Doctoral theses at NTNU, 2023:390

David Pasdeloup

Deep Learning in the Echocardiography Workflow: Challenges and Opportunities

Doctoral thesis

NTNU
Norwegian University of Science and Technology
Thesis for the Degree of
Philosophiae Doctor
Faculty of Medicine and Health Sciences
Department of Circulation and Medical Imaging

NTNU
Norwegian University of
Science and Technology

David Pasdeloup

Deep Learning in the Echocardiography Workflow: Challenges and Opportunities

Thesis for the Degree of Philosophiae Doctor

Trondheim, November 2023

Norwegian University of Science and Technology
Faculty of Medicine and Health Sciences
Department of Circulation and Medical Imaging

NTNU
Norwegian University of
Science and Technology

NTNU
Norwegian University of Science and Technology

Thesis for the Degree of Philosophiae Doctor

Faculty of Medicine and Health Sciences
Department of Circulation and Medical Imaging

ISBN 978-82-326-7480-0 (printed ver.)
ISBN 978-82-326-7479-4 (electronic ver.)
ISSN 1503-8181 (printed ver.)
ISSN 2703-8084 (online ver.)

IMT-report 2023:390

Doctoral theses at NTNU, 2023:390

Printed by NTNU Grafisk senter

Dyp læring i ekkokardiografi: Utfordringer og muligheter

Ekkokardiografi er den mest brukte avbildningsteknikken for hjerte, med høy tilgjengelighet, lave kostnader og sanntidsmuligheter. Med mange indikatorer for hjertets funksjon muliggjør ekkokardiografi nøyaktig diagnose av hjertesykdommer. Imidlertid praktiseres ekkokardiografi primært innad på sykehus av erfarne operatører og er begrenset av arbeidskrevende tolkninger og høy operatøravhengighet. På den andre siden har håndholdt ultralyd gjort sin inntreden i markedet, og tilbyr nye muligheter for bruk utenfor sykehus og i utviklingsland. Likevel kan nye brukere av håndholdte enheter ha begrenset ekspertise sammenlignet med sykehusleger, noe som kan hindre bruk av slike enheter i feltet.

Den overordnede målsetningen med denne avhandlingen er å undersøke bruken av dyp læring (DL) som en løsning til utfordringer knyttet til ekkokardiografi, både i og utenfor sykehuset. Først undersøkte vi utfordringene knyttet til bruk av DL til å automatisere tolkningen av ekkokardiografiske bilder. Videre ble potensialet for å bruke DL til å hjelpe operatører ved å veilede dem under skanning utforsket. Til slutt ble det utført en klinisk studie av den utviklede veiledningsapplikasjonen.

Resultatene indikerer at bruk av DL til å automatisere tolkningen av bilder i ekkokardiografi er utfordrende, og at utviklingen av klinisk verdifulle DL-modeller krever en tverrfaglig kompetanse innen ultralyd, medisin og statistikk. Disse funnene og de foreslåtte løsningene for å håndtere utfordringene kan ha implikasjoner for utvikling og evaluering av fremtidige DL verktøy innen ekkokardiografi. Videre viser den utviklede sanntidsapplikasjonen for veiledning av operatører at DL gir en utmerket mulighet til å lette og forbedre bildeakkvisisjon. Den kliniske studien viser at metoden er nyttig i sykehussammenheng, med mulighet for redusert operatøravhengighet i ekkokardiografi. Videre studier bør undersøke verdien av sanntidsveiledning for ikke-eksperter utenfor sykehusene.

David Pasdeloup
Institutt for sirkulasjon og bildediagnostikk, NTNU
Hovedveileder: Lasse Løvstakken
Biveiledere: Erik Smistad & Håvard Dalen
Finansieringskilde: SFI CIUS (Centre for Innovative Ultrasound Solutions)

Ovennevnte avhandling er funnet verdig til å forsvares offentlig for graden Philosophiae Doctor (PhD) i medisinsk teknologi. Disputas finner sted i auditorium MTA, Fred Kavli-bygget, tirsdag 28. november 2023, kl. 11.30

Abstract

Echocardiography is the most widely used imaging technique for cardiac imaging due to its availability, low-cost and real-time capabilities. With numerous indicators of heart function, it now enables precise diagnosis of heart diseases. However, echocardiography is primarily practiced within hospitals by expert operators and has some limitations, such as labor-intensive interpretations and high operator dependence. On the other hand, portable echocardiography devices have entered the market, offering new possibilities for point-of-care applications and increasing access to echocardiographic assessment in developing countries. Nonetheless, new operators using portable devices may have limited expertise compared to hospital residents, which could potentially hinder the full utilization of portable devices in the field.

The overall goal of this thesis is to investigate the use of deep learning (DL) techniques to address the limitations of echocardiography, from both the hospital and point-of-care perspectives. We first investigated the challenges when using DL to automate the interpretation of echocardiographic images. Secondly, the potential of using DL to assist operators by guiding them during scanning is explored. Finally, a clinical study of the developed guiding application is proposed.

Results indicate that using DL to automate echocardiographic image interpretation is challenging, and that the development of clinically valuable DL models requires expertise in ultrasound, clinical, and statistical domains. These findings and the proposed mitigation solutions may have implications for the development and evaluation of future DL tools for interpretation of echocardiographic images. Further, the proposed real-time application for guiding operators demonstrates that DL offers an excellent opportunity to facilitate and improve image acquisition. The clinical study show that the developed method is beneficial in the hospital setup, with the possibility to reduce the operator dependence of echocardiography. Further studies should investigate the value of real-time guiding for non-expert users in point-of-care settings.

Preface

This thesis is submitted in partial fulfillment of the requirements for the degree of *Philosophiae Doctor* (Ph.D.) at the Faculty of Medicine of the Norwegian University of Science and Technology (NTNU). The research was funded by the Centre for Innovative Ultrasound Solutions (CIUS) and was carried out at the Department of Circulation and Medical Imaging (ISB), NTNU. The main supervisor has been Professor Lasse Løvstakken, and co-supervisors have been Håvard Dalen and Erik Smistad, both from ISB, NTNU.

Acknowledgements

I want to express my gratitude to everyone who contributed to the completion this PhD. First of all, I want to thank my supervisor, Lasse, who welcomed me in the Ultrasound Group for my master project in 2019. Your combination of experience and dedication has been a driving force for my research, and I take pride in our achievements. I also want to thank Erik and Andreas for their continuous support and inspiring programming skills. Sindre, thank you for all our discussions and the fruitful collaboration that emerged from the synergy between a clinician and an engineer. Also, thank you, Sigbjørn, for your commitment and your invaluable help in the key moments. My appreciation also extends to the cardiologists - Håvard, Espen, Bjørnar and Stian - for their understanding and reactivity when acquiring new data. I would also like to thank the Faculty of Medicine, CIUS and the industry partners for giving me the opportunity to conduct this work.

All my colleagues in the Ultrasound Group deserve my profound gratitude. Sharing office and going on conferences with you has been both enriching and enjoyable. I am immensely proud to be a part of a community where ultrasound excellence converges with a warm and friendly atmosphere.

I would like to thank my parents for their support in all these years and for allowing me to pursue my dreams. Finally, to you, Ingeborg, I am so grateful for your patience, encouragements and love. The joy you bring into my life is immeasurable.

Table of Contents

Abbreviations

2D	Two-dimensional
3D	Three-dimensional
A4C	Apical four chamber view
A2C	Apical two chamber view
ALAX	Apical long-axis view
AI	Artificial intelligence
AUROC	Area under the receiver operating characteristic
BA	Bland-Altman
CNN	Convolutional neural network
DL	Deep learning
DOF	Degree of freedom
ECG	Electrocardiogram
ED	End-diastole
ES	End-systole
HF	Heart Failure
HFrEF	Heart failure with reduced ejection fraction
HFmrEF	Heart failure with mildly reduced ejection fraction
HFpEF	Heart failure with preserved ejection fraction
LoAs	Bland-Altman limits of agreement
LV	Left ventricle
LVEF	Left ventricular ejection fraction
MAE	Mean absolute error
ML	Machine learning
R2	Coefficient of determination
TTE	Transthoracic echocardiography
US	Ultrasound

Introduction

1

According to the 2023 World Heart Report [1], cardiovascular diseases were the leading cause of death worldwide in 2021, with 80% of these deaths occurring in low- and middle-income countries. Improving patient care necessitates a focus on all disease stages, from early diagnosis to follow-up. In this context, imaging plays a vital role, enabling healthcare professionals to make informed clinical decisions throughout the entire continuum of disease management.

Among several heart imaging techniques, echocardiography is the most used [2]. Based on the propagation of ultrasound waves in the body, it is today used to examine several hundred thousands patients worldwide everyday. The reasons for such a large adoption of echocardiography are, among others, its availability, its relative low cost, no harmful radiation and its real-time functionality which is invaluable when looking at a moving organ like the heart. Echocardiography has a wide range of applications, from quick patient screening in emergency units to quantification of the heart function and its evolution over time with measurements of clinical parameters such as left ventricular ejection fraction (LVEF). Today, the guidelines [2] recommend to perform a large number of quantitative measurements to assess cardiac function. However, echocardiographic images can be suboptimal, and, in worst case, only allow for qualitative measurements and assessments. This can be explained by the high workload relying on echocardiography laboratories and the lack of skilled clinicians to carry out these examinations, with large disparities across hospitals and continents. These challenges are expected to remain in the future with an increasingly aging population.

Meanwhile, we witness two technological advances with a potential impact on echocardiography. The first one relates to ultrasound imaging itself, with the entry of handheld ultrasound devices. A key enabler for portable ultrasound was the development of three-dimensional (3D) ultrasound imaging in the 1990s, that made possible to miniaturize hardware and to place it within the transducer, removing the need for

external processing units. Today, it is possible to connect an ultrasound transducer to a regular smartphone and to control it within an application downloaded from one of the main *app stores*. The portability of hand-held devices makes them even more available than cart-based scanners, opening opportunities for decentralizing echocardiography or using them in emergency vehicles. Their low price (<10'000€) compared to high-end cart-based scanners (>100'000€) can also contribute to spread ultrasound imaging in developing countries. However, out of all the organs that can be imaged with ultrasound, the heart is among the most difficult to image due to its complexity, dynamic nature and location within the thoracic cage. Achieving diagnostic relevant images requires training and practice, and the lack of skilled operators out of the hospital prevents the technology to be used to its full potential even though it is available. Lowering the bar to use the technology is necessary to enable more health personnel perform echocardiography with portable devices. Making echocardiographic images easier to acquire and interpret could contribute to redefine the care path of heart disease, with better prioritization and earlier detection of cardiac dysfunction.

The second technical change relates to the wide and rapid spread of methods in Artificial Intelligence (AI). Although the theory behind these methods was established in 1943 [3], AI became a major research topic in the 2010s, particularly with deep learning (DL). A key enabler for the recent advances in the field is the availability of powerful enough hardware, especially graphic processing units (GPUs). DL has shown to surpass human performance at some tasks, for example at games [4], and to be able to automate certain tasks like driving cars. However, AI and DL development have been accompanied by notorious failure cases, for example by discriminating people [5] or by being involved in infamous traffic fatalities [6]. This has raised a legitimate societal concerns about the use, and possible misuse, of these technologies, also nourished by a lack of understanding from the general public on how they work. In medical imaging, DL has successfully been applied to some tasks like skin disease classification and MRI image reconstruction, but was not free from failures there either. In a review study [7], it was showed that out of 62 DL algorithms for COVID-19 detection, none were of potential clinical use. Some methodological failures [8] such as the lack of metrics that reflect the clinical outcome, the use of biased datasets or unnecessary complex models could explain this finding.

DL methods have also been applied to echocardiography with promising results. They can be used at all stages of the echocardiography workflow:

during acquisition to obtain better images [9], after acquisition to automate image interpretation and reduce echocardiography laboratories workload [10–14], and finally in the clinical decision path [15]. Another promise of DL in ultrasound is to reduce the large inter- and intra-observer variability associated with ultrasound image acquisition and manual interpretation [16–19]. This could make it possible to detect smaller changes in the measured clinical indices, which is especially relevant for repeated examinations and patient follow-up. Finally, in the context of portable echocardiography at the point of care, DL can be used to emulate the knowledge of an echocardiography expert, and contribute to breaking the skills barrier of non-expert users. However, considering the low number of DL solutions currently available from the main echocardiography equipment vendors, there are reasons to think that DL for echocardiographic image analysis has not yet reached its maturity stage.

1.1 Thesis outline

This first chapter introduces the current echocardiography workflow and its limitations, providing the motivations and aims of this thesis. Each contribution is briefly summarized, highlighting how DL techniques were used to achieve the predefined aims. Furthermore, the discussion section offers additional insights into the presented work and contextualizes the contributions.

In Chapter 2, the essential background information required to understand this thesis is provided. This chapter offers relevant fundamentals and terminology related to echocardiography, DL, and statistics, and their interconnections within the context of this study. Additionally, where necessary, further elaboration on the methods employed in the contributions is also presented.

In Chapters 3 and 4, the two technical contributions of this thesis are included. Finally, Chapter 5 presents a clinical investigation of the method introduced in Chapter 4.

1.2 Echocardiography workflow

The workflow of a complete trans-thoracic echocardiography (TTE) examination in the echocardiography laboratory has three principal steps: the image acquisition at the bed side, the measurement of clinical indices on these images, and finally the diagnosis based on both the echocardiography findings and other inputs. Other forms of echocardiography such as emergency echocardiography or operative echocardiography does not belong to the workflow described below. The focus of this thesis is TTE, which is further referred to as echocardiography for simplicity.

Image acquisition

Echocardiography starts with image acquisition, with patients lying on the bed and the heart being imaged with a transducer placed on their chest, pointing towards the heart. Image acquisition is a challenging task for the operators for multiple reasons.

First, they have to deal with an adversarial combination of human physiology and ultrasound physics, with the heart being placed within the thoracic cage, behind the ribs and surrounded by the lungs. Given that the ultrasound waves do not propagate well through bones and the air contained in the lungs, echocardiography challenges the operators in finding the acoustic window giving the best possible image quality, with a high signal to noise ratio and limited occurrence of artifacts such as reverberations or acoustic shadows.

In addition to the challenge of finding a suitable acoustic window, operators have to follow guidelines and recommendations on view standardization to visualize the relevant structures. This is most easy to achieve using 3D echocardiography where the user have a direct access to the spatial context, at the cost of a limited image quality and frame rate. However, low-end and portable echocardiography devices are restricted to two-dimensional (2D) imaging, relying on numerous *standard views*, corresponding to predefined anatomical cut-planes of the heart and the presence, or absence, of certain structures in the images. Obtaining standard views is highly important and is a necessary condition for performing future quantitative interpretation. Protocols define which measurements should be performed on which standard view, so that they can be compared with normal values or with measurements from previous examinations. View standardization is, however, difficult to achieve and require from the operators a good mental representation of the heart to maneuver the

transducer towards the optimal imaging plane and the intended standard view.

In this thesis, image quality and view standardization are considered as two different concepts, and only the latter is addressed. In practice, they are interconnected as the operators have to maintain a satisfying image quality while maneuvering the transducer towards optimized views. The echocardiographic image acquisition step is consequently dependent on both operators skills and patient body-habitus.

Image interpretation

Acquired images are often stored in hospital's PACSs (picture archiving and communication system) while the patients leave the echocardiography laboratory. Measurements of clinical indices are performed later on separate work stations, either by the operator who acquired the images or another one. From a complete examination, multiple measurement can be obtained, representing sizes, volumes, blood velocities or tissue velocities. They form the basis for further clinical decision-making.

Manual interpretation is a labor intensive task that requires multiple actions from the operator. The most common actions are choosing the appropriate recording for the measurement to be performed, choosing the suitable time point within the recording and manually tracing contours and distances. It is worth noting that some clinical indices like the left ventricular ejection fraction are obtained from a combination of multiple measurements, eventually from different ultrasound recordings, increasing the interpretation time and the number of error sources. Asking multiple operators to do the same measurements often leads to a large variation known as inter-observer variability, and there is also even a certain variability when having the same operator do the measurements multiple times known as intra-observer variability. A high inter- and intra-observer variability can lead to high uncertainty of the measurements and thereby the diagnosis.

Diagnosis

Finally, clinicians use echocardiographic measurements to make a diagnosis, often in combination with other clinical inputs such as symptoms, family history, and measurements based on auscultation, ECG or blood samples. These measurements can be used to both make an initial diagnosis, and to follow up patients, for example after heart failure to evaluate the reaction to

medical treatments and surgery. Another application of echocardiography is follow up of patients undergoing cardiotoxic cancer treatments, in which a degradation of the heart function can lead to the decision of stopping cancer treatment.

1.3 Limitations of the echocardiography workflow

Due to the difficulties inherent to image acquisition, echocardiography recordings can be suboptimal, both in terms of image quality and image standardization, which negatively impacts the workflow downstream. Low image quality makes interpretation more difficult, for example when the operators trace the border between blood and heart muscles, limiting the precision of subsequent measurements. Lack of image standardization can lead to measurements of the wrong heart structures. Together, poor image quality and the lack of standardization introduce additional errors, further called *acquisition error* (ϵ_{acq}), causing the measured values to deviate from their *true* values. Since the patients have left the examination room, improved recordings cannot be acquired, resulting in measurements performed on suboptimal recordings, or not performed at all.

Independent of the error introduced during the acquisition, there is also some variability introduced at the interpretation step, where the possible uncertainty sources (choice of recording and timepoint, tracings, etc.) accumulates. This results in large intra- and inter-observer variability in the interpretation, which indicates the presence of what is called further *interpretation error* (ϵ_{int}).

Manual measurements of any clinical indices M_{CI} are therefore an estimation of the *true* value T_{CI} rather than the *true* value itself. M_{CI} can be expressed as:

$$M_{CI} = T_{CI} + \epsilon_{acq} + \epsilon_{int} \tag{1.1}$$

According to [19], both the acquisition and the interpretation error accounts for a large part of the final measurement in echocardiography. For the LVEF with the Simpson's method, they reported a 95% upper limit of the absolute measurement value difference of 7.8% for the acquisition variability and of 7.1% for the interpretation variability. This suggests that the variability introduced in the acquisition step and interpretation step can be of similar magnitude. However, as the variability is not purely additive, the effect of reducing variability in one step of the workflow can be hidden by the remaining variability in the other step. To effectively reduce variability in the echocardiography workflow, one should reduce the errors

in both the acquisition and interpretation steps.

Despite their high variability, echocardiography measurements forms the basis for clinical decisions. The variability is particularly challenging in the case of repeated examinations since it does not allow to detect small changes in the heart function. For LVEF, the minimum detectable change is about 8.5% to 10% [20, 21], which is relatively large compared to the range of physiologically possible LVEF, between 10% and 70%.

1.4 Motivations and aims of the thesis

The overall aim of this thesis is to investigate the use of DL to address the aforementioned limitations of the echocardiography workflow. Overcoming these limitations can benefit expert users within the hospital echocardiography laboratory and enable non-expert users to perform echocardiography at the point of care.

From a non-expert user perspective, an aim of this thesis is to make echocardiography image acquisition easier in order to use the hand-held devices to their full potential. From an expert user perspective, the aim is to ensure highly standardized views in order to reduce the *acquisition error* for more informed clinical decisions.

Regarding the interpretation step of the workflow, using DL has been a topic of high interest in the last years. However, recent review studies [7, 8] invite to moderate the expectations towards the use of DL for medical image analysis. Identifying the strengths and weaknesses of DL in the context of automatic echocardiographic measurements is therefore of utmost importance.

On this basis, the specific aims of this thesis are:

- **Aim 1:** Investigate the challenges related to the use of DL for automatic echocardiography measurements, and propose strategies to mitigate these challenges
- **Aim 2:** Investigate the use and potential benefits of DL to facilitate image acquisition and to improve view standardization

1.5 Summary of the presented work

This section summarizes the three contributions included in this thesis. The first contribution is a technical contribution that relates to the use of DL to automate interpretation of echocardiographic recordings and the associated challenges. The second and third contributions investigate the use of DL during echocardiographic image acquisition to improve view standardization, from technical and clinical perspective respectively.

1.5.1 Challenges and Strategies for Automatic Measurements with Deep Learning in Cardiovascular Imaging

Deep Learning has shown potential to automate interpretation of echocardiographic recordings and to extract measurements from them, by removing the observer variability [22] and possibly reducing the *interpretation error*. However, automatic measurements with DL have not yet fully reached commercial systems and clinical practice, suggesting that some challenges remain unsolved.

In this contribution, the left ventricular ejection fraction (LVEF) measurement serves as the use case to investigate the challenges related to the use of DL to fully automate the interpretation of echocardiographic data in and end-to-end fashion, where the measurement is directly estimated from the ultrasound data. First a baseline DL framework is established. This framework is similar to [23, 24], where the LVEF is estimated from an apical four chamber recording. The baseline model was then tested on two additional external datasets to evaluate its relevance in a clinical context.

Performance of the established baseline framework was in line with the one originally reported in [23, 24] when using the same testing setup. However, performance on external evaluation datasets was below what could be expected, suggesting that further investigations were needed. By analyzing the shortcomings of the baseline framework, three main challenges related to the use of DL to automate echocardiographic measurements were identified. An overarching challenge was related to the evaluation of the DL algorithm, where we showed that common metrics may not be robust enough to assess whether a DL algorithm has clinical value. The second challenge was related to the training data, where imbalance and lack of diversity resulted in suboptimal performance of the trained model. The last challenge was related to the generalizability and robustness of DL models, where we showed that they can perform well on a single dataset within an initial research setting, but that performance degrades when using

the models on new datasets, thereby limiting their clinical relevance.

Through the investigation of these shortcomings, some possible mitigation solutions emerged. For the evaluation challenge, we proposed an extended Bland-Altman analysis that is more independent of the test data demographics, hence allowing the comparison of algorithm performance across datasets, and that better captures the clinical problem at hand. To mitigate the training data challenge, we proposed to optimize patient selection with two strategies: oversampling and enrichment. Finally, the generalization challenge was addressed by introducing domain knowledge into the DL framework, mainly in the form of data augmentations. Overall, this contribution highlights the gap between a proof-of-concept DL model working in a research setting and a clinically useful measurement method that is expected to perform properly on all patient populations.

This paper has been submitted for review to Journal of the American College of Cardiology: Cardiovascular Imaging, and is presented here in its current form. The candidate was the main contributor to all aspects of the work, except for acquisition of ultrasound data.

1.5.2 Real-Time Echocardiography Guidance for Optimized Apical Standard Views

Finding the correct 2D cut-plane of the heart represents a cognitive challenge for the echocardiography operators who do not have access to the spatial context, but solely rely on a mental representation of the heart. To simplify the task of apical standard views acquisition, we propose to use DL to estimate the position of the 2D cut-plane relatively to the heart, and further guide the echocardiography operators by proposing adequate transducer movements towards an improved cut-plane. The overall aim of the method is to optimize the standardization of the structures visualized at the mitral valve depth.

This second contribution builds upon the method proposed by [11] which uses 2D slices of 3D ultrasound volumes to generate 2D training data with a reference position. The strength of this approach is that an exact reference position for each 2D slice can be obtained directly from the 3D volume. Extensions from [11] are multiple, with an annotation method that takes account for variation in the heart morphology and an improved position feedback that relates to both the rotation and tilt of the transducer. The trained deep neural networks were extensively tested on multiple external datasets representative of clinical routine, showing

robustness to new data and ability to detect minor view standardization errors. The same deep neural networks were also implemented in a real-time application, further called scanning assistant, where the current position of the 2D cut-plane was estimated and recommendations to obtain better views were displayed on a dedicated user-interface. Preliminary testing of the scanning assistant showed the potential to detect and quantify the correct and sub-optimal positioning of apical standard views in real-time, as well as proposing the correct transducer movements to obtain better standard views when possible, as shown in Video 4.1.

This paper was published in Ultrasound in Medicine and Biology (UMB), Volume 49, Issue 1, pages 333-346, January 2023. It is presented here in its original form. The candidate was the main contributor to all aspects of the work, except for acquisition of ultrasound data.

1.5.3 Real-time Guiding by Deep Learning of Experienced Operators to Improve Standardization of Echocardiographic Acquisitions

In the previous contribution, we presented the technical building blocks of a DL based scanning assistant to guide echocardiography users during the acquisition of apical standard views. This work is a clinical study of the previously described method, in the context of expert echocardiography users. In the first inclusion period of the study, sonographers were told to acquire apical standard views for quantification of heart function, without knowledge of the actual study aims. After the first inclusion period, the sonographers were made aware of the study aims and got to familiarize with the scanning assistant and its user interface on 10 patients each. In the second inclusion period, each patient got three examinations, a normal one by a sonographer, one by a sonographer with the scanning assistant and one by a cardiologist. The latter was used as a reference. This protocol allowed to compare view standardization for sonographers before they got informed of the study aims, sonographers who have previously been trained with the scanning assistant and sonographers actively using the assistant while scanning. Additionally, this protocol allowed comparison of sonographers with cardiologists which were considered as the reference in terms of standardization.

The study had two end-points. For the primary end-point, an expert cardiologist scored the standardization of all recordings for both periods, blinded to inclusion period, operator, and use of the scanning assistant

or not. The proportion of standardized recordings increased by at least 80% (depending on the view, $p<0.01$) when using the scanning assistant in the second period compared to the first period in which the sonographers did not used the assistant and were not informed of the study aims. Considering the primary end point being dependent on the subjectivity of the cardiologist who scored standardization, we introduced a secondary end-point, in which the view standardization was measured on all recordings by the DL method itself. Results from this secondary end-point also indicated a significant increase ($p<0.05$) in standardization when the sonographer used the scanning assistant. Overall, the agreement between the manual scoring by expert cardiologist and automatic DL analysis reinforces the conclusion on the ability of the trained deep neural networks to increase view standardization. It further suggests that the presented method could be used in real-time to increase standardization of apical standard views, but also retrospectively as a standardization control tool.

This paper has been submitted for review to Journal of the American College of Cardiology: Cardiovascular Imaging, and is presented here in its current form. The candidate was the second author and contributed to development of the employed DL method, the statistical analysis, as well as writing the manuscript. Sigbjørn Sæbø was the first author and primal investigator.

1.6 Discussion of results

This thesis focuses on exploring the challenges and opportunities associated with utilizing DL to enhance the echocardiography workflow. In the first contribution, three key challenges related to the translation of DL methods from research to clinical practice were identified, and corresponding solutions were proposed. Subsequently, this thesis demonstrates that DL can effectively aid echocardiography operators during image acquisition, representing a significant opportunity for workflow improvement.

1.6.1 Challenges and Strategies for Automatic Measurements with Deep Learning in Cardiovascular Imaging

The contribution presented in Chapter 3 highlighted three of the challenges of using DL to automate echocardiographic measurements. These challenges were illustrated with the specific use case of left ventricular ejection

fraction (LVEF) estimation with end-to-end DL, with focus on the clinical task of heart failure management. Although the chosen end-to-end learning approach was potentially amplifying the challenges, it is believed that the highlighted challenges and their mitigation strategies are also relevant to other approaches such as multi-step approaches, which automate each individual step of the manual interpretation. The three identified challenges are discussed in details in the subsections below.

Evaluation challenge

Regarding the evaluation challenge, the proposed extended Bland-Altman analysis was an attempt to make metrics more independent on the data, allowing a more fair comparison of the performance of the DL methods on multiple datasets and benchmarking of the DL methods themselves. The proposed analysis does not have the ambition of setting a new standard for evaluation of the DL. It may have some flaws and is more complex than a regular Bland-Altman analysis, which limits its understandability. The main objective of the proposed extended Bland-Altman analysis was to illustrate that DL measurement methods, especially end-to-end ones, may not fulfill the assumptions for the regular Bland-Altman analysis and that there is need for an unified approach to quantify and benchmark the performance of DL methods, independently of the data they are being tested on.

In echocardiography, the evaluation procedure is often limited by the lack of gold standard for the reference values, making the *true* measurement value inaccessible. In such situation, it is not straightforward to separate the error in the human reference values from the error the DL estimated values. In the worst possible case, one may have an ideal DL model estimating the exact *true* measurement values, but observe unsatisfactory performance metrics due to the error in the human reference values. This makes it difficult to demonstrate that a DL measurement method can surpass human performance. Specifically for LVEF estimation, [25] claim superiority of their DL method compared to sonographers. Their results are however questionable as DL measurements themselves were used as the reference for most of the subjects. One approach to obtain better manual references could be to use multiple observers to analyze the images in order to obtain a consensus reference values. The limitation of this approach is the amount of manual work required, which hampers the possibility to evaluate DL measurement methods on large datasets and multiple patient subgroups. Another worthwhile approach is to obtain LVEF reference values from MRI

images and to compare them to DL estimations from echocardiographic recordings. MRI measurements of LVEF rely on manual tracing of the LVEF contour, which also may limit the accuracy of the reference values. LVEF calculation without manual tracing and potentially better reference values, could be obtained with multi-gated acquisition (MUGA) scanning, where the volumes of the left ventricle are obtained by counting radioactive particles with a gamma camera. This approach however involves the use of a radiotracer, which limits prospective acquisition of evaluation data. Retrospective use of MUGA references values is not fully possible either as MUGA is recommended for only a certain type of pathology (mostly for patients undergoing cardiotoxic cancer treatment), hence limiting the diversity in the evaluation data.

Finally, it was observed in our work a correlation between the difference and average in the Bland-Altman plot, which may be due to a reduced DL measurement error compared to the manual measurement error. Whether this is an artifact or a real consequence of reduced measurement error warrants further investigations, with the possibility to establish a formalism to demonstrate that one measurement method has a reduced error compared to the other.

Training data challenge

The second challenge highlighted by this contribution was related to the training data, and its lack of diversity. As mitigations for these issues, two techniques were proposed: data enrichment and data upsampling. The former is relevant for prospective acquisition or annotation of data, where one intend to maximize data diversity relatively to the task to be solved. The latter is relevant for retrospective use of data.

These two techniques intend to make the deep neural network learn the relevant heart features in the images rather than the underlying statistic of the training dataset. For the presented use case, both these techniques contributed to improve the performance of the trained deep neural network, without increasing the data amount. These two techniques were especially useful for small dataset sizes, suggesting that increasing the data amount may not necessarily be the first action to increase DL performance. The effect of these techniques vanished for larger dataset sizes, but it was not found that they degraded the performance.

It is also worth noting that end-to-end learning methods are trained to directly estimate a measurement value. This makes them prone to learn underlying statistical patterns from of the training data, and may

introduce issues such as heteroscedasticity. In that case, data resampling and enrichment strategies were particularly useful to mitigate this type of issue. On the other hand, in multi-step DL pipelines, each component is trained to perform a specific action, independently of the measurement value. Thus, multi-step DL pipelines are less prone to learn the underlying statistics in the training data, and the value of data resampling and enrichment strategies can be expected to be less prominent than in end-to-end learning.

Another approach for increasing training data diversity could be to use data from multiple datasets. However, this introduces the risk of the dataset acting as a confounding factor, in a situation where the DL method learns shortcuts and associates its output to the dataset rather than to the images features [26]. Mixing of multiple datasets warrants extreme carefulness and rigorous testing.

Generalization challenge

Generalizability and robustness of the DL method was the third challenge investigated in this contribution. To start with, it is worth noting that the available data in fundamental DL research is much more heterogeneous than in DL applied to echocardiography. Size of the datasets plays an important role in heterogeneity. ImageNet [27], one of the largest image datasets contains >14'000'000 images whereas the largest echocardiography datasets in 2023 contains around 10'000 recordings [24]. Beyond the size of the datasets, heterogeneity also depends on the way the datasets are constructed. ImageNet is constructed with images from the web, compiling mostly images from individuals taken with their own camera. On the other hand echocardiography datasets are constructed with images often acquired within a single hospital and/or with devices from the same manufacturer. Additionally, echocardiography datasets follow strict privacy regulations. As a consequence, the data domain is larger for general purpose datasets than for echocardiography datasets, making generalization to never seen data a greater challenge for the latter.

Limited generalization capabilities were confirmed in the presented echocardiography use case, where the initial baseline DL model showed good performance on internal testing data, but significantly degraded performance on external testing datasets from other centers acquired with scanners from other manufacturers. This motivated the use of data augmentations, a way of introducing echocardiography domain knowledge into the DL framework, to mitigate the observed generalization issue.

