, and y_a and y_b are defined as follows:

$$y_a = \alpha x \tag{2.6}$$

$$y_b = \beta(x-a) + y_a \tag{2.7}$$

2.2.2.2 Histogram equalization

Histogram equalization is another technique based on the image's histogram. It is a popular method used for contrast adjustment that provides a simple and effective way to enhance an image by redistributing its gray levels based on the probability distribution of the gray levels [47]. The main goal of histogram equalization is to achieve a more uniform distribution of pixel intensities. In traditional histogram equalization techniques, the histograms of all gray levels were equalized on average, resulting in a more balanced distribution of pixel values across the entire range. However, these techniques can sometimes overstretch the gray levels, leading to wide histogram boxes and potential loss of image details [45]. Histogram equalization adjusts the pixel intensities to be more uniformly distributed, enhancing the overall contrast of the image. This method can reveal more details and improve the visual appearance.

Suppose X=$\{$(X(i,j)$\}$ a discrete grayscale image of L discrete gray levels, represented as $\{X_0, X_1, X_2,...X_L\}$. For a given image X, the probability density function $P(X_a)$ is defined as:

$$P(X_k) = \frac{n_k}{n} \tag{2.8}$$

Where n_k is the number of occurrence of gray level X and n is the total number of pixels in the image. Additionally, let's determine in the equation (2.9) the cumulative distribution function C in accordance with $P(X_k)$.

$$C(x) = \sum_{j=0}^{k} P(X_j) \tag{2.9}$$

where x is X_k for $k = 0, 1, 2, ...L-1$ and $0 \leq C(X_k) \leq 1$

Histogram equalization uses the cumulative distribution function as a level transformation function to map the input image into the complete dynamic range (X_0, X_{L-1}). Equation (2.10) presents a transformation function $f(x)$ based on the cumulative distribution function. $f(x)$ indicates the intended histogram equalization output image.

$$f(x) = X_0 + (X_{L-1} - X_0)C(x) \tag{2.10}$$

2.2.2.3 Contrast Limited Adaptive Histogram Equalization (CLAHE)

As opposed to the traditional histogram equalization that uses a single histogram for the whole image, the adaptive histogram equalization method [48] computes many histograms. Each histogram corresponds to a different image's section. They are all used to redistribute the image's brightness values. A generalization of adaptive histogram equalization is the Contrast Limited Adaptive Histogram Equalization (CLAHE) [49]. It makes hidden features of the image more visible by balancing the distribution of the gray values.

2.2.2.4 morphological operations

Morphological operations are techniques used to eliminate imperfections in an image. Morphology is an image enhancement method based on mathematical set theory that removes noise while keeping relevant objects of interest. The main idea of these operations is to apply a structuring element to an input image and produce an output image of the same size. Morphology operates based on the shape and form of objects. The pixel values of the output image are obtained by comparing the corresponding pixel in the input image with its neighboring pixels. The fundamental morphological operations are erosion, dilation, opening, and closing. These operations remove small unwanted elements, fill gaps, smooth boundaries, and highlight or suppress certain image features depending on the specific operation used.

2.3 Image segmentation

For precise image analysis, robust image segmentation is a required step. It is a crucial component of computer vision technologies and algorithms. Image segmentation is the process of dividing the image into various parts. Hence, it locates image objects and boundaries [50]. It assigns labels to pixels in the image to classify them and distinguish between different elements. Three main types of segmentation tasks result from this

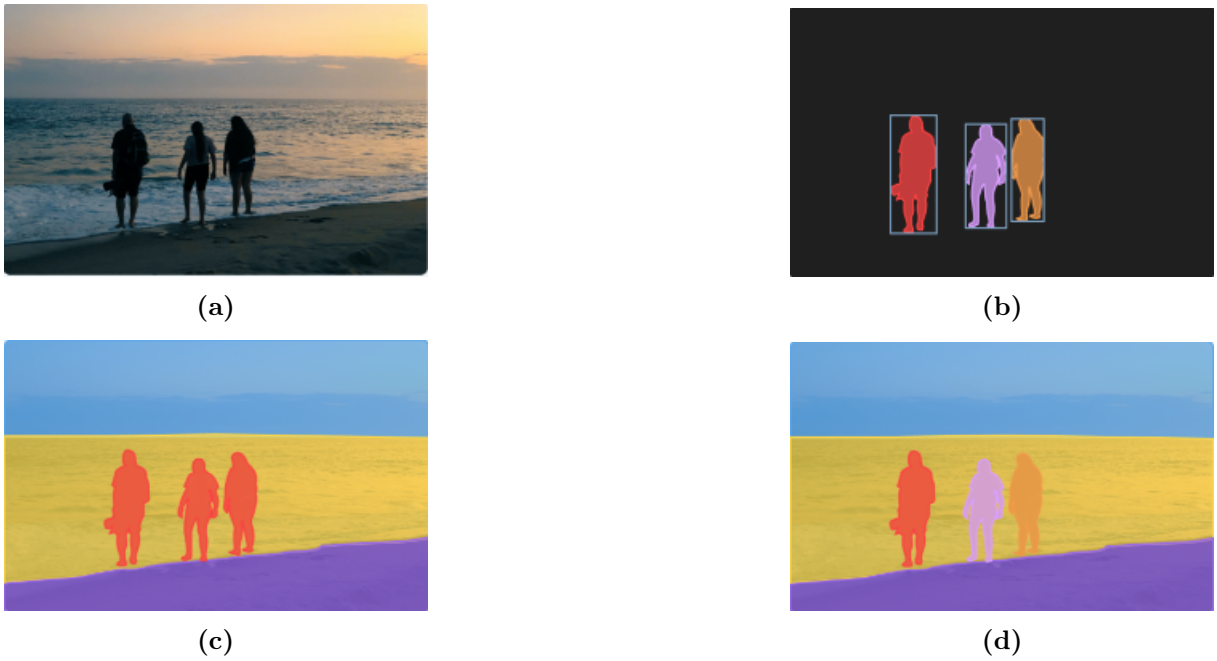


Figure 2.2: Image segmentation techniques [7]. (a) original image. (b) Instance segmentation (per-object mask and class label). (c) Semantic segmentation (per-pixel class labels). (d) Panoptic segmentation (per-pixel class+instance labels).

distinction. Figure 2.2 shows a typical example of each type of segmentation applied to a given image.

2.3.1 Types of image segmentation

- Instance segmentation (Figure 2.2b): assigns distinct labels for separate instances of objects belonging to the same class [51]. It classifies the pixels based on these instances of an object. Instead of knowing the region class, instance segmentation divides comparable or overlapping regions based on the boundaries of objects.

- Semantic segmentation (Figure 2.2c): classifies the pixels in an image based on the semantic classes. It distinguishes between object categories regardless of their particular instances. Every pixel in this model belongs to a single class. The segmentation model does not refer to any other context or data.

- Panoptic segmentation (Figure 2.2d): this is the more recent segmentation method that combines the typically distinct tasks of semantic segmentation (assign a class label to each pixel) and instance segmentation (detect and segment each object instance). In panoptic segmentation, the goal is to give a unique class label to each pixel in the image and indicate the class of the object and the instance it belongs

to. It predicts the identity of each object, separating each instance of each object in the image.

Semantic segmentation seems to be the method of choice for biomedical segmentation challenges. Since no object of the same type appears more than once. Especially in echocardiography, semantic segmentation is more suitable than the other segmentation type in this context. It assigns distinct labels to different cardiac structures. By labeling each pixel according to the object it belongs to, semantic segmentation enables the accurate delineation and localization of these structures. In this work, any pixel not belonging to the LV structure is assigned to the background class.

2.3.2 Medical image segmentation metrics

In medical image segmentation, several metrics can assess the performance of automatic segmentation approaches compared to manual ground truth annotations. These metrics help quantify the accuracy and quality of the segmentation results.

Many segmentation metrics are computed based on the confusion matrix for a binary segmentation task. All these indices are based on the cardinalities of the confusion matrix: True Positive (TP), False Positive (FP), True Negative (TN), and False Negative (TN). Some of these metrics for binary segmentation between the surface manually annotated (S_m) and the surface automatically segmented (S_a) are listed below:

- Dice Similarity Coefficient (DSC) [52]: is also known as Sørensen-Dice index. It is the most used metric. It calculates the overlap between S_m and S_a. DSC gives a value between 0 (poor segmentation) and 1 (perfect segmentation). This metric is calculated using the following formula:

$$DSC = 2 \times \frac{|S_m \cap S_a|}{|S_m| + |S_a|} = \frac{2TP}{2TP + FP + FN} \tag{2.11}$$

- Jaccard Coefficient (JC) [53]: is also known as Intersection Over Union (IOU). It evaluates the similarity and diversity of the segmented region. This coefficient is determined by dividing the intersection of two segmented areas by their union (See eq.(2.12)). The IoU metric penalizes under- and over-segmentation more than DSC [54].

$$JC = \frac{|S_m \cap S_a|}{|S_m \cup S_a|} = \frac{TP}{TP + FP + FN} \tag{2.12}$$

- Accuracy (Acc): is also known as Rand index or pixel accuracy. It presents the proportion of correct positive and negative predictions to all other predictions. In medical image segmentation, using accuracy metrics for evaluation is severely discouraged. Since medical images often contain a single object of interest with a small area of pixels, the accuracy metric always produces a high score because of the TN inclusion (pixels annotated as the background).

$$Acc = \frac{TP + TN}{TP + TN + FN + FP} \tag{2.13}$$

- Receiver Operating Characteristic (ROC): is a graph that shows the performance of a classification model at all classification thresholds. The performance of a model is measured through the true positive rate (TPR) against the false positive rate (FPR) where:

$$TPR = \frac{TP}{TP + FN} \text{ and } FPR = \frac{FP}{FP + TN} \tag{2.14}$$

  The Area Under the ROC Curve (AUC) [55] measures the entire two-dimensional area under the ROC curve from (0.0) to (1.1). AUC is also used to validate machine learning classifiers and binary segmentation methods. The AUC ranges in value from 0 to 1. A model whose predictions are wrong has an AUC of 0. However, a model that gives correct predictions has an AUC of 1.

In addition to the region overlap-based indices, other metrics based on the spatial distance provide a more accurate evaluation of contouring accuracy. They are frequently limited to the pixels of the contour manually delineated $C_m = u1, u2, ..., u_n$ and automatically segmented $C_a = v1, v2, ..., v_n$.

- Hausdorff Distance (HD): is a metric that measures the dissimilarity between two sets of points. In the case of image segmentation, it quantifies the difference between the segmented region and the ground truth region. It represents the maximum distance between a point in one set (e.g., the segmented region) and the closest point in the other set (e.g., the ground truth region). A smaller HD indicates a better alignment between the two regions, implying a more accurate segmentation. The following formula defines the HD:

$$HD = max(max_i\{d(u_i, C_a)\} + max_j\{d(v_j, C_m)\}) \tag{2.15}$$

  Where $d(u_i, C_a) = min_j \parallel v_j - u_i \parallel$ presents the minimum of the Euclidean distances.

- Mean Absolute Distance (MAD): is determined as the average of the absolute distances using the Euclidean distance d as used in the HD. It indicates the global disagreement between two contours. The MAD is the average error in segmentation, whereas the HD is the maximum error. The MAD is calculated as follows:

$$MAD = \frac{1}{2}\left(\frac{1}{n}\sum_{i=1}^{n} d(u_i, C_a) + \frac{1}{n}\sum_{j=1}^{n} d(v_j, C_m)\right) \tag{2.16}$$

  Where $d(u_i, C_a) = min_j \parallel v_j - u_i \parallel$ presents the minimum of the Euclidean distances.

2.4 Overview of neural networks

AI encompasses any action in which machines imitate the cognitive behaviors of humans. Machines process large amounts of data to identify patterns and perform activities similar to those performed by humans. Machine learning (ML) is a branch of AI that aims to teach computers how to learn without being programmed for specific tasks [56]. ML aims to develop algorithms capable of both learning from and predicting data. In ML, a subset of interconnected neurons known as ANNs can solve complex problems. ANNs are sometimes called neural networks. These models simulate the electrical activity of the brain and nervous system [57]. They consist of connected processing elements called artificial neurons or nodes. These neurons are organized in layers or vectors, where the output of one layer serves as input to the next layer or even other layers. Figure 2.3 illustrates a typical example of an ANN. Generally, ANN includes an input layer, one or more hidden layers, and an output layer. Each neuron connects to every neuron in the subsequent layer or only to specific ones. Neurons are associated with weights and thresholds. When the output of a neuron exceeds the defined threshold value, it becomes activated and sends data to the next layer. This process of transmitting data from the input to the output layers is known as forward propagation. The mathematical formula for the forward propagation of each neuron is:

$$y_j = f\left(b_j + \sum_{i=1}^{i=n} w_{ij} x_i\right) \tag{2.17}$$

Where y_j is the output of the j neuron, w_{ij} the weight of each input variable, b_j the bias term, n the number of input variables, and x_i the input variable. The final output is obtained next by applying the selected activation function f to each node output. Different activation functions exist, such as Heaviside, piecewise linear, sigmoid, and Gaussian [58]. Take for example, the sigmoid function. It is defined as follows:

$$f(x) = \frac{1}{1 + e^{(-x)}} \tag{2.18}$$

As the neural network consists of many neurons that are arranged into successive layers, the output layer's value is calculated as follows:

$$Y_k = c_k + \sum_{i=1}^{nX} a_{jk} X_j \tag{2.19}$$

Where Y_k is the normalized output variables, c_k the bias value of k output, $a_j k$ the weight of each neuron, and nX the number of neurons. The number of neurons in the input and output layers depends on the data. It is fixed by the users in the hidden layer. Generally, it varies from one to the whole number of input variables in the hidden layer.

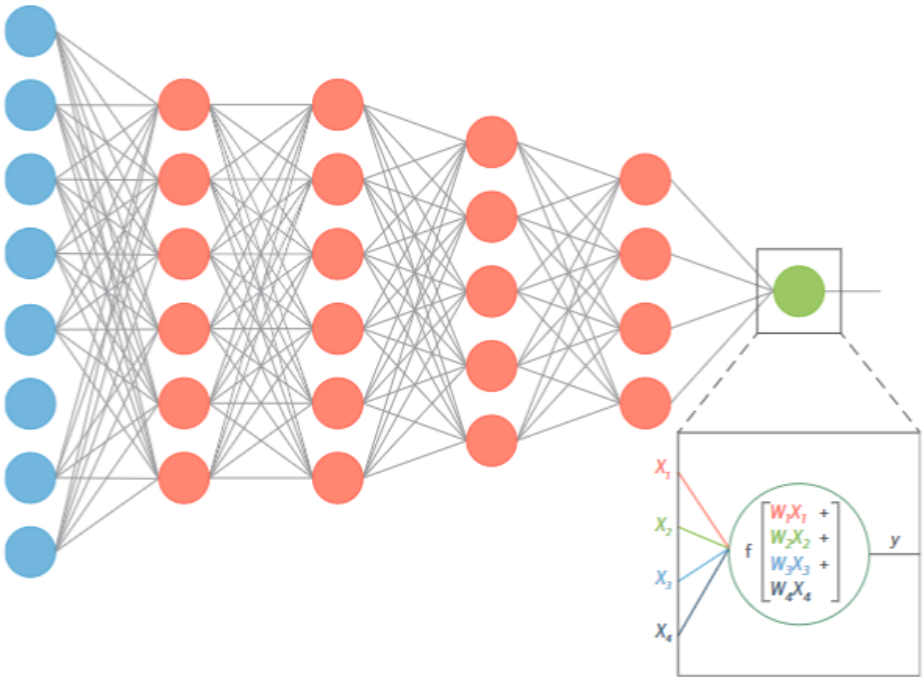


Figure 2.3: Artificial neural network with 5 layers [8]. The input and output layers are shown in blue and green, respectively, while the hidden layers are shown in red.

In the beginning, all the weights of the networks have a random assignment. The values are propagated through the layers when the network is activated. The propagation is forward from the input to the output layer that makes a prediction, passing by the hidden layers. The network has to be trained on data to make more accurate predictions. Since the model knows the observed value in the training set, it is possible to calculate the error obtained in the predicted output. A chosen loss function (cost function) computes the error during training. The network learns from this error by updating weights and

biases. This process is called backpropagation. The goal for backtracking is to propagate the loss backward and utilize an appropriate optimizer technique such as Stochastic Gradient Descent (SGD), Adaptive Moment Estimation (Adam) [59], and Root Mean Square Propagation (RMS-Prop). The goal of using the optimizers is to change and adjust the weights and biases of the neural network to lessen the error. The following equations designate the SGD optimizer:

$$w_{ij}^{l} \leftarrow w_{ij}^{l} - \varepsilon * \frac{\partial C}{\partial w_{ij}^{l}} \tag{2.20}$$

$$b_{j}^{l} \leftarrow b_{j}^{l} - \varepsilon * \frac{\partial C}{\partial b_{j}^{l}} \tag{2.21}$$

where ε is the learning rate, C the loss function, w_{ij}^{l} the weight of each input variable at layer l, and b_{j}^{l} the bias value of j neuron at layer l.

The principle of the training phase of a neural network is to repeat the forward propagation from input to output and backward propagation several times until the error gets below a predefined threshold ($C(w, b)$ converges for all weights w_{ij}^{l} and biases b_{j}^{l}). However, the model prediction in the testing phase for new data is obtained by performing a forward pass.

A model may become more complex to capture all the relationships that training data inherently expresses. The complexity of an ANN model may have two unfavorable effects. The first is that running a sophisticated model could take a long time. Second, a large model can perform very well on training data because it can memorize all the underlying relationships in trained data; However, it performs poorly on validation data because it cannot generalize to unseen data. This phenomenon has the consequence of making the model learn the statistical noise present in the training data, which has the negative effect of causing the model to poor performance when tested on new data (validation or test datasets). Hence, the generalization error increases because of the overfitting problem. Overfitting occurs When a machine learning model predicts outcomes accurately for training data but not for unseen data. There are multiple regularization techniques to give more reliability and stability and lower the risks of overfitting, such as L1 regularization, L2 regularization, Dropout [60], and Early stopping.

2.5 Convolutional neural networks

ANNs are rarely applied to images due to the higher dimensionality of 2D data. The high number of features yields dense connectivity is computationally expensive for ANNs.

CNNs are well-known deep learning architectures that are analogous to traditional ANNs [61]. Compared to ANNs, CNNs are better at handling numerical arrays with two or more dimensions. Less learning parameters are significantly needed for CNNs to perform better on tasks involving images. LeCun et al. [62] built for the first time a CNN for handwritten zip code recognition in 1989 and used the word (convolution), which is the original version of LeNet [63]. CNNs identify the visual patterns directly from pixel images.

2.5.1 Main components of CNN

CNNs (See Figure 2.4) consist of multiple stacked layers such as a convolutional layer, batch normalization, activation layer, pooling layer, and Fully Connected layer (FC):

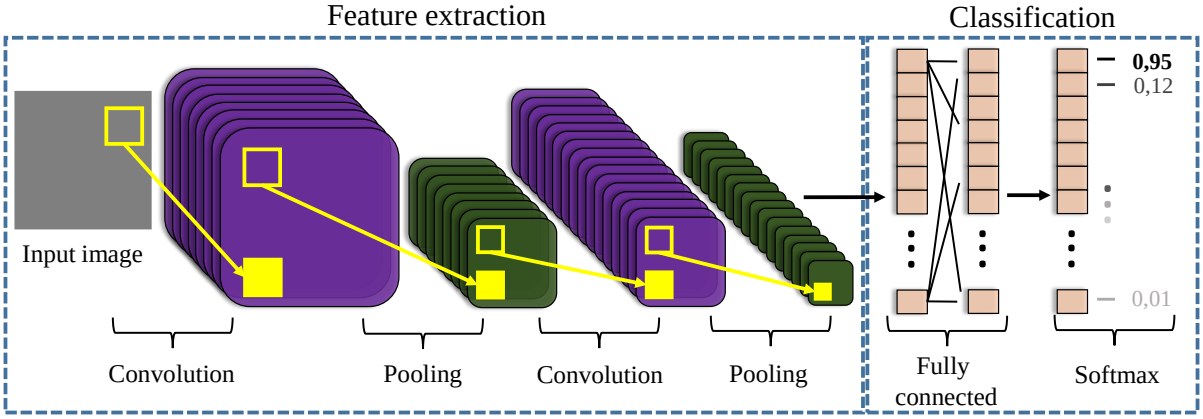


Figure 2.4: Illustration of convolutional neural network.

- Convolutional layer: is the core component used in CNN. It extracts the features from the input image using fixed-size matrices called kernels or filters. Kernels are moving over the input image to compute an element-wise multiplication between the values in the kernel matrix and the input image value. The following formula represents the convolution operation for an input image x:

$$X[u,v] = \sum_{i=0}^{m-1}\sum_{j=0}^{m-1}\sum_{c=1}^{C} w[i,j,c] * x[u+i, v+j] \tag{2.22}$$

  Where $[i,j]$ indicates a value of row i and column j of an array, w the weights of the filter of size $m \times m$, C the number of channels of the input image, and X the output layer (feature map).

  Figure 2.5 shows the convolution operation as an illustration. The size of the feature maps is determined by the number of filters (depth), the number of pixels that shift

over the original image (stride), and the process of adding zeros to the input image (padding).

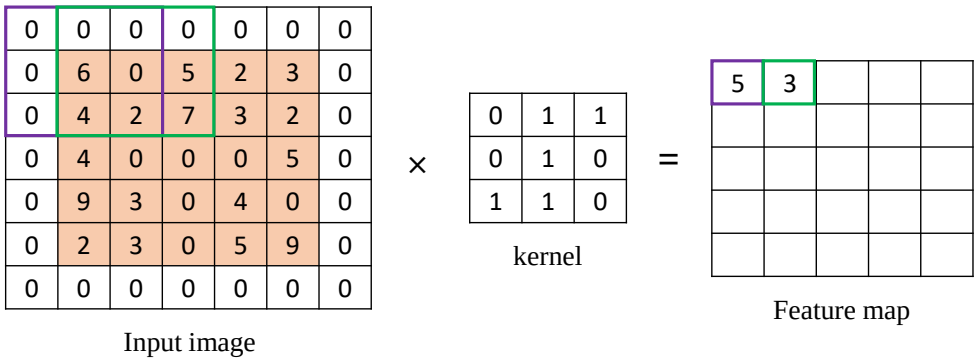


Figure 2.5: Typical example of a convolution operation.

- Batch normalization: has been introduced by Ioffe et al. [64]. The main idea behind this technique is to increase the stability and return the input of each layer to zero mean and constant standard deviation. It makes deep neural network training faster [65]. Batch normalization can act as a regularization technique [64].

- Activation layer: is an element-wise function applied to the convolved result. It adds nonlinearities to the CNN, which allows a multi-layer network to detect nonlinear features. Many types of activation functions are used in CNN, e.g., hyperbolic tangent activation function (tanh) [66], sigmoid, and Rectified Linear Unit (ReLU) [67]. The ReLU activation function is the most used because it learns faster than the other activation functions [68]. Furthermore, it performs better and avoids the problem of vanishing gradient. The weights of a network with this vanishing gradient issue cannot be updated, which decreases the network's performance. When the input is a negative value, the ReLU activation function returns 0, but when it is a positive value, it returns x as follows:

$$f(x) = \begin{cases} 0, & x < 0 \\ x, & x \geq 0 \end{cases} \tag{2.23}$$

- Pooling layer: is commonly placed between two convolutional layers. It reduces the resolution of the feature maps. The spatial size of each feature map is decreased independently by two typical pooling strategies: max-pooling [69] and average pooling [70] (See Figure 2.6). Pooling operation reduces the size of the region covered by the filter to a single value (maximum or average, respectively). The objective of

adding pooling layers in a CNN is to avoid the problem of overfitting by reducing the number of parameters.

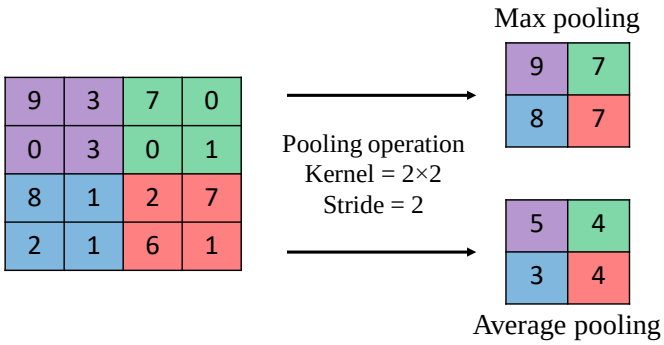


Figure 2.6: Typical example of pooling operations.

- Fully connected layer: is placed after the convolution and pooling layers, which result in high-level features of the original image. FC uses these features to classify the input image based on the training dataset into distinctive classes. Every neuron in the FC layer is connected to each activation of the previous layer to produce global semantic information. The classification probability is the last step in a CNN. It uses a Softmax activation function after the FC layer to generate normalized values in the range [0,1].

2.5.2 Standard segmentation architectures

Many researchers have explored deep learning-based segmentation algorithms, particularly in medical imaging. The remarkable success of CNNs in solving classification problems has motivated the application of these networks to image semantic segmentation. A design in this domain is the Fully-Convolutional Network (FCN) (Long et al., 2015), which can generate a segmentation map for an entire input image with a single forward pass. However, deep CNNs for semantic segmentation can significantly increase the number of learning parameters. In general, semantic segmentation involves pixel-level classification of an input image. Researchers proposed organizing the layers into an encoder-decoder architecture to address the challenge of increased parameter count while maintaining high segmentation performance. These architectures represent improved versions of FCNs. The encoder-decoder networks utilize an encoder network to downsample the input and extract image features, while the decoder network performs upsampling to restore the extracted features to the original image size. This process leads to the final segmentation image. Our focus is on U-shaped segmentation architectures in this thesis. The following sections will discuss these networks.

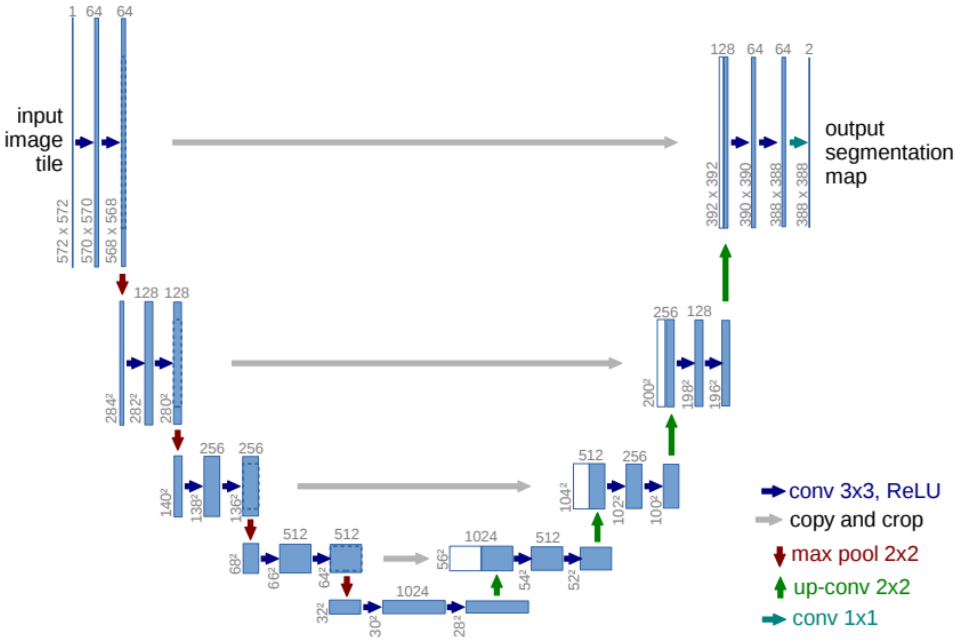


Figure 2.7: U-Net architecture [9].

2.5.2.1 U-Net

U-Net architecture is an encoder-decoder network invented by Ronneberger et al. [9] in 2015 (See Figure 2.7). It has been used to solve a variety of medical image segmentation issues successfully. The original architecture of U-Net is composed of four encoder layers that create a contraction path, a bottom layer that acts as a bottleneck, and four decoder layers that set up the expansion path. Each encoder layer is connected to its corresponding decoder block by a skip connection to merge high-resolution local features with low-resolution global features. The components of this architecture produce a U-like structure to the network. The encoder path consists of two 3×3 convolution layers with the same number of filters (ranging from 64 to 512 at each level). ReLU activation function follows each convolution layer. Then a 2×2 max-pooling operation is applied to reduce the spatial dimensions. The bottleneck consists of two layers of 3×3 convolutions, with a ReLU activation function after each convolution (with 1024 kernels). The decoder part starts with a 2×2 transposed convolution layer to re-cover the resolution of the input image. After that, two 3×3 convolution layers are applied, followed by ReLU activation functions. The decoder output goes through a 1×1 convolution layer with a sigmoid or softmax activation function. Many advancements in U-Net architecture have been proposed since the first version released by Ronneberger et al. [9].

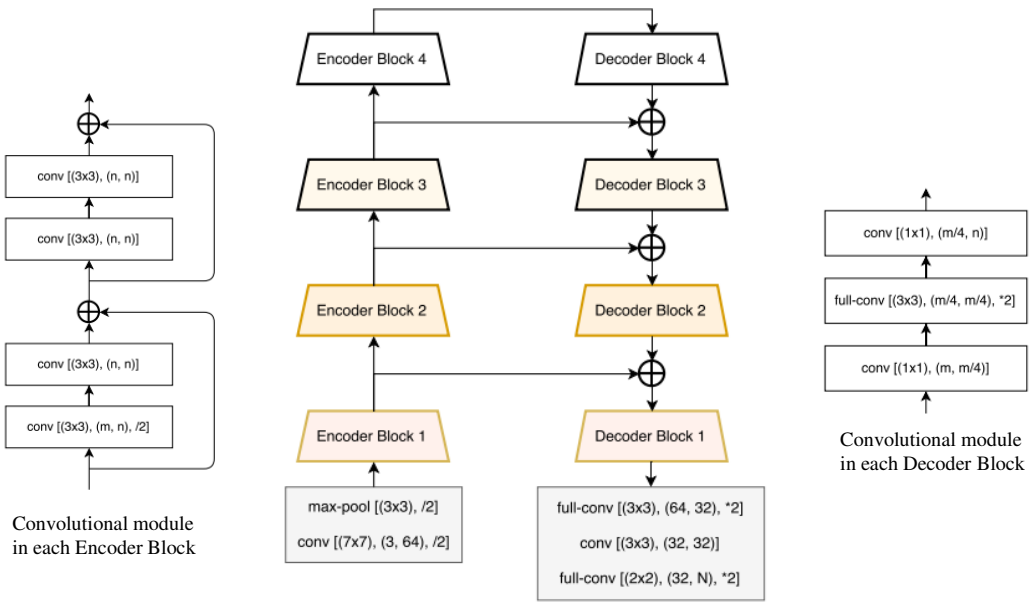


Figure 2.8: Overview of the LinkNet framework [10]. [Left]: Convolutional module in each Encoder Block of LinkNet architecture. [center]: LinkNet architecture. [right]: Convolutional module in each Decoder Block of LinkNet architecture.

2.5.2.2 LinkNet

LinkNet is an encoder-decoder architecture proposed by Chaurasia et al. [10] (illustrated in Figure 2.8). This network connects the input of each encoder block to the output of its corresponding decoder block. After each downsampling block, LinkNet tries to share the information learned by the encoder with the decoder. By doing this, the decoder has few parameters and can retrieve the lost information used by the decoder and its upsampling operations. The low number of parameters improves the time needed for segmentation in this network. LinkNet uses an initial block that contains a convolution layer with a kernel size of 7×7 and a stride of 2, followed by a max-pooling layer of window size 2×2 and stride of 2. The next portion of the encoder is a series of residual blocks (equivalent to those included in ResNet-18 [71]). After that, each decoder block uses the method of full convolution [72]. This technique is also implemented later in the final decoder block to reduce the feature maps from 64 to 32, followed by 2D convolution. The network contains at the end a full convolution as a classifier with a 2×2 kernel size.

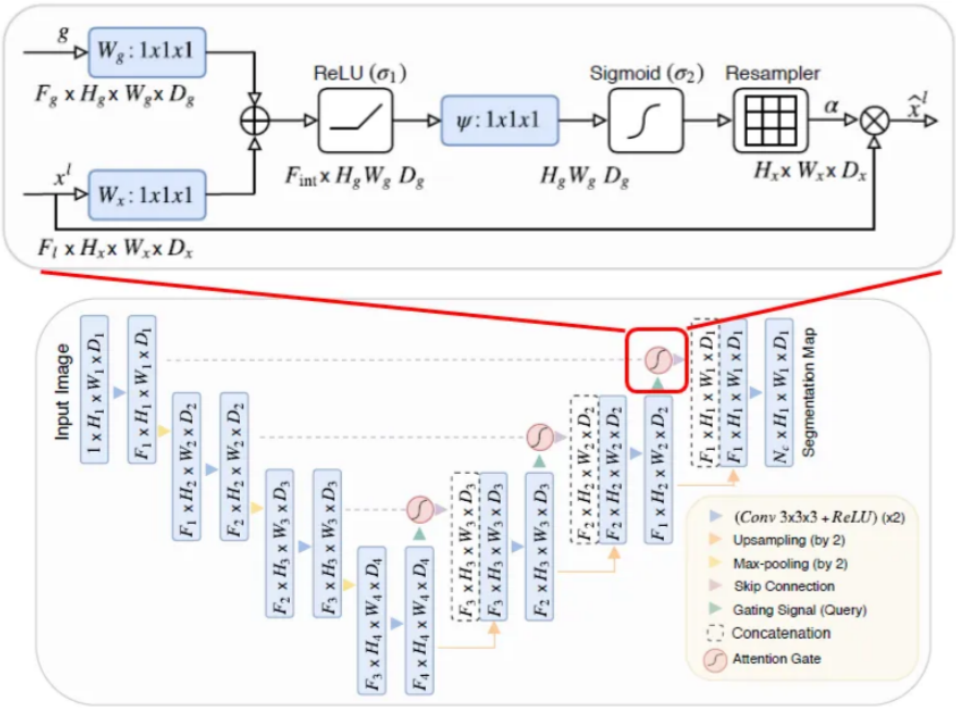


Figure 2.9: Attention U-Net architecture [11]. [Top]: Attention gate. [Bottom] Attention U-Net.

2.5.2.3 Attention U-Net

This network was proposed by Oktay et al. in 2018 [11] (presented in Figure 2.9). The main idea of the attention U-Net is to integrate attention gates in the skip connections of the original U-Net. These components give the network the ability to focus on relevant features. They allow paying attention to the object of interest without using explicit localization modules. Each attention gate has two input vectors: x and g. The vector g has better performance and smaller dimensions because it comes from a coarser scale than the input query signal x. First, the two input feature maps pass into individual $1\times1\times1$ convolution, after which they are added together. The ReLU activation function is applied to the summation output. Second, another $1\times1\times1$ convolution is carried out but with sigmoid as the activation function. A resampling phase (trilinear interpolation) is applied after the convolution operations to change the size of the attention coefficients to be element-multiplied with the low-level query signal. Finally, the upsampled feature maps at the lower level of the decoder are concatenated with the features received from the attention gates. Due to the attention mechanism, attention U-Net focuses on learning the target structures, even if they are small and with varying shapes.

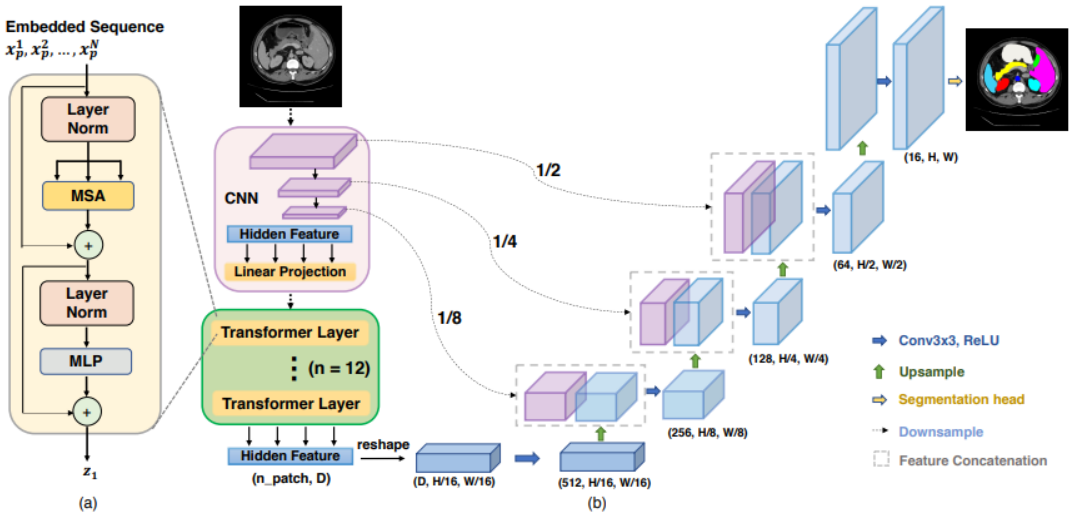


Figure 2.10: Overview of TransUNet architecture. (a) Graphic of the Transformer layer. (b) TransUNet architecture.

