UNIVERSITY
OF OSLO

Cristiana Ferreira Tiago

Deep Generative Models Applied to 2D and 3D Echocardiography

Image Generation and Analysis

Thesis submitted for the degree of Philosophiae Doctor

Department of Informatics
Faculty of Mathematics and Natural Sciences

GE Vingmed Ultrasound AS

2023

Series of dissertations submitted to the
Faculty of Mathematics and Natural Sciences, University of Oslo
No. 2641

ISSN 1501-7710

Cover: UiO.
Print production: Graphic center, University of Oslo.

If you can't explain it simply, you don't understand it well enough.
- Albert Einstein

Preface

This thesis is submitted in partial fulfillment of the requirements for the degree of *Philosophiae Doctor* at the Department of Informatics, Faculty of Mathematics and Natural Sciences, University of Oslo. The research presented here was conducted at the University of Oslo and at GE Vingmed Ultrasound in the scope of the MARie Curie Intelligent UltraSound european innovative training network, under the supervision of Doctor Kristin McLeod and co-supervision of Doctor Jurica Sprem, Doctor Sten Roar Snare, and Doctor Eigil Samset between March 2020 and May 2023.

This work has received funding from the European Union's Horizon 2020 research and innovation programme under the Marie – Skłodowska – Curie grant agreement No 860745.

Acknowledgements

I have a few individuals to thank to for all the help and support, both academic and social, during these three years. I would like to express my sincere gratitude to the following people without whom this thesis would not have been possible.

First and foremost, I would like to thank my understanding and knowledgeable thesis supervisors: Krissy, Jurica, Sten Roar and Eigil. Their expertise, feedback, and encouragement have been invaluable to me, and I am truly grateful for their mentorship. Not only they encouraged me to follow my research ideas but also kept me down to earth and cheered my spirit when I needed the most. I would also like to express my gratitude towards Svein Arne, who always gave me his honest feedback, support and guidance throughout this entire process. A special thanks goes to Krissy for bringing me into this adventure during a world pandemic and keeping up confidence levels high despite all my kamikaze ideas. Also a big shout-out to Jurica, who became a great friend after this time.

I would also like to thank my colleagues and friends at GE Vingmed. Müjde, Nuno, Jurica, Jan, Benjamin, Pravda and everyone else in the Oslo office. Thanks for the laughs and this good run! I also want to dedicate a word to Andy which was immensely helpful for me since the beginning of my PhD, when we were the only 2 people in the office everyday during the pandemic. And of course, a thank you word to the colleagues that the University of Oslo gave me, particularly Børge, Tollef and Sarina. Their knowledge and dedication to research was of great help and inspiration to me but, most importantly, we got to share our experiences as PhD students (with all the good and terrible things this brings!).

I am also grateful for the friends I have made by jumping into the MARCIUS project. Thank you to Müjde, Nitin, Paulo, Ahmed and Claudia for all the MARCIUS team buildings we got to have, all the trips around Europe, and

for the crazy seasonal schools! All of those left me with amazing memories. I would also like to thank the rest of the MARCIUS people who made this project possible and for putting up together such a nice group of humans, especially Krissy, Jurica, Joost, Jan, Marcus, and of course Lieven, for describing our group as "the tequila drinkers". I will take every moment with me.

I am deeply indebted to my family for their love, support, and encouragement. When I moved away from them, when the whole world stopped and all countries closed, they have always been there for me, through thick and thin, and I could not have completed this thesis without their understanding and patience. My parents, Luísa and Eduardo, are the most resilient couple I have met, since they handled both of their "kids" leaving the house and the country at the same time, in the most spectacular way. They raised me to be kind and honest, sacrificing a lot, and still were always supportive even when nothing looked promising or I had no hope in my future. I want to leave a word for my grandparents, uncles, aunts, godfather and godmother for also giving me their love during these three years away from them. I cannot forget my cousins, from the youngest to the eldest, and of course to the one looking down for me.

If there is someone I cannot put into enough words how much he means to me is my brother, Ivo. I have to thank him for all the love, support, messages, phone calls, trips, relax moments and damn spectacular memories and stories we have built together.

Finally, I would like to thank all the new friends I made when I moved to Oslo. I came without knowing anyone but now my life is way more full after meeting all of them. Nuno, Nando, Maria, Målin, and everyone else who had an impact in my life in Norway. Of course a very specially thanks goes to Müjde. Thank you for the motivational talks, for making Norway more bearable, for your friendship, and the amazing blue mosque moments! Solid. There is no way you are leaving my life now and road to Israel! A final thank you goes to all my long time friends who influenced me throughout this experience in one way or another. I am grateful for their friendship.

Cristiana Ferreira Tiago
Oslo, July 2023

Scientific Summary

Echocardiography, the ultrasound imaging of the heart, is the go-to medical imaging modality when it comes to visualizing the heart. Among all its advantages and disadvantages, the high temporal resolution of echocardiography undoubtedly emerges as a valuable strength for visualizing and evaluating the dynamic behavior of a beating organ as the heart is.

In a world where cardiovascular diseases are the main cause of death, it is of utmost importance to develop automatic algorithms and methods which help to improve the quality of treatment offered to patients. The growing and aging population poses a pressure problem on healthcare systems, therefore automatic approaches to analyze medical imaging scans, saving clinicians' time for more severe cases, is a need.

Relevant work has been made in the field, from using mathematical and physical models to simulate the heart mechanics and the ultrasound image acquisition processes, to developing automatic and, each time more accurate, Deep Learning models. Deep Learning algorithms have shown high accuracy not only when helping the clinician to deliver a final diagnosis to a patient, but also when using them for simulation tasks. These kinds of models revealed a great ability to generate synthetic echocardiography images, with high quality and variability, which are helpful for researchers and healthcare personnel. These images are not only a valuable educational resource to train clinicians, but also a useful asset to train and improve new Deep Learning algorithms, advancing the state of the art of the echocardiography imaging field.

The development of deep generative models, such as Generative Adversarial Networks (GANs) and Diffusion Models, has revolutionized the field of medical image synthesis by enabling the generation of realistic and diverse 2D and 3D images. These models have overcome challenges associated with limited data availability, privacy restrictions, and the complexity of capturing the full range of anatomical and physiological variations in echocardiography.

GANs consist of a generator and a discriminator network engaged in an adversarial training process. By training the generator network on a large dataset of real echocardiography images, GANs can generate synthetic images that closely resemble real patient data. This has facilitated tasks such as data augmentation, anomaly detection, and echocardiography image synthesis for clinical use.

Diffusion models, on the other hand, have emerged as powerful tools for generating high-quality synthetic images by simulating a diffusion process. These models iteratively refine an initial noise source to generate images that capture the target image distribution. Diffusion models have shown promising results, better than GANs, in generating realistic and diverse images. Due to its novelty, these

type of deep generative models has not so far been applied to echocardiography images.

Using these synthetically generated images, both 2D and 3D, to develop consequent downstream Deep Learning models to perform varying tasks, for example segmentation, showed promising results. The final volume of the heart can be predicted with higher accuracy when using the synthetic images together with real ones, to train these models.

The application of deep generative models in echocardiography has several important implications. Firstly, it offers a solution to the scarcity of labeled data, which is often a challenge in medical imaging. By generating synthetic images, these models provide a means to augment training data and improve the performance of automated algorithms.

Secondly, deep generative models facilitate the exploration of rare or complex cardiac conditions that may not be adequately represented in the available datasets, due to its smaller prevalence in society. By generating diverse images that capture different pathologies and anatomical variations, these models enhance the understanding and analysis of complex cardiac cases.

Furthermore, the generation of realistic synthetic echocardiography images enables the development and evaluation of novel image processing techniques and algorithms. Researchers can test and optimize image analysis algorithms, evaluate the performance of novel image reconstruction methods, and investigate the effects of different imaging parameters on diagnostic accuracy.

While deep generative models have demonstrated great potential, there are challenges and limitations that need to be addressed. These include the need for large and diverse training datasets, ensuring generalizability across different patient populations and imaging modalities, and the validation and interpretation of synthetic images in a clinical setting.

In conclusion, the development of deep generative models, including GANs and diffusion models, has significantly advanced the field of echocardiography by enabling the generation of realistic and diverse 2D and 3D images. These models have the potential to enhance data availability, improve automated analysis, and facilitate research and development in echocardiography.

In this thesis, both GANs and diffusion models have made significant contributions to 2D and 3D echocardiography image synthesis. They have helped overcome data limitations and have provided a means to generate large datasets with known ground truth information. This has been particularly valuable for training and evaluating Deep Learning algorithms used in echocardiography analysis and diagnosis. Furthermore, the clinical interpretation of the synthetic images revealed the utility and valuable information comprised in them.

Sammendrag

Ekkokardiografi, ultralydavbildning av hjertet, er den foretrukne metoden for medisinsk avbildning av hjertet. Med tanke på alle dens fordeler og ulemper, kommer den høye tidsoppløsningen til ekkokardiografi utvilsomt frem som en verdifull styrke for å visualisere og evaluere den dynamiske atferden til et hjerte som slår.

I en verden der hjerte- og karsykdommer er de dominerende dødsårsakene, er det av aller største betydning å utvikle automatiske algoritmer og metoder som bidrar til å forbedre kvaliteten på behandlingen som tilbys pasientene. Den økende og aldrende befolkningen legger et pressproblem på helsevesenet, derfor er det behov for automatiske metoder for analyse av medisinske avbildninger så klinikerne kan bruke mer på alvorlige tilfeller.

Det har vært gjort relevant arbeid på dette fagfeltet, fra bruk av matematiske og fysiske modeller for å simulere hjertemekanismen og avbildningsprosessene for ultralyd, til utvikling av automatiske og stadig mer nøyaktige dyplæringsmodeller. Dyplæringsalgoritmer har vist høy nøyaktighet ikke bare når de hjelper klinikeren med å stille en endelig diagnose for en pasient, men også når de brukes til simulering. Denne typen modeller har vist seg å være veldig egnet til å generere syntetiske ekkokardiografiske bilder med høy kvalitet og variasjon, som er nyttige for forskere og helsepersonell. Disse bildene er en god pedagogisk ressurs for å trene klinikere, men de er også verdifulle for å trene og forbedre nye dyplæringsalgoritmer og for å fremme kunnskapsnivået innen ekkokardiografisk avbildning.

Utviklingen av dype generative modeller, som Generative Adversarial Networks (GAN) og diffusjonsmodeller, har revolusjonert feltet for syntetisering av medisinske bilder ved å muliggjøre generering av realistiske og varierte 2D- og 3D-bilder. Disse modellene har løst utfordringer knyttet til begrenset tilgang på data, personvernrestriksjoner og kompleksiteten ved å fange opp hele spekteret av anatomiske og fysiologiske variasjoner innen ekkokardiografi.

GAN-er består av et generator- og et diskriminatornettverk som er involvert i en treningsprosess der de to nettverkene konkurrer mot hverandre. Ved å trene generatornettverket på et stort datasett med ekte ekkokardiografiske bilder, kan GAN-er generere syntetiske bilder som ligner på ekte pasientdata. Dette har lagt til rette for oppgaver som dataaugmentering, anomalideteksjon og syntetisering av ekkokardiografiske bilder for klinisk bruk.

Diffusjonsmodeller har derimot vist seg å være kraftige verktøy for å generere syntetiske bilder av høy kvalitet ved å simulere en diffusjonsprosess. Disse modellene tar utgangspunkt i en støykilde og forbedrer iterativt for å generere bilder som oppnår den ønskede bildefordelingen. Diffusjonsmodeller har vist lovende resultater, bedre enn GAN-er, når det gjelder generering av realistiske

og varierte bilder. Siden dette er en ny metode, har ikke denne typen dype generative modeller blitt anvendt på ekkokardiografiske bilder tidligere.

Bruken av disse syntetisk genererte bildene, både 2D og 3D, som grunnlag for å utvikle dyplæringsmodeller som kan utføre ulike oppgaver, for eksempel segmentering, har vist seg å gi gode resultater. Hjertets volum kan beregnes med høyere nøyaktighet når de syntetiske bildene brukes sammen med ekte bilder for å trene disse modellene.

Bruken av dype generative modeller innen ekkokardiografi har flere viktige følger. For det første løser det problemet med å samle tilstrekkelige mengder med merkede data, noe som ofte er en utfordring innen medisinsk avbildning. Ved å generere syntetiske bilder åpner disse modellene for muligheten til å øke mengden opplæringsdata og forbedre ytelsen til automatiserte algoritmer.

For det andre vil dype generative modeller gjøre det mulig å utforske sjeldne eller komplekse hjertesykdommer som ofte ikke er tilstrekkelig representert i tilgjengelige datasett på grunn av lav forekomst i samfunnet. Ved å generere varierte bilder som inkluderer ulike patologier og anatomiske variasjoner, forbedrer disse modellene forståelsen og analysen av komplekse tilfeller av hjertesykdom.

Generering av syntetiske ekkokardiografiske bilder som er realistiske vil også legge grunnlag for utvikling og evaluering av nye bildebehandlingsteknikker og -algoritmer. Forskere kan teste og optimalisere bildeanalysealgoritmer, evaluere ytelsen til nye metoder for bilderekonstruksjon, og undersøke effekten av ulike bildeparametere på diagnostisk nøyaktighet.

Selv om dype generative modeller har vist stort potensial, så er det utfordringer og begrensninger som må tas høyde for. Dette inkluderer behovet for store og varierte treningsdatasett, samt generaliserbarhet på tvers av forskjellige pasientpopulasjoner og avbildningsmodaliteter, og validering og tolkning av syntetiske bilder i kliniske situasjoner.

Til oppsummering har utviklingen av dype generative modeller, inkludert GAN-er og diffusjonsmodeller, bidratt betydelig til å fremme fagfeltet ekkokardiografi ved å gjøre det mulig å generere realistiske og varierte 2D- og 3D-bilder. Disse modellene har potensial til å gjøre data lettere tilgjengelig, forbedre automatisert analyse og legge til rette for forskning og utvikling innen ekkokardiografi.

I denne avhandlingen har både GAN-er og diffusjonsmodeller bidratt betydelig til syntetisering av 2D- og 3D-ekkokardiografiske bilder. De har bidratt til å håndtere begrensninger ved data og har gitt mulighet for å generere store datasett med kjent "ground truth"-informasjon. Dette har vært spesielt verdifullt for trening og evaluering av dyplæringsalgoritmer som brukes til ekkokardiografisk analyse og diagnostikk. Videre har den kliniske tolkningen vist hvor nyttige de syntetiske bildene er, og hvor mye verdifull informasjon som finnes i de.

List of Papers

Paper I

Tiago, C., Gilbert, A., Salem Beela, A., Aase, S.A., Snare, S.R., Sprem, J., and McLeod, K. "A Data Augmentation Pipeline to Generate Synthetic Labeled Datasets of 3D Echocardiography Images Using a GAN". In: *IEEE Access*. Vol. 10, 2022, pp. 98803–98815. DOI: 10.1109/ACCESS.2022.3207177.

Paper II

Tiago, C., Snare, S.R., Sprem, J., and McLeod, K. "A Domain Translation Framework with an Adversarial Denoising Diffusion Model to Generate Synthetic Datasets of Echocardiography Images". In: *IEEE Access*. Vol. 11, 2023, pp. 17594–17602. DOI: 10.1109/ACCESS.2023.3246762.

Paper III

Tiago, C., Snare, S.R., McLeod, K., and Sprem, J.. "Denoising Diffusion Model for 3D Echocardiography Image Generation: Image Usability and Clinical Relevance". *Submitted for publication to* IEEE Open Journal of Engineering in Medicine and Biology.

Contents

List of Figures

List of Tables

Chapter 1

Introduction

1.1 Motivation

According to the most recent numbers from the World Health Organization, cardiovascular disease are the primary cause of death worldwide accounting for almost 18 million deaths every year [1]. All the conditions included in this group severely affect the heart's capability to deliver the blood through the body in an efficient manner. Therefore, it is of utmost importance to find and develop clinical tools that would allow to efficiently and accurately diagnose these cases in order to provide better treatment to the patients.

Amongst all the cardiac imaging modalities, cardiovascular ultrasound, or echocardiography, is unequivocally the most commonly used. Only relying on high frequency sound waves instead of ionizing radiation, echocardiography allows real-time imaging of the heart offering a complete assessment of this organ's anatomical structure and beating function, including the ability to visualize blood flow. An echocardiographic exam allows to quantify a wide range of clinical parameters related to the heart's performance, facilitating a more accurate and precise diagnosis.

To be able to save clinicians' time without reducing the quality of patient care, and at the same time facilitating clinical workflows within each healthcare system, there is a large need to develop automatic methods and algorithms, taking advantage of the fast evolving technological developments. As the global population continues to grow, the demand for more efficient and effective processes to acquire echocardiography images becomes increasingly critical. One solution to this challenge is automation, which can be facilitated through the use of deep learning algorithms. However, in order to build such algorithms, we need a large and diverse set of data. Obtaining this data can present significant challenges, including strict privacy and anonymization policies related to medical data, and requirement of skilled professionals to acquire echocardiography images, due to the various imaging scanners and diverse patient anatomies. Despite these hurdles, the need for automation remains pressing, and efforts to overcome these challenges are essential to continue advancing the capabilities of automatic methods in healthcare.

Deep learning replicates the behavior of the human brain by using artificial neural networks, which are inspired by the structure and function of biological neurons in the brain. These networks consist of layers of interconnected nodes that receive input signals, process them, and produce output signals. The neural network is fed with vast amounts of data, and the connections between the nodes are adjusted based on the patterns and relationships that the network learns from the data. This process is similar to the way that the human brain forms and

strengthens connections between neurons through repeated exposure to specific scenarios. Therefore, by using these artificial neural networks, deep learning algorithms are able to mimic the behavior of the human brain in processing and learning from data. Deep learning algorithms have been widely adopted for a range of applications, from image segmentation, which involves identifying and isolating specific features of interest in an image, to the quantification of cardiac anatomy and performance metrics in echocardiography exams. Thanks to their ability to achieve high levels of performance, deep learning algorithms have become an essential tool for facilitating the analysis of medical imaging data.

Training deep learning algorithms necessitates vast amounts of data. Acquiring echocardiography image datasets is complex, and publicly available image collections are scarce. To overcome this obstacle, one option is to employ deep learning tools to create images that appear genuine. After all, irony aside, while deep learning algorithms do require vast amounts of data to train effectively, they can also be used to generate synthetic data that mimics real-world examples. This intriguing aspect of deep learning highlights the potential for it to not only consume data voraciously but also contribute to data creation, opening up new possibilities for overcoming data sparsity challenges.

Deep generative models have demonstrated remarkable aptitude in generating synthetic image data in medical imaging [2]. These models can generate images with considerable accuracy. They are a powerful tool for generating synthetic medical images that can aid in the training of deep learning algorithms as they can create high-quality, realistic-looking medical images, even in cases where data is limited or scarce. This ability is particularly useful in medical fields where collecting large amounts of data is difficult. Additionally, these generative models can be used for data augmentation, increasing the diversity and amount of available data, thereby enhancing the accuracy and robustness of downstream deep learning models. Overall, deep generative models hold great promise for improving medical diagnosis, treatment, and research.

Studies demonstrated the capability of deep generative models to generate data and its utility when developing automatic deep learning methods to perform clinical tasks [3]. Nonetheless, different imaging domains present varying image characteristics which increase the complexity of the generative process. This is particularly evident in the echocardiography domain since these images have inherent speckle patterns which are difficult to replicate, to mention the most characteristic aspect. Similar to previous works done on medical image synthesis, there is room for improvement when trying to synthesize echocardiography images with relevant clinical information, meant to be used when developing helpful clinical deep learning algorithms.

1.2 Aims of the project

The main goal of the project was to develop deep generative models to synthesize large databases of synthetic echocardiography images, both 2D and 3D, and

to study and explore in further detail the quality and utility of such synthetic images.

The essential objectives of the project were the following:

- Describe the development of a 3D deep generative model.
- Use the synthetic 3D images to train a segmentation algorithm and prove their utility in creating more accurate deep learning algorithms.
- Create a more advanced and sophisticated 2D deep generative model and further explore its working principle, synthetic image quality, and relevant information present in them.
- Expand the previous 2D model to 3D and compare the 3D synthetic images with the ones from the first objective.

1.3 Context of the project

This thesis was conducted and completed as a part of the European Union funded MARie Curie Intelligent UltraSound (MARCIUS) project [1]. Being a European Innovative Training Network, the MARCIUS project is a consortium between industry and academia which includes several doctoral training programs. The project combines several research fields, such as cardiac imaging and physiology, and biomedical engineering, where an in silico simulation platform including both the generation of virtual patients and associated realistic image data is the main objective of the project. This thesis focuses on bringing together state of the art artificial intelligence algorithms to generate echocardiography data.

The work covered in this thesis explores in further detail the creation of synthetic 2D and 3D echocardiography images and datasets, relying on common but also on novel deep generative models, to achieve that purpose. It also explores the utility of such synthetic datasets in the development of automatic deep learning algorithms, aiming to facilitate clinical analysis of echocardiography exams, in light of the planned work packages (WPs) shown in Figure 1.1.

This project's work was essentially conducted at GE Vingmed Ultrasound, part of GE HealthCare, one of the leading companies when it comes to cardiac ultrasound imaging. A collaboration with Intelligent Ultrasound, one of the industrial beneficiaries of the MARCIUS project, led to the application of one of the proposed approaches described in this thesis in their products. The University of Oslo was the academic partner throughout the duration of this industrial PhD program.

[1] https://www.marcius-project.com/ - Marie Sklodowska-Curie grant agreement No 860745

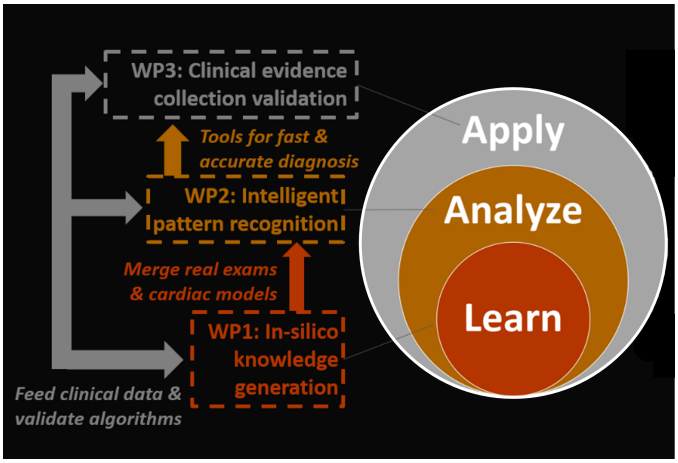

Figure 1.1: MARCIUS project overview and working plan, with defined WPs.

Chapter 2

Background

2.1 The heart

The heart is a muscular organ that plays a vital role in the circulatory system of the body. Its primary function is to pump blood through the body [4], supplying oxygen and nutrients to all the organs and tissues while removing waste products and carbon dioxide. The heart is divided into four chambers, the right and left atria (LA), and the right and left ventricles (LV). Additionally, the arteries are responsible for taking the blood away from the heart and the veins bring this fluid back to it. The LV is the largest and strongest chamber of the heart and is responsible for pumping oxygenated blood through the aortic valve to the aorta and then through the whole body. The blood then returns to the heart, specifically to the right atrium, and this now deoxygenated fluid is pumped to the lungs, from the right ventricle, where the carbon dioxide is eliminated and oxygen can enter the blood stream once again. After this oxygenation step, the blood returns to the LA and from here, passing to the LV, the whole cycle can start yet again. Between each atrium and ventricle there are valves which open and close allowing the blood to flow from the atria to the ventricles. The tricuspid valve allows the blood flow from the right atrium to the right ventricle, and the mitral valve, also known as bicuspid, lets the blood flow from the LA to the LV.