In practice, these augmentations intended to deconstruct the typical appearance of an echocardiography image and to remove the image attributes that could be linked to the manufacturer of the scanner, whereas they preserved the anatomy of the heart structure. The aim of this was to increase the focus of the deep neural network on the heart structures, regardless of the assumptions of what an ultrasound image looks like and of the post-processing techniques. The improved performance on the external datasets when using these augmentations suggests the validity of the approach.

In echocardiography, generalization is a more complex problem than only the image appearance. For applications using spatiotemporal data as in the presented use case, it could also have been relevant to consider the frame rate as a potential generalization issue since the internal and external datasets had limited overlap in the frame-rate of their respective recordings. Further, there are differences in the clinical protocols which may cause generalization issues and may be worth investigating. One example is the use of apical standard views focused on the left ventricle, by some centers, whereas other centers also include the left atrium. Operator experience and variable image standardization across datasets may also introduce performance disparities. To this regard, indication of image standardization, for example obtained with techniques such as those presented in Chapter 4, may help to quantify standardization at dataset level, and further interpret the generalization results across multiple datasets. Beyond these possible differences in the input image data, the precision and accuracy of the reference values plays a role in quantification of generalization. When comparing generalization across two datasets, if one have significantly less precise and accurate reference values than the other, differences in the performance metrics would arise independently of the image generalization capabilities of the DL model.

Finally, the limited heterogeneity of echocardiographic datasets questions the common DL practices of using a subset of the internal dataset to tune model hyperparameters and choose the best model parameters, independently of other external datasets. This practice relies on the assumption that the internal validation data is representative of future data, so that the hyperparameters choice and model parameters are optimal for future data. In the eventuality where the validation data is not heterogeneous enough, and consequently not representative of future clinical data, there is a risk for the DL model to be over-optimized to the internal data.

Future work

In the presented work, focus was placed on the challenges related to the use of DL to automate echocardiographic measurements, with emphasis on quantifying the improvements allowed by the proposed mitigation strategies. Establishing state of-the art performance was not the goal of this work.

In further work, emphasis could be placed on maximizing the performance for clinical use of the LVEF predictions. Adding other views than the apical four chamber to the framework could contribute to make more informed automatic estimations. Supporting low frame rate is also relevant to investigate considering that portable echocardiography devices operate at a limited frame rate and potentially miss the end-diastolic and the end-systolic times. Beyond the external testing related to new devices presented in this work, it is also important for clinical use to test thoroughly the method on different pathology subgroups in order to guarantee satisfying performance on a wide spectrum of patients.

Another direction worth investigating is to compare the presented end-to-end DL approach with multi-step approaches similar to [12] for the estimation of LVEF. Comparison at both the dataset level and of individual subjects could provide new insights on the strengths and limitations of each method. Further, end-to-end and multi-step approaches could be used in parallel to provide a single measurement with an estimation of the uncertainty. Such a redundant system could help detect failure cases when the two approaches disagree on the estimated measurement value and subsequently warn the operator. Ideally, redundant measurement methods should use different DL approaches, different training data and be designed by different teams to ensure their independence.

Beyond LVEF estimation, end-to-end DL approaches can also be used for other echocardiography interpretation tasks. Considering their ability to deal with spatiotemporal data, a potential use case could be to analyze aortic or mitral valve movements and the presence of valve insufficiency (leakage) or stenosis. In current hospital practice, these diseases are usually screened from a parasternal long-axis view and assigned an initial semi-quantitative measurement such as *healthy*, *mild*, *moderate* or *severe*. For subjects with findings, this semi-quantitative measurement is refined with the help of a more detailed examination, relying on multiple heart views, high-end scanners and highly qualified clinicians. From a DL perspective, this means that there exists data pairs where parasternal long-axis views are associated with a refined semi-quantitative measurement derived from

a comprehensive examination. A question worth further investigation is whether a DL model is able to retrieve the refined semi-quantitative measurement from a parasternal long-axis view only, using image features clinicians may not see. If proven possible, this could contribute to enable screening of valve disease at the point of care, with portable devices and without the need for expert users. A similar approach has recently been used to to detect heart failure with preserved ejection fraction from a single apical four chamber view (A4C) recording [28].

1.6.2 Real-Time Echocardiography Guidance for Optimized Apical Standard Views

The contribution presented in Chapter 4 describes a DL scanning assistant designed to help the echocardiography operators at the acquisition step of the echocardiography workflow. The core components of the scanning assistant are the DL models in charge of analyzing the images. The outputs of the DL models are then converted into a feedback which is presented to the operator through an intuitive user interface. The DL models can also be used independently of the real-time scanning assistant for retrospective analysis of echocardiographic recordings.

As of 2023, the concept of real-time guiding of echocardiography users to acquire standardized views of the heart has been addressed by few technical research publications (clinical publications are discussed later in Chapter 1). In 2012, [29] used a Kalman filter method to help the operators acquire the apical four chamber view. However, their approach is limited by the fact that only this view is studied. In 2019, [30] investigated the use of DL to guide echocardiography users. Their work was limited by the lack of prospective testing to demonstrate the validity of the proposed method. More recently, in 2018, [11] published what forms the base of the contribution presented in Chapter 4, where the key idea was to use 2D slices of 3D echocardiography recordings as the training material of a deep neural network.

The main advantage of using 2D slices from 3D recordings is the possibility to simulate transducer movements within the volume and to express these movements directly in a coordinate system attached to the heart. We believe this is more reliable and more practical than using an external positional sensor placed on the transducer, where the coordinate system is attached to the transducer. The use of external sensor is prone

to errors due to patient body and respiratory movements (the heart moves within the chest when the patient breathes) and drift of the positional sensor. For maximal context information in the 2D slices, only 3D recordings with a large field of view and increased scanning depth were used.

A consequence of using 2D slices from 3D volumes for the training is that the slices have a different appearance compared to the 2D data normally used in the clinic, with differences in the image shape and the characteristic speckle pattern. Similarly to the contribution from Chapter 3, generalization from 2D slices to regular 2D recordings was achieved by using data augmentation deconstructing the appearance of echocardiography images but preserving the heart structures in order to set the focus of the DL models onto the heart structures only. Additionally, a blurring layer was added straight after the deep neural network input as a way of smoothing the speckle pattern with the aim of making the 2D slices and regular 2D recordings more similar in appearance. The performance improvements related to the use of data augmentation and pre-processing suggests the validity of the proposed techniques, similarly to the previous contribution.

Another challenge was the evaluation of the deep neural networks as, for the present use case, 2D slices from 3D volumes were not representative of clinical use. Metrics obtained on such 2D slices have consequently limited clinical value. Further, there is a lack of clear guidelines that define what perfect apical standard views are, and are not, making the extraction of quantitative performance metrics challenging. Considering the clinical task - guiding the operator transducer movements towards optimal apical standard views - it was chosen to create a specific quantitative metric grasping whether the DL method was able to propose transducer movements in the correct direction, independently of the definition of a standard view. In the lack of clear consensus on how the correct standard views can be defined, the DL method was calibrated using two large retrospective datasets in which it was assumed that the recordings were standardized on average. Finally, the calibrated DL models were wrapped into the scanning assistant, which was shown to be able to detect both standard and non-standard apical views accurately, providing another evidence of the method validity.

The main limitation of the presented work lies in the assumptions that the operator had found the apex of the heart, removing the need for providing a feedback on transducer translations in the skin plane. This is a strong assumption knowing that non-expert users may have difficulties to

find the heart apex and the correct intercostal space. At the time being, the proposed method appears as more mature to help experienced users like sonographers rather than to guide non-expert users.

Future work

The limitations of the current method should be addressed in further work to allow transducer guidance along more degrees of freedom. Finding the apex of the heart is challenging as it involves out of plane translations of the transducer, for example from one intercostal space to another. For such movements, it may not possible to give a continuous feedback to the operator as the signal quality may fall dramatically when the transducer comes over a rib, hiding all heart structures and preventing any DL algorithm to make a consistent prediction. Additionally, it is unsure whether the approach of using slices from 3D recordings is usable to predict translations in the skin plane as 3D recordings provide little context information and have a degraded signal quality at low depth. Apical foreshortening detection, as proposed by [31] and [32], could help warn the used when the apex is not below the transducer. These approaches however do not indicate in which direction the transducer should be translated. It could also be worth investigating an approach involving classification of all the possible transducer position errors combined with heuristic rules to determine the required transducer movements.

Transducer tilt displacements in the imaging plane were not investigated in the presented work and also warrant further research. Unlike out-of-plane displacements, the required heart structures are visible, at least partly, in the image. The objective of the in-plane displacements is to optimize the position of these structures within the field of view, for example by aligning the ventricle axis with the transducer axis. The required transducer movements for such optimizations could for example be derived from the geometrical transformation between the ideal positioning of some heart landmarks within the field of view and the current position of these landmarks estimated by existing 2D segmentation tools of the left ventricle [33].

Finally, performing a pilot clinical study involving non-expert users could also be relevant in order to better align the clinical need and the foreseen necessary improvements of the method.

1.6.3 Real-time Guiding by Deep Learning of Experienced Operators to Improve Standardization of Echocardiographic Acquisitions

The contribution presented in Chapter 5 is a clinical study of the technical methods presented in Chapter 4. From a clinical perspective, guiding echocardiography users has been a topic of interest, especially for start-up companies surfing on the DL wave.

In line with the limitations of the technical method, this study targeted echocardiography expert users who are able to find the correct initial placement of the transducer. This is a notable difference with other studies which focused on non-experts users. Another difference is that the scanning assistant under investigation focuses only on the three apical standard views, whereas other studies and start-ups products have claims on more standard views. Our choice for the apical standard views was based on the high relevance of these views in clinical practice and the difficulty to acquire them, compared to views from the parasternal window which are more accessible. To the best of our knowledge, this presented study is the first using a control group of operators not using the DL tool, and hence the first in which the actual value of a DL guiding tool is quantified. It is unfortunate that all three studies investigating the value of the Caption Health guiding tool [34–36] have the same limitation, namely the absence of a control group of operators having the same level of experience, but do not use the tool.

In this work, the view standardization was addressed independently from the image quality, although they are coupled in practice. It was shown that, within a single patient, sonographers using the scanning assistant achieved better standardization than sonographers who did not. This suggests that there is room for view optimization, independently of the body habitus of the patients and how challenging they are to scan with regard to image quality.

A relevant insight gained in this study, confirming observations done during the technical development of the method, relates to the standardization of the apical two chamber views. The retrospective DL analysis of all the recordings indicated that operators not using the real-time DL tool, including cardiologists, had a smaller amount of rotation from apical four chamber to apical two chamber than sonographers using the real-time DL tool. Meanwhile, manual retrospective assessment of the recordings did not show that the amount of standardized recordings acquired by sonographers using the scanning assistant was significantly different than for recordings

acquired by cardiologists. These two observations together suggest that when acquiring the apical two chamber view, rotating the transducer further away from the apical four chamber view than what is sometimes done in clinical practice still results in acceptable two chamber views. From a clinical perspective, this could be interpreted by the fact that the operators, after having acquired the apical four chamber view, rotate the transducer until the right ventricle disappears and only two chamber remain visible. As exemplified in Video 4.2, this can lead to views with a thick septal wall or a slightly visible right ventricle during part of the heart cycle. Further rotation from the position where the right ventricle disappears, as suggested by the scanning assistant, mitigate this issue and lead to more anatomically correct apical two chamber views. This insight has potential implications for measurements using the apical two chamber view, especially left ventricle volumes and ejection fraction measurements. Indeed, the biplane method of disks to calculate these indices assumes two orthogonal imaging planes. In situations where the rotational position of the apical two chamber is too close to the the apical four chamber, this assumption is violated and the biplane volume calculation regress towards a repeated monoplane volume calculation. In this context, the proposed scanning assistant can be relevant to ensure a sufficient amount of transducer rotation between the apical four chamber and apical two chamber views.

One strength of the presented DL method was to help the operator acquire the three best cut planes corresponding the three optimal apical standard views, improving from the common assumption that the views are separated from each other by 60° transducer rotation. However, some clinicians rely on the 60° assumption between standard views as it easily connects to the bulleye plots, which are highly relevant to strain analysis and coronary perfusion. This illustrates that echocardiography practices are non-unified, which challenges the development and adoption of generic tools.

The present clinical study also revealed that sonographers using the assistant used around three times as much time compared to the sonographers not using the assistant for acquiring the three apical standard views. This increase in acquisition time could also have contributed to the increased standardization that was observed, independently of the use of the assistant. Further, increased acquisition time is not acceptable for current clinical practice with overloaded echocardiography laboratories. It is believed that integrating the scanning assistant into the scanner display rather than using a prototype with an external screen can reduce this time. Additionally, the inclusions were performed over two short time intervals

and each sonographer could familiarize with the scanning assistant with 10 patients, allowing for minimal learning only. It is expected that acquisition time decreases after a longer period as the operators get more familiar with the scanning assistant.

Future work

In this clinical study, the main focus was on evaluating the benefits for sonographers of using the real-time scanning assistant. Further work should investigate whether sonographers can improve their standardization skills over a longer period of time, without systematic use of the scanning assistant.

As mentioned earlier, studying the effect of the proposed real-time scanning assistant on non-expert users is relevant to establish preliminary results and identify directions for future technical developments. Ideally, such study should be performed using portable ultrasound devices and at the point of care to capture the clinical value and limitations of the proposed assistant when used out of the hospital.

Another use case of the presented method that should be investigated is teaching of echocardiography. For this use case, the method is not used in a diagnostic process, which reduces the risk in case of failure. It is believed that the intuitive feedback of the scanning assistant and its responsiveness can help medical students and doctors in specialization to learn echocardiography.

The present study also showed that the core deep neural networks also can provide reliable analysis of view standardization on retrospective settings. A potential use case of the method could be to generate retrospective reports of view standardization of individual users within entities such echocardiography laboratories. By identifying scanning habits and eventual mistakes of individual operators, standardization reporting could help to prioritize training for operators who need it the most and improve performance of echocardiography laboratories in the long run. In this use case, the risk associated with using the DL method is low as there is no direct interference with the patient diagnosis or treatment. If the method is proven working, this way of interacting with clinicians could also contribute to increase their trust into DL tools for echocardiography.

The last use case of the proposed method is data curation and quantification of view standardization of large datasets for research purposes. Better insights on the data at hand can indeed help to adapt study design and to interpret the results.

Finally, it is decisive for the adoption of the proposed method to demonstrate its added value for further measurements of the heart function. In a point of care scenario with handheld ultrasound, the method should enable non-expert operators to record images useful for initial measurements. In the hospital scenario, acceptance of the method is conditioned by its ability to reduce the measurement error related to the image acquisition. Proving this is however challenging due to the additional interpretation error introduced when the measurements are performed. Future research should therefore focus on quantifying the acquisition error independently of the measurement error.

1.7 Concluding remarks

In this thesis, the focus has been to investigate the use of DL in the echocardiography workflow. Results indicate that echocardiographic image interpretation with DL is challenging and that addressing these challenges requires ultrasound, medical and statistical knowledge. When aware of these challenges, DL represents immense opportunities for echocardiography. It has the potential to improve workflow in hospitals echocardiography laboratories, but also to increase the availability of echocardiography at the point of care in the context of low-cost handheld devices. This can contribute to improve patient care and to tackle the burden of cardiovascular disease worldwide.

1.8 Publication list

Through the course of this work, both written and oral contributions have been made to international conferences and peer reviewed journals. The following is a list of dissemination conducted in the period.

Contributions included in the thesis

1. **David Pasdeloup**, Sindre H. Olaisen, Andreas Østvik, Sigbjørn Sæbø, Håkon N. Pettersen, Espen Holte, Bjørnar Grenne, Stian B. Stølen, Erik Smistad, Svein Arne Aase, Håvard Dalen, and Lasse Lovstakken. "Real-Time Echocardiography Guidance for Optimized Apical Standard Views", *Ultrasound in medicine and biology*, Volume 49, Issue 1, pages 333-346, January 2023.

2. **David Pasdeloup**, Andreas Østvik, Sindre H. Olaisen, Eirik Skogvoll, Håvard Dalen, and Lasse Lovstakken. "Challenges and Strategies for Automatic Measurements with Deep Learning in Cardiovascular Imaging", submitted for review to *Journal of the American College of Cardiology: Cardiovascular Imaging*.

3. Sigbjørn Sæbø, **David Pasdeloup**, Håkon N. Pettersen, Erik Smistad, Andreas Østvik, Sindre H. Olaisen, Stian B. Stølen, Bjørnar L. Grenne, Espen Holte, Lasse Lovstakken, and Håvard Dalen. "Real-time Guiding by Deep Learning of Experienced Operators to Improve Standardization of Echocardiographic Acquisitions", submitted for review to *Journal of the American College of Cardiology: Cardiovascular Imaging*.

Other contributions in peer reviewed journals

1. Sigbjørn Sæbø, Håkon N. Pettersen, Erik Smistad, **David Pasdeloup**, Stian B. Stølen, Bjørnar L. Grenne, Lasse Løvstakken, Espen Holte, and Håvard Dalen. "Real-time guiding by deep learning during echocardiography to reduce left ventricular foreshortening and measurement variability", *European Heart Journal - Imaging Methods and Practice*, Volume 1, Issue 1, May 2023.

International conference presentations

1. **David Pasdeloup** (presenter), Sigbjørn Sæbø, Sindre H. Olaisen, Håkon N. Pettersen, Andreas Østvik, Espen Holte, Bjørnar Grenne,

Stian B. Stølen, Erik Smistad, Håvard Dalen, and Lasse Løvstakken. "Real-time Echocardiography Guidance for Optimized Apical Standard Views", Poster, *IEEE International Ultrasonics Symposium*, Venice, 2022.

2. Mohammad Mohajery (presenter), Morten S. Wigen, Solveig Fadnes, **David Pasdeloup**, Erik Smistad, Torvald Espeland, Sebastien Salles, and Lasse Løvstakken. "The variation of natural myocardial mechanical wave velocities due to 3D propagation", Poster, *Virtual IEEE International Ultrasonics Symposium*, 2021.

3. Håkon N. Pettersen (presenter), Sigbjørn Sæbø, **David Pasdeloup**, Andreas Østvik, Erik Smistad, Stian B. Stølen, Bjørnar Grenne, Lasse Løvstakken, Håvard Dalen, and Espen Holte. " The impact of real-time feedback by deep learning during echocardiography on test-retest variability of left ventricular systolic function measurements", Oral, *EuroEcho*, Berlin, 2021

4. Sigbjørn Sæbø (presenter), Håkon N. Pettersen, **David Pasdeloup**, Erik Smistad, Stian B. Stølen, Bjørnar Grenne, Lasse Løvstakken, Espen Holte, and Håvard Dalen. "Real-time automatic feedback by deep learning to reduce apical foreshortening in echocardiography", Poster, *EuroEcho*, Berlin, 2021

5. **David Pasdeloup** (presenter), Andreas Østvik, Erik Smistad, Thomas Grønli, Sindre H. Olaisen, Håvard Dalen, and Lasse Løvstakken. "Automatic Apical Standard Views extraction from 3D ultrasound volumes using Deep Learning", Poster, *Virtual IEEE International Ultrasonics Symposium*, 2020.

6. **David Pasdeloup** (presenter), Andreas Østvik, Erik Smistad, and Lasse Løvstakken. "Retraining strategies for improved generalization of deep neural networks", Poster, *IEEE International Ultrasonics Symposium*, Glasgow, 2019.

References

[1] "World Heart Report 2023 Confronting the World's Number One Killer," tech. rep., 2023.

[2] R. M. Lang, L. P. Badano, M. A. Victor, J. Afilalo, A. Armstrong, L. Ernande, F. A. Flachskampf, E. Foster, S. A. Goldstein, T. Kuznetsova, P. Lancellotti, D. Muraru, M. H. Picard, E. R. Retzschel, L. Rudski, K. T. Spencer, W. Tsang, and J. U. Voigt, "Recommendations for cardiac chamber quantification by echocardiography in adults: An update from the American Society of Echocardiography and the European Association of Cardiovascular Imaging," *Journal of the American Society of Echocardiography*, vol. 28, no. 1, pp. 1–39.e14, 2015.

[3] W. S. McCulloch and W. Pitts, "A logical calculus of the ideas immanent in nervous activity," *The Bulletin of Mathematical Biophysics*, vol. 5, pp. 115–133, dec 1943.

[4] D. Silver, A. Huang, C. J. Maddison, A. Guez, L. Sifre, G. Van Den Driessche, J. Schrittwieser, I. Antonoglou, V. Panneershelvam, M. Lanctot, S. Dieleman, D. Grewe, J. Nham, N. Kalchbrenner, I. Sutskever, T. Lillicrap, M. Leach, K. Kavukcuoglu, T. Graepel, and D. Hassabis, "Mastering the game of Go with deep neural networks and tree search," *Nature*, vol. 529, no. 7587, pp. 484–489, 2016.

[5] B. Marr, "The Problem With Biased AIs (and How To Make AI Better)," 2022.

[6] E. Helmore, "Tesla behind eight-vehicle crash was in 'full self-driving' mode, says driver," 2022.

[7] M. Roberts, D. Driggs, M. Thorpe, J. Gilbey, M. Yeung, S. Ursprung, A. I. Aviles-Rivero, C. Etmann, C. McCague, L. Beer, J. R. Weir-McCall, Z. Teng, E. Gkrania-Klotsas, A. Ruggiero, A. Korhonen, E. Jefferson, E. Ako, G. Langs, G. Gozaliasl, G. Yang, H. Prosch, J. Preller, J. Stanczuk, J. Tang, J. Hofmanninger, J. Babar, L. E. Sánchez, M. Thillai, P. M. Gonzalez, P. Teare, X. Zhu, M. Patel, C. Cafolla, H. Azadbakht, J. Jacob, J. Lowe, K. Zhang, K. Bradley, M. Wassin, M. Holzer, K. Ji, M. D. Ortet, T. Ai, N. Walton, P. Lio, S. Stranks, T. Shadbahr, W. Lin, Y. Zha, Z. Niu, J. H. Rudd, E. Sala, and C. B. Schönlieb, "Common pitfalls and recommendations for using machine learning to detect and prognosticate for COVID-19 using chest radiographs and CT scans," *Nature Machine Intelligence*, vol. 3, no. 3, pp. 199–217, 2021.

[8] G. Varoquaux and V. Cheplygina, "Machine learning for medical imaging: methodological failures and recommendations for the future," *npj Digital Medicine*, vol. 5, no. 1, 2022.

[9] F. Dietrichson, E. Smistad, A. Østvik, and L. Lovstakken, "Ultrasound Speckle Reduction Using Generative Adversial Networks," *IEEE International Ultrasonics Symposium, IUS*, vol. 2018-Octob, pp. 1–4, 2018.

[10] J. Zhang, S. Gajjala, P. Agrawal, G. H. Tison, L. A. Hallock, L. Beussink-Nelson, M. H. Lassen, E. Fan, M. A. Aras, C. Jordan, K. E. Fleischmann, M. Melisko, A. Qasim, S. J. Shah, R. Bajcsy, and R. C. Deo, "Fully Automated Echocardiogram Interpretation in Clinical Practice," *Circulation*, vol. 138, pp. 1623–1635, oct 2018.

[11] A. Østvik, E. Smistad, S. A. Aase, B. O. Haugen, and L. Lovstakken, "Real-Time Standard View Classification in Transthoracic Echocardiography Using Convolutional Neural Networks," *Ultrasound in Medicine and Biology*, vol. 45, pp. 374–384, feb 2019.

[12] E. Smistad, A. Østvik, I. Mjåland Salte, S. Leclerc, O. Bernard, and L. Lovstakken, "Fully Automatic Real-Time Ejection Fraction and MAPSE Measurements in 2D Echocardiography Using Deep Neural Networks," in *IEEE International Ultrasonics Symposium, IUS*, vol. 2018-Octob, IEEE Computer Society, dec 2018.

[13] S. Liu, Y. Wang, X. Yang, B. Lei, L. Liu, S. X. Li, D. Ni, and T. Wang, "Deep Learning in Medical Ultrasound Analysis: A Review," *Engineering*, vol. 5, pp. 261–275, apr 2019.

[14] A. Ostvik, I. M. Salte, E. Smistad, T. M. Nguyen, D. Melichova, H. Brunvand, K. Haugaa, T. Edvardsen, B. Grenne, and L. Lovstakken, "Myocardial Function Imaging in Echocardiography Using Deep Learning," *IEEE Transactions on Medical Imaging*, vol. 40, no. 5, pp. 1340–1351, 2021.

[15] A. Sánchez-Puente, P. I. Dorado-Díaz, J. Sampedro-Gómez, J. Bermejo, P. Martinez-Legazpi, F. Fernández-Avilés, J. Sánchez-González, C. Pérez del Villar, V. Vicente-Palacios, and P. L. Sanchez, "Machine Learning to Optimize the Echocardiographic Follow-Up of Aortic Stenosis," *JACC: Cardiovascular Imaging*, vol. 16, pp. 733–744, jun 2023.

[16] A. Hovland, U. H. Staub, H. Bjørnstad, J. Prytz, J. Sexton, A. Støylen, and H. Vik-Mo, "Gated SPECT Offers Improved Interobserver Agreement Compared With Echocardiography," *Clinical Nuclear Medicine*, vol. 35, pp. 927–930, dec 2010.

[17] R. Hoffmann, G. Barletta, S. von Bardeleben, J. L. Vanoverschelde, J. Kasprzak, C. Greis, and H. Becher, "Analysis of Left Ventricular Volumes and Function: A Multicenter Comparison of Cardiac Magnetic Resonance Imaging, Cine Ventriculography, and Unenhanced and Contrast-Enhanced Two-Dimensional and Three-Dimensional Echocardiography," *Journal of the American Society of Echocardiography*, vol. 27, pp. 292–301, mar 2014.

[18] C. Knackstedt, S. C. Bekkers, G. Schummers, M. Schreckenberg, D. Muraru, L. P. Badano, A. Franke, C. Bavishi, A. M. S. Omar, and P. P. Sengupta, "Fully Automated Versus Standard Tracking of Left Ventricular Ejection Fraction and Longitudinal Strain the FAST-EFs Multicenter Study," *Journal of the American College of Cardiology*, vol. 66, no. 13, pp. 1456–1466, 2015.

[19] C. Morbach, G. Gelbrich, M. Breunig, T. Tiffe, M. Wagner, P. U. Heuschmann, and S. Störk, "Impact of acquisition and interpretation on total inter-observer variability in echocardiography: results from the quality assurance program of the STAAB cohort study," *International Journal of Cardiovascular Imaging*, vol. 34, no. 7, pp. 1057–1065, 2018.

[20] J. E. Otterstad, G. Froeland, M. St John Sutton, and I. Holme, "Accuracy and reproducibility of biplane two-dimensional echocardiographic measurements of left ventricular dimensions and function," *European Heart Journal*, vol. 18, no. 3, pp. 507–513, 1997.

[21] P. Thavendiranathan, Z. B. Popović, S. D. Flamm, A. Dahiya, R. A. Grimm, and T. H. Marwick, "Improved interobserver variability and accuracy of echocardiographic visual left ventricular ejection fraction assessment through a self-directed learning program using cardiac magnetic resonance images," *Journal of the American Society of Echocardiography*, vol. 26, no. 11, pp. 1267–1273, 2013.

[22] I. M. Salte, A. Østvik, S. H. Olaisen, S. Karlsen, T. Dahlslett, E. Smistad, T. K. Eriksen-Volnes, H. Brunvand, K. H. Haugaa, T. Edvardsen, H. Dalen, L. Lovstakken, and B. Grenne, "Deep Learning for Improved Precision and Reproducibility of Left Ventricular Strain in Echocardiography: A Test-Retest Study," *Journal of the American Society of Echocardiography*, pp. 1–12, 2023.

[23] A. Ghorbani, D. Ouyang, A. Abid, B. He, J. H. Chen, R. A. Harrington, D. H. Liang, E. A. Ashley, and J. Y. Zou, "Deep learning interpretation of echocardiograms," *npj Digital Medicine*, vol. 3, no. 1, pp. 1–10, 2020.

[24] D. Ouyang, B. He, A. Ghorbani, N. Yuan, J. Ebinger, C. P. Langlotz, P. A. Heidenreich, R. A. Harrington, D. H. Liang, E. A. Ashley, and J. Y. Zou, "Video-based AI for beat-to-beat assessment of cardiac function," *Nature*, vol. 580, no. 7802, pp. 252–256, 2020.

[25] B. He, A. C. Kwan, J. H. Cho, N. Yuan, C. Pollick, T. Shiota, J. Ebinger, N. A. Bello, J. Wei, K. Josan, G. Duffy, M. Jujjavarapu, R. Siegel, S. Cheng, J. Y. Zou, and D. Ouyang, "Blinded, randomized trial of sonographer versus AI cardiac function assessment," *Nature*, vol. 616, no. 7957, pp. 520–524, 2023.

[26] A. J. DeGrave, J. D. Janizek, and S. I. Lee, "AI for radiographic COVID-19 detection selects shortcuts over signal," *Nature Machine Intelligence*, vol. 3, no. 7, pp. 610–619, 2021.

[27] J. Deng, W. Dong, R. Socher, L.-J. Li, Kai Li, and Li Fei-Fei, "ImageNet: A large-scale hierarchical image database," *2009 IEEE Conference on Computer Vision and Pattern Recognition*, pp. 248–255, 2010.

[28] A. Z. Arystan, G. C. Kane, S. V. Pislaru, and F. Lopez-jimenez, "Automated Echocardiographic Detection of Heart Failure With Preserved Ejection Fraction Using Arti fi cial Intelligence," vol. C, 2023.

[29] S. R. Snare, H. Torp, F. Orderud, and B. O. Haugen, "Real-time scan assistant for echocardiography," *IEEE Transactions on Ultrasonics, Ferroelectrics, and Frequency Control*, vol. 59, no. 3, pp. 583–589, 2012.

[30] G. Toporek, R. Naidu, H. Xie, A. Simicich, T. Gades, and B. Raju, "User guidance for point-of-care echocardiography using multi-task deep neural network," *Med Image Comput Comput Assist Interv*, vol. 1, pp. 309–317, 2019.

[31] E. Smistad, A. Østvik, I. M. Salte, D. Melichova, T. M. Nguyen, K. Haugaa, H. Brunvand, T. Edvardsen, S. Leclerc, O. Bernard, B. Grenne, and L. Løvstakken, "Real-Time Automatic Ejection Fraction and Foreshortening Detection Using Deep Learning," *IEEE Transactions on Ultrasonics, Ferroelectrics, and Frequency Control*, vol. 67, no. 12, pp. 2595–2604, 2020.

[32] W. J. C. Kim, A. Beqiri, A. J. Lewandowski, A. Mumith, R. Sarwar, A. King, P. Leeson, and P. Lamata, "Automated Detection of Apical Foreshortening in Echocardiography Using Statistical Shape Modelling," *Ultrasound in Medicine and Biology*, vol. 000, no. April, 2023.

[33] S. Leclerc, E. Smistad, J. Pedrosa, A. Ostvik, F. Cervenansky, F. Espinosa, T. Espeland, E. A. R. Berg, P. M. Jodoin, T. Grenier, C. Lartizien, J. Dhooge, L. Lovstakken, and O. Bernard, "Deep Learning for Segmentation Using an Open Large-Scale Dataset in 2D Echocardiography," *IEEE transactions on medical imaging*, vol. 38, no. 9, pp. 2198–2210, 2019.

[34] M. Schneider, P. Bartko, W. Geller, V. Dannenberg, A. König, C. Binder, G. Goliasch, C. Hengstenberg, and T. Binder, "A machine learning algorithm supports ultrasound-naïve novices in the acquisition of diagnostic echocardiography loops and provides accurate estimation of LVEF," *International Journal of Cardiovascular Imaging*, vol. 37, no. 2, pp. 577–586, 2021.

[35] A. Narang, R. Bae, H. Hong, Y. Thomas, S. Surette, C. Cadieu, A. Chaudhry, R. P. Martin, P. M. McCarthy, D. S. Rubenson, S. Goldstein, S. H. Little, R. M. Lang, N. J. Weissman, and J. D. Thomas, "Utility of a Deep-Learning Algorithm to Guide Novices to Acquire Echocardiograms for Limited Diagnostic Use," *JAMA Cardiology*, vol. 6, no. 6, pp. 624–632, 2021.