2.5.2.4 TransUNet

TransUNet [73] includes an encoder and a decoder for encoding and decoding image data to create a segmented image (See Figure 2.10), similar to the architectures described above. TransUNet, in contrast to conventional U-Nets, learns both the high-resolution spatial information from CNNs and the global context information from Transformers by introducing self-attention mechanisms. It achieves this by using a hybrid CNN-Transformer architecture as an encoder. First, CNN is used as a feature extractor to generate a feature map for the input. The output of each level of the CNN (high-level feature maps) is concatenated with the corresponding decoder level. The feature map of the CNN is vectorized into a sequence of flattened 2D patches using trainable linear projection. Patch embedding is applied to 1×1 patches extracted from the CNN feature map instead of the raw image. After that, the embeddings are passed into 12 transformers. Each transformer layer consists of multi-head self-attention and multi-layer perceptron modules. The output of the transformer layers has the shape of (n_patch, D), where D represents the total length of the embedding. Reshaping is applied to obtain (D, $H/16$, $W/16$) shape for the upsampling operations. $H/16$ and $W/16$ denote a reduction of the heights and widths by a factor of 16 in the earlier encoding operations. On the other side, the decoding process introduces a cascaded upsampler. It consists of multiple upsampling stages to decode the hidden feature and produce the final segmentation mask. Each block includes a 2 × upsampling operator, a 3×3 convolution layer, and a ReLU layer successively. By cascading these upsampling blocks, the output mask reaches the full resolution (C, H, W) with C: the number of objective classes, H: the image height,

and W: the image width. TransUNet outperforms several other techniques, including CNN-based self-attention techniques.

2.6 Conclusion

This chapter has provided the technical background for this thesis. The first section introduced image preprocessing, covering various conventional methods commonly used in the literature. The following part focused on the segmentation step, including its definition, types, and evaluation metrics. Then, we presented an overview of ANNs, highlighting their components and functioning. Finally, the last section emphasized the significance of deep learning in the segmentation process, discussing the main parts of CNNs and introducing standard segmentation architectures based on CNNs.

Chapter 3

Literature review

3.1 Introduction

Currently, cardiovascular diseases are considered the first cause of death in the world. Therefore, the search for a system of analysis of echocardiographic images for a reliable diagnosis is attracting the attention of many researchers. The importance of automatic heart function quantification to early detect cardiovascular pathologies has increased with the development of digital imaging and computing power. Indeed, precise quantitative evaluation of heart anatomy and function can determine the most effective treatments.

There are many clinical applications used in the echocardiography imaging modality. EF estimation and LV structure quantification are two clinical applications used in 2D and 3D echocardiography [74]. For that, the localization and segmentation of the LV in ED and ES frames are crucial tasks. The manual delineation of the LV contour is complicated and presents many drawbacks. Accordingly, several techniques have been proposed in the literature to automate the assessment of LV performance in 2D echocardiography [75,76], which is the focus of this thesis, and 3D echocardiograms [77–80]. The suggested methods for 2D LV segmentation are various. Most of the studies present approaches for the LV_{Endo} segmentation because the delineation of this structure is enough to estimate the LV_{EF}.

This chapter will present the different methods existing in the literature to segment the LV structure to assess cardiac function in echocardiographic images. Three categories can classify these approaches: conventional, shallow learning, and deep learning-based methods. At the end of this chapter, we will present a summary of this chapter.

3.2 Overview of the methods used to segment the left ventricle and evaluate its function in 2D echocardiography.

Numerous studies have been conducted to automate the segmentation of the LV structure in echocardiographic images. The segmentation of the LV_{Endo} boundary was the primary emphasis of the reported techniques. The following is an overview of the principal methods used to segment the LV and assess the cardiac function with a quick explanation of each approach. Figure 3.1 summarizes the set of the different works existing in the literature.

3.2.1 Conventional methods

Conventional methods encompass all traditional techniques that are not part of the machine learning domain. Previous surveys have extensively covered these methods used in medical B-mode images, as seen in the study published by Noble et al. [75].

3.2.1.1 Image-based models

Thresholding models Thresholding methods are pixel-based techniques for image segmentation. They are the most straightforward method because they rely on the intensity differences between background and object pixels. As a result, thresholding segmentation distinguishes areas of an image that correspond to the objects of interest. Ohyama et al. [81] proposed a method for LV_{Endo} detection in echocardiograms based on ternary double thresholding operation. Sigit et al. [82] also used thresholding with morphological operations and triangle equation to detect and reconstruct the LV border. The main disadvantage of thresholding is that we take pixel intensities and ignore any correlations between them. There is no assurance that the thresholding technique will identify contiguous pixels.

Edge-based techniques An edge signifies a transition in an image from one object or surface to another. It defines the border between two regions having distinct features. Edge-based segmentation is a method for processing images that detect the edges of different objects. Edge-based segmentation algorithms locate edges based on differences in contrast, texture, color, and saturation. Anwar et al. [83] used adaptive thresholding with a Canny edge detector to segment the heart wall cavity in two and four-chamber images. However, using just Canny edge detection cannot obtain the contour of the heart cavity. It was necessary to use the region area and collinear to enhance the results of edge identi-

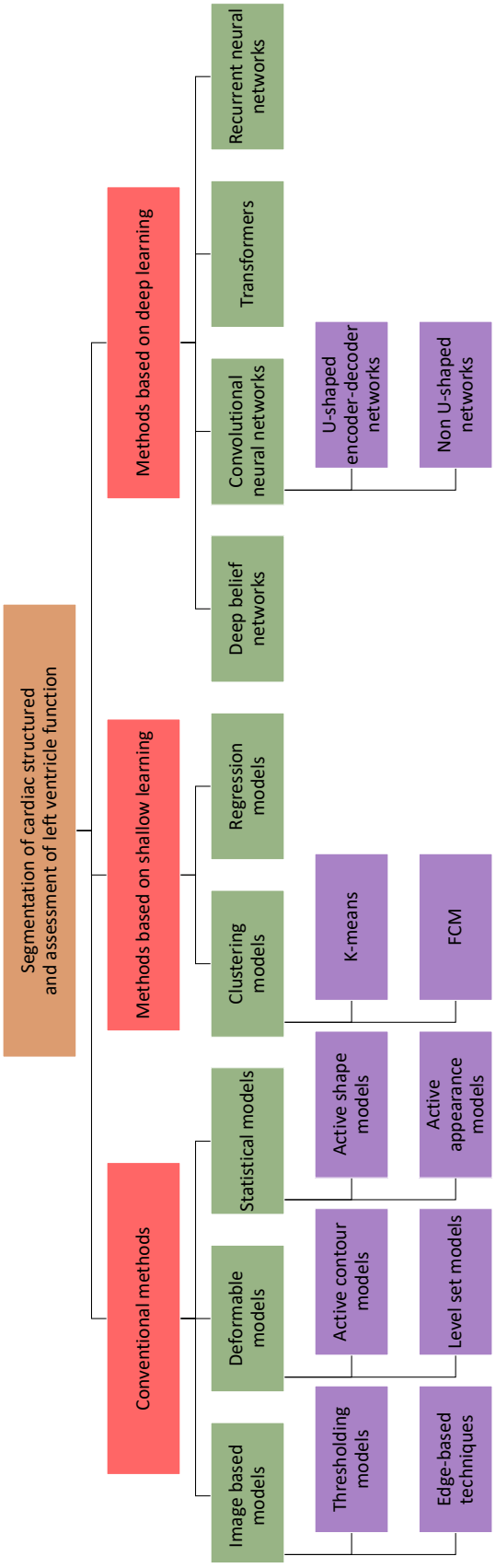


Figure 3.1: An illustration summarizing the principal methods proposed for the cardiac structure segmentation and the left ventricle function assessment.

fication. These two operations can remove minor contours that don't categorize as heart cavities. Laine et al. [84] proposed a method based on wavelet-based edge detection for border identification of the LV in 2D short-axis echocardiographic images. Another algorithm proposed in [85] achieved the same purpose. They used the anisotropic generalized Hough transform guided by a Gabor-like filtering. Although edge-based segmentation techniques are effective, they are noise-sensitive. Moreover, they cannot be applied to images with smooth transitions.

3.2.1.2 Deformable models

Active contour models Active contour models [86] are also known as snakes. They are sequential methods that separate the important pixels of the image using energy forces and constraints. An active contour model determines the real contour of objects by deforming the initial boundary within an image. The idea behind these approaches is to repeatedly minimize the energy function of the snakes to obtain smooth curves that fit the image features. The energy of the snakes E_{snake} combines internal E_{int} and external E_{ext} energies (See eq.(3.1)). The internal energy tries to impose a smoothness constraint on the model. Although, the external energy factor pulls the contour toward the features of the image, such as edges or lines.

$$E_{snake} = E_{int} + E_{ext} \tag{3.1}$$

Chalana et al. introduced a multiple active contour technique for segmenting the LV Endo and LV Epi in short-axis view images [87]. They used this method to characterize the LV by detecting two planar curves corresponding to the $\mathrm{LV_{Endo}}$ and $\mathrm{LV_{Epi}}$ borders. Mignotte and Meunier utilized statistical external energy in a discrete active contour for LV Endo segmentation in short-axis parasternal images [88]. The energy minimization was performed using Heitz's multiscale optimization strategy [89]. However, Mishra et al. proposed an active contour solution where optimization was achieved using a genetic algorithm [90]. Hammo et al. [91] suggested external energy based on gradient vector flow, which utilized optical flow to provide information about the heart's mechanical movement to the active contour model [91]. Singh et al. used optical flow to guide the propagation of a fitted contour from one frame to another in conjunction with an active contour approach [92]. A Gaussian-smoothed gradient of intensity was the basis for the image force in the active contour. Despite the advantages of active contour techniques in image segmentation, they have some limitations. They need a proper starting contour. Moreover, the snake-deformation process takes a long time.

Level set models Level set models are deformable implicit models proposed by Osher and Sethian [93] in 1988. They are similar to active contours utilized as a numerical methodology for tracking forms and interfaces. The main difference between active contours and level sets is that active contours move predefined snake points directly, according to an energy minimization strategy. However, the level set approaches move contours implicitly as a specific level of a function. The level set function ϕ is determined by an evolution equation known as an implicit active contour expressed as:

$$\frac{\partial \phi}{\partial t} = F|\nabla \phi|, \phi_0(x, y) = \phi(0, x, y) \tag{3.2}$$

Where $F = div(\frac{\nabla \phi}{|\nabla \phi|})$ indicates the speed function that controls the motion of the contour, $|\nabla \phi|$ denotes the normal direction, and ϕ_0 indicates the initial contour.

The level set approach has been widely used for LV segmentation in echocardiographic images. Lin et al. [94] presented a variant of this method concept for the LV_{Endo} boundaries segmentation at each frame of the echocardiographic image sequence. They proposed a multi-scale level set framework integrating edge and region information across spatial scales (pyramid levels). Yan et al. [95] applied the level set method to echocardiographic images using an adapted fast marching technique. They employed an average intensity gradient-based measure in the speed term to minimize mistakes caused by local feature (intensity gradient) measurements. The authors used a parasternal short axis and an A4C view sequence to evaluate the proposed method describing only qualitative findings. Sarti et al. [96] reported a level set maximum likelihood technique for ultrasound image segmentation, utilizing the Rayleigh probability distribution to model the gray levels of ultrasound images. They employed an energy function with a density probability distribution and smoothness restrictions to create a partial differential equation-based flow. They developed a level set formulation to find the minimum value of the model and segment the image accordingly. Fang et al. [97] proposed incorporating temporal information into the level set method. This method regularizes the curve evolution and overcome leakage boundary problems caused by dropouts of the inner heart wall boundary. When segmenting both LV_{Endo} and LV_{Epi}, the contours may overlap due to variable contrast and haziness. Dietenbeck et al. [98] introduced a constraint that enforces spacing between the LV_{Endo} and LV_{Epi} to address this issue, utilizing a level set model constrained by a shape formulation to segment both contours.

3.2.1.3 Statistical models

Active shape models Active shape models (ASMs) [99] are statistical models that can represent the shape and textural data in a specific region. They involve iteratively

deforming an initial form to fit the desired boundaries in an image. ASMs rely on a training set of samples to create a statistical model that captures the variance in shape. A fixed number of landmark points are used to define different forms. The shape space is constructed as a statistical model centered around the average shape. The shape boundaries are determined during the training process based on local image attributes using the training samples. ASM-based techniques often employ shape deformations that follow a Gaussian distribution. Principal Component Analysis (PCA) is used to limit the degree of shape deviations from the mean shape and capture the main patterns of variation in the shape space. In summary, ASMs combine statistical modeling techniques with iterative shape deformation to accurately represent and align shapes in images, allowing for shape-based analysis and segmentation tasks.

A combination of ASM and ACM was proposed by Hamarneh et al. [100] for LV_{Endo} segmentation in 2D echocardiographic images. The proposed technique incorporates ASM's ability to produce structures comparable to those in a training set and ACM's skill in constructing connected and smooth borders. Paragios et al. [101] developed a composite time-consistent 2D+time active shape model. The proposed model consists at first of a training step where the shape model was created from a linear combination of a diastolic and a systolic model obtained from a PCA applied to registered curves. Next, there were two primary steps in the segmentation process. For each image in the sequence, the LV boundary was segmented. As a result, it was possible to recover the similarity transformation parameters and register the shape model in each frame. The ASMs on echocardiogram video sequences were also used in [102]. The authors located and tracked the LV region over a heart cycle by detecting and propagating the expert annotations using ASM. Another framework was developed by Ali et al. [103] for 2D echocardiography segmentation that incorporates ASM, Nakagami distribution, and means squared eigenvalues error. Since ASM can deform and continually alter an initial shape to fit the intended boundary by employing a set of points, the authors used it to handle speckle noise and shadows. In this study, the Nakagami distribution was employed to increase the visibility of the echocardiography borders. However, the mean squared eigenvalues error reduces the total variance necessary for each landmark. ASMs were widely used for recognizing and delineating anatomic structures. However, it has some limitations [104]: low delineation accuracy, need for many landmarks, sensitivity to search range, sensitivity to initialization, and inability to properly utilize the unique information inherent in the image to be segmented.

Active appearance models Active Appearance Models (AAMs) are an extension of ASMs proposed by Cootes et al. [105]. As a combined statistical shape-appearance model,

an AAM characterizes image appearance and object shape over a set of samples. The Gaussian modes of variation are used to define the intensity distribution prior. AAMs can be used for image segmentation by reducing the difference between the model and an image along with statistically reasonable shape-intensity variations.

Basch et al. [106] presented a modified AAM named active appearance motion model applied on four-chamber echocardiographic image sequences of 129 patients. Nonlinear intensity normalization was a crucial phase in their process which proved to be of significant utility for accurate echocardiography results. The average distance between manual and automatic landmark points was 3.3 mm on the test set. The active appearance motion model gave more significant results than a classical AAM. Mitchell et al. [107] proposed a 3D AAM for 2D+T four-chamber echocardiography data segmentation. The method adapts itself in both space and time. During the model's training phase, manually segmented examples were used as input. The data for evaluating the proposed method are the same data used in [106]. The endocardial average distance error was 3.9 mm, which is slightly worse than the obtained results by Bosch et al. [106].

3.2.2 Methods based on shallow learning

Shallow learning refers to all machine learning algorithms and techniques that do not use deep multi-layer neural networks or multi-layer perceptrons.

3.2.2.1 Clustering models

Clustering methods are techniques for performing pixel-by-pixel image segmentation. In this kind of segmentation, we attempt to group adjacent pixels. A common exploratory pattern grouping technique for image analysis that separates the input space into regions is clustering. Some of the existing clustering methods include: K-means, improved K-means, fuzzy C-Means (FCM), and improved fuzzy C-means (IFCM) [108]. In the segmentation of the LV in the 2D echocardiography domain, K-means and FCM algorithms were used in [109, 110], respectively.

K-means K-means clustering is an iterative algorithm applied to unlabeled datasets to find particular groupings based on similarities between the data. The number K represents the number of the groups. It attempts to divide the dataset into K separate, non-overlapping clusters, with each data point belonging to just one group. In [109], they combined modified K-means with an active contour model to segment echocardiographic images in different views. The proposed method improved the computational time. The

segmentation takes just 16 seconds for images of 400×500 pixels with the proposed approach, while it takes 950 seconds with conventional k-means. Although the K-means approaches have advantages such as guaranteed convergence, guaranteed simplicity, and a low time complexity, there are some limitations [111], such as the need to know in advance the number of clusters, the dependence of solution quality on the number of formed clusters and the initial clusters, and the inadequacy of the method for non-convex data.

FCM FCM algorithm is an iterative clustering method that assigns data points to clusters based on similarity. It minimizes Euclidean distance as a cost function between each data point and the centroids of the various groups. The FCM produces effective medical image segmentation results. However, FCM is sensitive to noise, which can affect the accuracy of the segmentation results. To address this issue, Gupta et al. [110] proposed a hybrid method for ultrasound image segmentation. Their approach combines information from a modified FCM algorithm, called spatial constraint-based kernel FCM, and distance regularized level set-based edge features. The hybrid method aims to improve the segmentation accuracy by incorporating spatial constraints and edge information into the FCM algorithm. The authors evaluated their approach using real and synthetically generated ultrasound images to assess its performance.

3.2.2.2 Regression models

Regression analysis is a machine learning conception. It belongs to supervised learning because the system learns using input features and output labels. By calculating the impact of each variable on the others, regression analysis aids in creating a relationship between the variables. Data scientists can predict a continuous outcome (y) using regression in machine learning by applying mathematical techniques to the value of one or more predictor variables (x). Given its simplicity in predicting and forecasting, linear regression is the most widely used type of regression analysis.

Zhou et al. [112] proposed the generalized Image-Based Regression (IBR) for multiple-output scenarios. The authors utilized a boosting strategy to extract features from a redundant Haar-like feature set. The authors tested the proposed regressor on three tasks, i.e., age estimation, tumor detection, and LV endocardial wall localization. It improved regression performance while operating remarkably faster than traditional data-driven techniques like the support vector regressor [113]. Mean error achieved 2.148 versus 2.423 with support vector regressor in the endocardial wall detection task. In [114], the authors presented an enhanced version of IBR [112] called shape regression machine for the $\mathrm{LV}_{\mathrm{Endo}}$ segmentation in A4C echocardiograms. This approach contains two stages and segments

the ROI through statistical learning of the relationships between shape, appearance, and anatomy. It derives a regression solution to object detection in the first stage, which estimates a rigid form. In the second step, the algorithm estimates the nonrigid shape by learning a nonlinear regressor. The proposed algorithm detects and segments the LV_{Endo} in about 120 ms. In addition, Zhang et al. [115] developed a model that uses a regression task to learn the density probability from annotated training data. The authors used a regressor by choosing relative features to create an additive committee of weak learners based on a Haar-like feature, applying the gradient boosting method. The proposed algorithm performs over 60% better than the ASM model in the endocardial wall segmentation task.

3.2.3 Methods based on deep learning

The methods based on deep learning are the most used techniques for cardiac structure segmentation and heart function study. Most of the studies selected in this section apply their proposed methods to LV_{Endo} (See Figure 3.2) because of their importance in cardiac volumes and ejection fraction estimation. A summary of some relevant studies is given in Table 3.1.

3.2.3.1 Deep belief networks

Deep Belief Networks (DBNs) are a type of deep neural network. Unlike CNNs, which consist of one or more convolutional layers followed by fully connected layers, DBNs are probabilistic generative models that consist of multiple layers of hidden units [116].

Carneiro et al. [117] used a deep learning framework to segment the LV in apical long-axis echocardiograms. The proposed method utilized DBN [118] to forecast the rigid transformation and deformable model parameters for localization and segmentation. They evaluated the proposed method on two datasets: one with 400 annotated images from diseased patients (12 image sequences) and another with 80 annotated images from healthy cases (2 image sequences). The outcomes illustrated how DBN-based feature extraction was resistant to changes in image appearance. Nascimento and Carneiro [119] further improved the DBN-based architecture by incorporating sparse manifold learning in the rigid detection stage, reducing training and inference complexity. Carneiro and Nascimento [120] handled the coherence between temporally close frames to increase the precision and robustness of the LV segmentation. The segmentation of the current cardiac phase depends on earlier ones in the dynamic modeling method, which is based on a sequential monte carlo framework [121] with a transition model. The results demonstrated that this strategy outperforms their earlier study [122], which ignored temporal

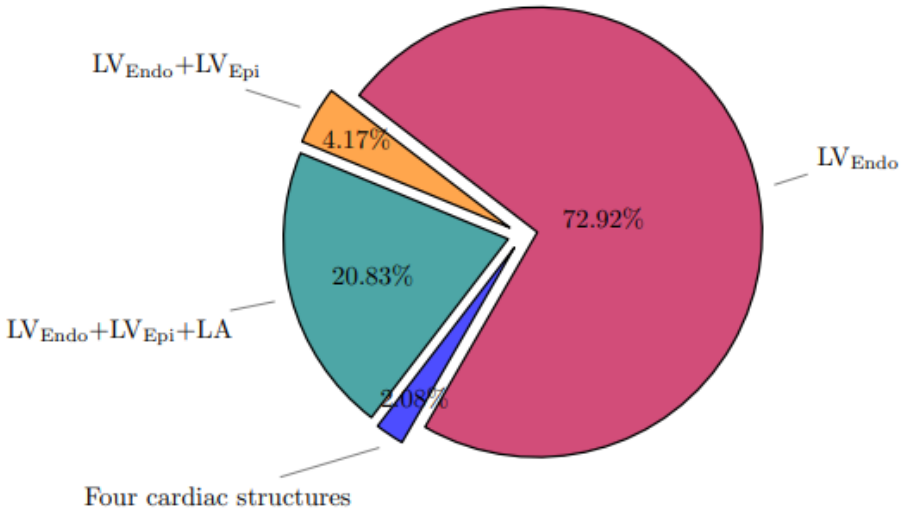


Figure 3.2: Pie chart demonstrating the segmentation of cardiac structures by deep learning based methods selected in this thesis.

information.

3.2.3.2 Convolutional neural networks

Most studies have demonstrated that deep learning can produce significantly better performance when compared with conventional machine learning methods. CNNs (Explained in Section 2.5) are the most used network for LV segmentation and quantitative analysis in 2D echocardiography.

Non U-shaped networks Lei et al. [123] introduced Cardiac-SegNet, an anchor-free mask CNN designed for multi-structure segmentation in echocardiographic images. The network consists of a backbone (2D ResNet), a fully convolutional one-stage object detector (FCOS) head, and a mask head. This backbone extracts multi-level and multi-scale features from the input image. The FCOS head, which is anchor-free, detects the region of interest (ROI) using the extracted features and represents the spatial relationship of the targets. The mask head utilizes a spatial attention technique to segment each detected ROI, highlighting relevant characteristics. Liu et al. [124] introduced a pyramid local attention module to capture local feature similarities and improve segmentation accuracy. This module enhances the network's ability to capture fine-grained details and local contextual information. Shen et al. [125] proposed an intermediate supervision deep neural network method to enhance LV segmentation from diverse data. This method utilizes intermediate supervision, which involves introducing auxiliary loss functions at different

intermediate layers of the network. This approach helps in guiding the training process and improving the segmentation performance. Zeng et al. [126] presented MAEF-Net, a multi-attention efficient feature fusion network, for automatic LV segmentation. The network automatically identifies the ED and ES frames to calculate the LV_{EF}. To properly collect heartbeat features while reducing noise, the MAEF-Net method used a multi-attention mechanism. It also utilizes a deep supervision mechanism and spatial pyramid feature fusion to enhance feature extraction capabilities. The proposed techniques leverage multiple architectural components, attention mechanisms, and supervision strategies to improve segmentation accuracy and robustness.

Other approaches provided an automatic segmentation methodology to address the LV analysis problem using CNNs in the spatiotemporal domain. Chen et al. [127] proposed a temporal affine network that performs three echocardiographic interpretation tasks: standard cardiac plane recognition, LV landmark detection, and LV segmentation. For evaluation, they gathered A4C sequences from 991 patients with 33047 images and A2C sequences from 723 patients with 32551 images. The suggested approach reached 91.14% of Dice with a lightweight MobileNetV2 network [128] as a backbone. Using their collected public dataset, EchoNet-Dynamic, Ouyang et al. [129] initially suggested segmenting the LV at the frame level using weak supervision from expert human tracings and before predicting the LV_{EF} for each cardiac cycle using spatiotemporal convolutions with residual connections. The model segmented the LV with a DSC of 0.92 and predicted ejection fraction with a mean absolute error of 4.1%. The model proposed in [130] aims to provide an adaptive calibration mechanism that uses the spatiotemporal coherence between adjacent frames to address the primary difficulties caused by background speckle noise in ultrasound image segmentation. The authors demonstrated that their method had benefits, but they also declared that it had certain drawbacks, particularly in the case of segmenting echocardiography videos with extremely irregular cardiac activity or low contrast between the ventricle and its surroundings.

In the literature, some studies have proposed frameworks for incorporating uncertainty into deep learning models, called Bayesian deep learning. This approach has been used in [131] for the automatic LV_{EF} evaluation in echo videos. Each weight in the deep learning model was described as a random variable with a Gaussian distribution. Uncertainties in the estimation of LV_{EF} can be modeled using this probabilistic approach. The results demonstrated the superior performance of the Bayesian technique over a (2+1)D architecture based on ResNet18 using the EchoNet-Dynamic dataset. Jafari et al. [132] also advanced Bayesian deep learning. The model was based on Bayesian U-Net to estimate the LV_{EF} based on segmentation of the LV in parasternal short-axis papillary muscles. Dahal et al. [133] examined three ensemble-based uncertainty models using four metrics to

get insight into uncertainty modeling for LV segmentation from ultrasound images. They demonstrated how uncertainty can be used to automatically reject poor quality images and enhance the results of segmentation methods.

U-shaped encoder-decoder networks The U-shaped networks are end-to-end CNNs which gained popularity as a technique for medical image segmentation due to their outstanding achievements [134]. U-Net was proposed by Ronneberger et al. [9] as the original architecture of U-shaped networks. To our knowledge, Smistad et al. [135] are the first to use U-Net to segment the left ventricle in 2D echocardiographic images. They trained a U-net using images labeled by a Kalman filter-based segmentation method [136]. Over 1,500 ultrasound recordings (100,000 2D apical frames) were acquired from 100 patients to evaluate the suggested technique. Authors reported that U-Net achieved comparable performance to the Kalman filter by using output data from this method to train the CNN. Referring to this architecture, Kulkarni et al. [137] developed a CNN incorporating feature extraction and denoising procedures for LV segmentation. It was trained on 70 patients and tested on 12 patients. A comparative study of U-Net, Residual U-Net [138], and Dense U-Net [139] algorithms was presented by Kim et al. [140] for the cardiac structure segmentation and clinical indices. The entire private dataset used in this study contains 500 patients (400 for training, 50 for validation, and 50 for testing). Through the obtained results, this study demonstrated the need for further technical development of fully automated deep learning methods to maintain clinical performance and the urgent need for well-designed clinical validation studies. Gomez et al. [141] used U-Net architecture to predict the LV_{Endo} and the key landmark points within this contour. The proposed approach makes use of a U-Net-based two-headed network (one for predicting the contour points and the other head for predicting a distance map to the contour). The performance advantages were up to 22% in terms of landmark localization ($<$4.5 mm) and distance to the ground truth contour ($<$3.0 mm).

Leclerc et al. [18] conducted a study comparing deep learning-based and non-deep learning methods for heart structure segmentation and clinical indices estimation using the CAMUS dataset. The findings revealed that encoder-decoder architectures (U-Net) outperformed non-deep learning approaches. Table 3.1 presents the evaluation results of their proposed method. Azarmehr et al. [142] evaluated the performance of a modified U-Net, as yet proposed by Leclerc et al. [18], for segmenting the LV_{Endo}. They used a private dataset of 61 patients with 992 annotated frames of A4C views. The results indicated that the original U-Net model outperformed the other modified U-Net models proposed by Leclerc et al., achieving an average DSC of 0.92$\pm$0.05 and a HD of 3.97$\pm$0.82. Several subsequent studies utilized the CAMUS dataset to evaluate proposed methods for

Table 3.1: Deep learning-based methods for LV segmentation and assessment. The acronyms DSC, HD, corr, mae, and rmse stand for: Dice Similarity Coefficient, Hausdorff Distance, correlation coefficient, mean absolute error, and root mean square error, respectively.

Study	Year	Deep learning model used	Dataset	Measure	Performance
[18]	2019	U-Net	CAMUS	DSC(ED) DSC(ES) HD(ED) HD(ES) corr(LV_{Endo}) mae(LV_{Endo})	0.939 ± 0.043 0.916 ± 0.061 5.3 ± 3.6 mm 5.5 ± 3.8 mm 0.823 4.3
[129]	2020	Spatiotemporal convolutions with residual connections + weak supervision	EchoNet-Dynamic	mae (LV_{EF}) rmse (LV_{EF})	7.35 9.53
[133]	2020	Uncertainty estimation techniques on convolutional network	CAMUS EchoNet-Dynamic	DSC(CAMUS(ED)) DSC(CAMUS(ES)) DSC(Echo-Net(ED)) DSC(Echo-Net(ES))	0.932 0.911 0.930 0.899
[135]	2017	U-Net + Kalman filter for generating labels	100 patients: 100,000 apical images	DSC HD	0.87 ± 0.06 5.9 ± 2.9 mm
[142]	2020	U-Net	61 patients: 992 images	DSC HD	0.92 ± 0.05 3.97 ± 0.82 mm
[143]	2020	FCN + adversarial training + post processing	CAMUS 100 patients: 1,395 images	DSC(CAMUS) / DSC HD	86.21% ± 9.9 / 92.13% ± 3.3 5.19 ± 7.6 mm
[144]	2019	U-Net + Anatomically Constrained CycleGAN	427 patients: 854 A4C images	Mean DSC (ED) Mean DSC (ES) Mean HD (ED) Mean HD (ES)	93.6% ± 2.9 90.2% ± 4.3 7.1 ± 3.2 mm 7.8 ± 3.0 mm
[145]	2021	TransBridge	Echonet-Dynamic	DSC HD	91.64% 4.185 mm
[146]	2018	U-Net + BiLSTM + optical flow	556 patients: 34,000 A4C images	Mean accuracy Mean DSC	97.7% 92.7%
[147]	2020	Dense pyramid and deep supervision network	100 patients: 10,858 A2C, A3C and A4C images	DSC HD	0.921 ± 0.046 5.75 ± 3.14 mm
[148]	2023	Bilateral lightweight deep neural network	6,000 A4C images	DSC Accuracy	0.9446 0.9742

cardiac structure assessment in 2D echocardiography. Some of these studies suggested improvements to the original U-Net architecture, for example, adding feature pyramids in each decoder level of the dilated U-Net [149, 150]. This modification addresses the limitation of ignoring the contribution of all semantic strengths during the segmentation process. In other studies [151–154], authors replaced U-Net blocks with residual units. Chernyshov et al. [155] tested 24 variations of U-Net and DeepLabV3+ [156] on the CAMUS and Cityscapes datasets. They modified the structural components, such as the receptive field, number of layers, and convolutional filters, to create different models. The experiments consistently showed high DSC values (0.86-0.90) on the CAMUS dataset and lower DSC values (0.48-0.67) on the Cityscapes dataset for all models. Hesse et al. [157] introduced active contour label correction to improve the segmentation performance of U-Net with imperfect labels. They corrected inaccurate ground-truth labels in the training set by incorporating active contour. Sfakianakis et al. [158] presented an ensemble of CNNs for segmenting the LV_{Endo}, LV_{Myo}, and LA. The ensemble combined five U-Nets trained on the CAMUS dataset. The framework incorporated customized data augmentation techniques to enhance learning capacity and generalization. They also introduced a modified loss function with soft Dice similarity and anatomical constraints. The segmentation accuracy on the CAMUS test set was reported as a DSC of 0.94/0.929 and an HD of 4.7/4.7 mm for ED and ES, respectively. Wei et al. [159] proposed CLAS, a method using 3D U-Net [160] as a shared feature extractor. CLAS produced segmentation maps over the entire cardiac cycle while aiming to maintain temporal consistency. During training, CLAS predicted deformation fields for propagating the annotations. The proposed approach achieved LV volume and EF estimation, with Pearson correlations of 0.958, 0.979, and 0.926, respectively, on the test set. In their subsequent work [161], Wei et al. presented a multi-task semi-supervised framework called MCLAS, which improved their previous CLAS method [159]. MCLAS allowed view identification, cardiac structure segmentation, and EF calculation. It outperformed the CLAS framework in LV volume and EF estimation, achieving Pearson correlations of 0.975, 0.983, and 0.946, respectively. Using a 10-fold cross-validation research design with data augmentation, Chen et al. [162] compared CLAS against a previous state-of-the-art frame-level segmentation technique [163]. They also assessed the generalizability of CLAS trained on the CAMUS dataset to the EchoNet-Dynamic dataset. With the CAMUS dataset, CLAS achieved correlations of 0.983, 0.969, and 0.883 for LV_{EDV}, LV_{ESV}, and LV_{EF}, respectively. On the EchoNet-Dynamic dataset's test set, which included 1274 patients, CLAS yielded a median absolute error of 4.9%±5.4 in EF estimation.

The EchoNet-Dynamic dataset has also been used in other studies to validate U-shape-based methods for automated segmentation of the LV_{Endo} and assessment of left

ventricular function in echocardiography [164]. Other works have used the CAMUS and EchoNet-Dynamic datasets, sometimes adding private datasets. Liu et al. [165] proposed a deep learning method based on a modified U-Net with symmetric architecture called DPS-Net. It was evaluated on 36890 frames of 2D echocardiography from 340 patients, CAMUS, and EchoNet-Dynamic datasets to demonstrate its adaptability to various echocardiographic systems. For 2D echocardiographic segmentation and landmark detection, Yang and Sermesant [166] examined how to incorporate shape constraints from global, regional, and pixel-level into an original U-Net architecture. Pixel-level shape constraint is more effective with 0.931/0.895 of DSC and 4.99/12.56 of HD, according to the evaluation results on the CAMUS/EchoNet-Dynamic datasets. Puyol-Antón et al. [167] presented a novel AI method for extracting sophisticated biomarkers from LV systolic and diastolic performance using the entire cardiac cycle segmentation. The nnU-Net model [168] was the basis for the proposed AI model, which was trained and tested using four different datasets. The framework achieved DSC of 0.931/0.922 with CAMUS and 0.935/0.926 with EchoNet-Dynamic in ED/ES, respectively. The authors declared that their method presented significant results but has some drawbacks. They intended to validate it more thoroughly by utilizing data annotated with ground truth along the cardiac cycle. Using U-Net and DeepLabV3 architectures, Saeed et al. [169] tested the effect of contrastive learning in segmenting the LV from echocardiography. The proposed solution reached a DSC of 0.9252 on EchoNet-Dynamic. The authors demonstrated that contrastive pretraining enhances LV segmentation ability, especially when annotated data is limited.

3.2.3.3 Generative adversarial networks

In recent years, a great deal of research has been done on Generative Adversarial Networks (GAN) [170], which are deep-learning-based generative models. They achieved significant progress in different challenges of the computer vision field, e.g., image generation, image-to-image translation, and facial attribute manipulation [171]. Two submodels in competition with each other, as depicted in Figure 3.3, construct the GAN model architecture: a generator model for creating new instances and a discriminator model for determining whether generated examples are real or fake. For LV segmentation in echocardiographic images, Arafati et al. [143] proposed a novel generalizable, fully automatic multi-label segmentation method for A4C view echocardiograms. It is based on deep FCNs and adversarial training. The GANs were used for pixel classification training. The authors used 1395 annotated images of 100 patients and the CAMUS dataset to validate the proposed algorithm. *DSC* results on the CAMUS dataset were:

86.21%±9.9 and 67.81%±27.8 of LV and LA, respectively. GANs were also utilized in LV segmentation context for echocardiographic image generation [172] and quality translation and improvement [144, 173–175]. Gilbert et al. [172] suggested synthesizing labeled 2D echocardiography images for LV segmentation using anatomical models and CycleGAN [176]. Jafari et al. [173] used a constrained cycleGAN to improve the image quality from a point-of-care ultrasound (POCUS) device. This study uses a total of 1089 echo studies from 841 patients. An average improvement of 30% and 34 mm of DSC and HD was obtained in the LV segmentation. Escobar et al. [174] presented a GAN architecture, called an UltraGAN, to enhance CAMUS images before segmenting the cardiac structures. The proposed network integrates frequency loss functions and an anatomical coherence constraint. The generator and discriminator in UltraGAN are CycleGAN [176] and PatchGAN [177], respectively. A simple U-Net was used to compare the segmentation results on the original and enhanced images. The authors found that training U-Net with enhanced data by Ultra-GAN can boost the echocardiogram segmentation results. A multi-space adaptation-segmentation-joint network for extracting the LV_{Endo} and LV_{Epi} in echocardiography, termed MACS, was proposed by Chen et al. [178]. It adopted a GAN architecture to handle the cross-domain echocardiography analysis. DSC and HD achieved 0.9033 and 5.65 for LV_{Endo} segmentation.

