Doctoral theses at NTNU, 2021:230

Andreas Østvik

Automatic analysis in echocardiography using machine learning

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

Andreas Østvik

Automatic analysis in echocardiography using machine learning

Thesis for the Degree of Philosophiae Doctor

Trondheim, June 2021

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 978-82-326-5899-2 (printed ver.)
ISBN 978-82-326-5982-1 (electronic ver.)
ISSN 1503-8181 (printed ver.)
ISSN 2703-8084 (online ver.)

Doctoral theses at NTNU, 2021:230

Printed by NTNU Grafisk senter

Automatisering av analyser i ekkokardiografi ved hjelp av maskinlæring

Ekkokardiografi er hjørnesteinen i moderne hjerteavbildning på grunn av tilgjengelighet, lave kostnader og sanntidsfunksjonalitet. Modaliteten har muliggjort sofistikerte og ikke-invasive vurderinger av hjertets morfofysiologi, med et bredt spekter av kliniske parametere med høy diagnostisk og prognostisk verdi. Til tross for klinisk innvirkning blir kvantitative målinger ofte utelatt i praksis fordi de er ressurskrevende og vanskelig å reprodusere. Automatisering kan redusere noen av disse begrensningene og omdefinere deler av den kliniske arbeidsflyten, men utformingen av generiske algoritmer er utfordrende på grunn av den iboende variasjonen i ekkokardiografidata og ekspertisen som kreves for tolkning.

Det overordnede målet med dette arbeidet var å undersøke bruk av *deep learning* metodikk for å helautomatisere flere trinn av bildeanalysen i en standard ekkokardiografi. Metodetilpasning for ultralyd ble vektlagt, samt adressering av grunnleggende domenebegrensning som støy og opptaksvariabilitet. Sanntidsstøtte og forbedring av arbeidsflyten var også viktige aspekter i utviklingen. I den første delen av avhandlingen presenteres metoder for automatisk klassifisering av hjertesnitt og deteksjon av hjertehendelser direkte fra ultralydbildene ved hjelp av kunstige nevrale nettverk. Videre presenteres en metode for estimering av hjertemuskelens bevegelse, samt integrasjonen av flere *deep learning* komponenter i en kaskade for helautomatiserte målinger av hjertemuskelens deformasjon. Den siste delen av avhandlingen omhandler en mulighetsstudie som sammenligner overnevnte metoder med en kommersielt tilgjengelig løsning.

Resultatene indikerer at de forskjellige komponentene i en ekkokardiografisk bildeanalyse kan være fordelaktig eller til og med forbedres ved bruk av *deep learning*. Fleksibiliteten av en læringsbasert tilnærming bidrar til å overgå konvensjonelle metoder på kjente begrensninger ved bruk av ultralyd. Integrasjonen av komponentene i en kaskade for helautomatiserte målinger var mulig, og ga oppmuntrende resultater ved å være sammenlignbar med variabiliteten mellom forskjellige kommersielle produsenter. Til tross for flere begrensninger, kan vi være optimistiske for fremtidig bruk av *deep learning* i ekkokardiografi.

Andreas Østvik
Institutt for sirkulasjon og bildediagnostikk, NTNU
Hovedveileder: Lasse Løvstakken
Biveileder: Erik Smistad
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 blir avviklet digitalt via Zoom, onsdag 16. Juni 2021 kl 12.15.

Abstract

Echocardiography is the cornerstone of modern cardiac imaging due to its availability, low cost and real-time functionality. The modality has enabled sophisticated non-invasive evaluation of the hearts morphophysiology, with a wide range of clinical parameters of high diagnostic and prognostic value. However, despite the clinical impact, quantitative measurements are often omitted in clinical practice by being labor intensive, time consuming and difficult to reproduce. Automation can reduce some of these limitations and redefine parts of the clinical workflow, but the design of generic algorithms is complex due to the inherent variability of echocardiography data and the expertise required for interpretation.

The overall goal of this work was to investigate the use of deep learning (DL) methods for fully automating several image analysis steps of an echocardiography exam. Emphasis was given to method adaptation for ultrasound (US) image processing, as well as addressing fundamental domain limitations such as noise and acquisition variability. Real-time support and workflow enhancements was also important features in the development. The thesis consists of three technical contributions and one clinical feasibility study. In the first part, a method for cardiac view classification with convolutional neural networks (CNNs) is presented. Further, we describe a recurrent CNN method for cardiac event detection. The third part presents a DL based motion estimator, and the integration of several DL components into a pipeline for automated longitudinal strain (LS) measurements. The last part is dedicated to a feasibility study comparing the latter with a commercially available solution.

Results indicate that the different components can benefit or even be improved with DL. The flexibility of learning-based approaches helps to surpass conventional methods on inherent limitations of US. Integrating DL components in a pipeline for fully automated measurements was feasible, and yielded encouraging results by being comparable to intervendor variability. Despite several limitations described in the thesis, we can be optimistic about the future employment of DL in echocardiography.

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-supervisor has been Erik Smistad, both from ISB, NTNU.

Acknowledgements

I want to express my sincere gratitude to everyone who contributed and supported me during the course of this PhD work. First of all, I would like to thank my main supervisor Lasse, for giving me this chance and believing in me. Your motivating encouragements, guidance, and profound experience have been vital to this work and its completion. I also wish to thank my co-supervisor Erik, for continuous support and valuable feedback. Your programming skills and efficiency have been an inspiration.

During my PhD I have been fortunate to work alongside fantastic colleagues from both the ultrasound group at ISB and the medical technology group at SINTEF. I am proud to be part of a great machine learning team, which has grown exponentially in recent years. Thank you for all the memorable moments, achievements and inspiring discussions, both at and outside of work. My co-authors deserve my gratitude, especially Adrian and Ivar who was the lead authors on two of the papers presented herein. Bjørnar for all the clinical insight and our fruitful collaboration. Special thanks goes to my dear friends Thomas, Stefano and Cristiana, whom I shared countless hours with through these years. Your presence has truly been invaluable to my well-being.

I would like to thank my family for their unconditional love and support. My parents, grandparents and sisters who always manage to detach me from work and remind me about other important things in life. Odd-Harald and Brit for always being there, and for being such wonderful grandparents to Anna. Finally, my dearest Oda Cathrine and Anna, I thank you for giving me an inexhaustible source of motivation. I am so grateful to have you in my life, and I am immensely thankful for the patience, love and encouragements through these years.

Table of Contents

Abbreviations and nomenclature

2D	Two-dimensional
3D	Three-dimensional
AI	Artificial intelligence
AV	Atrioventricular
CNN	Convolutional neural network
CV	Computer vision
DL	Deep learning
ECG	Electrocardiogram
ED	End-systole
EF	Ejection fraction
GLS	Global longitudinal strain
LV	Left ventricle
LVEF	Left ventricular ejection fraction
ME	Motion estimation
ML	Machine learning
ROI	Region of interest
SL	Semilunar
STE	Speckle tracking echocardiography
TDI	Tissue Doppler imaging
TTE	Transthoracic echocardiography
US	Ultrasound

Introduction

1

Through a comprehensive evolution from the early demonstrations by Edler and Hertz in 1953 [1] to the rich featured modality it is today, echocardiography remains the cornerstone of modern cardiac imaging. The combination of availability, low cost, portability and real-time functionality, makes it the most commonly used non-invasive tool in clinical cardiology [2]. Echocardiography has enabled possibilities for advanced quantification of the hearts morphophysiology, with clinical parameters such as left ventricular ejection fraction (LVEF), left atrial volume and global longitudinal strain. Today, several of these measurements are used in everyday routine by being readily available in commercial systems and included in the guidelines for cardiac chamber quantification. Undoubtedly this has significantly improved patient care and the assessment of the cardiovascular system.

Despite the impact, the introduction of quantitative measurements has not been without drawbacks [3]. They often require manual labor, increased time for examinations and reports, potentially delaying the diagnosis. Time constraints in the busy clinic can also affect the accuracy and variability of manual measurements. Incorporating automated measurements can potentially redefine the workflow in echocardiography laboratories, with potential benefits including time and cost savings, improved reproducibility, as well as streamlined acquisitions and reporting. However, quantitative echocardiography is complex, and the tacit knowledge of the operator is still necessary for extracting useful and accurate parameters from the acquisitions.

Today, we witness a paradigm shift in computer vision (CV) with modern *machine learning* (ML) algorithms, more specifically within the field of *deep learning* (DL). These techniques have surpassed human performance in a variety of problems, such as labeling images, mastering games and

classifying skin disease [4–6]. What is even more unique, is that DL algorithms have not only improved the accuracy on certain tasks, but also the time required to complete them [7]. This makes them more applicable for real-life deployment. DL methods have also been applied to a broad range of ultrasound (US) related tasks with success [8, 9]. From raw signal processing, advanced filtering and image formation to post-processing and image analysis tools. All have thrived as a result of progress in CV research, improved hardware and access to digital data, but analysis of display images from the traditional systems are so far the most common application of DL methods in US.

1.1 Automated quantification in echocardiography

Quantitative echocardiography involves the derivation of comprehensible measurements associated with an US recording, and is pivotal when evaluating the cardiac function. Usually this requires some form of software post-processing of the formed US image, but visual assessment, so-called *eyeballing*, is still being used extensively in clinical routine. This even applies for estimating clinical parameters such as LVEF. Eyeballing clinical parameters is not recommended due to its inherent subjectivity and high variability, but limitation in the alternatives still makes it a viable option [10]. Time is at a premium in echocardiographic laboratories, and manual measurements may be under-prioritized or not considered reliable. With todays technology, especially facilitated by DL, automated measurements can become as fast as eyeballing and as accurate as experts.

There are several ways to group quantitative methods. If the operator needs to interact partially with the software to produce the results, for instance to define anatomical landmarks, region of interest (ROI) or location of the image plane, we refer to the method as "semi automated". "Fully automated" algorithms can be performed end-to-end without any interaction from the user. Semi automated methods have better reproducibility compared to manual measurements, while fully automated measurements have zero variability for the same US images [10].

A typical software pipeline for quantification of a clinical measurement can be divided into a cascade of different tasks, as presented in Fig. 1.1. The composition would vary according to desired measurement, but DL can play

a key role in most of the steps.

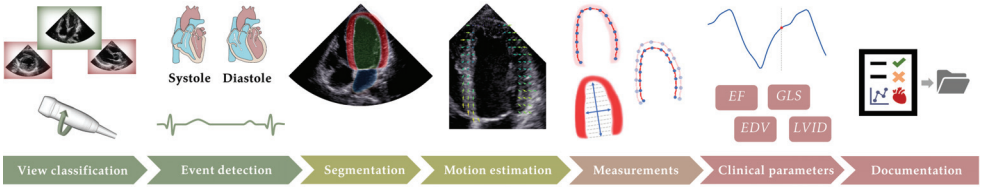


Figure 1.1: An example schematic of different steps that can constitute an automated echocardiography measurement.

The work herein mainly involves two-dimensional (2D) echocardiography, and some of the tasks would be less relevant in three dimensions. The following will be a brief introduction to the parts most relevant for this thesis, some proposed solutions to the different problems before the advent of DL, as well as fundamental limitations.

Classification of ultrasound acquisition

The first step of an analysis pipeline is to identify what is being imaged. A standard echocardiography exam consists of multiple video recordings from different acquisition postures, often called *views* [2, 11]. For an automated image analysis pipeline determining the current view can be regarded a mandatory step, as most measurements are specialized and thus require certain structures in the image. In practice, this is often implicit by the choice of measurement or specified by the operator or machine. In general, traditional algorithms have struggled to handle the possible diversity in the data, therefore clinicians mostly pick the required images for analysis and diagnosis manually [12]. This hampers the workflow, causing a demand for accurate and automated recognition of views in clinical software.

There are numerous aspects to consider when classifying an US acquisition, especially in terms of application and area of use. Images including the same content can have different quality depending on several factors, such as the equipment, patient condition and operator expertise. Some suitable for quantitative measurements, while others could produce poor or faulty results. For retrospective analysis, e.g. patient data post-visit, large databases, PACS organization and so on, detailed and flexible sorting of data with high detection rate independent of quality is desirable. Since the physical exam is finished, the quality of the raw acquisition can not

be improved and one must be content with the data. Also, large research studies of specific cohorts becomes more feasible with tools to sort views. In a prospective setting, however, feedback of image quality and guiding becomes more relevant. If the operator has the information necessary to improve the quality while scanning, for instance by quality metrics or suggestions on how to move the probe, this would be very beneficial [13].

Detection of cardiac events

For quantitative echocardiography, another important task is to handle the periodicity of the pumping heart, and define temporal measurement points. This is often done by dividing the acquisition into cardiac cycles, and defining the different phases with corresponding transitions. Several of the most common measurements in echocardiography are defined at specific time points, making it important to have accurate and reproducible timing detection methods. Also, the division into cardiac cycles facilitates standardization and effective storage.

Several commercial scanners rely on electrocardiograms (ECG) to define the cardiac cycle in a robust and automatic way [14]. However, it may be inconvenient with ECG cables, especially in point of care situations, and there are also shortcomings related to pathological ECG patterns. Visual inspections of ECG signals and the US sequence is also a recurrent approach, but the agreement between operators is quite low [15]. Using spectral Doppler over the valves is an accurate alternative, but in most cases it can not be performed simultaneously with the recordings used for measurements [16]. The synchronization across different recordings are complex and prone to error, partly due to the beat-to-beat variability. Detecting specific semantic time points also allows for using regression formulas to relate different subcycle events based on heart rate [17]. This is typically too general for the full population, and different formulas are derived based on gender, age, pathophysiology and more [18]. To avoid the mentioned limitations, the focus should thus be to detect the cardiac events for the actual recording based on image analysis alone.

Various methods have been proposed to automatically detect cardiac events directly from echocardiographic images. Some promising solutions have involved the use of segmentation methods [19]. However, an accurate segmentation of the left ventricle (LV) can not differentiate events with

similar area/volume. Also, segmentation methods are still not perfect, and segmentation errors can cause false event detections. Regional motion estimation (ME), specifically speckle tracking or tissue Doppler imaging (TDI) near the mitral annulus, have been successful compared to ECG and other methods [14, 16]. Despite good results, the feasibility on pathological cases and limited range of supported heart rates opens for further improvements.

Segmentation of cardiac structures

Cardiac image segmentation is one of the most important parts of many analysis workflows. The goal is to partition an image into semantically meaningful regions such as the ventricle lumen, the myocardium and atria. Further, this is used as a basis for numerous quantitative measurements, like the ejection fraction (EF), where accurate delineation of the endocardium is essential.

Traditionally, segmentation of heart structures in medical imaging modalities have been performed using methods like deformable models, active contours or atlases [20–22]. These methods have been extensively studied, and good performance have been shown. However, they are often tuned by hyperparameters, and require significant feature engineering or prior knowledge to achieve satisfactory accuracy [23]. Fully automatic methods available in clinical routine, some which are based on the latter, have limited adaption at many hospitals and can still be improved [24]. Due to this, manual or semi-automatic delineation of cardiac structures remains part of the daily work in echocardiography laboratories.

Myocardial motion estimation

Estimating the myocardial motion can serve as a rich descriptor of cardiac function. It can also be used to derive deformation metrics, such as myocardial strain. This allows for investigation of local wall motion and deformation, which is affected by many cardiac pathologies, for instance coronary artery disease. In clinical echocardiography, ME is typically done using speckle tracking algorithms or TDI, the prior being most common. Speckle tracking echocardiography (STE) is widely adopted, with methodological variants based on optical flow (OF). More specifically, block

matching methods have sought a lot of attention in the US community [25, 26]. In research, algorithms such as elastic image registration and phase sensitivity approaches [27–29], have been extensively investigated and achieved good results.

Two assumptions that often underlie the traditional algorithms is that the pixel intensity remains constant along the motion trajectory, and that motion is a pure translation in local regions. In general, this does not hold for 2D echocardiography. Cardiac tissue motion is a three dimensional (3D) phenomenon, involving both an apex to base shortening and a simultaneous twist. In 2D echocardiography, the motion of the tissue can thus be out of the image plane, both as a result of poor probe posture or the inherent myocardial fiber orientation with respect to the acoustic beam propagation. Therefore, given an optimal probe posture, it is still hard to tackle the problem in 2D. The effect complicates a lot of the traditional ME methods which assume consistency in local signal amplitude. The speckle pattern decorrelates, which reduces the trackable features. Despite improvements, any imposed assumptions will be a simplification of the actual problem. Methods have been proposed address this issue, such as incorporating conservation of the local phase signal [28] and elastic image registration. The latter is an optimization based method with the goal of finding a displacement field that minimizes some similarity metric between two images, where one is warped towards the other. One common problem with warping is that ambiguity can be caused by signal blocking artefacts, resulting in an ill-posed problem [30]. Often these methods involves the use of a priori regularization [26], which to some degree helps for the general case, but physical modelling of the cardiac muscle is complicated and a lot of simplifications have to be made.

Similar issues can occur due to noise common in echocardiography. In US there are several sources, such as reverberations, shadows and haze artefacts [31]. This hampers the ME accuracy by inflicting arbitrary and unstructured signals to the tracked regions. For conventional algorithms it is generally complicated to separate the useful data from noise, especially since there are many dissimilar origins and effects.

A pervasive problem with the algorithms is that they are often very complicated, and often require a high degree of manual hyperparameter tuning. This include the size of the search kernel, the range of the

search, smoothing factors and more. Their complexity also makes them computationally demanding, requiring expensive hardware to be fast enough for real-time use. Further, the heart can beat very fast, and assumptions made by current solutions require a high speckle consistency between frames, thus a very high frame rate on the US scanner.

Measurements and estimation of clinical parameters

An integration of the mentioned methods can be the basis of computing several clinical parameters automatically. For instance, the endocardial border of the ventricle could be extracted from the segmentation at several time points and views. Further, this could be used to estimate volume and EF. Anatomical landmarks, such as the apex and base points, can also be detected from the segmentation masks and be used to derive diagnostic parameters. For local deformation measurements, extracting useful points to track with ME is important.

Despite the possibilities and potential, automated measurements have not been widely embraced at many hospitals due to several limitations. As is understood from the previous sections, the task at hand is comprised of numerous steps that all can fail. Also, it is challenging to create algorithms that generalize to the extensive data variability. This includes image quality, pathology such as arrhythmia, abnormal chamber morphophysiology and more. Further, the different automatic measurements requires substantial studies, both in general population, but also in unique cohorts. The reproducibility of automatic measurements are often very high for large groups of patients, but on the individual level it can be suboptimal. Outliers must be handled with caution. Also, clinical adaption is not only about accurate and reproducible measurements. Intuitive presentation of results, user friendliness and accessibility are also essential. The workflow must be customized for daily routine and faster procedures.

As mentioned, the implementation of automatic measurements can be very beneficial in clinical routine. Today, measurements are mostly performed one single time per examination. Not using the average value over several cardiac cycles is a major limitation and not recommended as data quality and pathology can effect the measurements on a beat-to-beat basis [2, 3]. With fully automated methods, it becomes effortless to average over multiple heart cycles.

1.2 Limitations of myocardial strain imaging

Despite commends and reassuring experience with strain imaging, it has not been fully adopted in clinical practice. There could be numerous reasons for this, but robustness in real-life situations has been questioned [25, 32]. We also believe that the time required to perform these measurements is limiting. In 2D echocardiography, the pipeline of strain computation is composed of many components, and as mentioned earlier, these can all be sources of variation and inconsistencies.

The quality of the acquisition process is very important. It is influenced by several factors, including patient condition, operator expertise and equipment. For example, when the imaging plane transects the heart offset to the true apex, i.e. apical foreshortening, it will make the LV appear shorter and the apical region thicker. This leads to a geometric distortion which has a significant impact on measurements, resulting in overestimation of the LV function and underestimation of volume and length [33]. Further, the spatial and temporal resolution will have an effect on data quality, as local regions in the data will be less correlated between images if the resolution is too low. Lower temporal resolution will also lead to underestimation of strain [25].

Another important component of strain computation is the initialization of the region of interest (ROI), or tracking area. This can typically be points along the longitudinal of the ventricle. The placement of these points have high influence on strain measurements, with a significant gradient from the endocardial to the epicardial border. For standardization, this is typically done along one of the anatomic borders, or on the myocardial midline [34]. Naturally, manual contouring makes the operator variability quite high. In practice, segmentation is used to seed the tracking points, and the operator is allowed to adjust them upon measurement. The latter is a double-edged sword, on one side it allows the operator to adjust the worst case outliers, but on the other side it induces variability between measurements.

Global longitudinal strain is now recommended in the guidelines for chamber quantification in echocardiography [2]. However, regional strain measurements is not. The variability and reproducibility of these measurements are significantly higher compared to GLS [35]. The mentioned limitations are naturally a reason for this, but also regularization, such as smoothing, is suggested as one of the reasons for reduced reproducibil-

ity [34]. This typically lowers the resolution of detecting local changes, but helps on a global level.

Definition of cardiac events also have importance for strain parameters, especially regional deformation. Scars, reduced function and delayed polarization can lead to both early systolic lengthening and post-systolic shortening. It has been demonstrated that different surrogates for end diastole (ED) and end systole (ES) are unreliable for cases with regional pathology [36]. ES detection is most vital, and depends on a proper definition of the aortic valve closure (AVC), while wrong detection of ED can also result in false peak positive strain and bias in the strain peaks [35].

The training and experience of the operator is essential for strain imaging, not only to acquire proper data or adjust semi automated outputs such as ROI, but also to interpret the results. Interpretation of strain results are less intuitive for inexperienced operators, with an immense amount of values, variable representations and curves. Especially regional strain measurements where all the ventricle segments are considered individually and with respect to each other. Bull's eye plots and color anatomical M-Modes are good examples of intuitive summary representations of large amounts of data that helps the operator interpret the results [37, 38].

1.3 Aims of study

The overall aim of this work is to investigate the possibilities of using modern machine learning, namely deep learning, for fully automating several steps of an echocardiography examination. One key aspect is to address shortcomings of existing methods with respect to ultrasound and develop solutions for improved adaption. Investigations should also examine if these solutions can help tackle some of the aforementioned limitations of conventional methods. Another important asset to consider is the possibility of real-time processing and improved workflow. More specifically, the aim of this thesis is:

Aim 1: Investigate the use of deep learning for cardiac view classification, event detection and segmentation.

Aim 2: Investigate the use and potential benefits of using deep learning for myocardial motion estimation.

Aim 3: Integrate deep learning components into a pipeline for automatic strain measurements and compare its performance to state of the art solutions.

1.4 Summary of presented work

The following briefly summarize each contribution included in this thesis. The first three are focused towards technical method development and application. The last contribution is a clinical agreement study employing the integrated methods.

1.4.1 Real-time standard view classification in transthoracic echocardiography using convolutional neural networks

According to recommendations, a transthoracic echocardiography (TTE) exam should be performed with different probe postures to provide several standardized image views of the heart [2]. A standard view, such as the apical four-chamber (A4C), is usually a necessary prerequisite for quantitative measurements. Calculating biplane LVEF, for instance, would require the operator to acquire frames from both the A4C and apical two-chamber (A2C) views. Another aspect is that non-experts traditionally struggle to obtain these views in an optimal way. At worst, a suboptimal view can cause false interpretations of the data and patient diagnosis.

In this work we employ convolutional neural networks (CNNs) to develop a classification model for predicting cardiac views. We refer to our architecture as the cardiac view classification (CVC) network. It is composed of seven blocks of convolution filters, batch normalization, parametric rectified linear unit activations (PReLU) and max pooling [4, 39]. For the five last blocks Inception modules and a dense connection pattern are employed [40, 41]. Global average pooling layer was used before the final softmax activation. The network was trained on a dataset of 205 subjects with seven classes of the most common cardiac views. Further, we proposed the use of 2D image planes extracted from 3D US volume data acquired in the apical position to learn optimal probe orientations. The optimal angle for the three apical views were annotated in a probabilistic manner for 60 patients, and trained with the same network as a regression problem.

Results show that DL based methods provide state of the art results for

2D echocardiography with a sequence classification accuracy of 98.5% on the independent test data. With a runtime of 4.4 ms per frame it was also possible to run the network in real-time. For 3D data, the median deviation from optimal view was $4° \pm 3°$. This suggest that CNNs have the potential of being used for multiplanar reformatting and orientation guidance.

This paper was published in Ultrasound in Medicine and Biology (UMB), Volume 45, Issue 2, pages 374-384, February 2019. 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.

1.4.2 Detection of cardiac events in echocardiography using 3D convolutional recurrent neural networks

Another important task when assessing cardiac function is to determine various cardiac events. The most common measurement points are end-systole and end-diastole, which correspond to the time when the aortic and mitral valve closes respectively. Alternatively, the time points of lowest and highest ventricle volume. ED and ES are used extensively in quantitative echocardiography, for instance in the calculation of EF and global longitudinal strain (GLS).

In this work we proposed using a network composed of 3D CNNs followed by long short term memory (LSTM) layers to alleviate the spatio-temporal features in the image sequence. We argue that combined use of 3D CNN and LSTMs extends the context in both space and time, compared to using either individually or in combination with 2D convolutions. The network is trained to classify whether an image belongs to systole or diastole, and we use the switch between the states to define ES and ED. The network is trained on 300 patients of acquisitions from the A4C and A2C views, validated on 100 during training and tested on 100 patients post training.

Results indicated that the architecture combining 3D CNN and LSTM provided competitive results with state of the art solutions, and significantly better than combining 2D CNNs with LSTMs. The mean absolute error was roughly 1.5 frames for both views and events. In addition, runtime performance is fast with possibility of use in prospective pipelines.

This paper was published in IEEE International Ultrasonics Symposium (IUS),

pages 1-4, Oct 2018. It is presented here in its original form. The candidate was the second author and contributed to development of the employed neural network, parts of the annotation and data processing, as well as writing the manuscript. A. M. Fiorito was the first author and primal investigator.

1.4.3 Myocardial function imaging in echocardiography using deep learning

The deformation of the myocardium can be quantified, and this has shown beneficial for both diagnostic and prognostic evaluations of cardiac function. In echocardiography, we often refer to this as myocardial function imaging, or deformation imaging. Several markers are derived, such as strain and strain rate. Clinical use, however, still remains limited at many hospitals, which is partially believed to be due to its retrospective nature and questionable reproducibility. Motion estimation, commonly by speckle tracking, is a very important component for these measurements.

In this work, we develop a novel motion estimation method based on DL. The network is based on the PWC-Net architecture [42], with modifications to enhance performance on small and local displacements. This includes the removal of feature warping, higher level feature maps and flow estimation. A multi-scale loss with end-point error is employed, with contributions from all pyramid levels. We design a cascaded training regime with increasing resemblance to echocardiography data, and incorporate US relevant augmentation routines. The input of the model is two consecutive US images, and the output is the corresponding dense displacement field. Finally, we integrate the ME method in a pipeline with view classification, event detection and segmentation to fully automate longitudinal strain measurements.

The results show that learning-based ME has an unexploited potential both in terms of accuracy and runtime performance. We show that inducing US relevant augmentations can have a twofold benefit, firstly it increases the representation size of the data, but it also improves the models adaptability to image artifacts. *In vivo* results are promising, within expected limits of agreement seen in intervendor studies.

This paper has been accepted for publication in IEEE Transactions on Medical Imaging (TMI), Jan. 2021. It is presented here in its original form. The

candidate was the first author and contributed to all aspects of the work, except for acquisition of echocardiography data. I.M Salte performed the reference strain measurements, while E. Smistad and A.M. Fiorito was the primal investigators of the segmentation and cardiac event detection methods respectively.

1.4.4 Artificial intelligence for automatic measurement of left ventricular strain in echocardiography

In the previous contribution, we proposed a novel motion estimation method based on deep learning. We integrated this with view classification, event detection and segmentation in a pipeline for measuring longitudinal strain, and showed promising results on a limited amount of simulations and *in vivo* data. In this work, we investigated the agreement for GLS measurements between the proposed pipeline compared to a commercially available strain estimation software (2DS in EchoPAC v202, GE Vingmed Ultrasound AS) on a large *in vivo* dataset. The dataset consisted of 200 patients with a significant variation in LV function and demographic properties.

For all the patients, and in the majority of individual acquisitions, the DL pipeline succeeds to estimate GLS. The correspondence with the commercial system was comparable to intervendor studies. Further, the time required to analyse all the steps for the three apical views of on patient was less than 15 seconds, significantly faster than other proposed methods.

This paper has been submitted 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 all aspects of the technical method development and result generation, in addition to drafting of the manuscript. I.M. Salte was the first author and performed the statistical analysis and reference measurements, as well as writing of the manuscript.

1.5 Publication list

Through the course of this studies, 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. **Andreas Østvik**, Erik Smistad, Svein Arne Aase, Bjørn Olav Haugen and Lasse Løvstakken,"Real-time standard view classification in transthoracic echocardiography using convolutional neural networks", *Ultrasound in medicine and biology*, Volume 45, Issue 2, pages 374-384, February 2019.

2. Adrian Meidell Fiorito, **Andreas Østvik**, Erik Smistad, Sarah Leclerc, Olivier Bernard, and Lasse Løvstakken, "Detection of Cardiac Events in Echocardiography using 3D Convolutional Recurrent Neural Networks", *IEEE International Ultrasonics Symposium*, Kobe, 2018.

3. **Andreas Østvik**, Ivar Mjåland Salte, Erik Smistad, Daniela Melichova, Thuy Mi Nguyen, Kristina Haugaa, Harald Brunvand, Thor Edvardsen, Bjørnar Grenne and Lasse Løvstakken, "Myocardial function imaging in echocardiography using deep learning", accepted for publication in *IEEE Transactions on Medical Imaging*, January 2021.

4. Ivar Mjåland Salte, **Andreas Østvik**, Erik Smistad, Daniela Melichova, Thuy Mi Nguyen, Sigve Karlsen, Harald Brunvand, Kristina Haugaa, Thor Edvardsen, Lasse Løvstakken and Bjørnar Grenne, "Artificial intelligence for automatic measurement of left ventricular strain in echocardiography - Agreement between a novel fully automated deep learning pipeline and a commercially available semiautomatic reference method", submitted for review to *Journal of the American College of Cardiology: Cardiovascular Imaging*.

Other contributions in peer reviewed journals

1. Sarah Leclerc, Erik Smistad, João Pedrosa, **Andreas Østvik**, Frederic Cervenansky, Florian Espinosa, Torvald Espeland, Erik Andreas Rye Berg, Pierre-Marc Jodoin, Thomas Grenier, Carole Lartizien, Jan D'hooge, Lasse Løvstakken and Olivier Bernard, "Deep learning for segmentation using an open large-scale dataset in 2D echocardiography", *IEEE Transactions on Medical Imaging*, Volume 38, Issue 9, pages 2198-2210, September 2019.

2. Erik Smistad, **Andreas Østvik**, Ivar Mjåland Salte, Daniela Melichova,

Thuy Mi Nguyen, Kristina Haugaa, Harald Brunvand, Thor Edvardsen, Sarah Leclerc, Olivier Bernard, Bjørnar Grenne and Lasse Løvstakken, "Real-time automatic ejection fraction and foreshortening detection using deep learning", *IEEE Transactions on Ultrasonics, Ferroelectrics, and Frequency Control*, Volume 67, Issue 12, pages 2595-2604, December 2020.

3. Sarah Leclerc, Erik Smistad, **Andreas Østvik**, Frederic Cervenansky, Florian Espinosa, Torvald Espeland, Erik Andreas Rye Berg, Mourad Belhamissi, Sardor Israilov, Thomas Grenier, Carole Lartizien, Pierre-Marc Jodoin, Lasse Løvstakken and Olivier Bernard, "LU-Net: a multi-stage attention network to improve the robustness of segmentation of left ventricular structures in 2D echocardiography", *IEEE Transactions on Ultrasonics, Ferroelectrics, and Frequency Control*, Volume 67, Issue 12, pages 2519-2530, December 2020.

4. Erik Smistad, **Andreas Østvik**, and Andrè Pedersen, "High Performance Neural Network Inference, Streaming, and Visualization of Medical Images Using FAST", *IEEE Access*, Volume 7, pages 136310-136321, September 2019.

Conference proceedings

1. Fabian Sødal Dietrichson, Erik Smistad, **Andreas Østvik**, and Lasse Løvstakken, "Ultrasound speckle reduction using generative adversarial networks", *IEEE International Ultrasonics Symposium*, Kobe, 2018.

2. Sarah Leclerc, Erik Smistad, Thomas Grenier, Carole Lartizien, **Andreas Østvik**, Florian Espinosa, Pierre-Marc Jodoin, Lasse Løvstakken, and Olivier Bernard, "Deep learning applied to multi-structure segmentation in 2D echocardiography: A preliminary investigation of the required database size", *IEEE International Ultrasonics Symposium*, Kobe, 2018.

3. Sarah Leclerc, Erik Smistad, Thomas Grenier, Carole Lartizien, **Andreas Østvik**, Frederic Cervenansky, Florian Espinosa,Torvald Espeland, Erik Andreas Rye Berg, Pierre-Marc Jodoin, Lasse Løvstakken, and Olivier Bernard, "RU-Net: A refining segmentation network

for 2D echocardiography", *IEEE International Ultrasonics Symposium*, Glasgow, 2019.

4. Sarah Leclerc, Erik Smistad, **Andreas Østvik**, Frederic Cervenansky, Florian Espinosa, Torvald Espeland, Erik Andreas Rye Berg, Pierre-Marc Jodoin, Thomas Grenier, Carole Lartizien, Lasse Lovstakken, and Olivier Bernard, "Deep Learning Segmentation in 2D echocardiography using the CAMUS dataset: Automatic Assessment of the Anatomical Shape Validity", *International Conference on Medical Imaging with Deep Learning*, London, 2019.

5. **Andreas Østvik**, Erik Smistad, Torvald Espeland, Erik Andreas Rye Berg, and Lasse Løvstakken, "Automatic Myocardial Strain Imaging in Echocardiography Using Deep Learning", *Deep Learning in Medical Image Analysis and Multimodal Learning for Clinical Decision Support*, Granada, 2018.

6. **Andreas Østvik**, Lars Eirik Bø, and Erik Smistad. "EchoBot: An open-source robotic ultrasound system", *Information Processing in Computer Assisted Interventions*, Rennes, 2019.

7. Erik Smistad, **Andreas Østvik**, Bjørn Olav Haugen, and Lasse Løvstakken. "2D left ventricle segmentation using deep learning", *IEEE International Ultrasonics Symposium*, Washington DC, 2017.

8. Smistad, Erik, **Andreas Østvik**, Ivar Mjåland Salte, Sarah Leclerc, Olivier Bernard, and Lasse Lovstakken. "Fully automatic real-time ejection fraction and MAPSE measurements in 2D echocardiography using deep neural networks", *IEEE International Ultrasonics Symposium*, Kobe, 2018.

9. Smistad, Erik., Ivar Mjåland Salte, **Andreas Østvik**, Sarah Leclerc, Olivier Bernard, and Lasse Lovstakken. "Segmentation of apical long axis, four-and two-chamber views using deep neural networks", *IEEE International Ultrasonics Symposium*, Glasgow, 2019.

International conference presentations

1. **Andreas Østvik**, Ivar Mjåland Salte, Erik Smistad, and Lasse Løvstakken, "Adapting deep learning based motion estimation for my-

ocardial function imaging", Poster, *IEEE International Ultrasonics Symposium*, Glasgow, 2019.

2. **Andreas Østvik**, Lars Eirik Bø, and Erik Smistad, "EchoBot: An open-source robotic ultrasound system", Oral and poster, *International conference on Information Processing in Computer Assisted Interventions*, Rennes, 2019.

3. Adrian Meidell Fiorito, **Andreas Østvik** (presenter), Erik Smistad, Sarah Leclerc, Olivier Bernard, and Lasse Løvstakken, "Detection of Cardiac Events in Echocardiography using 3D Convolutional Recurrent Neural Networks", Poster, *IEEE International Ultrasonics Symposium*, Kobe, 2018.

4. **Andreas Østvik**, Erik Smistad, Torvald Espeland, Erik Andreas Rye Berg, and Lasse Løvstakken, "Automatic functional imaging in echocardiography using deep learning based segmentation and flow estimation", Oral, *IEEE International Ultrasonics Symposium*, Kobe, 2018.

5. **Andreas Østvik**, Erik Smistad, Torvald Espeland, Erik Andreas Rye Berg, and Lasse Løvstakken, "Automatic Myocardial Strain Imaging in Echocardiography Using Deep Learning", Poster, MICCAI 4th Workshop on Deep Learning in Medical Image Analysis, Granada, 2018.

6. Erik Smistad, **Andreas Østvik** (presenter), Bjørn Olav Haugen, and Lasse Løvstakken. "2D left ventricle segmentation using deep learning", Poster, *IEEE International Ultrasonics Symposium*, Washington DC, 2017.

7. **Andreas Østvik**, Erik Smistad, Svein Arne Aase, Bjørn Olav Haugen, Lasse Løvstakken, "Real-Time Classification of Standard Cardiac Views in Echocardiography using Neural Networks", Poster, *IEEE International Ultrasonics Symposium*, Washington DC, 2017.

1.6 Discussion of results

In this work, an investigation of using deep learning for various common image analysis steps in echocardiography was conducted. Initially this involved development of cardiac view classification, event detection and segmentation algorithms. Further, DL based motion estimation models adapted for echocardiography was proposed, followed by an integration of components into a pipeline for fully automated strain measurements. We show that DL methods can perform favourable compared to conventional methods and state of the art. In parallel with this work, several groups in the research community have presented related approaches to tackle similar problems, all supporting benefits of using DL. We try to incorporate some of the associated work into the further discussion.

Classification of cardiac views

The overall performance of the classification method shows that DL is an attractive method for cardiac view recognition. On seven of the most common cardiac views, state of the art results were achieved. Failures could in most cases be assigned to bad image quality, abnormal features or high similarity between views. However, the different classes included in our study represents distinct top level cardiac views with relatively high disparity between them. For instance, we did not differentiate between an A4C view with LV focus and one with RV focus. In practice, at least for quantitative measurements, this would greatly improve the value of the algorithm. Parallel studies show that similar results could be achieved on more views, and with less data [43].

The proposed CVC model was able to classify over 200 frames per second on a modern GPU at the time of writing. That is significantly higher than traditional methods, and could reduce number of clicks and time selecting data for analysis. Also, the method can be used in a prospective scenario on streamed US data without significant overhead. Real-time capabilities allowed for continuous acquisitions, which can be seen in work presented by our group [44]. This gives the operator the opportunity to acquire data from different views without clicking any button. In that work, we also propose metrics for estimating apical foreshortening based on the segmentation output and show that this significantly affects EF measurements. As

mentioned, this is also supported by other studies where they also show the effect on GLS [33]. The foreshortening detection can be run in real-time, giving the operator feedback if the view is suboptimal. However, further studies must be conducted to investigate if this improves the quality of the acquired data and subsequent measurements, and if continuous acquisition improves the workflow.

One major limitation with this kind of approaches is that classification models can become overly confident due to the inherent nature of the optimization procedure. For probe guidance and quality assurance they are not well suited, and more sensitive metrics must be employed, for instance, to differentiate a good acquisition from a bad one. In the work by Abdi *et al.* they derive a regression based quality score [13, 45] based on different scanner settings and visual features of the image. This includes centering of relevant objects, spatial settings, gain, as well as visibility of boundaries and valves. Results are promising, however, the number of views are limited and it is not known how these type of regression approaches would work on additional and more similar views. An alternative or supplementary direction related to quality assurance and guidance is discussed in contribution 1, where using 3D data for training the algorithms for use in a 2D acquisition scenario. With this type of approach, the operator could potentially get feedback on how to optimally align the probe. The work is limited to orientation, but extending it to tilting and position is also worth pursuing. The optimal way for further research is hard to determine, but a combination of a quality assurance metrics together with feedback to the operator on how to improve is reasonable.

Both the temporal and spatial resolution affects the measurement quality. An extension of this work could for instance include recommendation to the user regarding scanner configuration, such as width and depth adjustment. Reducing these parameters typically gives higher frame rate, and should thus be pursued for quantitative measurements where high temporal resolution is beneficial. In the future, these adjustments could potentially be performed without operator interaction enabled by communication between the ML algorithm and the scanner configurations.

Detection of cardiac events

Event detection using 3D CNNs with recurrent layers yielded promising results within interobserver variability. We recall that ED and ES represents

the time of mitral and aortic valve closure respectively. The data used in this study was limited to A4C and A2C views, and interestingly these does not include a visible aortic valve. Earlier it has been proposed that a notch/nadir in the velocity field close to the septal base before mitral valve opening is a recoil of the AVC [14]. This is visible with speckle tracking, and could also suggest that the DL models implicitly detects similar features. However, additional work must be conducted to verify this.

In the work by Dezaki *et al.* they use a regression based approach and apply a volume mimicking curve as image labels through the cardiac cycle [46]. Their best model architecture is composed of a DenseNet followed by gated recurrent units (GRU) [41, 47], and they design a loss function promoting ED and ES detection. Their average results on A4C surpass ours, despite our findings suggesting that 3D CNN followed by recurrent neural network (RNN) layers were superior to a 2D CNN to RNN approach. This could suggest that it is still possible to improve the results. Either way, there is a chance that their approach implicitly uses the ECG for labeling as they extract it from a conventional software, and as stated earlier, this should be avoided if possible.

Another potential issue with RNNs is that they often require a substantial amount of subsequent frames as input for the models to perform optimal. This can result in memory issues and limit the use on low-end systems. In recent work employing the model this also appeared as a bottleneck for real-time use [44], with a significant drop in frame rate when deployed in practice.

Currently, the models are limited to detecting diastole and systole, but an extension to detect additional cardiac events should be possible. The acquisition rate available in US scanners today, together with the capabilities of DL models, allows for the approximation of valve closures and openings, as well as the rapid filling phase, diastasis and atrial systole. This could be beneficial for several existing measurements, but also facilitate the development of new ones.

In conclusion, event detection using DL is very promising. Not only as a potential replacement for ECG when needed, but also for advancing current solutions. Detecting the valve closures from the images directly will remove the need for surrogates like QRS, which as mentioned is often affected by cardiac disease. This could potentially make quantitative measurements

more reliable, especially for regional measurements.