[36] D. Peck, J. Rwebembera, D. Nakagaayi, N. W. Minja, N. J. Ollberding, J. Pulle, J. Klein, D. Adams, R. Martin, K. Koepsell, A. Sanyahumbi, A. Beaton, E. Okello, and C. Sable, "The Use of Artificial Intelligence Guidance for Rheumatic Heart Disease Screening by Novices," *Journal of the American Society of Echocardiography*, vol. 36, no. 7, pp. 724–732, 2023.

Background

2

This chapter provides essential background information required to understand this thesis, establishing the necessary basis for the subsequent investigations. It begins with an overview of echocardiography to contextualize the work contained in the thesis and proceeds with an introduction to deep learning (DL) basics to understand the methodology used in the work. Additional DL concepts specifically relevant for echocardiographic image analysis are also presented. Finally, some statistical background is provided to interpret the results in the context of DL applied to echocardiography.

The concepts described in this chapter are restricted to that necessary for the understanding of this thesis, with some intentional omissions. A more extensive background on echocardiography can for instance be found in *Essentials Echocardiography* [1]. Regarding the DL concepts, one can refer to *Deep Learning* [2] which proposes a theoretical approach and to *Deep Learning with Python* [3] for a more practical approach. Finally, some relevant statistical concepts can be found in *Comparing Clinical Measurement Methods - A Practical Guide* [4].

2.1 Echocardiography

As a widely available, portable, non-ionizing and real-time imaging technique, echocardiography has become the cornerstone of cardiac imaging modalities, providing valuable information about the heart's chambers, valves, blood flow patterns, and overall cardiac function. With different devices, it is used throughout the whole cardiac care path, in emergency services, for routine, follow-up, and in the operating room.

The most practiced form of echocardiography, and subject of this thesis, is trans-thoracic echocardiography (TTE) in which an ultrasound transducer is placed on the patient's chest, allowing non-invasive imaging of the patient. Other invasive forms of echocardiography, mostly used in the operating room, are outside the scope of this thesis.

Echocardiography uses a transducer to emit sound waves that penetrate

the body. This wave emission is called transmission. These sound waves are reflected back by the medium (tissue, blood, bones, etc.) and received by the same transducer (reception). The received signals are then transformed into pixel data using beamforming algorithms. Based on this principle of transmission-reception, echocardiography has evolved throughout the years to become a versatile imaging technique, allowing real-time visualization of tissue and blood, in terms of both size and velocity.

According to the current guidelines [5], a typical echocardiography examination within the echocardiography laboratory involves a large number of quantitative measurements relying on many ultrasound recordings acquired from different transducer positions. Such extensive examinations allow precise quantification of the heart function but are labor intensive. Furthermore, the availability of skilled operators capable of performing echocardiography is limited, with the majority of them working within hospital settings.

The most common modality is brightness mode (B-mode), providing a grey-scale visualization of the heart structures. Available on most of ultrasound scanners, from high-end scanners (2.1a) to portable scanners (2.1b), two-dimensional (2D) B-mode is the central modality in ultrasound imaging. It is used by the operator to first find the correct cut-plane of the heart and to further measure the size of the heart structures. Three-dimensional (3D) B-mode imaging is currently reserved for high-end scanners. Echocardiography has multiple other modalities, many based on the Doppler effect to measure blood velocity with continuous wave Doppler (CW), pulsed wave Doppler (PW) or color flow imaging (CFI), and to measure tissue velocity with tissue Doppler imaging (TDI). In this thesis, B-mode is the modality of interest.

2.1.1 TTE B-Mode imaging

Due to the limited space between the ribs to image the heart, TTE B-mode imaging uses phased-array transducers which fits in between the ribs. To be able to image a wide region covering the width and depth of the heart with an array of small size, phased-arrays use electronic beam steering to send ultrasound waves in different directions. In practice, this is done by applying transmission and receive delays to the different elements of the transducer. Once the received signals are processed, the resulting pixel data is in either polar coordinates in the case of 2D imaging, or spherical coordinates in the case of 3D imaging. Further conversion to cartesian coordinates system recovers the physiological appearance and dimensions

(a) High-end scanner **(b)** Portable scanner

Figure 2.1: Left: GE Healthcare Vivid E95 cardiac ultrasound scanner, Right: GE VScan Air SL (GE Vingmed Ultrasound, Horten, Norway)

of the structures that were imaged.

For low-end echocardiography devices, a limitation is the lower available computational power to process the received signals in real-time, and the limited battery energy supply and thermal considerations required for portable devices. Currently, this results in a limited image quality and frame rate of the images acquired with these devices. Future technological developments increasing the available computational power for the same cost can be expected to mitigate this issue.

High-end scanners have more computational power available, allowing 3D imaging capabilities. For these scanners, the limitation in frame rate is a physical one, related to the maximum speed of (ultra)sound in human tissue. Similarly to using short aperture time for taking a clearly defined picture of a moving object with a camera, the time to take a single frame of the moving heart should be short relatively to the heart movement. However, and unlike light waves that propagates at the speed of light, the speed of sound for ultrasonic waves in the body is approximately c=1540ms^{-1}, which fundamentally limits the temporal resolution of ultrasound imaging. The time t_f to acquire an US frame is proportional to:

$$t_f \propto n_a \cdot n_e \cdot \frac{2z}{c} \tag{2.1}$$

where n_a (resp. n_e) are the number of beams in the azimuth (resp. elevation) directions and z the imaging depth. For 2D imaging, $n_e = 1$, and the limited speed of sound still allows frame rates >25 FPS, which is considered real-time for B-mode. However, the number of beams to be transmitted and received is squared (at constant spatial sampling) in 3D imaging, which significantly increase the time to acquire a frame. To get an impression, using 2.1, the time required to acquire a 14cm deep volume with a 60° field of view width in both the azimuth and elevation directions and a lateral sampling of 1° would for example be 0.65sec (corresponding to a frame rate of 1.53FPS), which is too slow considering that a heart beat typically lasts between 0.5 and 1.5 seconds. This leads to a trade-off between image quality and image frame rate, as a higher fundamental image resolution, achieved by increasing the imaging pulse frequency, bandwidth, and aperture, requires more image lines to result in a properly sampled image. Another solution is the use ECG gating, which consists in acquiring multiple sub-volumes through several heart beats, which are later synchronized and stitched using the ECG signal. However, ECG gating assumes constant heart rate of the patient (no arrhythmia) and that the transducer is held still. Deviations from these assumptions causes stitching artifacts. This limitation of 3D imaging makes that 2D imaging is still the preferred choice for some measurements such as strain.

2.1.2 Heart anatomy

The heart is the central component of the cardiovascular system, ensuring the circulation of blood and vital nutrients throughout the body, and the removal of waste products such as carbon dioxide. The heart has two sides that work synchronously: the right side and the left side. Each side is composed of an atria, a ventricle and two valves that ensure unidirectional flow as illustrated in Fig. 2.2. The apex region of the heart refers to the pointed tip or bottommost portion of the organ, while the base region corresponds to the atrium side, which is the top and broader part of the heart.

The function of the right side is to pump the blood to the lungs. The right atria receives deoxygenated blood from the body and transfer it to the right ventricle which eject it towards the lungs where the carbon dioxide gets filtered out and replaced by oxygen from inhaled air. The function of the left side is to pump the blood to the whole body. The left atrium then receives the newly oxygenated blood from the lungs, transfer it to the left ventricle which further pumps it out to the whole body. Of the two sides,

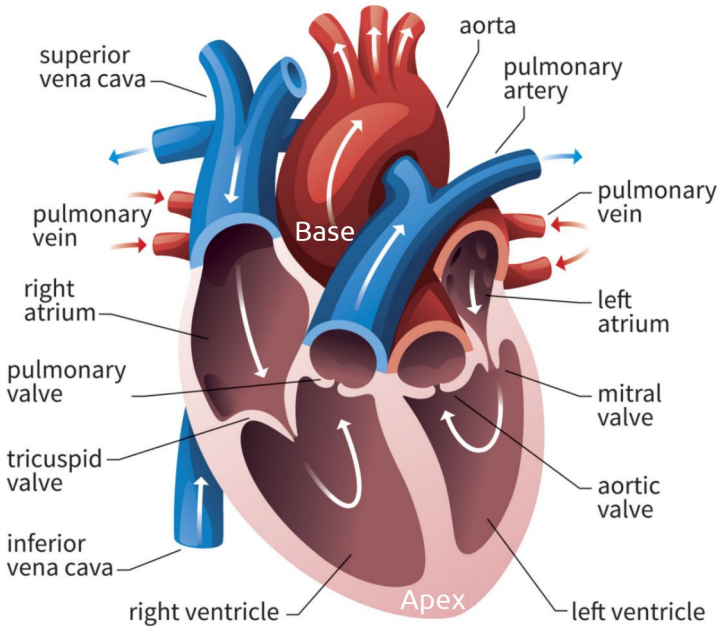


Figure 2.2: Cross-section of the heart showing chambers, valves and flow directions. Adapted from cardofmich.com

the left side is the most powerful and the one which receives most attention in clinical examinations.

The heart cycle is composed of two main phases: the diastole and the systole. During the diastole, the mitral and tricuspid valves are open, whereas the aortic and pulmonary valves are closed, and blood flows into the ventricles from the atria. The end of the diastole (ED) is marked by the mitral valve closure. During the systole, the mitral and tricuspid valves are closed and muscular contraction ejects the blood through the aortic and pulmonary valves that are open. The end of the systole corresponds to the closure of the aortic valve. The Wiggers diagram in Fig. 2.3, gives more details on the heart phases, and relations with pressures and volumes.

Various physiological factors such as age, gender and physical activity impact the heart function. Pathological factors such as ischemic heart disease or coronary artery disease can also impair the heart's ability to pump blood effectively. Additionally, congenital heart defects and valvular abnormalities can disrupt the normal flow of blood within the heart.

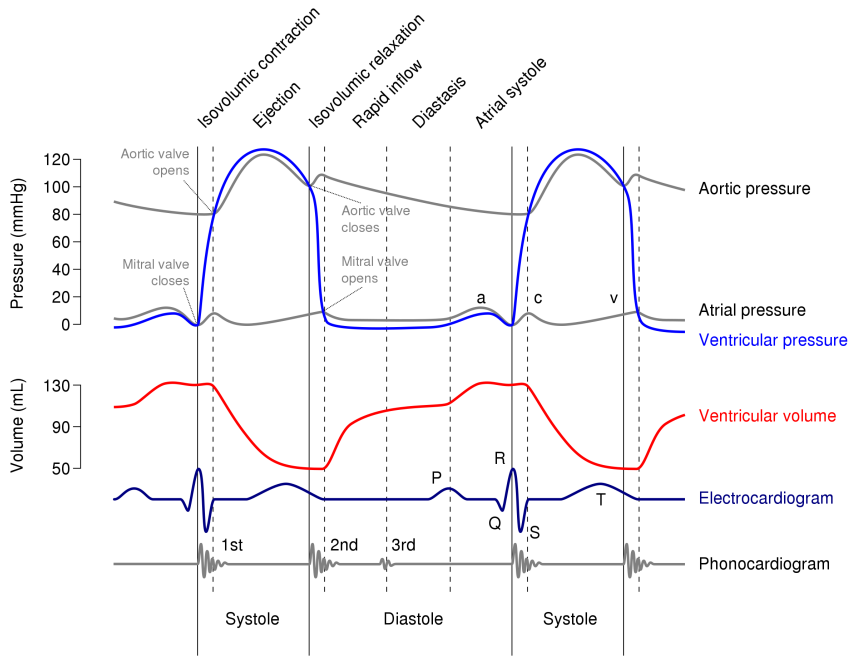


Figure 2.3: Wiggers diagram of the healthy heart, showing relation between left ventricular volume, pressure and valve events. Source: Wikimedia Commons

2.1.3 Apical Standard Views

TTE can be performed by placing the ultrasound transducer in different positions, called acoustic windows, on the chest of the patient. The four main windows to image the heart are the apical, parasternal, subcostal and suprasternal. These are illustrated in Fig. 2.4. In 2D echocardiography, different rotation, tilt and translation maneuvers from one of these acoustic windows allow for different cut-planes of the heart. The relevant cut-planes useful for establishing further measurements are defined as standard-views, which follow specific rules on which heart structures should or should not be visible in the images [5].

In this thesis, the focus is placed on the standard views acquired from the apical window, called apical standard views, from which many quantitative measurements such as left ventricular ejection fraction (LVEF) are obtained. The apical standard views are among the most difficult to acquire due to small intercostal space and respiratory movements. The three main apical standard views are:

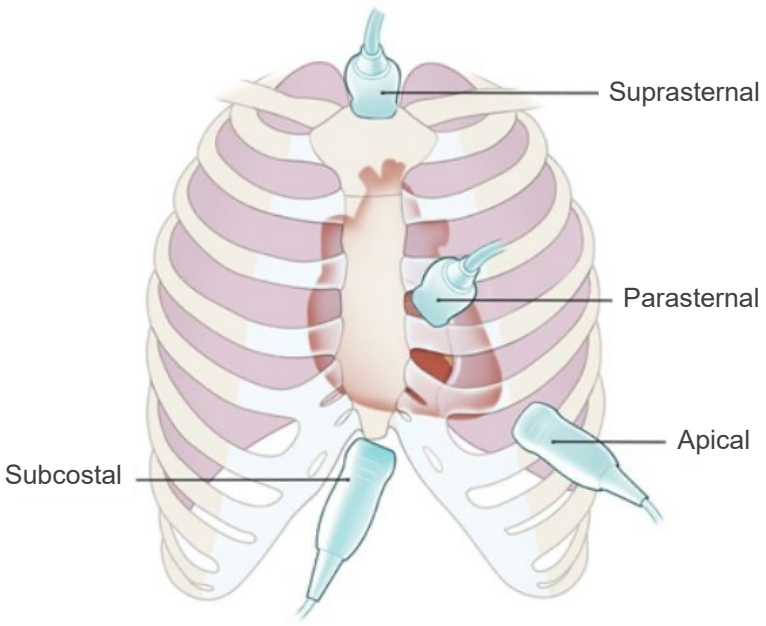
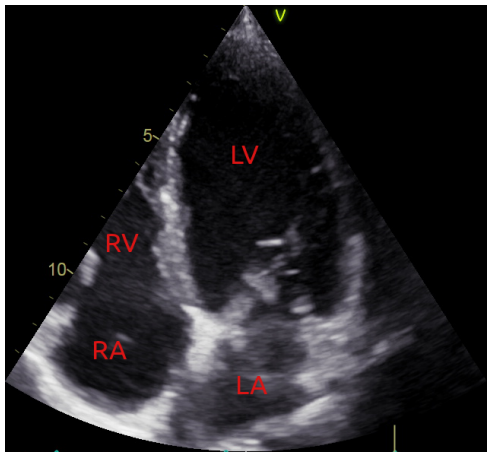


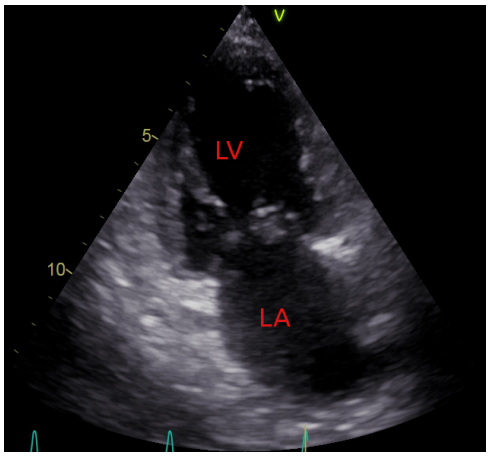
Figure 2.4: The four main acoustic windows for echocardiography. Source: SonoSim, Inc

- the apical four chamber view (A4C, Fig. 2.5a), in which all four chamber of the heart should be visible, together with the mitral and tricuspid valves.
- the apical two chamber view (A2C, Fig. 2.5b), in which only the left atrium and ventricle should be visible, together with the mitral valve
- the apical long-axis view (ALAX, Fig. 2.5c), in which only the left atrium and ventricle should be visible, together with the mitral valve and the aortic valve leaflets.

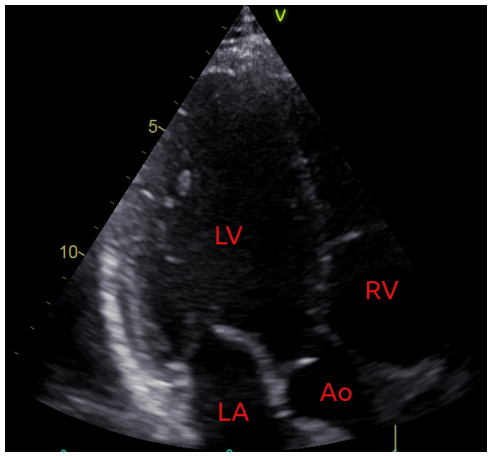
In the apical views, the transducer is usually placed on the 5th intercostal space, with the heart apex located near the transducer and the base further away from it. The left ventricle should ideally be placed at the center of the field of view (FOV), and the three apical standard views could theoretically be obtained with a simple rotation of the transducer [1]. Additionally, the transducer position should also be adjusted to maximize the left ventricle area [5]. However, these definitions of the apical standard views remain vague, and leave room for preferences, which may induce



(a) Apical Four Chamber



(b) Apical Two Chamber



(c) Apical Long-Axis

Figure 2.5: Examples of the three apical standard views. LV: Left ventricle, LA: Left atrium, RV: Right Ventricle, RA: Right atrium, Ao: Aorta

subjectiveness and interpretation errors, both at the individual and at the echocardiography laboratory level.

2.1.4 Left Ventricular Ejection fraction

Left ventricular ejection fraction (LVEF) is an important clinical measurement used to assess the heart function. It quantifies the percentage of blood ejected from the left ventricle at each contraction, providing information on cardiac pumping efficiency.

LVEF relates to the volume of blood ejected during the diastole (stroke volume) compared to the end diastolic volume and can be expressed as:

$$LVEF = \frac{\text{Stroke volume}}{LVV_{ED}} \cdot 100 = \frac{LVV_{ED} - LVV_{ES}}{LVV_{ED}} \cdot 100 \tag{2.2}$$

where LVV_{ED} (resp. LVV_{ES}) is the left ventricular volume at end diastole (resp. end systole)

Normal LVEF values typically range between 50% and 70%, and values below this range are generally indicative of reduced heart function [6, 7]. LVEF serves as a critical prognostic indicator in several cardiovascular diseases, including heart failure, coronary artery disease, and cardiomyopathies. LVEF can also be used to assess disease progression, response to interventions or response to cardiotoxic treatments. The current recommended method for calculating LVEF in 2D echocardiography is the 2D biplane method of disks. LVEF calculation from 3D ultrasound imaging is also possible, but not available on low-end and portable systems.

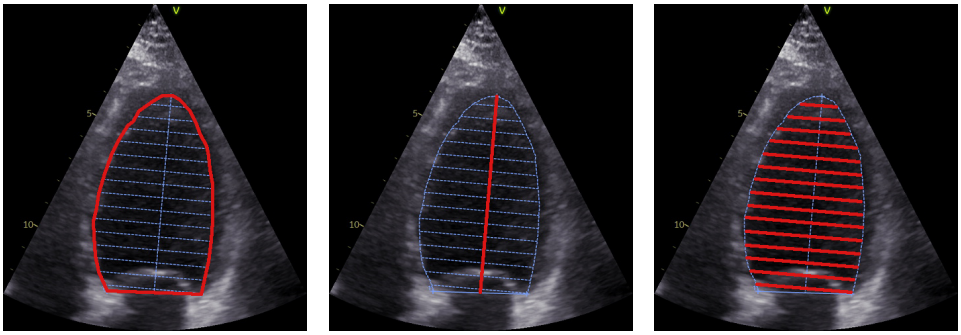
Single-plane method of disks

The single-plane method of disks (MOD), also called Simpson's rule, is used to estimate the volume of heart chambers from a 2D echocardiographic frame. Its algorithm is described below:

1. Manual delineation of the endocardial border (Fig. 2.6a). The two mitral annuluses are joined by a straight line.
2. Automatic detection of the ventricular long axis of length l_{LA}, with ends at the center of the mitral valve and at the apex (Fig. 2.6b).
3. Automatic tracing of N, typically 30, diameters of length d_i, equally spaced along the long axis and orthogonal to it (Fig. 2.6c)
4. Volume calculation following equation (2.3), assuming a circular shape of the left ventricle.

$$LVV_{single} = \frac{l_{LA}}{N} \sum_{i=1}^{N} \pi \cdot d_i^2 \tag{2.3}$$

This single-plane method has several drawbacks. One of them is the assumption that the disks are circular. In practice, this is not necessarily the case for the ventricle, and may lead to over- or under-estimation of the calculated volume. Another drawback is that the possible presence of apex foreshortening (the 2D imaging plane not cutting the heart at the apex) [8], which may cause an underestimation of the ventricle length l_{LA}



(a) Manual delineation of the endocardium

(b) Automatic tracing of the ventricle long-axis l_{LA}

(c) Automatic tracing of the diameters d_i

Figure 2.6: Method of disks exemplified on an apical four chamber view

and introduce error in the measured volumes. Finally, as the volume is proportional to the square of the diameters, small errors on the tracing of the endocardium may result in large volume under- or overestimations.

Biplane method of disks

The biplane MOD intend to mitigate the limitations of the single-plane MOD and is the current recommended method to calculate the left ventricular volumes and further establish LVEF from 2D echocardiography [5]. The biplane MOD uses two orthogonal imaging planes with long axis lengths $l_{LA,1}$ and $l_{LA,2}$ and diameters $d_{i,1}$ and $d_{i,2}$. Biplane volume calculation improves from single-plane by assuming an ellipsoidal instead of circular shape of the ventricle, and corrects for foreshortening by taking the maximum long axis length of the two views:

$$LVV_{biplane} = \frac{\max(l_{LA,1}, l_{LA,2})}{N} \sum_{i=1}^{N} \pi \cdot d_{i,1} \cdot d_{i,2} \tag{2.4}$$

Although this approach is the recommended for calculating left ventricular volumes, it still suffers from major limitations:

- The correction for foreshortening remains an approximation, and is inefficient if both imaging planes are foreshortened
- It is assumed that the two imaging planes are orthogonal. This is however not guaranteed in 2D imaging setups that use the A4C view for one plane and the A2C or ALAX view for the other plane.

- The biplane method requires two delineations of the endocardium at end diastole (ED), and two delineations at end systole (ES). This procedure is time consuming and has multiple error sources, in the choice of ES and ED frames and in the endocardium delineations. These volume errors further add up to each other in the LVEF calculation (2.2). To reduce these errors and account for interbeat variability, guidelines [5] recommend to perform LVEF calculation on three heart cycles, which is rarely done in practice.

These limitations translates to high intra- and inter-observer variability [9–12] and hamper the possibility to detect slight LVEF changes in repeated examinations.

2.2 Deep learning for echocardiographic image analysis

In this thesis, the use of DL within the echocardiography workflow is investigated. The core of DL is deep neural networks, which are algorithmic models that can learn complex relationships between input and output data through a training procedure. In the case of medical imaging, this training procedure can approximate the human learning process that associates input image data to an output, which can be for example a specific skill, a measurement or a diagnosis. The training procedure automatically finds the relevant features that contribute to predicting the correct output from the images, without the need for human engineered features. This can be regarded as a strength as DL may be able to discover and learn more complex features and relationships than a human could handle. On the other hand, deep neural networks are not as robust as humans with regard to confounders [13] and may not be easily explainable, which causes trust issues in the medical community.

The four elements required to train a deep neural network [2] are: 1) training data, 2) a loss function, 3) an optimization procedure and 4) a neural network architecture. Out of the four types of learning (supervised learning, semi-supervised learning, unsupervised learning and reinforcement learning), only supervised learning is used within the scope of this thesis. Supervised learning uses labeled training pairs, which means that every input $\boldsymbol{x}$ is associated with an output $\boldsymbol{y}$, the label. In the training pairs, the output label is often obtained by human interpretation and reasoning from the input, similarly to how clinicians would think or

do in their daily clinical practice. The training data hence embed the human knowledge in the form of many individual pairs. The objective of supervised learning is to establish a model capable of replicating the human interpretation and reasoning to be able to automatically predict output on future input data, as clinicians would have done it, in what is called *inference*.

The DL model can be regarded as a function $\boldsymbol{f}$ that learns the parameters $\boldsymbol{\theta}$ that ensure the best possible mapping between the input data $\boldsymbol{x}$ and the associated output label $\boldsymbol{y}$. For a given input $\boldsymbol{x}$, the output prediction of the DL model is denoted $\hat{\boldsymbol{y}}$.

$$\hat{\boldsymbol{y}} = \boldsymbol{f}(\boldsymbol{x}|\boldsymbol{\theta}) \tag{2.5}$$

In the field of DL, the labels are often referred as ground truth. In this thesis, the term *reference* is preferred since human labeling is subject to high intra- and inter-observer variability in echocardiography [9–12]. This variability is challenging for both the training and evaluation of DL models.

2.2.1 Neural network architectures

There exists different families of neural network architectures, and a large amount of variants within each family depending on the task to solve and on other requirements such as network size or runtime. Modern families of neural networks used for image analysis include Convolutional Neural Networks (CNNs), Transformers Neural Networks and Graph Neural Network. For all these families, the networks are organized in layers, with input layers, hidden layers and output layers.

In this thesis, only CNNs were used. Understanding the content of their hidden layers is not necessary for interpreting the results and will not be pursued here. Regarding a neural network as a function $\boldsymbol{f}$ that maps image or video data to an output probability vector is sufficient:

$$\boldsymbol{f} : \mathbb{R}^{x \times y \times z \times t} \rightarrow \mathbb{R}^{n} \tag{2.6}$$

where x and y are the number of pixels in the lateral directions, z in the depth direction, and t the number of frames. $y = 1$ for 2D imaging, and $t = 1$ for single frames without temporal information. The output vector represents a confidence score for each of the n classes, and all its elements should sum to 1. This is achieved by using a softmax function as the output layer. For each output element corresponding to class i, its confidence is

calculated using the softmax function defined as:

$$\sigma(x)_i = \frac{\exp(x_i)}{\sum_j^n \exp(e^{x_j})} \tag{2.7}$$

where x is the output of the second last layer. For pure classification tasks the predicted class p_c is the class having the maximum confidence:

$$p_c = \text{argmax}(\sigma(x)) \tag{2.8}$$

2.2.2 Loss function and optimization procedure

Training a neural network $\boldsymbol{f}$ consists in searching for the parameters $\boldsymbol{\theta}$ that optimize the mapping between the input $\boldsymbol{x}$ and the reference $\boldsymbol{y}$. These parameters $\boldsymbol{\theta}$ are refined through an optimization procedure that relies on a cost-function evaluating the goodness of the mapping. This cost-function is a function of the reference $\boldsymbol{y}$ and the output $\hat{\boldsymbol{y}}$ and is referred to as the loss $L(\boldsymbol{y}, \hat{\boldsymbol{y}})$ in the field of DL. This optimization procedure has three steps, which are repeated iteratively:

1. Loss calculation for iteration i
2. Backpropagation
3. Parameters update for next iteration $i + 1$

Each step is described more in details below. In practice iterations of this optimization procedure happens every N training pairs. N is called *batch size* and commonly ranges from 4 to 256, depending on the application and the size of each training pair in computer memory.

Loss function

The loss function is a distance measure between the reference $\boldsymbol{y}$ and the output $\hat{\boldsymbol{y}}$. There are many loss functions that depend on the purpose of the neural network. For classification problems, a common loss function is the cross-entropy. Specific loss functions used in this thesis are described later in Chapter 2.

Backpropagation

Backpropagation consists in computing the gradient of the loss L with regard to the parameters θ.

$$\frac{\mathrm{d}L}{\mathrm{d}\theta} = \frac{L_i - L_{i-1}}{\theta_i - \theta_{i-1}} \tag{2.9}$$

A key enabler for backpropagation is that the DL libraries such as Tensorflow [14] or PyTorch [15] do not only store the values of the model parameters θ and loss L, but also their derivatives $\mathrm{d}\theta$ and $\mathrm{d}L$, which facilitates backpropagation. This also implies that the loss function must be differentiable.

Parameters updates

Once the gradients are calculated, the parameters θ gets updated. The most basic way of updating the parameters is to directly use the gradient, in an optimization procedure called Gradient Descent (GD). The parameters θ for iteration $i+1$ can be calculated following:

$$\theta_{i+1} = \theta_i - \alpha \cdot \frac{L_i - L_{i-1}}{\theta_i - \theta_{i-1}} \tag{2.10}$$

where α is a positive term called learning rate, that controls the magnitude of the update. This formula implies that for each iteration, the parameters θ are updated in the same direction as the loss L, seeking for even better parameters when the loss improved and trying to reverse the degradation when the loss function worsened.

The convergence of Gradient Descent is however limited in practice as it can reach and stay at a local minima and is very dependent on the original choice of the learning rate. In this thesis, the Adam optimizer [16] was used in the contribution presented in Chapter 4. Adam is one of the most common used in the field of machine learning, with around 150'000 citations at the time of writing. It improves the GD method with an adaptive learning rate and a momentum to account for the gradient from *all* earlier iterations instead of the last iteration only. For the contribution presented in Chapter 3, the Adadelta optimizer [17] was found to converge faster than Adam and was therefore used. Adadelta addresses the issue of decaying learning rate during training and ensures that the network continues learning even after many iterations. This is done in practice by calculating the momentum and updating the learning rate based on a

moving window of gradient updates instead of all gradient updates from the start of the training. Aditionally, Adadelta does not need an initial learning rate to be specified at start-up, which avoids to start with a suboptimal learning rate.

Overfitting

According to the universal approximation theorem, there always exists a NN architecture that can be optimized to learn an optimal mapping between all the training pairs [18]. Although this is possible, it is not the aim of DL. DL instead aims to train NNs so that they can be predict the best possible output on *unseen* inputs. Performance on the training pairs alone does not matter from a clinical perspective.

There is therefore a need to find a compromise as to how much the NN should learn from the training data while keeping a good performance for unseen data. Over-optimization occurs when the neural network is optimized too well on training data and thus cannot reliably predict output on new data. In the field of DL, this over-optimization is called *overfitting* [3, Chap. 4]. A common technique to avoid overfitting is to keep part of the available data into a validation subset, not used for training. The NN parameters $\boldsymbol{\theta}$ are subsequently optimized solely on the training data, whereas the performance can be monitored throughout the training on both the training and validation data. In a well posed problem where the training procedure converges, one can generally observe a first period where both the training and validation performance improves. This mean that what the NN learns from the training data is relevant to make predictions on the unseen validation data. In the second period, the performance on the training data still improves while performance on the validation data stays stable or worsens. This means that the NN overfits to the training data by learning features that are not generalizable to new data. These two periods are illustrated in Fig. 2.7.

A common way to handle overfitting in practice is to monitor the loss value on the validation data, and save the model parameters that gives the best loss value on validation data. This procedure is called *early stopping* [3, Chap. 7]. The procedure however assumes that the validation data is representative of the future data the NN will be used on and that the loss value is representative of the clinical problem at stake. In both of the technical contributions presented in this thesis, these assumptions are not met.