The human heart works in a cyclic way, as earlier described and shown in Figure 2.1. Each cardiac cycle comprises 2 different phases: ventricular systole and ventricular diastole. The former is the contraction of the myocardium, the heart muscle, ejecting the blood to the aorta and then through the arteries, and the latter is the heart relaxation, when the ventricles have time to fill up again.

Considering the left heart, i.e. LA and LV, its ventricular systole starts when the mitral valve closes. The pressure inside the LV increases and while it is higher than the pressure in the aorta it causes the aortic valve to open and the blood to be ejected. As soon as the aortic pressure becomes higher than the left ventricular one, the aortic valve closes and the pressure inside the LV keeps decreasing until it is lower than the one in the left atrium. The diastole then starts and at this point in time, the mitral valve opens and the blood can pass from the atrium to the ventricle. When the pressure in this latter cavity becomes higher than the pressure in the atrium, the LA contracts to let the rest of the blood pass through, the mitral valve closes and a new cardiac cycle can start again. At the same time the same cycle happens on the right side of the heart.

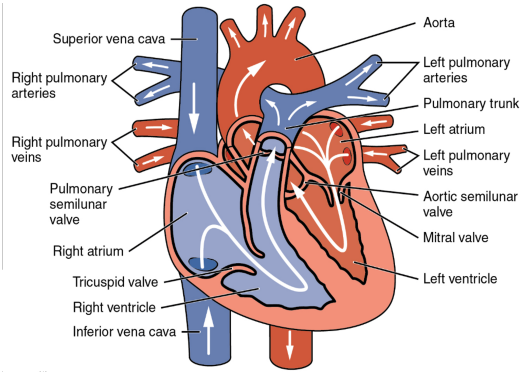

Figure 2.1: Human heart anatomy and blood circulation. Image taken from [4].

2.2 Cardiovascular disease and diagnosis

Worldwide, cardiac diseases or cardiac complications due to other conditions are the leading cause of death, affecting all levels of society [1]. A common way to detect them and reach a diagnosis is by imaging the heart and deriving certain measurements from the acquired images, provided that these reach the desired quality, both anatomical and temporal.

The heart performs its beating action throughout the whole life of a person, in a mechanic and rhythmic way. Several factors impact cardiac health such as stress, drug use, obesity and sedentarism, to name a few [1]. As in all mechanical systems, its performance can be assessed over time in order to detect any conditions that might occur. Besides mechanical, the heart also has an electric conducting system which can also be affected by different pathologies, constraining its pumping function.

Arrhythmias are a common cardiac condition where the heart shows an irregular beating pattern and are mainly caused by electrical dysfunctions. On the other hand, cardiomyopathies are morphological conditions which affect the heart muscle, the myocardium. This muscle can become thicker or stretched, either of them impairing the heart contraction and relaxation.

Cardiac imaging represents a great way to diagnose such pathologies, allowing to analyze the heart's anatomy. Many imaging modalities exist with different pros and cons. Besides allowing to visualize the heart, the different imaging modalities facilitate the estimation of different cardiac performance parameters which reflect the status of this organ.

Amongst the most relevant parameters, Ejection Fraction reflects the heart's capacity to pump blood efficiently. This value expresses how much blood is pumped out by the heart at each contraction of the myocardium. A small value suggests some malfunction with the heart, indicating the presence of a possible cardiomyopathy or even an arrhythmia [5].

2.3 Ultrasound imaging - Echocardiography

Regarding cardiac imaging, many modalities can and are used in clinical practice [6], and the clinical scenario is always take into account. Different clinical needs require the usage of different imaging modalities. Magnetic Resonance (MR) is the medical imaging modality that provides heart images with the best anatomical details. Besides providing clinicians with anatomically clear cardiac images, it does not use any ionizing radiation as magnetic fields are the only responsible element for the collected imaging signal. However these exams are time consuming and it is a very expensive imaging modality. The magnets and coils are expensive, the scanner is not portable and special installations for cooling and enclosing the magnetic fields are necessary. Computed Tomography (CT) is less expensive than MR and the images obtained with this modality also have a very high anatomical quality. Nonetheless, it relies on ionizing radiation, X-rays in this case, most of the scanners used in clinical sites are not portable and temporal resolution is usually low [6]. Ultrasound scanning represents the best compromise between these pros and cons. This imaging modality does not use any ionizing radiation as very high frequency sound waves are used to image the body, the scanner is fully portable, it is inexpensive to perform such examinations and temporal resolution is one of its main strengths [7]. This turns out to be an important detail when imaging a moving organ like the heart [8]. For these reasons, Ultrasound is the most common choice to visualize the human heart (Figure 2.2).

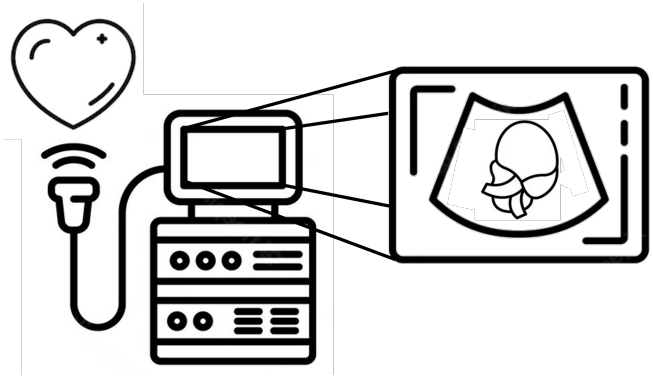

Figure 2.2: Ultrasound scan of the heart.

As mentioned, Ultrasound relies on high frequency sound waves, between 2 to 18 megahertz (MHz), which are sent through the body. Sound is a mechanical wave, which means that it needs a medium to propagate itself, and different mediums have different mechanical properties which affect the wave propagation in specific ways. When the ultrasound beam hits a barrier, i.e. passes from a medium with certain characteristics to another with different ones, it is reflected but not totally, with a part of it still propagating itself. The reflected wave is

called echo and is detected by the transducer, which is responsible for both the emission and the reception of the ultrasound beams. This process is illustrated in Figure 2.3.

When the heart is the organ being imaged with this modality, it is referred to as echocardiography, where the reception of different echos, with different frequencies, generated by the transmission of the ultrasound beam throughout the different tissues and chambers of the heart can generate an anatomical image of this organ [9].

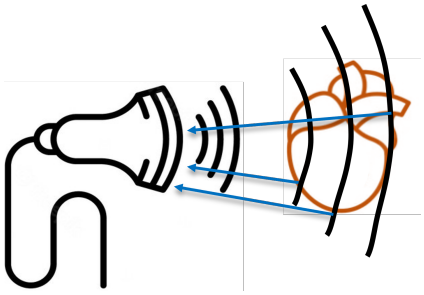

Figure 2.3: Ultrasound propagation through different physical barriers. The blue arrows represent the reflected ultrasound waves, which are detected by the transducer.

Echocardiography can acquire both 2D and 3D images of the heart together with temporal information about its contractile condition, where the anatomical resolution of the latter type of images is inferior. 3D echocardiography scans make the estimation of ventricles and atria volumes more accurate, comparing to the more conventional approaches that rely on 2D images [10]. With these volumes and other parameters, more complex performance parameters can be estimated such as the ejection fraction or the sphericity index, for example. However 3D echocardiography images are more challenging to acquire because the acquisition process takes longer time and special software is required to reconstruct and analyze the image [11] (Figure 2.4).

Therefore it is not unusual that a current lack of public 3D echocardiography datasets exists, due to the presented acquisition limitations. Nevertheless, such imaging modality still holds a high potential to be used as a data source since it is the main go-to cardiac imaging modality.

2.4 Simulation in Echocardiography

Over the years, the development of simulation methods has played a crucial role in advancing the field of echocardiography. Before the development of Deep Learning (DL) and its application to image synthesis, various simulation techniques were employed to generate synthetic echocardiography data for research, training, and educational purposes using for example biophysical models. Simulating ultrasound data is a standard practice for designing new ultrasound systems

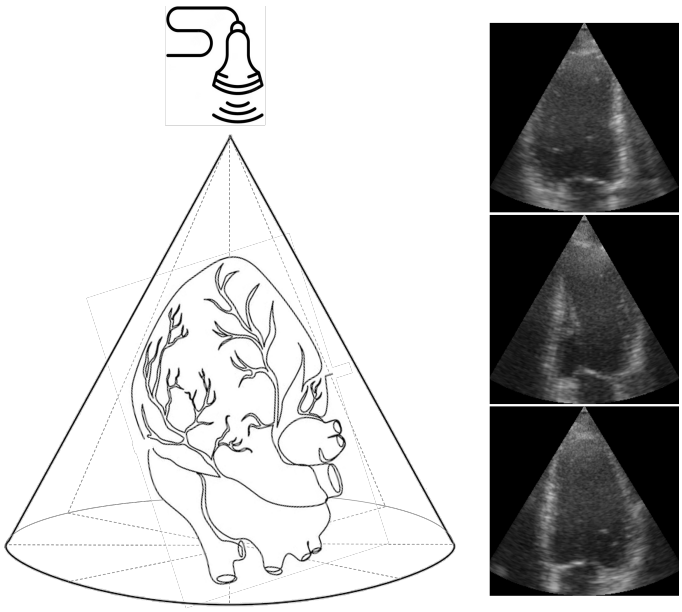

Figure 2.4: Echocardiography acquisition cone and different tilting imaging planes of the heart.

and validating novel ultrasound image processing techniques, particularly in the medical field [12], [13].

One of the earliest and more efficient approaches to simulating medical ultrasound data involved the use of mathematical and physics models. Developed by [14], Field II relied on linear acoustics to establish a gold standard to simulate ultrasound transducer fields and ultrasound imaging. This simulator is capable of calculating the emitted and echoed beams for a large number of different transducers, with different physical properties, and simulate realistic images of human tissue. In the linear domain Field II shows strong results but when the simulations need to be performed in 3D or in more complex domains, other simulators rise as a more accuracy-time efficient approach [15]. COLE [16] is able to simulate 2D and 3D images by decomposing complex convolution operations to a series of sequential 1D convolutions between the emitted ultrasound signal and the scatters present along each image line, and FUSK [17] applies the Fourier transform to generate both 2D and 3D ultrasound images.

Physical models comprised physical materials that mimicked the acoustic properties of cardiac tissues. By manipulating the model and using transducers, ultrasound-like images can be obtained. Physical models provided a hands-on learning experience, allowing practitioners to understand ultrasound principles and practice image interpretation.

With advancements in computer graphics and imaging technology, computer-based phantoms became popular in echocardiography simulation [18]. These

simulations use phantoms, mathematical algorithms and computational techniques to generate synthetic ultrasound images. By simulating the interaction between ultrasound waves and virtual cardiac structures, these phantoms are even able to generate realistic images with known ground truth information.

The Finite Element Method (FEM) is a numerical technique widely used in engineering and biomechanics. Despite the inherent complexity linked to these methods, in the context of echocardiography simulation, FEM models were employed to simulate the mechanical behavior of the heart. These models incorporate anatomical data and cardiac mechanics to predict cardiac motion and generate corresponding ultrasound images, mainly 3D images of the mitral valve [19], [20]. FEM simulations provide valuable insights into cardiac dynamics and contribute to the development of advanced echocardiography imaging techniques. However, due to its complexity levels, these are not as widely used as the physical models previously described, for example.

Moving towards more automatic image simulation techniques, Image-based rendering approaches leverage existing echocardiography images or datasets to generate new synthetic images. By manipulating the acquired images through geometric transformations, such as rotation, translation, and deformation, realistic synthetic images can be generated. These techniques allowed researchers to study image artifacts, test imaging algorithms, and simulate different imaging scenarios.

Simulation methods in echocardiography prior to DL offered several advantages. They provide a controlled environment for testing and validating imaging algorithms, avoiding ethical concerns and variability associated with clinical data. Simulations allow for the generation of large datasets with ground truth information, facilitating algorithm training and evaluation. Additionally, these methods offer educational benefits, enabling researchers and practitioners to gain hands-on experience without patient involvement. However, these simulation methods also have limitations. They often rely on simplified assumptions and mathematical models, which may not capture the full complexity and variability of real-world echocardiography data. Generating realistic motion and anatomical variations poses challenges, as does simulating realistic echocardiography image artifacts and noise. Furthermore, these methods require significant expertise and computational resources for their implementation, especially when the initial assumptions become more complex and precise.

The integration of DL has brought significant advancements, enabling more sophisticated and realistic simulation of echocardiography data, ultimately enhancing diagnostic accuracy and patient care in this important medical imaging domain.

2.5 Deep Learning

Within the large Artificial Intelligence (AI) domain, the DL concept is defined as a set of computational methods that are inspired in the human brain and by the way this organ works biologically. The working principle behind these

methods shows an attempt to learn from raw data, performing feature extraction without any user influence or bias, and make a prediction. DL algorithms reflect a high independence from the user and from any *a priori* constraints that would influence the final outcome of the learning process. Besides this major advantage, DL methods are more powerful than others in the AI field such as Machine Learning ones for example, as they represent a more specific subgroup. Their ability to mimic the human brain behavior makes them ready to deal with immensely large amounts and different types of data, as is the case of images, both 2D and 3D. At the cost of dealing with more complex types of data which can show a large variability amongst them, the computational power and learning time of DL methods is quite large.

Depending on data availability DL models can be trained under different scenarios. The most common one is supervised training, where the algorithm attempts to learn a relationship between the given input and expected output data. However, the expected output for a certain input is not always known or available and unsupervised training is as an attempt of the DL model to detect and learn patterns in the input data that seem relevant, even though there is no correct or wrong output. In between these two scenarios, there is also semi-supervised learning whose algorithms are trained on a combination of data with and without a known output.

In order to replicate human brain behavior, DL relies on Convolutional Neural Networks (CNNs), amongst other possibilities. As in an individual's brain, these are made of neurons which are arranged in layers where all the necessary calculations to the learning process are performed. CNN layers are arranged in a way such that neurons belonging to the same layer are not connected to each other but, instead, are connected to the neurons of the adjacent layer [21]. This structural organization makes the network more robust when dealing with more complex data, such as images. The CNN derives its name from its use of convolutions, a mathematical operation, to extract a multitude of features from the input image. As it happens in biology, the neurons in the visual cortex of the brain process limited receptive fields at a time overlapping them to create a full field with all the visual information. CNNs analyze the input images in this way as well. The artificial neurons of one layer learn from different small regions of one image, perform the convolutions, and their results are shared with the neurons of the following layer, this way reducing the number of parameters necessary for the model to learn, since it does not have to learn the same features at every layer.

Supervised training usually reaches the best results provided that enough data is available to train these models, since the learning process is usually long. However, large datasets are difficult and expensive to obtain, especially when the trained DL model focuses on a specific task that requires a certain type of data.

2.5.1 Deep Learning in Medical Imaging

Since DL models can closely mimic the human brain behavior, its utility in healthcare is of relevance. These can facilitate and/or automate several clinical procedures and workflows (Figure 2.5), saving clinicians' time which can be allocated for more important cases or tasks [22]; [23].

It has been proven that some DL models can actually achieve performance levels very close to the ones reached by humans [24], performing the most variate tasks such as image classification or segmentation. Going from a more simple binary classification evaluation of a patient (for example, priority or not priority) to a more demanding task as the segmentation of an anatomical structure, quantification of a certain performance metric and consequent patient evaluation.

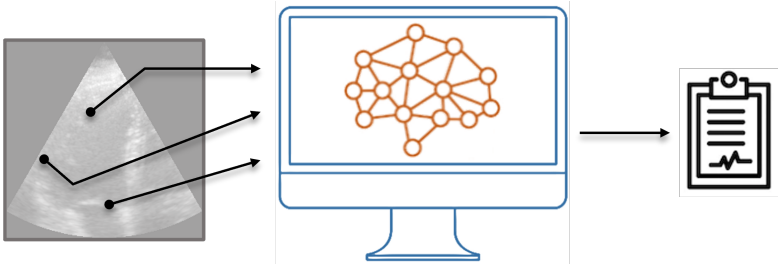

Figure 2.5: Deep Learning in medical imaging. Different models can mimic the brain behavior to analyze different types of medical information and ease the final diagnosis.

Clinicians also agree [25] that DL not only saves them time, which they can then dedicate to more complex clinical cases, but also provides a faster diagnosis to the patient without compromising its quality and accuracy, and sometimes even increasing the confidence levels.

DL in medical imaging is mostly used to perform segmentation tasks [26], which provide the clinicians with concrete delineations of the most relevant structures on an image. These segmentations can facilitate a final diagnosis by themselves, or are very often used downstream, since a variety of clinical performance parameters can be estimated from them.

Medical image domain translation is also an area where DL models have a large influence on [27], [28]. By performing domain translation, it becomes possible to transform an initial image into another that belongs to a different domain, i.e. group, and exhibits distinct characteristics compared to the original one. This is the case for noise reduction (remove unwanted noise from a noisy image [29]), super-resolution (increase the final image resolution [30]), modality conversion, and also image synthesis operations.

Image synthesis has two main purposes: data augmentation and generation of images not acquired due to clinical workflows. The data augmentation purpose allows to generate plausible images with sufficient variability. Considering that

the development of DL frameworks requires large and variate datasets, this image synthesis approach is commonly addressed in order to create synthetic datasets which can be used for training of CNNs for clinical usage [31]. This second purpose of image synthesis can be interpreted as modality conversion but where the final output is not known, i.e. the image from the modality we are converting to does not exist [32], [33].

2.6 Deep Generative Models

2.6.1 Generative Adversarial Networks

Generative Adversarial Networks (GANs) represent a specific subset of DL models, with a specific architecture, where it is possible to generate new data as it is needed [34]. These generative models are widely used to create new images since they are capable of doing so by learning from a given initial dataset.

A GAN is an unsupervised algorithm that has a partially supervised training process. As mentioned, supervised training requires the output for each input, i.e. the label, to be known during training, so the model can learn how to generalize, i.e. make predictions, on new data. GANs, on the other end, do not use labeled data since their final goal is not to make a prediction but generate new examples of the input data. During training, when the input data is an image for example, the GAN generates a fake image and attempts to discriminate if it is indeed fake or real, like the input images [35]. This way GANs set up a supervised training scenario to deal with an unsupervised learning task.

GANs are DL generative model architectures made of a generator and a discriminator. The principle behind this type of models' training is shown in Figure 2.6 and is such that the generator tries to create an image similar to the input ones, feeds it to the discriminator and this tries to distinguish if the image is indeed real or not. This training loop occurs until the generator can make a real looking image and the discriminator can be tricked to evaluate such image as real instead of synthetic. Both the generator and the discriminator are CNNs and the latter simply performs a binary classification task.

Since there is no labeled data when training a GAN, what this model attempts to do is generate images such that their probability distribution, p_{gan}, matches the input dataset one, p_{data}. To evaluate the similarity between these two probability distributions a loss function is calculated during each training step. The discriminator can be seen as a way to measure the generator loss as its performance depends on the generator output. However, this approach to calculate the loss assumes a fixed discriminator which can be easily misled by the generator as the generated images would present very small variations comparing to one single generated image in time which provided a very small loss value. This problem is known as mode collapse and happens when there is very little variance in the generated samples. To fix this issue the loss function has to depend on the discriminator performance as well, creating an adaptive and non-static generator loss. The generator, G, randomly generates a sample image, x, from a noise vector, z, in order to make $x = G(z) \implies p_{gan} \approx p_{data}$. On its

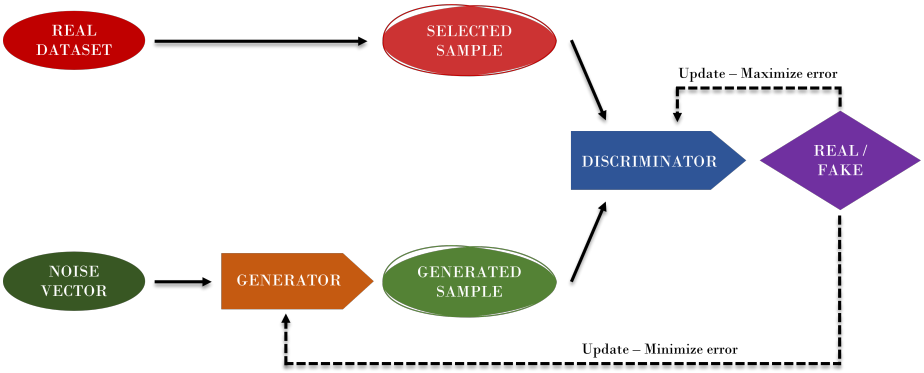

Figure 2.6: Training process of a GAN. The generator synthesizes samples, which are compared with real ones by the discriminator in order to distinguish between real and synthetic data.

turn, the discriminator, D, compares the generator output, $G(z)$, with the real sampled image, x.

The most common way to keep track of the difference between the two distributions and focus on both losses is by using the minimax loss function. The minimax loss function reaches its optimum value when the generator minimizes and the discriminator maximizes it. This function, described by (2.1) can be broken down in two parts, each of them reflecting the influence of the generator and the discriminator:

$$\min_G \max_D \mathcal{L} = \mathbb{E}_{x \sim p_{data}}[log(D(x))] + \mathbb{E}_{z \sim p_z}[log(1 - D(G(z)))] \tag{2.1}$$

During GAN training, the discriminator and generator are trained in alternating times, in an adaptive and flexible way. During the discriminator training, the generator is kept constant and the discriminator tries to learn how to distinguish between real and fake images, recognizing the limitations of the generator. Vice-versa, the discriminator is kept unchanged while the generator is training, otherwise this part of the GAN would be trying to achieve an optimal point that would always be moving, potentially leading to no convergence in the performance of the GAN model.

Conditional GANs rely on a condition to synthesize images [36], creating a variation of the traditional GAN model. In conditional GANs both the generator and discriminator's performance is subjected to extra information, which makes the generative process more specific. During the generator training and adding to the noise vector z previously described, it also receives a condition image y. This way, this deep generative model attempts to model the conditional probability $p_{gan}(z|y)$. The influence of this condition also affects the discriminator training and it is reflected on the new loss function, as shown in (2.2):

$$\min_G \max_D \mathcal{L} = \mathbb{E}_{x \sim p_{data}}[log(D(x|y))] + \mathbb{E}_{z \sim p_z}[log(1 - D(G(z|y)))] \tag{2.2}$$

The training datasets used to train GANs can be of two different types: paired or unpaired [37]. Conditional GANs rely on paired training datasets, i.e. the mapping function between the model's output and the input exists and is possible to predict.

The pix2pix model [32] is a type of conditional GAN specifically designed for image-to-image translation tasks. It was introduced as an effective framework for learning the mapping between input and output images. The training process involves a generator and a discriminator that engage in an adversarial game. The generator takes an input image and tries to transform it into an output image that resembles the target image. The discriminator, on the other hand, aims to distinguish between the generated output images and real target images being trained to classify them. It provides feedback to the generator by indicating the quality of the generated images and guiding the generator to produce more realistic outputs. As previously described, during training the generator is trained to minimize the difference between the generated output image and the target image. This is achieved through an adversarial loss, which encourages the generated images to be indistinguishable from the real target images (2.2). Additionally, a pixel-wise L1 loss (2.3) is commonly used to ensure that the generated images accurately match the target images at a pixel level.

$$\mathcal{L}_1 = \frac{1}{N} \sum_{i=1}^{N} (x_i - z_i)^2 \tag{2.3}$$

The adversarial training of the pix2pix model allows it to learn to generate high-quality and visually coherent output images that closely resemble the target images. It has been widely used in various domains, including computer vision, graphics, and image editing. Its versatility and effectiveness make it a powerful tool for tasks that require translating images from one domain to another, opening up possibilities for applications in areas such as medical image synthesis, style transfer, and data augmentation.