Segmentation of the left ventricle

The segmentation was a very important component in the later phases of this work, specifically for the automated strain measurements. It was used to segment the LV myocardium and extract the midline at initialization of tracking for strain measurements. The employed network was an U-Net based architecture [48] with modification emphasizing improved inference [49]. It supported segmentation of the LV lumen, myocardium and atrium in the A4C and A2C views. It was later used in the CAMUS study by Leclerc *et al.* [50], which was a collaboration between our group, the Creatis laboratory at the University of Lyon (France) and Katholieke Universiteit (KU) in Leuven (Belgium). More recently, the method was extended to support apical long-axis (APLAX) views by Smistad *et al.* [51].

On average we achieve good results on all classes and views. The worst performance is on the myocardium, which is the most important class in our strain pipeline. Segmentation of LV lumen has arguably been a priority, with many natural applications including volume and EF measurements. In general, this is also considered a simpler task with better visibility of endocardium compared to the epicardium in echocardiography. This is also prevalent from the results. In future work, emphasising myocardium segmentation by inferring shape regularization or class weighting promoting better delineation of the myocardium could be worth pursuing. Also, due to the fast inference, model complexity could be increased without affecting real-time capabilities.

Gradually, DL based segmentation algorithms have outperformed previous state of the art, and is now dominating the field [23]. The flexibility of the models and the profound performance on *in vivo* data supports clinical implementation and extensive use.

Myocardial motion estimation

For simulated data, the results show good correlation between the velocity of the underlying biomechanical model and the DL approach. The data is relatively homogeneous with limited variance between cases, and as expected the performance decreased when testing on *in vivo* data. It must be

taken into account that the reference from the commercial solution is not a ground truth, and despite proper validation, it can still produce suboptimal results for individual cases.

The loss and ground truth training data is restricted to optimization within the myocardium, and we can not assume that the motion estimator performs well outside of that region. Further studies must be conducted to evaluate the generalization. Also, in future work, the motion patterns in the data can be expanded significantly, and in this setting one would strive to generate a diverse representation for training. Multi-chamber electromechanical models are also becoming a possible direction for generating full heart motion patterns [52].

Contrary to a majority of ME methods, the DL approach does not infer any constraints regarding consistency of image intensity or phase. This makes the models more flexible and may opportune the handling of fundamental problems like out of plane motion and decorrelation of speckle. Existing methods such as elastic image registration can bypass these limitations to some extent, but the use of image warping and model regularization are not necessarily optimal as mentioned earlier. The conventional models will not be specialized for the data they are used for, which is one of the main reasons for pursuing learning-based methods. However, for ME in echocardiography the ground truth displacement maps are not practical to extract from *in vivo* data. So far the most immediate solution is to use simulated US. This is not optimal, but in light of the conducted work we believe it is propitious with an increasing degree of realism and an infinite supply of underlying motion patterns. A promising alternative to supervised optical flow is unsupervised learning approaches. These methods have produced competitive results on common benchmarks [53]. Also, estimating 3D displacement fields from 2D images can also be a possible direction, and even though its an extremely ill-posed problem, DL based methods have shown promising results [54]. In any case, finding solutions for improved *in vivo* validation will be essential.

A related pilot study using neural networks derived from the FlowNet 2.0 architecture [55] also suggests the use of simulated data to deal with lack of ground truth for ME [56]. They also show that the simulation to real transfer is feasible, with competitive result on a rotating phantom versus a state of the art conventional method. We join them in supporting the

high potential of these methods, both in terms of simplicity by reduced hyperparameter tuning and adaptability.

If incorporated in the learning, ME methods with DL have the potential of being more adaptive if faced by noise. We propose using relevant noise inducing data augmentation for this, and show that this helps in controlled experiments. Defining noise for the purpose of augmentation is a paradox problem, and finding exact descriptions have been a topic of widespread research in the US community [31]. In DL, augmentations have an explicit regularization effect, and the variance in random noise application may end up in the range of realistic noise and thus have a positive effect on the final model. It will be important in future work to systematically investigate what kind of augmentation effects improves ME adaptability and performance on *in vivo* data.

Despite fast processing, the inference time of roughly 15 frames per second is relatively far away from the real-time limit. However, compared to other optical flow methods these methods are fast, and with additional pruning and optimization we believe it is possible to use in prospective applications. This part will also be facilitated by the rapid development of new hardware and optimized inference engines.

Automating strain measurements

Results within proposed intervendor variability indicate that DL based measurement pipelines for automatic strain estimation is promising. The patient material in contribution 4 is relatively inhomogeneous with a wide strain measurement distribution. However, each subgroup is relatively small, and larger intervendor studies with a broader population is required to map the general performance. There is still some deviation from interobserver variability, but we argue that closing the gap is possible with the aforementioned improvements.

Tackling apical foreshortening and out of plane motion is complex. Our proposed foreshortening measurement [44] can potentially supplement the strain calculations as a quality metric. It could also be interesting to investigate if it is possible to derive compensation functions or variance estimation based on this. In some subjects, avoiding foreshortening is impossible, and alternatives in such situations are in demand [33].

As stated in contribution 4, the average time spent by an operator in

order to conduct a single GLS analysis using commercial semi-automatic software is 5-10 minutes. The proposed pipeline for fully automated measurement can perform the full analysis in less than 15 seconds. In the pipeline, the bottleneck for improving the runtime performance is the ME network. As suggested in contribution 3, there are several opportunities to optimize and prune the ME network for improved runtime performance. A goal should be to surpass the recommended frame rate of 60 frames per second [2], and enable prospective velocity and strain measurements of the moving tissue while scanning. However, it is not known if this is required for learning-based approaches, as the classical assumptions regarding speckle consistency can be bypassed by DL methods. Nonetheless, we believe real-time support can be very valuable to the operator, similar to way visual inspection is almost mandatory when doing Doppler acquisitions.

For clinical implementation, several factors can minimize the impact on measurement variability and we believe that the flexibility in component-based approaches can create more trust among clinicians compared to an fully end-to-end solution. As it is today, operator quality assurance should still be possible, with an interactive user interface and opportunity of adjusting the ROI and assessing tracking quality. Adjustments will cause variability, and should be avoided unless necessary, but an option would be to present or store results from both tracks. Further, ML based anomaly detection can be a way to capture potential errors.

Robust regional assessment of myocardial tissue will potentially be incremental to global measurements, and facilitate more detailed and patient-specific diagnostics. However, regional strain measurements are still not a reliable tool [2, 35]. Reasons for this include apical angle distortion, reverberation and reduction in lateral resolution with depth. The effects will often cause inhomogeneous tracking conditions from apex to base. This comes on top of the fundamental noise in US discussed earlier, and makes it very complicated to design conventional algorithms for good performance in general. Learning-based algorithms have the potential of bypassing these limitations by embedding the problem into the optimization procedure. This is supported by our model adaption results in contribution 3, where the DL based method adapts to local noise and abnormality along the myocardium, while the conventional method struggles. This makes DL very attractive for regional measurements, but efforts are still necessary to reduce uncertainties

and potential flaws and misinterpretations along the analysis pipeline [57].

1.7 Concluding remarks

In this work, the focus has been to investigate the use of deep learning for several image analysis tasks in echocardiography. This includes cardiac view classification, event detection, segmentation and motion estimation. The results from the studies indicate that all of them can benefit or even be improved using DL. Further, the flexibility of data driven models surpass conventional methods on inherent limitations in US due to noise and acquisition variability. Possibilities within echocardiography are therefore immense. The integration of methods in a fully automated pipeline for strain measurements was feasible and yielded optimistic results. It is believed that such pipelines can facilitate accelerated diagnosis within echocardiography in the future, and potentially improve the robustness and accuracy of clinical measurements. We believe that the research community has not fully exploited these powerful tools, and expect them to be widespread in clinical echocardiography routine within short time.

1.8 Thesis outline

The thesis outline is as follows: In Chapter 2 the relevant background and terminology for echocardiography, deep learning and motion estimation is given. This should give the unfamiliar reader an introduction to the matter and capability of understanding the problems and work presented in the following chapters. In Chapter 3-5 three technical contributions are included as originally published, but adapted to the book layout. A clinical agreement study based on the developed technical methods is presented in the Chapter 6, and is included here as submitted to the journal.

References

[1] I. Edler and C. H. Hertz, "The use of ultrasonic reflectoscope for the continuous recording of the movements of heart walls.," *Clinical physiology and functional imaging*, vol. 24, no. 3, pp. 118–136, 1954.

[2] R. M. Lang, L. P. Badano, V. Mor-Avi, J. Afilalo, A. Armstrong, L. Ernande, F. A. Flachskampf, E. Foster, S. A. Goldstein, T. Kuznetsova, *et al.*, "Recommendations for cardiac chamber quantification by echocardiography in adults: an update from the american society of echocardiography and the european association of cardiovascular imaging," *European Heart Journal-Cardiovascular Imaging*, vol. 16, no. 3, pp. 233–271, 2015.

[3] M. T. Nolan and P. Thavendiranathan, "Automated quantification in echocardiography," *JACC: Cardiovascular Imaging*, vol. 12, no. 6, pp. 1073–1092, 2019.

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

[5] D. Silver, J. Schrittwieser, K. Simonyan, I. Antonoglou, A. Huang, A. Guez, T. Hubert, L. Baker, M. Lai, A. Bolton, *et al.*, "Mastering the game of go without human knowledge," *nature*, vol. 550, no. 7676, pp. 354–359, 2017.

[6] Y. Liu, A. Jain, C. Eng, D. H. Way, K. Lee, P. Bui, K. Kanada, G. de Oliveira Marinho, J. Gallegos, S. Gabriele, *et al.*, "A deep learning system for differential diagnosis of skin diseases," *Nature Medicine*, pp. 1–9, 2020.

[7] E. Ilg, N. Mayer, T. Saikia, M. Keuper, A. Dosovitskiy, and T. Brox, "Flownet 2.0: Evolution of optical flow estimation with deep networks," in *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, Jul 2017.

[8] R. J. Van Sloun, R. Cohen, and Y. C. Eldar, "Deep Learning in Ultrasound Imaging," *Proceedings of the IEEE*, vol. 108, pp. 11–29, jan 2020.

[9] 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," vol. 5, pp. 261–275, apr 2019.

[10] 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, pp. 1456–1466, sep 2015.

[11] G. Wharton, R. Steeds, J. Allen, H. Phillips, R. Jones, P. Kanagala, G. Lloyd, N. Masani, T. Mathew, D. Oxborough, B. Rana, J. Sandoval, R. Wheeler, K. O'gallagher, and V. Sharma, "A minimum dataset for a standard adult transthoracic echocardiogram: a guideline protocol from the British Society of Echocardiography York Teaching Hospital NHS Foundation Trust,"

[12] H. Khamis, G. Zurakhov, V. Azar, A. Raz, Z. Friedman, and D. Adam, "Automatic apical view classification of echocardiograms using a discriminative learning dictionary," *Medical Image Analysis*, vol. 36, pp. 15–21, feb 2017.

[13] A. H. Abdi, C. Luong, T. Tsang, G. Allan, S. Nouranian, J. Jue, D. Hawley, S. Fleming, K. Gin, J. Swift, R. Rohling, and P. Abolmaesumi, "Automatic Quality Assessment of Echocardiograms Using Convolutional Neural Networks: Feasibility on the Apical Four-Chamber View," *IEEE transactions on medical imaging*, vol. 36, no. 6, 2017.

[14] S. A. Aase, S. R. Snare, H. Dalen, A. Stoylen, F. Orderud, and H. Torp, "Echocardiography without electrocardiogram," *European Journal of Echocardiography*, vol. 12, pp. 3–10, jan 2011.

[15] M. Zolgharni, M. Negoita, N. M. Dhutia, M. Mielewczik, K. Manoharan, S. M. A. Sohaib, J. A. Finegold, S. Sacchi, G. D. Cole, and D. P. Francis, "Automatic detection of end-diastolic and end-systolic frames in 2D echocardiography," *Echocardiography*, vol. 34, pp. 956–967, jul 2017.

[16] S. A. Aase, A. Stoylen, C. B. Ingul, S. Frigstad, and H. Torp, "Automatic timing of aortic valve closure in apical tissue Doppler images," *Ultrasound in Medicine and Biology*, vol. 32, pp. 19–27, jan 2006.

[17] A. M. Weissler, W. S. Harris, and C. D. Schoenfeld, "Systolic Time Intervals in Heart Failure in Man," *Circulation*, vol. 37, pp. 149–159, feb 1968.

[18] R. P. Lewis, S. E. Rittogers, W. F. Froester, and H. Boudoulas, "A critical review of the systolic time intervals.," *Circulation*, vol. 56, pp. 146–158, aug 1977.

[19] S. Darvishi, H. Behnam, M. Pouladian, and N. Samiei, "Measuring Left Ventricular Volumes in Two-Dimensional Echocardiography Image Sequence Using Level-set Method for Automatic Detection of End-Diastole and End-systole Frames," *Research in Cardiovascular Medicine*, vol. 1, pp. 39–45, sep 2012.

[20] V. Tavakoli and A. A. Amini, "A survey of shaped-based registration and segmentation techniques for cardiac images," *Computer Vision and Image Understanding*, vol. 117, pp. 966–989, sep 2013.

[21] C. Petitjean and J. N. Dacher, "A review of segmentation methods in short axis cardiac MR images," vol. 15, pp. 169–184, apr 2011.

[22] J. A. Noble and D. Boukerroui, "Ultrasound image segmentation: A survey," vol. 25, pp. 987–1010, aug 2006.

[23] C. Chen, C. Qin, H. Qiu, G. Tarroni, J. Duan, W. Bai, and D. Rueckert, "Deep learning for cardiac image segmentation: A review," *Frontiers in Cardiovascular Medicine*, vol. 7, p. 25, 2020.

[24] A. C. Armstrong, E. P. Ricketts, C. Cox, P. Adler, A. Arynchyn, K. Liu, E. Stengel, S. Sidney, C. E. Lewis, P. J. Schreiner, J. M. Shikany, K. Keck, J. Merlo, S. S. Gidding, and J. A. Lima, "Quality Control and Reproducibility in M-Mode, Two-Dimensional, and Speckle Tracking Echocardiography Acquisition and Analysis: The CARDIA Study, Year 25 Examination Experience," *Echocardiography*, vol. 32, pp. 1233–1240, aug 2015.

[25] "Myocardial strain imaging: review of general principles, validation, and sources of discrepancies," *European Heart Journal - Cardiovascular Imaging*, vol. 20, pp. 605–619, jun 2019.

[26] B. Heyde, R. Jasaityte, D. Barbosa, V. Robesyn, S. Bouchez, P. Wouters, F. Maes, P. Claus, and J. D'Hooge, "Elastic image registration versus speckle tracking for 2-d myocardial motion estimation: A direct comparison in vivo," *IEEE Transactions on Medical Imaging*, vol. 32, no. 2, pp. 449–459, 2013.

[27] K. McLeod, A. Prakosa, T. Mansi, M. Sermesant, and X. Pennec, "An incompressible log-domain demons algorithm for tracking heart tissue," in *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, vol. 7085 LNCS, pp. 55–67, 2012.

[28] M. Alessandrini, A. Basarab, H. Liebgott, and O. Bernard, "Myocardial motion estimation from medical images using the monogenic signal," *IEEE Transactions on Image Processing*, vol. 22, no. 3, pp. 1084–1095, 2013.

[29] B. Heyde, R. Jasaityte, D. Barbosa, V. Robesyn, S. Bouchez, P. Wouters, F. Maes, P. Claus, and J. D'Hooge, "Elastic image registration versus speckle tracking for 2-d myocardial motion estimation: A direct comparison in vivo," *IEEE Transactions on Medical Imaging*, vol. 32, no. 2, pp. 449–459, 2013.

[30] S. Zhao, Y. Sheng, Y. Dong, E. I.-C. Chang, and Y. Xu, "MaskFlownet: Asymmetric Feature Matching with Learnable Occlusion Mask," *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, pp. 6277–6286, mar 2020.

[31] A. Fatemi, E. A. R. Berg, and A. Rodriguez-Molares, "Studying the Origin of Reverberation Clutter in Echocardiography: In Vitro Experiments and In Vivo Demonstrations," *Ultrasound in Medicine and Biology*, vol. 45, pp. 1799–1813, jul 2019.

[32] K. E. Farsalinos, A. M. Daraban, S. Ünlü, J. D. Thomas, L. P. Badano, and J. U. Voigt, "Head-to-Head Comparison of Global Longitudinal Strain Measurements among Nine Different Vendors: The EACVI/ASE Inter-Vendor

Comparison Study," *Journal of the American Society of Echocardiography*, vol. 28, pp. 1171–1181.e2, oct 2015.

[33] 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, L. P. Badano, 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, pp. 337–343, jul 2019.

[34] J. U. Voigt, G. Pedrizzetti, P. Lysyansky, T. H. Marwick, H. Houle, R. Baumann, S. Pedri, Y. Ito, Y. Abe, S. Metz, J. H. Song, J. Hamilton, P. P. Sengupta, T. J. Kolias, J. D'Hooge, G. P. Aurigemma, J. D. Thomas, and L. P. Badano, "Definitions for a common standard for 2D speckle tracking echocardiography: consensus document of the EACVI/ASE/Industry Task Force to standardize deformation imaging," *European heart journal cardiovascular Imaging*, vol. 16, no. 1, pp. 1–11, 2015.

[35] O. Mirea, E. D. Pagourelias, J. Duchenne, J. Bogaert, J. D. Thomas, L. P. Badano, J. U. Voigt, J. Hamilton, S. Pedri, P. Lysyansky, G. Hansen, Y. Ito, T. Chono, J. Vogel, D. Prater, S. Park, J. Y. Lee, H. Houle, B. Georgescu, R. Baumann, B. Mumm, Y. Abe, and W. Gorissen, "Variability and Reproducibility of Segmental Longitudinal Strain Measurement: A Report From the EACVI-ASE Strain Standardization Task Force," *JACC: Cardiovascular Imaging*, vol. 11, pp. 15–24, jan 2018.

[36] R. O. Mada, P. Lysyansky, A. M. Daraban, J. Duchenne, and J. U. Voigt, "How to define end-diastole and end-systole?: Impact of timing on strain measurements," *JACC: Cardiovascular Imaging*, vol. 8, pp. 148–157, feb 2015.

[37] D. Liu, K. Hu, P. Nordbeck, G. Ertl, S. Störk, and F. Weidemann, "Longitudinal strain bull's eye plot patterns in patients with cardiomyopathy and concentric left ventricular hypertrophy," vol. 21, p. 21, may 2016.

[38] L. A. Brodin, J. Van der Linden, and B. Olstad, "Echocardiographic functional images based on tissue velocity information," *Herz*, vol. 23, no. 8, pp. 491–498, 1998.

[39] S. Ioffe and C. Szegedy, "Batch normalization: Accelerating deep network training by reducing internal covariate shift," in *International Conference on Machine Learning*, pp. 448–456, 2015.

[40] C. Szegedy, V. Vanhoucke, S. Ioffe, J. Shlens, and Z. Wojna, "Rethinking the inception architecture for computer vision," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 2818–2826, 2016.

[41] G. Huang, Z. Liu, K. Q. Weinberger, and L. van der Maaten, "Densely connected convolutional networks," *arXiv preprint arXiv:1608.06993*, 2016.

[42] D. Sun, X. Yang, M.-Y. Liu, and J. Kautz, "PWC-Net: CNNs for optical flow using pyramid, warping, and cost volume," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 8934–8943, 2018.

[43] A. Madani, R. Arnaout, M. Mofrad, and R. Arnaout, "Fast and accurate classification of echocardiograms using deep learning," p. 8, jun 2018.

[44] E. Smistad, A. Østvik, I. Mjåland Salte, D. Melichova, T. Mi Nguyen, K. Haugaa, H. Brunvand, T. Edvardsen, S. Leclerc, O. Bernard, B. Grenne, L. Løvstakken, A. Østvik are, and T. Mi Nguyen are, "Real-Time Automatic Ejection Fraction and Foreshortening Detection Using Deep Learning," *IEEE transactions on ultrasonics, ferroelectrics, and frequency control*, 2020.

[45] A. H. Abdi, C. Luong, T. Tsang, J. Jue, K. Gin, D. Yeung, D. Hawley, R. Rohling, and P. Abolmaesumi, "Quality assessment of echocardiographic cine using recurrent neural networks: Feasibility on five standard view planes," in *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, vol. 10435 LNCS, pp. 302–310, Springer Verlag, 2017.

[46] F. T. Dezaki, Z. Liao, C. Luong, H. Girgis, N. Dhungel, A. H. Abdi, D. Behnami, K. Gin, R. Rohling, P. Abolmaesumi, and T. Tsang, "Cardiac Phase Detection in Echocardiograms with Densely Gated Recurrent Neural Networks and Global Extrema Loss," *IEEE Transactions on Medical Imaging*, vol. 38, pp. 1821–1832, aug 2019.

[47] K. Cho, B. Van Merriënboer, C. Gulcehre, D. Bahdanau, F. Bougares, H. Schwenk, and Y. Bengio, "Learning phrase representations using RNN encoder-decoder for statistical machine translation," in *EMNLP 2014 - 2014 Conference on Empirical Methods in Natural Language Processing, Proceedings of the Conference*, pp. 1724–1734, Association for Computational Linguistics (ACL), jun 2014.

[48] O. Ronneberger, P. Fischer, and T. Brox, "U-net: Convolutional networks for biomedical image segmentation," in *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, vol. 9351, pp. 234–241, Springer Verlag, may 2015.

[49] E. Smistad, A. Ostvik, B. O. Haugen, and L. Lovstakken, "2D left ventricle segmentation using deep learning," in *IEEE International Ultrasonics Symposium, IUS*, IEEE Computer Society, oct 2017.

[50] 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, pp. 2198–2210, sep 2019.

[51] E. Smistad, I. M. Salte, A. Ostvik, S. Leclerc, O. Bernard, and L. Lovstakken, "Segmentation of apical long axis, four- and two-chamber views using deep neural networks," in *IEEE International Ultrasonics Symposium, IUS*, vol. 2019-Octob, pp. 8–11, IEEE Computer Society, oct 2019.

[52] C. M. Augustin, A. Neic, M. Liebmann, A. J. Prassl, S. A. Niederer, G. Haase, and G. Plank, "Anatomically accurate high resolution modeling of human whole heart electromechanics: A strongly scalable algebraic multigrid solver method for nonlinear deformation," *Journal of Computational Physics*, vol. 305, pp. 622–646, jan 2016.

[53] L. Liu, J. Zhang, R. He, Y. Liu, Y. Wang, Y. Tai, D. Luo, C. Wang, J. Li, and F. Huang, "Learning by Analogy: Reliable Supervision from Transformations for Unsupervised Optical Flow Estimation," *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, pp. 6488–6497, mar 2020.

[54] J. Hur and S. Roth, "Self-Supervised Monocular Scene Flow Estimation," *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, pp. 7394–7403, apr 2020.

[55] E. Ilg, N. Mayer, T. Saikia, M. Keuper, A. Dosovitskiy, and T. Brox, "FlowNet 2.0: Evolution of Optical Flow Estimation with Deep Networks," *Proceedings - 30th IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2017*, vol. 2017-January, pp. 1647–1655, dec 2016.

[56] E. Evain, K. Faraz, T. Grenier, D. Garcia, M. De Craene, and O. Bernard, "A Pilot Study on Convolutional Neural Networks for Motion Estimation from Ultrasound Images," *IEEE Transactions on Ultrasonics, Ferroelectrics, and Frequency Control*, vol. 67, pp. 2565–2573, dec 2020.

[57] N. Duchateau, A. P. King, and M. De Craene, "Machine Learning Approaches for Myocardial Motion and Deformation Analysis," vol. 6, p. 190, jan 2020.

Background

2

This chapter aims to provide the unfamiliar reader the necessary prerequisites for understanding the relevant concepts and terminology of this work. A brief introduction to ultrasound and echocardiography are given initially, together with a fundamental description of myocardial function imaging. Further, a technical overview of deep learning and motion estimation for image analysis is included.

2.1 Ultrasound

Ultrasound is defined as mechanical waves with frequencies higher than the upper limit of the human audible range [1]. Mechanical waves transfer energy through a medium by oscillation of its particles. Ultrasound (US) images are made by transmitting waves into a medium using a transducer, wherein they propagate and scatter as a result of discontinuity in acoustic impedance. The energy of the backscattered US is registered on a transducer surface, and used to form images based on the intensity of the backscattered echo and the presumed location of the scatterers. The waves propagate with a velocity

$$c = \sqrt{\frac{K}{\rho}}, \tag{2.1}$$

where K is the bulk modulus and ρ is the density of the medium. This velocity c is often referred to as the speed of sound, and its relationship to the waves frequency f and wavelength λ is

$$c = \lambda f. \tag{2.2}$$

US can traverse fluids and most soft tissues, but has difficulties with bone and air. This is mainly due to the large difference in impedance and attenuation properties. For water and blood, and most soft tissue in the human body, the speed of sound is often set to roughly $1540\,\mathrm{m\,s^{-1}}$. Assuming a constant wave velocity, it is possible to estimate the distance z between the source of the scatterer and the probe,

$$z = \frac{ct}{2}, \tag{2.3}$$

where t is the time from transmission. The ability to distinguish small structures, i.e. the axial resolution, is dependent on the pulse length, which is inverse proportional to the frequency of the transmitted pulse. Hence higher frequency gives better spatial resolution.

As the US waves propagate through a medium, the energy attenuates due to absorption, scattering loss and spreading. The attenuation is also frequency dependent, reducing the waves energy with higher frequency. This effect limits the penetration depth and results in a compromise between resolution and the size of the sonified region.

Imaging

Conventional US imaging is a pulse-echo technique, which means that the US waves are transmitted from the same transducer (commonly called probe) as they are received. The acoustic waves, or beam, are typically generated and detected by an array of elements situated on this transducer. These are commonly made up by piezoelectric crystals that exploits the piezoelectric effect for converting energy from mechanical stress to an electric charge and vice versa. An alternative concept is the capacitive micromachined ultrasonic transducer (CMUT), where the energy is converted due to change in capacitance [2].

Images are formed by steering the beam over a region of interest (scanning), and registering the backscattered signal. Steering is mainly done electronically using the transducer array, but applications of mechanical steering also exist. Scanning can be performed in several ways, for instance sector scanning that uses transmission delays on the elements to focus and steer the beam in a desired direction. This is often referred to as phased array scanning, and is the most widely used imaging technique in cardiac

US. Further, we have different modalities in US with properties facilitating tissue or blood imaging. The most common is brightness mode (B-mode), which is two-dimensional gray scale images of tissue. Motion mode (M-mode) is another tissue imaging technique, which displays the envelope signal along a specific beam direction over time. In addition, numerous Doppler modalities exist, such as continuous wave (CW), pulsed wave (PW) Dopper and color flow imaging (CFI). In this thesis the focus is B-mode imaging.

2.2 Echocardiography

The human heart is a muscular organ situated behind the sternum, slightly offset to the left side of the chest. Its main objective is to pump blood through the circulatory system, providing oxygen and nutrients to the body and removing waste products such as carbon dioxide through the lungs. The pace of the pump in a normal resting adult is generally between 40 and 120 beats per minute (bpm).

The heart is enclosed by a sac called the pericardium, and the heart wall consists of the layers endocardium, myocardium and epicardium. The myocardium is the cardiac muscle, built up by muscle and pacemaker cells. An illustration of the heart can be seen in Fig. 2.1. The adult heart has an average long axis length of approximately 12 cm [3], width of 8.5 cm and thickness of 6 cm. This is significantly affected by factors such as age, gender and physical activity. The heart consists of four chambers, two on the right and left side for the pulmonary and systemic circulation respectively. Both sides have one atrium, which receives blood, and a ventricle which ejects blood into the body and lungs. The chambers are separated by septum walls of cardiac tissue and four valves.

The cardiac cycle is periodic and consists of the diastolic and systolic phases. Diastole is the period of muscle relaxation and refilling of blood, initiated by the semilunar (SL; aortic and pulmonary valve) valves closing and an isovolumetric relaxation phase where all valves are closed. When the atrial pressure rises above the ventricular pressure, the atrioventricular (AV; mitral and tricuspid valve) valves open and blood flows passively into the ventricles. This is followed by electrical depolarization of the atrium, where the atria starts to contract forcing additional blood to flow across the

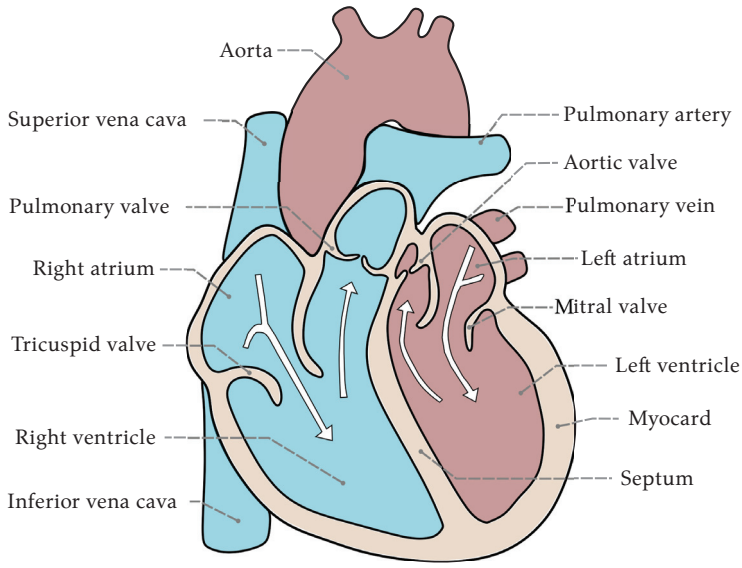


Figure 2.1: Schematic of the human heart showing the four chambers, valves and other relevant landmarks.

AV valves. After the atria contraction ends, the pressure drops and the AV valves close. The next stage is systole, which is the ventricular contraction phase with blood pumping out of the heart and into the body and lungs. It is initiated with the isovolumetric contraction, which begins with the ventricular depolarization and all valves closed. The pressure then increases in the ventricle till it exceeds the aortic and pulmonary pressure, and the ejection stage starts with the SL valves opening and blood pumping out of the ventricles. Towards the end of the ejection stage, the ventricle repolarizes resulting in a reduction in tension and pressure generation till the SL valves close and the cardiac cycle is back to start.

Echocardiography is US imaging of the heart. In principle, it does not differ from other types of diagnostic US, but the intrinsic properties of the organ and its location requires some specialized adaption. The most common way of acquiring US images of the heart is by placing a probe on intercostal locations of a subject. This procedure is called transthoracic echocardiography (TTE). An examination is typically composed of image sequences from different standardized probe postures, referred to as acoustic windows. An example is the apical window, where the probe is positioned close to the apex of the heart, usually in the 5th intercostal space,

and oriented along its long axis. See Fig. 2.2 for an illustrative sketch, where some relevant landmarks are indicated. Different probe postures in the same window results in US images from different planes called views. For an echocardiographic examination, the views are often standardized with semantic names.

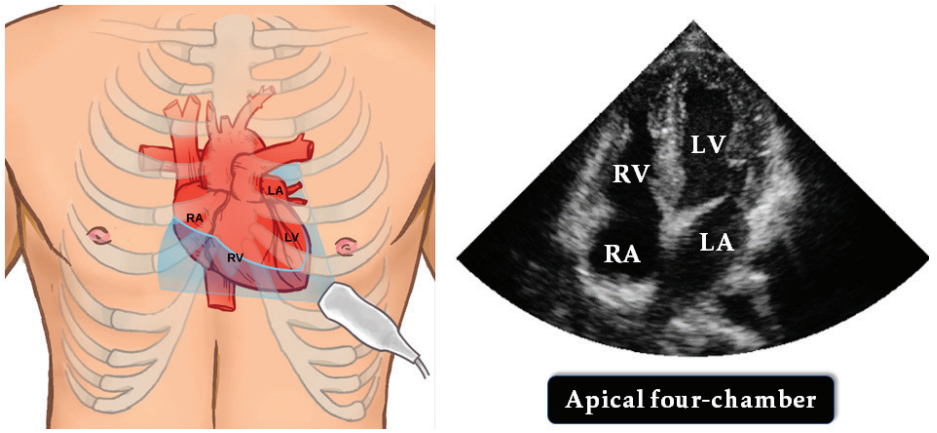


Figure 2.2: Illustration of probe posture when acquiring an ultrasound acquisition of the apical four-chamber view. Corresponding US image is shown on the right side. Left/right ventricle (LV, RV) and left/right atrium (LA, RA) are indicated. Illustration courtesy of H. E. Mørk (www.helemork.com).

The intercostal area, i.e. the spacing between the ribs, does not permit the use of US probes with large apertures. Phased array transducers with electrical steering and a small footprint are typically used for adult echocardiography. Due to the compromise between penetration and resolution, the frequency range is roughly 2-4 MHz, reaching depth covering the whole heart. Another important factor to consider in echocardiography is the dynamics, and to avoid underestimation of tissue motion or similar, the frame rate must be sufficient. For adult 2D echocardiography, the frame rate is typically in the range 45-100 Hz for high-end scanners.

In echocardiography several factors can corrupt the image quality. This can often be traced back to fundamental physical phenomenons, natural artefacts and system design. A selection of noise sources includes reverberations, aberrations, acoustic shadows, specular reflections, attenuation, as well as side- and grating lobes [4, 5]. In US imaging we assume that only one scattering process occurs during wave propagation, however, in reality

the wave will be scattered multiple times and the backpropagated signal from a specific scatterer will also be received several times. This is called reverberation, and will cause a ghosting effect that degrades the contrast of the image. Another assumption is that the sonified medium is homogeneous with constant speed of sound. As mentioned earlier, the speed of the US wave is dependent on the properties of the medium which will vary in practice. This causes a distortion of the wavefront, so-called aberration, as it travels through different types of tissue. Aberration degrades the focus convergence, and makes the resolution worse as a result of broadened mainlobe. Further, attenuation can for instance be caused by the frequency dependency, and a shift of center frequency as a result of wave propagation. Acoustic shadows are characterized by a signal void behind structures that either strongly absorb or reflect the waves. Specular reflections caused by the ribs are also assumed to cause hazy clutter noise in the image [6].

2.2.1 Myocardial function imaging

Myocardial function imaging often refers to the principle of quantifying the cardiac muscle function by velocity and deformation measurements by a medical imaging modality. In echocardiography, strain is used to describe local shortening, thickening and lengthening of the myocardium [7], and is commonly measured by tissue Doppler imaging (TDI) or speckle tracking echocardiography (STE). Myocardial strain measurements using TDI was first introduced in the late 1990s [8, 9], while STE later emerged as the most widely used technique [10]. The dominance of STE compared to TDI can mainly be attributed to less angle dependency and opportunity for measurements on conventional B-mode images.

Strain is the deformation produced by the application of a stress, which is the force per unit cross-sectional area [11]. Strain is dimensionless, and represents the fractional change from the original to the unstressed state. As illustrated in Fig. 2.3, strain allows for the investigation of different spatial components of the contractile function along the anatomical directions of the ventricle. This includes longitudinal strain (LS), circumferential strain and radial strain. LS is computed along the long axis of the ventricle, while circumferential along the short axis. Perpendicular to LS, the radial strain is computed between the endo- and epicardial borders of the ventricle.

The term strain originates from continuum mechanics, where it is used to

describe deformation of an object normalized to its original shape and size. There are several ways to describe this, the most common being Lagrangian or Eulerian (often called natural) strain. Both are illustrated in Fig. 2.3. Lagrangian strain ε corresponds to fractional change of length L at a given point in time t with respect to some reference length L_0. L_0 can for instance be the longitudinal ventricular length at end diastole. The Eulerian strain ε_E is the sum of ratio between the instantaneous deformation and the length, i.e. the instantaneous length change. They can be found by the formulas,

$$\varepsilon(t) = \frac{L(t) - L_0}{L_0} \qquad \varepsilon_E(t) = \int_{t_0}^{t} \frac{1}{L(t)} \frac{dL(t)}{dt} dt = \ln\left(\frac{L(t)}{L_0}\right). \tag{2.4}$$

The relationship between Lagrangian and Eulerian strain is given by,

$$\varepsilon(t) = \exp(\varepsilon_E(t)) - 1. \tag{2.5}$$

In clinical practice Lagrangian strain with STE is most common. Eulerian strain is often used with TDI.

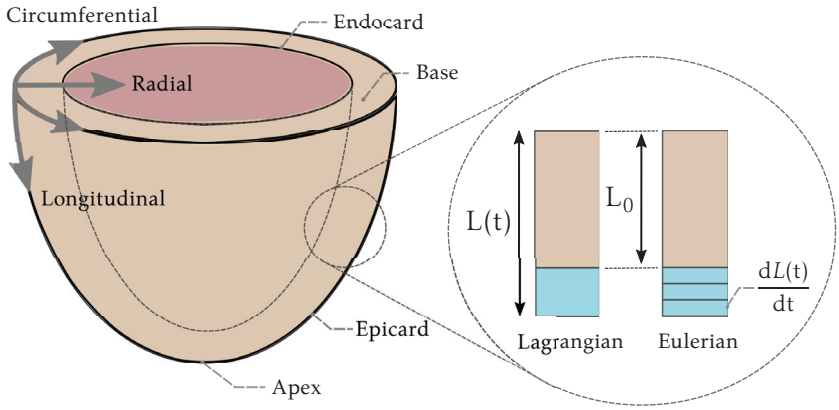


Figure 2.3: Illustration of a heart ventricle. The apex at the bottom cap, and base on top. The myocardium is delimited by the epicard and endocard borders. To the top left, the different anatomical directions are indicated. On the right a representation of the difference between Lagrangian and Eulerian strain is shown.

As with ejection fraction, strain measured by echocardiography is also load dependent and is thus not able to describe the true myocardial contractility. These clinical parameters alone can not be used to fully describe the cardiac function.

2.3 Deep learning and neural networks

In our modern society, machine learning algorithms are embedded into many aspects of our everyday life. From low risk recommendation services for music and television, to the aid of health personnel making critical decisions in medical procedures. Machine learning is a broad term in the subfield of *artificial intelligence* (AI), and is defined as algorithms that can learn from data. They include well known algorithms such as decision trees, support vector machines and regression analysis [12]. In recent years the field has been dominated by an approach called deep learning, which is a family of representation learning algorithms. We can refer to the latter as the automatic formation of useful representations from data.

The building blocks of learning algorithms usually consist of (i) specification of data, (ii) an objective function, (iii) an optimization procedure and (iv) a model [13]. Further, we can differentiate between four types of learning, namely supervised, semi-supervised, unsupervised and reinforcement. In this work supervised learning was most relevant, thus being the main focus of this chapter. Supervised learning involves using labeled data to conduct the learning process and produce a model that takes some input $\boldsymbol{x}$, for example an image, and map it to a paired output $\boldsymbol{y}$. The output $\boldsymbol{y}$ is often referred to as label or ground truth [14].

In DL, the cardinal model is called *feedforward neural network* (NN) [13]. The main goal of a NN is to approximate a function f, by defining a mapping $\boldsymbol{y} = \boldsymbol{f}(\boldsymbol{x}; \boldsymbol{\Theta})$ and learning the parameters $\boldsymbol{\Theta}$ that results in the best estimation. The input data flows through the function of intermediate calculations to the final output with no feedback connection, hence *feedforward*. With feedback the model is called *recurrent neural network* (RNN). A feedforward neural network can be divided into three overarching components; an input layer accepting some inputs $\boldsymbol{x}$, an arbitrary amount of hidden layers and the output layer producing the prediction $\hat{\boldsymbol{y}}$. Most hidden layers can be described as an affine transformation followed by element-wise application of an activation function σ. This is defined as

$$\boldsymbol{f}(\boldsymbol{x}; \boldsymbol{W}, \boldsymbol{b}) \stackrel{\text{def}}{=} \sigma(\boldsymbol{W}^{\mathrm{T}}\boldsymbol{x} + \boldsymbol{b}), \tag{2.6}$$

where $\boldsymbol{W}$ are the weights and $\boldsymbol{b}$ is the bias. For a neural network of K layers,

this is a composite function or *network* [15],

$$\hat{\boldsymbol{y}} = \boldsymbol{f}(\boldsymbol{x}) = (\boldsymbol{f}_K \circ \boldsymbol{f}_{K-1} \circ \cdots \circ \boldsymbol{f}_0)(\boldsymbol{x}) = \boldsymbol{f}_K(\boldsymbol{f}_{K-1}(\cdots(\boldsymbol{f}_0(\boldsymbol{x})\cdots)), \tag{2.7}$$

where $\hat{\boldsymbol{y}}$ are the predictions. Every function $\boldsymbol{f}_i$ possesses its own parameters $\boldsymbol{W}_i$ and $\boldsymbol{b}_i$, often referred to as learnable parameters $\boldsymbol{\Theta} = \{\boldsymbol{W}_0, \boldsymbol{b}_0, ..., \boldsymbol{W}_K, \boldsymbol{b}_K\}$. The number of layers correspond to model *depth*, and it is from this terminology that the term *deep* originates.

Activation functions

Activation functions are important in order for a network to learn complex patterns in the inputs. Most practical problems have a high degree of complexity, and to that end the functions need to be non-linear. Without a non-linear activation function the neural network would become a combination of linear functions, and thus linear itself. They also restrict the output values from the layer to a certain limit.

One of the most widely used activation functions for DL is the rectified linear unit (ReLU) [16]. It is defined as,

$$\sigma_{\text{ReLU}}(x) = \max(0, x) = \begin{cases} x & \text{if } x > 0, \\ 0 & \text{otherwise.} \end{cases} \tag{2.8}$$

One issue with this activation function is that it is zero for all negative outputs, which potentially causes layers to degenerate and not learn anything. Two improvements that allows for positive gradients when the output is negative is the leaky ReLU and the parametric ReLU (PReLU). They are defined as $\sigma(x) = \max(ax, x)$, where a is a preset constant for leaky ReLU and a learnable parameter for PReLU [17, 18].