For the LVEF problem in Chapter 3, there was three test datasets, from

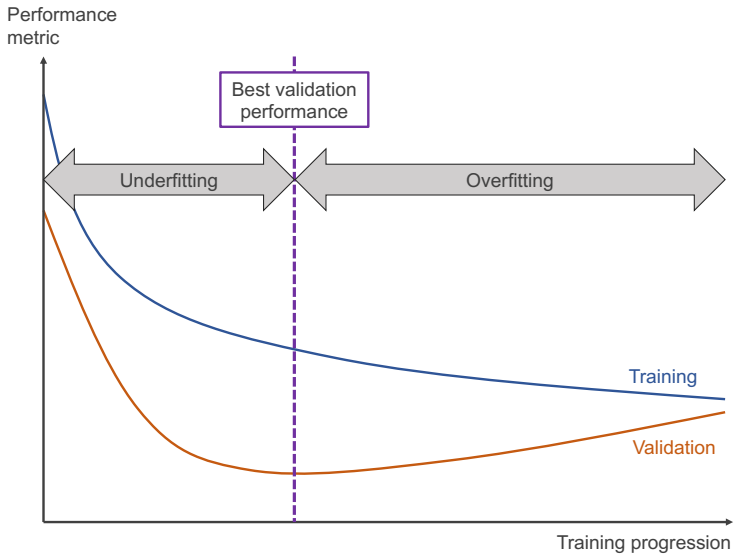
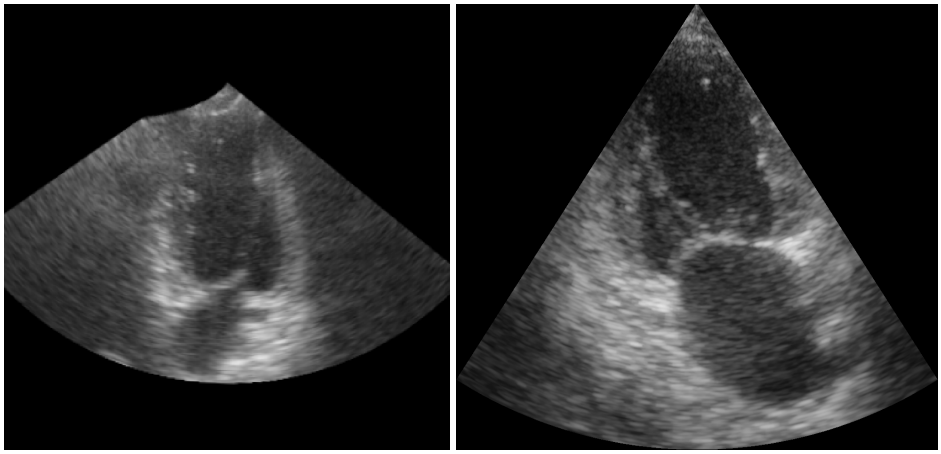


Figure 2.7: Evolution of the performance metric on the training and validation data throughout the training. Early training shows underfitting, where both the performance on the validation and training data improves. Overfitting starts when the validation performance stops improving.

different populations and acquired with different ultrasound machines, bringing an issue for defining what was representative validation data. Further, focus was placed on the clinical problem of heart failure management whereas the NN was optimized with a loss function based on the LVEF, introducing a discrepancy between the loss and the clinical problem. There is no guarantee that optimizing the deep neural network for LVEF performance will lead to improvements in heart failure management.

For the position regression problem in Chapter 4, training data was made of slices from 3D US volumes whereas the NN was meant to finally be used on 2D data, here again introducing a discrepancy between the training data and data representative of the use scenario as shown in Fig. 2.8. Also the final clinical problem was to propose the operator transducer displacements whereas the loss function was to predict a relative position along either the rotation or tilt degree of freedom (DOF). The loss function was consequently only a proxy for the clinical problem of guiding the operator, but not directly representative of it.

Both contributions details approaches to measure performance on representative data and according to the clinical problem.



(a) Slice from a 3D volume (Training) **(b)** 2D frame (Inference)

Figure 2.8: *Left:* Slice from a 3D volume acquired specifically for the purpose of the study presented in Chap. 4. The field of view is wide and has a skewed shape. *Right:* 2D frame acquired with a 2D transducer, showing a triangular field of view rounded at the bottom.

2.2.3 Automated measurements with deep learning

Deep learning has shown a high potential to automate measurements from echocardiographic data. Such measurements can be extracted from one or several recordings, either 1D, 2D or 3D. One of the inherent advantages of DL is the removed intra- and inter-observer variability from the interpretation step, meaning that repeated automatic measurements on the same recording return the same value. However, it is worth mentioning that, for a given recording, removing human variability is a necessary but not sufficient condition for removing the measurement error as the DL method can still have a systematic error with regard to the *true* value. Automatic DL approaches also allow real-time (or close to real-time) measurements, permitting the operators to optimize the recordings and significantly reducing the interpretation time.

There has been two main approaches for automating measurements with supervised DL. The first one can be referred as *multi-step approach* and the second one as *end-to-end approach*.

Multi-step measurements approach

Multi-step measurement approaches aims to reproduce the manual measurement pipeline by replacing manual actions by DL components. Examples of manual actions that can be automatized with DL are detection of standard views [19, 20], cardiac events [21], or cardiac structures [22]. The possibility to visualize the output of each component of the pipeline benefits interpretability. However, having numerous automatic components multiplies the number of error sources that add to each other, similarly to the manual pipeline.

Such fully automatic multi-step approaches have been conceptualized for left ventricular ejection fraction [23] and left ventricular strain [24]. Multi-step approaches are not considered in this thesis.

End-to-end measurements approach

The other approach to automate measurements can be referred to as *end-to-end*. It consists in using a single DL model taking echocardiographic data as input and providing an estimation of the measurement as output, without human interaction. This kind of approach is often called "black-box" as it is not easily explainable, and can be regarded as automatic "eye-balling". The potential advantage of end-to-end approaches compared to multi-step approaches is that they do not implement any human *a priori* and are particularly suited to discover image features that human operators may not see. End-to-end measurements is the approach used in the contribution presented in Chapter 3.

2.2.4 Regression from images

Both contributions in Chapters 3 and 4 are regression problems in a continuous space of outputs. In Chap. 3, a neural network is used to regress a clinical measurement, the left ventricular ejection fraction (LVEF), from echocardiographic cine-loops (2D+time). In Chap. 4, a NN is used to regress the relative position of the ultrasound transducer relatively to the heart.

The strategy adopted was to use classifiers networks to address these two regression problems. A similar approach has been used previously for age estimation from facial pictures [25] or brain MRI [26]. By using classifiers for regression, one can make the training procedure aware of the continuous nature of the problem, as well as a possible uncertainty in the reference values.

Continuity and uncertainty

Continuity can be introduced in the training procedure by using non-binary labels that represents a probability distribution. This is also relevant in the presence of uncertain reference values, which are better represented by a probability distribution than a single value.

LVEF problem

For the LVEF problem presented in Chap. 3, the manual reference consists in LVEF values in a continuous space, which physiologically range between approx. 10% and 70%. To automatically regress these values, a NN with $N = 100$ output classes was used, and a continuous space of possible LVEF between 0% and 100% was discretized into 100 bins of 1% width each. The chosen bin size was therefore small relatively to the magnitude of the measurements. The reference values were further transformed into a Gaussian reference probability distribution, centered on the reference value and with a standard deviation of 10 LVEF % points, in line with the observer variability of LVEF measurements [9–12]. The predicted LVEF value was recovered by using the dot products of the output confidence vector C_{LVEF}, analogous to a probability distribution, and the corresponding LVEF bins.

Position regression problem

For the position regression problem presented in Chap. 4, not all the input frames initially had an associated label. Indeed, the 2D training data was generated by simulating a scanning "swipe" by slicing a 3D US volume to simulate rotation and tilt movements of the transducer.

For the tilt degree of freedom (DOF), the tilt span, and by extension the number of slices, was defined by the heart anatomy to obtain constant heart structures relatively to the slice location within the span. NNs with $N_{tilt} = 11$ classes were used for the tilt direction, and labels were automatically assigned to each slice. The predicted tilt position $\hat{p}_{\text{tilt}}$ of any frame X was then expressed as a value in [0;10] calculated using expression (2.11), where C^X_{tilt} is the output confidences vector of the tilt NN for the frame X.

$$\hat{p}_{\text{tilt}}(X) = \sum_{n=0}^{N_{\text{tilt}}-1} C^X_{\text{tilt}}(n) \cdot n \tag{2.11}$$

For the rotational DOF, 360 slices were generated with a one degree rotational increment, simulating a 360° rotation of the transducer. The last slice hence neighbored the first slice, giving to the problem a circular nature.

Out of these 360 slices, 12 were manually annotated as Characteristic Cross Sections (CCSs) related to the presence or absence of specific heart structures. It was therefore chosen to use a neural network with $N_{rot} = 12$ output classes and to label all the slices *relatively* to these 12 classes, thus accounting for heart shape variability among patients. This was done by using linear interpolation relatively to the two nearest manually annotated CCSs. The predicted rotational position $\hat{p}_{\text{rot}}$ of any frame X accounts for the circular nature of the problem. It is expressed relatively to the 12 CCSs, taking values in [0;12[calculated using expression (2.12), where C^X_{rot} is the output confidences vector of the rotational NN for the frame X.

$$\hat{p}_{\text{rot}}(X) = \frac{N_{\text{rot}}}{2\pi} \arctan2\left(\sum_{n=0}^{N_{\text{rot}}-1} C^X_{\text{rot}}(n) \cdot \sin\left(\frac{n \cdot 2\pi}{N_{\text{rot}}}\right), \sum_{n=0}^{N_{\text{rot}}-1} C^X_{\text{rot}}(n) \cdot \cos\left(\frac{n \cdot 2\pi}{N_{\text{rot}}}\right)\right) \tag{2.12}$$

Specific loss functions

Considering that the output's confidence vectors C also are probability distributions, it was chosen to use loss functions that quantify the similarity between two probability distributions.

Kullback–Leibler divergence

For the LVEF problem presented in Chap. 3, the Kullback–Leibler (KL) divergence [25] was used to measure the similarity between the predicted probability function and the reference probability function. The KL divergence D_{KL} between two discrete probability distributions P and Q of length N can is expressed as:

$$D_{KL}(P,Q) = \sum_{x=0}^{N-1} P(x) \log\left(\frac{P(x)}{Q(x)}\right) \tag{2.13}$$

For the use case of LVEF, the loss L was then expressed as:

$$L_{\text{LVEF}} = D_{KL}(ref_{\text{LVEF}}, C_{\text{LVEF}}) \tag{2.14}$$

where ref_{LVEF} represents the probability distribution associated to the manual reference of LVEF and C_{LVEF} the probability distribution predicted by the NN. Fig. 2.9 shows two predictions examples.

Squared Earth Mover's Distance

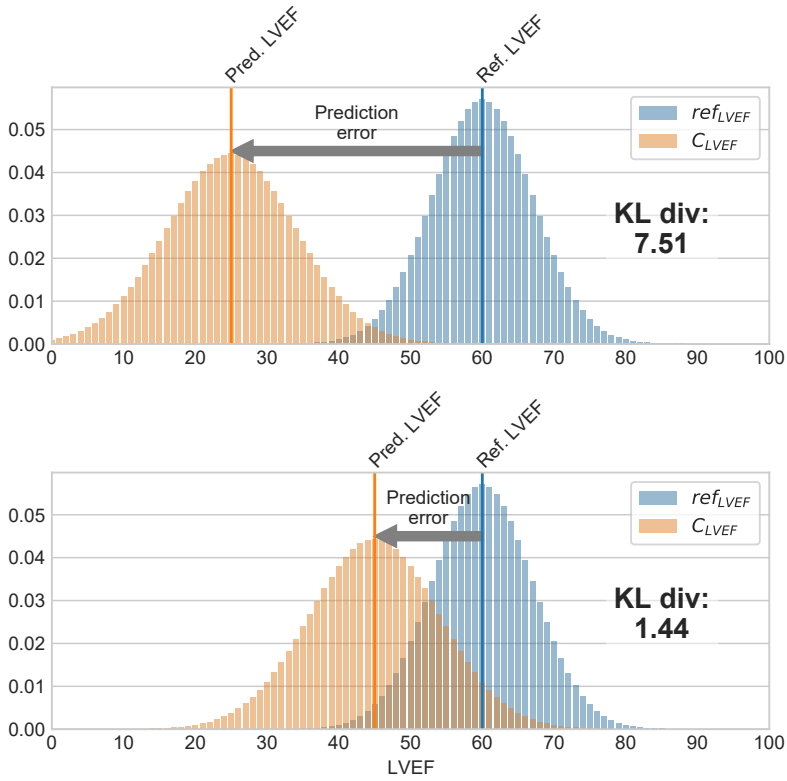


Figure 2.9: Illustration of the KL divergence. KL divergence increases with distance between the two probabilities distributions. In converging training procedures, the KL divergence decreases throughout the training and the prediction error improves.

For the position regression problem presented in Chap. 4, the Squared Earth Mover's Distance was used as the loss function [27]. For normalized one dimensional distributions, it can be reduced to the Wasserstein metric which can be expressed in its discrete version as:

$$W(P,Q) = \frac{1}{N}\sum_{i=0}^{N-1}\left|\sum_{k=0}^{i} p_k - \sum_{k=0}^{i} q_k\right| \tag{2.15}$$

where P and Q are two discrete probability distributions of length N. For the rotational case, there was a further need to account for the circular nature of the problem. This was done by calculating W for different shifts

and taking the minimum value:

$$W_{circular}(P,Q) = \min_{j\in[0;N[}(W(\text{shift}_j(P),\text{shift}_j(Q))) \tag{2.16}$$

The loss L for the tilt NN was then calculated as:

$$L_{\text{tilt}}(ref^X_{\text{tilt}}, C^X_{\text{tilt}}) = W(ref^X_{\text{tilt}}, C^X_{\text{tilt}})^2 \tag{2.17}$$

whereas the loss for the rotational NN was calculated as

$$L_{\text{rot}}(ref^X_{\text{rot}}, C^X_{\text{rot}}) = W_{circular}(ref^X_{\text{rot}}, C^X_{\text{rot}})^2 \tag{2.18}$$

where ref^X is the reference probability distribution of frame X and C^X its predicted probability distribution.

2.2.5 Data augmentation

Image data augmentation is a widely employed technique in DL to augment the size of the training dataset artificially. By applying various transformations to existing images, data augmentation enhances the model's capacity to generalize and achieve better performance on unseen data.

The primary objective of data augmentation is to introduce variations in the training images that mirror real-world scenarios encountered during inference. This process strengthens the model's robustness and mitigates overfitting, where the model memorizes the training data without effectively learning the underlying patterns.

Some classical image data augmentation techniques employed in DL are:

- *Rotation:* Images can be rotated to simulate different viewpoints
- *Flipping:* Images can be horizontally or vertically flipped
- *Translation:* Horizontal or vertical shifts can be applied to images, simulating object movement within the image. This augmentation cultivates the model's invariance to object location.
- *Scaling:* Images can be scaled up or down, emulating objects at different distances. Zooming can also focus on specific regions of interest.
- *Cropping and Padding:* Images can be cropped or padded to different aspect ratios or resolutions, simulating variations in image sizes or aspect ratios within the training data.
- *Dropout:* Part of images can be removed

These augmentation techniques can be combined to generate a diverse range of training examples. In this thesis, flipping, cropping and padding were not used as flipped images or images with wrong aspect ratio create physiologically wrong heart structures. On the other hand, rotation, translation, scaling or dropout preserve the physiological appearance of the heart structures. It is however worth noting that they break the appearance of what an US image looks like.

Further augmentations specific to the echocardiography domain were used in the presented contributions. Among them were augmentations to simulate changes in scanning parameters such as dynamic range, gain, or field-of-view width. Augmentations that aims to reproduce acoustics phenomenons such haze artifacts, shadows or depth attenuation were also used. Examples of augmented data as used in the position regression contribution from Chap. 4 are visible in Fig. 2.10.

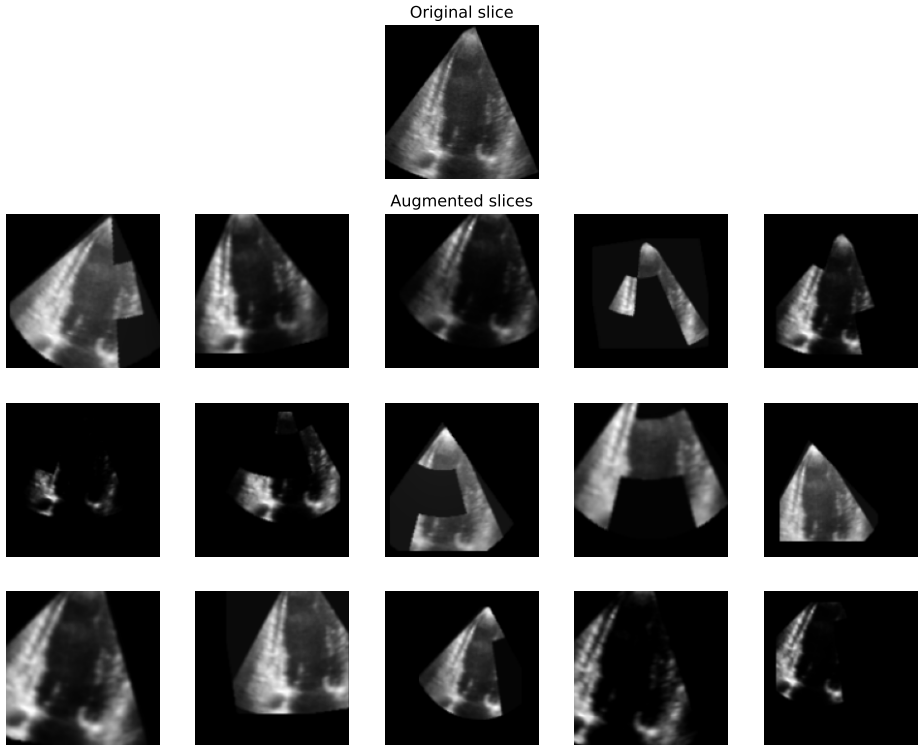


Figure 2.10: Examples of multiple augmented frames generated from a single frame. These augmentations are representative of the ones used in the position regression contribution from Chap. 4

2.3 Bland-Altman analysis

Bland-Altman analysis is a statistical technique often used to compare two different measurement methods and conclude on whether they are interchangeable or not [28]. It was developed by J. Martin Bland and Douglas G. Altman in 1986 and has since become widely used in various fields, including medicine, biology, and engineering with their original article totaling over 50'000 citations.

The purpose of Bland-Altman analysis is to compare two quantitative measurement methods and determine the level of agreement and bias between them. It is the reference statistical approach when there is no ground truth for comparison (i.e. when the *true* value is not accessible). The analysis helps identify any systematic differences between the methods and provides insights into the magnitude and direction of any discrepancies.

The Bland-Altman plot visualizes the agreement between the two methods on a scatter plot, with the differences between the measurements plotted on the y-axis and the average of the measurements plotted on the x-axis. An horizontal line representing the mean difference and two parallel lines representing the limits of agreement (mean difference ± 1.96 times the standard deviation of the differences) are also plotted. An example of a Bland-Altman plot is shown in Fig. 2.11.

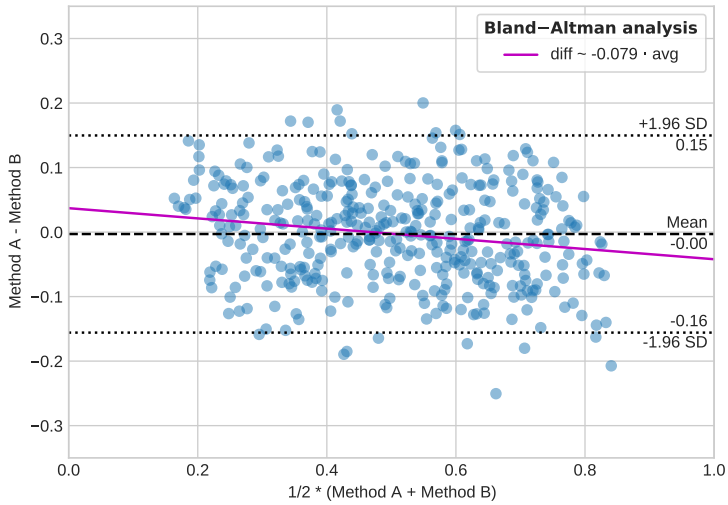


Figure 2.11: Example of a Bland-Altman plot

By examining the plot, several key characteristics can be observed:

- *Fixed bias:* The mean difference between the two methods is represented by the horizontal dashed line on the plot. If this line is close to zero, it suggests there is no significant bias. A deviation from zero indicates a systematic difference between the two methods.
- *Limits of agreement:* The parallel dotted lines on the plot represent the limits of agreement, which provide a range within which most differences between the two methods are expected to fall. The wider the limits, the larger the variability or disagreement between the methods.
- *Outliers:* Any data points that fall outside the limits of agreement or exhibit a substantial deviation from the mean difference are considered outliers and may warrant further investigation.

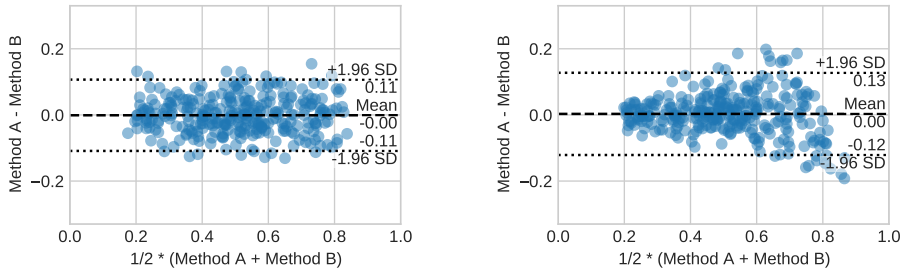
Additionally, the correlation between the difference and the average of the two methods provides insights on an eventual proportional bias between the two methods. In this thesis, Bland-Altman plots are completed with a regression line between the difference and the average to show this possible proportional bias. The existence of a correlation between the difference and the average however does not necessarily imply a correlation of the measurement method with the *true* values.

Assumptions

Bland-Altman analysis assumes that the differences between the two methods are normally distributed. Another assumption is homoscedasticity, meaning that the variability of differences should be constant across the range of measurements as shown in Fig. 2.12a. Violations of these assumptions, for example the presence of heteroscedacity (Fig. 2.12b) may affect the validity of the analysis. In the contribution presented in Chap. 3, the homoscedasticity assumption in not met, which should be considered when interpreting the results.

Relation to precision and accuracy

Within the field of machine learning, accuracy and precision are often used to describe the performance of a model by comparing a *true* value to the model predictions, as illustrated in Fig. 2.13. A model can be considered as valid if it is both accurate and precise.



(a) Homoscedasticity **(b)** Heteroscedasticity

Figure 2.12: Bland-Altman analysis showing a) homoscedasticity, the magnitude of the difference is constant relatively to the average and b) heteroscedacity, with the magnitude of the differences depending on the average, violating one of the assumptions for the Bland-Altman analysis

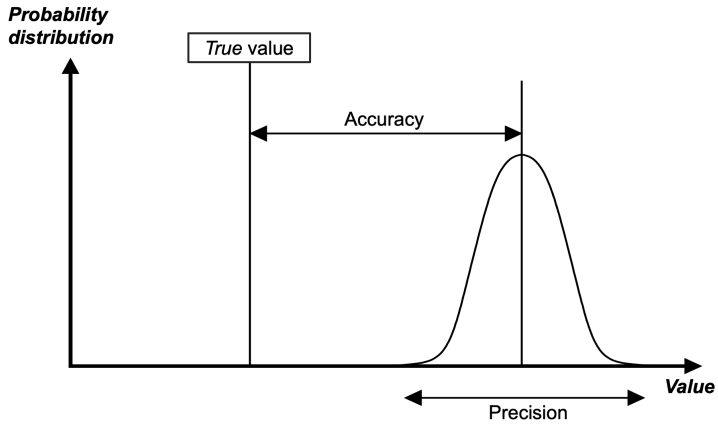


Figure 2.13: Quantification of accuracy and precision of a measurement method relatively to the *true* value

In echocardiography, the *true* value of a measurement is typically not accessible, leading the DL predicted measurements to be compared to the manual reference measurements. In this context of uncertainty in both measurement methods, accuracy can be approximated by the mean error and correlation between the difference and average from the Bland-Altman analysis, and precision can approximated by the width of the limits of agreement. Such approach to quantify accuracy and precision is however limited by the purpose of the Bland-Altman analysis itself, which is to assess interchangeability of the two methods.

Correlation between the difference and the average

According to [29], the difference and the average of two measurements methods X and Y can be correlated, without necessarily meaning that there is a bias between the measurement methods and the *true* values. In [30], this correlation between the difference and the average of two measurements sets X and Y can be expressed as:

$$Corr(X-Y, \frac{X+Y}{2}) = \frac{\tau_X^2 - \tau_Y^2}{\sqrt{\left(\tau_X^2 - \tau_Y^2\right)^2 - 4\rho^2\tau_X^2\tau_Y^2}} \tag{2.19}$$

where τ_X^2 and τ_X^2 are the variances of the two measurements sets and ρ the correlation between X and Y. If one set $\tau_X^2 = k \cdot \tau_Y^2$, the correlation between the difference and the average becomes:

$$Corr(X-Y, \frac{X+Y}{2}) = \frac{k-1}{\sqrt{(k+1)^2 - 4k\rho^2}} \tag{2.20}$$

There is therefore a theoretical correlation between the difference and average values when the variances of the two sets of measurements are unequal ($k \neq 1 \Leftrightarrow \tau_X^2 \neq \tau_Y^2$).

The variance of the measurements sets τ_X^2 and τ_Y^2 can be decomposed into the variance of the *true* values τ_{true}^2 and the variance of the measurement errors $\tau_{err\_X}^2$ and $\tau_{err\_Y}^2$:

$$\begin{aligned} \tau_X^2 &= \tau_{true}^2 + \tau_{err\_X}^2 \\ \tau_Y^2 &= \tau_{true}^2 + \tau_{err\_Y}^2 \end{aligned} \tag{2.21}$$

Thus k can be expressed as:

$$k = \frac{\tau_{true}^2 + \tau_{err\_X}^2}{\tau_{true}^2 + \tau_{err\_Y}^2} \tag{2.22}$$

It appears then that when the range of the true values is much wider compared to the measurement error ($\tau_{true}^2 \gg \tau_{err\_X}^2$ and $\tau_{true}^2 \gg \tau_{err\_Y}^2$), k is close to 1. This is what is expected in realistic circumstances. As noted by Carstensen [4, Chap. 4], it would otherwise "mean that the study was either badly designed (too narrow a range of the true values) or that (at least) one of the methods was so imprecise that it would be clinically useless".

In the context of DL applied to echocardiography, where X denotes the set of manual reference values and Y the corresponding estimated values by DL, two specific aspects should be accounted for when comparing a manual measurement reference method with an automatic DL measurement method:

1. The human measurement error may not be negligible compared to the range of true values. This is for example the case for LVEF where the observer variability has the same order of magnitude as the measurement. This means $\tau^2_{true} \sim \tau^2_{err\_X}$
2. DL effectively reduces measurement error compared to human operators, such that $\tau^2_{err\_Y} < \tau^2_{err\_X}$

If the two aspects occur together, k is different from 1 and a correlation between the differences and the average could appear, even in the absence of a proportional bias between the DL measurement method and the *true* values. Bland-Altman analyses should therefore be interpreted carefully in the context of echocardiographic measurements and DL.

References

[1] S. D. Solomon, *Essential Echocardiography*. Elsevier, 2019.

[2] I. Goodfellow, Y. Bengio, and A. Courville, *Deep Learning*. MIT Press, 2016.

[3] F. Chollet, "Deep Learning with Python," Manning, 2018.

[4] B. Carstensen, *Comparing Clinical Measurement Methods*. Wiley, aug 2010.

[5] R. M. Lang, L. P. Badano, M. A. Victor, J. Afilalo, A. Armstrong, L. Ernande, F. A. Flachskampf, E. Foster, S. A. Goldstein, T. Kuznetsova, P. Lancellotti, D. Muraru, M. H. Picard, E. R. Retzschel, L. Rudski, K. T. Spencer, W. Tsang, and J. U. Voigt, "Recommendations for cardiac chamber quantification by echocardiography in adults: An update from the American Society of Echocardiography and the European Association of Cardiovascular Imaging," *Journal of the American Society of Echocardiography*, vol. 28, no. 1, pp. 1–39.e14, 2015.

[6] T. A. McDonagh, M. Metra, M. Adamo, R. S. Gardner, A. Baumbach, M. Böhm, H. Burri, J. Butler, J. Celutkiene, O. Chioncel, J. G. Cleland, A. J. Coats, M. G. Crespo-Leiro, D. Farmakis, M. Gilard, and S. Heymans, "2021 ESC Guidelines for the diagnosis and treatment of acute and chronic heart failure," *European Heart Journal*, vol. 42, no. 36, pp. 3599–3726, 2021.

[7] P. A. Heidenreich, B. Bozkurt, D. Aguilar, L. A. Allen, J. J. Byun, M. M. Colvin, A. Deswal, M. H. Drazner, S. M. Dunlay, L. R. Evers, J. C. Fang, S. E. Fedson, G. C. Fonarow, S. S. Hayek, A. F. Hernandez, P. Khazanie, M. M. Kittleson, C. S. Lee, M. S. Link, C. A. Milano, L. C. Nnacheta, A. T. Sandhu, L. W. Stevenson, O. Vardeny, A. R. Vest, and C. W. Yancy, "2022 AHA/ACC/HFSA Guideline for the Management of Heart Failure: A Report of the American College of Cardiology/American Heart Association Joint Committee on Clinical Practice Guidelines," *Journal of the American College of Cardiology*, vol. 79, no. 17, pp. e263–e421, 2022.

[8] E. Smistad, A. Østvik, I. Mjåland Salte, S. Leclerc, O. Bernard, and L. Lovstakken, "Fully Automatic Real-Time Ejection Fraction and MAPSE Measurements in 2D Echocardiography Using Deep Neural Networks," in *IEEE International Ultrasonics Symposium, IUS*, vol. 2018-Octob, IEEE Computer Society, dec 2018.

[9] C. Knackstedt, S. C. Bekkers, G. Schummers, M. Schreckenberg, D. Muraru, L. P. Badano, A. Franke, C. Bavishi, A. M. S. Omar, and P. P. Sengupta, "Fully Automated Versus Standard Tracking of Left Ventricular Ejection Fraction and Longitudinal Strain the FAST-EFs Multicenter Study," *Journal of the American College of Cardiology*, vol. 66, no. 13, pp. 1456–1466, 2015.

[10] A. Hovland, U. H. Staub, H. Bjørnstad, J. Prytz, J. Sexton, A. Støylen, and H. Vik-Mo, "Gated SPECT Offers Improved Interobserver Agreement Compared With Echocardiography," *Clinical Nuclear Medicine*, vol. 35, pp. 927–930, dec 2010.

[11] R. Hoffmann, G. Barletta, S. von Bardeleben, J. L. Vanoverschelde, J. Kasprzak, C. Greis, and H. Becher, "Analysis of Left Ventricular Volumes and Function: A Multicenter Comparison of Cardiac Magnetic Resonance Imaging, Cine Ventriculography, and Unenhanced and Contrast-Enhanced Two-Dimensional and Three-Dimensional Echocardiography," *Journal of the American Society of Echocardiography*, vol. 27, pp. 292–301, mar 2014.

[12] C. Morbach, G. Gelbrich, M. Breunig, T. Tiffe, M. Wagner, P. U. Heuschmann, and S. Störk, "Impact of acquisition and interpretation on total inter-observer variability in echocardiography: results from the quality assurance program of the STAAB cohort study," *International Journal of Cardiovascular Imaging*, vol. 34, no. 7, pp. 1057–1065, 2018.

[13] A. J. DeGrave, J. D. Janizek, and S. I. Lee, "AI for radiographic COVID-19 detection selects shortcuts over signal," *Nature Machine Intelligence*, vol. 3, no. 7, pp. 610–619, 2021.

[14] M. Abadi, P. Barham, J. Chen, Z. Chen, A. Davis, J. Dean, M. Devin, S. Ghemawat, G. Irving, M. Isard, M. Kudlur, J. Levenberg, R. Monga, S. Moore, D. G. Murray, B. Steiner, P. Tucker, V. Vasudevan, P. Warden, M. Wicke, Y. Yu, and X. Zheng, "TensorFlow: A system for large-scale machine learning," *Proceedings of the 12th USENIX Symposium on Operating Systems Design and Implementation, OSDI 2016*, pp. 265–283, 2016.