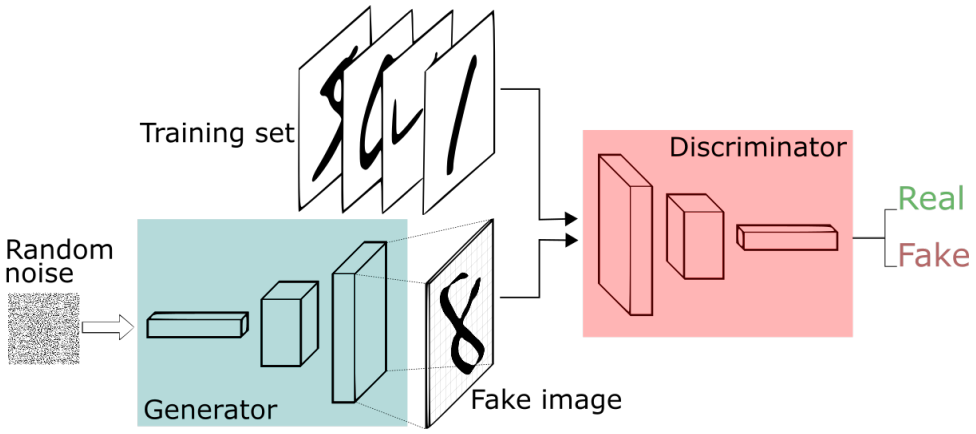


Figure 3.3: Typical example of GAN architecture[a]. The generator transforms noise into an imitation of the data to try to trick the discriminator network. The discriminator tries to identify real data from fakes created by the generator network.

[a]https://mgubaidullin.github.io/deeplearning4j-docs/generative-adversarial-network.html

3.2.3.4 Transformers

The idea of transformers was first presented by Vaswani et al. [12] in 2017. They developed the transformers as a new attention-driven building block for machine translation. These

attention blocks are specific layers of neural networks that combine data from the entire input sequence [179]. Although the Transformer architecture has an encoder-decoder structure, it does not use convolutions or recurrence to produce an output (See Figure 3.4). Vision Transformers (ViTs) [180] are the most common transformers architecture. ViTs are constructed by cascading many transformer layers to capture the overall context of an input image. A sequence of vectors is created by dividing an image into fixed-size patches, linearly embedding each one, adding position embeddings, and then feeding the assembled vectors to a conventional Transformer encoder.

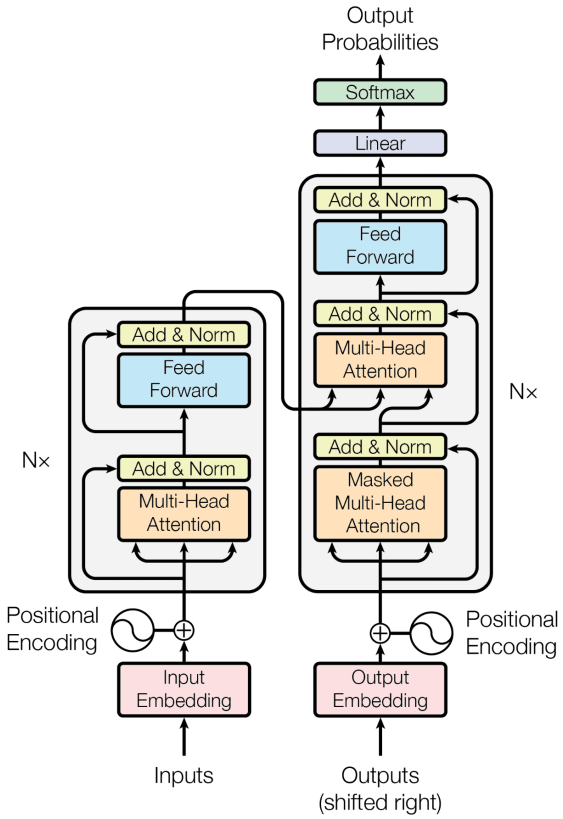


Figure 3.4: The encoder-decoder structure of the Transformer architecture [12].

A lightweight Transformer for LV segmentation in echocardiography called Trans-Bridge was proposed by Deng et al. [145]. The architecture combines a CNN encoder-decoder structure and a transformer structure. To combine the multi-level characteristics extracted by the CNN encoder and create global and inter-level dependencies, the transformer layers connect the CNN encoder and decoder. The patch embedding layer of the transformer has been redesigned using the shuffling layer [181] and group convolutions to minimize the number of parameters. The architecture efficiency was tested on the EchoNet-Dynamic dataset. The reached value of DSC was 91.4%. The same concept

of merging the transformer and CNN to segment the LV in echocardiographic images was applied in [182]. The Transformer and CNN branches receive a single-frame image, respectively. The derived multi-scale feature maps are then combined in the fusion module. The bridge attention module calculates the segmentation map after calculating the attention using three-layer fusion features. The authors also used the EchoNet-Dynamic dataset to evaluate the proposed method. The obtained DSC increased to 92.4%. An approach proposed by Reynaud et al. [183] directly and automatically computed the LV_{EF} after detecting the ED and ES frames from ultrasound videos. The framework includes a residual auto-encoder network and a BERT model [184]. It estimated the LV_{EF} with a mae of 5.95 in the EchoNet-Dynamic dataset.

3.2.3.5 Recurrent neural networks

A recurrent neural network (RNN) is a neural sequence model capable of learning features and long-term dependencies from sequential and time-series data [185]. RNN can simulate a data series so that prior samples are reliant on the current ones. Figure 3.5 shows that RNNs are best suited for sequence data where a point in the series can be paired with information from earlier points since they feed their output back as input. Gradient vanishing and exploding problems are the main drawbacks of RNNs. A version of RNN called Long Short-Term Memory (LSTM) [186] networks was developed to solve these disadvantages. The LSTM model is trained using back-propagation.

A recurrent neural network (RNN) (Figure 3.5) is a neural sequence model designed to process sequential and time-series data by capturing dependencies between different elements in the sequence. Unlike feed-forward neural networks, RNNs have feedback connections that allow information to be passed from one sequence step to the next. The main advantage of RNNs is their ability to capture long-term dependencies and contextual information in sequential data. However, traditional RNNs suffer from the problems of gradient vanishing and exploding during training. These problems occur when the gradients become very small or large, making it difficult for the model to learn effectively. As a result, RNNs may struggle to capture long-term dependencies or make accurate predictions. The Long Short-Term Memory (LSTM) network was introduced in [186] to overcome these limitations. LSTMs are a variant of RNNs designed to alleviate the gradient vanishing and exploding problems. They achieve this by incorporating memory cells and gating mechanisms that regulate the flow of information within the network.

Most RNN-based methods for LV segmentation and assessment of its function in echocardiograms analyze temporal information during the cardiac cycle. Jafari et al. [146] developed a deep learning framework by combining U-Net, stacked bidirectional convolu-

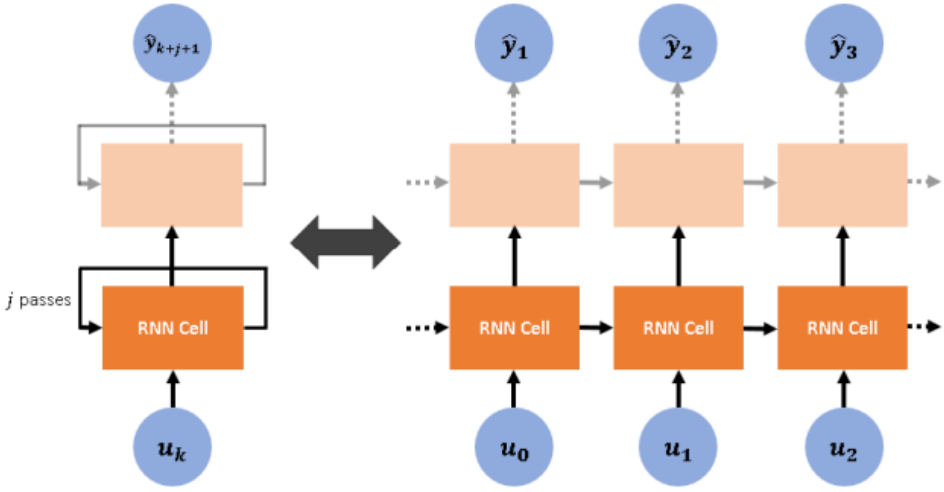


Figure 3.5: A recurrent neural network architecture [13].

tional LSTM, and inter-frame optical flow [187] to segment one target frame (ED and ES frames) using multiple frames. The authors used a dataset with 648 A4C echocardiograms collected from 566 patients for evaluation. The results showed that the suggested methodology segmented data with a notable high accuracy of 97.9% and a standard deviation of less than 1%. Ge et al. [188] proposed a k-shaped network of end-to-end deep neural networks for multiview segmentation and multidimensional quantification of the LV in A4C and A2C echocardiography sequences. In this framework, spatial-temporal information in the cardiac activity was effectively captured by the Bi-ResLSTM implemented in the bidirectional recurrent [189] and the shortcut connection of convolutional LSTMs. The authors used 2000 2D echo images of 50 participants from 2 hospitals for evaluation and achieved a DSC of 91.44±4.02 and 90.44±4.20 in A4C and A2C, respectively. Li et al. [147] developed a multiview recurrent aggregation network for echocardiographic recordings segmentation with cardiac cycle analysis. This method was evaluated on the CAMUS dataset and a private dataset of 13500 2D echocardiographic images collected from 150 patients, reaching an AUC of 0.9997 on CAMUS data. Lin et al. [190] proposed a convolutional long-short-term-memory attention-gated U-Net (CLA-U-Net) for automatic LV segmentation in 2D echocardiograms. In the encoder part of the U-Net, they added a convolutional long-short-term memory block to capture temporal information between the frames and integrated a channel attention mechanism in the skip connections. From the EchoNet-Dynamic test set, the LV was segmented with a DSC of 0.9311.

3.3 Conclusion

In recent years, there has been a growing interest in LV assessment in 2D echocardiography, which has captured the attention of numerous researchers. This chapter provides a literature review in this field of research, categorizing the state-of-the-art methods into three groups: conventional techniques, shallow learning-based algorithms, and deep learning-based frameworks. Our survey reveals that deep learning-based approaches, particularly U-shaped encoder-decoder networks, are the most commonly utilized techniques for LV segmentation and cardiac function assessment in 2D echocardiography.

Chapter 4

Impact of attention mechanism on U-Net architecture for the Left Ventricle segmentation

4.1 Introduction

Over time, cardiovascular research has advanced to help in the early diagnosis of cardiac disorders. The segmentation of the cardiac structures in 2D echocardiographic images, especially the LV, is the subject of intense research. Generally, the image segmentation technique is an important and challenging step in image processing because it can locate the relevant objects of the image. Accurate segmentation of the LV_{Endo} in ED and ES frames facilitates the cardiac function assessment. It can successfully replace the manual delineation of the LV_{Endo} in clinical routine, which suffers from different complications. The adoption of automatic segmentation techniques may aid in resolving these problems, as well as decrease intra- and inter-user discrepancy.

The automatic methods used to automate the task of LV segmentation are based on deep CNN-based models. The convolutional neural network U-Net, one of the most promising deep learning algorithms for the segmentation of medical images, is the basis of our research in this chapter. Various developments have been made in the original U-Net architecture to get the best version that fits the investigated domain. An important property sought in image processing networks is their ability to concentrate on relevant objects while disregarding irrelevant areas, which can be achieved through attention mechanisms. This chapter examines the impact of attention gates on two modified U-Net architectures for LV segmentation in echocardiographic images.

In the subsequent sections of this chapter, we will describe the principles of the pro-

posed method, including image preprocessing and segmentation. We will then present the design of experimental setups and the results obtained from each experiment. These results will be discussed in the subsequent section. Finally, we will provide the main conclusions and summarize the essential points of this chapter.

4.2 Methods and procedure

We present in this section the proposed CNN method and procedures used to segment the LV in echocardiographic images.

4.2.1 Image preprocessing

Images must be scaled to a specific size before being fed into a CNN, as CNNs receive inputs of the same size. Large images take up significant memory space and require large neural networks. As a result, the complexity of the memory space and the computational time increase. Thus, the size of input images is chosen based on a compromise between accuracy and processing efficiency. Images might lose information when we apply the resizing operations on them. Cropping the border pixels or scaling down using interpolation are two approaches for resizing down to a fixed size [191]. In both cases, it is possible to lose information. Scaling runs the probability of distorting features or patterns across the image. However, cropping may delete features that exist in the border areas. Scaling is a better option for reducing the dimension of larger images to the desired size because it presents less risk than losing patterns. The first step in the proposed method is to resize the original images using scaling down instead of cropping. For training a CNN, the resolutions of the input images commonly range between 64×64 and 256×256. In this thesis, We chose the dimension of 256×256 as the desired size of the input data.

Among the causes that prevent the extraction of important feature information is the degradation of image quality. An image enhancement improves the quality of echocardiographic images. This preprocessing technique can better display the structure of the heart. The main goal of image enhancement is to reveal hidden details or to boost the contrast of poor images. Furthermore, it is an essential step to improve the segmentation results and offers a wide range of options for enhancing the quality of images. Histogram equalization is one of the most widely used contrast-enhancing algorithms. Therefore, we applied this technique because of the poor contrast and heterogeneity of the echocardiographic images. Figure 4.1 presents examples of the three types of CAMUS images (good, medium, and poor quality) before and after histogram equalization. Figure 4.2 illustrates the histogram graphs of each image.

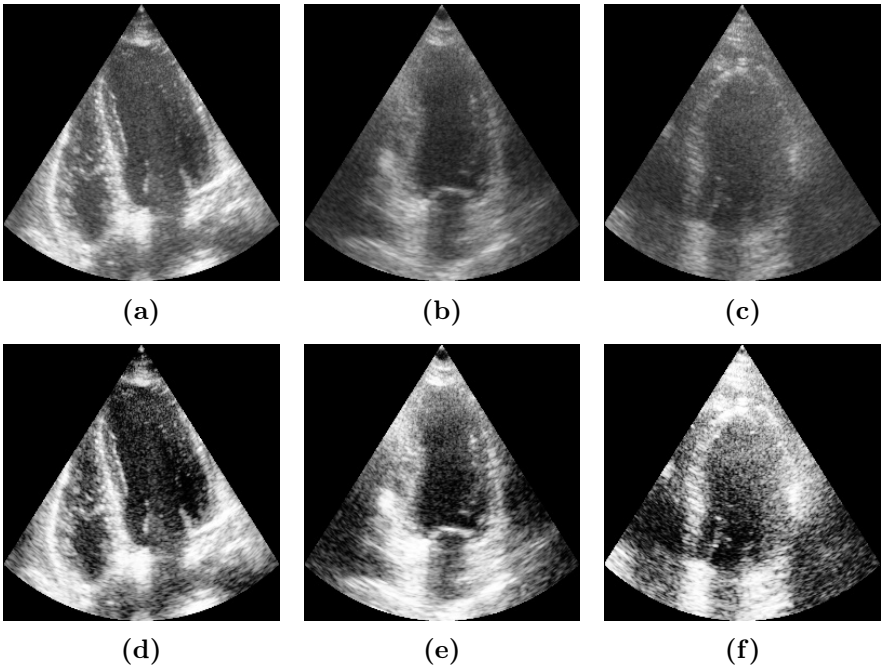


Figure 4.1: Histogram equalization of CAMUS images. (a-b-c) Good, medium, and poor qualities, respectively. (d-e-f) Histogram equalization of (a-b-c) images, respectively.

Data augmentation is a technique used to prepare data before feeding it into a CNN. It involves creating modified copies of existing data. The created data artificially increase the training set size. Data augmentation offers several benefits, such as improving the generalization and robustness of the trained model. It can also help prevent overfitting by introducing variations and diversifying the training samples. However, data augmentation also has certain limitations, such as the need for reliable evaluation systems to assess the quality of augmented datasets.We didn't implement data augmentation in the experiments conducted in this thesis. Data augmentation can introduce noise and artifacts into the augmented data. Echocardiographic images are already noisy, augmenting it further may not be advisable, as it could exacerbate these issues.

4.2.2 Proposed segmentation architecture

4.2.2.1 U-Net 1 and U-Net 2

Due to the success of the U-Net architectural network in the segmentation of medical images, our proposed method is based on U-Net architecture. We tested the impact of attention mechanisms on two modified U-Net.

There are many different U-Net designs presented in the literature. Leclerc et al. [18] proposed two modified U-Net implementations for the cardiac structure segmentation

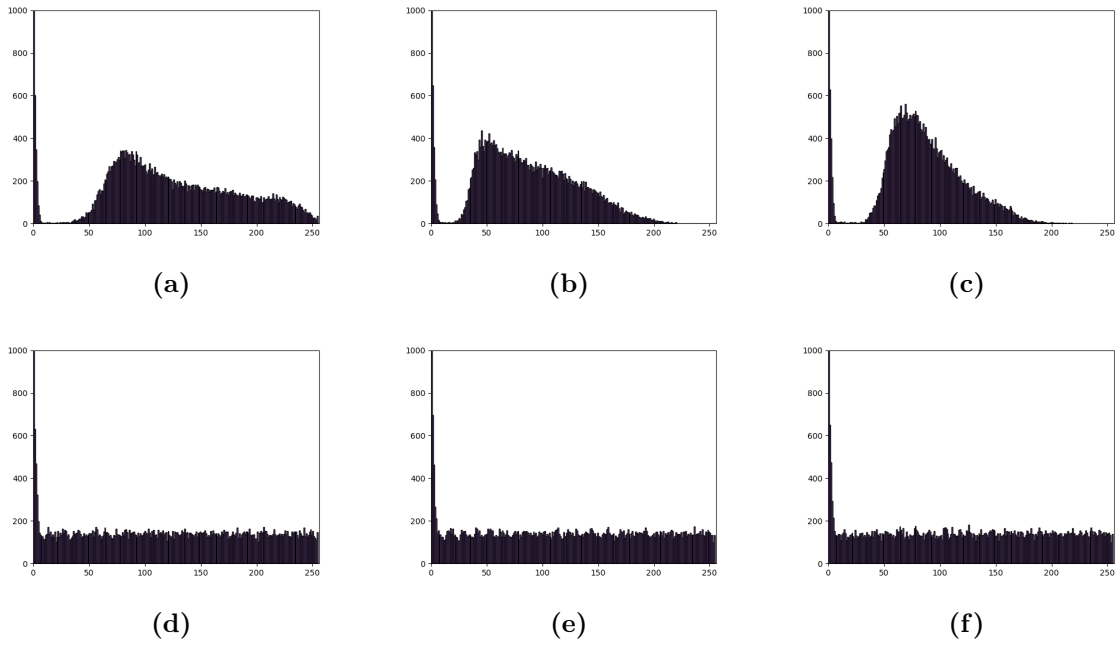


Figure 4.2: Histogram graphs corresponding to images in Figure 4.1.

in the CAMUS dataset. These two architectures are called U-Net 1 and U-Net 2. U-Net 1 design was based on the model proposed by Smistad et al. [135], and U-Net 2 was adapted from their previous work [192]. U-Net 1 and U-Net 2 were optimized for speed and accuracy, respectively. The comparison between these two architectures authorizes the examination of the effects of hyper-parameter and architecture selections on the quality of the segmentation outcomes. As depicted in Table 4.1, the main differences between U-Net 1 and U-Net 2 are:

- U-Net 1 has six levels and fewer filters, while U-Net 2 has only five levels and more filters. Thus, U-Net 1 is a deeper model, while U-Net 2 is a broader architecture.
- The resolution in the last level of the encoder part of U-Net 1 and U-Net 2 are 8×8 and 16×16, respectively.
- In the decoder path, U-Net 1 uses an upsample layer (UpSampling2D) with a kernel size of 2×2. U-Net 2 uses a transpose convolutional layer (Conv2DTranspose) that applies an inverse convolution operation (Deconvolution).
- Batch normalization has not been used in U-Net 1 but has been added after each convolution of U-Net 2.
- Both models do not include spatial dropout as it tends to reduce performance.

Table 4.1: The main differences between U-Net 1 and U-Net 2 architectures. Reproduced from [18].

Network characteristic	U-Net 1	U-Net 2
Feature maps	32 ↓ 128 ↑ 16	32 ↓ 512 ↑ 32
Lowest resolution	8 × 8	16 × 16
Upsampling scheme	2 × 2 repeats	Deconvolutions
Normalization scheme	/	BatchNorm
Spatial dropout	/	/

4.2.2.2 Attention mechanism

Attention modules have been widely utilized in CNN networks. Models with these gates learn to focus and amplify the relevant features while suppressing the irrelevant regions [193]. They are added into the skip connections right before the concatenation layers in the case of U-Net architecture [11]. They make it possible to process feature maps transmitted via skip connections and propagate only the most crucial spatial data to the expansion path. In this chapter, we investigate the impact of attention mechanisms on U-Net 1 and U-Net 2 models for the task of LV_{Endo} segmentation in echocardiographic images. Specifically, we examine U-Net 1, U-Net 2, attention U-Net 1, and attention U-Net 2.

We used soft attention gates presented in [14] and adapted from [11]. They exhibited an improvement in segmentation performance. The attention gates with U-Net architecture are depicted in Figure 4.3. The attention gate consists of several successive operations and takes two inputs: the input signal from the encoder and the gating signal collected from a coarser scale in the decoder. Element-wise sum and 1×1 convolution are applied to these input tensors to obtain the attention coefficients transformed then using ReLu and Sigmoid functions. An element-wise multiplication is performed between the input signal and attention coefficients. The concatenation layer joins the relevant features to the output of upsampling. Following the approach proposed by Abraham et al. [14], attention gates are applied in all skip connections except the last one to avoid over-suppression (See Figure 4.3).

We incorporated the attention modules into U-Net 1 and 2 to evaluate the effectiveness of the attention mechanism on U-shaped networks. Table 4.2 illustrates the number of parameters before and after integrating the attention gates.

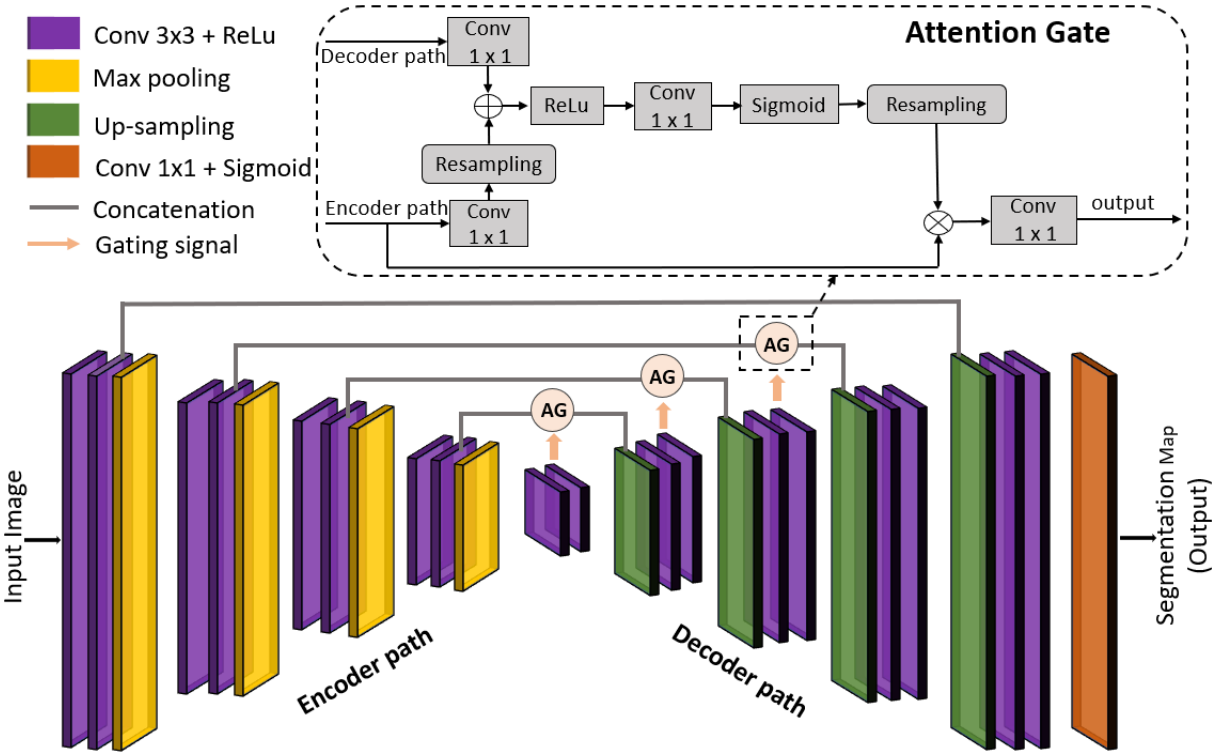


Figure 4.3: Structure of U-Net with attention mechanism proposed in [14].

Table 4.2: Number of parameters of each model

Network	Total parameters	Trainable parameters	Non-trainable parameters
U-Net 1	1.976.593	1.976.593	0
Attention U-Net 1	3.504.793	3.501.657	3.136
U-Net 2	17.476.657	17.466.385	10.272
attention U-Net 2	21.236.471	21.222.039	14,432

4.2.2.3 Deep supervision

The concept of deep supervision was proposed by Lee et al. [194]. In the U-Net architecture, deep supervision has been added to the output map of the decoder channel. The idea behind this strategy is to introduce each final layer of each decoder stage into a 3×3 convolutional layer followed by an upsampling operation and a sigmoid function. After that, it averages the feature maps of all the outputs of the expansion path. Deep supervision requires semantic discrimination at all scales from intermediate levels of U-Net

architecture. In attention U-Net, it ensures that the attention module can affect the responses to different visual foreground content.

4.3 Experiments and results

4.3.1 CAMUS dataset description

Cardiac Acquisitions for Multi-structure Ultrasound Segmentation (CAMUS)[1], is an open-access dataset used for multi-structure ultrasound segmentation. The dataset was collected from the University Hospital of Saint-Étienne (France) using GE Vivid E95 ultrasound scanners equipped with a GE M5S probe and EchoPAC analysis software. This dataset incorporates 2D A2C and A4C sequences from 500 patients in total. The image sequences are represented in polar coordinates and exported from the GE system as sets of B-mode images. The same interpolation method was used to express each image in Cartesian coordinates with a unique grid resolution. The resolution is $\lambda/2 = 0.3$ mm along the x-axis parallel to the probe and $\lambda/4 = 0.15$ mm along the z-axis perpendicular to it, where λ represents the wavelength of the ultrasound probe. Each view for every patient consists of at least one complete cardiac cycle to enable the manual annotation of cardiac structures at ED and ES. The manual annotations were provided only in the training folder for 450 patients (1800 images). The labeled cardiac chambers in each image are $\mathrm{LV}_{\mathrm{Endo}}$, $\mathrm{LV}_{\mathrm{Myo}}$, and LA (as presented in Fig4.4). CAMUS is a highly varied dataset in image quality (good, medium, and poor qualities) and disease cases. It contains clinical metrics ($\mathrm{LV}_{\mathrm{EDV}}$, $\mathrm{LV}_{\mathrm{ESV}}$, and $\mathrm{LV}_{\mathrm{EF}}$) for each patient. Additionally, it provides information about the age and sex of each patient..

4.3.2 Contrast enhancement using histogram equalization

In general, echocardiographic images often exhibit low contrast. We utilize the CAMUS dataset in this chapter to validate the proposed method. The dataset consists of images with different levels of quality, with 35% (good quality), 46% (medium quality), and 19% (poor quality). The percentage of medium and poor quality images is higher than than the good quality images percentage. This observation motivated us to investigate the application of histogram equalization as a contrast enhancement technique before applying deep learning algorithms for segmentation. In this section, we compare histogram equalization and other contrast enhancement methods.

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

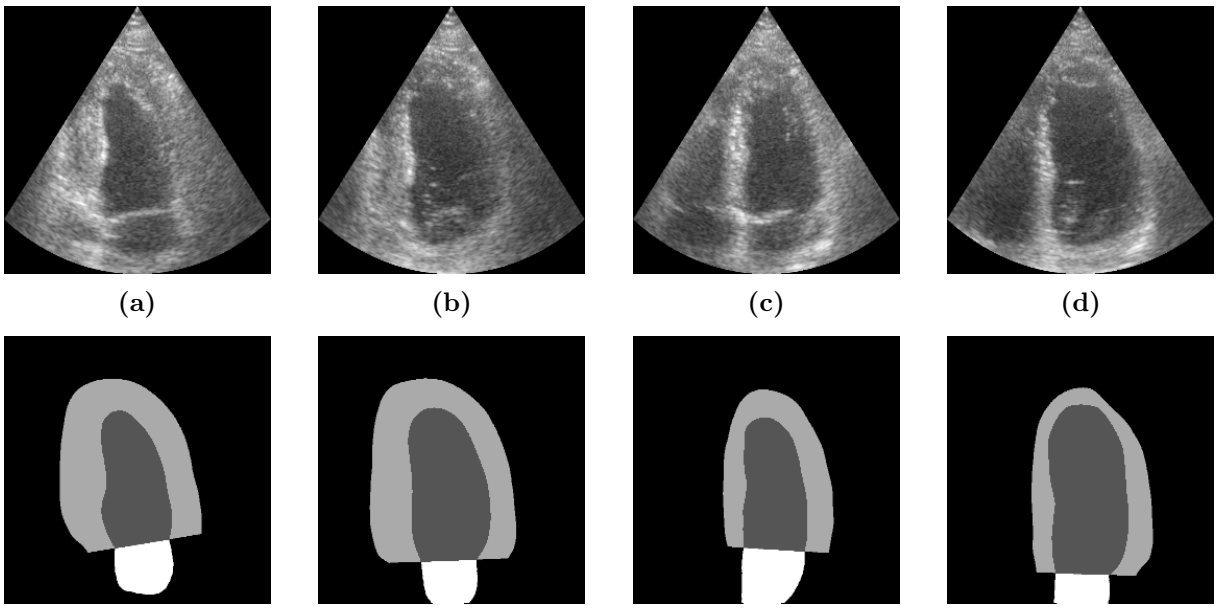


Figure 4.4: Typical images from the CAMUS dataset with their respective ground truths of the same patient. (a) image A2C in the ES. (b) image A2C in the ED. (c) image A4C in the ES. (d) image A4C in the ES. (e) image mask of (a). (f) image mask of (b). (g) image mask of (c). (h) image mask of (h). The outlined structures are $\mathrm{LV}_{\mathrm{Endo}}$ (dark gray), $\mathrm{LV}_{\mathrm{Myo}}$ (light gray), and LA (white).

4.3.2.1 Performance metrics

The performance metrics quantify the effectiveness of an image enhancement operation. They can quantitatively measure the features of an image. The performance of the pre-processing methods tested in this chapter is assessed using: Mean Squared Error (MSE), Peak Signal to Noise Ratio (PSNR), Structural Similarity Index (SSIM), entropy, Absolute Mean Brightness Error (AMBE), and visual appearance. Each evaluation metric is described as follows:

- Mean Squared Error (MSE): is the mean squared difference between the original image and the resulting image by an estimate. A higher value of MSE indicates a higher disparity between the original image and the processed image. MSE is defined by the following formula, where m and n are the width and height of the images, A is the enhanced image, B is the original image, and (i, j) is the row and column pixels of the original and enhanced images.

$$MSE = \frac{1}{m * n} \sum_{i=0,j=0}^{m-1,n-1} [A(i,j) - B(i,j)]^2 \tag{4.1}$$

- Peak Signal to Noise Ratio (PSNR): measures the ratio of the peak signal power (the maximum possible power of the signal) to the power of the distortion or noise introduced during the compression or reconstruction process. PSNR is usually expressed in decibels (dB) to account for the logarithmic scale of human perception. A higher PSNR value indicates less distortion or noise in the reconstructed image compared to the original, implying better quality. The PSNR is calculated based on the MSE, which quantifies the average squared difference between the pixel values of the original and reconstructed images. The formula for calculating PSNR is as follows:

$$PSNR = 20 \log_{10} \left(\frac{MAX_a}{\sqrt{MSE}} \right) \tag{4.2}$$

  Where MAX_a is the maximum pixel value of the image. It is 255 when the pixels are represented with 8 bits per sample.

- Structural Similarity Index Measure (SSIM): is a technique to measure the similarity between two images. It is based on the visible structures in the image and can quantify the degradation of image quality. The following equation defines the SSIM:

$$SSIM = \frac{(2\mu_x\mu_y + C_1)(2\sigma_{xy} + C_2)}{(\mu_x^2 + \mu_y^2 + C_1)(\sigma_x^2 + \sigma_y^2 + C_2)} \tag{4.3}$$

- Entropy: it is possible to describe the texture of the input image using entropy, a well-known statistical measure of randomness. It allows us to evaluate the degree of detail of the enhanced image. Higher entropy indicates better preservation of visual features. The following equation defines the entropy:

$$Entropy = \sum_{i=1}^{n} p_i \log_2(p_i) \tag{4.4}$$

  Where p_i value is the occurrence probability of a particular pixel.

- The Absolute Mean Brightness Error (AMBE): is a metric used to evaluate the quality of an image enhancement. It quantifies the absolute difference between the mean brightness of the original image and the enhanced image. A lower AMBE value indicates that the enhanced image has preserved the original brightness more accurately. The AMBE can provide insights into the preservation of brightness, but it may not be sufficient to evaluate the overall quality of an image enhancement algorithm. It is defined as follows:

$$AMBE = |E(A) - E(B)| \tag{4.5}$$

  where A and B the input and output images, respectively and $E(.)$ stands for the statistical mean.

- Visual appearance: by visually examining the enhanced image and comparing it to the original, one can directly observe and assess the differences in brightness, contrast, color, sharpness, and other visual attributes. Visual inspection allows for a more comprehensive evaluation of the overall improvement or degradation in image quality.

4.3.2.2 Evaluation

We compare histogram equalization to other contrast enhancement methods. The approaches are tested here to assess how much the optical appearance of the preprocessed images is optimized visually and digitally. These algorithms are developed to improve the contrast measurement directly. We evaluate and compare the following methods: contrast stretching, histogram equalization, CLAHE, and morphological operations. In the case of

morphological operations, we apply the technique presented by Kushol et al. [195]. The authors developed the approach to enhance different X-ray images. It consists of a combination of morphological transformations that are top-hat and bottom-hat transforms.

The contrast enhancement methods tested in this thesis are applied to all images of the CAMUS dataset. All these images are used for training and testing the deep learning methods. Each performance metric is calculated for each image. Then, the mean of the resulting values is computed. The comparison results of the image enhancement techniques on CAMUS images are depicted in Table 4.3.

By analyzing this table, the contrast stretching produces better results than other technics in terms of MSE, PSNR, and SSIM. However, the analysis based only on these metrics is not reliable in the case of image enhancement. The histogram equalization gives an SSIM of 80.61%, which outperforms the results of the CLAHE method. This result demonstrates that the images enhanced by the histogram equalization are similar to the original images at a percentage of 80.61% versus 64.40% with CLAHE. The richness of details in an image is assessed using the entropy performance metric. We can observe from the table that the histogram equalization performs the best with 5.15 of entropy, and the second best method is the CLAHE. The AMBE metric presents the performance of brightness preservation of the tested methods. The histogram equalization has a higher value of AMBE, which may be good, especially in ultrasound images, because they can have low brightness.

In addition, the visual appearance is a significant factor for evaluating the image quality and comparing it before and after the contrast enhancement. Figure 4.5 presents the qualitative results of applying image enhancement methods on different samples. The images illustrated in this figure present different echocardiographic views and have various qualities. From the visual inspection, contrast stretching does not introduce significant changes in the visual appearance of the images. The original and contrast-stretched images appear very similar. Morphological operations do not seem to improve the visual quality of the images and generate noisy pixels, which can be visually distracting. For the CLAHE method, an improvement in the contrast of the images is evident. However, the histogram equalization enhances the images' contrast better than the morphological transformations and allows more visibility of the LV structure. Overall, based on the visual appearance, we can conclude that histogram equalization is a more effective contrast enhancement method than contrast stretching, morphological operations, and CLAHE for the echocardiographic images in this dataset.