For the output layer, a sigmoid or softmax (SM) function is commonly used. For n classes, the softmax function for class i is defined as,

$$\sigma_{\text{SM}}(x)_i = \frac{\exp(x_i)}{\sum_j^n \exp(e^{x_j})}, \tag{2.9}$$

where x is the input. This will result in a categorical distribution of the n different classes with values ranging from zero to one, with the sum of all class predictions equal to one.

Loss

Learning algorithms in ML rely on optimization of an objective function, e.g. "minimize the mean squared error loss". The function we want to minimize is called the loss function, which computes the error on a single example. The cost function is typically the average over several examples, or the whole dataset. These terms are often used interchangeably.

The choice of loss function is an essential part of training a neural network, and depends on the problem at hand. For classification tasks a commonly used loss function is cross-entropy (CE), where the idea is to give a logarithmic penalty while training based on how far the predicted class is from ground truth. CE is defined as

$$L_{\text{CE}}(\boldsymbol{y}, \hat{\boldsymbol{y}}) = -\sum_{i=1}^{n} y_i \cdot \log(\hat{y}_i), \tag{2.10}$$

where $\boldsymbol{y}$ is the ground truth, $\hat{\boldsymbol{y}}$ is the class prediction and n is the number of classes. For multiclass classification, i.e. each sample belongs to one class, categorical CE is usually employed. Here, the softmax function is used as the final activation of $\hat{\boldsymbol{y}}$, which gives

$$L_{\text{CCE}}(\boldsymbol{y}, \hat{\boldsymbol{y}}) = -\sum_{i=1}^{n} y_i \cdot \log(\sigma_{\text{SM}}(\hat{y})_i) = -\log\left(\frac{\exp(\hat{y}_p)}{\sum_j^n \exp(e^{\hat{y}_j})}\right), \tag{2.11}$$

where $\hat{y}_p$ is the prediction corresponding to the GT class. Note that for a multiclass classification, only the GT term is nonzero and kept in the final expression.

While the classification task seeks to automatically assign a label to an unlabeled example, regression is a problem of predicting values. Mean absolute error (MAE), for instance, is a measure of error between the true and predicted values for a specific example. It can be formulated as

$$L_{\text{MAE}}(\boldsymbol{y}, \hat{\boldsymbol{y}}) = \frac{1}{m}\sum_{i=1}^{m} |\boldsymbol{y}_i - \hat{\boldsymbol{y}}_i|, \tag{2.12}$$

where m is the length of the output vectors. MAE is often used as a loss function for regression problems.

Backpropagation and optimization

Neural networks learn using gradient-based algorithms. The learning consists of two parts called backpropagation and optimization. Backpropagation refers to the method of computing the gradient of the loss function $\nabla_{\boldsymbol{\Theta}} L(\boldsymbol{y}, \hat{\boldsymbol{y}}; \boldsymbol{\Theta})$ with respect to the parameters $\boldsymbol{\Theta}$. The optimization performs the learning using this gradient. Gradient descent, for instance, is a way to optimize a loss function L parameterized by the models parameters $\boldsymbol{\Theta}$. This happens by updating the parameters in the opposite direction of the gradient of the loss function. We separate into three different variants based on the amount of data used for each update; gradient descent, stochastic gradient descent (SGD) and mini-batch SGD. Gradient descent utilizes the whole dataset for one single update of parameters, while SGD performs an update for every example. In DL practice, the most common is mini-batch SGD, which performs updates for every N training examples. The common size of N range from 4 to 256, but varies significantly between different applications and model capacity.

Gradient descent algorithms proceeds in *epochs*, which correspond to using the entire training set once to update the learnable parameters. The learnable parameters $\boldsymbol{\Theta}$ are initialized before the first epoch, typically by randomized values, zeros or with some known distribution. A few common initialization methods include He and Glorot/Xavier [18, 19]. Moving on to the training, the learnable parameters are, as mentioned, updated using the gradient of the loss and the output after forward propagation. This can be formulated as

$$\boldsymbol{\Theta} \leftarrow \boldsymbol{\Theta} - \alpha \nabla_{\boldsymbol{\Theta}} L(\boldsymbol{\Theta}) \quad \Longrightarrow \quad \boldsymbol{W} \leftarrow \boldsymbol{W} - \alpha \frac{\partial L}{\partial \boldsymbol{W}}, \qquad \boldsymbol{b} \leftarrow \boldsymbol{b} - \alpha \frac{\partial L}{\partial \boldsymbol{b}}, \tag{2.13}$$

where the learning rate α controls the size of the update.

There are several challenges with SGD that limits good convergence. This includes choosing a proper learning rate and avoiding suboptimal local minima as a result of highly non-convex loss functions. One problem with SGD is related to poor updates and oscillations in parts of the loss surface with large curvatures in one direction. To accelerate learning and limit such situations, incorporating momentum in the SGD has been proposed [20, 21]. The momentum algorithm introduces a new variable $\boldsymbol{\nu}$ that accumulates an exponential decaying moving average of previous gradients. The new update

rule is

$$\boldsymbol{\nu} \leftarrow \eta \boldsymbol{\nu} - \alpha \nabla_{\boldsymbol{\Theta}} L(\boldsymbol{\Theta}), \tag{2.14}$$

$$\boldsymbol{\Theta} \leftarrow \boldsymbol{\Theta} + \boldsymbol{\nu}. \tag{2.15}$$

The contribution of the accumulated gradients are determined by a preset factor η relative to the learning rate α. The parameter update will then be dependent of the previous gradients and their alignment.

SGD with momentum is one of many ways of improving the optimization process with respect to the mentioned limitations. An extensive description of alternative methods is out of scope for this thesis, but we want to briefly mention the adaptive moment estimation (Adam) [22]. It is one of the most popular in practice, and combines the use of adaptive learning rate and momentum. The Adam optimizer stores an exponentially decaying average of both the past gradient and the past squared gradient. The prior being similar to momentum. The first moment estimate $\boldsymbol{m}_t$ and second moment estimate $\boldsymbol{v}_t$ is calculated as,

$$\boldsymbol{m}_t = \beta_1 \boldsymbol{m}_{t-1} + (1 - \beta_1) \cdot \nabla_{\boldsymbol{\Theta}} L(\boldsymbol{\Theta}) \tag{2.16}$$

$$\boldsymbol{v}_t = \beta_2 \boldsymbol{v}_{t-1} + (1 - \beta_2) \cdot \nabla_{\boldsymbol{\Theta}} L(\boldsymbol{\Theta}) \odot \nabla_{\boldsymbol{\Theta}} L(\boldsymbol{\Theta}), \tag{2.17}$$

where β_1 and β_2 are the exponential decay rates. To correct for the induced bias by initialization the moments with zeros, corrected bias estimate for the first and second moment is calculated as

$$\hat{\boldsymbol{m}}_t = \frac{\boldsymbol{m}_t}{1 - \beta_1^t}, \qquad \hat{\boldsymbol{v}}_t = \frac{\boldsymbol{v}_t}{1 - \beta_2^t}. \tag{2.18}$$

Finally, the parameters are updated according to

$$\boldsymbol{\Theta} \leftarrow \boldsymbol{\Theta} - \alpha \frac{\hat{\boldsymbol{m}}_t}{\sqrt{\hat{\boldsymbol{v}}_t} + \epsilon}, \tag{2.19}$$

where ϵ is a small constant used for numerical stability. The second term makes the learning rate adaptive by modification based on past gradients.

2.3.1 Convolutional neural networks

Convolutional neural networks (CNNs) are a type of DL algorithms that extensively employ convolutional operations in combination with other

characteristic methods. The key idea is to exploit local connections in the data at several levels of context resolution together with concepts such as pooling and sharing of learnable parameters. CNNs are especially effective for data with a grid-like topology, such as 2D images.

CNNs have been used successfully for many applications since the early 1990s, but was somewhat forsaken by the mainstream CV community until the ImageNet competition in 2012 [23]. Here, Krizhevsky *et al.* proposed the use of a deep CNN called AlexNet, which almost halved the error rates compared to the best competing image recognition methods [24]. Their success can to some extent be attributed to efficient use of graphics processing units (GPU), ReLU activation, as well as regularization methods such as *dropout* and *data augmentation*. Now, CNNs constitute the state of the art for almost any image recognition and detection task, and is dominating the CV field.

Convolution operation

The convolution operation can be used in place of the general matrix multiplication in the standard NN mentioned earlier. For the two-dimensional image I, the 2D convolution is given as

$$S(i,j) = (I * K)(i,j) = \sum_m \sum_n I(m,n)K(i-m, j-n), \tag{2.20}$$

where K is the *kernel*, and S is referred to as the *feature map*. The convolution operator can be extended to any dimensions, for instance to 3D, which is relevant for volumetric and video data.

In the context of ML, the learning algorithm will learn the appropriate values of the kernel. In a standard NN, each weight is multiplied once by an element of the input, and never reused. For CNNs, however, we make use of the concept of shared weights. This means that the same weights (or kernel) is used for more than one function in a model, e.g. same kernel for each position in an image or feature map. This means that, rather than learning one set of weights for each location in an image, we learn one set to be used on the entire image. The latter causes equivariance to translation.

Pooling

Pooling refers to a statistic downsampling of the different neighbourhoods in a feature map, which commonly is an essential part of a CNN. The feature

map is partitioned into several windows of a specified size, and for each window a filtering process is performed where the output is a reduced representation. The most common is max pooling, which outputs the maximum value of the window. A common alternative is average pooling, where the values of the window is averaged. The main role of pooling is to reduce the spatial resolution, and hence increase the global context. This reduces the memory consumption and number of computations, as well as the capability of invariance to small translations in the input.

Regularization

One of the recurring challenges in DL is to get the algorithms to generalize, i.e. achieve good performance on unseen data. DL model performance are often measured by how well it performs on test data, and its ability to make the training and test error as small as possible, as well as the gap between them. If the model is not able to achieve sufficiently low error on the training data, we call it *underfitting*. *Overfitting*, on the other hand, happens when the gap between training error and test error becomes very large. Common ways to tackle this is by increasing the dataset used for training and reduce the number of learnable parameters (representational capacity) of the neural network. Other strategies that aim to increase generalization, but with limited expense to the training error, are collectively referred to as regularization techniques. A lot of approaches exist, some by adding direct penalties to the cost function or learnable parameters, other emphasise manipulating the data or training procedure.

Weight decay One of the simplest and most common parameter penalties are the L^2 and L^1 norms, which is a regularization strategy that seeks to bring the weights closer to the origin by adding a term proportional to the weights to the cost function [13]. This is also known as weight decay, and can be added individually to each layer.

Dropout The core idea of *dropout* is to randomly zero out (drop) the output of hidden neurons in a neural network with a given probability per input [25]. These neurons will not contribute to the forward pass, nor be included in the backpropagation. For every sample this corresponds to randomly sampling a network with a different weight composition,

effectively forcing the feature extractors to become more robust. Dropout can be implemented into many model types, but fully connected layers are the most relevant. For convolution layers it will introduce a random noise effect, but can not guarantee weights to be excluded from updates due to the inherent cross dependency composition of the matrix as a result of the convolution operation.

Data augmentation Artificially enlarging the dataset using input transformations that preserve labels is an effective way of reducing the generalization error. The broad term is *data augmentation*, and includes numerous methods such as image flipping, color manipulation, scaling, rotation and more. The transformations are applied either offline upon training, or online while training with a given probability. Adding noise to the input can also be regarded as a form of data augmentation. The total effect is a larger data representation that limits overfitting, but it can also make the models more robust to real-life artefacts.

Early stopping Through the course of training, a common scenario is that the training and validation error decrease simultaneously in a correlated fashion, but at some point the two metrics starts to diverge with the validation error rising again. Returning to a parameter selection at an earlier time step would then yield a model with better validation accuracy. This strategy is called *early stopping*, and is a very popular regularization technique in DL. In practice, the common way is to track the validation accuracy between epochs and store the best performing model. The patience (number of epochs) of the stopping routine is set as a hyperparameter, and determines how many training epochs should be conducted without any improvement before ending the learning procedure.

Batch normalization

Batch normalization (BN) was proposed by Ioffe and Szegedy in 2015 as a transform to improve the training procedure by normalizing the inputs of a layer [26]. Empirical advantages have included faster convergence and more stable training. Due to practical limitations with stochastic optimization and impracticality retrieving global information, the normalization is restrained to each mini-batch B of size m. The mini-batch mean μ_B and variance σ_B is

then given as,

$$\mu_B = \frac{1}{m}\sum_{i=1}^{m} x_i \qquad \sigma_B^2 = \frac{1}{m}\sum_{i=1}^{m}(x_i - \mu_B)^2. \tag{2.21}$$

The normalized $\hat{x}$ input is then,

$$\hat{x}_i = \frac{x_i - \mu_B}{\sqrt{\sigma_B^2 + \epsilon}}, \tag{2.22}$$

where ϵ is an arbitrarily small value added to the denominator for numerical stabilization. The BN transform is then defined as

$$y_i = \gamma \hat{x}_i + \beta \equiv \mathrm{BN}(x_i; \gamma, \beta), \tag{2.23}$$

where γ and β are learned parameters from the optimization procedure, and y is the output passed to other network layers.

For inference the BN transform is modified to use the expected population mean and variance,

$$\mathrm{E}(x_i) = \mathrm{E}_B(\mu_B) \qquad \mathrm{Var}(x_i) = \frac{m}{m-1}\mathrm{E}_B(\sigma_B^2). \tag{2.24}$$

Substituting this into (2.23) gives

$$\mathrm{BN}(x_i; \gamma, \beta) = \frac{\gamma x_i}{\sqrt{\mathrm{Var}(x_i) + \epsilon}} + \left(\beta - \frac{\gamma \mathrm{E}(x_i)}{\sqrt{\mathrm{Var}(x_i) + \epsilon}}\right), \tag{2.25}$$

which is a linear transform of x_i.

2.3.2 Network architectures

Following the success of Krizhevsky *et al.* in the ImageNet competition in 2012 [24], advancing research on CNNs have steadily improved results on various image classification tasks. Initially, the trend was cursory addition of layers, which led to increased problems with vanishing gradients and impractical growth in resource demands [27]. This motivated researchers to discover new ways of effectuating network architectures.

One of the most influential proposals came from the authors of "Network In Network" (NIN) [28], who suggested the use of convolutions with kernel

size (1×1) to combine features between layers. Feature pooling or bottlenecks, proved to increase effectiveness per feature, and became a viable option for parameter reduction. Another interesting finding stems from the classification part of their network. The feature maps were spatially averaged, instead of adding the more typical fully connected layers. Further, the output was fed directly into the softmax activation. This reduces the parameter count, and it is also claimed to make the network less prone to overfitting.

The key insights from NIN inspired Szegedy *et al*. [27] creating the Inception architecture (introduced as GoogleNet). The principal difference to other networks was the building blocks referred to as Inception modules. Each of these blocks consist of parallel routes of convolutions with varying kernel size, in addition to a pathway with pooling. Several editions of Inception has been presented after its introduction, and especially the third edition of the Inception architecture sought a lot of attention for significant improvements on benchmark datasets [29]. The fundamental philosophy of parallel routes in depth is the same, and the major architectural difference compared to the original topology is spatial factorization of large spatial filters. Based on this, three different modules are designed and used throughout the network. In the lower parts, i.e where the feature maps are relatively large, the (5×5) convolutions are factorized into two layers of (3×3) convolutions. The other modules utilize asymmetric convolutions, e.g. a (3×1) followed by a (1×3) convolution. In addition to factorization, batch normalization is utilized after convolution layers.

Rethinking connection of layers: ResNet and DenseNet

With increasing network depth, a prevalent problem referred to as degradation may also occur [30]. This is observed by a subpar saturation of training accuracy followed by a rapid degrade. He *et al.* addresses this problem by using a residual connectivity pattern (introduced in the ResNet architecture). Essentially, this involves implementing a connection pattern where the output of a component block/layer is added to the input before further propagation. See Fig. 2.4 for a conceptual example. The connections proved to reduce the degradation effect and allow a forthright backpropagation of gradient signal towards the bottom layers, thus enabling training very deep networks.

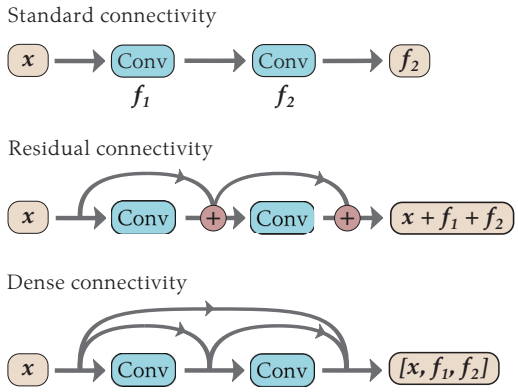


Figure 2.4: Example of three common layer connection patterns. The residual connection pattern is formulated such that layers learn residual functions with reference to the input using identity mappings. Input is added to the convolution block (blue) output. Characteristics of dense connectivity are that every layer has a direct connection to every subsequent layer with the same feature map size. Instead of summation, this pattern relies on channel-wise concatenation.

Short connections between layers have naturally sought attention, for instance, it inspired the proposal of the densely connected convolutional network (DenseNet) [31]. Here, they take the insight a bit further and concatenate the output of every preceding component block with equally sized feature maps and use it as input into all subsequent layers. An example of the concept is also illustrated in Fig. 2.4. Such a dense connectivity pattern is claimed to further alleviate the vanishing gradient problem, and perhaps more importantly; it can improve the feature propagation and reusability.

2.3.3 Recurrent neural networks

Recurrent neural networks (RNNs) have shown promising performance on sequential data such as text and speech recognition [32]. What sets RNNs apart from the ordinary feedforward networks, is that they have internal hidden states for retaining temporal information when modelling data over time. It can be considered a feedforward network dependent on all previous states. Extending (2.6) to cover time we get

$$\boldsymbol{f}_t(\boldsymbol{f}_{t-1}, \boldsymbol{x}_t) = \sigma(\boldsymbol{W}^{\mathrm{T}} \cdot [\boldsymbol{f}_{t-1}, \boldsymbol{x}_t] + \boldsymbol{b}), \tag{2.26}$$

where t is the timestep. Despite being powerful in theory, the vanilla RNN struggles with practical difficulties of long-term dependencies while

training [33, 34]. This is especially prevalent if the temporal gap between the relevant information and the point of prediction becomes very large. The main reasons for the problem is vanishing or exploding gradients. Fortunately, several methods have been proposed to alleviate the issue with long-term dependencies.

Long short term memory

A special type of RNN explicitly designed to tackle the long-term dependency problem is the Long Short-Term Memory (LSTM) networks [35]. A key components of the LSTM is the cell state c and the three gates, namely input gate, forget gate and output gate. An illustration of the LSTM module can be seen in Fig. 2.5. The information added to the cell state is regulated by the semantically named gates. The forget gate controls what to keep from the previous cell state $\boldsymbol{c}_{t-1}$ by element-wise multiplication with a gate factor $\boldsymbol{g}_{\mathrm{f}}$, which is dependent on the input $\boldsymbol{x}_t$, the previous output $\boldsymbol{f}_{t-1}$ and some learned parameters $\{\boldsymbol{W}_{\mathrm{f}}, \boldsymbol{b}_{\mathrm{f}}\}$. This can be formulated as

$$\boldsymbol{g}_{\mathrm{f}} = \sigma_{\mathrm{s}}(\boldsymbol{W}_{\mathrm{f}} \cdot [\boldsymbol{f}_{t-1}, \boldsymbol{x}_t] + \boldsymbol{b}_{\mathrm{f}}), \tag{2.27}$$

where σ_{s} is the sigmoid activation function. Further, the input gate decides what information should be added to the cell state by the term

$$\boldsymbol{g}_{\mathrm{i}} = \sigma_{\mathrm{s}}(\boldsymbol{W}_{\mathrm{i}} \cdot [\boldsymbol{f}_{t-1}, \boldsymbol{x}_t] + \boldsymbol{b}_{\mathrm{i}}) \odot \tanh(\boldsymbol{W}_{\mathrm{c}} \cdot [\boldsymbol{f}_{t-1}, \boldsymbol{x}_t] + \boldsymbol{b}_{\mathrm{c}}). \tag{2.28}$$

The output of the hyperbolic tangent $\tanh(\cdot)$ controls the sign of the update. The new cell state $\boldsymbol{c}_t$ is then

$$\boldsymbol{c}_t = \boldsymbol{g}_{\mathrm{f}} \odot \boldsymbol{c}_{t-1} + \boldsymbol{g}_{\mathrm{i}}. \tag{2.29}$$

Finally, the output gate controls the output $\boldsymbol{f}_{\mathrm{t}}$ of the LSTM by a multiplicative weighting between the gate factor $\boldsymbol{g}_{\mathrm{o}}$,

$$\boldsymbol{g}_{\mathrm{o}} = \sigma_{\mathrm{s}}(\boldsymbol{W}_{\mathrm{o}} \cdot [\boldsymbol{f}_{t-1}, \boldsymbol{x}_t] + \boldsymbol{b}_{\mathrm{o}}), \tag{2.30}$$

and the filtered cell state $\boldsymbol{c}_t$. We end up with

$$\boldsymbol{f}_{\mathrm{t}} = \boldsymbol{g}_{\mathrm{o}} \odot \tanh(\boldsymbol{c}_t). \tag{2.31}$$

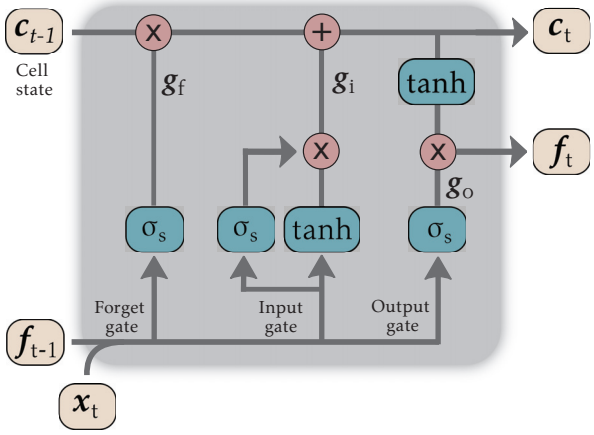


Figure 2.5: Data flow in the Long short-term memory (LSTM) layer. The cell state is updated on the top line, and the different gates are labeled at the bottom. The beige boxes are the inputs and outputs of the layer, while the red boxes are mathematical operations with × representing element-wise multiplication. The blue boxes indicate the weighted transforms with sigmoid activation σ_s and tanh activation. Learnable weights are not shared between the transforms.

The above description is a relatively plain LSTM layer, but several variants exists. For instance, another popular variant in the same methodological family is the gated recurrent unit (GRU), which proposes to combine the input and forget gates into one single update gate, as well as merging the cell state and the output [36].

2.4 Motion estimation

Motion is an integral part of our visual perception, and a rich source of information. Estimating the motion is a crucial task in many applications, including surveillance, action recognition and autonomous vehicles [37–39]. It is also very important in a biomedical context for tasks such as image registration, blood flow estimation and organ deformation [40–42].

Motion is a 3D phenomenon, and a 2D imaging sensor (e.g. camera, ultrasound probe) only captures a projection or slice of the 3D scene onto an image plane. Motion estimation is thus an ill-posed problem in 2D, and accurate estimation remains difficult both due to theoretical and practical limitations [43]. The perspective projection of the true velocity onto the plane is referred to as the *motion field* [44]. Pure motion parallel to this

plane can not be guaranteed, and the data available in 2D is explicitly only the variations of the image brightness pattern. Thus on images, only the apparent motion or *optical flow* (OF) is detectable, i.e. the displacements of the intensity values.

The concept of optical flow was introduced by the psychologist Gibson in the 1940s to describe how motion is fundamental to our visual perception of the world [45]. The term later gained proper foothold in the CV community after the pioneering work by Horn and Schunck [46], as well as Lucas and Kanade [47]. Horn and Schunck defined OF as the apparent motion of the brightness pattern in an image, and used the assumption that pixel intensity remains constant during displacement. In general, OF methods have achieved considerable success in tackling the fundamental problem of estimating the apparent motion $\boldsymbol{v}$ between two adjacent images $\boldsymbol{I}_t$ and $\boldsymbol{I}_{t+1}$. However, there is still several limitations that hampers the performance, including highly variable motion, lack of texture, noise, non-uniform illumination, reflections, transparency and occlusions. The trade-off between minimizing an optimization criteria and regularization have been challenging, especially for corner cases and outliers. Recent progress have been facilitated by learning-based approaches, which can bypass this formulation [48]. DL methods are now dominating the field and have surpassed traditional approaches on common benchmark data [49, 50].

In the following a brief introduction to traditional approaches, including some basic assumptions, key concepts and constraints, is presented. A thorough review of the different methodological branches are outside the scope of this thesis, but the interested reader is referred to relevant literature for a survey [51]. Further, some of the seminal learning-based approaches are described. The latter category of methods is highly influenced by traditional concepts, and we therefore try to draw some lines between them.

2.4.1 Traditional optical flow

Following the work of Horn and Schunk, many approaches for OF computation have been proposed. We often divide them into classes based on their assumptions, constraints and optimization procedure. Common naming includes region based, feature based, differential and energy based

methods [51–53][1]. Among these, the most widely used techniques are variational methods.

Variational methods

A lot of the traditional optical flow methods can be derived from the *brightness constancy constraint*,

$$I(x, y, t) = I(x + \Delta x, y + \Delta y, t + \Delta t). \tag{2.32}$$

Assuming small displacements, this can be developed as a first order Taylor series,

$$I(x + \Delta x, y + \Delta y, t + \Delta t) \approx I(x, y, t) + \frac{\partial I}{\partial x}\Delta x + \frac{\partial I}{\partial y}\Delta y + \frac{\partial I}{\partial t}\Delta t. \tag{2.33}$$

Dividing by Δt and truncating higher order terms yields,

$$\frac{\partial I}{\partial x}\frac{\Delta x}{\Delta t} + \frac{\partial I}{\partial y}\frac{\Delta y}{\Delta t} + \frac{\partial I}{\partial t} = 0 \implies \nabla I \cdot \boldsymbol{v} = -\frac{\partial I}{\partial t}, \tag{2.34}$$

which is referred to as the *optical flow constraint*. The system is underdetermined by being an equation with two unknowns. This is also known as the *aperture problem* of OF, which states that motion can only be determined if neighboring context is taken into account [54]. Therefore, additional assumptions and constraints must be applied for the system to be solvable. This has resulted in a wide range of well-known methods which introduce additional conditions for estimating the displacement, such as the mentioned method by Horn and Schunck. They combine a data term based on the the OF constraint with a global spatial smoothness term in a energy-based formulation, which they minimize. This regularizes neighbouring points to have similar motion profiles [46]. Variational methods includes the range of extensions and modifications made to the original Horn and Schunk formulation.

The above approximation is good when the displacements are low (e.g. less then one pixel), but in real image sequences this is rarely a satisfied condition. To cope with a larger range of motions, a *coarse-to-fine* strategy is often employed [55]. Here, a multi-resolution pyramid, as shown in Fig. 2.6,

[1] In literature, the terminology can be slightly inconsistent and overlapping.

is exploited to extract sub-sampled representations of the original image and hence lower motion fields in a sequence.

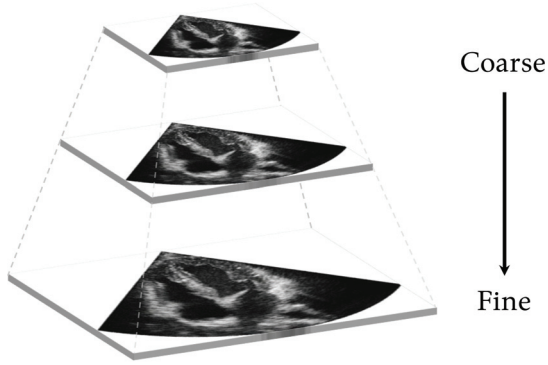


Figure 2.6: Illustration of a coarse-to-fine image pyramid. For image sequences, the pixel displacement between coarser level representations will be lower than the full resolution images.

The analysis typically starts from the coarsest resolution level where the velocity is smallest. The velocity is estimated and projected up to a finer level where it is used to guide the finer resolved computation. At this stage, *warping* is often conducted [56, 57], which involves propagating one of the subsequent images towards the other via forward or backward propagation using the estimated motion. The warped image can then be used to estimate more accurate submotion at the given level in an iterative fashion. In Fig. 2.7 an example of image warping with nearest neighbour interpolation is shown.

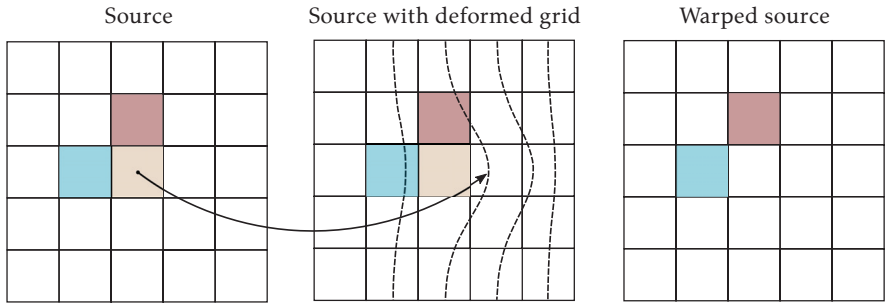


Figure 2.7: Illustration of image warping. The warped representation is constructed by looking up the pixel value of the source image with the given deformed grid (e.g. from the estimated flow). Noticeably, this is a nearest neighbour interpolation, but bilinear and cubic interpolation is often used to obtain intensities at non-integer coordinates.

Region-based matching

Numerical differentiation can by error prone due to noise, large displacement and limited number of frames in the image sequence [51]. An alternative is region-based matching, which aims to locate matching regions between images, and use this to estimate the displacement field $\boldsymbol{v}$. The key idea is to divide the current image $\boldsymbol{I}_t$ into a set of macro blocks with a given size $\boldsymbol{b} = (m, n)$, and compare it to the corresponding block and its spatial neighbourhood in the adjacent image $\boldsymbol{I}_{t+1}$. For matching, a cost function is employed to estimate the similarity between blocks. There are several common functions, such as the normalized cross correlation (NCC), sum of squared differences (SSD) and sum of absolute differences (SAD) [52]. SSD can for instance be formulated as

$$\mathrm{SSD}(\boldsymbol{x}, \Delta\boldsymbol{x}) = \sum_{\Delta\boldsymbol{b}} \left(I_t(\boldsymbol{x} + \Delta\boldsymbol{b}) - I_{t+1}(\boldsymbol{x} + \Delta\boldsymbol{x} + \Delta\boldsymbol{b})\right)^2, \tag{2.35}$$

where $\boldsymbol{x}$ is the pixel location, $\Delta\boldsymbol{b}$ corresponds to the macro block size and $\Delta\boldsymbol{x}$ is the offset. In its most naive implementation, often called exhaustive search, the macro blocks for each point in $\boldsymbol{I}_t$ is compared to every point in the adjacent image. To reduce the number of computations, a smaller search range around $\boldsymbol{x}$ in $\boldsymbol{I}_{t+1}$is often imposed. Regardless, the result is a *cost volume* with total size equal to the multiple of the image size and search range. The offset $\Delta\boldsymbol{x}$ with the lowest value per image location $\boldsymbol{x}$ corresponds to an estimate of the displacement. Further refinement is often performed with subpixel peak detection [58] or similar algorithms. In Fig. 2.8 an example of the cost volume created from the correlation between two images is shown.

Farnebäck method

Another popular algorithm in the family of optical flow is the Farnebäck method [59]. The idea is to approximate some neighborhood of each pixel with a quadratic polynomial. The local intensity model is then

$$\boldsymbol{I}_1(x, y) = \boldsymbol{I}_1(\boldsymbol{x}) = \boldsymbol{x}^T \boldsymbol{A}_1 \boldsymbol{x} + \boldsymbol{b}_1^T \boldsymbol{x} + c_1, \tag{2.36}$$

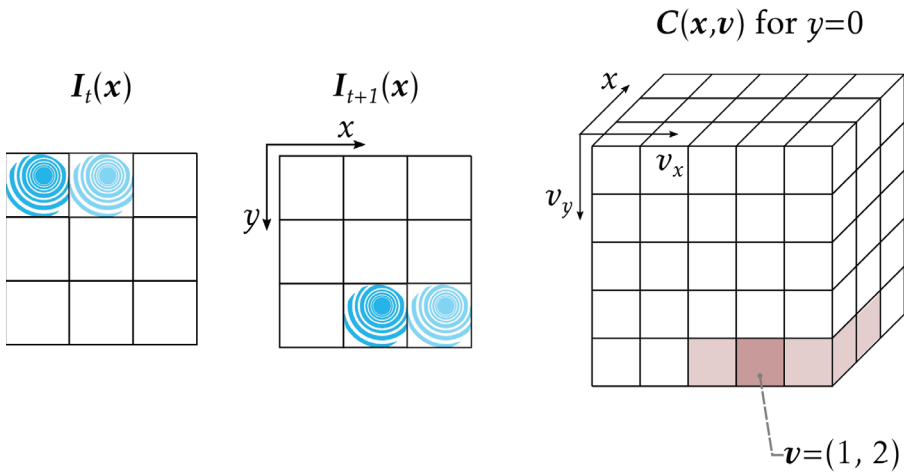


Figure 2.8: Illustration of the cost volume C created by the correlation between the images I_t and I_{t+1}. Here, a search range of ±2 in each direction around every pixel location x is used, and a kernel size of 3×3. The cube illustrates the resulting correlation values for the first row $y = 0$ in the reference image.

where A is a symmetric matrix of size equal to the dimensions of x. Further, b is a vector and c is a scalar. We then construct the model for the adjacent image by the displacement field v,

$$I_2(x) = I_1(x - v) = (x - v)^T A_1 (x - d) + b_1^T (x - v) + c_1, \tag{2.37}$$
$$= x^T A_1 x + (b_1 - 2A_1 v)^T x + v^T A_1 v - b_1^T v + c_1, \tag{2.38}$$
$$= x^T A_2 x + b_2^T x + c_2. \tag{2.39}$$

Which gives the following relationships,

$$A_2 = A_1, \qquad b_2 = b_1 - 2A_1 v, \qquad c_2 = v^T A v - b_1^T v + c_1. \tag{2.40}$$

Noticably, the middle term of the equation above can be solved with respect to v if A_1 is invertible. Solving for the displacement field v gives

$$v = -\frac{1}{2} A_1^{-1} (b_2 - b_1). \tag{2.41}$$

Naturally, the coefficients do not not hold for the entire image, so they are expanded to have a representation for each location, i.e. $A_i(x)$, $b_i(x)$ and

$c_i(\boldsymbol{x})$. In practice $\boldsymbol{A}_1(\boldsymbol{x}) \neq \boldsymbol{A}_2(\boldsymbol{x})$, so we make the approximation

$$\boldsymbol{A}(\boldsymbol{x}) \approx \frac{\boldsymbol{A}_1(\boldsymbol{x}) + \boldsymbol{A}_2(\boldsymbol{x})}{2}. \tag{2.42}$$

Finally, this gives the constraint

$$\boldsymbol{A}(\boldsymbol{x})\boldsymbol{v}(\boldsymbol{x}) = \Delta\boldsymbol{b}(\boldsymbol{x}), \tag{2.43}$$

where $\Delta\boldsymbol{b}(\boldsymbol{x}) = -(\boldsymbol{b}_2(\boldsymbol{x}) - \boldsymbol{b}_1(\boldsymbol{x}))/2$ and $\boldsymbol{v}(\boldsymbol{x})$ is spatially varying. Equation (2.43) is a least squares problem and can be solved accordingly.

The solution of equation (2.43) can be found pointwise, but in practice this can result in suboptimal and noisy results. To make the algorithm more robust numerous extensions and post-processing have been proposed, including the use of neighbourhood windows, parameterized motion model, gaussian intensity smoothing and image pyramids [55].

2.4.2 Learning-based motion estimation

As mentioned, CNNs have been applied successfully to a wide range of computer vision problems, and have become the *de facto* algorithms for modern image analysis. Traditionally, this involved tasks such as classification, object detection and segmentation, and in that regard a natural extension would be optical flow. Early demonstrations applying CNNs for OF estimation used it as a feature extractor substituting the data term in the variational formulation, with similar smoothing and optimization strategies [60, 61]. Concurrently, Fisher *et al.* proposed an end-to-end CNN regression approach for OF estimation which could side-step the classical formulation with energy minimization and regularization [62, 63]. The CNN architectures named FlowNet resulted in a shift in OF research, which in later years have been dominated by DL based methods.

The CNN based methods learn to compute OF from pairs of input images, either supervised with labeled datasets of corresponding motion patterns or unsupervised minimizing a proxy loss. The latter requires careful design of the loss function, but is very beneficial as no annotated data is required for training [64]. So far supervised methods have achieved best results, and we briefly describe some prominent work in the following.

FlowNet

Initially, two architectures named FlowNetSimple (FlowNetS) and FlowNetCorr (FlowNetC) was introduced [62]. Both are end-to-end approaches with an U-Net type of structure [65]. The inputs are two consecutive images, while the output is a dense displacement map. In FlowNetS, the images are concatenated before being fed into the network, while for FlowNetC, the images are fed into two parallel routes of feature extraction. The features of each route is then joined in a correlation layer before further propagation through the network. The encoder part of both networks consists of nine convolution layers, but in FlowNetC a correlation between the output of the two parallel feature extractors occurs after the third layer. They use strides of two in six of the layers for pooling, and ReLU activations. A refinement decoder consisting of deconvolutions layers is also used for both networks. The predicted displacement map is one-quarter size compared to the input, and from that stage a bilinear upsampling is performed for the output to be equal to the input shape. Skip connections are used between the encoder and decoder part. For training they use *endpoint error* (EPE) loss, which is the Euclidean distance between the predicted motion vector and the ground truth. Adam was used as the optimization method, and it was trained on a dataset they created named FlyingChairs. Further, they achieved better performance by fine-tuning on the MPI Sintel dataset [66]. The final results were slightly below state of the art, but demonstrated that CNNs have a significant potential for OF estimation.

Based on their findings, the same group extended their work with a new architecture called FlowNet 2.0 [49], and reported competitive results compared to top ranking methods. It combines five CNNs into a large model, mainly based on the FlowNetC and FlowNetS architectures. Two parallel branches are dedicated to large and small displacements respectively. The branch for large displacements stacks three networks in a cascade, first one FlowNetC followed by two FlowNetS. The branch for small displacements is a modified FlowNetS called FlowNetSD. Here, they reduce the filter size of the first convolution layers, and remove the first stride. They also add convolution layers between the deconvolutions in the refinement part to alleviate issues with noise. The inputs to both branches are two concecutive images, but the intermediate inputs to the FlowNetS models consists of the predicted flow, the pair of images where the second is warped towards the

first with the predicted flow, and the brightness error. The brightness error correspond to the difference between the first image and the second image warped with the flow prediction from the previous CNN. Finally, the outputs of the two branches are fused together in another CNN before the final flow prediction. Here, the input is the first image, as well as the flow, flow magnitude and brightness error from both branches.

FlowNet 2.0 is a very large model that requires sequential training of each subnetwork. They propose a curriculum learning strategy using several different datasets in the training schedule. The conducted ablation studies shows the effect of using multiple datasets, as well as their ordering, and is an important finding regarding the impact of the training schedule.

Spatial pyramid network (SPyNet) and PWC-Net

To address some of the limitations with the FlowNet models, a trend in the following work was to integrate classical principles in the architectures. For instance, in the spatial pyramid network (SPyNet) they use a coarse-to-fine approach with spatial pyramids consisting of shallow CNNs performing OF estimates to reduce the model size significantly [67]. From the coarsest level and up to the top, the estimates are upsampled and used to warp the target images towards the source for refined output. They achieved similar results as the base FlowNet models with significantly fewer parameters on several benchmarks. The results did not surpass FlowNet 2.0, but showed that integrating traditional concepts into the DL algorithm have potential.

In the PWC-Net architecture they follow this line of research, and designed a network combining CNNs, spatial pyramids, warping and cost volumes [50]. The main difference between PWC-Net and the other networks are that they employ feature warping with the upsampled flow from lower resolution levels, and construct partial cost volume estimations between the first image features and the warped features of the second image at every pyramid level. At the time of publication, it was the best performing OF method on several benchmarks. Moreover, compared to FlowNet 2.0, it was significantly smaller in terms of learnable parameters and achieved faster runtime performance.

References

[1] B. Angelsen, *Ultrasound Imaging: Waves, Signals, and Signal Processing*. Emantec AS, 2000.

[2] Ö. Oralkan, A. S. Ergun, J. A. Johnson, M. Karaman, U. Demirci, K. Kaviani, T. H. Lee, and B. T. Khuri-Yakub, "Capacitive micromachined ultrasonic transducers: Next-generation arrays for acoustic imaging?," *IEEE Transactions on Ultrasonics, Ferroelectrics, and Frequency Control*, vol. 49, pp. 1596–1610, nov 2002.

[3] H. Gray, *Gray's Anatomy: The Anatomical Basis of Medicine and Surgery, 39th edition*. Churchill-Livingstone, 2004.

[4] Løvstakken, Lasse, *Signal processing in diagnostic ultrasound: Algorithms for real-time estimation and visualization of blood flow velocity*. PhD thesis, Norwegian University of Science and Technology, 2007.

[5] A. Fatemi, E. A. R. Berg, and A. Rodriguez-Molares, "Studying the Origin of Reverberation Clutter in Echocardiography: In Vitro Experiments and In Vivo Demonstrations," *Ultrasound in Medicine and Biology*, vol. 45, pp. 1799–1813, jul 2019.

[6] A. Fatemi, H. Torp, S. Aakhus, and A. Rodriguez-Molares, "Increased clutter level in echocardiography due to specular reflection," in *Medical Imaging 2017: Ultrasonic Imaging and Tomography* (N. Duric and B. Heyde, eds.), vol. 10139, p. 101391D, SPIE, mar 2017.

[7] O. A. Smiseth, H. Torp, A. Opdahl, K. H. Haugaa, and S. Urheim, "Myocardial strain imaging: How useful is it in clinical decision making?," vol. 37, pp. 1196–1207b, apr 2016.