[15] A. Paszke, S. Gross, F. Massa, A. Lerer, J. Bradbury, G. Chanan, T. Killeen, Z. Lin, N. Gimelshein, L. Antiga, A. Desmaison, A. Köpf, E. Yang, Z. DeVito, M. Raison, A. Tejani, S. Chilamkurthy, B. Steiner, L. Fang, J. Bai, and S. Chintala, "PyTorch: An imperative style, high-performance deep learning library," *Advances in Neural Information Processing Systems*, vol. 32, no. NeurIPS, 2019.

[16] D. P. Kingma and J. L. Ba, "Adam: A method for stochastic optimization," *3rd International Conference on Learning Representations, ICLR 2015 - Conference Track Proceedings*, pp. 1–15, 2015.

[17] M. D. Zeiler, "ADADELTA: An Adaptive Learning Rate Method," 2012.

[18] K. Hornik, M. Stinchcombe, and H. White, "Multilayer feedforward networks are universal approximators," *Neural Networks*, vol. 2, no. 5, pp. 359–366, 1989.

[19] A. Østvik, E. Smistad, S. A. Aase, B. O. Haugen, and L. Lovstakken, "Real-Time Standard View Classification in Transthoracic Echocardiography Using Convolutional Neural Networks," *Ultrasound in Medicine and Biology*, vol. 45, pp. 374–384, feb 2019.

[20] J. Zhang, S. Gajjala, P. Agrawal, G. H. Tison, L. A. Hallock, L. Beussink-Nelson, M. H. Lassen, E. Fan, M. A. Aras, C. Jordan, K. E. Fleischmann, M. Melisko, A. Qasim, S. J. Shah, R. Bajcsy, and R. C. Deo, "Fully Automated Echocardiogram Interpretation in Clinical Practice," *Circulation*, vol. 138, pp. 1623–1635, oct 2018.

[21] A. M. Fiorito, A. Ostvik, E. Smistad, S. Leclerc, O. Bernard, and L. Lovstakken, "Detection of Cardiac Events in Echocardiography Using 3D Convolutional Recurrent Neural Networks," *IEEE International Ultrasonics Symposium, IUS*, vol. 2018-Janua, 2018.

[22] S. Leclerc, E. Smistad, J. Pedrosa, A. Ostvik, F. Cervenansky, F. Espinosa, T. Espeland, E. A. R. Berg, P. M. Jodoin, T. Grenier, C. Lartizien, J. Dhooge, L. Lovstakken, and O. Bernard, "Deep Learning for Segmentation Using an Open Large-Scale Dataset in 2D Echocardiography," *IEEE transactions on medical imaging*, vol. 38, no. 9, pp. 2198–2210, 2019.

[23] E. Smistad, A. Østvik, I. M. Salte, D. Melichova, T. M. Nguyen, K. Haugaa, H. Brunvand, T. Edvardsen, S. Leclerc, O. Bernard, B. Grenne, and L. Løvstakken, "Real-Time Automatic Ejection Fraction and Foreshortening Detection Using Deep Learning," *IEEE Transactions on Ultrasonics, Ferroelectrics, and Frequency Control*, vol. 67, no. 12, pp. 2595–2604, 2020.

[24] A. Ostvik, I. M. Salte, E. Smistad, T. M. Nguyen, D. Melichova, H. Brunvand, K. Haugaa, T. Edvardsen, B. Grenne, and L. Lovstakken, "Myocardial Function Imaging in Echocardiography Using Deep Learning," *IEEE Transactions on Medical Imaging*, vol. 40, no. 5, pp. 1340–1351, 2021.

[25] B. B. Gao, C. Xing, C. W. Xie, J. Wu, and X. Geng, "Deep Label Distribution Learning with Label Ambiguity," *IEEE Transactions on Image Processing*, vol. 26, no. 6, pp. 2825–2838, 2017.

[26] H. Peng, W. Gong, C. F. Beckmann, A. Vedaldi, and S. M. Smith, "Accurate brain age prediction with lightweight deep neural networks," *Medical Image Analysis*, vol. 68, p. 101871, 2021.

[27] L. Hou, C.-P. Yu, and D. Samaras, "Squared Earth Mover's Distance-based Loss for Training Deep Neural Networks," no. November, 2016.

[28] J. Martin Bland and D. Altman, "STATISTICAL METHODS FOR ASSESSING AGREEMENT BETWEEN TWO METHODS OF CLINICAL MEASUREMENT," *The Lancet*, vol. 327, pp. 307–310, feb 1986.

[29] M. A. Mansournia, R. Waters, M. Nazemipour, M. Bland, and D. G. Altman, "Bland-Altman methods for comparing methods of measurement and response to criticisms," *Global Epidemiology*, vol. 3, p. 100045, 2021.

[30] J. Bland and D. Altman, "Comparing methods of measurement: why plotting difference against standard method is misleading," *The Lancet*, vol. 346, pp. 1085–1087, oct 1995.

Challenges and Strategies for Automatic Measurements with Deep Learning in Cardiovascular Imaging

3

David Pasdeloup[1], Andreas Østvik[1,2], Sindre H. Olaisen[1], Eirik Skogvoll[1], Håvard Dalen[1,3], and Lasse Løvstakken[1]

[1] Dept. of Circulation and Medical Imaging, NTNU, Trondheim, Norway
[2] SINTEF Digital, Medical Image Analysis, Trondheim, Norway
[3] Clinic of Cardiology, St. Olav's Hospital, Trondheim, Norway

Real-Time Echocardiography Guidance for Optimized Apical Standard Views

4

David Pasdeloup[1], Sindre H. Olaisen[1], Andreas Østvik[1,2], Sigbjørn Sæbø[1], Håkon N. Pettersen[1], Espen Holte[1,3], Bjørnar L. Grenne[1,3], Stian B. Stølen[3], Erik Smistad[1,2], Svein Arne Aase[4], Håvard Dalen[1,3], and Lasse Løvstakken[1]

[1] Dept. of Circulation and Medical Imaging, NTNU, Trondheim, Norway
[2] SINTEF Digital, Medical Image Analysis, Trondheim, Norway
[3] Clinic of Cardiology, St. Olav's Hospital, Trondheim, Norway
[4] GE Vingmed Ultrasound AS, Horten, Norway

Measurements of cardiac function such as left ventricular ejection fraction and myocardial strain are typically based on 2D ultrasound imaging. The reliability of these measurements depends on the correct pose of the transducer such that the 2D imaging plane properly aligns with the heart for standard measurement views, and is thus dependent on the operator's skills. We propose a deep learning tool that suggests transducer movements to help users navigate towards the required standard views while scanning. The tool can simplify echocardiography for less-experienced users and improve image standardization for more experienced users. Training data was generated by slicing 3D ultrasound volumes, which permits to simulate the movements of a 2D transducer. Neural networks were further trained to calculate the transducer position in a regression fashion. The method was validated and tested on 2D images from several datasets representative of a prospective clinical setting. The method proposed the adequate transducer movement 75% of the time when averaging over all degrees of freedom, and 95% of the time when considering transducer rotation solely. Real-time application examples demonstrates the direct relation between the transducer movements, the ultrasound image, and the provided feedback.

4.1 Introduction

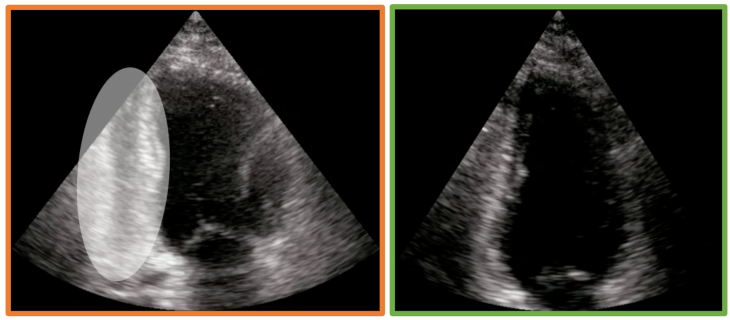
Echocardiography is the cornerstone image modality for measuring and evaluating cardiac function, today primarily based on 2D image acquisition following a defined protocol of standard views and measurements used for subsequent diagnostics [1]. The acquisition of these standard views is challenging for inexperienced users, and is prone to substantial variation even among experienced users. This limits the availability of echocardiography for the patients, and also decreases measurements reproducibility [2].

The acquisition of the three apical standard views, apical four-chamber (A4C), two-chamber (A2C) and long-axis (ALAX) can be broken down into three steps, as illustrated in Fig. 4.1. First, the users need to find the correct intercostal acoustic window where the anatomical apex of the heart is closest to the transducer. In the second step, users need to rotate and tilt the transducer around the left ventricular centerline (between the apex and the center of the mitral valve) to produce the three standard apical imaging planes with minimal foreshortening. The last step consists in optimizing the images for both anatomical correctness and image quality to establish a more detailed cardiac examination. We here define *Anatomical Correctness* (*AC*) to depend only on the location of the 2D imaging plane. *AC* is not related to the quality of the ultrasound (US) signal.

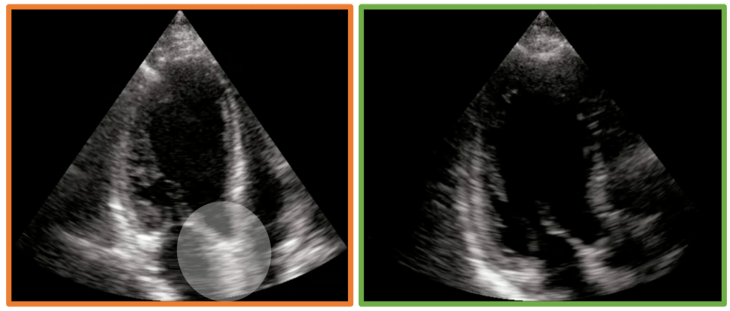
Figure 4.1: Procedure steps and experience required for acquisition of the apical standard views.

The starting point of our work is that the standard apical views acquired by both inexperienced and experienced users have an operator variability in *AC*, influencing the measurements that follow. Two examples of standard views of different quality are shown in Fig. 4.2. These are recordings from the echo lab, thus representing clinical practice, where anatomical landmarks reveal differences in *AC*.

3D US imaging can be part of a future solution, as volume measurements can be more directly extracted, and standard planes automatically extracted [3, 4]. However, 3D and multi-plane imaging still suffers from limitations



(a) A2C view – The suboptimal recording (left) has the right ventricle partly visible



(b) ALAX view – The aortic valve is not clearly visible in the sub-optimal recording (left)

Figure 4.2: Example of the inter-operator variability in the imaging plane of clinical recordings. For a given target view on the same patient, the left recording acquired by operator A is a suboptimal standard view whereas the right recording from operator B complies with the guidelines recommendations. Full cine-loops are available in Video 4.3.

such as a lower frame rate and suboptimal image quality compared to 2D imaging, and optimizing each view is often needed to get good image quality. Further, 3D US is not readily available for hand-held systems. We thus believe that 2D echocardiography will remain important for the foreseeable future.

Some research and development has previously been done to facilitate the echocardiography procedure described in Fig. 4.1. In step 1, finding a suitable intercostal space can be seen as minimizing apical foreshortening, which was previously approached using image segmentation [5]. Finding the imaging plane in step 2 has been addressed before by both classical and machine learning (ML) based approaches [6, 7]. However, these methods were only able to grade the *AC*, and did not give the user feedback on the required transducer movements to acquire the standard views.

In Østvik et al. [7], we proposed an exploratory extension of the view classification method with 3D data for training to enable feedback to the user on rotational movement of the transducer. However, the solution was preliminary, limited to the rotational movement, and struggled to generalize sufficiently for real-time 2D imaging. To our best knowledge, Toporek et al. [8] is the only technical description of a DL-based echocardiography guidance system with training data generated with a 2D transducer and an external positional sensor. Their approach is, however, limited by a low number of patients used in the training dataset, the accuracy not being quantified for the A2C and ALAX views, and lack of prospective testing. Navigation using an inertial measurement unit could be used to provide feedback, and thus guide the user. A solution would, however, still be needed to recognize the target view and the direction to it. Li et al. [9] used a neural network to calculate the required geometric transformation to obtain a standard plane within a 3D fetal US volume and Droste et al. [10] proposed a neural network that predicts an *angle to target* value given a 2D fetal image. However, echocardiography introduces additional challenges such as the ribs limiting the acoustic window or anatomical variability. It is therefore not straightforward to apply methods from fetal US to heart US.

ML echocardiography assistance is also a topic of interest for several companies who have developed proprietary solutions. Narang et al. (Caption Health Inc., Brisbane, CA, USA) [11] showed that their algorithm can help nurses to acquire recordings of diagnostic quality, but did not reported results on the correctness of each individual standard view. The same algorithm was tested on medical students by Schneider et al. [12] who reported the algorithm being helpful to acquire the A4C and A2C views. However, these two studies lack a control group of users who scan without the algorithm assistance. The technical challenge of echocardiography guidance is therefore not solved in its entirety.

In this work, we propose a real-time guidance method for 2D echocardiography which estimates the transducer rotation and tilt in relation to the cardiac anatomy based only on the 2D image input. The method provides feedback on how to adjust the transducer position in order to reach the target standard apical views. This tool can be valuable 1) to improve image view consistency and help limit subsequent measurement variability for experienced users in the echo lab, and 2) to unleash the potential of echocardiography for non-expert users using hand-held devices, thus unlocking a potential tool for detecting heart disease prior to hospital referral. Finally, the tool can 3) be used to train novice users in echocardiography.

Main contributions

We have developed and evaluated a software-based echocardiography guidance system based on positional regression directly from 2D images. The main contributions of this paper with respect from previous work from Østvik et al. [7] are:

- An improved protocol for defining echocardiographic anatomical views, including non-standard apical views used for image guiding
- A semi-automated approach for extracting 2D training data from 3D US recordings with an accurate reference position
- An more robust neural network for estimating the transducer rotational position
- New neural networks for estimating the transducer tilt position
- Extensive validation and test on several representative 2D datasets from the clinic
- A real-time prototype application that shows evidence of the method validity

4.2 Methods

Our main approach is to use deep convolutional neural networks to predict the transducer position relative to the heart in the form of a regression task based on the 2D US image input, where training data was generated from 3D US volumes, as detailed in the following sections.

4.2.1 Degrees of freedom and problem formulation

Two-dimensional echocardiography involves multiple degrees of freedom (DOF), with two translation DOFs in the skin plane and three rotation DOFs (transducer rotation, in-plane tilt, out-of-plane tilt). Additional challenges can be attributed to the motion of the heart, patient breathing, and the restricted acoustic window between the ribs, limiting the possibility to obtain good image quality and aligned cardiac views.

To reduce the complexity of the problem, we initially assumed that the heart apex is located at a shallow depth and regarded it as an anchor, meaning that the correct intercostal space is used and that no significant translations at the skin surface are required. Considering that the out-of-plane transducer rotations are difficult for the users who lack spatial representation of the heart, we address in this work the rotation and out-

of-plane tilt DOFs (referred further as rotation and tilt DOFs) and leave the in-plane tilt DOF to separate work.

We considered rotation and tilt separately and trained four individuals models, one for the transducer rotation and three for the transducer tilt in the different apical views. The approach of using multiple models mimicks the sonographers' workflow who iteratively adjust the transducer tilt and rotation. It is also convenient to isolate method failures, and improves explainability.

A particular challenge for position regression is the intrinsic variation of the hearts shape. This is exemplified in Fig. 4.3, which shows the patient variability in the amount of rotation from the A4C to the A2C and ALAX views. Due to this variation, we did not design a neural network that outputs an *angle to target* value as in Droste et al. [10], but rather a position relative to the characteristic heart structures defined below.

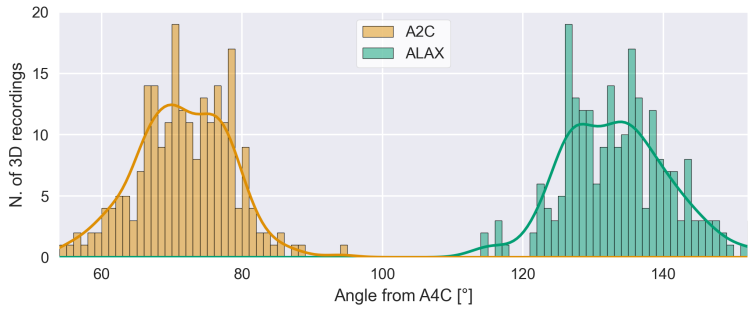


Figure 4.3: The angle between A4C-A2C and A4C-ALAX standard views obtained from the manually annotated 3D data.

4.2.2 Training data generation

The 2D training images were generated from 3D US recordings following the procedure shown in Fig. 4.4. Starting from a 3D US recording, we automatically extracted the apex and base landmarks and the maximum left ventricle (LV) radius (steps 1. and 2.). The landmark and radius estimates were obtained from the segmentation mask of our 3D-Unet [13], and were quality controlled and corrected by a clinical expert if needed.

Once the LV axis (going through apex and base landmarks) was defined, 360 slices were automatically generated (step 3.) at end diastole (ED) for each patient to simulate a transducer rotation movement around the heart's long axis. A clinical expert annotated 6 of the 360 slices as

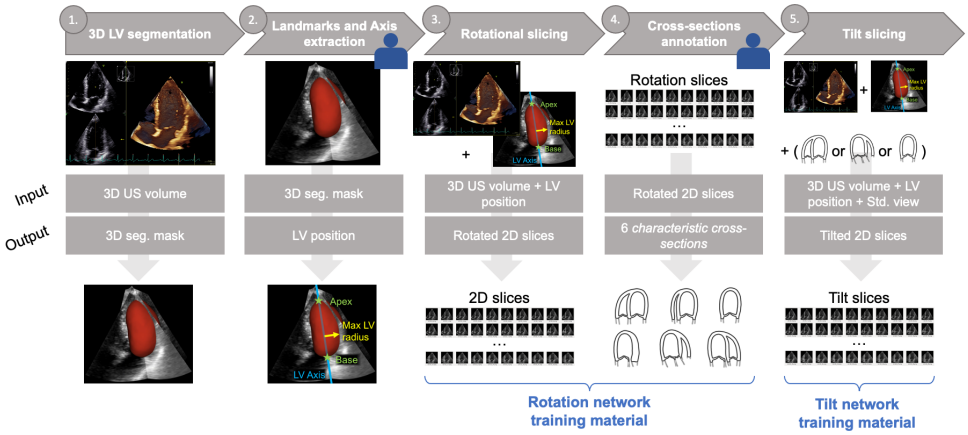


Figure 4.4: Training data generation pipeline. Step 1. takes a 3D US volume at end diastole (ED) as input and outputs a 3D mask of the LV. Step 2. generates from the mask the apex and base landmarks which together form the LV geometric rotational axis. The maximum LV radius is also extracted. The axis and the radius were manually corrected by a clinical expert. Step 3. automatically generates 360 slices of the volume around the defined rotational axis. At step 4., a clinical expert annotated six slices as *CCS*s. Step 5. automatically generates the slices to train the tilt models based on the ED rotational position (either A4C, A2C or ALAX) and the LV axis and radius. The annotations (LV axis and *CCS*s) were made at the ED frame whereas the training slices were generated from multiple frames throughout the cardiac cycle.

characteristic cross-sections (*CCS*) (step 4.). Three were the apical standard views (Fig. 4.5b), while three were non-standard (Fig. 4.5c), inserting additional clinical knowledge in the rotation position regression. The characteristic features of the additional views are termed 1) the A42C view (mid way between A4C and A2C) which has the presence of a thin right ventricle (RV) but no tricuspid valve visible, or a thickened inferoseptal wall, 2) The ANS (Annulus Start) view and 3) ANE (Annulus End) view, which are recognisable by the left ventricular outflow tract (LVOT) being slightly visible, while the aorta valve leaflets are not visible. ANS further typically has a thick anterior wall, whereas ANE has the RV visible.

For the tilt DOF, the corresponding *CCS*s were extracted automatically at the rotational position of the standard views (either A4C, A2C or ALAX). Knowing the LV axis length l_{LV} and the LV radius r_{LV}, as shown in Fig. 4.6a,

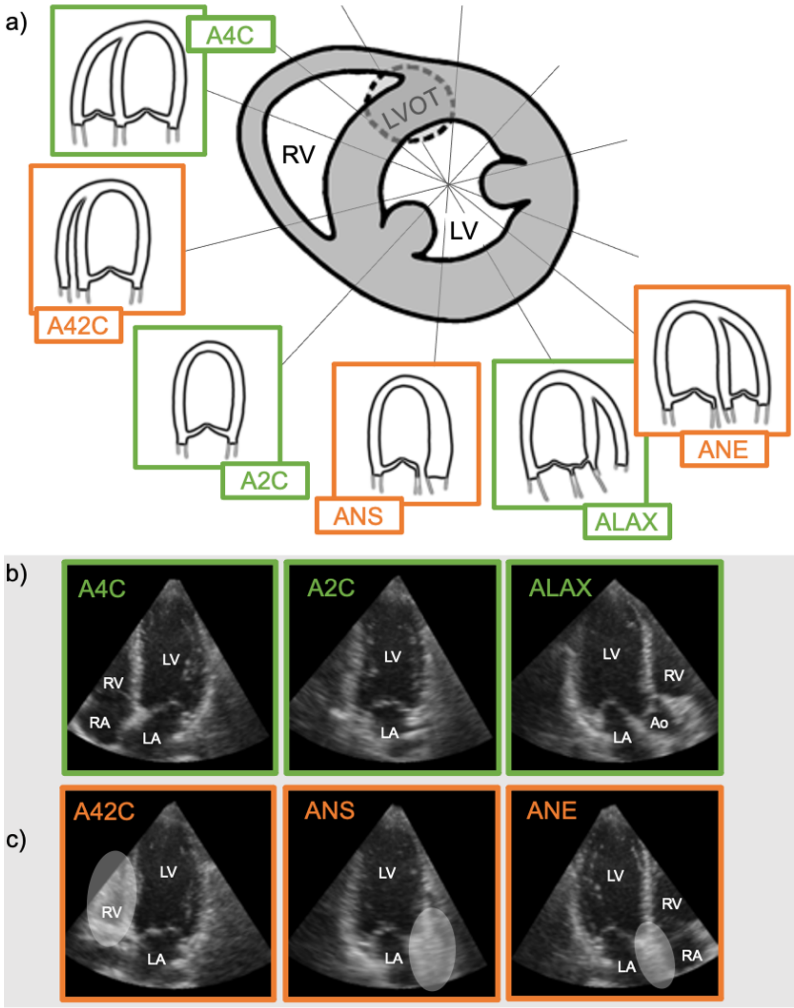


Figure 4.5: (a) Position of the six rotational *CSS*s in the parasternal short-axis plane. (b) The three *CSS*s corresponding to the apical standard views, and (c) for non-standard views. Characteristic features of the A42C view is the presence of a thin RV without visible tricuspid valve or a thick inferior septal wall. ANS and ANE are recognised by LVOT being slightly visible, but the aorta valve leaflets not visible. ANS typically has a thick anterior wall, whereas ANE has the RV visible.

the tilt span S_w can be expressed as,

$$S_w = 2 \cdot \tan^{-1} \frac{\alpha \cdot r_{\mathrm{LV}}}{l_{\mathrm{LV}}}. \tag{4.1}$$

The α is a coefficient chosen larger than 1 so that the slices at the outer edges of the span neither include the mitral valve (MV) nor the left atrium (LA). The tilt slices were evenly distributed over the tilt span, which should correspond to anatomical features derived from any LV size rather than an absolute tilt angle, thus accounting for patient variability.

Once annotation at ED is done, the training slices are generated from several frames assuming the LV axis and *CCS*s annotated at ED are stable throughout a single cardiac cycle.

4.2.3 Image position regression with Deep Learning

Considering the continuity along the slices for each DOF, the position can be posed as a regression problem and thus addressed with DL-based architectures. Continuity was included in the training procedure using the Earth mover's distance as the loss function [14], which measures the distance between the true and predicted positions, and penalizes predictions with larger distance error. To account for patient variability, the predicted position $\hat{p}$ is given relatively to the labeled *CCS*s and thus to the characteristic heart structures presented in Fig. 4.5 and 4.6.

Separate rotation and tilt networks were trained. The rotation network has 12 output categories, corresponding to the 6 rotational *CCS*s and their flipped counterparts. Since the different *CCS*s are not evenly spaced along the rotational DOF as shown in Fig. 4.3, the number of rotational slices is sampled to balance the categories.

The tilt network has 11 output categories sampled equidistantly by the automated tilt annotation procedure described in Fig. 4.4. Since there are more slices than labelled *CCS*s for both rotational and tilt DOFs, the slices in-between are assigned non-binary labels obtained through linear interpolation of the two closest *CCS*s so that the models learn the continuous nature of the task.

The predicted relative position $\hat{p}$ for any input frame X is obtained by the dot product of the network output vector C and class index vector as

$$\hat{p}(X) = \sum_{n=0}^{N-1} C^X(n) \cdot n. \tag{4.2}$$

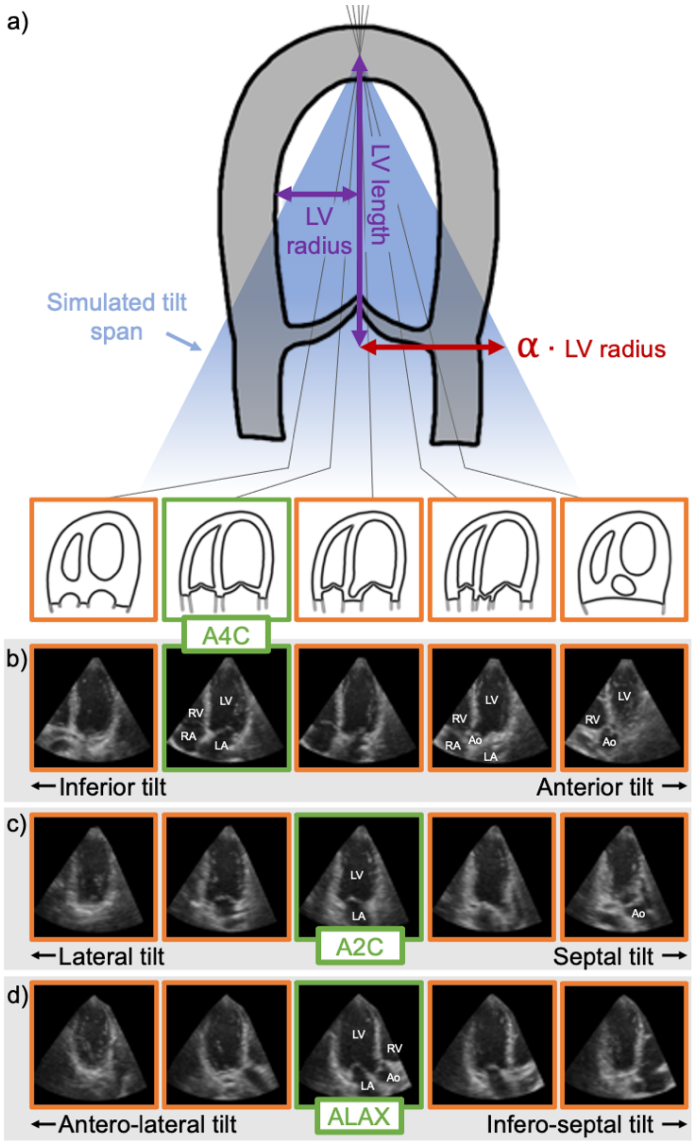


Figure 4.6: (a) Description of cross-sections for the A4C tilting (b) A4C tilt samples (c) A2C tilt samples (d) ALAX tilt samples

We can then define the categorical distance d of any frame X to the target CCS by

$$d(X, CCS) = \hat{p}(X) - p(CCS), \tag{4.3}$$

where $p(CCS)$ is the index of the target cross-sectional view.

We used our CVCNet topology [7] which is optimized for US image classification and real-time performance. Other model architectures are benchmarked in the supplementary material.

In the real-time application, the target CCS corresponds to the three standard views A4C, A2C and ALAX. The sign of d then indicate the direction of the required transducer movement along the corresponding DOF and low values of d indicate a correct view.

4.2.4 Data augmentation and pre-processing

Rotation, translation, scaling and gamma intensity transformation augmentations form our *Baseline* set of augmentations used in all experiments. Additionally, we used several US specific augmentations, termed *US augmentations*:

- *Gaussian shadows* to mimick signal dropouts or acoustic wave propagation artefacts
- *Depth attenuation* to add varying depth-dependent dampening of the acoustic waves
- *Haze artifacts* to mimick acoustic haze typically originating from reverberations
- *Non-linear color-maps* to make the models robust to different color transformations
- *Contrast scaling and brightness transform* to increase robustness to gain and dynamic range
- *Field of view (FOV) masking* to make the models disregard the FOV shape and avoid unintended effects related to the sector width and depth settings.

Specifically for our image guiding task, we finally introduced and evaluated two pre-processing techniques related to the use of 3D volumes as the training data source:

- *Spatial Reference Noise*: In the data generation process, we established a reference LV axis position (Fig. 4.4, step 2.). To increase the robustness of the networks to minor errors in transducer tilt and translation, we introduced stochastic deviations to the actual LV axis and rotational

position before generating the slices. The amount of deviation remains small compared to the LV dimensions so that the annotations can still be considered valid.
- *Gaussian Blurring*: This pre-processing step aims to improve generalization from 3D slices to 2D US images at inference. Blurring is implemented as a Gaussian filter with standard deviation two orders of magnitude lower than the image size. The goal is to make images more similar by smoothing details like speckle while preserving the heart structures. Blurring is applied upstream of the convolutional layers for both training and inference.

4.3 Datasets

We use datasets of 2D and 3D recordings to train, validate and test our models. Regional ethics board approval and written consent was obtained for all studies and patients involved.

4.3.1 Training and internal validation

Three-dimensional recordings were used for training the models and were also used as part of the validation procedure. The FOV of 3D recordings is often reduced to achieve higher frame rate in clinical practice. Therefore, slices from such volumes are not necessarily representative of what is acquired with 2D echocardiography. Consequently, we only included 3D volumes that cover the whole LV and its surrounding structures (RV and LA). The 3D recordings acquired using ECG gating with visible stitching artifacts were excluded. This ensures that we have sufficient context to generate slices with structures similar to those visible in 2D echocardiography which is the data used during inference.

We included 3D US recordings from three different datasets, in total N=162 patients:

- *CETUS*: An open dataset of N=45 patients [15], both pathological and healthy. The recordings are acquired with scanners from three different vendors.
- *NTNU 3D A*: Recordings from N=36 patients acquired at St Olavs Hospital, Trondheim, Norway, using a GE E95 scanner and 4Vc transducer (GE Vingmed Ultrasound, Horten, Norway). The population is representative of the daily routine at the echo lab, with both healthy and diseased patients.

- *NTNU 3D B*: N=81 recordings acquired with a GE E95 scanner at institutions spread over six different countries.

20% of the patient data were set aside for internal validation and are not used for training. Slices from 3D US volumes have generally lower image quality, especially in the near field. To ensure that the trained models do not overfit the training data, we also validated the method on an external dataset made of 2D US images representative of clinical use as described in the following.

4.3.2 External 2D validation dataset

To validate our method against 2D images, we specifically acquired the *NTNU 2D Guiding* dataset, composed of N=47 patients where each include 15 recordings. For each standard view the following was acquired by a clinical expert:

- The standard view,
- Two views with negative and positive rotation angle compared to the standard one,
- Two views with negative and positive tilt angle compared to the standard one.

The non-standard recordings have a position error relative to the standard recording that is qualitatively known, allowing to evaluate the ordering metrics given by equations 4.4 and 4.5.

4.3.3 External 2D test datasets

We tested our method on two larger retrospective datasets composed of standard view recordings, without information on the *AC*. The *CAMUS* dataset (500 patients) [16] which features A4C and A2C views and the *NTNU-LVD* (Left Ventricular Disease) dataset (168 patients) [17] which features in addition the ALAX view. Both datasets are representative of daily clinical practice in terms of image quality and pathological cases.

Finally, we tested our method on a repeatability study dataset of 88 patients who underwent three consecutive exams containing each of the three apical views, carried out by a panel of three sonographers and three cardiologists. We call this dataset *NTNU 2D Repeatability*.

4.4 Experimental setup

4.4.1 Training and validation

Our method gives feedback to help users adjust the transducer position towards a more optimized view. At the end of each training epoch, we ran two validation procedures:

1. For the first procedure the metric of interest is the validation loss, calculated on the internal validation subset of the annotated slices from 3D. This validation loss is calculated similarly to the training loss. Neither *Spatial Reference Noise* nor *US augmentations* are applied when calculating this validation loss.