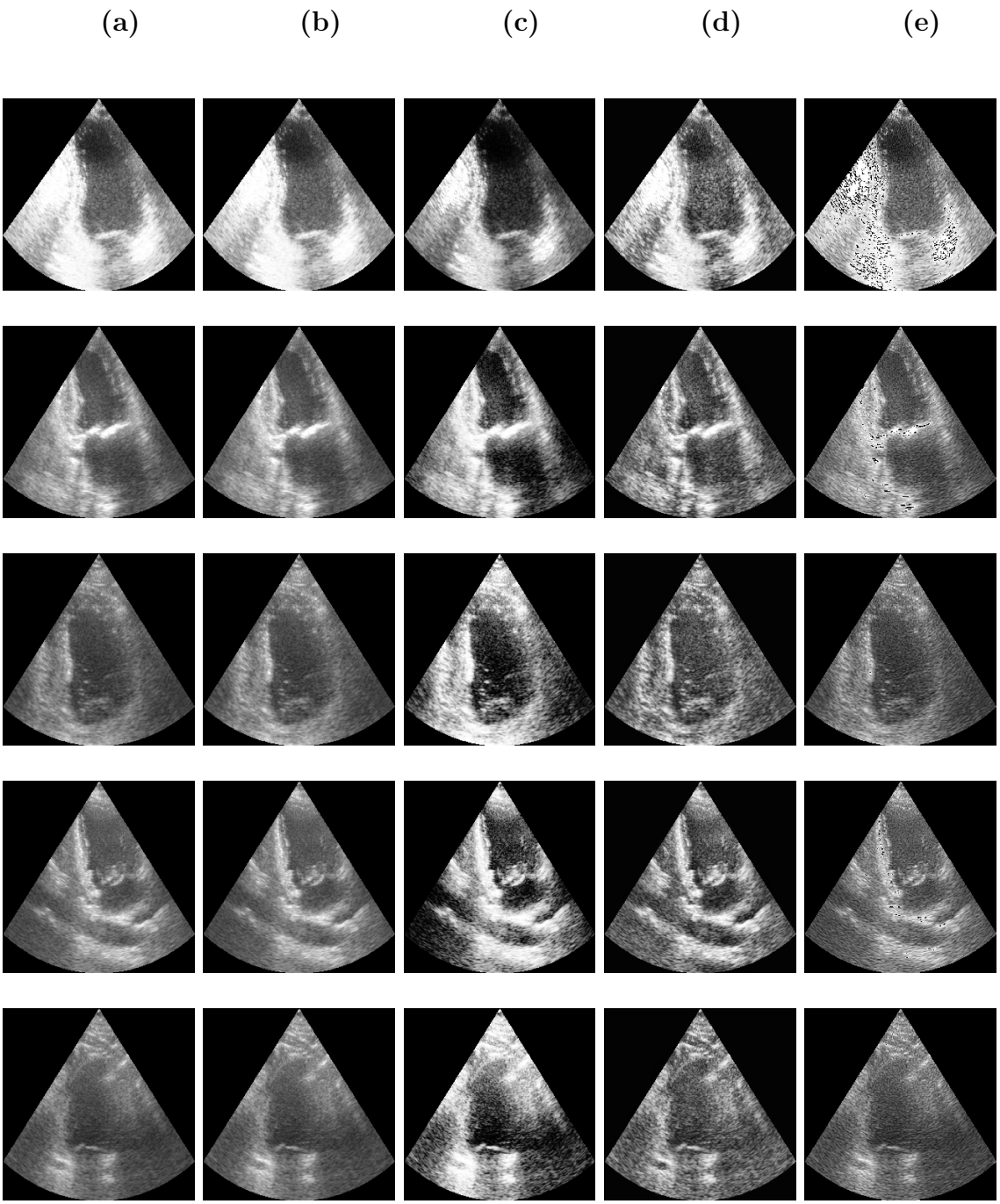


Figure 4.5: Contrast enhancement of different images taken from CAMUS dataset. (a) Original images. (b) Contrast stretching. (c) Histogram equalization. (d) CLAHE. (e) Morphological operations.

Table 4.3: Comparison of four contrast enhancement techniques on images of CAMUS dataset

Performance metric	Contrast stretching	Histogram equalization	CLAHE	Morphological operations
Mean MSE	16.55	180.90	430.27	154.71
Mean PSNR	70.74	30.57	30.65	33.73
Mean SSIM	99.78	80.61	64.40	91.04
Mean Entropy	4.82	5.15	4.93	4.72
Mean AMBE	1.43	20.98	10.28	0.38
Visual appearance	Not Good	Very Good	Good	Not Good

4.3.3 Segmentation of Left Ventricle structure on CAMUS

4.3.3.1 Experiment setup

In this section, the experiments were implemented in Python using Tensorflow[2] and Keras[3] libraries. The networks were trained for 250 epochs with batch size of 8 on a Linux workstation equipped with a double Intel Xeon 2.2GHz, 3 GHz CPU, and two 24Go Nvidia Quadro P6000 GPUs. Adam optimizer [196] with a learning rate of 1e-4 was used. The weights were initialized with glorot-uniform [197] and the Dice objective function was used to minimize the parameters. In the last prediction layer, the Sigmoid activation function was utilized.

4.3.3.2 Distribution of the dataset

The CAMUS dataset consists of 500 patients of A4C and A2C view sequences with the complete cardiac cycle. The manual annotation by experts of each view sequence in the ED and ES from 450 patients is available. However, the remaining view sequences from 50 patients are not annotated. The total number of annotated frames in this dataset is 1800. Our proposed framework is based on deep learning, which requires the data and the annotation by experts as input. Thus, we use the 1800 labeled images.

The authors in [18] proposed a dataset division for the standard cross-validation. All images were divided into 10 folds. Using the same distribution in terms of image quality and LV_{EF} as the entire dataset, each fold has 50 patients. The 450 patients (9 folds) were

[2] https://www.tensorflow.org/?hl=fr
[3] https://keras.io/

utilized for the training/validation stages of the machine learning algorithms. 8 folds (400 patients) were employed for training and 1 fold (50 patients) for validation to optimize the parameters. The testing phase was conducted using the remaining sub-sample.

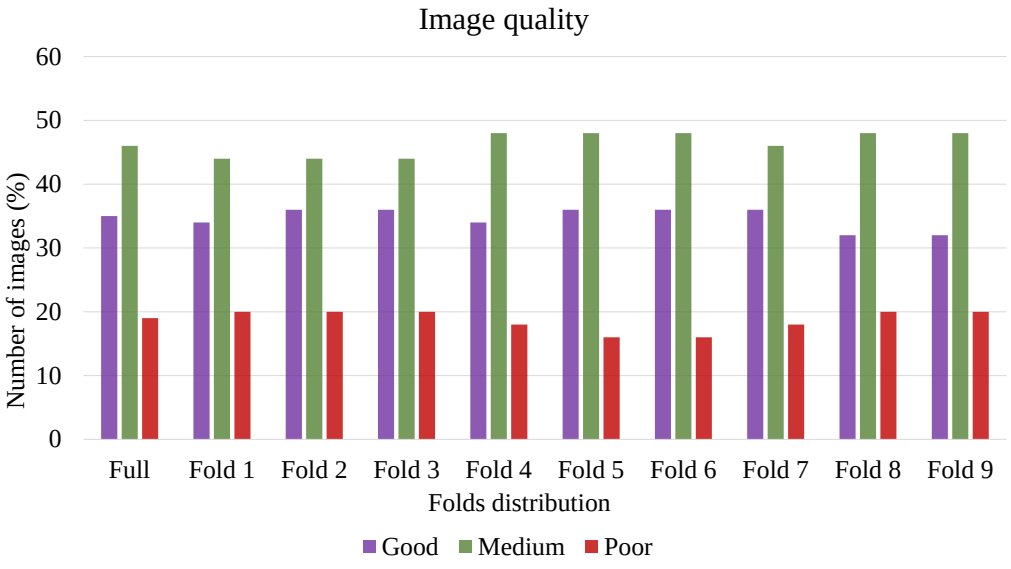
On the other hand, Zyuzin et al. [151] followed the same division technique. Nevertheless, they split the data into 9 folds of 50 patients each because there are only 450 patients in open access. The results of this distribution are presented in Figure 4.6. The authors announced that the partitioning of images by quality criteria is similar to the division result in [18]. However, the distribution according to the LV_{EF} was noticeably different. In addition, the graphs in Figure 4.6 show that the division allows having the same distribution of the image quality in each fold, which is not the case in terms of the LV_{EF}. Hence, the 9 folds have the same partitioning only according to the image quality as the dataset had.

In this thesis, we manually divide the CAMUS dataset following the strategy proposed by Zyuzin et al. [151] because we use the available data of the 450 patients. These folds allow us to apply k-fold cross-validation (k=9) for evaluating deep learning models. Each fold is used for validation, while the remaining 8 folds are used to train the model in case of cross-validation.

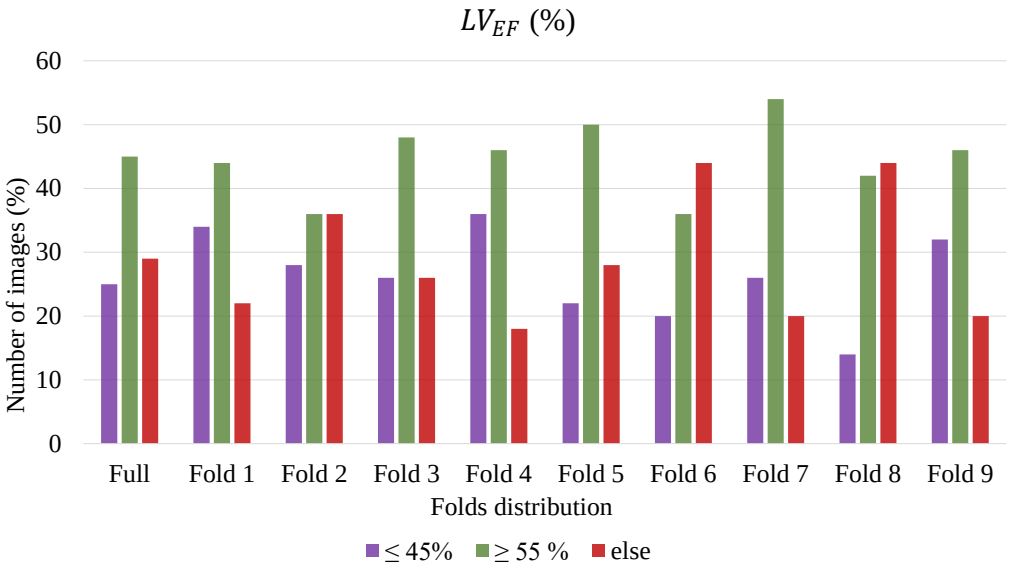
4.3.3.3 Quantitative evaluation

Table 4.4 shows the average DSC and the HD results of the 9 models created for each method. The results are presented according to the two cardiac cycle periods (ED/ES). For a fair comparison, these models are trained with the same parameters. Note that poor-quality images are not eliminated during training and validation. We obtain a mean DSC of 0.939/0.916 for ED/ES in the case of attention U-Net 1 versus 0.928/0.899 for ED/ES with U-Net 1. In addition, the mean Dice of attention U-Net 2 is 0.940/0.919 for ED/ES versus 0.930/0.907 for ED/ES with U-Net 2 without attention. Also, the best results in terms of HD are reported when both U-Net 1 and 2 are implemented with attention gates.

Furthermore, we present the results of the standard deviation in Table 4.4. This measure can reveal the dispersion of the data. A low standard deviation implies that the data are clustered around the mean, while a high standard deviation indicates that the data are spread out more from the mean value. In other words, a high or low standard deviation indicates that the data points are above or below the mean. A standard deviation near zero implies that the data points are close to the mean value. The standard deviation is calculated for each observation in the data set. Therefore, it is a sensitive measure, as outliers can affect it. An outlier is an observation abnormally different from other values



(a)



(b)

Figure 4.6: Distribution of CAMUS dataset for 450 patients based on the main characteristics. (a) According to the image quality. (b) According to the LV_{EF}.

in a random population-based sample. We see in Table 4.4 that the attention modules improve the values of the standard deviation of the two models. Particularly, the standard deviation of HD of U-Net 2 and attention U-Net 2 are 2.98/3.68 mm and 1.92/1.86 mm in ED and ES, respectively.

Table 4.4: Comparison of DSC and HD metrics in ED and ES expressed as (mean ± standard deviation) for 9-fold cross validation of the four models.

Network	ED		ES	
	DSC	**HD (mm)**	**DSC**	**HD (mm)**
U-Net 1 [18]	0.928±0.040	6.16±2.4	0.899±0.061	6.32±2.34
Attention U-Net 1	0.939±0.029	5.25±1.66	0.916±0.048	5.44±2.04
U-Net 2 [18]	0.930±0.040	5.86±2.98	0.907±0.068	5.84±3.68
Attention U-Net 2	0.940±0.031	4.11±1.92	0.919±0.049	5.45±1.86

4.3.3.4 Statistical results

Statistical analysis is essential in the interpretation of the quantitative results. Therefore, we present the obtained results in the form of box plots which can provide a visual representation of the data distribution. A box plot is also known as a whisker plot based on a five-number summary: the minimum, first quartile, median, third quartile, and maximum. The box plots of Figure 4.7 present the DSC of validation from fold 1 data for a more detailed interpretation of the results. To determine the influence of the quality images and the cardiac cycle frames (ED or ES) on the segmentation results of the investigated models, each sub-figure in Figure 4.7 presents the segmentation results on different images. The Figures (4.7a/4.7b/4.7c/4.7d) illustrate the box plots computed from the DSC of each model for the following images' type (ED with good and medium quality/ES with good and medium quality/ED with poor quality/ES with poor quality), respectively.

We can observe that adding attention improves median performance, especially with U-Net 1 in both ED and ES. Moreover, the attention units improve the segmentation of U-Net 2, especially in the case of poor images in ED and ES. Conversely, there is not a noticeable improvement in the case of the median values of U-Net 2. However, they tighten the interquartile of the box plots of good and medium images in ED and ES. Accordingly, we can say that the attention units are more effective with U-Net 1 in all

validation samples and U-Net 2 in case of poor images. However, they give more consistent segmentation over the good and medium quality with attention U-Net 2 compared to U-Net 2 without attention.

Additionally, we corroborate these results with the radar chart in Figure 4.8 for a more detailed results presentation. The difference between the images in the CAMUS dataset is not only in terms of ED and ES frames or image quality but also regarding the type of acquisition of the cardiac view. The CAMUS dataset includes A2C and A4C view acquisition. Figure 4.8 presents the DSC of the segmentation performance of some images for each network. These images are taken randomly from fold 1 and are different in terms of cardiac cycle period (ED/ES), view chambers (2CH, 4CH), and image quality (good, medium, and poor). From the values along each axis, attention U-Net 1 and attention U-Net 2 give the best scores in most cases. Attention U-Net 2 doesn't perform well, especially with the ED images of A2C view of good, medium, and poor quality. However, the improvement of DSC with this network is remarkable with A4C images.

4.3.4 Additional experiments

4.3.4.1 Influence of the size of the training dataset

The performance of a machine learning model is thought to be significantly influenced by the size of the training dataset. Specifically, deep learning algorithms need big data to perform well. In general, more training data increases the performance of these models, but a lack of training data can give a poor approximation and may lead to the overfitting problem. However, data collection is the most challenging step, mainly in medical imaging. Moreover, data annotation in the medical domain must be carried out only by experts. This task is very rough and takes more time. For these reasons, it is advantageous to develop a deep learning model that performs effectively and generalizes better with less training data. In this section, we analyze the impact of modifying the size of the training set on the performance of a U-Net architecture containing attention mechanisms.

Herein, We investigate the model which gave the best results on segmenting the $\mathrm{LV}_{\mathrm{Endo}}$ in DSC and HD using 9-fold cross-validation, which is attention U-Net 2. We choose fold 1 for the testing phase and the other folds for training the model. We realize four experiments by modifying the size of the training set each time as follows:

- Experiment 1: fold (1) as the test set; folds (2 and 3) as the training set (25% of the training dataset).
- Experiment 2: fold (1) as the test set; folds (2, 3, 4, and 5) as the training set (50% of the training dataset).

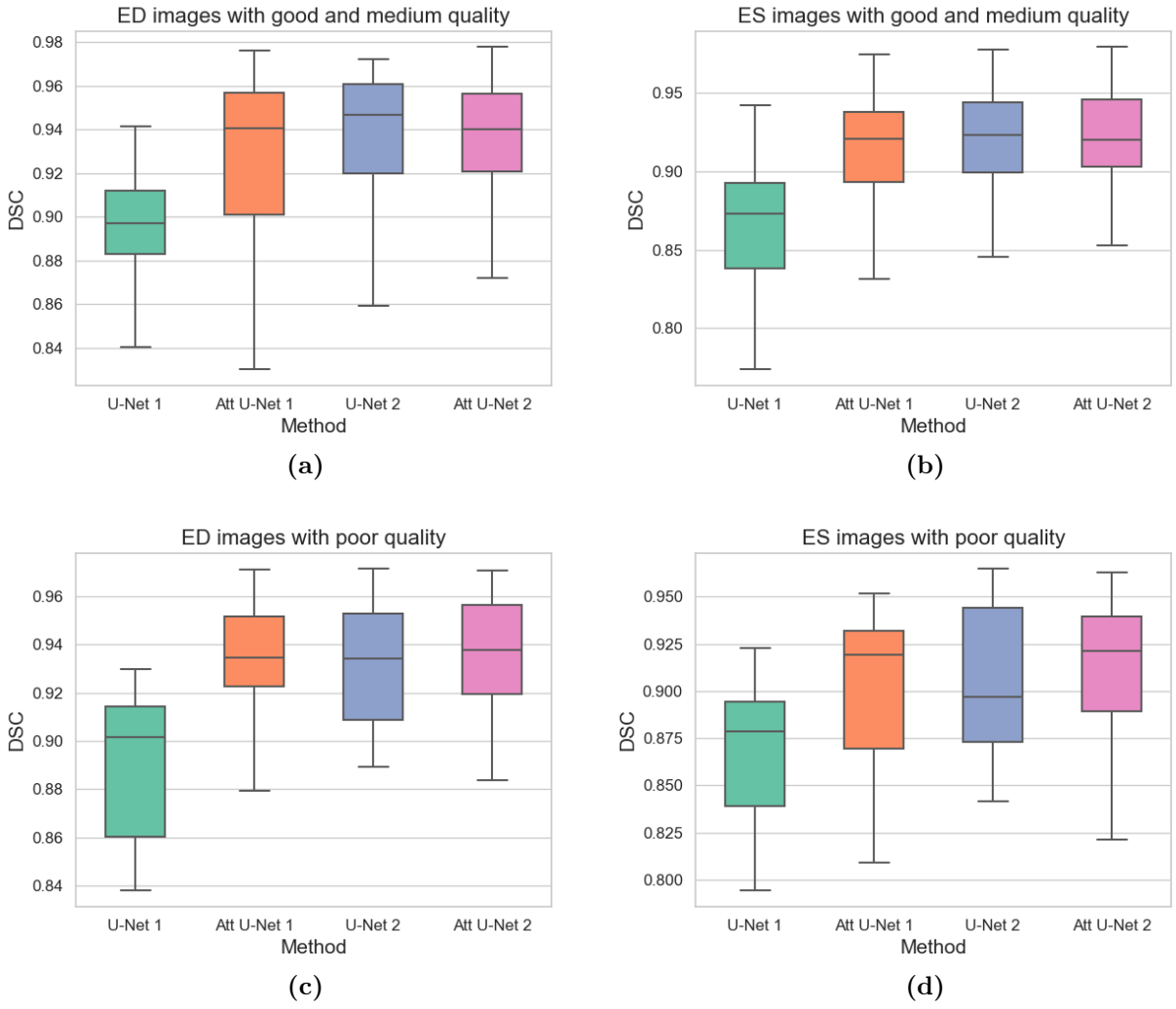


Figure 4.7: Results of DSC box plots of the networks from fold 1 in ED and ES. (a) DSC box plots on ED images having good and medium quality. (b) DSC box plots on ES images having good and medium quality. (c) DSC box plots on ED images having poor quality. (d) DSC box plots on ES images having poor quality.

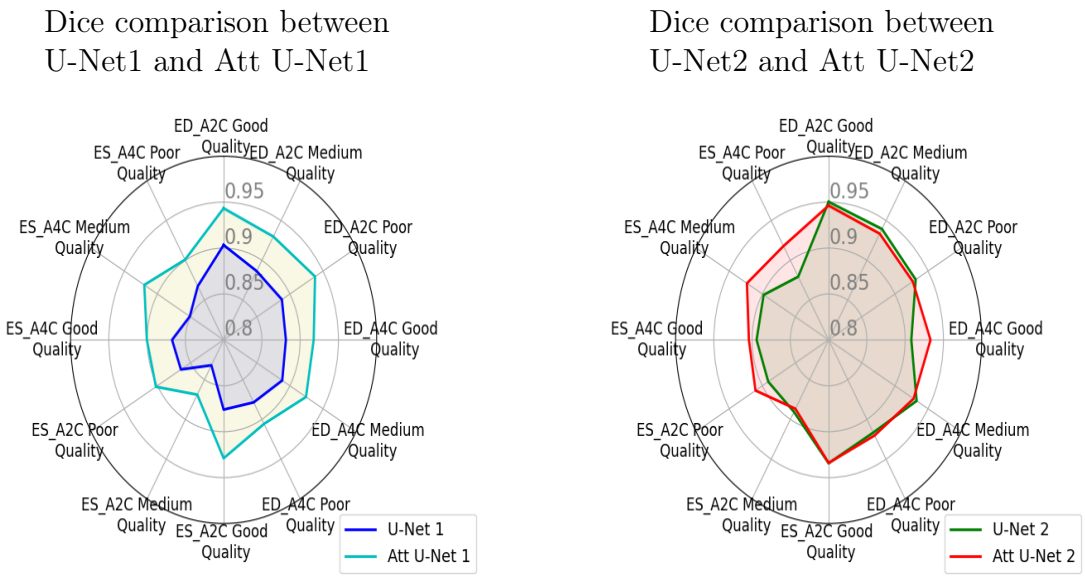


Figure 4.8: Radar chart presenting Dice coefficient results of segmentation of different samples from fold 1.

- Experiment 3: fold (1) as the test set; folds (2, 3, 4, 5, 6, and 7) as the training set (75% of the training dataset).
- Experiment 4: fold (1) as the test set; folds (2, 3, 4, 5, 6, 7, 8, and 9) as the training set (100% of the training dataset).

Table 4.5 reveals the DSC and HD of attention U-Net 2 with various training set sizes in ED and ES results. For a fair comparison, each time attention U-Unet 2 is retrained from scratch and tested on the same fold, which is fold 1. We see in Table 4.5 that the results of 25% are the worst. In general, the improvement of the segmentation metrics is noticed by increasing the number of images in the training set. Specifically, the difference is more pronounced between 25% and 100% training sets (for instance, an increase of DSC in ED from 0.917 to 0.941 and HD in ES from 8.53 mm to 5.20 mm. However, there isn't much difference between 75% and 100% training sets.

The results are also plotted in Figure 4.9 for a more elaborate representation. Our findings are that a 100% training set gives the best results in terms of DSC in ED and ES. By increasing the size of the training set, the DSC results improve. The 75% training set gets HD results comparable to what is obtained with 100% in ED and ES. The interquartile ranges and whiskers of the box plots of the 25% training set are longer than the others, especially in HD. This observation demonstrates that the 25% fragment used as training data gives more dispersed results from the mean values, which proves the high standard deviation values. In this experiment, the minimum values of the median in DSC box plots

Table 4.5: Comparison between different pieces of training set used when training attention U-Net 2 model in ED and ES.

Training set size	ED		ES	
	DSC	HD (mm)	DC	HD (mm)
25%	0.917±0.055	6.48±5.52	0.900±0.091	8.53±8.58
50%	0.923±0.044	4.97±2.39	0.911±0.063	5.82±3.09
75%	0.929±0.040	4.36±1.92	0.911±0.076	5.45±2.08
100%	0.941±0.042	4.36±1.86	0.915±0.093	5.20±1.98

and the maximum median values in HD box plots demonstrate the worst results of the 25% fragment presented in Table 4.5.

4.3.4.2 Influence of the localization of the LV region

In this sub-section, we study the influence of the localization of the LV on the accuracy of the segmentation results produced by attention U-Net 2 architecture. Herein, the localization of the LV_{Endo} involves applying the crop operation around the region containing the LV_{Endo} in the images of the CAMUS dataset. Image cropping is a technique that removes irrelevant components while keeping a portion of the image containing the target object. Image cropping aims to improve the overall composition and achieve better visual perception. Furthermore, It authorizes the investigation of the contextual elements in terms of an area of interest in the image.

Before training the network, we apply cropping to all images and masks of the training and testing sets. After that, the images are resized to 256×256 dimensions. The localization of the LV region is applied using the concept of a bounding box. Therefore, we apply the bounding box on the label images to precisely crop the desired region of interest using its boundaries. This strategy allows us to designate and extract the pixels representing the LV. The same bounding box obtained from the masks is applied to the original images to select the LV region with the same number of pixels. In this section, we examine the influence of the localization technique on the results of LV segmentation through four experiments. In each experiment, we apply a different value of margin. These values represent the number of pixels between the maximum or minimum pixels of the region containing the LV and the pixels representing the boundary of the cropped images. The added margin offers context around the desired LV region for the segmentation task. Al-

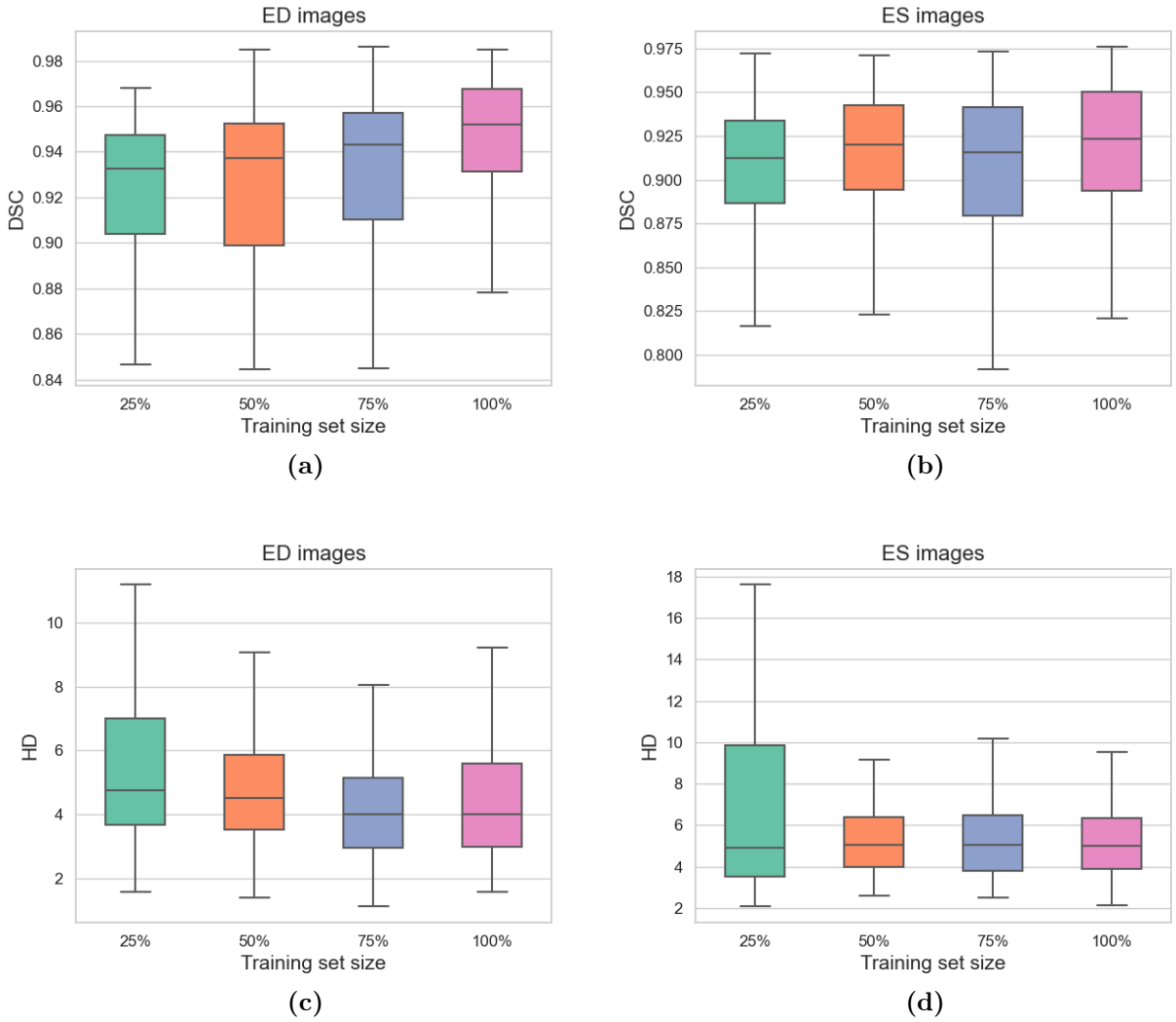


Figure 4.9: Box plots of attention U-Net 2 performance by modifying the training set size. (a) Box plots of Dice coefficient in ED. (b) Box plots of Dice coefficient in ES. (c) Box plots of Hausdorff distance in ED. (d) Box plots of Hausdorff distance in ES.

gorithm 4.1 provides a detailed description of the function, which we developed to crop CAMUS images IMG based on their ground truths GT. The values of the margins M investigated in the experiments are 0 (no margin), 10, 30, and 50 pixels. Figure 4.10 illustrates how the LV structure is localized in this thesis with different margin values in the original and mask images.

Algorithm 4.1: Algorithm for Cropping Original Images Based on Their Ground Truths

Input: Original Images IMG, Ground Truths Images GT, Margin M
Output: Cropped Original Images IMG_C, Cropped Ground Truths Images GT_C
Function `crop_img_and_gt`(IMG, GT, M)**:**
  foreach $img_i \in IMG$ & $gt_i \in GT$ **do**
    $(X, Y) \longleftarrow$ coordinates of pixels where $(gt_i > 0)$;
    if $M = 0$ **then**
      $maxx \longleftarrow \max(X)$;
      $maxy \longleftarrow \max(Y)$;
      $minx \longleftarrow \min(X)$;
      $miny \longleftarrow \min(Y)$;
    else
      $maxx \longleftarrow \max(X) + M$;
      $maxy \longleftarrow \max(Y) + M$;
      $minx \longleftarrow \min(X) - M$;
      $miny \longleftarrow \min(Y) - M$;
    end
    $img\_c_i \longleftarrow img\_c_i(minx : maxx, miny : maxy)$;
    $gt\_c_i \longleftarrow gt\_c_i(minx : maxx, miny : maxy)$;
  end
  return IMG_C, GT_C;
End Function

Table 4.6 exhibits the results of DSC and HD of attention U-Net 2 with various margin values applied in ED and ES images. The evaluated margins are labeled in the table as follows: M-0 (no margin), M-10 (margin = 10 pixels), M-30 (margin = 30 pixels), M-50 (margin = 50), and w/o loc indicates that we don't use any localization strategies. The results present the testing outcomes of the images of fold 1. Each time attention U-Unet 2 is retrained from scratch for a fair comparison. From this table, we see that the results of DSC and HD are not consistent. The best result of DSC in ED is when applying the cropping without margin. In ES images, the favorable DSC result is in the case of the M-30 cropping. It gave 0.967 and 0.952 in ED and ES, respectively. However, attention U-Net 2 performs well in terms of HD when we don't use the cropping technique. The results are 4.36 mm and 5.20 mm in ED and ES, respectively.

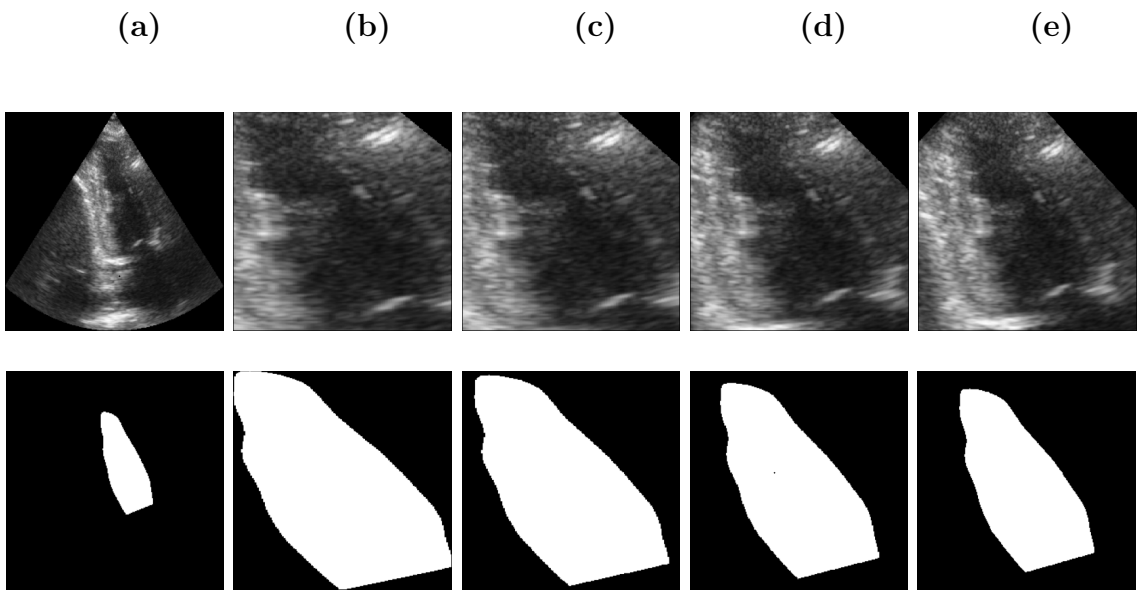


Figure 4.10: An example of the localization of the left ventricle structure. [Top of the figure]: original images. [Bottom of the figure]: corresponding ground truth images. (a) Images without cropping. (b) Images with cropping and margin = 0. (c) Images with cropping and margin = 10. (d) Images with cropping and margin = 30. (e) Images with cropping and margin = 50.

Figure 4.11 demonstrates the observations of the obtained results presented in Table 4.6. The median values and the length of the interquartile of the box plots of the DSC results in the case of w/o loc are the worst, while they are the best in terms of HD in both ED and ES.

To give more precise results because of the inconsistency in the results of the segmentation parameters (DSC and HD) used for evaluating the performance of attention U-Net 2. We must add another effective evaluation metric to represent our findings. This metric is the ROC graphical plot. This curve can show the binary segmentation ability of the proposed model. It denotes the trade-offs between the TPR and FPR. The estimated ROC curves with the AUC for the images of fold 1 with different margins are displayed in Figure 4.12. The obtained results of the ROC curves follow those of the HD. The mean AUC scores of 0.9667 and 0.9598 in ED and ES, respectively, indicate that the optimal performance of the proposed model is when we don't apply the localization of the LV in the images of the dataset. However, the attention U-Net 2 is the least efficient when using the bounding box around the LV without any margin in ED and ES.

4.3.4.3 Influence of the deep supervision

In this section, we apply the deep supervision strategy to the proposed model to study its influence on the segmentation of LV structure in echocardiographic images. As in the

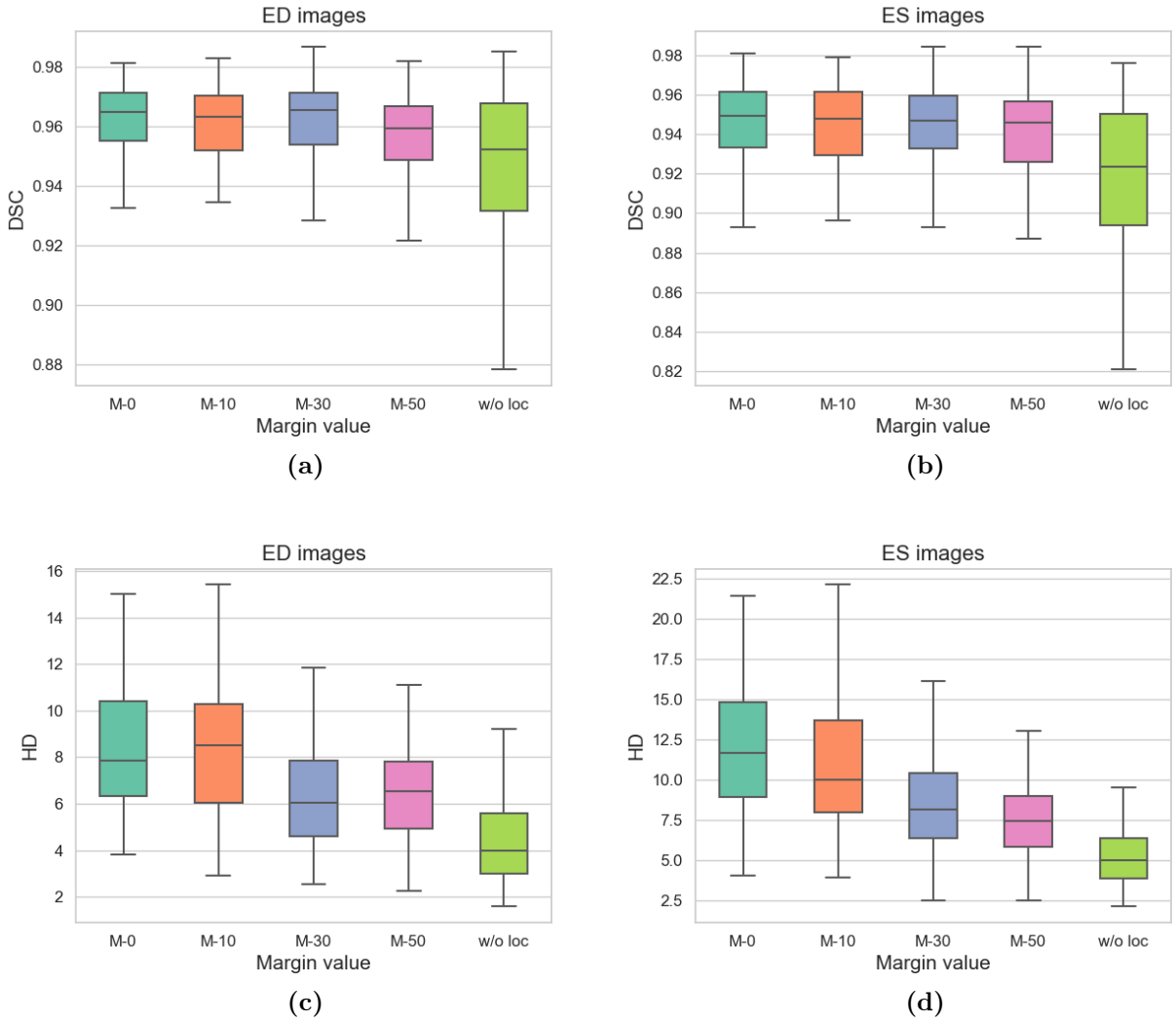


Figure 4.11: Box plots of attention U-Net 2 performance by modifying the margin size applied for the LV localization. (a) Box plots of Dice coefficient in ED. (b) Box plots of Dice coefficient in ES. (c) Box plots of Hausdorff distance in ED. (d) Box plots of Hausdorff distance in ES.