However, the existence of these paired datasets is limited. To deal with unpaired datasets, when the mapping between the input and output does not exist, a different type of GAN is used, the cycle GAN (CycleGAN). CycleGANs are trained in an unsupervised fashion [38], combining adversarial training with a cyclic loss function based on auto-encoders [39]. These models train two pairs of generator and discriminator, one for each training dataset, since these are not related. During training, the generated image from the first generator is used as input to the second generator, whose output should match the initial image (the input to the first generator). The cycle loss quantifies the difference between the image synthesized by the second generator and the original one. The same reverse process also occurs.

Both types of GANs can be used to synthesize new images and perform domain translation, regardless of the use of paired or unpaired training datasets.

As mentioned, GANs synthesize high quality image samples but suffer from lack of variability. In order to address this downside and generate high-quality and diverse images by capturing the fine-grained details and variability in the data, [40] presented a Vector Quantized Generative Adversarial Network (VQ-GAN). This generative network combines elements of both GANs and vector quantization, a compression technique used to represent data points using a predefined set of representative vectors called a codebook or dictionary. It is commonly used to reduce the dimensionality of data while preserving important features. By using vector quantization, the original image dataset can be efficiently represented using a smaller number of vectors from the codebook, reducing the memory or storage requirements.

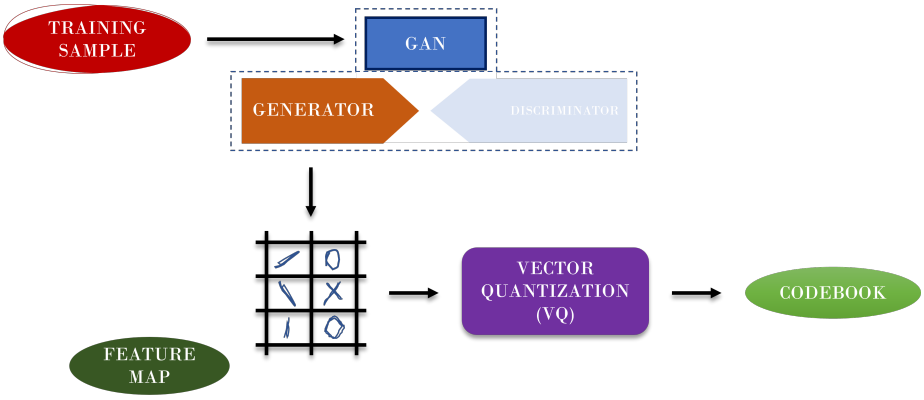

Figure 2.7: Vector quantization operation. The codebook is generated from the image features extracted by the GAN generator. The larger it is, the more variable the synthetic images will be.

In the VQ-GAN the generator takes random noise as input and produces an image. However, instead of directly generating pixel values, it generates a feature map of visual parts of the images. The vector quantization operation then transforms these image features into the codebook, in a similar way as a clustering algorithm would do, i.e. grouping the features in a predefined set of vectors that represent different image characteristics. During training, the VQ-GAN incorporates a quantization loss that encourages the generated codes to match the nearest neighbor vectors in the codebook. This loss promotes the generation of diverse and representative codes, ensuring that the generated images capture the essential characteristics of the training data. Then, as in an adversarial training scenario, to enforce the generated images' realism, the discriminator evaluates the quality of the generated images and provides feedback to the generator.

The combination of the quantization loss and the adversarial training enables the VQ-GAN to generate realistic and diverse images. By leveraging the discrete codes and the codebook, it captures both local and global features of the images,

allowing for fine-grained control over the generated content. The quality of the generated images depends on the size and quality of the codebook. A larger codebook can better represent the data, but it also requires more memory and computational resources. Designing an optimal codebook involves striking a balance between representation accuracy and efficiency.

2.6.2 Denoising Diffusion Models

Denoising Diffusion Models (DDMs) [41] are also part of the Deep Generative models' group and its application appeared very recently [42], outperforming GANs [43]. These models work in a destructive way, as they take an image sample and progressively add noise to it, corrupting its initial information, until the model ends up having an image made of pure Gaussian noise. The training process of a diffusion model aims to recover the original image starting from this noisy input image, by learning the reverse denoising process across a sufficiently large number of steps. Once the model is trained, it is possible to generate any number of image samples just by inputting a random noisy image.

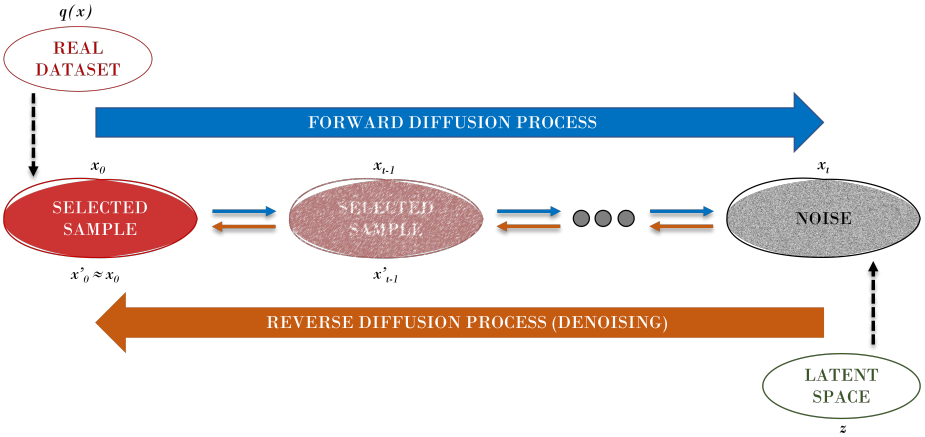

Figure 2.8: Training process of a DDM. The model progressively adds Gaussian noise to the input sample and reverses the process in order to recover the initial input.

DDMs are latent variable models [44]. A statistical latent variable model maps observable variables to a latent space made of latent variables, that can only be deduced from the initial observable ones. This is where the noisy images, obtained after destroying the training samples, lay.

Figure 2.8 shows the training procedure of a DDM. The forward process consists of iteratively adding Gaussian noise to the input training image, $x_0 \sim q(x_0)$, during a sufficiently large number of steps, T, until a sample x_T, a latent variable, is obtained belonging to a Gaussian distribution $z \sim \mathcal{N}(0, I)$, the latent space. This process creates a Markov chain which is reversed during

the reverse diffusion process [45]. During this process, each reverse step relies on a neural network to provide estimates for μ and/or Σ. These parameters describe the data distribution of the generated images, $p_\theta(x_t)$, whose optimal values should be the closest as possible to the training dataset distribution's. Due to the iterative nature of this reverse process, the DDMs sampling time is consequently long.

Therefore, the loss function used by such models is based on the Kullback-Leibler (KL) divergence between two normal distributions and can be minimized in order to find the set of parameters that provide the smallest divergence between both distributions, as shown in (2.4) [46]:

$$\min_\theta \mathcal{L} = \min_\theta \sum_{t>=1} \mathbb{E}_{q(x_t)} \left[D_{KL}(q(x_{t-1}|x_t)||p_\theta(x_{t-1}|x_t)) \right] \tag{2.4}$$

where D_{KL} represents the KL divergence, $\mathbb{E}$ denotes the mathematical expectation, and p_θ the neural network estimate for the generated samples distribution parameters.

This way, DDMs do not require any type of adversarial training. i.e. paired datasets, to generate high quality image samples, not rising image variablity issues as the ones described in Section 2.6.1, common to occur when training GANs.

2.6.3 Adversarial Diffusion Models

Both GANs and DDMs have strengths and weaknesses. On the one hand GANs can sample high quality synthetic images in a relatively fast time interval, on the other hand, the diversity of the generated samples is not large. Oppositely, DDMs generate high quality images with a lot more variability, at the cost of needing a larger time window to do so.

Combining the strengths of these two deep generative models [46], the adversarial diffusion models are capable of generating widely diverse and high quality image samples, with a small sampling time. These more robust models use a DDM in combination with a GAN, creating adversarial diffusion models.

The forward diffusion step is performed in the same way, i.e. from the analysis of the input images adding noise until the creation of the latent space, but the reverse denoising process uses a conditional GAN, instead of a regular CNN, to learn the statistical parameters associated with the probability distribution implicit on the training dataset. The conditioning, y, of the sampling process is done during the reverse denoising step [47] and can be applied via class labels, text, or images, for example. During each denoising step, the model attempts to predict the statistical distribution parameters, including the conditioning information [48], $p_\theta(x_t|y)$. At the same time, and in order to save time during training and reduce the sampling time, using a GAN during the reverse process allows only one or very few denoising steps to be learned.

Ozbey et al. [49] showed that these adversarial diffusion models can be trained using a guide image, at each denoising step, to help the model learn the

reverse diffusion process, this way increasing the anatomical information and quality of the generated images.

Depending on the type of neural network used to learn the reverse diffusion process, i.e. a CNN or a GAN, DDMs are also capable of generating new images and perform domain translation operations.

2.6.4 Deep Generative Models in Medical Imaging and Echocardiography

The application of deep generative models such as GANs and DDMs represents a large field of research in medical imaging. In [50] the authors described how GANs have been widely used to generate realistic medical images. A large part of these developments result from using these DL techniques on imaging modalities other than echocardiography, since these modalities have a less challenging acquisition process.

However, recently Gilbert *et al.* [51] developed a pipeline to generate 2D echocardiography images with corresponding anatomical labels, relying on a GAN to perform unpaired image translation [38]. Generating 3D echocardiography images presents different challenges, due to the image acquisition process technicalities addressed on Section 2.3. Using deep generative models to synthesize such type of images is therefore more demanding and did not previously set any reproducible results, yet.

More recently, the usability of DDMs brought an alternative to GANs. Due to the generated samples quality and diversity, its usage in medical imaging is recent. From image reconstruction [52], to segmentation [53], these models showed great potential. Its ability to generate medical images has been explored in [49], both for 2D images [54] and 3D [55], and [56], where 3D + time cardiac MR images were synthesized. Similarly to what happens with GANs, these models' capacity to generate echocardiography images, of any dimensionality, has not been fully explored though.

2.7 Summary of Papers

This thesis comprises three scientific papers focused on both 2D or 3D echocardiography image generation, using deep generative models, and their impact on clinical use. Figure 2.9 gives an overview of the MARCIUS project, and shows how the papers written on this thesis contribute to the overall project.

Paper I focuses on the creation of a data augmentation tool based on a 3D GAN to synthesize 3D echocardiography images, also showing the utility of generated images to train DL segmentation algorithms.

Paper II demonstrates how adversarial DDMs can be a more efficient approach to generate medical image data, specifically for 2D echocardiography images.

Paper III further investigates the performance of a 3D DDM combined with a particular type of 3D GAN, responsible for extracting the real image characteristics. Synthetic 3D echocardiography image samples, generated by the diffusion model are further analyzed and validated by experienced users.

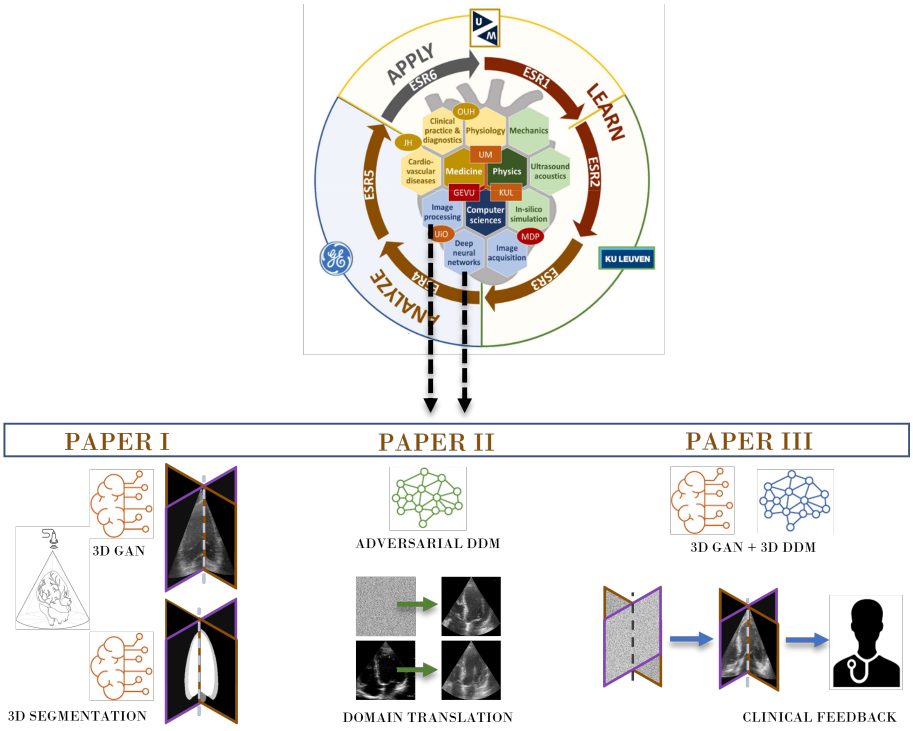

Figure 2.9: Summary of written papers in light of the PhD project overview.

2.7.1 Paper I: A Data Augmentation Pipeline to Generate Synthetic Labeled Datasets of 3D Echocardiography Images Using a GAN

Cristiana Tiago, Andrew Gilbert, Ahmed Salem Beela, Svein Arne Aase, Sten Roar Snare, Jurica Sprem, and Kristin McLeod, *IEEE Access*, 2022.

In this scientific paper, we designed an automatic pipeline to generate 3D echocardiography images with corresponding anatomical masks, using a 3D GAN. These images were then used to train a 3D segmentation DL algorithm.

Privacy concerns related to the usage of medical data, together with the lack of publicly available echocardiography images datasets, and the large requirement

of data to develop DL models, creates the need for a tool which can generate synthetic medical images with relevant clinical information associated with them. The proposed pipeline allows to generate 3D echocardiographic images with corresponding ground truth labels, this way alleviating the need to acquire the images and the subsequent laborious and error-prone manual labeling process.

We constructed a 3D GAN model, named 3D Pix2pix, based on the original 2D version of a paired GAN. As mentioned in Section 2.6.1, a paired GAN learns how to generate images belonging to either of two imaging domains. We trained our 3D deep generative model on a initial dataset which included detailed anatomical segmentations of the heart as ground truth label sources, provided by a cardiologist. This dataset was combined with a second one made up of the corresponding real 3D echocardiography scan images. At inference time, i.e. after training and fine-tuning the model, and to generate the synthetic 3D dataset, high resolution anatomical models from CT were used as the ground truth label image, for which the GAN generated the corresponding echocardiography image.

The synthetic images were created and then post-processed in order to increase their quality. Their qualitative analysis revealed that there is a good delineation of the main structures of the heart, namely the LV, LA, and myocardium (MYO), closely following the anatomical labels extracted from the anatomical models. A further analysis of the obtained results was made after using these synthetic datasets to train a 3D segmentation model, where the previously enumerated cardiac structures were detected. Several datasets with different percentages of synthetic images mixed with real ones were created and the volume similarity between the original and the segmented volumes showed that including synthetically generated images in the training datasets improves the final segmentation quality.

Therefore, the results demonstrate the creation of a 3D GAN model and its utility as an automatic tool to perform data augmentation, and also the applicability of such images to train DL algorithms, tackling the current lack of public 3D echocardiography datasets.

2.7.2 Paper II: A Domain Translation Framework with an Adversarial Denoising Diffusion Model to Generate Synthetic Datasets of Echocardiography Images

Cristiana Tiago, Sten Roar Snare, Jurica Sprem, and Kristin McLeod, *IEEE Access*, 2023.

The second paper describes the work we conducted on generating 2D echocardiography images using adversarial DDMs, and using this novel deep generative model to perform medical image domain translation.

As addressed in Section 2.5.1, domain translation is a common DL application in the medical image field. This operation can also be considered as a data augmentation technique, as it is capable of creating new images belonging to different domains but still relevant for clinical application. It also favors time

saving in clinical workflows, since certain patient images can be generated from a set of previously acquired ones.

We used deep generative models as a resource to translate images from one domain to another, exploring the image synthesis capabilities of DDMs when used together with a GAN. We constructed an adversarial DDM which is capable of generating varying and high quality echocardiography images with a quick sampling time. In the proposed generative model, the GAN is responsible for learning the reverse denoising diffusion process working on a paired fashion, where guide images, anatomical masks in this case, are used to guide this learning process. These guide images ensure that the most relevant anatomical structures of each echocardiography image were kept and represented on the generated image samples.

Besides exploring this image augmentation capability of DDMs, several domain translation operations allowed to create different echocardiography images datasets, with different vendors and image acquisition characteristics. The obtained results proved that adversarial DDMs are indeed capable of generating highly variable image samples (in opposition to GANs), keeping the anatomical structures present on the guide image on the synthetic sample. These synthetically generated images also show high quality and the sampling duration is shortened when we use a GAN to learn the reverse diffusion process.

The proposed method showed high generalization ability, introducing a framework to create 2D echocardiography images suitable to be used for clinical research purposes.

2.7.3 Paper III: Denoising Diffusion Model for 3D Echocardiography Image Generation: Image Usability and Clinical Relevance

Cristiana Tiago, Sten Roar Snare, Kristin McLeod, and Jurica Sprem, submitted to *IEEE Open Journal of Engineering in Medicine and Biology*, 2023.

In paper III we combined two deep generative models to synthesise 3D echocardiography images. First, a 3D Vector Quantized (VQ) GAN was trained to capture the features inherent to the training dataset, and a following 3D DDM was used to actually generate the synthetic images, based on the image characteristics previoulsly learned by the VQ-GAN.

3D echocardiography image datasets are challenging to acquire, as explained previously, and getting permission to access medical images of different patients is a complicated process. Medical image synthesis has been proving to be a good asset to tackle these problems, as there are several methods that can generate synthetic medical images with a very high realism level.

Following the work we previously conducted, in this third paper we took another step in generating echocardiography image datasets using deep generative models. Similarly to paper I, we synthesized 3D echocardiography image samples (3 spatial dimensions), using a 3D diffusion model, similarly to paper II. However,

in paper III, the diffusion model is trained on the results obtained from a previously trained 3D VQ-GAN. In this work, the VQ-GAN is trained on a real 3D echocardiography dataset and generates a latent space where the image characteristics are encoded. Then, the 3D DDM attempts to generate realistic and variate 3D echocardiography volumes, based on this VQ-GAN encoded latent space.

After training the generative model, a synthetic dataset was created and it was evaluated by experienced clinicians and sonographers regarding it image realism, anatomical correctness, and frame consistency. The synthetic images were also compared with a synthetic dataset generated via a 3D GAN (the one presented in paper I).

Finally, the results showed that the proposed image synthesis model, 3D VQ-GAN and DDM, is able to generate 3D echocardiography images with a very high realism level and relevant anatomical information. This diffusion model also proved to be better than a 3D GAN at synthesizing such type of image samples. The 3D DDM revealed itself as a good tool to perform variate and realistic medical image synthesis, particularly on 3D echocardiography.

Chapter 3

Discussion

3.1 Data for Deep Learning

To achieve the best results, DL models need to be trained on large amounts of high quality medical images. The quality of training images is critical in DL, as it can significantly impact the model's accuracy and performance. Additionally, most of times, the images should be annotated by expert clinicians to provide accurate and consistent labels for the DL model to learn from (supervised learning). By using high quality training images, the DL model can learn to recognize complex patterns and features in the data, leading to improved performance in various clinical tasks.

The diversity of training images is also important in DL in healthcare, as it ensures that the DL model can generalize and make accurate predictions on new unseen data. The use of a diverse set of training images can help the DL model learn to recognize variations in echocardiography images, such as differences in patient anatomies, imaging acquisition protocols, and disease presentations. Therefore, the selection of training images should include a wide range of medical conditions, patient populations, and image acquisition settings.

For this reason, it is critical to invest in the acquisition and preparation of a representative set of training images to ensure the success of DL in healthcare. Since data holds an important role in DL, the successful findings from paper I support this assumption by proving that synthetic data does indeed improve the performance of DL algorithms. Furthermore, this data can be generated from deep generative models such as GANs and DDMs, as described in all the 3 papers collected in this thesis.

3.2 Medical Image Synthesis and Clinical Relevance of Synthetic Images

Medical image synthesis is a rapidly growing field that has immense potential in clinical applications. It involves generating artificial medical images that closely resemble real ones, using DL techniques [57] such as GANs and, more recently, DDMs. One important aspect of medical image synthesis is ensuring clinical relevance: the generated images must accurately represent the anatomical structures being studied, and should also be realistic enough to aid in the diagnosis and treatment of patients. By producing high quality medical images, this technology has the potential to improve the way new deep learning algorithms are implemented and utilized.

Synthetic medical data generated using deep learning algorithms has several benefits for medical research and healthcare applications. It allows researchers

to generate large datasets with varied characteristics and complexities that can be used to develop and train DL models. This helps overcome the problem of data scarcity that is often faced in medical DL research due to privacy concerns and the difficulty of collecting and annotating large amounts of data. These problems are particularly evident in the echocardiography imaging domain, as there are no public databases of 3D echocardiography datasets and only few exist for 2D images.

Synthetic images can be used to augment existing datasets, thereby improving the accuracy and generalization of DL models. Image synthesis can be used to address the issue of data imbalance that is common in medical datasets. This is because synthetic data can be generated to address specific data gaps, thereby improving the performance of DL models on underrepresented classes.

Creating a data generation tool capable of generate data that is not easily obtainable through traditional methods, is of great utility. For instance, domain translation operations can be used to generate medical images with different modalities or resolutions, which can be used to train DL models for improved image analysis and diagnosis.

Clinicians and sonographers can be trained on these large and variate synthetic datasets, which provide them a diverse range of training scenarios to practice on [58], regardless of their expertise level. Deep generative models proved to be a reliable and efficient tool when it comes to generating realistic looking data, capable of misleading the trained eye of several observers, with clinical relevance. Additionally, synthetic medical images can be manipulated to simulate different diseases or to demonstrate the effects of various treatments in time, allowing clinicians to gain a deeper understanding of medical concepts.

3.3 Synthetic Images Quality

High quality medical images are characterized by their sharpness, contrast, and spatial resolution, which allow clinicians to visualize and accurately diagnose pathological changes in the body. The synthesized images should have minimal noise, artifact, and distortion, ensuring that the anatomical structures are clearly visible and distinguishable from the surrounding tissue. In addition, good quality medical images should also be reproducible, meaning that they can be reliably acquired and interpreted across different image viewing platforms. Overall, good quality medical images play an indirect but crucial role in the diagnosis and treatment of various medical conditions. These widely variate synthetic datasets improve the performance of downstream DL algorithms used to facilitate the final diagnoses and patient outcomes.

The main results of papers II and III support these statements. Both 2D and 3D echocardiography synthetic images proved to have high quality, accurately representing the heart's anatomy and speckle patterns.

3.4 DDMs VS GANs

While GANs are faster and require less computation power, DDMs are more robust to different types of noise, making them better suited for medical image synthesis [59] where image quality and accuracy are of utmost importance.

Both models present viable options to create a data augmentation tool, however a DDM brings more advantages to echocardiography image synthesis, as stated in papers I and III, since the detail and quality of the synthetic images is more relevant than the image sample generation time.

Echocardiography is an essential tool in the diagnosis and treatment of many cardiac conditions, and the quality of medical images is highly relevant. In recent years, diffusion models have emerged as a powerful technique for generating high quality medical images. These models utilize complex mathematical algorithms to model the diffusion of molecules within biological tissue, providing a detailed and accurate representation of the tissue's structure. The result is echocardiography images with exceptional clarity, contrast, spatial and temporal resolution, which can aid in the detection of abnormalities and guide treatment decisions. The quality of the generated image samples with diffusion models has the potential to significantly enhance the accuracy of diagnoses, making them a valuable tool in healthcare.

Figure 3.1: Deep generative models (GAN and DDM) can be used to synthesize high quality echocardiography images, both 2D and 3D. These images can be then used to train and improve DL algorithms.

3.5 Limitations

Despite the promising results presented, there are possible limitations to the work presented through this thesis. Mainly, these include access to real medical data, computational resources, and use of synthetic data to different tasks.