[8] A. Heimdal, A. Stoylen, H. Torp, and T. Skjaerpe, "Real-time strain rate imaging of the left ventricle by ultrasound," *Journal of the American Society of Echocardiography*, vol. 11, no. 11, pp. 1013–1019, 1998.

[9] S. Urheim, T. Edvardsen, H. Torp, B. Angelsen, and O. A. Smiseth, "Myocardial strain by Doppler echocardiography: Validation of a new method to quantify regional myocardial function," *Circulation*, vol. 102, pp. 1158–1164, sep 2000.

[10] H. Geyer, G. Caracciolo, H. Abe, S. Wilansky, S. Carerj, F. Gentile, H. J. Nesser, B. Khandheria, J. Narula, and P. P. Sengupta, "Assessment of Myocardial

Mechanics Using Speckle Tracking Echocardiography: Fundamentals and Clinical Applications," vol. 23, pp. 351–369, apr 2010.

[11] I. Mirsky and W. W. Parmley, "Assessment of Passive Elastic Stiffness for Isolated Heart Muscle and the Intact Heart," tech. rep., 1973.

[12] A. Burkov, *The Hundred-Page Machine Learning Book*. Andriy Burkov, 2019.

[13] I. Goodfellow, Y. Bengio, and A. Courville, *Deep Learning*. MIT Press, 2016. http://www.deeplearningbook.org.

[14] Y. Lecun, Y. Bengio, and G. Hinton, "Deep learning," vol. 521, pp. 436–444, may 2015.

[15] M. P. Deisenroth, A. A. Faisal, and C. S. Ong, *Mathematics for machine learning*. Cambridge University Press, 2020.

[16] V. Nair and G. E. Hinton, "Rectified linear units improve Restricted Boltzmann machines," in *ICML 2010 - Proceedings, 27th International Conference on Machine Learning*, pp. 807–814, 2010.

[17] A. L. Maas, A. Y. Hannun, and A. Y. Ng, "Rectifier nonlinearities improve neural network acoustic models," in *ICML Workshop on Deep Learning for Audio, Speech and Language Processing*, 2013.

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

[19] X. Glorot and Y. Bengio, "Understanding the difficulty of training deep feedforward neural networks," in *Journal of Machine Learning Research*, vol. 9, pp. 249–256, 2010.

[20] B. T. Polyak, "Some methods of speeding up the convergence of iteration methods," *USSR Computational Mathematics and Mathematical Physics*, vol. 4, pp. 1–17, jan 1964.

[21] D. E. Rumelhart, G. E. Hinton, and R. J. Williams, "Learning representations by back-propagating errors," *Nature*, vol. 323, no. 6088, pp. 533–536, 1986.

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

[23] O. Russakovsky, J. Deng, H. Su, J. Krause, S. Satheesh, S. Ma, Z. Huang, A. Karpathy, A. Khosla, M. Bernstein, A. C. Berg, and L. Fei-Fei, "ImageNet Large Scale Visual Recognition Challenge," *International Journal of Computer Vision*, vol. 115, pp. 211–252, sep 2014.

[24] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "ImageNet Classification with Deep Convolutional Neural Networks," *Advances in neural information processing systems*, pp. 1097–1105, 2012.

[25] N. Srivastava, G. Hinton, A. Krizhevsky, and R. Salakhutdinov, "Dropout: A Simple Way to Prevent Neural Networks from Overfitting," tech. rep., 2014.

[26] S. Ioffe and C. Szegedy, "Batch normalization: Accelerating deep network training by reducing internal covariate shift," in *International Conference on Machine Learning*, pp. 448–456, 2015.

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

[28] M. Lin, Q. Chen, and S. Yan, "Network in network," *arXiv preprint arXiv:1312.4400*, 2013.

[29] C. Szegedy, V. Vanhoucke, S. Ioffe, J. Shlens, and Z. Wojna, "Rethinking the inception architecture for computer vision," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 2818–2826, 2016.

[30] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 770–778, 2016.

[31] G. Huang, Z. Liu, K. Q. Weinberger, and L. van der Maaten, "Densely connected convolutional networks," *arXiv preprint arXiv:1608.06993*, 2016.

[32] A. Graves, A. R. Mohamed, and G. Hinton, "Speech recognition with deep recurrent neural networks," in *ICASSP, IEEE International Conference on Acoustics, Speech and Signal Processing - Proceedings*, pp. 6645–6649, oct 2013.

[33] Y. Bengio, P. Simard, and P. Frasconi, "Learning Long-Term Dependencies with Gradient Descent is Difficult," *IEEE Transactions on Neural Networks*, vol. 5, no. 2, pp. 157–166, 1994.

[34] R. Pascanu, T. Mikolov, and Y. Bengio, "On the difficulty of training recurrent neural networks," in *30th International Conference on Machine Learning, ICML 2013*, no. PART 3, pp. 2347–2355, 2013.

[35] S. Hochreiter and J. Schmidhuber, "Long Short-Term Memory," *Neural Computation*, vol. 9, pp. 1735–1780, nov 1997.

[36] K. Cho, B. Van Merriënboer, C. Gulcehre, D. Bahdanau, F. Bougares, H. Schwenk, and Y. Bengio, "Learning phrase representations using RNN encoder-decoder for statistical machine translation," in *EMNLP 2014 - 2014 Conference on Empirical Methods in Natural Language Processing, Proceedings of the Conference*, pp. 1724–1734, 2014.

[37] V. Kastrinaki, M. Zervakis, and K. Kalaitzakis, "A survey of video processing techniques for traffic applications," *Image and Vision Computing*, vol. 21, pp. 359–381, apr 2003.

[38] S. Ali and M. Shah, "Human action recognition in videos using kinematic features and multiple instance learning," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 32, pp. 288–303, feb 2010.

[39] M. Menze and A. Geiger, "Object scene flow for autonomous vehicles," in *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, vol. 07-12-June, pp. 3061–3070, 2015.

[40] B. C. Vemuri, S. Huang, S. Sahni, C. M. Leonard, C. Mohr, R. Gilmore, and J. Fitzsimmons, "An efficient motion estimator with application to medical image registration," *Medical Image Analysis*, vol. 2, pp. 79–98, mar 1998.

[41] L. N. Bohs, B. J. Geiman, M. E. Anderson, S. C. Gebhart, and G. E. Trahey, "Speckle tracking for multi-dimensional flow estimation," *Ultrasonics*, vol. 38, pp. 369–375, mar 2000.

[42] S. Mondillo, M. Galderisi, D. Mele, M. Cameli, V. S. Lomoriello, V. Zacà, P. Ballo, A. D'Andrea, D. Muraru, M. Losi, E. Agricola, A. D'Errico, S. Buralli, S. Sciomer, S. Nistri, and L. Badano, "Speckle-tracking echocardiography: A new technique for assessing myocardial function," *Journal of Ultrasound in Medicine*, vol. 30, pp. 71–83, jan 2011.

[43] S. S. Beauchemin and J. L. Barron, "The Computation of Optical Flow," *ACM Computing Surveys (CSUR)*, vol. 27, pp. 433–466, sep 1995.

[44] A. Verri and T. Poggio, "Motion Field and Optical Flow: Qualitative Properties," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 11, no. 5, pp. 490–498, 1989.

[45] J. J. Gibson, "The perception of the visual world.," 1950.

[46] B. K. Horn and B. G. Schunck, "Determining optical flow," *Artificial Intelligence*, vol. 17, pp. 185–203, aug 1981.

[47] B. D. Lucas and T. Kanade, "An iterative image registration technique with an application to stereo vision," in *Proceedings of Imaging Understanding Workshop*, vol. 2, pp. 674–679, 1981.

[48] Z. Teed and J. Deng, "RAFT: Recurrent All-Pairs Field Transforms for Optical Flow," in *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, vol. 12347 LNCS, pp. 402–419, 2020.

[49] E. Ilg, N. Mayer, T. Saikia, M. Keuper, A. Dosovitskiy, and T. Brox, "Flownet 2.0: Evolution of optical flow estimation with deep networks," in *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, Jul 2017.

[50] D. Sun, X. Yang, M.-Y. Liu, and J. Kautz, "PWC-Net: CNNs for optical flow using pyramid, warping, and cost volume," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 8934–8943, 2018.

[51] J. L. Barron, D. J. Fleet, and S. S. Beauchemin, "Performance of optical flow techniques," *International Journal of Computer Vision*, vol. 12, pp. 43–77, feb 1994.

[52] M. Z. Brown, D. Burschka, and G. D. Hager, "Advances in computational stereo," *IEEE transactions on pattern analysis and machine intelligence*, vol. 25, no. 8, pp. 993–1008, 2003.

[53] D. Fleet and Y. Weiss, "Optical flow estimation," in *Handbook of Mathematical Models in Computer Vision*, pp. 237–257, 2006.

[54] D. Fortun, P. Bouthemy, and C. Kervrann, "Optical flow modeling and computation: A survey," *Computer Vision and Image Understanding*, vol. 134, pp. 1–21, may 2015.

[55] P. Anandan, "A computational framework and an algorithm for the measurement of visual motion," *International Journal of Computer Vision*, vol. 2, no. 3, pp. 283–310, 1989.

[56] M. J. Black and P. Anandan, "The robust estimation of multiple motions: Parametric and piecewise-smooth flow fields," *Computer Vision and Image Understanding*, vol. 63, pp. 75–104, jan 1996.

[57] T. Brox, A. Bruhn, N. Papenberg, and J. Weickert, "High accuracy optical flow estimation based on a theory for warping," *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, vol. 3024, pp. 25–36, 2004.

[58] R. B. Fisher and D. K. Naidu, "A Comparison of Algorithms for Subpixel Peak Detection," in *Image Technology*, pp. 385–404, Springer Berlin Heidelberg, 1996.

[59] G. Farnebäck, "Two-frame motion estimation based on polynomial expansion," *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, vol. 2749, pp. 363–370, 2003.

[60] M. Bai, W. Luo, K. Kundu, and R. Urtasun, "Exploiting semantic information and deep matching for optical flow," in *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, vol. 9910 LNCS, pp. 154–170, 2016.

[61] D. Gadot and L. Wolf, "PatchBatch: A Batch Augmented Loss for Optical Flow," in *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, vol. 2016-Decem, pp. 4236–4245, 2016.

[62] A. Dosovitskiy, P. Fischer, E. Ilg, P. Häusser, C. Hazırbaş, V. Golkov, P. v.d. Smagt, D. Cremers, and T. Brox, "Flownet: Learning optical flow with convolutional networks," in *IEEE International Conference on Computer Vision (ICCV)*, 2015.

[63] J. Hur and S. Roth, "Optical flow estimation in the deep learning age," in *Modelling Human Motion*, pp. 119–140, Springer, 2020.

[64] S. T. H. Shah and X. Xuezhi, "Traditional and modern strategies for optical flow: an investigation," *SN Applied Sciences*, vol. 3, p. 289, mar 2021.

[65] O. Ronneberger, P. Fischer, and T. Brox, "U-net: Convolutional networks for biomedical image segmentation," in *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, vol. 9351, pp. 234–241, Springer Verlag, may 2015.

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

[67] A. Ranjan and M. J. Black, "Optical flow estimation using a spatial pyramid network," in *Proceedings - 30th IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2017*, vol. 2017-Janua, pp. 2720–2729, 2017.

Real-time Standard View Classification in Transthoracic Echocardiography using Convolutional Neural Networks

3

Andreas Østvik[1], Erik Smistad[1,2], Svein Arne Aase[3], Bjørn Olav Haugen[1], and Lasse Løvstakken[1]

[1] Dept. of Circulation and Medical Imaging, NTNU, Trondheim, Norway

[2] SINTEF Medical Technology, Trondheim, Norway

[3] GE Vingmed Ultrasound AS, Horten, Norway

Transthoracic echocardiography examinations are usually performed according to a protocol comprising different probe postures providing standard views of the heart. These are used as a basis when assessing cardiac function, and it is essential that the morphophysiological representations are correct. Clinical analysis is often initialized with the current view, and automatic classification can thus be useful in improving today's workflow. In this article, convolutional neural networks (CNNs) are used to create classification models predicting up to seven different cardiac views. Data sets of 2-D ultrasound acquired from studies totaling more than 500 patients and 7000 videos were included. State-of-the-art accuracies of (98.3±0.6)% and (98.9±0.6)% on single frames and sequences, respectively, and real-time performance with (4.4±0.3) ms per frame was achieved. Further, it was found that CNNs have the potential for use in automatic multiplanar reformatting and orientation guidance. Using 3-D data to train models applicable for 2-D classification, we achieved a median deviation of (4±3)° from the optimal orientations.

3.1 Introduction

Transthoracic echocardiography (TTE) is widely used for assessment of cardiac function. The examinations are usually performed according to protocols involving different probe postures providing several views of the

heart [1]. Image quality varies substantially between patients and is operator dependent, which increases inter-observer variability and decreases the feasibility of detailed quantitative measurements in the clinic. Cardiac view classification (CVC), that is, determining the image plane through the heart, is the essential first interpretation step in any TTE examination. Clinical implementation of automatic solutions is currently limited, but we believe it could affect several elements of everyday practice.

Finding valid cardiac views has traditionally been difficult for apprentices. The European Association of Echocardiography recommends a minimum of 350 examinations to acquire basic competence for standard TTE [2]. Together with the requirement for expert resources, didactic tools using real-time CVC can potentially reduce this number by providing standardization through active quality assurance and probe alignment guidance. Further, a new group of users are adopting echocardiography through the introduction of hand-held devices, making ultrasound (US) more available in general. An implementation with low hardware requirements can be used on such devices and thus provide support in point-of-care situations where cardiologists normally are absent [3].

Tools used when diagnosing cardiac diseases are often initialized with specification of current view, and in most cases this must be done manually by the operator. Automatic classification can improve the workflow and adaptivity of quantitative tools and allow continuous scanning and on-site analysis of several quantitative parameters without pushing a single button on the ultrasound scanner. In addition, such a solution could enhance user experience in 3-D US acquisitions by improving automatic extraction of relevant 2-D image planes from volumes [4].

Finally, CVC can also complement patient database archives by automatically labeling recordings and thus enable better search functionality, data mining and categorization utilities. In turn, this could, for example, improve follow-up by automatically extracting corresponding views at different stages of patient care.

3.1.1 Related work and state of the art

[5] reviewed cardiac view classification for TTE up to 2013. Most studies consider a selection of three or four of the most common cardiac views: apical two chamber (A2C), apical four-chamber (A4C) and apical long-axis

(ALAX), as well as the parasternal long-axis (PLAX) and parasternal short-axis (PSAX). Some consider additional views, such as apical five chamber, subcostal four-chamber (SC4C) and vena cava inferior (SCVC), together with a class for unknown data. Examples of relevant views are shown in Fig. 3.1.

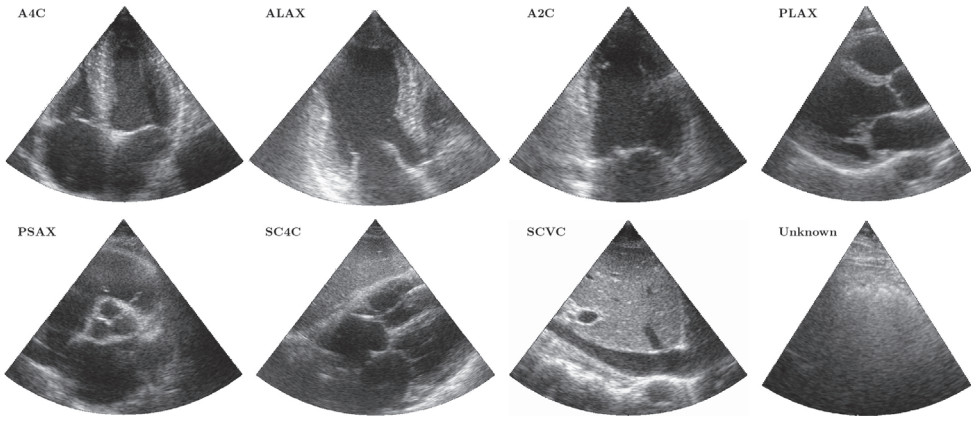


Figure 3.1: Seven cardiac views in transthoracic echocardiography obtained in arbitrary stages of the heart cycle. Examples of the apical four chamber (A4C), long-axis (ALAX), two chamber (A2C), parasternal long axis (PLAX), short-axis (PSAX), subcostal four-chamber (SC4C) and vena cava inferior view (SCVC) is illustrated, in addition to a nonassignable sample labeled unknown.

Prior studies claim to achieve overall accuracies as high as 98% on image sequences, such as reported by [6]. In general, inclusion of more views have reduced accuracy considerably. To the best of our knowledge, [7] reported the largest data set, containing 1080 and 223 image sequences for training and validation, respectively.

Most previous studies have used a support vector machine classifier on features extracted with various methods. Recently, deep convolutional neural networks (CNNs) have had great success in many image classification tasks [8]. As opposed to traditional machine learning approaches with hand-crafted features, these methods learn both the feature extraction and classification directly from the training data. After [9] won the annual ImageNet challenge (ILSVRC) in 2012 using a CNN [10], it has become the predominant approach for solving computer vision and recognition tasks.

CNNs have attracted significant attention from the US image analysis community, where hand-crafting generic features can be difficult. [11] was among the first to report use of CNNs for US view classification, more

specifically for locating the fetal abdominal standard plane. Currently, a body of related work in the domain of fetal US image classification exists that methodologically also uses CNNs [12–14]. In addition, much research for TTE currently involves the use of CNNs. [15] and [16] have used it for evaluation of cardiac function. [17] used it to automatically assess the quality of up to five views using a regression-based recurrent approach. Recently, [18] used CNNs for classifying eight different cardiac views using a method fusing hand-crafted and learned features. Their database consisted of 432 image sequences, and they achieved an average accuracy of 92.1% validating on 152 image sequences.

3.1.2 Main contributions

In the work described here, our aim was to develop fully automated and robust methods for real-time CVC using CNNs and facilitate their use in a clinical setting. We also investigated the potential for applying these methods to automatic extraction of 2-D views from 3-D volumes and orientation guidance for finding optimal views in 2-D US. Compared with previous studies, the contributions of this paper are as follows:

- Annotation and training on significantly more patient data than previously included, with extensive patient-based cross-validation and testing ensuring unbiased results
- Consideration of up to seven of the most common cardiac views: A2C, A4C, ALAX, PSAX, PLAX, SC4C and SCVC, in addition to a class for unknown data
- Analysis of two common network topologies and a proposed network design based on recent work in the field with the aim of being both accurate and effective
- Experiments on orientation guidance for finding optimal apical views and a comparison between models trained with either 2-D or 3-D data
- Analysis of computational requirements and performance

3.2 Convolutional neural networks

Three CNNs were investigated for cardiac view classification. Compared to the problems in which typical image classification networks are designed, we consider CVC easier. The consensus on increasing network depth to achieve better results does not necessarily hold for such tasks, and we believe that competitive performance can be achieved with less complex networks. We therefore address this issue by combining observations from relevant work and propose a network that aims to balance the accuracy and effectiveness for this use.

For extensive details of the investigated networks, the reader is referred to relevant articles [9,19]. Herein, we introduce them briefly and emphasize their differences and our changes to the original topology. Furthermore, we accentuate the background of our design choices.

3.2.1 AlexNet architecture

The winner of ILSVRC 2012 is a CNN referred to as AlexNet. It is a simple feed-forward network with five blocks of convolutional layers followed by rectified linear activation units (ReLU) and maximum pooling. Local response normalization is used after the first two convolution layers. The final part of the network is composed of two fully connected layers with ReLU and dropout regularization, whereas the final classifier is a fully connected layer followed by softmax activation.

Compared with the original topology, the local response normalization layers were removed for this study, and batch normalization [20] was used instead for additional regularization, as suggested by [21].

3.2.2 Inception architecture

Some of the most influential proposals after AlexNet came from the authors of "Network In Network" (NIN) [22], who suggested using bottlenecks (e.g., convolutions with kernel size 1×1) to combine features between layers. The key insights from their article inspired [23] to create the Inception architecture (introduced as GoogleNet). The principal difference from other networks is the building blocks, referred to as Inception modules. Each block consists of parallel routes of convolutions with varying kernel size,

in addition to a pathway with pooling. Fig. 3.2 is a schematic of a typical module with bottlenecks and batch normalization. Several editions of Inception have been presented since its introduction, but the fundamental philosophy of parallel routes in depth is the same.

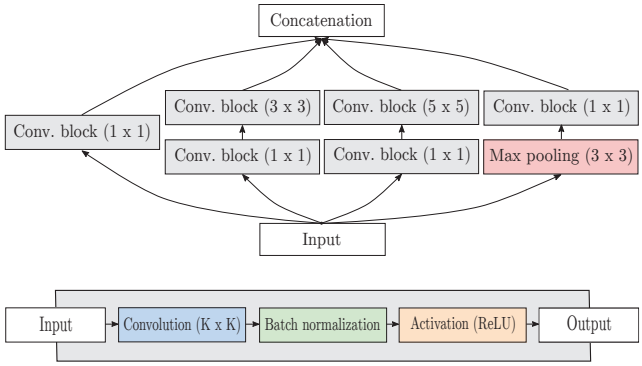


Figure 3.2: The inception module is a combination of parallel convolution blocks with different kernel sizes and a pooling branch concatenated into a single output. Each convolution block (Conv.) consists of convolutions followed by batch normalization and non-linear activation.

In this study, the third edition of the Inception architecture [19] was employed. The major architectural difference compared with the original topology is spatial factorization of large spatial filters. On this basis, three different modules were designed and used throughout the network. In the lower parts, where the feature maps are relatively large, the module is similar to that in Fig. 3.2, except that the (5×5) convolutions are factorized into two layers of (3×3) convolutions. The other modules use asymmetric convolutions, for example, a (3×1) followed by a (1×3) convolution. In addition to factorization, batch normalization is used after convolution layers. Here we use smaller input images than intended for this architecture, and to allow better information flow and avoid convolution filters larger than the feature maps, we removed the second max-pooling layer.

3.2.3 Cardiac view classification architecture

The network we propose resembles that in the discussed work and employs a combination of introduced concepts. Fig. 3.3 is an overview of the architecture. The fundamental building blocks consist of convolutions, batch normalization [20] and non-linear activation units. Batch normalization was

added to speed up training by allowing higher learning rates and avoiding use of network resources to compensate for outlying filter weights during backpropagation. Parametric rectified linear units (PReLU) were chosen as the activation unit in all blocks [24]. Compared with the frequently used ReLU, which is zero for negative values, PReLU allows non-zero gradients for inactive units. The negative part is a linear function with a learned slope.

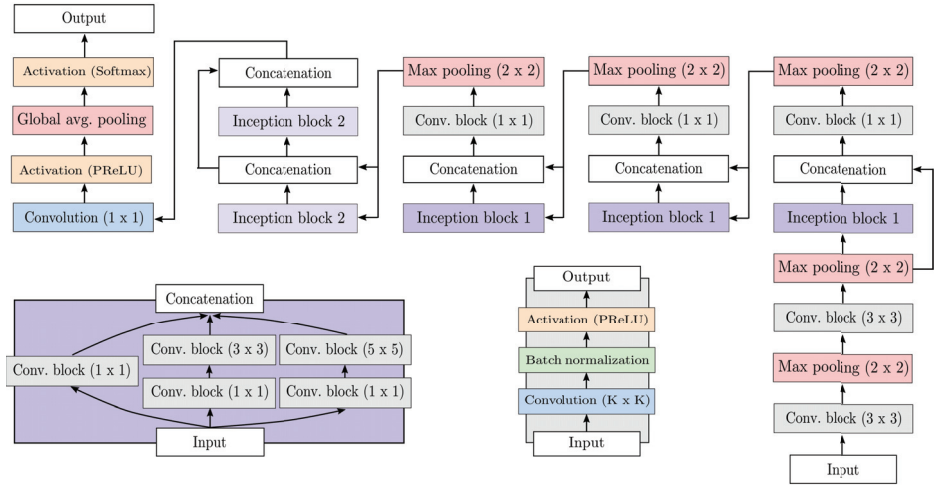


Figure 3.3: Schematic of the proposed network architecture used for cardiac view classification. Convolution blocks (gray boxes) are composed of convolutions, batch normalization and PReLU activations. Two versions of the Inception module are employed: the illustrated one being used in the lower part of the network (dark purple) and a simplified one without the (5×5) route in higher parts of the network (bright purple). The final classifier block consist of another compressing convolution layer with kernel size (1×1) and filter amount equal to the number of views. The output is activated with a PReLU layer. Finally, global average pooling followed by softmax activation yields a prediction vector as output.

Initially, input is propagated through two component blocks with (3×3) convolution kernels, followed by max pooling. The first and second convolution layer have 16 and 32 filters, respectively. We use pooling of size (2×2) and equal strides to downsample without overlap. After the second pooling layer, data are processed through an Inception module with three parallel routes. Each route consists of a bottleneck, two of which were followed by blocks with larger convolution kernels, (3×3) and (5×5), respectively. This is equivalent to the module in Fig. 3.2 without the pooling route. The bottlenecks in the Inception module reduce the number of filters

by 25%, 25% and 50% in the order of small to large convolution kernels, respectively. Furthermore, the number of filters is increased by 25% in the following convolution block.

Inspired by the connection scheme in DenseNet [25], the input of the Inception module is concatenated with the output and processed into a transition module with bottleneck and max pooling. This step is repeated three times, and as emphasized by [26] in the base classifier of the YOLO object detection system, we doubled the number of filters before every new pooling layer. As opposed to their implementation, we control this behavior in the bottleneck of the transition block. The dense connectivity pattern further alleviates the vanishing gradient problem, and perhaps more importantly, it can enhance feature propagation and reusability.

After the third transition, the data are processed through two Inception blocks with a constant number of filters and no pooling. The route with (5×5) convolution kernels was omitted in these modules, and dropout regularization was used between them. The final classification block consisted of a compressing convolution layer with (1×1) kernels and number of filters equal to the class count. This was activated with another PReLU, before features were spatially averaged and fed into a softmax activation as in NIN. The spatial pooling replaces the more typical fully connected layers. This reduces the parameter count and, it is also claimed, makes the network less vulnerable to overfitting [22].

3.3 Experimental setup

Experiments were divided into two parts. First, training and evaluation on annotated 2-D data were conducted using three different CNNs: AlexNet, Inception and the proposed CVC architecture. Afterward, 2-D data extracted from 3-D volumes were included and used to train new models using the CVC architecture. Three-dimensional data were then evaluated, together with a comparison between the models trained with the same architecture on 2-D data.

3.3.1 Database and annotation

Three different data sets of anonymous US data were included in this study. All data originated from patient studies approved by the Regional Commit-

tee for Medical Research Ethics and conducted according to the Helsinki Declaration. Written informed consent was obtained from all patients. The sample data are considered representative of a regular cardiological clinic and give a distribution of both healthy and ill participants in the relevant age groups.

2D US image sequences

The first data set consists of 4582 US videos with varying numbers of frames from 205 patients. Acquisition was performed by three senior cardiologists according to a standard protocol for echocardiography using a GE Vivid E9 US scanner (GE Vingmed Ultrasound, Horten, Norway) with a GE M5S phased-array transducer. Fifty-six of the patients were diagnosed with systolic or diastolic cardiac dysfunction. The population age ranged from 20 to 91 years with an average age of 64 years. The second data set was randomly drawn from the Nord-Trøndelag Health Study (HUNT) population study [27] and consisted of 2559 US videos from 265 subjects. Acquisition was performed by one senior cardiologist according to the same protocol using a GE Vivid 7 scanner with a GE M5S phased-array transducer. All subjects were free from known cardiac dysfunction, and the population had an average age of 49 years.

The videos were annotated manually and categorized into seven different classes: A4C, ALAX, A2C, PLAX, PSAX, SC4C and SCVC. Subcostal acquisitions were not included in the HUNT study. Fig. 3.4. summarizes the data indicating the class balance. Non-assignable images were labeled unknown, but the number was not considered sufficient for training relative to the other classes. Thus, samples from a laboratory experiment with the goal of acquiring arbitrary US images without clinical relevance were added. The total was 41,450 images from 460 videos.

Considerable variations in image quality were discovered, and in a parallel annotation task, the images were labeled as poor, acceptable or good. The relative distribution labeled by an expert cardiologist pre-analysis was (32%, 41%, 27%) from poor to good respectively.

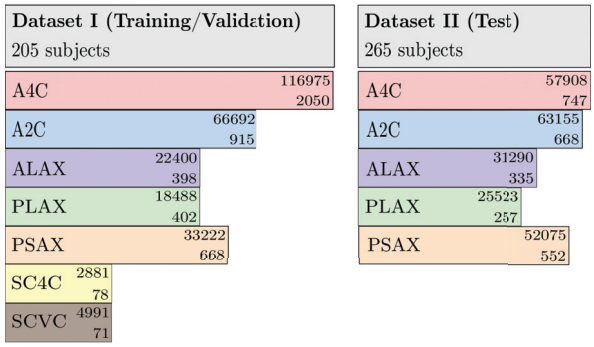


Figure 3.4: Overview of the two 2-D data sets. The upper value is the number of frames in the given class, and the lower value is the number of videos.

3D US volume sequences

The 3-D data set consists of 60 anonymous US volumetric exams with varying numbers of volumes from the same number of patients. Acquisition was performed by two senior cardiologists by placing the probe in the apical position using a GE Vivid E9 US scanner with a 3V 4D sector array transducer probe.

Data were generated by extracting 2-D images, or planes, from the 3-D volume around a fixed depth axis placed in the frustum center. This mimics the scenario of rotating a 2-D probe in an apical position, generating all possible views oriented with respect to the depth axis. Here we extracted one frame per degree, yielding a total of 360 images per volume. Eyeballing and simple caliper measures were used to choose three different frames (angles) from each volume as optimal apical views (A4C, A2C, ALAX). These angles were further used as the peak of an asymmetric Gaussian weighting when labeling the data. The tail of the Gaussian label l is determined by the distance between adjacent peaks and is given by

$$l^{\text{view}}_{\leftarrow,\rightarrow} = \exp\left\{-\left(\frac{\Delta\theta_{\leftarrow,\rightarrow}}{\sqrt{2}\sigma_{\leftarrow,\rightarrow}}\right)^2\right\}. \tag{3.1}$$

Here, $\Delta\theta_{\leftarrow,\rightarrow}$ is the angular distance from the peak of a specific view in a given direction. The standard deviation is the fractional distance to the nearest adjacent peak in either direction, that is, $\sigma_{\leftarrow,\rightarrow} = |\theta^{\text{view}} - \theta^{\text{adj.view}}_{\leftarrow,\rightarrow}|/3$. An example annotation with reference to the 17-segment

left ventricle model [1] is provided in Fig. 3.5. This annotation scheme was chosen to allow a connection between adjacent peaks. Unlike a binary classification, this enables a more robust transition region between optimal views and may be more suitable for orientation guiding and quality assurance while scanning. It could also be used to extract the desired 2-D planes automatically from 3-D volumes.

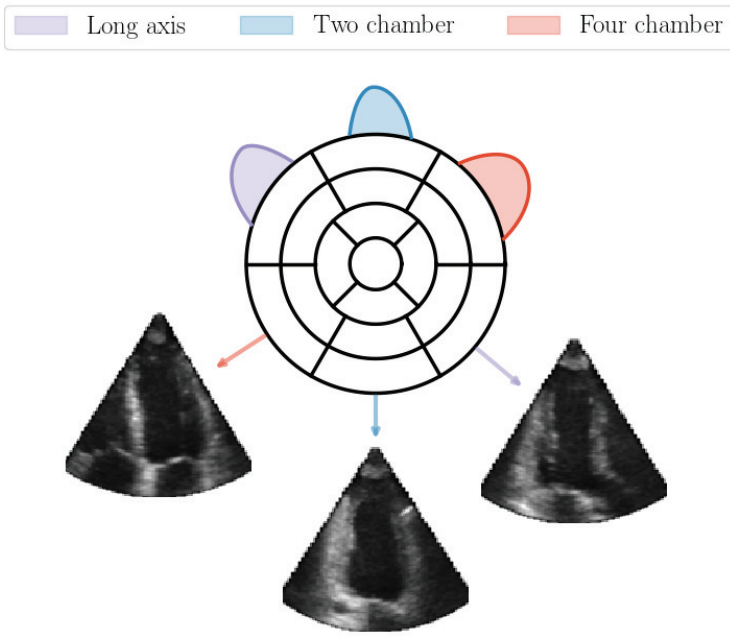


Figure 3.5: Sketch of an example annotation with reference to the left ventricle segment model (17 divisions). The curves correspond to the confidence label of a specific cardiac view, where higher values suggests optimal orientation.

3.3.2 Preprocessing

The data were scan converted from beamspace data stored in DICOM format. The 3-D data were stitched when necessary. No image enhancement filters were applied. For training, the images were intensity normalized and downsampled to a size of (128×128) pixels. No data augmentation was applied.

3.3.3 Learning details

Training was performed over a maximum of 100 epochs using mini-batch gradient descent with a batch size of 64. In machine learning, one epoch is defined as a complete pass of training data, whereas the batch size is the number of examples shown for each weight update. We used the categorical cross-entropy and mean absolute error (MAE) loss functions [28] for training on 2-D and 3-D data, respectively. An adaptive moment estimation method for stochastic optimization named Adam [29] was used with a maximum learning rate of 10^{-4}. Uniform Glorut initialization [30] was used on the convolution layers before training. The model was evaluated on unknown data between epochs, where the best model was saved underway. The data were fully shuffled after every epoch. To avoid unnecessary training time and overfitting, early stopping routines based on validation accuracy were used with a patience of 20 epochs.

As seen in Fig. 3.4, the training data are clearly unbalanced, with a ratio of 1:29 between the least and most represented class. To combat possible bias toward high representations, the training data were downsampled before every new epoch by randomly drawing frames from each US acquisition based on its ratio compared with the least represented class. This allows training on equal amounts of data from each class and every epoch; by performing this on a per-sequence basis, representations from each acquisition are also included. Note that we still use the term epoch, although it breaks the definition of passing the entire dataset.

To setup the learning environment, the framework Keras was utilized with Tensorflow [31] as backend. Experiments was carried out on a workstation installed with an Ubuntu 16.04 operating system. The hardware consisted of an Intel Core i7-6820HK CPU with a clock speed of 4.10 GHz, 32 GB RAM and a NVIDIA GeForce GTX 1070 GPU with 8GB of memory.

3.3.4 Methods and metrics for evaluation

A 10-fold patient-based cross-validation technique was performed, separating the first data set into training and validation partitions. For each run, this corresponds to omitting 20 or 21 patients from the 2-D data. The same was done for the 3-D data, in which each fold consists of six patients. Such patient-based model validation will give a better impression of the

expected results on new patient data. To the best of our knowledge, this is the first publication on the topic in which patient-based cross-validation is extensively used. To further evaluate the model we included an independent data set for testing purposes only.

In addition to accuracy, validation metrics such as precision and recall were used because of the imbalanced class frequency in the 2-D data. They are defined as $TP/(TP+FP)$ and $TP/(TP+FN)$, respectively, where TP is the true positives, FP the false positive and FN the false negative. The model accuracy is defined as the ratio of true predictions to all predictions.

Further, for validation on 3-D data, the MAE is calculated over the angle interval for all volumes of every subject as

$$\mathrm{MAE} = \frac{\sum_{\theta=0}^{\theta_{\max}} |l_\theta^{\mathrm{true}} - l_\theta^{\mathrm{pred}}|}{\theta_{\max}}, \tag{3.2}$$

where the angle $\theta \in [0, \theta_{\max}] = [0, 2\pi)$, and l_θ^{true}, l_θ^{pred} is the true and predicted labels for a given angle. In addition, we performed a qualitative inspection comparing the predictions to ground truth by visualizing them together.

To determine the classification time per incoming image in a deployed setting, an experiment in which images are loaded individually in a loop and classified with the trained models was conducted. This mimics a clinical scenario in which frames are acquired and processed one by one. A total of 30,000 images were processed for each experiment, and for every model we investigated the change in inference time using the GPU. As a hardware invariant measure for runtime, the number of floating point operations was added. This was calculated using the profiler tool released through the Tensorflow framework.

Together with the network definition, the storage requirements are determined mainly by the number of parameters needed to initialize the network. This number is calculated using the Keras framework.

3.4 Results

3.4.1 Analysis on 2D data

Experimental results from patient-based cross-validation using three different network topologies trained on 2-D data are given in Table 3.1. The trained models were tested on an independent and unknown test set, yielding the results outlined in Table 3.2. The sequence validation was performed using a majority vote approach. The CVC model yielded competitive results despite having significantly fewer learned parameters. Compared with the other models, the model variance is lower for the CVC network. Low average inference time per image was achieved for all networks using the frameworks Tensorflow and FAST [32]. This was without any emphasis on inference optimization. Using the GPU, the CVC network classifies approximately 230 frames per second, whereas AlexNet manages twice that number. This is well within the limits of real-time view classification in this context.

Table 3.1: Experimental results from cross-validation on dataset I using three different network topologies. Validations are per single frame and image sequence (in parenthesis for precision and recall). Bold metric indicate best score. Runtime measurements, number of floating point operations and trainable parameters are also given.

(a) AlexNet with BN

	Precision (%)	Recall (%)
A4C	97.7 (98.4)	96.0 (97.6)
ALAX	92.1 (95.0)	95.9 (97.7)
A2C	94.7 (96.5)	96.2 (97.0)
PLAX	96.4 (97.6)	98.1 (99.0)
PSAX	95.7 (97.8)	95.2 (96.7)
SC4C	88.4 (93.6)	96.8 (97.3)
SCVC	94.3 (98.5)	92.2 (94.4)
Unknown	99.2 (95.8)	98.7 (**100.0**)
Overall accuracy(%):		
Frame		96.4 ± 1.2
Sequence		97.5 ± 1.3
Runtime [ms]:		
GPU		$\mathbf{2.0 \pm 0.2}$
CPU		$\mathbf{8.1 \pm 0.2}$
Operations [GFLOPS]:		**0.25**
Parameters:		~20.6M

(b) Inception ver. 3 (Modified)

	Precision (%)	Recall (%)
A4C	97.9 (98.8)	97.8 (99.0)
ALAX	96.8 (99.0)	95.6 (96.2)
A2C	96.5 (**97.5**)	96.7 (98.2)
PLAX	97.0 (97.8)	98.4 (99.5)
PSAX	96.5 (98.6)	97.0 (97.6)
SC4C	92.6 (**96.1**)	96.3 (94.9)
SCVC	97.7 (100.0)	92.9 (**95.8**)
Unknown	99.1 (99.1)	98.8 (**100.0**)
Overall accuracy(%):		
Frame		97.4 ± 1.1
Sequence		98.5 ± 0.8
Runtime [ms]:		
GPU		10.7 ± 0.6
CPU		20.4 ± 0.5
Operations [GFLOPS]:		1.45
Parameters:		~21.8M

(c) Proposed CVC Network

	Precision (%)	Recall (%)
A4C	**98.5** (**99.0**)	**98.5** (**99.3**)
ALAX	**98.1** (**99.2**)	**96.2** (**98.0**)
A2C	**96.9** (**97.5**)	**97.8** (**98.3**)
PLAX	**98.5** (**99.5**)	**99.1** (**100.0**)
PSAX	**98.7** (**100.0**)	**97.9** (**98.2**)
SC4C	**92.7** (94.0)	**99.1** (**100.0**)
SCVC	**99.4** (**100.0**)	**95.3** (94.4)
Unknown	**99.6** (**99.8**)	**99.6** (**100.0**)
Overall accuracy(%):		
Frame		$\mathbf{98.3 \pm 0.6}$
Sequence		$\mathbf{98.9 \pm 0.6}$
Runtime [ms]:		
GPU		4.4 ± 0.3
CPU		15.9 ± 0.4
Operations [GFLOPS]:		0.80
Parameters:		~**10.6M**

To the best of our knowledge, the results surpass current state of the art on 2-D B-mode data and indicate that neural networks are well suited for ultrasound view classification tasks. Accessible benchmark data would be

Table 3.2: Experimental results on test dataset II using three different network topologies. Validations are per single frame and image sequence (in parenthesis for precision and recall). Bold metric indicate best score.

(a) AlexNet with BN

	Precision (%)	Recall (%)
A4C	93.7 (96.0)	99.3 (99.7)
ALAX	97.3 (99.0)	90.7 (93.1)
A2C	95.4 (96.4)	93.1 (95.2)
PLAX	93.1 (96.6)	98.3 (99.4)
PSAX	98.9 (99.7)	95.9 (98.3)
Overall accuracy(%):		
Frame		95.5 ± 0.7
Sequence		97.3 ± 0.6

(b) Inception ver. 3 (Modified)

	Precision (%)	Recall (%)
A4C	94.7 (96.7)	99.5 (99.8)
ALAX	97.7 (99.1)	90.0 (92.2)
A2C	95.1 (96.1)	94.3 (96.0)
PLAX	94.6 (96.6)	98.9 (99.6)
PSAX	99.1 (99.7)	96.9 (98.3)
Overall accuracy(%):		
Frame		96.1 ± 1.6
Sequence		97.5 ± 1.4

(c) Proposed CVC Network

	Precision (%)	Recall (%)
A4C	**96.2** (**97.8**)	**99.6** (**99.8**)
ALAX	**98.6** (**99.5**)	**93.1** (**95.3**)
A2C	**96.6** (**97.6**)	**96.0** (**97.4**)
PLAX	**97.5** (**99.3**)	**98.7** (**99.7**)
PSAX	**99.4** (**99.9**)	**98.3** (**99.5**)
Overall accuracy(%):		
Frame		**97.4 ± 0.6**
Sequence		**98.5 ± 0.5**

needed for a proper comparison with related work, but it is believed that the diversity and size of the data set used in this study at worst yield an equal baseline.

3.4.2 Analysis on 3D data

The averaging of MAE over all volumes of every subject is illustrated in Fig. 3.6. This is calculated from the classification of 360 angles/images per volume. The worst and best cases are indicated, together with the medians of all subjects. The predictions and the ground truth from these cases are illustrated in Fig 3.7. The median MAE of all subjects was $(3.8 \pm 2.4)\%$, and the median deviation from true to predicted peak was $(4 \pm 3)^\circ$. The median MAE using the CVC model trained on 2D data only was $(11.3 \pm 9.7)\%$.