Table 4.6: Comparison between different margin values used to localize the LV region when training and testing attention U-Net 2 model in ED and ES.

Margin size	ED		ES	
	DSC	HD (mm)	DC	HD (mm)
M-0	0.967±0.015	8.48±2.63	0.952±0.022	12.90±6.13
M-10	0.960±0.016	8.33±3.01	0.952±0.023	11.25±4.60
M-30	0.962±0.014	6.29±2.07	0.953±0.020	8.53±3.26
M-50	0.957±0.015	6.58±2.12	0.948±0.023	7.73±3.09
w/o loc	0.941±0.042	4.36±1.86	0.915±0.093	5.20±1.98

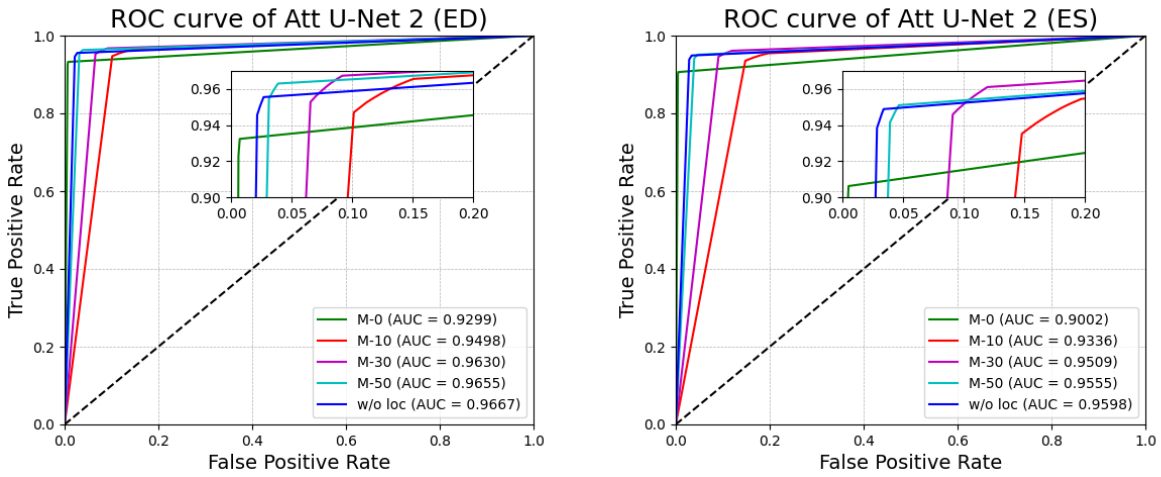
previous experiments, we keep fold 1 as the test set while the other folds are for training the networks. Table 4.7 presents the results of the proposed network with and without deep supervision in terms of DSC and HD in ED and ES. From this table, we can notice that the segmentation results without the deep supervision technique outperform those with it, except for the standard deviation of HD in ED and ES. There isn't a noticeable improvement in mean DSC and mean HD. These results can reveal the ineffectiveness of deep supervision operation with attention U-Net 2 for LV segmentation.

Table 4.7: Comparison of segmentation accuracy for attention U-Net 2 with deep supervision.

Att U-Net 2 network	ED		ES	
	DSC	HD (mm)	DC	HD (mm)
w/ deep supervision	0.941±0.043	4.54±1.24	0.914±0.093	5.64±1.35
w/o deep supervision	0.941±0.042	4.36±1.86	0.915±0.093	5.20±1.98

4.4 Discussion

A methodology for automatic LV segmentation in echocardiographic images is explored in this chapter. The first step involves preprocessing to improve image quality before training the segmentation models. Image enhancement techniques are applied to enhance



(a) **(b)**

Figure 4.12: Comparison of ROC curves of attention U-Net 2 by modifying the margin value each time. (a) ROC curve of attention U-Net 2 in ED. (b) ROC curve of attention U-Net 2 in ES.

the contrast of the images, allowing for more accurate and successful segmentation. The performance of various contrast enhancement methods is evaluated using different metrics and visual appearance assessment. Among the techniques tested, histogram equalization yielded the best results. It effectively enhances the contrast of the echocardiographic images, particularly those of poor quality. Histogram equalization is a simple yet powerful technique that modifies the overall distribution of pixel intensities, unlike methods such as contrast stretching that scale the intensity range. The histogram equalization as preprocessing can significantly improve the visual appearance of echocardiographic images and enhance their contrast. This improvement is crucial for subsequent LV segmentation tasks, as it aids in accurately delineating the boundaries of the LV. The simplicity and effectiveness of histogram equalization make it a viable option for contrast enhancement in echocardiographic image processing.

A fully automated technique based on attention mechanisms was proposed to address LV segmentation. Attention gates were added to the skip connections of the U-Net 1 and U-Net 2 architectures. These two networks were optimized and adapted for CAMUS images. Table 4.4 demonstrates that attention U-Net 2 gives the best results with 9-fold cross-validation. Moreover, it outperforms the inter-observer variability and existing approaches reported on the CAMUS dataset. Table 4.8 shows the DSC results on ED and ES. Figure 4.13 qualitatively shows the attention U-Net 2 and other investigated networks. It exhibits the results of the contour predictions for the LV by the two architectures with

and without attention mechanism of four different sample subjects (a, b, c, and d). It indicates that the predicted LV boundary with the attention units into U-Net 1 and 2 matches the annotation boundary by the expert more accurately. The LV boundaries with U-Net 1 are less smooth than those predicted by U-Net 1 with attention gates. We explain this observation by the fact that the attention modules allow the U-Net 1 to converge better and thus predict the contour of the LV more precisely. This way, the results of the architectures containing attention gates to focus on the target regions are more interesting than the original architectures.

Table 4.8: Dice result of Attention U-Net 2 comparing with the state of the art in ED and ES jointly.

Method	**DC**
inter-observer	0.896 ± 0.047
U-Net ++ [198]	0.912 ± 0.048
ACNN [199]	0.915 ± 0.041
UltraGAN [174]	0.924 ± 0.051
SegAN [200]	0.917 ± 0.071
Attention U-Net2	**0.930 ± 0.040**

The influence of the size of the training dataset on the attention U-Net 2 performance was conducted. The CAMUS training data was split into different subsets. These experiments reveal that using few data in the training set is insufficient for deep learning model convergence. The high standard deviation values presented in Table 4.5 in the case of using only 25% of the training set highlight how challenging to segment the images of the test set consistently using the proposed model. Moreover, Figure 4.9 tells us that increasing the size of the training set will not make a difference at a given time. For instance, there isn't a big difference between 75% and 100% training set sizes.

The experiments conducted in Section 4.3.4.2 provide insights into the impact of LV localization on the segmentation results of the attention U-Net 2 architecture. Surprisingly, the results indicate that the localization of the LV structure does not lead to improved segmentation performance with this particular architecture. One possible explanation for this observation is that attention mechanisms focus on small objects within an image. These mechanisms learn to selectively amplify important features and suppress irrelevant regions, enhancing the model's ability to capture intricate details. However, the

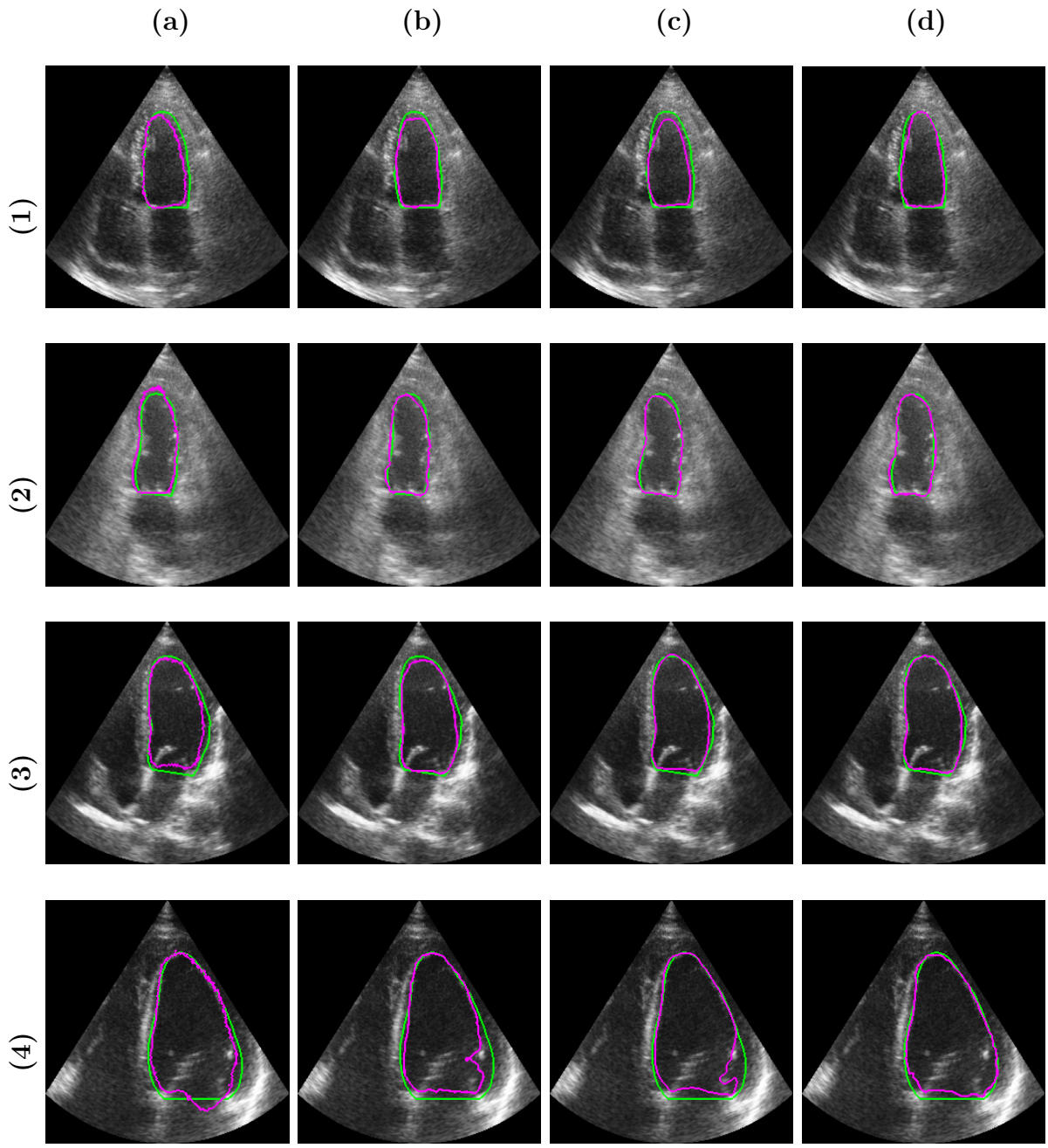


Figure 4.13: Visual segmentation comparison between the 4 networks for three different subjects taken from fold 1. (a) U-Net 1 (b) Att U-Net 1 (c) U-Net 2 (d) Att U-Net 2. The green curve is the reference annotation with the cardiologist, and the magenta curve is the prediction result of each architecture.

attention modules already adequately highlight the desired area, so the additional localization operation of the LV structure in B-mode echocardiographic images is redundant and avoidable. The effectiveness of the localization operation is often prominent when the cropped region includes multiple objects of interest. In such cases, localizing specific structures helps guide the attention mechanism toward the relevant areas and improves the segmentation results. However, in the case of a network that already incorporates attention modules and is designed to focus on relevant features, the additional localization step may not provide significant benefits. Based on these findings, localizing the LV structure in B-mode echocardiographic images may not be necessary when using a network architecture incorporating attention modules. Therefore, combining attention mechanisms with explicit localization steps may not provide additional advantages and could introduce unnecessary complexity to the model.

The experiment conducted to investigate the impact of deep supervision on the proposed model for LV segmentation reveals that it does not improve the segmentation performance. One possible explanation for the lack of improvement with deep supervision is the use of batch normalization in the proposed architecture. Batch normalization is a regularization technique that helps mitigate the effects of overfitting and stabilize the training process by normalizing the activations within each mini-batch. It is designed to address the same issues of gradient instability that deep supervision aims to alleviate. Given that the proposed architecture already incorporates batch normalization, which provides regularization and gradient stabilization, the additional deep supervision signals may not offer many advantages. In this case, the optimal implementation of the attention U-Net 2 architecture would be to retain the final output of the decoder as the final result of the segmentation. This implementation simplifies the model and avoids the need for intermediate objectives, allowing the attention mechanisms to focus on optimizing the segmentation task without the added complexity of deep supervision. It's worth noting that the lack of improvement with deep supervision in this specific experiment does not diminish its potential effectiveness in other contexts or for different tasks. Deep supervision can still be a valuable technique in other scenarios where it aids in improving gradient flow, regularization, or handling specific challenges associated with training deep neural networks.

4.5 Conclusion

In this chapter, we applied the attention mechanism to the U-Net 1 and 2 architectures to examine their performance for the LV contour delineation. U-Net 1 and U-Net 2 are

derived from the original U-Net architecture. They have been adapted and improved to produce the best cardiac structure segmentation results on the CAMUS dataset. The results demonstrated that the insertion of attention units into the skip connections of these two architectures improves LV segmentation. The established framework consisting of attention U-Net 2 outperformed the methods proposed in the literature. Experiments were also conducted on the effects of the training set size, the localization of the LV, and the deep supervision technique on the network performance. The performance of a deeper neural network degrades with few training data. Moreover, the suggested model showed its resilience and ability to segment the LV region without any localization scheme or deep supervision design.

Chapter 5

Echocardiographic images analysis for Left Ventricle assessment with transfer learning

5.1 Introduction

Recently, the use of deep CNN, a development in computer vision technology, for medical image segmentation has been considered a significant contribution to medical image analysis. It is necessary to find the optimal values of the filters of a CNN during training to recognize the correct class of an input image after it has passed through all its layers. Unfortunately, this task can require considerable data and computing power. In most domains, the data available for deep CNN learning is scarce, especially in medical images. Therefore, the main challenge is to develop computer-aided diagnostic systems with limited available data.

Transfer learning strategy can mitigate the performance. It is convenient for medical image segmentation, especially when few samples are available to train convolutional neural networks. In this chapter, our goal is to segment the LV region in B-mode echocardiographic images and to analyze its function with transfer learning. Due to the limited data available in this domain, we concentrate on transfer learning, a widely relevant strategy to lessen the requirement for annotated data. Transfer learning applies previously learned information to solve new, related issues quickly and effectively. It makes it possible to train the model on less data, which makes it very beneficial for segmenting medical images. We test this strategy on U-Net architecture and other networks derived from it. We choose U-Net as the foundational architecture for the proposed framework due to its proven effectiveness in medical image segmentation tasks. U-Net has demonstrated superior per-

formance compared to other segmentation architectures in various studies. Furthermore, it is an end-to-end, fully convolutional network that enables effective learning of an entire image.

An overview of the proposed framework and the main contribution of this study will be highlighted in Section 5.2. Then, the experimental design and findings will be presented in Section 5.3, followed by a discussion of the results in Section 5.4. Finally, the conclusion will provide a summary of the relevant points of the chapter.

5.2 Methods and procedure

This part of the work presents the methodology and procedure we suggest for the echocardiographic image analysis to assess the LV structure.

5.2.1 Transfer learning

Transfer learning is a machine learning technique that allows reusing a model created for one task as the basis for another. It helps AI systems to apply the knowledge they've learned from one task (source task) to another (target task). It aims to utilize samples, models, or model parameters gained while solving the problem in the source domain to solve another related but different one in the target domain [201]. This approach requires fewer resources and less labeled data relevant to train new models for the target task. It is primarily used to enhance the algorithm's generalizability and overall effectiveness. In addition, it can speed up the overall process of training the new model with the transferred knowledge. Transfer learning has been widely applied to several medical imaging applications, including, but not limited to, segmentation, object identification, and disease categorization [202].

Pratt et al. [203] demonstrated the effectiveness of transfer learning in neural networks for the first time. It was applied later to solve problems in computer vision [204]. Among transfer learning methods, there is a method that adapts a pre-trained model to a new task called fine-tuning. This model has been previously trained using a large dataset from another domain, such as the ImageNet[1] dataset [205]. ImageNet is a large dataset of images with annotations created for computer vision research. This dataset was created to provide benchmarking data to support research and to implement enhanced computer vision algorithms [206]. The ImageNet Large Scale Visual Recognition Challenge is an annual competition using subsets of the ImageNet dataset to encourage the development

[1] https://image-net.org/

of better computer vision techniques. The typical challenge tasks for most years are image classification, single-object localization, and object detection.

Many competition models have been made available to be used as starting points for transfer learning in numerous computer vision applications. The reuse of these models offers several advantages. They have acquired the ability to recognize generic features due to being trained on over 1,000,000 images across 1,000 classes, resulting in excellent performance. Additionally, numerous libraries provide straightforward application programming interfaces (APIs) for obtaining and directly using various standard pre-trained models. The model weights are available as free downloads in different deep learning frameworks, including Keras[2].

These CNNs models pre-trained on ImageNet can be used as feature extraction models (called backbones). The early and middle convolutional layers contain low-level features such as lines, edges, and curves, while the layers far from the input interpret high-level features to distinguish between output classes in the context of a classification task. It is possible to select the level of detail for feature extraction from a pre-trained model. For instance, the output of the pre-trained model after a few layers might be adequate if the new task is very different from the source task. Although, the model output from layers considerably deeper in the pre-trained model may be employed if the new task is relatively comparable to the source task. The pre-trained model or the desired part of this model can be incorporated directly into a new neural network architecture. In the segmentation task with deep learning, the backbones can be integrated into the encoder part of the encoder-decoder network to process images and extract relevant features. The layers of the decoder part are then randomly initialized and trained to realize the target task. The weights of the pre-trained model can be frozen, meaning that they are not updated as the new model is trained. However, the transferred feature layers can be fine-tuned to the new task by backpropagating the errors from the new model into the base features of the pre-trained model [204]. Freezing the backbone weights or fine-tuning them depends on the size of the target dataset and the number of parameters of the backbone layers included in the encoder path.

5.2.2 Backbones for transfer learning

Several deep CNNs are used to learn and offer relevant features in image recognition and related computer vision tasks. These pre-trained models are widely used for transfer learning because of their performance. The following is an overview of some deep CNN architectures used as feature extractors (backbones of U-Net architectures):

[2] https://keras.io/

5.2.2.1 VGG-Net

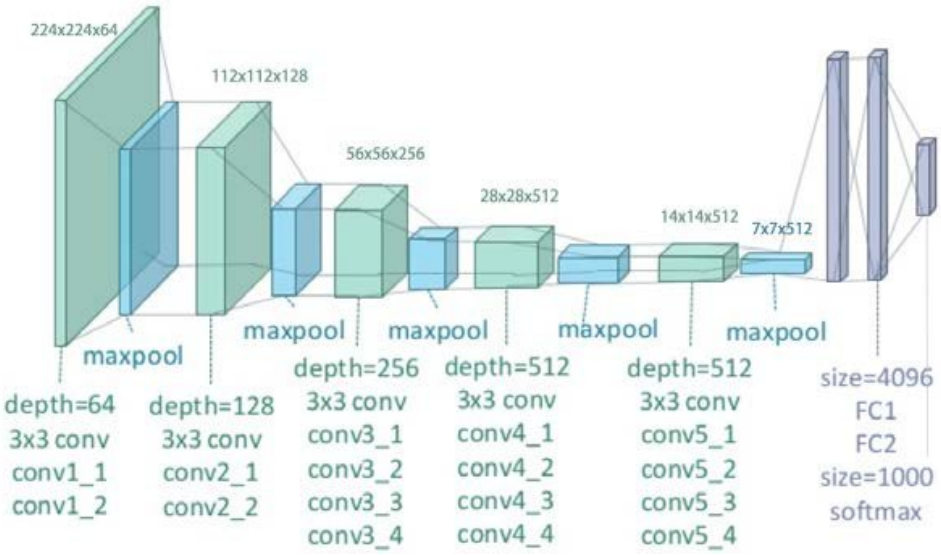


Figure 5.1: VGG19 architecture [15], conv, maxpool, and FC imply convolution, fully connected, and max-pooling layers, respectively.

The Visual Geometry Group (VGG-Net) [207] was proposed by K. Simonyan and A. Zisserman. In the ILSVRC 2014, it received a top 5 test accuracy ranking. Multiple stack convolutional blocks are the foundation of VGG-Net design. Each block consists of two convolutional layers with a kernel size of 3×3 dimension followed by pooling layers. Finally, the convolutional building blocks of VGGV-Net are followed by three FC layers. The goal is to investigate the impact of raising a CNN's depth on the accuracy by including more convolutional layers. VGG-Net comes in two models, VGG16 and VGG19 (illustrated in Figure 5.1). The numbers 16 and 19 relate to the number of convolutional layers included in each model.

5.2.2.2 Deep residual learning networks

He et al. [16] presented a Deep Residual Learning Network (ResNet). This model was the winner of ILSVRC 2015. The goal was to address the issue of vanishing gradients and parameter explosions. According to experiments, the performance of a CNN would get worse when the number of stacked convolutional layers increases without modifying its structure. In other words, the gradients of network parameters vanish as the depth increases. To solve this problem, ResNet presents a residual learning framework using identity-mapping shortcuts that connect the input of one layer with the output of the next layer. It can have a variety of sizes based on how many layers are in the model. All variants of ResNet presented in [16] contain four principal modules comprising residual

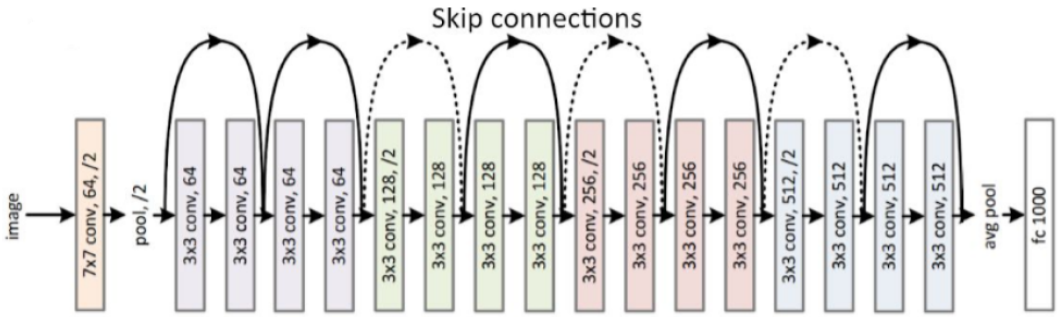


Figure 5.2: ResNet-18 architecture [16].

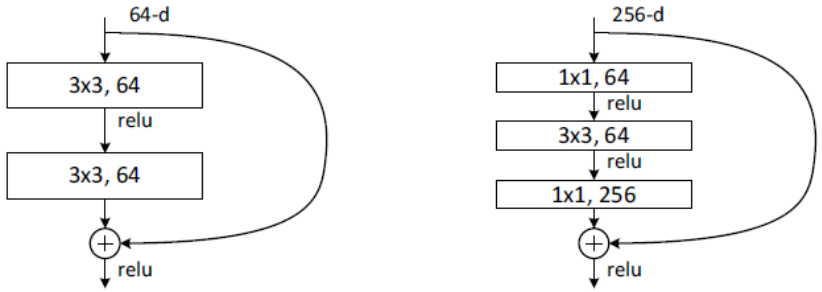


Figure 5.3: Residual block structures presented in [16]. [Left] 3 × 3 standard structure for ResNet-18/34. [Right] bottleneck structure for ResNet-50/101/152.

blocks. In the case of ResNet-18/34, each residual block has two 3×3 convolutional layers followed by a batch normalization layer and a ReLU activation function, except the last operation of a block, which does not use the ReLU function. Figure 5.2 illustrates the ResNet-18 architecture. In ResNet-50/101/152, the authors present a bottleneck block instead of the basic blocks of two operations. The main idea of the bottleneck is to use three layers consisting of 1×1, 3×3, and 1×1 convolutions, respectively, in each block. The difference between the standard residual block and the bottleneck structure is illustrated in Figure 5.3.

5.2.2.3 Densely connected convolutional networks

Densely Connected Convolutional Networks (DenseNet), an expanded architecture for the ResNet concept, was proposed by Huang et al. [17]. As shown in Figure 5.4, the main idea of this architecture is to connect each layer to every other layer. The design of the DenseNet model consists of successive dense blocks connected by transition layers. The transition Layer applies a batch normalization layer, 1×1 convolution operation followed by a 2×2 average pooling layer. The size of the feature maps within a dense block is constant, while the number of filters varies. Instead of summing, the authors applied a

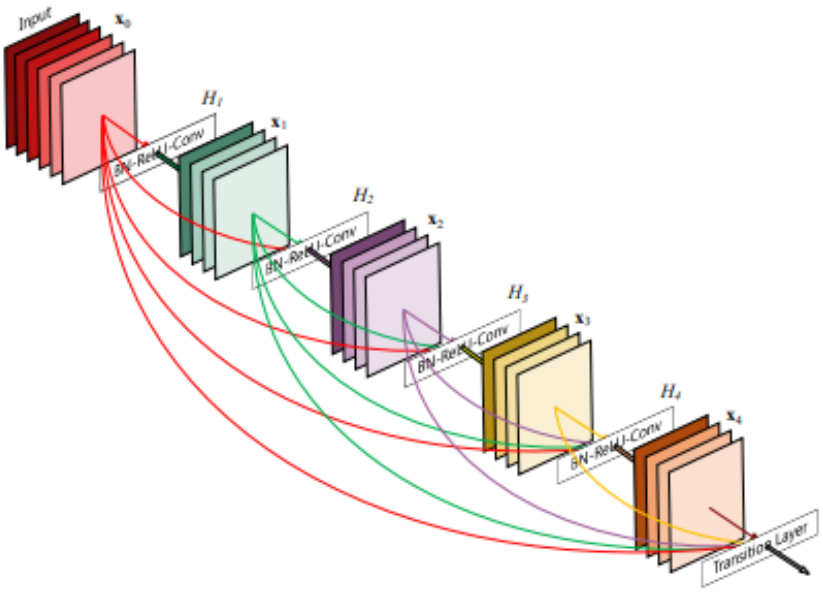


Figure 5.4: Illustration of 5-layer dense block [17].

concatenation operation to combine the feature maps learned by different layers and pass them to a new layer. This feature map combination improves computational efficiency and memory efficiency. This concept presents the main difference between DenseNets and ResNets architectures. Compared to other equivalent CNN, DenseNet requires fewer parameters since there are no redundant feature maps to train. There are various versions of the DenseNet, including DenseNet121/160/201. The numerical values represent the number of DenseNet layers.

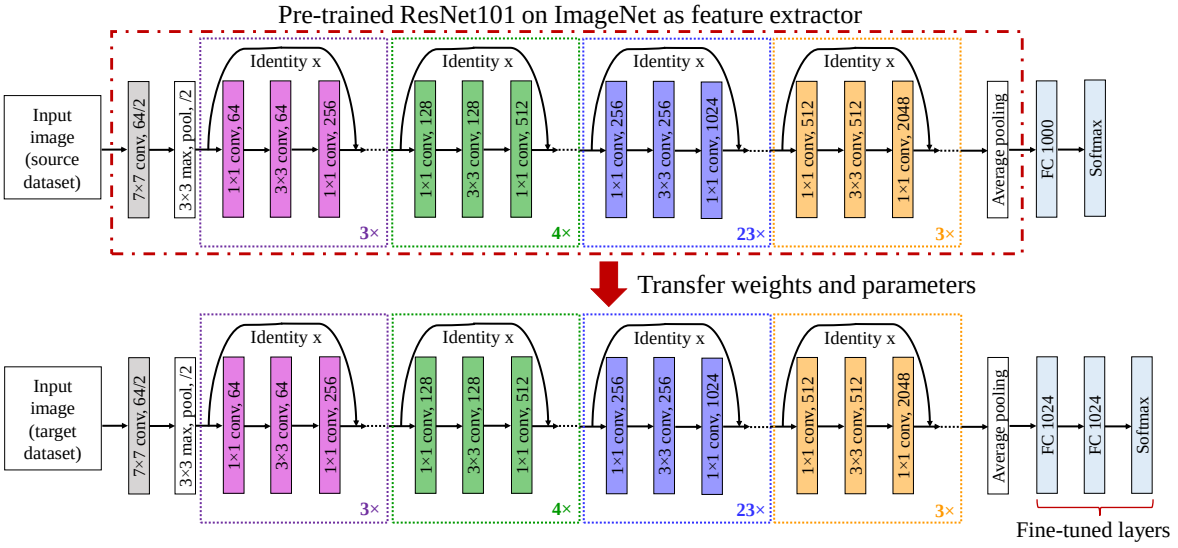


Figure 5.5: Typical example of transfer learning for classification task

5.2.3 Example of transfer learning for classification

Figure 5.5 illustrates an example of the transfer learning process for classification. The top network presents the pre-trained ResNet101 as a feature extractor. The bottom architecture is trained on the target dataset, while the highest architecture was initially trained on ImageNet. The weights and parameters of the pre-trained CNN are transferred to the bottom architecture. The final fully connected layer, considered the classifier part of the pre-trained CNN networks, is removed. The bottom design is then expanded with two other fully connected layers adapted to the new task. Each contains 1024 hidden units, followed by the Softmax activation function for fine-tuning.

5.2.4 Proposed Segmentation framework

The suggested methodology enables LV delineation using feature extraction from echocardiographic images based on an end-to-end learning process of the shape and texture of the heart chambers in each B-Mode echocardiographic apical view. Additionally, the created framework is based on transfer learning. It allows using prior knowledge from the extensive ImageNet dataset by including pre-trained deep CNNs in the segmentation architecture. As a result, a new level of feature extraction was established. In addition, this method provides much deeper architectures that can efficiently complete the segmentation process without encountering the vanishing gradient problem. Figure 5.6 illustrates the proposed framework. Later steps explain each part of it.

5.2.4.1 Preprocessing

Any changes made to the raw data before feeding it into a machine learning or deep learning algorithm is a preprocessing operation. Preprocessing enables more effective data analysis and helps achieve optimal results even with low-quality data. In the case of image data, preprocessing is necessary before it can be used as input for the model. For example, fully connected layers in convolutional neural networks require that all images are in arrays of the same size. In this chapter, we follow the same procedure presented in Chapter 4 and resize CAMUS images and labels to 256×256. This resizing step was necessary because the original images in the dataset had different sizes. The chosen resolution is suitable for encompassing the left ventricle (LV) region while requiring less memory. By reducing the size of the input images, the time necessary to train the model is also reduced. Model preprocessing can accelerate inference and decrease training time without impacting the model's performance.

Furthermore, image preprocessing is applied to improve particular qualities in the

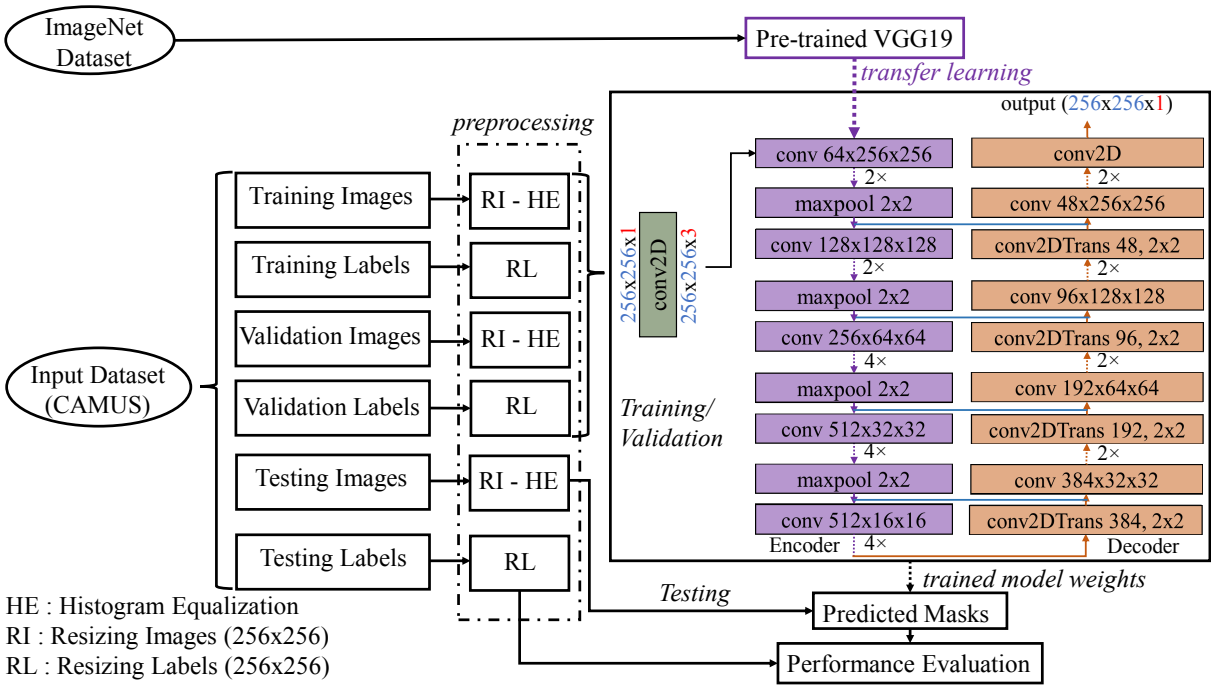


Figure 5.6: An overview of the proposed methodology to segment LV in echocardiography images consists of the U-Net 2 architecture with pre-trained VGG19 as the encoder.

image that are crucial for the application we are developing. Depending on the characteristics of the application, we must enhance the image quality to improve the performance. Following the experiments presented in Chapter 4, histogram equalization was applied to the input images after resizing operation. This enhancement can raise the quality of CAMUS images since the low contrast and heterogeneity in the electrocardiography images make the heart structures less obvious.

5.2.4.2 Channel adaptation

After the preprocessing step, we applied channel adaptation. This operation adjusts the input images to the input of the networks. The pre-trained backbones have been trained using red/green/blue (RGB) images as input. Conversely, the images of the CAMUS dataset are encoded in one channel (grayscale). Therefore, the input images cannot be passed into the pre-trained architectures, which prompts the question of how best to utilize these models. This issue may be solved using multiple methods to improve the usage of pre-trained models. For instance, we can duplicate the grayscale intensities across all three channels for color images. However, this process may not be the best solution because it is time-consuming. We used another technique to avoid this issue. We added an extra 1D convolutional layer with three convolutional filters before the the encoder path beginning. This layer scales the pixel intensities and converts grayscale to RGB to make

better use of the power of the pre-trained convolutional blocks. The backpropagation algorithm determines the parameters of this additional layer during training.

5.2.4.3 Segmentation models based on transfer learning

The main contribution of this chapter is to discover the best collection of the set of neural architecture and pre-trained models that presents the best segmentation performances. The best combination corresponds to the framework through which ED and ES echocardiographic images will be well segmented. Good segmentation allows suitable estimation of the LV function. We attempt to find the best combination by investigating the influence of various modules integrated into the encoder part of the segmentation architectures. That's why we examine the performance of different backbones containing different layers and parameters.