Firstly, to train any DL model a large amount of training data is necessary, and deep generative models are no exception. The quality of data used for training significantly influences the final outcome of the model. It is known that access to large quantities of variable echocardiography images is constrained by the data privacy limitations imposed that apply to the use of real patients' medical data. In the first place, to train the GANs and the DDMs introduced earlier it is necessary to obtain real 2D and 3D echocardiography images, since the trained model is only as good at predicting new images as the images present

in the training dataset. This way, the variability present in training dataset is a limitation to the results possible to obtain.

To extend the application of these models and reduce the effect of this limitation, the development of more generative models is needed, so that the lack of plausible synthetic data can no longer be an issue.

Secondly, to train such generative models with improved levels of reliability, large computational resources are necessary. Using large datasets, dealing with 3D images, and requiring fine tuning of the final models can take long periods of time, with this time decreasing if the computer processors become sufficiently powerful. The access to such resources is, therefore, a limitation to training and obtaining the best generative model as possible. However, attempting different training strategies offers a solution to this.

Finally, another limitation is linked to the downstream application of the synthetic images. The type of images generated by the models might not be suitable for all the possible image analysis applications. As an example, the images synthesized by the model described in Paper II can be used for segmentation of the LV, LA, and myocardium, but they might not be very useful if the structures intended to segment are the mitral valve leaflets or papillary muscles. Thus, it is important to initially define the application case and then generate synthetic images accordingly.

3.6 Future Work

The benefits of synthetic medical data and DL are numerous and varied, as discussed in this section and presented in the research papers included in this thesis. These technologies have the potential to revolutionize medical research and healthcare by providing healthcare professionals with the tools to make more accurate and personalized diagnoses and treatment plans, and by enabling researchers to develop and test new medical technologies and interventions in a safe and ethical manner.

Synthetic medical data can be used to simulate complex medical conditions and procedures, providing healthcare professionals with the opportunity to train and practice in a risk-free environment. This reduces the risk of errors and complications during real medical procedures, which can have serious consequences for patients. This opens the door to generate not only echocardiography images linked to normal subjects but also generate such type of images where different health conditions could be present.

Furthermore, the application of these deep generative models can also be of great utility when it comes to generate different types of images [60], with characteristics linked to other imaging modalities and also different human organs. Future work can aim to create a publicly available large dataset of medical images from different imaging modalities, vendors, organs and health conditions, opening the door to create a large and diverse public repository suitable to be used for research and move healthcare systems forward.

Chapter 4

Conclusion

This thesis is a compilation of three research papers focused on addressing the topic of medical image synthesis using deep generative models. The work presented throughout the thesis proves the relevance of the topic applied to cardiac ultrasound imaging. The work presented in this thesis shows the advances made on creating different deep generative model architectures capable of synthesizing realistic 2D and 3D echocardiography images.

The ability to generate synthetic medical images has significant clinical relevance, particularly in the field of cardiology. Accurate and timely diagnosis of heart disease is critical to ensuring optimal patient outcomes, and synthetic medical images can provide clinicians with valuable insights that may be difficult or impossible to obtain through other means.

4.1 3D Generative Adversarial Network to synthesize echocardiography images and train 3D segmentation models

In conclusion, paper I focuses on extending a two dimensional GAN model to three dimensions and use it to generate different 3D echocardiography volumes, associated with anatomical annotations of the heart structures. These images proved to be clinically relevant and meaningful as they showed to be a good data augmentation resource to train DL models.

4.2 2D Denoising Diffusion Model to synthesize echocardiography images

Secondly, paper II covers the usage of a different and more efficient and accurate generative model. It describes the application of an adversarial diffusion model to generate 2D echocardiography images, which proved its performance by generating realistic images with variate image domain characteristics. The results obtained on medical image synthesis using deep generative models opened up exciting new possibilities for generating synthetic medical images, specifically 2D and 3D echocardiography images. These synthetic images offer a range of benefits, including increased data availability, improved data quality, reduced reliance on sometimes cumbersome imaging protocols, and no privacy issues linked to the data.

4.3 Clinical usability of synthetic echocardiography images generated with a 3D Denoising Diffusion Model

Finally, paper III shows the usability of a 3D DDM to synthesize 3D heart volumes, offering a comparison to the results from the first paper. Furthermore, the use of advanced and novel deep generative models in image synthesis has the potential to revolutionize the field of medical imaging by enabling researchers to generate high quality synthetic images on demand, tailored to the specific needs of study groups. This could have significant implications for the diagnosis and treatment of a wide range of medical conditions.

Overall, however, the potential benefits of image synthesis using deep generative models are enormous [61], and this technology is likely to play an increasingly important role in the field of medical imaging in the years to come. As researchers continue to explore the possibilities of synthetic medical images, it is clear that this field holds enormous promise for improving patient outcomes and advancing our understanding of the human body.

Bibliography

[1] Organization, W. H., *Cardiovascular diseases (CVDs)*, en.

[2] Garcea, F., Serra, A., Lamberti, F., and Morra, L., "Data augmentation for medical imaging: A systematic literature review," *Computers in Biology and Medicine*, vol. 152, p. 106 391, Jan. 2023.

[3] Uzunova, H., Wilms, M., Forkert, N. D., Handels, H., and Ehrhardt, J., "A systematic comparison of generative models for medical images," *International Journal of Computer Assisted Radiology and Surgery*, vol. 17, no. 7, pp. 1213–1224, Jul. 2022.

[4] OpenStax, *19.1 heart anatomy - anatomy and physiology |openstax.*

[5] Association, A. H., *Ejection fraction heart failure measurement.*

[6] Rehman, R., Yelamanchili, V. S., and Makaryus, A. N., "Cardiac imaging," in Treasure Island (FL): StatPearls Publishing, 2022.

[7] Hasegawa, H., "Very high frame rate ultrasound for medical diagnostic imaging," *AIP Conference Proceedings*, vol. 2173, no. 1, Nov. 2019.

[8] Omar, A. M. S., Bansal, M., and Sengupta, P. P., "Advances in echocardiographic imaging in heart failure with reduced and preserved ejection fraction," *Circulation Research*, vol. 119, no. 2, pp. 357–374, Jul. 2016.

[9] Klibanov, A. L. and Hossack, J. A., "Ultrasound in radiology: From anatomic, functional, molecular imaging to drug delivery and image-guided therapy," *Investigative radiology*, vol. 50, no. 9, pp. 657–670, Sep. 2015.

[10] Subramani, S., "Comparison between 2d and 3d echocardiography for quantitative assessment of mitral regurgitation: Current status," *Annals of Cardiac Anaesthesia*, vol. 25, no. 2, pp. 198–199, 2022.

[11] Turton, E. W. and Ender, J., "Role of 3d echocardiography in cardiac surgery: Strengths and limitations," *Current Anesthesiology Reports*, vol. 7, no. 3, pp. 291–298, 2017.

[12] Platts, D. G., Humphries, J., Burstow, D. J., Anderson, B., Forshaw, T., and Scalia, G. M., "The use of computerised simulators for training of transthoracic and transoesophageal echocardiography. the future of echocardiographic training?" *Heart, Lung & Circulation*, vol. 21, no. 5, pp. 267–274, May 2012.

[13] Biswas, M., Patel, R., German, C., *et al.*, "Simulation-based training in echocardiography," *Echocardiography (Mount Kisco, N.Y.)*, vol. 33, no. 10, pp. 1581–1588, Oct. 2016.

[14] Jensen, J. A., "A model for the propagation and scattering of ultrasound in tissue," *The Journal of the Acoustical Society of America*, vol. 89, no. 1, pp. 182–190, Jan. 1991.

[15] Gao, H., Hergum, T., Torp, H., and D'hooge, J., "Comparison of the performance of different tools for fast simulation of ultrasound data," *IEEE Ultrasonics Symposium*, vol. 52, no. 5, pp. 573–577, Jul. 2012.

[16] Gao, H., Choi, H. F., Claus, P., *et al.*, "A fast convolution-based methodology to simulate 2-d/3-d cardiac ultrasound images," *IEEE transactions on ultrasonics, ferroelectrics, and frequency control*, vol. 56, no. 2, pp. 404–409, Feb. 2009.

[17] Hergum, T., Crosby, J., Langhammer, M. J., and Torp, H., "The effect of including fiber orientation in simulated 3d ultrasound images of the heart," ISSN: 1051-0117, Oct. 2006, pp. 1991–1994.

[18] Yao, C., Simpson, J., Schaeffter, T., and Penney, G., "Multi-view 3d echocardiography compounding based on feature consistency," *Physics in medicine and biology*, vol. 56, pp. 6109–28, Sep. 2011.

[19] Votta, E., Caiani, E., Veronesi, F., Soncini, M., Montevecchi, F., and Redaelli, A., "Mitral valve finite-element modelling from ultrasound data: A pilot study for a new approach to understand mitral function and clinical scenarios," *Philosophical transactions. Series A, Mathematical, physical, and engineering sciences*, vol. 366, pp. 3411–34, Jul. 2008.

[20] Verhey, J. F., Nathan, N. S., Rienhoff, O., Kikinis, R., Rakebrandt, F., and D'Ambra, M. N., "Finite-element-method (fem) model generation of time-resolved 3d echocardiographic geometry data for mitral-valve volumetry," *Biomedical Engineering Online*, vol. 5, p. 17, Mar. 2006.

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

[22] Gandhi, V. C. and Gandhi, P. P., "A survey - insights of ml and dl in health domain," in *2022 International Conference on Sustainable Computing and Data Communication Systems (ICSCDS)*, Apr. 2022, pp. 239–246.

[23] Scheetz, J., Rothschild, P., McGuinness, M., *et al.*, "A survey of clinicians on the use of artificial intelligence in ophthalmology, dermatology, radiology and radiation oncology," *Scientific Reports*, vol. 11, no. 1, p. 5193, Mar. 2021.

[24] Asch, F. M., Poilvert, N., Abraham, T., *et al.*, "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, no. 9, Sep. 2019.

[25] Shorten, C. and Khoshgoftaar, T. M., "A survey on image data augmentation for deep learning," *Journal of Big Data*, vol. 6, no. 1, Jul. 2019.

[26] Zhang, D., Lin, Y., Chen, H., *et al.*, *Deep learning for medical image segmentation: Tricks, challenges and future directions*, arXiv:2209.10307 [cs], Sep. 2022.

[27] Kaji, S. and Kida, S., "Overview of image-to-image translation by use of deep neural networks: Denoising, super-resolution, modality conversion, and reconstruction in medical imaging," *Radiological Physics and Technology*, vol. 12, no. 3, pp. 235–248, Sep. 2019.

[28] Uzunova, H., Ehrhardt, J., and Handels, H., "Memory-efficient gan-based domain translation of high resolution 3d medical images," *Computerized Medical Imaging and Graphics*, vol. 86, Dec. 2020.

[29] Yang, Q., Yan, P., Zhang, Y., *et al.*, "Low-dose ct image denoising using a generative adversarial network with wasserstein distance and perceptual loss," *IEEE Transactions on Medical Imaging*, vol. 37, no. 6, pp. 1348–1357, Jun. 2018.

[30] Güngör, A., Askin, B., Soydan, D. A., Saritas, E. U., Top, C. B., and Çukur, T., "Transms: Transformers for super-resolution calibration in magnetic particle imaging," *IEEE Transactions on Medical Imaging*, vol. 41, no. 12, pp. 3562–3574, Dec. 2022.

[31] Lustermans, D. R. P. R. M., Amirrajab, S., Veta, M., Breeuwer, M., and Scannell, C. M., "Optimized automated cardiac mr scar quantification with gan-based data augmentation," *arXiv:2109.12940 [cs, eess]*, Sep. 2021.

[32] Isola, P., Zhu, J.-Y., Zhou, T., and Efros, A. A., "Image-to-image translation with conditional adversarial networks," *arXiv:1611.07004 [cs]*, Nov. 2018.

[33] Huo, Y., Xu, Z., Moon, H., *et al.*, "Synseg-net: Synthetic segmentation without target modality ground truth," *IEEE Transactions on Medical Imaging*, vol. 38, no. 4, pp. 1016–1025, Apr. 2019.

[34] Goodfellow, I., Pouget-Abadie, J., Mirza, M., *et al.*, "Generative adversarial networks," *Communications of the ACM*, vol. 63, no. 11, pp. 139–144, Oct. 2020.

[35] Arjovsky, M. and Bottou, L., *Towards principled methods for training generative adversarial networks*, arXiv:1701.04862 [cs, stat], Jan. 2017.

[36] Mirza, M. and Osindero, S., *Conditional generative adversarial nets*, arXiv:1411.1784 [cs, stat], Nov. 2014.

[37] Abu-Srhan, A., Almallahi, I., Abushariah, M. A. M., Mahafza, W., and Al-Kadi, O. S., "Paired-unpaired unsupervised attention guided gan with transfer learning for bidirectional brain mr-ct synthesis," *Computers in Biology and Medicine*, vol. 136, p. 104 763, Sep. 2021.

[38] Zhu, J.-Y., Park, T., Isola, P., and Efros, A. A., "Unpaired image-to-image translation using cycle-consistent adversarial networks," *arXiv:1703.10593 [cs]*, Aug. 2020.

[39] Kingma, D. P. and Welling, M., *Auto-encoding variational bayes*, arXiv:1312.6114 [cs, stat], Dec. 2022.

[40] Esser, P., Rombach, R., and Ommer, B., *Taming transformers for high-resolution image synthesis*, Jun. 2021.

[41] Sohl-Dickstein, J., Weiss, E. A., Maheswaranathan, N., and Ganguli, S., *Deep unsupervised learning using nonequilibrium thermodynamics*, arXiv:1503.03585 [cond-mat, q-bio, stat], Nov. 2015.

[42] Nichol, A. and Dhariwal, P., *Improved denoising diffusion probabilistic models*, arXiv:2102.09672 [cs, stat], Feb. 2021.

[43] Dhariwal, P. and Nichol, A., *Diffusion models beat gans on image synthesis*, arXiv:2105.05233 [cs, stat], Jun. 2021.

[44] Dodge, Y., *The Oxford Dictionary of Statistical Terms*. Oxford University Press, 2003.

[45] Ho, J., Jain, A., and Abbeel, P., "Denoising diffusion probabilistic models," in *Proceedings of the 34th International Conference on Neural Information Processing Systems*, ser. NIPS'20, Red Hook, NY, USA: Curran Associates Inc., Dec. 2020, pp. 6840–6851.

[46] Xiao, Z., Kreis, K., and Vahdat, A., *Tackling the generative learning trilemma with denoising diffusion gans*, arXiv:2112.07804 [cs, stat], Apr. 2022.

[47] Nichol, A., Dhariwal, P., Ramesh, A., *et al.*, *Glide: Towards photorealistic image generation and editing with text-guided diffusion models*, arXiv:2112.10741 [cs], Mar. 2022.

[48] Karagiannakos, S., *Deep Learning in Production*. Leanpub, Nov. 2021.

[49] Özbey, M., Dalmaz, O., Dar, S. U., *et al.*, *Unsupervised medical image translation with adversarial diffusion models*, arXiv:2207.08208 [cs, eess], Oct. 2022.

[50] Kazeminia, S., Baur, C., Kuijper, A., *et al.*, *Gans for medical image analysis*, arXiv:1809.06222 [cs, stat], Oct. 2019.

[51] Gilbert, A., Marciniak, M., Rodero, C., Lamata, P., Samset, E., and Mcleod, K., "Generating synthetic labeled data from existing anatomical models: An example with echocardiography segmentation," *IEEE Transactions on Medical Imaging*, vol. 40, no. 10, pp. 2783–2794, Oct. 2021.

[52] Güngör, A., Dar, S. U., Öztürk, Ş., *et al.*, *Adaptive diffusion priors for accelerated mri reconstruction*, arXiv:2207.05876 [cs, eess], Nov. 2022.

[53] Pinaya, W. H. L., Graham, M. S., Gray, R., *et al.*, *Fast unsupervised brain anomaly detection and segmentation with diffusion models*, arXiv:2206.03461 [cs, eess, q-bio], Jun. 2022.

[54] Pinaya, W. H. L., Tudosiu, P.-D., Dafflon, J., *et al.*, "Brain imaging generation with latent diffusion models," Mukhopadhyay, A., Oksuz, I., Engelhardt, S., Zhu, D., and Yuan, Y., Eds., ser. Lecture Notes in Computer Science, Springer Nature Switzerland, 2022, pp. 117–126.

[55] Dorjsembe, Z., Odonchimed, S., and Xiao, F., "Three-dimensional medical image synthesis with denoising diffusion probabilistic models," Jun. 2022.

[56] Kim, B. and Ye, J. C., "Diffusion deformable model for 4d temporal medical image generation," in *Medical Image Computing and Computer Assisted Intervention – MICCAI 2022*, Wang, L., Dou, Q., Fletcher, P. T., Speidel, S., and Li, S., Eds., ser. Lecture Notes in Computer Science, Springer Nature Switzerland, 2022, pp. 539–548.

[57] Shokraei Fard, A., Reutens, D. C., and Vegh, V., "From cnns to gans for cross-modality medical image estimation," *Computers in Biology and Medicine*, vol. 146, p. 105 556, Jul. 2022.

[58] Duong, M. T., Rauschecker, A. M., Rudie, J. D., *et al.*, "Artificial intelligence for precision education in radiology," *The British Journal of Radiology*, vol. 92, no. 1103, Nov. 2019.

[59] Khader, F., Mueller-Franzes, G., Arasteh, S. T., *et al.*, *Medical diffusion: Denoising diffusion probabilistic models for 3d medical image generation*, arXiv:2211.03364 [cs, eess], Jan. 2023.

[60] Yu, B., Wang, Y., Wang, L., Shen, D., and Zhou, L., "Medical image synthesis via deep learning," *Advances in Experimental Medicine and Biology*, vol. 1213, pp. 23–44, 2020.

[61] Svoboda, D. and Burgos, N., "Chapter 1 - introduction to medical and biomedical image synthesis," in *Biomedical Image Synthesis and Simulation*, ser. The MICCAI Society book Series, Burgos, N. and Svoboda, D., Eds., Academic Press, Jan. 2022, pp. 1–3.

Papers

Paper I

A Data Augmentation Pipeline to Generate Synthetic Labeled Datasets of 3D Echocardiography Images Using a GAN

I

Cristiana Tiago, Andrew Gilbert, Ahmed Salem Beela, Svein Arne Aase, Sten Roar Snare, Jurica Sprem, Kristin McLeod

Published in *IEEE Access*, 16 September 2022, volume 10, pp. 98803–98815.
DOI: 10.1109/ACCESS.2022.3207177.

Abstract

Due to privacy issues and limited amount of publicly available labeled datasets in the domain of medical imaging, we propose an image generation pipeline to synthesize 3D echocardiographic images with corresponding ground truth labels, to alleviate the need for data collection and for laborious and error-prone human labeling of images for subsequent Deep Learning (DL) tasks. The proposed method utilizes detailed anatomical segmentations of the heart as ground truth label sources. This initial dataset is combined with a second dataset made up of real 3D echocardiographic images to train a Generative Adversarial Network (GAN) to synthesize realistic 3D cardiovascular Ultrasound images paired with ground truth labels. To generate the synthetic 3D dataset, the trained GAN uses high resolution anatomical models from Computed Tomography (CT) as input. A qualitative analysis of the synthesized images showed that the main structures of the heart are well delineated and closely follow the labels obtained from the anatomical models. To assess the usability of these synthetic images for DL tasks, segmentation algorithms were trained to delineate the left ventricle, left atrium, and myocardium. A quantitative analysis of the 3D segmentations given by the models trained with the synthetic images indicated the potential use of this GAN approach to generate 3D synthetic data, use the data to train DL models for different clinical tasks, and therefore tackle the problem of scarcity of 3D labeled echocardiography datasets.

This work was supported by the European Union's Horizon 2020 Research and Innovation Programme through the Marie-Skłodowska Curie Grant under Agreement 860745.
The code is publicly available in: https://github.com/CristianaTiago/3D-echo-generation

Contents

I.1 Introduction

Medical imaging plays a crucial role in optimizing treatment pathways. Saving time when it comes to diagnosis and treatment planning enables the clinicians to focus on more complicated cases.

Many modalities are used to image the heart, such as Computed Tomography (CT), Magnetic Resonance (MR), and Ultrasound imaging, enabling several structural and functional parameters related to the organ's performance to be estimated. Such parameters are the basis of clinical guidelines for diagnosis and treatment planning.

Echocardiography is the specific use of ultrasound to image the heart. This imaging modality is widely used given its advantages of being portable, relatively low-cost, and the fact that the use of ionizing radiation is not required.

Deep Learning (DL), and specifically Convolutional Neural Networks (CNNs), have become extensively applied in medical image analysis because they facilitate the automation of many tedious clinical tasks and workflows such as estimation of ejection fraction, for example. These algorithms are capable of approaching human-level performance [1], thus potentially saving clinicians' time without decreasing the quality of care for patients. In fact, clinicians agree that using DL algorithms in the clinical workflow also improves patient access to disease diagnoses, increasing the final diagnosis confidence levels [2]. DL models can be developed to perform numerous medical tasks such as image classification, segmentation and even region/structure detection [3].

Echocardiography images can be acquired both in 2D and 3D. Time can also be taken into account, generating videos. 3D echocardiography images can be more difficult to assess than 2D images. However, for some specific application cases, 3D image acquisition brings great advantages since it can offer more accurate and reproducible measurements. One such case is ventricle and atrium volumes [4]. Amongst the causes of lack of annotated 3D echocardiography datasets are the higher complexity to acquire 3D echocardiography images and the fact that 3D is still not part of all echocardiography routine exams. Also, even when 3D images are recorded, delineating the structures in them is a challenging, time consuming, and user dependent task. Taken together and adding the fact that privacy regulations to access medical data are becoming stricter, these can explain why there is a lack of publicly available datasets of such type of images. Therefore, having an approach able to address this image scarcity is necessary. This current lack of 3D medical data and the great need of high

quality annotated data required by the DL models impacts the development of such algorithms and therefore the scientific and technological development of the 3D medical imaging field. Synthetic generation of labeled 3D echocardiography images is a DL based approach that provides a solution for this problem.

Synthetic data can help in the development of DL models for image analysis [5] and accurate labeling of these images. Furthermore, this approach works as a data augmentation strategy by generating additional data. It is known that creating datasets with a combination of real and synthetic images and use them to train algorithms that tackle medical challenges represents a successful solution to the image scarcity [6] problem. Such type of synthetic images even increase the heterogeneity present on these datasets, facilitating a more efficient performance of the trained models as they are exposed to a larger variety of images.

Generative Adversarial Networks (GANs) are specific DL architectures that create models capable of generating medical images closely resembling real images acquired from patients. These deep generative models rely on a generator and a discriminator. While the straightforward GAN discriminator distinguishes between real and fake, i.e., generated, images, the generator not only attempts to deceive the discriminator but also tries to minimize the difference between the generated image and the ground truth.

The generated synthetic images can even be associated with labels facilitating the acquisition of large labeled datasets, eliminating the need for manual annotation, and therefore the variabilities associated with the observer [7], which largely influences the final output [8]. 3D heart models are a great source of anatomical labels since they capture accurate information about the organ's structures [9]. Different types of models can be used for this purpose, such as animated models, biophysical models, or even anatomical models obtained from different imaging modalities [10], [11]. Recently, CT models were used as label sources to generate 2D echocardiography [12] and cardiac MR images [13], proving the utility of GANs for this task.

Developing a pipeline to generate synthetic data using GANs to create labeled datasets addresses the immense need for the large volume of data that DL algorithms require during training to perform an image analysis task, eliminates the need to acquire the images from subjects, and saves time of experienced professionals when annotating them, as the anatomical labels can be extracted from anatomical models. Usually, when developing such generative models, imaging artifacts are present and visible on the synthetically generated images. This widely common GAN performance drawback is addressed by applying some image post-processing operations [14] on the synthetically generated 3D echocardiography images.