3.5 Discussion

3.5.1 Technical considerations

Patient-based cross-validation indicates that the CVC network is best in terms of relevance and accuracy metrics. The standard deviation is almost halved for the cross-validation models, and it has the smallest parameter space. Testing on an independent and unknown data set also suggests better generalization. Excluding the subcostal and unknown views, the overall accuracy from cross-validation is $(98.1 \pm 0.7)\%$, making the test results within the calculated variation. The slight and consistent underestimation can have

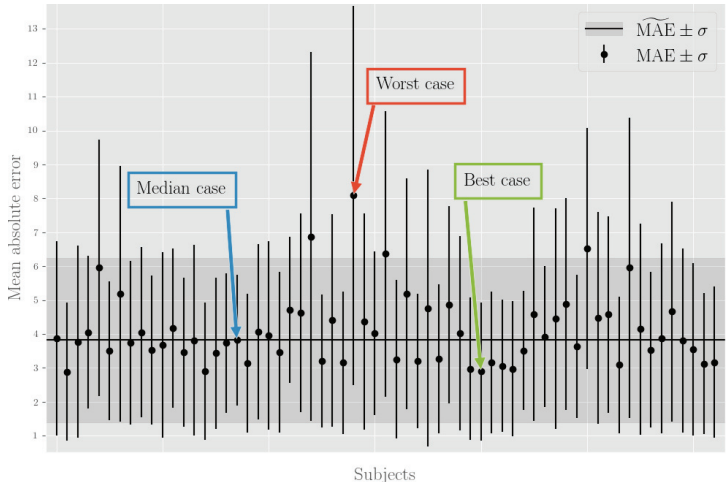
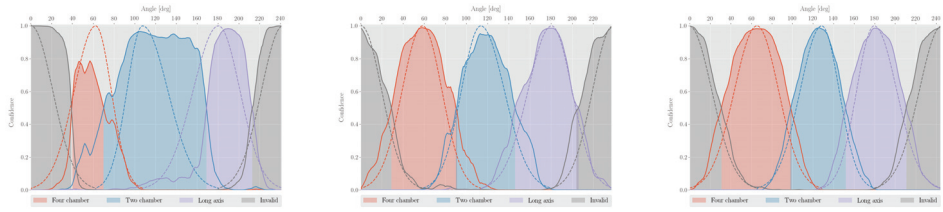


Figure 3.6: Mean absolute error (MAE) values of all subject volumes. The median MAE of all subjects is given by the horizontal line.



(a) Worst-MAE = $(8.1 \pm 5.6)\%$ **(b)** Median-MAE = $(3.8 \pm 1.9)\%$ **(c)** Best-MAE = $(2.9 \pm 2.0)\%$

Figure 3.7: Evaluation on 2D images extracted from 3D volumes with orientation angle with respect to the depth axis. The models used are trained with data from the 3D volumes. The dotted curves correspond to the assigned labels, while the filled curves is the model predictions.

multiple origins; for instance, it could be a small degree of overfitting toward the training/validation data set (e.g., scanner, probe and operators and their preferences). The trained models would probably benefit from a broader representation domain.

Compared with AlexNet, the other networks have smaller receptive fields and less coarse downsampling and, at least for the first layers, preserve more pixel information from the input image. On the other hand, less expressiveness is captured in the learned features. Though the Inception module can retain this to some extent by having a route with a semi-

large kernel, it may seem that adding features benefits generalization more in this scenario. Though it is hard to pinpoint the specific reason why CVC models surpass the results of the other networks on this task, we believe that the combination of Inception modules, dense connections, activation, bottlenecks and number of features (more than AlexNet, fewer than Inception) strengthens the generalization.

The subcostal window proves to be the hardest to classify; arguably, lack of training data is the probable cause. Even if this is the driving factor of these algorithms, still views with distinct characteristics tend to simplify the classification. For example, in Fig. 3.1, we see that the parasternal views seem to have more interclass variance than the apical views and have a higher success rate on unknown data despite learning from less.

Image quality is dependent on the acquisition environment and setup: the parameters used on the scanner, expertise of the physician and status of patient morphophysiology. On an abstract level, this information is embedded into the sequences from a specific patient, and by omitting the use of patient-based validation, the model would gain a fictitious advantage in predicting allegedly unknown data. Examples of poor images from the data set are provided in Fig. 3.8, where the model has predicted the views as indicated under every image. The variation in quality from Fig. 3.1 is apparent, and we discover that the model has more conflicts with ground truth when images are poor. Of 54 misclassified sequences in the cross-validation, 42 were classified in this category, whereas the remainder were acceptable.

Another interesting observation is that the images in Fig. 3.8 were acquired from two different patients and amount to approximately 15% of the total sequential error. By observation, sequences from the patient shown in the upper part of the figure have an abnormal artifact in the left ventricle. The other patients generally have noisy and virtually invisible structures. Both types of issues can cause classification problems and could potentially be present in all image sequences from a specific patient. By distributing sequences from the same patient in both the training and validation data sets, the model could effectively adapt to the irregularity. Patient-based cross-validation and independent tests should thus be emphasized when assessing results from generated models.

Compared with other work, our results seem promising. Methods

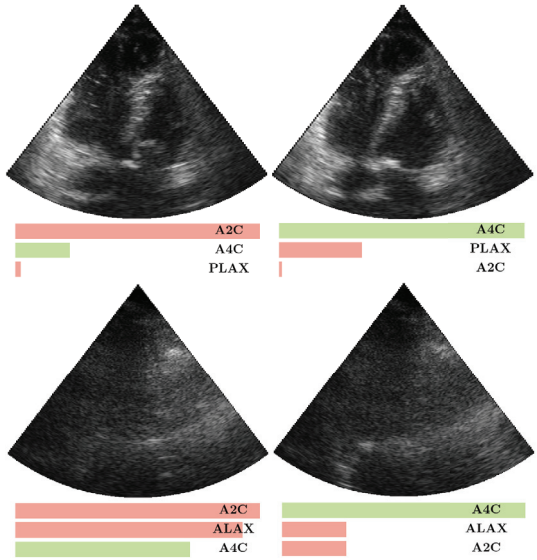


Figure 3.8: Example of poor cardiac ultrasound images from two different apical four chamber sequences classified by the proposed view detection model. Green label indicates the ground truth label, and the size corresponds to the fractional prediction of the model. The left side shows frames where the model has conflicts with ground truth.

and potential applications have some overlap with the research conducted by [17] on quality assessment of cine loops. However, their multistream regression network required 20 consecutive image frames to assess one label; to discriminate between views, every frame had to be passed through a shared layer architecture and into five different view-specific layers. This could be feasible for distinguishing views because Abdi *et al.* state that it is in real time, but their focus is on quality assessment of a given view; classification is not investigated.

In three dimensions, annotation of optimal views was difficult because variations in image features were insignificant for small angle intervals. This held especially for the four- and two-chamber views, whereas the long-axis view was easier because the diameter of the left ventricular outflow tract could be used as a reference in most cases. With this in mind, we still achieve a low median deviation of the predicted to true peaks in all patients, and by inspecting Fig. 3.7, we argue that the long-axis view appears more robust. In general, models trained with 3-D data achieve a low MAE. The performance of models trained with 2-D data, as expected, experiences more fluctuation,

and it can be difficult to detect the optimal view. The reason might be variations in image quality and views slightly off orientation. The latter are not distinguished in the 2-D data set, as we assumed that every examination contains the best possible view for every patient. Results could therefore be expected to have a saturating behavior around the optimal view.

3.5.2 Clinical perspective

As stated in the Introduction, automatic CVC has several clinical applications, such as improving workflow, enabling more automation and guiding inexperienced users. The results on the second independent data set in this study indicate that the accuracy of the proposed CVC methods based on CNNs is real even for data acquired with other scanners and by different operators. This accuracy, together with the measured low runtime and the real-time video, suggests that this method is ready for further testing in a clinical setting. Development in an end-to-end fashion allow low threshold deployment and applicability in many settings without any tuning or in-depth knowledge of the methods. No parameters are required; only an input image is needed. Results also indicate that including training data from heart volumes can improve guiding utilities and quality assurance while scanning. Despite this, models trained with 2-D data will probably be better suited for database utilities, such as data mining, search and categorization. It is easier to add more views, and the accuracy is very high.

Training opportunities for new health care personnel are limited, and expert knowledge is often captivated by workload or centralization. We believe these increasingly relevant problems could be addressed by tools such as automatic CVC. However, separate clinical studies on training effects, standardization and workflow must be induced to support this statement.

3.6 Conclusion

In the study described here, different neural networks were investigated for cardiac view classification. State-of-the-art results for standard 2-D echocardiography were achieved. The proposed network had a small number of trainable parameters and achieved real-time inference with high

accuracy. Although the demonstration looks robust when training on 2-D data, our initial experiments into apical view guidance based on 3-D data indicated room for further work. Using slices of 3-D volumes for training improved the results significantly, and we believe that further development toward real-time quality assurance and guidance from US images is plausible when including such data.

References

[1] R. M. Lang, L. P. Badano, V. Mor-Avi, J. Afilalo, A. Armstrong, L. Ernande, F. A. Flachskampf, E. Foster, S. A. Goldstein, T. Kuznetsova, *et al.*, "Recommendations for cardiac chamber quantification by echocardiography in adults: an update from the american society of echocardiography and the european association of cardiovascular imaging," *European Heart Journal-Cardiovascular Imaging*, vol. 16, no. 3, pp. 233–271, 2015.

[2] B. A. Popescu, M. J. Andrade, L. P. Badano, K. F. Fox, F. A. Flachskampf, P. Lancellotti, A. Varga, R. Sicari, A. Evangelista, P. Nihoyannopoulos, *et al.*, "European association of echocardiography recommendations for training, competence, and quality improvement in echocardiography," *European Journal of Echocardiography*, vol. 10, no. 8, pp. 893–905, 2009.

[3] A. E. Morris, "Point-of-care ultrasound: Seeing the future," *Current Problems in Diagnostic Radiology*, vol. 44, no. 1, pp. 3 – 7, 2015.

[4] X. Lu, B. Georgescu, Y. Zheng, J. Otsuki, and D. Comaniciu, "Autompr: Automatic detection of standard planes in 3d echocardiography," in *Biomedical Imaging: From Nano to Macro, 2008. ISBI 2008. 5th IEEE International Symposium on*, pp. 1279–1282, IEEE, 2008.

[5] O. A. B. Penatti, R. d. O. Werneck, W. R. de Almeida, B. V. Stein, D. V. Pazinato, P. R. Mendes Júnior, R. d. S. Torres, and A. Rocha, "Mid-level image representations for real-time heart view plane classification of echocardiograms," *Computers in Biology and Medicine*, vol. 66, pp. 66–81, 2015.

[6] H. Wu, D. M. Bowers, T. T. Huynh, and R. Souvenir, "Echocardiogram view classification using low-level features," in *IEEE 10th International Symposium on Biomedical Imaging*, pp. 752–755, 2013.

[7] J. H. Park, S. K. Zhou, C. Simopoulos, J. Otsuki, and D. Comaniciu, "Automatic cardiac view classification of echocardiogram," in *Proceedings of the IEEE International Conference on Computer Vision*, 2007.

[8] Y. LeCun, Y. Bengio, and G. Hinton, "Deep learning," *Nature*, vol. 521, no. 7553, pp. 436–444, 2015.

[9] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "ImageNet Classification with Deep Convolutional Neural Networks," *Advances in neural information processing systems*, pp. 1097–1105, 2012.

[10] O. Russakovsky, J. Deng, H. Su, J. Krause, S. Satheesh, S. Ma, Z. Huang, A. Karpathy, A. Khosla, M. Bernstein, A. C. Berg, and L. Fei-Fei, "ImageNet Large Scale Visual Recognition Challenge," *International Journal of Computer Vision (IJCV)*, vol. 115, no. 3, pp. 211–252, 2015.

[11] H. Chen, D. Ni, J. Qin, S. Li, X. Yang, T. Wang, and P. A. Heng, "Standard Plane Localization in Fetal Ultrasound via Domain Transferred Deep Neural Networks," *IEEE Journal of Biomedical and Health Informatics*, vol. 19, pp. 1627–1636, sep 2015.

[12] C. F. Baumgartner, K. Kamnitsas, J. Matthew, T. P. Fletcher, S. Smith, L. M. Koch, B. Kainz, and D. Rueckert, "Sononet: real-time detection and localisation of fetal standard scan planes in freehand ultrasound," *IEEE transactions on medical imaging*, vol. 36, no. 11, pp. 2204–2215, 2017.

[13] W. Huang, C. P. Bridge, J. A. Noble, and A. Zisserman, "Temporal heartnet: towards human-level automatic analysis of fetal cardiac screening video," in *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pp. 341–349, Springer, 2017.

[14] C. P. Bridge, C. Ioannou, and J. A. Noble, "Automated annotation and quantitative description of ultrasound videos of the fetal heart," *Medical image analysis*, vol. 36, pp. 147–161, 2017.

[15] D. P. Perrin, A. Bueno, A. Rodriguez, G. R. Marx, and J. Pedro, "Application of convolutional artificial neural networks to echocardiograms for differentiating congenital heart diseases in a pediatric population," in *Medical Imaging 2017: Computer-Aided Diagnosis*, vol. 10134, p. 1013431, International Society for Optics and Photonics, 2017.

[16] S. Narula, K. Shameer, A. M. S. Omar, J. T. Dudley, and P. P. Sengupta, "Machine-learning algorithms to automate morphological and functional assessments in 2d echocardiography," *Journal of the American College of Cardiology*, vol. 68, no. 21, pp. 2287–2295, 2016.

[17] A. H. Abdi, C. Luong, T. Tsang, J. Jue, K. Gin, D. Yeung, D. Hawley, R. Rohling, and P. Abolmaesumi, "Quality assessment of echocardiographic cine using recurrent neural networks: Feasibility on five standard view planes," in *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pp. 302–310, Springer, 2017.

[18] X. Gao, W. Li, M. Loomes, and L. Wang, "A fused deep learning architecture for viewpoint classification of echocardiography," *Information Fusion*, vol. 36, pp. 103–113, 2017.

[19] C. Szegedy, V. Vanhoucke, S. Ioffe, J. Shlens, and Z. Wojna, "Rethinking the inception architecture for computer vision," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 2818–2826, 2016.

[20] S. Ioffe and C. Szegedy, "Batch normalization: Accelerating deep network training by reducing internal covariate shift," in *International Conference on Machine Learning*, pp. 448–456, 2015.

[21] A. Canziani, A. Paszke, and E. Culurciello, "An analysis of deep neural network models for practical applications," *arXiv preprint arXiv:1605.07678*, 2016.

[22] M. Lin, Q. Chen, and S. Yan, "Network in network," *arXiv preprint arXiv:1312.4400*, 2013.

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

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

[25] G. Huang, Z. Liu, K. Q. Weinberger, and L. van der Maaten, "Densely connected convolutional networks," *arXiv preprint arXiv:1608.06993*, 2016.

[26] J. Redmon and A. Farhadi, "Yolo9000: Better, faster, stronger," *arXiv preprint arXiv:1612.08242*, 2016.

[27] H. Dalen, A. Thorstensen, S. A. Aase, C. B. Ingul, H. Torp, L. J. Vatten, and A. Stoylen, "Segmental and global longitudinal strain and strain rate based on echocardiography of 1266 healthy individuals: the hunt study in norway," *European Journal of Echocardiography*, vol. 11, no. 2, pp. 176–183, 2009.

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

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

[30] X. Glorot and Y. Bengio, "Understanding the difficulty of training deep feedforward neural networks," in *Proceedings of the Thirteenth International Conference on Artificial Intelligence and Statistics*, pp. 249–256, 2010.

[31] M. Abadi, A. Agarwal, P. Barham, E. Brevdo, Z. Chen, C. Citro, G. S. Corrado, A. Davis, J. Dean, M. Devin, S. Ghemawat, I. Goodfellow, A. Harp, G. Irving, M. Isard, Y. Jia, R. Jozefowicz, L. Kaiser, M. Kudlur, J. Levenberg, D. Mané, R. Monga, S. Moore, D. Murray, C. Olah, M. Schuster, J. Shlens, B. Steiner, I. Sutskever, K. Talwar, P. Tucker, V. Vanhoucke, V. Vasudevan, F. Viégas, O. Vinyals, P. Warden, M. Wattenberg, M. Wicke, Y. Yu, and X. Zheng, "TensorFlow: Large-scale machine learning on heterogeneous systems," 2015.

[32] E. Smistad, M. Bozorgi, and F. Lindseth, "FAST: framework for heterogeneous medical image computing and visualization," *International Journal of Computer Assisted Radiology and Surgery*, vol. 10, no. 11, pp. 1811–1822, 2015.

Detection of Cardiac Events in Echocardiography using 3D Convolutional Recurrent Neural Networks

4

Adrian Meidell Fiorito[1], Andreas Østvik[2], Erik Smistad[2], Sarah Leclerc[3], Olivier Bernard[3] and Lasse Løvstakken[2]

[1] Dept. of Engineering Cybernetics, NTNU, Trondheim, Norway

[2] Centre for Innovative Ultrasound Solutions, NTNU, Trondheim, Norway

[3] CREATIS, Universite de Lyon, Lyon, France

A proper definition of cardiac events such as end-diastole (ED) and end-systole (ES) is important for quantitative measurements in echocardiography. While ED can be found using electrocardiography (ECG), ES is difficult to extract from ECG alone. Further, on hand-held devices ECG is not available or cumbersome. Several methods for automatic detection of cardiac events have been proposed in the recent years, such as using a 2D convolutional neural network (CNN) followed by 1D recurrent layers. This structure may be suboptimal, as tissue movement has a spatio-temporal nature which is ignored in the CNN.

We propose using a 3D CNN to extract spatio-temporal features directly from the input video, which are fed to long short term memory (LSTM) layers. The joint network is trained to classify whether frames belong to either diastole or systole. ES and ED are then automatically detected as the switch between the two states. The 3D CNN + LSTM model performs favourably at detecting cardiac events on a dataset consisting of standard B-mode images of apical four- and two-chamber views from 500 patients. The mean absolute error between events in the apical four-chamber view is 1.63 and 1.71 frames from ED/ES reference respectively. Model inference is fast, using (30 ± 2) ms per 30 frame input sequence on a modern graphics processing unit.

4.1 Introduction

Detection of end-systole (ES) and end-diastole (ED) in echocardiography is an important step when assessing cardiac function. ED and ES are defined as the time points when the mitral valve and aortic valve closes respectively [1]. Several clinical metrics, such as ejection fraction and global longitudinal strain [2] are determined using the ES and ED images. The current approach for detecting ED usually involves finding the QRS-complex in additional measurements from electrocardiograms (ECG), or by visual inspection of the videos. Finding ES is more difficult in ECG alone, making visual inspection of ultrasound (US) images necessary. In clinical practice, this constitutes a significant amount of work that potentially could be automated. An additional benefit is that accurate detection of ES and ED solely using echocardiographic frames removes the need for applying ECG-patches, further reducing time and resources. This is especially useful for smaller devices such as the pocket-sized US scanners.

A multitude of machine learning methods have been proposed for learning video representations. Recently, deep learning have been able to perform on par or better than traditional approaches. These methods differ in the way spatial and temporal features are combined. In the two-stream network [3], one CNN is trained to extract features from still images, and another CNN is trained to capture motion patterns using a stack of optical flow frames. Several methods have been proposed to increase the temporal capacity of these models, such as extending the CNN to 3D [4]. Another popular approach is the Long-Term Recurrent Convolutional Network [5], which uses a CNN to extract features for individual frames. These features are input into a Long Short-term Memory (LSTM) [6] recurrent network for temporal fusion. Similarly, [7] use a shallow 3D CNN to extract features from short clips, which are passed to an LSTM network. Other methods use deeper 3D CNNs to learn spatio-temporal features [8, 9].

Several methods have been proposed for detecting cardiac events automatically in echocardiography. Cardiac cycle start and length are estimated without the use of ECG in [10]. To detect cycle start, the motion of a point near the mitral annulus is found using speckle tracking. This is compared to a database of left ventricle (LV) displacement curves to estimate the cycle start corresponding to the QRS complex in ECG. Other methods

explore manifold learning and dimensionality reduction [11, 12]. Frames in an echocardiogram are mapped to a learned manifold, and the fact that ED and ES occur in periods with small volumetric changes is used to detect these events as dense regions on the manifold. CNNs have been used to extract ED and ES with high precision in cine magnetic resonance imaging [13]. Here, a pretrained CNN is used as a feature extractor, and features are passed on to an LSTM layer. The model is trained to regress a typical volume curve of the LV over a single heartbeat. ED and ES is then identified as the largest and smallest regressed volume in the sequence, respectively. A similar approach applied to echocardiography replaced the pretrained CNN with a residual network [14].

In this work, we replaced the standard CNN with a 3D CNN for spatio-temporal feature learning. Further, we propose training the model on a target which is more suited for detecting ED and ES. The model is trained on variable length sequences, whereas previous deep learning approaches use fixed length input videos.

4.2 Methodology

4.2.1 Problem formulation

To train models for detecting ED and ES frames in a supervised manner, the target output must be generated. An intuitive approach involves posing this as classification with three classes: ED, ES, or neither. However, this introduces a class imbalance problem, as ED and ES frames are underrepresented. An easy way to achieve low loss is then to output neither for all frames.

In [13,14] the problem is formulated as a regression task. Here, the target is set to approximate a typical LV volume curve, by using a cubic function and normalizing the target 0 to 1. The representation is thus not the actual volume curve for a given sample, and therefore the model must attempt to learn a mapping which is not exactly present in the data. For some cases of pathology, such as in the event of post-systolic contraction, the volume might not be smallest at the time of ES. In addition, detecting ED/ES as extrema in the estimated volume curve might be difficult due to flat regions during isovolumetric periods, resulting in several candidates.

In this work, the problem is formulated as a binary classification task.

The target is set to 0 for frames belonging to systole, and 1 for frames in diastole. This alleviates the issue of class imbalance, as there is a comparable number of diastole and systole frames. ES and ED is detected as the frames where the output at the next timestep crosses 0.5 from below or above, respectively.

4.2.2 Network architecture

A 3D CNN architecture is presented which is capable of handling arbitrarily long sequences (until GPU memory is full). Due to high GPU memory utilization of 3D convolutions, the model contains few filters and use pooling frequently compared to state-of-the-art image recognition models. The CNN consists of five 3D convolutional layers, each followed by batch normalization, ReLU activation and max pooling layers. As one prediction should be made for each input frame, pooling is only performed along the spatial axes, and not along the temporal axis. For the same reason, each convolutional layer pads the input with zeroes to preserve the length of the data. As in [9], kernels have spatial and temporal size of 3, except from the first layer which uses a spatial size of 7. The number of feature maps double every convolutional layer, starting at 16 and ending at 256. At the output of the 3D CNN, dropout with a probability of 0.3 is performed to prevent overfitting. The output of shape $[t, 4, 2, 256]$ is then reshaped into t vectors of shape $[2048]$. LSTM layers are added to filter the CNN predictions and increase the capability to remember longer movements. Both LSTM layers have a cell state of size 32, resulting in 32 output features per timestep. An L^2 regularization of 1×10^{-4} is used for recurrent and convolutional kernels. A 1D convolutional layer with a sigmoid activation is placed at the end of the model, operating along the temporal axis. The layer has a single kernel of temporal size 3, with the aim of smoothing the output of the model and reduce the likelihood of the output erroneously crossing 0.5 as a result of noisy data. The model is implemented in Keras with the Tensorflow backend. Fig. 4.1 shows the overall layout.

4.2.3 The dataset

The dataset consists of apical four-chamber (A4C) and two-chamber (A2C) echocardiograms from 500 patients, acquired at the University Hospital of

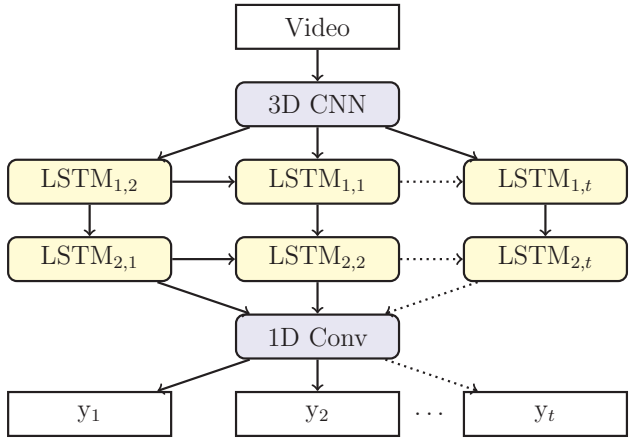


Figure 4.1: Schematic of the network architecture with a 3D CNN followed by LSTM layers and a 1D convolutional layer at the end.

St-Etienne (France) using a GE Vivid E95 ultrasound system (GE Vingmed Ultrasound, Horten, Norway) [15]. For most patients, a corresponding electrocardiogram (ECG) is aligned with each sequence, giving one ECG measurement per video frame. The data is representative for a typical outpatient clinic. The videos have varying sector geometries, sampling rates and duration. The sample time per frame is between 11.99 ms and 21.05 ms. Each video contains a varying number of cardiac cycles. For each video, one frame corresponding to ED and one frame corresponding to ES is labeled by an expert. The labeled ES and ED belongs to the same heart cycle, with ED labeled first for 498 of the A2C videos, and for 481 of the A4C videos. The dataset is split randomly into three folds, with 300 patients used for training, 100 for validation during training, and 100 for testing. Both the A4C and A2C videos for a single patient are placed in the same fold to avoid data leakage.

As only one ES and ED is labeled for each video in the dataset, no labeled input data contains a full heart cycle. In addition, a majority of the frames between the labeled ED/ES belongs to systole, as the labeled ED most commonly occurs before ES. To have training data for any part of the heart cycle, an additional ED is labeled by considering the accompanying ECG signal. The QRS-complex is used to label the ED that yields a fully labeled heart cycle. From 500 patients, 333 and 334 of the ECG signals corresponding to A4C and A2C videos respectively are considered of high

enough quality to accurately identify a second ED.

A number of frames before and after the labeled ED/ES are included to further expand the dataset size, and to make sure ED and ES does not occur at the first and last frames. The resulting dataset contains 26818 frames of A4C and 26170 frames of A2C echocardiograms, belonging to both diastole and systole. The frames are resized to size 128×80 using bicubic interpolation, and normalized by subtracting the mean and dividing by the standard deviation over all pixels in the training data.

4.2.4 Learning details

Training is done for 100 epochs with cross-entropy loss applied over each time step. The Adam optimizer with a learning rate of 1×10^{-4} was used. At the end of every epoch, the training data is shuffled. Both A4C and A2C views are used as training data. Model weights are saved at the epoch with the lowest mean absolute error (MAE) on the validation set. Training on only A4C views was also tested, but resulted in worse performance. The model is trained using mini-batches of four videos, and shorter videos and targets are padded at the end with zeroes. The loss is set to zero for padded frames before backpropagation.

Data augmentations were important for preventing overfitting. Sequences are downsampled temporally by a factor of 2 by discarding every other input frame with a probability of 0.2. Sequences are temporally cropped by randomly discarding between $0 - 80\%$ of the original duration, starting and ending at a random frame. After this, videos are rotated randomly between -10 to 10 degrees. Next, videos are randomly cropped spatially, removing between 0 and 20 pixels along each border. After cropping, the videos are resized to the input size expected by the model. Training and model evaluations were performed on a NVIDIA Titan V GPU with 12 GB RAM.

4.2.5 Evaluation

The error is defined as the difference between the time of a labeled event E and a detected event $\hat{E}$, either ED or ES. Using the notation of [13], the MAE

in frames is denoted the average frame difference (aFD),

$$\text{aFD} = \frac{1}{N}\sum_{1}^{N}|E - \hat{E}|, \tag{4.1}$$

where N is the number of events in the dataset. The mean (μ_e) and standard deviation (σ_e) of the error is also presented in milliseconds (ms).

In order to evaluate if the model is invariant to the cardiac cycle starting point, a variable number of additional frames are included at the beginning and end of the sequence. For each video in the test set, results are measured with 0%, 33% and 66% of the duration between the labeled ED and ES included at the beginning and end of the input data. The model output for the included frames are then discarded as there is no ground truth for these frames.

4.3 Results

Table 4.1 shows the resulting performance on the 100 patients in the test set. Three of the labeled ED and ES frames are not detected by the model for the A2C view, due to the event occurring near the first or last frames of the input. More than one detection of the same event occur three times for ED, and six times for ES. In all these cases, inspection reveals that the data is from a non-standard view or noisy. These cases are thus excluded in the result metrics. Fig. 4.2 shows a patient from the dataset along with labeled and detected events, while Fig. 4.3 shows the model output for the patient.

Table 4.1: Errors of detected ED and ES relative to labeled ED and ES

View	Event	aFD	$\mu_e(ms)$	$\sigma_e(ms)$
A2C	ED	1.40	-5.68	35.8
	ES	1.25	-1.94	29.9
A4C	ED	1.63	0.50	29.8
	ES	1.71	0.60	37.8

Table 4.2 shows the model compared to results reported in [14] for other deep learning approaches.

The time used to predict a single video consisting of 30 frames is

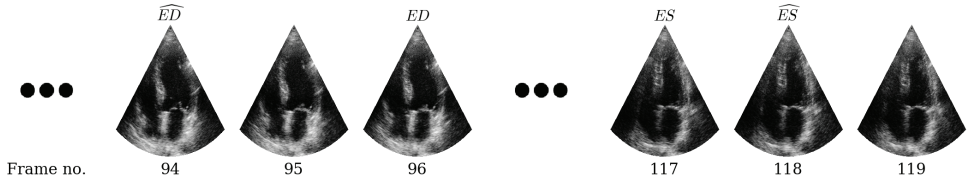


Figure 4.2: Example input sequence (apical four-chamber) along with the labeled frames (ED, ES) and frames detected by the model ($\widehat{ED}$, $\widehat{ES}$).

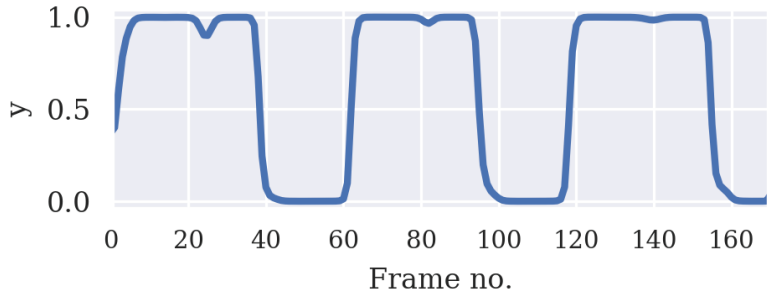


Figure 4.3: Output of the model on the sequence shown in Fig. 4.2. The model output, y, is close to 1 for frames corresponding to the diastole phase and 0 for frames in systole. ED and ES is detected as frames where the y crosses 0.5.

measured 100 times and averaged. This resulted in (30 ± 2) ms used on average for predicting 30 frames.

4.4 Discussion

The 3D CNN is able to detect both ED and ES accurately both for A4C and A2C views, as seen in Table 4.1. This suggests that the network has learned general features for both cardiac phases, such as movement of the atrioventricular valves and the contraction / relaxation of the myocardium. The model is suited for learning these features, as the 3D convolutional layers are able to learn motions between adjacent pixels. A 3D CNN alone might result in a noisy output, due to the noisy input data. This is where the LSTM layers can do a good job of filtering the CNN output. There are few visible difference between the labeled and detected ES frame in Fig. 4.3, and the most noticeable difference between the labeled and detected ED frames is the slightly more closed mitral valve for the labeled ED. As seen from Fig. 4.3, the model output closely resembles a square wave corresponding

Table 4.2: Comparison to metrics reported in [14] on the A4C view

Model	aFD (ED)	aFD (ES)
CNN + LSTM [13]	6.3	7.3
ResNet + LSTM [14]	3.7	4.1
3D CNN + LSTM	**1.6**	**1.7**

to systole and diastole frames, with only a few noticeable dips, showing how well the model separates between systole and diastole. As seen in Table 4.2, the aFD is less than half of [14]. Comparing the performance must however be performed with caution, due to the models being evaluated on different datasets.

An issue is that the labels are not guaranteed to be correct. Determining the exact moment of ED and ES can be difficult for a human annotator due to small differences between consecutive frames. These errors increase as the sampling rate increases. Therefore, it would be interesting to compare the variability of human annotators.

Frequent pooling and few convolutional kernels ensures that the model runs efficiently. It also has a regularization effect, as a small network is less likely to overfit to the training data. The approach has shown to work using a variable number of input frames, instead of limiting the input to a fixed number of frames. This means that the model can operate on an arbitrary long input sequence, and is not restricted to using a single heart cycle as input. Thus, the method may be used to automatically extract heart cycles when considering the distinct differences between output for diastole and systole frames.

4.5 Conclusion

In this paper, a novel method for detecting cardiac events in echocardiography using deep learning was proposed. A 3D CNN was employed followed by recurrent layers to facilitate the learning of spatio-temporal features. State-of-the-art results are achieved on a large dataset, which indicate that the chosen components enhances the solution of the task.

References

[1] R. M. Lang, L. P. Badano, V. Mor-Avi, 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. Rietzschel, 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," *European Heart Journal - Cardiovascular Imaging*, vol. 16, no. 3, pp. 233–271, 2015.

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

[3] K. Simonyan and A. Zisserman, "Two-stream convolutional networks for action recognition in videos," in *Advances in neural information processing systems*, pp. 568–576, 2014.

[4] J. Carreira and A. Zisserman, "Quo vadis, action recognition? a new model and the kinetics dataset," in *Computer Vision and Pattern Recognition (CVPR), 2017 IEEE Conference on*, pp. 4724–4733, IEEE, 2017.

[5] J. Donahue, L. Anne Hendricks, S. Guadarrama, M. Rohrbach, S. Venugopalan, K. Saenko, and T. Darrell, "Long-term recurrent convolutional networks for visual recognition and description," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 2625–2634, 2015.

[6] S. Hochreiter and J. Schmidhuber, "Long short-term memory," *Neural computation*, vol. 9, no. 8, pp. 1735–1780, 1997.

[7] M. Baccouche, F. Mamalet, C. Wolf, C. Garcia, and A. Baskurt, "Sequential deep learning for human action recognition," in *Human Behavior Understanding*, (Berlin, Heidelberg), pp. 29–39, Springer Berlin Heidelberg, 2011.

[8] S. Ji, W. Xu, M. Yang, and K. Yu, "3d convolutional neural networks for human action recognition," *IEEE transactions on pattern analysis and machine intelligence*, vol. 35, no. 1, pp. 221–231, 2013.

[9] D. Tran, L. Bourdev, R. Fergus, L. Torresani, and M. Paluri, "Learning spatiotemporal features with 3d convolutional networks," in *Proceedings of the IEEE international conference on computer vision*, pp. 4489–4497, 2015.

[10] S. A. Aase, S. R. Snare, H. Dalen, A. Støylen, F. Orderud, and H. Torp, "Echocardiography without electrocardiogram," *European Journal of Echocardiography*, vol. 12, no. 1, pp. 3–10, 2010.

[11] P. Gifani, H. Behnam, A. Shalbaf, and Z. A. Sani, "Automatic detection of end-diastole and end-systole from echocardiography images using manifold learning," *Physiological Measurement*, vol. 31, no. 9, p. 1091, 2010.

[12] A. Shalbaf, Z. Alizadehsani, and H. Behnam, "Echocardiography without electrocardiogram using nonlinear dimensionality reduction methods," *Journal of Medical Ultrasonics*, vol. 42, Apr 2015.

[13] B. Kong, Y. Zhan, M. Shin, T. Denny, and S. Zhang, "Recognizing end-diastole and end-systole frames via deep temporal regression network," in *Medical Image Computing and Computer-Assisted Intervention – MICCAI 2016*, (Cham), pp. 264–272, Springer International Publishing, 2016.

[14] F. T. Dezaki, N. Dhungel, A. H. Abdi, C. Luong, T. Tsang, J. Jue, K. Gin, D. Hawley, R. Rohling, and P. Abolmaesumi, "Deep residual recurrent neural networks for characterisation of cardiac cycle phase from echocardiograms," in *Deep Learning in Medical Image Analysis and Multimodal Learning for Clinical Decision Support*, (Cham), pp. 100–108, Springer International Publishing, 2017.

[15] S. Leclerc, E. Smistad, T. Grenier, A. Østvik, F. Espinosa, L. Lovstakken, and O. Bernard, "Deep learning applied to multi-structures segmentation in 2D echocardiography: a preliminary investigation of the required database size," in *IEEE International Ultrasonics Symposium, IUS*, 2018.

Myocardial function imaging in echocardiography using deep learning

5

Andreas Østvik[1,2,4], Ivar Mjåland Salte[5,6], Erik Smistad[1,2,4], Thuy Mi Nguyen[5,6], Daniela Melichova[5,6], Harald Brunvand[5], Kristina Haugaa[6,7], Thor Edvardsen[6,7], Bjørnar Grenne[1,2,3], and Lasse Løvstakken[1,2]

[1] Centre for Innovative Ultrasound Solutions, NTNU, Trondheim, Norway
[2] Dept. of Circulation and Medical Imaging, NTNU, Trondheim, Norway
[3] Clinic of Cardiology, St. Olavs hospital, Trondheim, Norway
[4] Dept. of Health Research, SINTEF Digital, Trondheim, Norway
[5] Dept. of Medicine, Hospital of Southern Norway, Norway
[6] Faculty of Medicine, University of Oslo, Oslo, Norway
[7] Dept. of Cardiology, Oslo University Hospital, Oslo, Norway

Deformation imaging in echocardiography has been shown to have better diagnostic and prognostic value than conventional anatomical measures such as ejection fraction. However, despite clinical availability and demonstrated efficacy, everyday clinical use remains limited at many hospitals. The reasons are complex, but practical robustness has been questioned, and a large inter-vendor variability has been demonstrated. In this work, we propose a novel deep learning based framework for motion estimation in echocardiography, and use this to fully automate myocardial function imaging. A motion estimator was developed based on a PWC-Net architecture, which achieved an average end point error of (0.06 ± 0.04) mm per frame using simulated data from an open access database, on par or better compared to previously reported state of the art. We further demonstrate unique adaptability to image artifacts such as signal dropouts, made possible using trained models that incorporate relevant image augmentations. Further, a fully automatic pipeline consisting of cardiac view classification, event detection, myocardial segmentation and motion estimation was developed and used to estimate left ventricular longitudinal strain *in vivo*. The method showed promise by achieving a mean deviation of (-0.7 ± 1.6)% compared to a semi-automatic commercial

solution for $N = 30$ patients with relevant disease, within the expected limits of agreement. We thus believe that learning-based motion estimation can facilitate extended use of strain imaging in clinical practice.

5.1 Introduction

Motion estimation is an essential part of ultrasound imaging, especially in echocardiography, where it is used to assess cardiac function. Currently speckle tracking echocardiography (STE) is widely deployed, with many methodological variants such as variational optical flow (OF) and block-matching methods [1]. In research, conventional STE methods have been outperformed by phase sensitivity and elastic registration methods [2–4]. Despite being considered the standard, these methods have several unsolved challenges due to fundamental limitations of ultrasound (US) acquisitions. This includes dropouts, shadows, out-of-plane motion, drift sensitivity, foreshortening and more [5]. Several of these artifacts leads to a decorrelation of the US speckle pattern from frame to frame, thus complicating the tracking task.

Deformation imaging, such as measurement of myocardial strain, has shown great potential [6–8], and is claimed to have better diagnostic and prognostic value compared to conventional anatomical measurements such as ejection fraction (EF). Motion estimation (ME) is usually an essential part of these methods, and the measurements are dependent on its performance. Clinical use of deformation imaging is still limited, partly due to time constraints in the clinic, but also a lack of consensus about robustness and reproducibility. We also hypothesize that the retrospective nature of the analysis reduces its use, and believe that having the possibility to quality assure acquisitions while scanning would facilitate clinical implementation. Major efforts have been put into standardization of strain estimation techniques [8, 9]. Part of this involves developing common evaluation platforms and data, in which Alessandrini *et al*. [10] proposed a realistic *in silico* database of US sequences based on simulations with biomechanical models for comparison of STE algorithms.

Recently, motion estimation using convolutional neural networks (CNN) have shown promising results for general optical flow (OF) problems. Dosovitskiy *et al*. [11] demonstrated this, by learning to estimate motion

patterns directly from images using U-Net based architectures called FlowNet. Several flavours of the topology exists, such as FlowNetS, FlowNetC and FlowNet-SD, with decisive modifications, for instance to resolve issues with noisy artifacts and small displacements. By stacking several of these networks in a cascade and using complex training schedules, as in FlowNet 2.0, performance was on par or better than state of the art methods for traditional OF estimation. These methods introduced a shift in OF research, and in few years the work on the topic has increased dramatically, where the benchmarks have been dominated by deep learning (DL) based methods. One of the limitations of FlowNet 2.0 is the network complexity and inference speed. In PWC-Net, the developers succeeded in both increasing the accuracy and reducing the size of the CNN model by leveraging conventional OF components [12]. The PWC-Net and FlowNet architectures are currently the most common starting point for research on DL based OF estimation. The main difference between them is that FlowNet is encoder-decoder based, while PWC-Net use a spatial pyramid.

Using these type of network designs directly for ME in US imaging raises some concerns. Firstly, their design and training regime facilitate correlation between global image features, and an optimization for rigid motion patterns. This is not fully compatible with deformation imaging, where local coherent speckle is used to track local tissue motion and inherent non-rigid deformation patterns. Structures in US images do not have clear borders and traditionally STE has relied on tracking the local speckle pattern, rather than global texture features. On the other hand, speckle decorrelation occurs throughout the cardiac cycle, and this is a fundamental limit of static tracking kernels. Current CNN methods for ME use block matching between features of consecutive frames, but not between lower levels of the architecture [12,13]. This does not distinguish noise and speckle locally, and the cost volume will thus make limited use of coherent speckle between consecutive frames. We thus hypothesize that learning-based ME extended with knowledge from STE methods could improve robustness, and therefore be beneficial.