In this chapter, a total of 15 different architectures are investigated. Each method combines a pre-trained network (backbone) with a segmentation architecture. The examined segmentation networks include U-Net 1, U-Net 2, LinkNet, Attention U-Net, and TransUNet, previously described in Chapter 2. For this study, we replace the down-sampling part of each architecture with the pre-trained models defined in Section 5.2.2. These backbones consist of different units, such as standard convolutions in VGG19, residual units in ResNet101, and dense blocks in DenseNet121.

The core of the suggested deep learning framework is deep CNN. Each segmentation architecture is based on U-Net architecture. They involve two paths with two different tasks (encoder and decoder). In the encoder stage, we incorporate the pre-trained CNN. The input image is fed directly into it after the channel adaptation. The backbones extract the features from input images. On the other side of the network, the high-level contextual information is passed into the decoder path to be up-sampled. The final predicted mask is generated after the last block of the decoder.

Figure 5.6 shows the proposed approach for echocardiographic LV segmentation. The preprocessing step is depicted on the left side of the figure, and the fully convolutional network-based segmentation is shown on the right side using a combination example of U-Net 2 with the VGG19 backbone. Thus, the encoder extracts the features using standard convolution modules of the pre-trained VGG19, while the decoder part of the U-Net 2 produces the output mask. Hence, the standard convolution modules of the pre-trained VGG19 extract the features while the decoder part of the U-Net 2 reconstructs and creates the final predicted image from the compact representation.

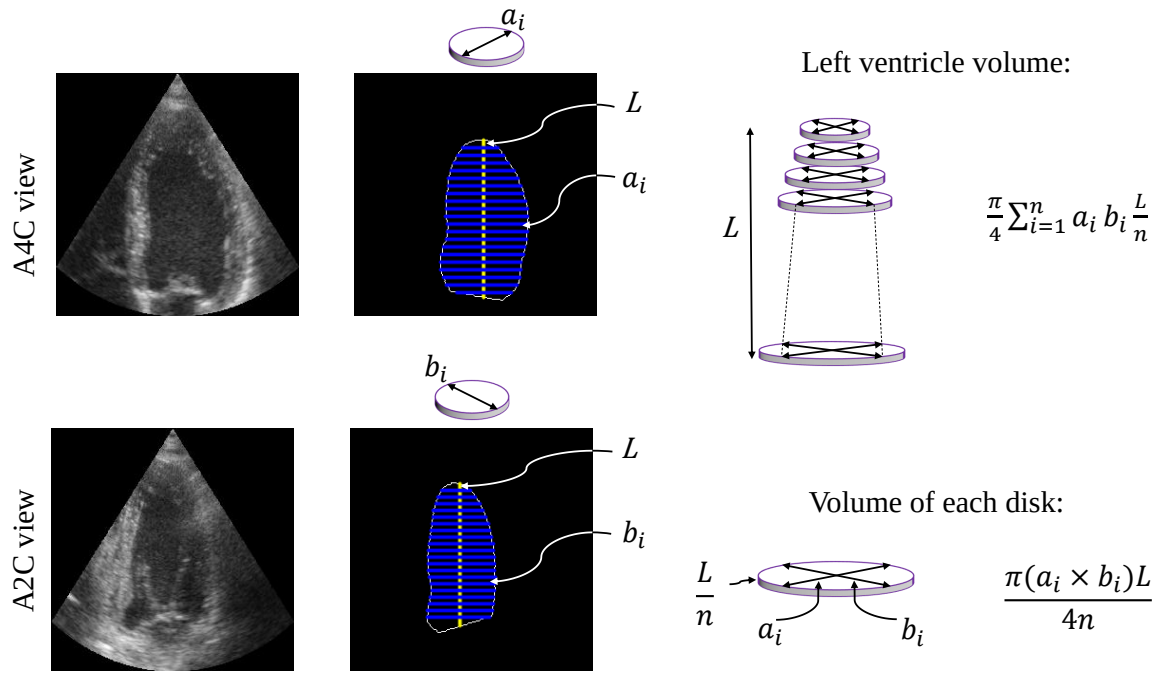


Figure 5.7: Estimation of LV volume from 2D echocardiography using the modified Simpson's rule approach.

5.2.5 Analysis of the left ventricular function

We applied the modified Simpson's rule [40] to approximate the corresponding volumes from the contours of the LV endocardium in 2D echocardiography. Figure 5.7 illustrates the principle of this approach for estimating chamber volume from B-mode echocardiography. It is utilized when the A4C and A2C contours are available. As this technique is less sensitive to geometric aberrations, it is the currently recommended two-dimensional method to assess the $\mathrm{LV_{EF}}$ [208]. The segmented surfaces of LV in ED and ES were divided into 20 discs of equal height. Hence, the left ventricular longest length L is split into 20 equal sections. The volume is estimated by summing the areas from diameters a_i and b_i of the 20 elliptic disks. Figure 5.7 also shows the form of each disk. The LV volume in ED and ES frames are calculated using the following equations:

$$LV_{EDV} = \frac{\pi}{4}\sum_{i=1}^{20} a(ED)_i b(ED)_i \frac{L(ED)}{20} \tag{5.1}$$

$$LV_{ESV} = \frac{\pi}{4}\sum_{i=1}^{20} a(ES)_i b(ES)_i \frac{L(ES)}{20} \tag{5.2}$$

Where $L(ED)$, $L(ES)$: the longest length of the left ventricular cavity in ED and ES, respectively.

By estimating the volumes of the LV in ED and ES frames following these methods, we can calculate the LV_{EF} directly by applying the corresponding mathematical formula.

We developed an algorithm to realize the modified Simpson's rule to analyze the left ventricular function. This algorithm allows us to trace the length of the left ventricular cavity from the middle of the mitral valve to the apex and the diameters of the 20 disks. Firstly, we applied a process based on the **regionprops** MATLAB function[3] to find the length. Secondly, we obtained the diameters by applying an operation of division. The limits of these segments were defined using the **bwperim** MATLAB function[4].

5.3 Experiments and results

5.3.1 Experimental setup

We employed k-fold cross-validation with k = 9 based on the distribution suggested by Zyuzin et al. [151]. Each fold contains 50 patients with the same distribution in terms of image quality. All backbones were pre-trained on the large public dataset CAMUS.

Every time we altered the segmentation architecture, we attempted to train the new models under the same conditions and parameters to make a fair comparison. We used Python to run the experiments in the Tensorflow and Keras environments. We developed the segmentation architectures using the following libraries Segmentation Models [209] and Keras-unet-collection [210].

We used Adam [196] as the optimizer throughout the training process. The learning rate hyper-parameter was 1e-4. All the networks were trained for 100 epochs with a batch size of 4 due to the GPU memory constraints. We set the initial weights of the backbones using pre-trained ImageNet initialization. The initialization of ImageNet weights aids in accelerating CNN convergence and training. We utilized the Dice loss function to calculate and minimize the models' errors. Since only one class is predicted in the final prediction layer, we used the Sigmoid activation function. All experiments were run on a Linux workstation with a dual Intel Xeon 2.2GHz and 3GHz CPU and two Nvidia Quadro P6000 GPUs with 24 Go each.

In the part of the analysis of the LV function, we used the MATLAB R2022a programming platform to develop the modified Simpson's rule approach.

[3] https://fr.mathworks.com/help/images/ref/regionprops.html
[4] https://fr.mathworks.com/help/images/ref/bwperim.html

5.3.2 Results on CAMUS dataset

5.3.2.1 Geometrical parameters results

The primary goal of the simulations is to evaluate the effectiveness of the suggested segmentation approach for LV identification in 2D echocardiographic images using the CAMUS dataset.

Table 5.1 shows the segmentation accuracy of the LV_{EF} with the previously described evaluated methods in ED and ES, respectively. The metrics are presented as a mean and standard deviation ($\mu \pm \sigma$). The results for these parameters were calculated using a 9-fold cross-validation. We highlight the best results for each parameter in bold.

U-Net1$_{VGG19}$ presents the best results for ED images with 0.942±0.062 of DSC and 0.982±0.045 of HD. However, this is not the case in terms of HD. Att U-Net$_{VGG19}$ gives the minimum value of HD. For ES images, TransUnet$_{VGG19}$ yields the greatest performance. LinkNet$_{VGG19}$ gives the worst results in terms of DSC and HD.

5.3.2.2 Clinical parameters results

The clinical parameters pertain to the analysis of images for left ventricular (LV) assessment, specifically the left ventricular volumes and EF (ejection fraction). The estimation results for these parameters are presented in Table 5.2, where the best scores for each corresponding index are highlighted in bold. Each architecture's evaluation of every clinical parameter is based on the following criteria:

- Pearson correlation coefficient (Corr): is the descriptive statistic to determine a linear correlation. It can compute the strength and direction of the relationship between two quantitative variables. Here is the formula to calculate the Corr for a sample:

$$Corr_{xy} = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^{n}(x_i - \bar{x})^2}\sqrt{\sum_{i=1}^{n}(y_i - \bar{y})^2}} \tag{5.3}$$

  Where x_i, y_i are the individual sample points (observed and predicted values, respectively), n is the sample size, $\bar{x} = \frac{1}{n}\sum_{i=1}^{n}(x_i)$ is the sample mean (similarly for $\bar{y}$).

- Bias: the definition of an estimate being biased or unbiased can be expressed using the mathematical formula (5.4). The bias results are presented in Table 5.2 as mean ± standard deviation.

Table 5.1: LV segmentation performance of the evaluated methods expressed as mean and standard deviation ($\mu \pm \sigma$). ED: End Diastole; ES End Systole; DSC: Dice Coefficient Similarity; JC: Jaccard Coefficient; HD: Hausdorff Distance.

Network	*ED*			*ES*		
	DSC	***JC***	***HD*** (mm)	***DSC***	***JC***	***HD*** (mm)
U-Net 1						
VGG19	**0.942 ±0.026**	**0.892 ±0.045**	4.11 ±2.39	0.924 ±0.035	0.860 ±0.059	3.91 ±1.95
ResNet101	0.941 ±0.032	0.890 ±0.053	4.17 ±2.41	0.915 ±0.050	0.846 ±0.080	4.32 ±2.79
DenseNet121	0.942 ±0.029	0.891 ±0.050	4.16 ±2.55	0.915 ±0.045	0.847 ±0.074	4.36 ±3.17
U-Net 2						
VGG19	0.940 ±0.033	0.888 ±0.054	4.77 ±4.32	0.919 ±0.052	0.854 ±0.080	4.75 ±5.57
ResNet101	0.939 ±0.032	0.886 ±0.051	4.50 ±3.20	0.915 ±0.045	0.847 ±0.073	4.55 ±3.83
DenseNet121	0.942 ±0.028	0.891 ±0.048	4.22 ±2.61	0.918 ±0.043	0.851 ±0.071	4.26 ±2.87
LinkNet						
VGG19	0.939 ±0,030	0.887 ±0.050	4.59 ±3.66	0.918 ±0.047	0.854 ±0.075	4.75 ±5.17
ResNet101	0.938 ±0.032	0.885 ±0.054	4.42 ±2.99	0.912 ±0.053	0.842 ±0.082	4.57 ±3.71
DenseNet121	0.940 ±0.033	0.889 ±0.054	4.29 ±2.85	0.917 ±0.047	0.849 ±0.075	4.25 ±2.55
Att U-Net						
VGG19	0.942 ±0.029	0.891 ±0.051	**4.06** ±2.79	0.923 ±0.040	0.860 ±0.066	3.92 ±2.54
ResNet101	0.940 ±0.029	0.889 ±0.049	4.20 ±2.93	0.920 ±0.042	0.854 ±0.069	4.20 ±3.65
DenseNet121	0.940 ±0.030	0.887 ±0.050	4.16 ±2.67	0.914 ±0.047	0.845 ±0.075	4.60 ±4.68
TransUNet						
VGG19	0.942 ±0.030	0.891 ±0.049	4.14 ±2.95	**0.925** ±0.038	**0.862** ±0.062	**3.88** ±3.22
ResNet101	0.939 ±0.030	0.887 ±0.050	4.24 ±2.66	0.918 ±0.043	0.852 ±0.070	4.11 ±2.88
DenseNet121	0.940 ±0.029	0.886 ±0.049	4.32 ±3.23	0.916 ±0.045	0.849 ±0.072	4.36 ±3.86

$$Bias(\hat{\theta}) = E(\hat{\theta}) - (\theta) \tag{5.4}$$

Where $\hat{\theta}$ serves as a statistic to estimate the population parameter θ. E stands for the expected value.

- Mean Absolute Error (MAE): measures the average size of absolute errors in a set of predictions. It is calculated as the absolute average difference between the predicted and actual values.

$$MAE = \frac{\sum_{i=1}^{n} |y_i - x_i|}{n} \tag{5.5}$$

Where $|y_i - x_i|$ is the absolute errors and n is the sample size.

Analyzing these three indexes for each clinical parameter allows us to know the reliability of each model in the proposed frameworks. However, the bias index is not taken into account because the best-performing method is not always the one with the lowest bias value.

We observe that U-Net1$_{\text{VGG19}}$ achieved the best results for LV_{ESV} and LV_{EF} in Table 5.2. Regarding the estimation of the LV_{EF}, U-Net1$_{\text{VGG19}}$ got a 0.813 correlation, a small bias (3.4%), a standard deviation (at most 8.3 %), and a mae of 6.6%. Moreover, the correlation, bias, and mae are 0.959, -1.1±7.6 ml, and 5.4 ml, respectively, for the LV_{ESV}. However, the U-Net2$_{\text{DenseNet121}}$ model attained the maximum value of the correlation of 0.974 and the minimum value of the mae of 7.4 ml for the LV_{EDV}.

5.3.2.3 Bland-Altman graphs

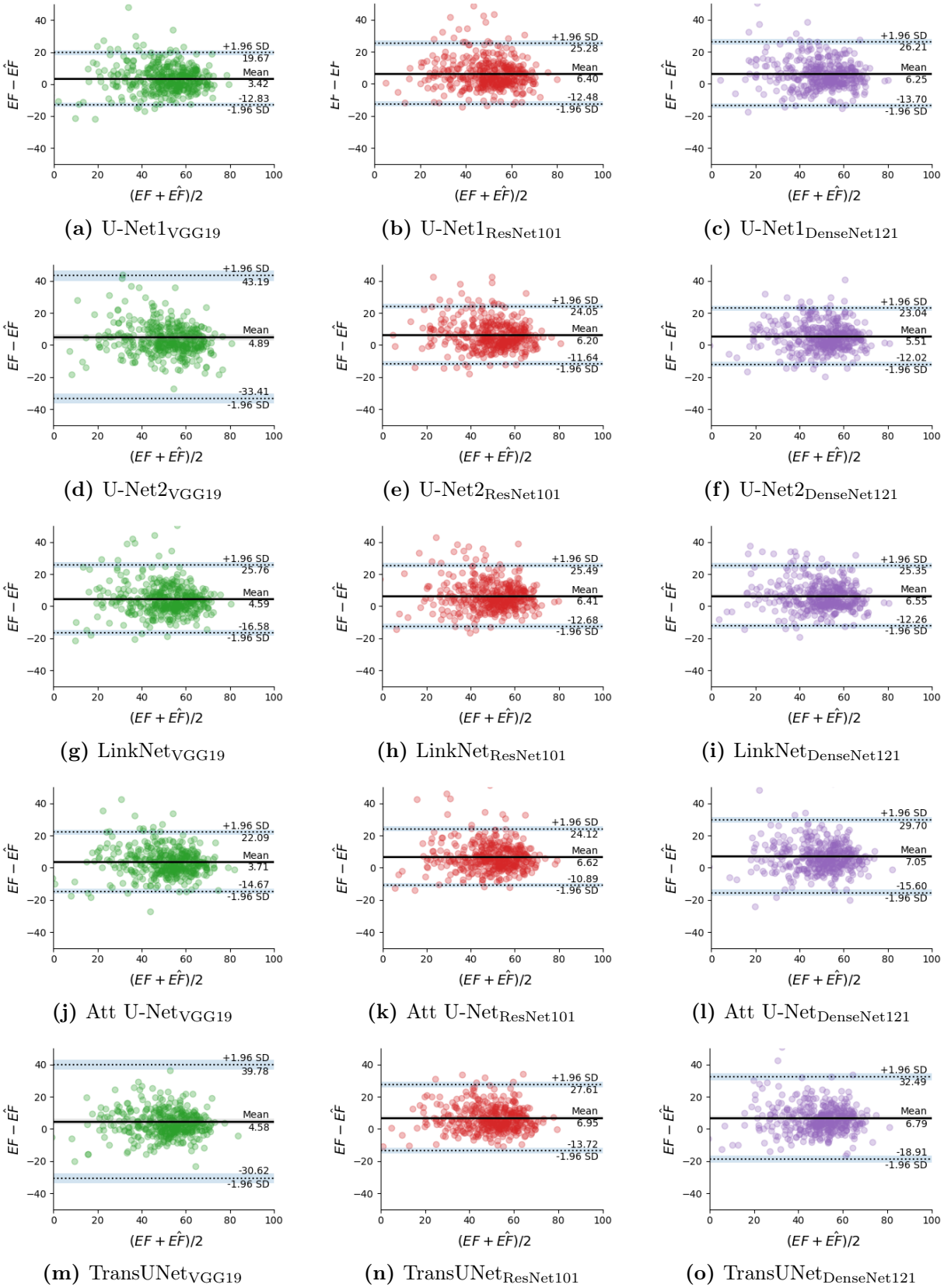
The Bland-Altman plot [211, 212] is a statistical analysis and an alternative methodology based on quantifying the degree of agreement between two quantitative measurements by calculating the limits of agreement and analyzing the mean difference. This method is valuable for identifying any bias in the mean differences and estimating the range of agreement. The mean and standard deviation between the two observations are used to calculate the statistical limitations. The Bland-Altman graph is a scatter plot xy, where the y-axis indicates the difference between the two paired measurements $(A - B)$ and the x-axis the average of these data $((A + B)/2)$. Alternatively, the graphic can be created using ratios or percentages. It is recommended that approximately 95% of the data points should be within ±1.96 standard deviation of the mean difference.

Statistical analysis on LV segmentation from 9-fold cross-validation was performed by representing Bland-Altman graphs in Figure 5.8. It enables investigation of the consis-

Table 5.2: Results of the clinical parameters of the evaluated methods. LV_{EDV}: Left Ventricular End Diastolic Volume; LV_{ESV}: Left Ventricular End Systolic Volume; LV_{EF}: Left Ventricular Ejection Fraction; corr: Pearson correlation coefficient; mae: mean absolute error.

Network	LV_{EDV}			LV_{ESV}			LV_{EF}		
	corr	***bias*$\pm\sigma$** (ml)	***mae*** (ml)	***corr***	***bias*$\pm\sigma$** (ml)	***mae*** (ml)	***corr***	***bias*$\pm\sigma$** (%)	***mae*** (%)
U-Net 1									
VGG19	0.969	1.8±11.2	7.9	**0.959**	-1.1±7.6	**5.4**	**0.813**	3.4±8.3	**6.6**
ResNet101	0.968	-1.9±11.4	7.8	0.951	-5.4±8.2	7.2	0.764	6.4±9.6	8.1
DenseNet121	0.968	-1.1±11.5	7.9	0.946	-4.9±8.5	7.1	0.756	6.3±10.2	8.2
U-Net 2									
VGG19	0.969	2.4±11.4	8.0	0.928	-1.6±9.8	5.8	0.526	4.9±19.5	8.5
ResNet101	0.969	-0.8 ±11.3	7.9	0.941	-4.7±9.0	7.0	0.789	6.2±9.1	8.0
DenseNet121	**0.974**	-1.3±10.4	**7.4**	0.952	-4.6±8.0	6.6	0.788	5.5±8.9	7.7
LinkNet									
VGG19	0.967	2.1±11.6	8.2	0.947	-1.9±8.4	5.6	0.719	4.6±10.8	7.6
ResNet101	0.963	-3.0±12.2	8.5	0.944	-6.2±8.8	7.7	0.758	6.4±9.7	8.2
DenseNet121	0.968	0.2±11.5	7.9	0.952	-4.4±8.0	6.7	0.769	6.5±9.6	8.3
Att U-Net									
VGG19	0.970	1.6±11.1	7.8	0.958	-1.3±7.5	5.4	0.783	3.7±9.4	7.0
ResNet101	0.971	2.0±11.1	7.9	0.948	-3.5±8.4	6.4	0.787	6.6±8.9	8.1
DenseNet121	0.970	0.0±11.1	7.8	0.943	-4.7±8.7	7.2	0.683	7.1±11.5	8.9
TransUNet									
VGG19	0.966	1.6±11.9	8.2	0.941	-1.8±8.9	5.4	0.541	4.6±17.9	7.7
ResNet101	0.968	2.3±11.6	8.1	0.951	-3.6±8.1	6.2	0.743	6.9±10.5	8.1
DenseNet121	0.967	0.9±11.8	8.1	0.944	-4.1±8.7	6.6	0.639	6.8±13.1	8.6

tency between $\mathrm{LV_{EF}}$ results derived from ground truth contours defined by the cardiologist and $\mathrm{LV_{EF}}$ results automatically computed by each examined method. The x-axis represents the mean of observed and predicted $\mathrm{LV_{EF}}$ scores and the y-axis represents the difference between observed and predicted $\mathrm{LV_{EF}}$ scores. Horizontal dotted lines show 95% confidence intervals, while black horizontal continuous lines present mean differences. Results are displayed for errors between -50 and 50 for easy comparison and suitable visualization. We observe that the Bland Altman diagrams presented in this figure confirm the robustness of $\mathrm{U\text{-}Net1_{VGG19}}$ compared with other approaches.



(a) U-Net1$_{\text{VGG19}}$ **(b)** U-Net1$_{\text{ResNet101}}$ **(c)** U-Net1$_{\text{DenseNet121}}$

(d) U-Net2$_{\text{VGG19}}$ **(e)** U-Net2$_{\text{ResNet101}}$ **(f)** U-Net2$_{\text{DenseNet121}}$

(g) LinkNet$_{\text{VGG19}}$ **(h)** LinkNet$_{\text{ResNet101}}$ **(i)** LinkNet$_{\text{DenseNet121}}$

(j) Att U-Net$_{\text{VGG19}}$ **(k)** Att U-Net$_{\text{ResNet101}}$ **(l)** Att U-Net$_{\text{DenseNet121}}$

(m) TransUNet$_{\text{VGG19}}$ **(n)** TransUNet$_{\text{ResNet101}}$ **(o)** TransUNet$_{\text{DenseNet121}}$

Figure 5.8: Bland Altman plots of the LV$_{\text{EF}}$ scores of CAMUS dataset. EF: Ejection Fraction scores calculated from masks manually segmented; $\hat{EF}$: Ejection Fraction scores calculated from masks automatically predicted.

5.3.2.4 Validation learning curves

Plots demonstrating model learning effectiveness across time or experience are known as learning curves. They are a diagnostic tool in machine learning for algorithms that use incremental learning from a training dataset. After each update during training, the model can be evaluated on the training dataset and a hold-out validation dataset. The graphs of the measured performance can be plotted to display learning curves. It is possible to identify learning issues, such as an underfit or overfit model, and determine if the training and validation datasets are sufficiently represented by examining the learning curves of models. There are two types of learning curves:

- Training learning curve: a measure of the model's learning efficiency derived from the training dataset.
- Validation learning curve: is used to measure the generalization performance of a model derived from the validation dataset.

Additionally, it is usually possible to create learning curves for various metrics based on the optimization and performance of the models.

- Optimization Learning Curves: are based on the metric used to optimize the model's parameters, such as loss.
- Performance Learning Curves: are based on the metric used to evaluate and choose the model, such as accuracy.

In this part of the work, we evaluate the validation learning curves of accuracy and loss metrics. Figure 5.9 shows the quality of U-Net 1 segmentation during the validation process. The green, red, and purple lines correspond to the VGG19, ResNet101, and DenseNet121 backbones. In the accuracy and loss diagrams, the U-Net1_{VGG19} model's curve outperforms the two others.

5.3.2.5 ROC curves

Furthermore, we present the ROC curves of U-Net1_{VGG19}, U-Net$1_{\text{ResNet101}}$, and U-Net$1_{\text{DenseNet121}}$ models in Figure 5.10. This graphical plot shows the true positive rate against the false negative rate to demonstrate the suggested methodology's capability for binary segmentation of the LV. The ROC curves for U-Net1_{VGG19} gain the optimal performance, with the area under the curve (AUC) values of 0.9719 in ED and 0.9709 in ES.

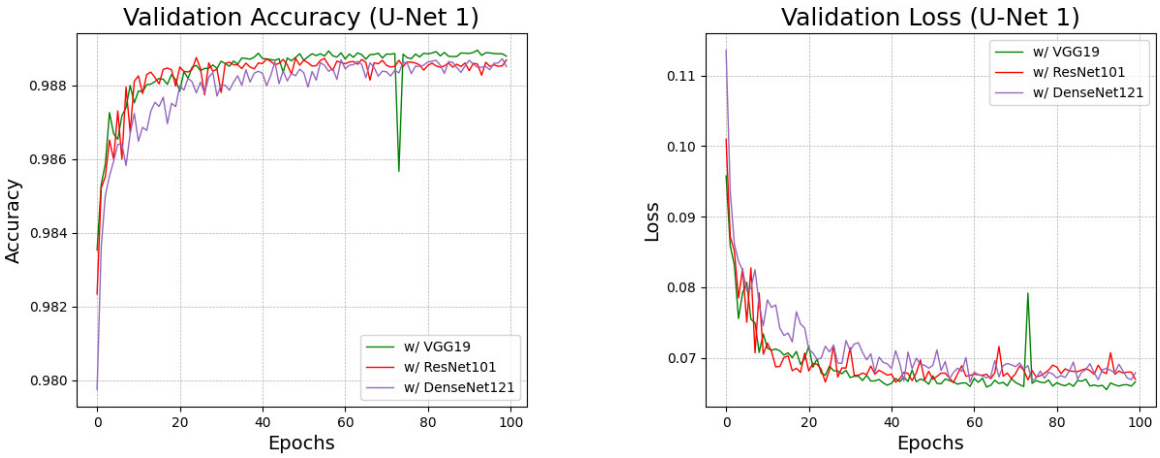


Figure 5.9: Accuracy and loss validation curves of U-Net 1 architecture pre-trained on VGG19 (green lines), ResNet101 (red lines), and DenseNet121 (purple lines).

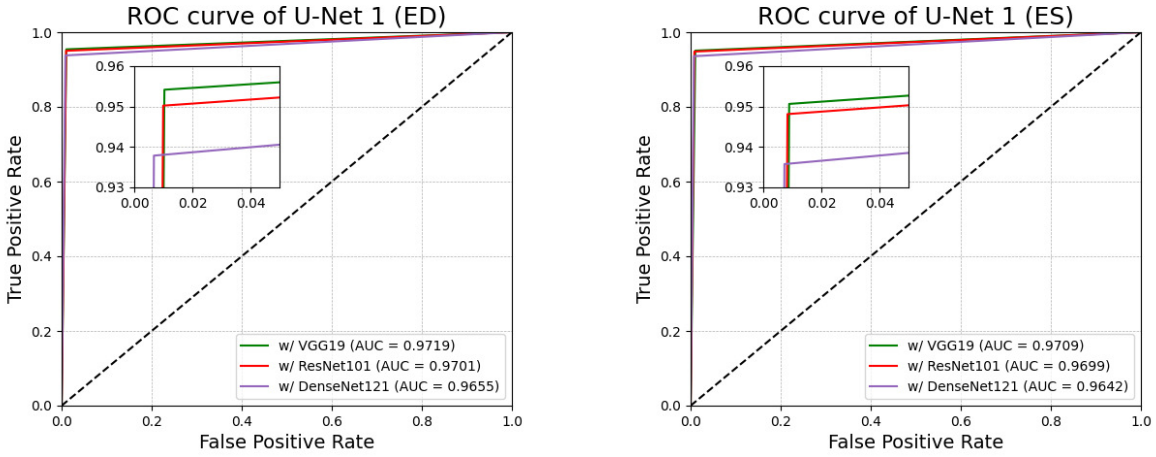


Figure 5.10: Comparison of ROC curves of the U-Net 1 architecture in ED and ES separately.

Table 5.3: Each method's total number of parameters and prediction time. #P is the number of parameters in million and #S denotes the prediction time in seconds.

Model	VGG19		ResNet101		DenseNet121	
	#P	**#S**	**#P**	**#S**	**#P**	**#S**
U-Net 1	27.3	1.75	47.5	8.22	9.7	6.48
U-Net 2	28.5	1.59	37.9	17.6	12.2	6.67
LinkNet	25.6	1.53	47.8	5.27	8.4	5.87
Att U-Net	21.8	1.14	30.0	11.2	6.2	5.44
TransUNet	24.2	5.45	31.9	10.2	8.6	9.90

5.3.2.6 Qualitative results

Figure 5.11 displays various examples of LV_{Endo} segmentation from the CAMUS dataset. In this figure, We present images from different patients, and they have different chamber views and various image qualities. The first, middle, and last columns correspond to the original, ground truth, and output images predicted by the best-proposed framework with $U\text{-}Net1_{VGG19}$. We observe that the segmentation approach delimited the LV region successfully, even in poor quality images such as the (E) case with missing borders. However, the suggested model seems to struggle with LV segmentation in some difficult samples, e.g., in the (f) sample.

5.3.3 Study of the generalizability

In machine learning, generalizability is the capacity of your model to fit correctly to additional, previously unseen data taken from the same distribution as the model's original data. In the following experiments, we assess the ability of the proposed framework to generalize on a completely different dataset. For this aim, we collected an echocardiographic image dataset.

5.3.3.1 Presentation of the private dataset

During the final two years of this thesis, we collected a private dataset from Dr. Fouad Belhachemi's cardiology clinic in Tlemcen, Algeria. The dataset comprises 100 echocar-

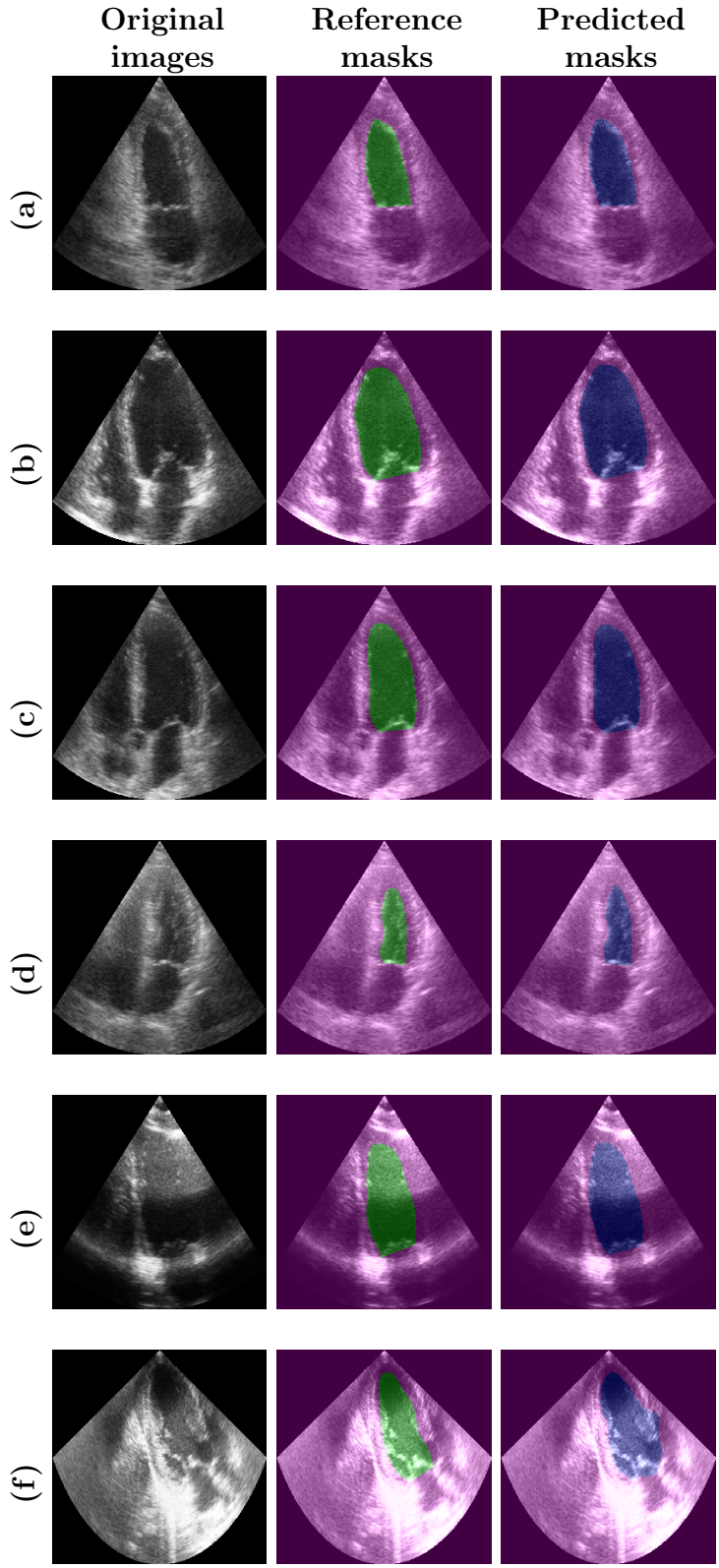


Figure 5.11: LV segmentation in different samples from CAMUS images by the best combination of U-Net 1 and VGG19 in the proposed methodology, compared to the reference masks. The last example (f) presents a prediction that failed in recovering the segmentation mask.

diography videos captured from A2C and A4C views obtained during routine medical examinations of 50 patients at the clinic. All patient data were de-identified. Any personally identifiable information was removed from the exported videos and images to ensure patient privacy. The dataset was acquired using the VIVID T8 R2 cardiovascular Doppler ultrasound unit incorporating the EchoPAC archiving software. We utilized the 3Sc-RS probe developed by GE Healthcare, an acoustic amplifier probe designed to enhance sensitivity across all modes and improve penetration (1.3-4.0 MHz).

Each exported video corresponds to a collection of B-mode images. At least one complete cardiac cycle is acquired in each view and for each patient, enabling the manual annotation of the endocardial border at two different time points representing the ED and ES by the cardiologist. These tracings allow the estimation of ventricular volumes. Figure 5.12 presents an example video of this dataset with manual annotation of $\mathrm{LV}_{\mathrm{Endo}}$ in ED and ES frames. Each image was cropped and masked to exclude text and material outside the scanning sector. Then, the generated images were downsampled into uniform 256×256 pixels. Additionally, the dataset contains corresponding labeled measurements, such as clinical measurements ($\mathrm{LV}_{\mathrm{EDV}}$, $\mathrm{LV}_{\mathrm{ESV}}$, and $\mathrm{LV}_{\mathrm{EF}}$) and information about the location of ED and ES frames in each video.

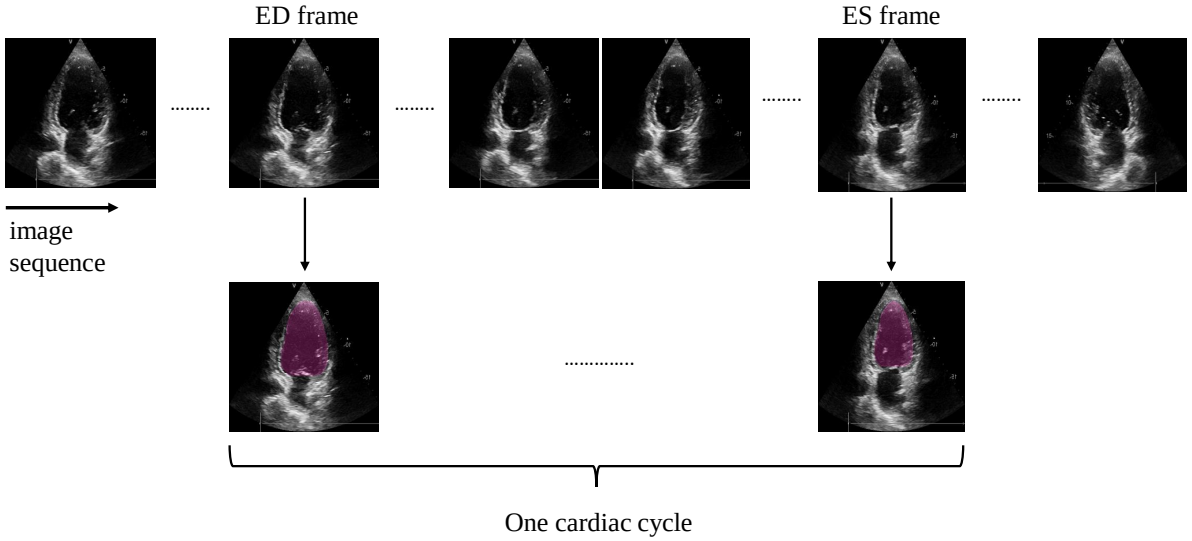


Figure 5.12: An illustration of a typical example of the private dataset with left ventricle annotation in ED and ES frames.