In practice synthetic images can be used to train DL models because they represent a good data augmentation strategy [15]. For instance, 3D medical image segmentation is the most common example of a medical task to which DL can turn out to be a good application. Labeled datasets made of real images combined with synthetic ones, which even include the respective anatomical labels, become the training dataset for 3D DL models, addressing the problem

of sparse 3D medical data availability [16].

I.1.1 State of the Art

DL has become widely used in medical imaging due to its potential in image segmentation, classification, reconstruction, and synthesis across all imaging modalities. Image synthesis has been a research topic for a few decades now, where some of the more conventional approaches use human-defined rules and assumptions like shape priors, for example [17]. Also, these image synthesis techniques depend on the imaging modality being considered to perform certain tasks. To tackle these shortcomings, CNNs are now becoming a widely used approach for image synthesis across many medical imaging modalities.

Many reasons motivate medical image generation, both 2D and 3D. Generative algorithms can perform domain translation, with a large applicability when converting images from one imaging modality to a different one, as Uzunova *et al.* [18] showed in their work converting 3D MR and CT brain images. GANs can also be used to generate a ground truth for a given input, as these DL models can be trained in a cyclic way, as is the case of the CycleGAN [19], for example. Additionally, generation of synthetic data used for DL algorithms also motivates the application and development of GAN architectures. Several research groups were able to generate medical images using this methodology as a data augmentation tool, even though most of them were developed under a 2D scenario and focused on a few imaging modalities, mainly MRI and CT. These imaging modalities raise less challenges when compared with Ultrasound due to the nature of the physics behind the acquisition process.

Ultrasound images have an inherent and characteristic speckle pattern and their quality is largely influenced by the scanner, the sonographer, and the patient anatomy. When it comes to generating 3D Ultrasound images a few more challenges arise, with the speckle pattern having to be consistent throughout the whole volume being the main one. The anatomical information present in the generated volume also has to hold this consistency feature.

Huo *et al.* [20] trained a 2D GAN model, SynSegNet, on CT images and unpaired MR labels using a CycleGAN. Similarly, Gilbert *et al.* [12] proposed an approach to synthesize labeled 2D echocardiography images, using anatomical models and a CycleGAN as well. The CycleGAN was proposed by Zhu *et al.* [19] and works under an unpaired scenario: the images from one training domain do not have to be related with the images belonging to the other domain. This GAN learns how to map the images from one to another and vice-versa. The paired version of this GAN is called Pix2pix. Isola *et al.* [21] proposed this image synthesis method which generates images from one domain to the other, and vice-versa, however the images belonging to the training domains are paired.

As mentioned, 3D echocardiographic data is sparser, but these images can be generated using GANs, and then used to train new algorithms. Both Gilbert *et al.* [12] and Amirrajab *et al.* [22] investigated the potential use of GAN synthesized datasets to train CNNs to segment different cardiac structures on different imaging modalities, but these methods were limited to 2D.

Hu *et al.* [23] attempted to generate 2D fetal Ultrasound scan images at certain 3D spatial locations. They concluded that common GAN training problems such as mode collapse occur. Abbasi-Sureshjani *et al.* [24] developed a method to generate 3D labeled Cardiac MR images relying on CT anatomical models to obtain labels for the synthesized images, using a SPADE GAN [25]. More recently, Cirillo *et al.* [26] adapted the original Pix2pix model to generate 3D brain tumor segmentations.

When dealing with medical images, U-Net [27] is a widely used CNN model to perform image segmentation, for example, since it provides accurate delineation of several structures on these images. More recently, Isensee *et al.* [28] proposed nnU-Net ("no new net"), which automatically adapts to any new datasets and enables accurate segmentations. nnU-Net can be trained on a 3D scenario and optimizes its performance to new unseen datasets and different segmentation tasks, requiring no human intervention.

Existing work to address the challenges of automatic image recognition, segmentation, and tracking in echocardiography has been mostly focused on 2D imaging. In particular, recent work indicates the potential for applying DL approaches to accurately perform measurements in echocardiography images. Alsharqi *et al.* [29] and Østvik *et al.* [30] used a DL algorithm to segment the myocardium in 2D echocardiographic images, from which the regional motion, and from this the strain, were measured. They showed that motion estimation using CNNs is applicable to echocardiography, even when the networks are trained with synthetic data. This work supports the hypothesis that similar approaches could also work for 3D synthetic data.

A large amount of work has been carried out on medical imaging generation and it still represents a challenge for the research community. To the best of our knowledge, the challenge of synthesizing 3D echocardiography images using GANs did not produce any reproducible results, therefore we propose a framework able to address this need.

I.1.2 Contributions

We propose an approach for synthesizing 3D echocardiography images paired with corresponding anatomical labels suitable as input for training DL image analysis tasks. Thus, the main contributions of the proposed pipeline beyond the state of the art include:

1. The extension of Gilbert *et al.* [12] work from 2D to 3D, adapting it from an unpaired to a paired framework (3D Pix2pix) and proposing an automatic pipeline to generate any number of 3D echocardiography images, tackling the lack of public 3D echocardiography datasets and corresponding labels.

2. The creation of a blueprint of heart models and post-processing methods for optimal generation of 3D synthetic data, creating a generic data augmentation tool, this way addressing the lack of 3D data generation works in echocardiography, since it significantly varies from 2D.

3. The demonstration of the usability of these synthetic datasets for training segmentation models that achieve high performance when applied to real images.

I.2 Methodology

The proposed pipeline is summarized in Fig. 1 and described in the following sections. Section II-A describes the preprocessing stage of annotation of the GAN training images to create anatomical labels for these. The training and inference stages are addressed in Section II-B describing how the GAN model was trained and used to synthesize 3D echocardiography images from CT-based anatomical labels and how different post-processing approaches, as described in Section II-C, were applied to these synthetic images. Next, on Section II-D, details regarding the creation of several synthetic datasets used to train 3D segmentation models are given, followed by Section II-E where the influence of adding real images to the synthetic datasets to train segmentation models is assessed.

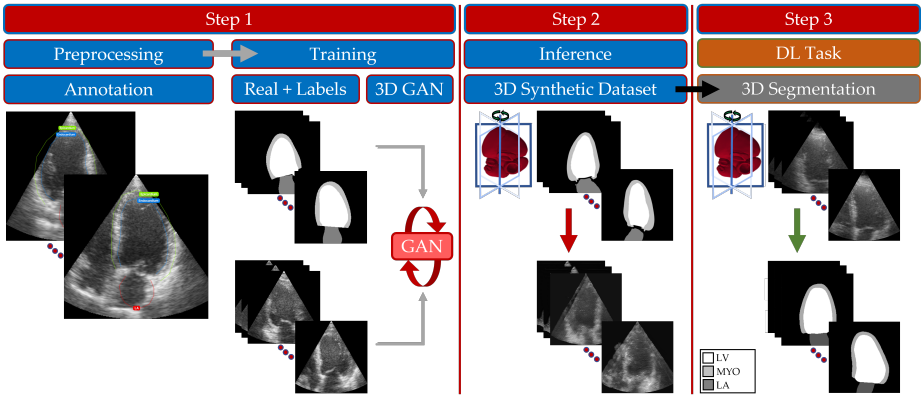

Figure I.1: 3D echocardiography image generation pipeline and inference results. Step 1: during the preprocessing stage, a set of 15 3D heart volumes were labeled by a cardiologist and anatomical labels for the LV, LA and MYO were generated. To train the 3D Pix2pix GAN model, the anatomical labels are paired together with the corresponding real 3D images. Step 2: at inference time, the GAN model generates one 3D image. An example obtained during this stage is shown. The proposed method is able to generate physiologically realistic images, giving correct structural features and image details. Step 3: to show the utility of the synthetic datasets, 3D segmentation models were trained using these GAN generated images (black arrow), but other DL tasks can be addressed.

I.2.1 Data Collection

To train the 3D image synthetization model, an annotated dataset was needed since this GAN set up works under a supervised scenario where two sets of images are used for training: a set containing real 3D echocardiography images and a second set with the correspondent anatomical labels manually performed by a cardiologist (see Fig.1, training stage).

To create the dataset of real 3D echocardiography images, these were acquired during one time point of the cardiac cycle of normal subjects, end-diastole in this work, when the left ventricle (LV) volume is at its largest value. Using GE Vivid Ultrasound scanners 15 heart volumes were acquired. The second set of images was made up of the anatomical labels corresponding to each of the 3D real images included in the set previously described. Each anatomical label image contains the label for the LV, left atrium (LA), and the myocardium (MYO).

To annotate the 3D echocardiography images a certified member of the American National Board of Echocardiography cardiologist, with more than 10 years of experience, used the V7 annotation tool [31] and contoured the three aforementioned structures (Fig. 1, preprocessing stage) on each of the volumes. These contours were then post-processed, applying a spline function to the contour points and resampling it, in order to generate gray scale labeled images. All the 3D images present on each training dataset were sized to $256 \times 256 \times 32$.

I.2.2 3D GAN Training

The Pix2pix model was proposed by Isola *et al.* [21] as a solution to image-to-image translation across different imaging domains. This model is capable of generating an output image for each input image by learning a cyclic mapping function across both training domains. The Pix2pix model works as a conditional paired GAN: given two training domains containing paired images, it learns how to generate new instances of each domain. The loss function was kept the same as presented in the original work – a combination of conditional GAN loss and the L1 distance. This way it is conditioning the GAN performance, assuring the information on the generated output image matches the information provided by the input.

This original work was constructed under a 2D scenario, but in this proposed work an extension to 3D was performed by changing the original architecture of the Pix2pix model.

We considered different architectures for the GAN generator and a 3D U-Net [32] was used to create a 3D version of the Pix2pix model. The discriminator architecture was kept the same, replacing only 2D layers with the correspondent 3D ones. During training of the GAN, data augmentation operations, including blurring and rotation, were performed on the fly, increasing the amount of 3D volumes used without the memory burden of having to save these. The 3D Pix2pix model used here was built using PyTorch [33] and its training was

performed over 200 epochs accounting for the images size and computational memory constraints, considering an initial learning rate of 0.0002 and the Adam optimizer.

At inference time, a common problem among image synthesis is the presence of checkerboard artifacts on the generated images. To tackle this problem, which decreases the quality of the synthesized images, we changed the generator architecture as suggested in [34] by replacing the transposed convolutions in the upsampling layers of the 3D U-Net with linear upsampling ones.

In order to generate synthetic echocardiography images for each of the inference cases, i.e., 3D CT-based heart models, the generator part of the GAN, which translates images from the anatomical labels domain to the echocardiography looking images domain, was used. Anatomical models of the heart [35] obtained from CT were used to create the inference gray scale labeled images, containing anatomical information about the LV, LA, and MYO. The main objective of this work was then accomplished by using the GAN as a data augmentation tool to generate synthetic datasets of 3D echocardiography images of size $256 \times 256 \times 32$ from these inference images, augmenting the quantity of 3D echocardiographic image data.

I.2.3 Synthetic Data Post-processing

During the post-processing stage of the synthetic images generated by the GAN, two different algorithms were experimented. The synthesized images were (a) filtered using the discrete wavelet transform, following Yadav *et al.* [36] work and (b) masked with an Ultrasound cone. The wavelet denoising operation uses wavelets that localize features in the data, preserving important image features while removing unwanted noise, such as checkerboard artifacts. An image mask representing the Ultrasound cone shape was applied to all synthesized images in order to match true Ultrasound data.

I.2.4 3D Segmentation

The GAN pipeline was able to generate labeled instances of 3D echocardiography images, as the model is capable of performing paired domain translation operations. To investigate the utility of the synthetic images, four 3D segmentation models were trained using the generated synthetic images as training set.

The trained model architecture for the 3D segmentation task was the 3D nnU-Net [28]. This network architecture was proposed as a self-adapting framework for medical image segmentation. This DL model adapts its training scheme, such as the loss function or slight variations on the model architecture, to the dataset being used and to the segmentation task being performed. It automates necessary adaptations to the dataset such as preprocessing, patch and batch size, and inference settings without the need of user intervention.

To train the first of four 3D segmentation models, $M_{Synthetic}$, described in this section, a labeled dataset made of 27 synthetically generated 3D echocardiography

images ($256 \times 256 \times 32$), $D_{Synthetic}$, was used. This dataset was obtained from the proposed 3D GAN pipeline at inference time, using anatomical labels from 27 CT 3D anatomical models.

To evaluate the effect of the post-processing operations on the synthesized images, three other datasets were created — $D_{Wavelet}$, D_{Cone}, and $D_{WaveletCone}$ —- and three additional segmentation models were trained using these – $M_{Wavelet}$, M_{Cone}, and $M_{WaveletCone}$, respectively (Fig. 2). $D_{Wavelet}$ was made of the original synthetic images from the $D_{Synthetic}$ dataset but where the wavelet denoising post-processing algorithm was applied, and D_{Cone}, was composed by the original synthetic images with the cone reshaping post-processing operation. Finally, a fourth dataset where both post-processing transformations — wavelet denoising and cone reshaping — were applied to the original synthetic images, $D_{WaveletCone}$, was created. All four datasets contained 27 3D echocardiography images with corresponding anatomical labels for the LV, LA and MYO.

All four 3D segmentation models, $M_{Synthetic}$, $M_{Wavelet}$, M_{Cone}, and $M_{WaveletCone}$, using nnU-Net, were trained on a 5-fold cross validation scenario during 800 epochs. The initial learning rate was 0.01 and the segmentation models were also built using PyTorch [33]. The loss function was a combination of dice and cross-entropy losses, as described in the original work by Isensee *et al.* [28].

Dice scores were used to assess the quality of the segmentations. This score measures the overlap between the predicted segmentation and the ground truth label extracted from the CT anatomical models. For each segmented structure the Dice score obtained at validation time is a value between 0 and 1, where the latter represents a perfect overlap between the prediction and the ground truth.

I.2.5 Real Data Combined with Synthetic – Data Augmentation

In their work, Lustermans *et al.* [16] showed that adding real data to GAN-generated synthetic datasets can help improve DL models train.

To facilitate a clearer analysis of the influence of using synthetic data to train DL models and the utility of this GAN as a data augmentation tool, three other segmentation models were trained on the datasets D_{Real}, $D_{17Real10Augmented}$, and $D_{17Real20Augmented}$. D_{Real} contained 17 real 3D echocardiography volumes acquired with GE Vivid Ultrasound scanners and labeled by a cardiologist.

$D_{17Real10Augmented}$ and $D_{17Real20Augmented}$ were made up of the same 17 real volumes just described together with 10 and 20 synthetic GAN-generated 3D echocardiography images, respectively. Thus allowing to assess the influence of using such type of images during DL models training (Fig. 2).

The 3D segmentation models trained on these datasets were M_{Real}, $M_{17Real10Augmented}$, and $M_{17Real20Augmented}$, respectively. All models used the nnU-Net architecture implemented with Pytorch. Similar to the ones described on Section II-D, they were trained for 800 epochs on a 5-fold cross validation scenario, with the same learning rate and loss function.

At inference time, a test set including real 3D echocardiography images was segmented by the three aforementioned models. To compare the segmentation

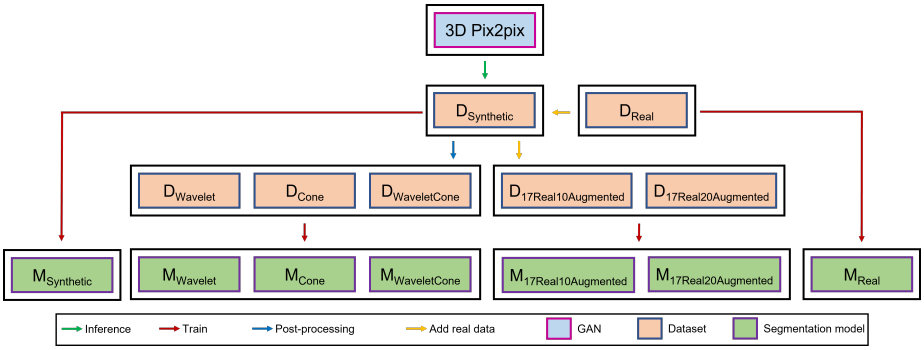

Figure I.2: Overview of all the created datasets and trained models in this work. The generative model, 3D Pix2pix, was trained in order to be used to generate synthetic 3D echocardiography datasets. This dataset, $D_{Synthetic}$, was post-processed applying different transformations and 3 other datasets were created —- $D_{Wavelet}$, D_{Cone}, and $D_{WaveletCone}$. A fifth dataset completely made of real images, D_{Real}, was created and to it, synthetic images from $D_{Synthetic}$ were added creating $D_{17Real10Augmented}$ and $D_{17Real20Augmented}$. All these 7 datasets were used to train 7 3D segmentation models —- $M_{Synthetic}$, $M_{Wavelet}$, M_{Cone}, $M_{WaveletCone}$, M_{Real}, $M_{17Real10Augmented}$, and $M_{17Real20Augmented}$.

results with the ones obtained from a cardiologist, Dice scores and Volume Similarity (VS) were calculated and used as comparison metrics. VS is calculated as the size of the segmented structures and is of high relevance in a 3D scenario since Dice score presents some limitations. Similarly to the Dice score, this evaluation metric takes values between 0 and 1 but is not overlap-based. Instead, it is a volume based parameter where the absolute volume of a region in one segmentation is compared with the corresponding region volume in the other segmentation [37].

I.3 Results

This work's results are presented as follows: Section III-A focuses on the GAN training, architectural modifications performed on the 3D Pix2pix model and their influence on the synthesized images. In Section III-B the influence of post-processing the synthetic images is shown. Finally, Sections III-C and III-D show the segmentation predictions from several models trained on different 3D echocardiography datasets (Fig. 2), as described in Sections II-C and II-D.

I.3.1 GAN Architecture and Training

The chosen GAN architecture influenced the final results. 3D U-Net was chosen as the generator architecture due to its good performance in the medical image

domain. The model was trained on a NVIDIA GeForce RTX 2080 Ti GPU and training took five days.

After applying the architectural changes described in Section II-B to remove the checkerboard artifacts, it seemed like these became less visible or even disappeared. However, this correction created some unwanted blurring on the generated images (Fig. 3), therefore the deconvolution layers were used instead of upsampling, and the synthesized images were post-processed to remove the checkerboard artifacts.

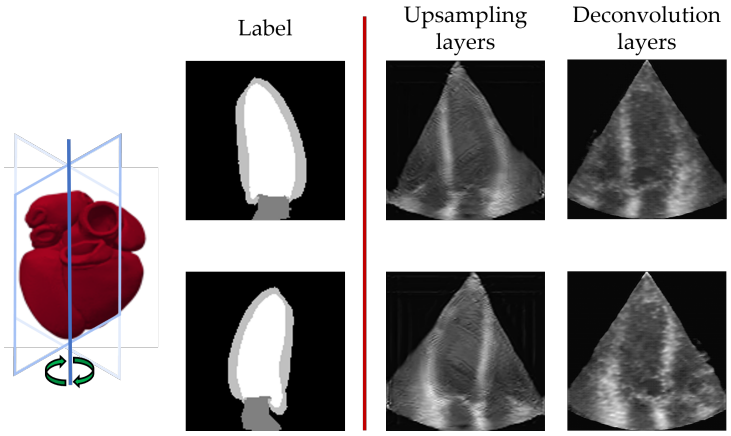

Figure I.3: Influence of architectural changes on the GAN generator to remove checkerboard artifacts. At inference time, a 3D anatomical model was used to extract the anatomical labels. The first column shows 2 different slices of this volume at different rotation angles. The middle column shows that synthesizing images using a GAN with upsampling layers smoothens the checkerboard artifacts but introduces blurring, which is not visible on the images when using a GAN with deconvolution layers (right column). Deconvolution layers are preferred to upsampling ones.

I.3.2 Synthetic Data Post-processing

After training the 3D GAN model and generating synthetic images corresponding to the input anatomical models, as described in Section II-C, the obtained 3D echocardiography images were post-processed in order to remove the aforementioned checkerboard artifacts.

The cone edges were slightly wavy in some cases and checkerboard artifacts were sometimes present. The post-processing experiment, where different transformations were applied to the synthesized images, showed that applying these can give a more realistic aspect to the GAN-generated images, ensuring that the anatomical information remained intact.

Performing these operations allowed to give a more realistic look to the generated echocardiography images (Fig. 4).

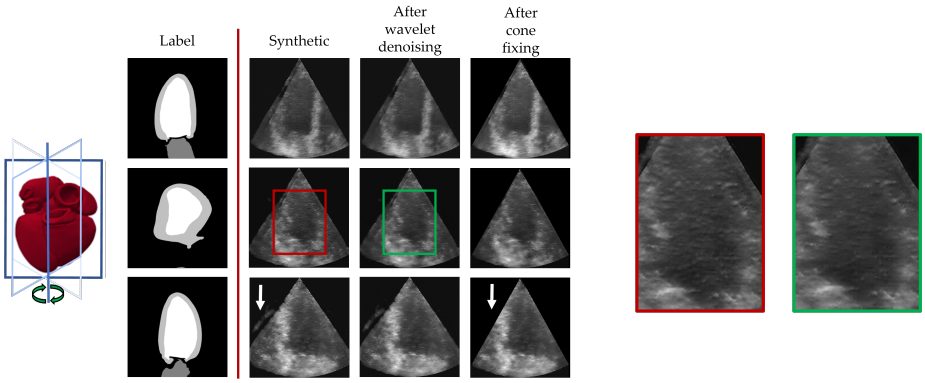

Figure I.4: 3D Pix2pix model inference results and post-processing step. At inference time, the anatomical labels were extracted from a 3D heart model. The first column shows 3 different rotation planes of this volume at different rotation angles. After generating the correspondent synthetic ultrasound image (second column) for this inference case, it was post-processed applying a wavelet denoising transformation to eliminate the checkerboard artifacts (third column) and also a cone reshaping step to smooth the wavy edges of the ultrasound cone (fourth column). Post-processing operations give a more realistic look to the synthesized images as indicated by the enlarged areas framed in red and green (wavelet denoise) and the white arrows (cone reshape).

I.3.3 Segmentation from Synthetic Datasets

Anatomical models were used in order to synthesize 27 3D echocardiography images. These were then used to create the synthetic datasets that were used to train 3D segmentation algorithms, as described in section II-D. Post-processing operations were performed on these images to create the $D_{Wavelet}$, D_{Cone}, and $D_{WaveletCone}$ datasets. Table 1 shows the average Dice scores (average $\pm$ standard deviation) of each segmented structure (LV, LA, and MYO) for each trained model — $M_{Synthetic}$, $M_{Wavelet}$, M_{Cone}, and $M_{WaveletCone}$, obtained from the validation dataset. Training took around five days for each fold using a NVIDIA GeForce RTX 2080 GPU, for all epochs. The complete table with all the Dice scores obtained for each training fold of each model can be found in Appendix — Table 5.

Adding to the Dice scores and to sustain the usability of synthetic images to train segmentation algorithms, Fig. 5 shows the 3D segmentation for an inference 3D echocardiography image acquired from a real subject. Each trained segmentation model was tested on real cases, at inference time.

Table I.1: Average validation dice scores (average ± standard deviation) of each segmented structure (LV, LA, and MYO) for each trained model on completely synthetic datasets – $M_{Synthetic}$, $M_{Wavelet}$, M_{Cone}, and $M_{WaveletCone}$. The best scores are highlighted.