The use of deep learning based ME in US, and especially in echocardiography, is limited [14]. Earlier, we demonstrated the use of FlowNet 2.0 out-of-the-box for estimating global longitudinal strain (GLS) in a pipeline with view classification, segmentation of the myocardium and state-estimation

techniques [15]. The results were promising, but both the training data and methods had several limitations, especially for regional motion patterns. In elastography, several studies have been conducted with use of FlowNet 2.0 to estimate the displacements [16]. In a recent pilot study, efforts were also made into benchmarking different networks components of FlowNet 2.0, with fine-tuning on simulated US data. The results were on par with current state of the art for flow estimation [17]. In sum, these studies indicate a potential and adaptability for CNN based ME in US image analysis.

Utilizing DL models in a cascade for fully automating clinical measurements have also become a popular research topic, for instance for measurements such as EF and strain [18]. We recently demonstrated an accuracy within interobserver variability on calculations of EF, with possibility for real-time analysis and quality assurance on-site [19]. In this study we aim to extend our work by incorporating ME in an automatic pipeline, in order to do fully automatic deformation measurements. Our goal is to develop DL based methods which may facilitate the implementation of functional imaging in the clinic by removing several steps of manual post-processing and enable real-time use. This could make the measurements more robust and less time consuming.

5.1.1 Main contributions

We propose a novel framework for motion estimation in echocardiography, and use this, together with other relevant components, to fully automate the estimation of longitudinal strain. The contributions of this paper are

- A motion estimator for echocardiography inspired by PWC-Net that incorporates domain knowledge from US, and constraints from relevant morphophysiology.
- A training setup with pretraining on synthetic data, and finetuning on more realistic US simulations with relevant augmentation routines.
- Analysis of deep learning based motion estimation with comparison on simulated and *in vivo* data.
- A fully automated pipeline for longitudinal strain measurements using cardiac view classification, event detection, segmentation and motion estimation by DL.

- Comparison between the automated pipeline and a commercial available system for GLS measurements.

5.2 Methods

5.2.1 Motion estimation with deep learning

Currently, PWC-Net is the most popular architecture for deep learning based optical flow estimation, and several variations exist [12, 20, 21]. It is inspired by conventional OF, including components such as pyramidal coarse-to-fine estimators, warping and cost volume in the design pattern. Still, it utilizes the strengths of CNNs by incorporating feature learning in several stages.

A simplified illustration of the network architecture can be seen in the left part of Fig. 5.1. The core method involves taking two consecutive images as input, and these are fed separately into a learnable CNN based feature extractor pyramid of L levels with shared weights. At each level l, the feature maps from the previous level is downsampled to half its size using strided convolutions. The features of level l of the second image is warped towards the first image using the upsampled flow from the consecutive level. A cost volume is estimated using correlation between the first image and warped features of the second image. For each layer, the cost volume, features from the first image and upsampled optical flow are input into a CNN which outputs a dense displacement map for the current pyramid level. The estimation is repeated upwards in size until the desired level. The output is then forwarded into a context network with dilated convolutions, which refines the flow, taking the estimated flow and features of the second last layer from the OF estimator as input. The final output is a dense displacement map resized to the the same spatial size as the input images.

The original PWC-Net implementation has seven feature pyramid extractor levels including the inputs. The output level is one-quarter of the inputs spatial size, and the flow is upsampled by bilinear interpolation after the context network. The basis of our implementation also has seven pyramid levels, but as opposed to the original implementation which produces flow estimation up to the second highest level, we extend our network to produce flow estimation up to the first level. This is further fed into a context network before the final upsampling as indicated in the right side of Fig. 5.1. Also, we include the final output in the loss function.

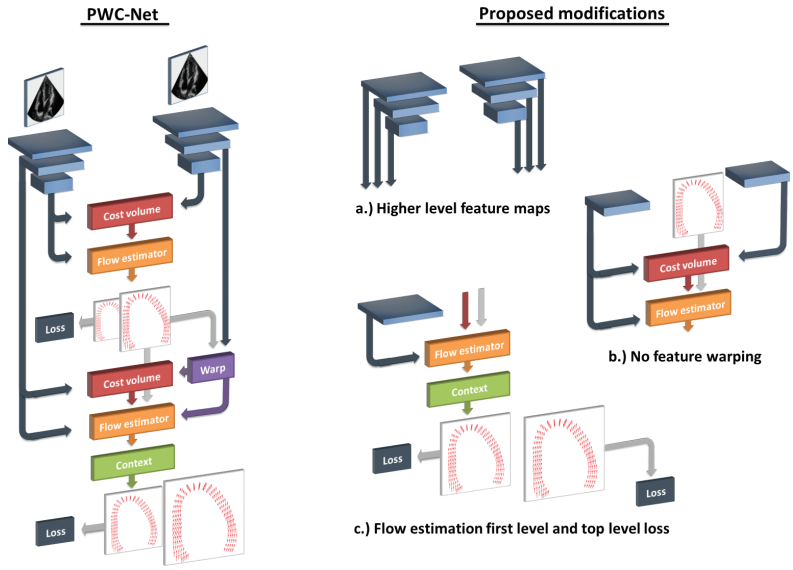


Figure 5.1: Sketch of a traditional PWC-Net architecture with three pyramid levels (left). Two consecutive images are fed into a pyramidal feature extractor. The cost volume is estimated between feature maps of the first image and the backward warped second image features (no warping at bottom). A CNN named Flow estimator, is used to estimate the flow at every level. At the top level a context network is used to refine the flow. The right part of the figure illustrates the modifications done for the EchoPWC-Net. Firstly, feature maps closer to the input is propagated through the cost volume and flow estimation routines (a). Warping of the features of the second image is removed and exchanged with a direct correlation between features (b) and flow at the two highest levels is also included in the loss (c).

This is to retain some of the useful speckle patterns lost when resampling from a low resolution level, and optimize for local variations and small displacements. We also hypothesise that the ambiguity caused by occlusions and out-of-plane motion during warping makes the original implementation problematic for echocardiography. One of the main motivations of using warping is to handle large motions and allow for smaller networks, but for echocardiography the typical motion between frames is small. We therefore remove the warping procedure, and instead estimate the cost volume directly between feature maps at every level. This is illustrated in Fig. 5.1b. We argue that this integrates some of the benefits of block matching between frames at every level of the pyramid, and thus have more resemblance to traditional STE. However, instead of locating the minima of

the cost volume as you would with STE, the full correlation map is passed to the flow estimator. We refer to our customized network as EchoPWC-Net. Additional implementation details are given in Section 5.3.3.

5.2.2 Pipeline for automated functional imaging

The proposed pipeline for myocardial function imaging is summarized in Fig. 5.2. Together with the discussed motion estimation, it consists of several in-house DL based methods, including cardiac view classification, event detection and myocardial segmentation. In addition, we initialize the tracking by extracting the mid ventricular centerline of the myocardium. We summarize the steps in the following, and the reader is referred to published work for more details about the DL networks [22–24].

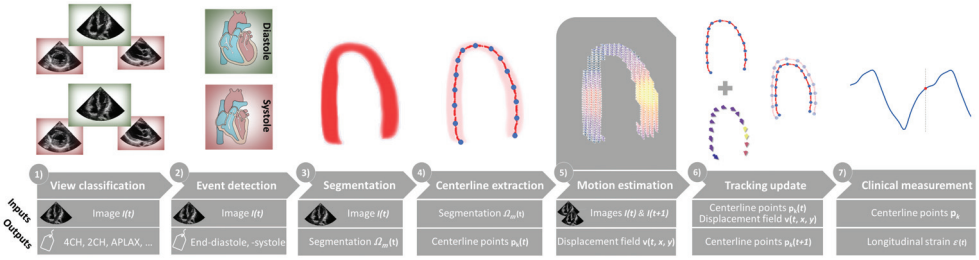


Figure 5.2: The measurement pipeline. Valid US images are forwarded through a segmentation network, and the resulting masks are used to extract the centerline and relevant parts of the image. The US data is further processed through the motion estimation network yielding a map of velocity vectors. The centerline is used to seed points which are used for tracking the myocard. The velocities of the myocardium are optionally used either directly to propagate the centerline points, or as a measurement update step of a Kalman filter. The results are used as a basis for strain measurements.

View classification

To ensure valid acoustic windows, we employ an in-house cardiac view classification (CVC) network [22]. The method recognizes up to eight different cardiac views, including the apical four-chamber (4CH), apical two-chamber (2CH), apical long-axis (APLAX), which are relevant for this study. The network topology is composed of seven block levels of convolution filters, batch normalization, PReLU activation and max pooling. Inception modules and a dense connectivity pattern are employed in the last five

blocks. A global average pooling layer was used before the final softmax activation. It was trained on a dataset of approximately 250 patients, and tested on a similarly sized independent dataset. The network input is standard scan converted B-mode images of size (128×128), and the output is a softmax activation yielding a confidence score for each class. This network has shown an accuracy of 98% and inference time of approximately 4 ms per frame.

Event detection

The cardiac phases are identified using a sequence-to-sequence CNN that can classify diastole and systole directly from B-mode images [23]. The network consists of five stacked levels of 3D convolutions, batch normalization, ReLU activation and max pooling. All convolution kernels have a temporal size of three, while the spatial kernel size is (7×7) for the first layer and (3×3) for the rest. The output of this stage is then propagated into two layers of long short-term memory (LSTM) modules with 32 units each. It was trained and validated on the CAMUS dataset of 500 patients [25]. The network handles variable number of frames with size (128×80) as input, and outputs a sequence of scalars in the interval zero to one. Zero indicates that the image is from the systolic phase, while one is the diastolic phase. End-diastolic (ED) and end-systolic (ES) frames were identified as the temporal points where the phase changes, i.e. cross-over from zero to one and vice versa. The method has shown an accuracy of (-5.5 ± 28.2) ms and (-0.6 ± 31.8) ms on ED and ES frames respectively, and mean absolute error of 1.53 and 1.55 frames from reference. For batch processing, a runtime of 16 ms per frame was measured.

Myocard segmentation

We utilize a segmentation network proven to work well in several studies. This is a slight modification of the U-Net architecture [26], with six levels in the encoder and decoder part. Each level is composed of (3×3) convolution filters and ReLU activation. Max pooling is performed in the encoder part, while upsampling with nearest neighbour interpolation is performed in the decoder part. Skip connections are used between the levels at each stage. The network was first described by Smistad *et al*. [27] and later used in the

CAMUS study of Leclerc *et al*. [25]. Recently, it has been used with success in an automatic measurement pipeline for ejection fraction and foreshortening detection [19]. In this study, we use the segmentation of the myocardium Ω_{m}. Initially, the network was trained for 4CH and 2CH views, but it was later extended to include the APLAX view [24]. It was designed for real-time performance, with 2 million parameters. Network input is an US image of size (256×256) together with a binary value indicating if it is an APLAX view or not. The output is a map of same size of the input image, where each pixel is classified as either LV lumen, myocard, left atrium or background. Data from the CAMUS dataset of 500 patients together with parts of an internal study were used for training. The network achieved a test dice score of 0.79 on the myocardium. A runtime of about 10 ms on a GPU was achieved.

Centerline extraction

The centerline $\mathcal{C}$ of the myocardium is defined by extracting the contour of the myocardial segmentation Ω_{m} and defining the endo- and epicardial borders. Further the base and apex points are defined as the points furthest away from the LV lumen centroid, in left bottom, right bottom and top direction respectively. The centerline is defined as the mid-point between two nearest endo- and epicardial points on the line perpendicular to the longitudinal. A total of k equidistant points $\mathbf{p}_k = \langle x, y \rangle$ along the longitudinal direction is then sampled, i.e. $\mathcal{C} = \{\mathbf{p}_1, \mathbf{p}_2, ..., \mathbf{p}_k\}$.

Motion estimation

The pipeline allows using different motion estimation methods. In this study, we employ four different variants, a traditional Farnebäck optical flow method [28], the FlowNet 2.0, the original PWC-Net and a modified PWCNet which we named EchoPWC-Net. With Farnebäck we use a grid based optimization minimizing average end point error (EPE) on simulated data to find the parameters for window size, pyramid levels, pyramid scale, iterations at pyramid scale, size of the kernel for polynomial expansion and smoothing factor for the derivative of the polynomial. All the methods produce a dense displacement map of velocity components $\mathbf{v}(x, y) = \left\langle v_x, v_y \right\rangle$ between two images $I(t)$ and $I(t+1)$.

Tracking update

The centerline points $\mathbf{p}_k$ can either be updated by propagating the points with the displacement field, i.e. $\mathbf{p}_k(t+1) = \mathbf{p}_k(t) + \mathbf{v}_k(t)$. Alternatively, it can be extracted from the segmentation directly without using the motion estimation method at all.

This step could also involve state estimation techniques such as the Kalman filter [29] or similar, but this was not pursued further in this study.

Clinical measurements

The centerline $\mathcal{C} \subseteq \Omega$ is used to calculate the longitudinal ventricular length ι, i.e. the arc length, for each timestep t. Further, this is used to estimate the Lagrangian strain

$$\epsilon(t) = (\iota(t) - \iota_0)/\iota_0, \tag{5.1}$$

along the center of the myocard. The reference length ι_0 is measured at the ED frame. The peak-systolic strain was used for both GLS and regional longitudinal strain (RLS) estimation, where the peak was defined as the minima between ED and ES strain values. For RLS, we divide the ED centerline at the apex and estimate three equally sized arcs on both sides and compute their strain individually [9].

5.3 Experiments

5.3.1 Datasets

Several datasets were used developing the methods, and in the following we will briefly describe the datasets used for modelling the motion estimation network, and testing the measurement pipeline. For information about data used to train other models, such as segmentation, view classification and event detection, the reader is referred to publications on the specific networks [22–24].

Synthetic data

We used three publicly available datasets commonly used for training and benchmarking of optical flow methods. All the datasets consist of image pairs and a corresponding dense displacement map.

- *FlyingChairs2D [11]:* Contains images of rendered 3D chair models moving in front of random backgrounds scraped from the photo management and sharing site Flickr. A total of 22872 images.
- *FlyingThings3D [30]:* Contains approximately 25000 stereo images sampled from a 3D scene of everyday objects flying along randomized trajectories on a textured background.
- *MPI SINTEL [31]:* Contains images from an open source animated short film. A total of 1628 frames from 35 different animation scenes.

Example image pairs from the datasets with corresponding flow can be seen in the upper part of Fig. 5.3.

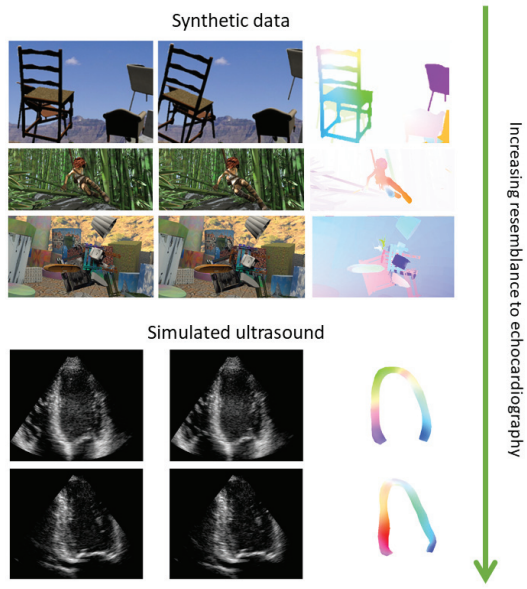


Figure 5.3: Examples from the datasets. Synthetic data from FlyingChairs2D [11], FlyingThings3D [30] and MPI SINTEL [31] and simulated ultrasound [10]. For each example, from left to right, we have two consecutive frames followed by the flow field from the first to the second frame.

Simulated ultrasound data

An open database of simulated echocardiography images created for quality assurance of speckle tracking algorithms [10] were employed. The data is

created with a complex simulation pipeline, where a 3D dataset of simulated US volumes of the heart and corresponding myocardial mesh is spatio-temporally aligned with a 2D template of real US data. Further, a synthetic motion field from a biomechanical model was used to propagate the mesh and aligned data. A scatter map was generated from the composition, and used to generate simulated US. In total, the data is composed of templates from seven different vendors and five motion patterns from the biomechanical model, including one healthy and four pathologies. Each with the three apical views, 4CH, 2CH and APLAX, resulting in a total of 105 sequences or 6165 frames. For each timestep, a set of 180 points divided among five longitudinal lines and six segments is provided by the authors. These points correspond to the underlying motion field of the biomechanical model aligned with the US data. An example of image pairs of 4CH and 2CH views from the dataset with corresponding flow can be seen in the lower part of Fig. 5.3. They also provide the view and the cardiac event timing for each sequence.

Clinical data

A dataset was collected from a clinical database of patients diagnosed with acute myocardial infarction (MI) or de-novo heart failure (HF) at a Norwegian hospital. The study was approved by the regional ethics committee (ref. 2013/573) and written consent was given by all patients. The images were acquired using GE Vingmed (Vivid 7, E9 or E95) scanners. Patients were included consecutively regardless of image quality. All exams were performed in clinically stable patients with sinus rhythm. Images were analyzed by a single clinician using clinical best practice as defined in [7] with the 2D strain (2DS) application in the clinical software EchoPAC release 202 (GE Vingmed AS, Horten, Norway.). This ensures that the proposed automated method is compared to actual clinical practice measurements techniques. To ensure a representative range of LV pathologies, a total of 30 patients from five different cohorts were randomly selected, resulting in six patients from each group. The groups were defined by a diagnosis of ST elevation MI (STEMI), non-ST elevation MI (NSTEMI), ischemic heart failure (HF), non-ischemic HF and no significant disease. The tracked mid ventricular points and corresponding strain values were exported from the software.

5.3.2 Data augmentation

Due to the unrealistic nature of simulated ultrasound and limited access to relevant (*in vivo*) data with a ground truth, we rely on several US-specific augmentation routines. Here we try to induce realistic artifacts common in echocardiography that usually hampers the success of speckle tracking, and describe a selection in the following.

Gaussian shadowing

Acoustic shadows often occur in US imaging due to structures that strongly reflect or absorb the US waves. This is often identified as a dark region behind the structure. We mimic this effect by placing random regions of intensity reductions in the image. Similar methods have been shown to have an effect on generalization for US segmentation tasks [32].

Haze artifact application

One artifact that is prevalent for some patient is acoustic haze. This can be identified as a semi-static noise band in the upper parts of the image. We randomly apply static high intensity artifacts with a Gaussian profile along the radial direction in polar coordinates.

Depth attenuation

The US wave looses energy as it travels through the body, and this can be identified as a gradual drop in intensity with distance from the probe. Similarly like the haze artifact application, we apply a varying degree of intensity attenuation along the radial direction. The attenuation does not consider depth independent noise, and is thus a simplification of the physical artifact.

Speckle reduction

The speckle pattern in images from different vendors often differ due to image enhancement and various filtering methods. To reproduce this effect, we smooth the images randomly using a bilateral filter, effectively reducing the speckle.

In addition to these US specific augmentations, we apply basic augmentations such as horizontal and vertical flipping, temporal reversing, frame skipping, rotation, random noise, scaling, image resampling artifacts, JPEG compression and gamma intensity transformations. Except for the flipping, reversing, scaling, skipping and rotation, the displacement map was not modified for any of the augmentation routines. All augmentations are applied in random combinations and on-line while training. Examples from some of the individual augmentations are given in the supplementary material. In addition, the effect on maximum flow distribution after five epochs with augmentation is visualized.

5.3.3 Implementation details

We implemented the machine learning environment using Tensorflow [33] version 2. The modelling and experiments were conducted on a workstation with an Ubuntu 16.04 operating system. The hardware consisted of an Intel Xeon CPU E5-2637 v2 with a clock speed of 3.50 GHz, 112 GB RAM and a NVIDIA Titan V GPU with 12 GB of memory.

Architecture parameters

For the feature extractor in EchoPWC-Net, we use one convolution layer in addition to the strided convolution, with equal amount of filters at each level. The amount of filters used was 16, 32, 64, 96, 128 and 192, from top to bottom level respectively. For the cost volume we use a search range of 4 for every level, and for the context network we use the architecture proposed in the original implementation [34]. This corresponds to a receptive field of 67×67 in the last layer.

Data preprocessing

To reduce the feature space and adapt for US, we converted all input data to grayscale, including the synthetic RGB data. The input was set to a fixed size of (448×576). As mentioned, the simulated US data is provided together with a set of 180 spatial points inside the myocardium for every timestep. We use these to generate a sparse displacement field, and we use cubic interpolation to convert to a dense displacement map with velocities $\mathbf{v}(x, y) = \left\langle v_x, v_y \right\rangle$ inside the myocardium Ω_{m}. To avoid boundary effects,

we extrapolate the epicardial points radially by 5% of the radial diameter, followed by masking by the concave hull enclosing the original points. The dense displacement map is used as the ground truth flow, and the units were set to pixel per frame.

Training procedures

Several training dataset schedules were investigated, resulting in four different setups:

- **Synthetic RGB:** Sequential training from scratch with FlyingChairs2D and FlyingThings3D RGB data.
- **Synthetic gray:** Sequential training from scratch with grayscale FlyingChairs2D and FlyingThings3D.
- **Synthetic gray → Simulated US:** Initialized with weights from synthetic gray followed by fine-tuning on simulated US.
- **Simulated US:** Trained from scratch on simulated US.

When training with synthetic data, we employ the basic augmentations mentioned in Section 5.3.2. In addition, we employ the US specific augmentation when training on simulated US.

Our models are trained with the Adam optimizer and a batch size of 4 for all experiments. The initial learning rate was set to 10^{-4}, with a halving schedule for each 100k and 20k iterations for synthetic and simulated US respectively. For fine-tuning, the initial learning rate was set to 10^{-5}. Training time from scratch was approximately three-four days for synthetic data with the fine-tuning schedule running over five days, and two days for simulated US. Early stopping with a patience of 30 epochs were used for all models.

Loss

We use a multi-scale loss function with end-point error. Since the labeled motion is sparse, we only optimize regionally where the input lies within the predefined segmented region Ω_{m}. We let $\mathbf{w}^l$ denote the dense flow field at the lth pyramid level. The loss is defined as

$$\mathcal{L}(\Theta) = \sum_{l=l_0}^{L} \beta_l \sum_{\mathbf{x}} |\tilde{\mathbf{w}}_{\Theta}^{l}(\mathbf{x}) - \mathbf{w}_{GT}^{l}| + \gamma|\Theta| \quad \forall \quad \mathbf{w} \in \Omega_{\mathrm{m}},$$

where Θ is the parameters and $\mathbf{x}$ is the inputs. The term β_l is set manually and used to weight the loss contribution from each layer. The second term regularizes the parameters, where γ is the regularization factor. This is similar to the loss used in FlowNet and PWC-Net, but restricted to regional optimization. In our implementation, we set the weights to $\beta_0 = 0.015$, $\beta_1 = 0.03$, $\beta_2 = 0.06$, $\beta_3 = 0.12$, $\beta_4 = 0.25$, $\beta_5 = 0.50$ and $\beta_6 = 1.0$. Based on the input size, this correspond to equally weighting each layers contribution to the loss. The regularization factor γ was set to 10^{-4}.

5.3.4 Evaluation

Metrics

We evaluate our methods using the end point error (EPE), which is a common metric for benchmarking optical flow performance. It is defined as the Euclidean distance between the ground truth velocity and the predictions, i.e. EPE = $\|\mathbf{v}_{\mathrm{GT}} - \mathbf{v}_{\mathrm{pred}}\|$. We also compute the strain values, as defined in (5.1). Regional strain is computed for each segment for each view, while global strain is computed for each view, and averaged over all views. In addition, we report correlation metrics, such as regression slope α and correlation coefficient ρ, as well as bias μ and 95 percentile limits of agreement (LOA).

Comparison I: Motion estimation

Nine different motion estimation methods are evaluated. The original PWC-Net, as well as different flavours of the EchoPWC-Net. For reference, the Farnebäck and FlowNet 2.0 methods are also included. The various methods are summarized in Table 5.1.

Comparison II: Automatic pipeline

For functional measurements on *in vivo* data, we also test two variants of the presented pipeline:

Table 5.1: Overview of different motion estimation models.

	Training dataset schedule			Augmentation	
	Synth. RGB	Synth. gray	Sim. US	Basic	US spec.
Farnebäck					
FlowNet 2.0	✓			✓	
PWC-Net	✓			✓	
PWC-Net-gray		✓		✓	
PWC-Net-gray-usft		✓	✓	✓	✓
PWC-Net-us			✓	✓	✓
EchoPWC-Net		✓		✓	
EchoPWC-Net-usft		✓	✓	✓	✓
EchoPWC-Net-us			✓	✓	✓

- **Segmentation only**: Recalculation of the centerline for every time point, and not using motion estimation. This refers to skipping part 5) of the measurement pipeline.

- **Tracking**: Initialization of centerline by segmentation, and propagation of points using the best performing motion estimation model. This refers to the full pipeline described earlier.

Comparison III: Model adaption

As mentioned, one of the limitations of traditional speckle tracking is the adaptability to various noise prevalent in US. To study the investigated methods ability to regularize, we design an evaluation strategy based on three of our US relevant augmentation routines, namely Gaussian shadow, haze artifact application and depth attenuation. More specifically, for Gaussian shadowing we apply a shadow region at the center of the mid septal segment of test data samples, and measure the change in relative EPE as a function of shadow amplitude, i.e. the relative degree of intensity signal. The size of the region was set to 20% of the image size in both directions, tuned to cover the whole segment for higher shadow amplitudes. For haze, we apply a localized band of haze mimicking noise in the upper half of the sector with an increasing intensity value. Finally, for depth attenuation we attenuate the intensity values gradually, with a fixed saturation area at the base level of the myocardium.

Table 5.2: Results on simulated ultrasound data. Average end point error (EPE) for (a) every vendor, (b) average over segments and (c) apical views. Units given in mm per timestep/frame ΔT^{-1}.

(a) Vendors

Method	ESAOTE [mm·ΔT^{-1}]	GE [mm·ΔT^{-1}]	Hitachi [mm·ΔT^{-1}]	Philips [mm·ΔT^{-1}]	Siemens [mm·ΔT^{-1}]	Toshiba [mm·ΔT^{-1}]	Samsung [mm·ΔT^{-1}]
Farnebäck	0.08 (0.06)	0.09 (0.07)	**0.06 (0.04)**	0.08 (0.06)	**0.06 (0.05)**	0.07 (0.05)	0.07 (0.05)
FlowNet 2.0	0.12 (0.10)	0.17 (0.13)	0.10 (0.08)	0.11 (0.08)	0.09 (0.08)	0.10 (0.08)	0.11 (0.09)
PWC-Net	0.12 (0.09)	0.13 (0.10)	0.10 (0.07)	0.10 (0.07)	0.09 (0.07)	0.10 (0.07)	0.10 (0.08)
PWC-Net-gray	0.19 (0.16)	0.21 (0.19)	0.13 (0.09)	0.15 (0.10)	0.12 (0.09)	0.15 (0.12)	0.17 (0.13)
PWC-Net-gray-usft	0.14 (0.10)	0.17 (0.12)	0.13 (0.09)	0.14 (0.10)	0.14 (0.10)	0.14 (0.11)	0.13 (0.09)
PWC-Net-us	0.10 (0.08)	0.12 (0.10)	0.10 (0.08)	0.10 (0.08)	0.10 (0.08)	0.09 (0.07)	0.09 (0.07)
EchoPWC-Net	0.17 (0.17)	0.19 (0.20)	0.12 (0.09)	0.14 (0.11)	0.12 (0.09)	0.13 (0.10)	0.13 (0.10)
EchoPWC-Net-usft	0.09 (0.11)	0.11 (0.12)	0.09 (0.08)	0.09 (0.08)	0.10 (0.08)	0.08 (0.06)	0.07 (0.07)
EchoPWC-Net-us	**0.07 (0.06)**	**0.07 (0.06)**	**0.06 (0.04)**	**0.06 (0.05)**	**0.06 (0.05)**	**0.06 (0.04)**	**0.05 (0.04)**

(b) Segments

Method	Base [mm·ΔT^{-1}]	Mid [mm·ΔT^{-1}]	Apical [mm·ΔT^{-1}]	Average [mm·ΔT^{-1}]
Farnebäck	0.10 (0.07)	0.07 (0.05)	0.05 (0.04)	0.07 (0.06)
FlowNet 2.0	0.15 (0.10)	0.12 (0.08)	0.07 (0.07)	0.12 (0.09)
PWC-Net	0.14 (0.09)	0.10 (0.07)	0.08 (0.07)	0.11 (0.07)
PWC-Net-gray	0.21 (0.16)	0.14 (0.11)	0.12 (0.11)	0.16 (0.13)
PWC-Net-gray-usft	0.19 (0.12)	0.15 (0.10)	0.09 (0.07)	0.14 (0.10)
PWC-Net-us	0.14 (0.09)	0.10 (0.07)	0.06 (0.05)	0.10 (0.08)
EchoPWC-Net	0.19 (0.16)	0.13 (0.10)	0.10 (0.08)	0.14 (0.11)
EchoPWC-Net-usft	0.12 (0.11)	0.09 (0.08)	0.08 (0.07)	0.10 (0.08)
EchoPWC-Net-us	**0.08 (0.06)**	**0.06 (0.04)**	**0.04 (0.03)**	**0.06 (0.04)**

(c) Views

4CH [mm·ΔT^{-1}]	2CH [mm·ΔT^{-1}]	APLAX [mm·ΔT^{-1}]	Average [mm·ΔT^{-1}]
0.07 (0.05)	0.08 (0.06)	0.08 (0.06)	0.07 (0.06)
0.10 (0.08)	0.12 (0.10)	0.12 (0.09)	0.12 (0.09)
0.10 (0.07)	0.11 (0.08)	0.11 (0.08)	0.11 (0.08)
0.15 (0.11)	0.17 (0.15)	0.16 (0.12)	0.16 (0.13)
0.13 (0.09)	0.14 (0.10)	0.15 (0.10)	0.14 (0.10)
0.10 (0.07)	0.10 (0.08)	0.10 (0.08)	0.10 (0.08)
0.13 (0.12)	0.14 (0.16)	0.14 (0.14)	0.14 (0.14)
0.11 (0.08)	0.10 (0.09)	0.10 (0.09)	0.10 (0.09)
0.06 (0.04)	**0.06 (0.04)**	**0.06 (0.05)**	**0.06 (0.04)**

5.4 Results

5.4.1 Simulated ultrasound

The PWC-Net model trained on grayscale FlyingChairs and FlyingThings3D achieved an average EPE of 4.80 and 6.41 on MPI Sintel Clean and Final respectively. Cross-validation was performed on the simulated US data by dividing into folds by vendor. This resulted in seven training sessions for each DL method. For the Farnebäck method, the grid based optimization yielded best EPE for 3 pyramid levels, with a scale of 0.5 and a window size of 69. A total of 5 iterations for each scale, a size of 5 pixels of the kernel for polynomial expansion and a smoothing factor of 1.1. The average EPE with corresponding standard deviation can be seen in Table 5.2a. On the respective test data, the results for segments and views are reported in Table 5.2b and 5.2c respectively. A correlation plot of the regional strain estimation for one of the vendors can be seen in Fig. 5.4, while the plots for all vendors can be found in the supplementary material. An example of

qualitative results for the different methods are shown in Fig. 5.5.

For reference, the correlation metrics of our implemented methods, compared to the work by Alessandrini *et al.*, is shown in Table 5.3, and also indicated in the correlation plots. The average over all vendors is used for the different metrics.

Results from our model adaption study is given in Fig. 5.6. Here, the EchoPWC-Net-us models adaptability is measured with respect to increasing application of different augmentation effects. The relative error of average EPE as a function of specified effect is given. It is worth noting that the baseline of the two models are different, i.e. the Farnebäck method has on average a higher average EPE than EchoPWC-Net-us with no shadows. The absolute deterioration is therefore higher for Farnebäck in all cases.

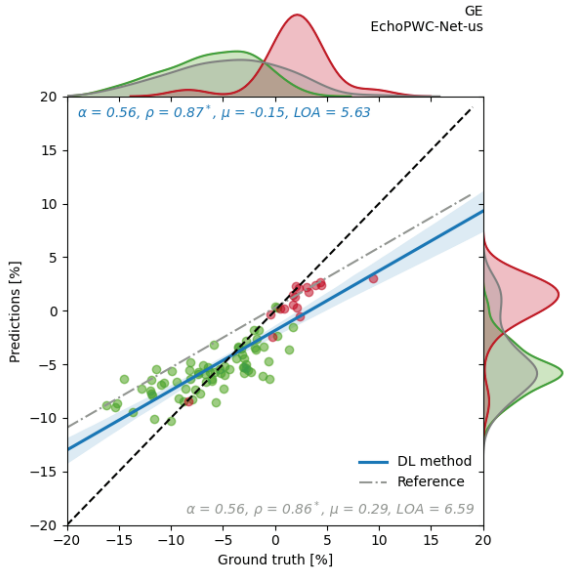


Figure 5.4: Correlation plot between the ground truth regional strain estimation and the DL method on simulated data from one selected vendor. Green dots represent healthy myocardial segments, while the red sick segments. In the top left corner, the slope α of the regression line, correlation coefficient ρ, bias μ and limits of agreement (LOA) is given. The corresponding reference values from Alessandrini *et al.* [10] are given in the bottom right corner.

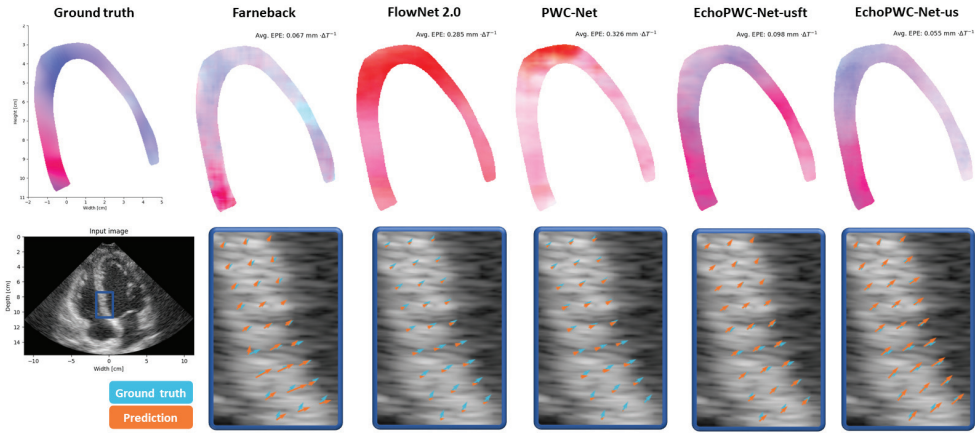


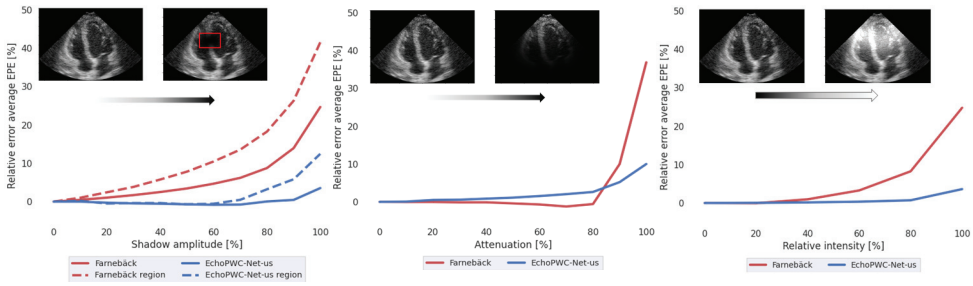
Figure 5.5: Example of predicted flow patterns for the different methods within the myocardium. The upper part is color coded with hue values, color and saturation indicates direction and magnitude respectively. The average EPE is given in the upper right corner of the image of each case. The bottom part of the image shows the velocity vector comparison between ground truth and the different methods inside the base septum segment as indicated by the blue bounding box in the US image. Light blue arrows are ground truth, while orange arrows are predictions.

Table 5.3: Comparison of the considered methods averaged over the different vendors in the simulated US data. Metrics include slope of the regression line α, correlation coeffisient ρ, bias μ and 95% limits of agreement (LOA).

Method	α	ρ	μ	LOA
Alessandrini *et al.* [10]	0.55 (0.14)	0.75 (0.13)	0.37 (0.65)	6.98 (1.53)
Farnebäck	0.42 (0.15)	0.65 (0.16)	-0.16 (0.42)	7.38 (1.73)
FlowNet 2.0	0.25 (0.15)	0.49 (0.19)	-1.70 (0.86)	8.76 (1.77)
PWC-Net	0.35 (0.12)	0.58 (0.15)	0.01 (0.40)	8.01 (1.51)
PWC-Net-gray	0.24 (0.09)	0.37 (0.18)	-0.58 (0.52)	10.07 (1.94)
PWC-Net-gray-usft	0.55 (0.11)	0.66 (0.14)	-0.32 (0.54)	7.63 (1.75)
PWC-Net-us	0.57 (0.12)	0.69 (0.14)	-0.30 (0.50)	7.40 (1.25)
EchoPWC-Net	0.38 (0.25)	0.44 (0.29)	0.96 (0.74)	10.53 (4.03)
EchoPWC-Net-usft	0.48 (0.13)	0.67 (0.14)	1.11 (0.78)	7.43 (1.43)
EchoPWC-Net-us	**0.60 (0.10)**	**0.84 (0.07)**	**0.11 (0.37)**	**5.45 (1.19)**

5.4.2 Clinical data

The ME methods were used on clinical *in vivo* data, and compared to a commercial system by estimating the average EPE. In Table 5.4 the results are shown for three cardiac views, and the corresponding average. We also tested the best performing model from the simulation study on this data



(a) Shadow **(b)** Depth attenuation **(c)** Haze

Figure 5.6: Testing of model adaptation abilities by measuring relative average end point error (EPE) as a function of fractional increase in augmentation effect. The Farnebäck method and EchoPWC-Net-us is plotted as red and blue lines respectively. **(a)** The shadow is applied in a specific region of the US image, as indicated by the red bounding box, and the EPE is calculated both regionally inside this box (dashed) and for the entire myocardium defined by the segmentation (solid). **(b)** Depth attenuation is applied with a fixed saturation area close to the base of the myocardium in radial coordinates. **(c)** Haze is applied to a fixed area in the upper half of the myocardium.

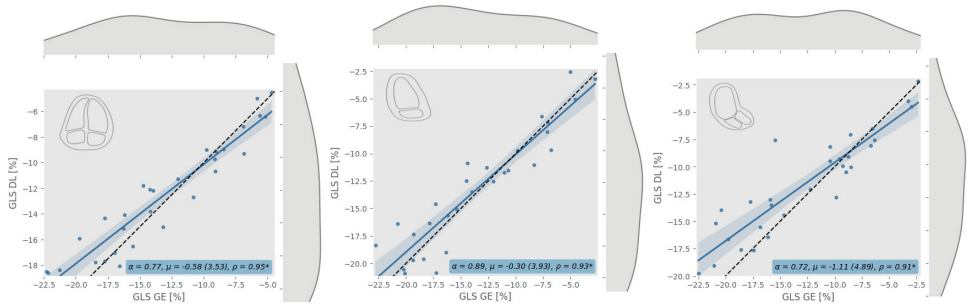
using our pipeline for automated functional imaging. The pipeline were tested with two different flavours as specified earlier, and a summary is given in Table 5.5. For the tracking method, a correlation plot of the GLS for each individual view is given in Fig. 5.7, while the average over all views is given in Fig. 5.8. For additional detail, Bland-Altman plots of the test data are also presented in the supplementary material.

Table 5.4: Average end point error and standard deviation (parenthesis) for every view on clinical data compared to a commercial method.

Method	A4C [mm·ΔT^{-1}]	A2C [mm·ΔT^{-1}]	APLAX [mm·ΔT^{-1}]	Average [mm·ΔT^{-1}]
Farnebäck	0.19 (0.08)	0.18 (0.08)	0.19 (0.09)	0.19 (0.08)
FlowNet 2.0	0.19 (0.08)	0.18 (0.07)	0.20 (0.09)	0.19 (0.08)
PWC-Net	0.24 (0.09)	0.24 (0.09)	0.25 (0.09)	0.24 (0.09)
PWC-Net-gray	0.25 (0.11)	0.26 (0.12)	0.27 (0.12)	0.26 (0.12)
PWC-Net-gray-usft	0.19 (0.08)	0.19 (0.08)	0.19 (0.09)	0.19 (0.08)
PWC-Net-us	0.19 (0.08)	0.18 (0.08)	0.19 (0.08)	0.19 (0.08)
EchoPWC-Net	0.23 (0.10)	0.24 (0.11)	0.26 (0.12)	0.25 (0.11)
EchoPWC-Net-usft	0.17 (0.07)	0.18 (0.07)	0.18 (0.08)	0.18 (0.07)
EchoPWC-Net-us	**0.16 (0.07)**	**0.16 (0.07)**	**0.17 (0.08)**	**0.16 (0.07)**

Table 5.5: Average difference and standard deviation (paranthesis) for global longitudinal strain on clinical data comparing to a commercial method.

Method	A4C [%]	A2C [%]	APLAX [%]	Average [%]
Segmentation only	-0.54 (2.51)	-0.34 (3.45)	-0.03 (5.00)	-0.28 (2.36)
Tracking	-0.58 (1.79)	-0.30 (1.99)	-1.11 (2.50)	-0.71 (1.63)



(a) 4CH **(b)** 2CH **(c)** APLAX

Figure 5.7: Correlation plot of global longitudinal strain (GLS) estimates between commercial system and deep learning based method for specific views. Each dot represents one subject. In the bottom right corner, the slope α of the regression line, bias μ with limits of agreement (LOA) of 1.96σ in parenthesis, and correlation coefficient ρ is given.