2. For the second procedure we introduce the *Ordering Success Rate (OSR)*. This metric is more appropriate for evaluating our models from the user perspective of 2D image guiding. The procedure uses views from the *NTNU 2D Guiding* external validation dataset. For each combination of patient, standard view, rotation or tilt, we use a set of three recordings corresponding to three different orientations:

   - *Recording A:* Anatomically correct standard view,
   - *Recording B:* Slight positive transducer angulation along the DOF from Recording A,
   - *Recording C:* Slight negative transducer angulation along the DOF from Recording A.

   The true ordering is then expressed as:

   $$p(C) < p(A) < p(B), \tag{4.4}$$

   and the ordering success is then defined by:

   $$\text{Outcome} = \begin{cases} \text{Success}, & \text{if} \quad \hat{p}(C) < \hat{p}(A) < \hat{p}(B) \\ \text{Fail}, & \text{otherwise} \end{cases} \tag{4.5}$$

   where p and $\hat{p}$ are the ground truth and predicted relative positions of the 2D imaging plane, respectively. The *OSR* is defined as the ratio of success cases over all cases.

We run several experiments using the *US augmentations*, the *Spatial Reference Noise* and the *Gaussian Blurring* separately and all together. The performance is compared to using only the *Baseline* set of augmentations. For each DOF, the best model is chosen considering the *OSR* metric, which reflects the guiding performance on 2D images. The validation loss remains informative to avoid overfitting to the training data when selecting the model for further testing.

4.4.2 Measure of image standardization

To test our models at a larger scale, we retrospectively run them on the CAMUS and NTNU-LVD external test datasets made of 2D standard views representative of clinical practice. These datasets contain data labelled as standard views. While the labels do not contain information on the true *AC*, this experiment allows us to qualitatively investigate the degree of standardization of recordings acquired in a clinical setup.

Additionally, we study the inter-operator *AC* variability in the image acquisition by quantifying the *AC* on recordings from the *NTNU 2D Repeatability* dataset.

4.4.3 Real-time image guiding application

We designed and developed a prototype application for image guidance using the FAST framework [18] which combines image streaming from a GE E95 clinical scanner (GE Vingmed Ultrasound, Horten, Norway), and real-time DL inference and visualization. In the application, as shown in Fig. 4.10 and in Video 4.1, the user chooses a standard view to acquire and a target box corresponding to this view gets displayed on the short-axis schematic of the heart. The current position of the transducer relative to the heart is updated in real-time based on the models and shown as a blue line. When the current view is predicted as anatomically correct by both the rotation and tilt networks, the blue line is inside the target box which changes color from red to green to indicate a valid view.

Video 4.1: Click or scan to watch online

4.5 Results

4.5.1 Patient variability

From the output of step 4 in the annotation procedure in Fig. 4.4, we could investigate the variability in rotation angle from A4C to A2C and ALAX views. Results are presented in Fig. 4.3 and demonstrates variability in heart shapes. This justify the need for a method that calculates the transducer position relatively to some characteristic heart features rather than an *angle to target* value.

4.5.2 Method validation

Transducer rotation estimation

Fig. 4.7a shows the predictions through the cardiac cycle for the six rotational *CCS*s and suggests good convergence of the training procedure on the internal validation set of 3D slices when using only the baseline augmentations. These predictions are made on 2D slices from 3D volumes, as for training, and are expected to yield the best case results. As observed, the rotational model separates the different classes throughout the cardiac cycle although the LV axis and the rotational positions are annotated at ED. Fig. 4.8a is made from the same images, but using a model trained with additional US-augmentations and specific pre-processing. To evaluate performance on 2D data during training, the *OSR* was calculated on recordings from the *NTNU 2D Guiding* dataset at the end of each epoch. Results are presented in the left column of Table 4.1 and corresponds to the average *OSR* for the three groups of standard views. The best model from

the baseline experiment ordered the standard and non-standard 2D views correctly in 92% of the cases, while an improved score of 95% was obtained when using our additional augmentations and pre-processing steps. This improvement in the *OSR* is also visible in Fig. 4.8b, where the predictions are more accurate and precise through the heart cycle than in Fig. 4.7b.

Table 4.1: Success rate for ordering non-standard views. Values in bold show improvement using specific augmentations compared to the baseline training procedure.

	Rotation (all views)	Tilt A4C	Tilt A2C	Tilt ALAX
Baseline	0.92	0.81	0.53	0.44
US augmentations	**0.93**	**0.95**	0.53	0.33
Gaussian blurring	**0.93**	0.81	**0.60**	0.37
Spatial Ref. Noise	0.90	**0.84**	**0.60**	0.42
All	**0.95**	**0.91**	**0.67**	**0.47**

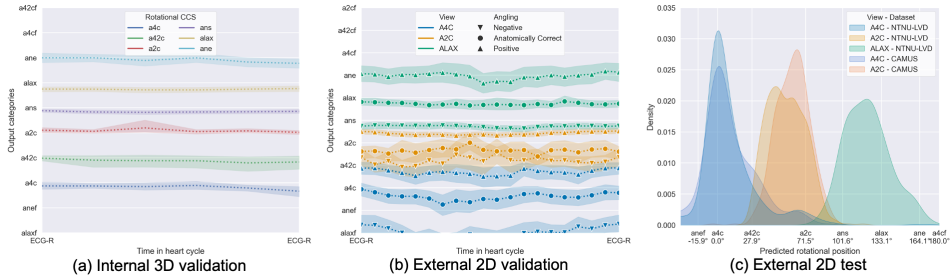


Figure 4.7: Rotational results - Model with the best validation loss from the baseline experiment. (a) Model predictions for the annotated *CCS*s, (b) model predictions for the *NTNU 2D Guiding* recordings, (c) density plot of the *AC* for recordings from *CAMUS* and *NTNU-LVD*

Transducer tilt estimation

On the sliced 3D data (Fig. 4.9, left column), the A4C tilt model predictions are consistent with the ground truth *CCS*s, with stable predictions along the cardiac cycle. Although all three models converge, the variability and

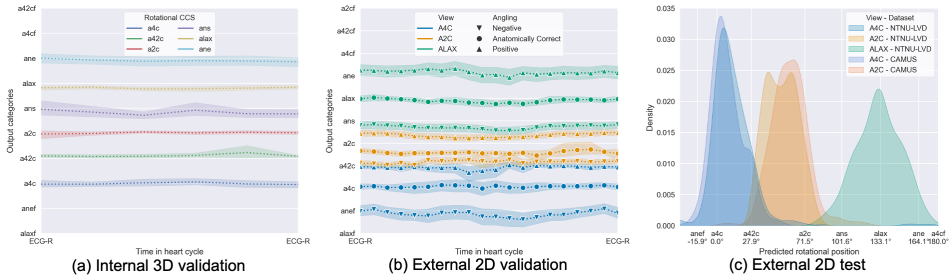


Figure 4.8: Rotational results - Model with the best *OSR* on 2D non-standard views validation data from the experiment featuring all the pre-processing and augmentations. (a) Model predictions for the annotated *CCS*s, (b) model predictions for the *NTNU 2D Guiding* recordings, (c) density plot of the *AC* for recordings from *CAMUS* and *NTNU-LVD*

stability are slightly worse for the A2C and ALAX tilt models, with more classes overlapping. On the 2D validation data (center column), the A4C tilt model has the best *OSR* of 91% whereas being 67% (resp. 47%) for the A2C (resp. ALAX) view. These scores can be compared against a random choice scenario leading to a 16.6% correct ordering, meaning that all the models performs significantly better than a random baseline. One can also note from Table 1 that the improvements obtained with each augmentation or pre-processing taken separately are not additive when combined.

4.5.3 Real-time image guiding application

In addition to the quantitative offline assessment, we assessed the method qualitatively using our real-time application. The Video 4.1 shows the real time application being used to acquire the three apical standard views. To the best of our knowledge, this video is the first one showing evidence of a correlation between the transducer movements, the ultrasound image, and the given feedback for all three apical views throughout a large number of heart cycles. For all standard and non-standard views, the position of the blue feedback line on the short axis schematic was responsive to the transducer movements and consistent with the structures visible on the US image.

Fig. 4.10a shows a suboptimal A4C view with visible LVOT. The blue feedback line is consequently moved away from the target towards the anterior wall. In Fig. 4.10b the operator acquired a good A4C view. This is confirmed by the application displaying the feedback line into the target box which consequently switched color to green.

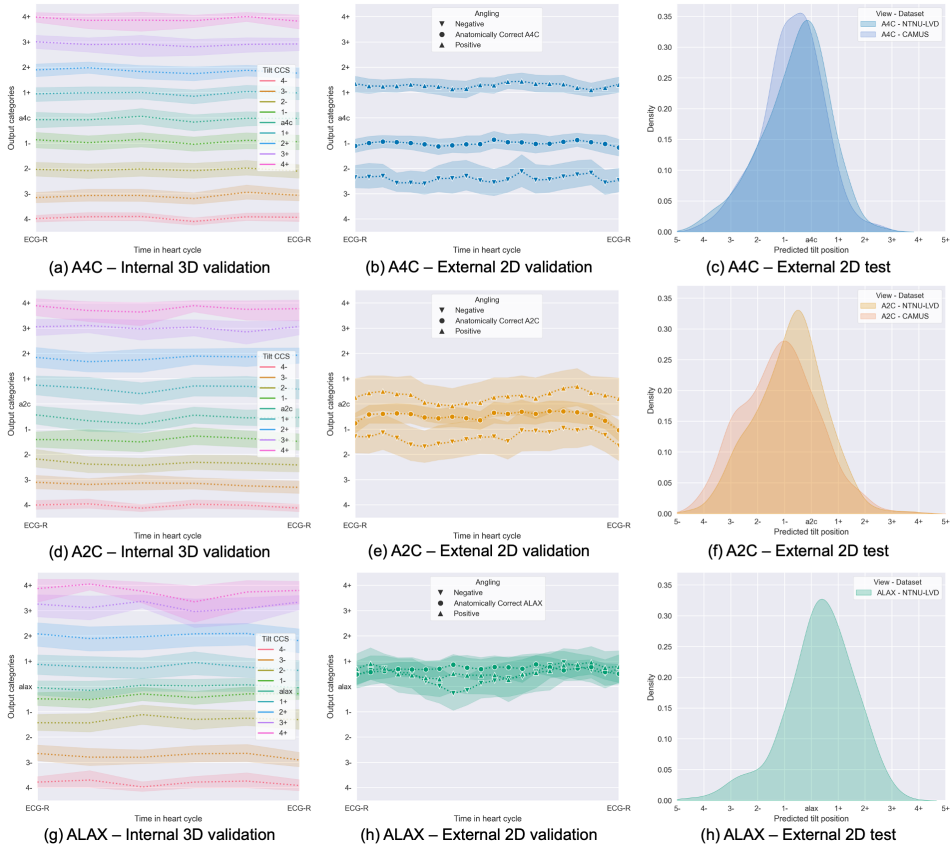
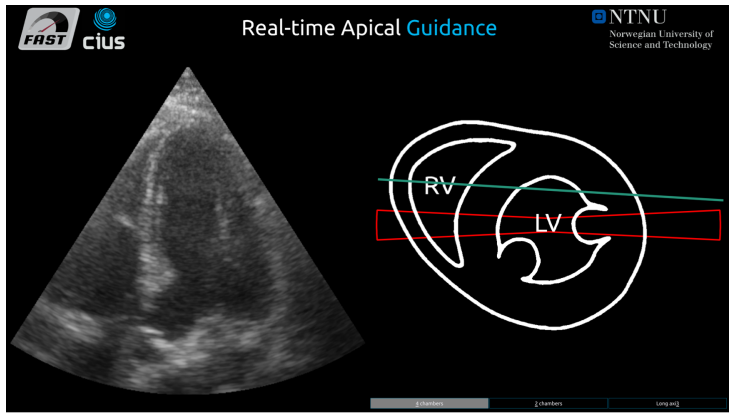


Figure 4.9: Tilt results - Model with the best *OSR* on 2D non-std. views validation data from the experiment featuring all the pre-processing and augmentations. (left column) Model predictions for the annotated *CCS*s, (center column) model predictions for the *NTNU 2D Guiding* recordings, (right column) density plot of the *AC* for recordings from *CAMUS* and *NTNU-LVD*

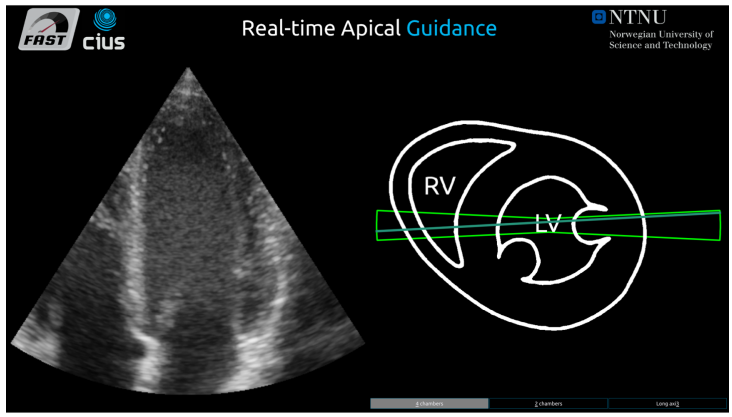
4.5.4 Measure of image standardization

Retrospective assessment of recordings

The predicted view rotation and tilt for recordings from the CAMUS and the NTNU-LVD datasets are shown in the right column of Figs. 4.8 and 4.9. The corresponding visual inspection is available in Video 4.2 which shows that for each standard view, suboptimal recordings are located at the tails of the distributions whereas standard ones are located at the center of the distributions. This suggests that our method is robust enough to predict the correct transducer position over a large number of patients.



(a) Incorrect A4C view with visible LVOT – Target box is red



(b) Correct A4C view – Target box is green

Figure 4.10: Screen captures of the real-time application prototype. Left side of the screen shows the US image streamed from a GE E95 scanner. Right side of the screen shows the calculated feedback based on the neural networks output.

Additionally, one can observe that the positional biases for the external test datasets are consistent with the positional biases on the 2D validation data from the *NTNU 2D Guiding* dataset, with for example the A2C rotational position being shifted towards the A4C rotational position.

Inter-Operator Variability

We evaluated the *AC* of the assumed optimal apical standard views from the *NTNU 2D Repeatability* data set with our method, with predictions showing significant differences among operators (Fig. 4.11). For the A2C rotational

Video 4.2: Click or scan to watch online

position the method suggested that operator 3 tends to acquire the A2C views closer to the A4C view. This was confirmed by the visual inspection available in Video 4.3 where the RV was typically partly visible. For the ALAX rotational position, the results suggested that the ALAX views of operators 1 and 3 are similar to the A4C view with a vertical flip. Visual inspection from Video 4.3 revealed that these views included four chambers (RA is not expected in ALAX) and the LVOT (the aorta valve leaflets are expected to be visible in addition to the LVOT).

Video 4.3: Click or scan to watch online

4.6 Discussion

We developed and evaluated a deep learning-based method capable of estimating the anatomical orientation of apical views in echocardiography. Multiple deep neural networks were trained to regress the position of the transducer using 2D slices generated from 3D US volumes as the training material. So far we considered the transducer rotation and out-of-plane tilt movements, assuming the user has already positioned the transducer

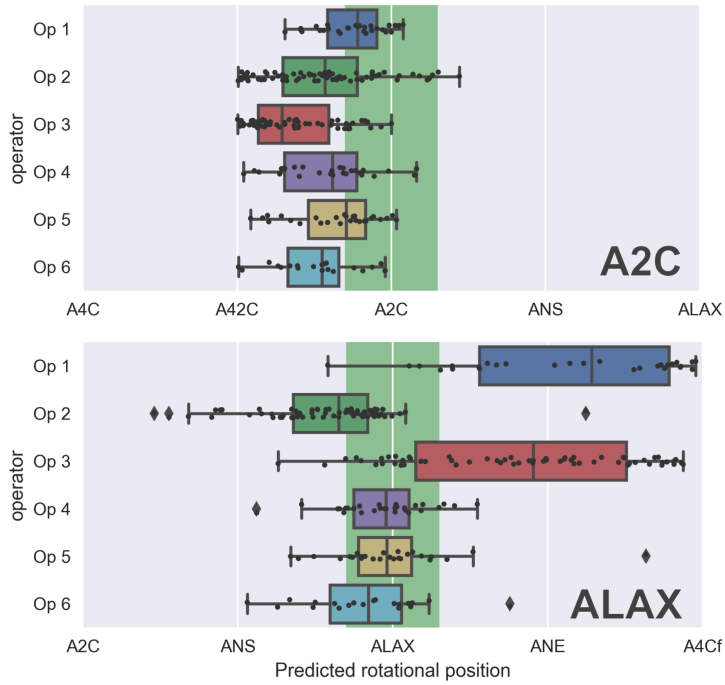


Figure 4.11: Predicted rotational position by operator for 2D recordings for the *NTNU 2D Repeatability* dataset

such that the image plane passes approximately through the apex of the heart. The method was evaluated using 2D ultrasound recordings and showed promising results in terms of robustness and accuracy. The real-time implementation of the method showed ability to provide the user correct feedback on the required movements needed to acquire the valid standard views.

The method is suitable for real-time inference both on off-the-shelf GPUs as as well as less powerful hand-held devices. Thus, it may benefit less experienced users to obtain anatomically correct standard views for diagnostics, and can contribute to standardize image views acquired in the echo lab. The method can also be used for quality control, for instance to help clinicians automatically selecting the recording and cardiac cycle most suitable for a given measurement within an exam consisting of many recordings. This can also be highly useful in a research setting, for instance when data mining large amounts of patient examinations.

For both online and offline use cases, the method can reduce the measurement variability introduced during image acquisition, and thereby

contribute to a more accurate patient management. Used retrospectively, the method can be used as a tool to analyze and improve scanning habits in individual operators.

An important aspect of our work is the use of 3D US recordings as the primary data source, which combined with our semi-automatic annotation procedure allowed us to use data from 162 patients from several echo labs. Using 3D data, the labelled reference position is given relative to the heart anatomy. Our method therefore compensate for the heart movements inside the chest that occurs while breathing, contrarily to Toporek et al. [8] which uses an external sensor to acquire the reference position.

Using 3D US volumes as the primary data source comes at the price of a discrepancy between the training data made of 3D slices and the inference data consisting of 2D recordings. We therefore introduced an additional validation metric, the *OSR*, which is directly related to the ability to detect small transducer movements on 2D data and thus has more clinical relevance than the traditional validation loss.

The present work divided the method into simpler sub problems, which allowed us to better identify some of the challenges related to the apical standard views *AC* assessment with DL networks. For instance, we found that the estimated rotational position was more accurate than the estimated tilt position. The average predictions of the rotational model over the cardiac cycle on the *NTNU 2D Guiding* shown in Fig. 4.8b revealed a bias of the A2C rotation towards the A4C. This bias is also present in the CAMUS and NTNU-LVD datasets, as shown in Fig. 4.8c. Visual inspection available in Video 4.2 revealed that many recordings labelled as standard A2C views partly include the RV or the coronary sinus vein. This suggests that the rotational position reference is correct and that the A2C 2D recordings from the *NTNU 2D Guiding*, *CAMUS* and *NTNU-LVD* datasets does not fully comply with the expected A2C standard view. Our method thus has the potential to improve *AC* for experienced users.

Visual inspection of the tilt recordings from the *NTNU 2D Guiding* dataset in the A2C and ALAX positions showed that the three recordings are similar, explaining the lower *OSR* results for the A2C and ALAX tilt models compared to the rotational and A4C tilt models. The lower *OSR* on the 2D data is also consistent with the results on the 3D slices (Figs. 4.9d and 4.9g) which have a slightly higher variability than the A4C tilt model (Fig. 4.9a). This suggests that tilt regression is more difficult to learn in the A2C and ALAX than for A4C views, although the data generation and training procedures are identical. Visual inspection of the tilt training data revealed that the slices in the A4C direction are potentially richer in features than in

the A2C and ALAX directions.

For the A4C tilt model, one can notice from Figs. 4.9b and 4.9c a consistent bias for the 2D A4C view compared to the A4C sliced in a 3D volume. Visual inspection of the training material suggested a bias in the reference due to the automatically generated LV axis. Indeed, we used the geometrical LV axis for the rotational axis during training data generation, whereas the axis used in practice by clinicians crosses the mitral plane slightly closer to the anterior wall. Nevertheless, such reference biases are not an issue for our method as they can be quantified from the validation dataset and compensated in post-processing for further inference.

For all DOFs, the position predictions on both validation slices and the *NTNU 2D Guiding* dataset did not significantly vary throughout the cardiac cycle, suggesting that the annotations made at the ED frame are valid for a complete cycle.

Using the US-specific augmentations improved or maintained the accuracy on the *OSR*. This supports the hypothesis that adding domain knowledge from US through data augmentation improves the robustness of the trained networks. Further, the blurring step applied at both training and inference seems to be beneficial for performance, suggesting that local image features are less relevant for the present task.

The CVCNet neural network topology was benchmarked against other topologies (benchmark available in the supplementary material). Neither smaller networks (MobileNet V2) nor larger networks (Inveption V3, ResNet50) provided significantly different results. Since carrying a clinical *superiority test* as proposed by Varoquaux et al. [19] is not realistic during early technical development, we assumed our method being topology agnostic and focused our efforts a careful preprocessing of the data and an in-depth evaluation on several external datasets to demonstrate robustness.

Although designed to quantify the guiding abilities of our models, the *OSR* metric is limited by the fact that we could not control the amount of positive and negative rotation or tilt introduced by the clinicians who acquired the *NTNU 2D Guiding* dataset. This, in addition to the small size of the *NTNU 2D Guiding* dataset, makes conclusions on improvements obtained with data augmentation and pre-processing difficult.

Despite limited quantitative results for our method due to the aforementioned, the Videos 4.1, 4.2, and 4.3 show evidence of the accuracy and robustness of our method by showcasing the association between the transducer movements, the US images and the predicted position.

In use, the main limitation of our guidance tool is the need for the correct intercostal point as a starting point. While this is achievable for experienced

users, positioning the transducer over the apex can be challenging for non-experts. Further clinical testing will identify the improvements required to make the application usable by most users. Future work will address the explainability of the method as this is required for the method to be adopted by the medical community. The same approach should be applicable to make a guiding tool for the parasternal long- and short-axis views. However, since 3D US is more often performed in the apical window, 3D data might be less available.

4.7 Conclusion

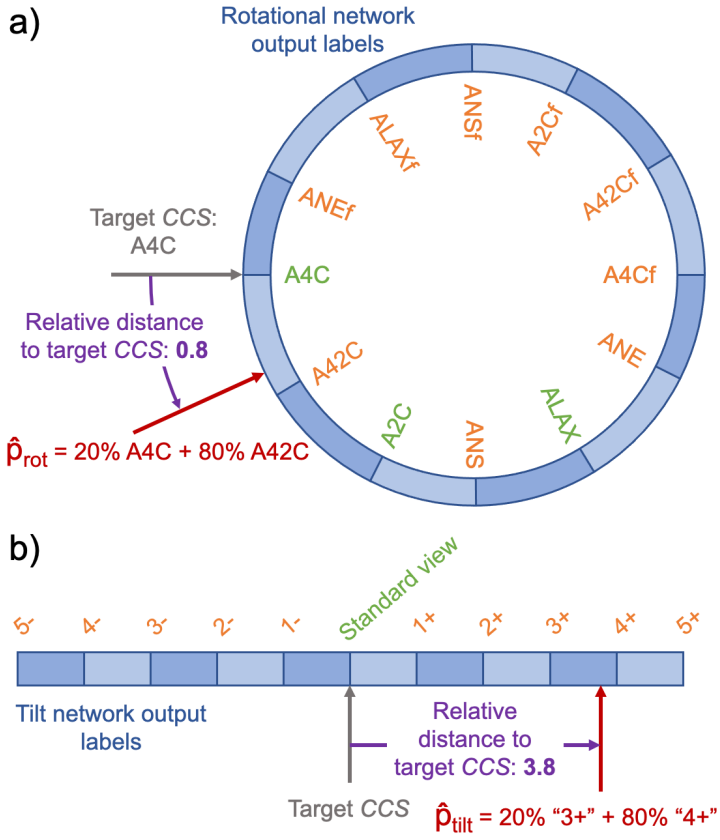
We have described a method to help ultrasound users acquire apical standard views of the heart. The backbone of the method is based on deep neural networks performing regression of the transducer position relative to the heart. The networks were trained on 2D slices from 3D US volumes, where reference positions were obtained using a semi-automated approach. Testing on multiple external datasets of 2D recordings showed that the method could detect suboptimal image planes and unveil individual operators' scanning habits. This suggests that our method is sufficiently robust and accurate to be of clinical value. A real-time application that supports inference of images streamed from a clinical US scanner and displays an intuitive feedback about view position was developed. Examples demonstrate the expected relation between the transducer movements, the ultrasound image and the machine learning calculated feedback. Further work will quantify the clinical value of the method inside and outside the echo lab, and map the potential benefit for expert and non-expert users.

4.8 Supplemental material

This supplemental material provides additional content from the study. Section 4.8.1 provides a graphical explanation of the *OSR* metric together with its formula for the rotational case. Section 4.8.2 provides a comparison of four different classifier topologies for the guiding task.

4.8.1 Metrics

Graphical explanation of the position calculation is presented in Supplemental Fig. 4.1.



Supplemental Figure 4.1: Position estimation calculation from networks output for the (a) rotation DOF and (b) the tilt DOF

For the rotational model, the predicted relative position $\hat{p}_{\text{rot}}$ accounts for the circular nature of the problem. Form the network output vector C_{rot} of

length N, the predicted position of any frame X can be expressed as:

$$\hat{p}_{\text{rot}}(X) = \frac{N}{2\pi} \arctan2\Big(\sum_{n=0}^{N-1} C_{\text{rot}}^{X}(n) \cdot \sin\big(\frac{n \cdot 2\pi}{N}\big), \sum_{n=0}^{N-1} C_{\text{rot}}^{X}(n) \cdot \cos\big(\frac{n \cdot 2\pi}{N}\big)\Big) \tag{4.6}$$

We can then define the categorical distance d to the target *CCS* by:

$$d(X, CCS) = \hat{p}_{\text{rot}}(X) - p_{\text{rot}}(CCS), \tag{4.7}$$

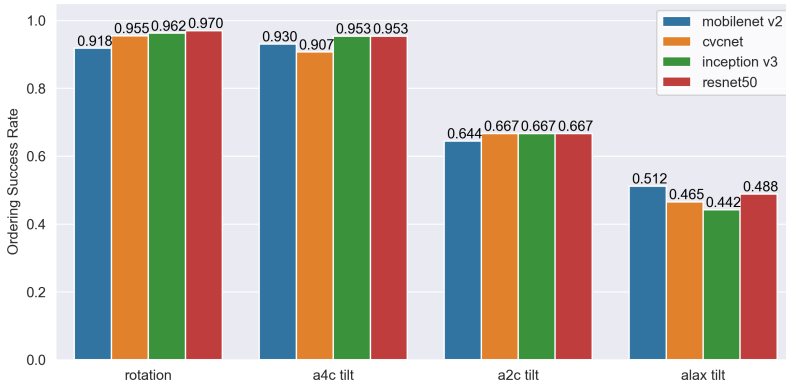
where $p(CS)$ here is the index of the target cross-sectional view in L_{rot}.

Finally, we account for the circular nature of the problem and define the minimal categorical distance d_{rot} to the target *CS* by:

$$d_{\text{rot}}(X, CCS) = \begin{cases} d(X, CCS) - N, & \text{if } d(X, CCS) > N/2 \\ d(X, CCS) + N, & \text{if } d(X, CCS) < -N/2 \\ d(X, CCS), & \text{otherwise} \end{cases} \tag{4.8}$$

4.8.2 Comparison of Neural Network topologies

We investigated the performance of four model topologies (MobileNet V2, CVCNet, Inception V3 and ResNet50) for the ordering accuracy for the non-standard views (Fig. 4.2). Main findings was that larger networks slightly improved the *OSR* for the rotational and A4C tilt models. However, none of the network topologies performed significantly better than others for the A2C and ALAX tilt DOFs. This suggests that the neural network topology is not a key factor for the method performance.



Supplemental Figure 4.2: Comparison of the best ordering score on 2D non standard-views for different network topologies

References

[1] R. M. Lang, L. P. Badano, M. A. Victor, J. Afilalo, A. Armstrong, L. Ernande, F. A. Flachskampf, E. Foster, S. A. Goldstein, T. Kuznetsova, P. Lancellotti, D. Muraru, M. H. Picard, E. R. Retzschel, L. Rudski, K. T. Spencer, W. Tsang, and J. U. Voigt, "Recommendations for cardiac chamber quantification by echocardiography in adults: An update from the American Society of Echocardiography and the European Association of Cardiovascular Imaging," *Journal of the American Society of Echocardiography*, vol. 28, no. 1, pp. 1–39.e14, 2015.

[2] C. Morbach, G. Gelbrich, M. Breunig, T. Tiffe, M. Wagner, P. U. Heuschmann, and S. Störk, "Impact of acquisition and interpretation on total inter-observer variability in echocardiography: results from the quality assurance program of the STAAB cohort study," *International Journal of Cardiovascular Imaging*, vol. 34, no. 7, pp. 1057–1065, 2018.

[3] K. Chykeyuk, M. Yaqub, and J. Alison Noble, "Class-specific regression random forest for accurate extraction of standard planes from 3D echocardiography," *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, vol. 8331 LNCS, no. Lv, pp. 53–62, 2014.

[4] J. S. Domingos, E. Lima, P. Leeson, and J. A. Noble, "Local phase-based fast ray features for automatic left ventricle apical view detection in 3d echocardiography," *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, vol. 8331 LNCS, pp. 119–129, 2014.

[5] E. Smistad, A. Østvik, I. M. Salte, D. Melichova, T. M. Nguyen, K. Haugaa, H. Brunvand, T. Edvardsen, S. Leclerc, O. Bernard, B. Grenne, and L. Løvstakken, "Real-Time Automatic Ejection Fraction and Foreshortening Detection Using Deep Learning," *IEEE Transactions on Ultrasonics, Ferroelectrics, and Frequency Control*, vol. 67, no. 12, pp. 2595–2604, 2020.

[6] S. R. Snare, H. Torp, F. Orderud, and B. O. Haugen, "Real-time scan assistant for echocardiography," *IEEE Transactions on Ultrasonics, Ferroelectrics, and Frequency Control*, vol. 59, no. 3, pp. 583–589, 2012.

[7] A. Østvik, E. Smistad, S. A. Aase, B. O. Haugen, and L. Lovstakken, "Real-Time Standard View Classification in Transthoracic Echocardiography Using

Convolutional Neural Networks," *Ultrasound in Medicine & Biology*, vol. 45, pp. 374–384, feb 2019.

[8] G. Toporek, R. Naidu, H. Xie, A. Simicich, T. Gades, and B. Raju, "User guidance for point-of-care echocardiography using multi-task deep neural network," *Med Image Comput Comput Assist Interv*, vol. 1, pp. 309–317, 2019.

[9] Y. Li, B. Khanal, B. Hou, A. Alansary, J. J. Cerrolaza, M. Sinclair, J. Matthew, C. Gupta, C. Knight, B. Kainz, and D. Rueckert, "Standard plane detection in 3D fetal ultrasound using an iterative transformation network," *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, vol. 11070 LNCS, pp. 392–400, 2018.

[10] R. Droste, L. Drukker, A. T. Papageorghiou, and J. A. Noble, "Automatic Probe Movement Guidance for Freehand Obstetric Ultrasound," *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, vol. 12263 LNCS, pp. 583–592, 2020.

[11] A. Narang, R. Bae, H. Hong, Y. Thomas, S. Surette, C. Cadieu, A. Chaudhry, R. P. Martin, P. M. McCarthy, D. S. Rubenson, S. Goldstein, S. H. Little, R. M. Lang, N. J. Weissman, and J. D. Thomas, "Utility of a Deep-Learning Algorithm to Guide Novices to Acquire Echocardiograms for Limited Diagnostic Use," *JAMA Cardiology*, vol. 6, no. 6, pp. 624–632, 2021.