	Models			
	$M_{Synthetic}$	$M_{Wavelet}$	M_{Cone}	$M_{WaveletCone}$
LV	0.926 ± 0.006	**0.927 ± 0.005**	0.926 ± 0.006	0.924 ± 0.008
LA	**0.818 ± 0.011**	0.816 ± 0.010	0.816 ± 0.021	0.814 ± 0.016
MYO	**0.808 ± 0.016**	0.808 ± 0.017	0.803 ± 0.018	0.801 ± 0.023

I.3.4 Segmentation from Combined Datasets

In Table 2 one can see the average Dice scores (average ± standard deviation), obtained at validation time, of each segmented structure (LV, LA, and MYO) for each trained model on the combined datasets: M_{Real}, $M_{17Real10Augmented}$, and $M_{17Real20Augmented}$. In Appendix – Table 4 the complete table with all the Dice scores for each trained fold of all three models can be found.

Fig. 6 shows the predicted segmentations given by these trained models, next to the ground truth segmentation provided by a cardiologist. The models were tested on a test set made of 3D echocardiography images from real subjects.

To compare the output segmentation from the DL models, the Dice scores and VS were calculated based on the predicted segmentations and the anatomical labels from a cardiologist and the results are in Table 3.

Table I.2: Average validation dice scores (average ± standard deviation) of each segmented structure (LV, LA, and MYO) for each trained model on combined datasets – M_{Real}, $M_{17Real10Augmented}$, and $M_{17Real20Augmented}$. The best scores are highlighted.

	Models		
	M_{Real}	$M_{17Real10Augmented}$	$M_{17Real20Augmented}$
LV	**0.938 ± 0.008**	0.928 ± 0.006	0.927 ± 0.007
LA	**0.862 ± 0.023**	0.830 ± 0.016	0.826 ± 0.017
MYO	0.724 ± 0.028	**0.767 ± 0.027**	0.763 ± 0.025

I.4 Discussion

In this work we built a pipeline to generate synthetic 3D labeled echocardiography images using a GAN model. These realistic-looking synthetic datasets were used to train 3D DL models to segment the LV, LA, and MYO.

Moreover, combined datasets including synthetic and real 3D images were created, with the VS metric supporting that generated 3D echocardiography

Table I.3: Average test set dice scores (average ± standard deviation) of each segmented structure (LV, LA, and MYO) and Volume Similarity of the segmented volume for the M_{Real}, $M_{17Real10Augmented}$, and $M_{17Real20Augmented}$ models. The best scores are highlighted.

	Models		
	M_{Real}	$M_{17Real10Augmented}$	$M_{17Real20Augmented}$
	Dice score		
LV	0.924 ± 0.019	**0.929 ± 0.020**	0.922 ± 0.017
LA	**0.876 ± 0.021**	0.874 ± 0.020	0.867 ± 0.022
MYO	0.666 ± 0.041	**0.708 ± 0.053**	0.680 ± 0.063
	Volume Similarity		
Heart Volume	0.831 ± 0.038	**0.844 ± 0.047**	0.836 ± 0.041

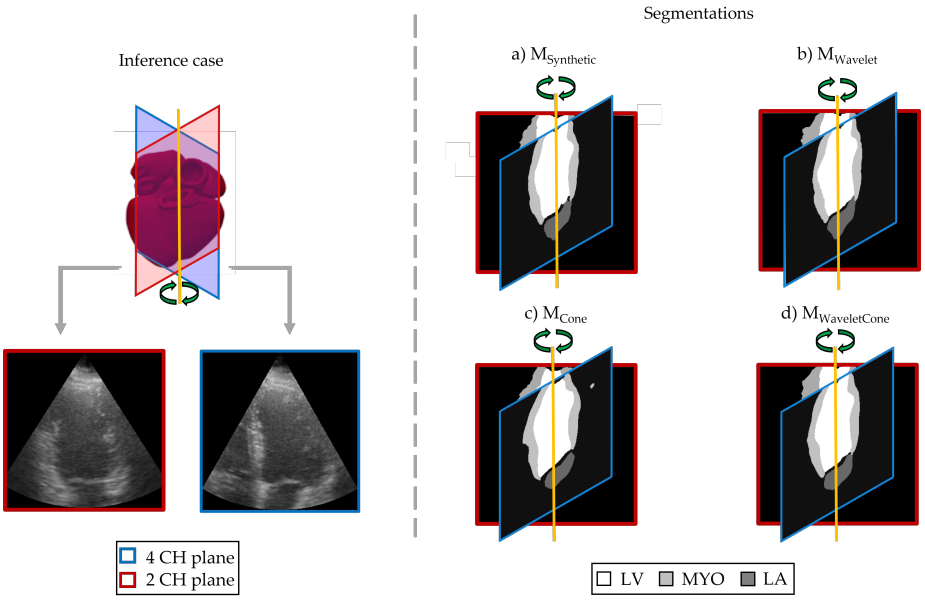

Figure I.5: Inference segmentation results from each trained model on synthetic datasets. On the left is shown a schematic representation of the heart and 2 cutting planes correspondent to a real 3D echocardiography image from the test set: the 4-chamber (CH), with blue frame, and the 2-CH, with red frame. On the right, the LV, LA, and MYO segmentation results provided by each of the 4 segmentation models: a) $M_{Synthetic}$, b) $M_{Wavelet}$, c) M_{Cone}, and d) $M_{WaveletCone}$ follow. A qualitative analysis of the segmentation results from each of the models, shows that the one where the training data was not post-processed, $M_{Synthetic}$, gives the best output due to a smoother segmentation of the relevant structures.

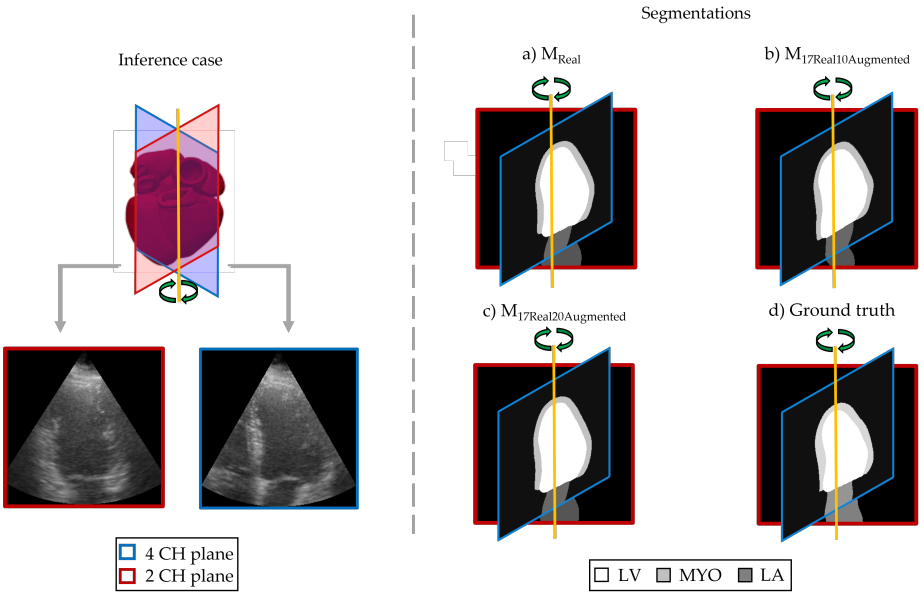

Figure I.6: Inference segmentation results from the trained models on augmented datasets with synthetic images. On the left is shown a schematic representation of the heart and 2 cutting planes correspondent to a real 3D echocardiography image from the test set: the 4-CH, with blue frame, and the 2-CH, with red frame. On the right, the LV, LA, and MYO segmentation results provided by the following 3 segmentation models: a) M_{Real}, b) $M_{17Real10Augmented}$, and c) $M_{17Real20Augmented}$, follow. To allow comparison and measure the Dice score and VS, d) shows the ground truth segmentation performed by a cardiologist. A qualitative analysis of the segmentation results from each of the models, shows that combining synthetic with real data improves the segmentation output due to a more accurate segmentation of the relevant structures.

images can be used to train 5 DL models, as data augmentation. Segmentation tasks were considered to exemplify the utility of the synthesized data, however the pipeline is generic and could be applied to generate other imaging data and train any DL tasks with anatomical labels as input, as further discussed in this section. A brief discussion on future applications and modifications of this approach is also presented.

I.4.1 3D Pix2pix GAN – Qualitative Analysis

The pipeline synthesizes 3D echocardiographic datasets with corresponding labels delineating different structures in the images.

After training the 3D Pix2pix GAN model, a qualitative analysis of the synthesized images indicated that the main structures of the heart were well

delineated in the generated images (Fig. 1, inference stage). Moreover, image details such as the cone, noise, and speckle patterns are also present and are continuous throughout each volume.

I.4.2 Post-processing and 3D Segmentation – Synthetic Datasets

To evaluate the utilization of synthetic images for research purposes and the extent to which the post-processing transformations affected the final results, four segmentation models were trained using four different datasets, as described earlier in Section III-C.

Despite the very small differences in the Dice scores shown in Table 1 and in Appendix – Table 5, the inference segmentations (Fig. 5) support the idea that the model trained on the dataset whose images were not post-processed, $M_{Synthetic}$, provided the best segmentation prediction.

The results regarding the influence of the post-processing step on the synthetically generated images supported the fact that applying a wavelet denoising transformation or cone reshaping, or even both transformations together, to these, in order to make the synthetic images look even more realistic, does not necessarily lead to better results when segmenting the LV, LA, and MYO (Fig. 5). This result shows some dependence on the DL task being performed. We segmented large volumes of the 3D image, comparing to its whole content. For this reason, the subtle differences in the voxels intensities that create the checkerboard artifacts do not seem to affect the prediction of the segmentation model.

To create the used synthetic datasets, CT acquired 3D anatomical models of the heart were used to extract the anatomical labels and create the input cases to the 3D GAN. The segmentation results and the echocardiography-looking aspect of the synthetic images pointed towards the generalization of this pipeline, as it can synthesize 3D echocardiography images, having as labels source different types of 3D models of the heart. The methodology to generate synthetic datasets can be generalized to other modalities, diseases, organs, as well as structures within the same organ (sub-regions of the heart, for example).

Shin *et al.* [5] and Shorten and Khoshgoftaar [38] showed that GANs can be widely used to perform data augmentation of medical image datasets. The work from these authors, together with the presented results, encourage the main contributions of this work stating that GANs can be used to generate synthetic images with labels, working as a data augmentation strategy, and tackling the concern of scarcity of 3D echocardiography labeled datasets, especially if there are underrepresented data samples within the available real datasets.

I.4.3 3D Segmentation – Combined Datasets

Further results on the usage of synthetic datasets were explored and presented in Section III-D. Here, three datasets made of GAN generated and real 3D images were used to train more segmentation models and further evaluate the influence of the presence of real data in these datasets, as demonstrated in [16].

Fig. 6 a), b), and c) showed the anatomical segmentations of the LV, LA, and MYO predicted by the best trained fold of each model – M_{Real}, $M_{17Real10Augmented}$, and $M_{17Real20Augmented}$. From the qualitative analysis, the segmentations delineate well the anatomical structures in consideration throughout the whole volume. At the same time, and similarly to what was discussed on Section IV-B, the average Dice scores presented in Table 2 led to the conclusion that having a dataset of real images combined with synthetic ones leads to more accurate final segmentations.

From the obtained results is also possible to assess the influence of using combined datasets with different percentages of synthetic data. Table 2 and Table 3 show that adding synthetic data to the initial real dataset improves the 3D segmentation of real 3D echocardiography images. They also show that adding larger amounts of synthetic data does not improve the results to a large extent.

Fig. 6 d) showed the ground truth inference case segmentation performed by a cardiologist. From these ground truth segmentations available for all the cases in the test set, the Dice scores and the VS in Table 3 were calculated.

Given the 3D nature of the task and due to the Dice metric limitations, the VS was additionally calculated and used as comparison metric. In particular, $M_{17Real10Augmented}$ showed to perform better at segmenting when the Dice score was considered as performance metric. On the other hand, $M_{17Real20Augmented}$ performed better in terms of VS metric. These results showed that the models trained on the combined datasets, i.e., with real and synthetic images, provided more accurate segmentation outputs (the 3D volume), relatively to the model trained with only real data, M_{Real}. The results support the previous work done by [16], confirming that including synthetic data on datasets made of real data improves and helps the final outcome of the DL models.

Additionally, this result reinforces that the proposed pipeline, relying on a 3D GAN model, can be used as a data augmentation tool. This framework arises as a solution to the lack of publicly available medical labeled datasets.

I.4.4 Further Applications

The presented pipeline has the potential to be further explored. As the demand for medical images is increasing, the proposed approach can be extended to synthesize images from other imaging modalities other than Ultrasound, such as MR or CT. It can also generate images where other organs are represented or even fetuses [13].

Another extension of this work would be to use different types of 3D models from which ground truth anatomical labels could be extracted. Besides anatomical models, animated or biophysical models represent other options that can be considered. The usage of anatomical models of pathological hearts are another possible extension, in order to generate pathological 3D echocardiography scans. Depending on the type of 3D model being considered, different annotations can be extracted, increasing the amount of clinically relevant tasks where these synthetic datasets can be used.

The generated 3D echocardiography images illustrated a heart volume during one time step of the whole cardiac cycle (end-diastole). It would be of great interest to generate 3D images of the heart during other cardiac cycle events and even to generate a beating volume throughout time, as high temporal resolution is one of the main strengths of Ultrasound imaging. On the other hand, a limitation to the Ultrasound images generation is that different scanning probe combinations lead to the acquisition of images with different quality levels. This large variability makes the GAN learning process more complex.

In this work we explored, to an extent, the effects that architectural changes of the GAN model have on the final synthesized images. We used different architectures for the GAN generator but more 3D CNNs exist and are showing up every day. These can be used to train the generative models, since DL strategies are becoming extremely common to use as medical image synthesis and analysis tools. Once the images were synthesized, we used wavelet denoising and an in-house developed algorithm to fix the Ultrasound cone edges. However, there are other denoising transformations and cone reshaping algorithms that can be experimented to post-process the images.

We trained several DL models to perform 3D segmentation to show that synthesized images can be used as input to train DL models. Nevertheless, the pipeline is generic and could be applied to other DL tasks that automatically assign anatomical labels to images, e.g., structure/feature recognition or automatic structural measurements. Furthermore, the GAN-generated labeled datasets are not only useful as input to train DL models but also could be used to train researchers and clinicians on image analysis.

Finally, during this pipeline development, computational memory constraints were present, mainly due to the large size of 3D volumes, complicating the process of developing a framework adapted to these. Future work will include study strategies to overcome these limitations.

I.5 Conclusion

An automatic data augmentation pipeline to create 3D echocardiography images and corresponding anatomical labels using a 3D GAN model was proposed. DL models are becoming widely used in clinical workflows and large volumes of medical data is a fundamental requirement to develop such algorithms with high accuracy. Generating synthetic data that could be used for the purpose of training DL models is of utmost importance since this generative model can become a widely used tool to address the existent lack of publicly available data and increasing challenges with moving data due to privacy regulations. Furthermore, the proposed methodology not only generates synthetic 3D echocardiography images but also associates labels to these synthetic images, eliminating the need for experienced professionals to do so, and without adding potential bias in the labels.

The proposed GAN model shows a generalization component since it can generate synthetic echocardiography images using 3D anatomical models of the

heart obtained for imaging modalities other than from Ultrasound.

The obtained results in this work indicate that synthetic datasets made up of GAN-generated 3D echocardiography images, and respective labels, are a good data augmentation resource to train and develop DL models that can be used to perform different medical tasks in the cardiac imaging domain, such as heart segmentation, where real patients' data is analyzed.

Appendix

See Table 4 and Table 5.

Table I.4: Validation dice scores of each segmented structure (LV, LA, and MYO) for each trained model on combined datasets -— M_{Real}, $M_{17Real10Augmented}$, and $M_{17Real20Augmented}$. The higher the score, the better the agreement between the model prediction and the ground truth segmentation. The best training fold of each model is highlighted.

	Models														
	M_{Real}					$M_{17Real10Augmented}$					$M_{17Real20Augmented}$				
	1	2	3	4	5	1	2	3	4	5	1	2	3	4	5
LV	0.933	0.932	**0.950**	0.930	0.943	0.924	0.929	0.919	**0.938**	0.930	0.917	0.928	**0.935**	0.934	0.923
LA	0.837	0.869	0.873	0.837	**0.896**	0.830	0.841	0.820	0.808	**0.853**	0.831	**0.841**	0.838	0.829	0.793
MYO	0.710	0.699	**0.766**	0.697	0.750	0.745	0.745	0.779	**0.815**	0.753	0.715	0.771	0.766	0.780	**0.785**

Table I.5: Validation dice scores of each segmented structure (LV, LA, and MYO) for each trained model on completely synthetic datasets – $M_{Synthetic}$, $M_{Wavelet}$, M_{Cone}, and $M_{WaveletCone}$. The higher the score, the better the agreement between the model prediction and the ground truth segmentation. The best training fold of each model is highlighted.

	Models																			
	$M_{Synthetic}$					$M_{Wavelet}$					M_{Cone}					$M_{WaveletCone}$				
	1	2	3	4	5	1	2	3	4	5	1	2	3	4	5	1	2	3	4	5
LV	0.924	0.930	0.924	0.918	**0.934**	0.927	0.930	0.923	0.919	**0.934**	0.926	0.930	0.921	0.918	**0.933**	0.928	0.928	0.914	0.914	**0.935**
LA	**0.833**	0.831	0.809	0.810	0.807	0.821	**0.831**	0.806	0.815	0.805	0.837	**0.842**	0.805	0.811	0.787	**0.834**	0.832	0.803	0.793	0.807
MYO	**0.824**	0.822	0.794	0.784	0.816	**0.829**	0.822	0.791	0.785	0.814	**0.824**	0.819	0.783	0.780	0.809	**0.828**	0.813	0.775	0.773	0.816

Acknowledgements. This work was supported by the European Union's Horizon 2020 Research and Innovation Programme through the Marie-Skłodowska Curie Grant under Agreement 860745.

References

[1] 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," *Circulat., Cardiovascular Imag.*, vol. 12, no. 9, Sep. 2019, Art. no. e009303, doi: 10.1161/CIRCIMAGING.119.009303.

[2] J. Scheetz, P. Rothschild, M. McGuinness, X. Hadoux, H. P. Soyer, M. Janda, J. J. J. Condon, L. Oakden-Rayner, L. J. Palmer, S. Keel, and P. van Wijngaarden, "A survey of clinicians on the use of artificial intelligence in

ophthalmology, dermatology, radiology and radiation oncology," *Sci. Rep.*, vol. 11, no. 1, Mar. 2021, Art. no. 1, doi: 10.1038/s41598-021-84698-5.

[3] A. Aljuaid and M. Anwar, "Survey of supervised learning for medical image processing," *Social Netw. Comput. Sci.*, vol. 3, no. 4, p. 292, May 2022, doi: 10.1007/s42979-022-01166-1.

[4] L. P. de Isla, D. V. Balcones, C. Fernández-Golfín, P. Marcos-Alberca, C. Almería, J. L. Rodrigo, C. Macaya, and J. Zamorano, "Three dimensional-wall motion tracking: A new and faster tool for myocardial strain assessment: Comparison with two-dimensional-wall motion tracking," *J. Amer. Soc. Echocardiogr.*, vol. 22, no. 4, pp. 325–330, Apr. 2009, doi: 10.1016/j.echo.2009.01.001.

[5] H.-C. Shin, N. A. Tenenholtz, J. K. Rogers, C. G. Schwarz, M. L. Senjem, J. L. Gunter, K. P. Andriole, and M. Michalski, "Medical image synthesis for data augmentation and anonymization using generative adversarial networks," in *Simulation and Synthesis in Medical Imaging*, Cham, Switzerland: Springer, 2018, pp. 1–11, *doi* : $10.1007/978 - 3 - 030 - 00536 - 8_1$.

[6] R. J. Chen, M. Y. Lu, T. Y. Chen, D. F. K. Williamson, and F. Mahmood, "Synthetic data in machine learning for medicine and healthcare," *Nature Biomed. Eng.*, vol. 5, no. 6, pp. 493–497, Jun. 2021, doi: 10.1038/s41551-021-00751-8.

[7] M. J. M. Chuquicusma, S. Hussein, J. Burt, and U. Bagci, "How to fool radiologists with generative adversarial networks? A visual Turing test for lung cancer diagnosis," in *Proc. IEEE 15th Int. Symp. Biomed. Imag. (ISBI)*, Apr. 2018, pp. 240–244, doi: 10.1109/ISBI.2018.8363564.

[8] A. Thorstensen, H. Dalen, B. H. Amundsen, S. A. Aase, and A. Stoylen, "Reproducibility in echocardiographic assessment of the left ventricular global and regional function, the HUNT study," *Eur. J. Echocardiogr.*, vol. 11, no. 2, pp. 149–156, Mar. 2010, doi: 10.1093/ejechocard/jep188.

[9] I. Banerjee, C. E. Catalano, G. Patané, and M. Spagnuolo, "Semantic annotation of 3D anatomical models to support diagnosis and follow-up analysis of musculoskeletal pathologies," *Int. J. Comput. Assist. Radiol. Surgery*, vol. 11, no. 5, pp. 707–720, May 2016, doi: 10.1007/s11548-015-1327-6.

[10] W. P. Segars, G. Sturgeon, S. Mendonca, J. Grimes, and B. M. W. Tsui, "4D XCAT phantom for multimodality imaging research," *Med. Phys.*, vol. 37, no. 9, pp. 4902–4915, Sep. 2010, doi: 10.1118/1.3480985.

[11] W. Kainz, E. Neufeld, W. E. Bolch, C. G. Graff, C. H. Kim, N. Kuster, B. Lloyd, T. Morrison, P. Segars, Y. S. Yeom, M. Zankl, X. G. Xu, and B. M. W. Tsui, "Advances in computational human phantoms and their applications in biomedical engineering—A topical review," *IEEE Trans. Radiat. Plasma Med. Sci.*, vol. 3, no. 1, pp. 1–23, Jan. 2019, doi: 10.1109/TRPMS.2018.2883437.

[12] A. Gilbert, M. Marciniak, C. Rodero, P. Lamata, E. Samset, and K. Mcleod, "Generating synthetic labeled data from existing anatomical models: An example with echocardiography segmentation," *IEEE Trans. Med. Imag.*, vol. 40, no. 10, pp. 2783–2794, Oct. 2021, doi: 10.1109/TMI.2021.3051806.

[13] C. W. Roy, D. Marini, W. P. Segars, M. Seed, and C. K. Macgowan, "Fetal XCMR: A numerical phantom for fetal cardiovascular magnetic resonance

imaging," *J. Cardiovascular Magn. Reson.*, vol. 21, no. 1, p. 29, May 2019, doi: 10.1186/s12968-019-0539-2.

[14] A. Perperidis, "Postprocessing approaches for the improvement of cardiac ultrasound B-mode images: A review," *IEEE Trans. Ultrason., Ferroelectr., Freq. Control*, vol. 63, no. 3, pp. 470–485, Mar. 2016, doi: 10.1109/TUFFC.2016.2526670.

[15] L. Chai, J.-Y. Zhu, E. Shechtman, P. Isola, and R. Zhang, "Ensembling with deep generative views," 2021, *arXiv:2104.14551*.

[16] D. R. P. R. M. Lustermans, S. Amirrajab, M. Veta, M. Breeuwer, and C. M. Scannell, "Optimized automated cardiac MR scar quantification with GAN-based data augmentation," 2021, *arXiv:2109.12940*.

[17] J. Pedrosa, S. Queirós, O. Bernard, J. Engvall, T. Edvardsen, E. Nagel, and J. D'hooge, "Fast and fully automatic left ventricular segmentation and tracking in echocardiography using shape-based b-spline explicit active surfaces," *IEEE Trans. Med. Imag.*, vol. 36, no. 11, pp. 2287–2296, Nov. 2017, doi: 10.1109/TMI.2017.2734959.

[18] H. Uzunova, J. Ehrhardt, and H. Handels, "Memory-efficient GAN-based domain translation of high resolution 3D medical images," *Computerized Med. Imag. Graph.*, vol. 86, Dec. 2020, Art. no. 101801, doi: 10.1016/j.compmedimag.2020.101801.