5.4.3 Runtime performance

The EchoPWC-Net achieves a runtime of (18.9±0.7) frames per second (FPS), while the Farnebåck method can process at (8.8±0.1) FPS. The frame rates of the different pipelines for estimating global longitudinal strain was (30.8 ± 0.9) FPS and (15.6 ± 0.3) FPS for segmentation and tracking respectively.

5.5 Discussion

We have presented a method for motion estimation using DL and integrated this successfully in a pipeline for longitudinal strain measurements. The ME method is inspired by PWC-Net and relevant training strategies, but with modifications to make it more compliant for myocardial tracking. Our choices have an intuitive motivation, firstly to increase the resemblance to echocardiography for the training data by using simulated US and relevant

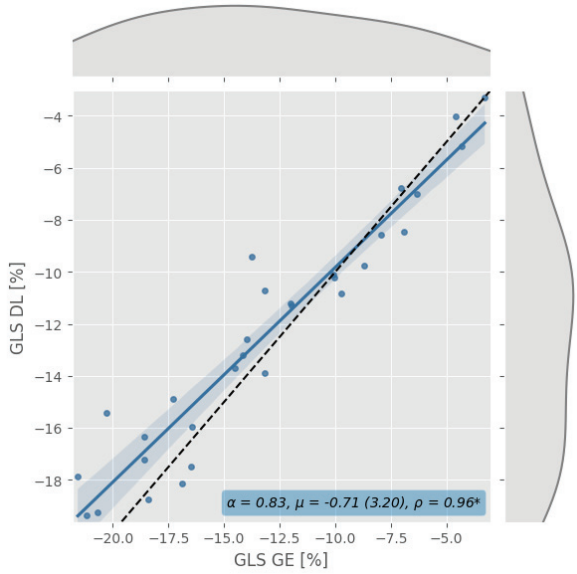


Figure 5.8: Correlation plot of global longitudinal strain (GLS) estimates between commercial system and deep learning based method averaged over the three apical views. Each dot represents one subject. In the bottom right corner, the slope α of the regression line, bias μ with limits of agreement (LOA) of 1.96σ in parenthesis, and correlation coefficient ρ is given.

augmentations. Secondly, to improve the tracking task by incorporating a more direct correlation between features and loss optimization for low level feature learning, including the cost volumes at every pyramid level.

The motion magnitude of common datasets used in OF research, such as *FlyingChairs2D* and *FlyingThings3D*, is on average much higher than the displacement between frames in typical echocardiography data. This is illustrated in the supplementary material. We thus question the validity of these datasets for pretraining. As shown in Table 5.2, training on simulated US data alone gives significantly better results. Using pretraining datasets with lower average flow could improve the results of fine tuning, but was not pursued here. We further observed a mismatch between the simulated US data and the clinical *in vivo* data, where the average maximum flow distribution for the latter was about twice as high. We used data augmentations to tackle this problem, but expect that an improvement of the training data quality and size will further improve the models in later iterations.

One motivation for using warping in CNNs is to mitigate the need of a large search range in the cost volume estimation. Due to the lower flow magnitudes between frames in echocardiography compared to general OF problems, incorporating the direct cost volume between features was feasible. In addition to increasing the general performance of the network for deformation imaging, the modifications introduced in EchoPWC-Net may also cause less ambiguity for occluded areas resulting from warping [35]. We believe further optimizations can be made, for instance an adaptive search range when calculating the cost volume would potentially reduce the runtime. Also the size of the pyramid can probably be reduced.

Table 5.2 suggest that the ME method producing best results is the EchoPWC-Net-us, which is trained from scratch with simulated data and several US-specific augmentation routines. Results were consistent across vendors and views. For segments, the absolute error is decreasing towards the apex, which is expected. The distribution of velocity vectors is limited for the dataset which may influence the trained models ability to generalize. Compared to the FlyingChairs dataset, the typical maximum velocity magnitude of the simulated US data is more than ten times lower. This partially explains the mismatch between the fine-tuned model and the model trained from scratch, as the latter will be biased to a lower velocity field. The qualitative results in Fig 5.5 further suggest the mismatch, where FlowNet 2.0 and PWC-Net yield similar results, but relatively far from the ground truth. The Farnebäck method, as well as the models trained on US data, yields good results across the entire myocardium. Noticeably, the prior has a more noisy pattern compared to the DL methods.

For strain values, the tendency of EchoPWC-Net-us is a slight underestimation for healthy segments, and a slight overestimation for sick segments. This is evident from Fig. 5.4, and also from the correlation plots in the supplementary material. Noticeably, the majority of peak strain values are below 10%, and is generally low compared to clinical data. Again this indicate some limitations in the training data. A comparison to the average strain values reported by Alessandrini *et al*. [10] for the same data is given in Table 5.3. The EchoPWC-Net-us method performs slightly better on average across vendors, especially considering the variance. Although there is potential for further improvement, these findings suggest that learning

based methods can perform on par or better compared to state of the art on simulated data.

One of the major motivations of investigating the use of DL based methods, is their ability to adapt to the data representation used while training. The use of augmentation routines mimicking typical image artifacts in this scenario is therefore very appealing, as it could also address some of the big challenges with traditional methods. The result presented in Fig. 5.6 shows a significant improvement over a traditional OF method. For Gaussian shadowing the relative regional error is increased by less than 10% for EchoPWC-Net-us, while over 40% for the Farnebäck method. Similar effects can be seen for depth attenuation and haze application. This suggest that the ME model actively uses features from lower levels of the pyramid, and potentially the context network, in order to get the necessary global context for filling in parts of missing data. A qualitative comparison of the methods can also be found in the supplementary material. Here, the predicted flow is visualized with and without artifact for both methods. The results from the model adaption study show that the benefits of augmentation routines are twofold; in addition to increasing the effective size of the dataset, the models become more robust to image artifacts. Further studies must be conducted to evaluate the effect *in vivo*, but we emphasize the advantages of incorporating relevant augmentations in the training stage.

The average EPE on clinical data is significantly higher than for simulated data. We also notice that the relative improvement by training on simulated data is less effective *in vivo*. This may be due to the limited range of displacements present for the training data. Further, as the underlying biomechanical motion model is equal across vendors, we also suspect a slight overfit to the motion model. As can be observed in Table 5.4, performance improves as more distinctly relevant data is included, and the lack of relevant training data is thus believed to be a limitation which needs to be addressed.

For calculation of GLS using the pipeline we see from Table 5.5 that the measurement variance improves significantly using tracking instead of segmentation alone. The segmentation model is trained on ED and ES frames, thus calculating end-systolic strain instead of peak strain is expected to yield more similar results. We also note that the results from

the APLAX view is worse than for A4C and A2C for both approaches. As the motion estimation is rather consistent across views, the myocard segmentation and the centerline extraction is the main source of this discrepancy. As shown in previous work, the segmentation performs worse on the ALAX view [24]. Also, the asymmetry of the view can complicate the centerline extraction. Better overall results can therefore be achieved by improving these components. As seen in Fig. 5.7 and Fig. 5.8 it is a significant correlation between the methods. However, we also notice a similar tendency as for simulated data, with an underestimation of larger strain values, and an overestimation of lower. The range of strain values for the DL method is therefore slightly smaller compared to the commercial method. The most probable reason for this is again the training data, where low strain values are highly over-represented [10].

In the vendor comparison study [8], the commercial system used for our *in vivo* data overestimates strain values by an average of 1.6% compared to the mean of all vendors. Also, software only methods from Epsilon and TomTec achieve a mean difference for GLS of $(2.50 \pm 1.94)\%$ and $(-0.70 \pm 1.68)\%$ respectively when comparing to GE. This suggests that our average difference of $(0.71 \pm 1.63)\%$ is within limits of agreement of what can be expected from different commercial systems when evaluated on the same data.

The average runtime of the networks and pipelines are reasonable compared to previously reported findings [12, 19]. The CNN is fast compared to the Farnebäck implementation used, but more work must be conducted to achieve real-time performance in echocardiography. As mentioned, we believe the network can be pruned substantially, for instance by making the search range in the cost volume adaptive for the different pyramid levels.

Although the results are encouraging, we believe there are several points that can be highlighted as a recommendation for further work in addition to what is already mentioned. Post-processing and regularization are a common part of the general workflow of strain computation [4]. This includes drift compensation, temporal and spatial smoothing, as well as state estimation or recurrent methods. In this work this was not extensively investigated, but we believe it could enhance the results if important factors are considered. Post-processing can reduce the noise, but it can

also limit the range of strain values and thus reduce the ability to detect local abnormalities. Further, an investigation of regional motion patterns and strain *in vivo* is hard to validate, but still a direction that should be pursued to establish robust methods that allows for extended clinical use of these sensitive measurements. Representative data is the key, and continued efforts should be made to establish a larger database for training and validating motion estimation methods in echocardiography.

5.6 Conclusion

In this paper we present a novel pipeline for myocardial function imaging in echocardiography using deep learning. We demonstrate that a modified PWCNet motion estimation network named EchoPWC-Net can perform on par or better compared to other known methods when training on simulated ultrasound data. Results are within limits of agreements of relevant work and commercial systems, both on *in silico* and *in vivo* data. We argue that the main limitations stems from limited training data, and that the results can be further improved by increased data volume, and resemblance to clinical echocardiography. Our pipeline is able to estimate longitudinal strain automatically in a prospective nature. By being simple and robust, we believe these methods can facilitate the use of deformation imaging in the clinic.

5.7 Appendix

In this supplementary appendix we provide additional examples and results from the study. Section 5.7.1 contains illustrations of some metrics of the used datasets, while in Section 5.7.2 some examples of US relevant augmentations are displayed. In Section 5.7.3 we show some visualization results from our adaption study, and detailed results from the vendor comparison of simulated data and clinical analysis.

5.7.1 Datasets

In Fig. 5.9a and Fig. 5.9b the maximum flow distributions of the simulated dataset without and with augmentations are visualized. The simulated data

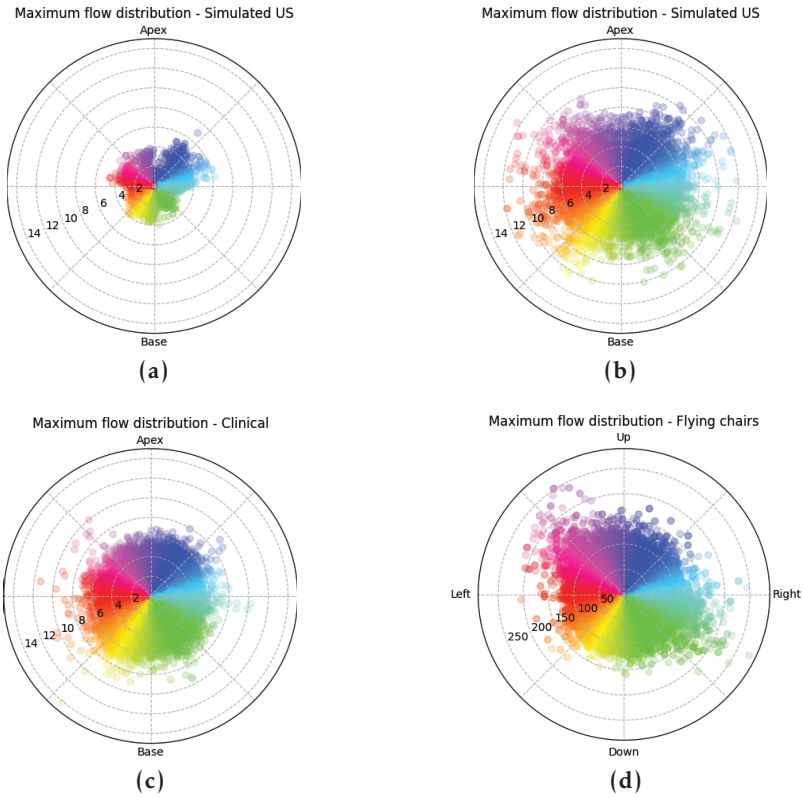


Figure 5.9: Maximum flow distribution of (**a**) simulated US without augmentation, (**b**) simulated US with augmentations after five epochs, (**c**) clinical data used in the study without augmentation and (**d**) the flying chairs dataset used in the study without augmentation. There is a significant mismatch between the clinical and simulated dataset without augmentation. The plots are in polar coordinates, where motion towards the heart apex is upwards, and towards the base is downwards (**a**-**c**). The color hue indicates the flow direction, and the transparency the density of samples. The magnitude is in pixel per frame.

is limited, which is also prevalent by inspecting the magnitude and direction of the motion.

Augmentations that affect the motion field are horizontal and vertical flipping, temporal reversing, skipping of frames and rotation. The effect is noticeable comparing to the example without augmentation. By applying augmentations, the maximum flow distribution is more aligned with the clinical data, which is illustrated in Fig. 5.9c. We can also argue that the clinical data is more complex, as it is also more evened in all directions.

5.7.2 Example augmentations

In Fig. 5.10 an example of the depth attenuation applied to a raw image is visualized. Similarly, Fig. 5.11 shows an example of the haze application, while Fig. 5.12 shows an example of shadowing. The mappings are given for every example.

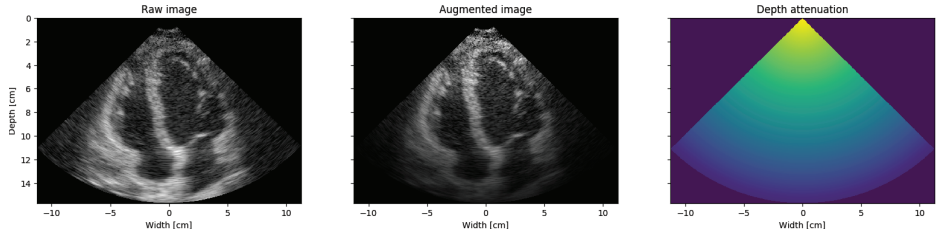


Figure 5.10: Example of depth attenuation augmentation. From left to right, the raw image, the augmented image and the map used to augment the image are displayed respectively. The raw images is multiplied by the attenuation map.

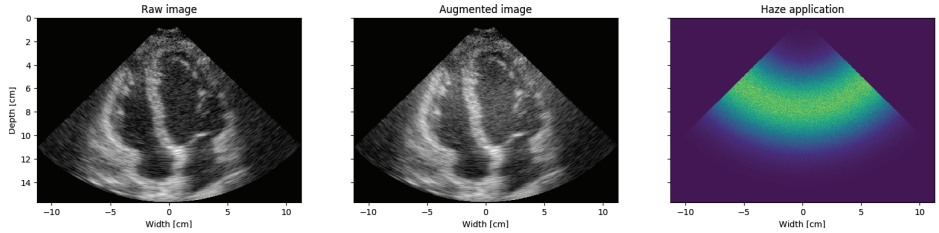


Figure 5.11: Example of haze application augmentation. From left to right, the raw image, the augmented image and the map used to augment the image are displayed respectively. The haze is applied in an additive manner.

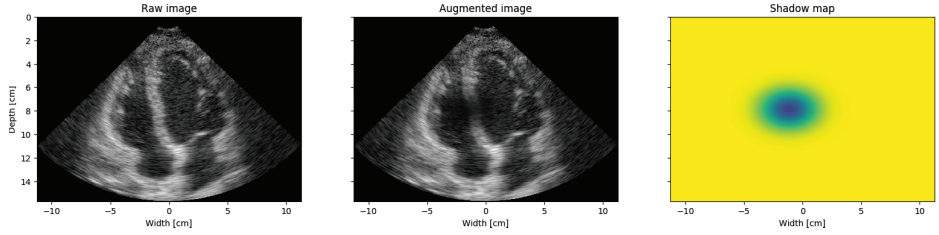


Figure 5.12: Example of shadow augmentation. From left to right, the raw image, the augmented image and the map used to augment the image are displayed respectively. The raw images is multiplied by the shadow map.

5.7.3 Results

In Fig. 5.13 the correlation plots between the ground truth regional strain estimation and the Echo-PWC-Net-us method is given. Some qualitative results from the model adaption study is given in Fig. 5.14 and Fig. 5.15 using the Farnebäck and Echo-PWC-Net-us model respectively. For additional detail about the clinical results, Bland-Altman plots are given for the average GLS, and each apical view in Fig. 5.16-5.19.

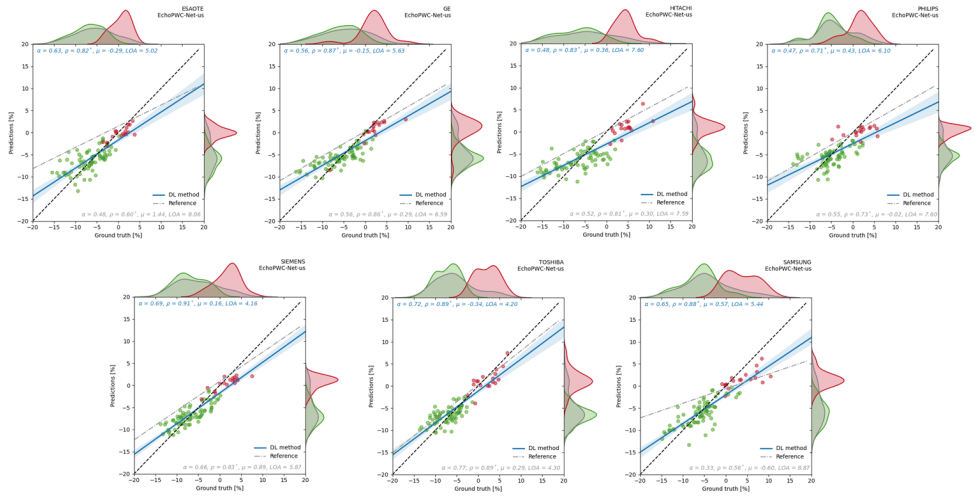


Figure 5.13: Correlation plots between the ground truth regional strain estimation and the DL method on simulated data for all vendors. Green dots represent healthy myocardial segments, while the red sick segments. In the top left corner, the slope α of the regression line, correlation coefficient ρ, bias μ and limits of agreement (LOA) is given. The corresponding reference values from Alessandrini *et al.* [10] are given in the bottom right corner.

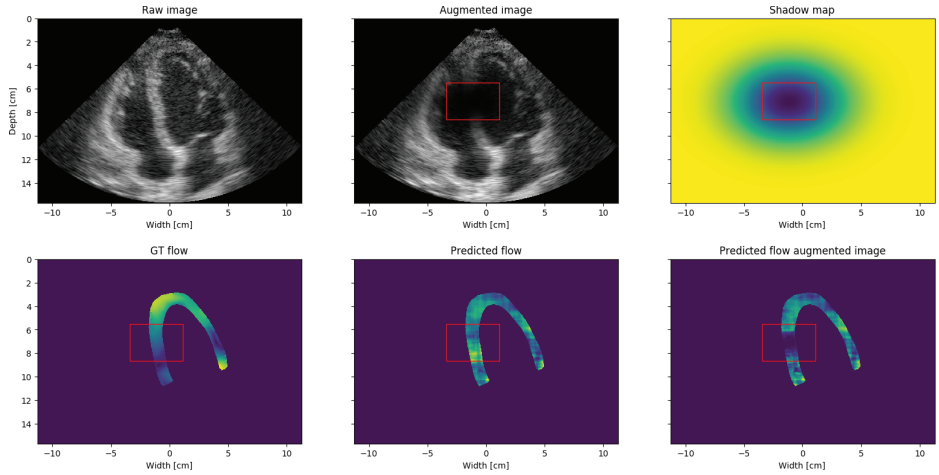


Figure 5.14: Qualitative results from the model adaption study using the Farnebåck method. On the top, the image, corresponding degraded image and shadow map is given. In the bottom, the flow magnitude of the ground truth is visualized, together with the predicted flow with and without the artefact applied.

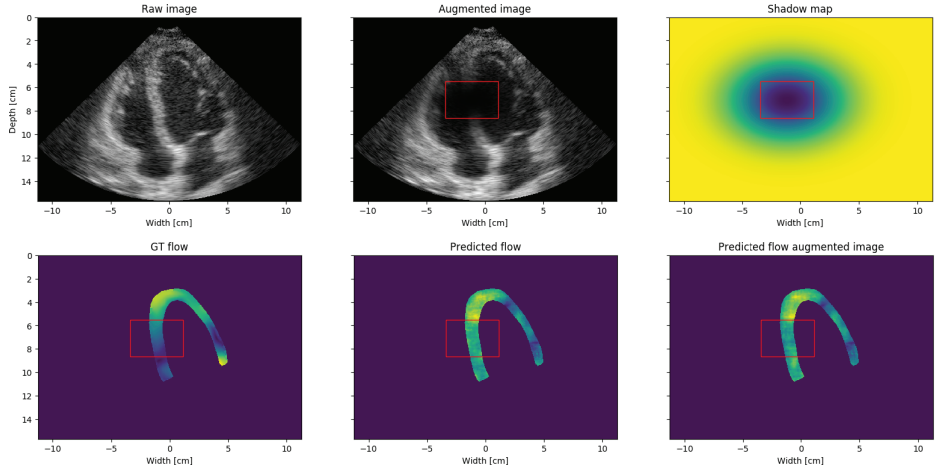


Figure 5.15: Qualitative results from the model adaption study using the PWCNet method trained with US relevant augmentation. On the top, the image, corresponding degraded image and shadow map is given. In the bottom, the flow magnitude of the ground truth is visualized, together with the predicted flow with and without the artefact applied.

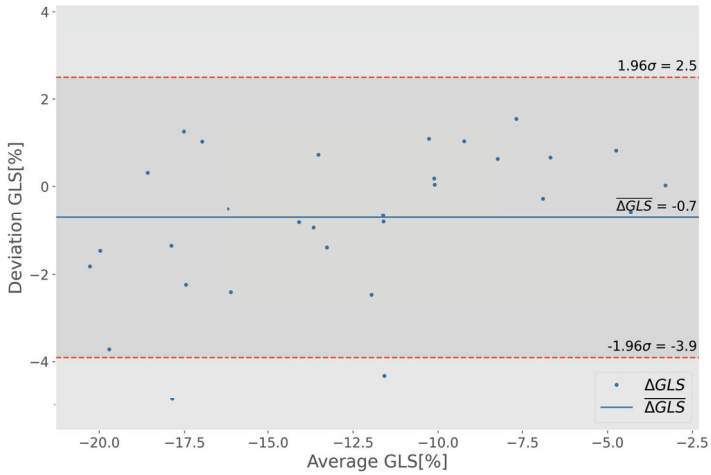


Figure 5.16: Bland-Altman plot of global longitudinal strain (GLS) estimates between commercial system and deep learning based method averaged over the three apical views. Each dot represents one subject.

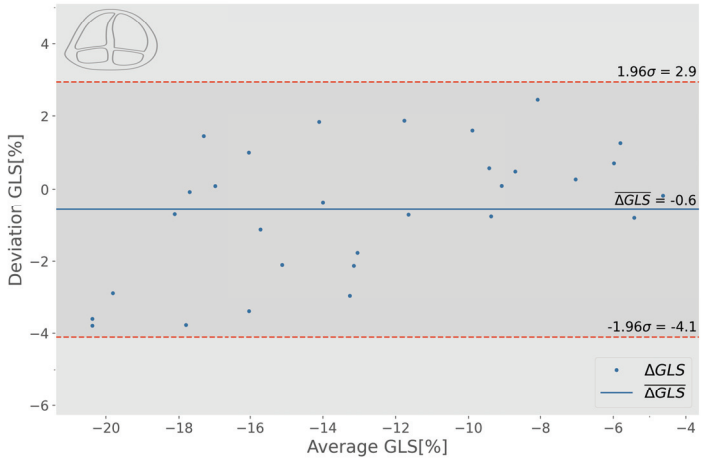


Figure 5.17: Bland-Altman plot of global longitudinal strain (GLS) estimates between commercial system and deep learning based method averaged on the apical four chamber view (4CH). Each dot represents one subject.

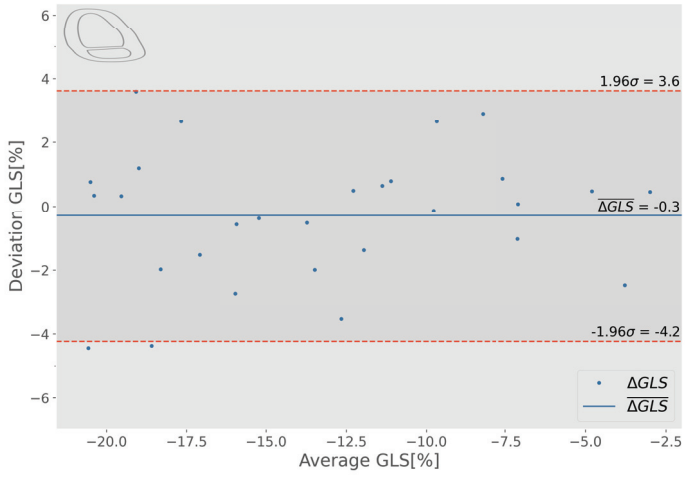


Figure 5.18: Bland-Altman plot of global longitudinal strain (GLS) estimates between commercial system and deep learning based method averaged on the apical two chamber view (2CH). Each dot represents one subject.

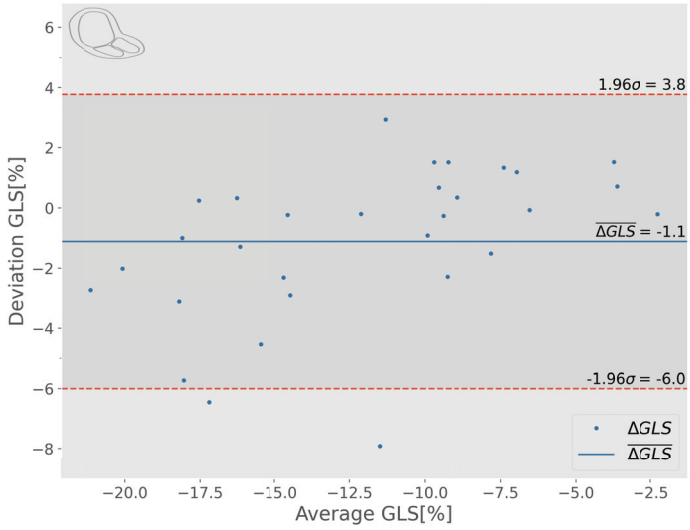


Figure 5.19: Bland-Altman plot of global longitudinal strain (GLS) estimates between commercial system and deep learning based method averaged on the apical long-axis view (APLAX). Each dot represents one subject.

References

[1] H. Geyer, G. Caracciolo, H. Abe, S. Wilansky, S. Carerj, F. Gentile, H.-J. Nesser, B. Khandheria, J. Narula, and P. P. Sengupta, "Assessment of myocardial mechanics using speckle tracking echocardiography: fundamentals and clinical applications," *Journal of the American Society of Echocardiography*, vol. 23, no. 4, pp. 351–369, 2010.

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

[3] B. Heyde, R. Jasaityte, D. Barbosa, V. Robesyn, S. Bouchez, P. Wouters, F. Maes, P. Claus, and J. D'hooge, "Elastic image registration versus speckle tracking for 2-d myocardial motion estimation: A direct comparison in vivo," *IEEE transactions on medical imaging*, vol. 32, no. 2, pp. 449–459, 2012.

[4] M. S. Amzulescu, M. De Craene, H. Langet, A. Pasquet, D. Vancraeynest, A. C. Pouleur, J. L. Vanoverschelde, and B. L. Gerber, "Myocardial strain imaging: review of general principles, validation, and sources of discrepancies," *European Heart Journal - Cardiovascular Imaging*, vol. 20, no. 6, pp. 605–619, 2019.

[5] H. Blessberger and T. Binder, "Two dimensional speckle tracking echocardiography: basic principles," *Heart*, vol. 96, no. 9, pp. 716–722, 2010.

[6] S. Urheim, T. Edvardsen, H. Torp, B. Angelsen, and O. A. Smiseth, "Myocardial strain by doppler echocardiography: validation of a new method to quantify regional myocardial function," *Circulation*, vol. 102, no. 10, pp. 1158–1164, 2000.

[7] R. M. Lang, L. P. Badano, V. Mor-Avi, J. Afilalo, A. Armstrong, L. Ernande, F. A. Flachskampf, E. Foster, S. A. Goldstein, T. Kuznetsova, *et al.*, "Recommendations for cardiac chamber quantification by echocardiography in adults: an update from the american society of echocardiography and the european association of cardiovascular imaging," *European Heart Journal-Cardiovascular Imaging*, vol. 16, no. 3, pp. 233–271, 2015.

[8] K. E. Farsalinos, A. M. Daraban, S. Ünlü, J. D. Thomas, L. P. Badano, and J.-U. Voigt, "Head-to-head comparison of global longitudinal strain measurements among nine different vendors: the eacvi/ase inter-vendor comparison study,"

Journal of the American Society of Echocardiography, vol. 28, no. 10, pp. 1171–1181, 2015.

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

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

[11] A. Dosovitskiy, P. Fischer, E. Ilg, P. Häusser, C. Hazırbaş, V. Golkov, P. v.d. Smagt, D. Cremers, and T. Brox, "Flownet: Learning optical flow with convolutional networks," in *IEEE International Conference on Computer Vision (ICCV)*, 2015.

[12] D. Sun, X. Yang, M.-Y. Liu, and J. Kautz, "PWC-Net: CNNs for optical flow using pyramid, warping, and cost volume," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 8934–8943, 2018.

[13] E. Ilg, N. Mayer, T. Saikia, M. Keuper, A. Dosovitskiy, and T. Brox, "Flownet 2.0: Evolution of optical flow estimation with deep networks," in *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2017.

[14] N. Duchateau, A. P. King, and M. De Craene, "Machine Learning Approaches for Myocardial Motion and Deformation Analysis," 2020.

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

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

[17] E. Evain, K. Faraz, T. Grenier, D. Garcia, M. De Craene, and O. Bernard, "A pilot study on convolutional neural networks for motion estimation from ultrasound images," *IEEE Transactions on Ultrasonics, Ferroelectrics, and Frequency Control*, 2020.

[18] 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, no. 16, pp. 1623–1635, 2018.

[19] E. Smistad, A. Østvik, I. Mjåland Salte, D. Melichova, T. M. Nguyen, H. Brunvand, T. Edvardsen, S. Leclerc, O. Bernard, B. Grenne, and L. Lovstakken, "Real-Time automatic ejection fraction and foreshortening detection using deep learning," *IEEE Transactions on Ultrasonics, Ferroelectrics, and Frequency Control*, 2020.

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

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

[22] 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, no. 2, pp. 374–384, 2019.

[23] A. M. Fiorito, A. Østvik, E. Smistad, S. Leclerc, O. Bernard, and L. Løvstakken, "Detection of Cardiac Events in Echocardiography using 3D Convolutional Recurrent Neural Networks," in *2018 IEEE International Ultrasonics Symposium (IUS)*, 2018.

[24] E. Smistad, I. Salte, A. Østvik, S. Leclerc, O. Bernard, and L. Løvstakken, "Segmentation of apical long axis, four-and two-chamber views using deep neural networks," in *IEEE International Ultrasonics Symposium, IUS*, 2019.

[25] S. Leclerc, E. Smistad, J. Pedrosa, A. Østvik, F. Cervenansky, F. Espinosa, T. Espeland, E. A. R. Berg, P.-M. Jodoin, T. Grenier, C. Lartizien, J. Drhooge, L. Lovstakken, and O. Bernard, "Deep Learning for Segmentation using an Open Large-Scale Dataset in 2D Echocardiography," *IEEE Transactions on Medical Imaging*, 2019.

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

[27] E. Smistad, A. Østvik, B. Haugen, and L. Lovstakken, "2D left ventricle segmentation using deep learning," in *IEEE International Ultrasonics Symposium, IUS*, 2017.

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

[29] R. E. Kalman, "A new approach to linear filtering and prediction problems," *Journal of basic Engineering*, vol. 82, no. 1, pp. 35–45, 1960.

[30] N. Mayer, E. Ilg, P. Häusser, P. Fischer, D. Cremers, A. Dosovitskiy, and T. Brox, "A large dataset to train convolutional networks for disparity, optical flow, and scene flow estimation," in *IEEE International Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016.

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

[32] E. Smistad, K. F. Johansen, D. H. Iversen, and I. Reinertsen, "Highlighting nerves and blood vessels for ultrasound-guided axillary nerve block procedures using neural networks," *Journal of Medical Imaging*, vol. 5, no. 4, p. 044004, 2018.

[33] M. Abadi, P. Barham, J. Chen, Z. Chen, A. Davis, J. Dean, M. Devin, S. Ghemawat, G. Irving, M. Isard, *et al.*, "Tensorflow: A system for large-scale machine learning.," in *OSDI*, vol. 16, pp. 265–283, 2016.

[34] F. Yu and V. Koltun, "Multi-scale context aggregation by dilated convolutions," *arXiv preprint arXiv:1511.07122*, 2015.

[35] S. Zhao, Y. Sheng, Y. Dong, E. I. Chang, Y. Xu, *et al.*, "Maskflownet: Asymmetric feature matching with learnable occlusion mask," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 6278–6287, 2020.

Artificial Intelligence for Automatic Measurement of Left Ventricular Strain in Echocardiography

6

Ivar Mjåland Salte[1,2], Andreas Østvik[3,4], Erik Smistad[3,4], Daniela Melichova[1,2], Thuy Mi Nguyen[1,2], Sigve Karlsen[1], Harald Brunvand[1], Kristina Haugaa[2,5], Thor Edvardsen[2,5], Lasse Lovstakken[3,4] and Bjørnar Grenne[3,4,6]

[1] Dept. of Medicine, Hospital of Southern Norway, Norway
[2] Faculty of Medicine, University of Oslo, Oslo, Norway
[3] Centre for Innovative Ultrasound Solutions, NTNU, Trondheim, Norway
[4] Dept. of Circulation and Medical Imaging, NTNU, Trondheim, Norway
[5] Dept. of Cardiology, Oslo University Hospital, Oslo, Norway
[6] Clinic of Cardiology, St. Olavs hospital, Trondheim, Norway

Background: Global longitudinal strain (GLS) is an important parameter in the evaluation of left ventricular function. However, analyses of GLS are time consuming and demands expertise, and are thus underused in clinical practice.

Objectives: To examine if fully automated measurements of GLS using a novel motion estimation technology based on deep learning and artificial intelligence (AI) are feasible and comparable to a conventional speckle-tracking application.

Methods: 200 patients with a wide range of LV function were included. Three standard apical cine-loops were analyzed using the AI pipeline. The AI method measured GLS and was compared to a commercially available semi-automatic speckle-tracking software (EchoPAC v202, GE Healthcare).

Results: The AI method succeeded to correctly classify the three standard apical views and perform timing of cardiac events in 97% and 96% of the cine-loops, respectively. Furthermore, the method successfully performed automatic segmentation, motion estimates and measurements of GLS in all examinations, across different cardiac pathologies and throughout the spectrum of LV function. GLS was (-12.0±4,1)% for the AI method and

(-13.5±5.3)% for the reference method. Bias was (-1.4±0.3)% (95% limits of agreement 2.3 to -5.1), which is comparable to intervendor studies. The AI method eliminated measurement variability and was fast enough to allow for real-time analysis.

Conclusions: Through the range of LV function this novel AI method succeeds, without any operator input, to automatically identify the three standard apical views, perform timing of cardiac events, trace the myocardium, perform motion estimation and measure GLS. Fully automated measurements based on AI could facilitate the clinical implementation of GLS.

6.1 Introduction

Assessment of left ventricular (LV) function is fundamental for diagnosis, risk stratification and guiding of treatment in patients with cardiac disease. There are multiple echocardiographic parameters available to measure and quantify LV function. Left ventricular ejection fraction (LVEF) is the most widely used method and accepted as a gold standard [1]. However, LVEF has only modest reproducibility and is limited by geometric assumptions that results in poor measurements of LV function in regional pathologies and with concentric remodeling [1, 2]. Moreover, LVEF has limited ability to detect subtle deprivation of LV function [3]. Global longitudinal strain (GLS) by speckle-tracking echocardiography has emerged as a promising parameter to improve assessment of LV function. Multiple studies have shown that GLS has incremental prognostic value compared to LVEF [4], as well as being more sensitive detecting subtle changes in LV function and providing more reproducible measurements [3, 5]. The European association of cardiovascular imaging (EACVI) and American Society of Echocardiography (ASE) now recommend GLS as a supplement to LVEF when evaluating LV function [6]. As a result, vendors have developed software enabling measurement of GLS [7]. Although these methods are semi-automatic, analyses are still time consuming and demands expertise, and thus underused in everyday clinical practice.

Deep learning, the most recent advancement in artificial intelligence (AI), now enables computers to learn from annotated images and perform fully automated image analysis without any operator input [8]. Previous

AI and machine learning techniques required explicitly designed pattern recognition features to be created by the designers of the AI, while the novel deep learning techniques enables the AI to independently learn the patterns and combinations of patterns in the dataset needed to make accurate predictions, thereby allowing for fully automated calculations previously only possible with extensive manual work. This has caused deep learning neural networks to become the most successful and state of the art method of current AI research [9,10]. Deep learning neural networks has successfully been adapted to perform several specific tasks in echocardiographic image analysis that previously would have needed human input, such as view classification [11,12], timing of events [13,14], and image segmentation [15]. Even though AI and deep learning in echocardiography are still in its infancy, there are at present commercially available software solutions that have implemented neural networks for tasks such as view classification and segmentation to provide automatic measurement of GLS. However, the core task of strain imaging, namely motion estimation, is still performed by traditional speckle tracking algorithms. We have recently demonstrated that a deep learning neural network could also be trained to estimate motion in 2D-echocardiography, and that such a network could be implemented in an end-to-end deep learning AI pipeline for automatic measurements of GLS [16]. Compared to traditional speckle tracking, far more sophisticated motion estimation algorithms can be constructed by using deep learning. A deep learning based motion estimation network could learn to integrate information about different moving speckle patterns and global features of an image and independently learn to differentiate artifacts from true motion. Fully automated GLS measurements based on deep learning for motion estimation have the potential to both reduce time spent on manual tracing and improve reproducibility, and due to the processing speed of optimized deep learning algorithms this could eventually enable on-screen measurements in real-time while the operator acquires images. Thus, the field of deep learning represents a paradigm shift in medical imaging and could change how we perform clinical measurements in cardiology.

We hypothesized that a fully automated AI method based on deep learning could, without any operator input, identify and classify the three standard apical views, perform event timing, trace the myocardium, perform motion estimation and calculate GLS, producing comparable results to a

commercially available semi-automated speckle-tracking method. The aim of this study was to test this hypothesis in echocardiographic examinations from patients with a wide range of LV function, different cardiac pathologies, and varying image quality.

6.2 Methods

6.2.1 Study design

A measurement system comparison study (MCS) was performed by analysis of 200 echocardiographic examinations. Each examination represented a test for each method, resulting in two paired GLS measurements for each examination. The first measurement system consisted of a single experienced observer using a commercially available semi-automatic method for GLS measurements. The second measurement system was a novel AI method measuring GLS without any observer input. A single heart cycle was chosen from each view and the exact same recording and cardiac cycle was used for both methods. Analyses were performed without knowledge of clinical data or previous measurement results. To assess if agreement between methods was affected by LV function, subgroup analyses were performed by categorizing the 200 between method differences by LV function measured by LVEF (normal LVEF >50%, mildly reduced LVEF 40-59%, moderately reduced LVEF 30-39% and severely reduced LVEF <30%). Finally, subgroups were evaluated according to image quality (good, fair or poor). The proposed deep learning AI pipeline automatically estimates end diastole and end systole using a deep learning AI timing network. To explore how this automatic event timing affected the GLS measurements, all examinations were analyzed twice by the AI pipeline. First, the deep learning AI pipeline was performed as proposed including automatic event timing using the AI timing network, Secondly, analyses using the same AI pipeline was repeated using the event timing defined by the reference method. Intra- and interobserver reproducibility was assessed in a random subset of 25 patients to illustrate the variability observed when measuring conventional LVEF and GLS by the reference method as compared to the novel AI method. Intraobserver reanalyzes of these examinations were performed by the same observer four weeks after the initial measurements. An experienced second observer at a different

hospital analyzed the same examinations to assess interobserver variability. All reanalyzes were performed using the exact same heart cycle and blinded to previous measurements and clinical data.

6.2.2 Material

To achieve a study population with a wide range of cardiac function and different pathologies, we included five predefined patient groups: 35 patients with Non-ST-elevation myocardial infarction (NSTEMI), 35 patients with ST-elevation myocardial infarction (STEMI), 50 patients with ischemic heart failure, 50 patients with non-ischemic heart failure, and 30 patients admitted for chest pain where neither clinical examination, laboratory tests, ECG, echocardiography or coronary angiography revealed any evidence of cardiac origin. Patients were included regardless of the quality of echocardiographic recordings. Myocardial infarction was defined according to the universal definition [17]. Patients were included consecutively for each group and regardless of image quality. Exclusion criteria were significant valvular disease, atrial fibrillation, age below 18 or inability to give written informed consent. The study was approved by the Regional Committee for Medical and Health Research Ethics and was conducted in compliance with the ethical principles of the Declaration of Helsinki.

6.2.3 Echocardiographic Examinations

The echocardiographic examinations were recorded using GE Vivid E7/E9/E95 ultrasound systems (GE Ultrasound, Horten, Norway). Echocardiographic examinations and measurements were performed in accordance with EACVI guidelines [18]. LV focused echocardiographic recordings were performed in the three standard apical views with simultaneous ECG tracing. Frame rate was 67±9 frames/second. LVEF was measured by the Simpson biplane disc summation method using tracings from apical four-chamber and two-chamber views. Image quality was assessed based on visual assessment of each of the 18 individual myocardial segments of the three apical views. A segment was considered missing if partly outside the image sector, or if the myocardium was indistinguishable from surrounding structures due to artifacts. Good quality examination was defined as no missing segments in neither of the three apical views, fair quality was defined as 1-2 missing

segments, and poor quality was defined as >2 missing segments.