[12] M. Schneider, P. Bartko, W. Geller, V. Dannenberg, A. König, C. Binder, G. Goliasch, C. Hengstenberg, and T. Binder, "A machine learning algorithm supports ultrasound-naïve novices in the acquisition of diagnostic echocardiography loops and provides accurate estimation of LVEF," *International Journal of Cardiovascular Imaging*, vol. 37, no. 2, pp. 577–586, 2021.

[13] E. Smistad, E. N. Steinsland, and L. Lovstakken, "Real-time 3D left ventricle segmentation and ejection fraction using deep learning," *IEEE International Ultrasonics Symposium*, pp. 2021–2023, 2021.

[14] L. Hou, C.-P. Yu, and D. Samaras, "Squared Earth Mover's Distance-based Loss for Training Deep Neural Networks," no. November, 2016.

[15] O. Bernard, B. Heyde, M. Alessandrini, D. Barbosa, S. Camarasu-pop, F. Cervenansky, S. Valette, E. Galli, M. Geleijnse, A. Papachristidis, G. Johan, and D. Jan, "Challenge on Endocardial Three-dimensional Ultrasound Segmentation (CETUS)," pp. 1–8, 2014.

[16] S. Leclerc, E. Smistad, J. Pedrosa, A. Ostvik, F. Cervenansky, F. Espinosa, T. Espeland, E. A. R. Berg, P. M. Jodoin, T. Grenier, C. Lartizien, J. Dhooge, L. Lovstakken, and O. Bernard, "Deep Learning for Segmentation Using an Open Large-Scale Dataset in 2D Echocardiography," *IEEE transactions on medical imaging*, vol. 38, no. 9, pp. 2198–2210, 2019.

[17] J. F. Grue, S. Storve, H. Dalen, Ø. Salvesen, O. C. Mjølstad, S. O. Samstad, H. Torp, and B. O. Haugen, "Automatic Measurements of Mitral Annular Plane Systolic Excursion and Velocities to Detect Left Ventricular Dysfunction," *Ultrasound in Medicine & Biology*, vol. 44, pp. 168–176, jan 2018.

[18] E. Smistad, A. Ostvik, and A. Pedersen, "High Performance Neural Network Inference, Streaming, and Visualization of Medical Images Using FAST," *IEEE Access*, vol. 7, pp. 136310–136321, 2019.

[19] G. Varoquaux and V. Cheplygina, "Machine learning for medical imaging: methodological failures and recommendations for the future," *npj Digital Medicine*, vol. 5, no. 1, 2022.

Real-time Guiding by Deep Learning of Experienced Operators to Improve Standardization of Echocardiographic Acquisitions

5

Sigbjørn Sæbø[1,3], David Pasdeloup[1], Håkon N. Pettersen[1], Erik Smistad[1,2], Andreas Østvik[1,2], Sindre H. Olaisen[1], Stian B. Stølen[1,3], Bjørnar L. Grenne[1,3], Espen Holte[1,3], Lasse Løvstakken[1], and Håvard Dalen[1,3]

[1] Dept. of Circulation and Medical Imaging, NTNU, Trondheim, Norway
[2] SINTEF Digital, Medical Image Analysis, Trondheim, Norway
[3] Clinic of Cardiology, St. Olav's Hospital, Trondheim, Norway

Background Impaired standardization of echocardiograms may increase inter-operator variability.

Objectives We aimed to study whether real-time guiding of experienced sonographers by deep learning (DL) could improve standardization of apical recordings.

Methods Patients (n=88) in sinus rhythm referred for echocardiography were included. All participants underwent three echocardiograms, whereof two were performed by sonographers, and the third by cardiologists. In the first study period (Period 1), the sonographers were instructed to provide echocardiograms for analyses of left ventricular (LV) function. Subsequently, after brief training, the DL guiding was used in Period 2 by the sonographer performing the second examination. View standardization was quantified retrospectively by a human expert as the primary endpoint and the DL algorithm as the secondary endpoint. All recordings were scored in rotation and tilt both separately and combined, and categorized as standardized or not.

Results Sonographers using DL guiding had more standardized acquisitions for the combination of rotation and tilt than sonographers without guiding in both periods (all $p<.05$) when evaluated by the human expert and DL (except for A2C by DL evaluation). When rotation and tilt were analyzed individually, A2C and ALAX rotation and A2C tilt were significantly

improved and the others were numerically improved when evaluated by the echocardiography expert and all, except for A2C rotation, were significantly improved when evaluated by DL ($p<.01$).

Conclusion Real-time guiding by DL improved standardization of echocardiographic acquisitions by experienced sonographers. Future studies should evaluate the impact with respect to improved reproducibility and when used by less experienced operators.

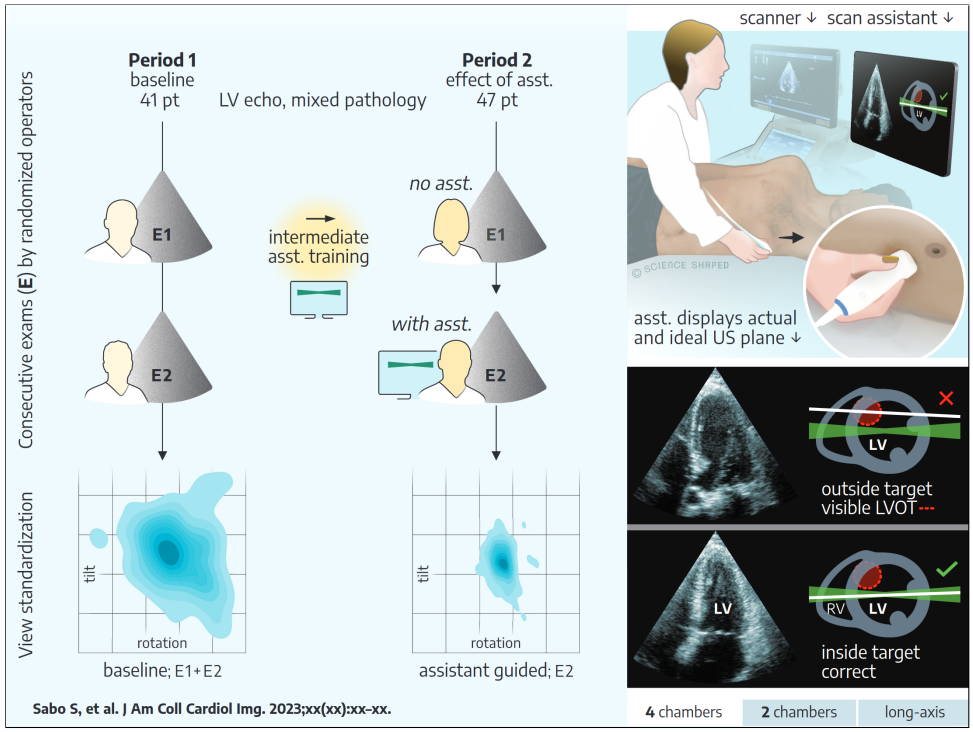
5.1 Introduction

Echocardiography provides essential information about cardiac morphology and function, making it a central tool for clinical decision-making in cardiology [1]. However, echocardiographic measurements have considerable operator variability, even in the hands of experienced operators [2, 3]. This reduces the sensitivity to detect subtle changes in cardiac function. Impaired view standardization is an important source of measurement variability, as standardized views form the basis of obtaining reliable analyses. It has been shown that the variability introduced by the recordings is equal to that of the analyses for most left ventricular function measurements [4–6]. Thus, expert consensus emphasizes the importance of reporting image quality and improving standardization of recordings for optimal diagnostics and treatment in cardiology [1, 7]. However, until now, evaluation of view standardization has been subjective and rarely quantified in the clinic and research. Novel ultrasound imaging analysis by deep learning (DL) techniques can process echocardiographic images in real-time [8,9]. Beyond measurements of cardiac size and function, such methods have shown the potential to improve view standardization of echocardiographic recordings during scanning [10, 11]. Thus, we aimed to study whether the use of a DL based scan assistant to guide experienced sonographers in real-time during scanning could optimize the three standard apical views to better comply with the current recommendations. This was first evaluated using retrospective assessment of view standardization by a human expert as reference. Secondly, the effect of the scan assistant was evaluated using retrospective DL standardization assessment as the reference.

5.2 Methods

5.2.1 Study population

Patients with mixed cardiac pathology were prospectively included from the St. Olavs University Hospital echocardiography laboratory (Figure 5.1). Inclusion criteria were 1) referral with indication for comprehensive echocardiography and 2) ability to provide written informed consent. Patients with non-sinus rhythm or indication for contrast echocardiography were excluded. Inclusion was restricted to predefined dates based on the availability of location and personnel. The study was approved by the regional ethical committee (REC Central Norway 2019/7160) and performed according to the Helsinki Declaration. All participants provided their written informed consent.



Central illustration: Study design, core results and the DL scan assistant
Left: characteristics of the two study periods and the repeated examinations of each patient by different sonographers. Plots depict spread of view standardization of apical views in rotation and tilt, as evaluated retrospectively by the DL algorithm. *Right:* Setup during patient examination and the scan assistant's feedback to cases of incorrect versus correct apical four-chamber view acquisitions.
DL: deep learning, US: ultrasound, LV: left ventricle, RV: right ventricle

5.2.2 Study design

All participants underwent three consecutive echocardiographic examinations without leaving the examination bench. The participants stayed in supine position between the examinations, and the time delay between examinations was minimized to the change of operators only. The first and second examinations were performed by two (of three) sonographers, and the third examination by one (of four) experts in echocardiography (all cardiologists). Operators were only present during their respective examinations and blinded to all information made available by the others. The selection of operators for the different examinations was random.

The data collection was divided into two separate periods (Figure 5.1).

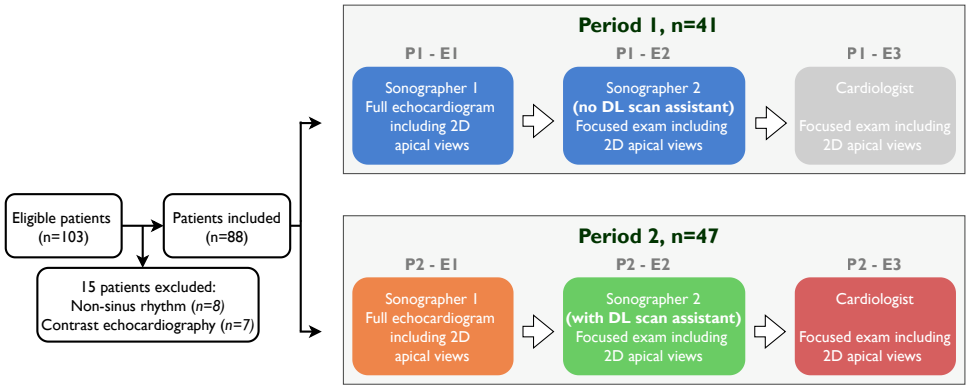


Figure 5.1: Flowchart of the study population
The same color coding for operator groups are also used in Figures 5.3 to 5.5 and Tables 5.2 and 5.3.
DL: deep learning, E1: Examination 1, E2: Examination 2, P1: Period 1, P2: Period 2

In the first period (Period 1), the sonographers were instructed to provide optimal echocardiograms for comparative analyses of left ventricular (LV) function. However, they were not explicitly informed about the aim of the study, the content of the training period, or the second data collection period. After Period 1 of data collection, the sonographers were introduced to the real-time DL scan assistant and trained on ten patients each. In the second period (Period 2), the sonographer performing the first examination (Sonographer 1) did this similarly as in Period 1, except for now being aware of the study aims and recently being trained in the use of the real-time DL scan assistant. The sonographer who performed the second examination (Sonographer 2) used the real-time DL scan assistant during scanning. The three sonographers participating in the study were randomly allocated to the role of Sonographer 1 or 2, ensuring that the sonographer examination with guiding was not performed by the same individual in all inclusions. This design was chosen to account for inter-operator differences in standardization of acquisitions.

5.2.3 Echocardiographic image acquisition

Examinations were performed with the patient in a left lateral decubitus position using a Vivid E95 scanner (version 202, GE Vingmed Ultrasound, Horten, Norway) with a phased array transducer (4Vc). For each recording, three consecutive cardiac cycles were obtained. The first examination was

a comprehensive echocardiogram according to ASE/EACVI guidelines [1]. The second and third exams were focused to the study aims and included among others two-dimensional (2D) B-mode (grayscale) apical two-chamber (A2C), four-chamber (A4C), and long-axis recordings (ALAX) for analyses of LV size and function. All study-specific recordings were blinded and de-identified before storing on the hospital's imaging server.

5.2.4 Real-time DL scanning assistant

The technical aspects of the real-time DL scan assistant used in this study has previously been described in detail [10]. It has two main components: a core DL algorithm automatically analyzing the images and a software application with a graphical user interface running the DL algorithm in real-time while interacting with the operator (Central Illustration). The core DL algorithm estimates the position of the 2D imaging plane relative to the heart for the rotation and tilt degrees of freedom (DOFs), where rotation indicates the rotational movement of the transducer, e.g., from A4C towards A2C or ALAX. Similarly, tilt indicates the tilting movement of the imaging plane away from the centerline, like anterior or posterior movement in A4C and lateral or septal movement in A2C. Both tilt and rotation are exemplified in Figure 5.2. As the heart anatomy varies from patient to patient, the position of each 2D image plane is calculated relative to the position of predefined and invariant 2D imaging planes, including the planes corresponding to three apical standard views (Supplemental Figure 1). Unlike traditional classification DL methods, which are only trained on standard views [8, 12, 13], this method is also trained using non-standard views to enhance spatial understanding. This approach allows to 1) estimate the deviation from the target standard view and 2) suggest transducer tilt and rotation movements to achieve the respective standard view. The core DL algorithm is implemented in the FAST framework [14], to form a complete real-time application that receives ultrasound images from a commercial ultrasound scanner and generates intuitive feedback on how to move the transducer to obtain the standardized view. A screen capture of the application is shown in the Central Illustration, and Video 5.1 shows an example of the method used in real-time, with both standardized and non-standardized apical views.

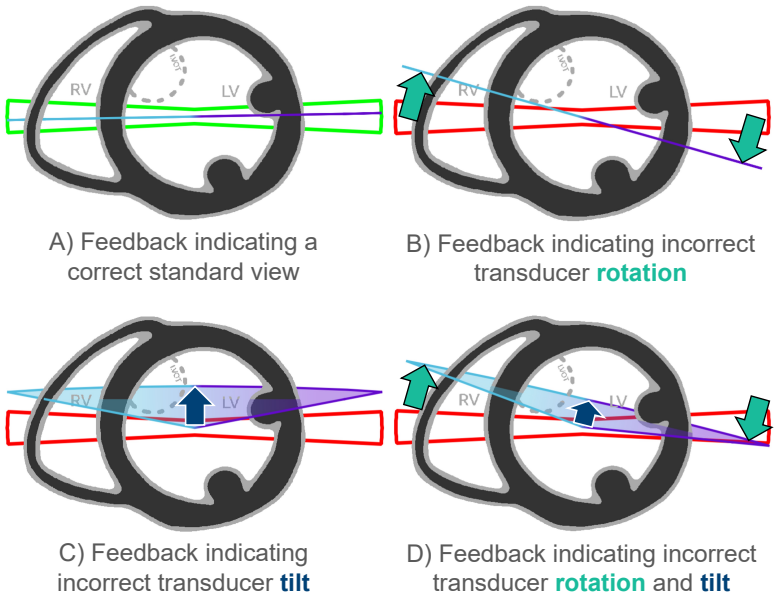


Figure 5.2: Rotation and tilt degrees of freedom in correct and incorrect apical four-chamber imaging planes
The imaging plane orientation is illustrated on a mid-ventricular parasternal short-axis cross-section.
LV: Left ventricle, LVOT: Left ventricular outflow tract, RV: Right ventricle

Video 5.1: Click or scan to watch online

5.2.5 Retrospective view standardization evaluation by a human expert

One echocardiography expert cardiologist (H. D.) manually evaluated view standardization in all recordings blinded to the study period, operator details, and whether the scan assistant was used or not. Rotation and tilt deviations were evaluated individually on a continuous scale from –3 to 3 in all three cycles for A2C, A4C, and ALAX recordings (Figure 5.3). A score of 0 indicated a perfect view, and negative or positive values reflected the direction and amount of deviations from optimal alignment of angle and tilt (Supplemental Table 5.1). Correct view standardization was arbitrarily predefined as a score between -0.5 and 0.5. The human expert who performed retrospective evaluation was not involved in the development of the scan assistant.

5.2.6 Retrospective view standardization evaluation by DL

All recordings were retrospectively analyzed by the core DL algorithm, which assigned them a relative tilt and rotational position on a continuous scale. We defined correctly standardized recordings for the automatic DL standardization evaluation to be within the center 80% along the relative positional axis of the cardiologists' recordings (Figure 5.3). The 80% position interval was arbitrarily predefined.

5.2.7 Statistics

Following the study aim, four null hypotheses were tested for evaluation of the effect of real-time DL guiding:

- *H0-A:* The proportion of standardized recordings is the same for sonographers before and after learning to use the scan assistant.
- *H0-B:* The proportion of standardized recordings is the same for sonographers actively using the scan assistant and those who have never used the scan assistant.
- *H0-C:* The proportion of standardized recordings is the same for sonographers who are actively using the scan assistant and those with training in, but not actively using the scan assistant.
- *H0-D:* The proportion of standardized recordings is the same for sonographers actively using the scan assistant and for cardiologists not using the scan assistant.

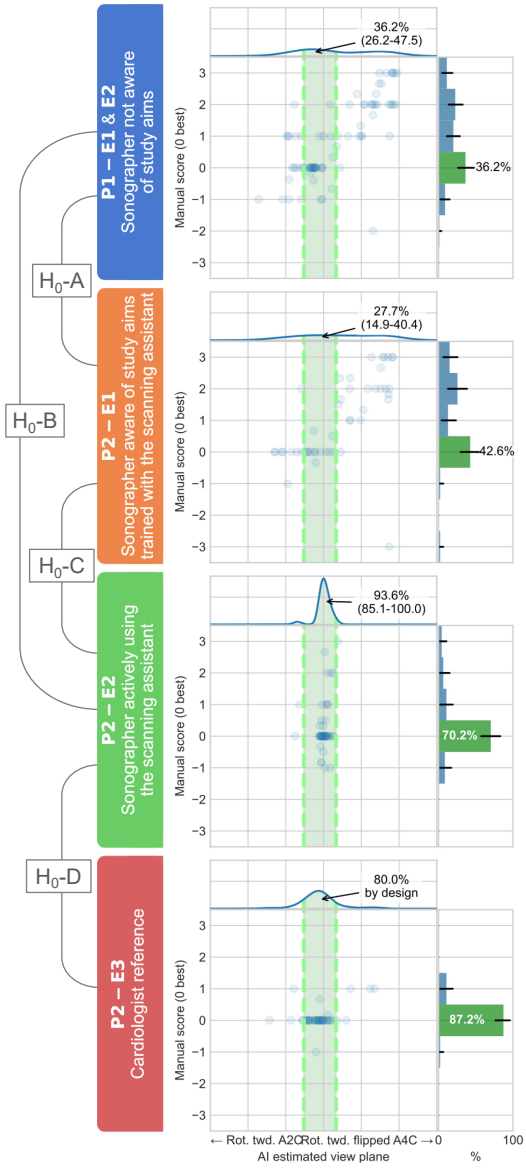


Figure 5.3: Visualization of the standardization assessment according to examinations, study periods and null hypotheses
Here exemplified by the rotational transducer position of apical long-axis view recordings. The distribution of DL algorithm scores is shown along the x-axes, while the distribution of expert's manual scores is shown along the y-axes. The color-coded left panel indicates operators, periods, whether the DL guiding algorithm was used or not, and relation to the four null hypotheses. The reference interval for the DL standardization evaluation was based on the center 80% of the cardiologists' distribution and is shown by the light green colored interval.
A4C: apical four chamber view, A2C: apical two chamber view, ALAX: apical long-axis view, E: Examination, H0-*X*: Null hypotheses tested, P: Inclusion period.

We investigated these null hypotheses according to the primary and secondary endpoint. The primary endpoint for hypotheses A-D was improvement in view standardization as evaluated in the human expert image analysis, and the secondary endpoint was similar but based on the DL image analysis. All hypotheses were tested in the rotational DOF, tilt DOF, and combination of rotation and tilt DOFs. Data obtained by the cardiologists in Period 1 were not included in these analyses. Statistical testing for differences in proportions between groups were calculated using the chi-squared tests. Bootstrapping (10 000 resamples) was used for both the manual and automatic analyses to estimate 95% confidence intervals (CI) for proportions of standardized recordings. All analyses were performed with Python 3.9 using Scipy 1.10.10. P-values <0.05 were considered statistically significant. Using SamplePower (version 3; IBM Statistics, NY, USA), inclusion of 38 individuals in each group provided 80% power to detect an increase in the proportion of standardized recordings from 50% to 80% (α 0.05) by use of the scan assistant. To account for potential drop-outs 15% more individuals were finally included.

5.3 Results

5.3.1 Study population

Of 103 recruited participants, those not in sinus rhythm (n=8) or with indication for contrast echocardiography (n=7) were excluded. Thus, a total of 88 patients (54% women) were included. Baseline characteristics of the study population and the key echocardiographic measurements are shown in Table 5.1. In short, the mean (SD) age was 61 (17) years, 34 (38%) had heart failure or previous myocardial infarction, while 16 (18%) had moderate or severe valvular disease. Only 3 (3%) had chronic obstructive pulmonary disease. Sonographers using the scan assistant used more time to obtain the three apical views, with a mean of 175.7 seconds per examination compared to 61.7 seconds in sonographers without assistance (p<.001).

5.3.2 Primary endpoint: Impact on view standardization evaluated by a human expert

The proportions of standardized recordings as evaluated by the human echocardiography expert are presented in Table 5.2 and visualized in Figure 5.4. For rotation and tilt combined, sonographers using the scan assistant had significantly more standardized recordings than sonographers

Table 5.1: Basic characteristics of the study population according to study periods.

	Period 1	Period 2
Included study participants	41	47
Age	62 (18)	61 (15)
Women	19 (45%)	29 (38%)
Clinical characteristics:		
Heart failure	8 (20%)	8 (17%)
Acute myocardial infarction	10 (24%)	8 (17%)
Moderate or severe valvular disease	7 (17%)	9 (19%)
Chronic obstructive pulmonary disease	2 (5%)	1 (2%)
Body mass index, kg/m²	26 (4)	26 (3)
Heart rate, beats per minute	66 (13)	69 (10)
Systolic blood pressure, mmHg	143 (25)	136 (23)
Echocardiographic characteristics		
Left ventricular ejection fraction, %	54 (13)	55 (11)
LV end-diastolic volume, biplane, ml	124 (53)	120 (49)

Values are presented as mean (SD) or numbers (proportions)

without assistance in Period 1 for all three apical views (p<.01). Similar results were also found for individual DOFs for A2C tilt, A2C rotation, and ALAX rotation (p<.01), and also A4C rotation and A4C tilt (p=.069 and p=.065, respectively). ALAX tilt was numerically improved, but the difference was not statistically significant (p=.192).

Compared to cardiologists, sonographers using the scan assistant were not significantly less standardized in A4C and A2C views for combined rotation and tilt (p=.239 and p=.182, respectively). In ALAX combined rotation and tilt, cardiologists were significantly more standardized than sonographers using the scan assistant (p<.05).

5.3.3 Secondary endpoint: Impact on view standardization evaluated by DL

The proportions of recordings with correct standardization as evaluated by the core DL algorithm are presented in Table 5.3. Figure 5.5 provides a qualitative visualization of the distribution of the estimated position

Table 5.2: Proportion of standardized acquisitions according to human expert evaluation

	Period 1	Period 2			p-values			
	1: P1-E1&E2	2: P2-E1	3: P2-E2	4: P2-E3	1 vs. 2 H0-A	1 vs. 3 H0-B	2 vs. 3 H0-C	3 vs. 4 H0-D
Rotation								
A4C	71.6 [61.7, 81.5]	74.5 [61.7, 87.2]	87.2 [76.6, 95.7]	83.0 [72.3, 93.6]	0.885$^{\Delta}$	0.069*	0.190*	0.772*§
A2C	56.8 [45.7, 67.9]	85.1 [74.5, 93.6]	85.1 [74.5, 93.6]	83.0 [72.3, 93.6]	<.005$^{\Delta}$	<.005*	1.000	1.000*§
ALAX	36.2 [26.2, 46.2]	42.6 [27.7, 57.4]	70.2 [57.4, 83.0]	87.2 [76.6, 95.7]	0.606$^{\Delta}$	<.001*	<.05*	0.078§
Tilt								
A4C	30.9 [21.0, 40.7]	40.4 [25.5, 55.3]	48.9 [34.0, 63.8]	76.6 [63.8, 87.2]	0.366$^{\Delta}$	0.065*	0.534*	<.05
A2C	71.6 [61.7, 81.5]	68.1 [53.2, 80.9]	100.0 [100.0, 100.0]	85.1 [74.5, 93.6]	0.826	<.001*	<.001*	<.05*§
ALAX	81.2 [72.5, 88.8]	91.5 [83.0, 97.9]	91.5 [83.0, 97.9]	97.9 [93.6, 100.0]	0.192$^{\Delta}$	0.192*	1.000	0.358§
Rotation and tilt								
A4C	12.0 [7.9, 16.2]	23.4 [16.3, 30.5]	44.7 [36.9, 53.2]	59.0 [50.4, 67.6]	0.153$^{\Delta}$	<.001*	0.050*	0.239§
A2C	39.1 [32.9, 45.3]	61.0 [53.2, 68.8]	85.1 [79.4, 90.8]	71.6 [63.8, 78.7]	<.05$^{\Delta}$	<.001*	<.05*	0.182*§
ALAX	35.1 [29.3, 41.0]	40.4 [32.6, 48.9]	63.8 [56.0, 71.6]	85.8 [80.1, 91.5]	0.686$^{\Delta}$	<.005*	<.05*	<.05

Data shown as proportion [95% confidence interval] % of correctly standardized views.
Δ: Proportion of standardized views numerically higher after sonographers got aware of study aims and trained with the scanning assistant
*: Proportion of standardized views numerically higher when using the scanning assistant
§: Sonographer recordings not significantly less standardized compared to cardiologists

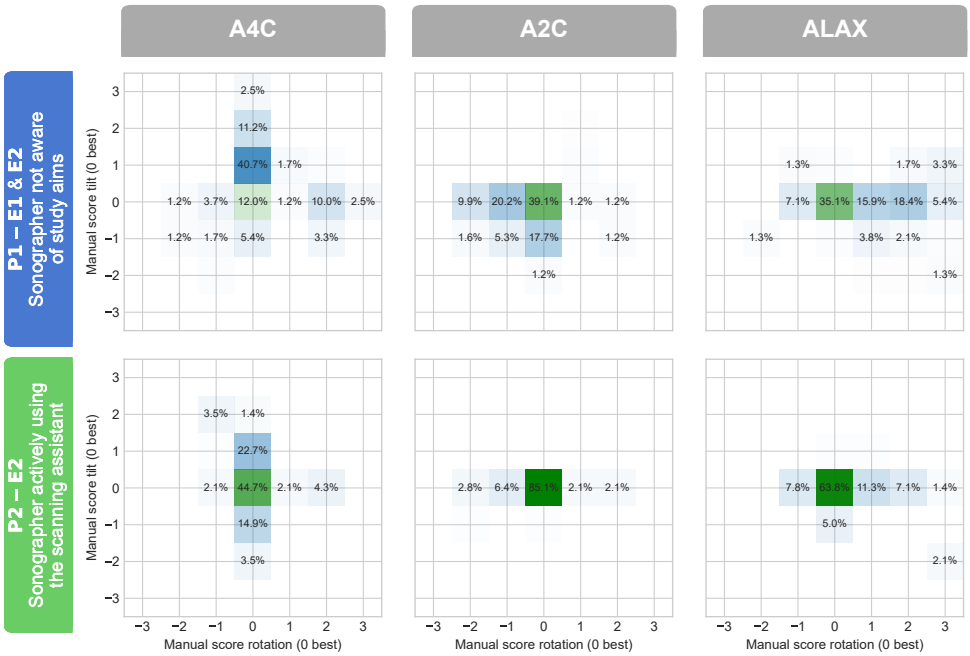


Figure 5.4: Blinded view standardization evaluation by an echocardiography expert
Rotation is presented along the x-axis and tilt along the y-axis. The percentages inside the boxes indicate the proportions of recordings with a given score.
A2C: Apical two-chamber view, A4C: Apical four-chamber view, ALAX: Apical long axis view, E1: Examination 1, E2: Examination 2, P1: Period 1, P2: Period 2.

along each DOF. For rotation and tilt combined, sonographers using the scan assistant in Period 2 significantly improved standardization of A4C and ALAX acquisitions compared to sonographers in Period 1 ($p<.001$). For A2C combined rotation and tilt, the difference was not statistically significant ($p=.936$). However, A2C tilt was significantly improved ($p<.001$), while A2C rotation had a significantly lower proportion of correct recordings ($p=.002$). For other views and DOFs individually, sonographers using the scan assistant significantly improved proportions of correct acquisitions compared to sonographers in Period 1 in A4C and ALAX rotation ($p<.001$) and all three views for tilt ($p<.05$). Figure 5.6 illustrates operator-specific distributions for the rotational DOF in ALAX recordings. Similar distributions for the other apical views and DOFs are included in Supplemental Figure 5.2. Compared to cardiologists, sonographers using the scanning assistant were not significantly less standardized in combined rotation and tilt for all the three apical standard views ($p>.099$)

Table 5.3: Proportion of standardized acquisitions according to the core DL algorithm analysis

	Period 1	Period 2			p-values			
	1: P1-E1&E2	2: P2-E1	3: P2-E2	4: P2-E3	1 vs. 2 H0-A	1 vs. 3 H0-B	2 vs. 3 H0-C	3 vs. 4 H0-D
Rotation								
A4C	66.7 [56.8, 76.5]	70.2 [57.4, 83.0]	97.9 [93.6, 100]	80.0 by design	0.827$^{\Delta}$	<.001*	<.001*	0.116*§
A2C	80.2 [71.6, 88.9]	72.3 [59.6, 85.1]	53.2 [38.3, 68.1]	80.0 by design	0.417	<.005	0.088	<.01
ALAX	36.2 [26.2, 47.5]	27.7 [14.9, 40.4]	93.6 [85.1, 100]	80.0 by design	0.425	<.001*	<.001*	0.073*§
Tilt								
A4C	81.5 [72.8, 90.1]	83.0 [72.3, 93.6]	95.7 [89.4, 100]	80.0 by design	1.000$^{\Delta}$	<.05*	0.094*	0.094*§
A2C	72.8 [63.0, 82.7]	78.7 [66.0, 89.4]	100.0 [100, 100]	80.0 by design	0.597$^{\Delta}$	<.001*	<.005*	<.005*§
ALAX	71.2 [61.3, 81.2]	66.0 [53.2, 78.7]	93.6 [85.1, 100]	80.0 by design	0.671	<.01*	<.005*	0.199*§
Rotation and tilt								
A4C	49.8 [43.6, 56.0]	50.4 [42.6, 58.9]	85.8 [80.1, 91.5]	80.0 by design	1.000$^{\Delta}$	<.001*	<.001*	0.217*§
A2C	54.3 [48.1, 60.9]	47.5 [39.0, 56.0]	56.7 [48.9, 64.5]	80.0 by design	0.576	0.936*	0.491*	0.944§
ALAX	30.5 [24.7, 36.4]	19.9 [13.5, 26.2]	85.8 [80.1, 91.5]	80.0 by design	0.268	<.001*	<.001*	0.099*§

Data shown as proportion [95% confidence interval] % of correctly standardized views.
Δ: Proportion of standardized views numerically higher after sonographers got aware of study aims and trained with the scanning assistant
*: Proportion of standardized views numerically higher when using the scanning assistant
§: Sonographer recordings not significantly less standardized compared to cardiologists

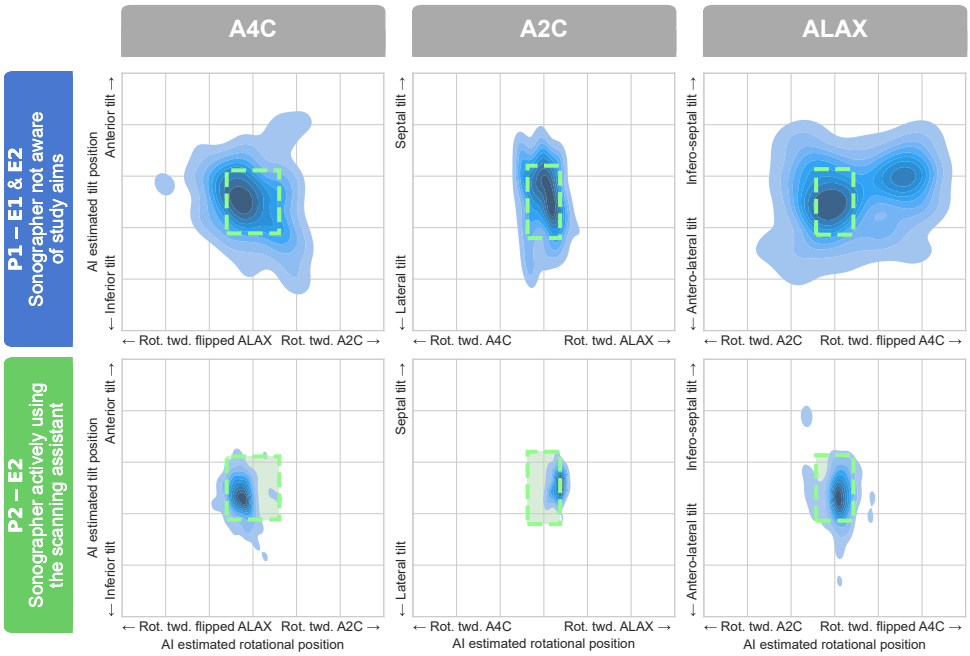


Figure 5.5: Retrospective automatic view standardization estimated by the DL core models
The green box indicates the standardization reference, as defined by the center 80% of view orientation in the cardiologist recordings.
A2C: Apical two-chamber view, A4C: Apical four-chamber view, ALAX: Apical long axis view, E1: Examination 1, E2: Examination 2, P1: Period 1, P2: Period 2, Rot. twd.: Rotation towards.