5.3.3.2 clinical parameters results

To evaluate clinical generalizability, we apply the proposed CAMUS-trained image segmentation framework, without tuning, to the external private dataset. We choose the

combination U-Net1$_{\text{VGG19}}$ because this architecture gives the best results on the CAMUS images. As indicated previously, the model was trained with a 9-fold cross-validation strategy. We tested the nine models on the new data and selected the model validated on the fold1, which gives the best results of generalizability.

From Table 5.4, we observe that the proposed framework with LV$_{\text{EDV}}$ as segmentation architecture clinically generalizes well, especially in the case of LV$_{\text{EDV}}$ with a correlation of 0.953 and a mae coefficient of 7.8. However, the model doesn't present very good generalizability regarding the LV$_{\text{EF}}$.

Table 5.4: Clinical parameters results of the testing of U-Net1$_{\text{VGG19}}$ on the private dataset. LV_{EDV}: Left Ventricular End Diastolic Volume; LV_{ESV}: Left Ventricular End Systolic Volume; LV_{EF}: Left Ventricular Ejection Fraction; corr: Pearson correlation coefficient; mae: mean absolute error.

Clinical parameter	Evaluation index	Result
LV$_{\text{EDV}}$	**corr**	0.953
	bias$\pm\sigma$**(ml)**	3.7$\pm$6.9
	mae(ml)	7.8
LV$_{\text{ESV}}$	**corr**	0.871
	bias$\pm\sigma$**(ml)**	2.6$\pm$7.0
	mae(ml)	8.7
LV$_{\text{EF}}$	**corr**	0.513
	bias$\pm\sigma$**(%)**	-1.6$\pm$17.0
	mae(%)	9.6

5.3.3.3 Bland-Altman graphs

We added statistical analysis of the results presented by Bland-Altman graphs in Figure 5.13. This figure illustrates the consistency between LV$_{\text{EDV}}$, LV$_{\text{ESV}}$, and LV$_{\text{EF}}$ results derived from contours manually delimited by the cardiologist and LV$_{\text{EDV}}$, LV$_{\text{ESV}}$, and LV$_{\text{EF}}$ results obtained from the LV$_{\text{EDV}}$ method, respectively. Graphs 5.13a and 5.13b demonstrate a good agreement between LV$_{\text{EDV}}$ and LV$_{\text{EDV}}$ estimated through the gener-

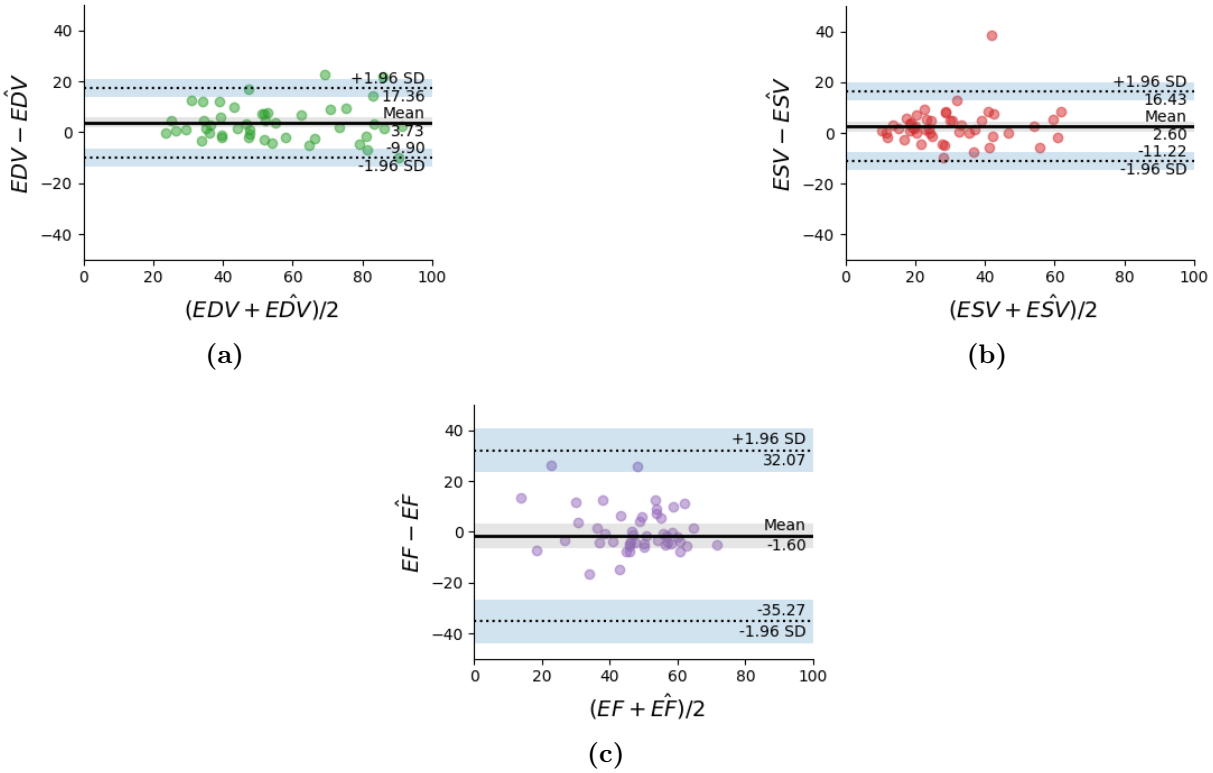


Figure 5.13: Bland Altman plots of: (a) LV_{EF}, (b) LV_{EDV}, (c) and LV_{ESV} scores of the private dataset. EF: Ejection Fraction scores calculated from masks manually segmented; $\hat{EF}$: Ejection Fraction scores calculated from masks automatically predicted. EDV: End Diastolic scores calculated from masks manually segmented; $\hat{EDV}$: End Diastolic scores calculated from masks automatically predicted. ESV: End systolic scores calculated from masks manually segmented; $\hat{ESV}$: End systolic scores calculated from masks automatically predicted.

alizability of U-Net1$_{\text{VGG19}}$ on the external dataset. We observe from Graph 5.13c that there is an acceptable agreement between LV_{EF} and LV_{EF}.

5.3.3.4 Qualitative results

For more visualization and interpretation of the results of LV_{Endo} segmentation from the private dataset, we present in Figure 5.14 different samples. The first, middle, and last columns display the original images, the ground truth masks, and the output images predicted by U-Net1$_{\text{VGG19}}$ trained on CAMUS images. The segmentation results demonstrate that this model generalizes well in most external images. However, there are some failed segmentation cases. An example is shown in the last row of the figure.

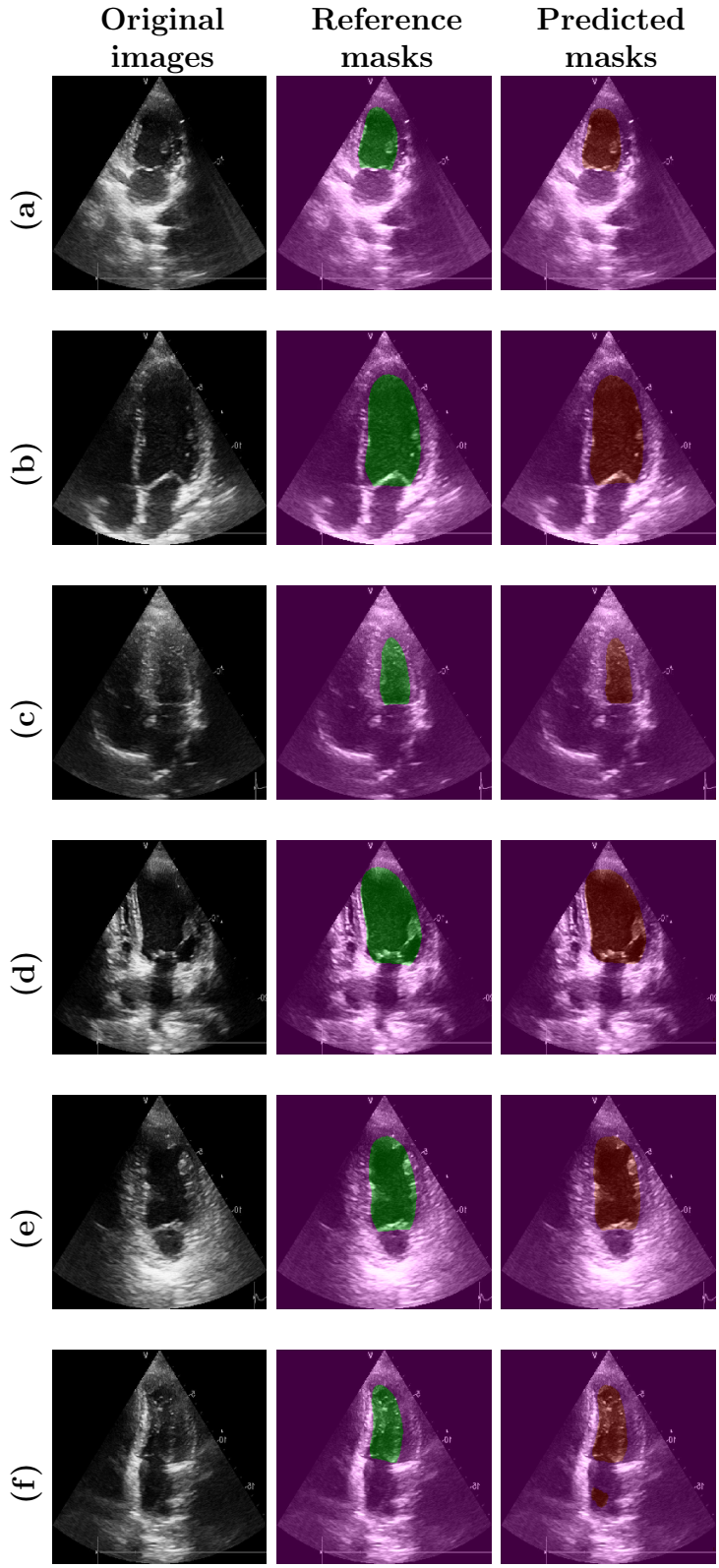


Figure 5.14: LV segmentation of different samples from the external dataset using U-Net1$_{\text{VGG19}}$ trained on CAMUS images. The last sample (f) presents a prediction that failed in recovering the corresponding segmentation mask.

5.4 Discussion

The developed framework for automated LV delineation from 2D echocardiographic images has demonstrated effectiveness and efficiency in this study. The utilization of the openly available and extensively annotated CAMUS dataset has allowed for a thorough assessment of the proposed methodology. The constructed approach consists of convolutional neural networks as feature extractors and transfer learning techniques. CNNs have shown promising performance in various medical image processing tasks and have proven the ability to reuse learned knowledge in different computer vision applications, including detection, segmentation, and classification. Hence, we selected some deep CNNs as the feature extractors in this study. By employing encoder-decoder architectures and leveraging pre-trained feature extractors, the proposed framework addresses the limitations of conventional segmentation techniques, even when faced with limited training data. This approach enables accurate and efficient LV delineation, reducing the workload for cardiologists.

Table 5.5: Comparison of the proposed method's performance with current state-of-the-art approaches in both ED and ES on CAMUS dataset.

Study	**DSC(%)**	**HD(mm)**
Inter-observer	89.60	6.30
Leclerc et al. [18]	92.75	5.40
Kim et al. [140]	92.0	4.92
Dahal et al. [133]	92.36	-
Escobar et al. [174]	92.39	-
Yang and Sermesant [166]	93.10	4.99
Saeed et al. [169]	93.11	-
Proposed(U-Net1$_{\mathbf{VGG19}}$)	**93.30**	**4.01**

The suggested methodology incorporates preprocessing procedures that involve resizing the images and applying histogram equalization before applying the transfer learning strategy. One of the prerequisites for neural networks is that all input images must have the same size. Hence, the input images were resized and reduced to a standardized dimension of 256×256 pixels. Preprocessing plays a vital role in the overall pipeline, particularly

in the case of echocardiographic images, which often suffer from low contrast. To address this issue, we employed histogram equalization to enhance the contrast of the images. The differences in pixel intensities across the image are redistributed by equalizing the histogram, resulting in a more balanced distribution and improved contrast. This enhancement process facilitates the detection of the endocardium wall, making it easier to discern and delineate.

LV segmentation results may be affected by the architecture simplification when using simple convolutions, although VGG19 achieved the best LV_{EF} prediction using the U-Net 1 design. U-Net 1 is a version of the U-Net network already optimized for CAMUS image segmentation accuracy [18]. U-$Net1_{VGG19}$ also and globally demonstrated the best clinical parameter performances, particularly LV_{ESV} and LV_{EF} (See Table 5.2 and Figure 5.8a). Figure 5.11 shows that the proposed method has successfully segmented the LV endocardium for different image types and patients, except for some images with low contrast or high luminosity in some areas of the heart, such as image (f).

From Figure 5.8, the LV_{EF} parameters calculated by U-$Net1_{VGG19}$ exhibit great agreement with the manually obtained scores (Figure 5.8a). However, U-$Net2_{VGG19}$ (Figure 5.8d) and $TransUNet_{VGG19}$ (Figure 5.8m) present bad agreement despite showing good geometrical results for ED and ES cardiac phases, respectively. The LV_{EF} error reaches very high values in some cases. This observation may be due to the poor segmentation of these networks in the other cardiac phase, for which they do not perform well. Consequently, the results of the LV_{EF} calculation for these cases are considered to be outliers. Hence, U-$Net1_{VGG19}$ is more stable in estimating the clinical indices, i.e., LV_{EDV} and LV_{ESV}.

It is important to note that a neural network's hyper-parameters, such as the number of layers and neurons, can significantly affect network performance. We present the number of parameters and prediction time for each design examined in this study in Table 5.3. It becomes apparent upon this table analysis that the VGG19 feature extractor demonstrates superior prediction speed compared to other backbone architectures. However, DenseNet121 contains fewer parameters for each segmentation network. Therefore, this is due to the design of DenseNet, where each 3×3 convolution is enhanced by a bottleneck consisting of a 1×1 convolution. As a result, DenseNet produces fewer feature maps per convolution than other architectures like ResNet101. This characteristic of DenseNet121 contributes to improved computational and memory efficiency, allowing networks utilizing DenseNet121 as an encoder to predict test images within a shorter time frame.

The LV was successfully segmented and analyzed using the suggested methods in most test images. Based on the similarity of the ultrasound dataset for LV segmentation, we compare our results to the most recent state-of-the-art techniques. Table 5.5 demonstrates

that our transfer learning technique outperforms competing methods in terms of DSC and HD in both ED and ES. However, due to various factors, segmentation algorithms may perform less effectively in particular images. For instance, the shadows, speckle noise, and artifacts are characteristics of the nature of ultrasound images. These shortcomings reduce the ability of feature extractor modules to define the endocardial border of the LV.

To clinically assess the generalizability of the proposed framework, we also deploy the best-performing model (U-Net1$_{\text{VGG19}}$) to another external dataset without tuning. We collected this dataset to evaluate the robustness of the proposed algorithm based on transfer learning on different datasets exported from various scanners tools. The dataset contains 100 B-mode echocardiographic videos from 50 patients. The experiments demonstrate the generalizability of the segmentation framework for echocardiography analysis. We report good correlation and narrower limits of agreement on clinical parameters, especially LV_{EDV} and LV_{ESV}. This observation demonstrates that U-Net1$_{\text{VGG19}}$ reached consistent performance on clinical data. Many tracks in future work should further enhance the LV_{EF} prediction by developing strong deep learning, which adapts better to challenging circumstances.

5.5 Conclusion

This chapter presented an in-depth analysis and study of a transfer learning-based methodology for LV analysis in 2D echocardiographic images. The proposed framework incorporates U-shaped encoder-decoder networks and employs a preprocessing technique to enhance the quality of input images for LV segmentation. Given the low contrast nature of echocardiographic images, histogram equalization was utilized as a preprocessing technique to address this issue. Multiple encoder-decoder architectures, namely U-Net 1, U-Net 2, LinkNet, Attention U-Net, and TransUNet, were investigated in combination with three pre-trained backbones: VGG19, ResNet101, and DenseNet121. The transfer learning strategy is effective, improves the segmentation performance, and gives a shorter prediction time. Among the various networks created and evaluated, the U-Net 1 network with VGG19 as the encoder, which had already been pre-trained on the ImageNet dataset, yielded the best results. This combination demonstrated superior reliability and stability of geometrical and clinical parameters such as LV_{EDV}, LV_{ESV}, and LV_{EF}. The proposed methodology exhibited good generalizability performance to an external private dataset, confirming its consistency and robustness. The findings of this study highlight the effectiveness of the proposed transfer learning-based methodology for LV analysis in 2D echocardiographic images. U-Net 1 and VGG19 combination provides reliable and

accurate LV segmentation results, with potential applications in clinical practice and computer-aided diagnosis systems.

General Conclusion

Summary of contributions

Cardiovascular diseases are the leading cause of death in the world. They involve all disorders of the heart and blood vessels. The LV is an essential component of the cardiovascular system. Through its function, it connects almost all organs of the system. Most cardiovascular diseases predominantly affect the LV cavity. Early detection and diagnosis of these illnesses are inevitable. Hence, developing an automated system can significantly improve clinical processes by giving cardiologists the decision-support tools they need for the early detection, diagnosis, and treatment of cardiovascular diseases.

This thesis aims to develop an automatic system for echocardiographic image analysis based on deep learning to evaluate the performance of the LV chamber. Echocardiography is the modality used in clinical routine to assess the anatomy and physiology of the heart. It has multiple advantages, i.e., non-invasive, fast, real-time, and inexpensive. In this research, we focus on the B-mode apical images produced from the TTE diagnostic examination. This kind of echocardiographic image allows the cardiologists to estimate multiple relevant clinical parameters, e.g., $\mathrm{LV}_{\mathrm{EDV}}$, $\mathrm{LV}_{\mathrm{ESV}}$, and $\mathrm{LV}_{\mathrm{EF}}$, to assess the performance of the LV cavity.

The clinical parameters enable a quantification of the LV function from an accurate segmentation of the LV region. The manual delineation of the LV routinely performed by cardiologists is a critical and rough task that presents various drawbacks. Thus, the segmentation process must be essentially automatic. The thesis was conducted in two main directions to build a reliable system for echocardiographic image analysis assisting cardiologists in daily clinical practice.

The thesis consists of five main chapters. It begins with a comprehensive introduction that sets the stage for the research. The first two chapters provide the necessary background information. The first chapter covers the clinical aspects relevant to echocardiography modality and cardiac function assessment, while the second chapter delves into the technical aspects of image processing and deep learning techniques. The third chapter

was devoted to the state-of-the-art to define the current knowledge about our research. The last two chapters presented the contributions of this thesis.

In the first research axis, we investigated the influence of the attention mechanism on two improved U-Net architectures for LV segmentation from the echocardiographic images dataset, namely CAMUS. Based on the realized experiments and the obtained findings, we demonstrated that attention mechanisms improve the segmentation of the LV. The second area of investigation consisted of proposing an automatic framework based on transfer learning for echocardiographic image analysis. The main idea was to search for the best combination between U-shaped encoder-decoder networks and pre-trained backbones. The algorithm was trained and tested on the CAMUS dataset. We tested the approach on an external dataset to evaluate the generalizability. The proposed deep learning framework corresponds to the best combination between the segmentation architecture and the pre-trained backbone, produced a better performance with geometrical and clinical parameters, shorter prediction time, and generalized well.

The comparison between the performances with those reported in the literature demonstrated that the obtained results were satisfactory for all the algorithms we have developed. The proposed methods presented accurate predictions of the LV region. However, we have identified certain limitations and drawbacks:

- Speckle noise: Ultrasound images often suffer from speckle noise, which can degrade image quality and obscure fine details. The presence of speckle noise may impact the accuracy of LV segmentation results.

- Reliability of clinical parameters: Accurate estimation of clinical parameters, such as LV volumes and EF, relies heavily on the segmentation results of the LV in both ED and ES frames.

- Limited GPU memory: The computational requirements of deep learning models, especially when training large networks, can be demanding in terms of GPU memory. Limited GPU memory may impose constraints on increasing the batch size during training, which can affect the convergence and overall performance of the investigated algorithms.

The research presented in this thesis has shown encouraging results that indicate that the proposed echocardiographic image analysis system is robust for segmenting the LV structure to evaluate cardiac function. It can assist cardiologists in assessing LV function performance and early diagnosing cardiovascular diseases.

Perspectives

The extension of this thesis work opens up several research directions and perspectives. Some potential avenues for further investigation include:

- Image enhancement and denoising: Exploring machine learning or deep learning techniques to improve the quality of echocardiographic images through image enhancement and denoising methods. These operations can involve developing models that learn to enhance image contrast, reduce noise, and enhance the visibility of fine details in echocardiographic images.

- Speckle noise reduction: Investigating deep learning approaches, such as generative adversarial networks (GANs), for effectively reducing speckle noise in echocardiographic images. Designing and training GAN models tailored for despeckling operations can help enhance the echocardiographic images' visual quality and clarity.

- Temporal coherence modeling: Studying the temporal coherence and relationships between frames throughout the cardiac cycle, particularly between ED and ES frames. Exploring network architectures such as LSTM or 3D versions of U-shaped networks can capture spatiotemporal information and improve the segmentation and analysis of echocardiographic image sequences.

- Automatic ED and ES frame detection: Developing methods for automatically detecting the ED and ES frames from echocardiographic cine series. This operation eliminates the need for ED and ES frames' manual identification by observing changes in LV dimensions. Automated frame detection can facilitate efficient and standardized analysis of cardiac function.

- Analysis of other heart cavities: Extending the analysis to other heart cavities, such as the LA or RV, using similar segmentation and analysis techniques. Exploring the applicability of the proposed methods for the segmentation and assessment of other cardiac structures can provide a more comprehensive understanding of cardiac function.

- Analysis of LV function using speckle tracking: Investigating alternative techniques and parameters, such as speckle tracking, for assessing LV function. Speckle tracking methods estimate LV myocardial velocities and deformation variables, such as strain and strain rate, which can provide additional insights into LV function beyond traditional segmentation-based approaches.

Bibliography

[1] Karin Hammer. High resolution 3d-imaging of the physiology and morphology of isolated adult cardiac myocytes from rat and mice. 2009.

[2] Hans-Hinrich Sievers, Kathrin Schubert, Ashkan Jamali, and Michael Scharfschwerdt. The influence of different inflow configurations on computational fluid dynamics in a novel three-leaflet mechanical heart valve prosthesis. *Interactive CardioVascular and Thoracic Surgery*, 27(4):475–480, 2018.

[3] Brigham and Women's Virtual Heart Failure Clinic. What is an echocardiogram? | Virtual Heart Failure Clinic.

[4] Sarah Marie-Solveig Leclerc. *Automatisation de la segmentation sémantique de structures cardiaques en imagerie ultrasonore par apprentissage supervisé*. PhD thesis, Université de Lyon, 2019.

[5] Thomas Dietenbeck. *Segmentation of 2D-echocardiographic sequences using level-set constrained with shape and motion priors*. Theses, INSA de Lyon, November 2012.

[6] Clément Papadacci. *Imagerie échographique ultrarapide du cœur et des artères chez l'homme : Vers l'imagerie ultrarapide 3D et l'imagerie du tenseur de rétrodiffusion ultrasonore*. Theses, Université Paris-Diderot Paris 7, November 2014.

[7] Daniel Mechea. What is Panoptic Segmentation and why you should care., January 2019.

[8] Geert Litjens, Francesco Ciompi, Jelmer M Wolterink, Bob D de Vos, Tim Leiner, Jonas Teuwen, and Ivana Išgum. State-of-the-art deep learning in cardiovascular image analysis. *JACC: Cardiovascular imaging*, 12(8 Part 1):1549–1565, 2019.

[9] Olaf Ronneberger, Philipp Fischer, and Thomas Brox. U-net: Convolutional networks for biomedical image segmentation. In *International Conference on Medical image computing and computer-assisted intervention*, pages 234–241. Springer, 2015.

[10] Abhishek Chaurasia and Eugenio Culurciello. Linknet: Exploiting encoder representations for efficient semantic segmentation. In *2017 IEEE Visual Communications and Image Processing (VCIP)*, pages 1–4. IEEE, 2017.

[11] Ozan Oktay, Jo Schlemper, Loic Le Folgoc, Matthew Lee, Mattias Heinrich, Kazunari Misawa, Kensaku Mori, Steven McDonagh, Nils Y Hammerla, Bernhard Kainz, et al. Attention u-net: Learning where to look for the pancreas. *arXiv preprint arXiv:1804.03999*, 2018.

[12] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. *Advances in neural information processing systems*, 30, 2017.

[13] Wee Chin Wong, Ewan Chee, Jiali Li, and Xiaonan Wang. Recurrent neural network-based model predictive control for continuous pharmaceutical manufacturing. *Mathematics*, 6(11):242, 2018.

[14] Nabila Abraham and Naimul Mefraz Khan. A novel focal tversky loss function with improved attention u-net for lesion segmentation. In *2019 IEEE 16th international symposium on biomedical imaging (ISBI 2019)*, pages 683–687. IEEE, 2019.

[15] Yufeng Zheng, Clifford Yang, and Alex Merkulov. Breast cancer screening using convolutional neural network and follow-up digital mammography. In *Computational Imaging III*, volume 10669, page 1066905. SPIE, 2018.

[16] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778, 2016.

[17] Gao Huang, Zhuang Liu, Laurens Van Der Maaten, and Kilian Q Weinberger. Densely connected convolutional networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 4700–4708, 2017.

[18] Sarah Leclerc, Erik Smistad, Joao Pedrosa, Andreas Østvik, Frederic Cervenansky, Florian Espinosa, Torvald Espeland, Erik Andreas Rye Berg, Pierre-Marc Jodoin, Thomas Grenier, et al. Deep learning for segmentation using an open large-scale dataset in 2d echocardiography. *IEEE transactions on medical imaging*, 38(9):2198–2210, 2019.

[19] World Health Organization. Cardiovascular diseases (CVDs).

[20] Michelle N. Berman, Connor Tupper, and Abhishek Bhardwaj. *Physiology, Left Ventricular Function*. StatPearls Publishing, Treasure Island (FL), 2022.

[21] Rameez Rehman, Varun S. Yelamanchili, and Amgad N. Makaryus. *Cardiac Imaging*. StatPearls Publishing, Treasure Island (FL), 2022.

[22] Ms Aayushi Bansal, Dr Rewa Sharma, and Dr Mamta Kathuria. A systematic review on data scarcity problem in deep learning: solution and applications. *ACM Computing Surveys (CSUR)*, 54(10s):1–29, 2022.

[23] Ziyang Wang. Deep learning in medical ultrasound image segmentation: A review. *arXiv preprint arXiv:2002.07703*, 2020.

[24] Shinelle Whiteman, Yusuf Alimi, Mark Carrasco, Jerzy Gielecki, Anna Zurada, and Marios Loukas. Anatomy of the cardiac chambers: A review of the left ventricle. *Translational Research in Anatomy*, 23:100095, June 2021.

[25] Ali Ostadfar. Biofluid Dynamics in Human Organs. In *Biofluid Mechanics*, pages 111–204. Elsevier, 2016.

[26] S. Y. Ho. Anatomy and myoarchitecture of the left ventricular wall in normal and in disease. *European Journal of Echocardiography*, 10(8):iii3–iii7, December 2009.

[27] Ateet Kosaraju, Amandeep Goyal, Yulia Grigorova, and Amgad N. Makaryus. *Left Ventricular Ejection Fraction*. StatPearls Publishing, Treasure Island (FL), 2022.

[28] Zhonghua Sun. Cardiac ct imaging in coronary artery disease: Current status and future directions. *Quantitative Imaging in Medicine and Surgery*, 2(2):98, 2012.

[29] Karima Addetia, Tatsuya Miyoshi, and et all. Normal Values of Left Ventricular Size and Function on Three-Dimensional Echocardiography: Results of the World Alliance Societies of Echocardiography Study. *Journal of the American Society of Echocardiography*, 35(5):449–459, May 2022.

[30] Thomas Binder. 1.6.2 Ultrasound Probe. *123 Sonography*, May 2012.

[31] Vincent Chan and Anahi Perlas. Basics of ultrasound imaging. In *Atlas of ultrasound-guided procedures in interventional pain management*, pages 13–19. Springer, 2011.

[32] Marvin C Ziskin, Paul S Lafollette Jr, Kostas Blathras, and Varkey Abraham. Effect of scan format on refraction artifacts. *Ultrasound in medicine & biology*, 16(2):183–191, 1990.

[33] Chapter 3 - attenuation. In Andrew T. Gray, editor, *Atlas of Ultrasound-Guided Regional Anesthesia (Third Edition)*, pages 5–6. Elsevier, third edition edition, 2019.

[34] Alexander Ng and Justiaan Swanevelder. Resolution in ultrasound imaging. *Continuing Education in Anaesthesia Critical Care & Pain*, 11(5):186–192, 2011.

[35] Huong T Le, Nicholas Hangiandreou, Robert Timmerman, Mark J Rice, W Brit Smith, Lori Deitte, and Gregory M Janelle. Imaging artifacts in echocardiography. *Anesthesia & Analgesia*, 122(3):633–646, 2016.

[36] Harvey Feigenbaum. Role of m-mode technique in today's echocardiography. *Journal of the American Society of Echocardiography*, 23(3):240–257, 2010.

[37] Alaa A Mohamed, Ahmed A Arifi, and Ahmed Omran. The basics of echocardiography. *Journal of the Saudi Heart Association*, 22(2):71–76, 2010.

[38] Islam Aly, Asad Rizvi, Wallisa Roberts, Shehzad Khalid, Mohammad W Kassem, Sonja Salandy, Maira du Plessis, R Shane Tubbs, and Marios Loukas. Cardiac ultrasound: an anatomical and clinical review. *Translational Research in Anatomy*, 22:100083, 2021.

[39] Roberto M Lang, Karima Addetia, Akhil Narang, and Victor Mor-Avi. 3-dimensional echocardiography: latest developments and future directions. *JACC: Cardiovascular Imaging*, 11(12):1854–1878, 2018.

[40] E D Folland, A F Parisi, P F Moynihan, D R Jones, C L Feldman, and D E Tow. Assessment of left ventricular ejection fraction and volumes by real-time, two-dimensional echocardiography. A comparison of cineangiographic and radionuclide techniques. *Circulation*, 60(4):760–766, October 1979.

[41] Alenrex Maity, Anshuman Pattanaik, Santwana Sagnika, and Santosh Pani. A comparative study on approaches to speckle noise reduction in images. In *2015 International Conference on Computational Intelligence and Networks*, pages 148–155. IEEE, 2015.

[42] Guang Deng and LW Cahill. An adaptive gaussian filter for noise reduction and edge detection. In *1993 IEEE conference record nuclear science symposium and medical imaging conference*, pages 1615–1619. IEEE, 1993.

[43] Lizhe Tan and Jean Jiang. Chapter 13 - image processing basics. In Lizhe Tan and Jean Jiang, editors, *Digital Signal Processing (Third Edition)*, pages 649–726. Academic Press, third edition edition, 2019.

[44] Jyoti Jaybhay and Rajveer Shastri. A study of speckle noise reduction filters. *signal & image processing: An international Journal (SIPIJ)*, 6(3):71–80, 2015.

[45] Yunliang Qi, Zhen Yang, Wenhao Sun, Meng Lou, Jing Lian, Wenwei Zhao, Xiangyu Deng, and Yide Ma. A comprehensive overview of image enhancement techniques. *Archives of Computational Methods in Engineering*, 29(1):583–607, 2022.

[46] Ching-Chung Yang. Image enhancement by modified contrast-stretching manipulation. *Optics & Laser Technology*, 38(3):196–201, 2006.

[47] Mohammad Abdullah-Al-Wadud, Md Hasanul Kabir, M Ali Akber Dewan, and Oksam Chae. A dynamic histogram equalization for image contrast enhancement. *IEEE Transactions on Consumer Electronics*, 53(2):593–600, 2007.

[48] Stephen M Pizer, E Philip Amburn, John D Austin, Robert Cromartie, Ari Geselowitz, Trey Greer, Bart ter Haar Romeny, John B Zimmerman, and Karel Zuiderveld. Adaptive histogram equalization and its variations. *Computer vision, graphics, and image processing*, 39(3):355–368, 1987.

[49] Karel Zuiderveld. Contrast limited adaptive histogram equalization. *Graphics gems*, pages 474–485, 1994.

[50] Ying Tan. Chapter 11 - applications. In Ying Tan, editor, *Gpu-Based Parallel Implementation of Swarm Intelligence Algorithms*, pages 167–177. Morgan Kaufmann, 2016.

[51] Abdul Mueed Hafiz and Ghulam Mohiuddin Bhat. A survey on instance segmentation: state of the art. *International journal of multimedia information retrieval*, 9(3):171–189, 2020.

[52] Lee R Dice. Measures of the amount of ecologic association between species. *Ecology*, 26(3):297–302, 1945.

[53] Raimundo Real and Juan M Vargas. The probabilistic basis of jaccard's index of similarity. *Systematic biology*, 45(3):380–385, 1996.

[54] Dominik Müller, Iñaki Soto-Rey, and Frank Kramer. Towards a guideline for evaluation metrics in medical image segmentation. *arXiv preprint arXiv:2202.05273*, 2022.

[55] James A Hanley and Barbara J McNeil. The meaning and use of the area under a receiver operating characteristic (roc) curve. *Radiology*, 143(1):29–36, 1982.

[56] Christopher M Bishop and Nasser M Nasrabadi. *Pattern recognition and machine learning*, volume 4. Springer, 2006.

[57] Steven Walczak and Narciso Cerpa. Artificial neural networks. In Robert A. Meyers, editor, *Encyclopedia of Physical Science and Technology (Third Edition)*, pages 631–645. Academic Press, New York, third edition edition, 2003.

[58] Anil K Jain, Jianchang Mao, and K Moidin Mohiuddin. Artificial neural networks: A tutorial. *Computer*, 29(3):31–44, 1996.

[59] Diederik P Kingma and Jimmy Ba. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014.

[60] Geoffrey E Hinton, Nitish Srivastava, Alex Krizhevsky, Ilya Sutskever, and Ruslan R Salakhutdinov. Improving neural networks by preventing co-adaptation of feature detectors. *arXiv preprint arXiv:1207.0580*, 2012.

[61] Keiron O'Shea and Ryan Nash. An introduction to convolutional neural networks. *arXiv preprint arXiv:1511.08458*, 2015.

[62] Yann LeCun, Bernhard Boser, John S Denker, Donnie Henderson, Richard E Howard, Wayne Hubbard, and Lawrence D Jackel. Backpropagation applied to handwritten zip code recognition. *Neural computation*, 1(4):541–551, 1989.

[63] Zewen Li, Fan Liu, Wenjie Yang, Shouheng Peng, and Jun Zhou. A survey of convolutional neural networks: analysis, applications, and prospects. *IEEE transactions on neural networks and learning systems*, 2021.

[64] Sergey Ioffe and Christian Szegedy. Batch normalization: Accelerating deep network training by reducing internal covariate shift. In *International conference on machine learning*, pages 448–456. PMLR, 2015.

[65] Shibani Santurkar, Dimitris Tsipras, Andrew Ilyas, and Aleksander Madry. How does batch normalization help optimization? *Advances in neural information processing systems*, 31, 2018.

[66] Yann A LeCun, Léon Bottou, Genevieve B Orr, and Klaus-Robert Müller. Efficient backprop. In *Neural networks: Tricks of the trade*, pages 9–48. Springer, 2012.

[67] Vinod Nair and Geoffrey E Hinton. Rectified linear units improve restricted boltzmann machines. In *Icml*, 2010.

[68] Jiuxiang Gu, Zhenhua Wang, Jason Kuen, Lianyang Ma, Amir Shahroudy, Bing Shuai, Ting Liu, Xingxing Wang, Gang Wang, Jianfei Cai, et al. Recent advances in convolutional neural networks. *Pattern recognition*, 77:354–377, 2018.

[69] Y-Lan Boureau, Jean Ponce, and Yann LeCun. A theoretical analysis of feature pooling in visual recognition. In *Proceedings of the 27th international conference on machine learning (ICML-10)*, pages 111–118, 2010.

[70] Tao Wang, David J Wu, Adam Coates, and Andrew Y Ng. End-to-end text recognition with convolutional neural networks. In *Proceedings of the 21st international conference on pattern recognition (ICPR2012)*, pages 3304–3308. IEEE, 2012.

[71] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778, 2016.

[72] Jonathan Long, Evan Shelhamer, and Trevor Darrell. Fully convolutional networks for semantic segmentation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 3431–3440, 2015.

[73] Jieneng Chen, Yongyi Lu, Qihang Yu, Xiangde Luo, Ehsan Adeli, Yan Wang, Le Lu, Alan L Yuille, and Yuyin Zhou. Transunet: Transformers make strong encoders for medical image segmentation. *arXiv preprint arXiv:2102.04306*, 2021.