[19] J.-Y. Zhu, T. Park, P. Isola, and A. A. Efros, "Unpaired image-to image translation using cycle-consistent adversarial networks," 2017, *arXiv:1703.10593*.

[20] Y. Huo, Z. Xu, H. Moon, S. Bao, A. Assad, T. K. Moyo, M. R. Savona, R. G. Abramson, and B. A. Landman, "SynSeg-Net:Synthetic segmentation without target modality ground truth," *IEEE Trans. Med. Imag.*, vol. 38, no. 4, pp. 1016–1025, Apr. 2019, doi: 10.1109/TMI.2018.2876633.

[21] P. Isola, J.-Y. Zhu, T. Zhou, and A. A. Efros, "Image-to-image translation with conditional adversarial networks," 2016, *arXiv:1611.07004*.

[22] S. Amirrajab, S. Abbasi-Sureshjani, Y. A. Khalil, C. Lorenz, J. Weese, J. Pluim, and M. Breeuwer, "XCAT-GAN for synthesizing 3D consistent labeled cardiac MR images on anatomically variable XCAT phantoms," 2020, *arXiv:2007.13408*.

[23] Y. Hu, E. Gibson, L.-L. Lee, W. Xie, D. C. Barratt, T. Vercauteren, and J. A. Noble, "Freehand ultrasound image simulation with spatially conditioned generative adversarial networks," in *Molecular Imaging, Reconstruction and Analysis of Moving Body Organs, and Stroke Imaging and Treatment*, vol. 10555, M. J. Cardoso et al., Eds. Cham, Switzerland: Springer, 2017, pp. 105–115, *doi* : $10.1007/978 - 3 - 319 - 67564 - 0_1 1$.

[24] S. Abbasi-Sureshjani, S. Amirrajab, C. Lorenz, J. Weese, J. Pluim, and M. Breeuwer, "4D semantic cardiac magnetic resonance image synthesis on XCAT anatomical model," 2020, *arXiv:2002.07089*.

[25] T. Park, M.-Y. Liu, T.-C. Wang, and J.-Y. Zhu, "Semantic image synthesis with spatially-adaptive normalization," 2019, *arXiv:1903.07291*.

[26] M. D. Cirillo, D. Abramian, and A. Eklund, "Vox2Vox: 3D-GAN for brain tumour segmentation," 2020, *arXiv:2003.13653*.

[27] O. Ronneberger, P. Fischer, and T. Brox, "U-Net: Convolutional networks for biomedical image segmentation," 2015, *arXiv:1505.04597*.

[28] F. Isensee, P. F. Jaeger, S. A. A. Kohl, J. Petersen, and K. H. Maier-Hein, "nnU-Net: A self-configuring method for deep learning-based biomedical image segmentation," *Nature Methods*, vol. 18, no. 2, pp. 203–211, Feb. 2021, doi: 10.1038/s41592-020-01008-z.

[29] M. Alsharqi, W. J. Woodward, J. A. Mumith, D. C. Markham, R. Upton, and P. Leeson, "Artificial intelligence and echocardiography," *Echo Res. Pract.*, vol. 5, no. 4, pp. R115–R125, Dec. 2018, doi: 10.1530/ERP-18-0056.

[30] 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, Switzerland: Springer, 2018, pp. 309–316, *doi* : $10.1007/978 - 3 - 030 - 00889 - 5_3 5$.

[31] V7 Ltd. *V7—AI Data Platform for ML Teams.* Accessed: Feb. 25, 2022. [Online]. Available: https://www.v7labs.com/

[32] Ö. Çiçek, A. Abdulkadir, S. S. Lienkamp, T. Brox, and O. Ronneberger, "3D U-Net: Learning dense volumetric segmentation from sparse annotation," 2016, *arXiv:1606.06650*.

[33] A. Paszke, S. Gross, F. Massa, A. Lerer, J. Bradbury, G. Chanan, T. Killeen, Z. Lin, N. Gimelshein, L. Antiga, and A. Desmaison, "PyTorch: An imperative style, high-performance deep learning library," 2019, *arXiv:1912.017030*.

[34] A. Odena, V. Dumoulin, and C. Olah, "Deconvolution and checkerboard artifacts," *Distill*, vol. 1, no. 10, p. e3, Oct. 2016, doi: 10.23915/distill.00003.

[35] C. Rodero, M. Strocchi, M. Marciniak, S. Longobardi, J. Whitaker, M. D. O'Neill, K. Gillette, C. Augustin, G. Plank, E. J. Vigmond, P. Lamata, and S. A. Niederer, "Linking statistical shape models and simulated function in the healthy adult human heart," *PLOS Comput. Biol.*, vol. 17, no. 4, Apr. 2021, Art. no. e1008851, doi: 10.1371/journal.pcbi.1008851.

[36] A. K. Yadav, R. Roy, A. P. Kumar, C. S. Kumar, and S. K. Dhakad, "De-noising of ultrasound image using discrete wavelet transform by symlet wavelet and filters," in *Proc. Int. Conf. Adv. Comput., Commun. Informat. (ICACCI)*, Aug. 2015, pp. 1204–1208, doi: 10.1109/ICACCI.2015.7275776.

[37] A. A. Taha and A. Hanbury, "Metrics for evaluating 3D medical image segmentation: analysis, selection, and tool," *BMC Med. Imag.*, vol. 15, no. 1, p. 29, Aug. 2015, doi: 10.1186/s12880-015-0068-x.

[38] C. Shorten and T. M. Khoshgoftaar, "A survey on image data augmentation for deep learning," *J. Big Data*, vol. 6, no. 1, p. 60, Jul. 2019, doi: 10.1186/s40537-019-0197-0.

Authors' addresses

Cristiana Tiago GE Vingmed Ultrasound, GE Healthcare, 3183 Horten, Norway, and Department of Informatics, University of Oslo, 0373 Oslo, Norway, cristiana.tiago@ge.com

Andrew Gilbert, Svein Arne Aase, Sten Roar Snare, Jurica Sprem, and Kristin McLeod

GE Vingmed Ultrasound, GE Healthcare, 3183 Horten, Norway, andrew.gilbert@ge.com, sveinarne.aase@ge.com, stenroar.snare@ge.com, jurica.sprem@ge.com, kristin.mcleod@ge.com

Ahmed Salem Beela CARIM School for Cardiovascular Diseases, Maastricht University Medical Center, 6229 ER Maastricht, The Netherlands, and Faculty of Medicine, Suez Canal University, Ismailia, Egypt, a.salembeela@maastrichtuniversity.nl

Paper II

A Domain Translation Framework with an Adversarial Denoising Diffusion Model to Generate Synthetic Datasets of Echocardiography Images

Cristiana Tiago, Sten Roar Snare, Jurica Sprem, Kristin McLeod

Published in *IEEE Access*, 20 February 2023, volume 11, pp. 17594–17602.
DOI: 10.1109/ACCESS.2023.3246762.

II

Abstract

Currently, medical image domain translation operations show a high demand from researchers and clinicians. Amongst other capabilities, this task allows the generation of new medical images with sufficiently high image quality, making them clinically relevant. Deep Learning (DL) architectures, most specifically deep generative models, are widely used to generate and translate images from one domain to another. The proposed framework relies on an adversarial Denoising Diffusion Model (DDM) to synthesize echocardiography images and perform domain translation. Contrary to Generative Adversarial Networks (GANs), DDMs are able to generate high quality image samples with a large diversity. If a DDM is combined with a GAN, this ability to generate new data is completed at an even faster sampling time. In this work we trained an adversarial DDM combined with a GAN to learn the reverse denoising process, relying on a guide image, making sure relevant anatomical structures of each echocardiography image were kept and represented on the generated image samples. For several domain translation operations, the results verified that such generative model was able to synthesize high quality image samples: MSE: 11.50 $\pm$ 3.69, PSNR (dB): 30.48 $\pm$ 0.09, SSIM: 0.47 $\pm$ 0.03. The proposed method showed high generalization ability, introducing

This work was supported by the European Union's Horizon 2020 Research and Innovation Programme through the Marie-Skłodowska Curie Grant under Agreement 860745.
The code is publicly available in: https://github.com/CristianaTiago/2D-diffusion-echo

a framework to create echocardiography images suitable to be used for clinical research purposes.

Contents

II.1 Introduction

Echocardiography is the application of ultrasound imaging to the heart. This imaging modality is the most frequently used to image this organ, because it carries several advantages: there's a relative low cost and the equipment is portable, in comparison with Computed Tomography (CT) and Magnetic Resonance (MR). Ultrasound imaging also has the benefit of not using any ionizing radiation, this way not being harmful to the patient.

One other big advantage of echocardiography is its temporal resolution. When investigating cardiac motion, this modality still holds an advantage over others. Its wide usage in clinical practice and workflows make echocardiography a first port of call to detect pathological cases and assess the anatomy and function of the heart.

To optimize treatment pathways and spare clinicians' time to go over more severe cases, DL in healthcare has been proving its utility during the past years [1]. Besides these mentioned advantages, DL helps clinicians to reach the final diagnosis quicker without compromising the confidence level of it [2], reaching human-level performance [3].

In fact, DL has many and varied applications in the medical imaging domain. Image classification, anatomical structures segmentation and even detection of regions of interest are some of the most common usages of these mathematical methods. However, more recently other applications have been gaining terrain such as image generation [4] and image domain translation/adaptation [5], which help extend the usability in this domain, where there are increasing challenges in collecting sufficient and variable datasets.

DL algorithms learn functions and patterns from data, either from time series or images. Even though with echocardiography being such a widely used cardiac imaging modality, the access to medical image data became more complicated due to all the current anonymization and privacy regulations. Consequently, there is a current need for medical data specially to train DL algorithms.

Several studies, including [6] and [7] showed that synthetically generated images have a positive influence in the research and development of DL algorithms. Adding synthetic data to datasets made of real images adds variety to these

medical image datasets and presents a solution to data scarcity, a very real phenomenon existent in the medical DL field.

Generating synthetic data using deep generative models [8] provides a solution for this issue. Deep generative models are a subset of DL architectures trained to synthesize data. Within this group of neural networks, current methods include Variational Autoencoders and GANs. More recently, DDMs are also found under this category.

A GAN is a generative model based on a generator and a discriminator, where the former attempts to deceive the latter by minimizing the difference between the synthetically generated images and the real ones.

Diffusion models, similarly to all deep generative models, attempt to learn, by approximation, the probability distribution function representative of some training dataset. Particularly for these models, making them distinct from the rest, the generative procedure is based on the destruction of the input image by adding Gaussian noise to it, during a large enough number of steps, and consequently learn how to reverse these steps [9]. This way, it is possible to generate a synthetic image simply by denoising an initial randomly noisy input image. These models provide high fidelity/quality synthetic samples.

Creating a data augmentation tool to generate realistic echocardiography images is of need as it provides a solution to the scarcity of medical data.

II.1.1 State of the Art

Several image synthesis approaches are in practice today, with the choice of approach depending on the type of image being generated. When it comes to medical image synthesis, the choice of the imaging modality has a large impact on the selected models used to generate these images.

Most of the recent results and approaches adopt DL models to perform domain translation, with GANs being widely used since they can generate high quality samples, with a high level of realism [10], across several medical imaging modalities such as MRI [11] - [12], CT [13], and Ultrasound, namely echocardiography [14] - [15], with a fast sampling time. In a GAN, the generator tries to synthesize a sample that matches the target domain, which has an inherent data distribution function. The discriminator compares this synthesized image with the ones from the training dataset in order to distinguish them.

Echocardiography raises more challenges, when compared with other imaging modalities, due to the physics behind the acquisition and image reconstruction processes. Particularly [14] and [15] focused on generating 3D and 2D echocardiography, respectively. This type of medical image has inherent characteristics that strongly influence the final acquired image, namely the speckle pattern, the scanner functional characteristics, the patient's anatomy, and the sonographer's skills. Nevertheless, both works use GANs to synthesize the images, but the former considers a supervised GAN training and the latter an unsupervised approach.

However, GANs do not have a large diversity in the type of images they can generate, often leading the discriminator to converge too soon in training or

to mode collapse [16]. This phenomenon is very common when training GANs which drives the model to generate image samples with less quality and very little or even no variability at all.

DDMs, on the other hand, are capable of generating samples with a large variability without compromising its high quality [17]. These models were initially introduced by [18] in order to save time when sampling data from a training dataset, without having to learn a great number of training steps and parameters. These models destroy the input data distribution during a sufficiently large number of time steps and then use a neural network to learn how to reverse this process, restructuring the data.

In recent years, Ho et al. [9] and Song et al. [19] attempted to show an equivalence relationship between DDMs and score based generative models, which attribute a score to probability distributions based on the likelihood of data [20]. The work on training DDMs, based on original statistical physics theory, showed good results both in terms of synthetic image variability [21] and also of sample quality. Dhariwal and Nicol [17] demonstrated that DDMs are capable of outperforming GANs in terms of generated image quality. Furthermore, Nicol and Dhariwal [22] also showed that DDMs generate images with high likelihood values when such models are trained on datasets with a wide variety of images, what brings more complexity to the training dataset probability distribution.

To tackle the longer sampling time inherent to DDMs, both [22] and [23] presented contributions in terms of accelerating the forward diffusion process and adding noise to the input image over less steps. This way reducing the complexity of learning the reverse diffusion process and allowing to denoise image samples in a faster way, without compromising the image quality.

DDMs' application to medical image generation is yet not fully explored mainly due to its larger sampling time [9]. More recently, Xiao *et al.* [24] proposed to merge DDMs with GANs, in an attempt to make use of both generative models' strengths and tackle their individual weaknesses. This group proposed a denoising diffusion GAN, using a conditional GAN to model larger denoising steps during the reverse diffusion process.

Following the learning of conditional diffusion processes, Özbey *et al.* [25] proposed an adversarial diffusion model, SynDiff, where images from a source domain are used during training to guide the denoising diffusion process. This group applied such a model to perform medical image translation between brain MRI T1 and T2 weighted images. They were able to generate images of each domain, having an image from the other domain as a guide during the reverse diffusion process.

Taking the application of DDMs to extra dimensions, Kim and Ye [26] added a deformation module to the diffusion one and attempted to generate temporal volume images (3D + time) cardiac MRI images.

In this proposed work, we applied such deep generative models to generate echocardiography images. To keep a wide variety in the generated samples and decreasing the sampling time, without compromising the image quality, we propose a data augmentation tool based on a DDM and a GAN. The proposed adversarial diffusion model generates synthetic echocardiography images and

uses a GAN to learn the denoising process, whose performance is conditioned by anatomical masks of the heart. This way, these gray level masks guide the reverse diffusion process in order to maintain the anatomical information on the synthetic image. To the best of our knowledge, no previous work has presented reproducible results when it comes to generate such images using DDMs.

II.1.2 Summary of Contributions

We propose a data augmentation method to synthetically generate echocardiography images, using anatomical masks of the heart to guide the model during the image synthesis. These images are possible to use for research purposes in the medical image domain, such as the development of DL analysis tasks.

In summary and beyond the current state of the art, the main contributions of the proposed approach are:

1. The training of an adversarial diffusion model based on a DDM and a GAN, to generate synthetic echocardiography images.

2. The association of anatomical masks of the heart to the synthetically generated echocardiography image samples. This way, we tackle the lack of publicly available datasets, with labels, of echocardiography images.

3. The generation of echocardiography datasets belonging to different domains, such as from different scanners, using the proposed method to perform image domain translation.

II.2 Methodology

Fig. 1 illustrates the proposed approach. It is described in further detail in the following sections. Section II-A covers all the data collection and pre-processing steps, and Section II-B describes in further detail the working principle of the DDM and GAN behind the proposed adversarial diffusion model. Section II-C focuses on the creation of different image datasets and in Section II-D the image quality comparison metrics considered in this study are described.

II.2.1 Data Collection

The proposed adversarial diffusion model was trained on an already existing dataset of echocardiography images.

The CAMUS dataset, proposed by Leclerc *et al.* [27], includes 2D apical two and four chamber images, acquired at end-diastole (ED) and end-systole (ES) time instances of the cardiac cycle, with poor, good, and medium image quality levels. All images were acquired with a GE Vivid E95 ultrasound scanner. For our work, we selected only ED apical four chamber (4 CH) images with all levels of image quality, resulting in 450 images, resized to 256 x 256 pixels. All the images have associated anatomical masks for the Left Ventricle (LV),

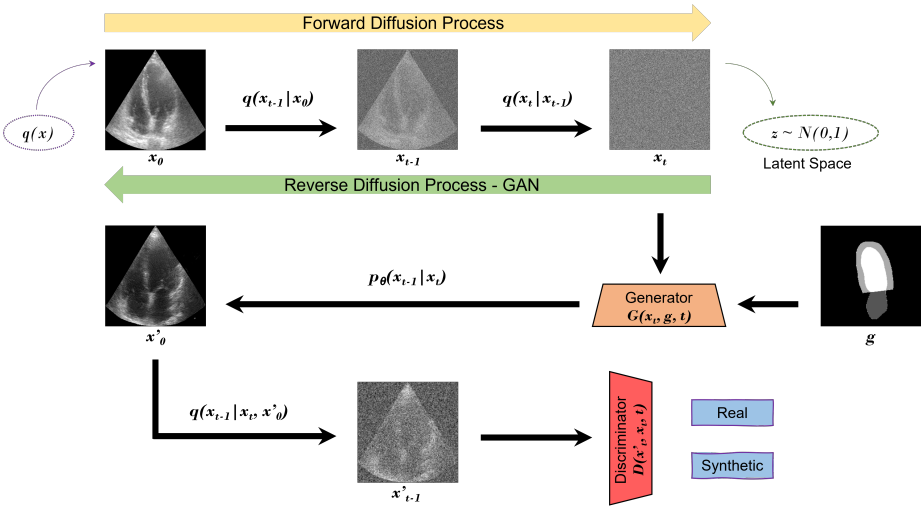

Figure II.1: Proposed pipeline to generate synthetic echocardiography images from a DDM and a GAN. Forward diffusion process: during this stage, the DDM module progressively adds Gaussian noise to the training image, x_0, belonging to the training dataset with a distribution $q(x)$, until a noisy image, x_t, is obtained after t time steps. This process creates a latent space, z, with a Gaussian distribution. The reverse diffusion process relies on a GAN to learn the reverse distribution, $p_\theta(x_t)$, and generate synthetic images, x'_0, in a conditional fashion.

Myocardium (MYO), and Left Atrium (LA). The dataset was split to train and validation sets by 90% and 10%, respectively.

Five other datasets were used for inference, to perform domain translation. All these were made of 256 x 256 apical 4CH images and included anatomical masks with the same structures considered in the CAMUS dataset. Table I summarizes all the considered datasets in this work.

First, the EchoNet-Dynamic dataset presented by Ouyang *et al.* [28], was used. This dataset contains more than ten thousand labeled echocardiogram videos. For the task of generating synthetic echocardiography images, only the ED frames of the echocardiographic videos were used. The anatomical masks associated with this dataset only showed the LV area. We then added the MYO and LA areas to them.

A second dataset was also made up of ED frames extracted from 3D (3 spatial dimensions) echocardiography images, acquired with different GE Vivid ultrasound scanners.

Two other datasets were also created using another two handheld GE ultrasound scanners: the Vscan Extend and the Vscan Air. The former is a pocket-sized scanner, and the latter is used to image the heart using a wireless probe and displaying the image on a smartphone. We created the anatomical labels for the second, third and fourth datasets.

A fifth dataset included ED frames extracted from 2D + time (2D + t) images, all of them acquired with GE Vivid ultrasound scanners, different from the GE Vivid E95. This one was previously labeled by a cardiologist.

Table II.1: Summary of all used datasets in this work.

Dataset	Origin	Acquisition scanner	Labels	Original image size	Final image size	Stage
CAMUS	Publicly available	GE Vivid E95	LA, LV, and MYO	Variable	256 x 256	DDM Training
Vscan	GE Healthcare	GE Vscan Extend	*	416 x 240		Inference (domain translation)
Vscan Air	GE Healthcare	GE Vscan Air	*	2040 x 1024		Inference (domain translation)
EchoNet	Publicly available	Multiple Philips and Siemens scanners	LV	112 x 112		Inference (domain translation)
2D + t	GE Healthcare	Multiple GE Vivid models (except E95)	*	1016 x 708		Inference (domain translation)
3D (spatial)	GE Healthcare	Multiple GE Vivid models (except E95)	*	Variable		Inference (domain translation)

* Labels for the LA, LV, and MYO were created for these datasets.

II.2.2 Adversarial Diffusion Model Training

The mathematical reasoning behind diffusion models was initially proposed by Sohl-Dickstein *et al.* [18]. This group showed that it is possible to reconstruct a noisy image in order to generate a sample belonging to a certain dataset with a defined probability distribution function. The denoising principle behind shows that these generative models offer higher quality and more variate image samples than others.

By themselves, DDMs are known to be based on unconditional diffusion processes applied during a large number of steps. However, the proposed adversarial diffusion model performs the reverse diffusion process in a conditional fashion. In order to synthesize image samples with similar statistical properties as the training dataset, the adversarial model uses images from a second domain to guide, i.e. condition, the reverse denoising algorithm. Furthermore, our adversarial DDM learns a faster reverse diffusion process which has a large step size instead of several small denoising instants. As represented on Fig. 1, the CAMUS dataset described on the previous section was used to train the proposed adversarial diffusion model. During the forward process of the training, an image x_0 is sampled from the training dataset with a probability distribution $q(x)$ and Gaussian noise is added to the image sample over T time steps. This process creates a Markov chain with a pre-defined variance β_t, defining the forward data distribution as:

$$q(x_t|x_{t-1}) = \mathcal{N}(x_t; \sqrt{1-\beta_t}x_{t-1}, \beta_t\mathbf{I}) \tag{II.1}$$

However, since the adversarial scenario allows the definition of a large step size, reducing the total number of denoising steps to be learned, the forward process can be re-written as:

$$q(x_t|x_{t-k}) = \mathcal{N}(x_t; \sqrt{1-\beta_t}x_{t-k}, \beta_t\mathbf{I}) \tag{II.2}$$

where k is the step size and $k >> 1$, as defined in [25].

On the other hand, the reverse denoising process is also a Markov chain approximated by a Gaussian distribution $p_\theta(x_{0:T})$, where θ are the predicted parameters of the reverse diffusion probability distribution, estimated by the GAN:

$$p_\theta(x_{t-k}|x_t) = \mathcal{N}(x_{t-k}; \mu_\theta(x_t, t), {\sigma_t}^2\mathbf{I}) \tag{II.3}$$

The training process of our adversarial DDM aims to minimize the difference between the conditional GAN predicted probability distribution p_θ, and the original training distribution $q(x)$:

$$\min_\theta \mathcal{L} = \min_\theta \sum_{t>=1} \mathbb{E}_{q(x_t)} \left[D(q(x_{t-k}|x_t)||p_\theta(x_{t-k}|x_t))\right] \tag{II.4}$$

where D represents the Kullback-Leibler divergence used in this loss function [9].

In the proposed architecture x'_0 is reconstructed by the generator of the GAN from the latent space z, where the feature information about the training data is encoded, and which follows a normal distribution.

Associated with the echocardiography images from the CAMUS dataset, there are anatomical masks which were used to guide the denoising process. This way, the GAN performance is conditioned when estimating the denoising distribution p_θ.

Given a source image y to guide the reverse diffusion process, the generator G attempts to estimate $p_\theta(x_{t-k}|x_t, y)$ by synthesizing x'_{t-k} such that $x'_{t-k} \sim p_\theta(x_{t-k}|x_t, y)$. The discriminator $D(x'_{t-k}, x_t, t)$ distinguishes between samples from either the real probability distribution, $q(x)$, or the predicted $p_\theta(x)$.