5.4 Discussion

This study investigated the effect of a real-time scanning assistant to optimize apical views standardization by experienced sonographers. The effect was quantitatively evaluated using retrospective human expert analysis and DL analysis. The main finding was that experienced sonographers using the scanning assistant significantly improved the proportion of standardized apical recordings.

5.4.1 Comparison to prior studies of real-time guiding

To the best of our knowledge, this is the first study to quantify standardization of recordings by experienced echocardiographers. Furthermore, it is the first study to show the distribution of view standardization and the

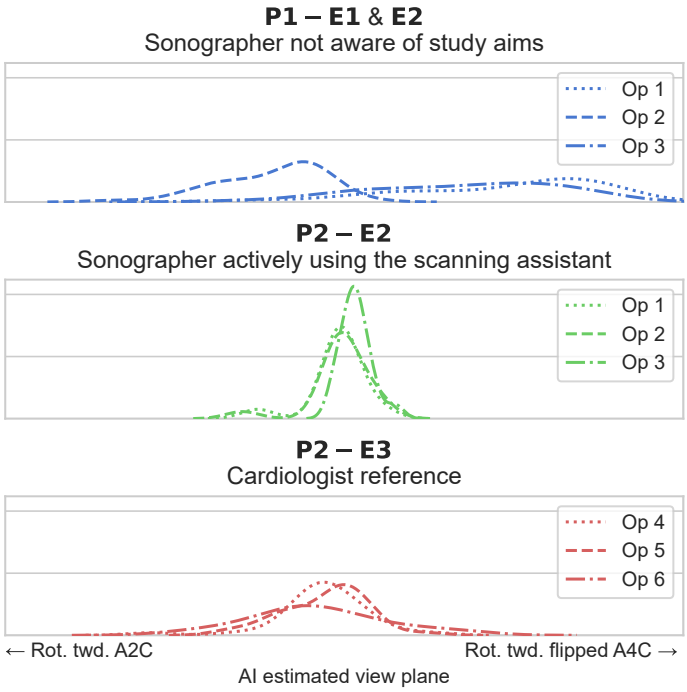


Figure 5.6: Operator specific distributions for relative view standardization with respect to transducer rotation in the apical long axis view
The green box indicates the standardization reference, as defined by the center 80% of view orientation in the cardiologist recordings.
A2C: Apical two-chamber view, A4C: Apical four-chamber view, E1: Examination 1, E2: Examination 2, E3: Examination 3, Op: Operator, P1: Period 1, P2: Period 2, Rot. twd.: Rotation towards.

potential benefits of real-time guiding of experienced echocardiographers to optimize standardization of apical views. Even though comparative studies are lacking, previous studies from our center and elsewhere do not show poorer standardization of recordings at our center [5, 15, 16], indicating a potential to improve the quality in echocardiographic laboratories across the world.

Some studies have presented DL-based software for real-time guiding of operators. Narang et al. [17] and Schneider et al. [18] assessed guiding of inexperienced operators (nurses or medical students without any prior echocardiographic experience). In comparison, the experienced sonographers in the presented study had more than five years of echocardiographic experience with a minimum of >2,000 comprehensive echocardiograms

performed each. In the studies by Narang and Schneider, machine learning-based metrics were used to evaluate image quality and shown as a single parameter displayed as a bar on the screen. In the presented study, the real-time scan assistant guided the operator by visualizing the view plane in real-time and intuitively suggested adjustments of the transducer position to obtain the well standardized imaging plane. Narang and Schneider also evaluated image quality retrospectively concerning its applicability for qualitative estimation of cardiac volumes and systolic parameters, whereas the retrospective human scoring in the presented study evaluated view standardization based on predefined parameters related to the presence of specific heart structures in the images. Thus, the previously published studies differed from the presented by evaluating more liberal endpoints in less experienced operators. Furthermore, Narang [17] and Schneider [18] did not randomize the operators to the use of guiding, but used an experienced sonographer without guiding as reference. Thus, these two studies did not fully evaluate the isolated effect of guiding itself, as performed in the presented study.

5.4.2 Improving view standardization by real-time guiding

For experienced sonographers using the scan assistant, the proportion of standardized recordings as evaluated by the blinded human expert (primary endpoint) was significantly improved for the combination of tilt and rotation in all three apical views compared to sonographers not using the assistant. Thus, the proposed real-time scan assistant contributed to better standardization. This finding adds to previous studies by showing the effect of the guiding procedure itself and by proving that it is possible to improve view standardization even among experienced operators. Importantly, this also strengthens the possibility of reducing test-retest variation in clinical echocardiography performed across high-standard echocardiographic laboratories. In line with the results from manual expert analyses, the retrospective DL analyses (secondary endpoint) showed that sonographers using the scan assistant had significantly higher proportions of well standardized acquisitions than those without the assistant in all views and DOFs, except for A2C rotation and A2C combined rotation and tilt. The finding of similar results using the DL analysis and the human expert as reference indicates the robustness of the DL method to assess image standardization. Furthermore, it concludes that the results were not provided by methodological concerns as the automatic retrospective DL image analysis used the same core algorithm as the

Video 5.2: Click or scan to watch online

real-time DL guiding scan assistant. As shown by the operator-specific distributions presented in Figure 6 and Supplemental Figures 2-6, the individual view-specific preferences in echocardiographic imaging planes were distinct in Period 1 where the scan assistant was not used. In contrast, the distributions were narrower and more consistent across the operators in Period 2 where the scan assistant was used. This effect was most pronounced for the rotational position of the transducer in the ALAX view.

For the A2C view, the DL standardization analysis revealed a reduced proportion of recordings with correct rotational alignment for sonographers using the scan assistant compared sonographers in Period 1 and cardiologists. At first glimpse, this could indicate that the scan assistant had a negative effect on standardization in the A2C view. However, as shown by the distributions across and within operator groups, the sonographers using the scan assistant were the most consistent in the rotational DOF. The choice of reference method used in this study, based on the central 80% of the cardiologists' distribution, could explain the reduced proportion of images classified as standardized. The cardiologists' A2C recordings were rotated more towards A4C compared to what was proposed by the DL algorithm as a correctly standardized A2C (Figure 5.5 and Video 5.2). This finding could indicate that optimizing alignment is equally important among experienced cardiologists and that the sonographers using the scan assistant obtained more standardized recordings than the reference by cardiologists.

5.4.3 Clinical implications

Slight variations in angulation and position of the ultrasound transducer may influence the recorded view, and thus, cause variations in measurements from repeated recordings [3]. The recommended strategy has been to use the same operator for patient follow-up to reduce test-retest

variability. However, ensuring the same operator during follow-up is logistically challenging in hospitals, and intra-operator variability is also significant [19]. Thus, ensuring optimal standardization of acquisitions is of the highest interest for echocardiographic laboratories worldwide and may improve test-retest variability of measurements and in-hospital workflow. The presented study shows that the proposed scan assistant has the potential to improve standardization of echocardiographic recordings across experienced users.

Diagnostic ultrasound by medical residents with basic echocardiography training has been shown to improve in-hospital diagnostic quality [20], and improving standardization of acquisitions by inexperienced operators may broaden the clinical benefits of echocardiography. Currently, the learning structure in ultrasound training is based on feedback from human mentors, which may have different subjective preferences. A consistent feedback like the one provided by the DL scan assistant could improve acquisition habits and aid in echocardiographic training. The impact of such training should be evaluated in future studies.

Large echocardiographic databases are commonly used for research purposes and training of novel DL tools [13,21]. The quality of such datasets has usually been evaluated without objective quantitative evaluation and assumed to be of sufficient quality by citing the source or addressing the experience of the operators. Automatic standardization assessment could provide a quantitative and objective insight on the quality of echocardiographic databases, and the consistent results between the human expert's and the DL algorithm's view standardization analyses in this study shows the robustness of the DL algorithm for this purpose. In a recent study, our group also showed the high validity of fully automatic quantification of LV foreshortening [22], another key parameter indicating the standardization of echocardiographic recordings. Several vendors and developers have created methods that automatically recognize the different echocardiographic views. Together, such methods form the basis for a standardized evaluation of the quality of echocardiographic databases used for research purposes. Hopefully, this may aid interpretation across studies in the future.

5.4.4 Strengths and limitations

The study has some noticeable strengths. Firstly, the operators in our study were experienced sonographers and cardiologists working in an echocardiography laboratory accredited by the European Association of Cardiovascular Imaging. As most echocardiograms are performed in well-qualified echocardiography laboratories, this adds to the generalizability of the study results. Secondly, the use of two different endpoints to evaluate view standardization and similar findings across these endpoints strengthens the results.

The study also has some limitations. Firstly, the DL scan assistant assumes correct positioning of the transducer at the apical LV point, indicating no LV foreshortening. LV foreshortening is a common problem in echocardiography [23], but a recent study from our research group showed that experienced sonographers had little foreshortening (2-5 mm) [22]. Nevertheless, future development of the DL scan assistant should also include foreshortening guidance to further optimize the anatomical orientation of apical views. Secondly, the study sample was modest, and including a larger population could have made minor differences between operator groups more evident. Only patients with non-sinus rhythm or need for contrast echocardiography were excluded, and inclusion was planned to restricted dates based on the availability of operators to serve both the clinical needs and align with the study methodology. Thus, despite the modest sample size, we believe the results are generalizable to populations commonly found in echocardiography laboratories.

The nature of the real-time scan assistant implies that learning to use the software increases attention to view standardization, which could have biased the results. To anticipate this issue, we performed inclusions in Period 1 before the sonographers were informed about the study aims and trained in using the scan assistant. Furthermore, sonographers were randomized at each patient inclusion to use the scan assistant in Period 2. This enabled us to differentiate the learning effect, the direct guiding effect, and the combined learning and guiding effect. However, accounting for the learning effect resulted in different patients being included across the two periods. Furthermore, a slight variation in the proportion of recordings for each sonographer between periods constitutes a limitation. Nevertheless, analyses within Period 2 (i.e., learning effect vs. direct guiding effect) showed significant improvements in standardization in sonographers with the scan assistant, indicating a positive effect of real-time guiding on top of the learning effect.

Furthermore, sonographers who used guiding used approximately two minutes more to obtain the three apical views. However, the sonographers were trained on just ten patients each, and we expect that the time spent using the scan assistant will be reduced with long-term use. Lastly, the thresholds of -0.5 to 0.5 for manual and 80% interval for DL analysis were arbitrarily predefined before analyzing the results, and other thresholds could have influenced the results.

5.4.5 Future perspectives

Future studies should evaluate the clinical impact of real-time guiding of operators at different levels of experience and patient populations in multicenter studies. Additionally, future studies should assess the effect of guiding combined with automatic measurements to reduce test-retest variability of echocardiographic measurements. With the refinement of DL methods able to both guide operators and quantify view standardization of stored data there is a need for a broad consensus on correct view standardization and how to report such findings. Ultimately, guiding of experienced operators in high quality echocardiography laboratories may improve standardization and the sensitivity to detect subtle changes in cardiac anatomy and function.

5.5 Conclusion

Sonographers using the proposed real-time DL scan assistant to optimize recordings of the three standard apical views had significantly higher proportions of standardized acquisitions than those not using the scan assistant. Our findings show that real-time guiding by DL can improve standardization of echocardiographic acquisitions when used by experienced personnel in high-quality echocardiographic laboratories. The impact on test-retest variability of measurements, the performance of the scan assistant in less experienced operators and the clinical impact must be addressed in future studies.

5.6 Clinical Perspectives

Clinical Competencies

Standardization of acquisitions is a core competency component in echocardiographic training. Real-time guiding of operators could lead to improved standardization, and possibly reduce test-retest variability of measurements. Furthermore, it could also aid in training of less experienced operators in echocardiography and improve quality control in echocardiography laboratories.

Translational Outlook

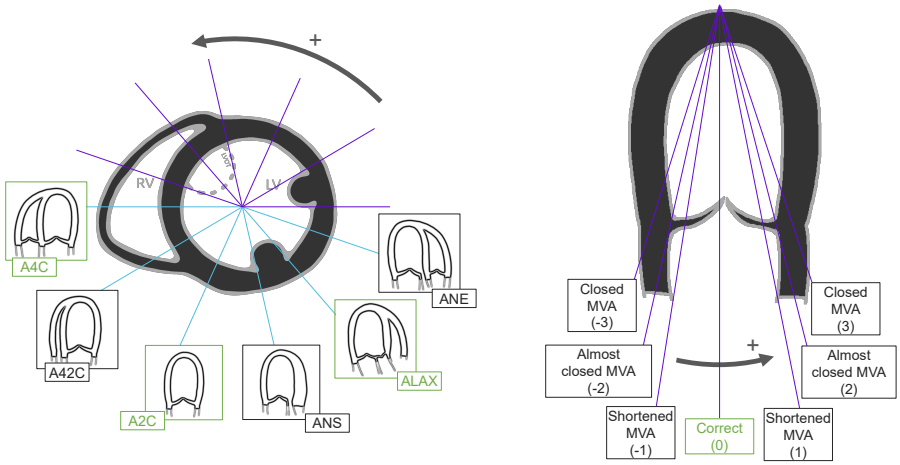
Guiding of experienced operators using DL could improve quality of recordings, optimize workflow and aid training in the future of echocardiography. Continuous inter-disciplinary innovations to improve accuracy of deep learning methods and multicenter studies are needed before full-scale clinical implementation.

Acknowledgments

We sincerely thank sonographers Gunn Catrin Fossum Bøen, Even Olav Jakobsen, and Per Kristian Langøy for their efforts in image acquisition and measurements.

5.7 Supplemental material

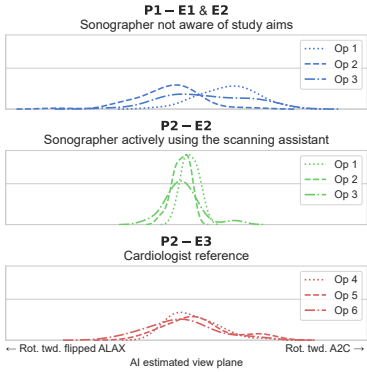
5.7.1 Supplemental Figures



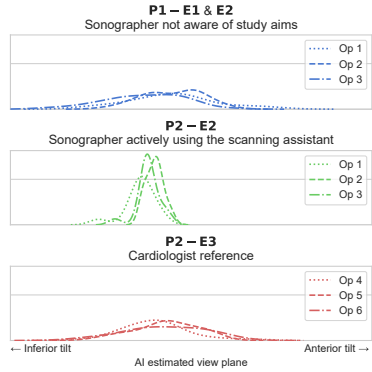
(a) Rotational DOF

(b) Tilt DOF

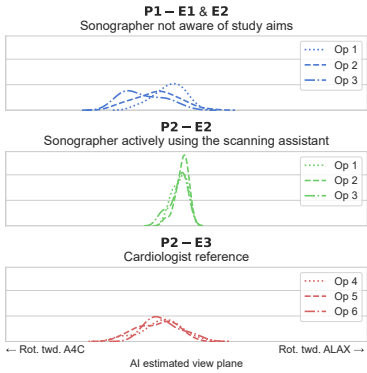
Supplemental Figure 5.1: Degree of rotation and tilt for the human expert scoring system
The scoring corresponds with the contents of Supplemental Table 5.1. The arrow marked with the + sign indicates the direction of the scoring scale.
A4C: Apical four-chamber, A2C: Apical two-chamber, A42C: rotational position where the right ventricle appears/disappears, ALAX: Apical long-axis, ANE: Aortic valve annulus end, ANS: Aortic valve annulus start, MVA: Mitral valve annulus



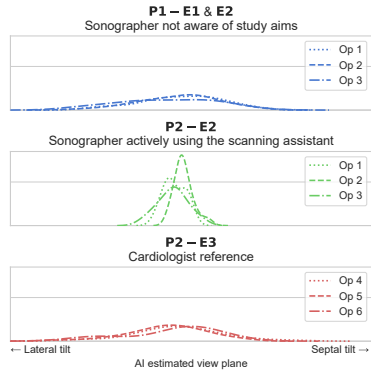
(a) A4C rotation



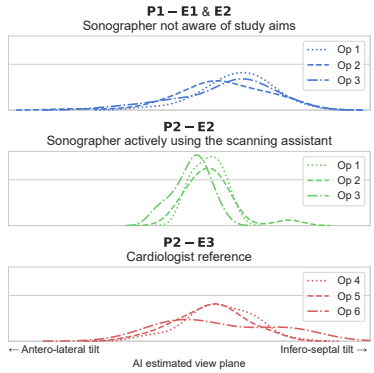
(b) A4C tilt



(c) A2C rotation

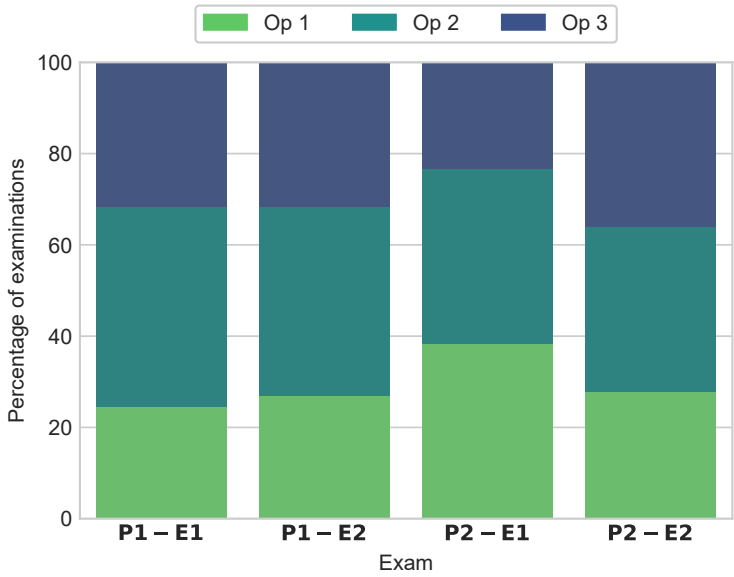


(d) A2C tilt



(e) ALAX tilt

Supplemental Figure 5.2: Operator-specific distributions for relative view standardization with respect to transducer rotation and tilt degrees of freedom
A2C: Apical two-chamber view, A4C: Apical four-chamber view, ALAX: Apical long axis view, E1: Examination 1, E2: Examination 2, E3: Examination 3, Op: Operator, P1: Period 1, P2: Period 2, Rot. twd: Rotation towards.



Supplemental Figure 5.3: Percentage of examinations per sonographer according to study period and examination
E1: Examination 1, E2: Examination 2, Op: Operator, P1: Period 1, P2: Period 2.

5.7.2 Supplemental Tables

Supplemental Table 5.1: Scoring system for the human expert evaluation of standardization

Rotational DOF		Tilt DOF	
Apical four-chamber view:			
3	> A42C	3	Apical five-chamber
2	A42C	2	Anteriorly, AV visible
1	< A42C	1	Anteriorly, LVOT visible
0	A4C	0	Correct tilt
-1	< ANE flipped	-1	Shortened MVA
-2	ANE flipped	-2	Almost closed MVA
-3	> ANE flipped	-3	Closed MVA
Apical two-chamber view:			
3	> ANS	3	Closed MVA, open AV
2	ANS	2	Almost closed MVA
1	< ANS	1	Shortened MVA/LVOT
0	A2C	0	Correct tilt
-1	< A42C	-1	Lateral, some anterior PM
-2	A42C	-2	More lateral, more PM
-3	> A42C	-3	Closed MVA
Apical long-axis view:			
3	> ANE	3	Closed MVA
2	ANE	2	Almost closed MVA
1	< ANE	1	Shortened MVA
0	ALAX	0	Correct tilt
-1	< ANS	-1	Some anterolateral, shortened AV
-2	ANS	-2	More anterolateral, no AV
-3	> ANS	-3	Closed MVA, no AV

The integers are landmarks, but the scale was continuous. The rotational and tilt direction and landmarks are illustrated in Supplemental Figure 5.1
A4C: Apical four-chamber, A2C: Apical two-chamber, A42C: rotational position where the right ventricle appears/disappears, ALAX: Apical long-axis, ANE: Aortic valve annulus end, ANS: Aortic valve annulus start, MVA: Mitral valve annulus

References

[1] R. M. Lang, L. P. Badano, M. A. Victor, J. Afilalo, A. Armstrong, L. Ernande, F. A. Flachskampf, E. Foster, S. A. Goldstein, T. Kuznetsova, P. Lancellotti, D. Muraru, M. H. Picard, E. R. Retzschel, L. Rudski, K. T. Spencer, W. Tsang, and J. U. Voigt, "Recommendations for cardiac chamber quantification by echocardiography in adults: An update from the American Society of Echocardiography and the European Association of Cardiovascular Imaging," *Journal of the American Society of Echocardiography*, vol. 28, no. 1, pp. 1–39.e14, 2015.

[2] A. Thorstensen, H. Dalen, B. H. Amundsen, S. A. Aase, and A. Stoylen, "Reproducibility in echocardiographic assessment of the left ventricular global and regional function, the HUNT study," *European Journal of Echocardiography*, vol. 11, no. 2, pp. 149–156, 2010.

[3] J. E. Otterstad, G. Froeland, M. St John Sutton, and I. Holme, "Accuracy and reproducibility of biplane two-dimensional echocardiographic measurements of left ventricular dimensions and function," *European Heart Journal*, vol. 18, no. 3, pp. 507–513, 1997.

[4] T. Baron, L. Berglund, E. M. Hedin, and F. A. Flachskampf, "Test–retest reliability of new and conventional echocardiographic parameters of left ventricular systolic function," *Clinical Research in Cardiology*, vol. 108, no. 4, pp. 355–365, 2019.

[5] J. M. Letnes, T. Eriksen-Volnes, B. Nes, U. Wisløff, Ø. Salvesen, and H. Dalen, "Variability of echocardiographic measures of left ventricular diastolic function. The HUNT study," *Echocardiography*, vol. 38, no. 6, pp. 901–908, 2021.

[6] C. Morbach, G. Gelbrich, M. Breunig, T. Tiffe, M. Wagner, P. U. Heuschmann, and S. Störk, "Impact of acquisition and interpretation on total inter-observer variability in echocardiography: results from the quality assurance program of the STAAB cohort study," *International Journal of Cardiovascular Imaging*, vol. 34, no. 7, pp. 1057–1065, 2018.

[7] M. Galderisi, B. Cosyns, T. Edvardsen, N. Cardim, V. Delgado, G. Di Salvo, E. Donal, L. E. Sade, L. Ernande, M. Garbi, J. Grapsa, A. Hagendorff, O. Kamp, J. Magne, C. Santoro, A. Stefanidis, P. Lancellotti, B. Popescu, G. Habib, F. A.

Flachskampf, B. Gerber, A. Gimelli, and K. Haugaa, "Standardization of adult transthoracic echocardiography reporting in agreement with recent chamber quantification, diastolic function, and heart valve disease recommendations: An expert consensus document of the European Association of Cardiovascular Imag," *European Heart Journal Cardiovascular Imaging*, vol. 18, no. 12, pp. 1301–1310, 2017.

[8] A. Østvik, E. Smistad, S. A. Aase, B. O. Haugen, and L. Lovstakken, "Real-Time Standard View Classification in Transthoracic Echocardiography Using Convolutional Neural Networks," *Ultrasound in Medicine and Biology*, vol. 45, pp. 374–384, feb 2019.

[9] R. Varudo, F. A. Gonzalez, J. Leote, C. Martins, J. Bacariza, A. Fernandes, and F. Michard, "Machine learning for the real-time assessment of left ventricular ejection fraction in critically ill patients: a bedside evaluation by novices and experts in echocardiography," *Critical Care*, vol. 26, no. 1, pp. 1–5, 2022.

[10] D. Pasdeloup, S. H. Olaisen, A. Østvik, S. Sabo, H. N. Pettersen, E. Holte, B. Grenne, S. B. Stølen, E. Smistad, S. A. Aase, H. Dalen, and L. Løvstakken, "Real-Time Echocardiography Guidance for Optimized Apical Standard Views," *Ultrasound in Medicine and Biology*, vol. 49, no. 1, pp. 333–346, 2023.

[11] E. Smistad, A. Østvik, I. M. Salte, D. Melichova, T. M. Nguyen, K. Haugaa, H. Brunvand, T. Edvardsen, S. Leclerc, O. Bernard, B. Grenne, and L. Løvstakken, "Real-Time Automatic Ejection Fraction and Foreshortening Detection Using Deep Learning," *IEEE Transactions on Ultrasonics, Ferroelectrics, and Frequency Control*, vol. 67, no. 12, pp. 2595–2604, 2020.

[12] A. Madani, R. Arnaout, M. Mofrad, and R. Arnaout, "Fast and accurate view classification of echocardiograms using deep learning," *npj Digital Medicine*, vol. 1, no. 1, pp. 1–8, 2018.

[13] J. Zhang, S. Gajjala, P. Agrawal, G. H. Tison, L. A. Hallock, L. Beussink-Nelson, M. H. Lassen, E. Fan, M. A. Aras, C. Jordan, K. E. Fleischmann, M. Melisko, A. Qasim, S. J. Shah, R. Bajcsy, and R. C. Deo, "Fully Automated Echocardiogram Interpretation in Clinical Practice," *Circulation*, vol. 138, pp. 1623–1635, oct 2018.

[14] E. Smistad, A. Ostvik, and A. Pedersen, "High Performance Neural Network Inference, Streaming, and Visualization of Medical Images Using FAST," *IEEE Access*, vol. 7, pp. 136310–136321, 2019.

[15] I. M. Salte, A. Østvik, S. H. Olaisen, S. Karlsen, T. Dahlslett, E. Smistad, T. K. Eriksen-Volnes, H. Brunvand, K. H. Haugaa, T. Edvardsen, H. Dalen, L. Lovstakken, and B. Grenne, "Deep Learning for Improved Precision and Reproducibility of Left Ventricular Strain in Echocardiography: A Test-Retest Study," *Journal of the American Society of Echocardiography*, pp. 1–12, 2023.

[16] T. Eriksen-Volnes, J. F. Grue, S. Hellum Olaisen, J. M. Letnes, B. Nes, L. Løvstakken, U. Wisløff, and H. Dalen, "Normalized Echocardiographic Values From Guideline-Directed Dedicated Views for Cardiac Dimensions and Left Ventricular Function," *JACC: Cardiovascular Imaging*, vol. lic, 2023.

[17] A. Narang, R. Bae, H. Hong, Y. Thomas, S. Surette, C. Cadieu, A. Chaudhry, R. P. Martin, P. M. McCarthy, D. S. Rubenson, S. Goldstein, S. H. Little, R. M. Lang, N. J. Weissman, and J. D. Thomas, "Utility of a Deep-Learning Algorithm to Guide Novices to Acquire Echocardiograms for Limited Diagnostic Use," *JAMA Cardiology*, vol. 6, no. 6, pp. 624–632, 2021.

[18] M. Schneider, P. Bartko, W. Geller, V. Dannenberg, A. König, C. Binder, G. Goliasch, C. Hengstenberg, and T. Binder, "A machine learning algorithm supports ultrasound-naïve novices in the acquisition of diagnostic echocardiography loops and provides accurate estimation of LVEF," *International Journal of Cardiovascular Imaging*, vol. 37, no. 2, pp. 577–586, 2021.

[19] D. Medvedofsky, K. Kebed, L. Laffin, J. Stone, K. Addetia, R. M. Lang, and V. Mor-Avi, "Reproducibility and experience dependence of echocardiographic indices of left ventricular function: Side-by-side comparison of global longitudinal strain and ejection fraction," *Echocardiography*, vol. 34, no. 3, pp. 365–370, 2017.

[20] O. C. Mjølstad, G. N. Andersen, H. Dalen, T. Graven, K. Skjetne, J. O. Kleinau, and B. O. Haugen, "Feasibility and reliability of point-of-care pocket-size echocardiography performed by medical residents.," *European heart journal cardiovascular Imaging*, vol. 14, no. 12, pp. 1195–1202, 2013.

[21] F. M. Asch, V. Mor-Avi, D. Rubenson, S. Goldstein, M. Saric, I. Mikati, S. Surette, A. Chaudhry, N. Poilvert, H. Hong, R. Horowitz, D. Park, J. L. Diaz-Gomez, B. Boesch, S. Nikravan, R. B. Liu, C. Philips, J. D. Thomas, R. P. Martin, and R. M. Lang, "Deep Learning-Based Automated Echocardiographic Quantification of Left Ventricular Ejection Fraction: A Point-of-Care Solution," *Circulation: Cardiovascular Imaging*, vol. 14, no. 6, p. E012293, 2021.

[22] S. Sabo, H. N. Pettersen, E. Smistad, D. Pasdeloup, S. B. Stølen, B. L. Grenne, L. Lovstakken, E. Holte, and H. Dalen, "Real-time guiding by deep learning during echocardiography to reduce left ventricular foreshortening and measurement variability," *European Heart Journal - Imaging Methods and Practice*, vol. 1, no. 1, pp. 1–9, 2023.

[23] S. Ünlü, J. Duchenne, O. Mirea, E. D. Pagourelias, S. Bézy, M. Cvijic, A. S. Beela, J. D. Thomas, L. P. Badano, J. U. Voigt, J. D. Thomas, J. Hamilton, S. Pedri, P. Lysyansky, G. Hansen, Y. Ito, T. Chono, J. Vogel, D. Prater, J. H. Song, J. Y. Lee, H. Houle, B. Georgescu, R. Baumann, B. Mumm, Y. Abe, and W. Gorissen, "Impact of apical foreshortening on deformation measurements: A report from the EACVI-ASE Strain Standardization Task Force," *European Heart Journal Cardiovascular Imaging*, vol. 21, no. 3, pp. 337–343, 2020.

ISBN 978-82-326-7480-0 (printed ver.)
ISBN 978-82-326-7479-4 (electronic ver.)
ISSN 1503-8181 (printed ver.)
ISSN 2703-8084 (online ver.)

NTNU
Norwegian University of
Science and Technology