[74] Karen Andrea Lara Hernandez, Theresa Rienmüller, Daniela Baumgartner, and Christian Baumgartner. Deep learning in spatiotemporal cardiac imaging: A review of methodologies and clinical usability. *Computers in Biology and Medicine*, 130:104200, 2021.

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

[76] Vilson Soares de Siqueira, Moisés Marcos Borges, Rogério Gomes Furtado, Colandy Nunes Dourado, and Ronaldo Martins da Costa. Artificial intelligence applied to support medical decisions for the automatic analysis of echocardiogram images: A systematic review. *Artificial intelligence in medicine*, 120:102165, 2021.

[77] Daniel Barbosa, Denis Friboulet, Jan D'hooge, and Olivier Bernard. Fast tracking of the left ventricle using global anatomical affine optical flow and local recursive block matching. *MIDAS J*, 10, 2014.

[78] Michael Bernier, P Jodoin, and Alain Lalande. Automatized evaluation of the left ventricular ejection fraction from echocardiographic images using graph cut. *MICCAI Challenge Echocardiogr. Three-Dimensional Ultrasound Segmentation (CETUS)*, pages 25–32, 2014.

[79] Suyu Dong, Gongning Luo, Kuanquan Wang, Shaodong Cao, Qince Li, and Henggui Zhang. A combined fully convolutional networks and deformable model for automatic left ventricle segmentation based on 3d echocardiography. *BioMed research international*, 2018, 2018.

[80] Suyu Dong, Gongning Luo, Kuanquan Wang, Shaodong Cao, Ashley Mercado, Olga Shmuilovich, Henggui Zhang, and Shuo Li. Voxelatlasgan: 3d left ventricle segmentation on echocardiography with atlas guided generation and voxel-to-voxel discrimination. In *Medical Image Computing and Computer Assisted Intervention–MICCAI 2018: 21st International Conference, Granada, Spain, September 16-20, 2018, Proceedings, Part IV 11*, pages 622–629. Springer, 2018.

[81] Wataru Ohyama, Tetsushi Wakabayashi, Fumitaka Kimura, Shinji Tsuruoka, and Kiyotsugu Sekioka. Automatic left ventricular endocardium detection in echocardiograms based on ternary thresholding method. In *Proceedings 15th International Conference on Pattern Recognition. ICPR-2000*, volume 4, pages 320–323. IEEE, 2000.

[82] Riyanto Sigit, Mohd Marzuki Mustafa, Aini Hussain, Oteh Maskon, and Ika Faizura Mohd Noh. Automatic border detection of cardiac cavity images using boundary and triangle equation. In *TENCON 2009-2009 IEEE Region 10 Conference*, pages 1–4. IEEE, 2009.

[83] Anwar Anwar, Riyanto Sigit, Achmad Basuki, and I Putu Adi Surya Gunawan. Automatic segmentation of heart cavity in echocardiography images: Two & four-chamber view using iterative process method. In *2019 International Electronics Symposium (IES)*, pages 177–182. IEEE, 2019.

[84] Andrew Laine and Xuli Zong. Border identification of echocardiograms via multiscale edge detection and shape modeling. In *Proceedings of 3rd IEEE international conference on image processing*, volume 3, pages 287–290. IEEE, 1996.

[85] Darian M Onchis, Codruta Istin, Cristina Tudoran, Mariana Tudoran, and Pedro Real. Timely-automatic procedure for estimating the endocardial limits of the left

ventricle assessed echocardiographically in clinical practice. *Diagnostics*, 10(1):40, 2020.

[86] Michael Kass, Andrew Witkin, and Demetri Terzopoulos. Snakes: Active contour models. *International journal of computer vision*, 1(4):321–331, 1988.

[87] Vikram Chalana, David T Linker, David R Haynor, and Yongmin Kim. A multiple active contour model for cardiac boundary detection on echocardiographic sequences. *IEEE Transactions on Medical Imaging*, 15(3):290–298, 1996.

[88] M Mignotte and J Meunier. A multiscale optimization approach for the dynamic contour-based boundary detection issue. *Computerized Medical Imaging and Graphics*, 25(3):265–275, 2001.

[89] Fabrice Heitz, Patrick Pérez, and Patrick Bouthemy. Multiscale minimization of global energy functions in some visual recovery problems. *CVGIP: image understanding*, 59(1):125–134, 1994.

[90] Abhishek Mishra, PK Dutta, and MK Ghosh. A ga based approach for boundary detection of left ventricle with echocardiographic image sequences. *Image and vision Computing*, 21(11):967–976, 2003.

[91] Ali K Hamou and Mahmoud R El-Sakka. Optical flow active contours with primitive shape priors for echocardiography. *EURASIP journal on advances in signal processing*, 2010:1–10, 2009.

[92] Ivana Mikic, Slawomir Krucinski, and James D Thomas. Segmentation and tracking in echocardiographic sequences: Active contours guided by optical flow estimates. *IEEE transactions on medical imaging*, 17(2):274–284, 1998.

[93] Stanley Osher and James A Sethian. Fronts propagating with curvature-dependent speed: Algorithms based on hamilton-jacobi formulations. *Journal of computational physics*, 79(1):12–49, 1988.

[94] Ning Lin, Weichuan Yu, and James S Duncan. Combinative multi-scale level set framework for echocardiographic image segmentation. *Medical Image Analysis*, 7(4):529–537, 2003.

[95] JiaYong Yan and TianGe Zhuang. Applying improved fast marching method to endocardial boundary detection in echocardiographic images. *Pattern Recognition Letters*, 24(15):2777–2784, 2003.

[96] Alessandro Sarti, Cristiana Corsi, Elena Mazzini, and Claudio Lamberti. Maximum likelihood segmentation of ultrasound images with rayleigh distribution. *IEEE transactions on ultrasonics, ferroelectrics, and frequency control*, 52(6):947–960, 2005.

[97] Wen Fang, Kap Luk Chan, Sheng Fu, Shankar Muthu Krishnan, et al. Incorporating temporal information into level set functional for robust ventricular boundary detection from echocardiographic image sequence. *IEEE Transactions on Biomedical Engineering*, 55(11):2548–2556, 2008.

[98] Thomas Dietenbeck, Martino Alessandrini, Daniel Barbosa, Jan D'hooge, Denis Friboulet, and Olivier Bernard. Detection of the whole myocardium in 2d-echocardiography for multiple orientations using a geometrically constrained level-set. *Medical image analysis*, 16(2):386–401, 2012.

[99] Timothy F Cootes, Christopher J Taylor, David H Cooper, and Jim Graham. Active shape models-their training and application. *Computer vision and image understanding*, 61(1):38–59, 1995.

[100] Ghassan Hamarneh. *Towards intelligent deformable models for medical image analysis*. Department of Signals and Systems, School of Electrical and Computer ..., 2001.

[101] Nikos Paragios, Marie-Pierre Jolly, Maxime Taron, and Rama Ramaraj. Active shape models and segmentation of the left ventricle in echocardiography. In *Scale Space and PDE Methods in Computer Vision: 5th International Conference, Scale-Space 2005, Hofgeismar, Germany, April 7-9, 2005. Proceedings 5*, pages 131–142. Springer, 2005.

[102] David Beymer, Tanveer Syeda-Mahmood, Arnon Amir, Fei Wang, and Scott Adelman. Automatic estimation of left ventricular dysfunction from echocardiogram videos. In *2009 IEEE Computer Society Conference on Computer Vision and Pattern Recognition Workshops*, pages 164–171. IEEE, 2009.

[103] Yasser Ali, Soosan Beheshti, and Farrokh Janabi-Sharifi. Echocardiogram segmentation using active shape model and mean squared eigenvalue error. *Biomedical Signal Processing and Control*, 69:102807, 2021.

[104] Jiamin Liu and Jayaram K Udupa. Oriented active shape models. *IEEE Transactions on medical Imaging*, 28(4):571–584, 2008.

[105] Timothy F. Cootes, Gareth J. Edwards, and Christopher J. Taylor. Active appearance models. *IEEE Transactions on pattern analysis and machine intelligence*, 23(6):681–685, 2001.

[106] Johan G Bosch, Steven C Mitchell, Boudewijn PF Lelieveldt, Francisca Nijland, Otto Kamp, Milan Sonka, and Johan HC Reiber. Automatic segmentation of echocardiographic sequences by active appearance motion models. *IEEE transactions on medical imaging*, 21(11):1374–1383, 2002.

[107] Steven C Mitchell, Johan G Bosch, Boudewijn PF Lelieveldt, Rob J Van der Geest, Johan HC Reiber, and Milan Sonka. 3-d active appearance models: segmentation of cardiac mr and ultrasound images. *IEEE transactions on medical imaging*, 21(9):1167–1178, 2002.

[108] Yasir Aslam, N Santhi, N Ramasamy, and K Ramar. A review on various clustering approaches for image segmentation. In *2020 fourth international conference on inventive systems and control (ICISC)*, pages 679–685. IEEE, 2020.

[109] S Nandagopalan, BS Adiga, C Dhanalakshmi, and N Deepak. Automatic segmentation and ventricular border detection of 2d echocardiographic images combining k-means clustering and active contour model. In *2010 Second International Conference on Computer and Network Technology*, pages 447–451. IEEE, 2010.

[110] Deep Gupta and Radhey Shyam Anand. A hybrid edge-based segmentation approach for ultrasound medical images. *Biomedical Signal Processing and Control*, 31:116–126, 2017.

[111] Himanshu Mittal, Avinash Chandra Pandey, Mukesh Saraswat, Sumit Kumar, Raju Pal, and Garv Modwel. A comprehensive survey of image segmentation: clustering methods, performance parameters, and benchmark datasets. *Multimedia Tools and Applications*, pages 1–26, 2021.

[112] Shaohua Kevin Zhou, Bogdan Georgescu, Xiang Sean Zhou, and Dorin Comaniciu. Image based regression using boosting method. In *Tenth IEEE International Conference on Computer Vision (ICCV'05) Volume 1*, volume 1, pages 541–548. IEEE, 2005.

[113] Vladimir N Vapnik. The nature of statistical learning. *Theory*, 1995.

[114] Shaohua Kevin Zhou. Shape regression machine and efficient segmentation of left ventricle endocardium from 2d b-mode echocardiogram. *Medical image analysis*, 14(4):563–581, 2010.

[115] Jingdan Zhang, Shaohua Kevin Zhou, Dorin Comaniciu, and Leonard McMillan. Conditional density learning via regression with application to deformable shape segmentation. In *2008 IEEE Conference on Computer Vision and Pattern Recognition*, pages 1–8. IEEE, 2008.

[116] Amelia Ritahani Ismail, Abdullah Ahmad Zarir, et al. Comparative performance of deep learning and machine learning algorithms on imbalanced handwritten data. *International Journal of Advanced Computer Science and Applications*, 9(2), 2018.

[117] Gustavo Carneiro, Jacinto C Nascimento, and António Freitas. The segmentation of the left ventricle of the heart from ultrasound data using deep learning architectures and derivative-based search methods. *IEEE Transactions on Image Processing*, 21(3):968–982, 2011.

[118] Geoffrey E Hinton, Simon Osindero, and Yee-Whye Teh. A fast learning algorithm for deep belief nets. *Neural computation*, 18(7):1527–1554, 2006.

[119] Jacinto C Nascimento and Gustavo Carneiro. Deep learning on sparse manifolds for faster object segmentation. *IEEE Transactions on Image Processing*, 26(10):4978–4990, 2017.

[120] Gustavo Carneiro and Jacinto C Nascimento. Combining multiple dynamic models and deep learning architectures for tracking the left ventricle endocardium in ultrasound data. *IEEE transactions on pattern analysis and machine intelligence*, 35(11):2592–2607, 2013.

[121] Arnaud Doucet, Nando De Freitas, Neil James Gordon, et al. *Sequential Monte Carlo methods in practice*, volume 1. Springer, 2001.

[122] Gustavo Carneiro, Jacinto Nascimento, and António Freitas. Robust left ventricle segmentation from ultrasound data using deep neural networks and efficient search methods. In *2010 IEEE International Symposium on Biomedical Imaging: From Nano to Macro*, pages 1085–1088. IEEE, 2010.

[123] Yang Lei, Yabo Fu, Justin Roper, Kristin Higgins, Jeffrey D Bradley, Walter J Curran, Tian Liu, and Xiaofeng Yang. Echocardiographic image multi-structure segmentation using cardiac-segnet. *Medical Physics*, 48(5):2426–2437, 2021.

[124] Fei Liu, Kun Wang, Dan Liu, Xin Yang, and Jie Tian. Deep pyramid local attention neural network for cardiac structure segmentation in two-dimensional echocardiography. *Medical Image Analysis*, 67:101873, 2021.

[125] Ying Shen, Heye Zhang, Yiting Fan, Alex Puiwei Lee, and Lin Xu. Smart health of ultrasound telemedicine based on deeply represented semantic segmentation. *IEEE Internet of Things Journal*, 8(23):16770–16778, 2020.

[126] Yan Zeng, Po-Hsiang Tsui, Kunjing Pang, Guangyu Bin, Jiehui Li, Ke Lv, Xining Wu, Shuicai Wu, and Zhuhuang Zhou. Maef-net: Multi-attention efficient feature fusion network for left ventricular segmentation and quantitative analysis in two-dimensional echocardiography. *Ultrasonics*, 127:106855, 2023.

[127] Sihong Chen, Kai Ma, and Yefeng Zheng. Tan: temporal affine network for real-time left ventricle anatomical structure analysis based on 2d ultrasound videos. *arXiv preprint arXiv:1904.00631*, 2019.

[128] Mark Sandler, Andrew Howard, Menglong Zhu, Andrey Zhmoginov, and Liang-Chieh Chen. Mobilenetv2: Inverted residuals and linear bottlenecks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 4510–4520, 2018.

[129] David Ouyang, Bryan He, Amirata Ghorbani, Neal Yuan, Joseph Ebinger, Curtis P Langlotz, Paul A Heidenreich, Robert A Harrington, David H Liang, Euan A Ashley, et al. Video-based ai for beat-to-beat assessment of cardiac function. *Nature*, 580(7802):252–256, 2020.

[130] Huisi Wu, Jiasheng Liu, Fangyan Xiao, Zhenkun Wen, Lan Cheng, and Jing Qin. Semi-supervised segmentation of echocardiography videos via noise-resilient spatiotemporal semantic calibration and fusion. *Medical Image Analysis*, 78:102397, 2022.

[131] Mohammad Mahdi Kazemi Esfeh, Christina Luong, Delaram Behnami, Teresa Tsang, and Purang Abolmaesumi. A deep bayesian video analysis framework: towards a more robust estimation of ejection fraction. In *Medical Image Computing and Computer Assisted Intervention–MICCAI 2020: 23rd International Conference, Lima, Peru, October 4–8, 2020, Proceedings, Part II 23*, pages 582–590. Springer, 2020.

[132] Mohammad H Jafari, Nathan Van Woudenberg, Christina Luong, Purang Abolmaesumi, and Teresa Tsang. Deep bayesian image segmentation for a more robust ejection fraction estimation. In *2021 IEEE 18th International Symposium on Biomedical Imaging (ISBI)*, pages 1264–1268. IEEE, 2021.

[133] Lavsen Dahal, Aayush Kafle, and Bishesh Khanal. Uncertainty estimation in deep 2d echocardiography segmentation. *arXiv preprint arXiv:2005.09349*, 2020.

[134] Liangliang Liu, Jianhong Cheng, Quan Quan, Fang-Xiang Wu, Yu-Ping Wang, and Jianxin Wang. A survey on u-shaped networks in medical image segmentations. *Neurocomputing*, 409:244–258, 2020.

[135] Erik Smistad, Andreas Østvik, et al. 2d left ventricle segmentation using deep learning. In *2017 IEEE international ultrasonics symposium (IUS)*, pages 1–4. IEEE, 2017.

[136] Rudolph Emil Kalman et al. A new approach to linear filtering and prediction problems [j]. *Journal of basic Engineering*, 82(1):35–45, 1960.

[137] Pallavi Kulkarni and Deepa Madathil. Fully automatic segmentation of lv from echocardiography images and calculation of ejection fraction using deep learning. *International Journal of Biomedical Engineering and Technology*, 40(3):241–261, 2022.

[138] Foivos I Diakogiannis, François Waldner, Peter Caccetta, and Chen Wu. Resunet-a: A deep learning framework for semantic segmentation of remotely sensed data. *ISPRS Journal of Photogrammetry and Remote Sensing*, 162:94–114, 2020.

[139] Steven Guan, Amir A Khan, Siddhartha Sikdar, and Parag V Chitnis. Fully dense unet for 2-d sparse photoacoustic tomography artifact removal. *IEEE journal of biomedical and health informatics*, 24(2):568–576, 2019.

[140] Sekeun Kim, Hyung-Bok Park, Jaeik Jeon, Reza Arsanjani, Ran Heo, Sang-Eun Lee, Inki Moon, Sun Kook Yoo, and Hyuk-Jae Chang. Fully automated quantification of cardiac chamber and function assessment in 2-d echocardiography: clinical feasibility of deep learning-based algorithms. *The International Journal of Cardiovascular Imaging*, 38(5):1047–1059, 2022.

[141] Alberto Gomez, Mihaela Porumb, Angela Mumith, Thierry Judge, Shan Gao, Woo-Jin Cho Kim, Jorge Oliveira, and Agis Chartsias. Left ventricle contouring of apical three-chamber views on 2d echocardiography. In *Simplifying Medical Ultrasound: Third International Workshop, ASMUS 2022, Held in Conjunction with MICCAI 2022, Singapore, September 18, 2022, Proceedings*, pages 96–105. Springer, 2022.

[142] Neda Azarmehr, Xujiong Ye, Faraz Janan, James P Howard, Darrel P Francis, and Massoud Zolgharni. Automated segmentation of left ventricle in 2d echocardiography using deep learning. *arXiv preprint arXiv:2003.07628*, 2020.

[143] Arghavan Arafati, Daisuke Morisawa, Michael R Avendi, M Reza Amini, Ramin A Assadi, Hamid Jafarkhani, and Arash Kheradvar. Generalizable fully automated multi-label segmentation of four-chamber view echocardiograms based on deep convolutional adversarial networks. *Journal of The Royal Society Interface*, 17(169):20200267, 2020.

[144] Mohammad H Jafari, Zhibin Liao, Hany Girgis, Mehran Pesteie, Robert Rohling, Ken Gin, Terasa Tsang, and Purang Abolmaesumi. Echocardiography segmentation by quality translation using anatomically constrained cyclegan. In *Medical Image Computing and Computer Assisted Intervention–MICCAI 2019: 22nd International Conference, Shenzhen, China, October 13–17, 2019, Proceedings, Part V 22*, pages 655–663. Springer, 2019.

[145] Kaizhong Deng, Yanda Meng, Dongxu Gao, Joshua Bridge, Yaochun Shen, Gregory Lip, Yitian Zhao, and Yalin Zheng. Transbridge: A lightweight transformer for left ventricle segmentation in echocardiography. In *Simplifying Medical Ultrasound: Second International Workshop, ASMUS 2021, Held in Conjunction with MICCAI 2021, Strasbourg, France, September 27, 2021, Proceedings 2*, pages 63–72. Springer, 2021.

[146] Mohammad H Jafari, Hany Girgis, Zhibin Liao, Delaram Behnami, Amir Abdi, Hooman Vaseli, Christina Luong, Robert Rohling, Ken Gin, Terasa Tsang, et al. A unified framework integrating recurrent fully-convolutional networks and optical flow for segmentation of the left ventricle in echocardiography data. In *Deep Learning in Medical Image Analysis and Multimodal Learning for Clinical Decision Support: 4th International Workshop, DLMIA 2018, and 8th International Workshop, ML-CDS 2018, Held in Conjunction with MICCAI 2018, Granada, Spain, September 20, 2018, Proceedings 4*, pages 29–37. Springer, 2018.

[147] Ming Li, Chengjia Wang, Heye Zhang, and Guang Yang. Mv-ran: Multiview recurrent aggregation network for echocardiographic sequences segmentation and full cardiac cycle analysis. *Computers in biology and medicine*, 120:103728, 2020.

[148] Muhammad Ali Shoaib, Joon Huang Chuah, Raza Ali, Samiappan Dhanalakshmi, Yan Chai Hum, Azira Khalil, and Khin Wee Lai. Fully automatic left ventricle segmentation using bilateral lightweight deep neural network. *Life*, 13(1):124, 2023.

[149] Fisher Yu and Vladlen Koltun. Multi-scale context aggregation by dilated convolutions. *arXiv preprint arXiv:1511.07122*, 2015.

[150] Shakiba Moradi, Mostafa Ghelich Oghli, Azin Alizadehasl, Isaac Shiri, Niki Oveisi, Mehrdad Oveisi, Majid Maleki, and Jan Dhooge. Mfp-unet: A novel deep learning based approach for left ventricle segmentation in echocardiography. *Physica Medica*, 67:58–69, 2019.

[151] Vasily Zyuzin, Andrey Mukhtarov, Denis Neustroev, and Tatiana Chumarnaya. Segmentation of 2d echocardiography images using residual blocks in u-net architectures. In *2020 ural symposium on biomedical engineering, radioelectronics and information technology (USBEREIT)*, pages 499–502. IEEE, 2020.

[152] Alyaa Amer, Xujiong Ye, and Faraz Janan. Resdunet: a deep learning-based left ventricle segmentation method for echocardiography. *IEEE Access*, 9:159755–159763, 2021.

[153] Yasser Ali, Farrokh Janabi-Sharifi, and Soosan Beheshti. Echocardiographic image segmentation using deep res-u network. *Biomedical Signal Processing and Control*, 64:102248, 2021.

[154] Alyaa Amer, Tryphon Lambrou, and Xujiong Ye. Mda-unet: a multi-scale dilated attention u-net for medical image segmentation. *Applied Sciences*, 12(7):3676, 2022.

[155] Artem Chernyshov, Andreas Østvik, Erik Smistad, and Lasse Løvstakken. Segmentation of 2d cardiac ultrasound with deep learning: simpler models for a simple task. In *2022 IEEE International Ultrasonics Symposium (IUS)*, pages 1–4. IEEE, 2022.

[156] Liang-Chieh Chen, George Papandreou, Florian Schroff, and Hartwig Adam. Rethinking atrous convolution for semantic image segmentation. *arXiv preprint arXiv:1706.05587*, 2017.

[157] Linde S Hesse and Ana IL Namburete. Improving u-net segmentation with active contour based label correction. In *Medical Image Understanding and Analysis: 24th Annual Conference, MIUA 2020, Oxford, UK, July 15-17, 2020, Proceedings 24*, pages 69–81. Springer, 2020.

[158] Christoforos Sfakianakis, Georgios Simantiris, and Georgios Tziritas. Gudu: Geometrically-constrained ultrasound data augmentation in u-net for echocardiography semantic segmentation. *Biomedical Signal Processing and Control*, 82:104557, 2023.

[159] Hongrong Wei, Heng Cao, Yiqin Cao, Yongjin Zhou, Wufeng Xue, Dong Ni, and Shuo Li. Temporal-consistent segmentation of echocardiography with co-learning from appearance and shape. In *Medical Image Computing and Computer Assisted Intervention–MICCAI 2020: 23rd International Conference, Lima, Peru, October 4–8, 2020, Proceedings, Part II 23*, pages 623–632. Springer, 2020.

[160] Özgün Çiçek, Ahmed Abdulkadir, Soeren S Lienkamp, Thomas Brox, and Olaf Ronneberger. 3d u-net: learning dense volumetric segmentation from sparse annotation. In *Medical Image Computing and Computer-Assisted Intervention–MICCAI 2016: 19th International Conference, Athens, Greece, October 17-21, 2016, Proceedings, Part II 19*, pages 424–432. Springer, 2016.

[161] Hongrong Wei, Junqiang Ma, Yongjin Zhou, Wufeng Xue, and Dong Ni. Co-learning of appearance and shape for precise ejection fraction estimation from echocardiographic sequences. *Medical Image Analysis*, 84:102686, 2023.

[162] Yida Chen, Xiaoyan Zhang, Christopher M Haggerty, and Joshua V Stough. Assessing the generalizability of temporally coherent echocardiography video segmentation. In *Medical Imaging 2021: Image Processing*, volume 11596, pages 463–469. SPIE, 2021.

[163] Joshua V Stough, Sushravya Raghunath, Xiaoyan Zhang, John M Pfeifer, Brandon K Fornwalt, and Christopher M Haggerty. Left ventricular and atrial segmentation of 2d echocardiography with convolutional neural networks. In *Medical Imaging 2020: Image Processing*, volume 11313, pages 32–38. SPIE, 2020.

[164] Shunzaburo Ono, Masaaki Komatsu, Akira Sakai, Hideki Arima, Mie Ochida, Rina Aoyama, Suguru Yasutomi, Ken Asada, Syuzo Kaneko, Tetsuo Sasano, et al. Automated endocardial border detection and left ventricular functional assessment in echocardiography using deep learning. *Biomedicines*, 10(5):1082, 2022.

[165] Xin Liu, Yiting Fan, Shuang Li, Meixiang Chen, Ming Li, William Kongto Hau, Heye Zhang, Lin Xu, and Alex Pui-Wai Lee. Deep learning-based automated left ventricular ejection fraction assessment using 2-d echocardiography. *American Journal of Physiology-Heart and Circulatory Physiology*, 321(2):H390–H399, 2021.

[166] Yingyu Yang and Maxime Sermesant. Shape constraints in deep learning for robust 2d echocardiography analysis. In *Functional Imaging and Modeling of the Heart: 11th International Conference, FIMH 2021, Stanford, CA, USA, June 21-25, 2021, Proceedings*, pages 22–34. Springer, 2021.

[167] Esther Puyol-Antón, Bram Ruijsink, Baldeep S Sidhu, Justin Gould, Bradley Porter, Mark K Elliott, Vishal Mehta, Haotian Gu, Christopher A Rinaldi, Martin cowie, et al. Ai-enabled assessment of cardiac systolic and diastolic function from echocardiography. In *Simplifying Medical Ultrasound: Third International Workshop, ASMUS 2022, Held in Conjunction with MICCAI 2022, Singapore, September 18, 2022, Proceedings*, pages 75–85. Springer, 2022.

[168] Fabian Isensee, Paul F Jaeger, Simon AA Kohl, Jens Petersen, and Klaus H Maier-Hein. nnu-net: a self-configuring method for deep learning-based biomedical image segmentation. *Nature methods*, 18(2):203–211, 2021.

[169] Mohamed Saeed, Rand Muhtaseb, and Mohammad Yaqub. Contrastive pretraining for echocardiography segmentation with limited data. In *Medical Image Understanding and Analysis: 26th Annual Conference, MIUA 2022, Cambridge, UK, July 27–29, 2022, Proceedings*, pages 680–691. Springer, 2022.

[170] Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. Generative adversarial networks. *Communications of the ACM*, 63(11):139–144, 2020.

[171] Zhengwei Wang, Qi She, and Tomas E Ward. Generative adversarial networks in computer vision: A survey and taxonomy. *ACM Computing Surveys (CSUR)*, 54(2):1–38, 2021.

[172] Andrew Gilbert, Maciej Marciniak, Cristobal Rodero, Pablo Lamata, Eigil Samset, and Kristin Mcleod. Generating synthetic labeled data from existing anatomical models: an example with echocardiography segmentation. *IEEE Transactions on Medical Imaging*, 40(10):2783–2794, 2021.

[173] Mohammad H Jafari, Hany Girgis, Nathan Van Woudenberg, Nathaniel Moulson, Christina Luong, Andrea Fung, Shane Balthazaar, John Jue, Micheal Tsang, Parvathy Nair, et al. Cardiac point-of-care to cart-based ultrasound translation using constrained cyclegan. *International journal of computer assisted radiology and surgery*, 15:877–886, 2020.

[174] Maria Escobar, Angela Castillo, Andrés Romero, and Pablo Arbeláez. Ultragan: ultrasound enhancement through adversarial generation. In *Simulation and Synthesis in Medical Imaging: 5th International Workshop, SASHIMI 2020, Held in Conjunction with MICCAI 2020, Lima, Peru, October 4, 2020, Proceedings 5*, pages 120–130. Springer, 2020.

[175] Erqiang Deng, Zhiguang Qin, Dajiang Chen, Zhen Qin, Yi Ding, Ji Geng, and Ning Zhang. Engan: Enhancement generative adversarial network in medical image segmentation. 2022.

[176] Jun-Yan Zhu, Taesung Park, Phillip Isola, and Alexei A Efros. Unpaired image-to-image translation using cycle-consistent adversarial networks. In *Proceedings of the IEEE international conference on computer vision*, pages 2223–2232, 2017.

[177] Phillip Isola, Jun-Yan Zhu, Tinghui Zhou, and Alexei A Efros. Image-to-image translation with conditional adversarial networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1125–1134, 2017.

[178] Tongwaner Chen, Menghua Xia, Yi Huang, Jing Jiao, and Yuanyuan Wang. Cross-domain echocardiography segmentation with multi-space joint adaptation. *Sensors*, 23(3):1479, 2023.

[179] Fahad Shamshad, Salman Khan, Syed Waqas Zamir, Muhammad Haris Khan, Munawar Hayat, Fahad Shahbaz Khan, and Huazhu Fu. Transformers in medical imaging: A survey. *arXiv preprint arXiv:2201.09873*, 2022.

[180] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, et al. An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv preprint arXiv:2010.11929*, 2020.

[181] Qing-Long Zhang and Yu-Bin Yang. Sa-net: Shuffle attention for deep convolutional neural networks. In *ICASSP 2021-2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 2235–2239. IEEE, 2021.

[182] Songlin Shi, Palisha Alimu, Pazilai Mahemuti, Qingliang Chen, and Hao Wu. The study of echocardiography of left-ventricle segmentation combining transformer and cnn. *Available at SSRN 4184447*.

[183] Hadrien Reynaud, Athanasios Vlontzos, Benjamin Hou, Arian Beqiri, Paul Leeson, and Bernhard Kainz. Ultrasound video transformers for cardiac ejection fraction estimation. In *Medical Image Computing and Computer Assisted Intervention–MICCAI 2021: 24th International Conference, Strasbourg, France, September 27–October 1, 2021, Proceedings, Part VI 24*, pages 495–505. Springer, 2021.

[184] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*, 2018.

[185] Hojjat Salehinejad, Sharan Sankar, Joseph Barfett, Errol Colak, and Shahrokh Valaee. Recent advances in recurrent neural networks. *arXiv preprint arXiv:1801.01078*, 2017.

[186] Sepp Hochreiter and Jürgen Schmidhuber. Long short-term memory. *Neural computation*, 9(8):1735–1780, 1997.

[187] Balza Achmad, Mohd Marzuki Mustafa, and Aini Hussain. Inter-frame enhancement of ultrasound images using optical flow. In *Visual Informatics: Bridging Research and Practice: First International Visual Informatics Conference, IVIC 2009 Kuala Lumpur, Malaysia, November 11-13, 2009 Proceedings 1*, pages 191–201. Springer, 2009.

[188] Rongjun Ge, Guanyu Yang, Yang Chen, Limin Luo, Cheng Feng, Hong Ma, Junyi Ren, and Shuo Li. K-net: Integrate left ventricle segmentation and direct quantification of paired echo sequence. *IEEE transactions on medical imaging*, 39(5):1690–1702, 2019.

[189] Mike Schuster and Kuldip K Paliwal. Bidirectional recurrent neural networks. *IEEE transactions on Signal Processing*, 45(11):2673–2681, 1997.

[190] Zihan Lin, Po-Hsiang Tsui, Yan Zeng, Guangyu Bin, Shuicai Wu, and Zhuhuang Zhou. Cla-u-net: Convolutional long-short-term-memory attention-gated u-net for automatic segmentation of the left ventricle in 2-d echocardiograms. In *2022 IEEE International Ultrasonics Symposium (IUS)*, pages 1–4. IEEE, 2022.

[191] Mahdi Hashemi. Enlarging smaller images before inputting into convolutional neural network: zero-padding vs. interpolation. *Journal of Big Data*, 6(1):1–13, 2019.

[192] Sarah Leclerc, Erik Smistad, Thomas Grenier, Carole Lartizien, Andreas Ostvik, Florian Espinosa, Pierre-Marc Jodoin, Lasse Lovstakken, and Olivier Bernard. Deep learning applied to multi-structure segmentation in 2d echocardiography: A preliminary investigation of the required database size. In *2018 IEEE International Ultrasonics Symposium (IUS)*, pages 1–4. IEEE, 2018.

[193] Saumya Jetley, Nicholas A Lord, Namhoon Lee, and Philip HS Torr. Learn to pay attention. *arXiv preprint arXiv:1804.02391*, 2018.

[194] Chen-Yu Lee, Saining Xie, Patrick Gallagher, Zhengyou Zhang, and Zhuowen Tu. Deeply-supervised nets. In *Artificial intelligence and statistics*, pages 562–570. PMLR, 2015.

[195] Rafsanjany Kushol, Md Raihan, Md Sirajus Salekin, ABM Rahman, et al. Contrast enhancement of medical x-ray image using morphological operators with optimal structuring element. *arXiv preprint arXiv:1905.08545*, 2019.

[196] Diederik P Kingma and Jimmy Ba. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014.

[197] Xavier Glorot and Yoshua Bengio. Understanding the difficulty of training deep feedforward neural networks. In *Proceedings of the thirteenth international conference on artificial intelligence and statistics*, pages 249–256. JMLR Workshop and Conference Proceedings, 2010.

[198] Zongwei Zhou, Md Mahfuzur Rahman Siddiquee, Nima Tajbakhsh, and Jianming Liang. Unet++: A nested u-net architecture for medical image segmentation. In *Deep Learning in Medical Image Analysis and Multimodal Learning for Clinical Decision Support: 4th International Workshop, DLMIA 2018, and 8th International Workshop, ML-CDS 2018, Held in Conjunction with MICCAI 2018, Granada, Spain, September 20, 2018, Proceedings 4*, pages 3–11. Springer, 2018.

[199] Ozan Oktay, Enzo Ferrante, Konstantinos Kamnitsas, Mattias Heinrich, Wenjia Bai, Jose Caballero, Stuart A Cook, Antonio De Marvao, Timothy Dawes, Declan P O'Regan, et al. Anatomically constrained neural networks (acnns): application to cardiac image enhancement and segmentation. *IEEE transactions on medical imaging*, 37(2):384–395, 2017.

[200] Taeouk Kim, Mohammadali Hedayat, Veronica V Vaitkus, Marek Belohlavek, Vinayak Krishnamurthy, and Iman Borazjani. Automatic segmentation of the left ventricle in echocardiographic images using convolutional neural networks. *Quantitative Imaging in Medicine and Surgery*, 11(5):1763, 2021.

[201] Yuxia Geng, Jiaoyan Chen, Ernesto Jiménez-Ruiz, and Huajun Chen. Human-centric transfer learning explanation via knowledge graph. *arXiv preprint arXiv:1901.08547*, 2019.

[202] Padmavathi Kora, Chui Ping Ooi, Oliver Faust, U. Raghavendra, Anjan Gudigar, Wai Yee Chan, K. Meenakshi, K. Swaraja, Pawel Plawiak, and U. Rajendra Acharya. Transfer learning techniques for medical image analysis: A review. *Biocybernetics and Biomedical Engineering*, 42(1):79–107, 2022.

[203] Lorien Y Pratt. Discriminability-based transfer between neural networks. *Advances in neural information processing systems*, 5, 1992.

[204] Jason Yosinski, Jeff Clune, Yoshua Bengio, and Hod Lipson. How transferable are features in deep neural networks? *Advances in neural information processing systems*, 27, 2014.

[205] Olga Russakovsky, Jia Deng, Hao Su, Jonathan Krause, Sanjeev Satheesh, Sean Ma, Zhiheng Huang, Andrej Karpathy, Aditya Khosla, Michael Bernstein, et al. Imagenet large scale visual recognition challenge. *International journal of computer vision*, 115(3):211–252, 2015.

[206] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *2009 IEEE Conference on Computer Vision and Pattern Recognition*, pages 248–255, 2009.

[207] Karen Simonyan and Andrew Zisserman. Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556*, 2014.

[208] Ateet Kosaraju, Amandeep Goyal, Yulia Grigorova, and Amgad N Makaryus. Left ventricular ejection fraction. 2017.

[209] Pavel Yakubovskiy. Segmentation models. https://github.com/qubvel/segmentation_models, 2019.

[210] Yingkai Sha. Keras-unet-collection. https://github.com/yingkaisha/keras-unet-collection, 2021.

[211] Douglas G Altman and J Martin Bland. Measurement in medicine: the analysis of method comparison studies. *Journal of the Royal Statistical Society: Series D (The Statistician)*, 32(3):307–317, 1983.

[212] J. Martin Bland and Douglas G. Altman. Statistical methods for assessing agreement between two methods of clinical measurement. *International Journal of Nursing Studies*, 47(8):931–936, August 2010.