II.2.3 Domain Translation - Inference

Domain translation allows to transform images from a domain A to a domain B, so that the generated, i.e. domain-translated, images have similar characteristics to the ones belonging to the initial domain [29]. This operation learns how to do such translation by analyzing the probability distribution of the initial dataset and iteratively compare it with the statistical distribution of the target domain [30].

After training the adversarial diffusion model using the CAMUS dataset, at inference time, the datasets described on Section II-A were considered as input to the trained model. These inference steps allowed to perform domain translation and create synthetic datasets with characteristics similar to CAMUS.

II.2.4 Image Quality Comparison Metrics

To evaluate the quality of the generated image samples from all different synthesized datasets described before, several image quality metrics were calculated and compared.

The most commonly used image quality estimator is the Mean Squared Error (MSE), which quantifies the difference between two different images, measuring the differences pixel by pixel. If the synthetic image is similar to the ground truth one, then this error will be low.

The Peak Signal-to-Noise (PSNR) ratio takes into account the signal from the original image and the noise, i.e. error, of the generated sample. This metric is presented in dB and [31] considers values around and above 30 dB as representing good quality synthetic image samples.

Both these metrics are pixel based. To evaluate the quality of generated images using a method more similar to the human visual system, the Structural Similarity Method (SSIM) [32] was considered. SSIM takes into account the preserved and changed edges information between the original image and the generated one, and also the texture differences. This index takes values between 0 and 1, with higher values reflecting a larger image similarity.

Specifically created to measure the performance of GANs, the Fréchet Inception Distance (FID) was defined by Heusel *et al.* [33] to evaluate the quality of the generated samples from different datasets. Contrary to the already described metrics, the FID score does not directly compare generated and real images, but it measures the distance between the statistical distribution of synthetic and real datasets [34]. The lower this score is, the smaller the difference between the datasets.

II.3 Results

The training parameters and training time of the proposed adversarial diffusion models are described in Section III-A, and Section III-B details the results of the domain translation operation, together with the image quality comparison metrics obtained.

II.3.1 Adversarial Diffusion Model Training

The proposed adversarial DDM was trained during 500 epochs and for a total of four diffusion steps. The upper and lower bounds for the variance of the predicted distribution were kept the same as in [25]. The model was built using PyTorch [35] and it was trained on a computer equipped with four NVIDIA GeForce RTX 2080 GPUs (multiple GPU training). Training took approximately forty hours.

Fig. 2 gives an overview of the training results during the validation steps. It shows a generated sample with similar characteristics to the images in the training dataset, and keeping the anatomical information present in the guide image.

II.3.2 Domain Translation - Comparison Metrics

After training the adversarial diffusion model, it was used to perform domain translation operations. For each of the five previously created datasets with

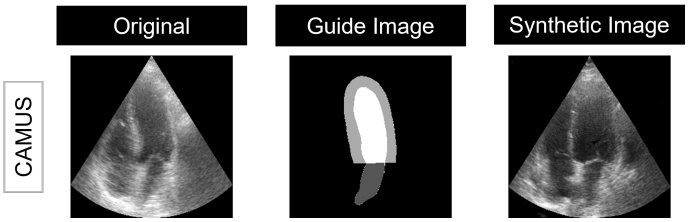

Figure II.2: Adversarial diffusion model training results. For the validation image shown on the left, the image on the right is the generated sample, outputted by the trained model.

different image characteristics, a synthetic dataset with properties similar to CAMUS was generated.

Fig. 3 shows the best generated image sample from each of the domain translation operations performed.

The FID score, in Table II, gives the overview of the complete dataset quality, instead of comparing individual image samples. Fig. 4 shows examples of the worst, median, and best generated images, in terms of the PSNR value, after the domain translation operation.

Table III lists the image comparison metrics calculated between the generated sample and the ground truth image, for each test image belonging to the inference datasets.

Table II.2: FID scores for each original inference dataset (before domain translation) and each synthetic dataset (after domain translation), compared with the training CAMUS dataset.The best scores are highlighted.

	Inference Datasets (before domain translation)					Synthetic Datasets (after domain translation)				
	Vscan	Vscan Air	EchoNet	2D + t	3D (spatial)	Vscan	Vscan Air	EchoNet	2D + t	3D (spatial)
FID	279.53	332.73	260.42	189.55	**61.08**	70.18	81.22	60.17	79.28	**50.87**

Table II.3: Comparison metrics (average ± standard deviation) for the domain translation operations. MSE, PSNR (dB), and SSIM were calculated for all the images in the 5 inference datasets. The best scores are highlighted.

	Metrics		
	MSE	PSNR (dB)	SSIM
Vscan	18.27 ± 9.26	30.09 ± 0.12	0.37 ± 0.01
Vscan Air	30.60 ± 7.79	28.94 ± 0.30	0.18 ± 0.03
EchoNet	22.39 ± 5.41	29.65 ± 0.20	0.31 ± 0.02
2D + t	11.95 ± 7.21	30.48 ± 0.13	0.40 ± 0.01
3D (spatial)	**11.50 ± 3.69**	**30.48 ± 0.09**	**0.47 ± 0.03**

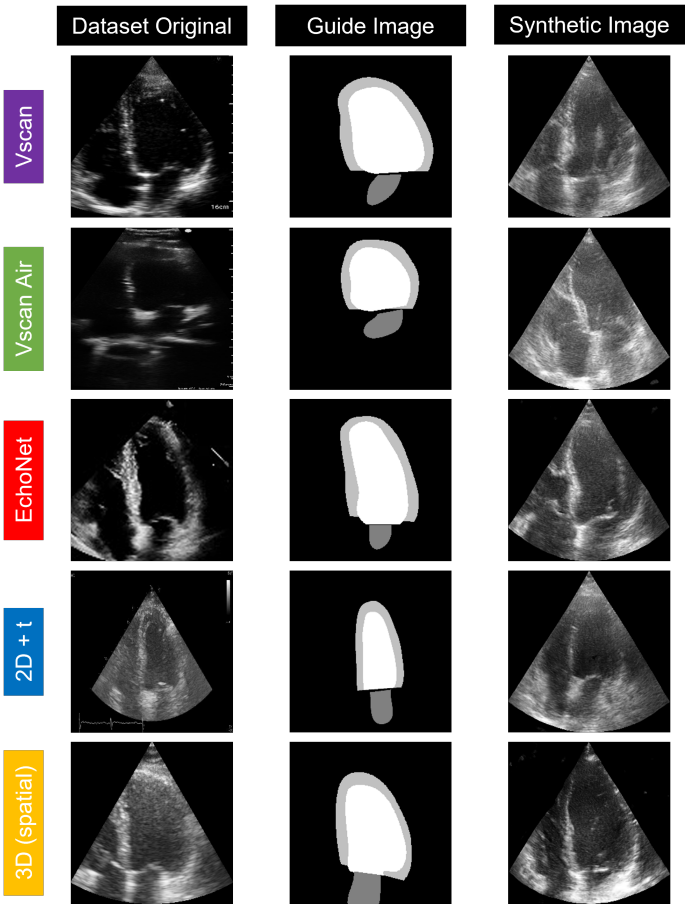

Figure II.3: Domain translation results. Best generated image from each of the inference datasets. All the synthetic images show characteristics of the CAMUS dataset and keep the anatomical information present in the guide image (white area – LV, dark gray area – LA, light gray area – MYO).

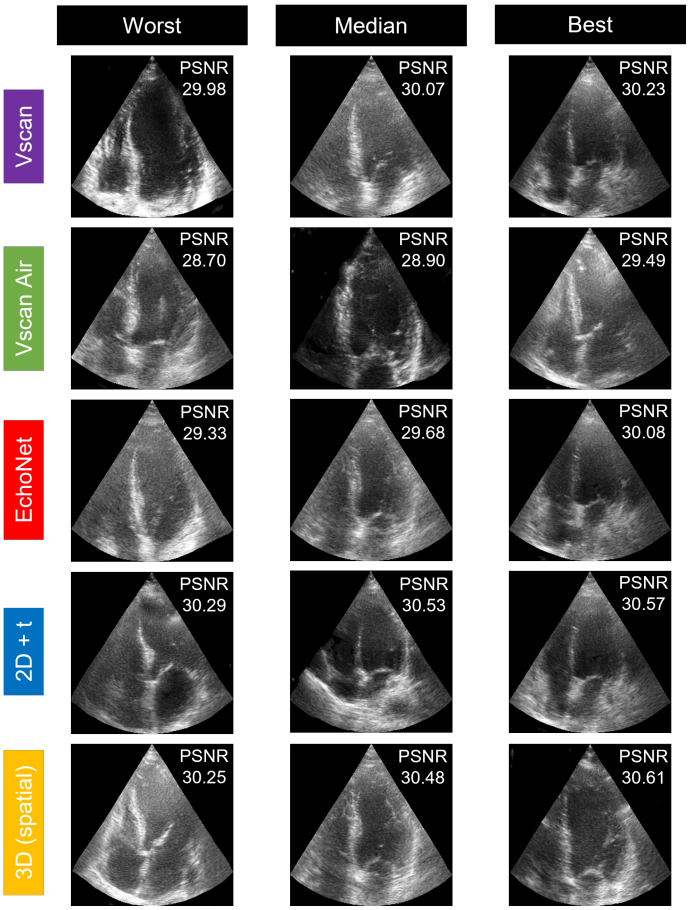

Figure II.4: Worst, median, and best synthesized image from each of the inference datasets, in terms of PSNR. The worst images are not totally discrepant from the best ones.

II.4 Discussion

The proposed adversarial diffusion model architecture, based on a DDM and a GAN, proved to be able to produce a wide variety of generated image samples with a fast sampling time. In fact, training such a complex model took less than two days. This result was expected since diffusion models were conceived to learn less training parameters, in comparison to other deep generative models such as GANs, making the training lighter and faster without compromising the final output quality.

None of the generated image samples required post-processing operations, for example to fix the cone shape [14] or remove unwanted noise, in opposition to what has been reported when generating images with other deep generative models, where these operations are often required. DDMs hold this advantage of generating more visually accurate image samples without requiring additional post-processing steps [17]. On the contrary to what happens with GANs, the image samples generated via the adversarial DDM show no artifacts.

After collecting data from five different echocardiography datasets with different image characteristics amongst them, the trained model was then used to perform different tasks of image domain translation.

In terms of image acquisition, the echocardiography scans acquired with the Vscan Air (Fig. 3) are substantially different from the images one would get if the GE Vivid E95 would be used, due to the nature of the ultrasound probe used by the former scanner. This dataset characteristic is supported by the results on Table II, where the FID score, for the Vscan Air dataset is the highest amongst all the inference datasets, when compared to CAMUS, reflecting this difference.

From Table II it is also visible that the inference dataset containing 2D apical 4CH echocardiography images extracted from 3D scans (where the 3 spatial dimensions were considered), 3D (spatial), is the most similar dataset to the CAMUS dataset, amongst all the five inference datasets, as it holds the lowest FID score. Consequent manual inspection confirmed that these datasets are visually the most similar.

Five domain translation operations were performed and shown in this work. During each of these, the trained adversarial diffusion model generated an image sample corresponding to each image in the considered dataset. The generated images were then compared to the ground truth and the MSE, PSNR, and SSIM were calculated (Table III).

The 3D (spatial) dataset showed the best results for all these three metrics. The high value of the PSNR indicates that the information present in the original inference images is preserved and visible on the synthetic images generated with the adversarial diffusion model.

The SSIM value for the Vscan Air dataset holds the lowest value reinforcing the conclusion described earlier, stating that this dataset images belong to a domain which is the most different from the CAMUS images domain. On the other hand, a PSNR close to 30 dB reflects that the domain translation operation was still able to synthesize images with meaningful information encoded on them.

After the domain translation operations, the FID score was calculated for each of the synthetic datasets (Table II). The 3D (spatial) synthetic dataset is still the one registering the lowest FID value amongst all the synthetic datasets. The difference between the FID scores obtained before and after the domain translation operations are indicative of the generalization ability of the proposed adversarial diffusion model. Table II shows a significant decrease in all datasets FID scores, after domain translation. The scores represent a smaller difference between the probability distribution of each synthetic dataset and the CAMUS. The EchoNet synthetic dataset has a smaller FID score than the 2D + t, even though, before domain translation, the opposite scenario, i.e. smaller FID for the 2D + t dataset, was verified.

The adversarial DDM trained was able to generate variate samples, closely depicting the LA, LV, and MYO, present on apical 4CH echocardiography images (Fig. 4). The images considered as worst, in terms of PSNR, still illustrate these structures and are not completely divergent from the best ones.

Presented results described and discussed in this section support the initial premises; namely that diffusion models are lighter and quicker to train and are able to generate high quality image samples. Creating an adversarial diffusion model, by using a GAN to learn the reverse diffusion process, brings the advantage of generating images with a small sampling time. The developed approach can be used to generate synthetic datasets of echocardiography image samples and also improve the quality of lower-resolution ones. This way, the adversarial DDM is a resource to generate images belonging to different image domains, helping in the development of DL models that perform equally well irrespective of the imaging scanner/vendor.

To the best of our knowledge, diffusion models were not yet used to generate clinically relevant echocardiography images, nor used to perform domain translation operations between substantially different medical image datasets. Our work demonstrated that such tasks are possible and the generated echocardiography images have high quality and include meaningful anatomical information, since anatomical masks were used to guide the reverse diffusion process.

In the future, the influence of the type of guide image used during the adversarial learning process will be further explored. Also, the analysis of the synthetic images in the clinical scenario will be assessed, by working closely with clinical end-users.

II.5 Conclusion

A domain translation framework based on an adversarial diffusion model was proposed, in order to generate synthetic datasets of echocardiography images. In the medical scenario, DL approaches outperform other methods for some tasks, including medical image generation and domain translation operations. These DL methods, however, require a large amount of data during their training and development.

The proposed framework relies on the usage of state of the art models and methods to both generate echocardiography images and also perform domain translation. These tasks allow to create a large amount of variate medical image data with clinical relevance which can be used for research and learning methodologies.

Furthermore, the proposed model showed a great generalization capacity, being able to synthesize echocardiography images with a large variability.

Acknowledgements. This work was supported by the European Union's Horizon 2020 Research and Innovation Programme through the Marie-Skłodowska Curie Grant under Agreement 860745.

References

[1] V. C. Gandhi and P. P. Gandhi, "A survey–insights of ML and DL in health domain," in *Proc. Int. Conf. Sustain. Comput. Data Commun. Syst. (ICSCDS)*, Apr. 2022, pp. 239–246, doi: 10.1109/ICSCDS53736.2022.9760981.

[2] A. Aljuaid and M. Anwar, "Survey of supervised learning for medical image processing," *Social Netw. Comput. Sci.*, vol. 3, no. 4, p. 292, May 2022, doi: 10.1007/s42979-022-01166-1.

[3] J. Scheetz, P. Rothschild, M. McGuinness, X. Hadoux, H. P. Soyer, M. Janda, J. J. J. Condon, L. Oakden-Rayner, L. J. Palmer, S. Keel, and P. van Wijngaarden, "A survey of clinicians on the use of artificial intelligence in ophthalmology, dermatology, radiology and radiation oncology," *Sci. Rep.*, vol. 11, no. 1, pp. 1–10, Mar. 2021, doi: 10.1038/s41598-021-84698-5.

[4] A. D. Schütte, J. Hetzel, S. Gatidis, T. Hepp, B. Dietz, S. Bauer, and P. Schwab, "Overcoming barriers to data sharing with medical image generation: A comprehensive evaluation," *NPJ Digit. Med.*, vol. 4, no. 1, Sep. 2021, doi: 10.1038/s41746-021-00507-3.

[5] *Domain Adaptation and Representation Transfer and Medical Image Learning With Less Labels and Imperfect Data.* Switzerland: Springer. Accessed: Nov. 4, 2022.

[6] A. Thorstensen, H. Dalen, B. H. Amundsen, S. A. Aase, and A. Stoylen, "Reproducibility in echocardiographic assessment of the left ventricular global and regional function, the HUNT study," *Eur. J. Echocardiogr.*, vol. 11, no. 2, pp. 149–156, Mar. 2010, doi: 10.1093/ejechocard/jep188.

[7] H. Uzunova, J. Ehrhardt, and H. Handels, "Memory-efficient GANbased domain translation of high resolution 3D medical images," *Computerized Med. Imag. Graph.*, vol. 86, Dec. 2020, Art. no. 101801, doi: 10.1016/j.compmedimag.2020.101801.

[8] L. Ruthotto and E. Haber, "An introduction to deep generative modeling," 2021, *arXiv:2103.05180.*

[9] J. Ho, A. Jain, and P. Abbeel, "Denoising diffusion probabilistic models," in *Proc. 34th Int. Conf. Neural Inf. Process. Syst.*, Red Hook, NY, USA, Dec. 2020, pp. 6840–6851.

[10] S. U. H. Dar, M. Yurt, L. Karacan, A. Erdem, E. Erdem, and T. Çukur, "Image synthesis in multi-contrast MRI with conditional generative adversarial

networks," *IEEE Trans. Med. Imag.*, vol. 38, no. 10, pp. 2375–2388, Oct. 2019, doi: 10.1109/TMI.2019.2901750.

[11] H. Li et al., "DiamondGAN: Unified multi-modal generative adversarial networks for MRI sequences synthesis," in *Medical Image Computing and Computer Assisted Intervention.* Cham, Switzerland: Springer, 2019, pp. 795–803, *doi* : $10.1007/978 - 3 - 030 - 32251 - 9_8 7$.

[12] S. Abbasi-Sureshjani, S. Amirrajab, C. Lorenz, J. Weese, J. Pluim, and M. Breeuwer, "4D semantic cardiac magnetic resonance image synthesis on XCAT anatomical model," 2020, *arXiv:2002.07089.*

[13] M. Selim, J. Zhang, B. Fei, G.-Q. Zhang, and J. Chen, "STAN-CT: Standardizing CT image using generative adversarial networks," in *Proc. AMIA Annu. Symp.*, Jan. 2021, pp. 1100–1109.

[14] C. Tiago, A. Gilbert, A. S. Beela, S. A. Aase, S. R. Snare, J. Sprem, and K. McLeod, "A data augmentation pipeline to generate synthetic labeled datasets of 3D echocardiography images using a GAN," *IEEE Access*, vol. 10, pp. 98803–98815, 2022, doi: 10.1109/ACCESS.2022.3207177.

[15] A. Gilbert, M. Marciniak, C. Rodero, P. Lamata, E. Samset, and K. Mcleod, "Generating synthetic labeled data from existing anatomical models: An example with echocardiography segmentation," *IEEE Trans. Med. Imag.*, vol. 40, no. 10, pp. 2783–2794, Oct. 2021, doi: 10.1109/TMI.2021.3051806.

[16] P. Isola, J.-Y. Zhu, T. Zhou, and A. A. Efros, "Image-to-image translation with conditional adversarial networks," 2016, *arXiv:1611.07004.*

[17] P. Dhariwal and A. Nichol, "Diffusion models beat GANs on image synthesis," 2021, *arXiv:2105.05233.*

[18] J. Sohl-Dickstein, E. A. Weiss, N. Maheswaranathan, and S. Ganguli, "Deep unsupervised learning using nonequilibrium thermodynamics," 2015, *arXiv:1503.03585.*

[19] Y. Song, J. Sohl-Dickstein, D. P. Kingma, A. Kumar, S. Ermon, and B. Poole, "Score-based generative modeling through stochastic differential equations," 2020, *arXiv:2011.13456.*

[20] Y. Song and S. Ermon, "Improved techniques for training score-based generative models," in *Proc. Adv. Neural Inf. Process. Syst.*, vol. 33, 2020, pp. 12438–12448. [Online]. Available: https://proceedings.neurips.cc/paper/2020/hash/92c3b916311a5517d9290576e3ea37ad-Abstract.html

[21] Y. Song, C. Durkan, I. Murray, and S. Ermon, "Maximum likelihood training of score-based diffusion models," 2021, *arXiv:2101.09258.*

[22] A. Nichol and P. Dhariwal, "Improved denoising diffusion probabilistic models," 2021, *arXiv:2102.09672.*

[23] J. Song, C. Meng, and S. Ermon, "Denoising diffusion implicit models," 2020, *arXiv:2010.02502.*

[24] Z. Xiao, K. Kreis, and A. Vahdat, "Tackling the generative learning trilemma with denoising diffusion GANs," 2021, *arXiv:2112.07804.*

[25] M. Özbey, O. Dalmaz, S. U. H. Dar, H. A Bedel, Ş. Özturk, A. Güngör, and T. Çukur, "Unsupervised medical image translation with adversarial diffusion models," 2022, *arXiv:2207.08208.*

[26] B. Kim and J. C. Ye, "Diffusion deformable model for 4D temporal medical image generation," in *Medical Image Computing and Computer Assisted Intervention.* Cham, Switzerland: Springer, 2022, pp. 539–548, *doi* : $10.1007/978 - 3 - 031 - 16431 - 6_51$.

[27] S. Leclerc, E. Smistad, J. Pedrosa, A. Østvik, F. Cervenansky, F. Espinosa, T. Espeland, and E. A. R. Berg, "Deep learning for segmentation using an open large-scale dataset in 2D echocardiography," *IEEE Trans. Med. Imag.*, vol. 38, no. 9, pp. 2198–2210, Sep. 2019, doi: 10.1109/TMI.2019.2900516.

[28] D. Ouyang, B. He, A. Ghorbani, N. Yuan, J. Ebinger, C. P. Langlotz, P. A. Heidenreich, R. A. Harrington, D. H. Liang, E. A. Ashley, and J. Y. Zou, "Video-based AI for beat-to-beat assessment of cardiac function," *Nature*, vol. 580, no. 7802, pp. 252–256, Apr. 2020, doi: 10.1038/s41586-020-2145-8.

[29] Z. Murez, S. Kolouri, D. Kriegman, R. Ramamoorthi, and K. Kim, "Image to image translation for domain adaptation," 2017, *arXiv:1712.00479.*

[30] Y. Zhu, Y. Tang, Y. Tang, D. C. Elton, S. Lee, P. J. Pickhardt, and R. M. Summers, "Cross-domain medical image translation by shared latent Gaussian mixture model," 2020, *arXiv:2007.07230.*

[31] O. S. Faragallah, H. El-Hoseny, W. El-Shafai, W. A. El-Rahman, and H. S. El-Sayed, "A comprehensive survey analysis for present solutions of medical image fusion and future directions," *IEEE Access*, vol. 9, pp. 11358–11371, 2021, doi: 10.1109/ACCESS.2020.3048315.

[32] G. P. Renieblas, A. T. Nogués, A. M. González, N. Gómez-Leon, and E. G. del Castillo, "Structural similarity index family for image quality assessment in radiological images," *J. Med. Imag.*, vol. 4, no. 3, Jul. 2017, Art. no. 035501, doi: 10.1117/1.JMI.4.3.035501.

[33] M. Heusel, H. Ramsauer, T. Unterthiner, B. Nessler, and S. Hochreiter, "GANs trained by a two time-scale update rule converge to a local Nash equilibrium," 2017, *arXiv:1706.08500.*

[34] Y. Skandarani, P.-M. Jodoin, and A. Lalande, "GANs for medical image synthesis: An empirical study," 2021, *arXiv:2105.05318.*

[35] A. Paszke et al., "PyTorch: An imperative style, high-performance deep learning library," 2019, *arXiv:1912.01703.*

Authors' addresses

Cristiana Tiago GE Vingmed Ultrasound, GE Healthcare, 3183 Horten, Norway, and Department of Informatics, University of Oslo, 0373 Oslo, Norway, cristiana.tiago@ge.com

Sten Roar Snare, Jurica Sprem, and Kristin McLeod GE Vingmed Ultrasound, GE Healthcare, 3183 Horten, Norway, stenroar.snare@ge.com, jurica.sprem@ge.com, kristin.mcleod@ge.com