6.2.4 Strain Measurements using the Reference Method

Conventional GLS was measured by speckle-tracking analyses using the semi-automatic analysis method (2DS) implemented in a widely used commercially available software (EchoPAC SWO version 202, GE Ultrasound). End diastole (ED) was defined by the automatic ECG trigger algorithm of the analysis software and only corrected if the automatic QRS detection failed. End-systole was manually defined by the aortic valve closure signal obtained by pulsed-wave Doppler in the left ventricular outflow tract or from continuous-wave Doppler through the aortic valve. The observer manually corrected the region of interest (ROI) by visual assessment of the endocardial and epicardial borders. Spatial and temporal smoothing were kept at default values. Drift compensation was applied as by default. No segments were excluded. A single heart cycle was analyzed for each of the three standard apical views and peak strain was obtained as calculated by the software. GLS was calculated as the average peak strain of the three apical views. The speckle-tracking analyses were performed in accordance with the consensus document of the EACVI/ASE/Industry Task Force to standardize deformation imaging [19].

6.2.5 Strain Measurements using a Deep Learning AI Pipeline

We used an in-house developed AI method based on deep learning consisting of a pipeline of four artificial neural networks (ANN), as illustrated in Fig. 6.1. A detailed technological description of the pipeline has been published separately [16]. The first network was based on the Inception and DenseNet architectures and used for image classification [12]. This network was trained to classify a presented image into one of multiple view classes, including: two-chamber, four-chamber and apical long axis. The second network was based on a recurrent ANN architecture and used for event timing [14]. This network was trained to classify series of presented images into systole or diastole. The third network was based on the U-net architecture and used for image segmentation [20, 21]. This network was trained to classify the image per pixel into four segmentation classes: lumen, left ventricle myocardium, left atrium or other/background. Per

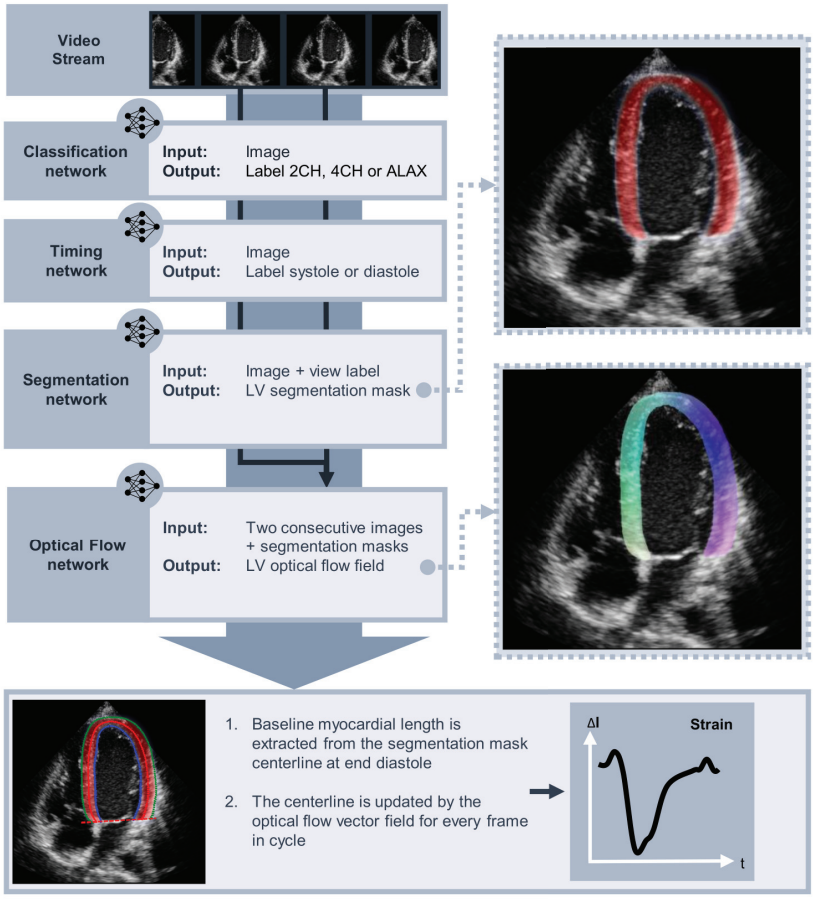


Figure 6.1: The AI pipeline for automatic measurement of global longitudinal strain consisting of four artificial neural networks. Visual feedback of the key steps involved in calculating the GLS is illustrated, such as the segmentation used to initiate the region of interest, an optical flow field visualizing the predicted local velocities, the extracted centerline from the segmentation mask and the points visualizing the motion used to calculate GLS.

pixel predictions were used to extract the position, size and shape of the ventricular myocardium, lumen and left atrium in an image. The myocardial segmentation performed by this network had a previously reported DICE score of 0.79±0.08 and was used to initialize the region of interest. The fourth network was based on a modified PWC-net optimized for estimation of motion in echocardiographic images [22]. This network learned to find patterns in two consecutive images and was trained to output an optical

flow vector field of equal size as the input images which when applied to the patterns in the first image would best reconstruct the location of the same patterns in the second image.

The view classification network was trained on an in-house dataset of out-patient examinations containing 424 hand-labeled echocardiographic recordings. The timing and segmentation networks were trained using the publicly available CAMUS dataset of 500 hand-labelled echocardiographic recordings [23]. The optical flow network was trained using synthetic echocardiography images where the true motion was known [24].

The AI method measured strain frame by frame based on the estimated movement of equally spaced points initialized along the centerline from myocardial segmentation at end diastole. The tracking was performed by updating the position of these points using the displacement fields from the optical flow network. A spline was fitted to the centerline points for each frame. The GLS was calculated for each view as the percentage change in length of the spline from end diastole to its shortest length through cycle. Similar to the reference method, Lagrangian peak strain was calculated for the specified heart cycle in each of the three apical views and GLS was calculated as the average of these three values.

6.2.6 Statistics

Association between methods was estimated by calculating the Pearson's correlation coefficient. The mean absolute difference between the two measurement systems was calculated by the mean value of the absolute difference between all measurement pairs. The agreement of the paired measurements was assessed using a Bland-Altman (B-A) analysis, which is the recommended statistical method in measurement comparison studies [25]. An a priori maximum limit of agreement (LOA) of ±4% was chosen based on known intervendor variability [7]. A sample size of 200 subjects was chosen in accordance with recommendations by JM Bland [26], author of the original B-A paper. This sample size provides sufficient accuracy, with 95% CI about the LOA of approximately ±0.24 standard deviations. Tests for normality were performed using Shapiro-Wilk and Kolmogorov-Smirnov tests. Brown-Forsythe test was used to assess if there was a statistically significant difference in variance between subgroups of measurement pairs when categorized by LVEF, presence of ischemic disease and image quality.

B-A statistical calculations and plot were performed using Python 3.7.4 (Python Software Foundation), where exact 95% CI limits of the LOA was calculated using code based on the method proposed by Shie [27]. All other statistical analyses were performed using open-source statistical Python packages (SciPy 1.5.4 and Statsmodels 0.12.1).

Table 6.1: Study Population. BMI = body mass index; bpm = beats per minute; LVEDV = left ventricular end-diastolic volume; LVEF = left ventricular ejection fraction; LVESV = left ventricular end-systolic volume; mmHg = millimeters of mercury; NSTEMI = Non ST-segment elevation myocardial infarction; SD = standard deviation; SBT = systolic blood pressure; STEMI = ST-segment elevation myocardial infarction.

Parameter	Overall Population (n=200)
Study Cohorts, n (%)	
NSTEMI	35 (17.5%)
STEMI	35 (17.5%)
Ischemic heart failure	50 (25%)
Nonischemic heart failure	50 (25%)
No significant cardiac disease	30 (15%)
Demographics, mean±SD (Range)	
Age, years	61±14 (22 – 91)
Male gender, n (%)	144 (72%)
Clinical Characteristics, mean±SD (Range)	
BMI, kg/m2	27±4 (18-43)
Heart rate, bpm	74±15 (44 – 132)
SBT, mmHg	125±21 (86-197)
Echocardiographic Measurements, mean±SD (Range)	
Echocardiographic LVEF, %	42±13 (7-70)
LVEDV, ml	128±66 (47-372)
LVESV, ml	80±57 (19-306)
LV Function by LVEF Category, n (%)	
Severely reduced: <30 %	29 (14.5%)
Moderately reduced: 30-39 %	60 (30%)
Mildly reduced: 40-49 %	41 (20.5%)
Normal: >50%	70 (35%)
Image Quality, n (%)	
Poor (>2 segment missing)	39 (19.5%)
Fair (1-2 segments missing)	71 (35.5%)
Good (0 segments missing)	90 (45%)

6.3 Results

Patient characteristics are summarized in Table 6.1. The view classification network succeeded to classify the correct view in 97% of the cine-loops (584/600). A confusion matrix summarizing classification results for each view is presented in Fig. 6.2.

		Prediction			
		4Ch	2Ch	Aplax	Other
View	4Ch	196	2	1	1
	2Ch	7	192	0	1
	Aplax	0	2	197	1
	Other	0	0	0	0

Figure 6.2: Confusion matrix presenting view classification results of the 200 echocardiographic examinations included in the study. Every row of the matrix corresponds with the actual view presented to the deep learning algorithm and each row sums up to a total of 200 subjects. The values in each column corresponds to the view prediction output by the deep learning network. Correctly classified views are accentuated with a darker shade of blue. 2Ch = apical two chamber; 4Ch = apical four chamber; Aplax = apical long axis view.

The timing network succeeded in estimation of both end-diastole and end-systole in 98% (593/600) of the cine-loops. Difference in timing of end systole and end diastole between the deep learning AI timing network and the reference method was 1.8±2.7 frames (17±42 ms) and 0.5±2.7 frames, respectively. Detailed results of the timing network for each view are presented in Table 6.2. When running the AI pipeline as proposed, with event timing defined by the deep learning AI network, the mean difference in measured GLS compared to using event timing defined by the reference method was 0.3%±0.3 (p=0.02). A BA-analysis presenting impact of timing method is presented in Fig. 6.3. Twenty-one patients (11%) had failures in either the view classification or timing. Both the reference method and the AI method succeeded in measuring GLS in all included recordings, when correct view and timing were verified.

The proposed method was run on a standard desktop computer with a modern graphics card and used approximately 4 ms per frame for view classification, 16 ms per frame for event timing, 10 ms per frame for myocardial segmentation and 30 ms per frame for motion estimation. Total processing time when running the entire pipeline was 4.3±0.7 seconds per

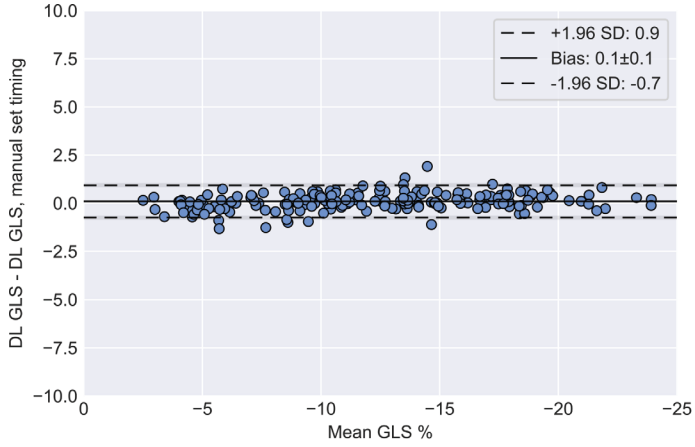


Figure 6.3: Bland Altman plot presenting the impact of cardiac event timing on global longitudinal strain (GLS) measurements by a deep learning (DL) Artificial intelligence (AI) pipeline. GLS measurements using the proposed AI pipeline, which includes a deep learning network for event timing, was compared to repeated measurements using the same AI pipeline with the exception that event timing was defined by the reference method.

Table 6.2: Cardiac event timing results of a deep learning network in 200 echocardiographic examinations, compared to a semi-automatic reference method. Feasibility is defined as the number of exams where the deep learning network was able to detect the cardiac event. Performance in frames and milliseconds were measured by mean difference and mean absolute difference (deep learning AI pipeline – reference method). 2Ch = apical two chamber view; 4Ch = apical four chamber view; Aplax = apical long axis view; SD = standard deviation.

Parameter	End diastole	End systole	All
Feasibility, n (%)			
4Ch	198 (99%)	200 (100%)	198 (99%)
2Ch	200 (100%)	200 (100%)	200 (100%)
Aplax	195 (98%)	200 (100%)	195 (98%)
Mean difference, frames±SD			
4Ch	0.7±2.5	0.75±2.7	0.7±2.6
2Ch	1.8±2.5	0.1±2.5	0.9±2.7
Aplax	2.9±2.7	0.6±2.9	1.7±3.0
All	1.8±2.7	0.5±2.7	1.1±2.8
Mean difference, msec±SD			
4Ch	10±38	13±40	11±39
2Ch	27±38	2±38	15±40
Aplax	43±40	10±43	26±45
All	26±40	8±40	17±42

view and 13.0±2.0 seconds for a full patient analysis including all three apical views.

6.3.1 In Between Methods Agreement

Mean GLS in the entire population was -12.1±5.0% and -13.5±5.3% for the AI method and the conventional method, respectively. The median absolute deviation was 1.4% and mean absolute difference was 1.8±1.5%. As shown in Fig. 6.4, there was a highly significant correlation between the methods (Pearson's coefficient 0.93, p<0.01). The B-A analysis presented in Fig. 6.5 between method differences revealed a bias of -1.4%±0.3 (p<0.01) with estimated LOA of ±3.7%.

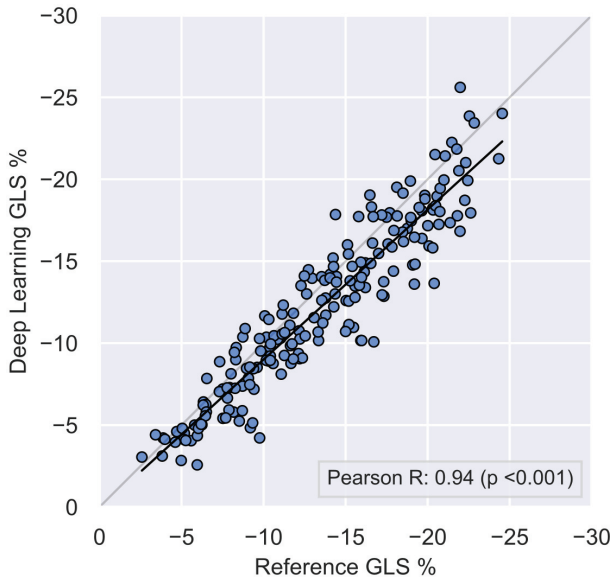


Figure 6.4: 200 echocardiographic examinations measured by both the reference method and the novel AI method. Each marker represents one examination. Solid black line represents the best fit line to the data by linear regression. Solid grey line represents the theoretical perfect correlation.

6.3.2 Agreement Categorized by LVEF and Image Quality

The spread of subjects across different categories of LVEF and image quality is presented in Table 6.1. There was no significant difference in variance

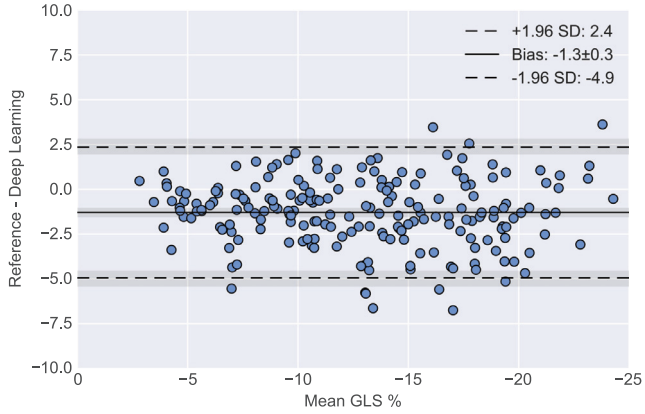


Figure 6.5: Bland Altman plot presenting measurement comparison results using 200 echocardiographic examinations comparing the reference method and the AI method. The figure show the limits of agreement (LOA) calculated assuming normal distribution of the differences between methods.

between measurement pairs from different subgroups categorized by LVEF (p=0.06). Moreover, no significant difference in variance was found between subgroups when categorized by image quality (p=0.58). Figure 6.6 presents the BA-plot where measurement pairs are categorically labeled by LVEF and image quality, illustrating the distribution of these categories throughout the range of GLS measured in the study population.

6.3.3 Intra- and Interobserver Variability and Agreement

Figure 6.7 show B-A-plots illustrating the relative difference in repeated GLS measurements when measured by two observers using the reference method, one observer using the reference method and repeated measurements by the automated deep learning AI pipeline. Importantly, the AI pipeline had no operator input and deep learning algorithms are deterministic in design, thus there were no variability when reanalyzing the exact same images. Assessment of intraobserver variability using the reference method resulted in no significant bias -0.1±0.2 (p=0.55) and LOA ±0.5%. However, there was a small interobserver bias observed when using the reference method 0.5±0.4 (p=0.04) and LOA ±2.1%. Visual representations of measurement agreement using the reference method are presented in B-A plots in Fig. 6.8 and Fig. 6.9.

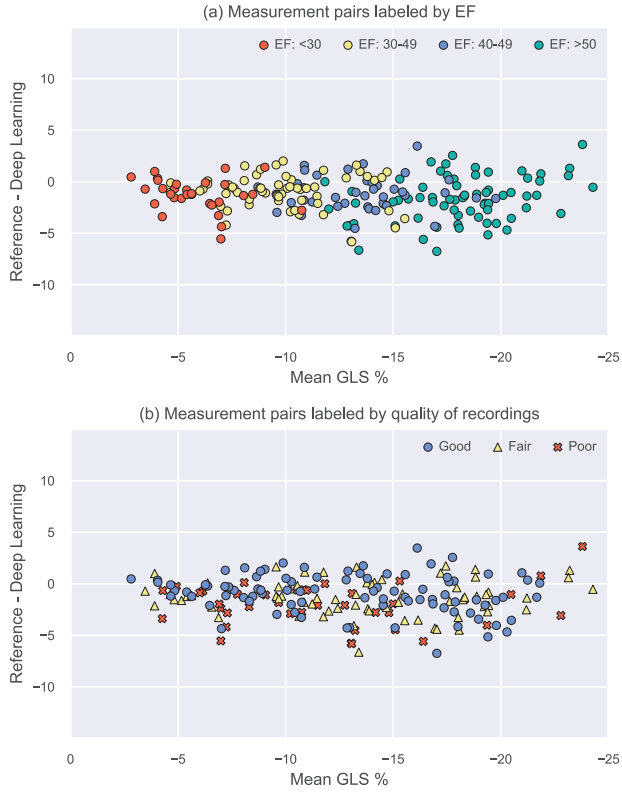


Figure 6.6: Bland Altman plot presenting measurement comparison results using 200 echocardiographic examinations comparing the reference method and the novel method based on artificial intelligence. Measurement pairs labelled by left ventricular ejection fraction (LVEF) (a), and by assessment of image quality (b).

6.4 Discussion

The current study presents, for the first time, the clinical feasibility of an end-to-end AI pipeline which incorporates a deep learning based artificial neural network specifically trained for motion estimation as an alternative to traditional speckle tracking based measures of strain. Through a wide range of LV function and image quality, the AI pipeline succeeded without any human input to correctly classify cardiac views, perform timing of cardiac events, and was able to trace myocardium, estimate motion and ultimately measure GLS.

The main motivation for developing an AI based pipeline for GLS

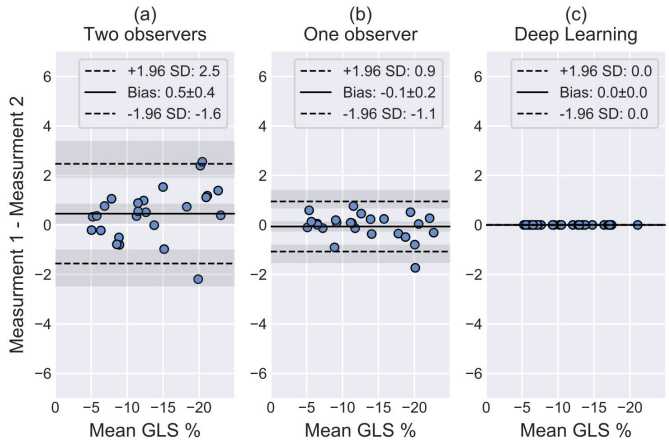


Figure 6.7: Bland-Altman plots of absolute GLS difference in repeated measurements using the exact same video clips in 25 echocardiographic examinations. The figure show agreement between repeated measurements using the reference method by two observers (a) and one observer (b), and for illustrative purposes the expected zero variability of repeated measurements by the fully automated and deterministic AI method (c). The grey shaded areas represent the 95% CI of estimates.

measurements is to provide a more robust and automatized method, with the potential to provide fully automatic real-time GLS measurements while performing the image acquisition, and with improved tracking accuracy and reduced measurement variability. The currently most widely used semi-automatic speckle-tracking methods need several steps of operator input, and time spent performing a single GLS analysis is reported to range from 5-10 minutes [28,29]. In contrast, all steps in the AI pipeline were performed in less than 15 seconds. The novel deep learning AI pipeline eliminates the need for time-consuming manual input, which make it effortless to acquire average measurements from multiple cardiac cycles as recommended in guidelines. Moreover, deep learning algorithms are deterministic. This means that the same input images always gives identical output, without variability, as seen in Fig. 6.7.

It is important to emphasize that removing the interpretation variability completely by a deterministic deep learning algorithm does make repeated measurements more reproducible but does not necessarily make the measurements more accurate. A measurement error made by a deep-

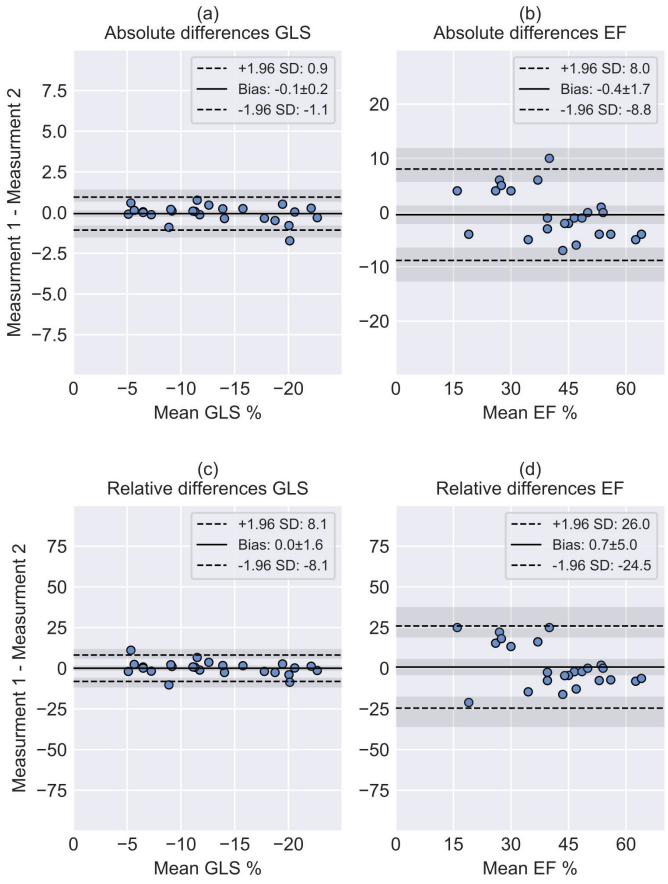


Figure 6.8: Intraobserver study using 25 echocardiographic examinations randomly picked from the total of 200 participants in study. Blinded reanalyzes were performed by the same observer using the exact same examination and heart cycle with 4 weeks between measurements. Bland Altman plot of absolute (a, b) and relative (c, d) differences of global longitudinal strain (GLS) and left ventricular ejection fraction (LVEF). Dotted line represents limits of agreement and the grey shaded area represents the 95% CI.

learning method will be reproduced every time the method is reapplied to analyze the exact same image. Poor image quality in echocardiography is a common problem and in the present study a total of 19% of subjects had more than 2 of 18 segments missing. The high percentage of examinations with suboptimal image quality could explain that 11% of subjects had failures of either the view classification network or the timing network. Suboptimal image quality is an unavoidable factor that limits the achievable

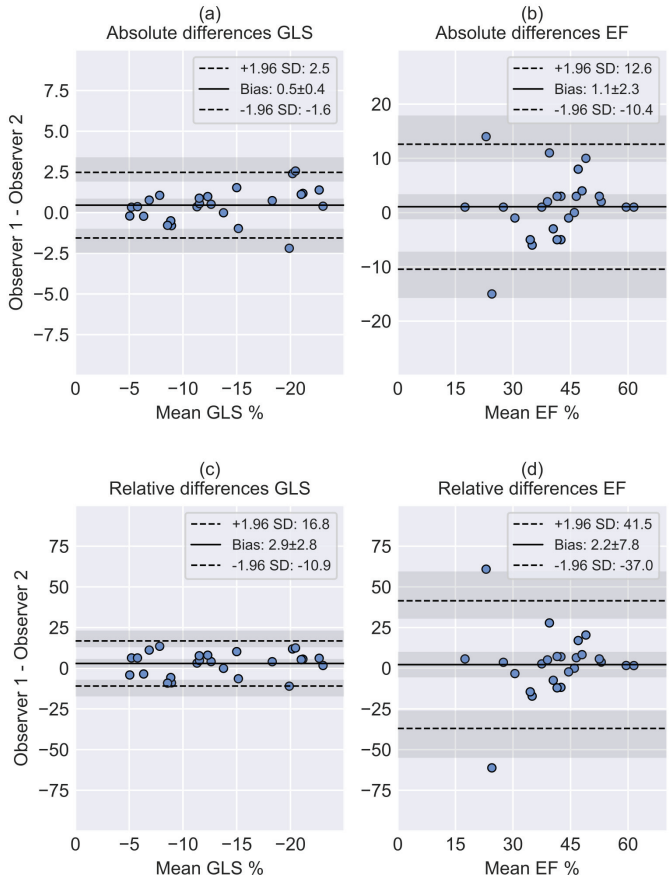


Figure 6.9: Interobserver study using the same 25 echocardiographic examinations as included in the intraobserver study. Blinded reanalyzes were performed by an external observer at another hospital using the exact same examination and heart cycle and compared to the first measurements in the intraobserver study. Bland-Altman plot of absolute (a, b) and relative (c, d) differences of global longitudinal strain (GLS) and left ventricular ejection fraction (LVEF). Dotted line represents limits of agreement and the grey shaded area represents the 95% CI.

accuracy of both manual and automatic measurements. Thus, measurement error or failure of deep learning algorithms is inevitable, even if deep learning algorithms were to outperform humans in terms of precision of measurements. This underlines the importance of not choosing a fully "black-box" AI method, such as directly predicting LVEF or GLS from images equivalent to "eyeballing" as proposed by some authors [30, 31]. An AI method should be designed to give visual feedback to the observer.

Misclassification of view or timing could easily be corrected by the observer. Motion estimation involves complex calculations that are not directly available for the user to inspect, both with the AI method and the semi-automatic reference method. However, an observer could visually inspect if tracking and motion estimates seems reasonable if provided a visual feedback. The method presented in this study was able to give visual feedback for each frame of the left ventricular segmentation, the motion estimation flow field, and the movement of the points used to calculate GLS (Central illustration).

Compared to a previously conducted intervendor comparison study by the EASCVI/ASE/Industry Task force to standardize deformation imaging [7], the two methods in the present study showed excellent correlation and high level of agreement. When assessing agreement, no significant difference in variance was found when measurement differences were categorized according to LV function measured by LVEF and by degree of image quality, suggesting that myocardial dysfunction and image quality had limited effect on agreement.

In the landmark intervendor study by Farsalinos *et al*. [7], two vendor independent software packages were compared to the same reference software as in our study. The bias reported was -0.7% and 2.5% strain units and LOA ±3.4% and ±3.8%, respectively. A study by Nagata *et al*. compared two different vendor independent software packages using the same reference vendor as in our study. They reported bias of -2.1% and LOA of ±4.1% [32]. Anwar *et al*. also compared a vendor independent software package to the same reference software as in our study, and found a bias of -2.9% strain units and LOA ±5.5% [33]. Thus, the bias of -1.4% and LOA of ±3.7% observed in our study are well within the overall range of the bias and LOA previously reported in intervendor agreement studies. This provides reasonable evidence to support that GLS measurements by the present deep learning AI pipeline are comparable to other clinically available semi-automatic methods.

To the best of our knowledge, there is currently no clinically available software application for fully automated GLS measurements in 2D echocardiography that have implemented a deep learning neural network specifically to estimate motion and produce a motion flow field as an alternative to traditional speckle tracking strain algorithms. Except for our

technical manuscript describing the present AI pipeline [16, 34], there are no published journal papers presenting a deep learning neural network for local motion estimation in 2D echocardiography that could produce flow field motion estimates of the entire myocardium. We are only aware of one other journal paper that presents an in detail description of a fully automated AI method for GLS measurements [35]. However, although the authors used deep learning to automatically initialize a ROI, conventional optical flow and not deep learning was used for motion estimation and calculation of GLS. Thus, they did not gain the full potential of deep learning to improve measurements of GLS. Direct comparison with our work is not possible as they do not present the values of LOA and use at least one other vendor as reference. They found median absolute deviation in GLS measurements of 1.4% in a population of 419 examinations and 1.6% in another population of 110 examinations. These findings are in line with our study where the median absolute deviation was 1.2%. Their method resulted in a GLS processing time of 1-4 minutes per view depending on number of frames and image size, while the pipeline in our study used less than 5 seconds per view. In addition, the individual networks used in the pipeline succeeded to process frames within milliseconds. Thus, if these deep learning methods are implemented into ultrasound machines, the individual steps of the AI pipeline could be computed during acquisition of images, enabling rapid bedside analysis, and even real-time measurements on the ultrasound scanner.

6.5 Study limitations

The study has some limitations. We only compared measurements against one reference method. There is no gold standard for GLS measurements and intervendor variability is a known problem. Moreover, the tracking software used by the vendors are not open source. Thus, we cannot conclude whether one measurement system in this study is more accurate than the other. We could only conclude that the measurement variability between these two measurement systems is within the range of previous intervendor studies. The total test-retest reliability of an echocardiographic measurement depends on multiple factors related to both image acquisition and observer interpretation. A recent study concluded that acquisition and

reader influenced the variability of both GLS and LVEF measurements to a similar extent [36]. This suggests that automation of measurements could substantially reduce the total variability in a test-retest setting by removing the individual observer interpretation. We focused on image interpretation and the two measurement systems analyzed the exact same images from one predefined cardiac cycle. Hence the present study was not designed to determine the effect of image acquisition on measurement reproducibility. Another limitation in the present study is that all examinations used for testing the deep learning algorithms were acquired using ultrasound machines from the same vendor. Consequently, we cannot conclude whether the AI method performs equally on images from different vendors. Another topic for further studies is whether deep learning based strain estimation is more accurate and robust in terms of capturing subtle differences in strain, or when exposed to image artifacts compared to currently used speckle-tracking methods.

Further research is needed to address the mentioned limitations before the deep learning measurements could be routinely used in a clinical setting. However, we find the present results promising both in terms of feasibility and agreement with the reference method.

6.6 Conclusion

Fully automated measurements of GLS using a novel deep learning AI based technology for motion estimation are feasible, fast and yields results comparable to the most widely used semi-automatic software. Deep learning networks remove the need for manual tracing and could both increase efficiency and improve reproducibility. The system can potentially be implemented in ultrasound scanners and allow for real-time GLS calculations. Fully automated measurements based on AI could be an important step to further facilitate the implementation of GLS in clinical practice.

References

[1] L. G. Klaeboe and T. Edvardsen, "Echocardiographic assessment of left ventricular systolic function," *Journal of Echocardiography*, vol. 17, pp. 10–16, mar 2019.

[2] T. H. Marwick, "Ejection Fraction Pros and Cons: JACC State-of-the-Art Review," *Journal of the American College of Cardiology*, vol. 72, pp. 2360–2379, nov 2018.

[3] B. Sjøli, S. Ørn, B. Grenne, T. Vartdal, O. A. Smiseth, T. Edvardsen, and H. Brunvand, "Comparison of Left Ventricular Ejection Fraction and Left Ventricular Global Strain as Determinants of Infarct Size in Patients with Acute Myocardial Infarction," *Journal of the American Society of Echocardiography*, vol. 22, pp. 1232–1238, nov 2009.

[4] K. Kalam, P. Otahal, and T. H. Marwick, "Prognostic implications of global LV dysfunction: A systematic review and meta-analysis of global longitudinal strain and ejection fraction," *Heart*, vol. 100, pp. 1673–1680, nov 2014.

[5] T. Negishi, K. Negishi, P. Thavendiranathan, G. Y. Cho, B. A. Popescu, D. Vinereanu, K. Kurosawa, M. Penicka, T. H. Marwick, S. Aakhus, M. Bansal, A. Calin, J. Čelutkienė, N. Fukuda, K. Hristova, M. Izumo, A. La Gerche, J. Lemieux, D. Mihalcea, P. Mottram, R. Morimoto Ichikawa, M. Nolan, T. Ondrus, S. Seldrum, M. Shirazi, E. Shkolnik, B. Thampinathan, L. Thomas, H. Yamada, and S. Yuda, "Effect of Experience and Training on the Concordance and Precision of Strain Measurements," *JACC: Cardiovascular Imaging*, vol. 10, pp. 518–522, may 2017.

[6] R. M. Lang, L. P. Badano, V. Mor-Avi, J. Afilalo, A. Armstrong, L. Ernande, F. A. Flachskampf, E. Foster, S. A. Goldstein, T. Kuznetsova, *et al.*, "Recommendations for cardiac chamber quantification by echocardiography in adults: an update from the american society of echocardiography and the european association of cardiovascular imaging," *European Heart Journal-Cardiovascular Imaging*, vol. 16, no. 3, pp. 233–271, 2015.

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

[8] G. Litjens, F. Ciompi, J. M. Wolterink, B. D. de Vos, T. Leiner, J. Teuwen, and I. Išgum, "State-of-the-Art Deep Learning in Cardiovascular Image Analysis," *JACC: Cardiovascular Imaging*, vol. 12, pp. 1549–1565, aug 2019.

[9] Y. Lecun, Y. Bengio, and G. Hinton, "Deep learning," vol. 521, pp. 436–444, may 2015.

[10] T. J. Sejnowski, "The unreasonable effectiveness of deep learning in artificial intelligence," *Proceedings of the National Academy of Sciences of the United States of America*, vol. 117, pp. 30033–30038, dec 2020.

[11] A. Madani, R. Arnaout, M. Mofrad, and R. Arnaout, "Fast and accurate classification of echocardiograms using deep learning," p. 8, jun 2018.

[12] 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, no. 2, pp. 374–384, 2019.

[13] F. T. Dezaki, Z. Liao, C. Luong, H. Girgis, N. Dhungel, A. H. Abdi, D. Behnami, K. Gin, R. Rohling, P. Abolmaesumi, and T. Tsang, "Cardiac Phase Detection in Echocardiograms with Densely Gated Recurrent Neural Networks and Global Extrema Loss," *IEEE Transactions on Medical Imaging*, vol. 38, pp. 1821–1832, aug 2019.

[14] A. M. Fiorito, A. Østvik, E. Smistad, S. Leclerc, O. Bernard, and L. Løvstakken, "Detection of Cardiac Events in Echocardiography using 3D Convolutional Recurrent Neural Networks," in *2018 IEEE International Ultrasonics Symposium (IUS)*, 2018.

[15] C. Chen, C. Qin, H. Qiu, G. Tarroni, J. Duan, W. Bai, and D. Rueckert, "Deep learning for cardiac image segmentation: A review," *Frontiers in Cardiovascular Medicine*, vol. 7, p. 25, 2020.

[16] 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*, pp. 1–12, 2021.

[17] K. Thygesen, J. S. Alpert, A. S. Jaffe, M. L. Simoons, B. R. Chaitman, H. D. White, H. A. Katus, F. S. Apple, B. Lindahl, D. A. Morrow, P. M. Clemmensen, P. Johanson, H. Hod, R. Underwood, J. J. Bax, R. O. Bonow, F. Pinto, R. J. Gibbons, K. A. Fox, D. Atar, L. K. Newby, M. Galvani, C. W. Hamm, B. F. Uretsky, P. G. Steg, W. Wijns, J. P. Bassand, P. Menasché, J. Ravkilde, E. M. Ohman, E. M. Antman, L. C. Wallentin, P. W. Armstrong, M. L. Simoon, J. L. Januzzi, M. S. Nieminen, M. Gheorghiade, G. Filippatos, R. V. Luepker, S. P. Fortmann, W. D. Rosamond, D. Levy, D. Wood, S. C. Smith, D. Hu, J. L. Lopez-Sendon, R. M. Robertson, D. Weaver, M. Tendera, A. A. Bove, A. N. Parkhomenko, E. J. Vasilieva, S. Mendis, H. Baumgartner, C. Ceconi, V. Dean, C. Deaton, R. Fagard, C. Funck-Brentano, D. Hasdai, A. Hoes, P. Kirchhof, J. Knuuti, P. Kolh, T. McDonagh, C. Moulin, B. A. Popescu,

Ž. Reiner, U. Sechtem, P. A. Sirnes, A. Torbicki, A. Vahanian, S. Windecker, J. Morais, C. Aguiar, W. Almahmeed, D. O. Arnar, F. Barili, K. D. Bloch, A. F. Bolger, H. E. Bøtker, B. Bozkurt, R. Bugiardini, C. Cannon, J. De Lemos, F. R. Eberli, E. Escobar, M. Hlatky, S. James, K. B. Kern, D. J. Moliterno, C. Mueller, A. N. Neskovic, B. M. Pieske, S. P. Schulman, R. F. Storey, K. A. Taubert, P. Vranckx, and D. R. Wagner, "Third universal definition of myocardial infarction," *Circulation*, vol. 126, pp. 2020–2035, oct 2012.

[18] 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 Imaging," *European Heart Journal - Cardiovascular Imaging*, vol. 18, pp. 1301–1310, dec 2017.

[19] J. U. Voigt, G. Pedrizzetti, P. Lysyansky, T. H. Marwick, H. Houle, R. Baumann, S. Pedri, Y. Ito, Y. Abe, S. Metz, J. H. Song, J. Hamilton, P. P. Sengupta, T. J. Kolias, J. D'Hooge, G. P. Aurigemma, J. D. Thomas, and L. P. Badano, "Definitions for a common standard for 2D speckle tracking echocardiography: consensus document of the EACVI/ASE/Industry Task Force to standardize deformation imaging," *European heart journal cardiovascular Imaging*, vol. 16, no. 1, pp. 1–11, 2015.

[20] E. Smistad, A. Østvik, B. Haugen, and L. Lovstakken, "2D left ventricle segmentation using deep learning," in *IEEE International Ultrasonics Symposium, IUS*, 2017.

[21] E. Smistad, I. Salte, A. Østvik, S. Leclerc, O. Bernard, and L. Løvstakken, "Segmentation of apical long axis, four-and two-chamber views using deep neural networks," in *IEEE International Ultrasonics Symposium, IUS*, 2019.

[22] D. Sun, X. Yang, M.-Y. Liu, and J. Kautz, "PWC-Net: CNNs for optical flow using pyramid, warping, and cost volume," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 8934–8943, 2018.

[23] S. Leclerc, E. Smistad, J. Pedrosa, A. Østvik, F. Cervenansky, F. Espinosa, T. Espeland, E. A. R. Berg, P.-M. Jodoin, T. Grenier, C. Lartizien, J. Drhooge, L. Lovstakken, and O. Bernard, "Deep Learning for Segmentation using an Open Large-Scale Dataset in 2D Echocardiography," *IEEE Transactions on Medical Imaging*, 2019.

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

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

[26] "How can i decide the sample size for a study of agreement between two methods of measurement?." https://www-users.york.ac.uk/~mb55/meas/sizemeth.htm. Accessed: 2019-08-01.

[27] G. Shieh, "The appropriateness of Bland-Altman's approximate confidence intervals for limits of agreement," *BMC Medical Research Methodology*, vol. 18, p. 45, may 2018.

[28] P. Barbier, O. Mirea, C. Cefalù, A. Maltagliati, G. Savioli, and M. Guglielmo, "Reliability and feasibility of longitudinal AFI global and segmental strain compared with 2D left ventricular volumes and ejection fraction: intra- and inter-operator, test–retest, and inter-cycle reproducibility," *European Heart Journal - Cardiovascular Imaging*, vol. 16, pp. 642–652, jun 2015.

[29] A. Manovel, D. Dawson, B. Smith, and P. Nihoyannopoulos, "Assessment of left ventricular function by different speckle-tracking software," *European Journal of Echocardiography*, vol. 11, pp. 417–421, jun 2010.

[30] F. M. Asch, N. Poilvert, T. Abraham, M. Jankowski, J. Cleve, M. Adams, N. Romano, H. Hong, V. Mor-Avi, R. P. Martin, and R. M. Lang, "Automated Echocardiographic Quantification of Left Ventricular Ejection Fraction Without Volume Measurements Using a Machine Learning Algorithm Mimicking a Human Expert," *Circulation: Cardiovascular Imaging*, vol. 12, p. 9303, sep 2019.

[31] 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, pp. 1–10, dec 2020.

[32] Y. Nagata, M. Takeuchi, K. Mizukoshi, V. C. C. Wu, F. C. Lin, K. Negishi, S. Nakatani, and Y. Otsuji, "Intervendor variability of two-dimensional strain using vendor-specific and vendor-independent software," *Journal of the American Society of Echocardiography*, vol. 28, pp. 630–641, jun 2015.

[33] S. Anwar, K. Negishi, A. Borowszki, P. Gladding, Z. B. Popović, F. Erenberg, and J. D. Thomas, "Comparison of two-dimensional strain analysis using vendor-independent and vendor-specific software in adult and pediatric patients," *JRSM Cardiovascular Disease*, vol. 6, p. 204800401771286, jan 2017.

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

[35] 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, no. 16, pp. 1623–1635, 2018.

[36] 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, pp. 355–365, apr 2019.

ISBN 978-82-326-5899-2 (printed ver.)
ISBN 978-82-326-5982-1 (electronic ver.)
ISSN 1503-8181 (printed ver.)
ISSN 2703-8084 (online ver.)

NTNU
Norwegian University of
Science and Technology